QUANTITATIVE APPROACHES TO PLANT BREEDING: CONCEPTS, STRATEGIES AND PRACTICAL APPLICATIONS

EDITED BY: Suchismita Mondal, Rodomiro Ortiz and Leonardo Abdiel Crespo Herrera

PUBLISHED IN: Frontiers in Plant Science

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88976-878-3
DOI 10.3389/978-2-88976-878-3

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

QUANTITATIVE APPROACHES TO PLANT BREEDING: CONCEPTS, STRATEGIES AND PRACTICAL APPLICATIONS

Topic Editors:
Suchismita Mondal, Montana State University, United States
Rodomiro Ortiz, Swedish University of Agricultural Sciences, Sweden
Leonardo Abdiel Crespo Herrera, International Maize and Wheat Improvement Center (Mexico), Mexico

Citation: Mondal, S., Ortiz, R., Herrera, L. A. C., eds. (2022). Quantitative Approaches to Plant Breeding: Concepts, Strategies and Practical Applications. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-878-3

Table of Contents

frontiers in Plant Science

ORIGINAL RESEARCH
published: 02 July 2021
doi: 10.3389/fpls.2021.692205

Genetic Dissection of Quantitative Resistance to Common Rust (*Puccinia sorghi*) in Tropical Maize (*Zea mays* L.) by Combined Genome-Wide Association Study, Linkage Mapping, and Genomic Prediction

Jiaojiao Ren[1,2†], Zhimin Li[2,3†], Penghao Wu[1], Ao Zhang[4], Yubo Liu[5], Guanghui Hu[6], Shiliang Cao[6], Jingtao Qu[7], Thanda Dhliwayo[2], Hongjian Zheng[5], Michael Olsen[8], Boddupalli M. Prasanna[8], Felix San Vicente[2] and Xuecai Zhang[2]**

[1] College of Agronomy, Xinjiang Agricultural University, Urumqi, China, [2] International Maize and Wheat Improvement Center (CIMMYT), Texcoco, Mexico, [3] College of Agronomy, Henan Agricultural University, Zhengzhou, China, [4] College of Bioscience and Biotechnology, Shenyang Agricultural University, Shenyang, China, [5] CIMMYT-China Specialty Maize Research Center, Crop Breeding and Cultivation Research Institute, Shanghai Academy of Agricultural Sciences, Shanghai, China, [6] Maize Research Institute, Heilongjiang Academy of Agricultural Sciences, Harbin, China, [7] Maize Research Institute, Sichuan Agricultural University, Chengdu, China, [8] International Maize and Wheat Improvement Center (CIMMYT), Nairobi, Kenya

Common rust is one of the major foliar diseases in maize, leading to significant grain yield losses and poor grain quality. To dissect the genetic architecture of common rust resistance, a genome-wide association study (GWAS) panel and a bi-parental doubled haploid (DH) population, DH1, were used to perform GWAS and linkage mapping analyses. The GWAS results revealed six single-nucleotide polymorphisms (SNPs) significantly associated with quantitative resistance of common rust at a very stringent threshold of *P*-value 3.70×10^{-6} at bins 1.05, 1.10, 3.04, 3.05, 4.08, and 10.04. Linkage mapping identified five quantitative trait loci (QTL) at bins 1.03, 2.06, 4.08, 7.03, and 9.00. The phenotypic variation explained (PVE) value of each QTL ranged from 5.40 to 12.45%, accounting for the total PVE value of 40.67%. Joint GWAS and linkage mapping analyses identified a stable genomic region located at bin 4.08. Five significant SNPs were only identified by GWAS, and four QTL were only detected by linkage mapping. The significantly associated SNP of S10_95231291 detected in the GWAS analysis was first reported. The linkage mapping analysis detected two new QTL on chromosomes 7 and 10. The major QTL on chromosome 7 in the region between 144,567,253 and 149,717,562 bp had the largest PVE value of 12.45%. Four candidate genes of *GRMZM2G328500*, *GRMZM2G162250*, *GRMZM2G114893*, and *GRMZM2G138949* were identified, which played important roles in the response of stress resilience and the regulation of plant growth and development. Genomic prediction (GP) accuracies observed in the GWAS panel and DH1 population were 0.61

OPEN ACCESS

Edited by:
Rodomiro Ortiz, Swedish University of Agricultural Sciences, Sweden

Reviewed by:
Claudia Teixeira Guimaraes, Brazilian Agricultural Research Corporation (EMBRAPA), Brazil
Darlene Lonjas Sanchez, Texas A&M AgriLife Research and Extension Center at Beaumont, United States

***Correspondence:**
Felix San Vicente
F.SanVicente@cgiar.org
Xuecai Zhang
xc.zhang@cgiar.org

† These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *07 April 2021*
Accepted: *08 June 2021*
Published: *02 July 2021*

Citation:
Ren J, Li Z, Wu P, Zhang A, Liu Y, Hu G, Cao S, Qu J, Dhliwayo T, Zheng H, Olsen M, Prasanna BM, San Vicente F and Zhang X (2021) Genetic Dissection of Quantitative Resistance to Common Rust (Puccinia sorghi) in Tropical Maize (Zea mays L.) by Combined Genome-Wide Association Study, Linkage Mapping, and Genomic Prediction. Front. Plant Sci. 12:692205. doi: 10.3389/fpls.2021.692205

and 0.51, respectively. This study provided new insight into the genetic architecture of quantitative resistance of common rust. In tropical maize, common rust could be improved by pyramiding the new sources of quantitative resistance through marker-assisted selection (MAS) or genomic selection (GS), rather than the implementation of MAS for the single dominant race-specific resistance gene.

Keywords: maize, common rust, quantitative resistance, genome-wide association study, linkage mapping, genomic prediction

INTRODUCTION

Common rust, caused by *Puccinia sorghi*, is one of the major foliar diseases in maize, which can cause up to 49% grain yield loss in susceptible varieties (Groth et al., 1983). The most sustainable strategy for controlling common rust is to develop and deploy resistant maize varieties, which requires the identification of the new source of resistance to common rust and the further understanding of the genetic basis and architecture of common rust resistance (Kibe et al., 2020).

In several recent studies, a broad genetic variation for common resistance was observed in tropical maize, and a few tropical maize inbred lines showing good resistance to common rust were identified (Rossi et al., 2020; Sserumaga et al., 2020). Among 50 tropical adapted maize breeding lines developed by International Maize and Wheat Improvement Center (CIMMYT), 12 lines with broad genetic diversity were identified as the potential donors of resistance alleles, and these lines are valuable breeding materials for the development and deployment of resistant hybrids to control common rust in tropical maize (Sserumaga et al., 2020). Furthermore, tropical maize germplasm is also an important source of resistance for improving common rust in temperate maize, and the six inbred lines developed by CIMMYT were identified as novel donors in Argentina for incorporating resistance to the local germplasm (Rossi et al., 2020). Those studies indicated the presence of genetic resistance to common rust in tropical maize germplasm. The donor lines identified in these studies are valuable donors for improving common rust resistance through breeding, which also are novel resistance sources for providing a better understanding of the genetic basis and architecture of common rust resistance.

Host-plant resistance, including both qualitative and quantitative resistances, had been identified as the most reliable and sustainable strategy for controlling common rust in maize (Zheng et al., 2018; Kibe et al., 2020). Previous efforts to exploit genetic resistance for common rust have largely been through dominant resistance (Rp) genes, and more than 26 Rp genes had been identified on maize chromosomes 3, 4, 6, and 10 (Hooker, 1985; Delaney et al., 1988). The Rp gene is qualitative and exhibits a high level of resistance to a specific *P. sorghi* race, and the resistance allele of Rp genes can be easily fixed into the breeding materials, but the resistance of Rp genes in some hybrids could break down due to the emerging *P. sorghi* race or multiple races caused infection happened in natural field condition (Zheng et al., 2018; Kibe et al., 2020). Quantitative resistance is due to partial or adult plant resistance, which is non-race-specific and often controlled by several genes to reduce the rate of fungal development on plant tissues (Olukolu et al., 2016). A few studies have been carried out on quantitative resistance to common rust mainly through linkage mapping (Lübberstedt et al., 1998; Kerns et al., 1999; Brown et al., 2001). Further studies are required to detect more sources of novel quantitative resistance alleles and exploit them to develop elite inbred lines or hybrids having stable and durable host-plant resistance to common rust.

Several linkage mapping analyses had been conducted in different genetic backgrounds to detect quantitative trait loci (QTL) associated with partial resistance to common rust (Lübberstedt et al., 1998; Kerns et al., 1999; Brown et al., 2001). These studies emphasized QTL detection in temperate maize germplasm, and QTL associated with partial resistance to common rust were distributed over all 10 chromosomes, without preference to chromosomes 3, 4, 6, and 10, which harbor qualitative Rp genes. Some QTL were overlapped in different studies and were consistent in different genetic backgrounds. These results suggest that major QTL associated with partial resistance from various elite backgrounds are possible to be pyramided for improving common rust resistance in temperate maize germplasm, and selection for multiple partial resistance alleles seems to be more promising than the marker-assisted selection (MAS) of the Rp genes.

Genome-wide association study (GWAS) is a useful tool for identifying molecular markers significantly associated with the target trait and exploring the underlying candidate genes (Yan et al., 2011; Wang et al., 2019). In a collection of 274 temperate maize inbred lines, the GWAS analysis was conducted to identify the SNPs significantly associated with common rust resistance; three loci significantly associated with common rust resistance were identified; and they were on chromosomes 2, 3, and 8. Candidate genes at these loci had predicted roles in cell wall modification and in regulating the accumulation of reactive oxygen species (Olukolu et al., 2016). The combined use of GWAS and linkage mapping can complement the strengths and weaknesses of each approach, and this approach has been successfully used in maize to dissect the genetic basis and architecture of complex traits (Li et al., 2016; Cao et al., 2017). In tropical maize germplasm, the combined use of GWAS and linkage mapping approach was applied to dissect the genetic basis of partial resistance to common rust recently (Zheng et al., 2018; Kibe et al., 2020). The results of these studies provide valuable

information on understanding the genetic basis of common rust resistance; the common stable QTL regions identified by both GWAS and linkage mapping, and the major QTL identified by GWAS or linkage mapping individually need to be explored further for developing functional molecular markers for MAS.

Genomic selection (GS), also known as genomic prediction (GP), is an extension of MAS that uses genome-wide markers to predict the genomic estimated breeding values (GEBVs) of the unphenotyped lines for selection (Meuwissen et al., 2001; Crossa et al., 2014). GP can greatly accelerate the genetic gain per unit time and the cost in plant breeding programs for complex traits, and it has been reported in many studies (Gowda et al., 2015; Zhang et al., 2015; Beyene et al., 2019; Wang et al., 2020a). To our knowledge, only one study has been reported evaluating the potential of GS and GP for improving common rust resistance in maize, where the GP accuracies ranged from 0.19 to 0.51 in different populations (Kibe et al., 2020).

In this study, a GWAS panel and a bi-parental DH population were used to perform GWAS, linkage mapping, and GP analyses, where both populations were phenotyped in multi-environment trials to evaluate their responses to common rust and genotyped with genotyping-by-sequencing (GBS) single-nucleotide polymorphisms (SNPs). The main objectives of this study were to: (1) detect the significantly associated SNPs, major QTL, and putative candidate genes conferring common rust resistance in tropical maize by the combined use of GWAS and linkage mapping; (2) explore the potential of GS and GP for improving common rust resistance; and (3) estimate the GP accuracies under different factors affecting the accuracy estimation.

MATERIALS AND METHODS

Plant Materials

A GWAS panel of 282 genetically diverse inbred lines was used for the GWAS and GP analyses in this study (**Supplementary Table 1**). The GWAS panel, Drought Tolerant Maize for Africa (DTMA), was collected by the Global Maize Program of CIMMYT. Based on the geographic information and environmental adaptation, the DTMA panel can be classified into nine subsets: (1) breeding lines from the lowland tropical maize breeding program in Mexico, (2) breeding lines from the highland tropical maize breeding program in Mexico, (3) breeding lines from the subtropical maize breeding program in Mexico; (4) inbred lines from the maize physiology breeding program in Mexico, (5) inbred lines from the maize entomology breeding program in Mexico, (6) breeding lines from the lowland tropical maize breeding program in Colombia, (7) breeding lines from the mid-altitude maize breeding program in Zimbabwe, (8) breeding lines from the highland tropical maize breeding program in Ethiopia, and (9) breeding lines from the maize breeding program of International Institute of Tropical Agriculture in Nigeria (Cairns et al., 2013; Yuan et al., 2019). A bi-parental DH population, DH1, was used for the linkage mapping and GP analyses. This DH population consisted of 189 DH lines, which were derived from the F_1 cross formed with two elite inbred lines of CML495 and La Posta Sequia C7 F64-2-6-2-2-B-B-B, CML495 shows good resistance to common rust, and La Posta Sequia C7 F64-2-6-2-2-B-B-B is susceptible to common rust.

Experimental Design

Both populations were evaluated for response to common rust under consistently high natural disease pressure at several locations in Mexico. The DTMA panel was evaluated at Agua Fria in the state of Puebla (110 masl; mega-environment: lowland tropical) in 2008, 2009, 2010, and 2012. Two tropical maize inbred lines (B.T.Z.T.R.L.BA90 12-1-1P-1P-1-1-1-1P-1-B/BTZTVCPR92A 27-7P-1-1P-1P-4P-B-B)-B-60TL-1-1-B-B-B-B and CML139 were used in all the trials as the resistant and susceptible checks, respectively. The population of DH1 was evaluated in two locations in 2013 at El Batan in the state of Mexico (2,249 masl; mega-environment: highland tropical) and Santa Catarina in the state of Nuevo Leon (680 masl; mega-environment: subtropical), respectively. For the DH1 population, the parental lines were used as the resistant and susceptible checks. A randomized complete block design with three replications was used for all trials. Each plot consisted of 11 plants in a 2 m row with a width of 0.75 m.

Disease Evaluation

Plants were visually evaluated for common rust three times at 7-day intervals, beginning 2 weeks after flowering. Disease severity was evaluated on a 1–5 scale based on the percentage of leaf area covered by lesions. A rating scale of 1 corresponds to high resistance covering 0–10% of the leaf surface, 2 corresponds to weak to moderate infection covering 10–25% of the leaf surface, 3 corresponds to moderate infection covering 25–50% of the leaf surface, 4 corresponds to moderate-to-severe infection covering 50–75% of the leaf surface, and 5 corresponds to severe infection covering $>$ 75% of the leaf surface. For each plot, the final highest score was used for further analysis. In both the DTMA panel and the DH1 population, the resistant and susceptible checks were used as controls to check for adequate levels of disease infection.

Phenotypic Data Analysis

The multi-environment trial analysis was conducted using META-R Version 6.04 (Alvarado et al., 2020). A mixed linear model was used to calculate the best linear unbiased predictors (BLUPs), variance components, and broad-sense heritability. The model used for data analysis was as follows:

$$Y_{ijk} = \mu + G_k + E_i + R_{j(i)} + EG_{ik} + \varepsilon_{ijk} \quad (1)$$

where Y_{ijk} is the observation of the kth genotype in the ith environment in the jth replicate, μ is the overall mean, G_k is the effect of the kth genotype, E_i is the effect of the ith environment, $R_{j(i)}$ is the effect of the jth replication nested on the ith environment, EG_{ik} is the effect of the interaction between the ith environment and kth genotype, and ε_{ijk} is the effect of experimental error. BLUPs across all environments were used for GWAS, linkage mapping, and GP analyses. Broad-sense

heritability across all environments was calculated as follows:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_{ge}^2/i + \sigma_e^2/ij} \quad (2)$$

where σ_g^2 is the genotypic variance, σ_{ge}^2 is the genotype × environment interaction variance, σ_e^2 is the error variance, i is the number of environments, and j is the number of replications in each environment. All of the factors were set as random effects when calculating heritability.

Genotyping and Genotypic Data Analysis

Young leaves of all the inbred lines and the parental lines were sampled for both populations. DNA extraction was performed using a CTAB method (CIMMYT, 2005). Genotypic data was generated using the GBS method at the Cornell University Biotechnology Resource Center (Ithaca, NY, United States). DNA sequencing was performed on Illumina HiSeq2000. TASSEL GBS Pipeline was used for SNP calling to align reads to maize B73 reference genome v2 (ZmB73_RefGen_v2). Imputation was carried out with the FILLIN method in TASSEL V5.0 (Bradbury et al., 2007; Swarts et al., 2014). The imputed GBS dataset was used for the GWAS and GP analyses, while the unimputed GBS dataset was used for the linkage mapping analysis (Wang et al., 2020b). A total of 955,690 SNPs were obtained for each inbred line, and 570 of them could not be mapped to any of the 10 maize chromosomes. The number of SNPs on each chromosome ranged from 148,752 on chromosome 1 to 67,126 on chromosome 10. SNPs with the missing rate (MR) of >20%, the heterozygosity rate of >5%, and the minor allele frequency (MAF) of <0.05 were excluded using the filter function in TASSEL V5.0.

Analyses of Linkage Disequilibrium, Population Structure, and GWAS

After filtering, 187,409 SNPs were obtained for GWAS in the DTMA panel. The linkage disequilibrium (LD) analysis was carried out using TASSEL V5.0 with a sliding window size of 50 SNPs. A squared Pearson correlation coefficient (r^2) between the vectors of SNP alleles was used to assess the level of LD decay across each chromosome, and $r^2 = 0.1$ was used as a cutoff. Population structure was conducted using the STRUCTURE V2.3.4 software (Hubisz et al., 2009) to estimate the number of subgroups in the DTMA panel, where one SNP per LD block was selected for the following analysis (Duggal et al., 2008). The parameters were set as follows: length of burn-in period = 30,000, number of MCMC reps after burn-in = 30,000, ancestry model = use admixture model, allele frequency model = allele frequency correlated, number of populations (K) = 1–10, and number of iterations = 10. STRUCTURE HARVESTER (Earl and vonHoldt, 2012) was used to visualize STRUCTURE V2.3.4 output, and delta K (ΔK) value was used to determine the K value of the number of subgroups.

Analysis of GWAS was conducted in the DTMA panel using the Fixed and random model Circulating Probability Unification (FarmCPU) method (Liu et al., 2016) in Genome Association and Prediction Integrated Tool-R (GAPIT) package (Lipka et al., 2012). The kinship matrix and the first three PCs were estimated by GAPIT to assess the population structure and control the false marker-trait association. The P-value of each SNP was calculated, and the threshold of P-value was determined at 3.70×10^{-6} by a false discovery rate correction method. The 100 bp source sequences of each significant SNP were used to do BLAST against the ZmB73_RefGen_v2 genome sequence in MaizeGDB (Portwood et al., 2019). Within the local LD block of significant SNPs, the annotated genes that are likely involved in disease resistance were identified as the putative candidate genes.

Linkage Map Construction and Linkage Mapping Analysis

A similarity/linkage (SL) method was used for bin map construction with high-quality unimputed SNPs in the DH1 population, and the details were previously described by Cao et al. (2017). In brief, 437 bins were constructed by 31,194 SNPs. Each bin was regarded as a genetic marker to construct the linkage map. Linkage map construction was conducted by MAP function in QTL IciMapping V4.2 software (Meng et al., 2015). The whole length of the linkage map of DH1 was 988.56 cM with an average marker (bin) density of 2.26 cM. An inclusive composite interval mapping (ICIM) approach was conducted for the linkage mapping analysis using the "BIP" function and the "ADD" mapping method in QTL IciMapping V4.2. A logarithm of the odds (LOD) score of 3.0 was used to declare the putative QTL. The additive effect and phenotypic variation explained (PVE) of each QTL were estimated.

Genomic Prediction Analysis

Genomic prediction analysis was conducted using the ridge regression best linear unbiased prediction (RRBLUP) model with the rrBLUP package (Endelman, 2011) within the DTMA panel and the DH1 population. In the imputed GBS dataset, TASSEL version 5.0 was used to filter the SNPs with a MAF > 0.05, a MR < 20%, and a heterozygosity rate < 5%. After filtering, 187,409 and 53,996 SNPs were used for GP in the DTMA panel and the DH1 population, respectively. In the DH1 population, 437 bins were also used for the GP analysis to estimate the prediction accuracy and compared it with the prediction accuracy estimated using all the 53,996 SNPs. To estimate the effect of marker density on GP accuracy, the number of SNPs varied from 100 to 50,000 (i.e., 10, 50, 100, 300, 500, 1,000, 3,000, 5,000, 10,000, and 50,000) were used to estimate the prediction accuracy in the DTMA panel and the DH1 population. In each marker density, SNPs were randomly selected 100 times. A fivefold cross-validation scheme repeated 100 times was used to estimate the prediction accuracy, where the prediction accuracy was defined as the average value of the correlations between the GEBVs and the observed breeding values. Training population size (TPS), ranged from 10 to 90% of the total population size, was selected to assess the effect of TPS on prediction accuracy in each of the two populations. The training set was randomly sampled to predict, and the remaining lines were used as the prediction set. The GP analysis was repeated 100 times in each population with different TPS.

TABLE 1 | Descriptive statistics, variance components, and broad-sense heritability (H^2) response to common rust in the Drought Tolerant Maize for Africa (DTMA) panel and the bi-parental doubled haploid (DH1) population.

Population	No. of lines	Mean	Min.	Max.	Median	SD[a]	Variance components[b]			h^{2c}
							σ_g^2	σ_{ge}^2	σ_e^2	
DTMA	282	2.32	1.26	4.13	2.30	0.52	0.33**	0.25**	0.24	0.80
DH1	189	2.25	1.73	3.10	2.20	0.23	0.10**	0.08**	0.20	0.57

[a] *SD, standard deviation.*
[b] σ_g^2, *genotypic variance.*
σ_{ge}^2, *genotype × environment interaction variance.*
σ_e^2, *error variance.*
**Significant at $P < 0.01$.*
[c] h^2, *broad-sense heritability.*

RESULTS

Phenotypic Variations

The descriptive statistics for the response to common rust in the DTMA panel and the DH1 population are presented in **Table 1** and **Supplementary Figure 1A**. The results indicated that there were abundant phenotypic variations within each population. In the DTMA panel, the disease scores ranged from 1.26 to 4.13, with a mean of 2.32. In the DH1 population, the disease scores ranged from 1.73 to 3.10, with a mean of 2.25. The most resistant (top 10%) and most susceptible lines (bottom 10%) for common rust in the DTMA panel and the DH1 populations are shown in **Supplementary Tables 2**, **3**, respectively. The mixed model analysis result revealed that the genotypic variance was statistically highly significant at $P < 0.01$ in both populations, as well as the variance of genotype-by-environment interaction. The estimated broad-sense heritabilities in the DTMA panel and the DH1 population were 0.80 and 0.57, respectively.

Basic Information of SNPs Before and After Filtering

The basic information about GBS data before and after filtering is shown in **Supplementary Table 4**. The number of SNPs after filtering decreased from 955,690 to 187,409 in the imputed dataset of the DTMA panel and from 955,690 to 31,194 in the unimputed dataset of the DH1 population. The MR after filtering decreased from 15.79 to 7.33% in the imputed dataset of the DTMA panel and from 42.53 to 9.73% in the unimputed dataset of the DH1 population. The heterozygosity rate increased in both populations after filtering, and the heterozygosity rates after filtering in the DTMA panel and the DH1 population were 2.83 and 3.17%, respectively. The average MAF after filtering increased from 0.09 to 0.18 in the DTMA panel and from 0.04 to 0.42 in the DH1 population.

Results of LD Decay Distance and Population Structure in the DTMA Panel

In the DTMA panel, the average LD decay distance across all the 10 chromosomes was 8.14 kb at an r^2 value of 0.1 (**Figure 1A**), and it ranged from 4.57 kb in chromosome 10–15.9 kb in chromosome 8. The population structure analysis showed that the delta K value reached a peak when the K value was 4, indicating that the DTMA panel can be divided into four subgroups (**Figures 1B,C**). The number of lines in subgroups 1, 2, 3, and 4 was 219, 13, 10, and 40, respectively. The different responses to common rust in the four subgroups are shown in **Supplementary Figure 1B**. The principal component analysis also revealed four subgroups, corresponding to the four subgroups identified by STRUCTURE analysis (**Figure 1D**).

Significantly Associated SNPs and Candidate Genes Revealed by GWAS

The GWAS results of the DTMA panel are presented in **Table 2** and **Figure 2**. At a very stringent threshold of P-value of 3.70×10^{-6}, a total of six SNPs at bins of 1.05, 1.10, 3.04, 3.05, 4.08, and 10.04 were identified to be significantly associated with common rust resistance in maize. The quantile–quantile (q–q) plot implied that the population structure and family relatedness were well controlled in the GWAS using the FarmCPU method (**Figure 2B**).

Among all the six SNPs, the two most significantly associated SNPs were identified on chromosome 1. The most significantly associated SNP of S1_278132829 was located at the bin of 1.10, it had the lowest P-value of 7.25×10^{-11}, and the MAF of this SNP was 0.25, with an additive effect of 0.13. The candidate gene of *GRMZM2G328500* (278,126,093–278,132,841 bp), encoding a UDP-glucose 6-dehydrogenase, contains the most significantly associated SNP of S1_278132829. The second most significantly associated SNP of S1_89238026 was located at the bin of 1.05, it had the second-lowest P-value of 9.81×10^{-10}, and the MAF of this SNP was 0.32, with an additive effect of 0.13. It neighbored with the candidate gene of *GRMZM2G114893* (89,236,681–89,237,918 bp), which encodes a zinc finger (C_2H_2 type) family protein.

On chromosome 3, two significantly associated SNPs were identified, i.e., S3_118933375 located at the bin of 3.04 and S3_147594533 located at the bin of 3.05. The SNP of S3_118933375 had a MAF of 0.10, with an additive effect of −0.17, and it was 587 bp away from the candidate gene of *GRMZM2G144004* (118,931,829–118,932,788 bp), encoding a putative uncharacterized protein. The SNP of S3_147594533 had a MAF of 0.11, with an additive effect of 0.15, and it was

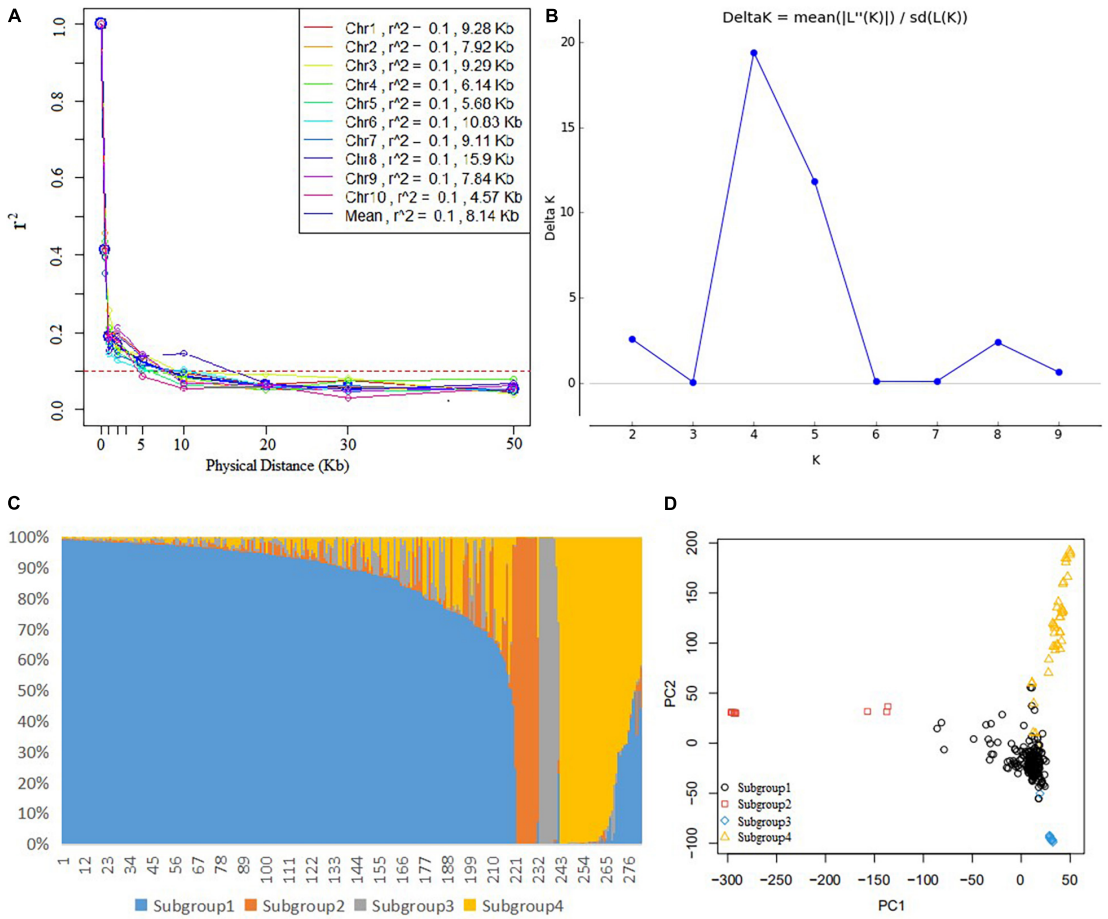

FIGURE 1 | Analysis of genetic diversity in the genome-wide association study (GWAS) panel. **(A)** Linkage disequilibrium decay across all 10 maize chromosomes, **(B)** the plot of delta *K*, **(C)** the estimated probability membership for each inbred line at *K* = 4, and **(D)** the principal component analysis plot showing four subgroups corresponding to the four subgroups by the STRUCTURE analysis.

TABLE 2 | Significantly associated single-nucleotide polymorphisms (SNPs) and candidate genes revealed by the genome-wide association study analysis.

SNP[a]	*P*-value	Allele[b]	MAF[c]	SNP effect[d]	Putative candidate gene	Annotation of candidate genes
S1_89238026	9.81×10^{-10}	A/G	0.32	0.13	*GRMZM2G114893*	Zinc finger (C2H2 type) family protein
S1_278132829	7.25×10^{-11}	A/T	0.25	0.13	*GRMZM2G328500*	UDP-glucose 6-dehydrogenase
S3_118933375	1.00×10^{-6}	C/T	0.10	−0.17	*GRMZM2G144004*	Unknown
S3_147594533	1.11×10^{-7}	A/T	0.11	0.15	*GRMZM2G162250*	*Zea mays ARGOS6*
S4_183913302	2.98×10^{-7}	G/C	0.17	0.13	*GRMZM2G138949*	BTB/POZ domain-containing protein
S10_95231291	1.32×10^{-7}	C/A	0.10	−0.16	*GRMZM2G131536*	Unknown

[a] *SNP name, chromosome_position, for example, S1_89238026 represents that the SNP is on chromosome 1, and the physical position is 89238026 bp.*
[b] *Letters to the left and right of the "/" refer to major allele and minor allele, respectively.*
[c] *MAF, minor allele frequency.*
[d] *Positive values indicate that the major allele is a resistance allele, and the negative values indicate that the minor allele is a resistance allele.*

located at the candidate gene of *GRMZM2G162250* (147,591,043–147,598,482 bp), which encodes a *Zea mays* ARGOS6 (auxin-regulated gene involved in organ size) protein.

On chromosome 4, the significantly associated SNP of S4_183913302 was located at the bin of 4.08, it had a MAF of 0.17, with an additive effect of 0.13, and this SNP was close to the candidate gene of *GRMZM2G138949* (183,909,192–183,910,514 bp), encoding a BTB/POZ domain-containing protein. On chromosome 10, the significantly associated SNP of S10_95231291 was located at the bin of 10.04, it had a MAF of 0.10, with an additive effect of −0.16, and this SNP was closely linked with the candidate gene of *GRMZM2G131536* (95,230,282–95,231,024 bp).

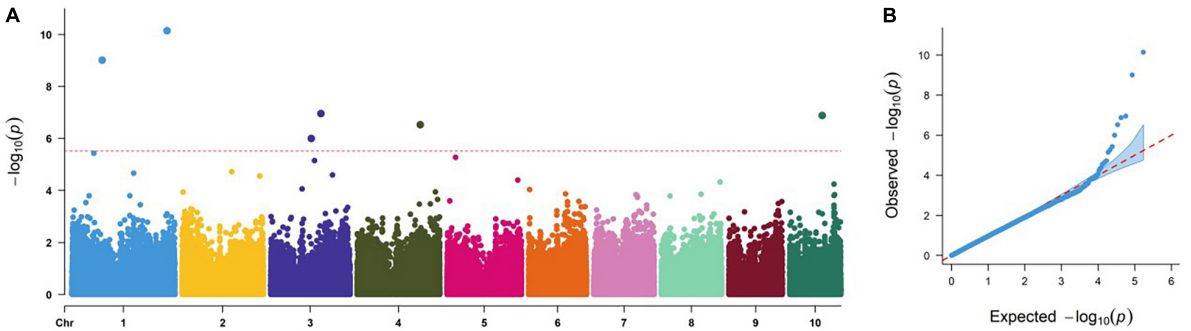

FIGURE 2 | Genome-wide association study Manhattan and quantile–quantile (q–q) plots for common rust resistance in the Drought Tolerant Maize for Africa (DTMA) panel. **(A)** Manhattan plot, the dashed line corresponds to the threshold level defined at $P = 3.70 \times 10^{-6}$ by a false discovery rate correction method; **(B)** q–q plot.

TABLE 3 | Quantitative trait loci detected from the linkage mapping analysis in the doubled haploid (DH1) population.

Chromosome	Position (cM)	Bin	Left marker[a]	Right marker	LOD[b]	PVE(%)[c]	Additive effect
1	28	1.03	S1_31252133	S1_34315390	6.77	10.34	−0.08
2	47	2.06	S2_183941772	S2_188133361	3.49	5.69	0.06
4	74	4.08	S4_184936775	S4_186039203	4.62	6.79	0.06
7	67	7.03	S7_144567253	S7_149717562	7.82	12.45	0.09
9	0	9.00	S9_1260192	S9_2825523	3.70	5.40	−0.06

[a] *Marker name, chromosome_position.*
[b] *LOD, logarithm of the odds.*
[c] *PVE, phenotypic variation explained.*

Quantitative Trait Loci Detected From Linkage Mapping Analysis

The linkage mapping results of the DH1 population are presented in **Table 3**. In total, five QTL located at bins 1.03, 2.06, 4.08, 7.03, and 9.00 were detected at the threshold of a LOD score of 3.0. The PVE value of the individual QTL ranged from 5.40 to 12.45%, and the total PVE value for all the five QTL was 40.67%. The QTL on chromosome 7 had the highest LOD score of 7.82 and the largest PVE value of 12.45%, indicating that it is a major QTL conferring the common rust resistance in maize. The common rust resistance alleles were derived from the resistant inbred line CML495 except for the two QTL located on chromosomes 1 and 9.

The significantly associated SNP of S4_183913302 identified by GWAS was closely linked with the QTL detected in DH1 on chromosome 4, it was flanked by the markers S4_184936775 and S4_186039203, and this QTL had a LOD score of 4.62 and a PVE value of 6.79%. However, the most significantly associated SNP of S1_278132829 identified by GWAS was not validated by the linkage mapping analysis. The major QTL on chromosome 7 detected from linkage mapping analysis was also not validated by the GWAS result.

Prediction Accuracies Estimated With the Different Marker Datasets, Marker Density, and Training Population Size

The GP accuracies estimated based on GBS SNPs were 0.61 and 0.51 in the DTMA panel and the DH1 population, respectively (**Figure 3A**). The GP accuracy based on bin markers was 0.53 in DH1 (**Figure 3B**). No significant difference in prediction accuracy was observed between GBS SNPs and bin markers. The effect of marker density and TPS on the GP accuracy is shown in **Figure 4**. In both the DTMA panel and the DH1 population, the prediction accuracy increased as the number of markers increased. In the DTMA panel, the prediction accuracy increased rapidly when the number of markers increased from 10 to 5,000, and then, the prediction accuracy increased slightly when the number of markers kept increasing. In the DH1 population, a sharp increase in the prediction accuracy was observed before reaching a plateau at about 300 markers, indicating that 300 SNPs were sufficient to achieve good accuracy of common rust resistance in the DH1 population. Prediction accuracy increased as the TPS increased in both populations. In the DTMA panel, the prediction accuracy increased rapidly when the TPS increased from 10 to 50%, and then, a little improvement in the prediction accuracy was observed when the TPS kept increasing. When 50% of the total genotypes were used as the training set, a relatively high prediction accuracy coupled with the smaller standard error was observed. A similar trend was observed in the DH1 population.

DISCUSSION

Common rust is a major disease of maize, causing 34% of the maize area to suffer economic losses (Zheng et al., 2018). Developing maize varieties with host plant resistance is the most sustainable strategy for the control of common rust,

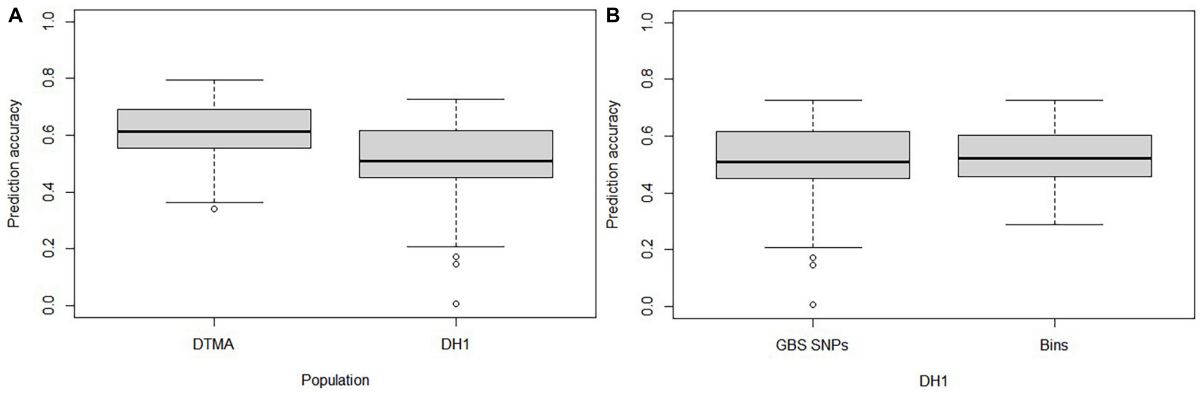

FIGURE 3 | Genomic prediction accuracy of common rust resistance in the DTMA panel and DH1 population. **(A)** In the DTMA panel and DH1 population estimated with genotyping-by-sequencing (GBS) single-nucleotide polymorphisms (SNPs); **(B)** in the DH1 panel estimated with GBS SNPs and bins.

which requires further understanding of the genetic basis and architecture of common rust resistance. Previous efforts to exploit genetic resistance for common rust have largely been through Rp genes, but the resistance of Rp genes could break down easily. Quantitative disease resistance controlled by several genes has proven to be highly durable, making it a better choice for long-term common rust resistance breeding. In this study, GWAS and linkage mapping analyses were applied to dissect the genetic base of quantitative resistance of common rust in maize. GWAS revealed six SNPs significantly associated with quantitative resistance of common rust at a very stringent threshold of *P*-value of 3.70×10^{-6}. Linkage mapping identified five QTL accounting for the total PVE value of 40.67%. These results provided new insight into the quantitative resistance of common rust, which implied that major QTL associated with quantitative resistance from various elite backgrounds are possible to be pyramided for improving common rust resistance, and the selection for multiple partial resistance alleles seems to be more promising than the MAS of the Rp genes in tropical maize germplasm.

In the GWAS, six SNPs distributed in bins 1.05, 1.10, 3.04, 3.05, 4.08, and 10.04 were associated with common rust resistance. Except for SNP of S10_95231291, all the SNPs were reported in previous GWAS and linkage mapping studies (Lübberstedt et al., 1998; Brown et al., 2001; Zheng et al., 2018; Kibe et al., 2020). The most and the second most significantly associated SNPs S1_278132829 and S1_89238026 detected in this study were also detected by linkage mapping in European flint germplasm (Lübberstedt et al., 1998). SNP S3_118933375 was in the same region of qCR3-113, a QTL for common rust (Kibe et al., 2020), and it was also close to SNP PZE-103072633 (115,864,889) (Zheng et al., 2018). Both qCR3-113 and PZE-103072633 were detected in tropical maize germplasm. SNPs S3_147594533 and S4_183913302 were mapped to the QTL intervals associated with common rust in sweet corn (Brown et al., 2001). QTL detected for a target trait are usually different due to the use of different genetic backgrounds and environments (Ren et al., 2020). Those common loci detected in different studies were stable QTL for common rust. SNP S10_95231291 was first reported, it had an additive effect of −0.16, and it was closely linked with the candidate gene of GRMZM2G131536. However, the function of the candidate gene of GRMZM2G131536 is still unknown.

In DH1, linkage mapping revealed five QTL distributed in bins 1.03, 2.06, 4.08, 7.03, and 9.00, respectively. Three of the five QTL were reported previously (Lübberstedt et al., 1998; Brown et al., 2001). The loci in bins 1.03 and 2.06 coincided with QTL identified by Lübberstedt et al. (1998). The locus in bin 4.08 was detected by both Lübberstedt et al. (1998) and Brown et al. (2001). The major QTL located on chromosome 7 was reported in this study for the first time, and it had the highest LOD score of 7.82 and the largest PVE value of 12.45%. It is a new source of resistance for common rust, which deserves further investigation.

Joint GWAS and linkage mapping can complement the advantages and disadvantages of each method (Li et al., 2016; Cao et al., 2017). In this study, GWAS and linkage mapping were implemented stepwise to detect loci associated with quantitative resistance of common rust. The genomic region located at bin 4.08 was detected by both GWAS and linkage mapping. SNP S4_183913302 was consistent with the locus identified between markers S4_184936775 and S4_186039203 in DH1. This locus was also reported by Lübberstedt et al. (1998) and Brown et al. (2001). The major QTL located on chromosome 7 identified by linkage mapping in DH1 was not detected through GWAS in the DTMA panel. This may be due to the very low frequency of one of the alleles of the relevant locus in the GWAS panel or the population structure related to the polymorphism at this locus (Famoso et al., 2011; Cadic et al., 2013). The most significantly associated SNP of S1_278132829 identified by GWAS was also not validated by the linkage mapping analysis. It may be because there is no genetic variation at this locus in the DH1 population. The major QTL identified by GWAS or linkage mapping individually, and the common stable QTL region identified by both methods need to be explored further for developing functional molecular markers for MAS.

The candidate gene analysis can lead to a better understanding of the genetic basis of common rust resistance. According to the results of GWAS, six candidate genes were identified in this study, and the function of four candidate genes

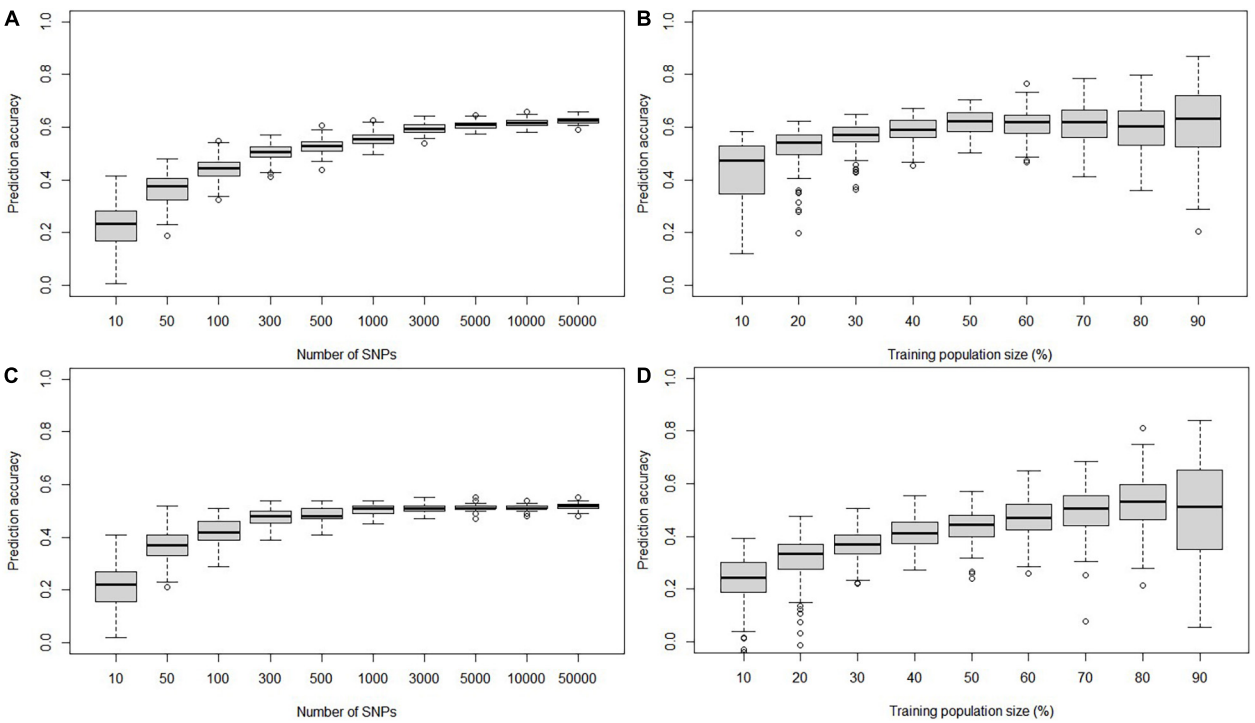

FIGURE 4 | Genomic prediction accuracy of common rust resistance in the DTMA panel and DH1 population, when the number of SNPs varied from 100 to 50,000 and training population size (TPS) ranged from 10 to 90% of the total population size. **(A)** In the DTMA panel estimated with different marker density, **(B)** in the DTMA panel estimated with different TPS, **(C)** in the DH1 population estimated with different marker density, and **(D)** in the DH1 population estimated with different TPS.

was annotated. These candidate genes were previously reported to play important roles in the response of stress resilience and the regulation of plant growth and development. *GRMZM2G328500* encodes a UDP-glucose 6-dehydrogenase, which is involved in the nucleotide-sugar interconversion process (Kost et al., 2020). *GRMZM2G162250* encodes a *Zea mays* ARGOS6 protein controlling plant growth, organ size, and grain yield. *GRMZM2G114893* encodes a zinc finger (C_2H_2 type) family protein, which is mainly involved in the regulation of plant growth, development, and tolerance to biotic and abiotic stresses (Kim et al., 2009; Xiao et al., 2009). *GRMZM2G138949* identified in bin 4.08 encodes a BTB/POZ domain-containing protein, which participates in a series of physiological and biochemical reactions and also plays an important role in resistance to plant disease (Cao et al., 1997; Silva et al., 2015). These results encourage fine-mapping and cloning of the candidate genes for controlling common rust in maize.

Genomic prediction and GS have been successfully used in several crops to accelerate genetic gain in breeding programs for improving complex traits, including resistance to major maize diseases (Gowda et al., 2015; Liu et al., 2021). A study on the potential of GS and GP to improve the common rust resistance in maize has been reported by Kibe et al. (2020), where the GP accuracies within populations ranged from 0.19 to 0.51, and the GP accuracies estimated from a larger population by combined several individual populations were higher than those estimated from the individual population with a smaller size. For implementing GP and GS to improve common rust resistance in tropical maize, an independent but related training set is encouraged to be built to predict the related populations not been phenotyped. These results were confirmed by this study. The GP accuracies observed in the DTMA panel and the DH1 population were 0.61 and 0.51, respectively. It indicates that common rust resistance in tropical maize could be improved by implementing GP and GS. Moreover, the factors affecting GP accuracy were also assessed in this study. Ten levels of marker density were used to assess the effect of marker density on prediction accuracy in the two populations. The results showed that the increase in marker density leads to an increase in prediction accuracy. The prediction accuracy reached a plateau when the marker density was 5,000 in the DTMA panel and 300 in the DH1 panel, which indicated that more makers are required to achieve good GP accuracy in populations with higher genetic diversity. A similar phenomenon was found for several traits in maize (Zhang et al., 2017; Guo et al., 2020; Liu et al., 2021). There was no significant difference between the prediction accuracy estimated based on the GBS SNPs and the bins in the DH1 population, which validated the high quality and accuracy of bins constructed in the bi-parental population. To assess the effect of TPS on prediction accuracy, nine levels of TPS were selected. As a result, increasing TPS leads to an increase in prediction accuracy. When 50% of the total genotypes were used as the training set, a relatively high prediction accuracy can be achieved. These results provide valuable information for improving common rust resistance in tropical maize by implementing GP and GS.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available. These data can be found at the CIMMYT Research Data & Software Repository Network: https://hdl.handle.net/11529/10548575.

AUTHOR CONTRIBUTIONS

BP, MO, TD, FS, and XZ conceived and designed the experiments. TD, FS, and XZ coordinated the phenotyping. XZ, BP, and MO coordinated the genotyping. JR, ZL, PW, AZ, YL, GH, SC, JQ, and HZ analyzed the data. JR, ZL, FS, and XZ drafted the manuscript. JR, ZL, MO, BP, FS, and XZ interpreted the results. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by the Mexico's Secretary of Agriculture and Rural Development (SADER), the Genomic Open-source Breeding Informatics Initiative (GOBII) (grant number: OPP1093167) supported by the Bill & Melinda Gates Foundation, and the CGIAR Research Program (CRP) on MAIZE. The CGIAR Research Program MAIZE receives W1&W2 support from the Governments of Australia, Belgium, Canada, China, France, India, Japan, South Korea, Mexico, Netherlands, New Zealand, Norway, Sweden, Switzerland, United Kingdom, United States, and the World Bank. This research was also funded by the CIMMYT-China Specialty Maize Research Center project funded by the Shanghai Municipal Finance Bureau, the National Natural Science Foundation of China (grant numbers: 32001561, 32060484, and 31771814), Xinjiang Youth Foundation (grant number: 2019D01A41), Chinese Postdoc Foundation (grant numbers: 2018M643774 and 2017M621318), the Chinese Postdoctoral International Exchange Program, the China Scholarship Council, Postdoctoral Science Foundation of Heilongjiang Province (grant number: LBH-Z17205), and the Fund for Distinguished Young Scholars of Heilongjiang Academy of Agricultural Sciences (grant number: 2019JCQN004).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.692205/full#supplementary-material

REFERENCES

Alvarado, G., Rodríguez, F. M., Pacheco, A., Burgueño, J., Crossa, J., Vargas, M., et al. (2020). META-R: a software to analyze data from multi-environment plant breeding trials. *Crop J.* 8, 745–756. doi: 10.1016/j.cj.2020.03.010

Beyene, Y., Gowda, M., Olsen, M. S., Robbins, K. R., Pérez-Rodríguez, P., Alvarado, G., et al. (2019). Empirical comparison of tropical maize hybrids selected through genomic and phenotypic selections. *Front. Plant Sci.* 10:1502. doi: 10.3389/fpls.2019.01502

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23, 2633–2635. doi: 10.1093/bioinformatics/btm308

Brown, A. F., Juvik, J. A., and Pataky, J. K. (2001). Quantitative trait loci in sweet corn associated with partial resistance to Stewart's wilt, northern corn leaf blight, and common rust. *Phytopathology* 91, 293–300. doi: 10.1094/PHYTO.2001.91.3.293

Cadic, E., Coque, M., Vear, F., Grezes-Besset, B., Pauquet, J., Piquemal, J., et al. (2013). Combined linkage and association mapping of flowering time in Sunflower (*Helianthus annuus L.*). *Theor. Appl. Genet* 126, 1337–1356. doi: 10.1007/s00122-013-2056-2

Cairns, J. E., Crossa, J., Zaidi, P. H., Grudloyma, P., Sanchez, C., Araus, J. L., et al. (2013). Identification of drought, heat, and combined drought and heat tolerant donors in maize. *Crop Sci.* 53, 1335–1346. doi: 10.2135/cropsci2012.09.0545

Cao, H., Glazebrook, J., Clarke, J. D., Volko, S., and Dong, X. (1997). The *Arabidopsis* NPR1 gene that controls systemic acquired resistance encodes a novel protein containing ankyrin repeats. *Cell* 88, 57–63. doi: 10.1016/s0092-8674(00)81858-9

Cao, S., Loladze, A., Yuan, Y., Wu, Y., Zhang, A., Chen, J., et al. (2017). Genome-wide analysis of tar spot complex resistance in maize using genotyping-by-sequencing SNPs and whole-genome prediction. *Plant Genome* 10:2. doi: 10.3835/plantgenome2016.10.0099

CIMMYT (2005). *CIMMYT Applied Molecular Genetics Laboratory, Laboratory Protocols*. Mexico: CIMMYT.

Crossa, J., Pérez, P., Hickey, J., Burgueño, J., Ornella, L., Cerón-Rojas, J., et al. (2014). Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity* 112, 48–60. doi: 10.1038/hdy.2013.16

Delaney, D. E., Webb, C. A., and Hulbert, S. H. (1988). A novel rust resistance gene in maize showing overdominance. *Mol. Plant Microbe Int.* 11, 242–245. doi: 10.1094/MPMI.1998.11.3.242

Duggal, P., Gillanders, E. M., Holmes, T. N., and Bailey-Wilson, J. E. (2008). Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. *BMC Genom.* 9:516. doi: 10.1186/1471-2164-9-516

Earl, D. A., and vonHoldt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv, Genet Resour.* 4, 359–361. doi: 10.1007/s12686-011-9548-7

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 4, 250–255. doi: 10.3835/plantgenome2011.08.0024

Famoso, A. N., Zhao, K., Clark, R. T., Tung, C., Wright, M. H., Kochian, L. V., et al. (2011). Genetic architecture of aluminum tolerance in rice (*Oryza sativa*) determined through genome-wide association analysis and QTL mapping. *PLoS Genet.* 7:e1002221. doi: 10.1371/journal.pgen.1002221

Gowda, M., Das, B., Makumbi, D., Babu, R., Semagn, K., Mahuku, G., et al. (2015). Genome-wide association and genomic prediction of resistance to maize lethal necrosis disease in tropical maize germplasm. *Theor. Appl. Genet.* 128, 1957–1968. doi: 10.1007/s00122-015-2559-0

Groth, J. V., Zeyen, R. J., Davis, D. W., and Christ, B. J. (1983). Yield and quality losses caused by common rust (*Puccinia sorghi* Schw.) in sweet corn (*Zea mays*) hybrids. *Crop Prot.* 2, 105–111. doi: 10.1016/0261-2194(83)90030-3

Guo, R., Dhliwayo, T., Mageto, E. K., Palacios-Rojas, N., Lee, M., Yu, D., et al. (2020). Genomic prediction of kernel zinc concentration in multiple maize populations using genotyping-by-sequencing and repeat amplification sequencing markers. *Front. Plant Sci.* 11:534. doi: 10.3389/fpls.2020.00534

Hooker, A. L. (1985). "Corn and sorghum rusts," in *The Cereal Rusts*, eds A. P. Roelfs and W. R. Bushnell (Orlando: Academic Press), 207–236. doi: 10.1016/b978-0-12-148402-6.50015-8

Hubisz, M. J., Falush, D., Stephens, M., and Pritchard, J. K. (2009). Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resour.* 9, 1322–1332. doi: 10.1111/j.1755-0998.2009.02591.x

Kerns, M. R., Dudley, J. W., and Rufener, G. K. (1999). QTL for resistance to common rust and smut in maize. *Maydica* 44, 37–45.

Kibe, M., Nyaga, C., Nair, S. K., Beyene, Y., Das, B., M, S. L., et al. (2020). Combination of linkage mapping, gwas, and gp to dissect the genetic basis of common rust resistance in tropical maize germplasm. *Int. J. Mol. Sci.* 21:6518. doi: 10.3390/ijms21186518

Kim, H. R., Chae, K. S., Han, K. H., and Han, D. M. (2009). The nsdC gene encoding a putative C2H2-type transcription factor is a key activator of sexual development in Aspergillus nidulans. *Genetics* 182, 771–783. doi: 10.1534/genetics.109.101667

Kost, M. A., Perales, H., Wijeratne, S., Wijeratne, A. J., Stockinger, E. J., and Mercer, K. L. (2020). Transcriptional differentiation of UV-B protectant genes in maize landraces spanning an elevational gradient in Chiapas, Mexico. *Evol. Appl.* 13, 1949–1967. doi: 10.1111/eva.12954

Li, X., Zhou, Z., Ding, J., Wu, Y., Zhou, B., Wang, R., et al. (2016). Combined linkage and association mapping reveals QTL and candidate genes for plant and ear height in maize. *Front. Plant Sci.* 7:833. doi: 10.3389/fpls.2016.00833

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., et al. (2012). GAPIT: genome association and prediction integrated tool. *Bioinformatics* 28, 2397–2399. doi: 10.1093/bioinformatics/bts444

Liu, X., Huang, M., Fan, B., Buckler, E. S., and Zhang, Z. (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet.* 12:e1005767. doi: 10.1371/journal.pgen.1005767

Liu, Y., Hu, G., Zhang, A., Loladze, A., Hu, Y., Wang, H., et al. (2021). Genome-wide association study and genomic prediction of *Fusarium* ear rot resistance in tropical maize germplasm. *Crop J.* 9, 325–341. doi: 10.1016/j.cj.2020.08.008

Lübberstedt, T., Klein, D., and Melchinger, A. E. (1998). Comparative quantitative trait loci mapping of partial resistance to *Puccinia sorghi* across four populations of European flint maize. *Phytopathology* 88, 1324–1329. doi: 10.1094/PHYTO.1998.88.12.1324

Meng, L., Li, H., Zhang, L., and Wang, J. (2015). QTL IciMapping: integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations. *Crop J.* 3, 269–283. doi: 10.1016/j.cj.2015.01.001

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. doi: 10.1093/genetics/157.4.1819

Olukolu, B. A., Tracy, W. F., Wisser, R., De Vries, B., and Balint-Kurti, P. J. (2016). A genome-wide association study for partial resistance to maize common rust. *Phytopathology* 106, 745–751. doi: 10.1094/PHYTO-11-15-0305-R

Portwood, J. C. L., Woodhouse, M. R., Cannon, E. K., Gardiner, J. M., Harper, L. C., Schaeffer, M. L., et al. (2019). MaizeGDB 2018: the maize multi-genome genetics and genomics database. *Nucleic Acids Res.* 47, D1146–D1154. doi: 10.1093/nar/gky1046

Ren, J., Boerman, N. A., Liu, R., Wu, P., Trampe, B., Vanous, K., et al. (2020). Mapping of QTL and identification of candidate genes conferring spontaneous haploid genome doubling in maize (*Zea mays* L.). *Plant Sci.* 293:110337. doi: 10.1016/j.plantsci.2019.110337

Rossi, E. A., Ruiz, M., Bonamico, N. C., and Balzarini, M. G. (2020). Identifying inbred lines with resistance to endemic diseases in exotic maize germplasm. *Crop Sci.* 60, 3141–3150. doi: 10.1002/csc2.20275

Silva, K. J., Brunings, A., Peres, N. A., Mou, Z., and Folta, K. M. (2015). The *Arabidopsis* NPR1 gene confers broad-spectrum disease resistance in strawberry. *Transgenic Res.* 24, 693–704. doi: 10.1007/s11248-015-9869-5

Sserumaga, J. P., Makumbi, D., Assanga, S. O., Mageto, E. K., Njeri, S. G., Jumbo, B. M., et al. (2020). Identification and diversity of tropical maize inbred lines with resistance to common rust (*Puccinia sorghi* Schwein). *Crop Sci.* 60, 2971–2989. doi: 10.1002/csc2.20345

Swarts, K., Li, H., Navarro, J. A. R., An, D., Romay, M. C., Hearne, S., et al. (2014). Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. *Plant Genome* 7:3. doi: 10.3835/plantgenome2014.05.0023

Wang, N., Liu, B., Liang, X., Zhou, Y., Song, J., Yang, J., et al. (2019). Genome-wide association study and genomic prediction analyses of drought stress tolerance in China in a collection of off-PVP maize inbred lines. *Mol. Breed.* 39:113. doi: 10.1007/s11032-019-1013-4

Wang, N., Wang, H., Zhang, A., Liu, Y., Yu, D., Hao, Z., et al. (2020a). Genomic prediction across years in a maize doubled haploid breeding program to accelerate early-stage testcross testing. *Theor. Appl. Genet.* 133, 2869–2879. doi: 10.1007/s00122-020-03638-5

Wang, N., Yuan, Y., Wang, H., Yu, D., Liu, Y., Zhang, A., et al. (2020b). Applications of genotyping-by-sequencing (GBS) in maize genetics and breeding. *Sci. Rep.* 10:16308. doi: 10.1038/s41598-020-73321-8

Xiao, H., Tang, J., Li, Y., Wang, W., Li, X., Jin, L., et al. (2009). *STAMENLESS 1*, encoding a single C2H2 zinc finger protein, regulates floral organ identity in rice. *Plant J.* 59, 789–801. doi: 10.1111/j.1365-313X.2009.03913.x

Yan, J., Warburton, M., and Crouch, J. (2011). Association mapping for enhancing maize (*Zea mays* L.) genetic improvement. *Crop Sci.* 51, 433–449. doi: 10.2135/cropsci2010.04.0233

Yuan, Y., Cairns, J. E., Babu, R., Gowda, M., Makubi, D., Magorokosho, C., et al. (2019). Genome-wide association mapping and genomic prediction analyses reveal the genetic architecture of grain yield and flowering time under drought and heat stress conditions in maize. *Front. Plant Sci.* 9:1919. doi: 10.3389/fpls.2018.01919

Zhang, A., Wang, H., Beyene, Y., Semagn, K., Liu, Y., Cao, S., et al. (2017). Effect of trait heritability, training population size and marker density on genomic prediction accuracy estimation in 22 bi-parental tropical maize populations. *Front. Plant Sci.* 8:1916. doi: 10.3389/fpls.2017.01916

Zhang, X., Pérez-Rodríguez, P., Semagn, K., Beyene, Y., Babu, R., López-Cruz, M. A., et al. (2015). Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and GBS SNPs. *Heredity* 114, 291–299. doi: 10.1038/hdy.2014.99

Zheng, H., Chen, J., Mu, C., Makumbi, D., Xu, Y., and Mahuku, G. (2018). Combined linkage and association mapping reveal QTL for host plant resistance to common rust (*Puccinia sorghi*) in tropical maize. *BMC Plant Biol.* 18:310. doi: 10.1186/s12870-018-1520-1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 25 August 2021
doi: 10.3389/fpls.2021.727077

Construction of Consensus Genetic Map With Applications in Gene Mapping of Wheat (*Triticum aestivum* L.) Using 90K SNP Array

*Pingping Qu[1,2], Jiankang Wang[1], Weie Wen[3], Fengmei Gao[4], Jindong Liu[1], Xianchun Xia[1], Huiru Peng[2] and Luyan Zhang[1]**

[1] The National Key Facility for Crop Gene Resources and Genetic Improvement, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, China, [2] State Key Laboratory of Agrobiotechnology, Key Laboratory of Crop Heterosis and Utilization, Beijing Key Laboratory of Crop Genetic Improvement, College of Agronomy and Biotechnology, China Agricultural University, Beijing, China, [3] Department of Cell Biology, Zunyi Medical University, Zunyi, China, [4] Crop Research Institute, Heilongjiang Academy of Agricultural Sciences, Harbin, China

OPEN ACCESS

Edited by:
Leonardo Abdiel Crespo Herrera, International Maize and Wheat Improvement Center, Mexico

Reviewed by:
Guriqbal Singh Dhillon, Thapar Institute of Engineering & Technology, India
Pasquale De Vita, Research Centre for Industrial Crops, Italy

***Correspondence:**
Luyan Zhang
zhangluyan@caas.cn

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *18 June 2021*
Accepted: *28 July 2021*
Published: *25 August 2021*

Citation:
Qu P, Wang J, Wen W, Gao F, Liu J, Xia X, Peng H and Zhang L (2021) Construction of Consensus Genetic Map With Applications in Gene Mapping of Wheat (Triticum aestivum L.) Using 90K SNP Array. Front. Plant Sci. 12:727077. doi: 10.3389/fpls.2021.727077

Wheat is one of the most important cereal crops worldwide. A consensus map combines genetic information from multiple populations, providing an effective alternative to improve the genome coverage and marker density. In this study, we constructed a consensus map from three populations of recombinant inbred lines (RILs) of wheat using a 90K single nucleotide polymorphism (SNP) array. Phenotypic data on plant height (PH), spike length (SL), and thousand-kernel weight (TKW) was collected in six, four, and four environments in the three populations, and then used for quantitative trait locus (QTL) mapping. The mapping results obtained using the constructed consensus map were compared with previous results obtained using individual maps and previous studies on other populations. A simulation experiment was also conducted to assess the performance of QTL mapping with the consensus map. The constructed consensus map from the three populations spanned 4558.55 cM in length, with 25,667 SNPs, having high collinearity with physical map and individual maps. Based on the consensus map, 21, 27, and 19 stable QTLs were identified for PH, SL, and TKW, much more than those detected with individual maps. Four PH QTLs and six SL QTLs were likely to be novel. A putative gene called *TraesCS4D02G076400* encoding gibberellin-regulated protein was identified to be the candidate gene for one major PH QTL located on 4DS, which may enrich genetic resources in wheat semi-dwarfing breeding. The simulation results indicated that the length of the confidence interval and standard errors of the QTLs detected using the consensus map were much smaller than those detected using individual maps. The consensus map constructed in this study provides the underlying genetic information for systematic mapping, comparison, and clustering of QTL, and gene discovery in wheat genetic study. The QTLs detected in this study had stable effects across environments and can be used to improve the wide adaptation of wheat cultivars through marker-assisted breeding.

Keywords: wheat (*Triticum aestivum* L.), consensus genetic map, QTL mapping, plant height, spike length, thousand-kernel weight

INTRODUCTION

Wheat (*Triticum aestivum* L.) is one of the most important cereal crops worldwide, providing about one-fifth of the total calories consumed by humans. Due to limited farmland and the rapid increase in human population, there is an urgent need to accelerate the genetic gain on grain yield through advanced genetic research and breeding activities in wheat. Genetic linkage map construction and quantitative trait locus (QTL) mapping are important areas in genetic research, as they provide fundamental information for gene cloning, marker-assisted breeding, and genome structure studies (Meng et al., 2015; Rasheed et al., 2016).

Linkage mapping approach based on individual populations has become routine in wheat genetic studies to dissect the genetic architecture of complex traits. However, a large number of co-localized markers and low marker density due to a limited genetic variation and a limited number of crossing-over events are commonly seen with linkage maps constructed in individual populations. Detected QTLs are usually specific to designated crosses with wide confidence intervals, hindering further genetic research on gene fine-mapping and cloning. Furthermore, linkage mapping in single populations can only identify QTLs with phenotypic variations from specific crosses, and each mapping population can only represent a small number of crossing-over events (Liu and Zeng, 2000). The narrow genetic basis associated with individual crosses and populations reduces both phenotypic and genotypic diversity. One way to solve these problems is to construct a consensus map as the connection across multiple populations.

A consensus genetic map combines genetic information from multiple populations, and therefore provides an effective alternative to improve genome coverage and marker density (Maccaferri et al., 2015; Allen et al., 2017). A higher marker density of the consensus map offers the chance to map more QTLs to narrower intervals and to identify more closely linked markers for the discovery of causal genes and marker-assisted selection (MAS) in breeding. Consensus maps can also be used to validate marker order, characterize genomic diversity, increase the power of genome-wide association studies, and conduct QTL meta-analysis (Cavanagh et al., 2013; Wang et al., 2014; Wingen et al., 2017; Liu et al., 2020).

Some computer tools that can be used for consensus map construction have been developed in the last 20 years, such as BioMercator (Arcade et al., 2004), JoinMap (Van Ooijen, 2006), MergeMap (Wu et al., 2010), MultiPoint (Ronin et al., 2012), and LPmerge (Endelman and Plomion, 2014). Using these tools, consensus maps have been developed for wheat. Somers et al. (2004) reported the first consensus map for wheat based on SSR markers from three doubled haploid (DH) and a recombinant inbred line (RIL) populations. Cavanagh et al. (2013) generated a high-density consensus map from seven populations, consisting of 7,504 single nucleotide polymorphism (SNP) markers. Wang et al. (2014) integrated six bi-parental DH populations to generate a consensus map using 40,267 markers. Liu et al. (2020) developed a consensus map with a total length of 4,080.5 cM containing 47,309 markers based on 21 individual linkage maps and three previously reported consensus maps.

In this study, a consensus genetic map was constructed using three bi-parental populations of RILs in wheat. QTL mapping was then conducted for plant height (PH), spike length (SL), and thousand-kernel weight (TKW) using the constructed consensus map. The mapping results were compared among populations, and with the results obtained using individual maps with the purpose of identifying stable and common QTLs. In addition, a simulation experiment was conducted to demonstrate the advantages of using a consensus map in QTL mapping.

MATERIALS AND METHODS

Plant Materials and Phenotyping Experimental Design

The three recombinant inbred line populations used in this study were derived from crosses Doumai × Shi 4185 (denoted as DS, 275 $F_{2:6}$ RILs), Gaocheng 8901 × Zhoumai 16 (denoted as GZ, 176 $F_{2:6}$ RILs), and Zhou 8425B × Chinese Spring (denoted as ZC, 245 $F_{2:8}$ RILs), which had been previously reported by Wen et al. (2017). Population DS and its parental lines were planted at Shunyi (Beijing, China) and Shijiazhuang (Hebei, China) in 2012–2013, 2013–2014, and 2014–2015 cropping seasons. Population GZ and its parental lines were planted at Anyang (Henan, China) and Suixi (Anhui, China) in 2012–2013 and 2013–2014 cropping seasons. Population ZC and its parental lines were planted at Zhengzhou and Zhoukou (Henan, China) in 2012–2013 and 2013–2014 cropping seasons. Randomized complete block designs with three replications were used in field trials. Each plot had three rows with 1.5 m in length and 0.2 m apart between rows. About 50 seeds were sown in each row. Field management was performed according to local practices.

Plant height was recorded as the average height based on 10 representative plants, measured from the base of the stem to the top of the spike excluding awns at the late grain-filling stage. SL was recorded as the average length of 20 representative spikes in populations DS and GZ, and five representative spikes in population ZC, measured from the base of the spike to the top of the spike excluding awns. TKW was evaluated by weighing three random samples of 500 kernels from each plot after harvest.

Genotyping and Marker Quality Control

Deoxyribonucleic acid was extracted from leaves of 15-day-old seedlings according to the cetyltrimethyl ammonium bromide (CTAB) protocol (Sharp et al., 1988). The populations were genotyped by the 90K wheat Infinium iSelect SNP array (Wang et al., 2014) at CapitalBio Corporation (http://www.capitalbio.com) in Beijing, China. Quality control of the genotypic data has been previously described in Wen et al. (2017), and described here briefly and as follows. First, heterozygous marker types were set as missing values. Then, markers with more than 10% of missing values were deleted. Finally, SNPs with minor allelic frequency lower than 0.3 were filtered out. The three individual linkage maps based on these markers were reported by Wen et al. (2017). SNPs on the three maps were used for consensus map construction. The R package VennDiagram (Chen and Boutros, 2011) was used to demonstrate the SNP numbers common among the three individual maps.

Statistical Analysis for Phenotypic Data

Analysis of variance and calculation of broad-sense heritability (H^2) from phenotypic data were performed using the AOV function in software QTL IciMapping V4.2 (Meng et al., 2015). Pearson correlation coefficients among traits were calculated using mean phenotypic values across environments.

Consensus Genetic Map Construction

First, markers from the three recombinant inbred line populations were grouped according to their chromosome information in individual maps reported by Wen et al. (2017). Markers that were present on the same chromosome in the three individual maps were treated as anchors. Then, an algorithm called combined linkage analysis (CLA, developed by the group of the authors) was used for consensus map construction. To assure the quality of the map, a limited number of markers were removed manually if they caused serious inconsistency in the marker order between the genetic and physical maps, or excessive expansion of the constructed genetic map. The R package LinkageMapView (Ouellette et al., 2018) was used to visualize the constructed consensus map.

Furthermore, four steps were involved in the CLA algorithm: step 1: derive the theoretical recombination frequencies of pairwise markers in each mapping population; step 2: estimate the recombination frequency between two linked markers and sampling variance of the estimated recombination frequency in each population. In addition to RIL populations, CLA is applicable to many other kinds of bi-parental populations, as described in Meng et al. (2015). For some kinds of mapping populations such as DH and RIL, the likelihood equation on recombination frequency has an explicit solution, so the maximum likelihood estimate can be calculated directly. For other kinds of mapping populations such as F_2 and F_3, the maximum likelihood estimate cannot be succinctly given. In this situation, either Newton iteration or the expectation-maximization (EM) algorithm has to be adopted in estimating the recombination frequency (Zhang et al., 2019). Step 3: estimate the combined recombination frequency using the estimates and their sampling variances from individual populations; reciprocal of sampling variance of the estimated recombination frequency is used as the weight of the corresponding population. Weight is set as zero for those populations where the pair-wise recombination frequency cannot be estimated. Step 4: construct the consensus linkage map based on the combined estimates of recombination frequencies between markers; a combination of the nearest-neighbor algorithm and a two-opt algorithm in solving the traveling salesman problem (TSP) was used in the marker ordering (Zhang et al., 2020a).

Comparison of Marker Orders in the Consensus Map, Physical Map, and Individual Genetic Maps

Spearman rank correlation was used to measure the collinearity of marker orders between the different maps, which was calculated by the R Software. Marker order in each chromosome in the consensus map was compared with the physical map order of the respective chromosome. To acquire the physical positions of the markers, sequences of SNPs were used to BLAST (Basic Local Alignment Search Tool) against the wheat genome IWGSC RefSeq v2.0 (https://urgi.versailles.inra.fr/download/iwgsc/IWGSC_RefSeq_Assemblies/v2.0/, International Wheat Genome Sequencing Consortium). The E-value threshold in BLAST was set at 10^{-10}. The markers were filtered out if their alignment lengths were lower than 80% of the query sequence length or the identities were lower than 0.85. If a marker was assigned to multiple chromosomes by BLAST, the position on the same chromosome as the consensus map was used in collinearity analysis. Marker order comparison was also conducted between the consensus map and individual maps, as well as among the three individual maps. For each comparison, only the common markers on two maps were used in the calculation of collinearity.

QTL Mapping Based on the Consensus Map

Quantitative trait locus mapping was conducted in the individual populations using the consensus map. The inclusive composite interval mapping (ICIM) implemented in the BIP function in QTL IciMapping V4.2 (Meng et al., 2015) was applied on the mean phenotypic values across blocks in each environment and best linear unbiased estimation (BLUE) values across multiple environments. Scanning step was set at 0.2 cM. Probabilities of adding and removing variables in stepwise regression were set at 0.001 and 0.002, respectively. Threshold logarithm of odd (LOD) score was set at 2.5, same as the QTL mapping studies on individual maps from the three populations (Gao et al., 2015; Li et al., 2018).

Quantitative trait loci and quantitative trait locus clusters were named with chromosomal locations, considering all populations together. QTLs detected in the same population were considered to be common if the distance between QTL positions was <20 cM in different environments. QTLs detected in different populations were considered to be common if the genetic and physical positions were close enough. In other words, distance in the linkage map was <20 cM in terms of QTL positions, and distance in the physical map was <25 Mb in terms of the minimum physical distances between flanking makers. In individual populations, QTLs are considered to be stable if they are identified in at least half of tested environments. Stable QTLs for different traits were classified into the same cluster if the minimum distance between the QTL confidence intervals was <15 cM. The shiny Circos tool (Yu et al., 2018) was used to visualize QTL positions on the consensus map. Stable QTLs detected with the consensus map in this study were compared with those detected by ICIM using individual maps (Gao et al., 2015; Li et al., 2018), according to physical and genetic positions of the flanking markers.

Genetic Models Used in Simulation

A simulation study was conducted to compare the QTL mapping results from the individual and consensus maps. We assumed that a chromosome has a length of 100 cM and contains 101 evenly distributed markers. Considering that the consensus map always has more markers than each individual map, we assume that the consensus map contained all the 101 markers, but that the individual map only contained half of them, i.e., 51 evenly distributed markers with marker density at 2 cM. Three QTL

TABLE 1 | Mean performance and heritability of plant height (PH), spike length (SL), and thousand-kernel weight (TKW) in the three RIL populations, Doumai × Shi 4185 (DS), Gaocheng 8901 × Zhoumai 16 (GZ), and Zhou 8425B × Chinese Spring (ZC), across multiple environments.

Population	Trait	Parent[a]		RIL population[b]			H^2_by_mean[c]	H^2_by_plot[d]
		P1	P2	Mean	SD	Range		
DS	PH	73.51	73.54	83.89	7.56	64.99–105.09	0.97	0.69
	SL	9.29	8.30	8.77	0.97	6.17–12.10	0.95	0.62
	TKW	50.30	35.35	43.56	4.95	30.52–60.10	0.96	0.75
GZ	PH	94.38	67.77	90.67	15.82	44.23–118.25	0.99	0.91
	SL	8.74	8.97	8.59	0.87	6.60–11.24	0.96	0.72
	TKW	43.83	48.17	46.52	3.81	33.46–55.55	0.91	0.59
ZC	PH	67.12	115.08	101.07	14.02	60.58–125.87	0.95	0.83
	SL	11.50	8.31	10.14	1.15	6.89–13.83	0.91	0.72
	TKW	52.63	29.10	37.12	4.16	26.52–48.83	0.94	0.81

[a] *Best linear unbiased estimation (BLUE) values across multiple environments. In population DS, P1 and P2 refer to Doumai and Shi 4185, respectively. In population GZ, P1 and P2 refer to Gaocheng 8901 and Zhoumai 16, respectively. In population ZC, P1 and P2 refer to Zhou 8425B and Chinese Spring, respectively.*

[b] *Values were based on BLUE across multiple environments.*

[c] *Heritability in broad sense based on replicated means.*

[d] *Heritability in broad sense based on plot level.*

SD, standard deviation.

distribution models were simulated (**Supplementary Table 1**). In model I, a QTL was located at 34.5 cM on the chromosome with an additive effect of 1. In model II, two QTLs were linked in the coupling phase, both with an additive effect of 1. In model III, two QTLs were linked in the repulsion phase with additive effects of −1 and 1, respectively. The broad sense heritability (H^2) was set at three levels, i.e., 0.05, 0.1, and 0.2 for model I, and 0.1, 0.2, and 0.4 for models II and III. One thousand RIL populations, each with a size of 200, were simulated for each model, and each heritability level by the BIP simulation functionality was implemented in QTL Ici Mapping V4.2 (Meng et al., 2015). The consensus map with 101 markers and the predefined QTLs were used to generate the simulated populations. Both the consensus and individual maps were used in QTL mapping. For QTL mapping using individual maps, genotypic data of the 51 markers were used. For QTL mapping using the consensus map, genotypic data of the 51 markers were the same as those in individual maps, but the other 50 markers only present in the consensus map were set as missing values. For the ICIM QTL mapping method on simulated populations, the scanning step was set at 0.1 cM and the threshold LOD score was set at 2.5. Probabilities for entering and removing variables in the stepwise regression were set at 0.001 and 0.002, respectively. QTL detection power was estimated according to a support interval of 5 cM centered at the position of true QTL. If multiple peaks occurred within the support interval, only the highest one was counted. QTLs identified out of the support interval were regarded as false positives (Li et al., 2012). The other parameters were set as default values.

RESULTS

General Information on Both Genotypic and Phenotypic Data

There were 10,986 markers on the linkage map constructed in population DS, 11,819 markers in population GZ, and 14,862 markers in population ZC. Populations DS and GZ shared 4,208 common markers; DS and ZC shared 4,420 common markers; GZ and ZC shared 5,183 common markers; the three populations had 1,880 markers in common (**Supplementary Figure 1**). A total of 25,736 unique markers on the three individual maps were used for consensus map construction.

Phenotypic means and heritability of the three traits are shown in **Table 1** for the three RIL populations across a number of environments. Frequency distributions in different populations and environments are shown in **Supplementary Figure 2** for PH, **Supplementary Figure 3** for SL and **Supplementary Figure 4** for TKW. For PH, Doumai was taller than Shi 4185 in four environments, but shorter in the other two environments in population DS; Gaocheng8901 was always taller than Zhoumai 16 in population GZ; Chinese Spring was always taller than Zhou 8425B in population ZC (**Supplementary Figure 2**). For SL, Doumai was longer than Shi 4185 in four environments, almost equal in one environment, and shorter in the other one environment; Zhoumai16 was longer than Gaocheng 8901 in three environments, and shorter in the other environment; Chinese Spring was always longer than Zhou 8425B (**Supplementary Figure 3**). For TKW, Doumai was always higher than Shi 4185; Zhoumai 16 was always higher than Gaocheng 8901; Zhou 8425B was always higher than Chinese Spring (**Supplementary Figure 4**). The three traits were continuously distributed in the three populations, similar to and typical in most QTL mapping studies. Much wider ranges were observed in the progenies in comparison with their parents, except for TKW in two environments in population ZC (**Supplementary Figures 2–4**). Heritability in the broad sense, based on the replicated means, was quite high for the three traits, ranging from 0.91 to 0.99 (**Table 1**), while heritability based on the plot level ranged from 0.59 to 0.91. Correlation coefficients between traits in the three populations are given in **Supplementary Table 2**. At a significance level of 0.01, PH was positively correlated with TKW in all three populations. SL showed a positive correlation

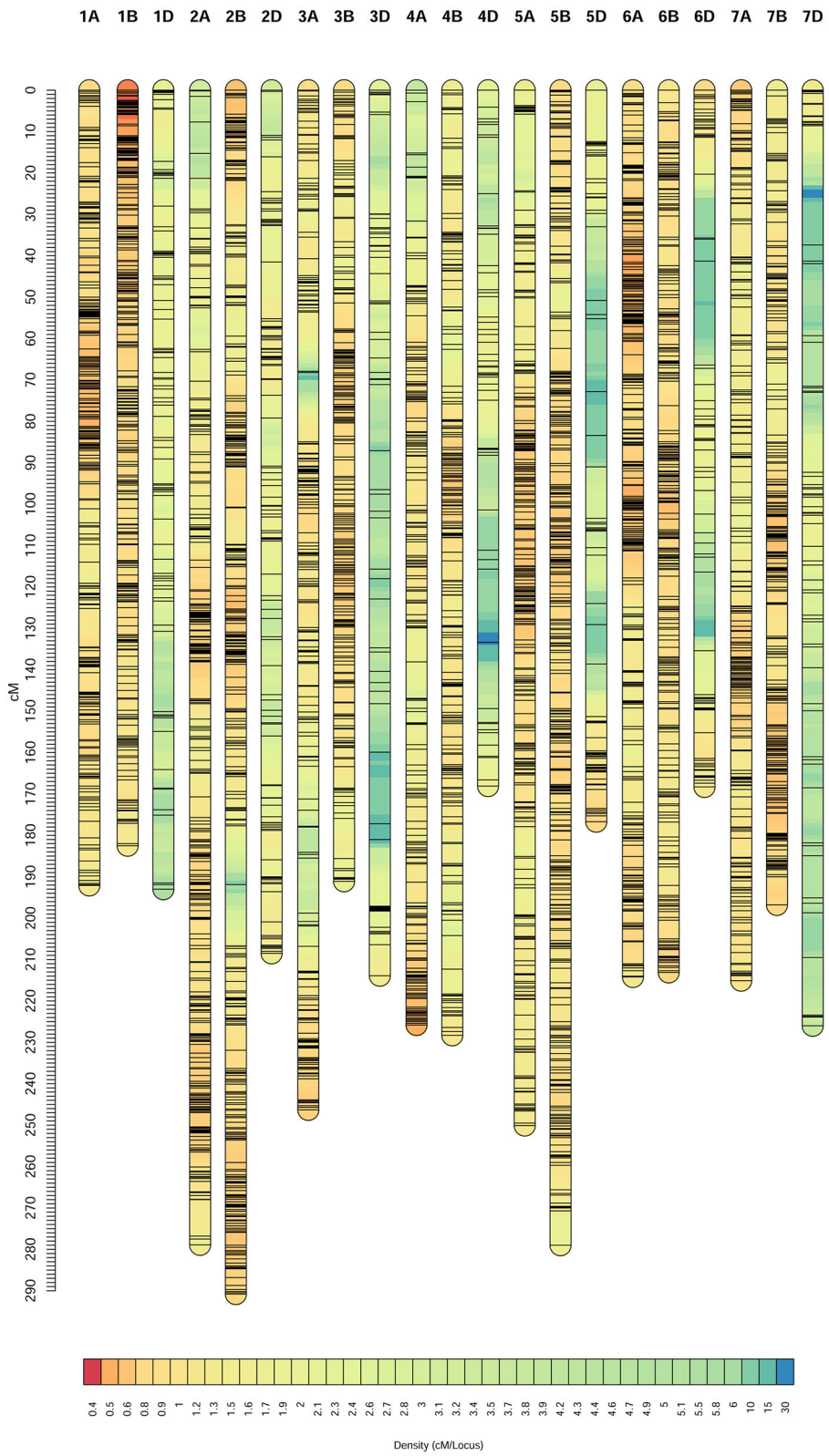

FIGURE 1 | Consensus genetic map constructed from the three recombinant inbred line (RIL) populations, Doumai × Shi 4185 (DS), Gaocheng 8901 × Zhoumai 16 (GZ), and Zhou 8425B × Chinese Spring (ZC).

TABLE 2 | Characteristics of the consensus genetic map constructed from the three RIL populations, DS, GZ, and ZC.

Chromosome	Length (cM)	Marker number	Bin number	Average BD (cM)[a]	Max BD (cM)[b]	Coefficient[c]	Consistent proportion (%)[d]
1A	192.09	1280	240	0.80	9.22	0.99	50.55
1B	182.47	2014	257	0.71	5.72	0.95	51.07
1D	192.98	700	82	2.35	19.42	0.96	57.93
2A	278.86	1685	268	1.04	10.08	0.99	51.99
2B	290.78	2431	324	0.90	14.73	0.87	45.67
2D	208.48	625	108	1.93	13.68	0.99	71.76
3A	246.30	1381	195	1.26	16.68	0.98	55.72
3B	191.12	1974	219	0.87	9.66	0.94	43.30
3D	213.91	294	60	3.57	23.97	0.74	46.04
4A	225.78	1304	193	1.17	12.36	0.98	47.22
4B	228.26	713	201	1.14	8.88	0.98	58.53
4D	168.06	106	53	3.17	17.71	0.95	75.68
5A	250.07	1238	255	0.98	18.50	0.96	62.41
5B	279.03	2475	330	0.85	8.33	0.95	46.34
5D	176.72	298	67	2.64	17.76	0.96	68.11
6A	214.24	1696	289	0.74	7.26	0.96	36.30
6B	213.08	1571	252	0.85	6.04	0.96	53.88
6D	168.33	350	73	2.31	25.08	0.96	54.44
7A	215.12	1701	213	1.01	8.00	0.99	62.07
7B	196.83	1577	236	0.83	7.77	0.98	45.75
7D	226.05	254	64	3.53	28.84	0.99	73.71
Genome							
A	1622.47	10285	1653	0.98	18.50	0.98	52.32
B	1581.57	12755	1819	0.87	14.73	0.98	49.22
D	1354.52	2627	507	2.67	28.84	0.94	63.95
Homeologous groups							
1	567.54	3994	579	0.98	19.42	0.97	53.19
2	778.12	4741	700	1.11	14.73	0.95	56.47
3	651.33	3649	474	1.37	23.97	0.89	48.35
4	622.10	2123	447	1.39	17.71	0.97	60.48
5	705.82	4011	652	1.08	18.50	0.95	58.95
6	595.64	3617	614	0.97	25.08	0.96	48.21
7	637.99	3532	513	1.24	28.84	0.99	60.51
Total	4558.55	25667	3979	1.15	28.84	0.95	55.17

[a] *Average distance between two adjacent bins.*
[b] *Maximum distance between two adjacent bins.*
[c] *Spearman rank correlation coefficient between the consensus map and IWGSC RefSeq v2.0.*
[d] *The proportion of SNPs arranged in the order same with those on the corresponding chromosomes of the physical map.*

with both PH and TKW in population DS. Other correlations were non-significant.

Characteristics of the Constructed Consensus Map

Of the 25,736 unique SNPs on the three individual linkage maps, 25,667 were assigned to the consensus map, resulting in 21 linkage groups corresponding to the 21 chromosomes in hexaploid wheat (**Figure 1**). General information on the consensus map is provided in **Table 2**, and positions of all the markers on both the genetic and physical maps are given in **Supplementary Table 3**. The consensus map spanned 4,558.55 cM in length, and the number of unique map positions (denoted as bins) was equal to 3,979. Lengths of the A, B, and D genomes were 1,622.47, 1,581.57, and 1,354.52 cM, respectively (**Table 2**). Chromosome 4D was the shortest, with a length of 168.06 cM, and had the least number of markers (i.e., 106) and the least number of bins (i.e., 53). Chromosome 2B was the longest with a length of 290.78 cM, and had the second largest number of markers (i.e. 2,431) and the second largest number of bins (i.e., 324). There were 18 gaps longer than 15 cM on the consensus map, 16 of which were located in the D genome (**Supplementary Table 3**). Average distance between adjacent bins was equal to 1.15 cM.

The single nucleotide polymorphism markers (SNPs) number was similar in the A and B genomes, i.e., 10,285 and 12,755 SNPs,

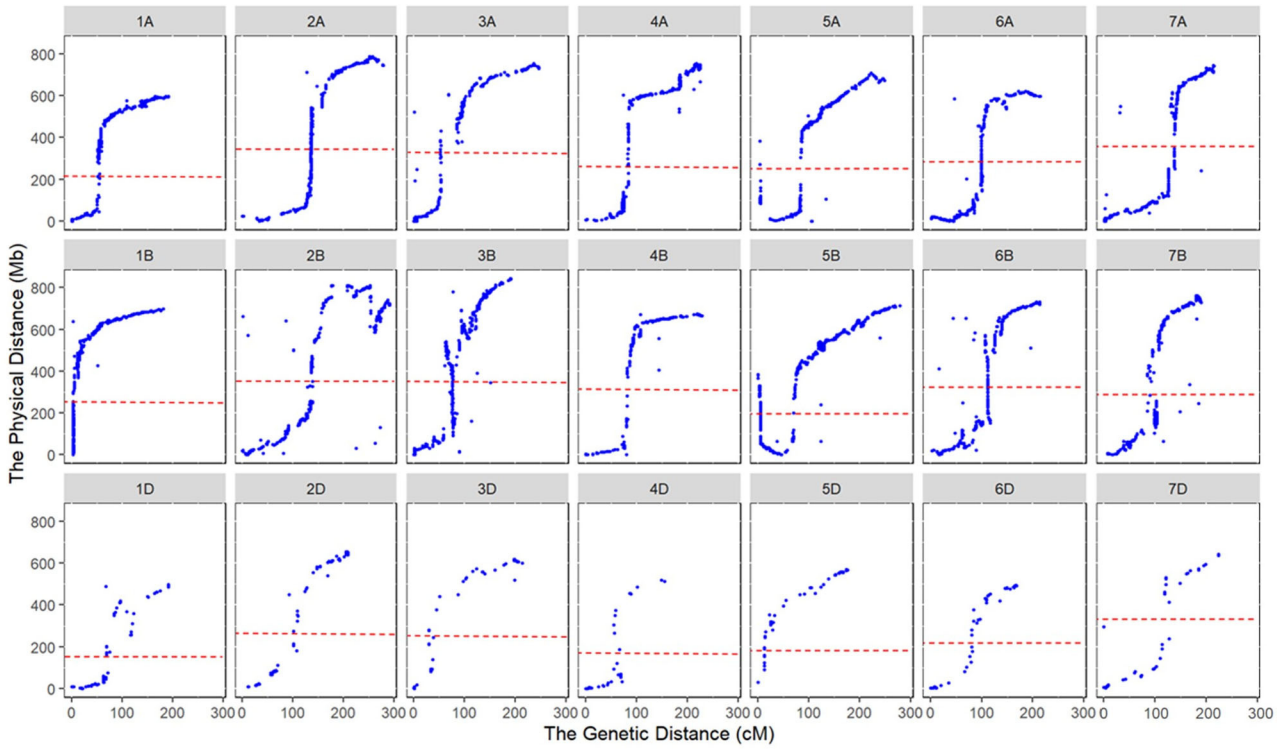

FIGURE 2 | Collinearity of marker orders between the consensus and physical maps. The dotted lines indicate the centromeres of chromosomes.

but the number was much lower in the D genome, i.e., 2,627 SNPs (**Table 2**). In comparison with the A and B genomes, the D genome was shorter and contained much fewer markers and bins, and more gaps, indicating that fewer crossing-over events happened on the D genome, which was also observed in the three individual maps. Although the marker number and bin number in the D genome were significantly lower than those in the A and B genomes, results from BLAST indicated that the constructed consensus map still had nearly complete coverage for chromosomes in the D genome.

Marker orders on the consensus map and physical map had high collinearity, with an average Spearman rank correlation coefficient of 0.95 across the 21 chromosomes (**Table 2**, **Figure 2**). Rank correlation coefficients were higher than 0.94 for all the chromosomes except 2B and 3D. The lower coefficient observed on 3D may be partly due to the much-reduced bin number when many markers were clustered in bins. Collinearity analysis between the consensus and physical maps also revealed that markers in large physical region around the centromeres of chromosomes tended to be clustered in a short genetic interval on consensus genetic map (**Figure 2**), indicating a much stronger recombination suppression occurred around the centromere than did that the distal regions.

Comparison of the Consensus Map With the Three Individual Maps

Wen et al. (2017) reported three linkage maps from three populations constructed with QTL IciMapping V4.0 (Meng et al., 2015), JoinMap 4.0 (Stam, 1993), and MapDisto 1.7 (Lorieux, 2012). Two of them had 21 linkage groups, and one had 31 linkage groups. The consensus map constructed in this study had 21 linkage groups, corresponding to the 21 chromosomes in hexaploid wheat. The marker and bin numbers on the consensus map were 1.73 and 1.15 times higher than the largest marker and bin numbers on the three individual maps. The length of the consensus map was 1.44 times longer than that of the longest individual map. Longer chromosomes on the individual maps also tended to be longer on the consensus map. For example, the two longest chromosomes on the consensus map, i.e., 2B and 5B, ranked first and third in mapping length in each of the three individual maps.

There were 616 markers with inconsistent chromosomes on the individual maps, but the inconsistent chromosomes for each marker were finalized to one unique chromosome on the consensus map (**Supplementary Table 4**). Among these markers, 540 were mapped to single chromosomes that they were located on the individual maps. For example, marker wsnp_Ex_c200_391015 was located on chromosomes 7A and 1A on individual maps of populations GZ and ZC, respectively, which was finalized on chromosome 1A on the consensus map. Forty-nine markers were mapped to one of the homeologous chromosomes. For example, marker Tdurum_contig28665_150 was located on chromosomes 1D, 1D, and 2A in populations DS, GZ, and ZC, respectively, and was finalized on chromosome 1A, a homeologous chromosome of 1D. Twenty-seven markers were mapped to neither the same chromosome nor homeologous

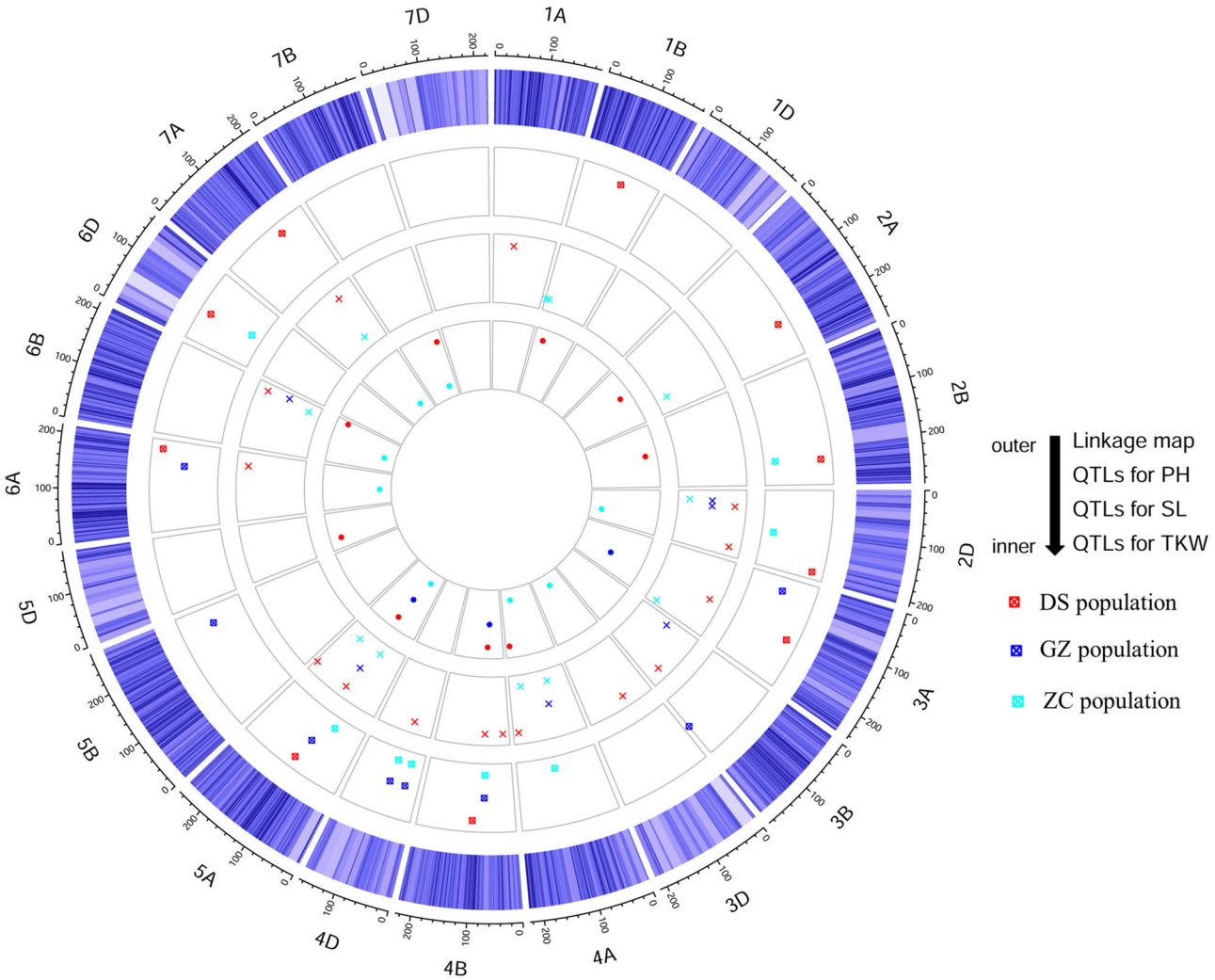

FIGURE 3 | Schematic representation of stable quantitative trait loci (QTLs) for plant height (PH), spike length (SL), and thousand-kernel weight (TKW) detected in three RIL populations, DS, GZ, and ZC from the consensus map.

chromosomes. For example, marker tplb0024a09_2369 was located on chromosomes 7D and 4A in populations DS and ZC, respectively, and was finalized on chromosome 2B on the consensus map (**Supplementary Table 4**).

The markers showed high collinearity across chromosomes between the consensus and individual maps, and the average Spearman rank correlation coefficient was similar to those between the individual maps (**Supplementary Table 5**). Fewer inconsistencies in orders between the consensus and individual maps were observed for closely linked markers.

QTLs for PH Detected From the Consensus Map and Comparison With Those From Individual Maps

Using the consensus map, a total of 40 QTLs were detected for PH (**Supplementary Table 6**), among which 10, 8, and 8 were stable in populations DS, GZ, and ZC, respectively (**Figure 3**, **Table 3**). Five QTLs were identified in two populations, i.e., *qPH-2B-2*, *qPH-4B-1*, *qPH-4D-1*, *qPH-4D-2*, and *qPH-5A-2*. *qPH-2B-2* were repeatedly detected in populations DS and ZC with LOD scores in the range of 3.62 to 22.98, explaining 1.63–8.05% of the phenotypic variance (PVE). *qPH-5A-2* was repeatedly detected in populations DS and GZ, with LOD scores ranging from 3.90 to 15.44, and PVE values ranging from 2.58 to 9.63%. *qPH-4B-1*, *qPH-4D-1*, and *qPH-4D-2* were repeatedly identified in populations GZ and ZC, taking the top three ranks in both populations by average LOD score, PVE value, and additive effect across environments. *qPH-4B-1* was mapped on chromosome 4B at the interval of 34.98–49.79 Mb on physical map with LOD scores ranging from 6.31 to 43.49, and PVE values ranging from 8.14 to 30.85%. *qPH-4D-1* was mapped on chromosome 4D at the interval of 14.14–17.01 Mb with LOD scores ranging from 6.54 and 17.10, and PVE values ranging from 8.06 to 16.48%.

TABLE 3 | Stable quantitative trait loci (QTLs) identified for PH in the three RIL populations, DS, GZ, and ZC using the consensus map.

QTL	Pop	Environments	Position (cM)	LOD	PVE (%)	Add
qPH-1B-2	DS	E2/E3/E4/E5/B	95.60–100.20	3.26–38.07	1.18–15.43	1.37 to 5.42
qPH-2A-2	DS	E1/E3/E4/E6/B	192.00–195.20	2.79–11.05	1.01–5.32	−2.28 to −1.15
qPH-2B-2	DS	E1/E2/E4/E5/B	224.40–239.60	3.62–22.98	1.63–8.05	−3.92 to −1.26
	ZC	E11/E13/E14/B	225.80–227.80	3.89–5.52	2.33–4.30	−3.20 to −2.31
qPH-2D-1	ZC	E13/E14/B	113.80–114.40	2.89–3.73	1.49–3.08	2.01 to 2.77
qPH-2D-3	DS	E1/E2/E3/E4/B	190.20–193.60	2.58–17.02	0.75–7.21	1.21 to 3.22
qPH-3A-1	GZ	E7/E8/E9/E10/B	34.20–36.80	2.85–4.00	3.48–5.41	2.74 to 4.47
qPH-3A-2	DS	E1/E2/E3/E5/B	130.40–148.60	16.23–43.54	6.67–28.31	−5.25 to −3.25
qPH-3B-2	GZ	E7/E10/B	190.60–190.60	2.85–3.33	3.53–4.06	3.08 to 3.77
qPH-4A-2	ZC	E11/E13/E14/B	114.60–132.80	3.27–4.43	1.33–4.47	2.36 to 2.60
qPH-4B-1	GZ	E7/E8/E9/E10/B	75.00–75.00	6.31–9.52	8.14–10.82	−5.72 to −3.92
	ZC	E11/E12/E13/E14/B	74.20–74.40	16.87–43.49	17.92–30.85	5.37 to 11.02
qPH-4B-2	DS	E2/E3/E4/E5/E6/B	100.40–102.40	3.12–5.91	1.64–2.54	−1.89 to −1.34
qPH-4D-1	GZ	E7/E8/E9/E10/B	33.20–36.00	6.68–8.86	8.06–13.62	4.71 to 6.87
	ZC	E11/E12/E13/E14/B	33.40–34.60	6.54–17.10	8.64–16.48	3.90 to 5.83
qPH-4D-2	GZ	E7/E8/E9/E10/B	73.20–73.80	6.20–7.83	8.39–10.90	4.57 to 5.58
	ZC	E11/E12/E13/E14/B	70.80–70.80	4.09–15.72	5.03–12.03	3.00 to 5.54
qPH-5A-1	ZC	E11/E12/B	76.40–86.20	2.60–3.72	2.07–4.60	1.87 to 2.84
qPH-5A-2	DS	E1/E2/E3/E4/E5	120.60–125.40	8.34–15.44	2.58–9.63	2.24 to 3.26
	GZ	E7/E8/E9/E10/B	113.20–135.60	3.90–5.52	4.79–6.87	−4.89 to −3.09
qPH-5B	GZ	E7/E8/E9/E10/B	234.80–237.40	3.60–4.46	4.09–5.54	2.91 to 4.36
qPH-6A-1	GZ	E7/E9/B	154.40–157.00	2.81–3.88	3.36–4.04	−3.35 to −3.03
qPH-6A-2	DS	E1/E3/B	192.60–192.60	4.10–9.47	1.45–4.48	−2.09 to −1.26
qPH-6D-1	DS	E1/E2/E4/E6/B	71.40–76.20	3.21–6.98	1.33–2.46	1.30 to 2.00
qPH-6D-2	ZC	E11/E13/E14/B	84.60–84.60	2.75–5.70	1.89–3.38	1.94 to 2.88
qPH-7A	DS	E1/E2/E3/E5/B	142.40–145.00	4.84–6.74	1.12–3.11	1.44 to 1.89

Pop, population; LOD, logarithm of odd; PVE, percentage of phenotypic variance explained; Add, additive effect; E1, 2012–2013 Beijing; E2, 2012–2013 Shijiazhuang; E3, 2013–2014 Beijing; E4, 2013–2014 Shijiazhuang; E5, 2014–2015 Beijing; E6, 2014–2015 Shijiazhuang; E7, 2012–2013 Anyang; E8, 2012–2013 Suixi; E9, 2013–2014 Anyang; E10, 2013–2014 Suixi; E11, Zhoukou2013; E12, Zhengzhou2013; E13, Zhoukou2014; E14, Zhengzhou2014; B, best linear unbiased estimation.

qPH-4D-2 was mapped on chromosome 4D at the interval of 32.97–65.01 Mb having LOD scores ranging from 4.09 to 15.72 and PVE values ranging from 5.03 to 12.03%. When the length of the confidence interval was set at 25 Mb, *qPH-4B-1* and *qPH-4D-1* were, respectively, coincident with dwarfing genes *Rht-B1* located at 33.61 Mb on 4B and *Rht-D1* located at 19.19 Mb on 4D (IWGSC RefSeq v2.0).

Quantitative trait locus mapping using the individual maps identified a total of 19 stable QTLs in the three populations, nine in population DS, and five each in populations GZ and ZC (Gao et al., 2015; Li et al., 2018). Sixteen of them were detected using the consensus map; Fifteen of which were stable across environments (**Supplementary Table 7**, **Supplementary Figure 5**). *qPH-2B-2* and *qPH-5A-2* were detected only in one population with the individual maps, but in two populations with the consensus map (**Table 3**, **Supplementary Table 7**), indicating the reliability of the two QTLs. With the consensus map, eight other stable QTLs were identified for PH, i.e., *qPH-2D-1*, *qPH-2D-3*, *qPH-3B-2*, *qPH-4D-2*, *qPH-6A-1*, *qPH-6A-2*, *qPH-6D-2*, and *qPH-7A*, three in population DS, two each in populations GZ and ZC, and one in populations GZ and ZC.

QTLs for SL Detected From the Consensus Map and Comparison With Those From the Individual Maps

Using the consensus map, a total of 54 QTLs were detected for SL (**Supplementary Table 6**), among which 15, 6, and 11 were stable in populations DS, GZ, and ZC, respectively (**Figure 3**, **Table 4**). *qSL-2D-1* was repeatedly identified in populations GZ and ZC with LOD scores ranging from 2.67 to 20.91, and PVE values ranging from 2.85 to 31.06%. *qSL-2D-2* was repeatedly detected in populations DS and GZ with LOD scores ranging from 3.20 to 6.40, and PVE values ranging from 1.60 to 6.68%. *qSL-5A-2* was repeatedly identified in populations DS and GZ with LOD scores ranging from 3.31 to 13.93, and PVE values ranging from 1.87 to 7.13%. *qSL-6B-4* was repeatedly detected in the three populations and mapped at chromosome 6B in the interval of 705.19–707.59 Mb on physical map, accounting for 3.36–21.30% of the phenotypic variance.

TABLE 4 | Stable QTLs identified for SL in the three RIL populations, DS, GZ, and ZC using the consensus map.

QTL	Pop	Environment	Position (cM)	LOD	PVE (%)	Add
qSL-1A-1	DS	E1/E2/E4/E6/B	53.60–75.80	2.96–6.05	1.47–4.43	−0.22 to −0.14
qSL-1B-1	ZC	E13/E14	2.60–3.80	4.53–6.20	4.13–5.67	−0.35 to −0.28
qSL-1B-2	ZC	E11/B	16.20–16.20	3.33–3.37	2.90–3.88	−0.22 to −0.29
qSL-2A-1	ZC	E11/E13/E14/B	210.80–228.60	2.66–6.55	2.34–5.72	−0.33 to −0.20
qSL-2D-1	GZ	E7/E8/E9/E10/B	32.00–43.20	11.74–20.91	9.84–31.06	0.36 to 0.59
	ZC	E11/E12/E13/E14/B	29.20–41.60	2.67–10.18	2.85–8.91	−0.41 to −0.24
qSL-2D-2	DS	E3/E5/B	50.60–55.00	3.20–5.97	1.60–3.75	−0.18 to −0.14
	GZ	E9/E10/B	55.80–57.00	3.97–6.40	2.82–6.68	0.21 to 0.27
qSL-2D-3	DS	E1/E2/E3/E4/E5/B	168.80–192.20	3.35–13.66	2.40–6.17	0.17 to 0.29
qSL-3A-4	DS	E2/E5/B	130.40–139.40	3.71–6.00	1.62–2.45	−0.15 to −0.15
qSL-3A-5	ZC	E12/E14/B	221.00–238.00	2.76–3.80	2.51–3.06	−0.22 to −0.22
qSL-3B-2	GZ	E7/E10	23.60–24.80	3.47–3.80	3.55–4.88	−0.23 to −0.20
qSL-3B-5	DS	E2/E4/E6/B	147.60–148.00	4.07–7.80	1.69–5.75	0.16 to 0.26
qSL-3D-2	DS	E1/E2/E5/B	85.20–87.00	4.05–6.81	2.50–2.72	−0.20 to −0.16
qSL-4A-1	ZC	E11/E12/E13/E14/B	73.00–83.20	3.07–12.13	3.27–11.74	−0.45 to −0.26
qSL-4A-2	GZ	E8/E9/E10/B	85.00–103.60	4.20–6.29	2.78–7.71	−0.26 to −0.22
qSL-4A-3	ZC	E11/E12/E13/E14/B	178.20–187.00	6.25–12.99	6.42–11.77	0.34 to 0.47
qSL-4A-4	DS	E1/E2/E5/B	203.40–215.20	5.00–5.90	2.01–3.34	0.14 to 0.23
qSL-4B-1	DS	E1/E2/B	21.00–21.40	4.50–7.41	1.77–3.25	0.14 to 0.23
qSL-4B-2	DS	E1/E2/E3/E4/E5/E6/B	75.00–80.00	4.90–31.94	3.60–16.99	0.21 to 0.45
qSL-4D	DS	E1/E2/E5/B	56.20–66.60	3.54–16.74	1.55–7.79	0.12 to 0.33
qSL-5A-1	ZC	E13/E14	89.80–93.60	2.92–3.83	2.60–3.13	0.21 to 0.24
qSL-5A-2	DS	E1/E2/E5/E6/B	124.20–124.60	4.28–13.93	1.87–7.13	0.15 to 0.34
	GZ	E7/E8	122.60–123.40	3.31–4.39	3.67–5.45	−0.21 to −0.20
qSL-5A-3	ZC	E11/E13/E14/B	190.20–191.00	6.80–11.15	6.25–9.96	0.33 to 0.43
qSL-5A-4	DS	E1/E3/E4	241.60–245.20	2.57–3.86	1.58–3.47	−0.17 to −0.15
qSL-6A-1	DS	E2/E3/E5/E6/B	165.80–177.60	4.37–9.06	2.27–4.17	0.17 to 0.21
qSL-6B-4	DS	E2/E3/E5/E6/B	180.80–198.60	8.93–35.46	4.00–21.30	−0.50 to −0.23
	GZ	E7/E9/E10/B	194.20–195.40	4.56–20.58	4.86–10.96	−0.43 to −0.23
	ZC	E13/E14	178.60–180.40	3.74–3.79	3.36–3.36	−0.26 to −0.25
qSL-7A-2	ZC	E13/E14	137.20–137.20	4.51–5.95	3.77–5.50	−0.30 to −0.27
qSL-7A-3	DS	E2/E5/B	152.00–152.20	4.15–17.49	1.62–8.52	0.13 to 0.32

Pop, population; LOD, logarithm of odd; PVE, percentage of phenotypic variance explained; Add, additive effect; E1, 2012–2013 Beijing; E2, 2012–2013 Shijiazhuang; E3, 2013–2014 Beijing; E4, 2013–2014 Shijiazhuang; E5, 2014–2015 Beijing; E6, 2014–2015 Shijiazhuang; E7, 2012–2013 Anyang; E8, 2012–2013 Suixi; E9, 2013–2014 Anyang; E10, 2013–2014 Suixi; E11, Zhoukou2013; E12, Zhengzhou2013; E13, Zhoukou2014; E14, Zhengzhou2014; B, best linear unbiased estimation.

In previous studies, QTL mapping using individual maps identified six, six, and nine stable QTLs in populations DS, GZ, and ZC, respectively (Gao et al., 2015; Li et al., 2018). This study detected all of them except *QSL.caas-5AL* in population ZC (**Supplementary Table 7**, **Supplementary Figure 6**). However, according to the linkage map constructed by Wen et al. (2017) for population ZC and the BLAST result, *QSL.caas-5AL* and *QSL.caas-5AL.1* tended to be the same. For the remaining 20 QTLs, 19 with stable effects were detected using the consensus map. *qSL-2D-1*, *qSL-2D-2,* and *qSL-5A-2* were detected only in one population using the individual maps, but all of them were detected in two populations using the consensus map (**Table 4**, **Supplementary Table 7**). With the consensus map, 10 other stable QTLs were identified for SL, i.e., *qSL-3A-4*, *qSL-3A-5*, *qSL-3B-5*, *qSL-4A-4*, *qSL-4B-1*, *qSL-4B-2*, *qSL-4D*, *qSL-5A-1*, *qSL-5A-4*, and *qSL-7A-3*, eight for population DS and two for population ZC.

QTLs for TKW Detected From the Consensus Map and Comparison With Those From the Individual Maps

Using the consensus map, a total of 53 QTLs were detected for TKW (**Supplementary Table 6**), among which nine, three, and eight were stable in populations DS, GZ, and ZC, respectively (**Figure 3**, **Table 5**). *qTKW-4B-2* was repeatedly identified in populations DS and GZ with LOD scores ranging from 3.08 to 49.22, explaining 7.57–36.51% of the phenotypic variance. *qTKW-4B-2* had the largest LOD score, PVE and additive effect across environments in population DS. This QTL was co-localized with *qPH-4B-1*, corresponding to the dwarfing

TABLE 5 | Stable QTLs identified for TKW in the three RIL populations, DS, GZ, and ZC using the consensus map.

QTL	Pop	Environment	Position (cM)	LOD	PVE (%)	Add
qTKW-1B-2	DS	E1/E2/E3/E6/B	44.40–45.00	4.70–45.92	2.37–22.43	0.93 to 3.72
qTKW-2A-2	DS	E2/E3/E4	125.00–125.00	2.57–7.27	1.32–5.90	0.70 to 1.15
qTKW-2B-2	DS	E4/E5/B	138.80–141.40	2.92–6.43	1.25–3.81	0.73 to 1.32
qTKW-2D-2	ZC	E11/E12/E13/E14/B	131.40–132.60	3.91–14.69	4.94–12.32	−1.64 to −1.00
qTKW-3A-2	GZ	E9/B	148.40–148.60	4.05–5.80	6.56–11.45	1.00 to 1.26
qTKW-3D	ZC	E11/E12/B	101.40–101.80	3.79–6.44	4.85–6.86	1.00 to 1.18
qTKW-4A-1	ZC	E11/E12/B	167.20–167.20	3.94–5.54	4.08–6.72	−1.17 to −0.91
qTKW-4A-2	DS	E1/E2/E6/B	200.00–213.40	2.51–4.57	0.89–3.47	−0.87 to −0.73
qTKW-4B-2	DS	E1/E2/E3/E4/E5/E6/B	75.00–76.00	13.28–49.22	8.81–36.51	1.60 to 3.78
	GZ	E7/E10/B	61.20–75.00	3.08–4.61	7.57–10.68	−1.43 to −1.09
qTKW-5A-1	ZC	E13/E14	65.40–84.60	3.92–4.67	3.02–3.53	−0.91 to −0.82
qTKW-5A-3	DS	E1/E2/E5/E6/B	116.40–124.40	3.83–7.46	2.53–5.05	1.03 to 1.27
qTKW-5A-4	GZ	E7/E8/E9/E10/B	106.20–112.00	2.69–4.68	6.36–8.64	−1.28 to −0.94
qTKW-5D-1	DS	E1/E2/E3/E4/E5/E6/B	49.20–54.80	6.23–11.44	2.38–7.58	1.12 to 1.66
qTKW-6A-3	ZC	E11/E12/E13/E14/B	98.60–102.20	5.67–16.80	7.37–14.28	−1.84 to −1.23
qTKW-6B-3	ZC	E11/E12/B	90.40–93.20	3.66–4.91	3.85–5.09	−1.01 to −0.82
qTKW-6B-5	DS	E1/E3/E5/E6/B	197.40–198.60	3.62–6.22	1.21–3.36	−1.17 to −0.86
qTKW-7A-1	ZC	E11/E12/E13/E14/B	140.40–141.20	2.72–4.77	2.04–5.71	−1.09 to −0.67
qTKW-7B-3	ZC	E13/E14	146.20–146.20	4.26–5.27	3.22–4.03	−0.98 to −0.85
qTKW-7B-4	DS	E1/E4/E5	168.80–171.40	5.79–18.41	3.10–12.11	1.06 to 2.36

Pop, population; LOD, logarithm of odd; PVE, percentage of phenotypic variance explained; Add, additive effect; E1, 2012–2013 Beijing; E2, 2012–2013 Shijiazhuang; E3, 2013–2014 Beijing; E4, 2013–2014 Shijiazhuang; E5, 2014–2015 Beijing; E6, 2014–2015 Shijiazhuang; E7, 2012–2013 Anyang; E8, 2012–2013 Suixi; E9, 2013–2014 Anyang; E10, 2013–2014 Suixi; E11, Zhoukou2013; E12, Zhengzhou2013; E13, Zhoukou2014; E14, Zhengzhou2014; B, best linear unbiased estimation.

gene *Rht-B1*. All stable QTLs detected with the individual maps were also stable when detected with the consensus map (**Supplementary Table 7**, **Supplementary Figure 7**). There were other three stable TKW QTLs identified using the consensus map, i.e., *qTKW-1B-2*, *qTKW-2D-2*, and *qTKW-6B-3*. *qTKW-1B-2* was mapped on chromosome 1B at the interval of 588.36–591.14 Mb on the physical map, with LOD scores ranging from 4.70 to 45.92, and PVE values ranging from 2.37 to 22.43% in population DS. *qTKW-2D-2* was mapped on chromosome 2D at the interval of 523.15–555.13 Mb with LOD scores ranging from 3.91 to 14.69, and PVE values ranging from 4.94 to 12.32% in population ZC. *qTKW-6B-3* was mapped on chromosome 6B at the interval of 157.21–162.58 Mb with LOD scores varying from 3.66 to 4.91, and PVE values varying from 3.85 to 5.09% in population ZC.

QTL Clusters for the Three Traits

As far as the stable QTLs across environments were concerned, 11 QTL clusters were identified and distributed on nine chromosomes (**Supplementary Table 8**), six of which affected two traits (i.e., *qClu-2D*, *qClu-4A-1*, *qClu-4A-2*, *qClu-4D*, *qClu-6A*, and *qClu-6B*), and five affected all the three traits (i.e., *qClu-3A-1*, *qClu-4B*, *qClu-5A-1*, *qClu-5A-2*, and *qClu-7A*). Eight clusters affected traits PH and SL. Among them, three clusters contained both PH and SL QTLs in population DS; one cluster contained both PH and SL QTLs in population ZC, and one cluster contained the closely linked PH and SL QTLs in populations DS and GZ. Each of the five clusters either increased or decreased both traits simultaneously. Genomic regions containing the stable QTLs for the three traits were located on chromosomes 3A, 4B, 5A, and 7A. The cluster on 4B was close to the Green Revolution gene *Rht-B1*. In cluster *qClu-5A-1*, QTLs affecting the three traits were consistently identified in populations DS and GZ, either increasing or decreasing the three traits simultaneously.

Potential Applications of the Detected QTLs in Wheat Breeding

To explore the potential applications of the detected QTLs in wheat breeding, QTL genotypes and genotypic values of each RIL in the three populations were predicted on the three traits with stable QTLs identified using BLUE values across environments (**Supplementary Tables 9–11**). For convenience, for the two alleles at each QTL, one is called positive and the other one is called negative. Parental sources of the two alleles can be determined from the sign of the estimated additive effect of the QTL. Due to the varied objectives on different traits in breeding, it should be noted that the positive allele is not always favored and that the negative allele is not always un-favored. For PH, nine, eight, and eight stable QTLs were used for prediction in populations DS, GZ, and ZC, respectively. The 10 highest RILs possessed at least eight, seven, and seven positive alleles in the three populations, respectively, whereas the 10 lowest RILs had no more than two positive alleles (**Supplementary Table 9**). For SL, 14, 7, and 4 stable QTLs were used for prediction. The 10 highest RILs possessed at least nine, seven, and four positive

alleles in the three populations, whereas the 10 lowest RILs had no more than four positive alleles in population DS, no positive allele in population GZ, and no more than 1 positive allele in population ZC (**Supplementary Table 10**). For TKW, seven, three, and six stable QTLs were used for prediction. The 10 highest RILs possessed at least six, three, and five positive alleles in the three populations, whereas the 10 lowest RILs had no more than 1 positive allele (**Supplementary Table 11**). RILs with the highest predicted genotypic values always had all the positive alleles for PH and TKW in the three mapping populations, and had all the positive alleles for SL in populations GZ and ZC. RILs with the lowest predicted genotypic values always had all the negative alleles for PH and SL in populations GZ and ZC, and had all the negative alleles for TKW in all the three mapping populations. For PH, all the 10 lowest RILs in population GZ and the 9 lowest RILs in population ZC contained the negative alleles at *qPH-4B-1* and *qPH-4D-1*, corresponding to genes *Rht-B1* and *Rht-D1*. For SL, *qSL-6B-4* was repeatedly identified in populations DS and GZ. Eighteen out of the 20 highest RILs in population DS and 36 highest RILs in population GZ possessed the positive allele at *qSL-6B-4*, while 17 out of the 20 lowest RILs in population DS and 38 lowest RILs in population GZ possessed the negative allele at *qSL-6B-4*. For TKW, *qTKW-4B-2* was consistently identified in populations DS and GZ. The 12.36% highest RILs in population DS and the 31.25% highest RILs in population GZ carried the positive allele at *qTKW-4B-2*, while the 17.45% lowest RILs in population DS and the 21.02% lowest RILs in population GZ carried the negative allele at *qTKW-4B-2*. Mean observed and predicted values of RILs having the positive allele at *qTKW-4B-2* were equal to 45.13 and 45.24 in population DS, and 47.11 and 47.95 in population GZ. In contrast, the observed means of RILs having the negative allele were equal to 42.68 and 40.4 in population DS, and 45.59 and 45.65 in population GZ.

Recombinant inbred lines with the predicted genotypic values on PH, SL, and TKW can serve for the choice of target genotypes meeting different breeding objectives, such as wheat cultivars with medium plant height, large spike length, and medium to high kernel weight. Given one target genotype, the predicted allelic combination of RILs can serve for the prediction of cross performance and the selection of suitable parental lines through simulation or other genomic prediction approaches (Yao et al., 2018).

QTL Mapping in Simulated Populations

In 1,000 simulated populations, the estimated QTL positions and effects using the individual and consensus maps are shown in **Table 6**. With the increase in heritability, QTL detection powers were increased and the false discovery rate (FDR) was decreased in the three models using either the individual or consensus maps. Approximately unbiased estimation of QTL positions and effects was obtained for each defined model and heritability level. The confidence intervals of QTLs detected from the consensus map were much narrower, and the associated standard errors were much smaller than those from individual maps. Detection power was much lower for QTLs in linkage models II and III than that in the unlinked model I at the same heritability levels for both the individual and consensus maps. FDR was much higher in models II and III than in model I, indicating the complexity and difficulty in dissecting linked QTLs in genetic studies.

DISCUSSION

Computer Tools in Consensus Map Construction

Two strategies have been adopted for consensus map construction in previous studies (Endelman and Plomion, 2014). The first one is based on the raw data of multiple mapping populations, and has been implemented in software MultiPoint (Ronin et al., 2012) and JoinMap (Van Ooijen, 2006). The second one is based on individual linkage maps previously constructed, and has been implemented in software BioMercator (Arcade et al., 2004), MergeMap (Wu et al., 2010), LPmerge (Endelman and Plomion, 2014), and QTL IciMapping (Meng et al., 2015). The first strategy is usually time-consuming when dealing with a large number of markers (Wu et al., 2010), which has drastically restricted the use of a large number of markers in the consensus map. The second strategy highly depends on the quality of individual maps and sometimes may result in maps with unreasonable length (Cavanagh et al., 2013; Wang et al., 2014; Wingen et al., 2017).

With the development of high-throughput sequencing technology, markers that can be used in genotyping mapping populations are growing rapidly. A large amount of markers brings a great challenge to consensus map construction, especially when raw genotypic data are used. The two raw data-based software packages mentioned above cannot deal with such a large number of markers used in this study. For example, both packages cannot generate a consensus map for chromosome 5B, which harbored 929, 1,406, and 1,508 SNPs in populations DS, GZ, and ZC, respectively. Map-based method only utilizes marker distances between adjacent markers, which may result in an inaccurate estimation of recombination frequency between markers especially when the order of markers changes on the consensus map. The CLA algorithm is a raw data-based method used in this study to deal with a large amount of markers. The combined recombination frequency between any pair of markers was calculated from the estimates in individual mapping populations. The estimated recombination frequencies are recorded in computer memory. Therefore, time can be greatly saved in computing.

Quality of the Consensus Map

The great number of markers and bins contained in the consensus map provided higher saturation of markers and better genome coverage, and expanded the length of the map. Previous studies have shown that increased recombination events and map resolution with an increased number of markers and density could contribute to longer map length (Ferreira et al., 2006; Wingen et al., 2017). The longer map length may also suffer from chromosomal structure differences in different mapping populations and the ordering algorithm used. Compared with the A and B genomes, the D genome had fewer unique markers, larger gaps, and shorter map length, which have been previously

TABLE 6 | Quantitative trait locus mapping results from 1,000 simulations using the individual and consensus maps in the three genetic models.

Model	H^{2a}	Map	QTL	Pos. ± SE (cM)[b]	Add ± SE[c]	CIL ± SE[d]	LOD ± SE[e]	Power (%)	FDR (%)[f]
I	0.05	Ind.	QTL1	34.32 ± 1.34	1.33 ± 0.22	3.55 ± 0.76	4.03 ± 1.41	31.5	44.44
		Cons.	QTL1	34.42 ± 1.36	1.32 ± 0.19	1.89 ± 0.35	3.97 ± 1.17	33.6	43.72
	0.1	Ind.	QTL1	34.42 ± 1.21	1.08 ± 0.21	3.30 ± 0.85	5.41 ± 2.10	71.1	25.16
		Cons.	QTL1	34.55 ± 1.27	1.09 ± 0.25	1.83 ± 0.41	5.62 ± 2.96	69.9	27.71
	0.2	Ind.	QTL1	34.40 ± 0.94	1.01 ± 0.16	2.85 ± 0.74	10.03 ± 3.02	89.3	12.45
		Cons.	QTL1	34.46 ± 1.12	1.01 ± 0.16	1.70 ± 0.41	10.07 ± 2.96	90.2	11.57
II	0.1	Ind.	QTL1	26.62 ± 1.37	2.03 ± 0.35	3.18 ± 1.05	5.61 ± 1.92	31.0	31.26
			QTL2	34.08 ± 1.32	1.99 ± 0.32	3.27 ± 0.97	5.27 ± 1.58	36.3	
		Cons.	QTL1	26.64 ± 1.29	2.07 ± 0.67	1.85 ± 0.37	6.00 ± 4.80	28.3	34.9
			QTL2	34.18 ± 1.37	1.98 ± 0.33	1.82 ± 0.42	5.33 ± 1.68	34.2	
	0.2	Ind.	QTL1	26.80 ± 1.21	1.88 ± 0.30	2.84 ± 0.83	9.93 ± 2.92	40.5	24.91
			QTL2	33.78 ± 1.11	1.84 ± 0.28	2.83 ± 1.02	9.54 ± 2.53	39.1	
		Cons.	QTL1	26.80 ± 1.21	1.87 ± 0.27	1.71 ± 0.42	9.96 ± 2.68	38.1	27.06
			QTL2	33.91 ± 1.20	1.83 ± 0.27	1.73 ± 0.46	9.52 ± 2.52	37.1	
	0.4	Ind.	QTL1	26.62 ± 1.09	1.33 ± 0.40	2.68 ± 0.84	13.06 ± 6.24	57.1	17.34
			QTL2	34.09 ± 1.03	1.39 ± 0.40	2.64 ± 0.86	13.99 ± 6.51	63.0	
		Cons.	QTL1	26.66 ± 1.130	1.36 ± 0.39	1.63 ± 0.40	13.67 ± 6.29	55.1	19.02
			QTL2	34.18 ± 1.180	1.39 ± 0.43	1.66 ± 0.39	14.19 ± 7.01	59.0	
III	0.1	Ind.	QTL1	26.11 ± 1.10	−1.04 ± 0.32	2.78 ± 0.95	9.51 ± 5.34	7.4	34.91
			QTL2	34.72 ± 1.11	1.04 ± 0.28	2.92 ± 0.87	9.26 ± 4.13	7.7	
		Cons.	QTL1	26.25 ± 1.04	−1.08 ± 0.35	1.68 ± 0.39	10.40 ± 6.04	7.3	40.93
			QTL2	34.82 ± 1.18	1.07 ± 0.31	1.71 ± 0.35	10.03 ± 4.85	8.0	
	0.2	Ind.	QTL1	26.19 ± 0.75	−0.98 ± 0.21	2.43 ± 0.63	16.07 ± 5.59	26.3	15.18
			QTL2	34.58 ± 0.77	0.99 ± 0.19	2.37 ± 0.67	16.24 ± 5.18	25.1	
		Cons.	QTL1	26.08 ± 0.75	−0.98 ± 0.2	1.54 ± 0.38	16.12 ± 5.36	27.3	16.05
			QTL2	34.58 ± 0.94	0.96 ± 0.18	1.51 ± 0.39	15.80 ± 5.17	27.1	
	0.4	Ind.	QTL1	26.33 ± 0.57	−0.95 ± 0.12	1.76 ± 0.49	30.92 ± 6.38	74.8	4.16
			QTL2	34.57 ± 0.59	0.94 ± 0.13	1.76 ± 0.51	30.80 ± 6.44	74.9	
		Cons.	QTL1	26.19 ± 0.66	−0.94 ± 0.13	1.25 ± 0.36	30.67 ± 6.83	77.2	5.23
			QTL2	34.51 ± 0.82	0.94 ± 0.13	1.27 ± 0.35	30.65 ± 6.51	76.9	

[a] *Heritability in broad sense.*
[b] *Position in cM and the associated standard error.*
[c] *Additive effect and the associated standard error.*
[d] *Confidence interval length and the associated standard error.*
[e] *LOD scores and the associated standard error.*
[f] *False discovery rate.*
Ind., individual map; Cons., consensus map.

reported in both consensus and individual maps in wheat (Wang et al., 2014; Li et al., 2015; Guan et al., 2018).

Collinearity was high between the genetic and physical positions. Marker order on the consensus and physical maps was highly correlated at the genome-wide level, but lower collinearity was sometimes observed in some chromosomal regions, which was also reported previously (Wingen et al., 2017). Of the 19,320 SNPs on the consensus map that had physical positions, on average there were 55.17% SNPs arranged in the same order as those on the corresponding chromosomes of the physical maps, ranging from 36.3 on chromosome 6A to 75.68% on chromosome 4D (**Table 2**). A higher proportion of the completely consistent marker order was found in the D genome (63.95%) than those in the A genome (52.32%) and the B genome (49.22%), which may be explained by the lower recombination on the D genome. The lower recombination events on the D genome contributed to lower sequence variability and had a weaker influence on the decay of syntenic block size. Some chromosomal structural variations were observed on the consensus map, such as intra-chromosomal translocation and inversion. For example, inversion happened around 22–25 Mb on chromosome 1A, and translocation occurred between regions around 88–93 and 106–109 Mb on chromosome 2A. The collinearity between marker orders in genetic and physical maps is often disturbed by the macrostructural variations in wheat, especially for consensus maps that are constructed from multiple populations. Local disorder of markers could also be caused by the variation of gene order in parents and genotyping errors.

The distribution of meiotic recombination events showed that recombination happened much more frequently in the distal

chromosomal regions, and that recombination tended to be suppressed near the centromeres, which was consistent with previous studies [Sourdille et al., 2004; International Wheat Genome Sequencing Consortium (IWGSC), 2018]. Collinearity analysis also showed that some markers might have conservative orders across populations, since their relative orders were consistent on the physical and genetic maps. Comparative analysis among the consensus, physical, and individual maps indicated the reliability of the consensus map constructed with the CLA algorithm.

Comparison of the Detected QTLs With Studies on Other Mapping Populations

In this study, eight stable PH QTLs were detected with the consensus map but not with the individual maps (**Table 7**). Guan et al. (2018) reported a PH QTL on chromosome 4D at the physical interval of 37.05–62.94 Mb, and Ren et al. (2021) reported a PH QTL on the same chromosome at the physical interval of 47.44–67.64Mb. *qPH-4D-2* (chr4D:32.97–65.01 Mb) was overlapped with the loci reported by Guan et al. (2018) and Ren et al. (2021). *qPH-6A-1* was located within the physical region as reported by Zanke et al. (2014). *qPH-6A-2* was mapped on chromosome 6A at the interval of 610.97–613.55 Mb. Similarly, Pang et al. (2020) detected a PH QTL on chromosome 6A at the interval of 609.3–609.9 Mb (IWGSC RefSeq v1.0). *qPH-6D-2* was located at the same marker interval of a PH QTL that was first reported and validated to be stable in two wheat populations by Wang et al. (2020). To the best knowledge of the authors, stable QTLs *qPH-2D-1*, *qPH-2D-3*, *qPH-3B-2*, and *qPH-7A* identified in this study were likely to be novel for PH. The increased marker density in the consensus map contributed to the detection of these novel QTLs.

For spike length, 10 QTLs were detected with the consensus map but not with the individual maps (**Table 7**). Among them, a stable QTL in population DS, i.e., *qSL-4B-2* explaining 3.60–16.99% of the phenotypic variance, was close to the Green Revolution gene *Rht-B1*. A number of previous studies have revealed that *Rht-B1* has a pleiotropic effect on PH, SL, and TKW (Schulthess et al., 2017; Sun et al., 2017; Li et al., 2018). QTL cluster *qClu-4B* in which *qSL-4B-2* was located affected all three traits (**Supplementary Table 8**). However, no stable PH QTL in *qClu-4B* was detected in population DS, indicating that *qSL-4B-2* may not be the same as *Rht-B1*. One SL QTL, i.e., *QSl.sdau-4B*, different from but close to *Rht-B1*, was precisely mapped and verified by Deng et al. (2011), which did not affect PH either. *SL-4B-2* was located in a similar position as *QSl.sdau-4B*, and was also in a similar physical position of *qSL4B.1* (chr4B: 36.7–37.8 Mb) reported by Pang et al. (2020). For the remaining nine QTLs, *qSL-3B-5* was mapped on chromosome 3B at the interval of 761.9–774.47 Mb, which was in the similar physical interval (chr3B: 771.94–788.06 Mb) as reported by Hu et al. (2020); *qSL-4A-4* and *qSL-5A-4* were close to those reported in Pang et al. (2020). Six SL QTLs were likely to be novel because of increased power when using the consensus map in QTL mapping, i.e., *qSL-3A-4*, *qSL-3A-5*, *qSL-4B-1*, *qSL-4D*, *qSL-5A-1*, and *qSL-7A-3*.

Compared with the individual maps, three other TKW QTLs were stably identified using the consensus map (**Table 7**), i.e., *qTKW-1B-2*, *qTKW-2D-2*, and *qTKW-6B-3*, which were in similar positions as those reported by Gerard et al. (2019), Zhang et al. (2020c), and Cook et al. (2021), respectively.

For the three traits, a total of 21 QTLs were identified using the consensus map but not the individual maps. Among them, 11 QTLs are consistent with those from previous studies on other mapping populations, and 10 QTLs are likely to be novel. Most of the 11 QTLs were first reported in recent years using high-density linkage maps, indicating that the increase in marker density improved the power of QTL detection. For the novel QTLs, six of them that control PH or SL were included in the cluster that harbored closely linked PH and SL QTLs (**Supplementary Table 8**). The PH of the wheat plant is equal to SL plus the lengths of all internodes above the ground. Theoretically, loci associated with SL may affect PH as well, which has been validated by some studies (Buerstmayr et al., 2011; Lv et al., 2014; Xu et al., 2014; Jahani et al., 2019; Chen et al., 2020). Furthermore, four novel SL QTLs were close to PH QTLs that have been reported using individual maps or other independent studies, indicating the reliability of the novel QTLs on SL or PH. Gene *TaERF8* was identified to be associated with PH and yield in wheat, and has been cloned from the wheat cultivar Chinese Spring (Zhang et al., 2020b), one parental line of population ZC. *TaERF8-2D* (chr2D: 368.21 Mb) was located in the flanking marker interval of *qPH-2D-1*, which was stably detected in population ZC in the three tested environments and in population DS in two tested environments. *TaERF8-2D* may be a candidate gene for *qPH-2D-1*. Annotations of gene functions were also performed for these novel QTLs based on the wheat reference sequence annotation database (IWGSC Annotation v1.1) as listed in **Supplementary Table 12**. The annotation information will facilitate the future fine mapping, map-based cloning, and functional analysis of the novel QTLs identified in this study.

Relationship Between QTLs for Phenotypically Correlated Traits PH and SL

Plant height is an important agronomic trait highly related to lodging resistance and harvest index in wheat. SL is highly related to grain yield by affecting kernel number and spike morphology (Donmez et al., 2001). Plants with suitable PH and larger spike are desirable in wheat breeding. Nine of the 21 stable PH QTLs were close to the stable SL QTLs (**Supplementary Table 8**), contributing to the genetic correlation between the two traits. PH and SL were positively correlated by phenotypic analysis in population DS, but the correlation was non-significant in the other two populations. In this study, closely linked PH and SL QTLs identified in the same population always had genetic effects at the same directions on both traits. Similar instances have been reported in previous studies (Buerstmayr et al., 2011; Lv et al., 2014; Xu et al., 2014; Jahani et al., 2019; Chen et al., 2020). Considering that some QTLs for SL may also affect PH, we speculated that the closely linked PH and SL QTLs are more likely to be the same genetic loci and have the same effect directions.

TABLE 7 | Quantitative trait loci for PH, SL, and TKW detected with the consensus map but not by the individual maps in the three RIL populations, DS, GZ, and ZC.

Trait	QTL	Pop	Environment	Physical interval (Mb)[a]	Neighboring loci in previous studies
PH	*qPH-2D-1*	ZC	E13/E14/B	344.29–426.06	*TaERF8-2D*, Zhang et al., 2020b
	qPH-2D-3	DS	E1/E2/E3/E4/B	617.78–631.92	
	qPH-3B-2	GZ	E7/E10/B	842.16–844.72	
	qPH-4D-2	GZ	E7/E8/E9/E10/B	32.97–65.01	*QPh.cau-4D.2*, Guan et al., 2018
		ZC	E11/E12/E13/E14/B		*QPh.sau-4D*, Ren et al., 2021
	qPH-6A-1	GZ	E7/E9/B	600.13–600.63	Zanke et al., 2014
	qPH-6A-2	DS	E1/E3/B	610.97–613.55	*qPH6A.4*, Pang et al., 2020
	qPH-6D-2	ZC	E11/E13/E14/B	337.17–361.16	*QPh.sicau-6D*, Wang et al., 2020
	qPH-7A	DS	E1/E2/E3/E5/B	611.92–621.35	
SL	*qSL-3A-4*	DS	E2/E5/B	656.58–663.11	
	qSL-3A-5	ZC	E12/E14/B	722.85–748.34	
	qSL-3B-5	DS	E2/E4/E6/B	761.90–774.47	*QSL-3B.2*, Hu et al., 2020
	qSL-4A-4	DS	E1/E2/E5/B	719.47–750.82	*qSL4A.3*, Pang et al., 2020
	qSL-4B-1	DS	E1/E2/B	6.94–10.81	
	qSL-4B-2	DS	E1/E2/E3/E4/E5/E6/B	34.98–49.80	*QSl.sdau-4B*, Deng et al., 2011 *qSL4B.1,* Pang et al., 2020
	qSL-4D	DS	E1/E2/E5/B	65.53–121.40	
	qSL-5A-1	ZC	E13/E14	437.35–445.46	
	qSL-5A-4	DS	E1/E3/E4	671.95–681.28	*qSL5A.2*, Pang et al., 2020
	qSL-7A-3	DS	E2/E5/B	647.11–648.26	
TKW	*qTKW-1B-2*	DS	E1/E2/E3/E6/B	588.36–591.14	*BS00039740_51*, Gerard et al., 2019
	qTKW-2D-2	ZC	E11/E12/E13/E14/B	523.15–555.13	*AX-109775854*, Zhang et al., 2020c
	qTKW-6B-3	ZC	E11/E12/B	157.21–162.58	*IWB61228-6B*, Cook et al., 2021

[a] *Physical positions for the flanking markers of QTLs based on IWGSC_RefSeq v2.0.*

E1, 2012–2013 Beijing; E2, 2012–2013 Shijiazhuang; E3, 2013–2014 Beijing; E4, 2013–2014 Shijiazhuang; E5, 2014–2015 Beijing; E6, 2014–2015 Shijiazhuang; E7, 2012–2013 Anyang; E8, 2012–2013 Suixi; E9, 2013–2014 Anyang; E10, 2013–2014 Suixi; E11, Zhoukou2013; E12, Zhengzhou 2013; E13, Zhoukou 2014; E14, Zhengzhou 2014; B, best linear unbiased estimation.

However, whether the closely linked QTLs on PH and SL belong to the same chromosomal loci with pleiotropic effects or different closely-linked loci needs further investigation and is beyond the scope of this study.

Further Analysis for a Major PH QTL Located on Chromosome 4DS

For plant height, only one QTL was detected on chromosome 4DS using the individual maps in populations GZ and ZC, but two stable QTLs, i.e., *qPH-4D-1* and *qPH-4D-2*, were identified using the consensus map in the same two populations, which were linked in the coupling phase (**Table 3**, **Supplementary Table 7**). The BLAST results indicated that *qPH-4D-1* was co-localized with the dwarfing gene *Rht-D1*. *qPH-4D-2* explained 8.39–10.9 and 5.03–12.03% of the phenotypic variance across environments in populations GZ and ZC, respectively. The alleles decreasing PH were from parents Zhoumai16 in population GZ and Zhou 8425B in population ZC.

Guan et al. (2018) reported two QTLs that were also linked in the coupling phase and located in similar positions as *qPH-4D-1* and *qPH-4D-2*. *qPH-4D-2* was detected in four environments and with BLUE values across eight environments in Guan et al. (2018). In addition, *qPH-4D-2* was closely linked with marker wsnp_Ex_c683_1341113, which was also observed in Guan et al. (2018). As reported by Ren et al. (2021), *qPH-4D-2* was detected in the similar position between SNPs AX-89692818 and AX-109606880 across environments. Therefore, it is highly possible that *qPH-4D-2* is a novel semi-dwarfing gene. The common marker wsnp_Ex_c683_1341113 was located at about 54.4 Mb on chromosome 4D (IWGSC RefSeq v1.0; IWGSC, 2018). A high confidence putative gene, *TraesCS4D02G076400* (50,888,586–50,889,461 bp), is located around the marker and in the confidence interval of *qPH-4D-2*, with the annotation of encoding gibberellin regulated protein (IWGSC RefSeq v1.1 annotation; IWGSC, 2018). Gibberellin is an essential endogenous regulator in plant growth. The well-known dwarfing genes *Rht-B1b* and *Rht-D1b* regulate DELLA proteins in gibberellin signaling to reduce the response to gibberellin (Peng et al., 1999). The gibberellin-sensitive gene *Rht8* was also widely used in regulating PH in wheat (Gasperini et al., 2012). Gene *TraesCS4D02G076400* in wheat was annotated to gene *GAST1* (UniProtKB/TrEMBL; Acc:C8C4P9), first reported in tomato to encode the gibberellins-stimulated transcript (Shi et al., 1992). *GAST1* belongs to the gibberellic acid-stimulated *Arabidopsis* (*GASA*) family, which plays important roles in plant growth and development, such as stem growth, plant height, and grain length, width, and weight (de la Fuente et al., 2006; Nahirñak et al., 2012a,b; Shi et al., 2020). Furthermore, *qPH-4D-2* was detected in two populations in this study, one from the cross between Zhou 8425B and Chinese Spring. *TraesCS4D02G076400* had high RNA expression levels in Chinese Spring in different tissues and development

stages (expVIP, http://www.wheat-expression.com/). Therefore, *TraesCS4D02G076400* is likely to be the candidate gene for *qPH-4D-2*. PH is a crucial trait for morphogenesis and grain yield in wheat. The newly discovered PH QTL on chromosome 4DS in this study may enrich the genetic resources in breeding for semi-dwarfing wheat. Reasons that *qPH-4D-2* was not identified by the individual could be the short distance between the QTL and *Rht-D1*, and the lower marker density around the two QTLs in individual mapping populations.

Advantages of Using Consensus Map in QTL Mapping

Due to the limited number of crossing-overs and limited genetic variation in individual populations, linkage maps constructed from individual mapping populations usually have a large number of co-localized markers and low marker density. A consensus map combines the genetic information included in multiple populations and provides a better genomic coverage with higher marker density (Maccaferri et al., 2015; Allen et al., 2017). A consensus map of higher density offers the chance to map QTLs to narrower chromosomal intervals, which will facilitate the discovery of causal genes and the identification of closely linked markers for MAS. Simulation results conducted in this study confirmed that the use of a consensus map with higher marker density reduced the confidence interval of detected QTLs.

Even for the same trait, QTLs detected in different populations using their own genetic maps sometimes are hardly compared and synthesized, because of the unshared markers and variations in the genetic background (Sukumaran et al., 2015). Comparisons on QTL positions estimated from different populations are usually conducted by anchoring the linked markers to the genome assembly. However, genome sequences usually have wide variations between parental varieties, and the anchor information to the genome sequence may not be completely accurate. A consensus map provides the direct comparison for QTLs detected from different populations, which is important, particularly in species lacking a completely sequenced reference genome. In this study, we demonstrated that QTL mapping using a consensus map can better identify common and stable QTLs across populations and environments. For example, *Rht-B1* and *Rht-D1*, which had been cloned, were the two genes reducing plant height in wheat (Peng et al., 1999). Each of them was located almost in the same position in two populations on the consensus map. *qPH-5A-2*, *qSL-2D-2*, *qSL-5A-2*, and *qTKW-4B-2* were detected in populations DS and GZ; *qPH-2B-2* was detected in populations DS and ZC; *qPH-4B-1*, *qPH-4D-1*, *qPH-4D-2*, and *qSL-2D-1* were identified in populations GZ and ZC; *qSL-6B-4* was detected in all the three populations. The common QTLs identified in multiple populations reflected the stable genetic effects of QTLs in different genetic backgrounds, which might be more valuable in breeding.

The genetic relationship among PH and SL QTLs as observed in this study, showed that QTL mapping using the consensus map can also facilitate the comparison across the correlated traits, and therefore provide the opportunity to understand the genetic correlation between phenotypically correlated traits and identify the QTL-rich genomic regions. Moreover, the consensus map also provides the chance to detect common QTLs with smaller effects occurring in different populations.

Further studies may still be needed to determine the key factors affecting the accuracy of consensus map construction and subsequent QTL mapping, such as proportion of common markers shared by multiple mapping populations, inconsistency degree of marker orders in individual populations, population-specific recombination frequencies, and the optimum algorithm used to construct the consensus map. In addition to bi-parental populations, as have been used in this study, multi-parental populations have been developed in recent years in crops together with suitable genetic analysis methods (Gardner et al., 2016; Zhang et al., 2017, 2019; Shi et al., 2019; Qu et al., 2020). In theory, a consensus map can also be constructed by combining a number of bi-parental and multi-parental populations, when common markers are shared by these populations.

In conclusion, the consensus map constructed for this study allows for systematic QTL mapping studies, and comparison and clustering of mapping results in wheat genetic studies. The QTL mapping based on the consensus map resulted in higher accuracy, narrower confidence interval, and a larger QTL number. The stable QTLs across tested environments and mapping populations, and the predicted QTL genotypes and genotypic values can be used to select wheat cultivars with suitable PH, large SL, and medium to high kernel weight. SNPs closely linked with these stable QTLs can be used to select suitable genetic materials and make suitable crosses in wheat breeding programs. SNPs closely linked to traits can also be converted into Kompetitive allele-specific PCR (KASP) markers (Kaur et al., 2021) and then used for large-scale genotyping to screen desirable individuals in segregating breeding populations.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

LZ and JW conceived and designed the research. PQ and LZ conducted data analysis. FG, WW, JL, and XX developed the populations, performed SNP genotyping, and conducted field trials. PQ, JW, and LZ wrote, drafted, and revised the manuscript. XX and HP provided guidance on data analysis and revised the manuscript. All the authors read and approved the final version of the manuscript for publication.

FUNDING

This study was supported by the National Natural Science Foundation of China (Project No. 31861143003).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.727077/full#supplementary-material

REFERENCES

Allen, A. M., Winfield, M. O., Burridge, A. J., Downie, R. C., Benbow, H. R., Barker, G. L., et al. (2017). Characterization of a Wheat Breeders' Array suitable for high-throughput SNP genotyping of global accessions of hexaploid bread wheat (*Triticum aestivum*). *Plant Biotechnol. J.* 15, 390–401. doi: 10.1111/pbi.12635

Arcade, A., Labourdette, A., Falque, M., Mangin, B., Chardon, F., Charcosset, A., et al. (2004). BioMercator: integrating genetic maps and QTL towards discovery of candidate genes. *Bioinformatics* 20, 2324–2326. doi: 10.1093/bioinformatics/bth230

Buerstmayr, M., Lemmens, M., Steiner, B., and Buerstmayr, H. (2011). Advanced backcross QTL mapping of resistance to Fusarium head blight and plant morphological traits in a *Triticum macha*× *T. aestivum* population. *Theor. Appl. Genet.* 123:293. doi: 10.1007/s00122-011-1584-x

Cavanagh, C. R., Chao, S., Wang, S., Huang, B. E., Stephen, S., Kiani, S., et al. (2013). Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proc. Natl. Acad. Sci. U.S.A.* 110, 8057–8062. doi: 10.1073/pnas.1217133110

Chen, H., and Boutros, P. C. (2011). VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Plant Biol.* 12:35. doi: 10.1186/1471-2105-12-35

Chen, W., Sun, D., Li, R., Wang, S., Shi, Y., Zhang, W., et al. (2020). Mining the stable quantitative trait loci for agronomic traits in wheat (*Triticum aestivum* L.) based on an introgression line population. *BMC Plant Biol.* 20:275. doi: 10.1186/s12870-020-02488-z

Cook, J., Acharya, R., Martin, J., Blake, N., Khan, I., Heo, H. Y., et al. (2021). Genetic analysis of stay-green, yield, and agronomic traits in spring wheat. *Crop Sci.* 61, 383–395. doi: 10.1002/csc2.20302

de la Fuente, J. I., Amaya, I., Castillejo, C., Sánchez-Sevilla, J. F., Quesada, M. A., Botella, M. A., et al. (2006). The strawberry gene *FaGAST* affects plant growth through inhibition of cell elongation. *J. Exp. Bot.* 57, 2401–2411. doi: 10.1093/jxb/erj213

Deng, S., Wu, X., Wu, Y., Zhou, R., Wang, H., Jia, J., et al. (2011). Characterization and precise mapping of a QTL increasing spike number with pleiotropic effects in wheat. *Theor. Appl. Genet.* 122, 281–289. doi: 10.1007/s00122-010-1443-1

Donmez, E., Sears, R., Shroyer, J., and Paulsen, G. (2001). Genetic gain in yield attributes of winter wheat in the Great Plains. *Crop Sci.* 41, 1412–1419. doi: 10.2135/cropsci2001.4151412x

Endelman, J. B., and Plomion, C. (2014). LPmerge: an R package for merging genetic maps by linear programming. *Bioinformatics* 30, 1623–1624. doi: 10.1093/bioinformatics/btu091

Ferreira, A., da Silva, M. F., and Cruz, C. D. (2006). Estimating the effects of population size and type on the accuracy of genetic maps. *Genet. Mol. Biol.* 29, 187–192. doi: 10.1590/S1415-47572006000100033

Gao, F., Wen, W., Liu, J., Rasheed, A., Yin, G., Xia, X., et al. (2015). Genome-wide linkage mapping of QTL for yield components, plant height and yield-related physiological traits in the Chinese wheat cross Zhou 8425B/Chinese Spring. *Front. Plant Sci.* 6:1099. doi: 10.3389/fpls.2015.01099

Gardner, K. A., Wittern, L. M., and Mackay, I. J. (2016). A highly recombined, high-density, eight-founder wheat MAGIC map reveals extensive segregation distortion and genomic locations of introgression segments. *Plant Biotechnol. J.* 14, 1406–1417. doi: 10.1111/pbi.12504

Gasperini, D., Greenland, A., Hedden, P., Dreos, R., Harwood, W., and Griffiths, S. (2012). Genetic and physiological analysis of *Rht8* in bread wheat: an alternative source of semi-dwarfism with a reduced sensitivity to brassinosteroids. *J. Exp. Bot.* 63:4419. doi: 10.1093/jxb/ers292

Gerard, G. S., Alqudah, A., Lohwasser, U., Börner, A., and Simón, M. R. (2019). Uncovering the genetic architecture of fruiting efficiency in bread wheat: a viable alternative to increase yield potential. *Crop Sci.* 59, 1853–1869. doi: 10.2135/cropsci2018.10.0639

Guan, P., Lu, L., Jia, L., Kabir, M. R., Zhang, J., Lan, T., et al. (2018). Global QTL analysis identifies genomic regions on chromosomes 4A and 4B harboring stable loci for yield-related traits across different environments in wheat (*Triticum aestivum* L.). *Front. Plant Sci.* 9:529. doi: 10.3389/fpls.2018.00529

Hu, J., Wang, X., Zhang, G., Jiang, P., Chen, W., Hao, Y., et al. (2020). QTL mapping for yield-related traits in wheat based on four RIL populations. *Theor. Appl. Genet.* 133, 917–933. doi: 10.1007/s00122-019-03515-w

International Wheat Genome Sequencing Consortium (IWGSC) (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361:eaar7191. doi: 10.1126/science.aar7191

Jahani, M., Mohammadi-Nejad, G., Nakhoda, B., and Rieseberg, L. H. (2019). Genetic dissection of epistatic and QTL by environment interaction effects in three bread wheat genetic backgrounds for yield-related traits under saline conditions. *Euphytica* 215:103. doi: 10.1007/s10681-019-2426-1

Kaur, J., Kaur, J., Dhillon, G. S., Kaur, H., Singh, J., Bala, R., et al. (2021). Characterization and mapping of spot blotch in *Triticum durum–Aegilops speltoides* introgression lines using SNP markers. *Front. Plant Sci.* 12:650400. doi: 10.3389/fpls.2021.650400

Li, F., Wen, W., He, Z., Liu, J., Jin, H., Cao, S., et al. (2018). Genome-wide linkage mapping of yield-related traits in three Chinese bread wheat populations using high-density SNP markers. *Theor. Appl. Genet.* 131, 1903–1924. doi: 10.1007/s00122-018-3122-6

Li, H., Vikram, P., Singh, R. P., Kilian, A., Carling, J., Song, J., et al. (2015). A high density GBS map of bread wheat and its application for dissecting complex disease resistance traits. *BMC Genomics* 16:216. doi: 10.1186/s12864-015-1424-5

Li, H., Zhang, L., and Wang, J. (2012). Estimation of statistical power and false discovery rate of QTL mapping methods through computer simulation. *Chin. Sci. Bull.* 57, 2701–2710. doi: 10.1007/s11434-012-5239-3

Liu, Y., Salsman, E., Wang, R., Galagedara, N., Zhang, Q., Fiedler, J. D., et al. (2020). Meta-QTL analysis of tan spot resistance in wheat. *Theor. Appl. Genet.* 133, 2363–2375. doi: 10.1007/s00122-020-03604-1

Liu, Y., and Zeng, Z.-B. (2000). A general mixture model approach for mapping quantitative trait loci from diverse cross designs involving multiple inbred lines. *Genet. Res.* 75, 345–355. doi: 10.1017/S0016672300004493

Lorieux, M. (2012). MapDisto: fast and efficient computation of genetic linkage maps. *Mol. Breed.* 30, 1231–1235. doi: 10.1007/s11032-012-9706-y

Lv, C., Song, Y., Gao, L., Yao, Q., Zhou, R., Xu, R., et al. (2014). Integration of QTL detection and marker assisted selection for improving resistance to Fusarium head blight and important agronomic traits in wheat. *Crop J.* 2, 70–78. doi: 10.1016/j.cj.2013.10.004

Maccaferri, M., Ricci, A., Salvi, S., Milner, S. G., Noli, E., Martelli, P. L., et al. (2015). A high-density, SNP-based consensus map of tetraploid wheat as a bridge to integrate durum and bread wheat genomics and breeding. *Plant Biotechnol. J.* 13, 648–663. doi: 10.1111/pbi.12288

Meng, L., Li, H., Zhang, L., and Wang, J. (2015). QTL IciMapping: integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations. *Crop J.* 3, 269–283. doi: 10.1016/j.cj.2015.01.001

Nahirñak, V., Almasia, N. I., Fernandez, P. V., Hopp, H. E., Estevez, J. M., Carrari, F., et al. (2012a). Potato snakin-1 gene silencing affects cell division, primary metabolism, and cell wall composition. *Plant Physiol.* 158, 252–263. doi: 10.1104/pp.111.186544

Nahirñak, V., Almasia, N. I., Hopp, H. E., and Vazquez-Rovere, C. (2012b). Snakin/GASA proteins: involvement in hormone crosstalk and redox homeostasis. *Plant Signal. Behav.* 7, 1004–1008. doi: 10.4161/psb.20813

Ouellette, L. A., Reid, R. W., Blanchard, S. G., and Brouwer, C. R. (2018). LinkageMapView—rendering high-resolution linkage and QTL maps. *Bioinformatics* 34, 306–307. doi: 10.1093/bioinformatics/btx576

Pang, Y., Liu, C., Wang, D., Amand, P. S., Bernardo, A., Li, W., et al. (2020). High-resolution genome-wide association study identifies genomic regions and candidate genes for important agronomic traits in wheat. *Mol. Plant* 13, 1311–1327. doi: 10.1016/j.molp.2020.07.008

Peng, J., Richards, D. E., Hartley, N. M., Murphy, G. P., Devos, K. M., Flintham, J. E., et al. (1999). 'Green revolution'genes encode mutant gibberellin response modulators. *Nature* 400, 256–261. doi: 10.1038/22307

Qu, P., Shi, J., Chen, T., Chen, K., Shen, C., Wang, J., et al. (2020). Construction and integration of genetic linkage maps from three multi-parent advanced generation inter-cross populations in rice. *Rice* 13:13. doi: 10.1186/s12284-020-0373-z

Rasheed, A., Wen, W., Gao, F., Zhai, S., Jin, H., Liu, J., et al. (2016). Development and validation of KASP assays for genes underpinning key economic traits in bread wheat. *Theor. Appl. Genet.* 129, 1843–1860. doi: 10.1007/s00122-016-2743-x

Ren, T., Fan, T., Chen, S., Li, C., Chen, Y., Ou, X., et al. (2021). Utilization of a Wheat55K SNP array-derived high-density genetic map for high-resolution mapping of quantitative trait loci for important kernel-related traits in common wheat. *Theor. Appl. Genet.* 134, 807–821. doi: 10.1007/s00122-020-03732-8

Ronin, Y., Mester, D., Minkov, D., Belotserkovski, R., Jackson, B., Schnable, P., et al. (2012). Two-phase analysis in consensus genetic mapping. *G3-Genes Genomes Genet.* 2, 537–549. doi: 10.1534/g3.112.002428

Schulthess, A. W., Reif, J. C., Ling, J., Plieske, J., Kollers, S., Ebmeyer, E., et al. (2017). The roles of pleiotropy and close linkage as revealed by association mapping of yield and correlated traits of wheat (*Triticum aestivum* L.). *J. Exp. Bot.* 68, 4089–4101. doi: 10.1093/jxb/erx214

Sharp, P., Kreis, M., Shewry, P., and Gale, M. (1988). Location of β-amylase sequences in wheat and its relatives. *Theor. Appl. Genet.* 75, 286–290. doi: 10.1007/BF00303966

Shi, C. L., Dong, N. Q., Guo, T., Ye, W. W., Shan, J. X., and Lin, H. X. (2020). A quantitative trait locus *GW6* controls rice grain size and yield through the gibberellin pathway. *Plant J.* 103, 1174–1188. doi: 10.1111/tpj.14793

Shi, J., Wang, J., and Zhang, L. (2019). Genetic mapping with background control for quantitative trait locus (QTL) in 8-parental pure-line populations. *J. Hered.* 110, 880–891. doi: 10.1093/jhered/esz050

Shi, L., Gast, R. T., Gopalraj, M., and Olszewski, N. E. (1992). Characterization of a shoot-specific, GA3-and ABA-regulated gene from tomato. *Plant J.* 2, 153–159. doi: 10.1046/j.1365-313X.1992.t01-39-00999.x

Somers, D. J., Isaac, P., and Edwards, K. (2004). A high-density microsatellite consensus map for bread wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* 109, 1105–1114. doi: 10.1007/s00122-004-1740-7

Sourdille, P., Singh, S., Cadalen, T., Brown-Guedira, G. L., Gay, G., Qi, L., et al. (2004). Microsatellite-based deletion bin system for the establishment of genetic-physical map relationships in wheat (*Triticum aestivum* L.). *Funct. Integr. Genomics* 4, 12–25. doi: 10.1007/s10142-004-0106-1

Stam, P. (1993). Construction of integrated genetic linkage maps by means of a new computer package: join map. *Plant J.* 3, 739–744. doi: 10.1111/j.1365-313X.1993.00739.x

Sukumaran, S., Dreisigacker, S., Lopes, M., Chavez, P., and Reynolds, M. P. (2015). Genome-wide association study for grain yield and related traits in an elite spring wheat population grown in temperate irrigated environments. *Theor. Appl. Genet.* 128, 353–363. doi: 10.1007/s00122-014-2435-3

Sun, C., Zhang, F., Yan, X., Zhang, X., Dong, Z., Cui, D., et al. (2017). Genome-wide association study for 13 agronomic traits reveals distribution of superior alleles in bread wheat from the Yellow and Huai Valley of China. *Plant Biotechnol. J.* 15, 953–969. doi: 10.1111/pbi.12690

Van Ooijen, J. (2006). *JoinMap® 4, Software for the Calculation of Genetic Linkage Maps in Experimental Populations. Kyazma B.V*. Wageningen, Netherlands.

Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B. E., et al. (2014). Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant Biotechnol. J.* 12, 787–796. doi: 10.1111/pbi.12183

Wang, Z., Hu, H., Jiang, X., Tao, Y., Lin, Y., Wu, F., et al. (2020). Identification and validation of a novel major quantitative trait locus for plant height in common wheat (*Triticum aestivum* L.). *Front. Genet.* 11:602495. doi: 10.3389/fgene.2020.602495

Wen, W., He, Z., Gao, F., Liu, J., Jin, H., Zhai, S., et al. (2017). A high-density consensus map of common wheat integrating four mapping populations scanned by the 90K SNP array. *Front. Plant Sci.* 8:1389. doi: 10.3389/fpls.2017.01389

Wingen, L. U., West, C., Leverington-Waite, M., Collier, S., Orford, S., Goram, R., et al. (2017). Wheat landrace genome diversity. *Genetics* 205, 1657–1676. doi: 10.1534/genetics.116.194688

Wu, Y., Close, T. J., and Lonardi, S. (2010). Accurate construction of consensus genetic maps via integer linear programming. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 8, 381–394. doi: 10.1109/TCBB.2010.35

Xu, Y., Wang, R., Tong, Y., Zhao, H., Xie, Q., Liu, D., et al. (2014). Mapping QTLs for yield and nitrogen-related traits in wheat: influence of nitrogen and phosphorus fertilization on QTL expression. *Theor. Appl. Genet.* 127, 59–72. doi: 10.1007/s00122-013-2201-y

Yao, J., Zhao, D., Chen, X., Zhang, Y., and Wang, J. (2018). Use of genomic selection and breeding simulation in cross prediction for improvement of yield and quality in wheat (*Triticum aestivum* L.). *Crop J.* 6, 353–365. doi: 10.1016/j.cj.2018.05.003

Yu, Y., Ouyang, Y., and Yao, W. (2018). shinyCircos: an R/Shiny application for interactive creation of Circos plot. *Bioinformatics* 34, 1229–1231. doi: 10.1093/bioinformatics/btx763

Zanke, C. D., Ling, J., Plieske, J., Kollers, S., Ebmeyer, E., Korzun, V., et al. (2014). Whole genome association mapping of plant height in winter wheat (*Triticum aestivum* L.). *PLoS ONE* 91:e113287. doi: 10.1371/journal.pone.0113287

Zhang, L., Li, H., Meng, L., and Wang, J. (2020a). Ordering of high-density markers by the k-Optimal algorithm for the traveling-salesman problem. *Crop J.* 8, 701–712. doi: 10.1016/j.cj.2020.03.005

Zhang, L., Liu, P., Wu, J., Qiao, L., Zhao, G., Jia, J., et al. (2020b). Identification of a novel ERF gene, *TaERF8*, associated with plant height and yield in wheat. *BMC Plant Biol.* 20:263. doi: 10.1186/s12870-020-02473-6

Zhang, L., Meng, L., and Wang, J. (2019). Linkage analysis and integrated software GAPL for pure-line populations derived from four-way and eight-way crosses. *Crop J.* 7, 283–293. doi: 10.1016/j.cj.2018.10.006

Zhang, N., Zhang, X., Song, L., Su, Q., Zhang, S., Liu, J., et al. (2020c). Identification and validation of the superior alleles for wheat kernel traits detected by genome-wide association study under different nitrogen environments. *Euphytica* 216:52. doi: 10.1007/s10681-020-2572-5

Zhang, S., Meng, L., Wang, J., and Zhang, L. (2017). Background controlled QTL mapping in pure-line genetic populations derived from four-way crosses. *Heredity* 119, 256–264. doi: 10.1038/hdy.2017.42

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 26 August 2021
doi: 10.3389/fpls.2021.711654

Genome-Wide Association Study of Waterlogging Tolerance in Barley (*Hordeum vulgare* L.) Under Controlled Field Conditions

*Ana Borrego-Benjumea[1], Adam Carter[1], Min Zhu[2], James R. Tucker[1], Meixue Zhou[3] and Ana Badea[1]**

[1] Brandon Research and Development Centre, Agriculture and Agri-Food Canada, Brandon, MB, Canada, [2] College of Agriculture, Yangzhou University, Yangzhou, China, [3] Tasmanian Institute of Agriculture, University of Tasmania, Hobart, TAS, Australia

Waterlogging is one of the main abiotic stresses severely reducing barley grain yield. Barley breeding programs focusing on waterlogging tolerance require an understanding of genetic loci and alleles in the current germplasm. In this study, 247 worldwide spring barley genotypes grown under controlled field conditions were genotyped with 35,926 SNPs with minor allele frequency (MAF) > 0.05. Significant phenotypic variation in each trait, including biomass, spikes per plant, grains per plant, kernel weight per plant, plant height and chlorophyll content, was observed. A genome-wide association study (GWAS) based on linkage disequilibrium (LD) for waterlogging tolerance was conducted. Population structure analysis divided the population into three subgroups. A mixed linkage model using both population structure and kinship matrix (Q+K) was performed. We identified 17 genomic regions containing 51 significant waterlogging-tolerance-associated markers for waterlogging tolerance response, accounting for 5.8–11.5% of the phenotypic variation, with a majority of them localized on chromosomes 1H, 2H, 4H, and 5H. Six novel QTL were identified and eight potential candidate genes mediating responses to abiotic stresses were located at QTL associated with waterlogging tolerance. To our awareness, this is the first GWAS for waterlogging tolerance in a worldwide barley collection under controlled field conditions. The marker-trait associations could be used in the marker-assisted selection of waterlogging tolerance and will facilitate barley breeding.

Keywords: barley, waterlogging tolerance, genome-wide associated study, marker-trait association, quantitative trait loci, candidate genes

OPEN ACCESS

Edited by:
Leonardo Abdiel Crespo Herrera, International Maize and Wheat Improvement Center, Mexico

Reviewed by:
Stefania Grando, Consultant, Ascoli Piceno, Italy
Hongliang Zheng, Northeast Agricultural University, China

****Correspondence:***
Ana Badea
ana.badea@agr.gc.ca

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *19 May 2021*
Accepted: *21 July 2021*
Published: *26 August 2021*

Citation:
Borrego-Benjumea A, Carter A, Zhu M, Tucker JR, Zhou M and Badea A (2021) Genome-Wide Association Study of Waterlogging Tolerance in Barley (Hordeum vulgare L.) Under Controlled Field Conditions. Front. Plant Sci. 12:711654. doi: 10.3389/fpls.2021.711654

INTRODUCTION

Waterlogging is a major abiotic stress that causes oxygen depletion and carbon oxide accumulation in the rhizosphere (Bailey-Serres and Voesenek, 2008) and has become one of the main concerns for crops limiting agricultural production globally. It is estimated that, worldwide, 10–16% of the arable soils are affected by waterlogging (Setter and Waters, 2003; Yaduvanshi et al., 2014). In western Canada, waterlogging has been identified as an important limiting factor for the crops grown, including barley. In the last decade, waterlogging was accountable for 52% of post-harvest

claims for crop losses by farmers in Manitoba and Saskatchewan [Manitoba Agricultural Services Corporation (MASC), 2017; Saskatchewan Crop Insurance Corporation (SCIC), 2017]. Waterlogging occurs when there is excess moisture in the soil caused by high precipitation combined with poor soil drainage, resulting in anoxic and hypoxia within roots (Arduini et al., 2016). Waterlogging also causes an excess of ethylene and carbon dioxide that also increases metabolic toxins and microelements such as iron and manganese in soil solution or roots, reduces respiration, root conductivity to water, and nutrient uptake, thus affecting plant growth and survival (Setter and Waters, 2003).

Barley (*Hordeum vulgare* L.) is the fourth most important cereal crop globally and Canada's fourth-largest crop and is primarily used for livestock feed, malting, and food (FAOSTAT Production, 2020; Statistics Canada, 2020). Canada is the fourth largest barley producer and the second-largest malt exporter in the world. On average, each year, ∼$1 billion is directly generated from the export of feed barley and malt [Canadian Agri-Food Trade Alliance (CAFTA), 2020]. Barley is more susceptible to waterlogging stress than other cereals (Setter and Waters, 2003). Waterlogging stress may cause significant yield losses in barley that vary from 10 to 50%, depending on factors such as the depth and duration of flooding, the development stage of the waterlogged plant, temperature (Setter et al., 1999) and type of soil (Pang et al., 2004). Waterlogging stress affects the genome-wide gene expression responses in barley roots, increasing the expression of many genes related to stress tolerance in barley roots, including glycolysis and fermentation-related genes, as well as ethylene-responsive element binding factors, and decreasing the expression of genes related to starch and sucrose metabolism, and nitrogen and amino acid metabolism (Borrego-Benjumea et al., 2020).

In barley, damages caused by soil waterlogging include chlorosis and premature leaf senescence, reduced root growth, tillering, dry matter accumulation, number and weight of kernels, and increased floral sterility (De San Celedonio et al., 2014, 2018; Masoni et al., 2016; Ploschuk et al., 2018; Sundgren et al., 2018). Under outdoor conditions in Argentina, Ploschuk et al. (2018) assessed tolerance to 14-days of early- or late-stage waterlogging of winter barley, which produced adventitious roots with 19% of aerenchyma. They showed that photosynthesis was reduced during waterlogging, but early-waterlogged plants were able to recover upon drainage with seed production reaching 85% of the controls, while late-waterlogged plants only attained 32% in seed production. Sayre et al. (1994) found that the growth stage of barley from leaf emergence to the booting stage is more sensitive to waterlogging, while Liu et al. (2020) reported that waterlogging close to heading is the most susceptible period, with yield losses primarily attributed to reductions in spikelet fertility and grain weight. In the Canadian Prairies, it has been projected increased precipitation in the coming years during May-June period (Blair et al., 2016). This is a critical period in the barley growing season in this region where increased precipitation reduces barley grain yield (Borrego-Benjumea et al., 2019). Therefore, it is important to develop cultivars tolerant to excess moisture and thus to increase the yield stability of barley.

Waterlogging tolerance is a complex quantitative trait under strong environmental influence with relatively low heritability of grain yield in barley (Hamachi et al., 1990). Due to this low heritability and dependency on environmental conditions, the direct selection of barley for waterlogging tolerance is time-consuming and less effective. Marker-assisted selection (MAS) is an effective approach that can improve the efficiency of breeding waterlogging-tolerant barley varieties and avoid environmental effects. MAS requires identifying appropriate quantitative trait loci (QTL) for traits associated with waterlogging tolerance, and the development of molecular markers closely linked to these traits. In barley, major QTL associated with waterlogging tolerance have revealed numerous genomic regions that affect important traits, such as chlorophyll fluorescence (Bertholdsson et al., 2015), root aerenchyma formation in cultivated and wild barley (Li et al., 2008; Zhang et al., 2016; Zhang X. et al., 2017), root membrane potential (Gill et al., 2017), root porosity (Broughton et al., 2015; Zhang et al., 2016), reactive oxygen species (ROS) formation (Gill et al., 2019), waterlogging score (Li et al., 2008; Zhou, 2011; Zhou et al., 2012), and yield components (Xue et al., 2010; Xu et al., 2012). All these major QTL have been mapped using doubled haploids (DH) populations from bi-parental crosses of contrasting phenotype parents for waterlogging. Although this approach has been the most applied and has been very successful in detecting many QTL for waterlogging tolerance in barley, few of the QTL reported have been successfully used in MAS.

Association mapping (AM) is another alternative to mapping QTL associated with complex traits in crops. The AM takes advantage of historic linkage disequilibrium to uncover genetic associations. Genome-wide association study (GWAS) requires high marker density because linkage disequilibrium (LD) is low in GWAS populations than in bi-parental populations. In GWAS, the mapping population consists of a diverse set of individuals or lines drawn from natural populations and breeding populations. GWAS has been used to detect QTL involved in response to waterlogging stress in various crops such as maize (Zhang et al., 2013), rice (Zhang M. et al., 2017), soybean (Cornelious et al., 2005) and wheat (Sundgren, 2018). In barley, GWAS has been used to identify QTL for not only agronomic traits, such as yield and yield components-related traits, using GWAS (Pasam et al., 2012; Locatelli et al., 2013; Tondelli et al., 2013; Pauli et al., 2014; Bellucci et al., 2017; Xu et al., 2018) but also tolerance to abiotic stresses such as salinity (Long et al., 2013; Fan et al., 2016; Mwando et al., 2020), drought (Varshney et al., 2012; Jabbari et al., 2018; Tarawneh et al., 2020), acid soil (Zhou et al., 2016), and low potassium (Ye et al., 2020) stress tolerance. However, no information is available for QTL mapping for waterlogging tolerance in barley by GWAS. In the present study, we assessed a worldwide barley collection for waterlogging stress tolerance under controlled field conditions. We evaluated

Abbreviations: BIO, above-ground dry Biomass; CABC, chlorophyll a+b content; CCC, chlorophyll carotenoids content; GP, number of grains per plant; GWAS, Genome-wide association study; KWP, kernel weight per plant; LD, linkage disequilibrium; PH, plant height; QTL, quantitative trait loci/locus; SNP, single nucleotide polymorphism; SP, number of spikes per plant; WLS, waterlogging score.

the phenotypic and genetic diversity and the patterns of LD decay across the barley genome. We conducted GWAS for waterlogging tolerant traits, aiming to uncover novel genomic regions and identify marker-trait associations for waterlogging tolerance and confirm the previously identified genomic regions and single nucleotide polymorphism (SNP) marker associated with waterlogging tolerance. To our awareness, this is the first AM study for waterlogging stress tolerance in a worldwide barley collection under controlled field conditions.

MATERIALS AND METHODS

Plant Material

A spring barley worldwide collection of 247 genotypes, including advanced breeding lines, cultivars, and landraces, was assembled and used in this study. The majority of genotypes were from Canada (30%), the USA (12%), China (10%), and Australia (8%). The rest were from 35 different countries.

Field Experiment

The barley genotypes were evaluated for waterlogging tolerance in controlled field conditions in one location at the experimental station of the Brandon Research and Development Centre (49°52′ N, 99°58′ W) in two consecutive years (2016 and 2017). This location is a place where water is prone to accumulate, creating excess moisture problems. The soil has a sandy loam texture. The field trial area was leveled before seeding to ensure that all plants would be under the same water level. A ridge was built on the treatment side and was encircled by a plastic film to avoid water escape. The experimental design used was a randomized complete block design with three replications. Each plot represented one experimental unit, consisting of a single-row plot of 0.92 m length containing 25 seeds evenly distributed with 0.31 m spacing between rows. Seeds were sown in late May or early June following standard agronomic practices. Waterlogging-tolerant genotype Deder2 and waterlogging-sensitive genotype Franklin were used as checks. The waterlogging stress treatment was initiated at the tillering stage on the treatment side by adding the water to 0.5–1 cm above the soil surface. Waterlogging treatment was maintained at the same level and continued until the susceptible checks showed considerable stress symptoms (around 70% leaf symptom yellowing) and genotypic differences were easily distinguishable. The treatment duration was 9 and 7 days in 2016 and 2017, respectively. Then water in the waterlogged plots was drained out, and the plants were allowed to grow to maturity. Standard agronomic and cultural practices were applied to the other side of the field, used as control. The precipitation during the growing season was 394 and 245 mm in 2016 and 2017, respectively.

After full maturity, three individual plants were randomly harvested from each plot for analytical measurements. The traits evaluated included above-ground dry Biomass (BIO), number of spikes per plant (SP), number of grains per plant (GP), kernel weight per plant (KWP), plant height (PH), chlorophyll a+b content (CABC), chlorophyll carotenoids content (CCC), and waterlogging score (WLS) and were measured for 2 years in both treatment and control conditions. WLS was determined based on plant survival and leaf chlorosis (1 = not affected by waterlogging, 9 = plants died from waterlogging) (**Supplementary Figure 1**) after drainage (Zhou, 2011). For chlorophyll content determination, the pooled upper second leaf samples of six plants per plot under waterlogging conditions and three plants per plot under control were collected after the last day of treatment. From each pooled tissue leaf sample per plot, three biological replicates of 50 mg leaf tissue each were incubated with methanol. The absorbance, at wavelengths 470, 653, and 666 nm, was read using a spectrophotometer (SpectraMax 190 Microplate Reader). The number of pigments was calculated according to the formula from Lichtenthaler and Wellburn (1983). The mean values (three plants from each replicate × three replicates) of each plot sampled were subjected to statistical analysis.

Statistical Analysis of Phenotypic Data

All data were analyzed using the statistical software JMP SAS version 14.1 (SAS Institute Inc., Cary, USA). The phenotypic data were analyzed using a mixed-effects model with genotype as a fixed effect, and year and replication nested within year as random effects. Least-squares means were estimated for waterlogging-treatment and control datasets within combined data across years. Pearson's correlation coefficient between pairs of traits was estimated to express the relationships between traits using the least-squares means across the combined years.

Genotyping

The barley collection was grown in the greenhouse to generate plant tissue for DNA extraction using a standard potting mix, standard photoperiod conditions (16 h light), and 70% humidity. Genomic DNA from each genotype was extracted from pooled leaf tissue samples of four seedlings per genotype using a Qiagen DNeasy Plant Mini Kit (Qiagen GMbH, Germany). Before normalization, the quality and quantity of the extracted DNA were verified using a NanoDrop 1000 spectrophotometer (Thermo Scientific, Wilmington, Delaware, USA) and agarose gel electrophoresis, respectively. The samples were genotyped using the Barley 50K iSelect SNP Array (Illumina Inc., San Diego, CA, USA), containing 44,040 working assays (Bayer et al., 2017). All these data is presented in **Supplementary Table 0**. The SNP markers were further filtered using thresholds for minor allele frequency (MAF) of 0.05, missing rate of 0.20, and heterozygosity of 0.01. The final, filtered set of 35,926 SNPs was subsequently used for GWAS. Genotypes showing more than 0.02 heterozygous loci and call rates below 0.95 were also excluded from further analysis. There were 3551, 5798, 5486, 3904, 6497, 4233, and 5017 SNPs located at chromosomes 1 to 7, respectively, with 1,440 markers of unknown position.

Population Structure, Kinship, and Linkage Disequilibrium Analyses

The population structure of the 247 barley genotypes, which represents the genetic similarity among genotypes, was assessed using the STRUCTURE program. Principal component analysis (PCA) (JMP Genomics 9.1) and neighbor-joining (NJ) (TASSEL

5.2.28) tree analysis were used as complementary approaches to confirm the results obtained using STRUCTURE. The STRUCTURE software version 2.3.4 (Pritchard et al., 2000) was used to estimate the most likely number of subpopulations (K) and the subpopulation coefficients (Q) by detecting allele frequency differences within the data and assigning individuals to those subpopulations based on analysis of likelihoods. A subset of 185 SNP markers, from the final filtered set of 35,926 SNP markers genotyped, were selected every ∼25,000,000 bp on each chromosome through the barley genome, to ensure that the sample was representative. A Bayesian-based analysis was run using the admixture ancestry model with correlated allele frequencies (Falush et al., 2003). The burn-in period was set at 100,000, and the Markov Chain Monte Carlo (MCMC) repetitions at 100,000. The number of assumed clusters (k) was set from $k = 1$–7, and for each k, five runs were performed separately. The output data from STRUCTURE were assessed using STRUCTURE Harvester (Earl and von Holdt, 2012), where the optimum number of subpopulations (K) was determined by the Evanno method (Evanno et al., 2005). The K value was considered to be optimum, while ΔK reaches the maximum. Data for the most likely number of determining clusters ($K = 3$) were run to correctly align the clusters labeled from all five replications in STRUCTURE to obtain Q coefficients. The Q matrix with the lowest variance for the most likely number of k populations was selected and used as the fixed covariate in GWAS models. PCA was performed in JMP Genomics version 9.1 (SAS Institute Inc., Cary, USA). A K matrix representing the proportion of shared alleles for all pairwise comparisons in each population was computed. The neighbor-joining phylogenetic tree was implemented in TASSEL version 5.2.28 (Bradbury et al., 2007), which uses simple parsimony substitution models and is displayed by Archaeopteryx software.

The pairwise kinship values (kinship K matrix) for the association panel were calculated using the Identity-by-Descent (IBD) method in JMP Genomics 9.1. The K matrix estimates the relationships among the lines using marker data, rather than pedigree information, and computes the relationship measures directly while also accounting for selection and genetic drift. This kinship matrix was used for the subsequent GWAS in JMP Genomics as a random factor. The kinship coefficient was calculated and plotted vs. its frequency in the association panel.

Linkage disequilibrium (LD) analysis of the whole-genome and each of the seven chromosomes was performed in JMP Genomics 9.1 using 35,926 SNPs. Squared correlation coefficients (r^2) were used to estimate the LD among the pairwise SNP markers using the maximum likelihood algorithm. To visualize the extent of LD, r^2 was plotted against the map distance (bp), and a smoothing spline was fitted ($\lambda = 100{,}000$). The baseline r^2 value was 0.1; an arbitrary value often used to describe LD decay (Zhu et al., 2008). The LD decay was estimated at the intersection point of the smoothing spline-fitting curve and the r^2 value and was considered to estimate the extent of LD in the genome. All LD values above this critical r^2 value were considered to be caused by genetic linkage.

Genome-Wide Association Mapping Analysis and SNP Markers Identification

A total of 247 spring barley genotypes were used in this study based on genotypic and phenotypic data availability. Genome-wide association (GWA) mapping was conducted on each group using a total of 35,926 SNPs in JMP Genomics 9.1. Based on the population structural analysis, the general linear model (GLM) and mixed linear model (MLM) were run to investigate best-fit models in the current study to search for SNP associations with the traits. The MLM model considers both population structure (Q) and relative kinship (K) effects, and showed the best approximation of the expected cumulative distribution of P-values, and therefore, more effective in controlling false positives, and it was used for GWAS. The population structure matrix (Q matrix) evaluated using STRUCTURE and the kinship matrix analyzed using JMP Genomics 9.1 were used for the model. Association analysis was performed for each trait in each treatment for the phenotypic mean value of 2016 and 2017. The estimated effects for each allelic class were obtained directly from the mixed linear model. Adjusted R^2 values were estimated from the linear regression model representing the percentage of phenotypic variation explained by the associated SNPs.

A GWAS threshold P-value of $< 1.6 \times 10^{-4}$ [$-\log_{10}(P\text{-value}) < 3.8$] was used for declaring significant-marker trait associations. They were based on the median of two threshold methods for determining significant P-values: a more stringent method of determining P-value (Wang et al., 2012), where the significance threshold is determined using the equation $\alpha = 1/m$ where m is the number of markers [$-\log_{10}(P\text{-value}) < 4.5$]; and a less stringent method (Chan et al., 2010) that is still widely accepted, where the bottom 0.1 percentile distribution of P-values is used as a threshold for significance [$-\log_{10}(P\text{-value}) < 3$]. Manhattan plots were constructed with the chromosome position on the X-axis against –log(P-value) of all SNPs, and quantile-quantile (QQ) plots of observed P-values were constructed against expected P-values using JMP Genomics 9.1. The distribution of the QQ plot was considered to select the best model for each trait. The optimum model for each variable was determined as the one with the QQ plot with a smaller deviation from the normal distribution.

The GWAS was performed with the control, waterlogging treatment and relative datasets. The relative dataset was calculated as the relative difference between trait performance at the control and waterlogging treatment conditions. The markers that were significantly associated were assigned to QTL regions based on the trait, their chromosomal positions, and the estimated LD decay (1.460 Mbp). The identified QTL regions under control conditions were compared with QTL reported in previous studies in barley dealing with agronomic traits (**Supplementary Table 1**), and the waterlogging treatment and relative datasets were compared with QTL reported in previous studies in barley for waterlogging stress tolerance-related traits (**Supplementary Table 2**). When possible, BarleyMap (http://floresta.eead.csic.es/barleymap/find/) was used to collect cM positions from the POPSEQ_2017 genome map (Mascher et al., 2013) for significant markers in our study, to enable an

approximate comparison between the physical and genetic map positions with the previous studies that reported QTL regions in genetic distance.

The phenotypic allele effect of each SNP locus, on the evaluated traits, was calculated through comparison of the average phenotypic value for each genotype for the specific allele with that of all genotypes (Mei et al., 2013).

Candidate Gene Prediction

We opted to investigate the genes in the vicinity of each significant marker-trait associations, using a pre-defined flanking window of 200-kb upstream and downstream, below the 1.46 Mb LD decay detected in the current barley mapping collection (Lei et al., 2019). The identified genes were manually screened for potential annotations. Predicted genes were extracted from the barley reference genome assembly (IBSC v2; Mascher et al., 2017). Annotations were downloaded from Ensembl (http://plants.ensembl.org/Hordeum_vulgare/Info/Index) and AmiGO Gene Ontology (amigo.geneontology.org). The role of the potential candidate genes in response to abiotic stresses, especially waterlogging, was further examined using published literature.

RESULTS

Phenotypic Data

Phenotypic variation was observed among genotypes for all traits in both control and waterlogging treatment (**Table 1**; **Supplementary Figure 2**). The frequency distribution of the genotypes for the investigated traits in the control and waterlogging treatment is presented in **Supplementary Figure 3**. In the control dataset, averaged over 2 years, BIO of the genotypes varied from 12.5 to 71.9 g, generated 5.4 to 22.4 SP, 9.2 to 312.2 GP, and weighted 0.2 to 14.5 g KWP. PH ranged from 18.5 to 95.8 cm, CABC varied from 0.89 to 1.54 mg/g leaf tissue, while CCC content varied from 0 to 0.17 mg/g leaf tissue (**Table 1**). After the exposure to waterlogging stress in the waterlogged dataset, averaged over 2 years, the genotypes varied in BIO from 1.7 to 36.3 g, generated 1.9 to 17.2 SP, 3.5 to 255.8 GP, and weighted 0.1 to 7.7 g KWP. PH ranged from 11.4 to 58.7 cm, CABC varied from 0.39 to 1.23 mg/g leaf tissue, while CCC varied from 0 to 0.12 mg/g leaf tissue (**Table 1**). As for WLS, the mean was 6.8, with a range from 4.7 to 8.8. Overall, for all genotypes, waterlogging stress reduced BIO, SP, GP, KWP, PH, CABC, and CCC by 72.1, 61.7, 67.5, 71.7, 45.1, 38.7, and 54.2%, respectively (**Supplementary Figure 3**). The coefficient of variation for the combined 2 years of data was higher for KWP (38.5 and 49.5% in control and waterlogging treatment, respectively), and lower for PH (16.0 and 17.4% in control and treatment conditions, respectively). There were highly significant ($P < 0.05$) genotypic differences both on individual and combined years for all traits except CABC and CCC (**Table 1**). The frequency distribution of all the traits generally fits a normal distribution (**Supplementary Figure 3**).

Correlations among traits under control and waterlogging treatment for 2016, 2017, and overall are shown in **Table 2**. In the combined 2 years of data, a negative correlation ($r = -0.14$ to -0.55; $P \leq 0.001$) was observed between the WLS and all the traits (**Table 2**). Yield component traits (BIO, SP, GP, KWP, and PH) had high correlations in both control ($r = 0.72$ to 0.94; $P \leq 0.001$) and waterlogging ($r = 0.50$–0.98; $P \leq 0.001$) treatment.

Population Structure, Kinship, and Linkage Disequilibrium Analyses

The Bayesian approach implemented in STRUCTURE revealed the presence of three subpopulations with the highest likelihood for $K = 3$ (**Supplementary Figure 4**) and partitioned the 247 genotypes into three principal groups composed of 96, 83, and 68 genotypes each. Furthermore, the PCA analysis displayed consistent results, confirming the existence of the three subpopulations in agreement with the population structure analysis by STRUCTURE (**Figure 1C**), with the first two coordinates accounting for 72.5% of the genotypic variation (**Figure 1A**). The phylogenetic analysis partitioned the 247 genotypes into three principal groups, following the results obtained with STRUCTURE and PCA analyses (**Figure 1B**). Subpopulation 1 is mainly composed of genotypes from the USA (21), Canada (16), and Australia (8), subpopulation 2 included genotypes mainly from China (23), Australia (10), Switzerland (9), and Ethiopia (8), while subpopulation 3 included genotypes from Canada (55), US (9), Australia (1), Brazil (1) China (1), and Japan (1).

Squared correlation coefficient (r^2) values among the marker pairs were used to estimate LD decay across all seven chromosomes (**Figure 2**) and each chromosome separately. The mean r^2 ranged from 0.0178 (chromosome 5H) to 0.0261 (chromosome 4H). The arbitrary baseline r^2 value was 0.1. The LD across all chromosomes decayed at 1,460,356 bp, whereas LD decay calculated for each chromosome separately ranged between 1,036,588 bp (chromosome 6H) and 2,290,772 bp (chromosome 1H). Based on the LD decay results, 35,926 SNPs (MAF > 0.05) will cover the entire barley genome and are adequate for GWAS with the assembled barley collection. Therefore, the mean window size of the QTL determined in this barley collection is ±1,460,356 bp from the highest peak of the significant marker-trait association.

Association Mapping Analysis

We performed GWAS using 35,926 SNPs (with MAF > 0.05) for the control and waterlogging treatment conditions, as well as the relative difference between them using the phenotypic overall field experiment (mean value of 2016 and 2017), and a threshold P-value of $< 1.6 \times 10^{-4}$ [$-\log_{10}(P\text{-value}) < 3.8$]. Manhattan plots showed the significance of markers associated with the evaluated traits for the overall control, waterlogging treatment and relative datasets in **Figures 3**–**5**. QQ plots displayed that the expected and observed P-values initially matched, but eventually, they were delineated and deviated to indicate a reasonable positive (**Supplementary Figures 5**–**7**). Thus, the GWAS analysis is reliable and not likely to give false negatives (**Figures 3**–**5**).

Control Dataset

In the overall control conditions, the GWAS analysis identified a total of 92 markers significantly associated with BIO (52

TABLE 1 | Mean values and standard deviations of waterlogging-related traits observed under control and waterlogging treatment in field conditions for 247 spring barley genotypes.

Trait	Year	Treatment	Mean	SD	Min	Max	Red.[a]	SE	CV (%)	G
BIO (g)	2016	Control	37.3	16.9	2.5	86.7	71.5%	1.07	45.3	***
		Waterlogged	10.6	6.5	2.4	50.4		0.41	61.3	***
	2017	Control	42.8	7.7	20.3	79.0	72.6%	0.49	18.0	***
		Waterlogged	11.7	5.4	0.4	31.3		0.34	45.9	***
	2016/17	Control	40.1	11.1	12.5	71.9	72.1%	0.70	27.7	***
		Waterlogged	11.2	5.0	1.7	36.3		0.32	44.6	***
SP	2016	Control	12.2	4.9	2.0	25.0	75.1%	0.31	39.9	***
(number)		Waterlogged	3.0	2.4	0.0	24.2		0.15	78.7	**
	2017	Control	13.8	2.7	6.7	23.4	49.8%	0.17	19.8	***
		Waterlogged	6.9	1.8	2.2	12.2		0.11	25.8	***
	2016/17	Control	13.0	3.0	5.4	22.4	61.7%	0.19	23.3	***
		Waterlogged	5.0	1.6	1.9	17.2		0.10	32.3	***
GP	2016	Control	150.8	84.3	2.0	430.5	90.0%	5.35	55.9	***
(number)		Waterlogged	15.0	26.2	0.0	329.0		1.7	174.4	***
	2017	Control	167.5	49.0	13.4	362.0	47.1%	3.11	29.3	***
		Waterlogged	88.6	35.1	1.0	196.9		2.2	39.6	***
	2016/17	Control	159.2	57.0	9.2	312.2	67.5%	3.62	35.8	***
		Waterlogged	51.8	24.2	3.5	255.8		1.5	46.8	***
KWP (g)	2016	Control	6.3	3.7	0.0	19.0	92.5%	0.23	58.6	***
		Waterlogged	0.5	0.9	0.0	9.1		0.1	182.8	***
	2017	Control	6.6	2.2	0.3	16.1	51.9%	0.14	33.4	***
		Waterlogged	3.2	1.4	0.0	8.6		0.1	43.8	***
	2016/17	Control	6.4	2.5	0.2	14.5	71.7%	0.16	38.5	***
		Waterlogged	1.8	0.9	0.1	7.7		0.1	49.5	***
PH (cm)	2016	Control	73.5	13.7	17.5	101.3	54.8%	0.87	18.7	***
		Waterlogged	33.2	10.7	12.3	65.0		0.7	32.3	NS
	2017	Control	72.5	11.0	19.5	104.0	35.3%	0.70	15.2	***
		Waterlogged	46.9	8.6	7.8	70.5		0.5	18.4	***
	2016/17	Control	73.0	11.6	18.5	95.8	45.1%	0.74	16.0	***
		Waterlogged	40.1	7.0	11.4	58.7		0.4	17.4	***
CABC	2016	Control	1.13	0.2	0.66	1.55	41.3%	0.01	13.75	NS
(mg/g leaf tissue)		Waterlogged	0.66	0.3	0.03	1.39		0.02	38.12	NS
	2017	Control	1.39	0.1	0.96	1.67	36.6%	0.01	9.63	NS
		Waterlogged	0.88	0.2	0.39	1.47		0.01	21.79	***
	2016/17	Control	1.26	0.1	0.89	1.54	38.7%	0.01	8.41	NS
		Waterlogged	0.77	0.2	0.39	1.23		0.01	21.22	**
CCC	2016	Control	0.06	0.02	0.00	0.12	10.5%	0.00	42.75	NS
(mg/g leaf tissue)		Waterlogged	0.05	0.02	0.00	0.09		0.00	34.52	NS
	2017	Control	0.14	0.03	0.01	0.22	71.8%	0.00	24.02	NS
		Waterlogged	0.04	0.03	0.00	0.16		0.00	82.08	NS
	2016/17	Control	0.10	0.03	0.00	0.17	54.2%	0.00	33.38	NS
		Waterlogged	0.04	0.02	0.00	0.12		0.00	58.30	NS
WLS	2016	Waterlogged	6.9	1.2	3.3	9.0		0.08	17.5	*
(1–9 rating)	2017	Waterlogged	6.7	0.7	4.7	9.0		0.05	10.8	***
	2016/17	Waterlogged	6.8	0.8	4.7	8.8		0.05	12.0	***

BIO, biomass; SP, spikes per plant; GP, grains per plant; KWP, kernel weight per plant; PH, plant height; CABC, chlorophyll a+b; CCC, carotenoids content; WLS, waterlogging score; SD, standard deviation; Red., Reduction; SE, standard error; CV, coefficient of variance; G, genotypic effect.

[a]*Reduction ratio of all genotypes relative to control.*

**Significant at P ≤ 0.05; **significant at P ≤ 0.01; ***significant at P ≤ 0.001; NS not significant.*

TABLE 2 | Pearson's phenotypic correlation coefficients among mean variables (least-squares entry means) of traits for control and waterlogging treatment measured in the spring barley collection in field conditions.

	Year		Control						
			BIO	**SP**	**GP**	**KWP**	**PH**	**CABC**	**CCC**
Waterlogging treatment	**2016**	BIO		0.83***	0.76***	0.70***	0.51***	−0.05 *NS*	0.07 *NS*
		SP	0.70***		0.87***	0.79***	0.44***	0.00 *NS*	0.07 *NS*
		GP	0.71***	0.79***		0.94***	0.48***	0.03 *NS*	0.04 *NS*
		KWP	0.71***	0.81***	0.96***		0.48***	0.02 *NS*	0.05 *NS*
		PH	0.25***	0.46***	0.28***	0.27***		−0.12**	0.15***
		CABC	0.23***	0.24***	0.12**	0.12**	0.50**		−0.55***
		CCC	0.18***	0.10**	0.08*	0.06 *NS*	0.29***	0.25***	
		WLS	−0.51***	−0.53***	−0.39***	−0.41***	−0.61***	−0.47***	−0.28***
	2017	BIO		0.71***	0.76***	0.80***	0.51***	−0.07 *NS*	0.10**
		SP	0.61***		0.73***	0.69***	0.22***	−0.01 *NS*	0.11**
		GP	0.61***	0.76***		0.96***	0.35***	−0.02 *NS*	0.04 *NS*
		KWP	0.55***	0.69***	0.96***		0.37***	−0.03 *NS*	0.07*
		PH	0.51***	0.32***	0.51***	0.49***		−0.03 *NS*	0.08*
		CABC	−0.22***	0.07 *NS*	0.12**	0.19***	−0.05 *NS*		0.02 *NS*
		CCC	0.05 *NS*	0.08*	0.05 *NS*	0.05 *NS*	0.04 *NS*	−0.11**	
		WLS	−0.52***	−0.31***	−0.32***	−0.27***	−0.42***	0.27***	−0.12**
	2016/17	BIO		0.79***	0.76***	0.72***	0.49***	0.02 *NS*	0.15***
		SP	0.61***		0.83***	0.76***	0.35***	0.06*	0.15***
		GP	0.53***	0.83***		0.94***	0.43***	0.05 *NS*	0.09**
		KWP	0.50***	0.80***	0.98***		0.44***	0.01 *NS*	0.06*
		PH	0.35***	0.55***	0.53***	0.52***		−0.10***	0.04 *NS*
		CABC	0.10***	0.30***	0.28***	0.30***	0.42***		0.18***
		CCC	0.08**	0.00 *NS*	−0.04 *NS*	−0.03 *NS*	0.06*	0.00 *NS*	
		WLS	−0.51***	−0.44***	−0.31***	−0.29***	−0.55***	−0.29***	−0.14***

Control above diagonal, waterlogging treatment below diagonal. The correlations are estimated by the REML method.
BIO, biomass; SP, spikes per plant; GP, grains per plant; KWP, kernel weight per plant; PH, plant height; CABC, chlorophyll a+b; CCC, carotenoids content.
**Significant at $P \leq 0.05$; **significant at $P \leq 0.01$; ***significant at $P \leq 0.001$; NS not significant.*

markers), SP (18 markers), GP (23 markers), KWP (15 markers), and PH (62 markers), with some markers associated with multiple traits (**Supplementary Table 3**). Based on their position on chromosomes, these 92 significant markers mapped on 28 QTL regions on chromosomes 2H, 3H, 5H, 6H, and 7H, with each QTL region consisting of 1 to 34 markers, which included two regions for KWP; four regions for SP and GP; 12 regions for BIO; and 20 regions for PH (**Figure 3**; **Supplementary Table 3**). Some genomic regions were associated with multiple traits, indicating possible shared QTL between traits. For BIO in the control conditions, we found six genomic regions, out of 12, consisting of clusters of significant markers that mapped at 27.8, 29.1, 515.6, 542.4, and 547.4 Mbp on chromosome 2H, and at 600.9 Mbp on 5H (**Table 3**; **Supplementary Table 3**; **Figure 3**); each region consisted of clusters from 2 to up to 34 markers and explained on average from 6.2 to 12.3% of the phenotypic variation. Chromosome 2H consisted of the highest number of markers significantly associated with BIO (52 SNPs), of which BOPA2_12_30872 had the lowest *P*-value (6.3×10^{-12}) with an allele effect size of 6.8 that individually explained 17.7% of phenotypic variation for BIO (**Supplementary Table 3**). The three genomic regions associated with SP in the control conditions were mapped at 29.7 Mbp on chromosome 2H, at 634.9 Mbp on chromosome 3H, and 35.4 Mbp on chromosome 6H and accounted on average for 5.8, 6.8, 6.9, and 6.4% of the phenotypic variation, respectively (**Table 3**; **Supplementary Table 3**). For GP in the control condition, we found two genomic regions at 29.7 Mbp (clusters of 14 SNPs) on chromosome 2H and 634.8 Mbp (7 SNPs) on 3H. On average, each genomic region explained between 6.9 and 7.1% of the phenotypic variation (**Table 3**; **Supplementary Table 3**). The two genomic regions associated with KWP in the control conditions were mapped at 29.7 Mbp on 2H (12 SNPs), and at 634.8 Mbp on 3H (3 SNPs). Each region explained, on average, from 6.1 to 6.7% of the phenotypic variation across the 2 years (**Supplementary Table 3**). For PH in the control conditions, we found nine genomic regions consisting of clusters of at least two significant markers that mapped at 28.5 Mbp (34 SNPs), 518.3

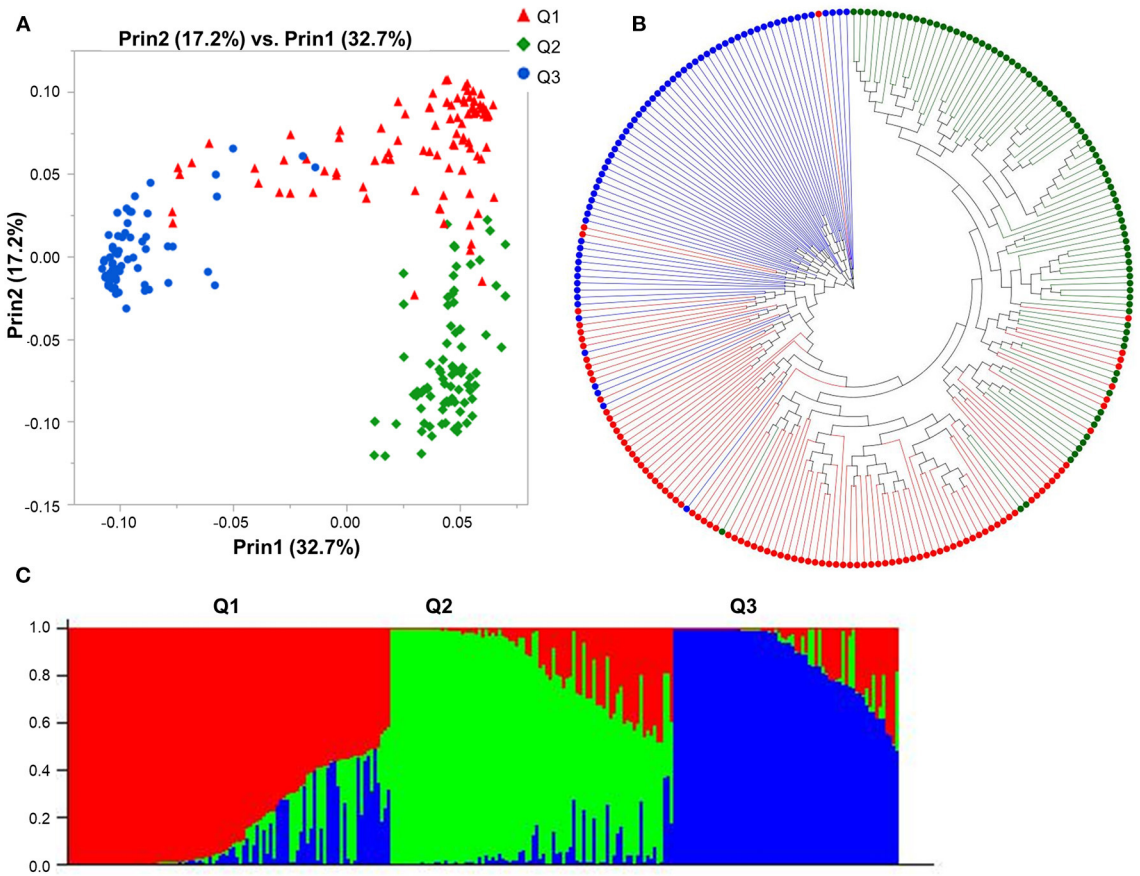

FIGURE 1 | Model-based populations of spring barley collection: **(A)** Two-dimension distribution analyzed by principal component analysis (PCA) by JMP Genomics 9.1, **(B)** phylogenetic tree constructed by neighbor-joining (NJ) of genetic distance by TASSEL 5.2.28, and **(C)** Classification of three populations using STRUCTURE 2.3.4. The color code indicates the distribution of the accessions to different populations (Q1: red, Q2: green, Q3: blue) consistent in **(A–C)**.

Mbp (2 SNPs), 523.4 Mbp (2 SNPs), 550.8 Mbp (2 SNPs), 723.7 Mbp (2 SNPs), and 727.6 Mbp (2 SNPs) on 2H, at 600.9 Mbp (2 SNPs), 613.3 Mbp on 5H (2 SNPs), and 75.1 Mbp on 7H (2 SNPs). Each region individually explained from 5.8 to 11.4% of the phenotypic variation (**Supplementary Table 3**).

Under control conditions, six marker-trait associations representing genomic regions were associated with different traits (**Table 3**). On chromosome 2H, the marker JHI-Hv50k-2016-69385 at 19.0 Mbp was associated with the traits BIO and PH, with similar effects in phenotype (6.9 and 5.8% phenotypic variation, respectively); the marker JHI-Hv50k-2016-72991 at 27.8 Mbp was coincidental for BIO, SP, and PH, although with different effects in each trait (from 5.8 to 11.9% phenotypic variation); the marker JHI-Hv50k-2016-73691 located at 29.6 Mbp was associated with the traits SP, GP, and KWP; and the marker JHI-Hv50k-2016-94875 at 496.6 Mbp was shared by the traits BIO and PH (6.9 and 5.8% phenotypic variation, respectively). On chromosome 3H, the traits GP and KWP were associated with the same marker JHI-Hv50k-2016-205562 located at 634.8 Mbp, with 8.2 and 6.8% phenotypic variation, respectively (**Table 3**). Finally, on chromosome 5H, the traits BIO and PH were associated with the marker JHI-Hv50k-2016-336773 mapped at 600.9 Mbp with similar effects for the two traits (6.2 and 7.5% phenotypic variation, respectively).

Waterlogging Treatment Dataset

In the overall waterlogging treatment conditions, the GWAS analysis identified a total of 63 markers significantly associated with BIO (33 markers), SP (11 markers), GP (10 markers), KWP (20 markers), PH (4 markers), and WLS (25 markers), with some markers associated with multiple traits (**Supplementary Table 4**). Based on their position on chromosomes, these 63 significant SNPs were assigned to 24 QTL regions on chromosomes 1H, 2H, 3H, 4H, 5H, 6H, and 7H, with each region consisting of 1–30 markers, which included three regions for BIO; seven regions for GP; nine regions each for SP and KWP, four regions for PH, and five for WLS (**Table 4**; **Figure 4**). Some QTL regions were associated with multiple traits, indicating possible shared QTL between traits. For BIO in the waterlogging treatment conditions, three genomic regions were detected at 27.8, 28.3, and 516.6 Mbp on chromosome 2H. The genomic region at 28.3 Mbp consisted of the highest number of markers significantly associated with BIO (32 SNPs),

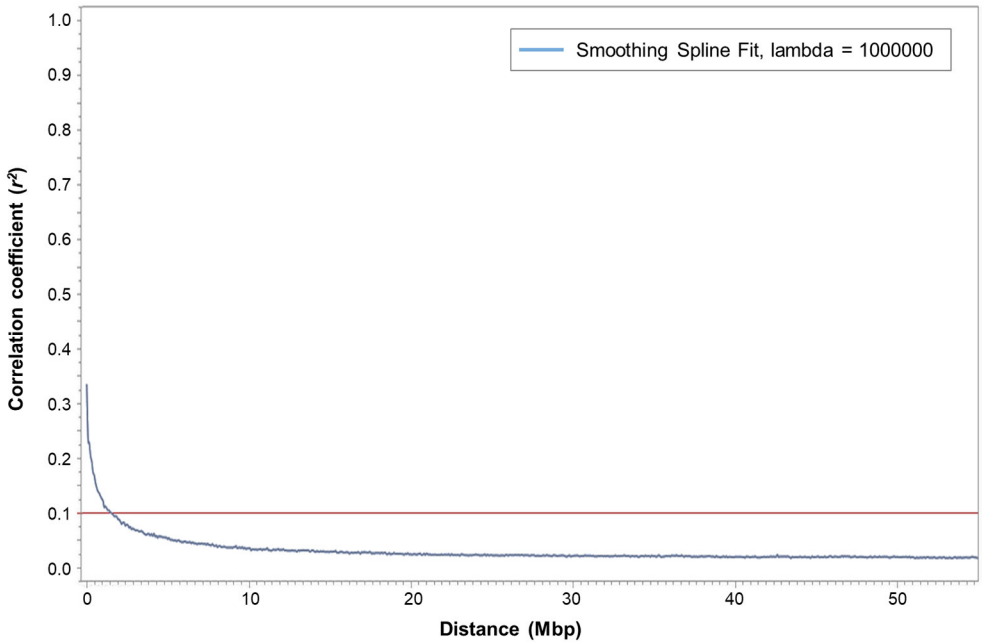

FIGURE 2 | Plot of pairwise SNP linkage disequilibrium (LD) r^2 value as a function of inter-marker genetic distances (Mbp) of 247 spring barley genotypes. The blue curve represents the smoothing spline regression model fit to LD decay. The red line represents the baseline r^2 value at 0.1. The intersection of the fitted smoothing spline and r^2 was observed at around 1,460,356 bp.

explaining on average 9.5% of the phenotypic variation of the trait (**Table 4**; **Figure 4**). The most significant SNP marker, BOPA2_12_30872 had the lowest *P*-value (3.3×10^{-8}) with an allele effect size of 2.8 that individually explained 11.8% of phenotypic variation for BIO (**Supplementary Table 4**). For SP in the waterlogging treatment conditions, we found two genomic regions consisting of clusters of two significant markers that mapped at 662.0 Mbp on 2H, and at 371,3 Mbp on 4H. Each region with an allele effect size of 1.2 individually explained from 7.4 to 8.57% of the phenotypic variation (**Table 4**). Clusters of two and three SNPs on chromosomes 2H at 29.6 Mbp and 5H at 568 Mbp, respectively, were significantly associated with GP in the waterlogging treatment conditions, which on average, accounted for 6.3 and 7.1% of phenotypic variation (**Table 4**; **Supplementary Table 4**). For KWP in the waterlogging treatment conditions, we found two genomic regions with at least two SNPs, at 16.8 Mbp (2 SNPs), and 29.7 Mbp on chromosome 2H (11 SNPs). On average, each genomic region explained between 6.1 and 6.9% of the phenotypic variation (**Supplementary Table 4**). The three genomic regions, with more than one SNP, associated with WLS in the waterlogging treatment conditions were found at 29.1 Mbp (17 SNPs) on chromosome 2H, and 0.37 and 569.8 Mbp (four and two SNPs, respectively) on 4H (**Table 4**; **Supplementary Table 4**; **Figure 4**); each region explained on average from 5.7 to 7.4% of the phenotypic variation. Chromosome 2H consisted of the highest number of markers significantly associated with WLS, of which BOPA2_12_30872 had the lowest *P*-value (7.5×10^{-6}) with an allele effect size of 0.4 that individually explained 7.9% of phenotypic variation for WLS (**Supplementary Table 4**).

Eight marker-trait associations associated with different traits were found in the waterlogging treatment conditions (**Table 4**). On chromosome 2H, the marker JHI-Hv50k-2016-68186 located at 16.8 Mbp was associated with the traits GP and KWP, although with different effects in each trait (from 6.1 to 7.6% phenotypic variation); the marker BOPA2_12_30872 located at 29.1 Mbp was coincidental for the traits BIO and WLS, with different effects on each trait (from 7.9 to 11.8% phenotypic variation); and the traits GP and KWP were associated to the same marker JHI-Hv50k-2016-73689 at 29.6 Mbp. On chromosome 4H, the traits SP and WLS were associated with the marker JHI-Hv50k-2016-225852 at 0.37 Mbp (7.3 and 6.8% phenotypic variation, respectively); and GP and KWP were associate to the same marker JHI-Hv50k-2016-249670 located at 512.9 Mbp (∼6.1% phenotypic variation). On chromosome 5H, the traits SP and GP were associated with the marker JHI-Hv50k-2016-322832 regions at 569.3 Mbp; and the marker BOPA2_12_11245 at 579.3 Mbp was coincidental for the traits SP, GP, and KWP, with a similar effect for the three traits, ∼6.2% phenotypic variation (**Table 4**). On chromosome 7H, the marker JHI-Hv50k-2016-449124 located at 13.6 Mbp was coincidental for the traits GP and KWP, with a similar effect.

Additionally, the analysis showed three markers on chromosome 2H co-localized in both control and waterlogging treatment conditions (**Tables 3**, **4**). The marker JHI-Hv50k-2016-72991 located at 27.8 Mbp was found to be associated with BIO, SP, and PH under control, and with BIO under waterlogging treatment conditions; the marker BOPA2_12_30872 at 29.1 Mbp

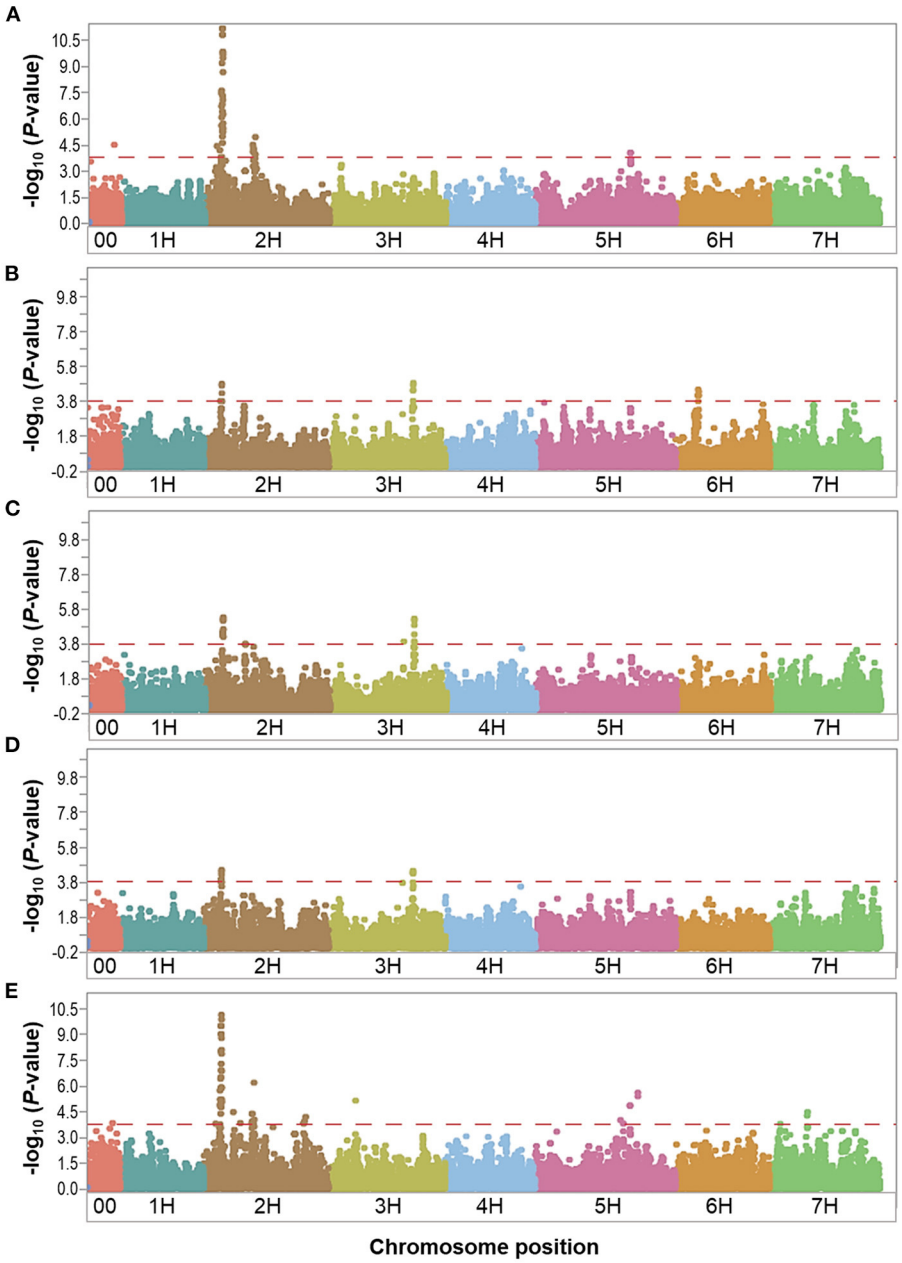

FIGURE 3 | Manhattan plots resulting from the SNP-based GWAS in overall control under field conditions. Manhattan plots for Biomass (BIO), Spikes per plant (SP), Grains per plant (GP), Kernel weight per plant (KWP), and Plant height (PH) are shown in **(A–E)**, respectively, and the x-axis shows SNP loci along the seven barley chromosomes. The horizontal red line shows the genome-wide significance threshold *P*-value of 1.6×10^{-4} or $-\log_{10}$ (*P*-value) value of 3.8. GWAS was performed using the MLM (Q + K) model in JMP Genomics for the field traits.

was identified in BIO under control, and BIO and WL under waterlogging treatment conditions; and the marker JHI-Hv50k-2016-80986 located at 73.5 Mbp was identified in PH under both control and waterlogging treatment conditions.

Relative Dataset

In order to find chromosomal regions that were significantly associated with waterlogging tolerance response, we analyzed the relative difference between the control and waterlogging treatment conditions. In the overall relative dataset, the GWAS analysis identified a total of 51 markers significantly associated with BIO (1 SNP), SP (17 SNPs), KWP (4 SNPs), PH (24 SNPs), and WLS (25 SNPs), with some markers associated with multiple traits (**Supplementary Table 5**). No significant markers were detected for GP in the relative dataset, unlike in the control and waterlogging treatment datasets. Based on

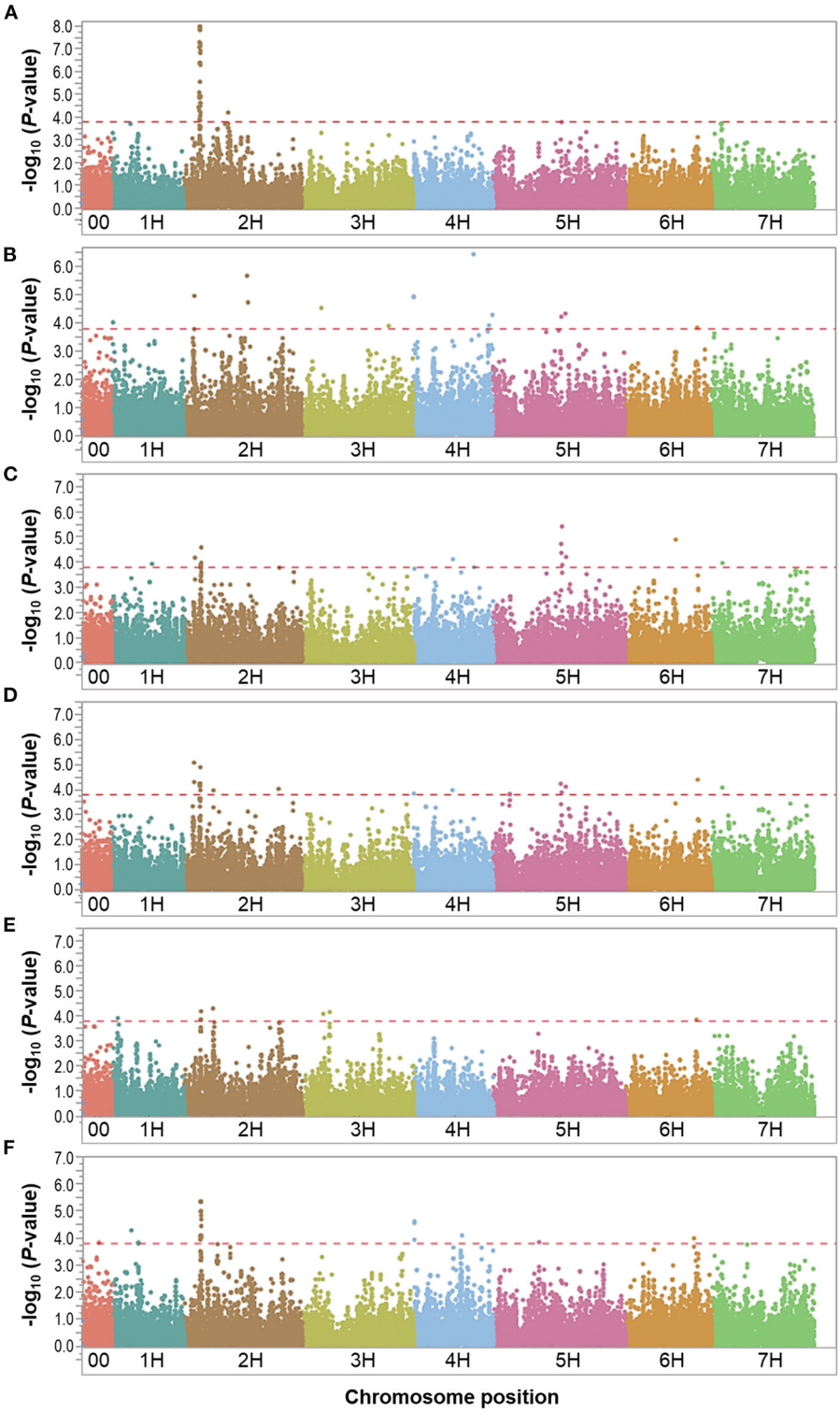

FIGURE 4 | Manhattan plots resulting from the SNP-based GWAS in waterlogging treatment under field conditions. Manhattan plots for Biomass (BIO), Spikes per plant (SP), Grains per plant (GP), Kernel weight per plant (GWP), Plant height (PH), and Waterlogging score (WLS) are shown in **(A–F)**, respectively, and the x-axis shows SNP loci along the seven barley chromosomes. The horizontal red line shows the genome-wide significance threshold *P*-value of 1.6×10^{-4} or $-\log_{10}$ (*P*-value) value of 3.8. GWAS was performed using the MLM (Q + K) model in JMP Genomics for the field traits.

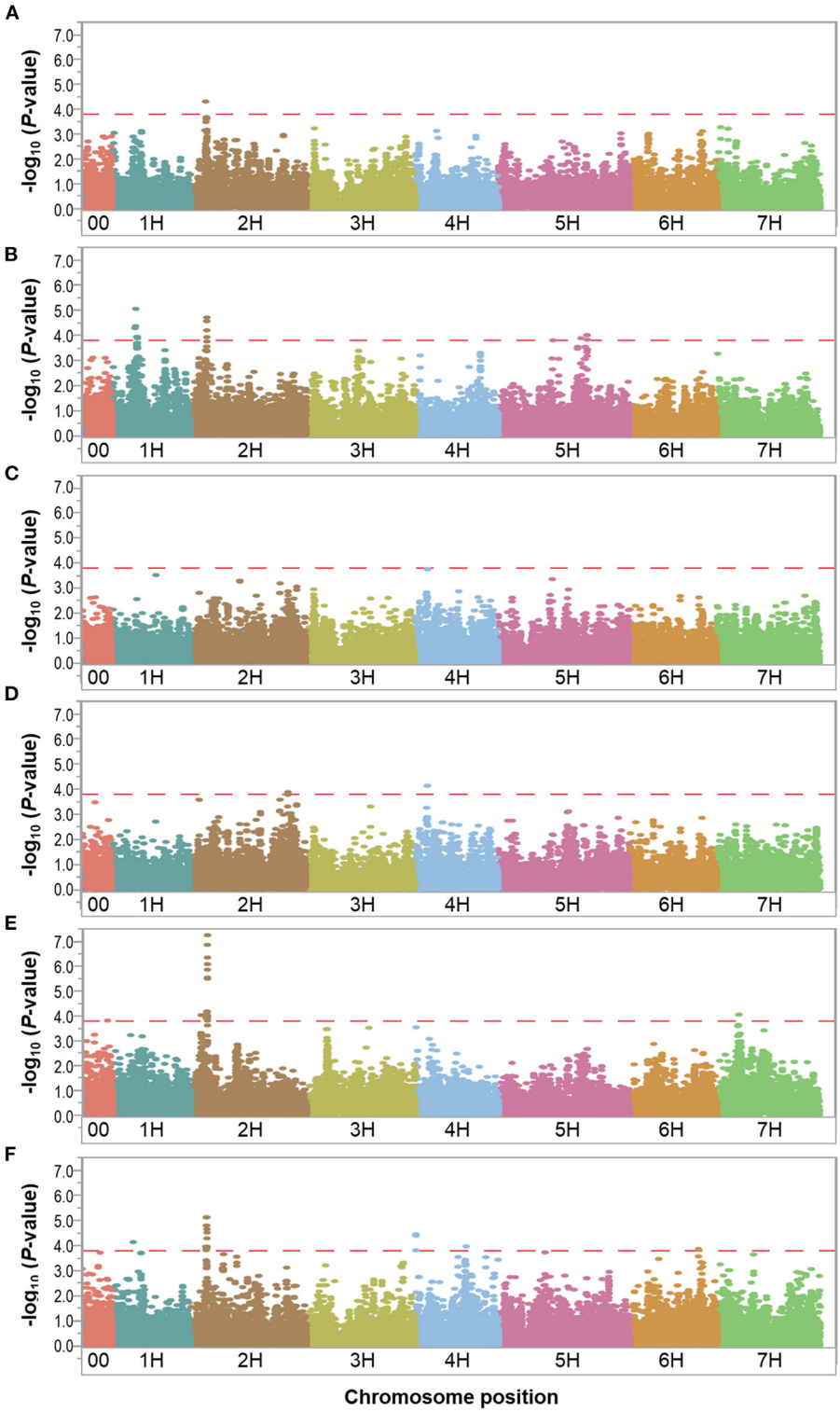

FIGURE 5 | Manhattan plots resulting from the SNP-based GWAS identified in the relative dataset. Manhattan plots for Biomass (BIO), Spikes per plant (SP), Grains per plant (GP), Kernel weight per plant (GWP), Plant height (PH), and Waterlogging score (WLS) are shown in **(A–F)**, respectively, and the x-axis shows SNP loci along the seven barley chromosomes. The horizontal red line shows the genome-wide significance threshold *P*-value of 1.6×10^{-4} or $-\log_{10}$ (*P*-value) value of 3.8. GWAS was performed using the MLM (Q + K) model in JMP Genomics for the field traits.

TABLE 3 | List of significant ($P < 1.6 \times 10^{-4}$) marker-trait associations detected by GWAS using the MLM (Q + K) model in JMP Genomics and favorable alleles (bold) for the assessed traits in the overall control conditions.

Trait	Marker[a]	Ch	Physical position (bp)[b]	Genetic position (cM)[c]	*P*-value	R^2 (%)[d]	MAF	Allele[e]	Additive effect
BIO	JHI-Hv50k-2016-69385[f,*]	2H	19,064,497	13.31	3.60E-05	6.90	0.11	**T**/G	−6.16
	JHI-Hv50k-2016-71792[f]	2H	23,485,824	15.37	6.10E-05	6.40	0.19	**T**/C	−3.19
	JHI-Hv50k-2016-72991*	2H	27,836,916	18.91	3.40E-08	11.90	0.10	**A**/T	−5.85
	BOPA2_12_30872	2H	29,124,597	19.90	6.30E-12	17.70	0.18	**A**/G	−6.77
	JHI-Hv50k-2016-94875[f,*]	2H	496,673,313	55.01	3.00E-05	6.90	0.08	T/**C**	−5.23
	BOPA1_ABC08774-1-1-752[f]	2H	508,786,535		7.60E-05	6.30	0.05	A/**C**	−6.07
	JHI-Hv50k-2016-95073[f]	2H	515,576,575	58.64	4.40E-05	6.70	0.08	T/**C**	−4.79
	SCRI_RS_127347[f]	2H	519,110,344	58.64	5.80E-05	6.40	0.11	**T**/C	−4.56
	JHI-Hv50k-2016-97672[f]	2H	542,384,101	59.42	1.10E-04	6.00	0.06	A/**T**	−5.98
	JHI-Hv50k-2016-98186[f]	2H	547,420,281	59.42	1.10E-04	6.00	0.06	**C**/G	−5.98
	JHI-Hv50k-2016-336773*	5H	600,914,687	126.30	8.70E-05	6.20	0.07	**A**/T	−5.97
	JHI-Hv50k-2016-336814	5H	600,979,263		8.70E-05	6.20	0.07	T/**G**	−5.97
SP	JHI-Hv50k-2016-72991*	2H	27,836,916	18.91	1.50E-04	5.80	0.10	**A**/T	−1.25
	JHI-Hv50k-2016-73691*	2H	29,669,343		1.60E-05	7.60	0.15	A/**G**	−1.40
	JHI-Hv50k-2016-205634	3H	634,932,524	109.80	1.40E-05	7.50	0.35	**T**/C	1.09
	JHI-Hv50k-2016-382988	6H	35,396,724	43.77	3.40E-05	6.90	0.25	A/**G**	−0.96
GP	JHI-Hv50k-2016-73691*	2H	29,669,343		4.70E-06	8.40	0.15	A/**G**	−26.84
	JHI-Hv50k-2016-88492[f]	2H	134,404,110	55.01	1.50E-04	5.80	0.13	**A**/G	25.01
	JHI-Hv50k-2016-200577	3H	609,227,175	90.16	1.20E-04	6.00	0.27	**A**/G	15.27
	JHI-Hv50k-2016-205562*	3H	634,801,729	108.90	5.60E-06	8.20	0.44	**T**/C	17.77
KWP	JHI-Hv50k-2016-73691*	2H	29,669,343		3.30E-05	7.00	0.15	A/**G**	−1.06
	JHI-Hv50k-2016-205562*	3H	634,801,729	108.90	3.60E-05	6.80	0.44	**T**/C	0.70
PH	JHI-Hv50k-2016-69385[f,*]	2H	19,064,497	13.31	1.60E-04	5.80	0.11	**T**/G	−6.65
	JHI-Hv50k-2016-72991*	2H	27,836,916	18.91	1.80E-06	9.00	0.10	**A**/T	−5.90
	JHI-Hv50k-2016-73085*	2H	28,455,236	18.91	1.10E-05	7.80	0.41	**T**/C	9.61
	JHI-Hv50k-2016-80986[f]	2H	73,504,389	49.73	3.30E-05	6.90	0.07	T/G	−7.98
	JHI-Hv50k-2016-86347[f]	2H	112,364,666		1.40E-04	5.80	0.08	T/**C**	5.80
	JHI-Hv50k-2016-94875[f,*]	2H	496,673,313	55.01	1.40E-04	5.80	0.08	T/**C**	−5.81
	JHI-Hv50k-2016-95379[f]	2H	518,293,896	58.00	4.10E-05	6.70	0.08	A/**G**	−6.81
	JHI-Hv50k-2016-95777[f]	2H	523,378,213	58.64	1.20E-04	5.90	0.12	**A**/T	−6.05
	JHI-Hv50k-2016-98273[f]	2H	548,916,905		6.30E-07	9.70	0.06	T/**C**	−8.85
	JHI-Hv50k-2016-98501[f]	2H	550,839,094	59.35	9.00E-05	6.20	0.18	C/**G**	−4.88
	JHI-Hv50k-2016-127739	2H	723,652,876	122.90	1.50E-04	5.80	0.13	**T**/G	−4.62
	JHI-Hv50k-2016-129870	2H	727,578,152	125.20	6.20E-05	6.40	0.07	**A**/G	−7.42
	BOPA2_12_10532[f]	3H	67,560,907	45.82	7.00E-06	8.00	0.05	C/**G**	−7.71
	JHI-Hv50k-2016-330643	5H	587,449,015	114.70	9.60E-05	6.10	0.08	**T**/C	−5.60
	JHI-Hv50k-2016-332746	5H	591,637,968	120.10	1.50E-04	5.90	0.07	**A**/G	−7.35
	JHI-Hv50k-2016-336773*	5H	600,914,687	126.30	1.40E-05	7.50	0.07	**A**/T	−8.13
	BOPA2_12_31234[f]	5H	613,268,086	134.70	2.40E-06	8.80	0.07	A/**G**	−7.10
	JHI-Hv50k-2016-447227[f]	7H	11,309,509	7.78	1.60E-04	5.70	0.05	**A**/T	−6.84
	JHI-Hv50k-2016-468495[f]	7H	71,962,797	58.04	5.20E-05	6.60	0.10	A/**T**	−5.03
	JHI-Hv50k-2016-468869[f]	7H	75,059,390	59.80	3.30E-05	6.90	0.09	**A**/G	−5.21

BIO, biomass; SP, spikes per plant; GP, grains per plant; KWP, kernel weight per plant; PH, plant height; Ch, chromosome number; MAF, minor allele frequency.

[a] *The marker with the highest R^2 in the genomic region is presented.*

[b] *Base pair positions of the marker in the chromosome based on a high-quality reference genome assembly for barley (Hordeum vulgare L.) (Mascher et al., 2017).*

[c] *Genetic marker positions (cM) of the marker obtained from the POPSEQ_2017 genome map in BarleyMap (http://floresta.eead.csic.es/barleymap/find/) (Mascher et al., 2013).*

[d] *R^2 (%) indicates the percentage of phenotypic variation explained by the significant marker.*

[e] *Allele that is in bold text is the favorable allele for the trait assessed.*

[f] *Marker-trait associations that have different positions than the previously identified QTL for yield and yield-related traits published on barley under unstressed conditions.*

**Putative QTL that may be associated with multiple traits.*

TABLE 4 | List of significant ($P < 1.6 \times 10^{-4}$) marker-trait associations detected by GWAS using the MLM (Q + K) model in JMP Genomics and favorable alleles (bold) for assessed traits in the overall waterlogging treatment conditions.

Trait	Marker[a]	Ch	Physical position (bp)[b]	Genetic position (cM)[c]	*P*-value	R^2 (%)[d]	MAF	Allele[e]	Additive effect
BIO	JHI-Hv50k-2016-72991	2H	27,836,916	18.99	1.30E-05	7.6	0.10	**A**/T	−2.51
	BOPA2_12_30872*	2H	29,124,597	19.90	3.30E-08	11.8	0.18	**A**/G	−2.75
	JHI-Hv50k-2016-95223	2H	516,581,410	57.72	8.80E-05	6.2	0.10	**T**/C	−2.42
SP	JHI-Hv50k-2016-3532	1H	3,453,791	4.96	1.30E-04	6.2	0.07	A/**G**	0.85
	JHI-Hv50k-2016-68266 [f]	2H	16,823,564	11.40	1.70E-05	7.4	0.08	**A**/G	1.34
	JHI-Hv50k-2016-109151	2H	662,018,769	82.51	3.70E-06	8.5	0.06	**A**/G	1.24
	JHI-Hv50k-2016-161633	3H	32,637,255	37.04	4.20E-05	6.7	0.06	A/**T**	0.99
	JHI-Hv50k-2016-225852[f,*]	4H	371,267	0.71	1.80E-05	7.3	0.06	T/**C**	1.17
	JHI-Hv50k-2016-262685	4H	607,200,114	85.84	7.40E-07	9.6	0.06	A/**G**	1.31
	JHI-Hv50k-2016-276624[f]	4H	645,759,577	117.30	7.20E-05	6.3	0.05	**T**/C	1.20
	JHI-Hv50k-2016-322832*	5H	569,308,558	97.51	8.20E-05	6.2	0.05	**A**/G	1.13
	BOPA2_12_11245*	5H	579,324,077		6.50E-05	6.4	0.06	C/**G**	1.07
GP	JHI-Hv50k-2016-68186*	2H	16,813,000	11.40	9.20E-05	6.1	0.11	**T**/C	10.47
	JHI-Hv50k-2016-73689*	2H	29,669,242		3.80E-05	6.8	0.14	A/**G**	−11.29
	JHI-Hv50k-2016-249670[f,*]	4H	512,990,076	54.32	1.10E-04	6.2	0.06	A/**G**	18.71
	JHI-Hv50k-2016-322832*	5H	569,308,558	97.51	6.50E-06	8.1	0.05	**A**/G	18.67
	BOPA2_12_11245*	5H	579,324,077		8.60E-05	6.2	0.06	C/**G**	15.07
	JHI-Hv50k-2016-410329[f]	6H	492,880,745	65.93	2.00E-05	7.3	0.07	**A**/C	18.63
	JHI-Hv50k-2016-449124[f,*]	7H	13,658,217	11.54	1.50E-04	5.8	0.35	**T**/C	7.27
KWP	JHI-Hv50k-2016-68186[f,*]	2H	16,813,000	11.40	1.30E-05	7.6	0.11	**T**/C	0.43
	JHI-Hv50k-2016-73689*	2H	29,669,242		2.00E-05	7.2	0.14	A/**G**	−0.44
	JHI-Hv50k-2016-82113	2H	79,456,923	49.73	1.40E-04	5.8	0.13	**T**/G	−0.34
	JHI-Hv50k-2016-127867	2H	724,202,574	120.80	1.30E-04	5.9	0.35	**A**/G	−0.26
	JHI-Hv50k-2016-249670[f,*]	4H	512,990,076	54.32	1.40E-04	6.0	0.06	A/**G**	0.68
	JHI-Hv50k-2016-322288	5H	568,058,046	97.51	8.10E-05	6.2	0.06	T/**G**	0.58
	BOPA2_12_11245*	5H	579,324,077		1.00E-04	6.0	0.06	C/**G**	0.55
	JHI-Hv50k-2016-424341[f]	6H	562,861,599	105.10	5.70E-05	6.5	0.06	**T**/G	0.56
	JHI-Hv50k-2016-449124[f,*]	7H	13,658,217	11.54	1.10E-04	6.0	0.35	**T**/C	0.27
PH	JHI-Hv50k-2016-73570	2H	29,307,953		9.00E-05	6.2	0.12	T/**C**	−3.30
	JHI-Hv50k-2016-80986	2H	73,504,389	49.73	7.00E-05	6.3	0.07	**T**/G	−5.31
	BOPA2_12_10968	3H	34,959,733	37.04	1.10E-04	6.0	0.06	**A**/G	−4.08
	JHI-Hv50k-2016-165725	3H	78,242,146		9.50E-05	6.2	0.30	**A**/G	3.61
WLS	JHI-Hv50k-2016-19217	1H	61,923,247		7.30E-05	6.3	0.07	T/**C**	−0.42
	BOPA2_12_30872*	2H	29,124,597	19.90	7.50E-06	7.9	0.18	**A**/G	0.39
	JHI-Hv50k-2016-225852[f,*]	4H	371,267	0.71	3.60E-05	6.8	0.06	T/**C**	−0.59
	BOPA1_3549-743[f]	4H	569,760,181	63.39	1.10E-04	6.0	0.40	**A**/G	0.26
	JHI-Hv50k-2016-421359[f]	6H	554,181,962	92.07	1.40E-04	5.9	0.08	A/**T**	−0.40

BIO, biomass; SP, spikes per plant; GP, grains per plant; KWP, kernel weight per plant; PH, plant height; WLS, waterlogging score; Ch, chromosome number; MAF, minor allele frequency.

[a] *The marker with the highest R^2 in the genomic region is presented.*

[b] *Base pair positions of the marker in the chromosome based on a high-quality reference genome assembly for barley (Hordeum vulgare L.) (Mascher et al., 2017).*

[c] *Genetic marker positions (cM) of the marker obtained from the POPSEQ_2017 genome map in BarleyMap (http://floresta.eead.csic.es/barleymap/find/) (Mascher et al., 2013).*

[d] *R^2 (%) indicates the percentage of phenotypic variation explained by the significant marker.*

[e] *Allele that is in bold text is the favorable allele for the trait assessed.*

[f] *Marker-trait associations that have different positions than the previously identified QTL for waterlogging stress-related traits published on barley under waterlogging conditions.*

**Putative QTL that may be associated with multiple traits.*

their position on chromosomes, these 51 significant SNPs were assigned to 17 QTL regions on chromosomes 1H, 2H, 4H, 5H, 6H, and 7H, with each region consisting of 1 to 42 markers (**Table 5**; **Figure 5**; **Supplementary Table 5**). Some QTL regions were associated with multiple traits, indicating possible shared QTL between traits.

Since the focus of our study is waterlogging tolerance in barley, and the QTL found in the relative dataset are stable, we centered the discussion on these QTL which we named following the rule: "Q," trait abbreviation, and chromosome number. One QTL associated with BIO, named QBIO.2H, was found on chromosome 2H and explained 6.6% of the phenotypic variation (**Table 5**; **Figure 5**; **Supplementary Table 5**; **Supplementary Figure 8**). This QTL also accounted for BIO under control and waterlogging treatment conditions (**Tables 3**, **4**). Nine QTL for SP were detected on chromosomes

TABLE 5 | List of significant ($P < 1.6 \times 10^{-4}$) marker-trait associations detected by GWAS using the MLM (Q + K) model in JMP Genomics and favorable alleles (bold) for assessed traits identified in the relative dataset.

QTL	Trait	Marker[a]	Ch	Physical position (bp)[b]	Genetic position (cM)[c]	*P*-value	R^2 (%)[d]	MAF	Allele[e]	Additive effect
QBIO.2H	BIO	JHI-Hv50k-2016-73118	2H	28,612,330	18.91	4.95E-05	6.64	0.43	A/**G**	6.17
QSP.1H-1	SP	JHI-Hv50k-2016-20766[f]	1H	107,293,686		5.17E-05	6.61	0.20	**T**/C	−6.33
QSP.1H-2		JHI-Hv50k-2016-20908	1H	187,645,763	47.94	4.31E-05	6.67	0.20	**T**/C	−6.38
QSP.1H-3		JHI-Hv50k-2016-21022	1H	241,516,420	47.94	8.79E-06	7.92	0.18	A/**G**	−6.71
QSP.1H-4		JHI-Hv50k-2016-22269	1H	296,548,971	47.94	1.15E-04	5.98	0.15	T/**G**	−6.12
QSP.1H-5		JHI-Hv50k-2016-22575	1H	303,086,870	47.94	1.15E-04	5.98	0.15	T/**C**	−6.12
QSP.2H		JHI-Hv50k-2016-73693*	2H	29,669,511		1.96E-05	7.24	0.06	**A**/C	13.31
QSP.5H-1		JHI-Hv50k-2016-312394[f]	5H	532,344,110		1.58E-04	5.79	0.08	T/**G**	10.59
QSP.5H-2		JHI-Hv50k-2016-332745	5H	591,637,898	120.07	1.30E-04	5.98	0.07	A/**G**	8.32
QSP.5H-3		JHI-Hv50k-2016-336773	5H	600,914,687	126.25	9.76E-05	6.07	0.07	A/**T**	8.15
QKWP.2H	KWP	JHI-Hv50k-2016-132004	2H	733,399,550	129.78	1.32E-04	5.85	0.06	T/**C**	6.98
QKWP.4H		JHI-Hv50k-2016-230103	4H	10,736,375	29.15	7.30E-05	6.31	0.06	A/**G**	9.95
QPH.2H-1	PH	BOPA2_12_30631[f]	2H	18,521,931	12.11	9.32E-05	6.10	0.50	**A**/G	2.91
QPH.2H-2		JHI-Hv50k-2016-73693*	2H	29,669,511		5.57E-08	11.46	0.06	A/**T**	12.99
QPH.7H		JHI-Hv50k-2016-457680	7H	32,776,909	29.96	8.89E-05	6.14	0.33	**A**/C	−4.13
QWLS.1H	WLS	JHI-Hv50k-2016-19217	1H	61,923,247	46.46	7.25E-05	6.29	0.07	T/**C**	−0.42
QWLS.2H		BOPA2_12_30872	2H	29,124,597	19.90	7.51E-06	7.94	0.18	**A**/G	0.39
QWLS.4H-1		JHI-Hv50k-2016-225850[f]	4H	370,915	0.71	4.05E-05	6.85	0.06	**T**/C	−0.58
QWLS.4H-2		BOPA1_3549-743[f]	4H	569,760,181	63.39	1.08E-04	5.99	0.39	**A**/G	0.26
QWLS.6H		JHI-Hv50k-2016-421359[f]	6H	554,181,962	92.07	1.36E-04	5.85	0.08	A/**T**	−0s.40

BIO, biomass; SP, spikes per plant; KWP, kernel weight per plant; PH, plant height; WLS, waterlogging score; Ch, chromosome number; MAF, minor allele frequency.
[a] The marker with the highest R^2 in the genomic region is presented.
[b] Base pair positions of the marker in the chromosome based on a high-quality reference genome assembly for barley (Hordeum vulgare L.) (Mascher et al., 2017).
[c] Genetic marker positions (cM) of the marker obtained from the POPSEQ_2017 genome map in BarleyMap (http://floresta.eead.csic.es/barleymap/find/) (Mascher et al., 2013).
[d] R^2 (%) indicates the percentage of phenotypic variation explained by the significant marker.
[e] Allele that is in bold text is the favorable allele for the trait assessed.
[f] Marker-trait associations that have different positions than the previously identified QTL for waterlogging stress-related traits published on barley under waterlogging conditions.
**Putative QTL that may be associated with multiple traits.*

1H (QSP.1H-1, QSP.1H-2, QSP.1H-3, QSP.1H-4 and QSP.1H-5), 2H (QSP.2H), and 5H (QSP.5H-1, QSP.5H-2, QSP.5H-3), and explained 5.8–7.9% of the phenotypic variance (**Table 5**; **Supplementary Table 5**). Two QTL for KWP were detected on chromosomes 2H (QKWP.2H) and 4H (QKWP.4H) and explained 5.9–6.3% of the phenotypic variance (**Table 5**; **Supplementary Table 5**). For PH, three QTL were identified, located on chromosomes 2H (QPH.2H-1 and QPH.2H-2) and 7H (QPH.7H). The QTL accounted for 6.1–11.5% of the phenotypic variance (**Table 5**; **Supplementary Table 5**). The QTL QWT.PH.2H-2 also accounted for PH under control and waterlogging treatment conditions (**Tables 3**, **4**). Five QTL affecting WLS were identified and they accounted for 5.9–7.9% of the phenotypic variance (**Table 5**; **Supplementary Table 5**). They were located in chromosomes 1H (QWLS.1H), 2H (QWLS.2H), 4H (QWLS.4H-1 and QWLS.4H-2) and 6H (QWLS.6H). These five QTL also accounted for WLS under waterlogging treatment (**Table 4**).

One genomic region was associated with various traits in the relative dataset (**Table 5**). On chromosome 2H, QTL QWT.BIO.2H, QWT.SP.2H and QWT.PH.2H-2 located at 28-29 Mbp were associated with BIO, SP, and PH, respectively, although with different effects in each trait (6.6–11.5% of phenotypic variation).

Candidate Genes

A total of 205, 190, and 156 genes were located within a 200-kb genomic region up- and down-stream centered from 32, 26 and 18 significant marker-trait associations in control (**Supplementary Table 6**), waterlogging treatment conditions (**Supplementary Table 7**) and relative dataset (**Supplementary Table 8**), respectively. Among those markers, 22, 19, and 14, from control, waterlogging treatment and relative datasets, respectively, were located inside genes. We focused on these genes and identified nine possible candidate genes associated with the measured traits under the control (**Table 6**), 13 possible candidate genes associated with these traits under the waterlogging treatment conditions (**Table 7**), and eight possible candidate genes associated with the measured traits in the relative dataset (**Table 8**).

Significant markers associated with BIO in control conditions were inside genes (HORVU2Hr1G013400, HORVU2Hr1G071330, HORVU2Hr1G072400, HORVU2Hr1G075950, HORVU5Hr1G096320, and HORVU2Hr1G070320) involved in the

TABLE 6 | Summary of potential candidate genes that contain significant markers associated with the assessed traits under control conditions.

Marker	Trait	Ch	Marker position (bp)	Gene ID	Start (bp)	End (bp)	Gene description
JHI-Hv50k-2016-71792	BIO	2H	23,485,824	HORVU2Hr1G011650	23,481,402	23,486,230	Undescribed protein
BOPA2_12_30872	BIO	2H	29,124,597	HORVU2Hr1G013400	29,123,724	29,127,894	Pseudo-response regulator 7
BOPA1_ABC08774-1-1-752	BIO	2H	508,786,535	HORVU2Hr1G071330	508,785,994	508,794,465	Glycine–tRNA ligase
JHI-Hv50k-2016-95073	BIO	2H	515,576,575	HORVU2Hr1G071980	515,568,391	515,580,748	Heparan-α-glucosaminide N-acetyltransferase
SCRI_RS_127347	BIO	2H	519,110,344	HORVU2Hr1G072400	519,108,149	519,110,415	Cytochrome P450 superfamily protein
JHI-Hv50k-2016-98186	BIO	2H	547,420,281	HORVU2Hr1G075950	547,420,245	547,422,120	Zinc finger homeodomain 1
JHI-Hv50k-2016-336773	BIO, KWP	5H	600,914,687	HORVU5Hr1G096320	600,914,511	600,916,443	UDP-Glycosyltransferase superfamily protein
JHI-Hv50k-2016-94875	BIO, PH	2H	496,673,313	HORVU2Hr1G070320	496,671,113	496,676,443	Yellow stripe like 6
JHI-Hv50k-2016-88492	GP	2H	134,404,110	HORVU2Hr1G033730	134,403,521	134,420,781	Proteasome maturation factor UMP1 family protein
JHI-Hv50k-2016-205562	GP, KWP	3H	634,801,729	HORVU3Hr1G091170	634,799,742	634,804,670	Receptor kinase 2
JHI-Hv50k-2016-73085	PH	2H	28,455,236	HORVU2Hr1G013020	28,452,211	28,456,166	Trichome birefringence-like 4
JHI-Hv50k-2016-86347	PH	2H	112,364,666	HORVU2Hr1G030520	112,360,955	112,366,308	Protein kinase superfamily protein
JHI-Hv50k-2016-95777	PH	2H	523,378,213	HORVU2Hr1G072750	523,377,399	523,379,178	Protein Terminal flower 1
JHI-Hv50k-2016-98501	PH	2H	550,839,094	HORVU2Hr1G076520	550,832,263	550,840,111	Pectinesterase family protein
JHI-Hv50k-2016-127739	PH	2H	723,652,876	HORVU2Hr1G111640	723,652,502	723,658,875	Plasma membrane ATPase
JHI-Hv50k-2016-129870	PH	2H	727,578,152	HORVU2Hr1G113190	727,572,166	727,583,311	Alpha-N-acetylglucosaminidase
BOPA2_12_10532	PH	3H	67,560,907	HORVU3Hr1G021150	67,560,410	67,562,131	Gigantea protein (GI)
JHI-Hv50k-2016-332746	PH	5H	591,637,968	HORVU5Hr1G093390	591,633,650	591,639,220	Solute carrier family 22 member 1
BOPA2_12_31234	PH	5H	613,268,086	HORVU5Hr1G101820	613,267,130	613,268,378	Undescribed protein
JHI-Hv50k-2016-447227	PH	7H	11,309,509	HORVU7Hr1G008690	11,307,419	11,313,973	Protein kinase superfamily protein
JHI-Hv50k-2016-468495	PH	7H	71,962,797	HORVU7Hr1G034400	71,959,645	71,963,636	Unknown function
JHI-Hv50k-2016-468869	PH	7H	75,059,390	HORVU7Hr1G034990	75,057,969	75,067,902	Kinesin-related protein 11

BIO, biomass; SP, spikes per plant; GP, grains per plant; KWP, kernel weight per plant; PH, plant height; WLS, waterlogging score; Ch, chromosome number.

regulation of the circadian clock, regulation of flowering time and development, embryogenesis, grain size and development, plant growth, development and senescence (**Table 6**). The role of the genes HORVU5Hr1G096320 and HORVU2Hr1G033730 harboring the markers JHI-Hv50k-2016-336773 and JHI-Hv50k-2016-88492, respectively, associated with GP and KWP traits were known to be essential in the regulation of seed development and grain size (**Table 6**). Several genes (HORVU2Hr1G013020, HORVU2Hr1G076520, and HORVU7Hr1G034990) associated with the significant markers for PH trait were known to be involved in cell wall processes, such as synthesis and deposition of secondary wall cellulose, modulation of cell wall mechanical stability during fruit ripening, cell wall extension during pollen germination and pollen tube growth, abscission, stem elongation, tuber yield and root development, microtubule-binding proteins involved in the microtubule control of cellulose microfibril order and cell wall strength. Some other genes (HORVU2Hr1G030520, HORVU5Hr1G093390, and HORVU7Hr1G008690) play a role in cell cycle regulation processes, such as modulating vesicle transport and channel activities, and specific transport of various substrates. Another group of genes (HORVU2Hr1G072750, HORVU2Hr1G111640, HORVU2Hr1G113190, and HORVU3Hr1G021150) regulate plant growth and reproductive development, flowering time and inflorescence architecture (**Table 6**).

Most of the genes harboring market-trait associations for the related traits in waterlogging treatment conditions are known to play a role in the regulation of waterlogging or other abiotic stress responses (**Table 7**). The genes HORVU2Hr1G072140, encoding Uridylate kinase, and HORVU2Hr1G013400, encoding Pseudo-response regulator 7 (PRR7), contain significant markers associated with BIO and are known to play a role in the response to abiotic stress, such as salinity, cold and oxidative stress (**Table 7**). The four genes HORVU6Hr1G070750 (annotated as E3 ubiquitin-protein ligase makorin), HORVU4Hr1G090640 (E3 ubiquitin-protein ligase RFWD3), HORVU4Hr1G000090

TABLE 7 | Summary of potential candidate genes that contain significant markers associated with the assessed traits under waterlogging treatment conditions.

Marker	Trait	Ch	Marker position (bp)	Gene ID	Start (bp)	End (bp)	Gene description
JHI-Hv50k-2016-95223	BIO	2H	516,581,410	HORVU2Hr1G072140	516,578,216	516,583,796	Uridylate kinase
BOPA2_12_30872	BIO, WLS	2H	29,124,597	HORVU2Hr1G013400	29,123,724	29,127,894	Pseudo-response regulator 7
JHI-Hv50k-2016-161633	SP	3H	32,637,255	HORVU3Hr1G014290	32,636,782	32,639,178	Delta(8)-Delta(7) sterol isomerase
JHI-Hv50k-2016-276624	SP	4H	645,759,577	HORVU4Hr1G090640	645,757,976	645,762,395	E3 ubiquitin-protein ligase RFWD3
JHI-Hv50k-2016-109151	SP	2H	662,018,769	HORVU2Hr1G094030	662,015,232	662,019,114	Ubiquitin-conjugating enzyme 3
JHI-Hv50k-2016-3532	SP	1H	3,453,791	HORVU1Hr1G001480	3,453,090	3,454,077	Undescribed protein
JHI-Hv50k-2016-322832	SP, GP	5H	569,308,558	HORVU5Hr1G083110	569,293,089	569,309,305	Leucine-rich repeat receptor-like protein kinase
BOPA2_12_11245	SP, GP, KWP	5H	579,324,077	HORVU5Hr1G087730	579,322,710	579,324,607	13S globulin seed storage protein 2
JHI-Hv50k-2016-225852	SP, WLS	4H	371,267	HORVU4Hr1G000090	369,520	374,029	RING/U-box superfamily protein
JHI-Hv50k-2016-410329	GP	6H	492,880,745	HORVU6Hr1G070750	492,878,969	492,884,688	E3 ubiquitin-protein ligase makorin
JHI-Hv50k-2016-249670	GP, KWP	4H	512,990,076	HORVU4Hr1G061070	512,989,821	512,992,961	C2H2-like zinc finger protein
JHI-Hv50k-2016-82113	KWP	2H	79,456,923	HORVU2Hr1G025510	79,452,094	79,457,099	B3 domain-containing protein
JHI-Hv50k-2016-424341	KWP	6H	562,861,599	HORVU6Hr1G087000	562,860,368	562,867,337	Heparanase-like protein 3
JHI-Hv50k-2016-127867	KWP	2H	724,202,574	HORVU2Hr1G111780	724,201,388	724,204,020	Receptor-like protein kinase 4
JHI-Hv50k-2016-322288	KWP	5H	568,058,046	HORVU5Hr1G082670	568,057,965	568,060,772	Undescribed protein
BOPA2_12_10968	PH	3H	34,959,733	HORVU3Hr1G015050	34,956,640	34,962,056	Enolase-phosphatase E1
JHI-Hv50k-2016-165725	PH	3H	78,242,146	HORVU3Hr1G022270	78,241,796	782,431,36	Pentatricopeptide repeat 336
BOPA1_3549-743	WLS	4H	569,760,181	HORVU4Hr1G069280	569,757,996	569,767,162	Alpha-L-fucosidase 2
JHI-Hv50k-2016-19217	WLS	1H	61,923,247	HORVU1Hr1G017900	61,919,204	61,923,605	Transcription factor PIF3

BIO, biomass; SP, spikes per plant; GP, grains per plant; KWP, kernel weight per plant; PH, plant height; WLS, waterlogging score; Ch, chromosome number.

(RING/U-box superfamily protein), and HORVU2Hr1G094030 (Ubiquitin-conjugating enzyme 3) associated with SP, GP, and KWP regulate abiotic stress signaling pathways, such as in waterlogging or flooding conditions (**Table 7**). Also, the associated genes HORVU5Hr1G083110 (Leucine-rich repeat receptor-like kinase family protein) and HORVU2Hr1G111780 (Receptor-like protein kinase 4) are known to be involved in abiotic stress responses, including drought, salt, cold, toxic metals and other stresses. The gene HORVU2Hr1G025510 (B3 domain-containing protein), associated with SP, is involved in abiotic stress and disease resistance signaling pathways. The gene HORVU4Hr1G061070 (C2H2 zinc finger protein) associated with GP and KWP, participates in mechanisms of tolerance to salinity, osmotic, cold, drought, oxidative and high-light stress response (**Table 7**). The gene HORVU3Hr1G022270 (Pentatricopeptide repeat 336), associated with PH, is known to regulate plant responses to abiotic stresses (**Table 7**). The significant markers associated with WLS were located inside the genes encoding PRR7 and RING/U-box superfamily protein, and the genes HORVU4Hr1G069280 (Alpha-L-fucosidase 2), involved in the response to waterlogging, drought and salinity stresses, and HORVU1Hr1G017900 (Phytochrome-interacting factor 3), which regulates the plant response to drought and salt stresses (**Table 7**).

In the relative dataset, the significant markers JHI-Hv50k-2016-20766 and JHI-Hv50k-2016-21022 associated with SP, were inside the genes HORVU1Hr1G024060 (Arginine/serine-rich splicing factor 35) and HORVU1Hr1G036060 (tRNA pseudouridine synthase A1), respectively, that play important roles in development and response to abiotic stresses (**Table 8**). The role of the gene HORVU2Hr1G114940, encoding Cyclic nucleotide-gated channel 8, contains significant markers associated with KWP and is known to play a crucial role in pathways related to cellular ion homeostasis, development, and defense against biotic and abiotic stresses. The gene HORVU7Hr1G022410, encoding RNA-binding protein mde7, was associated with PH and has functional roles during growth, development, and abiotic stress responses in plants

TABLE 8 | Summary of potential candidate genes that contain significant markers associated with the assessed traits identified in the relative dataset.

Marker	Trait	Ch	Marker position (bp)	Gene ID	Start (bp)	End (bp)	Gene description
JHI-Hv50k-2016-20766	SP	1H	107,293,686	HORVU1Hr1G024060	107,289,291	107,295,231	Arginine/serine-rich splicing factor 35
JHI-Hv50k-2016-20908	SP	1H	187,645,763	HORVU1Hr1G031370	187,632,592	187,656,006	tRNA pseudouridine synthase A1
JHI-Hv50k-2016-21022	SP	1H	241,516,420	HORVU1Hr1G036060	241,482,945	241,524,027	Cationic amino acid transporter 2
JHI-Hv50k-2016-22269	SP	1H	296,548,971	HORVU1Hr1G041530	296,548,421	296,553,191	Predicted protein
JHI-Hv50k-2016-22575	SP	1H	303,086,870	HORVU1Hr1G041960	303,085,891	303,088,081	Unknown function
JHI-Hv50k-2016-312394	SP	5H	532,344,110	HORVU5Hr1G071230	532,343,847	532,345,355	Unknown function
JHI-Hv50k-2016-332745	SP	5H	591,637,898	HORVU5Hr1G093390	591,633,650	591,639,220	Solute carrier family 22 member 1
JHI-Hv50k-2016-336773	SP	5H	600,914,687	HORVU5Hr1G096320	600,914,511	600,916,443	UDP-Glycosyltransferase superfamily protein
JHI-Hv50k-2016-132004	KWP	2H	733,399,550	HORVU2Hr1G114940	733,394,545	733,400,877	Cyclic nucleotide gated channel 8
JHI-Hv50k-2016-457680	PH	7H	32,776,909	HORVU7Hr1G022410	32,775,788	32,780,170	RNA-binding protein mde7
JHI-Hv50k-2016-19217	WLS	1H	61,923,247	HORVU1Hr1G017900	61,919,204	61,923,605	Transcription factor PIF3
BOPA2_12_30872	WLS	2H	29,124,597	HORVU2Hr1G013400	29,123,724	29,127,894	Pseudo-response regulator 7
JHI-Hv50k-2016-225850	WLS	4H	370,915	HORVU4Hr1G000090	369,520	374,029	RING/U-box superfamily protein
BOPA1_3549-743	WLS	4H	569,760,181	HORVU4Hr1G069280	569,757,996	569,767,162	Alpha-L-fucosidase 2

SP, spikes per plant; KWP, kernel weight per plant; PH, plant height; WLS, waterlogging score; Ch, chromosome number.

(**Table 8**). Additionally, the genes HORVU5Hr1G093390 and HORVU5Hr1G096320 were harboring markers associated with SP and were also identified in the control dataset harboring markers associated with PH, BIO, and KWP. The genes HORVU1Hr1G017900, HORVU4Hr1G000090, and HORVU4Hr1G069280 were harboring markers associated with WLS and also were identified in the waterlogging dataset associated with the same trait. Finally, the gene HORVU2Hr1G013400, encoding PRR7, contained markers associated with WLS in the waterlogging treatment and relative datasets, and BIO in the control dataset (**Tables 6–8**).

DISCUSSION

Waterlogging is becoming one of the challenging issues for modern agriculture globally. The development of tolerant cultivars with enhanced resilience to waterlogging stress has increasing importance to reduce the yield penalty. In this study, GWAS was performed based on linkage disequilibrium on a worldwide spring barley collection using control, waterlogging treatment and relative datasets for identifying QTL associated with yield-related traits and waterlogging tolerance.

Diverse Phenotypic Variation and Waterlogging Tolerant Barley Genotypes

In the present study, the barley collection assembled showed significant phenotypic variation, as well as highly genotypic differences, for all traits after waterlogging stress treatment, including BIO, SP, GP, KWP, PH, and WLS, except CABC and CCC. These results suggest that there is a good potential that these genotypes can be used to mine alleles for waterlogging tolerance for introgression into breeding barley lines for waterlogging tolerance improvement. Waterlogging stress considerably reduced BIO, SP, GP, KWP, PH, CABC, and CCC for all genotypes in response to waterlogging stress as expected, and it is consistent with earlier studies (Li et al., 2008; Xue et al., 2010). Significant negative correlations were found between WLS and all other traits.

The barley genotype Deder2 from Ethiopia showed a tolerant response to waterlogging stress, while the response of the genotypes Yerong from Australia, TR 587 and CDC Select from Canada, Champion, Xena, and TR 987 from the USA, and Harumaki Rokkakumugi from North Corea, was more moderate. Some of these barley genotypes (e.g., Deder2, Harumaki Rokkakumugi, and Yerong) were previously reported (Takeda, 1989; Li et al., 2008) to be tolerant to waterlogging stress while the others, which are modern cultivars (Canadian Food Inspection Agency, 2021; Washington State Crop Improvement Association, 2021; Westland Seed, 2021) and elite breeding lines, were not reported before and might represent novel sources of tolerance.

Genome-Wide Association Study Analysis

The GWAS is a powerful approach to locate common alleles associated with phenotypes with much higher resolution than linkage mapping because they reflect historical recombination events in broad-based diversity panels (Nordborg and Weigel, 2008). In this study, three statistical models were compared to

assess their ability to map QTL and identify SNPs associated with waterlogging tolerance. Finally, we selected the MLM + Q + K approach, which accounts for both population structure (STRUCTURE analyses) and K matrix, because of its statistical power to control false-positives associations, which has been used successfully in barley (Pasam et al., 2012; Fan et al., 2016; Jabbari et al., 2018) and maize (Yu et al., 2006). Population structure and familial relatedness can result in false positives in GWAS. Therefore, when GWAS is conducted, these parameters need to be considered in the model. In the present study, the level of the genetic structure of the panel was assessed by the NJ tree, PCA, and STRUCTURE analyses and all showed that the investigated genotypes are structured into three principal groups. This provided additional confidence given that most of the barley population structure studies use only two of these methods, STRUCTURE and PCA, to confirm their results (Varshney et al., 2012; Long et al., 2013; Fan et al., 2016; Zhou et al., 2016; Bengtsson et al., 2017; Jabbari et al., 2018; Thabet et al., 2018; Milner et al., 2019; Mwando et al., 2020; Ye et al., 2020). Moreover, the LD decay value identified (1.46 Mbp at $r^2 = 0.1$) suggested that the marker coverage is adequate for further GWAS analysis. A wide range of levels of LD decay, 2–10 cM, was reported by previous studies of worldwide barley collections (Comadran et al., 2009; Zhang et al., 2009; Pasam et al., 2012; Varshney et al., 2012; Long et al., 2013; Zhou et al., 2016). Comparison to any of these studies is hard to be made due to several factors such as size and diversity of the germplasm used, type and number of molecular markers, and measurement unit. Recently, Mwando et al. (2020) reported a LD decay of 3.5 Mbp ($r^2 = 0.2$) in 350 barley accessions using 24,138 DArTseq and SNP markers. While this time the measurement unit is the same (Mbp) the results are not directly comparable to our study either, mainly due to the different germplasm assessed. Nevertheless, the work conducted by Mwando et al. (2020) demonstrated successful association mapping was achieved with a lower number of molecular markers (24,138 vs. 35,926) than used in our study.

The overall GWAS was able to identify significant QTL in all control, waterlogging treatment and relative datasets for six (BIO, SP, GP, KWP, PH, and WLS) out of the eight traits measured. No significant QTL were detected for CABC and CCC in the tested conditions. Chlorophyll is one of the major chloroplast components for photosynthesis, and relative chlorophyll content has a positive relationship with photosynthetic rate (Guo et al., 2008). An earlier study reported the identification of QTL for chlorophyll fluorescence in barley under low oxygen concentration in hydroponics to simulate waterlogging but not for chlorophyll content or chlorophyll (Bertholdsson et al., 2015).

Identification of Known Waterlogging-Related QTL by GWAS

So far, several QTL mapping studies have been conducted using linkage mapping analysis in barley and many QTL associated with waterlogging tolerance have been successfully mapped using bi-parental linkage mapping based on various waterlogging related traits (Li et al., 2008; Xue et al., 2010; Zhou, 2011; Xu et al., 2012; Zhou et al., 2012; Bertholdsson et al., 2015; Broughton et al., 2015; Zhang et al., 2016; Gill et al., 2017, 2019; Zhang X. et al., 2017). These studies used DH populations from bi-parental crosses of contrasting phenotype parents for waterlogging. Direct comparisons of our GWAS findings with those studies are intricate, as the marker-trait linkages and chromosomal locations we identified were based on a worldwide barley collection not previously investigated for waterlogging traits. Moreover, different genotyping technologies and different linkage maps have been used in some of the previous studies, so the comparison is approximated. In general, our GWA mapping was highly consistent with those previous waterlogging tolerance QTL mapping studies conducted in bi-parental populations, and many QTL were identified for the same or related traits at similar positions, which confirmed the importance of the loci identified in the present study.

Some of the waterlogging-related QTLs detected in the waterlogging treatment dataset in our study are positioned closer to previously identified waterlogging stress-related QTLs for similar traits (Xue et al., 2010; Xu et al., 2012; Broughton et al., 2015; Ma et al., 2015). SP trait was associated with genomic regions related to the markers JHI-Hv50k-2016-3532 (at 3 Mbp on 1H), JHI-Hv50k-2016-109151 (at 662 Mbp on 2H) and JHI-Hv50k-2016-161633 (at 32 Mbp on 3H) were also associated with the related traits shoot fresh weight (QHSFW.1H) and tiller number (QHTiller.3H) in the Franklin x YYXT mapping population (Broughton et al., 2015), and grains per spike (GSw1.1 and GSw1.2) in Franklin x Yerong mapping population (Xue et al., 2010). The marker JHI-Hv50k-2016-3532 was also associated with the QTL for salinity and waterlogging tolerance (QSlww.YG.1H-1) in a DH population of Gairdner × YSM1 (Ma et al., 2015). The marker JHI-Hv50k-2016-109151 was also closely positioned near the QTL tfsur-1 which is associated with plant survival in the TX9425 × Franklin mapping population (Li et al., 2008). One of the genomic regions associated with KWP, related to the marker JHI-Hv50k-2016-127867 located at 724 Mbp on 2H was coincident with the previous identified QTL (SLw2.2) for spike length in the Franklin x Yerong population (Xue et al., 2010). Zhou (2011) also reported two QTL (QWL.YeFr.2H.2 and WL5.3) associated with waterlogging tolerance score, which is positioned near the marker JHI-Hv50k-2016-127867. WLS trait was associated with BOPA2_12_30872 located at 29 Mbp on 2H. This genomic region was previously detected in two different populations, TX9425 x Naso Nijo (Xu et al., 2012) and YSM1 x Gairdner (Ma et al., 2015), for the same trait. Additionally, in our study BIO was also associated with the same marker that was located on the genomic region 29.1–29.7 Mbp on chromosome 2H. Interestingly, in our study, this same region was also associated with the traits GP, KWP, and PH (JHI-Hv50k-2016-73570 and JHI-Hv50k-2016-73689).

Other waterlogging-related QTL detected in our study were identified in previous waterlogging stress studies but associated with different traits (Li et al., 2008; Xue et al., 2010; Zhou, 2011; Xu et al., 2012; Zhou et al., 2012; Broughton et al., 2015; Ma et al., 2015; Gill et al., 2017). For example, the traits SP, GP and KWP were associated with the genomic region 568.0–569.3 Mbp on 5H that was coincident for the QTL yfsur-2 for plant survival under waterlogging in the DH population of Yerong × Franklin (Li et al., 2008).

In our study, we identified five QTL in the relative dataset that were positioned closer to previously identified waterlogging stress-related QTLs for similar traits (Xu et al., 2012; Broughton et al., 2015; Ma et al., 2015). SP was associated with four QTL, QSP.1H-2, QSP.1H-3, QSP.1H-4, and QSP.1H-5, that were also associated with the related trait shoot dry weight (QHSDW.1H) in the Franklin x YYXT mapping population (Broughton et al., 2015). The QTL QWLS.4H-2 was associated with WLS and was also present in the waterlogging treatment dataset. Other waterlogging tolerance-related QTL detected in our study were identified in previous waterlogging stress studies but associated with different traits (Xue et al., 2010; Zhou, 2011; Xu et al., 2012; Broughton et al., 2015; Ma et al., 2015).

Identification of Novel Waterlogging-Related QTL by GWAS

Among the 37 QTL detected under waterlogging treatment conditions, 13 QTL were detected on genomic regions where no waterlogging-related QTL have been previously reported in barley. These 13 QTL located in 10 different genomic regions, probably represents novel loci for waterlogging stress. Two significant associated markers, JHI-Hv50k-2016-68186 and JHI-Hv50k-2016-68266, were identified on 2H at 16 Mbp. The first marker was associated with the trait KWP and the second with SP. On chromosome 4H at 0.37, 512, 569, and 645 Mbp, four markers, JHI-Hv50k-2016-225852, JHI-Hv50k-2016-249670, BOPA1_3549-743, and JHI-Hv50k-2016-276624, were identified. The first marker was associated with SP and WLS, the second marker with GP and KWP, the third marker with WLS and the last marker with SP. On chromosome 6H at 492, 554, and 562 Mbp, three markers, JHI-Hv50k-2016-410329, JHI-Hv50k-2016-421359, and JHI-Hv50k-2016-424341, were associated with GP, WLS, and KWP, respectively. The marker JHI-Hv50k-2016-449124 was associated with GP and KWP on 7H at 13 Mbp.

In the relative dataset, six QTL (QPH.2H-1, QSP.1H-1, QSP.5H-1, QWLS.4H-1, QWLS.4H-2, and QWLS.6H) out of 20 were detected on genomic regions that have not been reported in previous waterlogging-related QTL studies on barley conducted using bi-parental populations and they probably represent novel loci for waterlogging tolerance. SP was associated with the markers JHI-Hv50k-2016-20766 and JHI-Hv50k-2016-312394, located on chromosome 1H at 107 Mbp and 5H at 532 Mbp, respectively. The marker BOPA2_12_30631 was associated with PH on 2H at 18 Mbp. For WLS, three markers were found to be associated, JHI-Hv50k-2016-225850 and BOPA1_3549-743, located on 4H at 0.37 and 569 Mbp, respectively, and JHI-Hv50k-2016-421359 on 6H at 554 Mbp. The genomic regions at 0.37 and 569 Mbp on 4H and 554 Mbp on 6H were co-localized in waterlogging treatment and relative datasets, associated with WLS. Interestingly, QWLS.4H-2 is positioned relatively close to the QTL for aerenchyma formation (QTL-aerenchyma) and root porosity (QTL-rp4H) (Zhang et al., 2016).

Waterlogging-Related Candidate Genes

In the present study, 92 markers significantly associated with yield-related traits were identified in control conditions, which were located along 28 QTL regions on chromosomes 2H, 3H, 5H, 6H, and 7H; 63 significant markers were identified under waterlogging treatment conditions and mapped along 24 QTL regions on all chromosomes in the barley genome; while 51 significant markers located in 17 QTL regions distributed along chromosomes 1H, 2H, 4H, 5H, 6H, and 7H were identified in the relative data set. Among those QTL, we detected possible candidate genes that were associated with the measured traits under the different growing conditions, i.e., control, waterlogging treatment, and the relative difference between these two conditions.

Genes affected by waterlogging stress and involved in the tolerance of barley to this stress are most valuable in waterlogging breeding programs to develop and improve the efficiency of waterlogging-tolerant barley varieties. In our study, most of the potential candidate genes containing significant markers under waterlogging treatment conditions were detected on 2H and 4H associated with BIO, GP and PH. However, for the relative dataset, chromosome 1H contained most of the potential candidate genes, followed by 2H, 4H, and 5H. Four QTL that appears to harbor genes associated with abiotic stress tolerance were detected on both waterlogging treatment and relative datasets to be associated with WLS. The most significant two are QWLS.2H, harboring the gene PRR7 (HORVU2Hr1G013400) on 2H at 29.1 Mbp, is potentially similar to the reported QTL for membrane potential QMP.TxNn.2H (Gill et al., 2017); and the novel QWLS.4H-2, harboring the gene Alpha-L-fucosidase 2 (HORVU4Hr1G069280) on 4H at 569.7 Mbp, that is closely located to the reported QTL for aerenchyma formation (Zhang et al., 2016). PRR7 has a central role in the abiotic stress response and influences the regulation of flowering time and ABA-related processes, including control of genes affecting salinity, cold and oxidative stress response (Liu et al., 2013). This gene harbored the BOPA2_12_30872 marker that was also associated with BIO under waterlogging stress conditions. Alpha-L-fucosidase 2 is known to be involved in the breakdown of cell wall polymers and was previously reported to be upregulated in tolerant genotypes of maize, sesame, and chickpea in response to waterlogging, drought and salinity stresses, respectively (Thirunavukkarasu et al., 2013; Dossa et al., 2017; Kaashyap et al., 2018). These results indicated the reliability of the QTL in this study. The other two genes were detected on 1H and 4H. The Transcription factor PIF3 on QWLS.1H regulates the plant response to drought and salt stresses in maize (Gao et al., 2015) and plays a positive role in submergence-induced hypocotyl elongation in Arabidopsis (Wang et al., 2020). RING/U-box superfamily protein on the novel QWLS.4H-1 is involved in the ubiquitination reaction, a crucial mechanism that regulates signal transduction in diverse biological processes, including abiotic stress signaling pathways, such as in waterlogging or flooding conditions (Voesenek and Bailey-Serres, 2015; Loreti et al., 2016). This strong ubiquitin response is a robust indicator of changing physiological situation, by repurposing proteins through proteolysis. Additionally, the novel QWLS.6H detected only in waterlogging stress conditions harbored Receptor kinase 2 that belongs to the largest group within the receptor-like kinase (RLK) superfamily in plants and had been reported as having a main role in developmental processes, signaling networks and disease resistance. Many RLKs are involved in abiotic stress responses, including drought, salt, cold, toxic metals and other stresses (reviewed in Ye et al., 2017).

For example, a hypersensitive response was observed in response to salt and heat stress in Arabidopsis (Park et al., 2014). The homolog of the gene HORVU5Hr1G071230, harboring QSP.5H-1 on 5H at 532 Mbp, in Arabidopsis it is characterized as a cell wall integrity/stress response component.

Additionally, in our previous study, an RNA-Sequencing analysis was conducted to explore the mechanisms involved in the responses of two barley genotypes with tolerant, Deder2, and moderately-tolerant, Yerong, responses to waterlogging stress (Borrego-Benjumea et al., 2020). One of the top highly expressed differentially expressed genes (logFC $\geq \pm 4$ and adjusted $P < 0.05$) in the roots of waterlogged Deder2 and Yerong, was the upregulated gene Trichome birefringence-like 19 (8.47 logFC) which is very close to the marker JHI-Hv50k-2016-276624. This marker in the current study is associated with SP in the waterlogging treatment conditions. The underlying function of this gene is the ubiquitous modification of cell wall polymers by acetylation and is known to play a structural role in plant growth and microorganism and environmental stresses defenses (Nafisi et al., 2015), such as salinity and cold (Anantharaman and Aravind, 2010). The marker JHI-Hv50k-2016-249670, associated with GP and KWP in the waterlogging treatment conditions, is in the surroundings of the upregulated gene encoding the protein very-long-chain3-oxoacyl-CoA reductase 1 (5.26 logFC). This protein is required for the elongation of fatty acids precursors of sphingolipids, triacylglycerols, cuticular waxes and suberin, and play a role in the stress adaptation in rice. The downregulated gene Copalyl diphosphate synthase 2 (−7.34 logFC) is located very close to the marker JHI-Hv50k-2016-322288 associated with KWP in the waterlogging treatment conditions. This gene responds to arsenic detoxification in rice and it is involved in the plant adaptive responses to arsenic stress (Singh et al., 2017). The marker JHI-Hv50k-2016-3532, associated with SP in the waterlogging treatment conditions, is positioned in the surroundings of the downregulated gene encoding the protein Dirigent protein 21 (−4.76 logFC). This protein is involved in the defense response against salt and drought stress of pepper (Khan et al., 2018).

Further analysis is necessary to validate the associated candidate genes. However, this study represents the starting point of the discovery of candidate genes associated with waterlogging tolerance as well as the development of useful gene-based functional markers for barley breeding to speed up the development of waterlogging tolerant barley cultivars.

CONCLUSION

GWAS based on high-density SNP markers represents a powerful approach for dissecting complex quantitative traits. In this study, 247 worldwide spring barley genotypes were evaluated for yield components-related traits under control and waterlogging treatment conditions in the field, as well as the relative difference between these two conditions, and were genotyped using Barley 50K iSelect SNP Array. GWAS analysis showed that a total of 92, 63, and 51 markers were significantly associated with BIO, SP, GP, KWP, PH, and WLS traits in the control, waterlogging treatment, and relative datasets, respectively. Seventeen significant associations and eight potential candidate genes were detected for the relative dataset. Also, six novel QTL (QPH.2H-1, QSP.1H-1, QSP.5H-1, QWLS.4H-1, QWLS.4H-2, and QWLS.6H) were detected on genomic regions that have not been reported in previous waterlogging-related QTL studies on barley and they probably represent novel loci for waterlogging tolerance. These findings provide useful information for waterlogging tolerance in barley by marker-assisted selection in the future. For further research, it will be necessary the validation of the associated candidate genes and the development of markers based on associated SNPs.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

AB-B, AB, and MZhu: conceptualization. AB-B, AB, AC, JT, MZhu, and MZho: methodology and writing—review and editing. AB-B, AC, and JT: software. AB-B and AB: writing—original draft preparation. AB: supervision, project administration, and funding acquisition. All authors have read and agreed to the published version of the manuscript.

FUNDING

This research was part of the Barley Cluster project led by Alberta Barley with funding from the Western Grains Research Foundation and Agriculture and Agri-Food Canada (AAFC) under the Growing Forward 2 program.

ACKNOWLEDGMENTS

We gratefully thank Jatinder Singh Sangha, Rudy von Hertzberg, Bryan Graham, Chantel Lim, Elise Poole, Daniel Lysack, Bradley Rathwell, Kim Fleming, Audrey Bamber, and summer students of 2016 and 2017 from AAFC Brandon Research and Development Centre for technical assistance. We also thank Kazuhiro Sato, Professor Okayama University, Japan for providing the seeds for Deder2 and Harumaki Rokkakumugi, Harold E. Bockelman, Curator National Small Grains Collection for providing the seeds for about 50 of the accessions used in this study and Martin Ganal from TraitGenetics GmbH, Gatersleben Germany for providing the cluster file for SNP assignment.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.711654/full#supplementary-material

REFERENCES

Anantharaman, V., and Aravind, L. (2010). Novel eukaryotic enzymes modifying cell-surface biopolymers. *Biol. Direct* 5:1. doi: 10.1186/1745-6150-5-1

Arduini, I., Orlandi, C., Ercoli, L., and Masoni, A. (2016). Submergence sensitivity of durum wheat, bread wheat and barley at the germination stage. *Ital. J. Agron.* 11, 100–106. doi: 10.4081/ija.2016.706

Bailey-Serres, J., and Voesenek, L. A. (2008). Flooding stress: acclimations and genetic diversity. *Ann. Rev. Plant Biol.* 59, 313–339. doi: 10.1146/annurev.arplant.59.032607.092752

Bayer, M. M., Rapazote-Flores, P., Ganal, M., Hedley, P. E., Macaulay, M., Plieske, J., et al. (2017). Development and evaluation of a Barley 50k iSelect SNP Array. *Front. Plant Sci.* 8:1792. doi: 10.3389/fpls.2017.01792

Bellucci, A., Tondelli, A., Fangel, J. U., Torp, A. M., Xu, X., Willats, W. G., et al. (2017). Genome-wide association mapping in winter barley for grain yield and culm cell wall polymer content using the high-throughput CoMPP technique. *PLoS ONE* 12:e0173313. doi: 10.1371/journal.pone.0173313

Bengtsson, T., Manninen, O., Ahmed-Jahoor, A., and Orabi, J. (2017). Genetic diversity, population structure and linkage disequilibrium in Nordic spring barley (*Hordeum vulgare* L. subsp. vulgare). *Genet. Resour. Crop Evol.* 64, 2021–2033. doi: 10.1007/s10722-017-0493-5

Bertholdsson, N. O., Holefors, A., Macaulay, M., and Crespo-Herrera, L. A. (2015). QTL for chlorophyll fluorescence of barley plants grown at low oxygen concentration in hydroponics to simulate waterlogging. *Euphytica* 201, 357–365. doi: 10.1007/s10681-014-1215-0

Blair, D., Mauro, I., and Smith, R. (2016). *The Prairie Climate Atlas*. Winnipeg, MB: Prairie Climate Centre, University of Winnipeg. Available online at: http://climateatlas.ca/atlas.html (accessed March 6, 2018).

Borrego-Benjumea, A., Carter, A., Glenn, A. J., and Badea, A. (2019). Impact of excess moisture due to precipitation on barley grain yield in the Canadian Prairies. *Can. J. Plant Sci.* 99, 93–96. doi: 10.1139/cjps-2018-0108

Borrego-Benjumea, A., Carter, A., Tucker, J. R., Yao, Z., Xu, W., and Badea, A. (2020). Genome-wide analysis of gene expression provides new insights into waterlogging responses in barley (*Hordeum vulgare* L.). *Plants* 9:240. doi: 10.3390/plants9020240

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, R. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSELL software for association mapping of complex traits in diverse samples. *Bioinform.* 23, 2633–2635. doi: 10.1093/bioinformatics/btm308

Broughton, S., Zhou, G., Teakle, N., Matsuda, R., Zhou, M., O'Leary, R., et al. (2015). Waterlogging tolerance is associated with root porosity in barley (*Hordeum vulgare* L.). *Mol. Breed.* 35, 1–15. doi: 10.1007/s11032-015-0243-3

Canadian Agri-Food Trade Alliance (CAFTA) (2020). Available online at: http://cafta.org/agri-food-exports/cafta-exports/ (accessed September 1, 2020).

Canadian Food Inspection Agency (2021). Available online at: https://inspection.canada.ca/english/plaveg/pbrpov/cropreport/bar/app00002863e.shtml (accessed January 20, 2021).

Chan, E. K., Rowe, H. C., and Kliebenstein, D. J. (2010). Understanding the evolution of defense metabolites in Arabidopsis thaliana using genome-wide association mapping. *Genetics* 185, 991–1007. doi: 10.1534/genetics.109.108522

Comadran, J., Thomas, W. T., van Eeuwijk, F. A., Ceccarelli, S., Grando, S., Stanca, A. M., et al. (2009). Patterns of genetic diversity and linkage disequilibrium in a highly structured Hordeum vulgare association mapping population for the Mediterranean basin. *Theor. Appl. Genet.* 119, 175–187. doi: 10.1007/s00122-009-1027-0

Cornelious, B., Chen, P., Chen, Y., De Leon, N., Shannon, J., and Wang, D. (2005). Identification of QTLs underlying water-logging tolerance in soybean. *Mol. Breed.* 16, 103–112. doi: 10.1007/s11032-005-5911-2

De San Celedonio, R. P., Abeledo, L. G., and Miralles, D. J. (2014). Identifying the critical period for waterlogging on yield and its components in wheat and barley. *Plant Soil* 2018, 265–277. doi: 10.1007/s11104-014-2028-6

De San Celedonio, R. P., Abeledo, L. G., and Miralles, D. J. (2018). Physiological traits associated with reductions in grain number in wheat and barley under waterlogging. *Plant Soil* 429, 469–481. doi: 10.1007/s11104-018-3708-4

Dossa, K., Li, D., Wang, L., Zheng, X., Yu, J., Wei, X., et al. (2017). Transcriptomic, biochemical and physio-anatomical investigations shed more light on responses to drought stress in two contrasting sesame genotypes. *Sci. Rep.* 7:8755. doi: 10.1038/s41598-017-09397-6

Earl, D. A., and von Holdt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* 4, 359–361. doi: 10.1007/s12686-011-9548-7

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* 14, 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164, 1567–1587. doi: 10.1093/genetics/164.4.1567

Fan, Y., Zhou, G., Shabala, S., Chen, Z.-H., Cai, S., Li, C., et al. (2016). Genome-wide association study reveals a new qtl for salinity tolerance in barley (*Hordeum vulgare* L.). *Front. Plant Sci.* 7:946. doi: 10.3389/fpls.2016.00946

FAOSTAT Production (2020). Available online at: http://www.fao.org/faostat/en/#data/QC (accessed September 1, 2020).

Gao, Y., Jiang, W., Dai, Y., Xiao, N., Zhang, C., Li, H., et al. (2015). A maize phytochrome-interacting factor 3 improves drought and salt stress tolerance in rice. *Plant Mol. Biol.* 87, 413–428. doi: 10.1007/s11103-015-0288-z

Gill, M. B., Zeng, F., Shabala, L., Zhang, G., Fan, Y., Shabala, S., et al. (2017). Cell-based phenotyping reveals QTL for membrane potential maintenance associated with hypoxia and salinity stress tolerance in barley. *Front. Plant. Sci.* 8:1941. doi: 10.3389/fpls.2017.01941

Gill, M. B., Zeng, F., Shabala, L., Zhang, G., Yu, M., Demidchik, V., et al. (2019). Identification of QTL related to ROS formation under hypoxia and their association with waterlogging and salt tolerance in barley. *Int. J. Mol. Sci.* 20:699. doi: 10.3390/ijms20030699

Guo, P., Baum, M., Varshney, R. K., et al. (2008). QTLs for chlorophyll and chlorophyll fluorescence parameters in barley under post-flowering drought. *Euphytica* 163, 203–214. doi: 10.1007/s10681-007-9629-6

Hamachi, Y., Yoshino, M., Furusho, M., and Yoshida, T. (1990). Index of screening for wet endurance in malting barley. *Jpn J. Breed.* 40, 361–366. doi: 10.1270/jsbbs1951.40.361

Jabbari, M., Fakheri, B. A., Aghnoum, R., Nezhad, M. N., and Ataei, R. (2018). GWAS analysis in spring barley (*Hordeum vulgare* L.) for morphological traits exposed to drought. *PLoS ONE* 13:e0204952. doi: 10.1371/journal.pone.0204952

Kaashyap, M., Ford, R., Kudapa, H., Jain, M., Edwards, D., Varshney, R., et al. (2018). Differential regulation of genes involved in root morphogenesis and cell wall modification is associated with salinity tolerance in chickpea. *Sci. Rep.* 19:4855. doi: 10.1038/s41598-018-23116-9

Khahani, B., Tavakol, E., and Shariati, J. V. (2019). Genome-wide meta-analysis on yield and yield-related QTLs in barley (*Hordeum vulgare* L.). *Mol. Breed.* 39:56. doi: 10.1007/s11032-019-0962-y

Khan, A., Li, R. J., Sun, J. T., Ma, F., Zhang, H. X., Jin, J. H., et al. (2018). Genome-wide analysis of dirigent gene family in pepper (*Capsicum annuum* L.) and characterization of CaDIR7 in biotic and abiotic stresses. *Sci. Rep.* 8:5500. doi: 10.1038/s41598-018-23761-0

Lei, L., Poets, A. M., Liu, C., Wyant, S. R., Hoffman, P. J., Carter, C. K., et al. (2019). Environmental association identifies candidates for tolerance to low temperature and drought. *Genes Genomes Genet.* 9, 3423–3438. doi: 10.1534/g3.119.400401

Li, H., Vaillancourt, R., Mendham, N., and Zhou, M. (2008). Comparative mapping of quantitative trait loci associated with waterlogging tolerance in barley (*Hordeum vulgare* L.). *BMC Genom*e 9:401. doi: 10.1186/1471-2164-9-401

Lichtenthaler, H. K., and Wellburn, A. R. (1983). Determination of total carotenoids and chlorophylls a and b of leaf in different solvents. *Biochem. Soc. Trans.* 11, 591–592. doi: 10.1042/bst0110591

Liu, K. E., Harrison, M. T., Ibrahim, A., Manik,. S. M. N., Johnson, P., Tian,. X., et al. (2020). Genetic factors increasing barley grain yields under soil waterlogging. *Food Energy Secur.* 9: e238. doi: 10.1002/fes3.238

Liu, T., Carlsson, J., Takeuchi, T., Newton, L., and Farre, E. M. (2013). Direct regulation of abiotic responses by the Arabidopsis circadian clock component PRR7. *Plant J.* 76, 101–114. doi: 10.1111/tpj.12276

Locatelli, A., Cuesta-Marcos, A., Gutiérrez, L., Hayes, P. M., Smith, K. P., and Castro, A. J. (2013). Genome-wide association mapping of agronomic traits in relevant barley germplasm in Uruguay. *Mol. Breed.* 31, 631–654. doi: 10.1007/s11032-012-9820-x

Long, N. V., Dolstra, O., Malosetti, M., Kilian, B., Graner, A., Visser, R. G. F., et al. (2013). Association mapping of salt tolerance in barley (*Hordeum vulgare* L.). *Theor. Appl. Genet*. 126, 2335–2351. doi: 10.1007/s00122-013-2139-0

Loreti, E., van Veen, H., and Perata, P. (2016). Plant responses to flooding stress. *Curr. Opin. Plant Biol*. 33, 64–71. doi: 10.1016/j.pbi.2016.06.005

Ma, Y., Shabala, S., Li, C., Liu, C., Zhang, W., and Zhou, M. (2015). Quantitative trait loci for salinity tolerance identified under drained and waterlogged conditions and their association with flowering time in barley (*Hordeum vulgare* L). *PLoS ONE* 10:e0134822. doi: 10.1371/journal.pone.0134822

Manitoba Agricultural Services Corporation (MASC) (2017). *Tools and Data*. Available online at: https://www.masc.mb.ca/masc.nsf/tools_and_data.html (accessed March 15, 2018).

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S. O., Wicker, T., et al. (2017). A chromosome conformation capture ordered sequence of the barley genome. *Nature* 544, 427–433. doi: 10.1038/nature22043

Mascher, M., Muehlbauer, G. J., Rokhsar, D. S., Chapman, J., Schmutz, J., Barry, K., et al. (2013). Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). *Plant Sci. J.* 76, 718–727. doi: 10.1111/tpj.12319

Masoni, A., Pampana, S., and Arduini, I. (2016). Barley response to waterlogging duration at tillering. *Crop Sci*. 56, 2722–2730. doi: 10.2135/cropsci2016.02.0106

Mei, H., Zhu, X., and Zhang, T. (2013). Favorable QTL alleles for yield and its components identified by association mapping in Chinese Upland cotton cultivars. *PLoS ONE* 26:e82193. doi: 10.1371/journal.pone.0082193

Milner, S. G., Jost, M., Taketa, S., Mazon, E. R., Himmelbach, A., and Oppermann, M. (2019). Genebank genomics highlights the diversity of a global barley collection. *Nat. Genet*. 51, 319–326. doi: 10.1038/s41588-018-0266-x

Mwando, E., Han, Y., Angessa, T. T., Zhou, G., Hill, C. B., Zhang, X.-Q., et al. (2020). Genome-wide association study of salinity tolerance during germination in barley (*Hordeum vulgare* L.). *Front. Plant Sci.* 11:118. doi: 10.3389/fpls.2020.00118

Nafisi, M., Stranne, M., Fimognari, L., Atwell, S., Martens, H. J., Pedas, P. R., et al. (2015). Acetylation of cell wall is required for structural integrity of the leaf surface and exerts a global impact on plant stress responses. *Front. Plant Sci.* 6:550. doi: 10.3389/fpls.2015.00550

Nordborg, M., and Weigel, D. (2008). Next-generation genetics in plants. *Nature* 456, 720–723. doi: 10.1038/nature07629

Pang, J., Zhou, M., Mendham, N., and Shabala, S. (2004). Growth and physiological responses of six barley genotypes to waterlogging and subsequent recovery. *Aust. J. Agric. Res*. 55, 895–906. doi: 10.1071/AR03097

Park, S., Moon, J.-C., Park, Y. C., Kim, J.-H., Kim, D. S., and Jang, C. S. (2014). Molecular dissection of the response of a rice leucine-rich repeat receptor-like kinase (LRR-RLK) gene to abiotic stresses. *J. Plant Physiol*. 171, 1645–1653. doi: 10.1016/j.jplph.2014.08.002

Pasam, R. K., Sharma, R., Malosetti, M., Eeuwijk, F. A. V., Haseneyer, G., Kilian, B., et al. (2012). Genome-wide association studies for agronomical traits in a worldwide spring barley collection. *BMC Plant Biol*. 12:16. doi: 10.1186/1471-2229-12-16

Pauli, D., Muehlbauer, G. J., Smith, K. P., Cooper, B., Hole, D., Obert, D. E., et al. (2014). Association mapping of agronomic QTLs in US spring barley breeding germplasm. *Plant Genome* 7:3. doi: 10.3835/plantgenome2013.11.0037

Ploschuk, R. A., Miralles, D. J., Colmer, T. D., Ploschuk, E. L., and Striker, G. G. (2018). Waterlogging of winter crops at early and late stages: impacts on leaf physiology, growth, and yield. *Front. Plant Sci*. 9:1863. doi: 10.3389/fpls.2018.01863

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959. doi: 10.1093/genetics/155.2.945

Saskatchewan Crop Insurance Corporation (2017). *Sask Management Plus*. Available online at: http://www.saskcropinsurance.com/ (accessed March 15, 2018).

Sayre, K., Van Ginkel, M., Rajaram, S., and Ortiz-Monasterio, I. (1994). Tolerance to waterlogging losses in spring bread wheat: effect of time of onset on expression. *Ann. Wheat Newslett*. 40, 165–171.

Setter, T. L., Burgess, P., Waters, I., and Kuo, J. (1999). "Genetic diversity of barley and wheat for waterlogging tolerance in Western Australia," in *Proceedings of the 9th Australian Barley Technical Symposium* (Melbourne, VIC).

Setter, T. L., and Waters, I. (2003). Review of prospects for germplasm improvement for waterlogging tolerance in wheat, barley, and oats. *Plant Soil* 253, 1–34. doi: 10.1023/A:1024573305997

Singh, P. K., Indoliya, Y., Chauhan, A. S., Singh, S. P., Singh, A. P., Dwivedi, S., et al. (2017). Nitric oxide mediated transcriptional modulation enhances plant adaptive responses to arsenic stress. *Sci. Rep*. 7:3592. doi: 10.1038/s41598-017-03923-2

Statistics Canada (2020). *Estimated Areas, Yield, Production, Average Farm Price and Total Farm Value of Principal Field Crops*. Available online at: https://www150.statcan.gc.ca/n1/en/type/data?MM=1 (accessed September 1, 2020).

Sundgren, T. K. (2018). *A potential QTL for oxygen sensing detected in wheat subjected to waterlogging stress* (dissertation/master's thesis). Norwegian University of Life Sciences, Ås, Norway.

Sundgren, T. K., Uhlen, A. K., Waalen, W., and Lillemo, M. (2018). Field screening of waterlogging tolerance in spring wheat and spring barley. *Agron* 8:38. doi: 10.3390/agronomy8040038

Takeda, K. (1989). Varietal variation of flooding tolerance in barley seedlings, and its diallel analysis. *Jpn. J. Breed.* 39, 174–175.

Tarawneh, R. A., Alqudah, A. M., Nagel, and, M., and Börner, A. (2020). Genome-wide association mapping reveals putative candidate genes for drought tolerance in barley. *Environ. Exp. Bot*. 180:104237. doi: 10.1016/j.envexpbot.2020.104237

Thabet, S. G., Moursi, Y. S., Karam, M. A., Graner, A., and Alqudah, A. M. (2018). Genetic basis of drought tolerance during seed germination in barley. *PLoS ONE* 13:11. doi: 10.1371/journal.pone.0206682

Thirunavukkarasu, N., Hossain, F., Mohan, S., Shiriga, K., Mittal, S., Sharma, R., et al. (2013). Genome-wide expression of transcriptomes and their co-expression pattern in subtropical maize (*Zea mays* L.) under waterlogging stress. *PLoS ONE* 8:e70433. doi: 10.1371/journal.pone.0070433

Tondelli, A., Xu, X., Moragues, M., Sharma, R., Schnaithmann, F., Ingvardsen, C., et al. (2013). Structural and temporal variation in genetic diversity of European spring two-row barley cultivars and association mapping of quantitative traits. *Plant Genome* 6, 1–14. doi: 10.3835/plantgenome2013.03.0007

Varshney, R. K., Paulo, M. J., Grando, S., van Eeuwijk, F. A., Keizer, L. C. P., Guo, P., et al. (2012). Genome wide association analyses for drought tolerance related traits in barley (*Hordeum vulgare* L.). *Field Crops Res*. 126, 171–180. doi: 10.1016/j.fcr.2011.10.008

Voesenek, L. A. C. J., and Bailey-Serres, J. (2015). Flood adaptive traits and processes: an overview. *New Phytol*. 206, 57–73. doi: 10.1111/nph.13209

Wang, M., Yan, J., Zhao, J., Song, W., Zhang, X., Xiao, Y., et al. (2012). Genome-wide association study (GWAS) of resistance to head smut in maize. *Plant Sci*. 196, 125–131. doi: 10.1016/j.plantsci.2012.08.004

Wang, Q., Sun, G., Ren, X., Du, B., Cheng, Y., Wang, Y., et al. (2019). Dissecting the genetic basis of grain size and weight in barley (*Hordeum vulgare* L.) by QTL and comparative genetic analyses. *Front. Plant Sci*. 10:469. doi: 10.3389/fpls.2019.00469

Wang, X., Ma, Q., Wang, R., Wang, P., Liu, Y., and Mao, T. (2020). Submergence stress-induced hypocotyl elongation through ethylene signaling-mediated regulation of cortical microtubules in Arabidopsis. *J. Exp. Bot*. 71, 1067–1077. doi: 10.1093/jxb/erz453

Washington State Crop Improvement Association (2021). Available online at: http://washingtoncrop.com/2-row-barley/ (accessed January 20, 2021).

Westland Seed (2021). Available online at: https://westlandseed.com/seeds/champion-feed-barley/ (accessed January 20, 2021).

Xu, R., Wang, J., Li, C., Johnson, P., Lu, C., and Zhou, M. (2012). A single locus is responsible for salinity tolerance in a Chinese landrace barley (*Hordeum vulgare* L.). *PLoS ONE* 7:e43079. doi: 10.1371/journal.pone.0043079

Xu, X., Sharma, R., Tondelli, A., Russell, J., Comadran, J., Schnaithmann, F., et al. (2018). Genome-wide association analysis of grain yield-associated traits in a pan-European barley cultivar collection. *Plant Genome* 11:170073. doi: 10.3835/plantgenome2017.08.0073

Xue, D., Zhou, M., Zhang, X., Chen, S., Wei, K., Zeng, F., et al. (2010). Identification of QTLs for yield and yield components of barley under different growth conditions. *J. Zhejiang Univ. Sci.* 11, 169–176. doi: 10.1631/jzus.B0900332

Yaduvanshi, N., Setter, T., Sharma, S., Singh, K., and Kulshreshtha, N. (2014). Influence of waterlogging on yield of wheat (*Triticum aestivum*), redox

potentials, and concentrations of microelements in different soils in India and Australia. *Soil Res.* 50, 489–499. doi: 10.1071/SR11266

Ye, Y., Ding, Y., Jiang, Q., Wang, F., Sun, J., and Zhu, C. (2017). The role of receptor-like protein kinases (RLKs) in abiotic stress response in plants. *Plant Cell Rep*. 36, 235–242. doi: 10.1007/s00299-016-2084-x

Ye, Z., Zeng, J., Ye, L., Long, L., and Zhang, G. (2020). Genome-wide association analysis of potassium uptake and translocation rates under low K stress in Tibetan wild barley. *Euphytica* 216:35. doi: 10.1007/s10681-020-2556-5

Yu, J., Pressoir, G., Briggs, H., Vroh, I., Yamasaki, M., Doebley, J., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet*. 38, 203–208. doi: 10.1038/ng1702

Zhang, L. Y., Marchand, S., Tinker, N. A., and Belzile, F. (2009). Population structure and linkage disequilibrium in barley assessed by DArT markers. *Theor. Appl. Genet*. 119, 43–52. doi: 10.1007/s00122-009-1015-4

Zhang, M., Lu, Q., Wu, W., Niu, X., Wang, C., Feng, Y., et al. (2017). Association mapping reveals novel genetic loci contributing to flooding tolerance during germination in *Indica rice*. *Front. Plant Sci*. 8:678. doi: 10.3389/fpls.2017.00678

Zhang, X., Fan, Y., Shabala, S., Koutoulis, A., Shabala, L., Johnson, P., et al. (2017). A new major-effect QTL for waterlogging tolerance in wild barley (*H. spontaneum*). *Theor. Appl. Genet*. 130:1559. doi: 10.1007/s00122-017-2910-8

Zhang, X., Tang, B., Yu, F., Li, L., Wang, M., Xue, Y., et al. (2013). Identification of major QTL for waterlogging tolerance using genome-wide association and linkage mapping of maize seedlings. *Plant Mol. Biol. Rep*. 31:594. doi: 10.1007/s11105-012-0526-3

Zhang, X., Zhou, G., Shabala, S., Koutoulis, A., Shabala, L., Johnson, P., et al. (2016). Identification of aerenchyma formation related QTL in barley that can be effective in breeding for waterlogging tolerance. *Theor. Appl. Genet*. 129, 1167–1177. doi: 10.1007/s00122-016-2693-3

Zhou, G., Broughton, S., Zhang, X.-Q., Ma, Y., Zhou, M., and Li, C. (2016). Genome-wide association mapping of acid soil resistance in barley (*Hordeum vulgare* L.). *Front. Plant Sci*. 7:406. doi: 10.3389/fpls.2016.00406

Zhou, M. (2011). Accurate phenotyping reveals better QTL for waterlogging tolerance in barley. *Plant Breed*. 130, 203–208. doi: 10.1111/j.1439-0523.2010.01792.x

Zhou, M., Johnson, P., Zhou, G., Lic, C., and Lance, R. (2012). Quantitative trait loci for waterlogging tolerance in a barley cross of Franklin x YuYaoXiangTian Erleng and the relationship between waterlogging and salinity tolerance. *Crop Sci*. 52, 2082–2088. doi: 10.2135/cropsci2012.01.0008

Zhu, C., Gore, M., Buckler, E. S., and Yu, J. (2008). Status and prospects of association mapping in plants. *Plant Genome* 1, 5–20. doi: 10.3835/plantgenome2008.02.0089

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 09 September 2021
doi: 10.3389/fpls.2021.715910

Training Set Optimization for Sparse Phenotyping in Genomic Selection: A Conceptual Overview

Julio Isidro y Sánchez[1]* *and Deniz Akdemir*[2]*

[1] Centro de Biotecnologia y Genómica de Plantas, Instituto Nacional de Investigación y Tecnologia Agraria y Alimentaria, Universidad Politécnica de Madrid, Campus de Montegancedo, Madrid, Spain, [2] Animal and Crop Science Division, Agriculture and Food Science Centre, University College Dublin, Dublin, Ireland

Genomic selection (GS) is becoming an essential tool in breeding programs due to its role in increasing genetic gain per unit time. The design of the training set (TRS) in GS is one of the key steps in the implementation of GS in plant and animal breeding programs mainly because (i) TRS optimization is critical for the efficiency and effectiveness of GS, (ii) breeders test genotypes in multi-year and multi-location trials to select the best-performing ones. In this framework, TRS optimization can help to decrease the number of genotypes to be tested and, therefore, reduce phenotyping cost and time, and (iii) we can obtain better prediction accuracies from optimally selected TRS than an arbitrary TRS. Here, we concentrate the efforts on reviewing the lessons learned from TRS optimization studies and their impact on crop breeding and discuss important features for the success of TRS optimization under different scenarios. In this article, we review the lessons learned from training population optimization in plants and the major challenges associated with the optimization of GS including population size, the relationship between training and test set (TS), update of TRS, and the use of different packages and algorithms for TRS implementation in GS. Finally, we describe general guidelines to improving the rate of genetic improvement by maximizing the use of the TRS optimization in the GS framework.

Keywords: training set optimization, genomic selection, genome-wide markers, statistical design, sparse phenotyping, genomic prediction, mixed models

OPEN ACCESS

Edited by:
Suchismita Mondal, International Maize and Wheat Improvement Center, Mexico

Reviewed by:
Paulino Pérez-Rodríguez, Colegio de Postgraduados (COLPOS), Mexico
Gilles Charmet, INRAE Clermont-Auvergne-Rhône-Alpes, France

***Correspondence:**
Julio Isidro y Sánchez
j.isidro@upm.es
Deniz Akdemir
akdemir.work@gmail.com

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *27 May 2021*
Accepted: *10 August 2021*
Published: *09 September 2021*

Citation:
Isidro y Sánchez J and Akdemir D (2021) Training Set Optimization for Sparse Phenotyping in Genomic Selection: A Conceptual Overview. Front. Plant Sci. 12:715910. doi: 10.3389/fpls.2021.715910

1. INTRODUCTION

The rate of genetic gain in plant breeding must be enhanced to meet the demand of humanity for agricultural products in the next few decades (Xu et al., 2020). Tools, such as genomic assisted breeding (GAB), that improve the understanding of structural and functional aspects of plant genomes are key in modern breeding methods. GAB can be defined as the set of breeding tools (next-generation sequencing, omics information, and statistics) that study whole genomes by integrating multiple disciplines with new technology from informatics and robotic systems to improve selection and mating in plant breeding programs (Varshney et al., 2005, 2021). In GAB, other tools such as genetic transformation and genome editing are currently playing a key role to select better-adapted genotypes while pursuing faster genetic gains (Zhang et al., 2018). One of the emergent methodologies within GAB that have revolutionized plant and animal breeding is genomic selection (GS). GS is considered the most promising tool for genetic improvement of

the complex traits controlled by many genes, each with minor effects because (i) GS can increase the rates of genetic gain through increased accuracy of estimated breeding values (Heffner et al., 2009), (ii) significantly shorter breeding cycles (Crossa et al., 2017), and (iii) the better utilization of available genetic resources through genome-guided mate selection (Akdemir and Sánchez, 2016).

Breeders test candidate genotypes in multi-year and multi-location trials to select superior genotypes with high performance. This approach limits the number of variety candidates to be tested, and it is the main cause of the fact that plant breeding programs are time and cost-intensive. A breeding tool that combines the power of GS and the potential of an extensive collection of germplasm, assisted by new technologies, will offer promise in crop breeding to contribute to global food security (Xu et al., 2020) because it can accelerate the generation interval by reducing the generation time in plant breeding programs (Falconer and Mackay, 1996).

Bernardo (1994) was the first who proposed the use of genomic information as covariates for predicting untested genotypes but it Meuwissen et al. (2001) who came through with a new methodology to deal with the challenge of fitting prediction models when the number of genomic covariates (markers, p) is larger than the number of data points (n). Since then, simulations and empirical studies have demonstrated that GS could greatly accelerate the breeding cycle (Heffner et al., 2009), maintain genetic diversity within the breeding programs, and increase genetic gain beyond what is possible with phenotypic selection or quantitative trait loci (QTL) mapping approaches (Crossa et al., 2017). Genomic selection is a breeding tool that uses supervised machine learning approach with a training set (TRS) to predict genomic estimated breeding values (GEBVs) of an un-phenotyped test set (TS). (Isidro et al., 2016) of genotypes. The prediction of GEBVs involves a whole-genome regression model in which the known phenotypes are regressed on the markers. The GS models are trained on data that consists of both phenotypic and genome-wide markers data that is used to estimate marker (or lines) effects de los Campos et al. (2013). The combination of the marker effect estimates and the marker data from the TS is used to calculate GEBVs for the TS. The selection of individuals is based on the GEBVs as the selection criterion. The performance of the GS model is determined by calculating the correlation between GEBVs (genomic predictions) and the unknown true breeding value. As the true breeding values are never known, the available phenotypic records in the TRS are used by cross-validation values to evaluate GS. This is called prediction ability and should not be confused with prediction accuracy. The latter provides an estimate of the genotypic correlation and is estimated as the prediction ability divided by the square root of the heritability for the trait being predicted (Dekkers, 2007; Lee et al., 2008; Lorenzana and Bernardo, 2009; Riedelsheimer et al., 2012). Enhancing GS accuracy is very important for the success of GS breeding programs since the expected genetic gain from GS is directly proportional to the accuracy of GS models (Crossa et al., 2010; de los Campos et al., 2013).

There are many factors affecting the accuracy in GS by interacting in a complex network relationship (Zhong et al., 2009; Isidro et al., 2016; Liu et al., 2018; Zhang et al., 2019). Within these factors, there is one that is key to the accuracy of the prediction models in GS, and it is the design of the TRS since the predictability of a model is critical for the success of GS. In this study, the aim is to shed some light on the different TRS optimization criteria by covering the fundamentals of TRS optimization and its uses in GS, including selection strategies for long-term gains. We focus on reviewing the TRS methods from the literature that can be used as tools for designing a TRS and constructed an example to compare the TRS optimization strategies.

2. POPULATIONS IN GS

Genomic selection requires training of statistical models on available genotypic and phenotypic data from a TRS to make predictions about new genotypes. The selection of TRS involves different populations (**Figure 1**):

1. A calibration set (CS): is the group of genotypes available for the breeders from which the TRS is selected. If the individuals in this CS are phenotyped and genotyped, the populations for GS will be CS (TRS) and TS, and in theory, no need for optimization of the TRS (branch a in **Figure 1**). Nevertheless, a subset of the CS might be preferable, i.e., if very distant individuals (Lorenz and Smith, 2015) are present, to include or exclude extreme phenotypes (Lopez-Cruz and de Los Campos, 2021), or to remove irrelevant individuals (Brandariz and Bernardo, 2018). If only genotypic information is available and just a subset of them can be used for phenotyping due to budget restrictions, then a TRS will be carefully identified from the CS (branch b in **Figure 1**).
2. Training set (TRS): is where the prediction equation will be built. The TRS individuals present genotypic and phenotypic information. Under budget constraints, the aim is to select the minimum number of genotypes to phenotype, but that will assure an optimal accuracy on the TS population. The selection of the best genotypes to select from the CS to create the TRS is called optimization of the TRS. In TRS, the true response values are known (phenotypes). In this study, we used both the genotype and phenotype information from the TRS to obtain a prediction equation, which predicts the effect of each marker (or line) on the trait.
3. Remaining set population (RS): is the remaining genotypes in the CS that are used in the process of optimization. It could be also reserved for evaluating the performance of the statistical model before making predictions if the phenotypic information is available.
4. Test or Target set (TS): is the set of genotypes to predict. Only genotypic information is available in this population.

Therefore, the different populations in GS depend on whether or not the phenotypic information is available within the CS. **Figure 1** shows the distinction between the two major

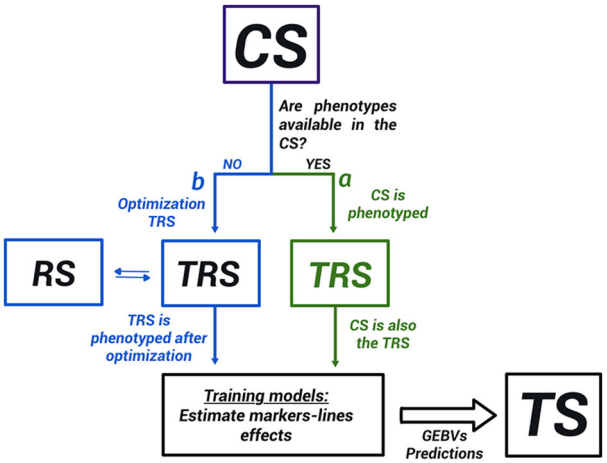

FIGURE 1 | Populations in genomic selection (GS). CS, calibration set, TRS, training set, RS, remaining set, and TS is the test set. When CS has all the phenotypic and genotypic information, CS and TRS are the same populations. Otherwise, we could have up to four different populations in the GS scheme. There are two different types of TRS selection problems: in one of these (a), the CS is already phenotyped and genotyped, and a subset of the CS is used as a TRS in the modeling stage. In the other (b), the CS is only genotyped and a TRS is constructed by phenotyping a subset of the CS.

groups of TRS optimization methods found in the literature. The first group of methods addresses the situation where the phenotypic information is already available in the CS (Neyhart et al., 2017; Brandariz and Bernardo, 2018; Lopez-Cruz and de Los Campos, 2021). They aim to use only a part of the CS when building a GS model excluding irrelevant genotypic and phenotypic information. For instance, constructing a TRS from only the individuals with high or low values of the phenotypes (Neyhart et al., 2017; Brandariz and Bernardo, 2018), or the more recently proposed sparse modeling approach Lopez-Cruz and de Los Campos (2021). The second group of methods, which is the main focus of discussion in this study, assumes that the phenotypic information is not available in the CS, and will be obtained after selecting a TRS. In this case, the resources of the breeding program are limited and just a subset of the individuals can be phenotype. In this situation, the TRS must be carefully built within the CS through an optimization process, and distinguish four different populations (CS, TRS, RS, and TS; **Figure 1**). In both groups of methods, the model validation is usually accomplished by cross-validation within the TRS (Heffner et al., 2009; Luan et al., 2009).

In general, within the TRS optimization framework, when the objective is to select a TRS to predict the remaining individuals from the same population we talk about *Un-targeted TRS*. Likewise, when a TS is first defined and genotyped, and then the TRS is optimized specifically around the TS then we define a *targeted TRS*. It is important to note, that not all optimization criteria are sensitive to this distinction, (i.e., refer next section, PAM, A-OPT, D-OPT), nevertheless, when it is so, this is reflected in how the optimization criteria are calculated (Lorenz and Smith, 2015; Akdemir and Isidro-Sánchez, 2019).

In addition, when there is heterogeneity within the environment such as row/column effects in the field, the optimal TRS of the phenotypic experiment involves not only the selection of the TRS but also the placement of genotypes in the environment (Heslot and Feoktistov, 2020). The experimental design might need blocking structure and environmental covariates and in these cases, the order in which the individuals are positioned in the environment will be important. We refer to this kind of optimization as the "ordered" optimization as opposed to the "unordered" optimization (Akdemir et al., 2021).

3. DESIGN OPTIMIZATION CRITERIA

The TRS optimization process is an optimal experimental design problem, and many aspects of GS implementation captured the attention of statisticians in the past (Smith, 1918; Kiefer, 1959; Fisher, 1960; Fedorov, 1972; Atkinson and Donev, 1992; Pukelsheim and Rosenberger, 1993; Fedorov and Hackl, 2012; Silvey, 2013). The design of the concept of the experiment should be more used to plan experimental designs in plant breeding programs and perform sets of well-selected optimization TRS to get the most informative combination out of the given factors.

The most common design optimization criteria method is indisputably the classical simple random or stratified sampling, mainly because of its simplicity and generality (Gentle, 2006), but also because of the difficulty to sample more efficiently when the number of candidate solutions is large. We classified the different design optimization criteria in to three major groups.

- Parametric design criteria are based on the assumption that the experimenter has specified a model before collecting the training data. These criteria usually depend on a scalar function of the information matrix for the model parameters which indicates the sampling variances and covariances of the estimated parameters or inferences of the model made from these models such as predictions for new individuals. Many popular designs such as the $A-$, $D-$, $E-$ criteria (Kiefer et al., 1985) are derived using a linear model as the underlying model. A linear model is a regression model where a response variable is modeled as a linear function of features that are functions of the explanatory variables plus some residual error:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{y}$ is the n dimensional vector for independent realizations of the response variable, $\boldsymbol{X}$ is the $n \times p$ design matrix for the corresponding explanatory variables and $\boldsymbol{X}$ is the $n \times q$ feature matrix, $\boldsymbol{\epsilon}$ is the n dimensional vector of independent residual terms which we assume to have mean zero and fixed variance σ_e^2 and finally, $\boldsymbol{\beta}$ is the q dimensional vector of regression coefficients. The least-squares estimator for the regression coefficients is given by $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ and for this estimator of the coefficients we can write the variance-covariance matrix as

$$Cov(\widehat{\boldsymbol{\beta}}) = \sigma_\epsilon^2((\boldsymbol{X}'\boldsymbol{X})^{-1}).$$

Now, suppose we have a certain design we want to evaluate which is expressed in a specific design matrix $\boldsymbol{X}_{TRS}$. Since

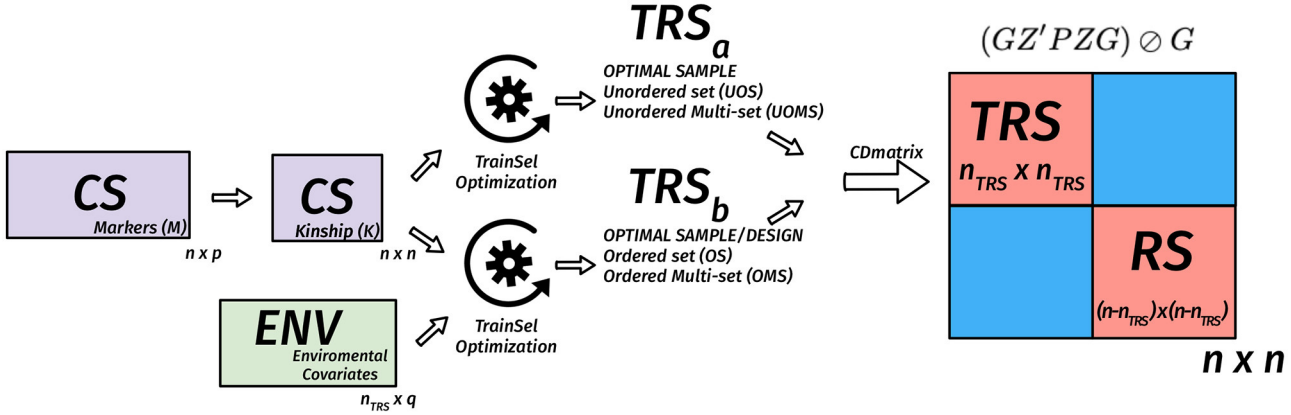

FIGURE 2 | Diagram of the matrix of the coefficient of determination (CD) criterion in the TRS optimization for unordered and ordered experiments for the un-targeted case. We assume that we have n genotypes in the candidate set, n_{TRS} genotypes are selected to the TRS. In this case, the $n \times n$ kinship matrix is used to calculate the $n \times n$ CD matrix for a given TRS, then the mean CD is calculated based on certain diagonals that correspond to TRS or remaining set (RS). The optimization algorithm is used to find the best TRS. When provided with a design matrix that has heterogeneous rows, then we are also looking for a design in addition to the selection of a TRS. In this case, both the kinship matrix and the $n_{TRS} \times q$ environmental covariates matrix are used to calculate the $n \times n$ CD matrix for a given TRS, then the mean CD is calculated based on certain diagonals that correspond to TRS or RS. The optimization algorithm is used to find the best TRS and best design with this TRS.

we can write the covariance of the estimated coefficients as $(\boldsymbol{X}'_{TRS}\boldsymbol{X}_{TRS})^{-1}$ up to a proportionality constant (which is the same for all other possible designs), we can use a function of this matrix to compare it with other designs. In general, a scalar function of this matrix is used to order the different designs. D-optimality criterion, for instance, can be expressed as $|(\boldsymbol{X}'_{TRS}\boldsymbol{X}_{TRS})|$, and designs with higher values are considered better. A-optimality criterion is expressed as $trace\left[(\boldsymbol{X}_{TRS})'\boldsymbol{X}_{TRS})^{-1}\right]$, and designs with lower values are considered better.

Some other criteria such as *CDmean*, *PEVmean*, (Laloë, 1993; Rincent et al., 2012; Isidro et al., 2015) rely on a mixed model as the underlying model: In the linear mixed-effects model of interest, the observations are assumed to result from a hierarchical linear model:

$$\boldsymbol{y} = \boldsymbol{E}\boldsymbol{\beta}_{env} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}$$

with $\boldsymbol{E}$ is the $n \times p$ design matrix for the environmental covariates, $\boldsymbol{\beta}_{env}$ is the p vector of the effects of the environmental covariates, $\boldsymbol{Z}$ is the $n \times N$ design matrix for the N genotypes in the candidate set, $\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \boldsymbol{R})$ is independent of $\boldsymbol{u} \sim N_q(\boldsymbol{0}; \boldsymbol{G})$. When using this mixed model in genomic prediction for a single environment, we use $\boldsymbol{G} = \sigma^2_k\boldsymbol{K}$ and $\boldsymbol{R} = \sigma^2_e\boldsymbol{I}$, where $\boldsymbol{K}$ is the relationship matrix of the genotypes (CS and if available the TS). When we use this mixed model with a multi-environmental genomic prediction, we assume $\boldsymbol{G} = \boldsymbol{V}_k \otimes \boldsymbol{K}$ and $R = \boldsymbol{V}_e \otimes \boldsymbol{I}$.

For this model, the CD matrix of $\widehat{\boldsymbol{u}}$ for predicting $\boldsymbol{u}$ is given by

$$(\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G}) \oslash \boldsymbol{G}$$

where $\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{E}(\boldsymbol{E}'\boldsymbol{V}^{-1}\boldsymbol{E})^{-1}\boldsymbol{E}'\boldsymbol{V}^{-1}$ is the projection matrix and $\oslash$ expresses the element-wise division. Usually, the mean of certain diagonal elements of the CD matrix is used to measure the quality of a sample. For instance, in a targeted design, the mean of the diagonal elements that correspond to the TS genotypes are used. When the design is un-targeted, we can use the mean over the diagonals that correspond to the remaining set. Another approach involves the calculation of the CD matrix for a given set of contrasts then taking the mean of the diagonals of this matrix (Rincent et al., 2012, 2017). In **Figures 2–4**, we diagrammatically illustrate the different populations, input matrices, the different parts of the CD matrix, and the process of optimization.

- Non-parametric designs criteria are model-free, i.e., they do not rely on models we intend to use with the resulting data. Some nonparametric designs are based on distance or similarity measures and aim to spread the TRS over the design space (space-filling design). Different measures or metrics quantify how a set of points is spread out. Some examples are: (i) partition around medoids (PAM) where the objective is to find a sequence of objects called medoids that are centrally located in clusters for a given distance measure, (ii) the maximin criteria are such that the minimum distance among the TRS is maximized, (iii) the minimax design (Johnson et al., 1990) where the TRS is such that the maximum of the minimum distances from the TRS to the rest of the CS or the TS is minimized, (iv) the Latin hypercube sampling divides the design region evenly into cubes and ensuring that the sample contains just one point in each such segment and aims at ensuring that each of the scalar inputs has the whole of its range well scanned, according to a probability distribution, and (v) the minimum spanning tree (MST) (Dussert et al., 1986). An MST is a tree that connects all the candidate design points and whose total edge lengths are minimal. Once a spanning tree of the candidate points is built, the mean and SD of edge lengths can be calculated. The spanning trees with the

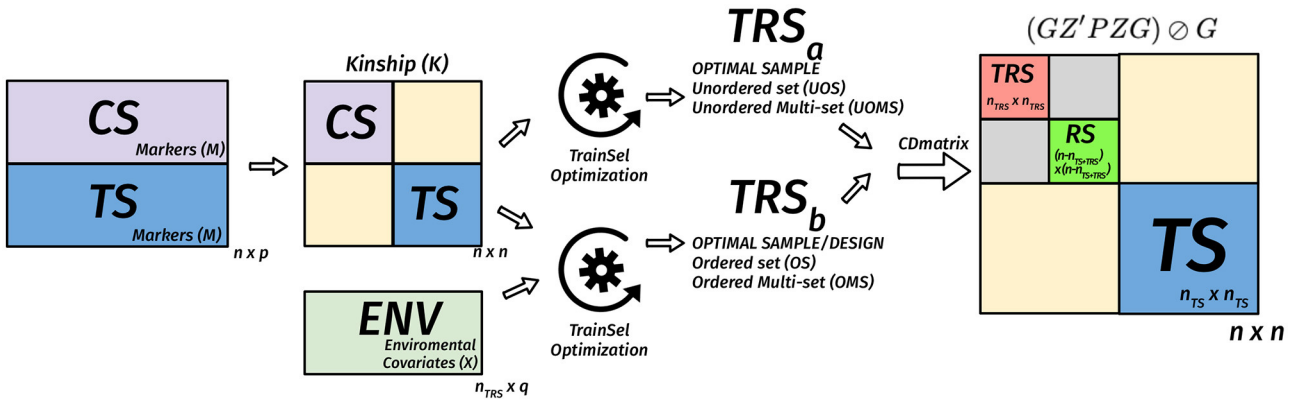

FIGURE 3 | Diagram of the matrix of the CD criterion in the TRS optimization for unordered and ordered experiments for the targeted case. We assume that we have n_{TS} genotypes in the TS, $n - n_{TS}$ genotypes in the CS, n_{TRS} genotypes are selected to the TRS from the CS. In this case, the $n \times n$ kinship matrix is used to calculate the $n \times n$ CD matrix for a given TRS, then the mean CD is calculated based on certain diagonals that correspond to TS. The optimization algorithm is used to find the best TRS. When provided with a design matrix that has heterogeneous rows, then we are also looking for a design in addition to the selection of a TRS. In this case, both the kinship matrix and the $n_{TRS} \times q$ environmental covariates matrix are used to calculate the $n \times n$ CD matrix for a given TRS, then the mean CD is calculated based on diagonals that correspond to TS. The optimization algorithm is used to find the best TRS and best design with this TRS.

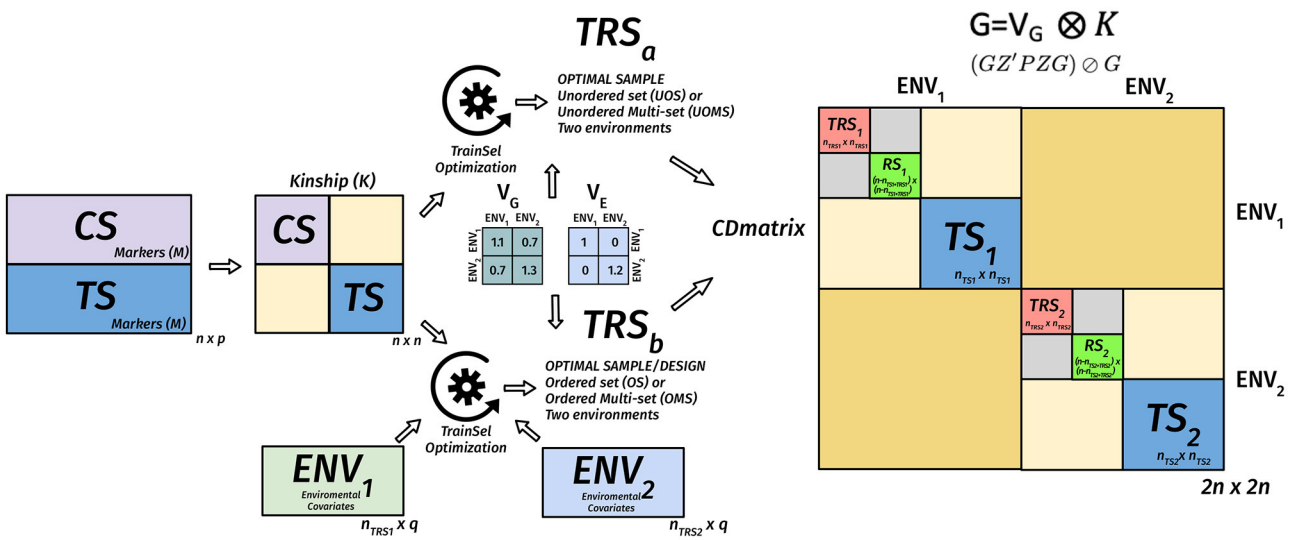

FIGURE 4 | Diagram of the matrix of the CD criterion in the TRS optimization for unordered and ordered experiments for the targeted case in a multi-environmental phenotypic experiment. We assume that we have n genotypes, n_{ts} of them are in the target set (TS), the remaining of in the candidate set, n_{TRS_1} genotypes are selected to the TRS in environment 1, n_{TRS_2} genotypes are selected to the TRS in environment 2. Two environments are assumed to have a positive genetic covariance, and this is expressed in V_k. The residual genetic covariance expressed in VE is diagonal, meaning that errors are uncorrelated between the two environments. These covariance matrices along with the genomic relationship matrix and if provided environmental covariates matrices for the environments are used to calculate the CD matrix ($2n \times 2n$) for a given design. The mean of the diagonals of this matrix that correspond to the TS is used as a criterion for evaluating different designs. The optimization algorithm tries to find the design that maximizes this criterion.

smallest mean are called minimal and among them, the ones with high variance are preferred. A TRS from an MST can be obtained by recursively pruning out, from the candidate set, the candidate points on the leaves of the MST with small edge lengths (Guo et al., 2019).

Non-parametric designs such as space-filling designs are well suited to the initial exploration objective. They can be used to select a smaller candidate set from a bigger candidate set to reduce the computational complexity of optimizing parametric design criteria.

- **Multiple design criteria**. Multiple models optimal experimental design criteria try to overcome the choice issue by combining more than one criteria into one *via* some type of averaging on multiple-objective optimization methods (Pukelsheim, 1993; Akdemir and Sánchez, 2016). In this approach, the Pareto front approach is used to evaluate several

criteria. The Pareto front is a set of non-dominated designs, i.e., as compared to the design points on the frontier, no other design point can be found that does not degrade at least one of these criteria values (as shown in **Figure 5**).

Many GS experiments will be performed in several environments and then the TRS optimization aims to find subsets of genotypes from the candidate set to be tested in each of the environments and perhaps the corresponding designs within some of these environments to address the heterogeneity within environments. The use of CD for this situation is illustrated in a diagram in **Figure 4**.

4. TRS OPTIMIZATION FOR SPARSE PHENOTYPING

The most important current bottleneck in plant breeding programs is the phenotypic evaluation (Crossa et al., 2017). Although genotyping is still costly, next-generation sequencing has decreased genotyping cost more than 100K folds in the last 20 years (National Human Genome Research Institute, 2020), and therefore, phenotyping needs to be optimized within a breeding program. The use of GS in breeding programs is potentially costly without the careful design of populations. When designing the implementation of the GS scheme into the breeding cycle, breeders need to focus first on several aspects: (i) to generate a specific breeding database for GS, (ii) to choose the filial generation to start GS, and (iii) to select the TRS to start GS modeling (Albrecht et al., 2011; Clark et al., 2012). The design of the TRS, also called optimization of the TRS, is the breeding process that uses the information from these aspects to create a TRS to start the GS process.

Training set optimization consists of choosing (within a panel of candidates) a set of training individuals that will better predict un-phenotyped germplasm in a TS. TRS optimization has attracted notable interest in the breeding community for several reasons (**Table 1**). First, the fact that predictions are based on markers or line effects calculated on the TRS raises the question of how to select the TRS to increase the efficiency and effectiveness of GS. Second, currently, the high cost of phenotyping makes the phenotype information the most important constraint in plant breeding programs. Better allocation of resources within plant breeding programs by observing a small size but representative TRS would reduce phenotypic cost and increase the quality of the phenotypic data by focusing on more expensive traits with more sophisticated instruments, or increasing complementary measurements of the same traits (sparse or selective phenotyping). Third, the traditional optimization process based on random sampling as a strategy to create the TRS does not always lead to an increase in predictive ability due to the under or over-representation of the genetic information in the TRS. The TRS optimization aims to enhance the process of sparse phenotyping, to reduce the cost of phenotyping while maintaining high prediction accuracy models.

Two important aspects within the TRS optimization are the fact that the TRS is a dynamic populations that must be updated through the breeding cycle program, and also that the TS needs to be into account when building the TRS (Akdemir et al., 2015).

The design of the TRS was initially started in animal breeding (Habier et al., 2007, 2010; Clark et al., 2012; Pszczola et al., 2012). These studies and others in plants (Windhausen et al., 2012; Wientjes et al., 2013) were focused on the importance of the relatives for the makeup of the TRS and on how to update the TRS to improve genomic prediction across generations. They highlighted how the TRS should be composed in terms of resemblance between TRS and TS, but they did not perform any optimization process, TRS was selected randomly. A random sampling of genotypes from a CS is a risky procedure because could lead to low-quality coverage of the total genetic space especially when the CS contains population structure (Windhausen et al., 2012; Isidro et al., 2015; Bustos-Korts et al., 2016). In the last decade, many studies (**Table 1**) examined the importance of optimization of the TRS by comparing specific selection criteria to random sampling.

The first study highlighting the importance of using statistical approaches to develop an optimal TRS was shown by Rincent et al. (2012) (**Table 1**). In this study, the objective was to define which individuals from a calibration (candidate) set are the optimal ones to predict a selection (TS) candidates. The idea was to use a criterion that could minimize genetic similarity within the TRS, because of the more similar the individuals within the TRS, the more duplication of the alleles, and therefore, more redundancy. Based on concepts from the mixed model equations introduced by Laloë (1993), Rincent et al. (2012) introduced criteria that aimed to maximize the reliability CD, the square correlation between GEBVs and true breeding values or minimized the prediction error variance (PEV) on the CS. In this study, they used a generalized version of CD and PEV (the contrast between breeding values). They showed that the optimization criteria improved prediction accuracy when comparing with random sampling. Rincent et al. (2012) have shown that mean of the coefficient of determination (CDmean) captured more genetic variability when building the TRS than mean of the prediction error variance (PEVmean) and that an optimized set of 100 lines achieved on average the same prediction accuracy as a set of 200 lines selected at random.

Isidro et al. (2015) proposed stratified sampling and stratified CD as alternative algorithms to improve the optimization of TRS under population structure effects. The optimization of the TRS based on genomic relationships resulted in higher prediction accuracies when compared with random sampling. In this study, they concluded that the optimization of the TRS depended on the interaction of trait architecture and population structure and on the ability of the algorithm to capture phenotypic variance. In the same year, Akdemir et al. (2015) derived a computationally efficient approximation to the PEV based on principal components of the genotypes as a criterion for TRS design that showed less computational burden than previous criteria. These studies were the first ones that open the door to other strategies to optimize the TRS. Bustos-Korts et al. (2016) proposed a TRS construction method that uniformly sampled the genetic space comprised by the target population (TS) of genotypes, although, the results were similar to CDmean.

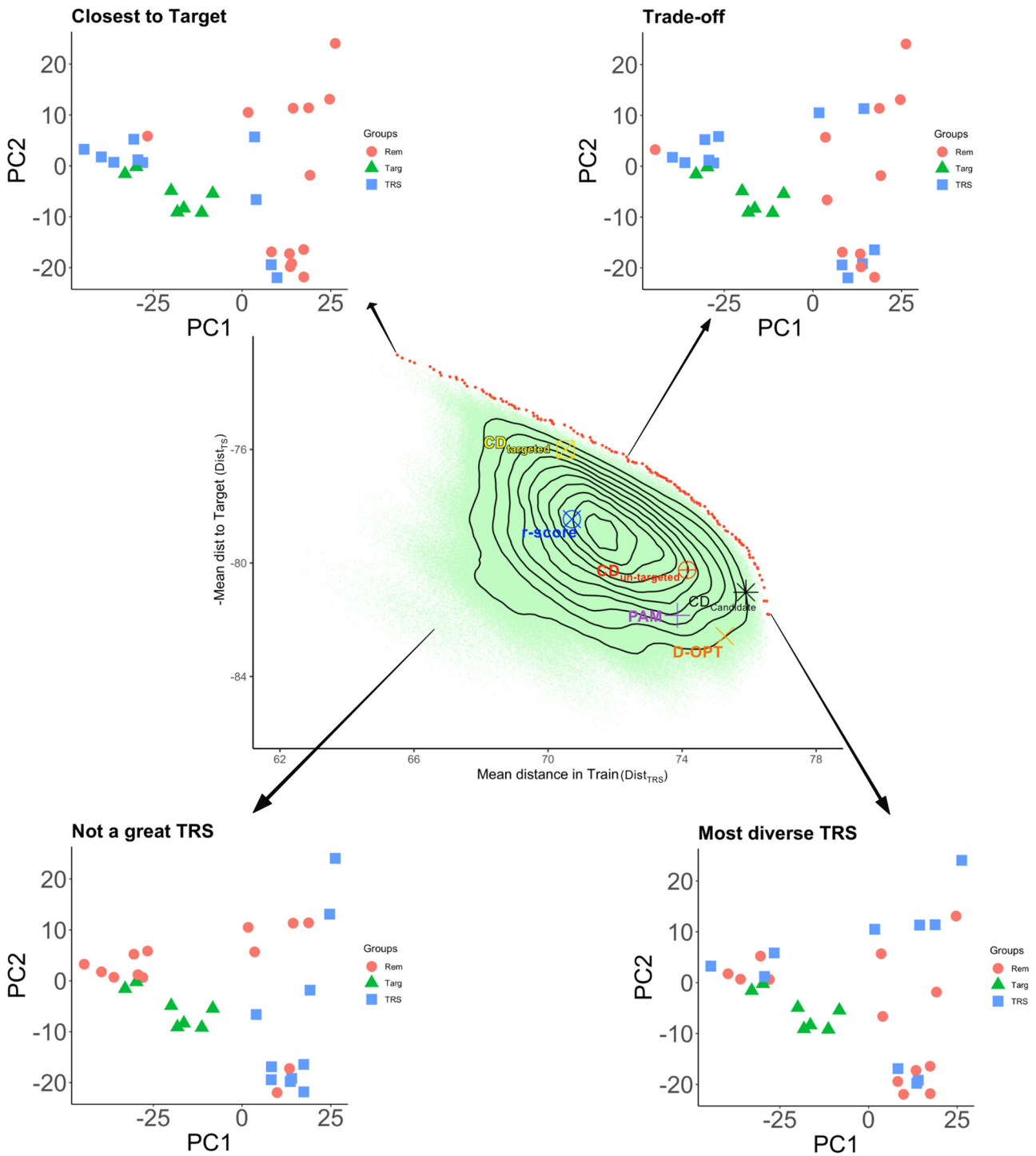

FIGURE 5 | Pareto front for minimizing the mean genomic distance of TRS genotypes to the TS genotypes (i.e., maximizing the negative of this quantity), and maximizing the mean genomic distance among the TRS genotypes. The figure represents a toy example where a sample of size 10 is selected from 23 candidate genotypes to predict 7 TS genotypes (the total number of different solutions is 23 choose 10 which is more than a million). Target genotypes along with selected TRS and the remaining sets are displayed in the genotypic space represented by the first two principal components of the marker matrix. With different symbols and colors, we indicate the optimal CD TRS's for targeted and un-targeted cases, D-optimal TRS, and the TRS selected by PAM. The red dots are the TRS that are on the Pareto front, i.e., no other TRS will be better than any of these for both criteria (non-dominated solutions). All the brown dots are dominated by the same two criteria. We get the most diverse set when the mean genetic distance in the TRS is maximal. We get a TRS closest to the TS when we minimize the mean genetic distance (maximize the negative) of TRS to TS. All of the parametric design criteria and PAM are dominated. Among those, CDmean targeted gives a TRS that is close to the TS. The remaining optimal TRS's are genetically diverse. The most genetically diverse set among the optimization criteria is the CDmean calculated for all genotypes in CS.

TABLE 1 | Key relevant scientific studies on training set (TRS) optimization.

Study	CDmean	PEVmean	Clustering	Other criteria	Package
Rincent et al. (2012)	✗	✗	–	–	Own code
Isidro et al. (2015)	✗	✗	✗	–	Own code
Akdemir et al. (2015)	✗	✗	✗	–	**STPGA**
Lorenz and Smith (2015)	–	–	–	Levels of TRS relationship	Own code
Bustos-Korts et al. (2016)	✗	–	✗	Uniform Sampling	Own code
He et al. (2016)	–	–	–	Random	–
Rincent et al. (2017)	✗	–	✗	CDpop and Crit_Kin	Own code
Neyhart et al. (2017)	✗	✗	–	Top and bottom proportion	Own code
Cericola et al. (2017)	–	–	–	Random sampling	Own code
Momen and Morota (2018)	✗	✗		Additive and Non-additive	Own code
Norman et al. (2018)	–	–	✗	Random	Own code
Akdemir and Isidro-Sánchez (2019)	✗	✗	–	D and A-OPT	**STPGA**
Ou and Liao (2019)	✗	✗	✗	r-score	**TSDFGS**
Mangin et al. (2019)	✗	✗	–	EthAcc	Own code
Guo et al. (2019)	✗	✗	PAM	FURS	**STPGA**
de Bem Oliveira et al. (2020)	–	–	–	Random, Family Random	Own code
Adeyemo et al. (2020)	–	–	✗	–	Own code
Mendonça and Fritsche-Neto (2020)	–	✗		–	**STPGA**
Olatoye et al. (2020)	✗	–	–	Random	Own code
Roth et al. (2020)	–	✗		Maximum and Mean relationship	**STPGA**
Sarinelli et al. (2019)	–	✗	✗	–	Own code
Tayeh et al. (2015)	✗	–	–	–	Own code
Atanda et al. (2021)	✗	–	–	Avg_GRM	Own code
Yu et al. (2020)	–	–	–	Upper Bound reliability	Own code
Ben-Sadoun et al. (2020)	✗	–	–	CDmean-multi	Own code
Heslot and Feoktistov (2020)		–	–	PEVridge	Own code
Akdemir et al. (2021)	✗	✗	✗	–	**TrainSel**
Kadam et al. (2021)	✗	✗	–	–	**STPGA**

CDmean, Mean of the coefficient of determination; PEVmean, Mean of the predictor error variance. A cross in the cell indicates that the criterion has been used for TRS optimization. Criteria different than CD, PEV, and Clustering are shown in the column Other Criteria. The software using R is specified in the Package column.

Other studies also stressed the importance of considering an other way to construct the TRS by random sampling (Lorenz and Smith, 2015; He et al., 2016; Cericola et al., 2017; Neyhart et al., 2017; Norman et al., 2018; de Bem Oliveira et al., 2020; Olatoye et al., 2020), clustering approaches (Akdemir et al., 2015; Isidro et al., 2015; Bustos-Korts et al., 2016; Rincent et al., 2017; Norman et al., 2018; Guo et al., 2019; Sarinelli et al., 2019; Adeyemo et al., 2020), by using different levels of relatedness between TRS and TS (Lorenz and Smith, 2015; Berro et al., 2019; Roth et al., 2020) or by using other alternatives algorithms to CD-mean and PEV-mean such as different design matrix algorithm (Akdemir and Isidro-Sánchez, 2019), estimated theoretical accuracy (EthAcc) (Mangin et al., 2019), upper bound reliability (Yu et al., 2020), or the Fast and Unique Representative Subset Selection (FURS) (Guo et al., 2014). A criterion that is derived directly from Pearson's correlation between GEBVs and phenotypic values of the TS derived from the GBLUP model showed higher predictive ability than CD and PEV (Ou and Liao, 2019). Most aforementioned approaches above, do not use information from the TS while building the TRS, which is detrimental for prediction accuracy (Lorenz and Smith, 2015; Akdemir and Isidro-Sánchez, 2019; Ou and Liao, 2019). The main reason for the decrease in accuracies is because the most informative TRS to predict the TS is the one where individuals are more closely related to the TS. This is because when pairs of individuals are closely related, they tend to inherit QTL blocks in the same linkage phase (Andreescu et al., 2007; Habier et al., 2010). This is especially critical when there is low marker density coverage because the assumption in GS of getting at least one marker in QTL with the trait of interest will not be perfectly met. The genetic relatedness between TRS and TS was addressed by Lorenz and Smith (2015), Rincent et al. (2017), and Akdemir and Isidro-Sánchez (2019). Recently, Atanda et al. (2021) used the average genomic relationship (Avg_GRM in **Table 1**) between a specific line in the TRS and all lines in the TS, and they statistically significant increase in the accuracies when compared with CD in some bi-parental populations. Nevertheless, this approach as in Rincent et al. (2017) did not consider the possible alleles duplication within the TRS.

Training optimization selection also has been used for pre-breeding discovery. Tanaka and Iwata (2018) proposed a strategy that used genomic prediction in pre-breeding for discovering

the best genotypes from a large number of candidates. They demonstrated by simulation that their Bayesian optimization could reduce the number of phenotyped accessions needed to find the best accession among a large number of candidates. Their strategy was based on predict uncertainty of the prediction rather than based only on high predicted values. Following this strategy, Tsai et al. (2021) used an augmented expected improvement for sequential phenotyping to identify the best individual from the CS. It is important to note that these studies are not focusing on building a TRS for GP, but on identifying the best candidate to be used for commercial or mating purposes. These approaches could be used when phenotyping is very expensive and not very time-consuming.

In the area of hybrid breeding, the optimization of the TRS is even more critical than in other breeding systems, since the selection of superior F1 hybrids (single crosses between fully inbred lines) implies developing first inbred lines and then identifying the best hybrid combinations between them. To facilitate this process, breeders typically split germplasm into complementary heterotic groups and select lines within each group for their ability to produce good hybrids when crossed to lines from a complementary group. The fullest assessment of single-cross performances would be a complete factorial mating design achieved by making all possible single crosses. However, the high number of lines to be evaluated per heterotic group makes this approach prohibitive (i.e., for 1,000 lines in each heterotic group, there would be 1 million possible crosses). Genomic models have been applied to hybrid prediction mainly in maize (Bernardo, 1994; Schrag et al., 2009; Technow et al., 2014; Kadam et al., 2016; Marulanda et al., 2016; Fristche-Neto et al., 2018; Seye et al., 2020), and wheat (Zhao et al., 2013, 2014, 2015; Longin et al., 2015; Marulanda et al., 2016; Schulthess et al., 2017), and less in other species such as rye (Wang et al., 2014) or sunflower (Reif et al., 2013; Mangin et al., 2017; Dimitrijevic and Horn, 2018; Heslot and Feoktistov, 2020). These studies have emphasized the interest in using TRS optimization compared to the traditional crossing designs.

In general, most of the TRS studies have used model-based parametric criteria (CDmean, PEVmean, and r-score), followed by non-parametric (i.e., PAM, FURS), and just a few studies used their own criteria (i.e., AvgGRM, U score) (**Table 1**). All these studies show that there is not a universal criterion to create a TRS. It will mainly depend on linkage disequilibrium between markers on TRS vs. TS, the relationship between TRS and TS (Habier et al., 2007; Goddard, 2009), the genetic architecture of the trait (McClellan et al., 2007; Jannink, 2010; Burstin et al., 2015), trait heritability (Hayes et al., 2009), and population structure effects (Isidro et al., 2015; Rincent et al., 2017).

To shed some light on the different TRS optimization criteria, we constructed a toy example where we compared several design criteria (CD, PAM, D-OPT, and r.score) with each other (**Figure 5**). In this example, there were 30 genotypes in total, seven of these genotypes were selected as the TS. The remaining 23 genotypes were used as the CS. We set the TRS size to 10, giving 23 choose 10 (1144066) different TRS possibilities. For each of these designs, we calculated the value of the mean genetic distance among the TRS ($Dist_{TRS}$), and the negative of the mean genomic distance from TRS to the TS ($Dist_{TS}$). In the Figure, the red dots are the TRS that are on the Pareto front, i.e., no other TRS will be better than any of these for both criteria (non-dominated solutions). Balancing the $Dist_{TRS}$ and $Dist_{TS}$ in the Pareto front gives you different TRS. For instance, when we minimize the mean genetic distance (maximize the negative) of TRS to TS, we obtained a TRS closest to the TS (top left graph). We get the most diverse TRS when the $Dist_{TRS}$ in the TRS is maximal (bottom right graph). If you balance both distances, then we get a TRS where there is a trade-off between $Dist_{TRS}$ and $Dist_{TS}$. The remaining TRS on the same plot is dominated with respect to the same two criteria. A TRS is dominated if we can find another TRS that improves at least one of these criteria without deteriorating the other criterion value. All of the design criteria and PAM are dominated with respect to $Dist_{TRS}$ and $Dist_{TS}$. Among those, CDmean targeted gives a TRS that is close to the TS, where CDmean calculated over the candidate set (CDMEAN-Cand) comes very close to the most diverse design. The contours of the density of $Dist_{TRS}$ and $Dist_{TS}$ over 1144066 different TRS possibilities show that a random design on average would be dominated by all of the optimal samples and would fall far away from the Pareto frontier. It is important to understand the different trade-offs involved in choosing a good TRS since this will help the experimenter to choose a suitable TRS or a TRS selection criterion among the alternatives.

Breeding programs usually deal CS's with 1,000's or 10,000's of genotypes. Although direct enumeration of all the possible TRS's is not possible in these cases, multi-objective optimization techniques can be utilized to approximate the frontier curves and single-objective optimization tools can be used to find optimal TRS's according to several single criteria. Then a plot similar to the one presented in **Figure 5** can be produced to evaluate the trade-offs among different designs. When the number of genotypes in the CS is so large that computationally intensive methods are prohibitive, we recommend using a less intensive method such as PAM or stratified sampling (Isidro et al., 2015; Guo et al., 2019), or one of the space-filling designs to reduce the number of CS to a manageable size ahead of comprehensive analysis. A practical overview of the statistical analysis needed to optimize the TRS using R and issues associated with the analysis have been addressed along with the R code in the study by Isidro y Sánchez et al. (2022). In addition, extra information can be found in the extensive vignette (https://github.com/TheRocinante-lab/TrainSel/blob/main/inst/TrainSelUsage.pdf).

5. SOFTWARE TOOLS FOR TRS OPTIMIZATION

While the practical use of TRS optimization in GS is supported by the literature, as shown above, the number of software tools for implementation is limited. As far as we are concerned, just three software have been developed and available for public use. The package STPGA Akdemir (2017) is an R package that uses a modified GA for solving subset selection problems but also allows users to chose from many predefined or user-defined criteria. Similarly, the package TSDFGS Ou and Liao (2019) is

an R package that focuses on optimization of the TRS by a genetic algorithm (GA) and can be used for TRS optimization based on three built-in design criteria [CDscore, PEVscore, and Pearson correlation (r-score)]. Recently, Akdemir et al. (2021) designed a new package called TrainSel to provide many more options than previous software. For example, TrainSel can select multiple sets from multiple candidate sets, users can specify whether or not the resulting set needs to be ordered, or the power to perform multi-objective optimization. In addition, TrainSel can be used for searching for solutions to a variety of TRS and experimental design problems, such as randomized complete block design, and lattice design, etc. Furthermore, it can be also used in combinatorial optimization problems for supervised and also unsupervised learning. The strength of TrainSel is that it combines TRS optimization with a particular experimental design, which has not been implemented in both of the above alternatives by Akdemir et al. (2021).

6. GENERAL GUIDELINES FOR A GOOD TRS

In this study, we highlight some of the guidelines learned from the literature when building an optimal TRS:

- When building the first TRS is key to keep, within the TRS, the historical germplasm used to generate the breeding populations. This will allow capturing the allelic diversity within the breeding program.
- The larger the TRS size the better predictions (Daetwyler et al., 2008; Zhong et al., 2009), since most characters are quantitative with a large number of loci and a very small effect size. The number of loci affecting quantitative characters likely ranges from 2,000 to 4,000 (MacLeod et al., 2016). Although adding genetically distant individuals might decrease accuracy (Lorenz and Smith, 2015), this is not a general rule. In addition, large TRS are needed to capture rare alleles at high frequencies to obtain a reliable estimate of their effects (MacLeod et al., 2016), even for highly quantitative traits if the rare allele is present in the sequencing or the genotyping is done from coding and regulatory regions.
- Markers can capture genetic relationships among genotypes, thereby affecting the accuracies of GEBVs (Habier et al., 2007). Therefore, a genetic relationship between TRS and TS is needed to obtain high accuracies. In general, a TRS should maximize the relationship with the TS (Albrecht et al., 2011; Pszczola et al., 2012; Akdemir and Isidro-Sánchez, 2019), but should minimize the relationship within the TRS (Clark et al., 2011; Lorenz, 2013; Bustos-Korts et al., 2016; Pszczola and Calus, 2016). That is to say, if TRS and TS come from different populations or breeding generations, a drop in accuracy is expected. The main reasons for the drop in accuracy are because LD between markers and QTL, or that QTL allele frequencies and/or effects can differ among populations (Hayes et al., 2009; Wientjes et al., 2015, 2017). The difference in allele frequencies between TRS and TS can affect prediction accuracy because allele frequencies can affect the estimated genomic relationship matrix when GBLUP models are implemented.
- The TRS must be updated with new genotyped and phenotyped individuals to assure the accuracy of GEBVs, is maintained over generations. Otherwise, recombination events will decrease LD between markers and QTL (Auinger et al., 2016). As phenotypes are the current bottleneck in plant breeding programs, the quality of the phenotypes is critical to the TRS optimization.
- The design of the TRS highly depends on the TS population. For example, if your TS is highly diverse, your TRS must be built to capture that diversity, otherwise, a significant drop in accuracy might occur. That is why targeted optimization approaches are chosen when building TRS (Akdemir and Isidro-Sánchez, 2019; Akdemir et al., 2021). From **Figure 5** we can observe that we get a TRS closest to the TS when we minimize the mean genetic distance (maximize the negative) of TRS to TS. Among the different TRS selection criteria, CDmean targeted gives a TRS that is close to the TS. The remaining optimal TRS's are genetically diverse but the most genetically diverse set among the optimization criteria is the CDmean calculated for all genotypes in CS. This type of evaluation of different design criteria together along with a frontier curve should shed some light on the selection of a particular TRS.
- If certain QTL with large effects for traits of interest exists, then these QTL can be given more influence while selecting the TRS. This could be done, for example in the mixed modeling framework by using the QTL as fixed effects (Spindel et al., 2016). In the non-parametric approach, more weights can be given when calculating the genetic distance matrix.
- In general, optimization criteria from mixed model theory (CDmean, PEVmean) performs better than random sampling under most scenarios, except for scenarios with a large population structure where these criteria might not be optimal (Isidro et al., 2015).

7. PERSPECTIVES FOR TRS OPTIMIZATION

Genomic selection is an emergent methodology that revolutionized plant and animal breeding, by using a statistical framework that uses genome-wide markers to predict breeding values for key breeding traits. In this framework, one critical step is how to select the best individuals to train the statistical models. As shown above, there has been quite a great research in this area, but there are still some questions to be answered. Following the literature, there is no "best" strategy to optimize the TRS, and therefore, a comparison between algorithms focusing on the different factors affecting the TRS on different populations would be helpful to answers some questions regarding TRS optimization.

We envision a substantial benefit applying TRS optimization methods to hybrid prediction, and also sparse testing in multi-environment, and multi-trait experiments (Jarquín et al., 2014; Akdemir et al., 2021; Crossa et al., 2021). For instance, in hybrid

prediction, TRS are traditionally constructed by methods such as top crosses, North Caroline design, etc. It has been shown that the TRS optimization methods improve hybrid prediction accuracies when comparing with the traditional design methods (Zhao et al., 2015, 2021; Fristche-Neto et al., 2018; Heslot and Feoktistov, 2020; Yu et al., 2020; Technow et al., 2021).

It is also expected that TRS selection methods will be used more commonly in multi-environmental phenotypic experiment design (Montesinos-López et al., 2019; McGowan et al., 2020) as more flexible and powerful tools such as the package R TrainSel becomes available for researchers. The use of genomic information in designing these experiments shifts the attention from replication of individuals to replication and representation of alleles in different environments.

In addition, more studies using haplotypes rather than just markers are needed, since accuracies are greater if TRS and TS share long-range haplotypes (Akdemir et al., 2015; Meuwissen et al., 2016; Scott et al., 2021). The decrease of whole genomic sequencing is allowing us to develop pan-genomes studies of many crops, which will allow us to switch from SNPs to longer more important haplotypes in the design of TRS populations. The development of haplotype-informed DNA markers will enable the selection of new haplotype combinations, which will increase the opportunity to attain optimized genetic combinations for improved performance and disrupt linkage drag (Varshney et al., 2021).

An unresolved issue in TRS optimization is the determination of the size of TRS. The size of TRS is usually dictated by the budget for the experiment, however, a breeder might need guidance for selecting a TRS size to avoid redundancy of individuals. For example, even though a breeder might have the resources to do 20 individuals, the breeder should know what is the optimal size to experiment. The optimal size of the TRS can be obtained from the multi-objective optimization framework Akdemir et al. (2019). The solutions on the Pareto front of an optimization problem Markowitz (1968), where one or more design criteria along with the TRS size are optimized, will provide the experimenter with a scenery of the optimal design space at each sample size. The usual methods of selecting a solution on a frontier can guide the determination of the TRS size.

Finally, a comparison of criteria with different populations, different genetic architectures, heritability values, and relationships among TRS and TS is needed, especially to evaluate if some previous claims in the TRS optimization area are true under the same population scenarios.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

AUTHOR CONTRIBUTIONS

JIS and DA: conception and design of the article, drafting the article, and critical revision of the article. Both authors contributed to the article and approved the submitted version.

FUNDING

Results have been achieved within the framework of the first transnational joint call for research projects in the SusCrop ERA-Net Cofound on Sustainable Crop production, with funding from the Department of Agriculture, Food and the Marine grant no. 2017EN104. This project has also received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 818144. JIS was supported by the Beatriz Galindo Program (BEAGAL18/00115) from the Ministerio de Educación y Formación Profesional of Spain and the Severo Ochoa Program for Centres of Excellence in R&D from the Agencia Estatal de Investigación of Spain, grant SEV-2016-0672 (2017-2021) to the CBGP.

ACKNOWLEDGMENTS

The authors thank all their funders as well as Breicon Genomics LTD for its expertise in genomic-assisted breeding. JIS thanks his family, Francisco Isidro Muñoz, Maria de Gracia Sánchez Sánchez, Antolin y María IS, his dearest friends "El oso" and Bradley, and "Mi collares" "JU&MA" (You are all the cheese of my macaroni). DA is indebted for all the support and help given to him by his dear parents, family, and friends, without which, none of this would be possible.

REFERENCES

Adeyemo, E., Bajgain, P., Conley, E., Sallam, A. H., and Anderson, J. A. (2020). Optimizing training population size and content to improve prediction accuracy of fhb-related traits in wheat. *Agronomy* 10, 543. doi: 10.3390/agronomy10040543

Akdemir, D. (2017). *STPGA: Selection of Training Populations by Genetic Algorithm*. R package version 5.2.1.

Akdemir, D., Beavis, W., Fritsche-Neto, R., Singh, A. K., and Isidro-Sánchez, J. (2019). Multi-objective optimized genomic breeding strategies for sustainable food improvement. *Heredity* 122, 672–683. doi: 10.1038/s41437-018-0147-1

Akdemir, D., and Isidro-Sánchez, J. (2019). Design of training populations for selective phenotyping in genomic prediction. *Sci. Rep*. 9, 1–15. doi: 10.1038/s41598-018-38081-6

Akdemir, D., Rio, S., and y Sánchez Julio, I. (2021). Trainsel: an r package for selection of training populations. *Front. Genet*. 12:607. doi: 10.3389/fgene.2021.655287

Akdemir, D., and Sánchez, J. I. (2016). Efficient breeding by genomic mating. *Front. Genet*. 7:210. doi: 10.3389/fgene.2016.00210

Akdemir, D., Sanchez, J. I., and Jannink, J.-L. (2015). Optimization of genomic selection training populations with a genetic algorithm. *Genet. Sel Evol*. 47:38. doi: 10.1186/s12711-015-0116-6

Albrecht, T., Wimmer, V., Auinger, H., Erbe, M., Knaak, C., Ouzunova, M., et al. (2011). Genome-based prediction of testcross values in maize. *Theor. Appl. Genet*. 123, 339–350. doi: 10.1007/s00122-011-1587-7

Andreescu, C., Avendano, S., Brown, S. R., Hassen, A., Lamont, S. J., and Dekkers, J. C. (2007). Linkage disequilibrium in related breeding lines of chickens. *Genetics* 177, 2161–2169. doi: 10.1534/genetics.107.082206

Atanda, S. A., Olsen, M., Burgue no, J., Crossa, J., Dzidzienyo, D., Beyene, Y., et al. (2021). Maximizing efficiency of genomic selection in cimmyt's tropical maize breeding program. *Theor. Appl. Genet*. 134, 279–294. doi: 10.1007/s00122-020-03696-9

Atkinson, A., and Donev, A. (1992). *Optimum Experimental Designs*. Clarendon. Oxford.

Auinger, H.-J., Schönleben, M., Lehermeier, C., Schmidt, M., Korzun, V., Geiger, H. H., et al. (2016). Model training across multiple breeding cycles significantly improves genomic prediction accuracy in rye (secale cereale l.). *Theor. Appl. Genet*. 129, 2043–2053. doi: 10.1007/s00122-016-2756-5

Ben-Sadoun, S., Rincent, R., Auzanneau, J., Oury, F., Rolland, B., Heumez, E., et al. (2020). Economical optimization of a breeding scheme by selective phenotyping of the calibration set in a multi-trait context: application to bread making quality. *Theor. Appl. Genet*. 133, 2197–2212. doi: 10.1007/s00122-020-03590-4

Bernardo, R. (1994). Prediction of maize single-cross performance using rflps and information from related hybrids. *Crop. Sci*. 34, 20–25. doi: 10.2135/cropsci1994.0011183X003400010003x

Berro, I., Lado, B., Nalin, R. S., Quincke, M., and Gutiérrez, L. (2019). Training population optimization for genomic selection. *Plant Genome* 12, 190028. doi: 10.3835/plantgenome2019.04.0028

Brandariz, S. P., and Bernardo, R. (2018). Maintaining the accuracy of genomewide predictions when selection has occurred in the training population. *Crop. Sci*. 58, 1226–1231. doi: 10.2135/cropsci2017.11.0682

Burstin, J., Salloignon, P., Chabert-Martinello, M., Magnin-Robert, J., Siol, M., Jacquin, F., et al. (2015). Genetic diversity and trait genomic prediction in a pea diversity panel. *BMC Genomics* 16:105. doi: 10.1186/s12864-015-1266-1

Bustos-Korts, D., Malosetti, M., Chapman, S., Biddulph, B., and van Eeuwijk, F. (2016). Improvement of predictive ability by uniform coverage of the target genetic space. *G3* 6, 3733–3747. doi: 10.1534/g3.116.035410

Cericola, F., Jahoor, A., Orabi, J., Andersen, J. R., Janss, L. L., and Jensen, J. (2017). Optimizing training population size and genotyping strategy for genomic prediction using association study results and pedigree information. a case of study in advanced wheat breeding lines. *PLoS ONE* 12:e0169606. doi: 10.1371/journal.pone.0169606

Clark, S. A., Hickey, J. M., Daetwyler, H. D., and van der Werf, J. H. (2012). The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genet. Sel Evol*. 44, 10–1186. doi: 10.1186/1297-9686-44-4

Clark, S. A., Hickey, J. M., and Van der Werf, J. H. (2011). Different models of genetic variation and their effect on genomic evaluation. *Genet. Sel Evol*. 43, 10–1186. doi: 10.1186/1297-9686-43-18

Crossa, J., de los Campos, G., Pérez, P., Gianola, D., Burgue no, J., Araus, J., et al. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics*. 186, 713–724. doi: 10.1534/genetics.110.118521

Crossa, J., Fritsche-Neto, R., Montesinos-Lopez, O. A., Costa-Neto, G., Dreisigacker, S., Montesinos-Lopez, A., et al. (2021). The modern plant breeding triangle: optimizing the use of genomics, phenomics, and enviromics data. *Front. Plant Sci*. 12:651480 doi: 10.3389/fpls.2021.651480

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci*. 22, 961–975. doi: 10.1016/j.tplants.2017.08.011

Daetwyler, H. D., Villanueva, B., and Woolliams, J. A. (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS ONE* 3:e3395. doi: 10.1371/journal.pone.0003395

de Bem Oliveira, I., Amadeu, R. R., Ferr ao, L. F. V., and Mu noz, P. R. (2020). Optimizing whole-genomic prediction for autotetraploid blueberry breeding. *Heredity* 125, 437–448. doi: 10.1038/s41437-020-00357-x

de los Campos, G., Hickey, J., Pong-Wong, R., Daetwyler, H., and Calus, M. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193, 327–345. doi: 10.1534/genetics.112.143313

Dekkers, J. (2007). Prediction of response to marker-assisted and genomic selection using selection index theory. *J. Anim. Breed. Genet*. 124, 331–341. doi: 10.1111/j.1439-0388.2007.00701.x

Dimitrijevic, A., and Horn, R. (2018). Sunflower hybrid breeding: from markers to genomic selection. *Front. Plant Sci*. 8:2238. doi: 10.3389/fpls.2017.02238

Dussert, C., Rasigni, G., Rasigni, M., Palmari, J., and Llebaria, A. (1986). Minimal spanning tree: a new approach for studying order and disorder. *Phys. Rev. B* 34:3528. doi: 10.1103/PhysRevB.34.3528

Falconer, D., and Mackay, T. (1996). *Introduction to Quantitative Genetics, Vol. 4*. Essex: Benjamin Cummings.

Fedorov, V. V. (1972). *Theory of Optimal Experiments*. Academic Press.

Fedorov, V. V., and Hackl, P. (2012). *Model-Oriented Design of Experiments, Vol. 125*. Springer Science Business Media.

Fisher, R. A. (1960). *The Design of Experiments*. New York, NY: Hafner.

Fristche-Neto, R., Akdemir, D., and Jannink, J.-L. (2018). Accuracy of genomic selection to predict maize single-crosses obtained through different mating designs. *Theor. Appl. Genet*. 131, 1153–1162. doi: 10.1007/s00122-018-3068-8

Gentle, J. E. (2006). *Random Number Generation and Monte Carlo Methods*. Springer Science Business Media.

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetics* 136, 245–257. doi: 10.1007/s10709-008-9308-0

Guo, T., Yu, X., Li, X., Zhang, H., Zhu, C., Flint-Garcia, S., et al. (2019). Optimal designs for genomic selection in hybrid crops. *Mol. Plant* 12, 390–401. doi: 10.1016/j.molp.2018.12.022

Guo, Z., Tucker, D., Basten, C., Gandhi, H., Ersoz, E., Guo, B., et al. (2014). The impact of population structure on genomic prediction in stratified populations. *Theor. Appl. Genet*. 127, 749–762. doi: 10.1007/s00122-013-2255-x

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

Habier, D., Tetens, J., Seefried, F.-R., Lichtner, P., and Thaller, G. (2010). The impact of genetic relationship information on genomic breeding values in german holstein cattle. *Genet. Sel. Evol*. 42:5. doi: 10.1186/1297-9686-42-5

Hayes, B., Bowman, P., Chamberlain, A., and Goddard, M. (2009). Invited review: Genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci*. 92, 433–443. doi: 10.3168/jds.2008-1646

He, S., Schulthess, A. W., Mirdita, V., Zhao, Y., Korzun, V., Bothe, R., et al. (2016). Genomic selection in a commercial winter wheat population. *Theor. Appl. Genet*. 129, 641–651. doi: 10.1007/s00122-015-2655-1

Heffner, E., Sorrells, M., and Jannink, J. (2009). Genomic selection for crop improvement. *Crop. Sci*. 49, 1–12. doi: 10.2135/cropsci2008.08.0512

Heslot, N., and Feoktistov, V. (2020). Optimization of selective phenotyping and population design for genomic prediction. *J. Agric. Biol. Environ. Stat*. 25, 579–600. doi: 10.1007/s13253-020-00415-1

Isidro y Sánchez, J., Akdemir, D., and Rio, S. (2022). *Hands on Training Optimization in Genomic Selection*. Springer.

Isidro, J., Akdemir, D., and Burke, J. (2016). "Genomic selection," in *The World Wheat Book: A History of Wheat Breeding, Vol. 3, Chapter 32*, eds A. William, B. Alain, and V. G. Maarten (Paris: Lavoisier), 1001–1023.

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

Jannink, J.-L. (2010). Dynamics of long-term genomic selection. *Genet. Sel. Evol*. 42, 35. doi: 10.1186/1297-9686-42-35

Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., et al. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theor. Appl. Genet*. 127, 595–607. doi: 10.1007/s00122-013-2243-1

Johnson, M. E., Moore, L. M., and Ylvisaker, D. (1990). Minimax and maximin distance designs. *J. Stat. Plan Inference* 26, 131–148. doi: 10.1016/0378-3758(90)90122-B

Kadam, D. C., Potts, S. M., Bohn, M. O., Lipka, A. E., and Lorenz, A. J. (2016). Genomic prediction of single crosses in the early stages of a maize hybrid breeding pipeline. *G3* 6, 3443–3453. doi: 10.1534/g3.116.031286

Kadam, D. C., Rodriguez, O. R., and Lorenz, A. J. (2021). Optimization of training sets for genomic prediction of early-stage single crosses in maize. *Theor. Appl. Genet*. 134, 687–699. doi: 10.1007/s00122-020-03722-w

Kiefer, J. (1959). Optimum experimental designs. *J. R. Stat. Soc. B* 21, 272–319.

Kiefer, J. C., Brown, L., Olkin, I., and Sacks, J. (1985). *Jack Carl Kiefer Collected Papers: Design of Experiments*. Springer.

Laloë, D. (1993). Precision and information in linear models of genetic evaluation. *Genet. Sel. Evol*. 25, 557–576. doi: 10.1186/1297-9686-25-6-557

Lee, S. H., Van Der Werf, J. H., Hayes, B. J., Goddard, M. E., and Visscher, P. M. (2008). Predicting unobserved phenotypes for complex traits from whole-genome snp data. *PLoS Genet*. 4:e1000231. doi: 10.1371/journal.pgen.1000231

Liu, X., Wang, H., Wang, H., Guo, Z., Xu, X., Liu, J., et al. (2018). Factors affecting genomic selection revealed by empirical evidence in maize. *Crop J*. 6, 341–352. doi: 10.1016/j.cj.2018.03.005

Longin, C. F. H., Mi, X., and Würschum, T. (2015). Genomic selection in wheat: optimum allocation of test resources and comparison of breeding strategies for line and hybrid breeding. *Theor. Appl. Genet*. 128, 1297–1306. doi: 10.1007/s00122-015-2505-1

Lopez-Cruz, M., and de Los Campos, G. (2021). Optimal breeding-value prediction using a sparse selection index. *Genetics* 210:iyab030. doi: 10.1093/genetics/iyab030

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

Lorenz, A. J., and Smith, K. P. (2015). Adding genetically distant individuals to training populations reduces genomic prediction accuracy in barley. *Crop Sci*. 55, 2657–2667. doi: 10.2135/cropsci2014.12.0827

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

Luan, T., Woolliams, J. A., Lien, S., Kent, M., Svendsen, M., and Meuwissen, T. H. (2009). The accuracy of genomic selection in norwegian red cattle assessed by cross-validation. *Genetics* 183, 1119–1126. doi: 10.1534/genetics.109.107391

MacLeod, I., Bowman, P., Vander Jagt, C., Haile-Mariam, M., Kemper, K., Chamberlain, A., et al. (2016). Exploiting biological priors and sequence variants enhances qtl discovery and genomic prediction of complex traits. *BMC Genomics*. 17:144. doi: 10.1186/s12864-016-2443-6

Mangin, B., Bonnafous, F., Blanchet, N., Boniface, M.-C., Bret-Mestries, E., Carrère, S., et al. (2017). Genomic prediction of sunflower hybrids oil content. *Front. Plant Sci*. 8:1633. doi: 10.3389/fpls.2017.01633

Mangin, B., Rincent, R., Rabier, C.-E., Moreau, L., and Goudemand-Dugue, E. (2019). Training set optimization of genomic prediction by means of ethacc. *PLoS ONE* 14:e0205629. doi: 10.1371/journal.pone.0205629

Markowitz, H. M. (1968). *Portfolio Selection: Efficient Diversification of Investments, Vol. 16*. Yale university press.

Marulanda, J. J., Mi, X., Melchinger, A. E., Xu, J.-L., Würschum, T., and Longin, C. F. H. (2016). Optimum breeding strategies using genomic selection for hybrid breeding in wheat, maize, rye, barley, rice and triticale. *Theor. Appl. Genet*. 129, 1901–1913. doi: 10.1007/s00122-016-2748-5

McClellan, J., Susser, E., and King, M. (2007). Schizophrenia: a common disease caused by multiple rare alleles. *Br. J. Psychiatry* 190, 194–199. doi: 10.1192/bjp.bp.106.025585

McGowan, M., Wang, J., Dong, H., Liu, X., Jia, Y., Wang, X., et al. (2020). Ideas in genomic selection with the potential to transform plant molecular breeding: a review. [Epub ahead of print].

Mendonça, L. D. F., and Fritsche-Neto, R. (2020). The accuracy of different strategies for building training sets for genomic predictions in segregating soybean populations. *Crop Sci*. 60, 3115–3126. doi: 10.1002/csc2.20267

Meuwissen, T., Hayes, B., and Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. doi: 10.1093/genetics/157.4.1819

Meuwissen, T., Hayes, B., and Goddard, M. (2016). Genomic selection: a paradigm shift in animal breeding. *Anim. Front*. 6, 6–14. doi: 10.2527/af.2016-0002

Momen, M., and Morota, G. (2018). Quantifying genomic connectedness and prediction accuracy from additive and non-additive gene actions. *Genet. Sel. Evolution* 50, 1–10. doi: 10.1186/s12711-018-0415-9

Montesinos-López, O. A., Montesinos-López, A., Tuberosa, R., Maccaferri, M., Sciara, G., Ammar, K., et al. (2019). Multi-trait, multi-environment genomic prediction of durum wheat with genomic best linear unbiased predictor and deep learning methods. *Front. Plant Sci*. 10:1311. doi: 10.3389/fpls.2019.01311

National Human Genome Research Institute (2020). *DNA Sequencing Costs: Data*. Available online at: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data (accessed May 6, 2021).

Neyhart, J. L., Tiede, T., Lorenz, A. J., and Smith, K. P. (2017). Evaluating methods of updating training data in long-term genomewide selection. *G3* 7, 1499–1510. doi: 10.1534/g3.117.040550

Norman, A., Taylor, J., Edwards, J., and Kuchel, H. (2018). Optimising genomic selection in wheat: Effect of marker density, population size and population structure on prediction accuracy. *G3* 8, 2889–2899. doi: 10.1534/g3.118.200311

Olatoye, M. O., Clark, L. V., Labonte, N. R., Dong, H., Dwiyanti, M. S., Anzoua, K. G., et al. (2020). Training population optimization for genomic selection in miscanthus. *G3* 10, 2465–2476. doi: 10.1534/g3.120.401402

Ou, J.-H., and Liao, C.-T. (2019). Training set determination for genomic selection. *Theor. Appl. Genet*. 132, 2781–2792. doi: 10.1007/s00122-019-03387-0

Pszczola, M., and Calus, M. (2016). Updating the reference population to achieve constant genomic prediction reliability across generations. *Animal* 10, 1018–1024. doi: 10.1017/S1751731115002785

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

Pukelsheim, F. (1993). *Optimal Design of Experiments, Vol. 50*. siam.

Pukelsheim, F., and Rosenberger, J. (1993). Experimental designs for model discrimination. *J. Am. Stat. Assoc*. 88, 642–649. doi: 10.1080/01621459.1993.10476317

Reif, J. C., Zhao, Y., Würschum, T., Gowda, M., and Hahn, V. (2013). Genomic prediction of sunflower hybrid performance. *Plant Breed*. 132, 107–114. doi: 10.1111/pbr.12007

Riedelsheimer, C., Czedik-Eysenberg, A., Grieder, C., Lisec, J., Technow, F., Sulpice, R., et al. (2012). Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nat. Genet*. 44, 217–220. doi: 10.1038/ng.1033

Rincent, R., Charcosset, A., and Moreau, L. (2017). Predicting genomic selection efficiency to optimize calibration set and to assess prediction accuracy in highly structured populations. *Theor. Appl. Genet*. 130, 2231–2247. doi: 10.1007/s00122-017-2956-7

Rincent, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., et al. (2012). Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (zea mays l.). *Genetics* 192, 715–728. doi: 10.1534/genetics.112.141473

Roth, M., Muranty, H., Di Guardo, M., Guerra, W., Patocchi, A., and Costa, F. (2020). Genomic prediction of fruit texture and training population optimization towards the application of genomic selection in apple. *Horticulture Res*. 7, 1–14. doi: 10.1038/s41438-020-00370-5

Sarinelli, J. M., Murphy, J. P., Tyagi, P., Holland, J. B., Johnson, J. W., Mergoum, M., et al. (2019). Training population selection and use of fixed effects to optimize genomic predictions in a historical usa winter wheat panel. *Theor. Appl. Genet*. 132, 1247–1261. doi: 10.1007/s00122-019-03276-6

Schrag, T., Frish, M., Dhillon, B., and Melchinger, A. (2009). Marker-based prediction of hybrid performance in maize single-crosses involving doubled haploids. *Maydica* 54, 353. doi: 10.1007/s00122-008-0934-9

Schulthess, A. W., Zhao, Y., and Reif, J. C. (2017). "Genomic selection in hybrid breeding," in *Genomic Selection for Crop Improvement* (Springer), 149–183.

Scott, M. F., Fradgley, N., Bentley, A. R., Brabbs, T., Corke, F., Gardner, K. A., et al. (2021). Limited haplotype diversity underlies polygenic trait architecture across 70 years of wheat breeding. *Genome Biol*. 22, 1–30. doi: 10.1186/s13059-021-02354-7

Seye, A., Bauland, C., Charcosset, A., and Moreau, L. (2020). Revisiting hybrid breeding designs using genomic predictions: simulations highlight the superiority of incomplete factorials between segregating families over topcross designs. *Theor. Appl. Genet*. 133, 1995–2010. doi: 10.1007/s00122-020-03573-5

Silvey, S. (2013). *Optimal Design: An Introduction to the Theory for Parameter Estimation, Vol. 1*. Springer Science Business Media.

Smith, K. (1918). On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika* 12, 1–85.

Spindel, J., Begum, H., Akdemir, D., Collard, B., Redo na, E., Jannink, J., et al. (2016). Genome-wide prediction models that incorporate de novo gwas are a powerful new tool for tropical rice improvement. *Heredity* 116, 395–408. doi: 10.1038/hdy.2015.113

Tanaka, R., and Iwata, H. (2018). Bayesian optimization for genomic selection: a method for discovering the best genotype among a large number of candidates. *Theor. Appl. Genet*. 131, 93–105. doi: 10.1007/s00122-017-2988-z

Tayeh, N., Klein, A., Le Paslier, M.-C., Jacquin, F., Houtin, H., Rond, C., et al. (2015). Genomic prediction in pea: effect of marker density and training population size and composition on prediction accuracy. *Front. Plant Sci*. 6:941. doi: 10.3389/fpls.2015.00941

Technow, F., Podlich, D., and Cooper, M. (2021). Back to the future: Implications of genetic complexity for hybrid breeding strategies. *G3* 5:jkab153. doi: 10.1093/g3journal/jkab153

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

Tsai, S.-F., Shen, C.-C., and Liao, C.-T. (2021). Bayesian optimization approaches for identifying the best genotype from a candidate population. *J. Agric. Biol. Environ. Stat*. 1–19. doi: 10.1007/s13253-021-00454-2

Varshney, R. K., Bohra, A., Yu, J., Graner, A., Zhang, Q., and Sorrells, M. E. (2021). Designing future crops: Genomics-assisted breeding comes of age. *Trends Plant Sci*. 26, 631–649. doi: 10.1016/j.tplants.2021.03.010

Varshney, R. K., Graner, A., and Sorrells, M. E. (2005). Genomics-assisted breeding for crop improvement. *Trends Plant Sci*. 10, 621–630. doi: 10.1016/j.tplants.2005.10.004

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

Wientjes, Y. C., Bijma, P., Vandenplas, J., and Calus, M. P. (2017). Multi-population genomic relationships for estimating current genetic variances within and genetic correlations between populations. *Genetics* 207, 503–515. doi: 10.1534/genetics.117.300152

Wientjes, Y. C., Calus, M. P., Goddard, M. E., and Hayes, B. J. (2015). Impact of qtl properties on the accuracy of multi-breed genomic prediction. *Genet. Sel. Evol*. 47, 1–16. doi: 10.1186/s12711-015-0124-6

Wientjes, Y. C., Veerkamp, R. F., and Calus, M. P. (2013). The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics* 193, 621–631. doi: 10.1534/genetics.112.146290

Windhausen, V. S., Atlin, G. N., Hickey, J. M., Crossa, J., Jannink, J.-L., Sorrells, M. E., et al. (2012). Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. *G3* 2, 1427–1436. doi: 10.1534/g3.112.003699

Xu, Y., Liu, X., Fu, J., Wang, H., Wang, J., Huang, C., et al. (2020). Enhancing genetic gain through genomic selection: from livestock to plants. *Plant Commun*. 1:100005. doi: 10.1016/j.xplc.2019.100005

Yu, X., Leiboff, S., Li, X., Guo, T., Ronning, N., Zhang, X., et al. (2020). Genomic prediction of maize microphenotypes provides insights for optimizing selection and mining diversity. *Plant Biotechnol. J*. 18, 2456–2465. doi: 10.1111/pbi.13420

Zhang, H., Yin, L., Wang, M., Yuan, X., and Liu, X. (2019). Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. *Front. Genet*. 10:189. doi: 10.3389/fgene.2019.00189

Zhang, Y., Massel, K., Godwin, I. D., and Gao, C. (2018). Applications and potential of genome editing in crop improvement. *Genome Biol*. 19, 1–11. doi: 10.1186/s13059-018-1586-y

Zhao, Y., Mette, M., Gowda, M., Longin, C., and Reif, J. (2014). Bridging the gap between marker-assisted and genomic selection of heading time and plant height in hybrid wheat. *Heredity* 112, 638–645. doi: 10.1038/hdy.2014.1

Zhao, Y., Mette, M. F., and Reif, J. C. (2015). Genomic selection in hybrid breeding. *Plant Breed*. 134, 1–10. doi: 10.1111/pbr.12231

Zhao, Y., Thorwarth, P., Jiang, Y., Philipp, N., Schulthess, A. W., Gils, M., et al. (2021). Unlocking big data doubled the accuracy in predicting the grain yield in hybrid wheat. *Sci. Adv*. 7:eabf9106. doi: 10.1126/sciadv.abf9106

Zhao, Y., Zeng, J., Fernando, R., and Reif, J. C. (2013). Genomic prediction of hybrid wheat performance. *Crop Sci*. 53, 802–810. doi: 10.2135/cropsci2012.08.0463

Zhong, S., Dekkers, J. C., Fernando, R. L., and Jannink, J.-L. (2009). Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. *Genetics* 182, 355–364. doi: 10.1534/genetics.108.098277

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Plant Science

SYSTEMATIC REVIEW
published: 16 September 2021
doi: 10.3389/fpls.2021.724517

A Systematic Narration of Some Key Concepts and Procedures in Plant Breeding

*Weikai Yan**

Ottawa Research and Development Center, Agriculture and Agri-Food Canada (AAFC), Ottawa, ON, Canada

The goal of a plant breeding program is to develop new cultivars of a crop kind with improved yield and quality for a target region and end-use. Improved yield across locations and years means better adaptation to the climatic, soil, and management conditions in the target region. Improved or maintained quality renders and adds value to the improved yield. Both yield and quality must be considered simultaneously, which constitutes the greatest challenge to successful cultivar development. Cultivar development consists of two stages: the development of a promising breeding population and the selection of the best genotypes out of it. A complete breeder's equation was presented to cover both stages, which consists of three key parameters for a trait of interest: the population mean (μ), the population variability (σ_G), and the achieved heritability (h^2 or H), under the multi-location, multi-year framework. Population development is to maximize $\mu\sigma_G$ and progeny selection is to improve H. Approaches to improve H include identifying and utilizing repeatable genotype by environment interaction (GE) through mega-environment analysis, accommodating unrepeatable GE through adequate testing, and reducing experimental error via replication and spatial analysis. Related concepts and procedures were critically reviewed, including GGE (genotypic main effect plus genotype by environment interaction) biplot analysis, GGE + GGL (genotypic main effect plus genotype by location interaction) biplot analysis, LG (location-grouping) biplot analysis, stability analysis, spatial analysis, adequate testing, and optimum replication. Selection on multiple traits includes independent culling and index selection, for the latter GYT (genotype by yield*trait) biplot analysis was recommended. Genomic selection may provide an alternative and potentially more effective approach in all these aspects. Efforts were made to organize and comment on these concepts and procedures in a systematic manner.

Keywords: heritability, genotype by environment interaction, optimum testing, optimum replication, multi-trait selection, biplot analysis, mega-environment analysis, breeder's equation

OPEN ACCESS

Edited by:
Suchismita Mondal, International Maize and Wheat Improvement Center, Mexico

Reviewed by:
Maryke T. Labuschagne, University of the Free State, South Africa
Vivi Noviati Arief, The University of Queensland, Australia

***Correspondence:**
Weikai Yan
weikai.yan@agr.gc.ca

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *13 June 2021*
Accepted: *09 August 2021*
Published: *16 September 2021*

Citation:
Yan W (2021) A Systematic Narration of Some Key Concepts and Procedures in Plant Breeding. Front. Plant Sci. 12:724517. doi: 10.3389/fpls.2021.724517

INTRODUCTION

Plant breeding plays a key role in meeting the human needs for more food, nutrition, and fiber under a changing climate. The goal of a plant breeding program is to develop new cultivars of a crop kind with improved yield and quality for its target region and end-use. All theories, concepts, processes, procedures, and analyses related to plant breeding are developed and implemented around this goal. A target region is the target population of environments, which is the sum of soil, climatic, biotic, and abiotic conditions plus common management practices that are likely to be encountered in the region. Improved yield means improved adaptation to the target region, which is reflected in improved mean performance and stability of performance across locations and years in the target region. Improved quality means improved adaptation to the end-uses that bring value and income to the growers in the target region. Both the target environments and the target end-uses may change over time, in a predictable or unpredictable manner. Yield is the result from integrating numerous traits including various yield components, agronomic traits, disease resistances, and tolerance to various abiotic stresses characteristic of the target region. Consequently, yield in different regions may mean different ways of packaging these traits and underlying gene alleles. Likewise, quality is a collective term of many parameters for a specific end-use. Thus, dealing with many traits simultaneously is an essential task of cultivar development, although most breeding-related publications deal with only a single trait, typically yield. The relation between yield and other traits is analogous to that between the skin and the hair of a fur or that between the trunk and the branches of a tree; other traits gain importance only when attached to (i.e., combined with) high yield (Yan and Frégeau-Reid, 2008; Yan et al., 2019a). Plant breeding is a mature discipline of applied sciences, with well-developed concepts and procedures. Nevertheless, a systematic combing and narration of the numerous, sometimes confusing, concepts and procedures should help both new and experienced breeders in their work toward developing superior cultivars. The concepts and procedures in plant breeding are indeed much easier to tackle for a single trait. So, much of the discussion will be on a single trait while keep in mind that multi-trait selection is essential to cultivar development, which is discussed in the last section. In addition, genomic selection (Goddard and Hayes, 2007; Heffner et al., 2009; Jannink et al., 2010) has become a growing point or integral part in most plant breeding programs. Its role will be briefly mentioned when the various concepts and procedures are discussed.

THE COMPLETE BREEDER'S EQUATION

The cultivar development process includes two stages: the development of a promising breeding population and the identification of the best progeny out of it. Breeding success can be measured by the following equation, referred to as the Complete Breeder's Equation (modified from Yan et al., 2019b),

$$B = (\mu + ih\sigma_G)/(YC), \quad (1)$$

in comparison with the well-known Breeder's Equation of Eberhart (1970),

$$\Delta G = ih\sigma_G/Y. \quad (2)$$

Here B stands for breeding success per unit time and cost and ΔG stands for selection gain over the population mean per unit time, for a trait of interest (typically yield). μ is the mean of the breeding population, σ_G is the square root of the genotypic variance of the population, i is the selection intensity in the unit of σ_G, h is the square root of achieved heritability (h^2 or H), Y is the length of the breeding cycle in years, and C is the operation cost per year. μ , σ_G *and* h are to be estimated from environments representing the target region. A target region may consist of multiple mega-environments, as will be discussed later. For the time being the target region is assumed to be a single mega-environment. A mega-environment is defined as a group of environments that share the same best cultivar(s) (Gauch and Zobel, 1997; Yan et al., 2000).

Relative to Equation 2, Equation 1 emphasizes the importance of population mean in cultivar development and serves as a reminder that any selection progress is on the basis of the population mean. The inclusion of C emphasizes that cultivar development is an enterprise that must consider the cost for the achieved genetic gain.

Cultivar development consists of two stages: population development and progeny selection. Practical breeders would agree that developing a promising breeding population, i.e., making a promising cross or crosses, is the crucial first step toward cultivar development. A promising breeding population is the basis for any meaningful selection effort. This point may be implied in Eberhart (1970) and by later researchers (e.g., Cobb et al., 2019; Rutkoski, 2019) when discussing the Breeder's Equation but its importance to cultivar development can never be overemphasized, thus implicitly indicated in Equation 1. The potential of a breeding population for cultivar development, shorted as population potential (P), depends on both the population mean (μ) and the population variability (σ_G):

$$P = \sqrt{\mu\sigma_G}. \quad (3)$$

Apparently, if there is no genetic variability, there would be no selection progress; if the population mean is low, it is unlikely to lead to any superior cultivars regardless of selection strategies. Practical plant breeders are well aware of the importance of the population mean. They cross best with best and look for recombinants better than both parents (Duvick, 1996). A high μ is usually achieved by using currently the most popular, usually the highest yielding, cultivars as parent(s), while a high level of σ_G is achieved by choosing parents that are different and complementary in yield components, agronomic traits, disease

resistances, and quality traits, and by use of a large enough breeding population. Crossing an adapted local cultivar with a geographically distant cultivar with desired traits led to some of the most important wheat cultivars in China (Zhao et al., 1981). In the era of genomic selection, μ and σ_G and therefore P can be predicted for any pair or set of potential parents for a trait of interest if reliable genomic models are available (Wang et al., 2018).

Usually, the genetic variability in a breeding population is created by crossing different parents, but it can also be created through induced mutations by treating a superior cultivar with γ radiation, chemical mutagen treatment, transposons, genetic transformation, or gene editing (e.g., van Harten, 1998; Kharkwal et al., 2004; Shu et al., 2012).

To maximize $\mu\sigma_G$ may suggest that μ and σ_G are equally important. In cultivar development, however, μ may be more important than σ_G although both are essential. The use of backcross, recurrent selection, and crosses between closely related breeding lines (e.g., Rasmusson and Phillips, 1997) are examples to ensure a high μ at the expense of σ_G. On the contrary, wide crosses (e.g., Baum et al., 1992) can bring much variability to the population at the expense of reduced population mean. Wide crosses are essential to introduce novel genes and traits from wild species (e.g., Ma et al., 2018; numerous research done worldwide for various crops), which are crucial to long-term crop improvement; however, they are unlikely to directly lead to superior cultivars.

SELECTION GAIN, SELECTION EFFICIENCY, SELECTION INTENSITY, CULLING RATE, AND HERITABILITY

Equation 2 or the second part Equation 1 consists of factors determining the selection gain and is known as the Breeder's Equation. It may be more accurately referred as the breeder's equation for progeny selection. Here σ_G is fixed for a given breeding population, i is a parameter artificially set, and h is the square root of achieved heritability. In fact, while i is the artificially set selection intensity, ih is the realized selection intensity. Cobb et al. (2019) discussed approaches to improving breeding efficiency in the framework of Equation 2, with the emphasis on reducing Y. Rutkoski (2019) reviewed the basis and approaches to achieve genetic gain.

Equation 1 can be better understood from **Figure 1**. Assume that the breeding population is normally distributed with a mean μ and a variability σ_G. The X-axis is the range of the phenotypic values and the Y-axis is the frequency density. The area under the curve is unity (1 or 100%). With a selection intensity ih, genotypes to be culled lie on the left side of the vertical line defined by $x = \mu + ih\sigma_G$, and genotypes to be retained lie on the right side of the line. The area α is the proportion of the population to be retained and $1 - \alpha$ is proportion to be culled. α is also the probability for $ih\sigma_G < 0$, while $1 - \alpha$ is the probability for $ih\sigma_G > 0$. In other words, α is the probability for a genotype with a phenotypic value of $\mu + ih\sigma_G$ to be no better than the population mean.

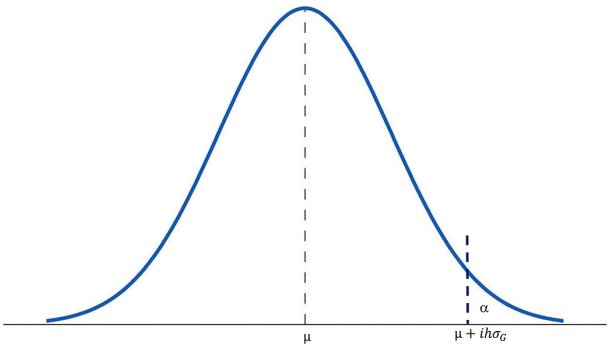

FIGURE 1 | A chart of normal distribution to show the relationships among various parameters in the Complete Breeder's Equation. μ is the mean of the breeding population; σ_G is the square root of the genotypic variance of the population; i is the artificially set selection intensity in the unit of σ_G; h is the square root of achieved heritability (h^2 or H); α is the portion of the population to be selected; it is also the probability that the best genotypes are not included in the selected portion.

An extended interpretation is that α is the risk that the best genotype in the population is *not* retained at the selection intensity ih. Apparently, the risk is reduced as ih is increased while h is the only objective variable. If $h = 0$, then $ih = 0$, and $\alpha = 50\%$. As h *approaches unity*, α approaches 0. This provides a clue for the choice of i. According to the normal distribution table, if α *is set to* 0.0001, then $z = ih = 3.7$. Therefore, it is rational to set $i = 3.7$ at $\alpha = 0.0001$. The relationships between heritability, ih, and α at $i = 3.7$ for some selected heritability values are listed in **Table 1**, assuming a population size of $n = 10,000$.

If α is interpreted as the percentage of the population that must be retained to ensure that the best genotype(s) is included, then the number of genotypes must be selected, N, will be:

$$N = n\alpha, \tag{4}$$

where n is the effective population size, i.e., the number of unique genotypes in the breeding population. The inverse of N may be defined as the rate of selection success (Yan et al., 2019b):

$$S = 1/N. \tag{5}$$

For example, for $n = 10,000$, $H = 0.9$, and $i = 3.7$, we have $\alpha = 0.02\%$ and $N = 2$ (**Table 1**). That is, for a population of 10,000 unique genotypes, an achieved heritability of 0.9 would guarantee that the best genotype is between the top two. A smaller N means less time (in years) and cost that are needed to single out the best genotype. In the extreme case, if a selection method (genomic prediction or any other approach) is accurate enough to identify the best genotype (i.e., $N = 1$) out of a breeding population, then all the time and cost associated with subsequent testing would be saved. In contrast, in the Ottawa oat breeding program, it takes about seven years of visual selection

TABLE 1 | The realized selection intensity ($z = ih$), the proportion of the population to be retained (α), and the number of genotypes to be retained (N) at different levels of heritability (H or h^2) assuming a population of $n = 10,000$ and a selection intensity of i=3.7.

$H = h^2$	h	$z = 3.7h$	(1-α)	α (%)	N (n = 10,000)	Corresponding breeding stages in Yan et al. (2019b)
0.0	0.00	0.00	0.5000	50.00	5000	
0.1	0.32	1.17	0.8790	12.10	1210	Stage 3.1 (yr1)
0.2	0.45	1.65	0.9505	4.95	495	
0.3	0.55	2.03	0.9788	2.12	212	Stage 3.2 (yr2)
0.4	0.63	2.34	0.9904	0.96	96	
0.5	0.71	2.62	0.9959	0.41	41	Stage 4.1 (yr3)
0.6	0.77	2.87	0.9980	0.21	21	Stage 4.2 (yr4)
0.7	0.84	3.10	0.9990	0.10	10	State 4.3 (yr5)
0.8	0.89	3.31	0.9995	0.05	5	Stage 4.4 (yr6)
0.9	0.95	3.51	0.9998	0.02	2	Stage 4.5 (yr7)
1.0	1.00	3.70	0.9999	0.01	1	Cultivar release

and yield trials to identify the best genotypes out of a breeding population (Yan et al., 2019b; last column of **Table 1**, this paper). Each year ∼10,000 F_2 derived breeding lines are planted in a hill nursery and 1,000 are visually selected in the field and the seed lab. Assuming that the best genotype is included in these selected lines, the rate of selection success for this stage (the "Hill Nursery" stage or Stage 3.1) is ∼1/1000, roughly corresponding to assuming an $H = 0.1$ (**Table 1**). The 1,000 selected lines are then planted in yield plots and ∼200 lines are visually selected (the "Observation Plot" stage or Stage 3.2). The accumulative rate of selection success for these two years of visual selection is, therefore, ∼1/200, corresponding to $H = 0.3$ (**Table 1**). It takes four to five additional years of multi-location test to single out the best few genotypes as potential new cultivars (Stages 4.1 to 4.5 in **Table 1**). Experience indicates that the top genotypes at the Stage 4.3 are usually the ones to be released as cultivars; this corresponds to $H = 0.7$ (**Table 1**). Trials in Stages 4.4 and 4.5 (years 2 and 3 of the Registration Test) are conducted mainly to verify the results and to obtain data required for official variety registration.

Genomic selection applied at the Hill Nursery stage (Stage 3.1) is expected to dramatically improve the rate of selection success so as to reduce the number of years spent in visual selection and yield trials (Y, Equations 1 and 2). A minimum requirement for a viable genomic selection procedure is to improve the selection efficiency to an extent that covers the extra cost spent in genotyping, phenotyping, and model development. Alternatively, genomic selection is justified if it can identify the best genotypes that may be discarded by the breeder's eye.

The parameter i should be set according to the population size such that $\alpha = 1/n$. This reflects the idea that a larger population allows a higher selection intensity at the same level of heritability and that the top genotype is the best genotype ($N = 1$) when $h = 1$. According to the normal distribution table, i should be set at 2.05, 2.33, 3.00, and 3.71 when $n = 50$, 100, 1,000, and 10,000, respectively. The relationship among heritability, selection intensity, and probability of false culling (α) is displayed in **Figure 2**. Incidentally, Singh and Chaudhary (1977) suggested setting $i = 2.063$ at $\alpha = 0.05$, in line with this idea.

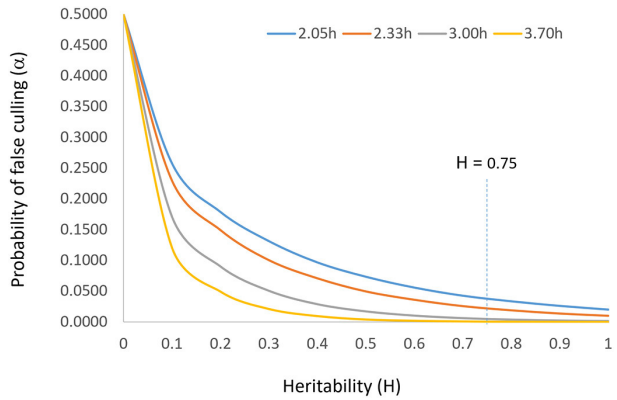

FIGURE 2 | The relationship between heritability (H or h^2) and probability of false culling (α) at four levels of selection intensity (i). It is suggested that the probability of false culling that can be tolerated be set according to the population size (n) such that $\alpha = 1/n$; i can then be determined by α by consulting the normal distribution table. The vertical line of H = 0.75 indicates the target heritability to be achieved for reliable selection.

Alternatively, the achieved heritability may be used as the culling rate when the population is small; the number of genotypes that must be retained can then be roughly estimated by:

$$N = n\left(1 - h^2\right). \tag{6}$$

When $h^2 = 0$, no genotype would be discarded because the selection is completely unreliable; and when $h^2 = 1$, all but the top performing genotype can be discarded because any observed difference is genetic and heritable. For example, if $n = 40$ and $h^2 = 0.95$, then 95% or 38 of the 40 entries can be discarded and the top two performing genotypes can be selected or recommended.

To summarize, for a given breeding population and a given target environment, the allowable culling rate, the allowable selection intensity, the achievable rate of selection success, and the expected selection gain are all determined solely by the achieved heritability, in a curvilinear fashion (**Figure 2**). Therefore, heritability is the single most important concept in progeny selection.

HERITABILITY UNDER THE MULTI-LOCATION, MULTI-YEAR FRAMEWORK

Cultivars are developed to adapt to a specific target region, i.e., to the environments that may be encountered across locations and years in a target region. Therefore, the heritability discussed so far must be defined in the multi-location, multi-year framework (Comstock and Moll, 1963; DeLacy et al., 1996; Atlin et al., 2000). According to the general linear model, a phenotype, i.e., an observed value, is a mixed effect of environmental main effect (E), genotypic main effect (G), genotype by environment interaction (GE), and experimental error (ε), where E is the sum of location main effect (L), year main effect (Y), and their interaction (LY). Assuming orthogonal experimental design i.e., the same set of genotypes are tested at the same set of locations each year with the same number of replicates, the phenotypic variance is $\sigma_P^2 = \sigma_G^2 + \frac{\sigma_{GL}^2}{l} + \frac{\sigma_{GY}^2}{y} + \frac{\sigma_{GLY}^2}{ly} + \frac{\sigma_\varepsilon^2}{lyr}$. The entry-mean heritability, i.e., the proportion of phenotypic variance that can be explained by the genetic variance at the entry mean level, is estimated by (Fehr, 1991; DeLacy et al., 1996):

$$H_{rly} = \frac{\sigma_G^2}{\sigma_P^2} = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_{GL}^2}{l} + \frac{\sigma_{GY}^2}{y} + \frac{\sigma_{GLY}^2}{ly} + \frac{\sigma_\varepsilon^2}{rly}}, \quad (7)$$

where H_{rly} stands for heritability across l locations in y years with r replicates; σ_{GL}^2, σ_{GY}^2, and σ_{GLY}^2 are the variances for the genotype by location interaction (GL), genotype by year interaction (GY), and genotype by location by year interaction (GLY), respectively; and σ_ε^2 is the variance for experimental error.

When trials are not conducted orthogonally regarding genotypes, location, years, or replicates, which is usually the case, each trial (location-year combination) may be considered as an environment, and the heritability can be estimated by

$$H_{rly} = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_{GE}^2}{\sum_{i=1}^{y} l_i} + \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{y}\sum_{j=1}^{l_i} r_{ij}}}, \quad (8)$$

where σ_{GE}^2 is the variance for genotype by environment interaction. For convenience, Equation 7 will be used in further discussions. Restricted maximum likelihood (REML) is the preferred method for estimating the various variances, particularly when the data are unbalanced (e.g., Gilmour et al., 1995). REML is implemented in all software packages with a mixed model procedure.

Heritability for a single trial can be estimated by

$$H_r = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_\varepsilon^2}{r}}, \quad (9)$$

However, H_r can be used to assess the data quality of a trial but not for making final selections. For making section decisions, Equation 7a below should be used instead:

$$H_{rly} = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_{GL}^2}{1} + \frac{\sigma_{GY}^2}{1} + \frac{\sigma_{GLY}^2}{1} + \frac{\sigma_\varepsilon^2}{r}}. \quad [7a]$$

That is, although the interaction terms cannot be estimated from a single trial, they must be factored in when making selection decisions. It can be seen that H_r is an inflated estimation of H_{rly} for a trial because the denominator in Equation 7a should be much larger than that in Equation 9.

Likewise, a heritability can be estimated for multi-location trials conducted in a year,

$$H_{rl} = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_{GL}^2}{l} + \frac{\sigma_\varepsilon^2}{rl}}, \quad (10)$$

but it is not to be used to make final selection decisions. Instead, equation 7b should be used,

$$H_{rly} = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_{GL}^2}{l} + \frac{\sigma_{GY}^2}{1} + \frac{\sigma_{GLY}^2}{l} + \frac{\sigma_\varepsilon^2}{rl}}. \quad [7b]$$

H_{rl} is an inflated estimation of H_{rly} for a single-year test. The definition of heritability in the form of Equation 7 is the only valid definition to be used in Equation 1; **Figure 1**, and **Table 1**, with $h = \sqrt{H_{rly}}$, even though H_{rly} cannot be directly estimated in some stages of the breeding cycle. It should be noted that the definition of heritability is in line with the concept of mixed models. It consists of variances for G, GE (= GL + GY + GLY), and experimental error but excludes that for E (= L + Y + LY), implying a mixed model. It implies that G, GE, and experimental error are considered as random effects but E as fixed effects (DeLacy et al., 1996). Researchers are often puzzled on which effects should be treated as random and which fixed when analyzing multi-environment trials data using mixed models (Piepho et al., 2003); for the purpose of genotype evaluation, this is clear from the definition of heritability. The definition of heritability is also consistent with the concept of GGE biplot analysis, which excludes E and focuses on G and GE for cultivar and test environment evaluation (Yan et al., 2000; Yan and Kang, 2002; Yan and Tinker, 2006; Yan, 2014).

All efforts made to improve selection efficiency are also efforts to improve the heritability as defined in Equation 7 or Equation 8. Put it differently, all possible approaches to

improve selection efficiency reside in the definition of heritability. These include approaches to deal with GE and approaches to minimize experimental error. Dealing with GE include two steps: (1) identifying and utilizing repeatable GE, a process often referred to as mega-environment analysis (Yan, 2014, 2015, 2016, 2019), and (2) accommodating unrepeatable GE through adequate testing (Yan et al., 2015; Yan, 2016, 2021). Dealing with experimental error includes adequate replication (Yan et al., 2015; Yan, 2021) and spatial variation adjustment (Cullis and Gleeson, 1991; Gilmour et al., 1997; Cullis et al., 1998; Burgueño et al., 2000; Qiao et al., 2000; Yang et al., 2004; Yan, 2014).

MEGA-ENVIRONMENT ANALYSIS AND UTILIZATION OF REPEATABLE GE

Repeatable GE vs. Unrepeatable GE

Mega-environment analysis is analysis of the G+GE patterns aiming at dividing a target region into meaningful subregions or mega-environments (subregions and mega-environments are used interchangeably in this article). Among the components of GE, GY and GLY are obviously unrepeatable because it is impossible to predict the environments of next year. It is possible, though, that some of the GL is repeatable as the soil and daylength at a location are fixed. Some management factors such as irrigation, fertilizer application, and fungicide application may also lead to repeatable GE (Cooper et al., 2021), which are lumped as "common management practices in the target region or mage-environment" for simplicity. Assuming that the test locations can be divided into two or more groups (subregions), the variance for GL will be divided into variance for genotype by subregion interaction (σ^2_{GS}), which is the repeatable part, and genotype by location interaction within subregions ($\sigma^2_{GL(s)}$), which is the unrepeatable part, of GL (Atlin et al., 2000; Yan, 2016):

$$\sigma^2_{GL} = \sigma^2_{GS} + \sigma^2_{GL(s)} \tag{11}$$

and the number of test locations l will also be divided among the subregions:

$$l = \sum_{k=1}^{s} l_k \tag{12}$$

where l_k is the number of test locations within subregion k. Subdivision of the target region into subregions will improve the overall heritability if the genotype by subregion interaction is sufficiently large, because the genotype by subregion interaction is converted into genotypic main effect within subregions when genotype evaluation is conducted by subregion:

$$H'_{rly} = \frac{\sigma^2_G + \sigma^2_{GS}}{(\sigma^2_G + \sigma^2_{GS}) + \frac{\sigma^2_{GL(S)}}{l} + \frac{\sigma^2_{GY}}{y} + \frac{\sigma^2_{GLY}}{ly} + \frac{\sigma^2_\varepsilon}{rly}} \tag{13}$$

where H'_{rly} is the entry-mean heritability when genotype evaluation is conducted by subregion. On the other hand, dividing a region into subregions may lead to reduced heritability within a subregion due to the smaller number of test locations (Equation 12). Thus, Atlin et al. (2000, 2011) warned that subdivision of a target region should be avoided if genotype by subregion interaction is small relative to G. They used the genetic correlation between divided subregions and the undivided whole region (r_G) as a measure to decide whether the target region should be divided, which is defined as:

$$r_G = \sqrt{\frac{\sigma^2_G}{\sigma^2_G + \sigma^2_{GS}}} \tag{14}$$

They suggested that subdivision should be avoided if r_G is high, although an explicit criterion was not given. In fact, the correlation between candidate subregions should be a more meaningful measure.

Nevertheless, if a subregion is found to be distinct from other subregions, it should be treated as such; if a subregion is economically important, it is justifiable to increase the number of test locations within it to achieve a sufficiently high heritability or selection reliability. The merit of dividing a target region into meaningful subregions is to allow selection and deployment of subregion-specific cultivars to achieve a higher genetic gain within each subregion and thereby the whole region. Annicchiarico (2021) presented a recent example that selection for mega-environment specific cultivars increased genetic gains, in addition to a good review on the subject matter. An essential condition for dividing a target region into subregions is the presence of substantial crossover genotype by subregion interactions (discussed below).

How to Reveal Repeatable GE

To investigate whether heritability can be improved by dividing a target region into subregions, the prerequisite is a good hypothesis on how to divide the target region. Various approaches have been used in dividing a jurisdictional region into agroclimatic regions as reviewed in Yan et al. (2011). A poor hypothesis will lead to the false conclusion that the target region cannot be divided and thereby miss the opportunity to utilize the repeatable GE. For example, Atlin et al. (2000) hypothesized that western Canada (including Alberta, Saskatchewan, and Manitoba) and eastern Canada (including Ontario, Quebec, and Maritime provinces) were two barley mega-environments and rejected the hypothesis. Based on the same dataset, however, Yan and Tinker (2005) showed two clear mega-environments, with locations in Alberta and Saskatchewan as one mega-environment and locations in Manitoba and the eastern Canadian provinces as the other. For another example, in analyzing the data of a set of maize hybrids tested at 24 sites in six African countries in 2009, Atlin et al. (2011) hypothesized that each country was a mega-environment and concluded that there was no mega-environment differentiation. However, a country is a political entity rather than an ecoclimatic region, so the hypothesis *per se* is questionable. A good hypothesis on mega-environment differentiation must be based on the G+GE patterns. Two methods have been developed to reveal repeatable GE patterns:

GGE + GGL biplot analysis (Yan, 2014, 2015, 2016) and LG (location-grouping) biplot analysis (Yan, 2019; Yan et al., 2019b, 2021).

GGE + GGL Biplot Analysis

As the definition of heritability in Equation 7 or Equation 8 suggested, data from multi-location, multi-year trials are required to conduct GGE + GGL biplot analysis, where GGE stands for G + GE (meaning fitting G + GE by principal components), and GGL for G + GL. In variety trials, usually the same set of genotypes are tested at all locations in a year but different sets of genotypes are tested in different years, because poor genotypes are dropped and new genotypes added each year. Consequently, multi-location, multi-year data are typically unbalanced. Nevertheless, usually a sizable number of common genotypes are tested in two or more consecutive years; this allows missing values in the genotype by environment two-way table to be imputed based on certain procedures (e.g., Yan, 2013). Data from such trials can be investigated using a GGE + GGL biplot (Yan, 2014, 2015, 2016), as shown in the example below.

The yield data from the 2013 to 2019 Quebec Oat Registration and Recommendation trials are used here as an example (data available from the author upon request). Each year 41 to 46 registered oat cultivars or breeding lines were tested at eight to 10 locations. The locations represent three ecoclimatic zones of Quebec (Yan et al., 2011; Yan, 2015). Zone 1 was represented by NDHY1 (St Hyacinthus) and STHU1 (St. Huber), Zone 2 by PRIN2 (Princeville), PINT2 (Pintendre), STAU2 (St. Augusta), and STET2 (St. Etienne), and Zone 3 by NORM3 (Normandin), HEBE3 (Hebertville), CAUS3 (Causapscal), and LAPO3 (La Pocatière); the number at the end of each location code indicates the zone it belongs. In addition, the trials were also conducted at OTT (Ottawa in Ontario), which is geographically close to Zone 1 of Quebec. A total of 118 genotypes and 67 trials (location by year combinations) were involved in these seven years, forming a 118-genotype by 67-trial two-way table, with 63% missing values. The first step of the analysis was to generate a GGE biplot containing the 118 genotypes and the 67 trials (**Figure 3**). The GGE biplot was constructed by the first two principal components from subjecting the trial-standardized genotype by

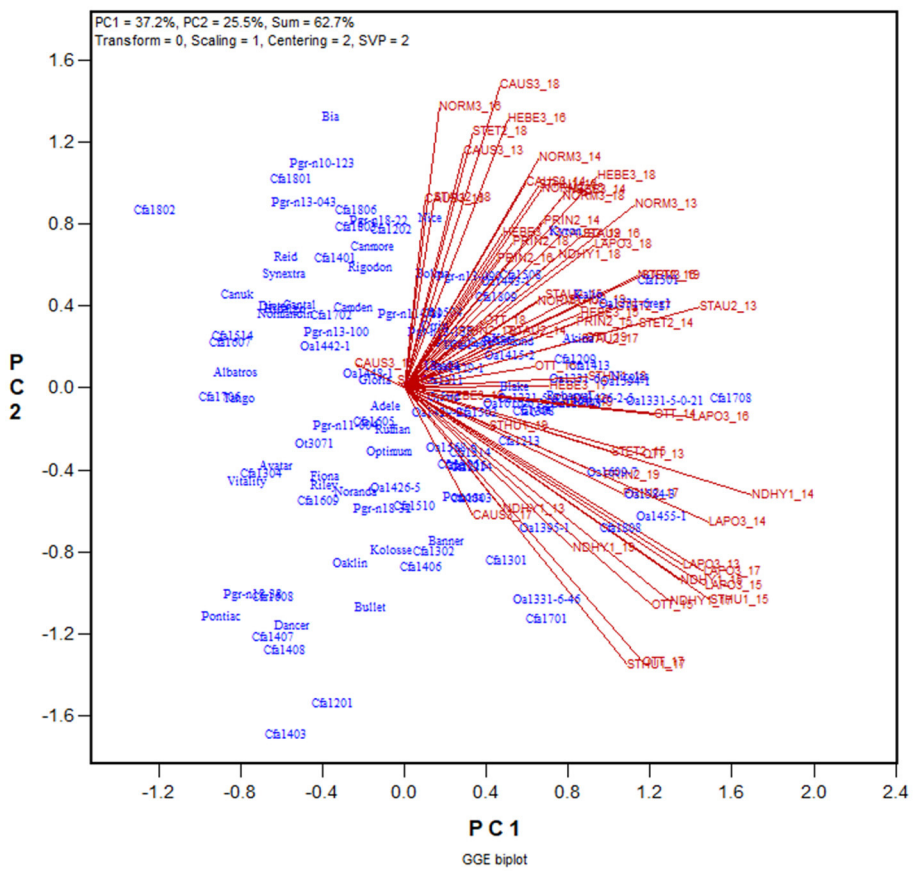

FIGURE 3 | GGE biplot to show the relative yield of 116 oat genotypes in 67 trials from the 2013–2019 Quebec provincial oat trials. The genotypes are displayed in blue and the trials in red. Each trial is displayed as a location-year combination. The Quebec locations are: NDHY1 (St Hyacinthus) and STHU1 (St. Huber) in Zone 1, PINT2 (Pintendre), PRIN2 (Princeville), STAU2 (St. Augusta), and STET2 (St. Etienne) in Zone 2, and NORM3 (Normandin), CAUS3 (Causapscal), HEBE3 (Hebertville), and LAPO3 (La Pocatière) in Zone 3. OTT (Ottawa) is a location in Ontario. PC1 and PC2 are the first two principal components from singular value decomposition of the trial-standardized yield data ("Scaling = 1," "Centering = 2"), with the singular values fully partitioned to the trial scores ("SVP = 2") for proper visualization of the correlations among trials.

trial two-way table to singular value decomposition, after proper singular value partition (Yan, 2002). The most obvious message from this fairly crowed biplot is that the trials placed on the upper portion of the biplot and those on the lower portion were negatively correlated, as indicated by the obtuse angles between them. This indicates existence of strong GE. The second step is to summarize the trials conducted at a location by a location marker, the placement of which is determined by the mean coordination of the trials (**Figure 4**). For example, the placement of the location LAPO3 (in red) was determined by the seven trials conducted at LAPO3, namely LAPO3_13, LAPO3_14, LAPO3_15, LAPO3_16, LAPO3_17, LAPO3_18, and LAPO3_19 (in black). The genotypes are represented by "+" for clarity. The biplot in **Figure 4** is both a GGE biplot and a GGL biplot, thus the term GGE+GGL biplot.

In **Figure 4** the 10 locations are clearly separated into two groups: group 1 include locations NORM3, HEBE3, CAUS3, PRIN2, PINT2, and STAU2 on the upper quadrant, and group 2 include locations NDHY1, STHU1, LAPO3, and OTT on the lower quadrant. Thereby the GE is divided into repeatable GE and unrepeatable GE. The genotype by location group interaction, i.e., the difference in the placement between the two location groups, is the repeatable GE; the genotype by trial interaction within groups, i.e., the variation in the placement among the trials within each of the two location groups, is the unrepeatable GE. The two location groups suggests two different mega-environments. All locations in mega-environment 1 (ME1) belong to Zone 2 or Zone 3 of Quebec; locations in mega-environment 2 (ME2) consists of two Zone 1 locations, a Zone 3 location, and OTT. Thus, the mega-environment differentiation was largely, but not completely, consistent with the agroclimatic zones.

LG Biplot Analysis

Presented in **Figure 5** is the LG biplot based on the same dataset that was used to generate the GGE + GGL biplot (**Figure 4**). The steps to construct the LG biplot follows. First, a genetic correlation matrix among locations was calculated for each year.

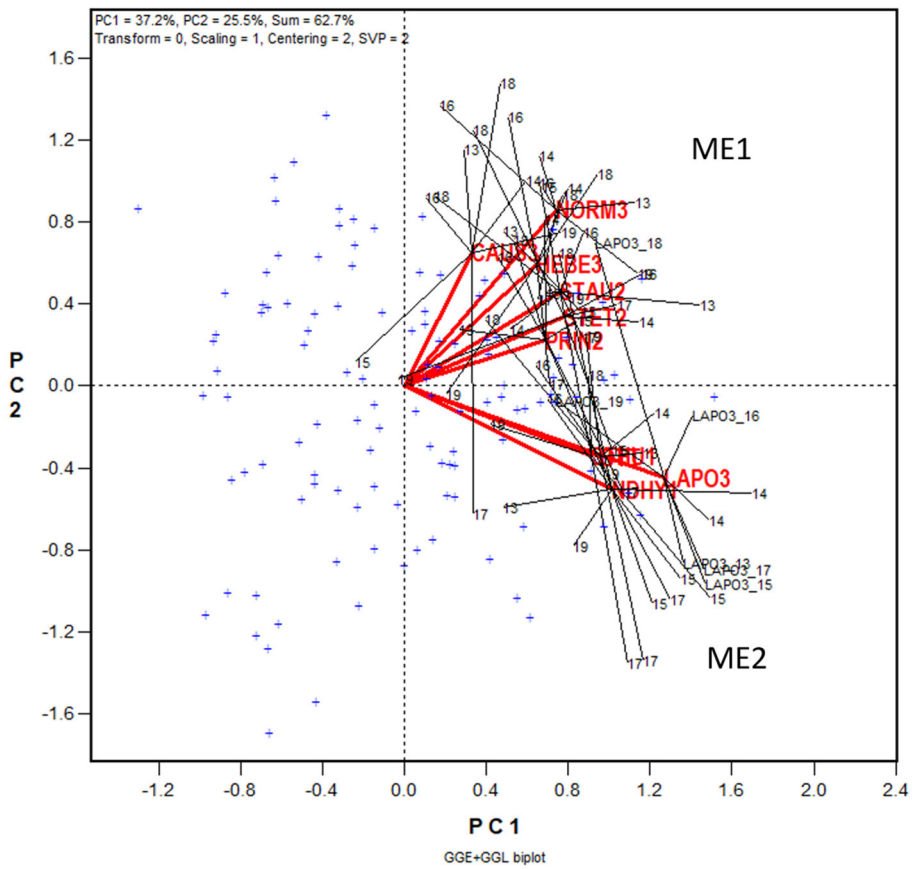

FIGURE 4 | GGE + GGL biplot modified from **Figure 3** to show two groups of locations or oat mega-environments (ME) in Quebec. Mega-environment 1 (ME1) consists of Zone 2 and Zone 3 locations PINT2, PRIN2, STAU2, STET2, CAUS3, HEBE3, and NORM3, and mega-environment 2 (ME2) includes locations NDHY1, STHU1, LAPO3, and OTT. The trials conducted at each location are presented as a cluster of trials, with the location name placed at the center and the individual trials, indicated by the last two digits of the year, placed around it, and the trials are connected to the center with straight lines. Two locations are regarded as belonging to the same mega-environment if their clusters overlap; they are regarded as belonging to different mega-environments otherwise. The variation in the placement of the locations between mega-environments represents repeatable GE and the variation in the placement of the trials within mega-environments represents unrepeatable GE. The genotypes are displayed as "+" in blue for clarity.

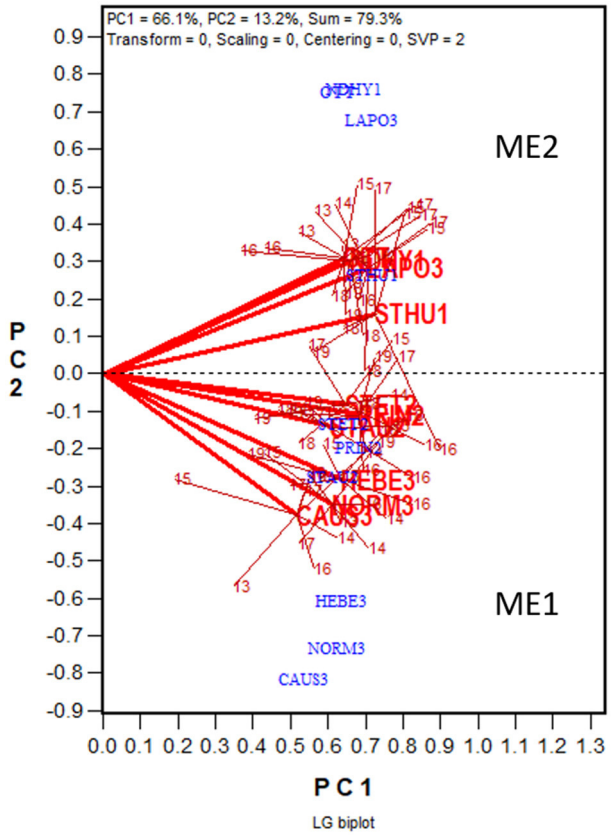

FIGURE 5 | LG biplot to show two oat mega-environments in Quebec. PC1 and PC2 are the first two principal components from singular value decomposition of the location by trial two-way table of correlations, without centering ("Centering = 0") or scaling ("Scaling = 0"). The LG biplot is a location by trial biplot, with the locations presented in blue and the trials in red. The trials conducted at each location are presented as a cluster of trials, with the location name placed at the center and the individual trials, indicated by the last two digits of the year, placed around it. The trials are connected to the location with straight lines. Two locations are regarded as belonging to the same mega-environment (ME) if their clusters overlap; they are regarded as belonging to different mega-environments otherwise. The same two MEs (ME1 and ME2) shown in **Figure 4** are shown in this LG biplot. The variation in the placement of the locations between mega-environments represents repeatable GE and the variation in the placement of the trials within mega-environments represents unrepeatable GE.

Second, the yearly correlation matrices were stacked to form a location by trial two-way table of correlations, each trial being a location-year combination. Third, the location by trial table was submitted to singular value decomposition, without entering or scaling ("centering = 0, scaling = 0"). Fourth, the resulting first two principal components were used to construct a location by trial biplot. Fifth, as in the GGE + GGL biplot, the trials conducted at a location in different years were summarized by a location marker defined by the mean coordination of the trials. Finally, the trials at a location are displayed as a cluster, with the location marker as the center and the trials in different years as members; the trials and the location are connected with straight lines. If the clusters of two locations overlap, they are regarded as belonging to the same mega-environment; if they do not overlap, they are considered as belonging to different mega-environments. In the LG biplot, the variation among trials and locations within a mega-environment, i.e., the variation among trials at overlapping locations, represents unrepeatable GE; the variation between two mega-environments represents repeatable GE. It can be seen that the same two mega-environments shown in the GGE+GGL biplot (**Figure 4**) are clearly separated in the LG biplot (**Figure 5**); thus, the two approaches are functionally equivalent or similar. Importantly, the LG biplot has the advantage that it does not require any common genotypes to be tested in different years. Therefore, it can be used to reveal repeatable GE and delineating mega-environments using multi-year trial data in which completely different sets of genotypes are tested in different years (Yan et al., 2021).

A general comment on the use of biplots follows. A 2-D biplot is usually used for data visualization for convenience and on the understanding that the most important patterns in the data are captured by the first two principal components. However, there may be cases where some important patterns exist in higher order principal components. This is usually indicated by the presence of vectors that are obviously shorter than others. When this is the case, variation not displayed in the biplot can be explored by biplots displaying a subset of the data. A recent example can be found in Yan et al. (2021).

Utilization of Repeatable GE by Selecting Mega-Environment Specific Cultivars

The approach to utilizing repeatable GE is to select separately for each mega-environment, preferably using the mean vs. stability view of the GGE biplot (**Figure 6**). The red line with a single arrow passes through the biplot origin and the mean environment (which has mean coordination of all environments) and is referred to as the average environment axis (AEA) or GGE-Mean axis; the arrow points to higher mean yield. The blue line with arrows at both ends points to greater instability in either direction; it can be referred as the GGE-stability axis (Yan, 2001; Yan and Kang, 2002; Yan and Tinker, 2006). This is an extended application of the inner-product property of a biplot (Gabriel, 1971). Thus, the three highest yielding cultivars in ME1 across 2013–2019 were Akina > Nicolas > Nice (**Figure 6A**), and those for ME2 were Nicolas > Akina > Richmond (**Figure 6B**). Therefore, the repeatable GE can be utilized by recommending different sets of cultivars in ME1 and ME2.

The similarity/dissimilarity in cultivar ranking between ME1 and ME2, along with that in the undivided whole region, are further presented in the which-won-where view of the GGE biplot in **Figure 7**. The polygon or which-won-where view of the GGE biplot (Yan et al., 2000) is also an extended application of the inner-product property of a biplot (Gabriel, 1971). The polygon was formed by connecting the genotypes that are placed away from the biplot origin in all directions. For each polygon side a line perpendicular to it was drawn from the biplot origin. These lines dissect the biplot into sectors. For each sector, the genotype at the vertex is the nominal highest yielder for the environments or mega-environments fell in it. In this case, Akina

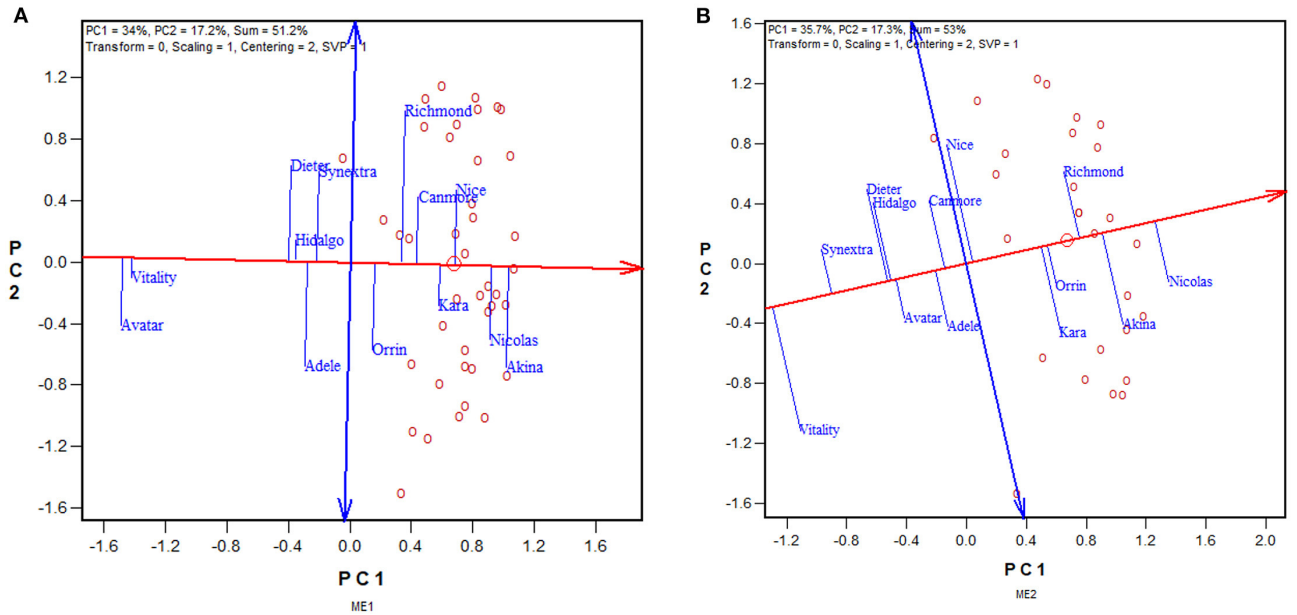

FIGURE 6 | GGE biplots to show the mean yield and instability of 13 oat cultivars in **(A)** mega-environment 1 (ME1) and **(B)** mega-environment 2 (ME2), across 2013–2019. See **Figure 4** and associated text for the definitions of ME1 and ME2. PC1 and PC2 are the first two principal components from singular value decomposition of trial-standardized yield data ("Centering = 2," "Scaling = 1"). The singular values were entirely partitioned to the genotypic scores ("SVP = 1") for proper genotype evaluation. The trials are represented by "o" for clarity. The red line with a single arrow is the average environment axis (AEA), the arrow pointing to higher mean yield. The blue line with arrows on both ends is the instability axis, the arrows pointing to greater instability in either direction.

was the highest yielder in ME1 while Nicolas was the highest yielder in ME2 and "ALL," indicating crossover genotype by subregion interaction. On the other hand, ME1 was placed close to the radiate line labeled "1," which separates ME1 from ME2; this means that Akina had higher yield than Nicolas in ME1 but not by much. The two mega-environments were moderately correlated ($r = 0.652$; **Figure 7**) and shared Akina and Nicolas as the top two yielding cultivars, though in a reversed order. Thus, the two oat mega-environments in Quebec were classified as sub mega-environments within one of the three major oat mega-environments in Canada (Yan et al., 2021).

ACCOMMODATION OF UNREPEATABLE GE THROUGH ADEQUATE TESTING

The solution to accommodating unrepeatable GE is to test adequately *within a target mega-environment*, i.e., to test at a sufficiently large number of locations in a sufficiently large number of years with sufficiently large number of replicates so as to sufficiently sample the environments and to achieve a sufficiently high heritability as defined in Equation 7 or Equation 8. It is obvious that more replicates, more locations, and more years will lead to a higher heritability. The solution to identify widely adapted cultivars (within a meg-environment) is to "test widely" (Troyer, 1996). However, each additional replicate, location, or year involves considerable cost. As a compromise between selection reliability and test cost, the concept "adequate test" was proposed and defined (Yan et al., 2015; Yan, 2016). The terms "adequate testing," "optimum testing," and "minimum testing" are used interchangeably in this article to indicate that a minimum level of testing in terms of years, locations, and replicates must be conducted to achieve sufficiently reliable selection. When tested inadequately, the selection intensity must be lowered according to the achieved heritability, to prevent superior genotypes from being mistakenly discarded. The "optimum" level of replicates, locations, or years was defined as one to achieve a heritability of 0.75, based on examining a heritability response curve (Yan et al., 2015; **Figure 2** this article). However, Cobb et al. (2019) suggested that a heritability of 0.5 was sufficient for reliable selection of the best 10 individuals to be used to start the next breeding cycle.

Optimum Number of Years

Based on Equation 7 and assuming neglectable GL and experimental error or unlimited number of locations and replicates, the minimum number of years required to achieve a heritability of 0.75 can be estimated by

$$y_{min} = \max[1,\ 3(\frac{\sigma^2_{GY}}{\sigma^2_G})] \tag{15}$$

For example, based on the yield data from three-year spans of Quebec provincial oat tests, the estimated minimum number of years to achieve a heritability of 0.75 was from 1.2 to 6.3 and averaged 3.2 (**Table 2**), while the officially required number of years to register a cultivar is three. So, the requirement for

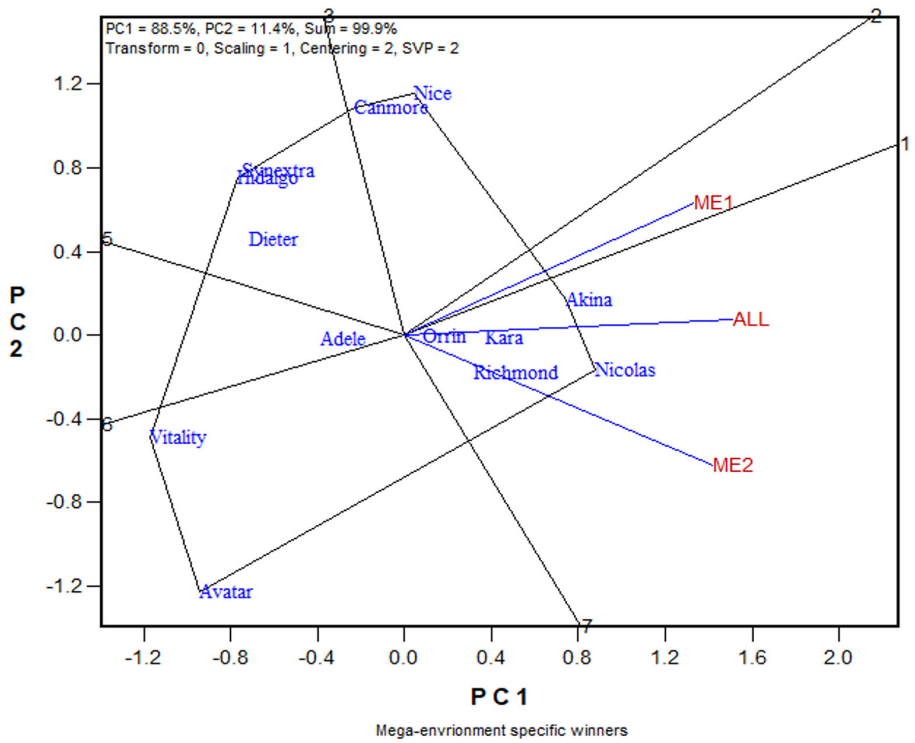

FIGURE 7 | The which-won-where view of the GGE biplot to show the relative yield of 13 oat cultivars in mega-environment 1 (ME1), mega-environment 2 (ME2), and the undivided Quebec oat growing regions (ALL). The polygon was formed by connecting the genotypes that are placed away from the biplot origin in all directions. For each polygon side a line perpendicular to it was drawn from the biplot origin. These lines dissect the biplot into sectors. For each sector, the genotype at the vertex is the nominal highest yielder for the environments or mega-environments fell in it. In this case, Akina was the highest yielder in ME1 while Nicolas was the highest yielder in both ME2 and "ALL."

three years of testing was adequate and appropriate in general. More years of testing were required for the 2016–2018 and the 2018–2020 (**Table 2**) spans due to reduced genetic variability and therefore achieved heritability.

Optimum Number of Locations

Yearly multi-location trials are usually balanced as the same set of genotypes are tested at all locations. Therefore, it is convenient to use yearly data to estimate the number of locations required for adequate testing. Assuming an infinite number of replicates or negligible experimental error, the heritability within a year (Equation 10) can be reduced to

$$H_{rl,max} = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_{GL}^2}{l}} \quad (16)$$

where $H_{rl,max}$ is the maximum achievable within-year heritability (Yan, 2021). Based on this equation, the minimum number of locations required to achieve a heritability of 0.75 can be estimated by (Yan et al., 2015; Yan, 2021)

$$l_{min} = \max[1,\ 3(\frac{\sigma_{GL}^2}{\sigma_G^2})] \quad (17)$$

The minimum number of locations so estimated is expected to differ with the year. Therefore, it should be estimated for a number of years to achieve a good understanding on the required number of test locations for a target mega-environment (Yan et al., 2015). Presented in **Table 2** are the estimated yearly minimum number of locations based on the 2013–2019 Quebec provincial oat trial data for the two mega-environments as well as for the undivided Quebec oat growing region. When estimated for the undivided region, the mean number was 8.4, in comparison to the actual number of locations of 9.6. Thus, the number of locations used was more than adequate in most years.

Interestingly, when estimated for each mega-environment, the estimated minimum number was ∼one location more than that actually used (7.3 vs. 5.9 for ME1 and 4.6 vs. 3.7 for ME2). Thus, even though there is a clear mega-environment differentiation, the trials in one mega-environment still provided useful information to selection for the other, because the two mega-environments were positively correlated (**Figure 7**). In contrast, the southern vs. northern oat mega-environments in eastern Canada were uncorrelated, and as a result, the total required number of locations was smaller when estimated separately for each mega-environment than that when estimated for the undivided region (Yan et al., 2015). In a Canada-wide study, the southern vs. northern mega-environments in eastern

TABLE 2 | The minimum number of years (y_{min}) required to achieve a heritability (H) of 0.75 estimated on the yield data of three-year spans from the Quebec provincial oat registration trials.

Three-year span	No. of genotypes	σ^2_G	σ^2_{GY}	σ^2_{GLY}	σ^2_P	H	y_{min}
2013–2015	26	0.63	0.25	0.50	0.71	0.88	1.2
2014–2016	23	0.70	0.34	0.58	0.81	0.86	1.5
2015–2017	27	0.48	0.53	0.73	0.65	0.73	3.3
2016–2018	27	0.30	0.64	0.80	0.52	0.59	6.3
2017–2019	27	0.52	0.49	0.70	0.68	0.76	2.8
2019–2020	30	0.29	0.42	0.64	0.43	0.68	4.3
Mean							3.2

TABLE 3 | The estimated minimum number of locations in comparison to that actually used for the Quebec provincial oat trials.

	The whole region		ME1[a]		ME2[a]		ME1 + ME2
Year	Actual	Estimated	Actual	Estimated	Actual	Estimated	Estimated[b]
2013	8	11.2	5	5.7	3	4.6	10.3
2014	9	6.8	6	4.6	3	2.4	7.0
2015	10	6.8	6	9.6	4	1.5	11.1
2016	10	5.9	6	3.4	4	8.3	11.7
2017	10	8.2	6	8.8	4	1.4	10.2
2018	10	7.4	6	6.9	4	6.1	13.0
2019	10	12.5	6	12.4	4	7.7	20.1
Mean	9.6	8.4	5.9	7.3	3.7	4.6	11.9

[a] *See **Figure 4** or **Figure 5** for the definition of mega-environment 1 (ME1) and mega-environment 2 (ME2);* [b] *The estimated number for ME1 + ME2 is the sum of the estimated number for ME1 and that for ME2.*

Canada were designated as ME1 and ME2, respectively, while the two Quebec mage-environments in **Figure 4** or **Figure 5** were designated as ME2a and ME2b (Yan et al., 2021). Given the results in **Table 3**, cultivar recommendation for each of the two Quebec mega-environments should consider performance both within the mega-environment and across the whole region, as shown in **Figure 7**.

Optimum Number of Replicates

Several classic studies investigated the optimum numbers of years, seasons, test locations, and replicates for the allocation of a fixed number of plots or fund according to the relative magnitudes of various variance components (Sprague and Federer, 1951; Hanson and Brim, 1963; Wricke and Weber, 1986; Swallow and Wehner, 1989). Conclusions from this "resource allocation" approach inevitably led to the suggestion to maximize the number of locations and/or years and to minimize the number of replicates (i.e., to use a single replicate) (McCann et al., 2012; Schmidt et al., 2018). However, this conclusion applies only when it is possible to increase the number of locations and/or years. For a breeding program or a regional crop recommendation committee, yield trials are conducted every year at a more or less fixed number of locations. Researchers need to know the minimum or optimum number of replicates under this scenario. To answer this question, an equation was derived from the definition of heritability on the single trial basis, in the form of Equations 15 and 17 (Yan et al., 2015). More recently, another equation was developed for estimating the optimum number of replicates in a multi-location context (Yan, 2021):

$$r_l = \max[1, 3\left(\frac{\sigma^2_\varepsilon}{\sigma^2_G}\right)\left(\frac{H_{rl,\,max}}{l}\right)] \quad (18)$$

where r_l is the optimum number of replicates given the number of locations l, and $H_{rl,\,max}$ is as defined in Equation 16. Equation 18 shows that the required number of replicates is determined by the relative magnitude of experimental error variance, $\frac{\sigma^2_\varepsilon}{\sigma^2_G}$, and is modified by the number of locations in a non-linear manner, because an increase in the number of locations also improves $H_{rl,\,max}$ (Equation 16). Applying this equation to the yield data of some oat trials conducted across Canada, it was determined that two replicates would suffice to identify the highest yielding oat cultivars (Yan, 2021). Applying this equation to the 2015–2019 yield data of barley, oat, spring wheat, and winter wheat trials conducted in Ontario also led to the conclusion that two replicates would suffice (Yan et al., 2000). It is recommended that similar analysis be conducted for other crops and regions. Regional variety trials are usually conducted with three or four replicates. Reducing the number of replicates to two, if supported,

can substantially reduce the evaluation cost or allow more genotypes to be evaluated with the same cost.

Importantly, reliable estimation of the various variances are a prerequisite to accurate estimation of the optimum number of years, optimum number of locations, and optimum number of replicates for adequate testing (Arief et al., 2015).

Adjust for Spatial Variation

The discussions on optimum testing and optimum replication above assumed that the field and management are uniform within each trial. However, spatial variation within trials has been recognized as a major source of experimental error. Traditionally it is controlled by blocking, i.e., dividing a replicate into blocks, such as the so-called incomplete blocks design (R.A. Fisher, from Street, 1990). This is referred to "dealing with spatial variation by design." In the last three decades, spatial analysis and adjustment becomes an increasingly popular research subject and a routine practice in the analysis of crop variety trials. The use of spatial analysis makes experimental design more flexible.

In a variety evaluation trial, g genotypes are usually allocated into a rectangular field of b rows (blocks) and c column (plots). The observed value in a plot, Y_{ij}, is, therefore, combined effects of the row, the column, the genotype, and the experimental error:

$$Y_{ij} = \mu + row_i + col_j + g_k + \varepsilon_{ij}, \tag{19}$$

μ being the mean of the trial. The effects of rows and columns can be modeled by various spatial analysis techniques (Cullis and Gleeson, 1991; Gilmour et al., 1997; Cullis et al., 1998; Burgueño et al., 2000; Qiao et al., 2000; Yang et al., 2004). Spatial analysis is a within-trial analysis so it is also referred to "local analysis" (Kempton et al., 1994; Grondona et al., 1996). A straightforward and intuitive approach is to use a polynomial regression to model any trend across the plots within each block (Yan, 2014), which is routinely used in the Ottawa oat breeding program. The order of the polynomial regression can be set according to the block size of the block. An iterative procedure can be used to adjust the Y_{ij} values so as to minimize the experimental error. Adjusted genotypic values are then calculated from the adjusted plot values at the final iteration. Spatial adjustment based on polynomial regression usually leads to reduced trial coefficient of variation and increased trial heritability (Yan, 2014). The plot values, and thus the genotypic values, will not be altered if no spatial trend is found. An example of spatial trend adjustment for a block of 36 plots in an oat trial conducted in Ontario in 2019 is presented in **Figure 8**. This procedure can also be used to fill missing plot values.

Genomic Selection: To Replace Multi-Environment Evaluation With Multi-Model Prediction

Some researchers believe that genomic selection will eventually replace breeders' visual selection and even alter the role of yield trials in making selection and recommendation decisions (Heffner et al., 2009; Jannink et al., 2010). Indeed, encouraging results of genomic selection have started to emerge as advanced genotyping, bioinformatics, and genomic modeling procedures have become available (e.g., Tinker et al., 2016; Bekele et al., 2018). The confidence on genomic selection comes from two aspects. First, genome-wide markers can sufficiently capture the genotypic variability of a relevant breeding population tested in a relevant environment (i.e., a trial at a location in a year). That is, the genetic variability of the breeding population observed in the trial can be accurately captured by a genomic model. Second, genomic models can be developed for a large number of trials that sufficiently to fully represent the target mega-environment. Assuming m models have been developed from m trials covering multiple locations and years, then predicting the performance of a breeding population using m models would be equivalent to testing the breeding population in m trials. This represents a great advantage of genomic selection over conventional selection because in a practical breeding program, it is impossible to test a large breeding population in replicated trials, let alone at multiple locations in multiple years. Assuming a genotype by model two-way table of predictions for a breeding population, the achievable heritability with genomic selection, H_m, can be estimated by

$$H_m = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_{GM}^2}{m}} \tag{20}$$

where σ_{GM}^2 is the variance for genotype by model interaction. As m increases, the achievable heritability with genomic selection and hence selection reliability can become much higher than what is achievable by conventional selection (Yan et al., 2019b). From this viewpoint, genomic selection is potentially a much more effective approach to dealing with unrepeatable GE.

On the other hand, yield trials aiming at genomic model development are large and expensive; it remains a question as to when model development can be considered complete and such trials are no longer needed (Yan et al., 2019b). If the year factor is completely random, then genomic model development may be considered complete at some point when sufficient data (years and locations) have been obtained. However, if there is a trend in climatic change, data from recent years would be more relevant for predicting future-year performances, and the trials must continue. It is also a question whether routinely conducted yield trials with a limited number of entries can be used to replace the large-scale trials for genomic model development or refinement, although it was so suggested (Heffner et al., 2009). Cost efficiency will continue to be a determinant factor to the application of genomic selection in plant breeding.

SELECTING FOR MEAN PERFORMANCE AND SELECTING AGAINST INSTABILITY

Superior cultivars must demonstrate high and stable yield cross the target mega-environment. Cultivars yielded well in some environments but poorly in others, relative to other cultivars, are said to be unstable and undesirable as they can cause unbearable losses to growers. Various stability indices have been developed in order to quantify instability. Lin et al. (1986) reviewed nine stability or instability parameters and classified them into four

groups. More parameters were proposed after that (e.g., Huehn, 1990). The initial idea of stability analysis was to select against unstable genotypes rather than to select for stable genotypes. This idea was somehow twisted to treating a stability index as a positive trait, which caused confusion among researchers. It would be less confusing to call these stability indices as instability indices. The large number of indices are also confusing to practical breeders. The purpose of this section is to reinstall the original idea of stability analysis and to clear up the confusions.

First, a stability index should reflect a genotype's susceptibility to GE, because it is GE that caused its unstable performance across environments. The numerous stability parameters may be classified according to its composition in terms of G, E, and GE. In the classification of Lin et al. (1986), stability indices in Group B involve GE only; they are suitable instability parameters. Indices in group A involve both E and GE; they are not suitable parameters because GE is confounded with E. The linear regression coefficient b against E in Eberhart and Russell (1966) (Group C) is a genotype's response to E; its usefulness depends on how well the linear regression fits the data (Lin et al., 1986), which is usually poor (Zobel et al., 1988). When the fit is sufficiently good, $b = 1$ means stable, $b > 1$ means good performance in high-yielding environments, and $b < 1$ means good performance in low-yielding environments (Ceccarelli, 1989). Deviation from the linear regression (Group D) is merely a measure of the goodness of fit of the linear regression and is not a useful measure of stability.

Second, stability analysis is a concept of selection within a mega-environment. So, it should be conducted within mega-environments rather than across mega-environments, unless the mega-environments are highly correlated. In such cases, the mega-environments should be merged and treated as a single mega-environment.

Third, a stability index representing GE must be used in combination with the mean performance (G), thus the term G+GE or GGE (Yan et al., 2000, 2007). High stability (less GE) is desirable only when combined with high mean performance. High stability is least desirable when combined with low mean yield because it means consistently low yielding (Yan et al., 2007). Parameters or procedures combining both G and GE include the superiority index of Lin and Binns (1988, 1994) and the stability index of Kang (1993). In addition, several graphical methods also combines G and GE. These include the AMMI1 biplot (Zobel et al., 1988) and the AEA view of the GGE biplot (**Figure 6**; Yan, 2001). Zobel et al. (1988) showed that AMMI analysis was superior to the joint regression of Eberhart and Russell (1966) and Alwala et al. (2010) concluded GGE biplot analysis to be a better platform than the joint regression. There is some debate on whether AMMI or GGE is a better approach in analyzing yield trial data (Gauch, 2006; Yan et al., 2007; Gauch et al., 2008; Yan, 2011). Many studies compared AMMI and GGE biplot analysis and concluded that GGE biplot analysis was superior (e.g., Badu-Apraku et al., 2012; Amira et al., 2013; Hoyos-Villegas et al., 2016; Oliveira et al., 2019). AMMI analysis was advocated as a means to separate "signal" (true GE) from "noise" (error) in GE and a means to use the GE-signal to adjust the genotypic means (Gauch and Zobel, 1988; Gauch, 2013). AMMI1 or AMMI2 (i.e., main effects plus the first one or two principal components of GE) is often found the best AMMI model, and the AMMI1 biplot is often used as a visual tool for genotype evaluation (Gauch, 2006).

FIGURE 8 | An example to show the plot values within a block as adjusted according to the field trend modeled by a polynomial regression.

Unfortunately, the AMMI1 biplot is not an effective graphical presentation of G + GE because G is often masked by the much larger E in it, because its G and GE axes are in different units, and because it does not have the inner-product property of a true biplot (Yan, 2011). Similar to AMMI analysis, GGE biplot analysis can also be viewed as a means to separate signal from noise. In a two-dimensional GGE biplot such as that in **Figure 6**, the first two principal components are considered as signals and the higher dimensions as noise. This GGE biplot displays the amount of G+GE in between that of AMMI1 and AMMI2; so, it should be close to the best model in most cases. In the AEA view of the GGE biplot (**Figure 6**), the GGE-mean axis represents the GE-adjusted genotypic means, while the GGE-stability axis represents the genotypes' susceptibility to unrepeatable GE (instability). Both axes pass through the biplot origin and are perpendicular to each other, meaning that they are independent parameters. Thus, the AEA view of the GGE biplot is a convenient tool for visual analysis of genotype-by-environment data and for visual selection for mean performance and against instability.

Finally, "test adequately" is much more important than any stability analysis. When tested adequately, genotypes with high mean performance should also be genotypes that are relatively stable, because it is not possible for a highly unstable genotype to achieve very high mean performance. However, when not tested adequately, as indicated by a low heritability across locations and years, neither the estimated mean nor the estimated stability is reliable, and a low selection intensity or culling rate must be applied. In such cases, selection should be mainly on mean performance, rather than on stability. Instead, effort should be made to understand the causes of the instability for a high-yielding genotype. For example, severe lodging may be the reason for its low yield in a severely lodged trial. If severe lodging rarely occurs in the target region, then the genotype is expected to show good mean yield and stability when tested adequately; if lodging is a common yield-limiting factor in the target region, then the genotype is expected to have low mean yield when tested adequately.

SELECTION FOR MULTIPLE TRAITS

While geneticists can focus on a single trait and ignore others, breeders must deal with multiple traits. In addition to high yield, which is always the most important breeding objective, a cultivar must meet a minimum requirement for each and every trait that is important to the relevant growers, processors, and end-users. In fact, the greatest challenge in plant breeding is to combine all desirable traits in a single genotype, because key breeding objectives are often adversely associated, due to either genetic linkage or pleiotropy (e.g., Tanksley, 1983; Yan and Wallace, 1995; Asins, 2002; Cooper et al., 2009; Hao et al., 2014; Crespo-Herrera et al., 2016). Strategies for multi-trait selection include independent culling and index selection (Simmonds and Smartt, 1999; Yan and Frégeau-Reid, 2008; Yan, 2014). Independent culling is to cull all genotypes that fail to meet the minimum requirement for any breeding objective, because such genotypes will not be accepted as cultivars. Index selection is to rank genotypes based on an index that is composed to reflect the perceived economic values of the genotypes. Independent culling can be implemented at all stages in the breeding cycle but it is more important in the early breeding stages when multi-location yield trials are not possible. Index selection is mainly implemented in the yield trial stage, at which all important traits can be determined. For cultivar development-oriented genomic selection, both independent culling and index selection should be conducted.

Independent Culling

Independent culling is important to ensure that selected high yielding genotypes will be accepted by growers and end-users; it is also an effective approach to reduce the breeding population size safely and speedily. Assume that t is the number of independently inherited breeding objectives, each with a heritability h_k^2, with $k = 1, \ldots t$. If h_k^2 is used as the culling rate for trait k, then the joint culling rate would be:

$$h_t^2 = 1 - \prod_1^t (1 - h_k^2) \tag{21}$$

and the number of genotypes must be retained to ensure that the best genotype is selected is:

$$N = n(1 - h_t^2), or \tag{22}$$

$$N/n = 1 - h_t^2,$$

N/n being the proportion of the breeding population that must be retained. For example, assume the culling rate for each of five traits is 0.3, then, according to Equation 21, the joint culling rate would be 0.83, and the retaining rate would be 0.17 or 17%. Therefore, a large proportion of the population can be safely culled by independent culling if multiple traits are considered, even though the heritability or culling rate is low for each trait. This explains the effectiveness of visual selection (culling) by an experienced breeder, who can visualize and select on many traits simultaneously.

Genomic selection for oat yield in eastern Canada proves effective (Bekele, Tinker, and Yan, unpublished results); it should also be effective for other traits that are more simply inherited than yield. Therefore, independent culling based on genomic models is expected to be more accurate than visual selection by even the most experienced breeder. If the traits under consideration are positively correlated, the overall culling rate would be lower than when they are independent; the overall culling rate would be higher if the traits are negatively correlated, which is often the case. The overall culling rate can be much higher if some of the target traits are simply inherited and less affected by GE and experimental error, for example, oil content in oat (Hizbai et al., 2012; Yan et al., 2016).

GYT (Genotype by Yield*trait) Analysis

A large portion of the genotypes that survived independent culling should be qualified as a cultivar if they are sufficiently high yielding. Therefore, the focus of selection following independent culling should be on yield although other target traits should also be considered. Selection based on a selection index is the common method for selection on multiple breeding objectives (note but: not any traits). Here, the GYT (genotype by yield*trait) analysis (Yan and Frégeau-Reid, 2018; Yan et al., 2019a) is recommended over the traditional index selection.

In traditional index selection, the superiority of genotype i, P_i, is calculated as

$$P_i = w_0 y_i + \sum_{j=1}^{t} (w_j x_{ij}), \tag{23}$$

where t is the number of breeding objectives that are to be selected in addition to yield, y_i is the standardized yield for genotype i, w_0 is the weight assigned for yield, w_j is the weight assigned for trait j, and x_{ij} is the standardized value of genotype i for trait j.

In the GYT approach, the superiority of a genotype may be presented as

$$P_i = y_i \sum_{j=1}^{t} (w_j x_{ij}). \tag{24}$$

The selection index for a genotype is usually presented as the standardized value of P_i. The difference between the traditional selection index (Equation 23) and the GYT approach (Equation 24) follows. In traditional index selection, the weight for a trait other than yield is a fixed value for all genotypes, while in the GYT approach it varies with the yield level of each genotype. In traditional index selection the emphasis is on the levels of the traits; in the GYT approach it is on the levels of yield-trait combinations. The GYT concept is better in reflecting the economic value of a trait. For example, superior lodging resistance (or high protein) has little value in a low yielding genotype but it is highly valuable in a high yielding genotype. Consequently, based on the traditional selection index, a low yielding genotype may be ranked the highest due to its superior levels in other traits; such genotypes will not be accepted as cultivars by growers, however (Yan et al., 2019a). This problem can be prevented with the GYT approach, as the genotypes ranked highest will always have high yield levels.

Another advantage of the GYT approach is that the superiority and the trait profiles of the genotypes can be visually investigated in a biplot, referred to as GYT biplot (Yan and Frégeau-Reid, 2018; Yan et al., 2019a,b). As an example, the mean values of the 13 oat cultivars for eight important traits from the 2013–2019 Quebec oat trials are presented in **Table 4**, ranked by their GYT index. The steps to construct a GYT biplot follow. First, standardize the genotype by trait table for each trait. Second, multiply yield with each trait to form a genotype by yield-trait combination two-way table. Third, subject the weighted genotype by yield-trait two-way able to singular value decomposition to obtain the principal components (PC). Fourth, multiply each of the yield-trait combination PC scores with the assigned weight. And finally, construct a biplot using the genotypic and trait combination scores of the first two principal components based on the yield-trait combination preserving singular value partition. Note that for milling oat a higher value is more desirable for all the traits listed in **Table 4** except lodging and oil content. For these two traits a smaller value is more desirable; they were therefore given a weight of "−1." The information contained in **Table 4** can be visualized in a GYT biplot (**Figure 9**). The biplot clearly shows the rank of the genotypes in their GYT index, i.e., Nicolas > Akina > Kara > Richmond >…> Avatar. Thus, Nicolas and Akina

TABLE 4 | Mean trait values of 13 oat cultivars tested in the 2013–2019 Quebec provincial oat trials and their GYT (Genotype by Yield*Trait) index.

Genotype	Traits and weights								
	Yield (kg ha^{-1})	β-glucan(%)	Groat (%)	Oil(%)	Protein (%)	Test weight (kg hl^{-1})	1000-Kernel Weight (g)	Lodging(0–9)	GYT Index
		1	1	−1	1	1	1	−1	
Nicolas	5,948	4.3	73.9	6.1	13.2	53.4	35.8	2.9	1.3
Akina	5,853	4.8	72.7	7.1	13.4	52.3	38.0	2.4	1.2
Kara	5,680	4.7	71.9	8.0	14.0	54.0	38.1	2.0	0.9
Richmond	5,631	3.8	71.7	5.4	12.6	55.1	39.2	3.0	0.5
Canmore	5,447	4.6	71.6	7.6	14.2	54.9	39.5	3.3	0.3
Nice	5,559	4.4	72.6	8.4	13.5	53.3	38.5	3.9	0.1
Orrin	5,468	4.4	71.1	6.9	13.3	54.2	38.5	3.5	0.1
Adele	5,354	4.6	75.3	8.4	12.8	54.0	38.7	4.5	−0.3
Dieter	5,183	4.1	73.4	5.7	14.0	54.3	38.9	3.6	−0.4
Synextra	5,171	4.3	72.0	7.3	14.9	56.0	37.3	3.7	−0.4
Vitality	5,049	4.0	75.7	7.9	13.6	53.9	40.3	3.9	−0.9
Hidalgo	5,158	4.7	74.3	8.0	13.1	53.1	34.5	4.0	−0.9
Avatar	5,060	3.9	74.8	7.9	12.2	56.6	36.2	4.9	−1.3

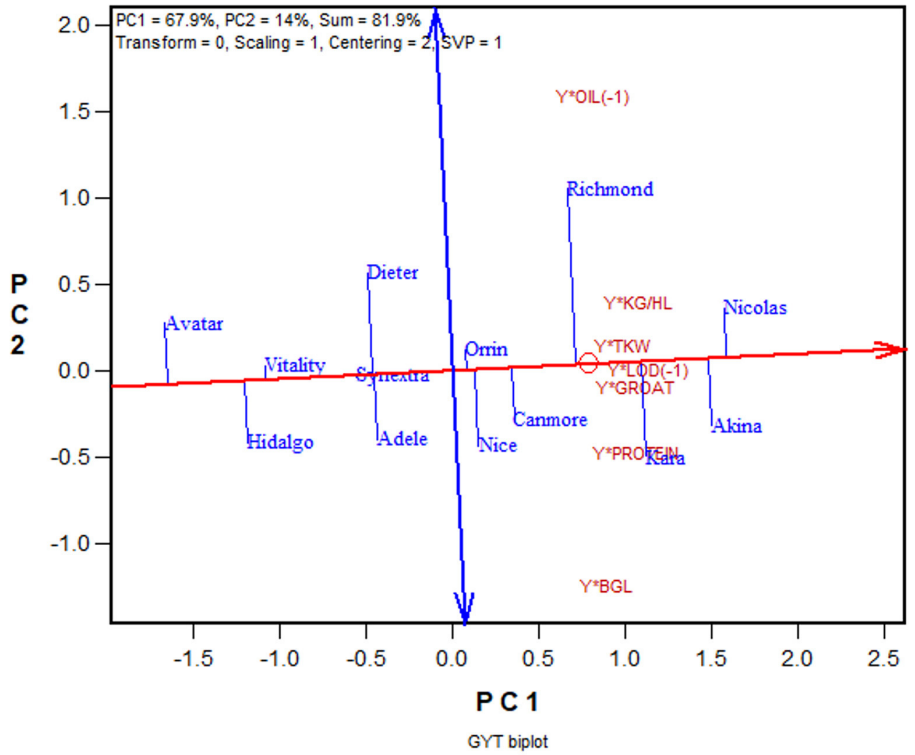

FIGURE 9 | GYT biplot to display the yield-trait combinations of 13 oat cultivars tested in the 2013–2019 Quebec provincial oat trials. The biplot was based on singular value decomposition of yield-trait combination standardized data ("Centering = 2, Scaling = 1"). The red line with a single arrow is the average yield-trait axis, the arrow pointing to higher GYT index. The blue line with arrows on both ends indicate contrasting trait profiles of the genotypes. For example, it showed Richmond to be strong yield-oil combination but weak in yield-β-glucan combination, while Kara had the opposite trait profile to that of Richmond.

should be selected and recommended to the Quebec oat growers without hesitation; they are in fact the most important two cultivars in Quebec. The biplot also shows the trait profiles of the genotypes. For example, it shows that Richmond is superior in having a low oil content but inferior in having a low β-glucan content. In fact, all cultivars placed above the red line (the GYT index axis) have relatively low oil and low β-glucan whereas the opposite is true for cultivars placed below the GYT index axis. "Y*Oil(−1)" and "Y*Lodging(−1)" indicate that oil content and lodging score were given a weight of "−1" because high oil content and high lodging are undesirable for milling oat (**Figure 9**).

The GYT biplot approach has been adopted in multi-trait selection for various crops (Boureima and Abdoua, 2019; de Oliveira et al., 2019; Hamid et al., 2019; Mohammadi, 2019; Gouveia et al., 2020; Mahmoud et al., 2020; Merrick et al., 2020; Badu-Apraku et al., 2021; Sofi et al., 2021; Tsenov et al., 2021; Xu et al., 2021).

CONCLUSIONS

Plant breeding plays a key role in meeting the increasing need for food, fiber, health, and comfort and in combating the adverse impacts of the changing climate. Plant breeding consists of two stages: breeding population development and progeny selection. For cultivar development, population development is more important than progeny selection but has largely been neglected in the literature. Hence, a "complete breeder's equation" was presented, which contains three key parameters: the population mean, the population variability, and the achieved heritability under the multi-location, multi-year framework. The value of a breeding population is measured by both the population mean and the population variability. For progeny selection, the key is to improve the heritability, i.e., selection reliability. Three aspects were identified to improve heritability: utilizing repeatable GE through mega-environment analysis, accommodating unrepeatable GE by adequate testing, and adequate replication and adjusting for spatial variation. Procedures for mega-environment analysis include GGE + GGL biplot analysis and LG biplot analysis. Adequate testing includes estimation and use of an optimum number of years, locations, and replicates. Cultivar evaluation within a mega-environment should select for mean performance and select against instability, with GGE biplot analysis being a preferred graphical method. A stability index is meaningful only when combined with high mean yield. Adequate testing is more important than any stability analysis. Last but not least, cultivar development must consider multiple traits; both

independent culling and index selection are essential. GYT biplot analysis is a preferred method for index selection. In addition, genomic selection is an alternative and potentially more effective approach in all stages and aspects of cultivar development if reliable models are developed and if it can be done cost-efficiently.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary materials, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

ACKNOWLEDGMENTS

The sample dataset used in this article was obtained from Denis Marois, the coordinator of the Regroupement des gestionnaires et copropriétaires du Québec (RGCQ) trials. The author thanks Denis and all sponsors of the oat trials for making the data available.

REFERENCES

Alwala, S., Kwolek, T., McPherson, M., Pellow, J., and Meyer, D. (2010). A comprehensive comparison between Eberhart and Russell joint regression and GGE biplot analyses to identify stable and high yielding maize hybrids. *Field Crops Res.* 119, 225–230. doi: 10.1016/j.fcr.2010.07.010

Amira, J. O., Ojo, D. K., Ariyo, O. J., Oduwaye, O. A., and Ayo-Vaughan, M. A. (2013). Relative discriminating powers of GGE and AMMI models in the selection of tropical soybean genotypes. *Afr. Crop Sci. J.* 21, 67–73.

Annicchiarico, P. (2021). Breeding gain from exploitation of regional adaptation: an Alfalfa case study. *Crop Sci.* 61, 2254–2271. doi: 10.1002/csc2.20423

Arief, V. N., DeLacy, I. H., Crossa, J., Payne, T., Singh, R., Braun, H. J., et al. (2015). Evaluating testing strategies for plant breeding field trials: redesigning a CIMMYT international wheat nursery. *Crop Sci.* 55, 164–177. doi: 10.2135/cropsci2014.06.0415

Asins, M. J. (2002). Present and future of quantitative trait locus analysis in plant breeding. *Plant Breed.* 121, 281–291. doi: 10.1046/j.1439-0523.2002.730285.x

Atlin, G. N., Baker, R. J., McRae, K. B., and Lu, X. (2000). Selection response in subdivided target regions. *Crop Sci.* 40, 7–13. doi: 10.2135/cropsci2000.4017

Atlin, G. N., Kleinknecht, K., Singh, K. P., and Piepho, H. P. (2011). Managing genotype x environment interaction in plant breeding programs: a selection theory approach. *J. Indian Soc. Agricu. Statistics* 65, 237–247.

Badu-Apraku, B., Bankole, F. A., Ajayo, B. S., Fakorede, M. A. B., Akinwale, R. O., Talabi, A. O., et al. (2021). Identification of early and extra-early maturing tropical maize inbred lines resistant to *Exserohilum turcicum* in sub-Saharan Africa. *Crop Protect.* 139:105386. doi: 10.1016/j.cropro.2020.105386

Badu-Apraku, B., Oyekunle, M., Obeng-Antwi, K., Osuman, A. S., Ado, S. G., Coulibay, N., et al. (2012). Performance of extra-early maize cultivars based on GGE biplot and AMMI analysis. *J. Agric. Sci.* 150:473. doi: 10.1017/S0021859611000761

Baum, M., Lagudah, E. S., and Appels, R. (1992). Wide crosses in cereals. *Annu. Rev. Plant Biol.* 43,117–143. doi: 10.1146/annurev.pp.43.060192.001001

Bekele, W. A., Wight, C. P., Chao, S., Howarth, C. J., and Tinker, N. A. (2018). Haplotype-based genotyping-by-sequencing in oat genome research. *Plant Biotechnol. J.* 16, 1452–1463. doi: 10.1111/pbi.12888

Boureima, S., and Abdoua, Y. A. O. U. (2019). Genotype by yield*trait combination biplot approach to evaluate sesame genotypes on multiple traits basis. *Turk. J. Field Crops* 24, 237–244. doi: 10.17557/tjfc.655165

Burgueño, J., Cadena, A., Crossa, J., Banziger, M., Gilmour, A. R., and Cullis, B. (2000). *User's Guide for Spatial Analysis of Field Variety Trials Using ASREML*. Texcoco: CIMMYT.

Ceccarelli, S. (1989). Wide adaptation: how wide? *Euphytica* 40, 197–205.

Cobb, J. N., Juma, R. U., Biswas, P. S., Arbelaez, J. D., Rutkoski, J., Atlin, G., et al. (2019). Enhancing the rate of genetic gain in public-sector plant breeding programs: lessons from the breeder's equation. *Theor. Appl. Genet.* 132, 627–645. doi: 10.1007/s00122-019-03317-0

Comstock, R. E., and Moll, R. H. (1963). Genotype-environment interactions. *Stat. Gen. Plant Breed.* 982, 164–196.

Cooper, M., van Eeuwijk, F. A., Hammer, G. L., Podlich, D. W., and Messina, C. (2009). Modeling QTL for complex traits: detection and context for plant breeding. *Curr. Opin. Plant Biol.* 12, 231–240. doi: 10.1016/j.pbi.2009.01.006

Cooper, M., Voss-Fels, K. P., Messina, C. D., Tang, T., and Hammer, G. L. (2021). Tackling G× E× M interactions to close on-farm yield-gaps: creating novel pathways for crop improvement by predicting contributions of genetics and management to crop productivity. *Theor. Appl. Genet.* 134, 1625–1644. doi: 10.1007/s00122-021-03812-3

Crespo-Herrera, L. A., Velu, G., and Singh, R. P. (2016). Quantitative trait loci mapping reveals pleiotropic effect for grain iron and zinc concentrations in wheat. *Ann. Appl. Biol.* 169, 27–35. doi: 10.1111/aab.12276

Cullis, B., Gogel, B., Verbyla, A., and Thompson, R. (1998). Spatial analysis of multi-environment early generation variety trials. *Biometrics* 54, 1–18. doi: 10.2307/2533991

Cullis, B. R., and Gleeson, A. C. (1991). Spatial analysis of field experiments-an extension to two dimensions. *Biometrics* 47, 1449–1460. doi: 10.2307/2532398

de Oliveira, T. R. A., de Amaral Gravina, G., de Moura Rocha, M., de Alcântara Neto, F., da Cruz, D. P., de Oliveira, G. H. F., et al. (2019). GYT biplot analysis: a new approach for cowpea line selection. *J. Exp. Agric. Int.* 41, 1–9. doi: 10.9734/jeai/2019/v41i530408

DeLacy, I. H., Basford, K. E., Cooper, M., Bull, J. K., and McLaren, C. G. (1996). "Analysis of multienvironment trialsa historical perspective," in *Plant Adaptation and Crop Improvement*, eds M. Cooper and G. L. Hammer (Wallingford: IRRI/CABI), 39124.

Duvick, D. N. (1996). Plant breeding, an evolutionary concept. *Crop Sci.* 36, 539–548. doi: 10.2135/cropsci1996.0011183X003600030001x

Eberhart, S. A. (1970). Factors affecting efficiencies of breeding methods. *Afr. Soils* 15, 655–680.

Eberhart, S. T., and Russell, W. A. (1966). Stability parameters for comparing varieties 1. *Crop Sci.* 6, 36–40. doi: 10.2135/cropsci1966.0011183X000600010011x

Fehr, W. (1991). *Principles of Cultivar Development: Theory and Technique*. Stuttgart: Macmillian Publishing Company.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika* 58, 453–467. doi: 10.1093/biomet/58.3.453

Gauch, G. H. Jr., and Zobel, R. W. (1997). Identifying mega- environments and targeting genotypes. *Crop Sci.* 37, 311–326.

Gauch, H. G., Jr., Piepho, H. P., and Annicchiarico, P. (2008). Statistical analysis of yield trials by AMMI and GGE: further considerations. *Crop Sci.* 48, 866–889. doi: 10.2135/cropsci2007.09.0513

Gauch, H. G. Jr. (2006). Statistical analysis of yield trials by AMMI and GGE. *Crop Sci.* 46, 1488–1500. doi: 10.2135/cropsci2005.07-0193

Gauch, H. G. Jr. (2013). A simple protocol for AMMI analysis of yield trials. *Crop Sci.* 53, 1860–1869. doi: 10.2135/cropsci2013.04.0241

Gauch, H. G. Jr., and Zobel, R. W. (1988). Predictive and postdictive success of statistical analyses of yield trials. *Theor. Appl. Genet.* 76, 1–10.

Gilmour, A. R., Cullis, B. R., and Verbyla, A. P. (1997). Accounting for natural and extraneous variation in the analysis of field experiments. *J. Agric. Biol. Environ. Stat.* 2, 269–293. doi: 10.2307/1400446

Gilmour, A. R., Thompson, R., and Cullis, B. R. (1995). Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* 51, 1440–1450. doi: 10.2307/2533274

Goddard, M. E., and Hayes, B. J. (2007). Genomic selection. *J. Anim. Breed. Genet.* 124, 323–330. doi: 10.1111/j.1439-0388.2007.00702.x

Gouveia, B. T., Barrios, S. C. L., do Valle, C. B., da Costa Gomes, R., Machado, W. K. R., de Sousa Bueno Filho, J. S., et al. (2020). Selection strategies for increasing the yield of high nutritional value leaf mass in Urochloa hybrids. *Euphytica* 216, 1–12. doi: 10.1007/s10681-020-2574-3

Grondona, M. O., Crossa, J., Fox, P. N., and Pfeiffer, W. H. (1996). Analysis of variety yield trials using two-dimensional separable ARIMA processes. *Biometrics* 52, 763–770. doi: 10.2307/2532916

Hamid, A. E., Aglan, M. A., and Hussein, E. (2019). Modified method for the analysis of genotype by trait (Gt) biplot as a selection criterion in wheat under water stress conditions. *Egypt. J. Agron.* 41, 293–312. doi: 10.21608/agro.2019.16580.1177

Hanson, W. D., and Brim, C. A. (1963). Optimal allocation of test material for two-stage testing with an application to evaluation of soybean lines. *Crop Sci.* 3, 43–49.

Hao, Y., Velu, G., Peña, R. J., Singh, S., and Singh, R. P. (2014). Genetic loci associated with high grain zinc concentration and pleiotropic effect on kernel weight in wheat (*Triticum aestivum* L.). *Mol. Breed.* 34,1893–1902. doi: 10.1007/s11032-014-0147-7

Heffner, E. L., Sorrells, M. E., and Jannink, J. L. (2009). Genomic selection for crop improvement. *Crop Sci.* 49, 1–12. doi: 10.2135/cropsci2008.08.0512

Hizbai, B. T., Gardner, K. M., Wight, C. P., Dhanda, R. K., Molnar, S. J., Johnson, D., et al. (2012). Quantitative trait loci affecting oil content, oil composition, and other agronomically important traits in oat. *Plant Genome* 5, 164–175. doi: 10.3835/plantgenome2012.07.0015

Hoyos-Villegas, V., Wright, E. M., and Kelly, J. D. (2016). GGE biplot analysis of yield associations with root traits in a Mesoamerican bean diversity panel. *Crop Sci.* 56, 1081–1094. doi: 10.2135/cropsci2015.10.0609

Huehn, M. (1990). Nonparametric measures of phenotypic stability. Part 2: applications. *Euphytica* 47, 195–201.

Jannink, J. L., Lorenz, A. J., and Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. *Brief. Funct. Genomics* 9, 166–177. doi: 10.1093/bfgp/elq001

Kang, M. S. (1993). Simultaneous selection for yield and stability in crop performance trials: Consequences for growers. *Agron. J.* 85, 754–757. doi: 10.2134/agronj1993.00021962008500030042x

Kempton, R. A., Seraphin, J. C., and Sword, A. M. (1994). Statistical analysis of two-dimensional variation in variety yield trials. *J. Agric. Sci.* 122, 335–342. doi: 10.1017/S0021859600067253

Kharkwal, M. C., Pandey, R. N., and Pawar, S. E. (2004). "Mutation breeding for crop improvement," in *Plant Breeding*, eds H. K. Jain, M. C. Kharkwal (Dordrecht: Springer), 601–645. doi: 10.1007/978-94-007-1040-5_26

Lin, C. S., and Binns, M. R. (1988). A superiority measure of cultivar performance for cultivar× location data. *Canad. J. Plant Sci.* 68, 193–198. doi: 10.4141/cjps88-018

Lin, C. S., and Binns, M. R. (1994). Concepts and methods for analyzing regional trial data for cultivar and location selection. *Plant Breed. Rev.* 12, 271–297. doi: 10.1002/9780470650493.ch10

Lin, C. S., Binns, M. R., and Lefkovitch, L. P. (1986). Stability analysis: where do we stand? *Crop Sci.* 26, 894–900.

Ma, F., Xu, Y., Ma, Z., Li, L., and An, D. (2018). Genome-wide association and validation of key loci for yield-related traits in wheat founder parent Xiaoyan 6. *Mol. Breed.* 38, 1–15. doi: 10.1007/s11032-018-0837-7

Mahmoud, M. W., Hussein, E., Aboelkassem, K. M., and Ibrahim, H. E. (2020). Graphical presentation of some peanut genotypes by comparing two patterns of biplot analysis. *J. Plant Produ.* 11, 697–705. doi: 10.21608/jpp.2020.112895

McCann, L. C., Bethke, P. C., Casler, M. D., and Simon, P. W. (2012). Allocation of experimental resources to minimize the variance of genotype mean chip color and tuber composition. *Crop Sci.* 52, 1475–1481. doi: 10.2135/cropsci2011.07.0392

Merrick, L. F., Glover, K. D., Yabwalo, D., and Byamukama, E. (2020). Use of genotype by yield*trait (GYT) analysis to select hard red spring wheat with elevated performance for agronomic and disease resistance traits. *Crop Breed. Genet. Genomics* 2:e200009. doi: 10.20900/cbgg20200009

Mohammadi, R. (2019). Genotype by yield* trait biplot for genotype evaluation and trait profiles in durum wheat. *Cereal Res. Commun.* 47, 541–551. doi: 10.1556/0806.47.2019.32

Oliveira, T. R. A. D., Carvalho, H. W. L. D., Oliveira, G. H. F., Costa, E. F. N., Gravina, G. D. A., Santos, R. D. D., et al. (2019). Hybrid maize selection through GGE biplot analysis. *Bragantia* 78, 166–174. doi: 10.1590/1678-4499.20170438

Piepho, H. P., Büchse, A., and Emrich, K. (2003). A hitchhiker's guide to mixed models for randomized experiments. *J. Agron. Crop Sci.* 189, 310–322. doi: 10.1046/j.1439-037X.2003.00049.x

Qiao, C. G., Basford, K. E., DeLacy, I. H., and Cooper, M. (2000). Evaluation of experimental designs and spatial analyses in wheat breeding trials. *Theor. Appl. Genet.* 100, 9–16. doi: 10.1007/s001220050002

Rasmusson, D. C., and Phillips, R. L. (1997). Plant breeding progress and genetic diversity from de novo variation and elevated epistasis. *Crop Sci.* 37, 303–310. doi: 10.2135/cropsci1997.0011183X003700020001x

Rutkoski, J. E. (2019). A practical guide to genetic gain. *Adv. Agron.* 157, 217–249. doi: 10.1016/bs.agron.2019.05.001

Schmidt, P., Möhring, J., Koch, R. J., and Piepho, H. P. (2018). More, larger, simpler: how comparable are on-farm and on-station trials for cultivar evaluation? *Crop Sci.* 58, 1508–1518. doi: 10.2135/cropsci2017.09.0555

Shu, Q. Y., Forster, B. P., Nakagawa, H., and Nakagawa, H. (2012). *Plant Mutation Breeding and Biotechnology*. Oxford: CAB International. doi: 10.1079/9781780640853.0000

Simmonds, N., and Smartt, J. (1999). *Principles of Crop Improvement, 2nd Edn*. Oxford: Blackwell Science Ltd. Press.

Singh, R. K., and Chaudhary, B. D. (1977). *Biometrical Methods in Quantitative Genetic Analysis*. Ludhiana: Kalyani.

Sofi, P. A., Saba, I., Ara, A., and Rehman, K. (2021). Comparative efficiency of GY*T approach over GT approach in genotypic selection in multiple trait evaluations: case study of common bean (*Phaseolus vulgaris*) grown under temperate Himalayan conditions. *Agric. Res.* 1–9. doi: 10.1007/s40003-021-00577-5

Sprague, G. F., and Federer, W. T. (1951). A comparison of variance components in corn yield trials. II. Error, year × variety, location × variety and variety components. *Agron. J.* 43, 535–541.

Street, D. (1990). Fisher's contributions to agricultural statistics. *Biometrics* 46, 937–945. doi: 10.2307/2532439

Swallow, W. H., and Wehner, T. C. (1989). Optimum allocation of plots to years, seasons, locations, and replications, and its application to once-over-harvest cucumber trials. *Euphytica* 43, 59–68. doi: 10.1007/BF00037897

Tanksley, S. D. (1983). Molecular markers in plant breeding. *Plant Mol. Biol. Rep.* 1, 3–8.

Tinker, N. A., Bekele, W. A., and Hattori, J. (2016). Haplotag: software for haplotype-based genotyping-by-sequencing analysis. *G3 (Bethesda)* 6, 857–873. doi: 10.1534/g3.115.024596

Troyer, A. F. (1996). Breeding widely adapted, popular maize hybrids. *Euphytica* 92, 163–174. doi: 10.1007/BF00022842

Tsenov, N., Gubatov, T., and Yanchev, I. (2021). Genotype selection for grain yield and quality based on multiple traits of common wheat (*Triticum aestivum* L.). *Cereal Res. Commun.* 49, 119–124. doi: 10.1007/s42976-020-00080-7

van Harten, A. M. (1998). *Mutation Breeding: Theory and Practical Applications*. Cambridge: Cambridge University Press.

Wang, L., Zhu, G., Johnson, W., and Kher, M. (2018). Three new approaches to genomic selection. *Plant Breed.* 137, 673–681. doi: 10.1111/pbr.12640

Wricke, G., and Weber, E. (1986). *Quantitative Genetics and Selection in Plant Breeding*. Berlin, NY: Walter de Gruyter

Xu, N. Y., Zhao, S. Q., Zhang, F., Fu, X. Q., Yang, X. N., Qiao, Y. T., and Sun, S. X. (2021). Retrospective evaluation of cotton varieties nationally registered for the Northwest Inland cotton growing regions based on GYT biplot analysis. *Acta Agron. Sinica* 47, 660–671. doi: 10.3724/SP.J.1006.2021.04135

Yan, W. (2001). GGEbiplot—A Windows application for graphical analysis of multienvironment trial data and other types of two-way data. *Agron. J.* 93, 1111–1118. doi: 10.2134/agronj2001.9351111x

Yan, W. (2002). Singular-value partitioning in biplot analysis of multienvironment trial data. *Agron. J.* 94, 990–996. doi: 10.2134/agronj2002.0990

Yan, W. (2011). GGE biplot vs. AMMI graphs for genotype-by-environment data analysis. *J. Indin Soc. Agric. Stat.* 65, 181–193.
Yan, W. (2013). Biplot analysis of incomplete two-way data. *Crop Sci.* 53, 48–57. doi: 10.2135/cropsci2012.05.0301
Yan, W. (2014). *Crop Variety Trials: Data Management and Analysis*. Hoboken, NJ: John Wiley & Sons. doi: 10.1002/9781118688571
Yan, W. (2015). Mega-environment analysis and test location evaluation based on unbalanced multiyear data. *Crop Sci.* 55, 113–122. doi: 10.2135/cropsci2014.03.0203
Yan, W. (2016). Analysis and handling of G×E in a practical breeding program. *Crop Sci.* 56, 2106–2118. doi: 10.2135/cropsci2015.06.0336
Yan, W. (2019). LG biplot: a graphical method for mega-environment investigation using existing crop variety trial data. *Sci. Rep.* 9, 1–8. doi: 10.1038/s41598-019-43683-9
Yan, W. (2021). Estimation of the optimal number of replicates in crop variety trials. *Front. Plant Sci.* 11:590762. doi: 10.3389/fpls.2020.590762
Yan, W., and Frégeau-Reid, J. (2008). Breeding line selection based on multiple traits. *Crop Sci.* 48, 417–423. doi: 10.2135/cropsci2007.05.0254
Yan, W., and Frégeau-Reid, J. (2018). Genotype by yield∗ trait (GYT) biplot: a novel approach for genotype selection based on multiple traits. *Sci. Rep.* 8:8242. doi: 10.1038/s41598-018-26688-8
Yan, W., Frégeau-Reid, J., Martin, R., Pageau, D., and Mitchell-Fetch, J. (2015). How many test locations and replications are needed in crop variety trials for a target region? *Euphytica* 202, 361–372.
Yan, W., Frégeau-Reid, J., Mountain, N., and Kobler, J. (2019a). Genotype and management evaluation based on genotype by yield*trait (GYT) analysis. *Crop Breed. Genet. Genomics* 1:e190002. doi: 10.20900/cbgg20190002
Yan, W., Frégeau-Reid, J., Pageau, D., and Martin, R. (2016). Genotype-by-environment interaction and trait associations in two genetic populations of oat. *Crop Sci.* 56, 1136–1145. doi: 10.2135/cropsci2015.11.0678
Yan, W., Hunt, L. A., Sheng, Q., and Szlavnics, Z. (2000). Cultivar evaluation and mega environment investigation based on the GGE biplot. *Crop Sci.* 40, 597–605.
Yan, W., and Kang, M. S. (2002). *GGE Biplot Analysis: A Graphical Tool for Breeders, Geneticists, and Agronomists*. CRC press.
Yan, W., Kang, M. S., Ma, B., Woods, S., and Cornelius, P. L. (2007). GGE biplot vs. AMMI analysis of genotype-by-environment data. *Crop Sci.* 47, 643–653. doi: 10.2135/cropsci2006.06.0374
Yan, W., Mitchell-Fetch, J., Beattie, A., Nilsen, K. T., Pageau, D., DeHaan, B., et al. (2021). Oat mega-environments in Canada. *Crop Sci.* 61, 1143–1153. doi: 10.1002/csc2.20426
Yan, W., Pageau, D., Frégeau-Reid, J., Lajeunesse, J., Goulet, J., Durand, J., and Marois, D. (2011). Oat mega-environments and test-locations in Quebec. *Canad. J. Plant Sci.* 91, 643–649. doi: 10.4141/cjps10139
Yan, W., and Tinker, N. A. (2005). An integrated biplot analysis system for displaying, interpreting, and exploring genotype× environment interaction. *Crop Sci.* 45, 1004–1016. doi: 10.2135/cropsci2004.0076
Yan, W., and Tinker, N. A. (2006). Biplot analysis of multi-environment trial data: principles and applications. *Canad. J. Plant Sci.* 86, 623–645. doi: 10.4141/P05-169
Yan, W., Tinker, N. A., Bekele, W. A., Mitchell-Fetch, J., and Frégeau-Reid, J. (2019b). Theoretical unification and practical integration of conventional methods and genomic selection in plant breeding. *Crop Breed. Genet. Genomics* 1:e190003. doi: 10.20900/cbgg20190003
Yan, W., and Wallace, D. H. (1995). Breeding for negatively associated traits. *Plant Breed. Rev.* 13, 141–177. doi: 10.1002/9780470650059.ch4
Yang, R. C., Ye, T. Z., Blade, S. F., and Bandara, M. (2004). Efficiency of spatial analyses of field pea variety trials. *Crop Sci.* 44, 49–55. doi: 10.2135/cropsci2004.0049
Zhao, H. Z., Zhang, H. F., and Song, Z. M. (1981). Several key issues in wheat breeding. *Shaanxi Agric. Sci.* 3, 1–8. In Chinese.
Zobel, R. W., Wright, M. J., and Gauch, H. G. Jr. (1988). Statistical analysis of a yield trial. *Agron. J.* 80, 388–393. doi: 10.2134/agronj1988.00021962008000030002x

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 23 September 2021
doi: 10.3389/fpls.2021.739072

Transcriptome Reveals Allele Contribution to Heterosis in Maize

Jianzhong Wu[1,2], Dequan Sun[2], Qian Zhao[3], Hongjun Yong[1], Degui Zhang[1], Zhuanfang Hao[1], Zhiqiang Zhou[1], Jienan Han[1], Xiaocong Zhang[1], Zhennan Xu[1], Xinhai Li[1], Mingshun Li[1]* and Jianfeng Weng[1]**

[1] Institute of Crop Science, Chinese Academy of Agricultural Sciences, Beijing, China, [2] Institute of Forage and Grassland Sciences, Heilongjiang Academy of Agricultural Sciences, Harbin, China, [3] Institute of Crop Cultivation and Tillage, Heilongjiang Academy of Agricultural Sciences, Harbin, China

Heterosis, which has greatly increased maize yields, is associated with gene expression patterns during key developmental stages that enhance hybrid phenotypes relative to parental phenotypes. Before heterosis can be more effectively used for crop improvement, hybrid maize developmental gene expression patterns must be better understood. Here, six maize hybrids, including the popular hybrid Zhengdan958 (ZC) from China, were studied. Maize hybrids created in-house were generated using an incomplete diallel cross (NCII)-based strategy from four elite inbred parental lines. Differential gene expression (DEG) profiles corresponding to three developmental stages revealed that hybrid partial expression patterns exhibited complementarity of expression of certain parental genes, with parental allelic expression patterns varying both qualitatively and quantitatively in hybrids. Single-parent expression (SPE) and parent-specific expression (PSE) types of qualitative variation were most prevalent, 43.73 and 41.07% of variation, respectively. Meanwhile, negative super-dominance (NSD) and positive super-dominance (PSD) types of quantitative variation were most prevalent, 31.06 and 24.30% of variation, respectively. During the early reproductive growth stage, the gene expression pattern differed markedly from other developmental stage patterns, with allelic expression patterns during seed development skewed toward low-value parental alleles in hybrid seeds exhibiting significant quantitative variation-associated superiority. Comparisons of qualitative gene expression variation rates between ZC and other hybrids revealed proportions of SPE-DEGs (41.36%) in ZC seed DEGs that significantly exceeded the average proportion of SPE-DEGs found in seeds of other hybrids (28.36%). Importantly, quantitative gene expression variation rate comparisons between ZC and hybrids, except for transgressive expression, revealed that the ZC rate exceeded the average rate for other hybrids, highlighting the importance of partial gene expression in heterosis. Moreover, enriched ZC DEGs exhibiting distinct tissue-specific expression patterns belonged to four biological pathways, including photosynthesis, plant hormone signal transduction, biology metabolism and biosynthesis. These results provide valuable technical insights for creating hybrids exhibiting strong heterosis.

Keywords: heterosis, transcriptome, maize, allele, DEGs

OPEN ACCESS

Edited by:
Leonardo Abdiel Crespo Herrera, International Maize and Wheat Improvement Center, Mexico

Reviewed by:
Hui Wang, Shandong Academy of Agricultural Sciences, China
Xiangfeng Wang, University of Arizona, United States

***Correspondence:**
Xinhai Li
lixinhai@caas.cn
Mingshun Li
limingshun@caas.cn
Jianfeng Weng
wengjianfeng@caas.cn

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *10 July 2021*
Accepted: *06 September 2021*
Published: *23 September 2021*

Citation:
Wu J, Sun D, Zhao Q, Yong H, Zhang D, Hao Z, Zhou Z, Han J, Zhang X, Xu Z, Li X, Li M and Weng J (2021) Transcriptome Reveals Allele Contribution to Heterosis in Maize. Front. Plant Sci. 12:739072. doi: 10.3389/fpls.2021.739072

INTRODUCTION

Maize is an important crop that contributes significantly to the production of food and industrial raw materials around the world. Genetic improvement of maize mainly depended on its abundant genetic diversity and strong heterosis. Heterosis is a ubiquitous biological phenomenon in which hybrids exhibit superior performance relative to the biparental value of their parents for one or more traits. Researchers have been studying heterosis in depth from diverse perspectives. Environmental influences on heterosis (Duvick, 2005), which emphasize the environmental adaptability of plants to overcome the bottleneck of stresses. Combining ability effect tends to suggest that parental combining ability selection is the effective approach in hybrid breeding (Makumbi et al., 2011; Zhang et al., 2015). Genetic distance dominance (Wu et al., 2016) of heterosis suggests that the combination of combing ability tests and genetic relationship analysis can be used to define heterotic groups for breeding selection and genetic improvement. Changes in chromosome dose in multiple regions of the genome regulate quantitative traits, which reflect the influence of dose variation of chromosome dosage effects (Birchler et al., 2016). QTL interactions was used to analyze component accumulation of related quantitative traits could reasonably explain the genetic basis of part of heterosis (Xiang et al., 2016; Zhu et al., 2016). Allelic variation in gene expression might play a protective role in defense against adverse environments (Hoecker et al., 2008; Waters et al., 2017). "omics effects" can also be used to interpret heterosis at different levels (Huang et al., 2016; Alonso-Peral et al., 2017). Low DNA methylation was found in maize hybrids (where methylation inhibits gene expression), but not in rice as a whole, revealing the DNA methylation and epigenetic effects (Groszmann et al., 2013; Dapp et al., 2015). All of the hypotheses above attempt to reveal the genetic mechanisms controlling heterosis.

Although the exact molecular mechanisms behind heterosis remain unknown (Schnable and Springer, 2013), three competing, but not mutually exclusive, hypotheses propose that dominance effects, over-dominance effects, epistatic effects, or some combination thereof explain the genetic control of heterosis. Dominant hypothesis emphasizes the introduction of new ideal type dominant allele in the process of plant growth and development, the dominant allele more popular than the recessive allele (Davenport, 1908; Bruce, 1910; Jones, 1917). Over-dominance hypothesis emphasizes the complementarity between alleles of the offspring of hybrids, and holds that the heterozygous state is more favorable than the homozygous state (East, 1908; Shull, 1908; Larièpe et al., 2012). Epistatic interactions hypothesis emphasizes the favorable alleles in the role of heterosis gain. For any site, the effect can be produced by additive dominant or over-dominance (Minvielle, 1987; Schnell and Cockerham, 1992; Frascaroli et al., 2007). It should be noted that heterosis in self-pollinated species (e.g., rice) may involve different genetic interactions from heterosis in cross-pollinated species (e.g., maize), so both dominant and epistatic hypotheses may be related (Garcia et al., 2008).

Plant breeding is the aggregation of superior alleles from different germplasm sources. The aggregation of favorable alleles from different parents in the hybrids can exhibit much greater effects and may result in elite varieties (Cai et al., 2014). At present, intervarietal hybrid advantage is the main technical term used to describe unknown underlying mechanism(s) responsible for superior hybrid maize characteristics, with few mechanistic clues obtained from past studies that focused on maize germplasm-associated heterosis groups and patterns (Reif et al., 2003; Aguiar et al., 2008; Zhang et al., 2018; Annor et al., 2020). Nevertheless, early maize varieties cultivated in China possessed limited genetic diversity that has hindered attempts to substantially increase hybrid maize yields in recent decades. A large number of alleles were deleted by long-term artificial selection, caused losses in diversity as reflected by reduced allele numbers through elimination of unfavorable alleles (Gao et al., 2017). Indeed, the lack of adequate parental genetic variation has caused breeding efforts utilizing numerous heterosis groups to fail to improve grain yield, prompting researchers to investigate heterosis mechanisms associated with strong predominant hybrid maize.

The superior allele is the basis of the dominant expression in hybrids. Allelic specific expression in hybrids is one of the mechanisms of heterosis (Ma et al., 2021). Researchers have speculated that heterosis-associated gene expression may be influenced by multiple biological processes, including DNA sequence variation, gene copy number change, histone modification, transcription factor regulation and DNA methylation (Zarayeneh et al., 2017). In one study, large numbers of DEGs detected during differentiation of maize spikelets and florets were identified based on gene expression profile differences between hybrid Zhengdan 958 and its parental lines Zheng58 and Chang7-2. Intriguingly, this set of DEGs encoded transcription factors or enzymes involved in biosynthesis of stress-related metabolites, suggesting that such genes are key players in heterosis expression (Li et al., 2012). Meanwhile, altered parental gene expression patterns observed in hybrid offspring indicate that intergenerational differential gene expression can generate hybrid phenotypes that are either superior or inferior to parental phenotypes (Stupar et al., 2008). Most DEGs showed inconsistent expression in tissues, which suggested extensive variation in the regulation of gene expression in a tissue-specific manner, about 97% of the single-parent expressed (SPE) genes exhibited intermediate or transgressive expression in hybrids, which might provide a wide range of opportunities for hybrid complementation through heterosis (Zhou et al., 2019). From the perspective of genome composition, genes of hybrid originate from these of parental inbred lines, thus, changes in gene expression and regulation in hybrid offspring must be responsible for heterosis, as demonstrated in previous studies that compared gene expression profiles between

Abbreviations: QTLs, quantitative trait loci; DEGs, differentially expressed genes; PCE, parental co-silence expression; PSE, parent-specific expression; HSE, hybrid-specific expression; SPE, single-parent expression; PAVs, presence-absence variations; PHCE, parent-hybrid co-expression; PSE, parent-specific expression; NSD, negative super-dominance; ND, negative dominance; PND, partial negative dominance; MP, mid-parent; PPD, partial positive dominance; PD, positive dominance; PSD, positive super-dominance.

hybrids and parent lines (Guo et al., 2004; Harrison et al., 2012; Paschold et al., 2012, 2014; Baldauf et al., 2016). However, studies comparing gene expression patterns among hybrids have not been reported.

The specific expression pattern of superior alleles is the molecular basis of heterosis to achieve. Hybrids can selectively express beneficial alleles under specific spatiotemporal conditions, thus showing superiority (Shao et al., 2019). Recently, transcriptomic analysis has served as a valuable tool for revealing regulatory mechanisms underlying differential expression of parental alleles and DEGs in hybrids vs. parental lines (Fu et al., 2015) and effects of parental genetic variations on expression of genes in hybrids (Hu et al., 2016; Li et al., 2016). Differential gene expression in hybrids may be responsible for heterosis, all possible patterns of gene action supports the hypothesis that multiple molecular mechanisms contribute to heterosis (Swanson-Wagner et al., 2006). The expression level of parental alleles in rice hybrids is unbalanced, which is considered to be one of the important mechanisms leading to heterosis (Shao et al., 2019; Lv et al., 2020). Therefore, the study of gene expression patterns in maize hybrids will play an important role in elucidating the heterosis mechanism. Although many hypotheses have been proposed regarding gene expression between parents and progeny that may reasonably explain one aspect of heterosis, but few studies investigating heterosis intensity of hybrids have been reported.

Regulation of gene expression runs through the whole development stage, but the significance of gene expression varies in stages of crop development. Many genes with specific functions are expressed only at a certain stage of plant growth and development, and their abnormal expression may lead to the generation of new phenotypes (Ferrándiz et al., 2000). Therefore, the study of gene expression changes in time and space is an important way to understand the mechanism of plant growth and development. Leaf, the most important site of photosynthesis in plants, is an essential source-to-sink sugar transport organ needed to sustain the entire reproductive process including grain yield. Indeed, leaf enlargement is a highly important characteristic of maize hybrids (Swanson-Wagner et al., 2006; Frascaroli et al., 2007). Maize spike differentiation, during which the plant meristem gradually shifts from vegetative to reproductive growth, is another key stage affecting maize yield. During maize spike differentiation, paired axillary spikelet meristem and floret meristem, which determine the number of rows and grains per ear, respectively, form and ultimately affect maize grain yield (Bommert et al., 2005; Pautler et al., 2013). Thus, leaves during the five-leaf stage, ears during the spikelet differentiation stage and seeds at 15 days post-pollination represent important transitions from maize vegetative to reproductive growth and are thus of great significance for breeding of high-yield maize.

Zhengdan958, a commercially successful hybrid variety created by Chinese maize breeding programs from parental inbred lines Zheng58 and Chang7-2 was studied here due to its expression of strong heterosis with regard to grain yield (Li et al., 2012; Kogelman et al., 2014; Yang et al., 2020). Over the past 10 years, this hybrid has been planted across China and is prized for its high and stable yield, high quality, resistance to high planting density, broad adaptability and resistance to multiple diseases and pests (Ma et al., 2018; Zhang et al., 2018). In this study, four typical maize inbred parental lines and six hybrids created from them using an incomplete diallel cross (NCII) strategy followed by comprehensive mRNA sequencing were used to investigate the effect of gene expression abundance of parental lines on hybrids, the variation of gene expression pattern among hybrids, and the specific expression gene and their pattern of Zhengdan958. Ultimately, results of this work should enhance our understanding of heterosis mechanisms so that heterosis can be harnessed to improve maize and other crops.

RESULTS

Identification of Gene Expression Abundance in Four Parental Lines

Numbers of differentially expressed genes (DEGs) were similar between hybrids and inbred lines, with ear (E) the source of the highest number of DEGs and leaf (L) and seed (S) yielding basically similar numbers of DEGs (**Supplementary Figure 1**). In particular, the number of DEGs in inbred parental line C seed was significantly lower than corresponding numbers found in other inbred parental lines. Notably, differences in transcript abundance between inbred parental lines for a given hybrid could be characterized as one of three difference types: expression abundance of the parental allele from one inbred line was greater than (>) that of the other, was less than (<) that of the other or was equal (=) to that of the other (**Figure 1**). In the present study, numbers of detected DEGs based on higher or lower mRNA-level abundance in leaf and ear organs of pairs of inbred parental maize lines were relatively similar. However, for some pairs of inbred lines as Z and C, C and M, C and Q, special ratios of parental gene expression levels in hybrid seeds relative to other tissues were observed as follows: $N_{Z>C}$:$N_{Z<C}$ ≈ 3:1, $N_{C>M}$:$N_{C<M}$ ≈ 1:3, $N_{C>Q}$:$N_{C<Q}$ ≈ 1:3. Thus, differences in allelic expression abundance between superior inbred parental lines and other lines suggests that superior inbred parental lines provide a more diverse and extensive selection background that supports a greater range of variation of gene expression in hybrid offspring. We might also infer from these results that hybrids with dominant gene action for a given trait are only produced when transcript abundance of genes derived from the superior inbred parental line reaches a certain proportion of the total transcript abundance derived from genes of both parental lines.

Partial Parental Expression of Alleles in Six Hybrids

Notably, no significant differences were found among total numbers of genes expressed among different hybrids (**Supplementary Figure 1**). However, presence-absence variations (PAVs) of DEGs that were detected in the four parental lines appeared as qualitative differences in hybrid differential gene expression profiles, with PAVs accounting for 19.18% of all differences detected in hybrid transcript expression

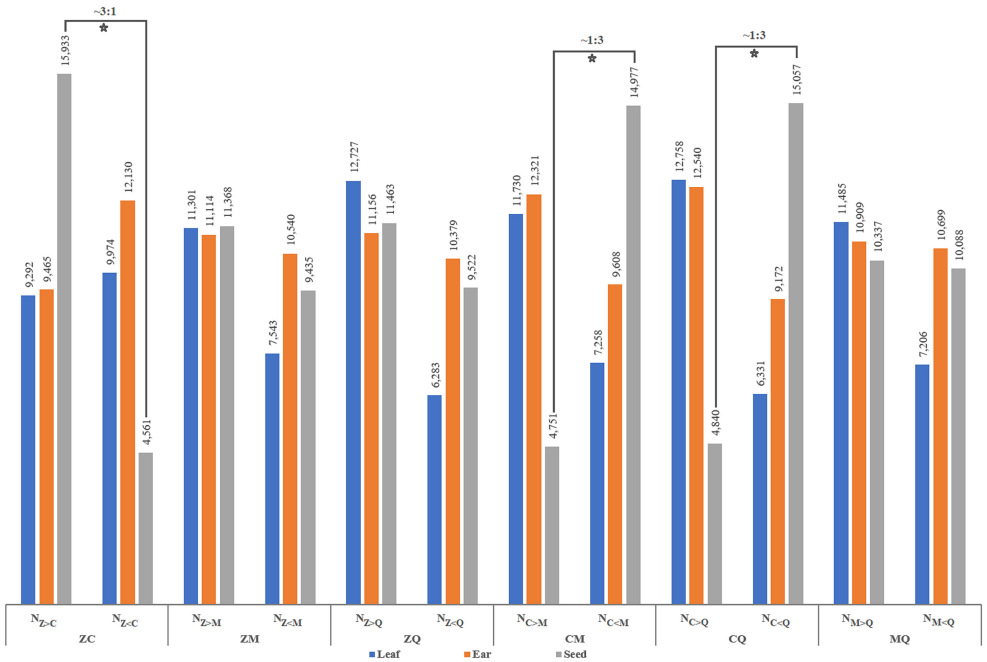

FIGURE 1 | Variation in the number of DEGs between parental inbred lines. $Nz_{>}c$ denotes the number of genes whose value of FPKM is greater in parent Zheng58 than in parent Chang7-2, and the same goes for other acronyms. The numbers at the top of the bar chart represent the number of different types of DEGs, the hollow five-pointed stars show special proportions of the number of genes, and the tissues [leaves (L), ears (E), and seeds (S)] shown in different colors are listed on bottom.

patterns. Of all PAVs, single-parent expression (SPE) and parent-specific expression (PSE) variations, the most extreme types of partial expression patterns involving high- or low-value parent alleles, were detected in DEGs in greatest proportions of 43.73 and 41.07%, respectively (**Figure 2A**). Expression patterns of most DEGs (80.82%) were of the parent-hybrid co-expression (PHCE) type (**Supplementary Table 2**). Among PHCE-type DEGs, we observed two predominant transgressive expression patterns, negative super-dominance (NSD) and positive super-dominance (PSD), with NSD and PSD proportions of all hybrid DEGs found to be 31.06 and 24.30%, respectively (**Figure 2B**). Therefore, we conclude that partial parental expression of alleles, which allele expressed biased toward to one of their parents, was responsible for qualitative variation in gene expression, while PHCE-type variations led to quantitative variation in gene expression. We thus presume that PHCE-type variations could be the main driving factor of heterosis expression in maize.

Partial parental expression of parental alleles in tissues of hybrids was consistent, but strength of partial expression was tissue-specific. With regard to qualitative variation, the number of SPE-DEGs with high parental partial expression exceeded the number of PSE-DEGs with low parental partial expression by 6.57%. Specifically, numbers of SPE-DEGs in leaves and ears were 50.85 and 45.12% greater, respectively, than corresponding numbers of PSE-DEGs, while in seeds the number of SPE-DEGs was only 61.60% of the number of PSE-DEGs. With regard to quantitative variation, the number of genes with expression biased toward that of the low-value parent (NSD-, ND-, and PND-DEGs) in hybrids was 26.52% higher than the number of genes whose expression was biased toward that of the high-value parent (PSD-, PD-, and PPD-DEGs). Notably, hybrids with biased gene expression toward alleles of the low-value parent exhibited significantly superior levels of quantitative variation.

Ultimately, both qualitative and quantitative variations were found to be associated with specific gene expression patterns in seeds as compared to leaves and ears, such as high proportions of PCE-, PSE- and NSD-DEGs and lower proportions of SPE-, PSD-, PD-, and PPD-DEGs in seeds as compared to corresponding proportions found in leaves and ears. Among qualitative variants, expression of alleles in hybrid seeds tended to be biased toward expression of low-value parent (PSE) alleles or was completely silent (PCE). However, with regard to quantitative variation, the number of NSD-DEGs in seeds was about 2.5 times the average number detected in leaves and ears, while the combined number of PSD-, PD- and PPD-DEGs in seeds was only 32.99% of the average number of DEGs in leaves and ears. Meanwhile, the number of DEGs exhibiting mid-parent allelic expression patterns was only about half of DEG numbers for other stages. Ultimately, hybrid gene expression patterns associated with the early reproductive growth stage significantly differed from patterns associated with other stages and were more highly influenced by alleles with partial to low parental expression patterns than by alleles with high parental expression.

Specific Patterns of Gene Expression in the Zhengdan958

To further clarify genes associated with various expression patterns in hybrids, we compared expression profiles of various

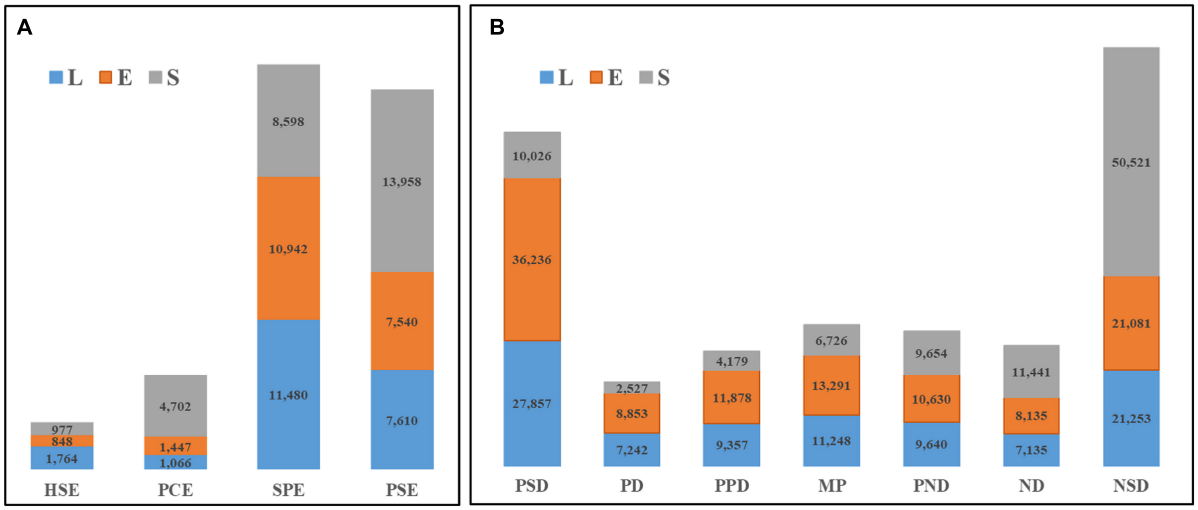

FIGURE 2 | Gene expression pattern in hybrids. Qualitative changes (PAVs) and quantitative changes (PHCE) in gene differential expression were represented in panels **(A,B)**, respectively. The numbers in the bars represent the DEGs in different gene expression types (horizontal classification), and the different colors indicate that each tissue part is listed in the figure.

genes exhibiting PAV and PHCE expression patterns in different tissues of ZC and other hybrids. ZC seeds yielded the highest number of SPE-DEGs and lowest number of PCE-DEGs, in opposition to gene expression patterns of hybrids overall. Although DEGs in ZC seeds with PAV-type expression patterns did not have a significant quantitative advantage, the total number of SPE-DEGs was highest, comprising 41.36% of all PAVs, a proportion that was 58.55% higher than the average proportion of SPE-DEGs for all other hybrids (28.36%). However, the proportion of ZC seed PCE-DEGs was only 45.60% of the overall average proportion for other hybrids (**Figure 3A**), while ZC seed PHCE-type expression trends were opposite to overall gene expression trends in hybrids. Specifically, lowest numbers of NSD-DEGs were found in ZC seeds, which comprised only 51.12% of the average number of NSD-DEGs detected in all other hybrids. Meanwhile, ND-DEGs, PND-DEGs, MP-DEGs, PPD-DEGs and PD-DEGs numbers in ZC seeds significantly exceeded average numbers in other hybrid seeds by 39.88, 105.08, 124.47, 94.42, and 60.14%, respectively, (**Figure 3B**). No other notable associations were found in leaves and ears in this study except for a result indicating the number of PSD-DEGs in ears of ZC was only 2/3 that of ears of other hybrids (**Supplementary Figure 2**). Therefore, we conclude that unique distribution patterns of DEGs in ZC are highly relevant to ZC grain yield-related heterosis.

Zhengdan 958-Specific Genes and Their Enrichment Pathway

To determine factors unique to ZC as the dominant hybrid, we identified 90, 118 and 137 ZC-specific DEGs in leaves (**Figure 4A**), ears (**Figure 4B**) and seeds (**Figure 4C**), respectively, that were not co-expressed in different tissues (**Supplementary Figure 3**). Details of ZC-specific genes are listed in **Supplementary Table 3**. The results indicated that these ZC-specific genes were expressed exclusively in different tissues, with synergistic expression of these genes forming the molecular basis for ZC comprehensive superiority relative to other hybrids.

In leaves, 33 HSE-DEGs, 48 SPE-DEGs and 9 PHCE-DEGs were identified, including 4, 1, 1, 2 and 1 DEGs with NSD-, ND-, PND-, MP- and PSD-type expression patterns, respectively. Among these DEGs, one HSE-DEG, Zm00001d024372, showed significant expression activity (**Figure 4D**) and was found to encode chloroplastic chlorophyll a-b binding protein. Thus, we inferred that Zm00001d024372 was an important ZC leaf protogene. Results of subsequent functional enrichment analyses indicated that specific ZC leaf DEGs were mainly enriched in biological pathways plant hormone signal transduction, MAPK signaling pathway-plant and photosynthesis—antenna proteins (**Figure 4G**). In ears, 29 HSE-DEGs, 62 SPE-DEGs and 27 PHCE-DEGs, including 9, 6, 2, 3, 1 and 6 DEGs with NSD-, ND-, PND-, PPD-, PD- and PSD-type expression patterns (**Figure 4E**), were enriched for photosynthesis, glutathione metabolism and brassinosteroid biosynthesis pathways (**Figure 4H**). In seeds, 43 HSE-DEGs, 80 SPE-DEGs and 14 PHCE-DEGs, including 7, 4, 1 and 2 DEGs with NSD-, ND-, MP- and PSD-type expression, were identified (**Figure 4F**). These DEGs were mainly enriched for pathways plant hormone signal transduction, arginine biosynthesis, glutathione metabolism and other biological pathways (**Figure 4I**). Taken together, these results indicate that ZC-specific DEGs were significantly and generally enriched for multiple biological pathways that included photosynthesis, plant hormone signal transduction, biology metabolism of glutathione, tryptophan, and others and biosynthesis of tropane, piperidine, pyridine alkaloid and diterpenoid (**Figure 5B**).

Gene Ontology (GO) pathway analysis showed that functions of these genes were mainly related to two of the three main GO categories (**Supplementary Table 4**). In the category of

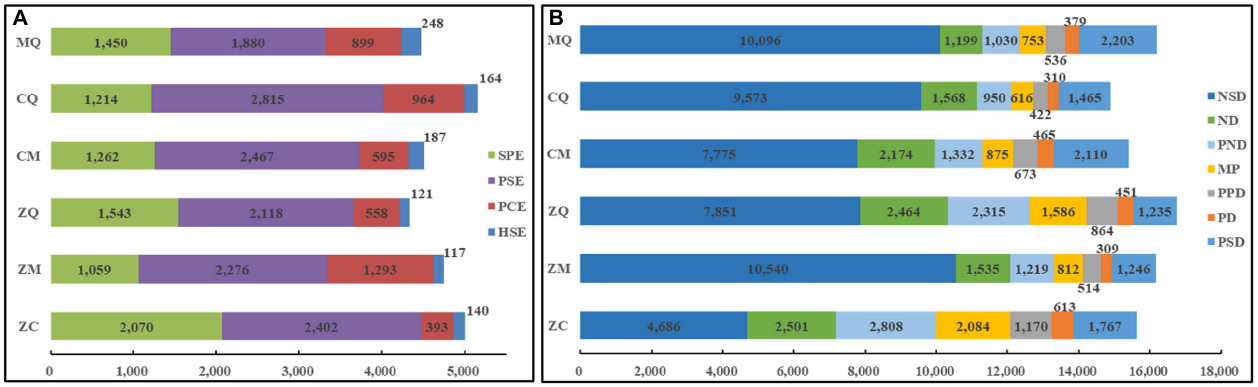

FIGURE 3 | Gene expression patterns in seeds of hybrids. The differential expression of DEGs in seeds of hybrids with PAV **(A)** and PHCE patterns **(B)**, where the expression quantity of different types of genes is marked on the bar, and different colors in the figure represent different types of DEGs listed on the right side of the figure. The abscissa is the number of differentially expressed genes.

biological process, genes detected within ZM00001D000361 (myb115), Zm00001d019230 (sid1), Zm00001d051465 (zag5), and ZM00001D053895 (bhlh51) play important roles in the morphogenesis of floral organs, such as the anther and anther wall tapetum. In the category of molecular function, the protein product of ZM00001D028264 participates in lipid IVA biosynthesis due to its UDP-3-*O*-[3-hydroxymyristoyl] *N*-acetylglucosamine deacetylase activity, the protein product of ZM00001D039634 (d1) participates in gibberellin biosynthesis due to its gibberellin 3-beta-dioxygenase activity, the protein product of Zm00001d036535 (oec33) has oxygen evolving activity, the protein product of Zm00001d048593 (rca1) has ribulose-1,5-bisphosphate carboxylase/oxygenase activator activity and the protein product of ZM00001D042122 participates in quercetin sulfate biosynthesis due to its brassinosteroid sulfotransferase activity (**Figure 5A**). Taken together, all of these protogene-associated enrichment pathways contribute to the overall dominance of ZC.

Among expression patterns of DEGs specifically expressed in each hybrid within the subset of PAV-type DEGs, SPE-DEGs showed a quantitative expression advantage over HSE-DEGs, while no notable variant-related expression advantage was found for PHCE-type DEGs (**Supplementary Figure 4**). These results thus suggest that impacts of advantageous effects due to dominant expression of PAV-type DEGs exceeded effects of co-dominant expression by hybrid-specific DEGs.

DISCUSSION

Comparisons of Transcript Abundance Among Inbred Lines Reveal Parental Allele-Associated Complementation Effects

Differential gene expressions are thought to play important roles in plant phenotypic development (Wallace et al., 2014), with intergenerational differential gene expression producing hybrid phenotypes that are either superior or inferior to parental phenotypes (Stupar et al., 2008). Furthermore, enrichment in hybrids of advantageous DEGs associated with certain parental alleles may explain superior hybrid qualities (Shao et al., 2019). Given the strong correlation between differential gene expression and hybrid performance, hybrid viability can be predicted by examining transcriptional activity at parental level (Thiemann et al., 2010). Intriguingly, here the number of DEGs varied significantly in different tissues, with greater numbers of genes found to be expressed in young ears in both hybrids and inbred parental lines, while numbers of expressed genes were similar in leaves and grains (**Supplementary Figure 1**). This result suggests that differences in gene expression abundance of parental inbred lines benefits the hybrid when the total numbers of DEGs are consistent or nearly consistent among parents and hybrid offspring. Notably, here we discovered that differences in transcript abundance between parental inbred lines and hybrid offspring varied among hybrids. Specifically, in seeds of dominant hybrids, the proportion of expressed genes was skewed to represent greater expression abundance of transcripts from one parent (**Figure 1**). This "asymmetry" of gene expression abundance between parental inbred lines may support a wider range of gene expression complementarity in hybrid offspring that can appear as heterosis. Therefore, we support the idea that specific expression of different alleles leads to a broader plasticity of gene expression (Shao et al., 2019).

Tissue-Specific Expression of Alleles Provides Opportunities for Heterosis at Various Stages

Tissue-specific genes refer to genes specifically expressed in different tissues that generate products conferring plants with specific morphological and structural characteristics and physiological functions. Tissue-specific gene expression can lead to phenotypic variation, especially in the case of dominant gene expression in hybrid tissues and organs that endows hybrids with special traits (Zhou et al., 2019). In this study, trends of

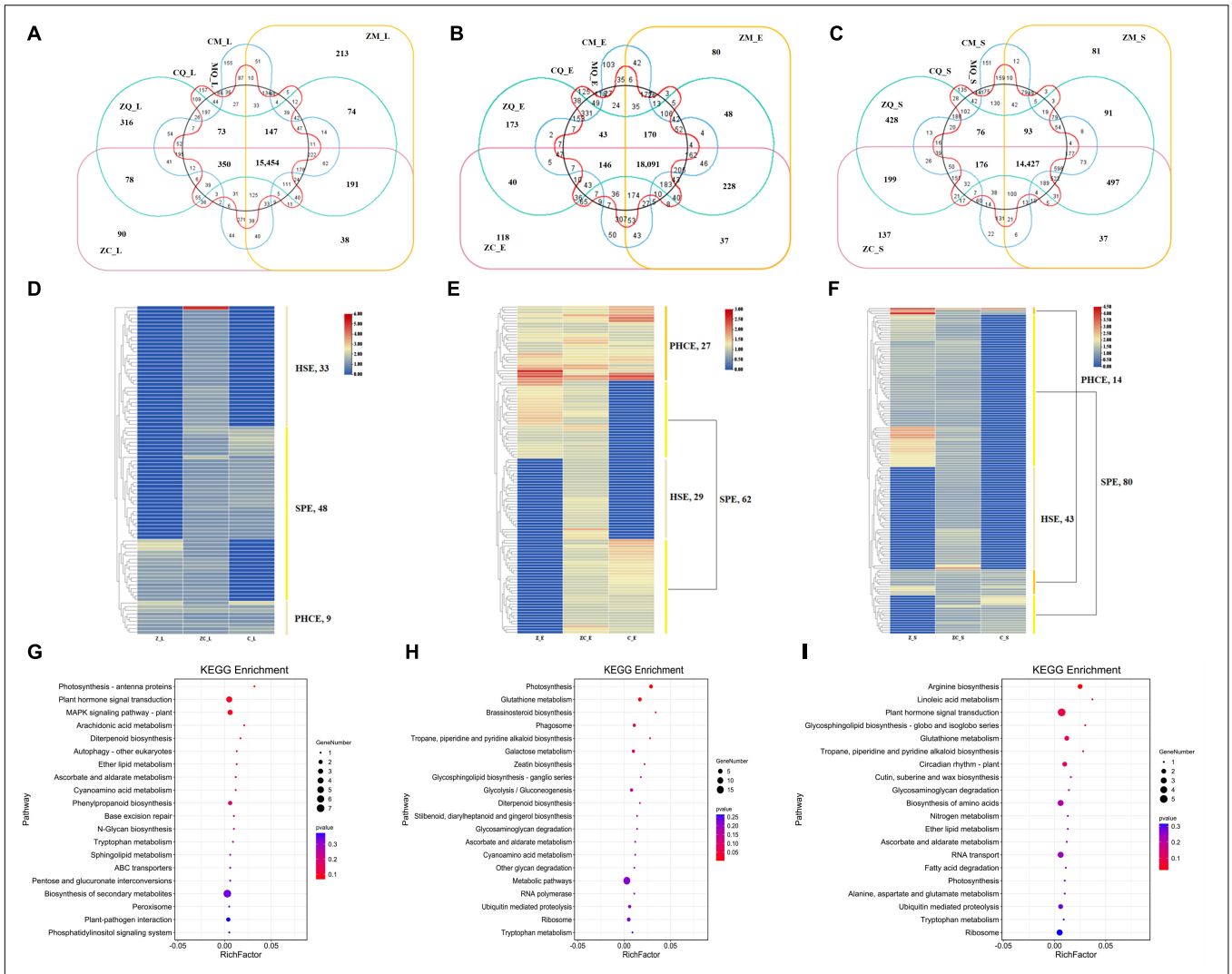

FIGURE 4 | Gene expression specificity of Zhengdan958 in different tissues. The DEGs of Zhengdan958-specific genes in leaves **(A)**, ears **(B)** and seeds **(C)**, and their expression patterns in leaves **(D)**, ears **(E)** and seeds **(F)** and the enrichment biological pathways in leaves **(G)**, ears **(H)** and seeds **(I)**. The Venn diagrams **(A–C)** show the distribution of DEGs of Zhengdan958-specific genes, and the overlapping regions represent the genes that are co-expressed among/between hybrids. The transition of colors from red to blue in heat maps **(D–F)** shows that the expression abundance of DEGs changes from high to low, the color bands on the right indicate different gene expression types, and the number of genes is listed near the type names in the figures. The size of dots in the bubble charts **(G–I)** represents the number of genes, and the color of the dots representing the *P*-value which listed in the figures.

variations of DEGs numbers among different tissues of hybrids and inbred parental lines were consistent (**Supplementary Figure 1**). Meanwhile, DEGs numbers varied significantly in different tissues, with numbers of genes expressed in ears significantly exceeding numbers expressed in leaves and seeds. This result could reasonably be attributed to the fact that switching of maize plants from vegetative growth to reproductive growth requires coordinated co-expression of a large number of genes.

In general, characterizations of tissue-specific expressed genes have improved our understanding of relationships between gene expression and tissue development (Xiao et al., 2010). Here, gene expression patterns differed significantly among different tissues. Regardless of whether the entire overall gene expression pattern of a hybrid resulted from qualitative variation or quantitative variation, significantly lower numbers of genes in hybrids exhibited expression patterns that were biased toward expression patterns of high-value parental genes than were biased toward expression patterns of low-value parental genes. We also found that hybrid seed-associated expression patterns significantly differed from expression patterns found in other tissues of hybrids, with numbers of PCE-DEGs, PSE-DEGs and NSD-DEGs detected in seeds that were respectively four times, two times and 2.5 times greater than corresponding numbers in ears or leaves. Meanwhile, numbers of PSD-DEGs, PD-DEGs, PPD-DEGs and MP-DEGs in seeds were respectively about one-third, one-third, one-half and one-half corresponding numbers associated with ears or leaves. Therefore, analysis of tissue-specific expression of

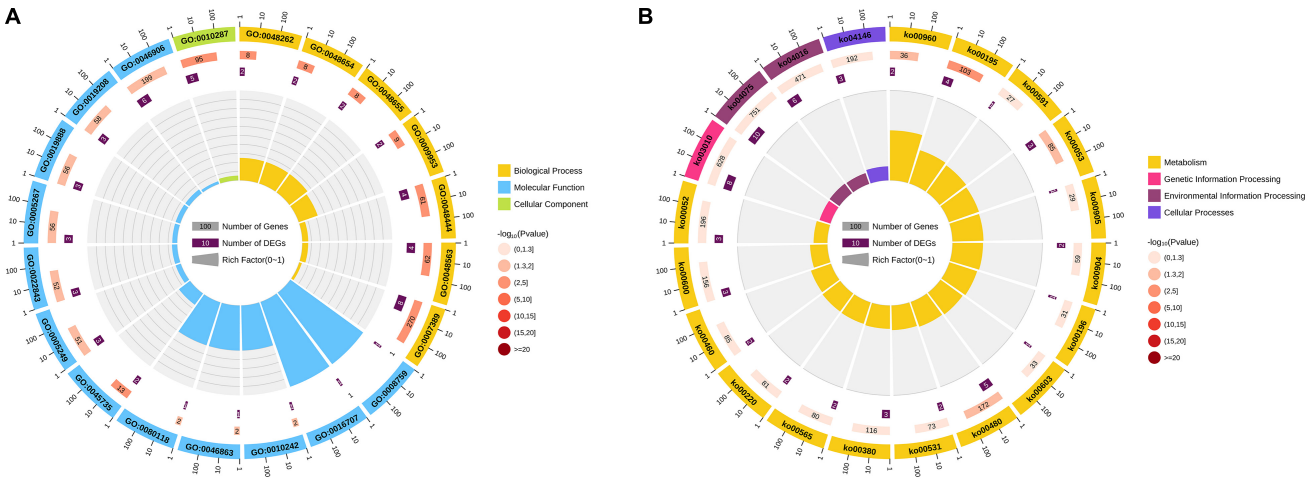

FIGURE 5 | Enrichment circle diagram of Zhengdan958-specific genes. The Top-20 pathway circles of significant enrichment difference were drawn on all the Zhengdan958-specific genes based on GO **(A)** and KEGG **(B)** enrichment pathways. Four circles in the circle diagram from outside to inside. The first circle is the enriched classification, and outside the circle is the scale of the number of genes, different colors represent different classifications. The second circle is the number of background genes in this category, the more genes the longer the bar, the smaller the value the redder the color. The third circle shows the total number of target genes. Rich factor value of each classification in the fourth circle.

DEGs in seeds, which represent the early reproductive growth stage, revealed underlying molecular mechanisms for heterosis in maize tissues that influenced grain yield, an observable trait. Validation of these results awaits further investigation.

Gene Expression Patterns of Allele Confers Heterosis Intensity Changes

A plethora of studies based on comparisons of differential gene expression profiles between maize hybrids and their parental inbred lines have elucidated mechanisms responsible for maize hybrid vigor. In one study, about 8.9–15.3% of genes were differentially expressed between hybrid and parental inbred lines in maize embryos at 6 days post-pollination (Meyer et al., 2007). In another study, Stupar and Springer found that F_1 progeny of inbred lines B73 and Mo17 exhibited significant differences in gene expression during seedling and young spike stages and in embryonic tissue as compared with expression patterns in parents (Stupar and Springer, 2006). In yet another study of the Zhongdan 808 hybrid, DEGs derived from parent NGS exhibited greater relative transcript abundance increases than did DEGs derived from parent CL11 (Hu et al., 2016), although these results were not related to the degree of heterosis, as studied in this work.

As mentioned above in the introduction section, we confirmed that the overall superior performance of ZC has led to large-scale ZC planting in China. To explore molecular mechanisms at the gene transcriptional level that are responsible for ZC superiority relative to other hybrids, we investigated ZC transcriptional gene expression patterns and compared them to expression patterns of other hybrids. Due to our main interest in grain yield-related heterosis, here we focused on ZC seed gene expression, not gene expression in leaves and ears, in order to study yield-related heterosis. Notably, ZC hybrid gene expression trends with regard to both qualitative and quantitative types of variation were diametrically opposed to trends observed for other hybrids, especially for expression trends in seeds (**Figure 3** and **Supplementary Figure 2**).

Differential gene expressions exhibiting SPE- and PCE-type expression patterns comprised the highest and lowest proportions, respectively, of DEGs identified in ZC seeds and respectively exhibited greatest biases toward expression of high-value and low-value parental alleles. Of PAV-type DEGs, SPE-DEGs comprised the highest proportion and PCE-DEGs comprised the lowest proportion of average hybrid DEGs. Such gene expression trends may provide clues to understanding why PAV-type DEGs are predominant in hybrid ZC. Importantly, NSD-type variation was the least prevalent type of variation observed for quantitative variants, while all other expression pattern types, except for those with an overdominance expression pattern, showed obvious quantitative advantages in ZC seeds. Ultimately, preferential expression in ZC seeds of high-value parental genes exhibiting qualitative variation played an important role in dominant expression of the hybrid. Meanwhile, quantitative-type expression of genes of the ultra-low-value parent (such as NSD) were lowest, thus illustrating that preferential expression of superior alleles derived from high-value parents played an important role development of hybrid advantage, which was consistent with the previous research in rice (Huang et al., 2015). Thus, overall gene expression patterns in hybrids in combination with comparative advantages should be helpful for guiding future breeding programs.

Enrichment of Superior Specific Alleles Promotes Heterosis

Differential gene expressions are thought to play important roles during plant phenotypic development (Wallace et al., 2014).

Furthermore, connections among advantageous alleles of parental genes have been proposed to explain hybrid vigor (Shao et al., 2019). In other words, higher levels of parental gene expression in hybrids could be due to enrichment of advantageous DEGs. The most direct strategy for identifying differences at the molecular level between ZC and other hybrids would be to screen for ZC-specific DEGs within the set of DEGs of all hybrids. Here, we analyzed expression patterns and enrichment pathways associated with ZC-specific DEGs detected in different tissues (**Figure 4**). The number of DEGs with expression patterns aligning with the PAVs model exceeded the number aligning with the PHCE model, regardless of whether the DEGs were unique to ZC or other hybrids. While analyzing expression pattern of ZC-specific genes, we assumed that these genes were already expressed in ZC and therefore were not HSE-type and PCE-type genes. Obviously, DEGs with PAV type expression patterns comprised the largest proportion of genes with altered expression associated specifically with hybrids, including the ZC.

It is noteworthy that a tissue-specific gene expressed in leaves (Stelpflug et al., 2016; Hoopes et al., 2019), chloroplastic chlorophyll a-b binding protein-encoding gene (ZM00001D024372, Lhca1), was identified as having significantly higher activity in ZC tissues (**Figure 4D**). This finding was important in that light-harvesting protein complexes (LHC I) form proteins encoded by Lhca1 and other genes (Lhca2, Lhca3, Lhca4, etc.) to trap sunlight within photosystem I (PSI), a key photosynthetic system in higher plants (Qin et al., 2015). Our suggestion is that Lhca1 should be targeted as a potentially useful gene source as part of future maize molecular breeding strategies to boost grain yield of maize.

In biological systems, various genes perform biological functions in coordination with one another. Coordinated gene functions associated with DEGs can be elucidated using KEGG pathway analysis to identify pathways with significantly enriched DEGs representation as compared to the genomic background. Here we found that enriched pathways associated with ZC-specific DEGs varied among different tissues, including plant hormone signal transduction in leaves, glutathione metabolism in ears, and arginine biosynthesis in seeds, all of which were key nodes of interconnected ZC-specific key biological pathways. Obviously, the enrichment of protogenes gradually shifted from photosynthetic signal to regulation of plant hormones and then to biosynthesis of amino acids at different developmental stages of maize from vegetative growth center to reproductive growth center. Overall, enrichment of ZC-specific DEGs was associated with pathways related to photosynthesis, glutathione metabolism and plant hormone signal transduction contributed to comprehensive phenotypic expression associated with heterosis. These results were validated by GO enrichment analysis, in which numerous genes were found to be functionally associated with biological processes related to biological regulation and response to stimulus and to molecular functions associated with binding and catalytic activity. Thus, superior alleles accumulate in the process of biosynthesis improvement, thereby promotes heterosis.

CONCLUSION

Heterosis has greatly boosted maize production and benefited human populations by providing greater food and economic security. Nevertheless, hybrid heterosis effects for a given combination of inbred parental lines cannot yet be predicted. In this study, six hybrids generated *via* incomplete diallel crosses of inbred parental lines were subjected to differential expression analysis to assess heterosis intensity. In general, the difference in gene expression abundance of alleles from parental inbred lines forms the basis of dominant performance of hybrids. Here, global gene expression patterns differed significantly between ZC and other hybrids, due to differences in both qualitative variation and quantitative variation of DEGs. By comparing gene expression patterns between ZC and other hybrids, the gene expression pattern of this strong predominant hybrid was revealed as was the important role of partial parental expression in heterosis. Meanwhile, we identified ZC-specifically expressed genes, as well as enrichment pathways associated with the set of identified genes. Annotation results indicated that enriched protogenes were associated with multiple biological pathways, such as photosynthesis, plant hormone signal transduction, ion channel regulation and other pathways associated with binding and catalytic activities. Intriguingly, we found that synergistic effects of expression of multiple distinct tissue-specific genes in ZC promoted comprehensive heterosis. This study provides new perspectives for elucidating molecular mechanisms that underlie heterosis and can serve as a technical reference for designing breeding programs that harness strong heterosis effects to improve maize and other crops.

MATERIALS AND METHODS

Plant Materials and Growth Conditions

Four elite maize inbred lines, Zheng58, Chang7-2, Mo17, and Qi319, were the typical lines of Reid, Sipingtou, Lancaster and P group, respectively. Six their hybrids, Zheng58 × Chang7-2, Zheng58 × Mo17, Zheng58 × Qi319, Chang7-2 × Mo17, Chang7-2 × Qi319, and Mo17 × Qi319, based on an incomplete diallel cross (NCII) were planted in April of 2019 at the Changping Station at the Institute of Crop Science, Chinese Academy of Agricultural Sciences (116°13′ E, 40°15′ N) in Beijing, China. Among these hybrids, Zhengdan958 is a commercially successful hybrid variety from Chinese maize breeding programs with the strong heterosis of yield (Li et al., 2012; Kogelman et al., 2014).

Tissue samples were collected as follows: (1) for leaf samples, three top leaves were collected from five-leaf stage seedlings and pooled; (2) for ear samples, more than 10 young ear tissues were collected at the stage of spikelet differentiation based on the phenotypic identification under a microscope and pooled; and (3) for seed samples, developing kernels containing embryo in the middle of ear were collected at 15 days after pollination (DAP). For each tissue, we sampled six biological replicates, three of which were used for subsequent analysis while the rest served as

a backup. Tissues were sampled, frozen immediately in liquid N_2, and then stored at −80°C until being used for RNA extraction.

For the convenience of subsequent description, the names of each maize line, hybrid, and sampled tissue are abbreviated as follows: Zheng58 (Z), Chang7-2 (C), Mo17 (M), Qi319 (Q), Zheng58 × Chang7-2 (ZC), Zheng58 × Mo17 (ZM), Zheng58 × Qi319 (ZQ), Chang7-2 × Mo17 (CM), Chang7-2 × Qi319 (CQ), Mo17 × Qi319 (MQ), Leaf (L), Ear (E), and Seed (S). Thus, the first biological replicate of a leaf sampled from the Zheng58 × Chang7-2 (ZC) hybrid is abbreviated as ZC_L1.

Construction and Sequencing of mRNA Libraries for RNA-Seq Analysis

A total of 90 samples of maize tissue (10 genotypes × 3 tissues × 3 biological replicates) were qualified for RNA sequencing (RNA-Seq) library construction. Tissues at the same stage of development were sampled from different plants and mixed to diminish differences between genotypes caused by sampling times.

We extracted total RNA with TRIzol Reagent (Invitrogen, CA, United States) according to the manufacturer's recommendations. The quantity and purity of total RNA samples were analyzed on the Bioanalyzer 2100 and RNA 1000 Nano Lab Chip Kit (Agilent, CA, United States) with RIN number > 7.0. We purified poly(A +) RNA from 5 mg of total RNA by two purification rounds on poly-T oligo magnetic beads. After mRNA was purified, it was fragmented in a fragmentation buffer containing divalent cations at high temperature. We then reverse-transcribed the RNA fragments following the procedures in the mRNA-Seq sample preparation kit (Illumina, San Diego, CA, United States) to generate the cDNA libraries. The libraries were then subjected to paired-end sequencing (2 bp × 150 bp, PE150) on an Illumina HiSeq 4000 platform at LC-BIOTECHNOLOGIES (LC Sciences, United States) Co., Ltd.,[1] according to the company's recommendations.

Analysis of Illumina RNA-Seq Data

We sequenced the transcriptomes of each genotype and tissue combination and generated a total of 3,899 billon raw paired-end 150-bp Illumina reads comprising a total of 584.86 gigabases (Gb) of sequence. We aligned the raw reads from each sample to the B73 maize genome[2] using Hisat2 (version 2.0.5) (Kim et al., 2016), which excludes some reads based on quality data and maps high-quality reads to the reference genome, with a minimum intron length of 20 bp and a maximum intron length of 50 kb, and other parameters set to defaults. We removed low-quality reads that contained sequencing adaptors or sequencing primers, reads containing more than 5% unascertained bases, and bases below Q20 before assembly. Up to 20 multiple alignments and at most two mismatches per read were allowed in HISAT when we mapped reads to the B73 reference genome. We used StringTie (version 1.3.4d) (Pertea et al., 2015) to calculate FPKM values (fragments per kilobase of exon model per million mapped reads) for RNAs in each sample. The genes with FPKM less than 1 were defined as non-expressed genes, so they were deleted in the calculation process. We used the R Bioconductor package "Mfuzz" (Kumar and Futschik, 2007) to cluster expressed genes according to the expression profiles of the parent-hybrid triplets. Short-read transcripts were reconstructed using Cufflinks (version 2.0) (Trapnell et al., 2012) and Trinity (version 0.1.0-alpha.14) (Haas et al., 2013), and then were mapped to the B73 maize genome using GMAP (Thomas and Serban, 2010) and were filtered for alignment coverage > 85% and alignment identity > 90%. A total of 3,257 billon cleaned paired-end reads comprising 488.57 Gbp of valid sequence data were produced. The results of sequence statistics and quality control were list in **Supplementary Table 1**. We have submitted the raw sequence data to NCBI (BioProject: PRJNA682889), and which has yet to be released.

Functional Enrichment Analysis

Enrichment analyses of DEGs were conducted using FuncAssociate 3.0 (Berriz et al., 2009) using Ensembl gene identifiers. GO enrichment analysis was performed using the OmicShare tools, a free online platform for data analysis[3]. Firstly, all DEGs were mapped to GO terms in the Gene Ontology database[4], gene numbers were calculated for every term, significantly enriched GO terms in DEGs comparing to the genome background were defined by hypergeometric test. The calculated *p*-value was gone through FDR correction, taking FDR ≤ 0.05 as a threshold. GO terms meeting this condition were defined as significantly enriched GO terms in DEGs. We used the ggplot2 program package to display the results of GO (Gene ontology) and KEGG (Kyoto Encyclopedia of Genes and Genomes,[5]) enrichment analyses. In addition, KO (KEGG Orthology) labels were used to represent the classification of proteins with similar functions in the same pathway.

Identification of Gene Expression Patterns

For each triplet comprising the set of parents and their hybrid, differences in gene expression could follow one of five patterns: (1) parental co-silence expression (PCE), in which the genes are expressed in parents but not in their hybrid; (2) parent-specific expression (PSE), in which genes are expressed in only one parent but not in the other parent and their hybrid; (3) hybrid-specific expression (HSE), in which genes are expressed only in the hybrid but not in its parents; (4) single-parent expression (SPE), in which genes are expressed in hybrid and one of its parents; and (5) parental-hybrid co-expression (PHCE), in which genes are expressed in the hybrid and both of its parents.

Models 1–4 (PCE, PSE, HSE and SPE) represent qualitative variations in differential gene expression that are essentially similar to presence-absence variations (PAVs) except that the gene is present in each genotype but are expressed only in a certain individual or genotype and not in another. As in the

[1] www.lc-bio.com/

[2] ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/plant/Zea_mays/latest_assembly_versions/GCA_000005005.6_B73_RefGen_v4

[3] www.omicshare.com/tools

[4] http://www.geneontology.org/

[5] www.genome.jp/kegg

quantitative gene expression pattern (PHCE) model, hybrids can be divided into three categories (high-value parents, middle-value parents, and low-value parents) and seven subcategories according to their A value (Stupar and Springer, 2006), where:

$$A = (P_{max} - F_1)/(P_{max} - P_{min})$$

In the formula, P_{max}, P_{min}, and F_1 represent levels of the parent with higher gene expression, the parent with the lower expression, and that of their hybrid, respectively. The categories of high-value parents, middle-value parents, and low-value parents include 2 ($A < 0.0$, $0 \leq A < 0.2$), 3 ($0.2 \leq A < 0.4$, $0.4 \leq A < 0.6$, $0.6 \leq A < 0.8$), and 2 ($0.8 \leq A < 1.0$, $A \geq 1.0$) subcategories with the A values corresponding to the interval in parentheses (Trapnell et al., 2009). Then all the genes of the aforementioned PHCE gene sets expressed in hybrids were divided into seven groups corresponding to positive super-dominance (PSD), positive dominance (PD), partial positive dominance (PPD), mid-parent (MP), partial negative dominance (PND), negative dominance (ND), and negative super-dominance (NSD) gene action.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/genbank/, PRJNA682889.

AUTHOR CONTRIBUTIONS

XL, JWe, and JWu conceived and designed the experiments. ML, ZX, JWe, and JH conducted the field experiments and collected the tissue samples. JWu and QZ performed the experiments and analyzed the data. JWu interpreted the data and drafted the manuscript. XL, JWe, and ML supervised the project and revised the manuscript. ZH, HY, DS, ZZ, XZ, and DZ provided helpful discussions on the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by National Natural Science Foundation of China (31971963); Innovation Project of Heilongjiang Academy of Agricultural Sciences (2019JCQN003). The funders had no role in study design, the collection, analysis, interpretation of data, writing of the manuscript, the preparation or decision of the manuscript to publication.

ACKNOWLEDGMENTS

We would like to thank Yunbi Xu for constructive suggestions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.739072/full#supplementary-material

Supplementary Figure 1 | Number of DEGs in hybrids and their parental inbred lines among tissues. Histogram depicting the numbers of DEGs in each genotype (inbred lines and hybrids) and tissues [leaves (L), ears (E), and seeds (S)] with colors listed on the right. The columns in the histogram are divided in two sections by wide gaps. The bottom half depicts the number of DEGs in L, E, and S of maize inbred lines Z, C, M, and Q, which represent Zheng58, Chang7-2, Mo17, and Qi319. The top half depicts the number of DEGs in L, E, and S of the hybrids, which include Zheng58 × Chang7-2 (ZC), Zheng58 × Mo17 (ZM), Zheng58 × Qi319 (ZQ), Chang7-2 × Mo17 (CM), Chang7-2 × Qi319 (CQ), and Mo17 × Qi319 (MQ). The different inbred and hybrid genotypes are shown on the *y*-axis, and the numbers of significant DEGS expressed by the different genotypes are shown on the *x*-axis.

Supplementary Figure 2 | Gene expression patterns in leaves and ears of hybrids. The differential expression of DEGs in leaves of hybrids with PAV **(A)** and PHCE patterns **(B)**, and the PAV **(C)** and PHCE **(D)** patterns in ears, where the expression quantity of different type of genes is marked on the bar, and colors in the figure represent different types of DEGs listed on the right side of the figure.

Supplementary Figure 3 | Distributions of ZC-specific DEGs in tissues. Venn diagram with colors represent leaves (L), ears (E), and seeds (S), the numbers of DEGs are listed in the corresponding section.

Supplementary Figure 4 | Expression patterns of specific genes expressed in different hybrids. The differential expression of DEGs in leaves (L), ears (E), and seeds (S) of hybrids with PAV **(A)** and PHCE **(B)** patterns. The composition of genes with different expression patterns in PHCE in leaves (L), ears (E) and seeds (S) was calculated in proportion to the total genes, respectively. Six hybrids: Zheng58 × Chang7-2 (ZC), Zheng58 × Mo17 (ZM), Zheng58 × Qi319 (ZQ), Chang7-2 × Mo17 (CM), Chang7-2 × Qi319 (CQ), and Mo17 × Qi319 (MQ).

Supplementary Table 1 | Descriptive statistics and quality control information for RNA-Seq data. Genotypes and tissue are abbreviated as follows: Zheng58 (Z), Chang7-2 (C), Mo17 (M), Qi319 (Q), Zheng58 × Chang7-2 (ZC), Zheng58 × Mo17 (ZM), Zheng58 × Qi319 (ZQ), Chang7-2 × Mo17 (CM), Chang7-2 × Qi319 (CQ), Mo17 × Qi319 (MQ), Leaf (L), Ear (E) and Seed (S). Thus, the first biological replicate of a leaf sampled from the Zheng58 × Chang7-2 (ZC) hybrid is abbreviated as ZC_L1. All genotypes were compared to the B73 maize genome with the mapped reads and ratio in the last column.

Supplementary Table 2 | Gene expression patterns in hybrids. Gene expressed with qualitative (PAVs) and quantitative (PHCE) variation in different tissues [leaves (L), ears (E), and seeds (S)] in the hybrids, which include Zheng58 × Chang7-2 (ZC), Zheng58 × Mo17 (ZM), Zheng58 × Qi319 (ZQ), Chang7-2 × Mo17 (CM), Chang7-2 × Qi319 (CQ), and Mo17 × Qi319 (MQ). The Gene expression patterns listed include different gene actions of parental co-silence expression (PCE), parent-specific expression (PSE), hybrid-specific expression (HSE), single-parent expression (SPE), negative super-dominant (NSD), negative dominant (ND), partial negative dominant (PND), mid-parent (MP), partial positive dominant (PPD), positive dominance (PD), and positive super-dominance (PSD).

Supplementary Table 3 | Details of Zhengdan958-specific DEGs. Distribution, expression patterns and enrichment pathways of Zhengdan958-specific genes in different tissues (leaves, ears, and seeds). The Gene expression patterns (Models) listed include different gene actions of hybrid-specific expression (HSE), single-parent expression (SPE), and parental/hybrid co-expression (PHCE).

Supplementary Table 4 | Details of Zhengdan958-specific genes in Gene ontology. Gene Ontology analysis were conducted with three categories of Biological Process, Cellular Component and Molecular Function in GO Term (level 1) by the Zhengdan958-specific genes, which were describes the molecular functions the gene products may perform, the cellular environment in which they are exposed, and the biological processes involved. The genes were assigned to the GO Term (level 2) listed with the rank in order of most to least of gene numbers.

REFERENCES

Aguiar, C. G., Schuster, I., Amaral, A. T., Scapim, C. A., and Vieira, E. S. (2008). Heterotic groups in tropical maize germplasm by test crosses and simple sequence repeat markers. *Genet. Mol. Res.* 7, 1233–1244. doi: 10.4238/vol7-4gmr495

Alonso-Peral, M. M., Trigueros, M., Sherman, B., Ying, H., Taylor, J. M., Peacock, W. J., et al. (2017). Patterns of gene expression in developing embryos of *Arabidopsis* hybrids. *Plant J.* 89, 927–939. doi: 10.1111/tpj.13432

Annor, B., Badu-Apraku, B., Nyadanu, D., Akromah, R., and Fakorede, M. (2020). Identifying heterotic groups and testers for hybrid development in early maturing yellow maize (*Zea mays*) for sub-Saharan Africa. *Plant Breed.* 139, 708–716. doi: 10.1111/pbr.12822

Baldauf, J. A., Marcon, C., Paschold, A., and Hochholdinger, F. (2016). Nonsyntenic genes drive tissue-specific dynamics of differential, nonadditive, and allelic expression patterns in maize hybrids. *Plant Physiol.* 171, 1144–1155. doi: 10.1104/pp.16.00262

Berriz, G. F., Beaver, J. E., Cenik, C., Tasan, M., and Roth, F. P. (2009). Next generation software for functional trend analysis. *Bioinformatics* 25, 3043–3044. doi: 10.1093/bioinformatics/btp498

Birchler, J. A., Johnson, A. F., and Veitia, R. A. (2016). Kinetics genetics: incorporating the concept of genomic balance into an understanding of quantitative traits. *Plant Sci.* 245, 128–134. doi: 10.1016/j.plantsci.2016.02.002

Bommert, P., Satoh-Nagasawa, N., Jackson, D., and Hirano, H. Y. (2005). Genetics and evolution of inflorescence and flower development in grasses. *Plant Cell Physiol.* 46, 69–78. doi: 10.1093/pcp/pci504

Bruce, A. B. (1910). The Mendelian theory of heredity and the augmentation of vigor. *Science* 32, 627–628. doi: 10.1126/science.32.827.627-a

Cai, D., Xiao, Y., Yang, W., Ye, W., Wang, B., Younas, M., et al. (2014). Association mapping of six yield-related traits in rapeseed (*Brassica napus* L.). *Theor. Appl. Genet.* 127, 85–96. doi: 10.1007/s00122-013-2203-9

Dapp, M., Reinders, J., Bédiée, A., Balsera, C., Bucher, E., Theiler, G., et al. (2015). Heterosis and inbreeding depression of epigenetic *Arabidopsis* hybrids. *Nat. Plants* 1:15092. doi: 10.1038/nplants.2015.92

Davenport, C. B. (1908). Degeneration, albinism and inbreeding. *Science* 28, 454–455. doi: 10.1126/science.28.718.454-b

Duvick, D. N. (2005). The contribution of breeding to yield advances in maize (*Zea mays* L.). *Adv. Agron.* 86, 83–145. doi: 10.1016/S0065-2113(05)86002-X

East, E. M. (1908). *Inbreeding in corn. In: Reports of the Connecticut Agricultural Experiment Station for Years 1907*. New Haven: Connecticut Agricultural Experiment Station, 419–428.

Ferrándiz, C., Liljegren, S. J., and Yanofsky, M. F. (2000). Negative regulation of the SHATTERPROOF genes by FRUITFULL during Arabidopsis fruit development. *Science* 289, 436–438. doi: 10.1126/science.289.5478.436

Frascaroli, E., Canè, M. A., Landi, P., Pea, G., Gianfranceschi, L., Villa, M., et al. (2007). Classical genetic and quantitative trait loci analyses of heterosis in a maize hybrid between two elite inbred lines. *Genetics* 176, 625–644. doi: 10.1534/genetics.106.064493

Fu, D., Xiao, M., Hayward, A., Jiang, G., Zhu, L., Zhou, Q., et al. (2015). What is crop heterosis: new insights into an old topic. *J. Appl. Genet.* 56, 1–13. doi: 10.1007/s13353-014-0231-z

Gao, L., Zhao, G., Huang, D., and Jia, J. (2017). Candidate loci involved in domestication and improvement detected by a published 90K wheat SNP array. *Sci. Rep.* 7:44530. doi: 10.1038/srep44530

Garcia, A. A., Wang, S., Melchinger, A. E., and Zeng, Z. B. (2008). Quantitative trait loci mapping and the genetic basis of heterosis in maize and rice. *Genetics* 180, 1707–1724. doi: 10.1534/genetics.107.082867

Groszmann, M., Greaves, I. K., Fujimoto, R., James Peacock, W., and Dennis, E. S. (2013). The role of epigenetics in hybrid vigour. *Trends Genet.* 29, 684–690. doi: 10.1016/j.tig.2013.07.004

Guo, M., Rupe, M. A., Zinselmeier, C., Habben, J., Bowen, B. A., and Smith, O. S. (2004). Allelic variation of gene expression in maize hybrids. *Plant Cell* 16, 1707–1716. doi: 10.1105/tpc.022087

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Philip, D., Bowden, J., et al. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* 8, 1494–1512. doi: 10.1038/nprot.2013.084

Harrison, P. W., Wright, A. E., and Mank, J. E. (2012). The evolution of gene expression and the transcriptome-phenotype relationship. *Semin. Cell Dev. Biol.* 23, 222–229. doi: 10.1016/j.semcdb.2011.12.004

Hoecker, N., Keller, B., Muthreich, N., Chollet, D., Descombes, P., Piepho, H. P., et al. (2008). Comparison of maize (*Zea mays* L.) F1-hybrid and parental inbred line primary root transcriptomes suggests organ-specific patterns of nonadditive gene expression and conserved expression trends. *Genetics* 179, 1275–1283. doi: 10.1534/genetics.108.088278

Hoopes, G. M., Hamilton, J. P., Wood, J. C., Esteban, E., Pasha, A., Vaillancourt, B., et al. (2019). An updated gene atlas for maize reveals organ-specific and stress-induced genes. *Plant J.* 97, 1154–1167. doi: 10.1111/tpj.14184

Hu, X., Wang, H., Diao, X., Liu, Z., Li, K., Wu, Y., et al. (2016). Transcriptome profiling and comparison of maize ear heterosis during the spikelet and floret differentiation stages. *BMC Genomics* 17:959. doi: 10.1186/s12864-016-3296-8

Huang, X., Yang, S., Gong, J., Zhao, Q., Feng, Q., Zhan, Q., et al. (2016). Genomic architecture of heterosis for yield traits in rice. *Nature* 537, 629–633. doi: 10.1038/nature19760

Huang, X., Yang, S., Gong, J., Zhao, Y., Feng, Q., Gong, H., et al. (2015). Genomic analysis of hybrid rice varieties reveals numerous superior alleles that contribute to heterosis. *Nat. Commun.* 6:6258. doi: 10.1038/ncomms7258

Jones, D. F. (1917). Dominance of linked factors as a means of accounting for heterosis. *Proc. Natl. Acad. Sci. U. S. A.* 2, 466–479. doi: 10.1073/pnas.3.4.310

Kim, D., Langmead, B., and Salzberg, S. L. (2016). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi: 10.1038/nmeth.3317.HISAT

Kogelman, L. J. A., Cirera, S., Zhernakova, D. V., Fredholm, M., Franke, L., and Kadarmideen, H. N. (2014). Identification of co-expression gene networks, regulatory genes and pathways for obesity based on adipose tissue RNA Sequencing in a porcine model. *BMC Med. Genomics* 7:57. doi: 10.1186/1755-8794-7-57

Kumar, L., and Futschik, M. E. (2007). Mfuzz: A software package for soft clustering of microarray data. *Bioinformation* 2, 5–7. doi: 10.6026/97320630002005

Larièpe, A., Mangin, B., Jasson, S., Combes, V., Dumas, F., Jamin, P., et al. (2012). The genetic basis of heterosis: multiparentalquantitative trait loci mapping reveals contrasted levels of apparent overdominance among traits of agronomical interest in maize (*Zea mays* L.). *Genetics* 190, 795–811. doi: 10.1534/genetics.111.133447

Li, D., Zeng, R., Li, Y., Zhao, M., Chao, J., Li, Y., et al. (2016). Gene expression analysis and SNP/InDel discovery to investigate yield heterosis of two rubber tree F1 hybrids. *Sci. Rep.* 6:24984. doi: 10.1038/srep24984

Li, Z. Y., Zhang, T. F., and Wang, S. C. (2012). Transcriptomic analysis of the highly heterotic maize hybrid zhengdan 958 and its parents during spikelet and floscule differentiation. *J. Integr. Agric.* 11, 1783–1793. doi: 10.1016/S2095-3119(12)60183-X

Lv, Q., Li, W., Sun, Z., Ouyang, N., Jing, X., He, Q., et al. (2020). Resequencing of 1,143 indica rice accessions reveals important genetic variations and different heterosis patterns. *Nat. Commun.* 11:4778. doi: 10.1038/s41467-020-18608-0

Ma, J., Zhang, D., Cao, Y., Wang, L., Li, J., Lübberstedt, T., et al. (2018). Heterosis-related genes under different planting densities in maize. *J. Exp. Bot.* 69, 5077–5087. doi: 10.1093/jxb/ery282

Ma, X., Xing, F., Jia, Q., Zhang, Q., Hu, T., Wu, B., et al. (2021). Parental variation in CHG methylation is associated with allelic-specific expression in elite hybrid rice. *Plant Physiol.* 186, 1025–1041. doi: 10.1093/plphys/kiab088

Makumbi, D., Betrán, J. F., Bänziger, M., and Ribaut, J. M. (2011). Combining ability, heterosis and genetic diversity in tropical maize (*Zea mays* L.) under stress and non-stress conditions. *Euphytica* 180, 143–162. doi: 10.1007/s10681-010-0334-5

Meyer, S., Pospisil, H., and Scholten, S. (2007). Heterosis associated gene expression in maize embryos 6 days after fertilization exhibits additive, dominant and overdominant pattern. *Plant Mol. Biol.* 63, 381–391. doi: 10.1007/s11103-006-9095-x

Minvielle, F. (1987). Dominance is not necessary for heterosis: a two-locus model. *Genet. Res.* 49, 245–247. doi: 10.1017/S0016672300027142

Paschold, A., Jia, Y., Marcon, C., Lund, S., Larson, N. B., Yeh, C. T., et al. (2012). Complementation contributes to transcriptome complexity in maize (*Zea mays* L.) hybrids relative to their inbred parents. *Genome Res.* 22, 2445–2454. doi: 10.1101/gr.138461.112

Paschold, A., Larson, N. B., Marcon, C., Schnable, J. C., Yeh, C. T., Lanz, C., et al. (2014). Nonsyntenic genes drive highly dynamic complementation of gene expression in maize hybrids. *Plant Cell* 26, 3939–3948. doi: 10.1105/tpc.114.130948

Pautler, M., Tanaka, W., Hirano, H. Y., and Jackson, D. (2013). Grass meristems I: shoot apical meristem maintenance, axillary meristem determinacy and the floral transition. *Plant Cell Physiol.* 54, 302–312. doi: 10.1093/pcp/pct025

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295. doi: 10.1038/nbt.3122

Qin, X., Suga, M., Kuang, T., and Shen, J. R. (2015). Photosynthesis. Structural basis for energy transfer pathways in the plant PSI-LHCI supercomplex. *Science* 348, 989–995. doi: 10.1126/science.aab0214

Reif, J. C., Melchinger, A. E., Xia, X. C., Warburton, M. L., Hoisington, D. A., Vasal, S. K., et al. (2003). Use of SSRs for establishing heterotic groups in subtropical maize. *Theor. Appl. Genet.* 107, 947–957. doi: 10.1007/s00122-003-1333-x

Schnable, P. S., and Springer, N. M. (2013). Progress toward understanding heterosis in crop plants. *Annu. Rev. Plant Biol.* 64, 71–88. doi: 10.1146/annurev-arplant-042110-103827

Schnell, F. W., and Cockerham, C. C. (1992). Multiplicative vs. arbitrary gene action in heterosis. *Genetics* 131, 461–469. doi: 10.1016/1050-3862(92)90005-P

Shao, L., Xing, F., Xu, C., Zhang, Q., Che, J., Wang, X., et al. (2019). Patterns of genome-wide allele-specific expression in hybrid rice and the implications on the genetic basis of heterosis. *Proc. Natl. Acad. Sci. U. S. A.* 116, 5653–5658. doi: 10.1073/pnas.1820513116

Shull, G. H. (1908). The composition of a field of maize. *J. Hered.* 4, 296–301. doi: 10.1093/jhered/os-4.1.296

Stelpflug, S. C., Sekhon, R. S., Vaillancourt, B., Hirsch, C. N., Buell, C. R., de Leon, N., et al. (2016). An Expanded Maize Gene Expression Atlas based on RNA Sequencing and its Use to Explore Root Development. *Plant Genome* 9, 1–15. doi: 10.3835/plantgenome2015.04.0025

Stupar, R. M., Gardiner, J. M., Oldre, A. G., Haun, W. J., Chandler, V. L., and Springer, N. M. (2008). Gene expression analyses in maize inbreds and hybrids with varying levels of heterosis. *BMC Plant Biol.* 8:33. doi: 10.1186/1471-2229-8-33

Stupar, R. M., and Springer, N. M. (2006). *Cis*-transcriptional variation in maize inbred lines B73 and Mo17 leads to additive expression patterns in the F1 hybrid. *Genetics* 173, 2199–2210. doi: 10.1534/genetics.106.060699

Swanson-Wagner, R. A., Jia, Y., DeCook, R., Borsuk, L. A., Nettleton, D., and Schnable, P. S. (2006). All possible modes of gene action are observed in a global comparison of gene expression in a maize F1 hybrid and its inbred parents. *Proc. Natl. Acad. Sci. U. S. A.* 103, 6805–6810. doi: 10.1073/pnas.0510430103

Thiemann, A., Fu, J., Schrag, T. A., Melchinger, A. E., Frisch, M., and Scholten, S. (2010). Correlation between parental transcriptome and field data for the characterization of heterosis in *Zea mays* L. *Theor. Appl. Genet.* 120, 401–413. doi: 10.1007/s00122-009-1189-9

Thomas, D. W., and Serban, N. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 7, 873–881. doi: 10.1093/bioinformatics/btq057

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–1111. doi: 10.1093/bioinformatics/btp120

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578. doi: 10.1038/nprot.2012.016

Wallace, J. G., Larsson, S. J., and Buckler, E. S. (2014). Entering the second century of maize quantitative genetics. *Heredity* 112, 30–38. doi: 10.1038/hdy.2013.6

Waters, A. J., Makarevitch, I., Noshay, J., Burghardt, L. T., Hirsch, C. N., Hirsch, C. D., et al. (2017). Natural variation for gene expression responses to abiotic stress in maize. *Plant J.* 89, 706–717. doi: 10.1111/tpj.13414

Wu, Y., San, V. F., Huang, K., Dhliwayo, T., Costich, D. E., Semagn, K., et al. (2016). Molecular characterization of CIMMYT maize inbred lines with genotyping-by-sequencing SNPs. *Theor. Appl. Genet.* 129, 753–765. doi: 10.1007/s00122-016-2664-8

Xiang, C., Zhang, H., Wang, H., Wei, S., Fu, B., and Xia, J. (2016). Dissection of heterosis for yield and related traits using populations derived from introgression lines in rice. *Crop J.* 4, 468–478. doi: 10.1016/j.cj.2016.05.001

Xiao, S. J., Zhang, C., Zou, Q., and Ji, Z. L. (2010). TiSGeD: a database for tissue-specific genes. *Bioinformatics* 26, 1273–1275. doi: 10.1093/bioinformatics/btq109

Yang, J., Liu, Z., Chen, Q., Qu, Y., Tang, J., Lübberstedt, T., et al. (2020). Mapping of QTL for grain yield components based on a DH population in maize. *Sci. Rep.* 10:7086. doi: 10.1038/s41598-020-63960-2

Zarayeneh, N., Ko, E., Oh, J. H., Suh, S., Liu, C., Gao, J., et al. (2017). Integration of multi-omics data for integrative gene regulatory network inference. *Intl. J. Data Min. Bioinform.* 18, 223–239. doi: 10.1504/IJDMB.2017.087178

Zhang, R., Xu, G., Li, J., Yan, J., Li, H., and Yang, X. (2018). Patterns of genomic variation in Chinese maize inbred lines and implications for genetic improvement. *Theor. Appl. Genet.* 131, 1207–1221. doi: 10.1007/s00122-018-3072-z

Zhang, X., Lv, L., Lv, C., Guo, B., and Xu, R. (2015). Combining ability of different agronomic traits and yield components in hybrid barley. *PLoS One* 10:e0126828. doi: 10.1371/journal.pone.0126828

Zhou, P., Hirsch, C. N., Briggs, S. P., and Springer, N. M. (2019). Dynamic patterns of gene expression additivity and regulatory variation throughout maize development. *Mol. Plant* 12, 410–425. doi: 10.1016/j.molp.2018.12.015

Zhu, Y. J., Huang, D. R., Fan, Y. Y., Zhang, Z. H., Ying, J. Z., and Zhuang, J. Y. (2016). Detection of QTLs for yield heterosis in rice using a RIL population and its testcross population. *Intl. J. Genomics* 2016:2587823. doi: 10.1155/2016/2587823

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 12 October 2021
doi: 10.3389/fpls.2021.709817

Comparative Genomic Analysis of Quantitative Trait Loci Associated With Micronutrient Contents, Grain Quality, and Agronomic Traits in Wheat (*Triticum aestivum* L.)

Nikwan Shariatipour[1†], *Bahram Heidari*[1*†], *Ahmad Tahmasebi*[1] *and Christopher Richards*[2†]

[1] *Department of Plant Production and Genetics, School of Agriculture, Shiraz University, Shiraz, Iran,* [2] *USDA ARS National Laboratory for Genetic Resources Preservation, Fort Collins, CO, United States*

Comparative genomics and meta-quantitative trait loci (MQTLs) analysis are important tools for the identification of reliable and stable QTLs and functional genes controlling quantitative traits. We conducted a meta-analysis to identify the most stable QTLs for grain yield (GY), grain quality traits, and micronutrient contents in wheat. A total of 735 QTLs retrieved from 27 independent mapping populations reported in the last 13 years were used for the meta-analysis. The results showed that 449 QTLs were successfully projected onto the genetic consensus map which condensed to 100 MQTLs distributed on wheat chromosomes. This consolidation of MQTLs resulted in a three-fold reduction in the confidence interval (CI) compared with the CI for the initial QTLs. Projection of QTLs revealed that the majority of QTLs and MQTLs were in the non-telomeric regions of chromosomes. The majority of micronutrient MQTLs were located on the A and D genomes. The QTLs of thousand kernel weight (TKW) were frequently associated with QTLs for GY and grain protein content (GPC) with co-localization occurring at 55 and 63%, respectively. The co- localization of QTLs for GY and grain Fe was found to be 52% and for QTLs of grain Fe and Zn, it was found to be 66%. The genomic collinearity within Poaceae allowed us to identify 16 orthologous MQTLs (OrMQTLs) in wheat, rice, and maize. Annotation of promising candidate genes (CGs) located in the genomic intervals of the stable MQTLs indicated that several CGs (e.g., *TraesCS2A02G141400*, *TraesCS3B02G040900*, *TraesCS4D02G323700*, *TraesCS3B02G077100*, and *TraesCS4D02G290900*) had effects on micronutrients contents, yield, and yield-related traits. The mapping refinements leading to the identification of these CGs provide an opportunity to understand the genetic mechanisms driving quantitative variation for these traits and apply this information for crop improvement programs.

Keywords: meta QTL analysis, comparative genomics, iron, zinc, candidate gene, GWAS

OPEN ACCESS

Edited by:
Leonardo Abdiel Crespo Herrera, International Maize and Wheat Improvement Center, Mexico

Reviewed by:
Velu Govindan, International Maize and Wheat Improvement Center, Mexico
Roi Ben-David, Agricultural Research Organization (ARO), Israel

***Correspondence:**
Bahram Heidari
bheidari@shirazu.ac.ir

†ORCID:
Nikwan Shariatipour
orcid.org/0000-0003-4174-4375
Bahram Heidari
orcid.org/0000-0002-5856-4592
Christopher Richards
orcid.org/0000-0002-9978-6079

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *14 May 2021*
Accepted: *06 September 2021*
Published: *12 October 2021*

Citation:
Shariatipour N, Heidari B, Tahmasebi A and Richards C (2021) Comparative Genomic Analysis of Quantitative Trait Loci Associated With Micronutrient Contents, Grain Quality, and Agronomic Traits in Wheat (Triticum aestivum L.). Front. Plant Sci. 12:709817. doi: 10.3389/fpls.2021.709817

INTRODUCTION

Most agricultural systems focus on increasing crop productivity and grain yield (GY) and fewer efforts have been devoted to the grain yield–quality tradeoff. However, a shift from prioritizing yield to more emphasis on quality, such as nutrient content is gaining ground in breeding programs (Khush et al., 2012). Extending the existing concepts for a simultaneous selection of GY, quality traits, and micronutrient contents seems necessary to facilitate the development of varieties with an effective combination of yield potential and end-use quality (Michel et al., 2019). A rapid increase in micronutrient deficiency in food grains has resulted in micronutrient malnutrition among consumers. Fe and Zn deficiencies are serious and prevalent sources of malnutrition in developing countries with high consumption of cereals, such as wheat (Black et al., 2013; Shahzad et al., 2014; Kumar et al., 2019). Diets based on staple food crops with Zn and Fe deficiency have been widely recognized as a major global health problem that affects almost three billion people (Murray and Lopez, 2013). Breeding crops through biofortification is a practical approach to cope with Fe and Zn deficiencies by increasing the grain Fe content (GFeC) and grain Zn content (GZnC) within the edible parts of staple food crops, especially cereals (Stein, 2010; Liu et al., 2019; Shariatipour and Heidari, 2020).

Wheat biofortification through breeding methods is a promising strategy to ameliorate Fe and Zn deficiencies in developing countries (Liu et al., 2019). One of the most important challenges in breeding for micronutrients are negative genetic trade-offs between yield and micronutrient traits (Flatt and Heyland, 2011; Fabian and Flatt, 2012). At the genetic level, such trade-offs are thought to be caused by alleles with antagonistic pleiotropic effects or by linkage disequilibrium between loci (Fabian and Flatt, 2012). While GFeC and GZnC biofortification is an important objective in wheat breeding programs, other important traits, such as GY and grain protein content (GPC) typically cannot be compromised. The wheat quality is measured by its rheological traits, such as GPC (Goel et al., 2019) since wheat is a major source of protein accounting for 19% of human protein intake in the developing countries (Braun et al., 2010). The genetic control of quality traits and GY is complex (Zilic et al., 2011; Velu et al., 2018; Giancaspro et al., 2019; Liu et al., 2019). A high priority of breeders is to develop high-quality genotypes that balance acceptable yield potential while maintaining quality characteristics, both of which are highly dependent on the co-variance between GY and the major quality (i.e., GZnC, GFeC, and GPC) criteria (Michel et al., 2019). However, a negative correlation between the GY and quality traits is challenging (Simmonds, 1995; Velu et al., 2016; Michel et al., 2019). In addition to this, our research indicates that protein concentration in grain decreases under elevated air CO_2 concentrations of 550 μmol/mol (Fernando et al., 2012). By the end of the twenty-first century, it is predicted that the global temperature will rise from 1.1 to −3.1°C and the atmospheric CO_2 concentration will reach above 550 ppm under intermediate scenarios (Pachauri et al., 2014) and that heat waves will occur with a higher frequency and longer duration (Pachauri et al., 2014). Given these future climate scenarios, it is critical to anticipate the effects of future growing environments and focus on breeding strategies that compensate for changes in grain quality. Likewise, it is important to have an inclusive breeding objective that tracks a portfolio of indirect traits responsible for grain quality and productivity, such as the assessment of dry matter accumulation, photosynthesis, coleoptile growth, carbon isotope discrimination, plant senescence, and rheological properties (Rebetzke et al., 2008; Liang et al., 2010; Vijayalakshmi et al., 2010; Goel et al., 2019).

Because of the large genome size and limited genome sequence information in wheat, the typical mapping intervals are quite large in most studies especially for the complex quality traits (Li Q. et al., 2020) and so further refinement is needed to narrow down the QTL intervals. Developing a statistically derived catalog of relevant loci is critical for developing marker-assisted selection (MAS) approaches in breeding programs. These markers can be applied to the quantitative trait loci (QTLs) that regulate the accumulation of high mineral nutrient concentration in grain along with QTLs for GY and grain quality traits. The QTL mapping method involves creating a QTL continuity map to identify genomic regions associated with quantitative traits (Mohan et al., 1997). Although QTL mapping is a powerful approach for detecting the genomic regions associated with complex traits, the genetic effects of QTL identified in different studies may not be present or are simply not tested in different genetic backgrounds and environments (Zhang L. Y. et al., 2010). In addition, the number of traits that can be measured in any single study is always resource limited (Acuña-Galindo et al., 2015). Overall, biparental populations are strongly influenced by different factors consisting of parents, the size and type of population, the choice of marker sets, and environmental conditions (Li et al., 2013; Izquierdo et al., 2018; Lei et al., 2018; Zhao et al., 2018).

In the last decade, an efficient approach called meta-QTL (MQTL) analysis has emerged in order to circumvent these restrictions. The MQTL method was initially developed by Goffinet and Gerber (2000) and was then improved by Veyrieras et al. (2007) is a method that gathers QTL data from independent experiments, years, location, and genetic backgrounds to detect stable QTLs (Goffinet and Gerber, 2000; Arcade et al., 2004; Hanocq et al., 2007; Sosnowski et al., 2012). The meta-QTL analysis integrates the information of QTLs from different population types and sizes identified in different environmental conditions to find stable MQTLs in a narrower genomic region with small CI (Goffinet and Gerber, 2000; Hanocq et al., 2007; Li et al., 2013).

The MQTL analysis allows for the dissection of genetic correlation among different traits (Truntzler et al., 2010; Danan et al., 2011; Xiang et al., 2012; Badji et al., 2018; Delfino et al., 2019). Hence, the MQTL analysis helps to enquire co-location

Abbreviations: AIC, akaike information criterion; AICc, corrected AIC; AWE, the average weight of evidence; BIC, Bayesian information criterion; CI, confidence interval; GCs, candidate genes; GO, gene ontology; GWAS, genome wide association studies; LOD, the logarithm of odds; MQTL, meta- quantitative trait loci; MTP, metal tolerance proteins; OrMQTL, orthologous MQTL; QTL, quantitative trait loci; SNP, single-nucleotide polymorphism.

of QTLs relying on dense marker maps that are responsible for different desirable traits including micronutrient contents, GY, and quality traits (Delfino et al., 2019). Currently, the MQTL analysis has become popular research in most studies (Goffinet and Gerber, 2000) related to micronutrients content (Raza et al., 2019), yield (Zhang L. Y. et al., 2010; Avni et al., 2018), and quality traits (Quraishi et al., 2017) to overcome the inconsistent QTL information reported for GZnC, GFeC, GPC, and yield traits. Further, MQTL analysis helps to identify the candidate genes (CGs) and refine genomic regions for yield and quality traits (Raza et al., 2019). However, little is known about the relation of micronutrients and quality-related MQTLs with agronomic traits in wheat.

The transferability of QTLs between cereals based on the analysis of syntenic regions and genomic collinearity helps to identify stable and important QTLs for use in breeding programs. The aims of this study were to perform a QTL meta-analysis to (1) identify QTLs that are consistently associated with grain quality, yield traits, and micronutrient content (2) explore the co-localized QTLs controlling GY, GPC, GZnC, and GFeC in the wheat genome, and (3) assess the transferability of QTLs between wheat, rice, and maize based on the comparative genomics and the orthologous MQTL (OrMQTS) mining. The outcome of this study will aid plant breeders in refining micronutrients, GY, and quality traits for crop improvement through marker-assisted breeding. Refining our understanding of the genetic architecture of micronutrients, GY, and quality traits often leads to the interrelationship between the regions of the genome that may be more challenging to breed independently. The MQTL is an analytic procedure that helps refine these relationships between CGs by providing for precision and statistical power.

MATERIALS AND METHODS

QTL Database Development

A database consisted of 735 quantitative trait loci (QTLs) derived from 27 independent mapping populations (assessed between 2006 and 2019) assigned to 70 traits (**Table 1**) was used for the meta-QTL (MQTL) analysis. The independent populations consisted of 20 recombinant inbred lines (RILs) and 7 double haploids (DH) populations, with population sizes ranging from 92 to 485 lines (**Supplementary Table 1**). The reported position, the proportion of the phenotypic variance (R^2), and the logarithm of odds (LOD score) of the initial QTLs were used for the analysis of meta-QTLs. For QTLs with missing LOD or R^2, the values were estimated by the following equation (Nagelkerke, 1991):

$$R^2 = 1 - 10^{(-2LOD/n)}$$

where n represents the size of the population.

Constructing Consensus Genetic Map and QTL Projection

The data files of the 27 maps were integrated with the Somers (Somers et al., 2004) reference map for the construction of a consensus genetic map. Attempts to use other mapping studies consisting of single-nucleotide polymorphism (SNP) were unsuccessful due to the lack of SNP density in the regions for the meta-QTL analysis. The constructed map file for each population consisted of information on cross-type, population size, map function, map units, and the position of different markers in different linkage groups. The individual QTLs derived from independent populations were projected onto the consensus genetic map consisted of 3,394 markers with a total length of 3,412.5 cM.

Meta-QTL Analysis

Meta-QTL analysis was performed in BioMercator v4.2 (Arcade et al., 2004; Sosnowski et al., 2012). For n QTLs, the BioMercator tests the most likely assumption based on Akaike information criterion (AIC), corrected AIC (AICc), AIC 3 candidate models (AIC3), Bayesian information criterion (BIC), and an average weight of evidence (AWE) criteria in which the prevalent value among them was considered as the best fit. The consensus QTL from the optimal model was reported as MQTL. Consequently, the MQTL position and distribution on each linkage group were presented as a heatmap using *pheatmap* R package (Kolde, 2013). Moreover, the initial QTLs with 95% confidence interval (CI), QTL density in the identified MQTL, and the distribution of MQTLs were drawn on the linkage groups using shinyCircos web tool based on the R program (Yu et al., 2017). The variation of QTL density for different traits toward centromeric and telomeric genomic regions was estimated following the approach by Martinez et al. (2016). The QTL density was determined by counting the number of QTLs for each trait on 50 cM intervals across the wheat genome, starting from the centromere region of a linkage group at position 0. The centromere position was retrieved from the study by Wan et al. (2017).

Functional Candidate Genes in MQTLs Intervals

The MQTLs containing more than five trait-QTLs from different experiments were considered as the most stable consensus regions and were analyzed for the detection of functional candidate genes (CGs). To identify the functional CGs, the sequences of the flanking markers for each MQTL were retrieved from "Grain Genes" database (https://wheat.pw.usda.gov/browse?class=probe;query=BARC%2A;begin=351) for the simple sequence repeat (SSR) flanking markers and "Diversity Array Technology" (https://www.diversityarrays.com/) (DArT) flanking markers. For flanking markers lacking a definite position on the wheat genome, the closest markers on the genetic consensus map were selected to determine the MQTL position. Additionally, for those flanking markers lacking sequence information in databases, the forward and reverse sequences were retrieved from the "Grain Genes" database and were used for the Basic Local Alignment Search Tool (BLAST) analysis against the newest wheat reference genome (IWGSC RefSeq v2.0) for detecting the genomic position of each MQTL. The annotation and gene ontology (GO) of genes lying at the MQTL interval were retrieved from EnsemblPlants (http://plants.ensembl.org/index.html) using the new wheat genome (IWGSC v2.0). Finally, the orthologous of genes located at each MQTL interval were

TABLE 1 | The list of assessed traits in meta-quantitative trait loci (MQTL) analysis.

Trait	Abbreviation	Trait	Abbreviation
200KW	200-kernel weight	GY	Grain yield
25%G	25% green leaf area	GZnC	Grain Zn content
50%G	50% green leaf area	HI	Harvest index
75%G	75% green leaf area	HW	Hectoliter weight
AGB	Above ground biomass	KH	Kernel hardness
BDT	Break down time	KL	Kernel length
BM	Biomass	KW	Kernel width
BY	Biological yield	LDMA	Leaves dry matter accumulation
CDMA	Culm dry matter accumulation	LL	Leaf length
CID	Carbon isotope discrimination	LS	Lodging score
DA	Days to anthesis	LW	Leaf width
DDT	Dough development time	LY	Leaf yellowing
DGC	Dry gluten content	MDR	Maturity date
DH	Days to heading	MRS	Maximum rate of senescence
DPM	Days to physiological maturity	MTI	Mixing tolerance index
DST	Dough stability time	NG	Number of grain per spike
DTF	Days to flowering	PDMA	Plants dry matter accumulation
FFD	Factor form density	PGMS	Percent green at maximum senescence
FLH	Flag leaf height	PH	Plant height
FWA	Flour water absorption	PLH	Penultimate leaf height
GAS	Grain area size	PT	Productive tillers/m^2
GCuC	Grain Cu content	SD	Seed diameter
GFD	Grain filling duration	SDS	Sedimentation rate
GFeC	Grain Fe content	SHS	Shattering score
GFR	Grain filling rate	SHZnC	Shoot Zn content
GL	Grain length	SL	Spike length
GL/GW	Grain length/grain width ratio	SN	Spike number
GMnC	Grain Mn content	SNS	Spikelet number per spike
GN	Grain number	SW	Spike weight
GPC	Grain protein content	TKW	Thousand kernel weight
GPL	Grain perimeter length	TMRS	Time to maximum rate of senescence
GSeC	Grain Se content	TN	Tiller number/m^2
GW	Grain width	UIH	Uppermost internode height
GWe	Grain weight/ear	WGC	Wet gluten content
GWs	Grain weight/spike	ZnE	Zn efficiency

investigated in rice to describe the functional CGs based on their reported functions in wheat or rice.

Identification of Traits Within MQTLs

To analyze traits within the MQTL regions, the MQTL results were converted into binary scores (0 or 1) on the basis of the absence/presence of an individual trait-QTL within an MQTL region. We tabulated the number of times a trait was present within an MQTL, the number of QTLs for a trait present within an MQTL (implying confirmation of the QTL), and the number of times the traits were able to be co-localized within an MQTL. A chi-squared test with one degree of freedom was performed to determine traits showing significant co-localization with grain protein content (GPC), grain zinc content (GZnC), grain Fe content (GFeC), and grain yield (GY) beyond what would be expected for a random distribution of QTL within MQTL throughout the genome. The expected number of MQTL associated with a trait and each GPC, GZnC, GFeC, and GY were separately calculated by multiplying the number of observed MQTL for a trait by the proportion of MQTL containing GPC, GZnC, GFeC, and GY QTL(s). Traits within the MQTL regions were also analyzed using IBM SPSS Statistics v.24. The simple regression analysis was performed using Minitab v. 18 to determine the effect of MQTLs on the association of GY, GPC, GFeC, and GZnC.

MQTLs and GWAS Comparison

The detected MQTLs were compared with the significant loci associated with different quantitative traits identified in wheat genome-wide association studies (GWAS). The mapped

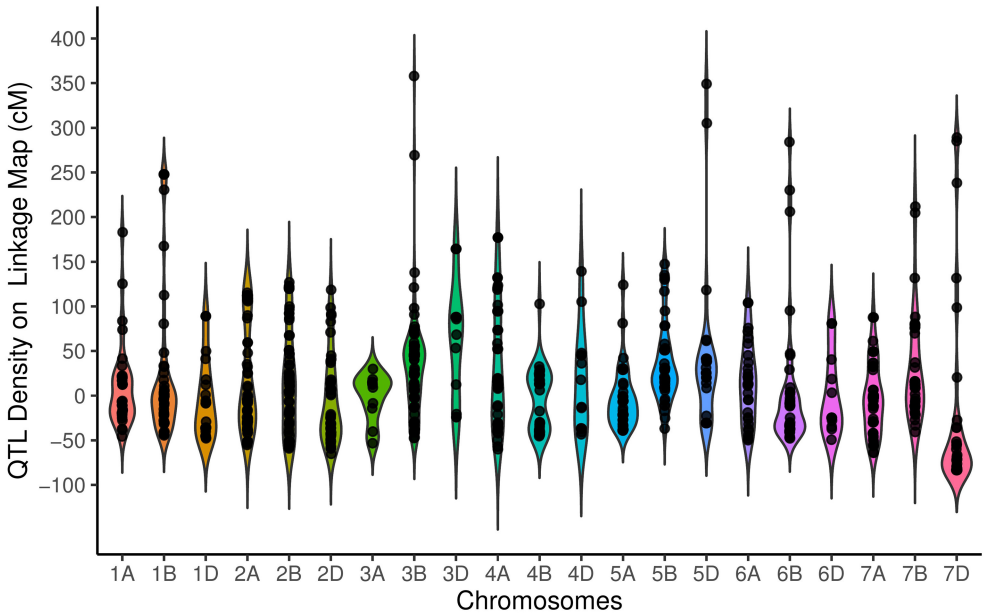

FIGURE 1 | Distribution of quantitative trait loci (QTLs) for traits on all chromosomes represented as the number of QTLs per distance (50 cM), starting from the centromeric region of each chromosome where it was considered at the position 0 cM. Each dot represents the exact location of each QTL.

coordinates of the identified significant loci through GWAS were compared to those found with the MQTL analysis.

Orthologous MQTL

Due to the high synteny among genes in Poacaea, the most stable and promising wheat MQTLs were evaluated for the detection of the orthologous MQTLs (OrMQTL) in rice (Lei et al., 2018; Raza et al., 2019; Khahani et al., 2020), and maize (Semagn et al., 2013; Wang et al., 2013, 2016). A set of orthologous genes at the MQTLs regions was considered as a criterion of a syntenic region using EnsemblPlants (http://plants.ensembl.org/index.htmldatabase).

RESULTS

Genomic Quantitative Trait Loci Distributions

The individual traits of quantitative trait loci (QTLs) used in this study are listed in **Supplementary Table 2**. The QTLs for thousand kernel weight (TKW) (18.03%), grain yield (GY), (13.11%), and a number of grains per spike (NS) (13.11%) were the most frequently reported agronomic QTLs identified in the tested mapping populations. The grain Fe content (GFeC) (33.33 %) and grain Zn content (GZnC) (28.21 %) of QTLs for micronutrient traits and grain protein content (GPC) (64.00 %) and sudden death syndrome (SDS) (10.00 %) of QTLs for quality traits were frequent.

Our results suggest a non-random distribution of QTLs within the wheat genome. Distribution of QTLs on the basis of physical size [$\chi^2_{(2)} = 60.17$, $P = 8.58E{-}14$] showed that 254, 326, and 155 QTLs were located on the A, B, and D genomes, respectively. The QTL distribution was significantly different among the seven chromosome groups [$\chi^2_{(6)} = 47.12$, $P = 1.77E{-}8$], ranging from as few as 76 QTLs on group 6 to as many as 165 QTLs on Group 2. Chromosome 3B with 75 QTLs had the highest number of QTLs, followed by chromosome 2B (62 QTLs) and 2A (61 QTLs), while chromosome 3A with 10 QTLs had the lowest QTL. The distribution of QTL over the genetic linkage map with respect to centromeric and telomeric regions was distinctly non-random (**Figure 1**). The non-telomeric region of each chromosome (−50 up to + 50 cM intervals) had the highest number of QTLs (**Figure 1**). We did not detect QTLs at 100 cM for the tested traits.

The distribution of meta QTLs (MQTLs) indicated that a cluster of MQTLs was mapped to the non-telomeric regions. Besides, MQTL_5B_4 and MQTL_6B_1 were located near the centromeric region of chromosomes 5B and 6B, respectively (**Figure 2**). There was a significant correlation between the number of initial QTLs and MQTLs ($r = 0.46$, $P < 0.03$). The number of MQTLs per chromosome varied from two (chromosomes 3A and 5D) to eight (chromosomes 1B, 2A, and 6B).

Meta-QTL Analysis

Of the 735 initial QTLs, 449 were successfully projected onto the genetics consensus map and used in the meta-QTL analysis (**Figure 3**). A total of 100 MQTLs were detected and the number of individual QTL per MQTL ranged from 1 to 43 (**Figures 4**, **5**; **Supplementary Figure 1**). The number of traits present per MQTL region ranged from 1 to 18. Among the identified MQTLs, MQTL_3B_1 that contained 43 QTLs had the highest number of initial QTLs followed by MQTL_7A_3 with 29 initial QTLs (**Supplementary Table 3**). These two MQTLs can be considered as the most stable QTLs under different experimental conditions. The detailed information of MQTLs consisted of the chromosome number, position, confidence interval (CI), flanking

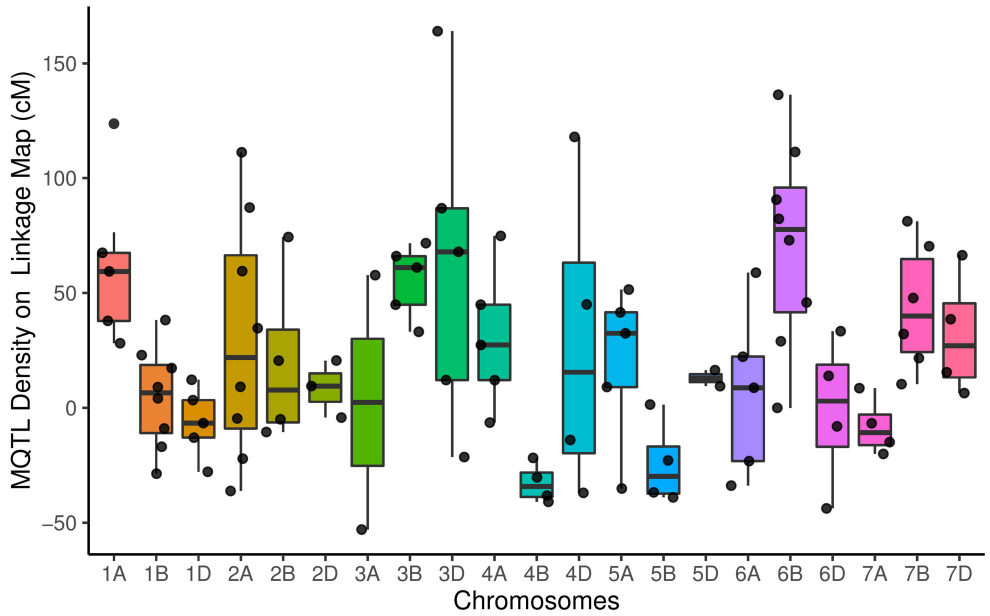

FIGURE 2 | Distribution of meta-quantitative trait loci (MQTLs) for assessed traits on all chromosomes represented as a number of MQTLs per distance (50 cM), starting from the centromeric region of each chromosome where it was considered at position 0 cM. Each dot represents the exact location of each MQTL.

markers, and traits which are shown in **Table 2**. The higher marker density of the consensus map compared with the lower marker density in the independent linkage maps helped to reduce the CI of QTLs up to three-fold with an average of 4.63 cM in MQTLs compared with the mean CI of 13.73 cM for the original QTLs. Among the detected MQTLs, the CI of 11 MQTLs was reduced up to <1 cM (**Table 2**).

Functional Identification of Candidate Genes

The genomic positions of the stable MQTLs and the number of functional candidate genes (CGs) in their intervals are reported in **Table 3**. The range for the number of the CGs annotated in the tested meta-QTLs was between 20 and 802. The MQTL_4D_1, MQTL_4B_3, MQTL_3B_1, and MQTL_5A_4 harbored the greatest number of CGs. Among the detected CGs on MQTL regions, several well-known genes including *psbL* (21.896318-21.896434 Mb), *psbT* (21.905522-21.905638 Mb), *rpl33* (21.899696-21.899896 Mb), and *rps4* (24.160684-24.161289 Mb) genes were located in the MQTL_3D_4 region. In addition, the *miR166* gene was detected in the MQTL_4D_1 (450.196302-450.196403 Mb) and MQTL_7A_4 (561.152174-561.152339 Mb) regions. Furthermore, several CGs with unknown annotation in wheat and were orthologous to genes in rice were identified (**Supplementary Tables 4**, **5**). Overall, some CGs, such as *TraesCS2A02G141400, TraesCS3B02G040900, TraesCS4D02G323700, TraesCS3B02G077100,* and *TraesCS4D02G290900* were uncovered with a possible role in micronutrient contents, yield, and yield-related traits.

Traits Analysis Within MQTLs

Our data show the unequal contribution of individual QTL across the detected MQTLs (**Figure 6** and **Supplementary Table 3**). Individual QTL for TKW were present in 41 of the 100 MQTL regions, the most for any agronomic trait, followed by GY (38 MQTL). Among quality traits, QTLs for GPC was the most distributed QTLs which were located among 19 MQTLs. The GFeC and GZnC QTLs were presented in 19 and 12 MQTLs, respectively (**Supplementary Table 3**).

Analysis for the co-localization of QTLs revealed that QTLs for GY and TKW were frequently co-localized with QTLs of the target traits (GPC, GFeC, and GZnC). The QTLs for TKW showed 52% co-localizations with QTLs for GPC (**Supplementary Table 6**). The GY QTLs showed 52% co-localization with the QTLs for GFeC. Results also indicated 66% co-localization between grain Fe and grain Zn QTLs. Co-localization frequency for GY (R^2 = 80.81%) was strongly associated with the total number of MQTL for a trait. The association of co-localization frequency with the overall number of MQTL for a trait was relatively strong for GPC (R^2 = 65.58%), GFeC (R^2 = 58.74%), and GZnC (R^2 = 56.00%). For most traits, association with target traits (GPC, GZnC, GFeC, and GY) did not differ from the expected on the basis of chi-squared analysis. However, the association of traits with GY, GPC, GFeC, and GZnC as target traits was more than expected (**Supplementary Table 6**).

Comparison of the Identified MQTLs and QTL Mapping in the Wheat GWAS

The comparison of the MQTL locations with genome-wide association studies (GWAS) QTL regions showed that 21 significant signals (SNPs-linked QTLs) of the available wheat

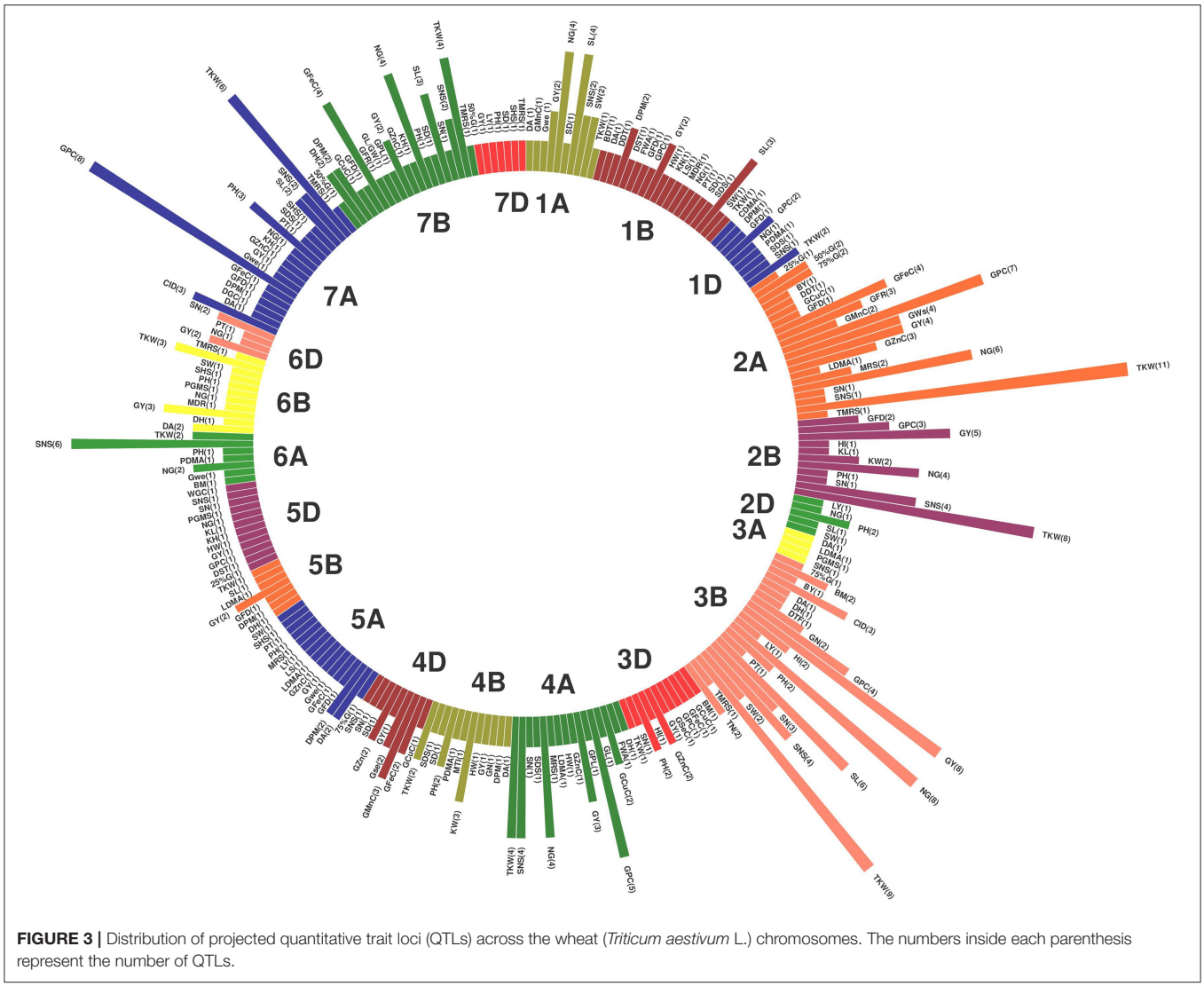

FIGURE 3 | Distribution of projected quantitative trait loci (QTLs) across the wheat (*Triticum aestivum* L.) chromosomes. The numbers inside each parenthesis represent the number of QTLs.

GWAS map were co-located with MQTLs of seven of the 35 traits tested in our study (**Table 4**). The results indicated the co-localization of significant single-nucleotide polymorphisms (SNPs) for GY (6 SNPs), a number of grains per spike (NG) (1 SNPs), plant height (PH) (4 SNPs), spike length (SL) (2 SNPs), spike number (SN) (1 SNPs), spike number per spike (SNS) (4 SNPs), and thousand kernel weight (TKW) (3 SNPs) traits in the wheat GWAS with the identified MQTLs in our study. For instance, the MQTL_4A_3, MQTL_4A_4, and MQTL_4B_3 identified for TKW in our study were positioned in the genomic regions of the major signal reported for TKW in the wheat GWSA map. Overall, the co-located MQTLs and significant GWAS signals were distributed on chromosomes 1B, 2B, 3B, 4A, 4B, 4D, 5A, 7A, and 7B (**Table 4**).

Orthologous MQTL Mining of Wheat in Rice and Maize

The comparative analysis for QTLs in wheat, rice, and maize resulted in the identification of orthologous MQTL (OrMQTL). Nine OrMQTLs were detected for wheat and rice including five OrMQTL for GY and two for PH, and GFeC/GZnC, respectively. Moreover, seven OrMQTL were identified in wheat and maize consisting of six and one OrMQTL for GY and PH, respectively. Among the uncovered OrMQTLs, the OrMQTL_10 was a cross-species QTL in wheat, rice, and maize (**Figure 7**; **Table 5**). The MQTL_7B_2, MQTL_3B_4, MQTL_4D_1, and MQTL_5A_4 for GY in wheat were in the co-linear regions for rice GY MQTLs on chromosome 6 (MQTL6-2), 1 (MQTL-YLD3), 3 (MQTL-YLD9), and 11 (MQTL-YLD19), respectively (**Table 5**). The wheat MQTL_4B_1 and MQTL_7B_5 were in the co-linear region of MQTLs for PH on chromosome 3 (MQTL-PH11) and 10 (MQTL-PH26) in rice (**Table 5**). In addition, wheat MQTL_2A_1, MQTL_4D_1 and MQTL_4D_1 were in the co-linear regions of MQTLs of GFeC and GZnC on chromosome 7 (rMQTL7.1), 6 (rMQTL6.3) and 7 (rMQTL7.2) in rice (**Table 5**). The MQTL_1B_1, MQTL_2A_1, MQTL_4A_3, MQTL_4B_2, and MQTL_4B_3 on wheat chromosomes 1B, 2A, 4A, and 4B

were in located in the co-linear regions of the MQTLs of GY on chromosomes 1 (MQTL7), 5 (MQTL29), 1 (MQTL10), 8 (MQTL44), and 2 (MQTL23) in maize (**Table 5**). Furthermore, an OrMQTL for MQTL_3D_3 on wheat chromosome 3D was detected on chromosome 10 (MQTL107) harboring the MQTL for PH in maize (**Table 5**). One of the wheat MQTLs (MQTL_7B_3) on chromosome 7B was located in the syntenic region of maize MQTLs chromosome 5 (MQTL5.5) and 6 (MQTL66), respectively, as well as rice MQTLs on chromosome 6 (MQTL-YLD14) for GY (**Figure 7**; **Table 5**). The genes located at the OrMQTL regions and their annotations were reported in **Supplementary Table 7**. Interestingly, the well-known and proved orthologous genes including *OsCYP72A18*, *OsFbox146*, *NAC22*, *SD37*, *BRD2*, *OsMAPK4,* and *ZmMPK5* were identified in rice and maize OrMQTLs.

DISCUSSION

Quantitative Trait Loci and Meta-Quantitative Trait Loci Distribution Over the Wheat Genome

Analysis of the genetic control of quantitative traits is a challenge in plant breeding that is due to the complexity and the

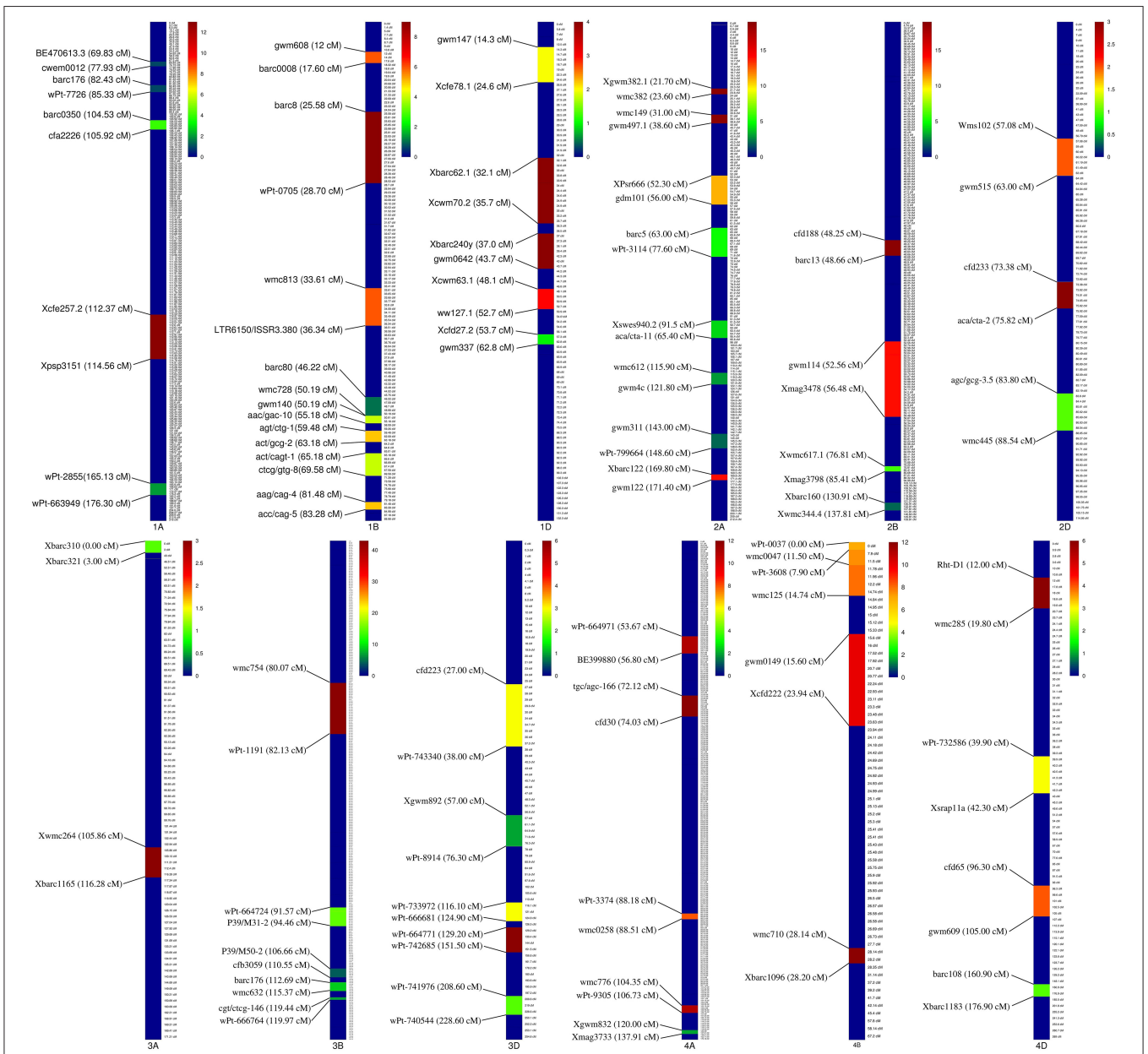

FIGURE 4 | Position of detected meta-quantitative trait loci (MQTLs) on the wheat genome associated with micronutrient content, grain quality, and quantitative traits with 95% confidence interval (CI). Each color in a different linkage group indicates the number of initial QTLs involved in each MQTL. The flanking markers for each MQTL are presented on the left side of the linkage groups.

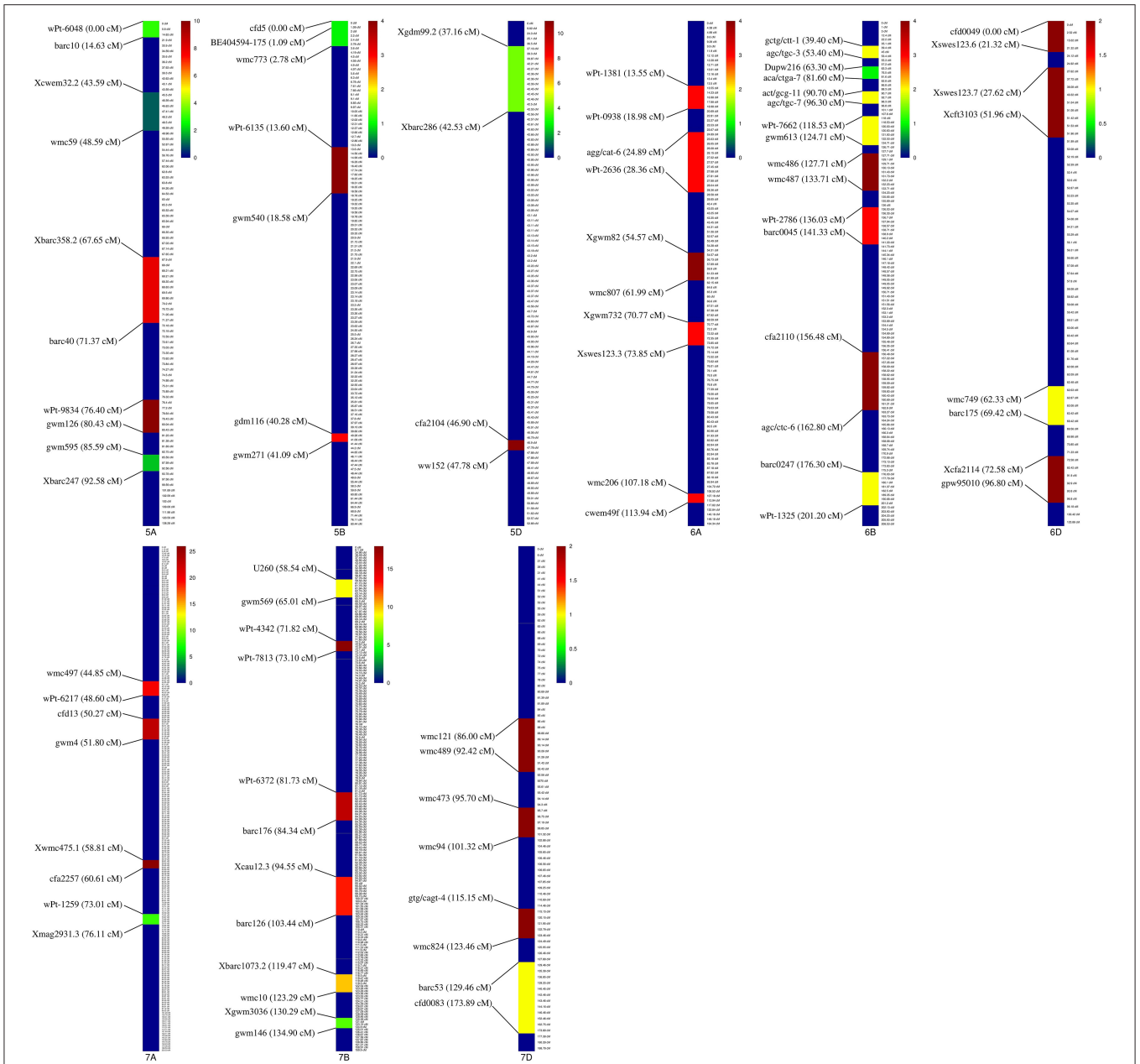

FIGURE 4 | Continued

difficulty of stacking numerous alleles for their control (Izquierdo et al., 2018). Meta-quantitative trait loci (MQTL) analysis made possible the consolidation of individual QTL into MQTL regions. The MQTL analysis and the comparative genomic approaches help to identify consensus QTLs and conserved genes across genomes and species. In the current MQTL analysis, we used information from over 60% of the individual QTLs to identify MQTLs that showed the statistical power of MQTL analysis for combining wheat QTLs that were in the range of 54–84% that have been reported in the previous studies of MQTL in wheat (Löffler et al., 2009; Acuña-Galindo et al., 2015; Soriano and Alvaro, 2019). In this study, we narrowed down the confidence interval (CI) of the detected MQTL regions compared to the initial QTLs used in our MQTL analysis. The efficacy of MQTL analysis in refining the CI for previously known QTLs as well as for validating their effects across different genetic backgrounds and environments is well-demonstrated (Goffinet and Gerber, 2000; Wu and Hu, 2012).

The results showed that QTLs and MQTLs had higher density in the non-telomeric and near-centromeric regions. The QTL

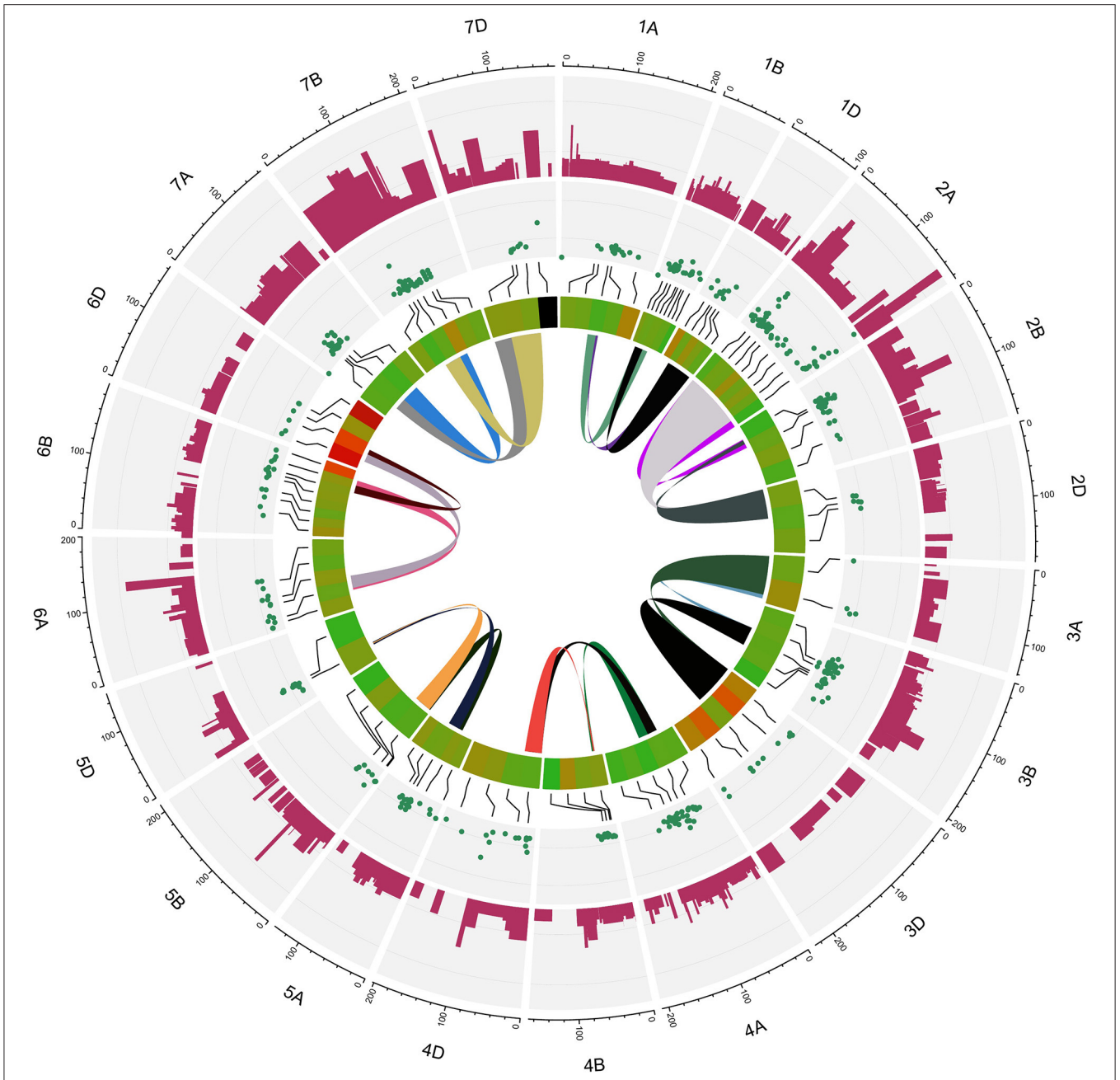

FIGURE 5 | Circus plot showing distribution of QTLs and MQTLs on 21 linkage group of wheat. The outermost circle indicates the length of each linkage map on the consensus genetic map. The second circle indicates the initial QTLs with 95% confidence intervals. The third circle displays the density of QTLs in each MQTLs. The fourth circle displays the position of detected MQTL related to assessed traits based on heatmap illustration. The last inner circle demonstrated common markers links between different genome (A, B, and D) of each linkage groups.

result from the genetic segregation of sequence polymorphisms at functional elements, such as regulatory sequences upstream of genes and/or coding sequences (Flint and Mackay, 2009; Salvi and Tuberosa, 2015). Therefore, it is expected that QTL density on a genetic map is driven by gene density, polymorphism rate at functional sites in genetic regions, and the frequency of recombination. This finding was in line with the result of Martinez et al. (2016), which illustrated higher QTLs densely mapped to the near centromeres on the genetic maps.

In the quality traits category, QTLs for grain protein content (GPC), grain Fe content (GFeC), and grain Zn content (GZnC) traits were the most observed QTLs in the identified

TABLE 2 | Description of detected meta-quantitative trait loci (MQTL).

MQTL name	Chromosome	Flanking markers	Position (cM)	CI (cM)	Individual QTL present in MQTL*
MQTL_1A_1	1A	BE470613.3–cwem0012	74.13	3.6	GWe
MQTL_1A_2	1A	barc176–wPt-7726	83.88	2.9	SNS
MQTL_1A_3	1A	barc0350–cfa2226	105.4	1	GY, SD, SL, SNS, TKW
MQTL_1A_4	1A	Xcfe257.2–Xpsp3151	113.51	2.1	DA, GY, NG, SD, SL, SW, TKW
MQTL_1A_5	1A	wPt-2855–wPt-663949	169.72	8.6	GMnC, SL
MQTL_1B_1	1B	gwm608–barc0008	15.34	4.31	DA, DPM, GFD, HW, SD, SDS, TKW
MQTL_1B_2	1B	barc8–wPt-0705	27.06	2.98	BDT, DA, DPM, GFD, GPC, HW, SDS, SL, TKW
MQTL_1B_3	1B	wmc813–LTR6150/ISSR3.380	35.00	2.75	BDT, GFD, GY, HW, NG, SDS, TKW
MQTL_1B_4	1B	barc80–wmc728	48.11	3.2	BDT, DDT, DST, FWA
MQTL_1B_5	1B	gwm140–aac/gac-10	53.01	3.19	DDT, DST, FWA, GY
MQTL_1B_6	1B	agt/ctg-1–act/gcg-2	61.31	2.49	DDT, DST, FWA, LS, SW
MQTL_1B_7	1B	act/cagt-1–ctcg/gtg-8	66.96	2.39	DDT, DST, FWA, MDR
MQTL_1B_8	1B	aag/cag-4–acc/cag-5	82.21	0.42	DDT, DST, FWA, NG, SL
MQTL_1D_1	1D	gwm147–Xcfe78.1	19.18	9.09	GFD, GPC, SDS, SNS
MQTL_1D_2	1D	Xbarc62.1–Xcwm70.2	34.03	2.94	DPM, GPC, SDS, TKW
MQTL_1D_3	1D	Xbarc240y–gwm0642	40.35	6.75	CDMA, DPM, GPC, PDMA, SDS
MQTL_1D_4	1D	Xcwm63.1–ww127.1	50.36	4.12	NG, PDMA
MQTL_1D_5	1D	Xcfd27.2–gwm337	59.23	2.18	CDMA, PDMA, TKW
MQTL_2A_1	2A	Xgwm382.1–wmc382	22.80	1.4	50%G, 75%G, GFeC, GPC, GY, GZnC, MRS, SNS, TKW, TMRS
MQTL_2A_2	2A	wmc149–gwm497.1	36.87	3.35	50%G, 75%G, GCuC, GFeC, GMnC, GPC, GWs, GY, MRS, NG, SNS, TMRS
MQTL_2A_3	2A	XPsr666–gdm101	54.37	3.09	25%G, 50%G, 75%G, DDT, DTF, GFD, GWs, GY, MRS, NG, TKW
MQTL_2A_4	2A	barc5–wPt-3114	68.17	6.85	BY, GFD, GY, MRS, NG
MQTL_2A_5	2A	Xswes940.2–aca/cta-11	93.68	3.15	GFR, GPC, LDMA, TKW
MQTL_2A_6	2A	wmc612–gwm4c	118.53	4.94	GFR, SN, TKW
MQTL_2A_7	2A	gwm311–wPt-799664	146.19	3	GMnC, NG
MQTL_2A_8	2A	Xbarc122–gwm122	170.28	0.39	GFR, GWs, GY, NG, KW
MQTL_2B_1	2B	cfd188–barc13	48.48	0.42	GFD, GY, KL, KW, NG, PH, SNS, TKW
MQTL_2B_2	2B	gwm114 -Xmag3478	54.01	2.68	GY, KL, KW, NG, SN, SNS, TKW
MQTL_2B_3	2B	Xwmc617.1–Xmag3798	79.57	4.22	GPC, HI, SNS, TKW
MQTL_2B_4	2B	Xbarc160–Xwmc344.4	133.37	1.29	GY, NG
MQTL_2D_1	2D	wms102–gwm515	60.77	3.82	PH, SL, SW
MQTL_2D_2	2D	cfd233–aca/cta-2	74.48	1.97	NG, LY, PH
MQTL_2D_3	2D	agc/gcg-3.5–wmc445	85.6	3.58	PH
MQTL_3A_1	3A	Xbarc310–Xbarc321	1.00	3	LDMA
MQTL_3A_2	3A	Xwmc264–Xbarc1165	111.77	7	DA, PGMS, SNS
MQTL_3B_1	3B	wmc754–wPt-1191	81.11	1.99	BM, CID, DA, DTF, GN, GPC, GY, HI, LY, NG, PH, SL, SN, SNS, SW, TKW, TN
MQTL_3B_2	3B	wPt-664724–P39/M31-2	92.92	2.65	GY, HI, LY, NG, PH, SL, SN, SNS, TKW, TN
MQTL_3B_3	3B	P39/M50-2–cfb3059	109.1	2.07	75%G, SNS, TKW
MQTL_3B_4	3B	barc176–wmc632	114.03	2.38	75%G, GY, NG, SNS, TKW, TMRS
MQTL_3B_5	3B	cgt/ctcg-146–wPt-666764	119.72	0.38	75%G, DH, GY, SL, SNS, TKW, TMRS
MQTL_3D_1	3D	cfd223–wPt-743340	32.58	9.44	GZnC, SN, TKW
MQTL_3D_2	3D	Xgwm892–wPt-8914	66.11	14.3	PH
MQTL_3D_3	3D	wPt-733972–wPt-666681	121.94	3.88	BM, GCuC, GFeC, GPC, GY, HI, PH
MQTL_3D_4	3D	wPt-664771–wPt-742685	140.91	13.53	BM, GFeC, GPC, GY, HI, PH
MQTL_3D_5	3D	wPt-741976–wPt-740544	218.00	9.27	GSeC, GZnC
MQTL_4A_1	4A	wPt-664971–BE399880	54.58	1.71	GL, GPC, GPL, GY, GZnC, NG, SNS
MQTL_4A_2	4A	tgc/agc-166–cfd30	73.07	1.94	DH, GL, GPL, GY, GZnC, HW, NG, SDS, SN, SNS, TKW

(Continued)

TABLE 2 | Continued

MQTL name	Chromosome	Flanking markers	Position (cM)	CI (cM)	Individual QTL present in MQTL*
MQTL_4A_3	4A	wPt-3374–wmc0258	88.37	0.3	FWA, GL, GPC, GPL, HW, MDR, NG, SDS, TKW
MQTL_4A_4	4A	wmc776–wPt-9305	105.92	1.62	FWA, GCuC, GL, GPL, HW, LDMA, NG, SDS, SNS, TKW
MQTL_4A_5	4A	Xgwm832–Xmag3733	135.87	0.71	FWA, GCuC
MQTL_4B_1	4B	wPt-0037–wmc0047	9.07	4.66	DA, DPM, HW, MTI, PH, SD, SDS
MQTL_4B_2	4B	wPt-3608–wmc125	11.84	3.76	DA, DPM, HW, MTI, PH, SD, SDS
MQTL_4B_3	4B	gwm0149–Xcfd222	19.71	7.9	GN, GY, HW, KW, MTI, PH, SD, SDS, TKW
MQTL_4B_4	4B	wmc710–Xbarc1096	28.20	0.01	GN, HW, KW, MTI, PDMA, PH, SD, SDS, TKW
MQTL_4D_1	4D	Rht-D1–wmc285	17.98	1.87	GCuC, GFeC, GMnC, GY, SN, SNS
MQTL_4D_2	4D	wPt-732586–Xsrap11a	41.00	2.06	GFeC, GMnC, GZnC, SN
MQTL_4D_3	4D	cfd65–gwm609	100.02	5.4	GFeC, GMnC, GSeC, SD
MQTL_4D_4	4D	barc108–Xbarc1183	172.97	6.6	GFeC, GZnC
MQTL_5A_1	5A	wPt-6048–barc10	1.87	4.02	DA, DPM, PT
MQTL_5A_2	5A	Xcwem32.2–wmc59	46.08	5.03	DPM, LDMA
MQTL_5A_3	5A	Xbarc358.2–barc40	69.46	3.35	75%G, DA, DPM, GFD, MRS, PH, SHS, SW
MQTL_5A_4	5A	wPt-9834–gwm126	78.56	3.22	75%G, DA, DPM, GFeC, GWe, GY, LS, LY, MRS, SHS
MQTL_5A_5	5A	gwm595–Xbarc247	88.51	5.49	GFeC, GZnC
MQTL_5B_1	5B	cfd5–BE404594-175	0.00	1.65	DPM
MQTL_5B_2	5B	BE404594-175–wmc773	2.20	1.07	GY
MQTL_5B_3	5B	wPt-6135–gwm540	16.10	4.55	DH, GY, SL, TKW
MQTL_5B_4	5B	gdm116–gwm271	40.43	0.28	DH, GFD, GY, LDMA
MQTL_5D_1	5D	Xgdm99.2–Xbarc286	40.43	4.17	25%G, KL, PGMS, SNS
MQTL_5D_2	5D	cfa2104–ww152	47.41	0.34	25%G, DST, GPC, HW, GY, KH, KL, NG, PGMS, SN, WGC
MQTL_6A_1	6A	wPt-1381–wPt-0938	16.14	5.04	GWe, PH, TKW
MQTL_6A_2	6A	agg/cat-6–wPt-2636	26.8	2.68	NG, PDMA, SNS
MQTL_6A_3	6A	Xgwm82–wmc807	58.75	5.2	BM, NG, SNS
MQTL_6A_4	6A	Xgwm732–Xswes123.3	72.28	2.11	SNS
MQTL_6A_5	6A	wmc206–cwem49f	108.86	3.32	SNS, TKW
MQTL_6B_1	6B	gctg/ctt-1–agc/tgc-3	48.00	7	GY, MDR
MQTL_6B_2	6B	Dupw216–aca/ctga-7	77.00	5.1	DH
MQTL_6B_3	6B	act/gcg-11–agc/tgc-7	93.87	2.71	SHS, SW
MQTL_6B_4	6B	wPt-7662–gwm613	120.97	4.47	DA, PH
MQTL_6B_5	6B	wmc486–wmc487	130.30	5.1	PGMS, PH, TKW
MQTL_6B_6	6B	wPt-2786–barc0045	138.66	4.87	GY, PGMS, TKW
MQTL_6B_7	6B	cfa2110–agc/ctc-6	159.43	5.9	GY, PGMS, TKW, TMRS
MQTL_6B_8	6B	barc0247–wPt-1325	184.32	15.62	NG, PGMS
MQTL_6D_1	6D	cfd0049–Xswes123.6	8.23	19.83	PT, SN
MQTL_6D_2	6D	Xswes123.7–Xcft3103	43.93	15.84	GY, SN
MQTL_6D_3	6D	wmc749–barc175	65.91	6.6	GY, NG, SN
MQTL_6D_4	6D	Xcfa2114–gpw95010	85.38	21.45	GY, SN
MQTL_7A_1	7A	wmc497–wPt-6217	45.93	1.82	CID, DA, DGC, DPM, GFD, GFeC, GPC, GZnC, KH, NG, PH, PT, SDS, SL, SNS, TKW
MQTL_7A_2	7A	cfd13–gwm4	51.07	1.48	CID, DA, DGC, DPM, GFD, GFeC, GPC, GZnC, KH, LS, NG, PH, PT, SDS, SHS, SNS, TKW, TMRS
MQTL_7A_3	7A	Xwmc475.1–cfa2257	59.28	0.75	CID, DA, DGC, DPM, GFD, GFeC, GPC, GY, GZnC, KH, NG, PH, PT, SDS, SHS, SL, SNS, TKW
MQTL_7A_4	7A	wPt-1259–Xmag2931.3	74.58	3.1	DGC, GPC, KH, PH, SDS, TKW
MQTL_7B_1	7B	U260–gwm569	61.32	4.04	50%G, DH, DPM, GFeC, GL/GW, GPL, NG, SL
MQTL_7B_2	7B	wPt-4342–wPt-7813	72.66	0.89	50%G, DH, DPM, GFD, GFeC, GFR, GL/GW, GPL, GY, GZnC, KH, NG, PH, SD, SL, TKW, TMRS

(Continued)

TABLE 2 | Continued

MQTL name	Chromosome	Flanking markers	Position (cM)	CI (cM)	Individual QTL present in MQTL*
MQTL_7B_3	7B	wPt-6372–barc176	83.18	2.37	50%G, DH, DPM, GCuC, GFeC, GFR, GL/GW, GPL, GY, KH, NG, PH, SL, TKW
MQTL_7B_4	7B	Xcau12.3–barc126	98.82	8.22	50%G, DPM, GFeC, GL/GW, GPL, GY, KH, PH, TKW
MQTL_7B_5	7B	Xbarc1073.2–wmc10	121.38	3.87	GFeC, GL/GW, GPL, KH, NG, SN, SNS, TKW
MQTL_7B_6	7B	Xgwm3036–gwm146	132.26	2.72	GFeC, GL/GW, GPL, NG, SNS, TKW
MQTL_7D_1	7D	wmc121–wmc489	89.41	4.94	50%G, SD, TMRS
MQTL_7D_2	7D	wmc473–wmc94	98.52	5.16	GY, SD
MQTL_7D_3	7D	gtg/cagt-4–wmc824	121.60	3.82	LY, PH
MQTL_7D_4	7D	barc53–cfd0083	149.45	35.94	SHS

** The full name of assessed traits are displayed in **Table 1**.*

MQTLs (**Supplementary Table 2**). In agronomic traits, QTLs for thousand kernel weight (TKW), grain yield (GY), and number of grains per spike (NG) were the most frequent QTLs identified in the MQTLs regions. The outcome of a MQTL analysis in tetraploid wheat revealed that more than 10 loci were associated with TKW (Avni et al., 2018). Besides, in another MQTL analysis in wheat, individual QTLs for TKW, GY, and kernel number had the highest number of individual QTLs in MQTL regions. This result shows the importance of these QTLs for tested traits. The probable assumptions for a higher frequency of QTLs for agronomic traits are easy to measure and frequent data for these traits in different genetic mapping studies. On the other hand, the TKW, GY, NG, GPC, GFeC, and GZnC traits are multi-genic, highly heritable, and relatively insensitive to the environment (especially, TKW) which suggests a high likelihood of different populations carrying different suites of relevant alleles (Cooper et al., 1995; Bezant et al., 1997; Avni et al., 2018; Velu et al., 2018; Zhang et al., 2020).

MQTLs and Functional Candidate Genes

Gene annotation analysis for the MQTL regions helps clarify our understanding of their genetic architecture and refining the targets of breeding for these traits. The results of functional genomics for the identified MQTLs showed that several well-known genes consisted of *psbL* for electron transfer in photosystem II (PSII) (Ozawa et al., 1997), *psbT* encoding a PSII subunit for maintaining optimal PSII activity under adverse growth conditions (Monod et al., 1994), *rpl33*f or structural constituent of ribosome and translation which confers tolerance to cold stress (Rogalski et al., 2008; Moin et al., 2017), and the *rps4* gene for the regulation of translational fidelity in wheat were located in the MQTl_3D_4 region. In addition, the *miR166* gene was detected in the MQTL_4D_1 and MQTL_7A_4 regions. Comparative genomic approaches start with making some form of alignment of genome sequences and looking for orthologous sequences in the aligned genomes and checking to what extent those sequences are conserved. Based on these, genome and molecular evolution are inferred and this may, in turn, be put in the context of phenotypic evolution or population genetics. Analysis of the conservation and diversification of the *miR166* family have shown that *miR166* members play a wide and important regulatory role in seed development. More recently, a short-tandem target mimic (STTM) method was used to verify if *miR166* regulates important agronomic traits in rice (Zhang et al., 2017). In a study, transgenic STTM165/166 plants showed significantly reduced seed number and sterile siliques in Arabidopsis, suggesting that *miR166* plays a vital role in seed development and it might be useful evidence to improve inferior grain size in wheat (Wang et al., 2018).

Among the detected and annotated genes in the MQTLs regions in this study, the *OsMED9* (*TraesCS3B02G077100*), which is an orthologous gene of rice, was identified in the MQTL_3B_1 region that was associated with GY and its components. A diverse array of MED genes has been identified for the regulation of GY and yield components in crop plants (Malik et al., 2016). The results indicated that the MQTL_4D_1 of our study was located in the regions of the rice orthologous *BG1* (*TraesCS4D02G290900*), *OsIDD1* (*TraesCS4D02G262500*), and *OsPAO* (*TraesCS4D02G309000*) genes in wheat. These genes are associated with metal ion binding, flowering time, days to heading, senescence, gravitropism, yield, and yield-related traits (Wu et al., 2008; Liu et al., 2015; Chen et al., 2016; Deng et al., 2017; Mishra et al., 2017). Overexpression of *BG1* leads to larger grain size in rice (Liu et al., 2015) and manipulation of *BG1* increases plant biomass, grain size, and GY in rice and Arabidopsis (Liu et al., 2015; Mishra et al., 2017). *OsIDD1* could rescue the never-flowering phenotype of *rid1* by a transition from vegetative to reproductive growth in rice (Deng et al., 2017), and the *PAOs* protein-encoding genes (Chen et al., 2016) regulate cellular polyamine levels which are critical for embryogenesis (Bertoldi et al., 2004; De-la-Pena et al., 2008), germination (Bethke et al., 2004; Liszkay et al., 2004), root growth (Cona et al., 2005), flowering and senescence (Kakkar and Sawhney, 2002), and mineral deficiency (Moschou et al., 2008, 2009). The orthologous *Ghd7* (*TraesCS5A02G541200*) that was among 427 genes located in the MQTL_5A_4 region in our study involves photoperiodism, flowering, days to heading, plant height (PH), and yield traits. Enhanced expression of *Ghd7* under long-day conditions delays heading and increases PH and panicle size in rice. The *Ghd7* gene plays crucial role in increasing the productivity and adaptability of rice globally (Xue et al., 2008). The uncovered rice orthologous *D27* (*TraesCS7B02G319100*) and *BRD2* (*TraesCS7B02G484200*) genes in wheat in chromosome 7B region are known to control tillering,

TABLE 3 | The genomic position of most stable meta-quantitative trait loci (MQTLs) on the wheat genome and number of genes in their genomic intervals.

MQTL	Chr. No	Genomic position on the wheat genome (Mb[¶])	Number of initial trait QTLs	Number of genes laying at the MQTL interval	MQTL	Chr. No	Genomic position on the wheat genome (Mb[¶])	Number of initial trait QTLs	Number of genes laying at the MQTL interval
MQTL_1A_4	1A	508.04–511.15	8	48	MQTL_4B_1	4B	601.95–640.98	7	348
MQTL_1B_1	1B	16.23–27.30	7	170	MQTL_4B_2	4B	644.87–652.88	7	113
MQTL_1B_2	1B	564.78–571.06	9	53	MQTL_4B_3	4B	509.04–576.50	9	479
MQTL_1B_3	1B	678.45–681.00	7	30	MQTL_4B_4	4B	597.03–608.25	9	105
MQTL_2A_1	2A	77.88–91.56	10	160	MQTL_4D_1	4D	425.23–490.12	6	802
MQTL_2A_2	2A	38.75–56.93	12	189	MQTL_5A_3	5A	439.58–444.92	8	46
MQTL_2A_3	2A	504.28–507.77	11	23	MQTL_5A_4	5A	671.39–702.96	10	427
MQTL_2B_1	2B	686.82–707.65	8	223	MQTL_5D_2	5D	28.96–34.08	11	38
MQTL_2B_2	2B	760.14–762.08	7	28	MQTL_7A_1	7A	1.47–4.99	16	86
MQTL_3B_1	3B	16.67–50.54	18	457	MQTL_7A_2	7A	49.58–53.07	18	30
MQTL_3B_2	3B	784.63–788.79	11	58	MQTL_7A_3	7A	25.06–41.48	18	233
MQTL_3B_4	3B	822.58–830.11	6	112	MQTL_7A_4	7A	560.02–563.50	6	35
MQTL_3B_5	3B	814.18–826.25	7	221	MQTL_7B_1	7B	32.55–36.92	9	55
MQTl_3D_3	3D	31.87–38.29	7	63	MQTL_7B_2	7B	669.71–693.34	17	241
MQTl_3D_4	3D	16.88–30.37	6	250	MQTL_7B_3	7B	557.05–582.70	14	170
MQTL_4A_1	4A	474.79–488.25	7	66	MQTL_7B_4	7B	38.87–41.30	9	20
MQTL_4A_2	4A	151.24–182.24	11	116	MQTL_7B_5	7B	741.47–744.92	8	107
MQTL_4A_3	4A	632.62–656.77	9	204	MQTL_7B_6	7B	729.40–740.72	6	93
MQTL_4A_4	4A	732.61–734.94	10	59	–	–	–	–	–

[¶] *Mb, represents mega base pair.*

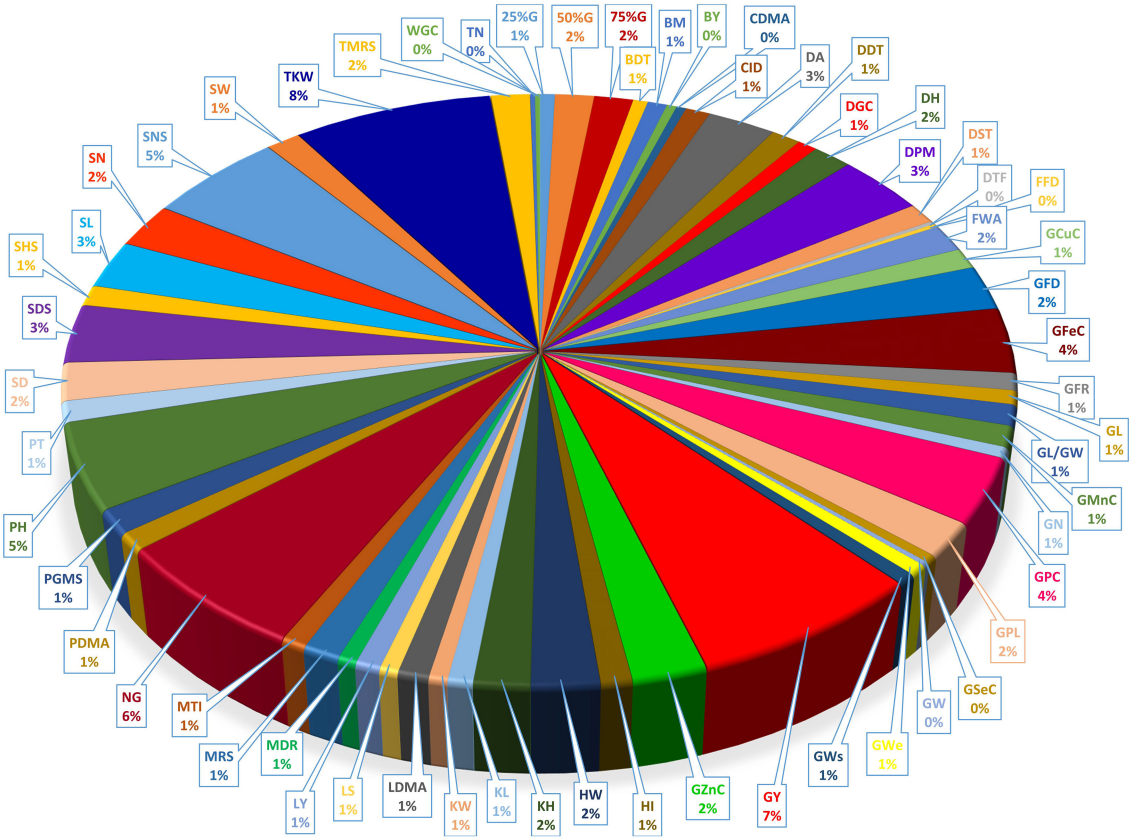

FIGURE 6 | Distribution of QTLs controlling different traits in detected MQTL regions.

heading date, PH, and yield traits in rice (Hong et al., 2005; Lin et al., 2009; Liu et al., 2016).

The identified rice orthologous *D18* (*TraesCS2B02G570800*), *OsRPK1* (*TraesCS5D02G034200*), *DRUS1,* and *DRUS2* (*TraesCS4A02G133800*) genes within the MQTL_2B_2, MQTL_4A_2, and MQTL_5D_2 regions of our MQTLs were associated with PH in wheat (Itoh et al., 2002; Zou et al., 2014; Pu et al., 2017). Manipulating these genes lead to varieties with dwarf and semi-dwarf phenotype and such phenotypes possess short and strong stalks, exhibit less lodging, and a greater proportion of assimilation partitioned into the grain, resulting in further yield increases (Hedden, 2003).

The identified rice orthologous *OsRLCK189* (*TraesCS1A02G317300*), *OsGH3-4* (*TraesCS1A02G320200*), *FT-L* (TraesCS2B02G511400), *OsRLCK57* (*TraesCS3B02G600300*), *OsHK1* (*TraesCS2B02G495500*), *AIM1* (*TraesCS3D02G077200*), and *WOX11* (*TraesCS2A02G100700*) genes in wheat was located in MQTL_1A_4, MQTL_2A_2, MQTL_2B_1, MQTL_3B_5, and MQTl_3D_3 regions show a role in root growth and inflorescence and floral development (Richmond and Bleecker, 1999; Izawa et al., 2002; Vij et al., 2008; Ogiso-Tanaka et al., 2013; Zhao et al., 2015, 2020; Cheng et al., 2016; Xu et al., 2017; Lehner et al., 2018; Kong et al., 2019). Inflorescence development in cereals directly affects grain number and size which are key determinants of yield (Yamburenko et al., 2017).

The rice orthologous *OsMTP12* (*TraesCS2A02G141400*), *OsMTP9* (*TraesCS3B02G040900*), and *OsMTP1* (*TraesCS4D02G323700*) belonging to the metal tolerance protein (MTP) gene family were identified in the MQTL_2A_1, MQTL_3B_1, and MQTL_4D_1 interval of the wheat genome in the present study. The MTP gene family plays a critical role in metal transport, mainly in Zn, Mn, Fe, Cd, Co, and Ni, metal homeostasis, and tolerance (Gustin et al., 2011; Zhang and Liu, 2017; Ram et al., 2019). The increased expression of MTP genes during seed development and their potential role in metal homeostasis during the seed filling stage had been documented (Ram et al., 2019). The *MTP1* and *MTP12* genes share a characteristic histidine-rich loop toward the c-terminal, which is known to have a role in Zn-binding (Ram et al., 2019).

The MQTL_1A_4, MQTL_3B_5, MQTL_4B_1, MQTL_4B_2, MQTL_4B_3, MQTL_4B_4, MQTL_7A_1, and MQTL_7A_4 interval regions harbored rice orthologous *VAL1* (*TraesCS4B02G314600*), *PROG1* (*TraesCS4B02G354000*), *D14* (*TraesCS4B02G258200*), *LPL2* (*TraesCS4B02G308000*), *OsINV3* (*TraesCS7A02G009100*), and *OsRLCK218* (*TraesCS7A02G385300*) genes with a role in diverse development and growth traits including leaf development, PH, shoot branching, panicle and tiller number, grain number, grain and spikelet size, and GY (Vij et al., 2008; Zhou et al., 2016; Morey et al., 2018; Wu et al., 2018; Yao et al., 2018; Zhang et al., 2018;

TABLE 4 | The collinear meta-quantitative trait loci (MQTLs) with the significant loci in wheat genome-wide association studies.

Trait*	Wheat MQTL	Chromosome (genomic position in Mb)	SNP marker name (genomic position in Mb)	Wheat GWAS references
GY	MQTL_3B_1	3B (16.67–50.54 Mb)	AX_109881378 (20.5–22.0)	Li et al., 2019
	MQTL_4A_2	4A (151.24–182.24 Mb)	M8680 (157.56)	Mathew et al., 2019
	MQTL_5A_4	5A (671.39–702.96 Mb)	AX-110458478 (692.17)	Hu et al., 2020
			AX-108839508 (692.16)	
			AX-109388349 (692.18)	
			AX-108829895 (692.39)	
NG	MQTL_7B_2	7B (669.71–693.34 Mb)	S7B_687521301 (687.52)	Jamil et al., 2019
PH	MQTL_4D_1	4D (425.23–490.12 Mb)	AX-95235641 (442.17)	Hu et al., 2020
	MQTL_7B_2	7B (669.71–693.34 Mb)	AX-110149206 (676.25)	Hu et al., 2020
			AX-95658823 (675.28)	
			AX-95149761 (680.08)	
SL	MQTL_1B_2	1B (564.78–571.06 Mb)	AX-109901032 (566.19)	Li Q. et al., 2020
	MQTL_7A_3	7A (25.06–41.48 Mb)	AX-109394807 (29.32)	Hu et al., 2020
SN	MQTL_4A_2	4A (151.24–182.24 Mb)	AX-109066809 (179.94)	Li Q. et al., 2020
SNS	MQTL_2B_1	2B (686.82–707.65 Mb)	AX-111551006 (706.93)	Li Q. et al., 2020
	MQTL_4D_1	4D (425.23–490.12 Mb)	AX-169338181 (433.00)	
			AX-111020167 (471.23)	
	MQTL_7A_1	7A (1.47–4.99 Mb)	AX-111542213 (4.55)	
TKW	MQTL_4A_4	4A (732.61–734.94 Mb)	S4A_733664972 (733.66)	Jamil et al., 2019
	MQTL_4A_3	4A (632.62–656.77 Mb)	AX-111600193 (642.37)	Li Q. et al., 2020
	MQTL_4B_3	4B (509.04–576.50 Mb)	AX-94402252 (564.39)	Hu et al., 2020

Full names of traits are displayed in **Table 1.*

Deng et al., 2020). These candidate genes (CGs) at the detected MQTL regions can potentially have the same function as their orthologous varieties in rice and therefore regulate various developmental and growth-related traits in wheat. Identification of these confines stable chromosomal regions and CGs that influence economically important traits in wheat can help to expedite wheat improvement in future breeding programs.

The MQTLs detected in this study will help identify CGs in these regions responsible for desirable traits and generate allele-specific markers through allele mining (Leung et al., 2015; Ogawa et al., 2018) for marker-assisted selection (MAS) application in pre-breeding population. Allele mining is a promising approach to dissect naturally occurring allelic variation at QTLs/CGs controlling desirable traits (e.g., yield, quality, and micronutrient content) which has potential applications in crop improvement programs (Kumar et al., 2010; Gokidi et al., 2017; Kumari et al., 2018). The data raised from this MQTL study help to refine genomic targets for validation and subsequent development of haplotypic markers in breeding programs. Information of the identified MQTLs can also be used for genetic transformation or allele screening in germplasm collections (ecoTilling) for finding new alleles capable of improving yield, quality, and micronutrient traits (Izquierdo et al., 2018; Belzile et al., 2020). Additionally, a promising application of the variation within these MQTL regions might be their introduction as fixed effects in genomic selection (GS) models to increase the accuracy of the prediction models in their use in wheat breeding programs (Spindel et al., 2016; Izquierdo et al., 2018).

MQTL and Traits Analysis

The QTLs projection on a consensus map allows for the inspection of co-location across traits and categories, which is especially relevant for complex traits (Delfino et al., 2019). The association between trait classes by analyzing the co-localization frequency of individual trait-QTLs demonstrated that TKW with 55% and 63% co-localization frequencies was frequently associated with GY and GPC. In addition, GY and GFeC had the highest co-localization frequency with GFeC (52%) and GZnC (66%), respectively. The results of MQTL analysis of our study confirmed correlations identified for QTLs of TKW and GY (An-Ming et al., 2011; Azadi et al., 2014; Mahdi-Nezhad et al., 2019), TKW and GPC (Wang L. I. et al., 2012; Goel et al., 2019), and GFeC and GZn (Roshanzamir et al., 2013; Pu et al., 2014; Tiwari et al., 2016; Liu et al., 2019) in other studies. Co-localization of QTLs for correlated traits has been identified by Wang et al. (2018). Co-localization of QTLs could be due to the pleiotropy or the presence of different linked genes in the same regions that can partly explain the correlation that exists between traits (Bhatta et al., 2018). The tightly linked genomic regions or pleiotropy can partly explain the correlation that exists between traits (Crespo-Herrera et al., 2016). The result of co-localization frequency of target traits (GY, GPC, GFeC, and GZnC) with the detected MQTLs in our study showed interacting MQTLs which affects the association of traits at the genomic level. Acuña-Galindo et al. (2015) demonstrated that TKW was most frequently associated with GY in MQTL regions, with 57% co-localization. The co-localization of TKW and GPC QTLs has been observed in the genetic analysis of wheat (Goel et al., 2019). The positive and

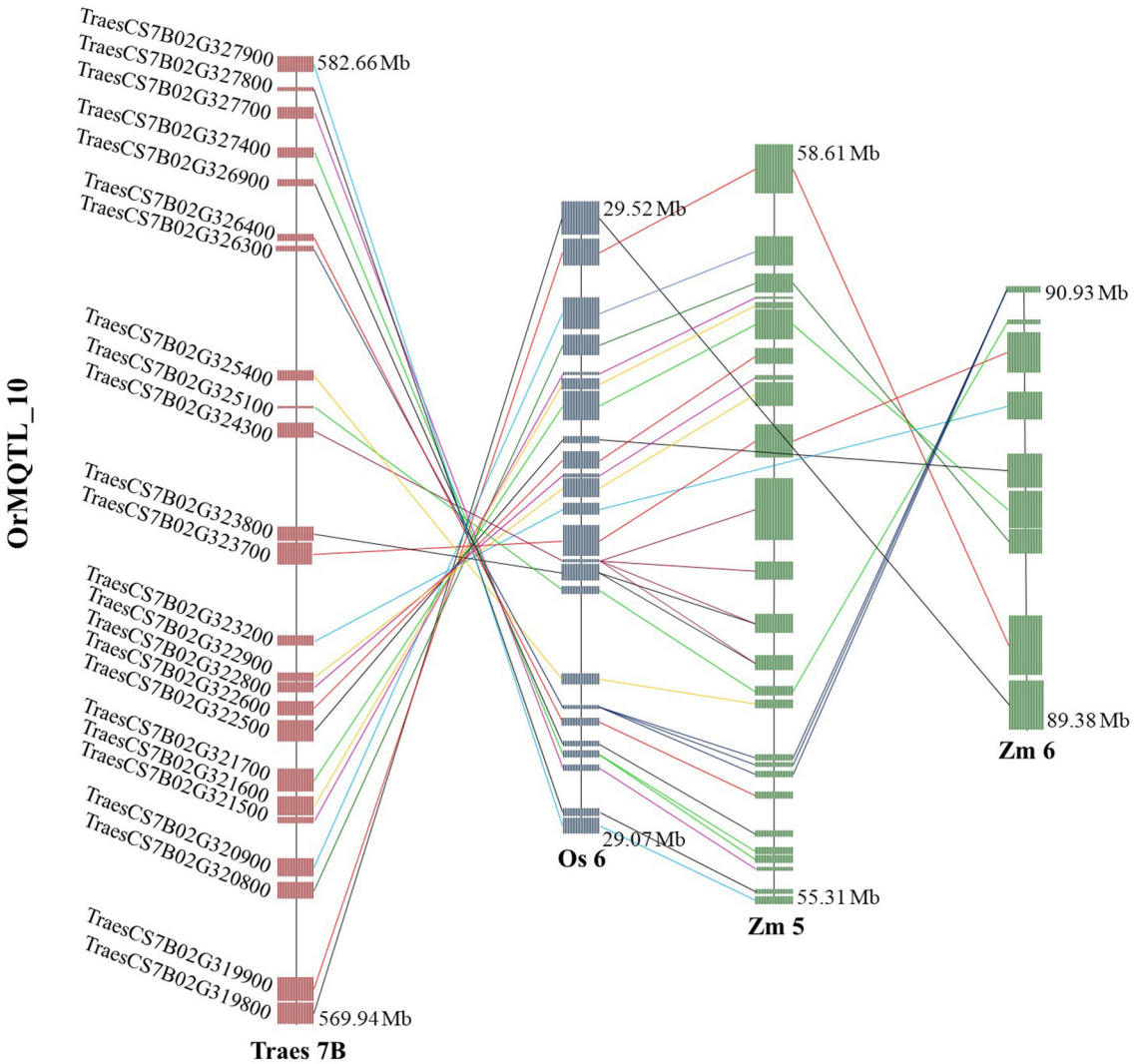

FIGURE 7 | The syntenic region of meta-quantitative trait loci (MQTLs) among the wheat, rice, and maize. Orthologous MQTL (OrMQTL)_10 indicates syntenic regions among identified MQTLs for grain yield (GY) control in wheat (MQTL_7B_3), rice (MQTL-YLD14), and maize (MQTL5.5 and MQTL66). The genomic position, chromosome number, and common genes among the wheat, rice, and maize are indicated.

highly significant associations of GPC and TKW traits have been reported in different wheat populations in other studies (Peleg et al., 2009; Badakhshan et al., 2013; Krishnappa et al., 2017). According to Liu et al. (2019), mineral nutrient concentrations and yield components have shown common QTLs which is in line with co-localized QTLs for grain Zn, Fe QTLs in the tetraploid and hexaploid wheat in other studies (Peleg et al., 2009; Xu et al., 2012; Crespo-Herrera et al., 2016; Krishnappa et al., 2017; Velu et al., 2017). In a recent QTL analysis study on rice, grain Fe and grain yield QTLs were found to be co-localized (Dixit et al., 2019). Crespo-Herrera et al. (2016) suggested the possibility of simultaneous breeding for GFeC and GZnC traits due to their co-localized QTLs. The current co-localization analysis within MQTLs regions indicates the possibility of simultaneous breeding of micronutrient content, grain quality, and GY by pyramiding the QTL regions through marker-assisted selection (MAS). Using -MAS, the specific regions can be transferred to the elite wheat genotypes to simultaneously increase the contents of various traits (Saini et al., 2020). Therefore, an attempt to pyramid QTLs responsible for GY, TKW, GPC, GZnC, and GFeC may accelerate progress in wheat variety development. The QTL pyramiding strategy has been used for simultaneous improvement of traits through MAS in wheat (Wang P. et al., 2012; Feng et al., 2018; Gautam et al., 2020; Muthu et al., 2020). A clear understanding of the co-location of QTLs and their effect on target traits, such as grain Fe and Zn concentrations and yield is very important for using the major effect of QTLs in marker-assisted breeding (Swamy et al., 2018). The results of MQTL analysis in our study showed that the location of 18 MQTLs was in agreement with the position of the QTLs in the meta-QTL studies by Zhang A. et al. (2010), Acuña-Galindo et al. (2015), and Kumar et al. (2020). However, our work adds to this body of genomic mapping research by identifying new MQTL specific for GY, grain quality, and micronutrient content.

TABLE 5 | The OrMQTLs detected in rice and maize according to the syntenic region with MQTLs in wheat.

Trait	Orthologous MQTL (OrMQTL)	Wheat MQTL	Wheat chr. no. (genomic position in Mb)	Rice/Maize original MQTL name	Rice chr. no. (genomic position in Mb)	Maize chr. no. (genomic position in Mb)	Rice/Maize MQTL reference
GY	OrMQTL_1	MQTL_7B_2	7B (674.1981–674.2014)	MQTL6-2	6 (3.1029–3.1075)	–	Lei et al., 2018
	OrMQTL_2	MQTL_3B_4	3B (828.4199–828.7654)	MQTL-YLD3	1 (25.0457–25.0494)	–	Khahani et al., 2020
	OrMQTL_3	MQTL_4D_1	4D (459.6856–459.6870)	MQTL-YLD9	3 (14.6804–14.6827)	–	Khahani et al., 2020
	OrMQTL_4	MQTL_5A_4	5A (700.1772–700.1789)	MQTL-YLD19	11 (10.4749–10.4777)	–	Khahani et al., 2020
	OrMQTL_5	MQTL_1B_1	1B (27.1500–27.1501)	MQTL7	–	1 (224.7216–224.7217)	Wang et al., 2013
	OrMQTL_6	MQTL_2A_1	2A (82.1848–82.1880)	MQTL29	–	5 (19.2226–19.2254)	Wang et al., 2013
	OrMQTL_7	MQTL_4A_3	4A (647.0874–647.0936)	MQTL10	–	1 (275.0211–275.0279)	Wang et al., 2013
	OrMQTL_8	MQTL_4B_2	4B (649.4698–649.4752)	MQTL44	–	8 (47.1438–47.1520)	Wang et al., 2013
	OrMQTL_9	MQTL_4B_3	4B (518.0952–518.0999)	MQTL23	–	2 (122.6755–122.6792)	Wang et al., 2016
	OrMQTL_10	MQTL_7B_3	7B (569.4388–582.6560)	MQTL-YLD14	6 (28.8459–29.5240)	–	Khahani et al., 2020
			7B (568.6510–582.6560)	MQTL5.5	–	5 (55.3131–58.7190)	Semagn et al., 2013
			7B (568.1069–579.8265)	MQTL66	–	6 (89.3129–90.9312)	Wang et al., 2016
PH	OrMQTL_11	MQTL_4B_1	4B (619.5886–635.8693)	MQTL-PH11	3 (1.8624–2.2252)	–	Khahani et al., 2020
	OrMQTL_12	MQTL_7B_5	7B (741.5702–741.5720)	MQTL-PH26	10 (13.3596–13.3632)	–	Khahani et al., 2020
	OrMQTL_13	MQTL_3D_3	3D (33.2949–33.2960)	MQTL107	–	10 (69.4836–69.4849)	Wang et al., 2016
GFeC, GZnC	OrMQTL_14	MQTL_2A_1	2A (88.2949–88.2975)	rMQTL7.1	7 (7.3945–7.3976)	–	Raza et al., 2019
	OrMQTL_15	MQTL_4D_1	4D (484.6841–484.6880)	rMQTL6.3	6 (21.3970–21.3993)	–	Raza et al., 2019
		MQTL_4D_1	4D (461.4899–461.4930)	rMQTL7.2	7 (19.9760–20.0812)	–	Raza et al., 2019

OrMQTL, orthologous MQTL; GY, grain yield; PH, plant height; GFeC, grain Fe content; GZnC, grain Zn content.

Comparative Genomic Analysis and Orthologous MQTL

The genome-wide association studies (GWAS) approach can help in gene discovery and in the analysis of the genetic basis of complex traits for the improvement of wheat. Comparative analysis of the wheat GWAS with the identified current MQTLs suggested the co-localization of 12 MQTLs (MQTL_1B_2, MQTL_2B_1, MQTL_3B_1, MQTL_4A_2, MQTL_4A_3, MQTL_4A_4, MQTL_4B_3, MQTL_4D_1, MQTL_5A_4, MQTL_7A_1, MQTL_7A_3, and MQTL_7B_2) and significant genomic regions of wheat traits that have been identified in the wheat GWAS studies (Jamil et al., 2019; Li et al., 2019; Mathew et al., 2019; Hu et al., 2020; Li X. et al., 2020). Some of the QTL hotspots suggested common genetic markers for wheat traits for further use in marker-assisted breeding that increase genetic gain in breeding programs. The identified OrMQTLs in this study can facilitate the detection of the underlying regulatory genes with evolutionary history and conservative function. The current MQTL analysis defines a genome-wide landscape on the most stable genetic markers and CGs related to micronutrient content, yield, and yield-related traits as the most economically important traits in wheat. The straight-up comparative genomics in our study indicated high synteny between wheat, rice, and maize QTLs especially for GY suggesting possible corroborating evidence of acting the same way in different species that was in line with the results of other MQTL studies (Ahn et al., 1993; Moore et al., 1995; Gale and Devos, 1998; Feuillet and Keller, 1999; Minx et al., 2005). Despite the high interest in the identification of genes involved in GY and yield-related traits in maize and wheat as two economically important crops, the responsible genes have largely remained unknown due to their complex genomes. Given a close evolutionary relation among grass genomes (Gaut, 2002), a synteny analysis of wheat, maize, and rice through the identification of OrMQTLs enabled us to broaden our genetic information and leads to uncover the possible function of unknown CGs among these species. Here we selected the most promising wheat MQTLs to explore their conserved syntenic regions reported in similar MQTLs studies on the same traits in rice and maize to identify OrMQTLs (**Table 5**). For the wheat MQTL_3B_4 of our study, there is an MQTL in the syntenic region in rice (Khahani et al., 2020) controlling OrMQTL_2 containing a rice *OsCYP72A18* orthologous gene (TraesCS3B02G609400 and TraesCS3B02G609600) for grGY (Swamy and Sarla, 2011). Moreover, the wheat MQTL_4D_1 and MQTL_7B_2 of our study were located in the syntenic regions of the rice genome on chromosomes 3 (OrMQTL_3) and 6 (OrMQTL_1), respectively. These two OrMQTLs encompass *OsFbox146* (TraesCS4D02G288900) and *OsFbox297* (TraesCS7B02G405900) orthologous genes for genetic control of GY in rice (Swamy and Sarla, 2011; Lei et al., 2018). Furthermore, the syntenic region of wheat MQTL_4B_1 possessed the two orthologous rice genes, *NAC22* (TraesCS4B02G328600) and *CYP96B4/SD37* (TraesCS4B02G342400) located on chromosome 3 (OrMQTL_11) which control plant growth (Tamiru et al., 2015; Hong et al., 2016). In the syntenic regions of the MQTL_2A_1 and MQTL_4D_1 of our study, there was orthologous *OsZIP1/OsZIP8* (TraesCS2A02G143400) and *OsFerroportin* (TraesCS4D02G323100) genes on chromosomes 7 (OrMQTL_14) and 6 (OrMQTL_15) of rice which are related to Zn and Fe transport/homeostasis (Morrissey and Guerinot, 2009; Bashir et al., 2011; Alagarasan et al., 2017) (**Table 5**).

More intriguingly, the syntenic regions of the wheat MQTL_7B_3 on both chromosome 6 of rice (OrMQTL_10) and chromosome 5 of maize, harbored the rice (*OsMAPK4, Os06g0699400)* and maize (ZmMPK5, Zm00001d014658) and orthologous gene in wheat (TraesCS7B02G322900). The mitogen-activated protein kinase (MAPK) cascades play important roles in regulating plant growth (PH), development, and stress responses (Zhu et al., 2020). The maize *ZmMPK5* is induced by various stimuli, involved in defense signaling pathways in various abiotic/biotic stress, confers tolerance to stresses (Zhang A. et al., 2010; Zhang et al., 2014), and subsequently leads to enhancement of plant growth and yield. Due to the key role of *OsMAPK4* in plant growth, grain development (Liu et al., 2018; Chen et al., 2021), and subsequently in GY, the results of the current syntenic analysis suggest the same function for *TraesCS7B02G322900* and *Zm00001d014658* genes located in OrMQTL_10 interval (**Figure 7**; **Table 5**; **Supplementary Table 7**).

CONCLUSION

The results of this study introduced several novel meta-quantitative trait loci (MQTLs) for improving wheat in multi-purpose breeding programs by identifying key genomic regions associated with agronomic performance, grain quality traits, and micronutrients content. The results of our MQTL analysis have significantly increased the power and precision of our ability to map wheat traits and will provide greater resolution for future fine mapping and marker development. Importantly, our data identify co-localization between grain yield (GY) QTLs with grain Zn content (GZnC), grain Fe content (GFeC), and grain protein content (GPC) QTLs that suggest an opportunity for simultaneous breeding for these traits. This study also provides an example for the value of comparative analysis between evolutionarily close cereal species for the identification of genomic regions and candidate genes (CGS) controlling quantitative traits. Our finding shows the utility of MQTL analysis for refining the location of genomic regions associated with a variety of traits and helps understand how their relative map positions can be exploited for crop improvement. Overall, these findings can lead to both increased selection efficiency and accuracy for breeding by providing the basis for marker development in a marker-assisted selection (MAS) program and for identifying a novel source of variation through allele mining efforts in genetic resource collections. Lastly, these refined MQTLs provide the basis for further focus on the genetic mechanisms controlling micronutrients, GY, and quality traits.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

NS conducted bioinformatics analysis and wrote the draft of the article. BH conceived and designed the project, the bioinformatics analysis, and complemented the writing of the article. AT helped in bioinformatics analysis. CR helped with reviewing and providing critical advice on the article. All authors have read and approved the manuscript.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the supports from Shiraz University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.709817/full#supplementary-material

Supplementary Figure 1 | The location of detected MQTLs with 95% confidence interval associated with quantitative traits in wheat chromosomes. The Lines on the left side of the linkage groups indicate the confidence interval (CI) of QTLs. The colored boxes on each linkage groups represent MQTLs region and the colors inside the vertical lines on the left of illustrate the best model of MQTLs. The molecular markers and their genetic distance (cM) over linkage groups are shown on the right side.

Supplementary Table 1 | The information of wheat population used for meta-quantitative trait loci (MQTL) analysis.

Supplementary Table 2 | The initial and projected QTL data on the genetic consensus map and identified MQTLs for the studied traits.

Supplementary Table 3 | The information of QTLs in the detected MQTLs.

Supplementary Table 4 | The list of uncovered candidate genes (CGS) and all annotated genes anchoring at each MQTL interval.

Supplementary Table 5 | The list of rice orthologous genes at each MQTL interval.

Supplementary Table 6 | The information of the chi-square test for co-localization frequency analysis of the QTLs of the tested traits in the detected MQTLs regions.

Supplementary Table 7 | The list of orthologous genes located at the syntenic regions of detected ortho-MQTLs between wheat, maize, and rice.

REFERENCES

Abdollahi-Sisi, N., Mohammadi, S. A., and Razeghi, J. (2018). Mapping QTLs for grain yield components in bread wheat under well-watered and rain-fed conditions. *J. Bio. Env. Sci.*13, 306–314.

Acuña-Galindo, M. A., Mason, R. E., Subramanian, N. K., and Hays, D. B. (2015). Meta-analysis of wheat QTL regions associated with adaptation to drought and heat stress. *Crop. Sci.* 55, 477–492. doi: 10.2135/cropsci2013.11.0793

Ahn, S., Anderson, J. A., Sorrells, M. E., and Tanksley, S. D. (1993). Homoeologous relationships of rice, wheat and maize chromosomes. *Mol. Gen. Genet.* 241, 483–490. doi: 10.1007/BF00279889

Alagarasan, G., Dubey, M., Aswathy, K. S., and Chandel, G. (2017). Genome wide identification of orthologous ZIP genes associated with zinc and Iron translocation in Setaria italica. *Front. Plant. Sci.* 8:775. doi: 10.3389/fpls.2017.00775

An-Ming, D. I., Jun, L. I., Fa, C. U., Chun-Hua, Z. H., Hang-Yun, M. A., and Hong-Gang, W. A. (2011). Mapping QTLs for yield related traits using two associated RIL populations of wheat. *Zuo Wu Xue Bao* 37, 1511–1524. doi: 10.3724/SP.J.1006.2011.01511

Arcade, A., Labourdette, A., Falque, M., Mangin, B., Chardon, F., Charcosset, A., et al. (2004). BioMercator: integrating genetic maps and QTL towards discovery of candidate genes. *Bioinformatics* 20, 2324–2326. doi: 10.1093/bioinformatics/bth230

Avni, R., Oren, L., Shabtay, G., Assili, S., Pozniak, C., Hale, I., et al. (2018). Genome based meta-QTL analysis of grain weight in tetraploid wheat identifies rare alleles of GRF4 associated with larger grains. *Genes* 9:636. doi: 10.3390/genes9120636

Azadi, A., Mardi, M., Hervan, E. M., Mohammadi, S. A., Moradi, F., Tabatabaee, M. T., et al. (2014). QTL mapping of yield and yield components under normal and salt-stress conditions in bread wheat (*Triticum aestivum* L.). *Plant Mol. Biol. Rep.* 33, 102–120. doi: 10.1007/s11105-014-0726-0

Badakhshan, H., Moradi, N., Mohammadzadeh, H., and Zakeri, M. R. (2013). Genetic variability analysis of grains Fe, Zn and beta-carotene concentration of prevalent wheat varieties in Iran. *Int. J. Agri. Crop Sci.* 6:57.

Badji, A., Otim, M., Machida, L., Odong, T., Kwemoi, D. B., Okii, D., et al. (2018). Maize combined insect resistance genomic regions and their co-localization with Cell Wall constituents revealed by tissue-specific QTL meta-analyses. *Front. Plant Sci.* 9:895. doi: 10.3389/fpls.2018.00895

Bashir, K., Ishimaru, Y., Shimo, H., Nagasaka, S., Fujimoto, M., Takanashi, H., et al. (2011). The rice mitochondrial iron transporter is essential for plant growth. *Nat. Commun.* 2, 1–7. doi: 10.1038/ncomms1326

Belzile, F., Abed, A., and Torkamaneh, D. (2020). Time for a paradigm shift in the use of plant genetic resources. *Genome* 63, 189–194. doi: 10.1139/gen-2019-0141

Bertoldi, D., Tassoni, A., Martinelli, L., and Bagni, N. (2004). Polyamines and somatic embryogenesis in two Vitis vinifera cultivars. *Physiol. Plant.* 120, 657–666. doi: 10.1111/j.0031-9317.2004.0282.x

Bethke, P. C., Gubler, F., Jacobsen, J. V., and Jones, R. L. (2004). Dormancy of Arabidopsis seeds and barley grains can be broken by nitric oxide. *Planta* 219, 847–855. doi: 10.1007/s00425-004-1282-x

Bezant, H., Laurie, D. A., Pratchett, N., Chojecki, J., and Kearsey, M. J. (1997). Mapping of QTL controlling NIR predicted hot water extract and grain nitrogen content in a spring barley cross using marker-regression. *Plant Breed.* 116, 141–145. doi: 10.1111/j.1439-0523.1997.tb02168.x

Bhatta, M., Baenziger, P., Waters, B., Poudel, R., Belamkar, V., Poland, J., et al. (2018). Genome-wide association study reveals novel genomic regions associated with 10 grain minerals in synthetic hexaploid wheat. *Int. J. Mol. Sci.* 19:3237. doi: 10.3390/ijms19103237

Bhusal, N., Sarial, A. K., Sharma, P., and Sareen, S. (2017). Mapping QTLs for grain yield components in wheat under heat stress. *PLoS ONE* 12:e0189594. doi: 10.1371/journal.pone.0189594

Black, R. E., Victora, C. G., Walker, S. P., Bhutta, Z. A., Christian, P., de Onis, M., et al. (2013). Maternal and child undernutrition and overweight in low-income and middle-income countries. *Lancet* 382, 427–451. doi: 10.1016/S0140-6736(13)60937-X

Braun, H. J., Atlin, G., and Payne, T. (2010). "Multi-location testing as a tool to identify plant response to global climate change," in Climate change and crop production, ed. M. P. Reynolds (CABI Press, Oxford), 115–138. doi: 10.1079/9781845936334.0115

Chen, B. X., Li, W. Y., Gao, Y. T., Chen, Z. J., Zhang, W. N., Liu, Q. J., et al. (2016). Involvement of polyamine oxidase-produced hydrogen peroxide during coleorhiza-limited germination of rice seeds. *Front. Plant. Sci.* 7:1219. doi: 10.3389/fpls.2016.01219

Chen, J., Wang, L., and Yuan, M. (2021). Update on the roles of rice MAPK cascades. *Int. J. Mol. Sci.* 22:1679. doi: 10.3390/ijms22041679

Cheng, S., Zhou, D. X., and Zhao, Y. (2016). WUSCHEL-related homeobox gene WOX11 increases rice drought resistance by controlling root hair

formation and root system development. *Plant Signal Behav.* 11:e1130198. doi: 10.1080/15592324.2015.1130198

Cona, A., Moreno, S., Cenci, F., Federico, R., and Angelini, R. (2005). Cellular re-distribution of flavin-containing polyamine oxidase in differentiating root and mesocotyl of *Zea mays* L. seedlings. *Planta* 221, 265–276. doi: 10.1007/s00425-004-1435-y

Cooper, M., Woodruff, D. R., Eisemann, R. L., Brennan, P. S., and DeLacy, I. H. (1995). A selection strategy to accommodate genotype-by-environment interaction for grain yield of wheat: managed-environments for selection among genotypes. *Theor. Appl. Genet.* 90, 492–502. doi: 10.1007/BF00221995

Crespo-Herrera, L. A., Velu, G., and Singh, R. P. (2016). Quantitative trait loci mapping reveals pleiotropic effect for grain iron and zinc concentrations in wheat. *Ann. Appl. Biol.* 169, 27–35. doi: 10.1111/aab.12276

Danan, S., Veyrieras, J. B., and Lefebvre, V. (2011). Construction of a potato consensus map and QTL meta-analysis offer new insights into the genetic architecture of late blight resistance and plant maturity traits. *BMC Plant Biol.* 11:16. doi: 10.1186/1471-2229-11-16

De-la-Pena, C., Galaz-Avalos, R. M., and Loyola-Vargas, V. M. (2008). Possible role of light and polyamines in the onset of somatic embryogenesis of *Coffea canephora*. *Mol. Biotechnol.* 39, 215–224. doi: 10.1007/s12033-008-9037-8

Delfino, P., Zenoni, S., Imanifard, Z., Tornielli, G. B., and Bellin, D. (2019). Selection of candidate genes controlling veraison time in grapevine through integration of meta-QTL and transcriptomic data. *BMC Genomics* 20, 1–19. doi: 10.1186/s12864-019-6124-0

Deng, L., Li, L., Zhang, S., Shen, J., Li, S., Hu, S., et al. (2017). Suppressor of rid1 (SID1) shares common targets with RID1 on florigen genes to initiate floral transition in rice. *PLoS Genet.* 13:e1006642. doi: 10.1371/journal.pgen.1006642

Deng, X., Han, X., Yu, S., Liu, Z., Guo, D., He, Y., et al. (2020). OsINV3 and its homolog, OsINV2, control grain size in rice. *Int. J. Mol. Sci.* 21:2199. doi: 10.3390/ijms21062199

Dixit, S., Singh, U. M., Abbai, R., Ram, T., Singh, V. K., Paul, A., et al. (2019). Identification of genomic region (s) responsible for high iron and zinc content in rice. *Sci. Rep.* 9:8136. doi: 10.1038/s41598-019-43888-y

El-Feki, W. M., Byrne, P. F., Reid, S. D., and Haley, S. D. (2018). Mapping quantitative trait loci for agronomic traits in winter wheat under different soil moisture levels. *Agronomy* 8:133. doi: 10.3390/agronomy8080133

Fabian, D., and Flatt, T. (2012). Life history evolution. *Nat. Edu. Knowl.* 3:24.

Feng, B., Chen, K., Cui, Y., Wu, Z., Zheng, T., Zhu, Y., et al. (2018). Genetic dissection and simultaneous improvement of drought and low nitrogen tolerances by designed QTL pyramiding in rice. *Front. Plant. Sci.* 9:306. doi: 10.3389/fpls.2018.00306

Fernando, N., Panozzo, J., Tausz, M., Norton, R., Fitzgerald, G., and Seneweera, S. (2012). Rising atmospheric CO2 concentration affects mineral nutrient and protein concentration of wheat grain. *Food. Chem.* 133, 1307–1311. doi: 10.1016/j.foodchem.2012.01.105

Feuillet, C., and Keller, B. (1999). High gene density is conserved at syntenic loci of small and large grass genomes. *Proc. Natl. Acad. Sci. USA.* 96, 8265–8270. doi: 10.1073/pnas.96.14.8265

Flatt, T., and Heyland, A. (2011). *Mechanisms of Life History Evolution: the Genetics and Physiology of Life History Traits and Trade-Offs*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199568765.001.0001

Flint, J., and Mackay, T. F. (2009). Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Res.* 19, 23–733. doi: 10.1101/gr.086660.108

Gahlaut, V., Jaiswal, V., Tyagi, B. S., Singh, G., Sareen, S., Balyan, H. S., et al. (2017). QTL mapping for nine drought-responsive agronomic traits in bread wheat under irrigated and rain-fed environments. *PLoS ONE* 12:e0182857. doi: 10.1371/journal.pone.0182857

Gale, M. D., and Devos, K. M. (1998). Comparative genetics in the grasses. *Proc. Natl. Acad. Sci. USA.* 95, 1971–1974. doi: 10.1073/pnas.95.5.1971

Gaut, B. S. (2002). Evolutionary dynamics of grass genomes. *New Phytol.* 154, 15–28. doi: 10.1046/j.1469-8137.2002.00352.x

Gautam, T., Dhillon, G. S., Saripalli, G., Singh, V. P., Prasad, P., Kaur, S., et al. (2020). Marker-assisted pyramiding of genes/QTL for grain quality and rust resistance in wheat (*Triticum aestivum* L.). *Mol. Breed.* 40, 1–14. doi: 10.1007/s11032-020-01125-9

Giancaspro, A., Giove, S. L., Zacheo, S. A., Blanco, A., and Gadaleta, A. (2019). Genetic variation for protein content and yield-related traits in a durum population derived from an inter-specific cross between hexaploid and tetraploid wheat cultivars. *Front. Plant Sci.* 10:1509. doi: 10.3389/fpls.2019.01509

Goel, S., Singh, K., Singh, B., Grewal, S., Dwivedi, N., Alqarawi, A. A., et al. (2019). Analysis of genetic control and QTL mapping of essential wheat grain quality traits in a recombinant inbred population. *PLoS ONE* 14:e0200669. doi: 10.1371/journal.pone.0200669

Goffinet, B., and Gerber, S. (2000). Quantitative trait loci: a meta-analysis. *Genetics* 155, 463–473. doi: 10.1093/genetics/155.1.463

Gokidi, Y., Bhanu, A. N., Chandra, K., Singh, M. N., and Hemantaranjan, A. (2017). Allele mining–an approach to discover allelic variation in crops. *J. Plant Sci. Res.* 33, 167–180.

Guo, Y., Du, Z., Chen, J., and Zhang, Z. (2017). QTL mapping of wheat plant architectural characteristics and their genetic relationship with seven QTLs conferring resistance to sheath blight. *PLoS ONE* 12:e0174939. doi: 10.1371/journal.pone.0174939

Gustin, J. L., Zanis, M. J., and Salt, D. E. (2011). Structure and evolution of the plant cation diffusion facilitator family of ion transporters. *BMC Evol. Biol.* 11:76. doi: 10.1186/1471-2148-11-76

Hanocq, E., Laperche, A., Jaminon, O., Lainé, A. L., and Le Gouis, J. (2007). Most significant genome regions involved in the control of earliness traits in bread wheat, as revealed by QTL meta-analysis. *Theor. Appl. Genet.* 114, 569–584. doi: 10.1007/s00122-006-0459-z

Hedden, P. (2003). The genes of the Green Revolution. *Trends Genet.* 19, 5–9. doi: 10.1016/S0168-9525(02)00009-4

Hong, Y., Zhang, H., Huang, L., Li, D., and Song, F. (2016). Overexpression of a stress-responsive NAC transcription factor gene ONAC022 improves drought and salt tolerance in rice. *Front. Plant Sci.* 7:4. doi: 10.3389/fpls.2016.00004

Hong, Z., Ueguchi-Tanaka, M., Fujioka, S., Takatsuto, S., Yoshida, S., Hasegawa, Y., et al. (2005). The rice brassinosteroid-deficient dwarf2 mutant, defective in the rice homolog of Arabidopsis DIMINUTO/DWARF1, is rescued by the endogenously accumulated alternative bioactive brassinosteroid, dolichosterone. *Plant Cell* 17, 2243–2254. doi: 10.1105/tpc.105.030973

Hu, P., Zheng, Q., Luo, Q., Teng, W., Li, H., Li, B., et al. (2020). Genome-wide association study of yield and related traits in common wheat under salt-stress conditions. *BMC Plant Biol.* 21:27. doi: 10.1186/s12870-020-02799-1

Itoh, H., Ueguchi-Tanaka, M., Sakamoto, T., Kayano, T., Tanaka, H., Ashikari, M., et al. (2002). Modification of rice plant height by suppressing the height-controlling gene, D18, in rice. *Breed. Sci.* 52, 215–218. doi: 10.1270/jsbbs.52.215

Izawa, T., Oikawa, T., Sugiyama, N., Tanisaka, T., Yano, M., and Shimamoto, K. (2002). Phytochrome mediates the external light signal to repress FT orthologs in photoperiodic flowering of rice. *Genes Dev.* 16, 2006–2020. doi: 10.1101/gad.999202

Izquierdo, P., Astudillo, C., Blair, M. W., Iqbal, A. M., Raatz, B., and Cichy, K. A. (2018). Meta-QTL analysis of seed iron and zinc concentration and content in common bean (*Phaseolus vulgaris* L.). *Theor. Appl. Genet.* 131, 1645–1658. doi: 10.1007/s00122-018-3104-8

Jamil, M., Ali, A., Gul, A., Ghafoor, A., Napar, A. A., Ibrahim, A. M., et al. (2019). Genome-wide association studies of seven agronomic traits under two sowing conditions in bread wheat. *BMC Plant Biol.* 19:149. doi: 10.1186/s12870-019-1754-6

Kakkar, R. K., and Sawhney, V. K. (2002). Polyamine research in plants—a changing perspective. *Physiol. Plant.* 116, 281–292. doi: 10.1034/j.1399-3054.2002.1160302.x

Khahani, B., Tavakol, E., Shariati, V., and Fornara, F. (2020). Genome wide screening and comparative genome analysis for meta-QTLs, ortho-MQTLs and candidate genes controlling yield and yield-related traits in rice. *BMC Genomics* 21, 1–24. doi: 10.1186/s12864-020-6702-1

Khush, G. S., Lee, S., Cho, J. I., and Jeon, J. S. (2012). Biofortification of crops for reducing malnutrition. *Plant Biotechnol. Rep.* 6, 195–202. doi: 10.1007/s11816-012-0216-5

Kolde, R. (2013). *pheatmap: Pretty Heatmaps. R package version 1.0.12*. Available online at: http://CRAN.R-project.org/package=pheatmap (accessed December 26, 2018).

Kong, W., Zhang, Y., Deng, X., Zhang, C., and Li, Y. (2019). Comparative genomic and transcriptomic analyses suggests the evolutionary dynamic of gh3 genes in gramineae crops. *Front. Plant Sci.* 10:1297. doi: 10.3389/fpls.2019.01297

Krishnappa, G., Singh, A. M., Chaudhary, S., Ahlawat, A. K., Singh, S. K., Shukla, R. B., et al. (2017). Molecular mapping of the grain iron and

zinc concentration, protein content and thousand kernel weight in wheat (*Triticum aestivum* L.). *PLoS ONE* 12:e0174972. doi: 10.1371/journal.pone.0174972

Kumar, A., Saripalli, G., Jan, I., Kumar, K., Sharma, P. K., Balyan, H. S., et al. (2020). Meta-QTL analysis and identification of candidate genes for drought tolerance in bread wheat (*Triticum aestivum* L.). *Physiol. Mol. Biol. Plants* 26, 1713–1725. doi: 10.1007/s12298-020-00847-6

Kumar, G. R., Sakthivel, K., Sundaram, R. M., Neeraja, C. N., Balachandran, S. M., Rani, N. S., et al. (2010). Allele mining in crops: prospects and potentials. *Biotechnol. Adv.* 28, 451–461. doi: 10.1016/j.biotechadv.2010.02.007

Kumar, S., Palve, A., Joshi, C., and Srivastava, R. K. (2019). Crop biofortification for iron (Fe), zinc (Zn) and vitamin A with transgenic approaches. *Heliyon* 5:e01914. doi: 10.1016/j.heliyon.2019.e01914

Kumari, R., Kumar, P., Sharma, V. K., and Kumar, H. (2018). Allele mining for crop improvement. *Int. J. Pure Appl. Biosci.* 6, 1456–1465. doi: 10.18782/2320-7051.6073

Lehner, K. R., Taylor, I., McCaskey, E. N., Jain, R., Ronald, P. C., Goldman, D. I., et al. (2018). A histidine kinase gene is required for large radius root tip circumnutation and surface exploration in rice. *bioRxiv*. 437012. doi: 10.1101/437012

Lei, L., Zheng, H. L., Wang, J. G., Liu, H. L., Sun, J., Zhao, H. W., et al. (2018). Genetic dissection of rice (*Oryza sativa* L.) tiller, plant height, and grain yield based on QTL mapping and metaanalysis. *Euphytica* 214, 1–7. doi: 10.1007/s10681-018-2187-2

Leung, H., Raghavan, C., Zhou, B., Oliva, R., Choi, I. R., Lacorte, V., et al. (2015). Allele mining and enhanced genetic recombination for rice breeding. *Rice* 8:34. doi: 10.1186/s12284-015-0069-y

Li, F., Wen, W., Liu, J., Zhang, Y., Cao, S., He, Z., et al. (2019). Genetic architecture of grain yield in bread wheat based on genome-wide association studies. *BMC Plant Biol.* 19:168. doi: 10.1186/s12870-019-1781-3

Li, Q., Pan, Z., Gao, Y., Li, T., Liang, J., Zhang, Z., et al. (2020). Quantitative trait locus (QTLs) mapping for quality traits of wheat based on high density genetic map combined with bulked segregant analysis RNA-seq (BSR-Seq) indicates that the basic 7S globulin gene is related to falling number. *Front. Plant Sci.* 11:1918. doi: 10.3389/fpls.2020.600788

Li, W. T., Liu, C., Liu, Y. X., Pu, Z. E., Dai, S. F., Wang, J. R., et al. (2013). Meta-analysis of QTL associated with tolerance to abiotic stresses in barley. *Euphytica* 189, 31–49. doi: 10.1007/s10681-012-0683-3

Li, X., Xu, X., Liu, W., Li, X., Yang, X., Ru, Z., et al. (2020). Dissection of superior alleles for yield-related traits and their distribution in important cultivars of wheat by association mapping. *Front. Plant Sci.* 11:175. doi: 10.3389/fpls.2020.00175

Liang, Y., Zhang, K., Zhao, L., Liu, B., and Meng, Q., Tian, J., et al. (2010). Identification of chromosome regions conferring dry matter accumulation and photosynthesis in wheat (*Triticum aestivum* L.). *Euphytica* 171, 145–156. doi: 10.1007/s10681-009-0024-3

Lin, H., Wang, R., Qian, Q., Yan, M., Meng, X., Fu, Z., et al. (2009). DWARF27, an iron-containing protein required for the biosynthesis of strigolactones, regulates rice tiller bud outgrowth. *Plant Cell* 21, 1512–1525. doi: 10.1105/tpc.109.065987

Liszkay, A., van der Zalm, E., and Schopfer, P. (2004). Production of reactive oxygen intermediates (O_2-, H_2O_2, and OH) by maize roots and their role in wall loosening and elongation growth. *Plant Physiol.* 136, 3114–3123. doi: 10.1104/pp.104.044784

Liu, J., Wu, B., Singh, R. P., and Velu, G. (2019). QTL mapping for micronutrients concentration and yield component traits in a hexaploid wheat mapping population. *J. Cereal Sci.* 88, 57–64. doi: 10.1016/j.jcs.2019.05.008

Liu, L., Tong, H., Xiao, Y., Che, R., Xu, F., Hu, B., et al. (2015). Activation of Big Grain1 significantly improves grain size by regulating auxin transport in rice. *Proc. Natl. Acad. Sci. USA.* 112, 11102–11107. doi: 10.1073/pnas.1512748112

Liu, X., Feng, Z. M., Zhou, C. L., Ren, Y. K., Mou, C. L., et al. (2016). Brassinosteroid (BR) biosynthetic gene lhdd10 controls late heading and plant height in rice (*Oryza sativa* L.). *Plant Cell Rep.* 35, 357–368. doi: 10.1007/s00299-015-1889-3

Liu, X., Li, J., Xu, L., Wang, Q., and Lou, Y. (2018). Expressing OsMPK4 impairs plant growth but enhances the resistance of rice to the striped stem borer Chilo suppressalis. *Int. J. Mol. Sci*. 19:1182. doi: 10.3390/ijms19041182

Löffler, M., Schön, C. C., and Miedaner, T. (2009). Revealing the genetic architecture of FHB resistance in hexaploid wheat (*Triticum aestivum* L.) by QTL meta-analysis. *Mol. Breed.* 23, 473–488. doi: 10.1007/s11032-008-9250-y

Mahdi-Nezhad, N., Kamali, M. J., McIntyre, C. L., Fakheri, B. A., Omidi, M., and Masoudi, B. (2019). Mapping QTLs with main and epistatic effect on Seri 'M82× Babax 'wheat population under salt stress. *Euphytica* 215:130. doi: 10.1007/s10681-019-2450-1

Malik, N., Dwivedi, N., Singh, A. K., Parida, S. K., Agarwal, P., Thakur, J. K., et al. (2016). An integrated genomic strategy delineates candidate mediator genes regulating grain size and weight in rice. *Sci. Rep.* 6, 1–12. doi: 10.1038/srep23253

Martinez, A. K., Soriano, J. M., Tuberosa, R., Koumproglou, R., Jahrmann, T., and Salvi, S. (2016). Yield QTLome distribution correlates with gene density in maize. *Plant Sci.* 242, 300–309. doi: 10.1016/j.plantsci.2015.09.022

Marza, F., Bai, G. H., Carver, B. F., and Zhou, W. C. (2006). Quantitative trait loci for yield and related traits in the wheat population Ning7840× Clark. *Theor. Appl. Genet.* 112, 688–698. doi: 10.1007/s00122-005-0172-3

Mathew, I., Shimelis, H., Shayanowako, A. I. T., Laing, M., and Chaplot, V. (2019). Genome-wide association study of drought tolerance and biomass allocation in wheat. *PLoS ONE* 14:e0225383. doi: 10.1371/journal.pone.0225383

Michel, S., Löschenberger, F., Ametz, C., Pachler, B., Sparry, E., and Bürstmayr, H. (2019). Combining grain yield, protein content and protein quality by multi-trait genomic selection in bread wheat. *Theor. Appl. Genet*. 132, 2767–2780. doi: 10.1007/s00122-019-03386-1

Minx, P., Cordum, H., and Wilson, R. (2005). Sequence, annotation, and analysis of synteny between rice chromosome 3 and diverged grass species. *Genome Res.* 15, 1284–1291. doi: 10.1101/gr.3869505

Mishra, B. S., Jamsheer, K. M., Singh, D., Sharma, M., and Laxmi, A. (2017). Genome-wide identification and expression, protein–protein interaction and evolutionary analysis of the seed plant-specific BIG GRAIN and BIG GRAIN LIKE gene family. *Front. Plant Sci.* 8:1812. doi: 10.3389/fpls.2017.01812

Mohan, M., Nair, S., Bhagwat, A., Krishna, T. G., Yano, M., Bhatia, C. R., et al. (1997). Genome mapping, molecular markers and marker-assisted selection in crop plants. *Mol. Breed*. 3, 87–103. doi: 10.1023/A:1009651919792

Moin, M., Bakshi, A., Madhav, M. S., and Kirti, P. B. (2017). Expression profiling of ribosomal protein gene family in dehydration stress responses and characterization of transgenic rice plants overexpressing RPL23A for water-use efficiency and tolerance to drought and salt stresses. *Front. Chem.* 5:97. doi: 10.3389/fchem.2017.00097

Monod, C., Takahashi, Y., Goldschmidt-Clermont, M., and Rochaix, J. D. (1994). The chloroplast ycf8 open reading frame encodes a photosystem II polypeptide which maintains photosynthetic activity under adverse growth conditions. *EMBO J.* 13, 2747–2754. doi: 10.1002/j.1460-2075.1994.tb06568.x

Moore, G., Devos, K. M., Wang, Z., and Gale, M. D. (1995). Cereal genome evolution: grasses, line up and form a circle. *Curr. Biol.* 5, 737–739. doi: 10.1016/S0960-9822(95)00148-5

Morey, S. R., Hirose, T., Hashida, Y., Miyao, A., Hirochika, H., Ohsugi, R., et al. (2018). Genetic evidence for the role of a rice vacuolar invertase as a molecular sink strength determinant. *Rice* 11:6. doi: 10.1186/s12284-018-0201-x

Morrissey, J., and Guerinot, M. L. (2009). Iron uptake and transport in plants: the good, the bad, and the ionome. *Chem. Rev.* 109, 4553–4567. doi: 10.1021/cr900112r

Moschou, P. N., Sanmartin, M., Andriopoulou, A. H., Rojo, E., Sanchez-Serrano, J. J., and Roubelakis-Angelakis, K. A. (2008). Bridging the gap between plant and mammalian polyamine catabolism: a novel peroxisomal polyamine oxidase responsible for a full back-conversion pathway in Arabidopsis. *Plant Physiol.* 147, 1845–1857. doi: 10.1104/pp.108.123802

Moschou, P. N., Sarris, P. F., Skandalis, N., Andriopoulou, A. H., Paschalidis, K. A., Panopoulos, N. J., et al. (2009). Engineered polyamine catabolism preinduces tolerance of tobacco to bacteria and oomycetes. *Plant Physiol.* 149, 1970–1981. doi: 10.1104/pp.108.134932

Murray, C. J., and Lopez, A. D. (2013). Measuring the global burden of disease. *N. Engl. J. Med.* 369, 448–457. doi: 10.1056/NEJMra1201534

Muthu, V., Abbai, R., Nallathambi, J., Rahman, H., Ramasamy, S., Kambale, R., et al. (2020). Pyramiding QTLs controlling tolerance against drought,

salinity, and submergence in rice through marker assisted breeding. *PLoS ONE* 15:e0227421. doi: 10.1371/journal.pone.0227421

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika* 78, 691–692. doi: 10.1093/biomet/78.3.691

Ogawa, D., Yamamoto, E., Ohtani, T., Kanno, N., Tsunematsu, H., Nonoue, Y., et al. (2018). Haplotype-based allele mining in the Japan-MAGIC rice population. *Sci. Rep.* 8:4379. doi: 10.1038/s41598-018-22657-3

Ogiso-Tanaka, E., Matsubara, K., Yamamoto, S. I., Nonoue, Y., Wu, J., Fujisawa, H., et al. (2013). Natural variation of the RICE FLOWERING LOCUS T 1 contributes to flowering time divergence in rice. *PLoS ONE* 8:e75959. doi: 10.1371/journal.pone.0075959

Ozawa, S., Kobayashi, T., Sugiyama, R., Hoshida, H., Shiina, T., and Toyoshima, Y. (1997). Role of PSII-L protein (psbL gene product) on the electron transfer in photosystem II complex. 1. Over-production of wild-type and mutant versions of PSII-L protein and reconstitution into the PSII core complex. *Plant Mol. Biol.* 34, 151–161. doi: 10.1023/A:1005800909495

Pachauri, R. K., Allen, M. R., Barros, V. R., Broome, J., Cramer, W., Christ, R., et al. (2014). Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change (p. 151). IPCC.

Peleg, Z., Cakmak, I., Ozturk, L., Yazici, A., Jun, Y., Budak, H., et al. (2009). Quantitative trait loci conferring grain mineral nutrient concentrations in durum wheat × wild emmer wheat RIL population. *Theor. Appl. Genet.* 119, 353–369. doi: 10.1007/s00122-009-1044-z

Pu, C. X., Han, Y. F., Zhu, S., Song, F. Y., Zhao, Y., Wang, C. Y., et al. (2017). The rice receptor-like kinases DWARF AND RUNTISH SPIKELET1 and 2 repress cell death and affect sugar utilization during reproductive development. *Plant Cell* 29, 70–89. doi: 10.1105/tpc.16.00218

Pu, Z. E., Ma, Y.u., He, Q. Y., Chen, G. Y., Wang, J. R., Liu, Y. X., et al. (2014). Quantitative trait loci associated with micronutrient concentrations in two recombinant inbred wheat lines. *J. Integr. Agric.* 13, 2322–2329. doi: 10.1016/S2095-3119(13)60640-1

Quraishi, U. M., Pont, C., Ain, Q. U., Flores, R., Burlot, L., Alaux, M., et al. (2017). Combined genomic and genetic data integration of major agronomical traits in bread wheat (*Triticum aestivum* L.). *Front. Plant Sci.* 8:1843. doi: 10.3389/fpls.2017.01843

Ram, H., Kaur, A., Gandass, N., Singh, S., Deshmukh, R., Sonah, H., et al. (2019). Molecular characterization and expression dynamics of MTP genes under various spatio-temporal stages and metal stress conditions in rice. *PLoS ONE* 14:e0217360. doi: 10.1371/journal.pone.0217360

Ramya, P., Chaubal, A., Kulkarni, K., Gupta, L., Kadoo, N., Dhaliwal, H. S., et al. (2010). QTL mapping of 1000-kernel weight, kernel length, and kernel width in bread wheat (*Triticum aestivum* L.). *J. Appl. Genet.* 51, 421–429. doi: 10.1007/BF03208872

Raza, Q., Riaz, A., Sabar, M., Atif, R. M., and Bashir, K. (2019). Meta-analysis of grain iron and zinc associated QTLs identified hotspot chromosomal regions and positional candidate genes for breeding biofortified rice. *Plant Sci.* 288:110214. doi: 10.1016/j.plantsci.2019.110214

Rebetzke, G. J., Condon, A. G., Farquhar, G. D., Appels, R., and Richards, R. A. (2008). Quantitative trait loci for carbon isotope discrimination are repeatable across environments and wheat mapping populations. *Theor. Appl. Genet.* 118, 123–137. doi: 10.1007/s00122-008-0882-4

Richmond, T. A., and Bleecker, A. B. (1999). A defect in β-oxidation causes abnormal inflorescence development in Arabidopsis. *Plant Cell* 11, 1911–1923. doi: 10.1105/tpc.11.10.1911

Rogalski, M., Schöttler, M. A., Thiele, W., Schulze, W. X., and Bock, R. (2008). Rpl33, a nonessential plastid-encoded ribosomal protein in tobacco, is required under cold stress conditions. *Plant Cell* 20, 2221–2237. doi: 10.1105/tpc.108.060392

Roshanzamir, H., Kordenaeej, A., and Bostani, A. (2013). Mapping QTLs related to Zn and Fe concentrations in bread wheat (*Triticum aestivum*) grain using microsatellite markers. *Iranian J. Genet. Plant Breed.* 2, 10–17.

Saini, D. K., Devi, P., and Kaushik, P. (2020). Advances in genomic interventions for wheat biofortification: a review. *Agronomy* 10:62. doi: 10.3390/agronomy10010062

Salvi, S., and Tuberosa, R. (2015). The crop QTLome comes of age. *Curr. Opin. Biotechnol.* 32, 179–185. doi: 10.1016/j.copbio.2015.01.001

Semagn, K., Beyene, Y., Warburton, M. L., Tarekegne, A., Mugo, S., Meisel, B., et al. (2013). Meta-analyses of QTL for grain yield and anthesis silking interval in 18 maize populations evaluated under water-stressed and well-watered environments. *BMC Genomics* 14, 1–6. doi: 10.1186/1471-2164-14-313

Shahzad, Z., Rouached, H., and Rakha, A. (2014). Combating mineral malnutrition through iron and zinc biofortification of cereals. *Compr. Rev. Food Sci. Food Saf.* 13, 329–346. doi: 10.1111/1541-4337.12063

Shariatipour, N., and Heidari, B. (2020). "Genetic-based biofortification of staple food crops to meet zinc and iron deficiency-related challenges," in *Plant Micronutrients, Deficiency and Toxicity Management*, ed. T. Aftab, K. R. Hakeem (Cham: Springer), 173–223. doi: 10.1007/978-3-030-49856-6_8

Shukla, S., Singh, K., Patil, R. V., Kadam, S., Bharti, S., Prasad, P., et al. (2014). Genomic regions associated with grain yield under drought stress in wheat (*Triticum aestivum* L.). *Euphytica* 203, 449–467. doi: 10.1007/s10681-014-1314-y

Simmonds, N. W. (1995). The relation between yield and protein in cereal grain. *J. Sci. Food Agric.* 67, 309–315. doi: 10.1002/jsfa.2740670306

Somers, D. J., Isaac, P., and Edwards, K. (2004). A high-density microsatellite consensus map for bread wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* 109, 1105–1114. doi: 10.1007/s00122-004-1740-7

Soriano, J. M., and Alvaro, F. (2019). Discovering consensus genomic regions in wheat for root-related traits by QTL meta-analysis. *Sci. Rep.* 9:10537. doi: 10.1038/s41598-019-47038-2

Sosnowski, O., Charcosset, A., and Joets, J. (2012). BioMercator V3: an upgrade of genetic map compilation and quantitative trait loci meta-analysis algorithms. *Bioinformatics* 28, 2082–2083. doi: 10.1093/bioinformatics/bts313

Spindel, J. E., Begum, H., Akdemir, D., Collard, B., Redoña, E., Jannink, J. L., et al. (2016). Genome-wide prediction models that incorporate *de novo* GWAS are a powerful new tool for tropical rice improvement. *Heredity* 116, 395–408. doi: 10.1038/hdy.2015.113

Stein, A. J. (2010). Global impacts of human mineral malnutrition. *Plant Soil* 335, 133–154. doi: 10.1007/s11104-009-0228-2

Swamy, B. P. M., Kaladhar, K., Anuradha, K., Batchu, A. K., Longvah, T., and Sarla, N. (2018). QTL analysis for grain iron and zinc concentrations in two *O. nivara* derived backcross populations. *Rice Sci.* 25, 197–207. doi: 10.1016/j.rsci.2018.06.003

Swamy, B. P. M., and Sarla, N. (2011). Meta-analysis of yield QTLs derived from interspecific crosses of rice reveals consensus regions and candidate genes. *Plant Mol. Biol. Rep.* 29, 663–680. doi: 10.1007/s11105-010-0274-1

Tamiru, M., Undan, J. R., Takagi, H., Abe, A., Yoshida, K., Undan, J. Q., et al. (2015). A cytochrome P450, OsDSS1, is involved in growth and drought stress responses in rice (*Oryza sativa* L.). *Plant Mol. Biol.* 88, 85–99. doi: 10.1007/s11103-015-0310-5

Tiwari, C., Wallwork, H., Arun, B., Mishra, V. K., Velu, G., Stangoulis, J., et al. (2016). Molecular mapping of quantitative trait loci for zinc, iron and protein content in the grains of hexaploid wheat. *Euphytica* 207, 563–570. doi: 10.1007/s10681-015-1544-7

Tiwari, V. K., Rawat, N., Chhuneja, P., Neelam, K., Aggarwal, R., Randhawa, G. S., et al. (2009). Mapping of quantitative trait loci for grain iron and zinc concentration in diploid A genome wheat. *J. Hered.* 100, 771–776. doi: 10.1093/jhered/esp030

Truntzler, M., Barriere, Y., Sawkins, M. C., Lespinasse, D., Betran, J., Charcosset, A., et al. (2010). Meta-analysis of QTL involved in silage quality of maize and comparison with the position of candidate genes. *Theor. Appl. Genet.* 121, 1465–1482. doi: 10.1007/s00122-010-1402-x

Velu, G., Guzman, C., Mondal, S., Autrique, J. E., Huerta, J., and Singh, R. P. (2016). Effect of drought and elevated temperature on grain zinc and iron concentrations in CIMMYT spring wheat. *J. Cereal Sci.* 69, 182–186. doi: 10.1016/j.jcs.2016.03.006

Velu, G., Singh, R. P., Crespo-Herrera, L., Juliana, P., Dreisigacker, S., Valluru, R., et al. (2018). Genetic dissection of grain zinc concentration in spring wheat for mainstreaming biofortification in CIMMYT wheat breeding. *Sci. Rep.* 8:13526. doi: 10.1038/s41598-018-31951-z

Velu, G., Tutus, Y., Gomez-Becerra, H. F., Hao, Y., Demir, L., Kara, R., et al. (2017). QTL mapping for grain zinc and iron concentrations and zinc efficiency in a tetraploid and hexaploid wheat mapping populations. *Plant Soil* 411, 81–99. doi: 10.1007/s11104-016-3025-8

Veyrieras, J. B., Goffinet, B., and Charcosset, A. (2007). MetaQTL: a package of new computational methods for the meta-analysis of QTL mapping experiments. *BMC Bioinform.* 8:49. doi: 10.1186/1471-2105-8-49

Vij, S., Giri, J., Dansana, P. K., Kapoor, S., and Tyagi, A. K. (2008). The receptor-like cytoplasmic kinase (*OsRLCK*) gene family in rice: organization, phylogenetic relationship, and expression during development and stress. *Mol. Plant* 1, 732–750. doi: 10.1093/mp/ssn047

Vijayalakshmi, K., Fritz, A. K., Paulsen, G. M., Bai, G., Pandravada, S., and Gill, B. S. (2010). Modeling and mapping QTL for senescence-related traits in winter wheat under high temperature. *Mol. Breed.* 26, 163–175. doi: 10.1007/s11032-009-9366-8

Wan, Y., King, R., Mitchell, R. A., Hassani-Pak, K., and Hawkesford, M. J. (2017). Spatiotemporal expression patterns of wheat amino acid transporters reveal their putative roles in nitrogen transport and responses to abiotic stress. *Sci. Rep.* 7:5461. doi: 10.1038/s41598-017-04473-3

Wang, L. I., Cui, F. A., Wang, J., Jun, L. I., Ding, A., Zhao, C., et al. (2012). Conditional QTL mapping of protein content in wheat with respect to grain yield and its components. *J. Genet.* 91, 303–312. doi: 10.1007/s12041-012-0190-2

Wang, P., Xing, Y., Li, Z., and Yu, S. (2012). Improving rice yield and quality by QTL pyramiding. *Mol. Breed.* 29, 903–913. doi: 10.1007/s11032-011-9679-2

Wang, Y., Huang, Z., Deng, D., Ding, H., Zhang, R., Wang, S., et al. (2013). Meta-analysis combined with syntenic metaQTL mining dissects candidate loci for maize yield. *Mol. Breed.* 31, 601–614. doi: 10.1007/s11032-012-9818-4

Wang, Y., Shi, C., Yang, T., Zhao, L., Chen, J., Zhang, N., et al. (2018). High-throughput sequencing revealed that microRNAs were involved in the development of superior and inferior grains in bread wheat. *Sci. Rep.* 8, 1–18. doi: 10.1038/s41598-018-31870-z

Wang, Y., Xu, J., Deng, D., Ding, H., Bian, Y., Yin, Z., et al. (2016). A comprehensive meta-analysis of plant morphology, yield, stay-green, and virus disease resistance QTL in maize (*Zea mays* L.). *Planta* 243, 459–471. doi: 10.1007/s00425-015-2419-9

Wu, C., You, C., Li, C., Long, T., Chen, G., Byrne, M. E., et al. (2008). RID1, encoding a Cys2/His2-type zinc finger transcription factor, acts as a master switch from vegetative to floral development in rice. *Proc. Natl. Acad. Sci. USA.* 105, 12915–12920. doi: 10.1073/pnas.0806019105

Wu, X. L., and Hu, Z. L. (2012). Meta-analysis of QTL mapping experiments. *Methods Mol. Biol.* 871, 145–171. doi: 10.1007/978-1-61779-785-9_8

Wu, Y., Zhao, S., Li, X., Zhang, B., Jiang, L., Tang, Y., et al. (2018). Deletions linked to PROG1 gene participate in plant architecture domestication in Asian and African rice. *Nat. Commun.* 9, 1–10. doi: 10.1038/s41467-018-06509-2

Xiang, K., Reid, L. M., Zhang, Z. M., Zhu, X. Y., and Pan, G. T. (2012). Characterization of correlation between grain moisture and ear rot resistance in maize by QTL meta-analysis. *Euphytica* 183, 185–195. doi: 10.1007/s10681-011-0440-z

Xu, L., Zhao, H., Ruan, W., Deng, M., Wang, F., Peng, J., et al. (2017). ABNORMAL INFLORESCENCE MERISTEM1 functions in salicylic acid biosynthesis to maintain proper reactive oxygen species levels for root meristem activity in rice. *Plant Cell* 29, 560–574. doi: 10.1105/tpc.16.00665

Xu, Y., An, D., Liu, D., Zhang, A., Xu, H., and Li, B. (2012). Molecular mapping of QTLs for grain zinc, iron and protein concentration of wheat across two environments. *Field Crops Res.* 138, 57–62. doi: 10.1016/j.fcr.2012.09.017

Xue, W., Xing, Y., Weng, X., Zhao, Y., Tang, W., Wang, L., et al. (2008). Natural variation in Ghd7 is an important regulator of heading date and yield potential in rice. *Nat. Genet.* 40, 761–767. doi: 10.1038/ng.143

Yamburenko, M. V., Kieber, J. J., and Schaller, G. E. (2017). Dynamic patterns of expression for genes regulating cytokinin metabolism and signaling during rice inflorescence development. *PLoS ONE* 12:e0176060. doi: 10.1371/journal.pone.0176060

Yao, R., Wang, L., Li, Y., Chen, L., Li, S., Du, X., et al. (2018). Rice DWARF14 acts as an unconventional hormone receptor for strigolactone. *J. Exp. Bot.* 69, 2355–2365. doi: 10.1093/jxb/ery014

Yu, Y., Ouyang, Y., and Yao, W. (2017). shinyCircos: an R/Shiny application for interactive creation of Circos plot. *Bioinformatics* 34, 1229–1231. doi: 10.1093/bioinformatics/btx763

Zhang, A., Zhang, J., Ye, N., Cao, J., Tan, M., Zhang, J., et al. (2010). ZmMPK5 is required for the NADPH oxidase-mediated self-propagation of apoplastic H_2O_2 in brassinosteroid-induced antioxidant defence in leaves of maize. *J. Exp. Bot.* 61, 4399–4411. doi: 10.1093/jxb/erq243

Zhang, D., Jiang, S., Pan, J., Kong, X., Zhou, Y., Liu, Y., et al. (2014). The overexpression of a maize mitogen-activated protein kinase gene (ZmMPK5) confers salt stress tolerance and induces defence responses in tobacco. *Plant Biol.* 16, 558–570. doi: 10.1111/plb.12084

Zhang, H., Zhang, J., Yan, J., Gou, F., Mao, Y., Tang, G., et al. (2017). Short tandem target mimic rice lines uncover functions of miRNAs in regulating important agronomic traits. *Proc. Natl. Acad. Sci. USA.* 114, 5277–5282. doi: 10.1073/pnas.1703752114

Zhang, L. Y., Liu, D. C., Guo, X. L., Yang, W. L., Sun, J. Z., Wang, D. W., et al. (2010). Genomic distribution of quantitative trait loci for yield and yield-related traits in common wheat. *J. Integr. Plant. Biol.* 52, 996–1007. doi: 10.1111/j.1744-7909.2010.00967.x

Zhang, M., Cui, Y., Liu, Y. H., Xu, W., Sze, S. H., Murray, S. C., et al. (2020). Accurate prediction of maize grain yield using its contributing genes for gene-based breeding. *Genomics* 112, 225–236. doi: 10.1016/j.ygeno.2019.02.001

Zhang, M., and Liu, B. (2017). Identification of a rice metal tolerance protein OsMTP11 as a manganese transporter. *PLoS ONE* 12:e0174987. doi: 10.1371/journal.pone.0174987

Zhang, T., Feng, P., Li, Y., Yu, P., Yu, G., Sang, X., et al. (2018). VIRESCENT-ALBINO LEAF 1 regulates leaf colour development and cell division in rice. *J. Exp. Bot.* 69, 4791–4804. doi: 10.1093/jxb/ery250

Zhao, H., Duan, K. X., Ma, B., Yin, C. C., Hu, Y., Tao, J. J., et al. (2020). Histidine kinase MHZ1/OsHK1 interacts with ethylene receptors to regulate root growth in rice. *Nat. Commun.* 11, 1–13. doi: 10.1038/s41467-020-14313-0

Zhao, X., Peng, Y., Zhang, J., Fang, P., and Wu, B. (2018). Identification of QTLs and meta-QTLs for seven agronomic traits in multiple maize populations under well-watered and water-stressed conditions. *Crop Sci.* 58, 507–520. doi: 10.2135/cropsci2016.12.0991

Zhao, Y., Cheng, S., Song, Y., Huang, Y., Zhou, S., Liu, X., et al. (2015). The interaction between rice *ERF3* and *WOX11* promotes crown root development by regulating gene expression involved in cytokinin signaling. *Plant Cell* 27, 2469–2483. doi: 10.1105/tpc.15.00227

Zhou, W., Wang, Y., Wu, Z., Luo, L., Liu, P., Yan, L., et al. (2016). Homologs of SCAR/WAVE complex components are required for epidermal cell morphogenesis in rice. *J. Exp. Bot.* 67, 4311–4323. doi: 10.1093/jxb/erw214

Zhu, D., Chang, Y., Pei, T., Zhang, X., Liu, L., Li, Y., et al. (2020). MAPK-like protein 1 positively regulates maize seedling drought sensitivity by suppressing ABA biosynthesis. *Plant J.* 102, 747–760. doi: 10.1111/tpj.14660

Zilic, S., Barac, M., Pešic, M., Dodig, D., and Ignjatovic-Micic, D. (2011). Characterization of proteins from grain of different bread and durum wheat genotypes. *Int. J. Mol. Sci.* 12, 5878–5894. doi: 10.3390/ijms12095878

Zou, Y., Liu, X., Wang, Q., Chen, Y., Liu, C., Qiu, Y., et al. (2014). *OsRPK1*, a novel leucine-rich repeat receptor-like kinase, negatively regulates polar auxin transport and root development in rice. *Biochim. Biophys. Acta* 1840, 1676–1685. doi: 10.1016/j.bbagen.2014.01.003

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 24 November 2021
doi: 10.3389/fpls.2021.771075

Training Set Construction for Genomic Prediction in Auto-Tetraploids: An Example in Potato

*Stefan Wilson[1], Marcos Malosetti[1], Chris Maliepaard[2], Han A. Mulder[3], Richard G. F. Visser[2] and Fred van Eeuwijk[1]**

[1] Biometris, Wageningen University & Research, Wageningen, Netherlands, [2] Plant Breeding, Wageningen University & Research, Wageningen, Netherlands, [3] Wageningen University & Research, Animal Breeding and Genomics, Wageningen, Netherlands

Training set construction is an important prerequisite to Genomic Prediction (GP), and while this has been studied in diploids, polyploids have not received the same attention. Polyploidy is a common feature in many crop plants, like for example banana and blueberry, but also potato which is the third most important crop in the world in terms of food consumption, after rice and wheat. The aim of this study was to investigate the impact of different training set construction methods using a publicly available diversity panel of tetraploid potatoes. Four methods of training set construction were compared: simple random sampling, stratified random sampling, genetic distance sampling and sampling based on the coefficient of determination (CDmean). For stratified random sampling, population structure analyses were carried out in order to define sub-populations, but since sub-populations accounted for only 16.6% of genetic variation, there were negligible differences between stratified and simple random sampling. For genetic distance sampling, four genetic distance measures were compared and though they performed similarly, Euclidean distance was the most consistent. In the majority of cases the CDmean method was the best sampling method, and compared to simple random sampling gave improvements of 4–14% in cross-validation scenarios, and 2–8% in scenarios with an independent test set, while genetic distance sampling gave improvements of 5.5–10.5% and 0.4–4.5%. No interaction was found between sampling method and the statistical model for the traits analyzed.

Keywords: training set construction, potato, sampling technique(s), genomic prediction (GP), auto-tetraploid

OPEN ACCESS

Edited by:
Rodomiro Ortiz, Swedish University of Agricultural Sciences, Sweden

Reviewed by:
Luis Felipe Ventorim Ferrão, University of Florida, United States
Marcio Resende, University of Florida, United States
John Edward Bradshaw, The James Hutton Institute, United Kingdom

***Correspondence:**
Fred van Eeuwijk
fred.vaneeuwijk@wur.nl

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *05 September 2021*
Accepted: *20 October 2021*
Published: *24 November 2021*

Citation:
Wilson S, Malosetti M, Maliepaard C, Mulder HA, Visser RGF and van Eeuwijk F (2021) Training Set Construction for Genomic Prediction in Auto-Tetraploids: An Example in Potato. Front. Plant Sci. 12:771075. doi: 10.3389/fpls.2021.771075

INTRODUCTION

The utilization of DNA marker information for selection in breeding programs has increased over the last two decades and can be attributed to two factors: the decrease of genotyping costs, and the advances in quantitative genetics methodology. Genomic prediction (GP) is an example of one such methodological breakthrough that estimates breeding or genotypic values (depending on the application) by regressing known phenotypes against high density molecular markers (Meuwissen et al., 2001). GP allows the prediction of phenotypes from marker information which speeds up the breeding cycle, as the performance of new material can be assessed prior to phenotype expression (Heffner et al., 2010).

The potential genetic gains from GP hinge on its ability to predict phenotypes accurately. This prediction accuracy is dependent on various factors including but not restricted to: trait heritability (Heffner et al., 2009), statistical models (de los Campos et al., 2013), genetic architecture of traits (Daetwyler et al., 2013), population structure (Asoro et al., 2011; Guo et al., 2014) as well as the size and composition of the training/calibration set (Pszczola et al., 2012; Rincent et al., 2012; Bustos-Korts et al., 2016; Akdemir and Isidro-Sanchez, 2019). This study focuses on the composition of the training set; those individuals with both phenotype and genotype information, that are used to train the model and estimate the marker effects used to make future predictions. Having both input and target information, the training provides the necessary data so that statistical models can learn and estimate the relationship between explanatory variables and the target (James et al., 2013). The training set should be constructed in a way that it covers a space which closely resembles the space occupied by future test sets. This is important for GP because in more recent times, due to relatively cheap genotyping, molecular marker information (explanatory variables), can often be collected more efficiently than phenotype information (target). The question is, which individuals should be phenotyped and thus be used to calibrate the model and generate reliable predictions for individuals without phenotypic information?

Various sampling strategies are available for training set construction. Simple random sampling allows each individual in the population an equal probability of being in the training set and does not utilize any prior information regarding the material. If population structure exists and the material is separated into sub-populations, this information can be included in a sampling method known as stratified sampling. Stratified random sampling selects individuals based on their sub-population membership. Studies in diploids have shown that this method is superior to simple random sampling, although the improvement depends on the extent of the separation between sub-populations (Isidro et al., 2015). When there is little population structure, uniform coverage of the genetic space may be more suitable, and this is achieved with genetic distance sampling (Jansen and van Hintum, 2007). This methodology was first introduced to define core collections for germplasm banks, but the principle can be extended to construction of the training set, because similar to core collections, the objective is to obtain a subset of individuals that contain the genetic diversity present in a larger population. Rincent et al. (2012) proposed another method for sampling the training set that evaluates the quality of prediction for a set of genotypes. An algorithm was developed that chooses a training set that maximizes prediction accuracy, based on prediction error variance (PEV) and coefficient of determination (CD) measures (Rincent et al., 2012).

Numerous comparative studies have evaluated different methods of training set construction (Asoro et al., 2011; Isidro et al., 2015; Bustos-Korts et al., 2016; Akdemir and Isidro-Sanchez, 2019). These past studies have been conducted on diploids (2 copies of each chromosome) whereas in this study, the focus is on tetraploids (4 copies of each chromosome). Plants often exhibit polyploidy, as seen in potato (*Solanum tuberosum*), which is an auto-tetraploid and the subject of this article. There is potential for genetic gain in applying genomic prediction to potato (Slater et al., 2016), and this was put into practice in recent studies (Habyarimana et al., 2017; Sverrisdóttir et al., 2017; Endelman et al., 2018). The current study seeks to investigate the first step of GP not emphasized in the aforementioned papers, which is the impact of training set construction on GP accuracies in tetraploid potato. A secondary aspect of this study is the investigation of genetic distance measures, as these will be required to implement genetic distance sampling. Various measures of genetic distance exist, and the effect it has on selection accuracy has not yet been evaluated. There are some proposed measures that are allegedly more suitable for polyploids by accounting for allele dosage in polyploid heterozygotes, and by considering the presence of unknown alleles, where the absence of one allele does not necessarily imply presence of the other (Dufresne et al., 2014).

To ensure that the training set construction method would be robust for many GP models, three types of statistical models were assessed to generate prediction accuracies. They belong to three general categories of GP models: no marker selection, marker selection and models that capture non-additive effects. This was included in the study to investigate the presence/absence of a relationship between the sampling method for constructing the training set and the statistical model. The aim is to uncover the most suitable method for constructing the training set when GP for tetraploids is performed, and whether suitable methods exhibit codependencies with other influences including statistical model, sample size and trait architecture.

MATERIALS AND METHODS

Plant Materials

Phenotypic and genotypic data were collected and made publicly available by The Solanaceae Coordinated Agricultural Project (SolCAP). The SolCAP North American potato diversity panel is a compilation of elite potato germplasm from breeding programs across the U.S., as well as historical varieties from the NRSP-6 potato gene bank (Hirsch et al., 2013), and includes tetraploid species, diploid species, wild species and some diploid and tetraploid genetic stocks. For this study only the 190 cultivated tetraploid lines that contained both phenotypic and genotypic data were analyzed. Additional information about these lines was provided including release dates and the classification of each variety into one of six market classes: French Fry processing, Chip Processing, Table Russet, Round White table, Yellow and Pigmented (Hamilton et al., 2011). Genotyping was done with an Infinium SNP array of 8303 markers, and analyses to determine allelic dosages were performed with GenomeStudio. Poor quality SNPs, and SNPs unable to distinguish between the heterozygous classes were removed, leaving 3763 bi-allelic SNPs with reliable information on allelic dosages (Hirsch et al., 2013). For all calculations utilizing SNP information, the marker matrix was coded categorically (AAAA, AAAB, AABB, ABBB, and BBBB) or as a numerical measure of the number of alternate alleles present (0,1,2,3, and 4), where "A" is the reference allele and "B" the alternative allele.

Genomic Prediction was conducted for the three quantitative traits, especially important to the French fry and potato chip markets: tuber length (millimeters), tuber fructose and sucrose content (milligrams $gram^{-1}$ fresh weight). Information on these traits were reported in the study by Rosyara et al. (2016), and were chosen so that for this study, we examine traits with high broad-sense heritabilities like tuber length and fructose content ($h^2 = 0.91$ and $h^2 = 0.85$, respectively), and sucrose content, a trait with intermediate heritability ($h^2 = 0.67$) (Rosyara et al., 2016). These traits, among others were measured in as many as four environments (New York 2010, Wisconsin 2010, New York 2011, and Wisconsin 2011) however not all traits were measured in all environments. The trials consisted of a randomized complete block design with two replicates in each environment and using a linear model accounting for experimental design variables, phenotypic values were generated as the best linear unbiased estimator (BLUE) (Rosyara et al., 2016).

Analyses

Population Structure

To assess population structure for the definition of strata, the marker data was analyzed using three methods: Principal Components Analysis (PCA), Discriminant Analysis of Principal Components (DAPC) and Analysis of Molecular Variance (AMOVA). In a population with distinct sub-divisions, a significant portion of the genetic variability of the population can be attributed to the differences between sub-populations. AMOVA estimates variance components of various factors, including the contribution of subgroups to a population's total variability (Excoffier et al., 1992). Population structure can also be visualized and quantified using Principal Components (Jombart, 2008). Market classes were given for the SolCAP North American diversity panel, and to visualize the extent of separation between these classes, DAPC was implemented. Unlike PCA which looks at overall variability (between and within classes), DAPC maximizes the between group variation with respect to the variation within groups (Jombart et al., 2010).

Sampling Methods

To evaluate training set construction methods, prediction accuracies were compared. Accuracy was defined as the correlation between observed phenotypic values and genotypic values of the validation/test set predicted by the corresponding genomic prediction model. The underlying hypothesis is that the prediction accuracy may be affected by the training set used to calibrate the model; a training set that does not cover the design space will result in poor predictions of the test set. In this study, four methods for constructing the training set were compared: simple random sampling, stratified random sampling, genetic distance sampling and the CDmean method.

- **Simple Random Sampling (SRS):** Training set construction is equivalent to taking a subset of a larger set. For simple random sampling, members of this subset are chosen randomly and completely by chance so that each individual from the panel has an equal probability to be selected for the training set.
- **Stratified Random Sampling (STRAT):** Using the population analysis results to define strata, this method randomly selects individuals from each sub-population, ensuring that every sub-population is represented in the sample, while maintaining the same strata proportions.

$$n_S = \frac{n}{N} \times N_S$$

For the above equation n_S is the number of individuals in the sample from stratum S, N_S is the number of individuals in the population from stratum S, while n and N are the total sample size and total number of individuals, respectively.

- **Genetic Distance Sampling (GD):** This method requires as input, the distances between genotypes calculated from the marker data. From the initial pool, one individual is randomly selected and all individuals within a radial distance r are discarded and will no longer be candidates for sampling. This ensures that the next individual sampled will not be genetically similar to the first individual. From the remaining set, a second individual is selected and again, all individuals within a genetic distance of r are discarded. This process is continued until the desired training set size is attained. The size of the sampling radius r, is dependent on the desired sample size. A larger sample size requires a smaller r and vice versa. The method is described in more detail in Jansen and van Hintum (2007), and is implemented in Genstat (VSN-International 2015). This implementation requires a similarity matrix, with a diagonal of $1's$ and the off-diagonals in the range of $[0, 1]$.

  This similarity matrix comprises of pairwise measures of genetic similarity between individuals, which Jansen and van Hintum calculated using the simple matching coefficient. The authors go on to suggest the Jaccard's similarity index as a suitable alternative (Jansen and van Hintum, 2007). Suggestions for calculating the genetic distance between polyploids have been made in literature (Dufresne et al., 2014), and include the Jaccard similarity index. As part of this study, four genetic distance measures were compared. These measures were chosen due to their suitability for SNP data, polyploids and their frequency of use.

  1. Nei's Genetic Distance makes the biological assumptions of an infinite alleles model and that genetic distances are a result of mutation and drift (Nei, 1972). A categorical marker matrix (AAAA, AAAB, AABB, ABBB, and BBBB) was used as input, and the Nei's distance between two individuals X and Y was calculated using the formula:

$$D_{XY} = -ln \frac{\sum_{i=1}^{2} \sum_{j=1}^{r} p_{ij,x} p_{ij,y}}{\sqrt{\sum_{i=1}^{2} (\sum_{j=1}^{r} p_{ij,x}^2) \sum_{i=1}^{2} (\sum_{j=1}^{r} p_{ij,y}^2)}}$$

  where r represents the total number of markers and $p_{ij,x}$, is the proportion of the i^{th} allele present at the j^{th} locus in individual X. For example, a particular locus with genotype AAAB has $p = 0.75$ for the reference allele "A." This study uses bi-allelic markers hence the summation over

the number of alleles is limited to two terms ($\sum_{i=1}^{2}$). The distance matrix was converted to a similarity matrix by subtracting from one, in accordance with the requirements of the genetic distance sampling algorithm.

2. Euclidean Distance makes no biological assumptions as it is purely a geometric distance measure. Using the numerical coding of the marker matrix (0,1,2,3, and 4) this measure calculates the distance between two individuals X and Y:

$$D_{XY} = \sqrt{\sum_{j=1}^{r}(X_j - Y_j)^2}$$

In this equation Y_j can be interpreted as the number of alternate alleles at the j^{th} marker in individual Y. The Euclidean distance matrix was converted to the similarity measure, and scaled to fit within the desired range $[0, 1]$ using the following transformation:

$$1 - \left(\frac{D_{XY}}{max(D_{XY})}\right)$$

3. Jaccard's Similarity Index does not make any biological assumptions and requires as input the numerical representation of the SNP data. The distance between two individuals X and Y is calculated as:

$$D_{XY} = \frac{\sum_{j=1}^{r}|X_j \cap Y_j|}{\sum_{j=1}^{r}|X_j \cup Y_j|}$$

In the above expression, $|X_j \cap Y_j|$ is the number of alternate alleles common to both individuals X and Y at the j^{th} marker, while the term $|X_j \cup Y_j|$ refers to the total number of alternate alleles at this same marker for individuals X and Y, without repetition (for tetraploids the maximum value for this term is 4). The resulting output was then converted to a similarity matrix.

4. Kosman and Leonard's Genetic Distance differs from previously mentioned genetic distance measures as it takes into account the ploidy level of the individuals (Kosman and Leonard, 2005). With the numerical marker matrix of allele dosages (0, 1, 2, 3, and 4) as input, this measure calculates the similarity between two individuals X and Y:

$$D_{XY} = \frac{1}{r}\sum_{j=1}^{r}\frac{X_j \cap Y_j}{q}$$

In this equation, $X_j \cap Y_j$ corresponds to the number of shared alleles at the j^{th} marker, which is divided by q the number of chromosome copies (4 for tetraploid), and averaged over all r markers.

- **Generalized coefficient of determination (CDmean):** The generalized coefficient of determination is a training set selection method based on the maximization of the precision of the prediction of differences (or contrast) between the average value of the entire population of candidate individuals and each individual in the test set (Rincent et al., 2012). Maximizing Equation 1 (below), leads to the maximization of the precision of contrasts.

$$CD(c) = diag\left[\frac{c'(A - \lambda(Z'MZ + \lambda A^{-1})^{-1})c}{c'Ac}\right] \quad (1)$$

Where c is the matrix of contrasts between each individual without phenotype information and the average of the candidate individuals, λ is the ratio between the residual and additive genetic variances, Z is a design matrix that will be used in GP models to relate observations to genomic values (seen in Equation 3 in a later section), and M is an orthogonal projector on the subspace spanned by the columns of the fixed effects design matrix, X (also seen in Equation 3), such that $M = I - X(X'X)^{-}X'$. A is the additive realized genomic relationship matrix as calculated by VanRaden (2008):

$$A = \frac{QQ'}{2\sum_{j=1}^{r}p_j(1 - p_j)} \quad (2)$$

Where Q is a matrix calculated from $Q_{ij} = W_{ij} + 1 - 2p_j$, with i individuals (rows) and j markers (columns). The term p_j is the frequency of the reference allele of the j^{th} marker and W is the numerical marker matrix, centered and scaled such that genotypes coded as allele dosages $\{0, 1, 2, 3, 4\}$ now become $\{-1, -0.5, 0, 0.5, 1\}$. The supporting literature (Rincent et al., 2012) reports negligible differences in selected samples, when different estimations of the genomic relationship matrix are used. This was confirmed in a small preliminary analysis where three different methods of calculating this matrix were tested, as prediction accuracies were similar between methods. Therefore, the VanRaden method was chosen as it is well-known in the context of genomic prediction.

From the description of λ above, its calculation requires an estimate of trait heritability (h^2) and though we have phenotypic data and can therefore estimate this value for the traits in question, this may not always be the case in practice. Often the decision of which genotypes are to be put in the field to garner phenotypic measurements, is made before estimates of heritability can be performed, as this calculation requires phenotypic data. Secondly, the individuals to be selected may not have to be chosen on the merit of one single trait, but rather by more traits with varying degrees of heritability. The supporting literature (Rincent et al., 2012), suggests and provides evidence that the use of an intermediate value of heritability (example 0.5), selects training sets very similar to those using more extreme values of heritability. A small preliminary analysis was performed and these results confirmed that the heritability estimate had little to no impact on prediction accuracy and therefore, for this study, the heritability input for the CDmean method was set at 0.5 for all traits.

The code for implementing both the CDmean method and genetic distance sampler, can be found in the **Supplementary Material**.

Prediction Scenarios

The training set selection methods were compared by two cross validation schemes: the training-validation (TV) scheme and the training-test (TT) scheme. The TV scheme follows a typical cross-validation approach where a portion of the individuals are used to train the model (training set) and those not part of the training set, used to evaluate model prediction accuracy (validation set). The effect of training set size was assessed by choosing 50, 75, 100, 125, and 150 individuals out of the total 190 with each sample size repeated 100 times. We must consider that the training and validation sets are complementary, therefore the size of the validation set depends on the size of the training set, so comparisons across training set sizes are not equally precise (see **Figure 1**). Additionally, when a diverse set of individuals are chosen, an equally diverse set of individuals are left behind, which may impose some bias. Another important consideration from an application point of view, is that in a real situation a breeder will have individuals that were not phenotyped at all, so we want to assess the performance of the sampling methods assuming that the information of some of the individuals is truly absent, which the TV scheme does not fully represent.

Therefore, a second approach (TT scheme) was used where the composition and size of the validation (test) set, is independent of the composition and size of the training set. In each realization of the TT scheme, we first randomly sampled 40 individuals as test set leaving the remaining 150 as the pool from which to sample the training set. Following the different sampling methods, we chose 25, 50, 75, and 100 genotypes from the remaining 150, as training set (see **Figure 2**). In turn, the sampling of the training set was repeated 50 times, making the accuracy of a particular realization the average of 50 repetitions of the same sampling method, that sample a certain number of individuals to train a particular statistical model and predict a given test set. This entire process was then repeated 50 times, each time with a new test set. This methodology ensures that all training set selection methods train a model that predicts the same test set and gives better assessment of training set selection methods. In addition, we investigated larger sizes of the test set (70, 95). For a test set of 70 individuals, training set sizes are the same as seen above (25, 50, 75, and 100), but for a test set of 95 individuals, the training sets evaluated were of sizes 30, 45, 60, and 75.

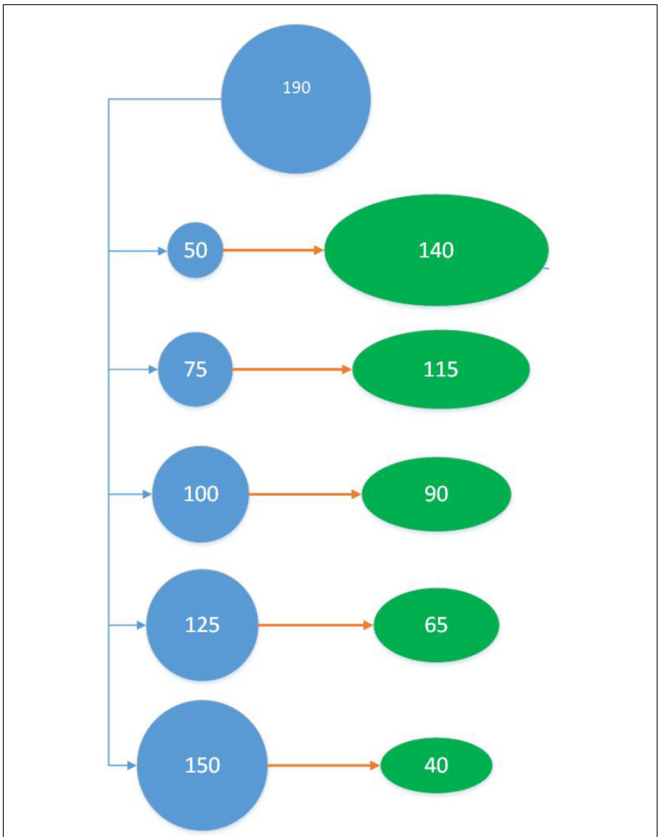

FIGURE 1 | Training-Validation (TV) Scheme: Out of 190 individuals, 50 individuals are sampled as the training set to train the model and validated using the remaining 140 individuals in the green oval. This is repeated 100 times for each training set sampling method. The entire process was then repeated for training set sizes 75, 100, 125, and 150 (which impacts the size of the validation set).

Genomic Prediction Models

The purpose of this study was to uncover a superior training set sampling method based on the accuracy of predictions. These predictions were generated with three different whole genome regression models, in order to investigate the presence/absence of an interaction between training set selection method and genomic prediction model.

- **GBLUP:**

$$y = X\beta + Zu + \epsilon \tag{3}$$

For Equation 3, y is a vector of phenotypic BLUEs, β is a vector of fixed effects (only the intercept in our case), u is a vector of genotypic values with distribution $u \sim N(0, A\sigma_g^2)$. A is the genomic relationship matrix as calculated in Equation 2 and σ_g^2 is the additive genetic variance. X and Z are design matrices as described previously and ϵ is the vector of residuals with distribution $\epsilon \sim N(0, \sigma_\epsilon^2)$. σ_ϵ^2 is the residual variance.

- **RKHS:** The model for Reproducing Kernel-Hilbert Spaces (RKHS) is the same as Equation 3, with one difference in that the genotypic values have a different distribution: $u \sim N(0, K\sigma_g^2)$. The genomic relationship matrix A, is replaced by the kernel matrix, $K = exp^{-\frac{D}{\theta}}$, where D is a Euclidean distance matrix and θ a tuning parameter. The tuning parameter controls how fast the relationship between two genotypes decays as the distance between the corresponding pairs of marker vectors increases (Jiang and Reif, 2015). To estimate θ, a grid search was conducted between $(0, 1]$ and the value that gave the maximum log-likelihood was chosen (Endelman, 2011). Applying RKHS in this study allows for the implicit modeling of non-additive effects.
- **BAYES Cπ:**

$$y = X\beta + Wb + \epsilon \tag{4}$$

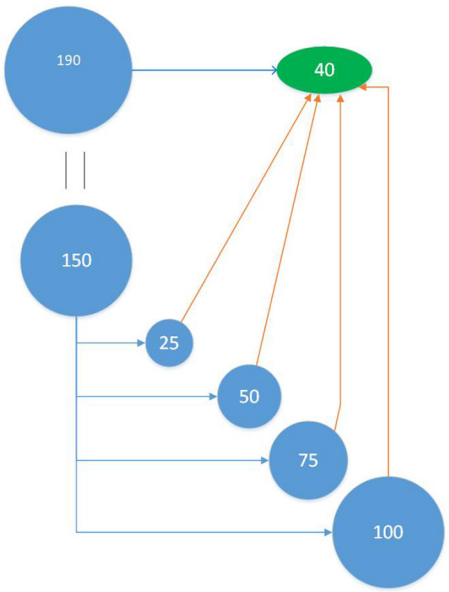

FIGURE 2 | Training-Test (TT) Scheme: Out of the 190 individuals a test set of 40 individuals were randomly selected. The remaining 150 individuals were now candidates for the training set and from this set, 25 individuals were selected and used to train the model that predicts the test set. This was performed 50 times for each training set sampling method. This process was repeated for 50 different test sets. The analysis was performed at varying training set sizes (25, 50, 75, and 100).

In Equation 4, where W is our matrix of marker information, b is a vector of marker effects. Bayes Cπ assumes that marker effects come from a mixture distribution with a proportion of markers (π) having zero effect and the remainder ($1-\pi$) having non-zero effects, such that for the j^{th} marker:

$$b_j = \begin{cases} 0 & : \text{with probability } \pi \\ \sim N(0, \sigma_b^2) & : \text{with probability } 1-\pi \end{cases}$$

The proportion of zero effect markers π, was estimated from the data. For this study, 5,000 iterations were performed with 2,500 discarded as burn-in, with the BGLR package (Prez and Campos, 2014). In preliminary analyses, larger number of iterations were tested and the outcomes were identical, in terms of prediction accuracy and convergence diagnostics.

Prediction Accuracy

As mentioned in previous sections, the ranking of the training set construction methods will be based on a measure of prediction accuracy. For both the TV and TT schemes, the observed phenotypic values of the training set are fed to the statistical models to estimate marker effects, while the phenotypic values of the validation (TV scheme) and the test set (TT scheme), are hidden from the model. Predictions are made on those individuals with hidden phenotypes, and the prediction accuracy is defined as the Pearson correlation between observed phenotypic values and the predicted genotypic values. Factors that may influence prediction accuracy are sample size, statistical model and the training set construction method, as well as various interactions between these factors. To answer this question, an Analysis of Variance (ANOVA) was carried out where the correlation (prediction accuracy) is treated as the response variable such that $accuracy = f(size, model, method)$ in a full factorial model. To conform to normality assumptions, these correlations (accuracies) were transformed using Fisher's z transformation, $z = \frac{1}{2}(ln(\frac{1+r}{1-r}))$.

All analyses were executed in R Core Team (2020), except for genetic distance sampling which was performed in Genstat as mentioned previously.

RESULTS

The 3,763 SNPs were reduced to 3,262 after the following filtering steps. For the 190 phenotyped tetraploid lines, monomorphic markers, unmapped markers, markers with a minor allele frequency of <5% and markers with missing values for more than 30 of the 190 individuals were removed.

Population Structure

The classification of the population into the six market classes, gives two subpopulations with <20 individuals. This is not ideal for stratified sampling as parameter estimates from these very small subgroups will produce large standard errors. Furthermore, based on past population structure results for this diversity panel, there are indications that some of these sub-populations can be merged.

PCA and DAPC results show that the six market classes can indeed be reduced to a smaller number of groups (**Figure 3**). Principal Components Analysis (**Figure 3A**) found that the first two principal components account for <10% of the explained variance with the 1st principal component capturing 5% of the variability, while the 2nd component explains only 3.55%. The decision on which classes should be merged were made by inspecting the results from DAPC (see **Figure 3B**). For this analysis, 100 principal components and three discriminant functions were chosen. From here we see that the French Fry processing and Table Russet market classes show considerable overlap, as well as the Chip processing and Round White table market classes.

The pigmented class is clearly separated but one question arose: Where does the yellow market class belong? AMOVA analyses found that genetic variation due to population structure was the highest (16.6%), when the yellow class was placed with chip processing and round white table classes, as suggested by the DAPC plot (**Figure 3B**). Other population structure configurations were analyzed, including each of the six separate market classes as its own sub-population, as well as maintaining the three clearly separated groups seen in **Figure 3B**, but placing the yellow market group with the pigmented class (see **Appendix**). Placing the yellow class with the chip and round-white class, instead of the pigmented class was supported by both AMOVA analyses and pairwise Fst statistics between the groups. Between Yellow and Pigmented, $Fst = 0.0165$, while between Yellow and Chip Processing-Round White table, $Fst = 0.0098$ (where Fst values closer to zero indicate populations that are

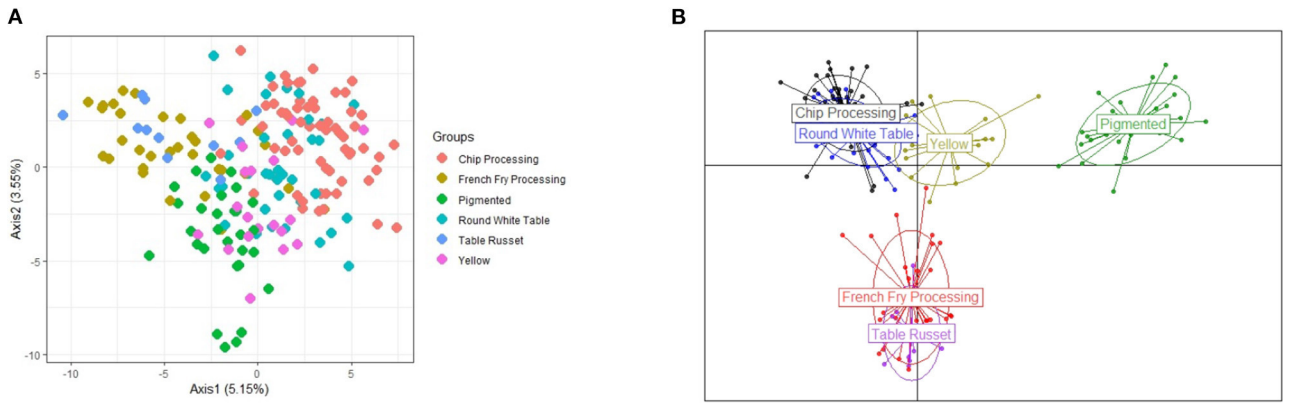

FIGURE 3 | Illustration of the population structure explained by the first 2 principal components (PCA) and discriminant analysis of the principal components (DAPC) showing the separation of the 6 market classes. **(A)** PCA with market classes. **(B)** DAPC of market classes.

TABLE 1 | Correlation between different genetic distance matrices.

	Euclidean	Nei	Jaccard
Nei	0.989		
Jaccard	0.927	0.941	
Kos.&Leo.	0.967	0.975	0.977

more genetically similar). For the remainder of the study, the discrete population structure used for stratification is defined by the three groups suggested in **Figure 3B**, with the yellow market class merged with the neighboring group of chip processing and round white table potatoes.

Genetic Distance Measures

Four different genetic distance measures were used to perform genetic distance sampling, and the sampled individuals were used to train the model. Prediction was performed on the left out individuals as described in the TV scheme. The similarities (correlations) between the different genetic distance matrices were assessed by a Mantel test (**Table 1**).

There is very little difference between the distance measures for the material in this study. The lowest correlations (0.927 and 0.941) occurred with the Jaccard distance measure, however this degree of similarity is still quite high.

The prediction accuracies from a common GBLUP model were quantified for three different traits (tuber length, fructose and glucose content), at sample sizes ranging from 50 to 150, for different genetic distance measures (**Figure 4**). It can be concluded that the choice of distance measure had a minor impact on prediction accuracy. Prediction accuracy is expected to increase as sample size increases and Euclidean distance was the most consistent measure across all traits. The remaining three measures displayed non-monotonically increasing prediction accuracies as sample size increased. Additionally, the Kosman Leonard distance, along with having very little application in literature, becomes computationally heavy when there are more than 10,000 markers. For this study, the Euclidean distance will be used henceforth when applying genetic distance sampling for training set construction.

Genomic Prediction: TV Scheme

After determining a suitable distance measure for genetic distance sampling, methods for acquiring the training set were compared (**Figure 5**).

Each row of **Figure 5** shows a single trait with the different genomic prediction models, and compares the prediction accuracies across sample sizes ranging from 50 to 150. For all traits, a difference is clearly observed between training set selection methods: with simple random and stratified random sampling (random methods) behaving similarly while genetic distance sampling and the CDmean method (analytical methods) sampled training sets, gave more accurate predictions. As expected, an increase in sample size increased prediction accuracy, but this was at a higher rate when using the analytical methods of selecting individuals. The lines above and below the points indicate the standard errors of the estimate of average accuracy, and the random sampling methods resulted in larger standard errors than the analytical methods. For all trait-statistical model combinations, the random methods of selecting the training set were not significantly different; stratifying the population before sampling, did not improve the accuracy of genomic prediction, in comparison to a simple random sample of the training set. Even though the analytical methods consistently performed better than the random methods, the comparative performance between the two analytical methods varied with traits. For tuber length, the genetic distance sampler selected a more optimal training than the CD method at lower sample sizes (50 and 75), but this difference diminished as the size of the training set increased. The CD mean method generally outperformed the genetic distance sampler in predicting fructose and sucrose,

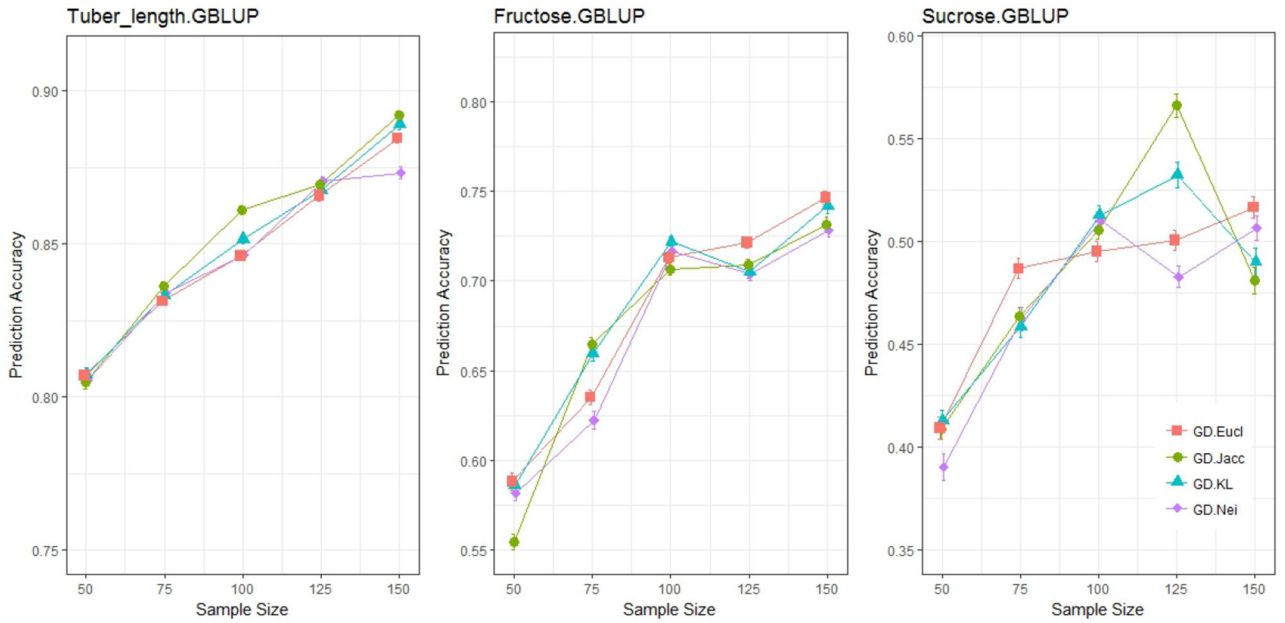

FIGURE 4 | Prediction accuracy under different distance measures for genetic distance sampling with a GBLUP statistical model. Each graphical window represents a different trait, with varying sample sizes on the *x*-axis and the outcome variable, prediction accuracy on the *y*-axis. The different colored lines represent different distance measures.

more noticeably so at higher sample sizes. Interestingly, at sample size 50 and 75, genetic distance sampling led to more accurate predictions of sucrose content, a result also observed for tuber length. Despite these minor differences, the results across all traits give clear support for utilizing analytical methods of selecting the training set, and some indication that the CDmean method is the better of the two analytical training set selection strategies.

The results shown in **Figure 5**, include information about the three different statistical models. The possibility of an interaction between statistical model and training set selection method was evaluated in this study, and results from an ANOVA analysis were used to quantify the impact of this interaction (**Table 2**).

The magnitude of the *F*-values in **Table 2** indicate how important a term is for predicting the outcome, which is the accuracy of genomic predictions in this case. The most important factor for driving genomic prediction accuracies is sample size, followed by the training set selection method and then the interaction between these two variables. The interaction of interest, between sampling method and statistical model, explains very little of the variation in prediction accuracy. There is no particular combination of sampling method and statistical model that results in more accurate predictions but rather, the main effects of these two variables. Results in **Table 2** are based on tuber length, and these results were consistent across all traits, with sampling method being highly significant, and its interaction with statistical model, non-significant. An interesting result is the significant interaction between sample size and sampling method which was consistent across all traits. This means that the sampling methods do not benefit equally from an increase in sample size, a result also observable from **Figure 5**.

For fructose, when the sample size is tripled (from 50 to 150), simple random sampling and stratified sampling improved by 19 and 23%, respectively, whereas genetic distance sampling and the CDmean method resulted in improvements of 27 and 31%, respectively. For sucrose, the CDmean method showed a 37% improvement by tripling the sampling size while simple random sampling improved by 25%. The relative improvement of using a analytical sampling method was greater for sucrose and fructose content. At the median sample size of 100, CDmean showed an improvement in prediction accuracy of 4, 14, and 13% for tuber length, fructose and sucrose content, respectively, when compared to simple random sampling. The genetic distance sampler for these traits (tuber length, fructose and sucrose content, respectively), showed improvements of 5.5, 10.5, and 10.5% in comparison to simple random sampling.

Genomic Prediction: TT Scheme

As discussed before, the objectives for using Genomic Prediction may vary. In many cases the objective is to predict new breeding lines (or clones) and for this scenario we have randomly selected a test set of 40 out of the 190 individuals. These 40 individuals represent the independent test set, and all sampling methods will construct the training set from the remainder of individuals. The trained model then performs predictions for the test set. In this way, each sampling method predicts the same test set.

Similar to the previous section, we looked at the prediction accuracy for three statistical models with sample sizes ranging

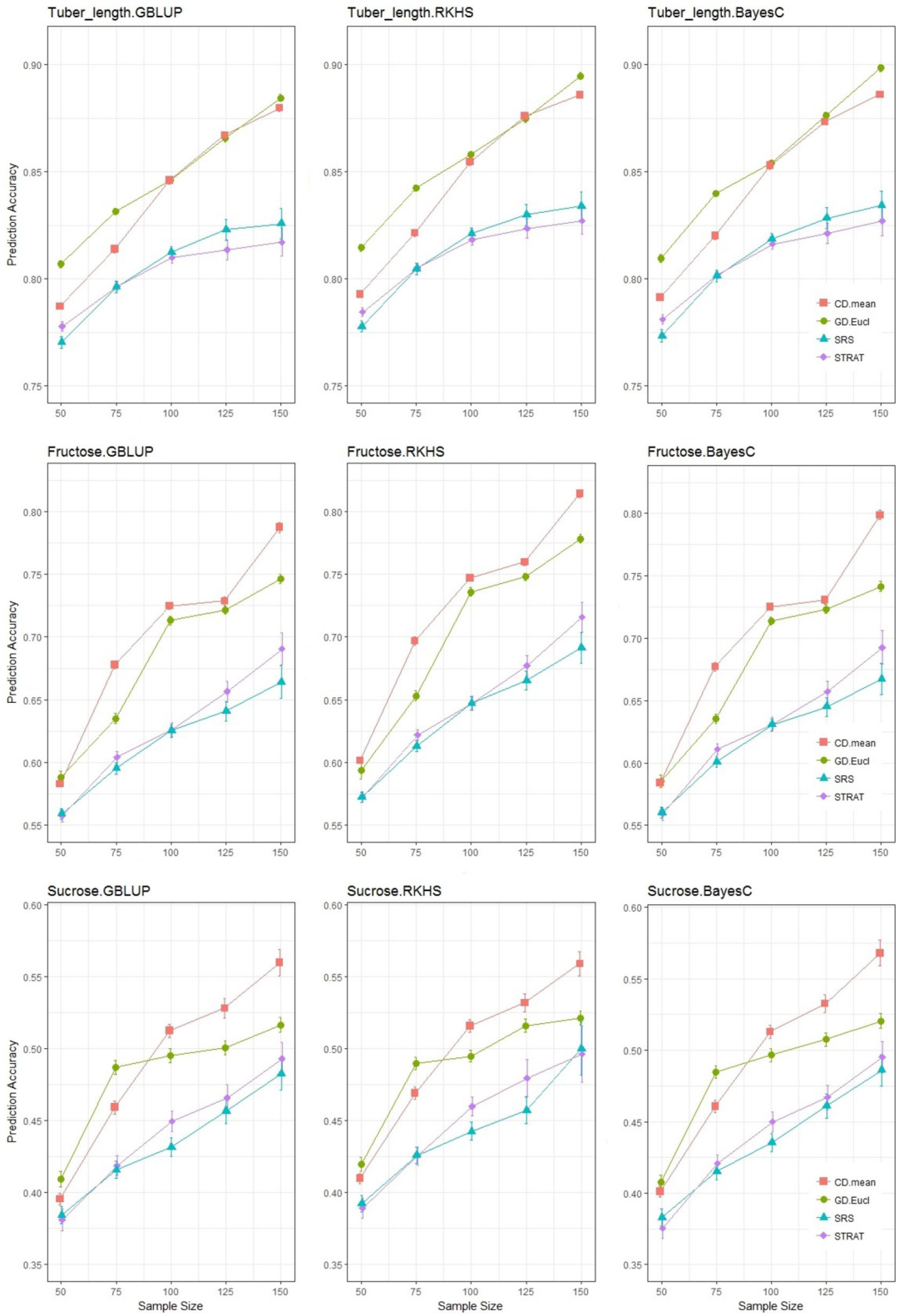

FIGURE 5 | Prediction accuracy for the 3 traits under the TV scheme (training and validation only). Each graphical window displays a different trait-statistical model combination, with varying sample sizes on the *x*-axis and prediction accuracy on the *y*-axis. The different colored lines represent different training set selection methods.

TABLE 2 | ANOVA table showing the significance of the statistical model, sample size, training set selection method, and interactions for the prediction accuracies of tuber length (TV Scheme).

	df	SS	MS	*F*-value	Pr(>F)
Method	3	0.278	0.0925	319	$< 2 \times 10^{-16}$
Sample size	1	0.386	0.386	1,330	$< 2 \times 10^{-16}$
Model	2	9.08×10^{-3}	4.54×10^{-3}	15.7	1.27×10^{-5}
Method: sample size	3	0.0605	0.0202	69.6	4.66×10^{-15}
Method: model	6	5.58×10^{-4}	9.30×10^{-5}	0.321	0.922
Sample size: Model	2	9.42×10^{-4}	4.71×10^{-4}	1.63	0.211
Method: sample size: model	6	3.53×10^{-4}	5.88×10^{-5}	0.203	0.974
Residuals	36	0.104	2.90×10^{-4}		

from 25 to 100, and compared the impact of the sampling method (**Figure 6**). The accuracies of the TT Scheme are a bit lower and are accompanied by larger standard errors than those observed in the TV Scheme, due to the application involving a test set, which is usually more difficult to predict but represents a more realistic scenario encountered by breeders. Nonetheless the decrease in accuracy was not drastic. The differences between sampling methods is still present, but less obvious than in the TV Scheme, especially at higher sample sizes where the accuracies of the various sampling methods converged as was expected, due to the significant overlap of individuals sampled in a limited population space of 150 varieties. This convergence is not observed in the TV Scheme and will be discussed in another section. At the lower sample sizes, where the potential overlap of training sets is reduced, the analytical methods give significantly higher accuracies than the random methods. For tuber length, genetic distance sampling and the CDmean method result in similar prediction accuracies for sample sizes $\geq$50, but for sucrose content, this similarity was dependent on the statistical model applied.

In comparison to the TV scheme, the results of the TT scheme exhibit a more significant impact due to statistical model, and to test whether there is an interaction with the sampling method an ANOVA analysis was conducted.

Similar to the results from the TV scheme shown in **Tables 2**, **3** shows that for the TT scheme, sample size was the most important factor driving prediction accuracy, and there was no interaction between the statistical model and the sampling method. It was noteworthy that the hierarchy of importance of predictive variables was quite different between schemes. Our factor of interest, sampling method, though still significant in the TT application, was not the second most important variable as seen before, but replaced by statistical model in the hierarchy. Also different to the TV scheme, the TT scheme results show no interaction between sample size and the sampling method. The results in **Figure 6** and **Table 3** were similar to those observed for fructose content, with CDmean only slightly outperforming the rest, but with even less differentiation between sampling methods. The ANOVA analysis for fructose content (not shown), showed that there was little to no impact of different training set construction methods.

Although this paper does not primarily focus on statistical models, it is still interesting to observe the differences in predictive performance (**Table 4**). For all traits, the GBLUP model gave the lowest accuracy of predictions, while the Bayes C model worked just as well as the RKHS model.

Application of the TT scheme to breeding programs, usually involves a test set of hundreds or even thousands of new potential cultivars. In this study it was impossible to emulate this application, still the impact of increasing the test size was investigated. For this investigation, we conducted the same analyses as seen in TT scheme but used a larger test set (70 and 90 individuals). There were no changes in the findings; the analytical methods, especially CDmean, sampled training sets that predicted the test sets with greater accuracy than the random methods (results not shown). Similar to the results seen above, these differences disappeared at larger sample sizes and were only evident at smaller training set sizes, where the overlap of sampled individuals between methods was minimal.

DISCUSSION

Training set construction has been proven to be important for GP in diploids and in this study, shown to be important for GP in tetraploids. Both ploidy levels benefit from incorporating genomic information into analytical methods of sampling the training set, when compared to random methods that do not directly utilize genomic information.

Only 190 varieties were included in this study which may limit the extrapolation of results to traditional breeding programs. Breeders often make selections within a particular market group. In these scenarios, one must decide if to train models using only individuals belonging to the target market group or allow for the borrowing of information from other market classes. Our study was too small to answer this question, however it has been shown that combining individuals from both within and across market classes, can lead to predictions that are as good as, and often better than predictions made from exclusively within the market class (Rio et al., 2019). This is especially valid when the population structure is less definitive, as seen in this study.

As we are predicting heterogeneous populations, the use of interaction models may be considered (Lehermeier et al., 2015), where population structure induces heterogeneity of marker effects. For the interaction models, sub-populations should be large enough and definitive enough to estimate marker effects, but in this study our sub-populations were small. As population structure and size increase in magnitude, the Sparse Selection Index is another promising alternative (Lopez-Cruz and Campos, 2021).

TV Scheme

For the training-validation scenario, results show a clear differentiation between the random methods (simple random sampling and stratified random sampling) and the analytical methods (genetic distance sampling and CDmean method). This separation between methods was not dependent on the statistical model used to make predictions which was confirmed by ANOVA analyses of prediction accuracies. As sample size

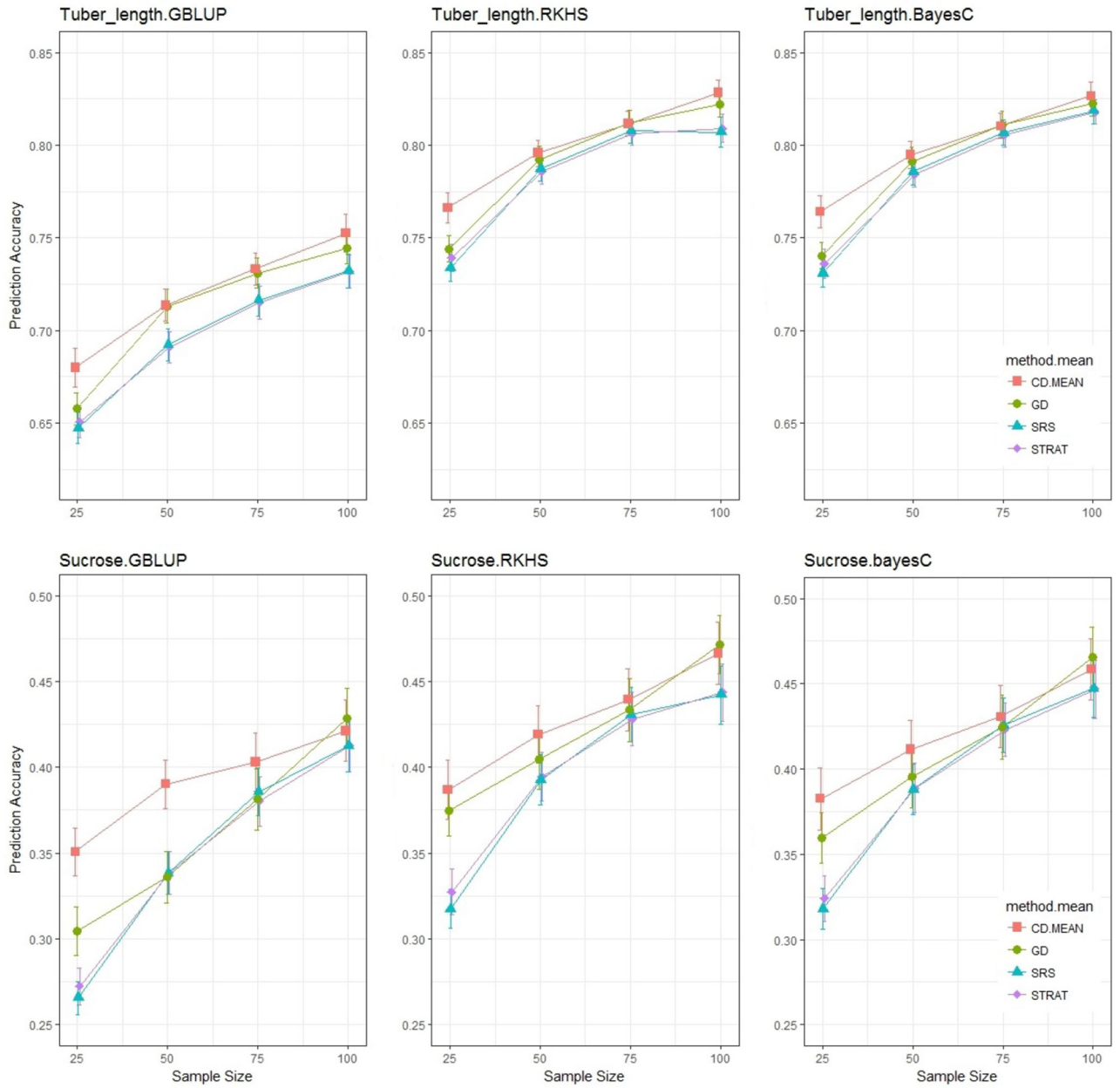

FIGURE 6 | Prediction Accuracy for tuber length and sucrose content with a test set of size 40. Each graphical window represents a different trait-statistical model combination, with varying sample sizes on the *x*-axis and prediction accuracy on the *y*-axis. The different colored lines represent different training set sampling methods.

increased so did prediction accuracy due to the fact that the estimation of marker effects is improved as the size of the training set increases, a finding also reported in studies of diploid crops (Rincent et al., 2012; Daetwyler et al., 2013; Bustos-Korts et al., 2016; Akdemir and Isidro-Sanchez, 2019). The improvement in accuracy awarded from a larger sample, is greater when applying an analytical method of sampling the training set. This result was supported by the significant interaction between sampling method and sample size. In training set construction for the TV scheme, we are essentially choosing a subset of individuals (randomly or analytically), that would calibrate the model used to make predictions on the subset of individuals not chosen for training; in essence the training set and validation sets are complements of each other. If we were to picture the population space spread evenly over four quadrants, and during training set construction, by chance all the members of a given quadrant belonged to the training set, then this quadrant would not be represented in the validation set. Our model would be trained

TABLE 3 | ANOVA table showing the importance of the statistical model, sample size and training set selection method and interactions for the prediction accuracies of tuber length (TT Scheme).

	df	SS	MS	*F*-value	Pr(>F)
Method	3	0.0136	4.52×10^{-3}	5.94	0.00352
Sample size	1	0.213	0.213	280	9.75×10^{-15}
Model	2	0.404	0.202	265	$< 2 \times 10^{-16}$
Method: sample size	3	8.79×10^{-4}	2.93×10^{-4}	0.385	0.765
Method: Model	6	3.71×10^{-4}	6.18×10^{-5}	0.0810	0.998
Sample size: Model	2	2.01×10^{-3}	1.01×10^{-3}	1.32	0.285
Method:sample size:model	6	3.45×10^{-4}	5.75×10^{-5}	0.0760	0.998
Residuals	24	0.0183	7.61×10^{-4}		

Higher F-values or Mean Sum Sq values indicate higher predictive power of a variable.

TABLE 4 | Marginal means and standard errors for prediction accuracy for varying combinations of statistical model (columns) and trait (rows).

	GBLUP	RKHS	BAYES.C
Tuber length (s.e. = 0.010)	0.708	0.792	0.792
Fructose (s.e. = 0.007)	0.450	0.580	0.571
Sucrose (s.e. = 0.005)	0.364	0.412	0.406

in a space where it is not making predictions, leading to poor predictive potential. As the size of the training set increases using random methods, there is a chance that we continue to calibrate the model using redundant misrepresentative information, and the gain from increasing sample size is contested by predicting individuals that are genetically distant from the members of the training set. For this reason, the predictive power gained by adding one individual to the training set, is greater when using an analytical method for selecting the training set over a random sampling method. Analytical methods of training set construction allow the space occupied by the training set to be similar to that of the validation set, and as we increase the size of the training set, the information provided for model calibration continues to describe the entire genetic space in more detail, and not randomly over-represent a few areas with redundant information.

Taking a closer look at the random methods, we see that stratifying our samples had very little impact on prediction accuracy in comparison to simple random sampling. Diploid studies have shown that stratification based on population structure information may not be beneficial to constructing the training set, when there is no extensive separation between sub-populations (Isidro et al., 2015; Bustos-Korts et al., 2016). The panel of tetraploid potatoes used in this study showed little population structure, with only 16% of the total variation due to population structure. Therefore, stratification before sampling did not improve the accuracy of GP in comparison to simple random samples, similar to the results of comparable studies of diploid species with little sub-population separation (Isidro et al., 2015).

For sucrose and fructose content, the CDmean method sampled training sets that lead to more accurate predictions, however for tuber length, the genetic distance sampler chose an equally optimal training set. The extra information that is incorporated by the CDmean method, may help in choosing a training set, better equipped for traits that are harder to predict. In a study comparing training set construction methods among various diploid species and different traits (Bustos-Korts et al., 2016), the results showed no significant difference between the CDmean method and genetic distance sampler. Genetic distance sampling establishes a radius that is used to exclude individuals that are genetically close to a previously chosen member of the training set, and only considers genomic information (genetic distance). The CDmean method though, makes use of more information than the genetic distance sampler: trait variability and heritability. For traits that are influenced by non-genetic (environmental) factors, like fructose and sucrose content (Kumar et al., 2004), genomic information alone will not be as beneficial as having both genomic and phenotypic information. The combined information of trait variability and heritability, as well as genomic relationships between individuals, allows the CDmean method to construct a training set that produces higher accuracies for these traits. However, this necessity for phenotypic input information, in addition to the increased computational load, can make the CDmean method less attractive than genetic distance sampling.

Distance Measures

The differences between distance measures is very small when compared by correlation diagnostics. We were not able to explain the unexpected behavior exhibited by the Nei's, Jaccard and Kosman and Leonard genetic distances, where for fructose and sucrose content, the accuracy of predictions did not monotonically increase as sample size increased. The fact that Euclidean distance produced accuracies that were monotonically increasing with sample size, motivates the use of this measure in this study. However, this finding is not conclusive for all tetraploid studies: only bi-allelic markers were available for this study, but tetraploid individuals can have up to four alleles (Silva et al., 2005; Salimi et al., 2016). The Kosman and Leonard distances can utilize this information as it considers the number of different alleles at a given marker, and this is expected to produce better measures of distance between individuals (Kosman and Leonard, 2005; Dufresne et al., 2014), whereas the Euclidean distance uses a count of one particular allele (reference allele) as input to calculate genetic distances. This study did not contain the multi-allelic marker information needed to truly test the differences between the distance measures, and for scenarios like this that are limited to bi-allelic markers, the difference between distance measures will not be relevant.

TT Scheme

To investigate the impact that the training set has on the prediction of new potential cultivars, the TT scheme was introduced which includes a randomly chosen test set. As expected, there was a decrease in overall prediction accuracy (Akdemir and Isidro-Sanchez, 2019). The divergence in accuracy

between the random and analytical methods as sample size increased, observed in the TV scheme was not seen in the TT scenario. This is due to the fact that all methods predict the same group of individuals, and leave a limited pool of candidates to be selected for training the model. As a result, there was overlap in the training sets sampled by the various sampling methods. Secondly, the composition of the trainng set had no affect on the individuals where predictions were made, an unavoidable situation with the TV scheme. The TT Scheme reveals that the differences between training set construction methods depend on the scenario for which these methods are applied; scenarios with an independent test set (new breeding material) or instances where it may be more cost and time efficient to phenotype a few individuals and predict the rest (phenotyping platforms, TV scheme). These results are not conclusive, due to the moderate number of individuals in this study. The performance at the smaller sample sizes for the TT scheme may give an impression of what an ideal situation would look like, where there is a large population thus minimizing the overlap of individuals in the training sets constructed by the different methods. At these low sample sizes, the CDmean method constructed training sets led to more accurate predictions. Similar to the TV scenario, there is evidence that the utilization of both genomic and phenotypic information by the CDmean method is more beneficial for predicting traits highly influenced by non-genetic (environmental) factors. The genetic distance sampler maintains its position as the second best sampler. In spite of the limitation created by the population size, the evidence is still substantial: for GP of tetraploids in a training-test scenario, analytical methods of sampling the training set lead to better predictions, as seen also in diploids (Bustos-Korts et al., 2016; Akdemir and Isidro-Sanchez, 2019).

Prediction Models

The performance of the prediction models can be explained by the architecture of the traits analyzed. GBLUP models work best for traits controlled by many small effects while models that perform marker selection are better suited for traits that are controlled by a few large effect QTL (de los Campos et al., 2013). A previous Genome Wide Association Study (GWAS) was conducted on the same diversity panel as this study, where significant QTLs were detected for tuber length, but not for sucrose and fructose content (Rosyara et al., 2016). Other studies have found that sucrose and fructose content are controlled by a small number of loci (Bradshaw et al., 2008; Sliwka et al., 2016; Rak et al., 2017). It is therefore not surprising that the BayesCπ model was able to make better predictions of all three traits in comparison to the GBLUP model.

Having four copies of each chromosome, one may expect that tetraploids exhibit more inter-locus interactions (epistasis) in comparison to diploids (Stich and Gebhardt, 2011). When non-additive effects like dominance and epistasis are present, they can be captured with the RKHS model (Gianola and van Kaam, 2008). Tuber length did not benefit from accounting for these effects while sucrose and fructose content showed little improvement. Fry color, strongly related to sugar content (Pritchard and Adam, 1994), can attribute the majority of its variability to additive effects, however there is a small contribution by non-additive effects (Endelman et al., 2018). This helps to explain the small but present improvement of the RKHS model over the BayesCπ model for these two traits.

CONCLUSIONS

- Genomic prediction of individuals with limited population structure requires a sampling method that uniformly covers the genetic space of the breeding population as opposed to stratified sampling based on discrete classifications into sub-populations.
- When GP is implemented to lessen the resources consumed by phenotyping, a portion of the population is phenotyped to train a model that predicts the remaining individuals. The TV scheme results show the value of explicitly using genomic information to sample the training set.
- The CDmean method of selecting a training set should be utilized for genomic prediction in potato, as it is robust to sample size, trait architecture, statistical model and application scenario.
- Further investigation has to be done before these results can be extrapolated to other traits and other polyploid crops. Testing on larger pools of varieties with more distinct subgroups is required.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

SW performed the analyses and drafted the manuscript. MM, CM, HM, and RV contributed to the discussion on analytical models and data preparation. FE guided analyses and was the general overseer for the project. All authors significantly contributed to the present study, read, and approved the final manuscript.

FUNDING

This study was funded by the following sources: Solynta, Meijer Potato, Pepsico, and the Dutch Research Council (NWO).

ACKNOWLEDGMENTS

This work is part of the research programme PredAPloid with project number 14520, which is financed by the NWO.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.771075/full#supplementary-material

REFERENCES

Akdemir, D., and Isidro-Sanchez, J. (2019). Design of training populations for selective phenotyping in genomic prediction. *Sci. Rep*. 9:1446. doi: 10.1038/s41598-018-38081-6

Asoro, F. G., Newell, M. A., Beavis, W. D., Scott, M. P., and Jannink, J.-L. (2011). Accuracy and training population design for genomic selection on quantitative traits in elite north american oats. *Plant Genome* 4, 132–144. doi: 10.3835/plantgenome2011.02.0007

Bradshaw, J., Hackett, C., Pande, B., Waugh, R., and Bryan, G. (2008). Qtl mapping of yield, agronomic and quality traits in tetraploid potato (solanum tuberosum subsp. tuberosum). *Theor. Appl. Genet*. 116, 193–211. doi: 10.1007/s00122-007-0659-1

Bustos-Korts, D., Malosetti, M., Chapman, S., Biddulph, B., and van Eeuwijk, F. (2016). Improvement of predictive ability by uniform coverage of the target genetic space. *G3* 6, 3733–3747. doi: 10.1534/g3.116.035410

Daetwyler, H. D., Calus, M. P. L., Pong-Wong, R., de Los Campos, G., and Hickey, J. M. (2013). Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics* 193, 347–365. doi: 10.1534/genetics.112.147983

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., and Calus, M. P. L. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193, 327–345. doi: 10.1534/genetics.112.143313

Dufresne, F., Stift, M., Vergilino, R., and Mable, B. K. (2014). Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. *Mol. Ecol*. 23, 40–69. doi: 10.1111/mec.12581

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with r package rrblup. *Plant Genome* 4, 250–255. doi: 10.3835/plantgenome2011.08.0024

Endelman, J. B., Carley, C. A. S., Bethke, P. C., Coombs, J. J., Clough, M. E., da Silva, W. L., et al. (2018). Genetic variance partitioning and genome-wide prediction with allele dosage information in autotetraploid potato. *Genetics* 209, 77–87. doi: 10.1534/genetics.118.300685

Excoffier, L., Smouse, P. E., and Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among dna haplotypes: application to human mitochondrial dna restriction data. *Genetics* 131, 479–491. doi: 10.1093/genetics/131.2.479

Gianola, D., and van Kaam, J. B. C. H. M. (2008). Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* 178, 2289–2303. doi: 10.1534/genetics.107.084285

Guo, Z., Tucker, D. M., Basten, C. J., Gandhi, H., Ersoz, E., Guo, B., et al. (2014). The impact of population structure on genomic prediction in stratified populations. *Theor. Appl. Genet*. 127, 749–762. doi: 10.1007/s00122-013-2255-x

Habyarimana, E., Parisi, B., and Mandolino, G. (2017). Genomic prediction for yields, processing and nutritional quality traits in cultivated potato (solanum tuberosum l.). *Plant Breed* 136, 245–252. doi: 10.1111/pbr.12461

Hamilton, J. P., Hansey, C. N., Whitty, B. R., Stoffel, K., Massa, A. N., Van Deynze, A., et al. (2011). Single nucleotide polymorphism discovery in elite north american potato germplasm. *BMC Genomics* 12:302. doi: 10.1186/1471-2164-12-302

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

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

Hirsch, C. N., Hirsch, C. D., Felcher, K., Coombs, J., Zarka, D., Van Deynze, A., et al. (2013). Retrospective view of north american potato (solanum tuberosum l.) breeding in the 20th and 21st centuries. *G3* 3, 1003–1013. doi: 10.1534/g3.113.005595

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

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. New York, NY: Springer.

Jansen, J., and van Hintum, T. (2007). Genetic distance sampling: a novel sampling method for obtaining core collections using genetic distances with an application to cultivated lettuce. *Theor. Appl. Genet*. 114, 421–428. doi: 10.1007/s00122-006-0433-9

Jiang, Y., and Reif, J. C. (2015). Modeling epistasis in genomic selection. *Genetics* 201, 759–768. doi: 10.1534/genetics.115.177907

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24, 1403–1405. doi: 10.1093/bioinformatics/btn129

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet*. 11:94. doi: 10.1186/1471-2156-11-94

Kosman, E., and Leonard, K. J. (2005). Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploid species. *Mol. Ecol*. 14, 415–424. doi: 10.1111/j.1365-294X.2005.02416.x

Kumar, D., Singh, B., and Kumar, P. (2004). An overview of the factors affecting sugar content of potatoes. *Ann. Appl. Biol*. 145, 247–256. doi: 10.1111/j.1744-7348.2004.tb00380.x,

Lehermeier, C., Schn, C.-C., and de Los Campos, G. (2015). Assessment of genetic heterogeneity in structured plant populations using multivariate whole-genome regression models. *Genetics* 201, 323–337. doi: 10.1534/genetics.115.177394

Lopez-Cruz, M., and Campos, G. (2021). Optimal breeding-value prediction using a sparse selection index. *Genetics* 218:iyab030. doi: 10.1093/genetics/iyab030

Meuwissen, T. H., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. doi: 10.1093/genetics/157.4.1819

Nei, M. (1972). Genetic distance between populations. *Am. Nat*. 106, 283–292. doi: 10.1086/282771

Prez, P., and Campos, G. (2014). Genome-wide regression prediction with the bglr statistical package. *Genetics* 198, 483–495. doi: 10.1534/genetics.114.164442

Pritchard, M. K., and Adam, L. R. (1994). Relationships between fry color and sugar concentration in stored russet burbank and shepody potatoes. *Am. Potato J*. 71, 59–68. doi: 10.1007/BF02848745

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

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: https://www.R-project.org/

Rak, K., Bethke, P. C., and Palta, J. P. (2017). Qtl mapping of potato chip color and tuber traits within an autotetraploid family. *Mol. Breed*. 37:15. doi: 10.1007/s11032-017-0619-7

Rincent, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., et al. (2012). Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: Comparison of methods in two diverse groups of maize inbreds (zea mays l.). *Genetics* 192, 715–728. doi: 10.1534/genetics.112.141473

Rio, S., Mary-Huard, T., Moreau, L., and Charcosset, A. (2019). Genomic selection efficiency and a priori estimation of accuracy in a structured dent maize panel. *Theor. Appl. Genet*. 132, 81–96. doi: 10.1007/s00122-018-3196-1

Rosyara, U. R., De Jong, W. S., Douches, D. S., and Endelman, J. B. (2016). Software for genome-wide association studies in autopolyploids and its application to potato. *Plant Genome* 9. doi: 10.3835/plantgenome2015.08.0073

Salimi, H., Bahar, M., Mirlohi, A., and Talebi, M. (2016). Assessment of the genetic diversity among potato cultivars from different geographical areas using the genomic and est microsatellites. *Iran J. Biotechnol*. 14, 270–277. doi: 10.15171/ijb.1280

Silva, D. H. N., Hall, A. J., Rikkerink, E., McNeilage, M. A., and Fraser, L. G. (2005). Estimation of allele frequencies in polyploids under certain patterns of inheritance. *Heredity* 95:327. doi: 10.1038/sj.hdy.6800728

Slater, A. T., Cogan, N. O. I., Forster, J. W., Hayes, B. J., and Daetwyler, H. D. (2016). Improving genetic gain with genomic selection in autotetraploid potato. *Plant Genome* 9. doi: 10.3835/plantgenome2016.02.0021

Sliwka, J., Sotys-Kalina, D., Szajko, K., Wasilewicz-Flis, I., Strzelczyk-yta, D., Zimnoch-Guzowska, E., et al. (2016). Mapping of quantitative trait loci for tuber starch and leaf sucrose contents in diploid potato. *Theor. Appl. Genet*. 129, 131–140. doi: 10.1007/s00122-015-2615-9

Stich, B., and Gebhardt, C. (2011). Detection of epistatic interactions in association mapping populations: an example from tetraploid potato. *Heredity* 107:537. doi: 10.1038/hdy.2011.40

Sverrisdóttir, E., Byrne, S., Sundmark, E. H. R., Johnsen, H., Ø., Kirk, H. G., et al. (2017). Genomic prediction of starch content and

chipping quality in tetraploid potato using genotyping-by-sequencing. *Theor. Appl. Genet.* 130, 2091–2108. doi: 10.1007/s00122-017-2944-y

VanRaden, P. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor declared a past collaboration with one of the authors RV.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

APPENDIX A: AMOVA TABLE TO ANALYSE VARIABILITY DUE TO POPULATION STRUCTURE

Appendix A1 | AMOVA analysis showing sources of variation from different configurations of population structure.

Source of variation	df	SS	MS	Est. Var.	Percentage
AMOVA with 6 market classes: | *CP* | *RWT* | *Y* | *P* | *FFP* | *TR* |					
Among subpops	5	0.208	4.16×10^{-2}	1.16×10^{-3}	14.78
Within subpops	184	1.234	6.71×10^{-3}	6.71×10^{-3}	85.22
Total	189	1.442			100
AMOVA with 3 market classes: | *CP*, *RWT*, *Y* | *P* | *FFP*, *TR* |					
Among subpops	2	0.154	7.72×10^{-2}	1.37×10^{-3}	16.61
Within subpops	187	1.288	6.89×10^{-3}	6.89×10^{-3}	83.39
Total	189	1.442			100
AMOVA with 3 market classes: | *CP*, *RWT* | *Y*, *P* | *FFP*, *TR* |					
Among subpops	2	0.152	7.62×10^{-2}	1.19×10^{-3}	14.71
Within subpops	187	1.290	6.90×10^{-3}	6.90×10^{-3}	85.29
Total	189	1.442			100
AMOVA with 4 market classes: | *CP*, *RWT* | *Y* | *P* | *FFP*, *TR* |					
Among subpops	3	0.178	5.93×10^{-2}	1.29×10^{-3}	15.93
Within subpops	186	1.264	6.80×10^{-3}	6.80×10^{-3}	84.07
Total	189	1.442			100
AMOVA with 2 market classes: | *CP* | *RWT*, *Y*, *P*, *FFP*, *TR* |					
Among subpops	1	0.075	7.51×10^{-2}	8.05×10^{-4}	9.97
Within subpops	188	1.367	7.27×10^{-3}	7.27×10^{-3}	90.03
Total	189	1.442			100

df, Degrees of freedom; SS, Sum of Squared deviations; MS, Mean Sum of Squared Deviations; Est. Var, Estimated Variance components; CP, Chip Processing; RWT, Round White Table; Y, Yellow (Y); P, Pigmented; FFP, French Fry Processing; TR, Table Russet. Classes grouped together between vertical lines (|).

frontiers in Plant Science

ORIGINAL RESEARCH
published: 30 November 2021
doi: 10.3389/fpls.2021.774478

Genetic Dissection of Hybrid Performance and Heterosis for Yield-Related Traits in Maize

Dongdong Li[1†], Zhiqiang Zhou[2†], Xiaohuan Lu[1,2†], Yong Jiang[3], Guoliang Li[1], Junhui Li[1], Haoying Wang[1], Shaojiang Chen[1], Xinhai Li[2], Tobias Würschum[4], Jochen C. Reif[3], Shizhong Xu[5], Mingshun Li[2]* and Wenxin Liu[1]**

[1] Key Laboratory of Crop Heterosis and Utilization, The Ministry of Education/Key Laboratory of Crop Genetic Improvement, Beijing Municipality/National Maize Improvement Center/College of Agronomy and Biotechnology, China Agricultural University, Beijing, China, [2] Institute of Crop Science, Chinese Academy of Agricultural Sciences, Beijing, China, [3] Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Stadt Seeland, Germany, [4] Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Stuttgart, Germany, [5] Department of Botany and Plant Sciences, University of California, Riverside, Riverside, CA, United States

OPEN ACCESS

Edited by:
Suchismita Mondal, International Maize and Wheat Improvement Center, Mexico

Reviewed by:
Xuecai Zhang, International Maize and Wheat Improvement Center, Mexico
Jihua Tang, Henan Agricultural University, China

***Correspondence:**
Wenxin Liu
wenxinliu@cau.edu.cn
Mingshun Li
lims2013@126.com
Shizhong Xu
shizhong.xu@ucr.edu

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *12 September 2021*
Accepted: *01 November 2021*
Published: *30 November 2021*

Citation:
Li D, Zhou Z, Lu X, Jiang Y, Li G, Li J, Wang H, Chen S, Li X, Würschum T, Reif JC, Xu S, Li M and Liu W (2021) Genetic Dissection of Hybrid Performance and Heterosis for Yield-Related Traits in Maize. Front. Plant Sci. 12:774478. doi: 10.3389/fpls.2021.774478

Heterosis contributes a big proportion to hybrid performance in maize, especially for grain yield. It is attractive to explore the underlying genetic architecture of hybrid performance and heterosis. Considering its complexity, different from former mapping method, we developed a series of linear mixed models incorporating multiple polygenic covariance structures to quantify the contribution of each genetic component (additive, dominance, additive-by-additive, additive-by-dominance, and dominance-by-dominance) to hybrid performance and midparent heterosis variation and to identify significant additive and non-additive (dominance and epistatic) quantitative trait loci (QTL). Here, we developed a North Carolina II population by crossing 339 recombinant inbred lines with two elite lines (Chang7-2 and Mo17), resulting in two populations of hybrids signed as Chang7-2 × recombinant inbred lines and Mo17 × recombinant inbred lines, respectively. The results of a path analysis showed that kernel number per row and hundred grain weight contributed the most to the variation of grain yield. The heritability of midparent heterosis for 10 investigated traits ranged from 0.27 to 0.81. For the 10 traits, 21 main (additive and dominance) QTL for hybrid performance and 17 dominance QTL for midparent heterosis were identified in the pooled hybrid populations with two overlapping QTL. Several of the identified QTL showed pleiotropic effects. Significant epistatic QTL were also identified and were shown to play an important role in ear height variation. Genomic selection was used to assess the influence of QTL on prediction accuracy and to explore the strategy of heterosis utilization in maize breeding. Results showed that treating significant single nucleotide polymorphisms as fixed effects in the linear mixed model could improve the prediction accuracy under prediction schemes 2 and 3. In conclusion, the different analyses all substantiated the different genetic architecture of hybrid performance and midparent heterosis in maize. Dominance contributes the highest proportion to heterosis, especially for grain yield, however, epistasis contributes the highest proportion to hybrid performance of grain yield.

Keywords: maize, hybrid performance, midparent heterosis, epistatic effect, pleiotropic loci, genomic selection

INTRODUCTION

Heterosis is the phenomenon that a hybrid outperforms its two parents (Birchler et al., 2006; Lippman and Zamir, 2007). Maize is the most successful example for the utilization of heterosis in crops to improve agricultural production, as single-cross varieties of maize have substantially contributed to the improvement of maize production in the past decades (Hochholdinger and Baldauf, 2018). There are three hypotheses to explain the genetic basis of heterosis: dominance (Bruce, 1910; Jones, 1917), overdominance (East, 1936) and epistasis (Powers, 1944). Many studies were performed to test these hypotheses, but the results often varied, depending on the populations and the traits studied, suggesting that heterosis is a complex genetic phenomenon. One commonly used design to study heterosis is the North Carolina Design III (NCIII) or Triple Testcross Design which allows to estimate the contribution of additive, dominance, and epistasis effects to heterosis (Melchinger et al., 2007b; Garcia et al., 2008). In a maize study, a total of 264 F_3 genotypes were generated by intercrossing B73 and Mo17, and the F_3 genotypes were then backcrossed to the two parents. The results showed that nearly all heterozygous individuals performed better than the homozygous individuals, supporting the overdominance (or pseudo-overdominance) hypothesis (Stuber et al., 1992). Conversely, the analysis of hybrid maize data from another NCIII design showed that dominance loci contributed the most to heterosis in maize, while the additive-by-additive effects contributed the most to the heterosis of rice (Garcia et al., 2008).

An alternative design is the North Carolina Design II (NCII) or factorial design, where a set of males is crossed with a set of females in a balanced or unbalanced way. In a partial NCII of maize, eight main effect (additive and dominance) QTL and 37 epistatic QTL pairs were identified (Bu et al., 2015). In addition to the NC mating designs, advanced maize populations were also developed and used for analysis of heterosis. For example, Wei et al. (2016) detected 36 heterotic loci from a series of single-segment substitution lines. Using near-isogenic lines for QTL detection, many additive QTL and additive-by-additive QTL pairs were detected (Melchinger et al., 2007a; Reif et al., 2009). An immortalized F_2 population (IMF_2) was also a promising mating design for dissecting the genetic basis of heterosis and epistasis QTL (Hua et al., 2003; Xu, 2013; Guo et al., 2014; Yi et al., 2019).

Linear mixed models (LMM) are a powerful tool for the genetic dissection of complex traits and are widely used in plant and animal breeding (Yu et al., 2006; Xu et al., 2014; Cui et al., 2020). In a hybrid population of rice, a LMM incorporating multiple polygenic covariance structures to control the genetic background was developed (Xu, 2013). In wheat, a quantitative genetics approach was proposed to dissect the genetic basis of grain-yield heterosis, allowing QTL mapping of dominance, epistasis and heterotic loci for midparent heterosis (MPH) (Jiang et al., 2017). In addition to QTL mapping, genomic selection (GS) has become a new tool for plant breeding and the genetic dissection of complex traits (Meuwissen et al., 2001) and has been applied to hybrid wheat (Zhao et al., 2013, 2015b; Jiang et al., 2017), hybrid rice (Cui et al., 2020) and hybrid maize (Albrecht et al., 2014; Technow et al., 2014).

The general combining ability (GCA) is a measure for the average performance of a line in different hybrid combinations, while the specific combining ability (SCA) describes the deviation of a hybrid from the performance expected based on the GCA of its two parental lines. The additive and additive-by-additive variances contribute to the variation of GCA, while the non-additive polygenic variances contribute to the variation of SCA (Reif et al., 2007). A two-step approach has been widely used to study the genetics underlying hybrid performance, where the first step consists of estimating the GCA, SCA and the MPH (Guo et al., 2014; Zhou et al., 2018; Yi et al., 2019) and the second step represents the QTL mapping step with the GCA, SCA and MPH treated as the traits of interest. In a previous genome-wide association study (GWAS) with an NCII population, different coding schemes for the genotypes were applied, namely the additive, dominance and recessive coding (Hyun et al., 2008; Liu et al., 2021). However, the additive model was usually not sufficient to explain hybrid performance and MPH. Thus, more elaborate models incorporating non-additive effects should be used to study heterosis.

In this study, we developed a NCII population of maize by crossing a set of 339 recombinant inbred lines (RILs) with two elite inbred lines, resulting in two populations of hybrids. A total of 10 traits were recorded in four to five environments and high-density genotypic data were obtained by genotyping-by-sequencing of the RILs and resequencing of the parents. The aims of this study were to (1) evaluate the heritability of MPH and the relative contribution of various traits to grain yield, (2) perform QTL mapping for main (additive and dominance) and non-additive effect loci for hybrid performance and MPH, (3) identify QTL hotspots for yield-related traits, (4) explore the mechanisms of heterosis and hybrid performance, and (5) assess the accuracy of genomic prediction in various breeding schemes.

MATERIALS AND METHODS

Plant Materials

A RIL population consisting of 365 F_{11} lines was developed by crossing inbred lines Qi319 as the male parent and Ye478 as the female parent originating from two different heterotic groups of maize (Zhou et al., 2016). Two hybrid populations were developed by crossing the RILs to two female testers, Chang7-2 and Mo17, and the two populations Chang7-2 × RIL and Mo17 × RIL were named TC and TM, respectively (Zhou et al., 2018). Different numbers of offspring were obtained from the two hybrid populations. A total of 339 common lines from the RIL, TC, and TM populations were retained for further analysis.

Experimental Design and Phenotypic Evaluation

The RIL, TC, and TM populations, their parents and the hybrids (Chang7-2 × Qi319, Mo17 × Qi319, Chang7-2 × Ye478, and Mo17 × Ye478) were field-evaluated in two different locations, Xinxiang (35.19°N and 113.53°E) and Shijiazhuang (37.27°N and

113.30°E), in two consecutive years, 2015 and 2016, resulting in 2 × 2 = 4 different environments. Traits recorded include plant height (PH, cm), ear height (EH, cm), row number per ear (RNPE, count), kernel number per row (KNPR, count), kernel thickness (KT, mm), kernel width (KW, mm), kernel length (KL, mm), volume weight (VW, g/L), hundred grain weight (HGW, g) and grain yield per plant (GY, g). Furthermore, in 2017, the VW trait was evaluated in the RIL population, traits HGW and GY were evaluated in all the three populations (RIL, TC, and TM) in one of the two locations, Xinxiang. Detailed descriptions of the traits evaluated can be found in a previous study (Lu et al., 2020).

We used a randomized incomplete block design with two replicates in each environment. To avoid competition, the RIL and the hybrid populations were planted separately. Each genotype was planted in two rows with a row interval of 0.6 m, a row length of 4 m and a plant interval of 0.25 m.

Phenotypic Data Analysis

The combination between year and location was considered as an environment (a total of 4 or 5 environments). The studentized residual razor method (Bernal-Vasquez et al., 2016) was used to remove outliers with a threshold of 2.8. The best linear unbiased estimations (BLUE) of the fixed effects and the variance components of the random effects were estimated using the following model:

$$y_{ijk} = \mu + G_i + E_j + G * E_{ij} + R_k(E_j) + \varepsilon_{ijk},$$

where y_{ijk} was the phenotypic value of the kth replicate of genotype i from the jth environment, μ was the overall mean, G_i (i = 1, 2,..., 339) was the effect of the ith genotype, E_j (j = 1, 2,..., 5) was the effect of the jth environment, $G * E_{ij}$ was the genotype-by-environment interaction effect, $R_k(E_j)$ (k = 1, 2) was the effect of the kth replicate nested in the jth environment, ε_{ijk} was the residual. For estimation of variance components, all random effects were assumed to be normally distributed with mean 0 and variances denoted by σ^2_G , $\sigma^2_{G\times E}$ and σ^2_ε for G_i, $G * E_{ij}$ and ε_{ijk}, respectively. The broad-sense heritability of a trait was defined as (Falconer and Mackay, 1996),

$$H^2 = \frac{\sigma^2_G}{\sigma^2_G + \frac{\sigma^2_{G\times E}}{N_E} + \frac{\sigma^2_\varepsilon}{N_E \times N_R}},$$

where N_E = 4 or 5 was the number of environments and N_R = 2 was the number of replicates within each environment.

Genetic analysis of MPH was conducted in two steps (Jiang et al., 2017). The first step was represented by BLUE of the trait value for each parent and each hybrid. The BLUE of the trait value obtained from the two replicates in one environment was calculated with the following formula:

$$y_{ik} = \mu + G_i + R_k + \varepsilon_{ik},$$

where y_{ik} was the trait value for the kth replicate of genotype i, μ was the mean of the trait under the current environment, G_i was the genetic value of the ith genotype and R_k was the effect of the kth replicate assumed to follow a $N(0, \sigma^2_R)$ distribution, ε_{ik} was assumed to follow a $N(0, \sigma^2_{ik})$ distribution.

The MPH was defined as (Melchinger et al., 2007b):

$$MPH = H - (P_1 + P_2)/2.$$

Where H was the BLUE value of hybrids, P_1 was the BLUE value of Chang7-2 or Mo17 (corresponding to female parent of hybrid), P_2 was the BLUE value of RIL (corresponding to the male parent of hybrid).

The second step in the MPH analysis required the following mixed model:

$$MPH_{ij} = \mu + G_i + E_j + \varepsilon_{ij},$$

where MPH_{ij} was the MPH value calculated in the first step for hybrid (genotype) i in environment j, G_i (i = 1, 2, . . ., 339) was the genetic effect of MPH for the ith hybrid, E_j was effect of the jth environment and ε_{ij} was the residual. Noted that G_i was treated as a fixed effect in the BLUE calculation or a random effect following a $N(0, \sigma^2_G)$ distribution in variance estimation, E_j was treated as a random effect following a $N(0, \sigma^2_E)$ distribution and ε_{ij} was assumed to be $N(0, \sigma^2_\varepsilon)$ distributed. The variance components of the above linear mixed model were implemented using the ASReml 3.0 package in R (Gilmour et al., 2009).

In addition, the hybrid performance was decomposed into GCA, SCA and interaction with the environment using a two-step method. In the first step, the BLUEs in RIL, TC and TM populations were calculated within each environment following the same formula above. In the second step, the following formula was applied to the hybrid performance (Zhao et al., 2015a):

$$y = \mu + E + GCA_{RIL} + GCA_{Tester} + SCA + GCA_{RIL} * E + GCA_{Tester} * E + SCA * E + \varepsilon.$$

Where y was the hybrid performance, μ was the mean, E was the environment effect, GCA_{RIL} was the GCA of RILs, GCA_{Tester} was the GCA of testers, the rest was the interaction between GCA, SCA, and environment, ε was the error. All effects were treated as random following normal distributions. The variances were estimated in ASReml 3.0 package in R (Gilmour et al., 2009).

Path analysis can be used to determine the relative contribution of independent variables to a response variable. Path analysis was implemented in the R package sem by taking GY as the response variable and the other traits as independent variables. Path coefficient p_i of variable X_i was obtained by $p_i = b_i\sqrt{SS_{X_i}/SS_Y}$, where b_i was the partial correlation, SS_{X_i} and SS_Y were sum of square for X_i and the response variable Y, respectively. Path diagrams were drawn with the semPlot package in R, where values above 0.14 (p = 0.01, n = 339) were displayed.

Genomic Data Analyses

The genotyping procedures for the RILs, the two parents of the RILs, and the two testers were described in a previous study (Zhou et al., 2016). In brief, for the four parents, the paired-end sequencing libraries were created with a fragment length of ∼500 bp and were sequenced on an Illumina HiSeq 2000 sequencer. The resequencing depth was ∼30×. For the RILs, a genotyping-by-sequencing (GBS) strategy was applied. A total of 137,699,000 reads were generated. On average, there were

357,376 reads per individual, which was approximately a 0.07-fold coverage of the maize genome. The cleaned reads were obtained after quality control.

The filtered high-quality reads of the four parents and the RILs were mapped to the reference genome (B73_RefGen_v4) with BWA (Li and Durbin, 2009). SAMtools (Li, 2011) were used to call SNPs with quantity over 20 and a total of 41,791,163 SNPs were finally produced. Details regarding the parameters for the SNP calling process can be found in a previous report (Zhou et al., 2016). After filtering of all SNPs for minor allele frequency < 0.05, missing rate > 0.1 and unknown physical positions, a total of 36,095 SNPs remained in the data set for analysis. Missing genotypes of SNPs were imputed using the BEAGLE software package (version v5) with the default parameters (Browning and Browning, 2016).

The low-coverage high-throughput sequence technologies like GBS generate sequences that are often error-prone, which might lead to errors for detection of genetic variants (Ma et al., 2019). Therefore, the hmm.vitFUN.rils function in the R package MPR.genotyping was used to correct the genotyping errors using a Hidden Markov model with errorRate = 0.05 (Xie et al., 2010). The SNPs with high error probabilities were either corrected or set to missing values.

The bin function in the ICIMapping package was used to bin redundant markers with missing rate > 0.2 and a distortion p-value < 0.001, while missing values and anchor information were considered at the same time (Meng et al., 2015). After the above imputation and correction, there were still a little proportion of missing values left, then the argmax method in qtl/R was used to perform the final imputation additive-by-dominance and (Broman et al., 2003). Finally, a total of 4,141 bins were discovered across the entire maize genome. The genetic map was constructed using the map function in the IciMapping package with the default parameter values (Meng et al., 2015).

Mapping Quantitative Trait Loci in Recombinant Inbred Line, TC, TM, and Pooled TC-TM Populations

To determine the contribution of each genetic component to hybrid performance and MPH variation and identify significant non-additive QTL, firstly, we combined the TC and TM populations to form a pooled population called TC-TM. For 341 lines (337RILs, 2 parents; 2 testers) lines, if the genotypes were the same as Ye478, it was coded as "1"; if the genotypes were the same as Qi319, it was coded as "−1", then the genotypes of the hybrids were inferred from their parents (the RILs and the testers). The additive and dominance coding matrices, Z and W, for individual j at marker k were coded as $Z_{jk} = \{1\ 0\ -1\}$ for the additive effect and $W_{jk} = \{0\ 1\ 0\}$ for the dominance effect.

The linear mixed model for variance component analysis was (Xu, 2013; Jiang et al., 2017):

$$y = X\beta + \xi_a + \xi_d + \xi_{aa} + \xi_{ad} + \xi_{dd} + \varepsilon, \quad (1)$$

where y was an $n \times 1$ vector of phenotypic values of the hybrids and $X\beta$ captured the fixed effects of the model that were not relevant to genetic effects. The design matrix for the fixed effects was $X = [X_0, X_1]$, where X_0 was an $n \times 1$ vector of unity (a vector with all elements being 1) and X_1 was an $n \times 1$ vector indicating one of the two populations, $X_{j1} = 0$ for TC and $X_{j1} = 1$ for TM. The last term of model (1) was a vector of residuals. The remaining terms in model (1) were various polygenic effects (each polygenic effect was an $n \times 1$ vector) and were defined below. $\xi_a = \sum_{k=1}^{m} Z_k a_k$ was the polygenic additive effect; $\xi_d = \sum_{k=1}^{m} W_k d_k$ was the polygenic dominance effect; $\xi_{aa} = \sum_{k=1}^{m-1} \sum_{k'=k+1}^{m} (Z_k \# Z_{k'})(aa)_{kk'}$ was the polygenic additive-by-additive effect; $\xi_{ad} = \sum_{k,k'=1,k'\neq k}^{m} (Z_k \# W_{k'})(ad)_{kk'}$ was the polygenic additive-by-dominance effect; $\xi_{dd} = \sum_{k=1}^{m-1} \sum_{k'=k+1}^{m} (W_k \# W_{k'})(dd)_{kk'}$ was the polygenic dominance-by-dominance effect. The operator $\#$ represented element-wise product of matrices. In the formulas above, a_k and d_k were the additive and dominance effect for marker k, $(aa)_{kk'}$, $(ad)_{kk'}$, and $(dd)_{kk'}$ were the additive-by-additive, additive-by-dominance and dominance-by-dominance effect between markers k and k', respectively. The distributions for the polygenic and residual effects were $\xi_a \sim N(0, K_a\sigma_a^2)$, $\xi_d \sim N(0, K_d\sigma_d^2)$, $\xi_{aa} \sim N(0, K_{aa}\sigma_{aa}^2)$, $\xi_{ad} \sim N(0, K_{ad}\sigma_{ad}^2)$, $\xi_{dd} \sim N(0, K_{dd}\sigma_{dd}^2)$, and $\varepsilon \sim N(0, I\sigma^2)$, where K_a, K_d, K_{aa}, K_{ad}, and K_{dd} were the corresponding kinship matrices calculated using the method given by Xu (2013). The six variance components (five genetic variance components and the residual variance) were estimated using the BGLR package in R (Pérez and De Los Campos, 2014) with the number of iterations set at 15,000 and the number of burn-in set at 5,000.

The variance-covariance matrix of y was

$$var(y) = K_a\sigma_a^2 + K_d\sigma_d^2 + K_{aa}\sigma_{aa}^2 + K_{ad}\sigma_{ad}^2 + K_{dd}\sigma_{dd}^2 + I\sigma^2.$$

Let $\lambda_x = \sigma_x^2/\sigma^2$, where σ_x^2 was one of the five genetic variance components and σ^2 was the residual variance. The above variance could be rewritten as

$$var(y) = (K_a\lambda_a + K_d\lambda_d + K_{aa}\lambda_{aa} + K_{ad}\lambda_{ad} + K_{dd}\lambda_{dd} + I)\sigma^2.$$

Define

$$K = K_a\lambda_a + K_d\lambda_d + K_{aa}\lambda_{aa} + K_{ad}\lambda_{ad} + K_{dd}\lambda_{dd},$$

so that

$$var(y) = (K + I)\sigma^2.$$

Let

$$\xi = \xi_a + \xi_d + \xi_{aa} + \xi_{ad} + \xi_{dd}.$$

Model (1) could be rewritten as

$$y = X\beta + \xi + \varepsilon, \quad (2)$$

which was the null model for the GWAS of main effect and epistatic effect detection. On this null model, we added a specific marker or marker pair to the model to test the putative effect.

To test the additive effect of marker k, we added $Z_k a_k$ to the null mode so that the linear mixed model became:

$$y = X\beta + Z_k a_k + \xi + \varepsilon. \quad (3)$$

Let $e = \xi + \varepsilon$, so that the model was rewritten as:

$$y = X\beta + Z_k a_k + e. \quad (4)$$

The expectation of model (4) was $E(y) = X\beta + Z_k a_k$ and the variance was:

$$var(y) = var(e) = var(\xi + \varepsilon) = (K + I)\sigma^2.$$

Let us perform eigenvalue decomposition for matrix K, $K = UDU^T$, where U was the eigenvector matrix and D was a diagonal matrix holding the eigenvalues. So,

$$var(e) = U(D + I)U^T\sigma^2.$$

Let $Q^T = \sqrt{(D + I)^{-1}}U^T$ and pre-multiply equation (4) by Q^T leading to

$$Q^T y = Q^T(X\beta + Z_k a_k + e) = Q^T X\beta + Q^T Z_k a_k + Q^T e. \quad (5)$$

Let $y^* = Q^T y$, $X^* = Q^T X$, $Z_k^* = Q^T Z_k$ and $e^* = Q^T e$. The above linear mixed model was

$$y^* = X^*\beta + Z_k^* a_k + e^*. \quad (6)$$

The variance of the transformed residuals was

$$\begin{aligned} var(e^*) &= var(Q^T e) \\ &= Q^T U(D + I)U^T Q\sigma^2 \\ &= \sqrt{(D + I)^{-1}}U^T U(D + I)U^T U\sqrt{(D + I)^{-1}}\sigma^2 \\ &= \sqrt{(D + I)^{-1}}(D + I)\sqrt{(D + I)^{-1}}\sigma^2 \\ &= \sqrt{(D + I)^{-1}}\sqrt{(D + I)}\sqrt{(D + I)}\sqrt{(D + I)^{-1}}\sigma^2 \\ &= I\sigma^2. \end{aligned}$$

The expectation and variance of y^* were $E(y^*) = X^*\beta + Z_k^* a_k$ and $var(y^*) = I\sigma^2$. Therefore, model (6) became a simple linear model with a homogeneous residual variance. The conventional least squares method could be used to estimate the parameters and test for the marker effect. Since the model of the transformed phenotypic values was very simple, the "lm" function in R was applied to estimate the marker effect and test the significance of the marker.

Considering the dominance and epistatic effects, we adopted a more general likelihood ratio test (LRT) for a particular effect. The likelihood ratio test for the additive effect of marker k was

$$LRT = -2\left[L_0(\hat{\beta}) - L_1(\hat{\beta}, \hat{a}_k)\right],$$

where $L_0(\hat{\beta})$ was the likelihood value evaluated from the null model given in equation (7) below,

$$y^* = X^*\beta + e^*, \quad (7)$$

and $L_1(\hat{\beta}, \hat{\alpha}_k)$ was the likelihood value evaluated from the full model given in equation (6). The LRT statistic was eventually converted into the log of odds (LOD) score using $LOD = LRT/4.61$. If the intervals of different QTL were overlapped or the genetic distance of peak SNP of two QTL was within 0.65 cM (the average density in the whole genome), we called such QTL as a pleiotropic QTL (a QTL affecting more than one trait).

Dominance effect of marker k was detected using the same model as the additive effect except that Z_k was replaced by W_k. In the following, we called the significant additive and dominance QTL as the main effect QTL.

The additive-by-additive effect was detected by the following likelihood ratio test,

$$LRT = -2\left[L_0(\hat{\beta}, \hat{a}_k, \hat{a}_{k'}) - L_1(\hat{\beta}, \hat{a}_k, \hat{a}_{k'}, (aa)_{kk'})\right],$$

where the null model was

$$y^* = X^*\beta + Z_k^* a_k + Z_{k'}^* a_{k'} + e^*, \quad (8)$$

and the full model was

$$y^* = X^*\beta + Z_k^* a_k + Z_{k'}^* a_{k'} + (Z_k^* \# Z_{k'}^*)(aa)_{kk'} + e^*, \quad (9)$$

Similarly, the additive-by-dominance effect was detected using

$$LRT = -2\left[L_0(\hat{\beta}, \hat{a}_k, \hat{d}_{k'}) - L_1(\hat{\beta}, \hat{a}_k, \hat{d}_{k'}, (ad)_{kk'})\right].$$

The null model and the full model were

$$y^* = X^*\beta + Z_k^* a_k + W_{k'}^* d_{k'} + e^*, \quad (10)$$

and

$$y^* = X^*\beta + Z_k^* a_k + W_{k'}^* d_{k'} + (Z_k^* \# W_{k'}^*)(ad)_{kk'} + e^*, \quad (11)$$

respectively. Similarly, the dominance-by-additive effect was detected using

$$LRT = -2\left[L_0(\hat{\beta}, \hat{a}_k, \hat{d}_{k'}) - L_1(\hat{\beta}, \hat{a}_k, \hat{d}_{k'}, (da)_{kk'})\right].$$

The null model and the full model were

$$y^* = X^*\beta + Z_k^* a_k + W_{k'}^* d_{k'} + e^*, \quad (12)$$

and

$$y^* = X^*\beta + Z_k^* a_k + W_{k'}^* d_{k'} + (W_k^* \# Z_{k'}^*)(da)_{kk'} + e^*, \quad (13)$$

respectively. Finally, the dominance-by-dominance effect was tested using

$$LRT = -2\left[L_0(\hat{\beta}, \hat{d}_k, \hat{d}_{k'}) - L_1(\hat{\beta}, \hat{d}_k, \hat{d}_{k'}, (dd)_{kk'})\right].$$

The corresponding null model and full model were

$$y^* = X^*\beta + W_k^* d_k + W_{k'}^* d_{k'} + e^*, \quad (14)$$

and

$$y^* = X^*\beta + W_k^* d_k + W_{k'}^* d_{k'} + (W_k^* \# W_{k'}^*)(dd)_{kk'} + e^*, \quad (15)$$

respectively. LOD scores were converted the same way as we did for the additive effect.

An empirical threshold of 2.5 for the LOD score was used to determine significance of an additive or a dominance effect. A LOD threshold of 5.0 was used to determine the significance of an epistatic effect (Churchill and Doerge, 1994; Xu, 2013). A confidence interval in the genome was determined for each

detected QTL with the following steps: (1) all significant SNPs passing the threshold were selected; (2) the most significant SNPs were kept within a 10 cM interval; (3) the QTL interval was formed using a 1.5-LOD drop-off method (Broman, 2001). The names of QTL referred to McCouch's method (McCouch et al., 1997), and a dash (–) was added to designate different datasets.

The estimated additive and dominance effects for each QTL were extracted from the estimated regression coefficients (a_k and d_k) from the models presented above. The proportion of the phenotypic variance explained (PVE) contributed by each QTL was calculated using (Utz et al., 2000; Garin et al., 2017),

$$PVE = 1 - \frac{RSS_{Full}}{RSS_{Null}},$$

where RSS_{Full} was the residual sum of squares of the full model and RSS_{Null} was the residual sum of squares of the null model.

We also performed QTL mapping in the RIL, TC and TM population separately. The model was the same as described above except that only the additive and additive-by-additive polygenic effects were used to control genetic background. QTL mapping for MPH was conducted using a similar linear mixed model to the original traits. Details of the MPH analysis can be found in a previous study (Jiang et al., 2017).

Genomic Selection

The genetic effects of single-cross hybrids can be dissected into additive, dominance and epistatic polygenic effects as mentioned before. Here, we only considered the first two components in the genomic prediction model. The linear mixed model was (Su et al., 2012; Xu et al., 2014).

$$y = X\beta + \xi_a + \xi_d + \varepsilon,$$

where y was the phenotype vector, $X\beta$ represented the fixed effect, ξ_a was the additive polygenic effect with an assumed distribution of $\xi_a \sim N(0, K_a\sigma_a^2)$, ξ_d was the dominance polygenic effect with a distribution of $\xi_d \sim N(0, K_d\sigma_d^2)$, K_a was the additive kinship matrix and K_d was the dominance kinship matrix.

Three genomic prediction schemes were proposed to mimic the scenarios in practical genomic hybrid breeding. Scheme (1), abbreviated as CV1: to predict the trait values for the TM population from the phenotypes and genotypes of the TC population or vice versa. Scheme (2), abbreviated as CV2: to select the hybrids sharing the same RILs in TC and TM population as the training set to predict the rest of the population. Scheme (3), abbreviated as CV3: to select the hybrids having the different RILs in TC and TM population as the training set to predict the rest of the population. Scheme (1) and (2) belong to the so-called T1 case, and scheme (3) is in the category of T2 (Technow et al., 2014; Zhao et al., 2015a). The three scenarios are illustrated in **Supplementary Figure 1**.

The across population prediction in scheme (1) was conducted using a model that contained only the additive polygenic effect. For schemes (2) and (3), the prediction models contained both the additive and the dominant polygenic effects. The predictions were implemented with the BGLR software package in R (Pérez and De Los Campos, 2014). The prediction accuracy was assessed with a two-fold cross-validation scheme. In each run, 1/2 of the lines were removed from the training set and then the correlation between the predicted values and the observed values of the removed lines was calculated. The two-fold cross-validation was repeated 200 times. In addition, significant SNPs were treated as fixed effects in the prediction model, which has been termed wGS (Bernardo, 2014; Würschum et al., 2018). For example, when the TC population was used to predict the TM population, the QTL detected in the TC population were treated as fixed effects in the linear mixed model used to predict the TM population. For schemes (2) and (3), QTL detected from the pooled population of TC and TM were treated as fixed effects included in the models to predict the rest of the population. For comparison, the additive model with kinship matrix inferred from RILs population was used to yield the prediction accuracy of 10 traits using a two-fold cross-validation scheme in TC and TM population, respectively. This process was repeated 200 times. Data visualization was done with the ggplot2 and ggpubr packages in R (Wickham, 2016).

RESULTS

Phenotypic Variation and Heritability in the Recombinant Inbred Line, TC, and TM Populations

The RIL population showed a larger variation for the 10 investigated traits than the TC and TM populations (**Table 1** and **Figure 1**). The genetic variance components were significant ($p < 0.01$) for all traits in the three populations. Except for the trait VW in the TM population, the variance of the genotype-by-environment interaction was also significant ($p < 0.01$) for all traits. The estimated broad-sense heritability ranged from 0.68 for VW to 0.95 for PH in the RIL population, from 0.57 for VW to 0.91 for PH and KT in the TC population, and from 0.60 for GY to 0.89 for PH in the TM population. In general, PH had the highest heritability and VW or GY had the lowest heritability. The obtained moderate to high heritability implied that the experimental designs and phenotyping procedures were appropriate and accurate.

The MPH for traits KT and VW was negative on average in both the TC and TM populations, which means that the hybrids often had phenotypic values lower than the mean of the two parents (**Figure 1**). In both the TC and TM populations, GY had the highest heterosis, followed by PH and EH (**Supplementary Table 1**), and the genetic variances of MPH were statistically significant for all traits. The heritability of MPH in the TC population ranged from 0.36 for VW to 0.81 for PH, while the heritability of MPH ranged from 0.27 for VW to 0.78 for PH in the TM population. The moderate to high heritability of MPH lay the foundation to dissect the genetic architecture of heterosis.

TABLE 1 | Summary statistics for 10 traits in the recombinant inbred line population developed by Ye478 × Qi319 (RIL), Chang7-2 × RIL (TC), and Mo17 × RIL (TM) populations.

Population	Traits	Min	Max	Mean	*SD*	CV (%)	σ^2_G	σ^2_{G*E}	σ^2_ε	N_E	H^2
RIL	PH	138.88	223.17	179.28	15.29	8.53	217.97**	26.64**	45.08	4	0.95
	EH	45.48	94.49	67.87	8.91	13.13	73.83**	9.39**	17.18	4	0.94
	RNPE	8.91	14.41	11.79	0.90	7.63	0.71**	0.13**	0.34	4	0.90
	KNPR	12.57	32.21	23.31	3.25	13.95	8.69**	3.72**	4.03	4	0.86
	KT	40.49	69.81	52.59	3.98	7.57	13.40**	2.84**	9.54	4	0.88
	KW	76.29	104.84	89.60	4.26	4.75	15.28**	3.07**	13.34	4	0.86
	KL	86.65	115.59	100.17	5.14	5.13	21.96**	5.76**	17.26	4	0.86
	VW	509.00	732.26	640.84	32.39	5.05	606.28**	399.41**	2026.25	5	0.68
	HGW	17.68	33.72	25.27	2.90	11.46	7.30**	2.41**	4.70	5	0.88
	GY	20.47	84.90	53.83	11.76	21.86	113.05**	59.47**	65.54	5	0.86
TC	PH	204.20	271.68	247.23	9.34	3.78	77.97**	8.67**	42.30	4	0.91
	EH	98.15	134.09	113.17	6.49	5.73	35.49**	5.53**	28.35	4	0.88
	RNPE	12.47	16.74	14.36	0.73	5.05	0.44**	0.07**	0.46	4	0.86
	KNPR	31.36	43.04	37.74	1.59	4.22	2.07**	1.00**	3.42	4	0.75
	KT	36.29	46.66	40.44	1.72	4.25	3.18**	0.15**	2.29	4	0.91
	KW	84.78	104.00	95.24	3.02	3.17	7.15**	2.16**	8.99	4	0.81
	KL	111.28	132.37	124.22	3.63	2.92	9.22**	4.60**	17.58	4	0.73
	VW	489.34	633.25	555.35	21.81	3.93	252.96**	143.69**	1238.20	4	0.57
	HGW	22.09	34.48	26.77	1.87	6.98	2.47**	1.33**	5.11	5	0.76
	GY	99.01	158.78	131.15	9.68	7.38	53.37**	35.86**	235.67	5	0.63
TM	PH	212.26	280.26	259.11	8.61	3.32	65.16**	15.32**	31.10	4	0.89
	EH	84.30	117.25	100.55	6.27	6.24	33.39**	8.11**	20.62	4	0.88
	RNPE	11.44	14.06	12.66	0.48	3.77	0.18**	0.03**	0.23	4	0.83
	KNPR	29.00	45.17	38.95	2.22	5.69	3.44**	2.86**	4.20	4	0.74
	KT	39.84	57.03	46.04	2.17	4.72	3.79**	0.69**	4.65	4	0.83
	KW	87.14	105.99	94.79	2.76	2.91	6.02**	0.98**	8.00	4	0.83
	KL	109.41	128.48	118.74	3.43	2.89	8.62**	1.98**	15.39	4	0.78
	VW	491.34	627.42	563.82	20.76	3.68	237.35**	68.33	1102.22	4	0.61
	HGW	24.16	34.25	28.39	1.75	6.16	2.17**	1.31**	4.68	5	0.75
	GY	94.99	152.12	125.65	8.74	6.95	40.76**	45.49**	180.99	5	0.60

*SD, standard deviation; CV, coefficient of variation; σ^2_G, genotypic variance; σ^2_{G*E}, genotype-by-environment interaction variance; σ^2_ε , error variance; N_E, the number of environments; H^2, broad-sense heritability; **, significance at 0.01 level; PH, plant height; EH, ear height; RNPE, row number per ear; KNPR, kernel number per row; KT, kernel thickness; KW, kernel width; KL, kernel length; VW, volume weight; HGW, hundred grain weight; GY, grain yield per plant.*

For PH, EH, RNPE, KT, and KL the GCA variance of testers ($\sigma^2_{GCA_{Tester}}$) had larger values than the GCA variance of RILs ($\sigma^2_{GCA_{RIL}}$), which indicated that the testers played an important role in hybrid performance. The SCA/GCA ratios indicating a relative contribution of additive and non-additive (dominance and epistasis) effects to phenotypic variation ranged from 0.04 for KT and 0.77 for GY (**Supplementary Table 2**). And for GY, the variance of SCA (σ^2_{SCA}) was higher than both $\sigma^2_{GCA_{Tester}}$ and $\sigma^2_{GCA_{RIL}}$, which was consistent with the large MPH variation in phenotype (**Figure 1J**).

Trait Correlation and Path Analysis in the Three Populations

Relatively high correlations between traits were observed in the three populations (**Figure 2A**). The correlation between traits KNPR and GY was $r = 0.69$ ($p < 0.01$) and the correlation between KL and GY was 0.57 ($p < 0.01$) in the RIL population, which were the highest among the correlations between GY and the other traits. In the TC population, the highest correlations of GY occured between GY and HGW ($r = 0.50$, $p < 0.01$) and between GY and KL ($r = 0.45$, $p < 0.01$). In the TM population, the highest correlations were between GY and KNPR ($r = 0.54$, $p < 0.01$) and between GY and KL ($r = 0.41$, $p < 0.01$).

It is difficult to determine which trait contributes the most to the variation of grain yield only through correlation analysis between GY and the other traits. We therefore next performed a path analysis of all traits with GY (**Figures 2B–D**). By taking GY as the response variable and all other traits as independent variables, we estimated the path coefficients for every trait. In the RIL population, the highest path coefficients occurred for KNPR (0.60) and for HGW (0.36). The trait HGW had the highest path coefficient (0.79), followed by KNPR (0.54) in the TC population. In the TM population, the highest path coefficient was 0.79 for KNPR, followed by 0.62 for HGW. In summary, KNPR and HGW contributed most to the variation of grain yield.

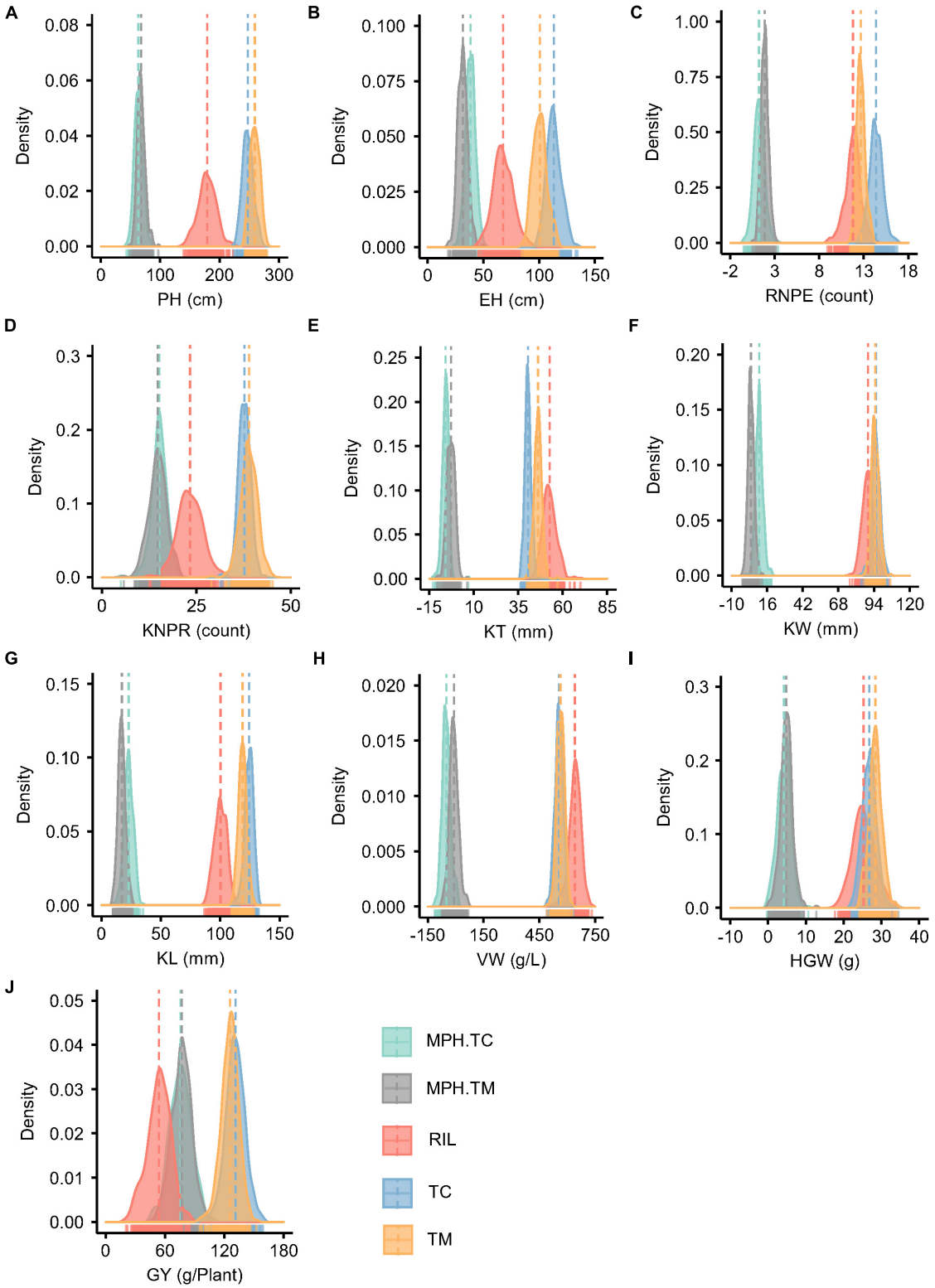

FIGURE 1 | Phenotype and midparent heterosis (MPH) distributions for 10 traits in the recombinant inbred line population developed by Ye478 × Qi319 (RIL), Chang7-2 × RIL (TC), and Mo17 × RIL (TM) populations. **(A)** PH, plant height; **(B)** EH, ear height; **(C)** RNPE, row number per ear; **(D)** KNPR, kernel number per row; **(E)** KT, kernel thickness; **(F)** KW, kernel width; **(G)** KL, kernel length; **(H)** VW, volume weight; **(I)** HGW, hundred grain weight; **(J)** GY, grain yield per plant; MPH.TC, MPH in TC population; MPH.TM, MPH in TM population.

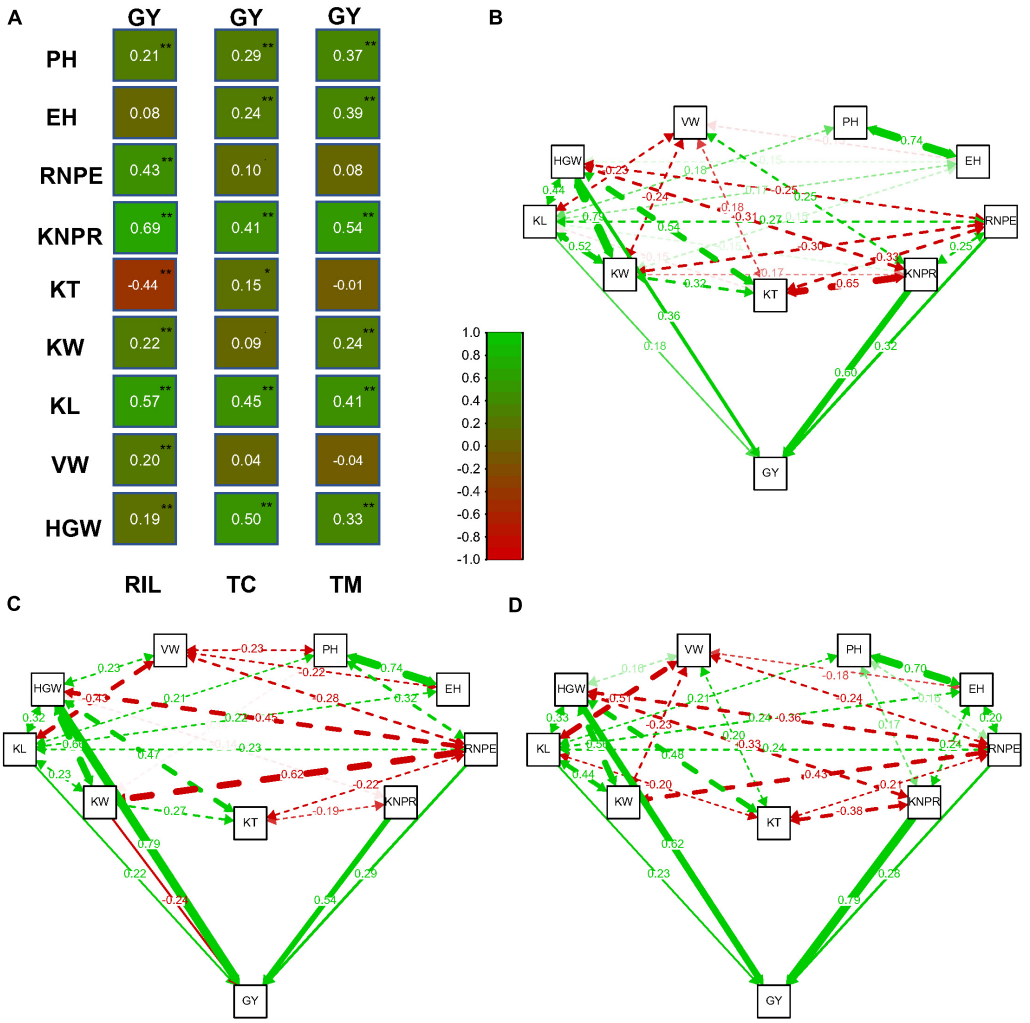

FIGURE 2 | Correlation and path analysis of 10 traits in the recombinant inbred line population developed by Ye478 × Qi319 (RIL), Chang7-2 × RIL (TC), and Mo17 × RIL (TM) populations. **(A)** Correlation coefficients between grain yield per plant (GY) and the other traits. **(B)** Correlation and path coefficients between GY and the other traits in the RIL population, **(C)** in the TC population, and **(D)** in the TM population. The lines toward GY are the path coefficients and the other lines among the traits are correlation coefficients. Only coefficients larger than 0.14 (p = 0.01, n = 339) are displayed. **, significance at 0.01 level; *, significance at 0.05 level; PH, plant height; EH, ear height; RNPE, row number per ear; KNPR, kernel number per row; KT, kernel thickness; KW, kernel width; KL, kernel length; VW, volume weight; HGW, hundred grain weight; GY, grain yield per plant.

Main Effect Quantitative Trait Loci Mapping in the Recombinant Inbred Line, TC, and TM Populations

A high-density genetic map was constructed using 4,141 bins, covering 2669.49 cM of the maize genome (**Supplementary Table 3** and **Supplementary Figure 2**). The average density of the marker map was 0.64 cM/bin in the whole genome, enabling a high resolution for QTL mapping.

To dissect the genetic architecture of the 10 traits, we first examined the additive model with the additive polygenic effect plus the additive-by-additive polygenic effect to control the genomic background (**Supplementary Table 4**). The additive (narrow-sense) heritability in the RIL population ranged from 0.25 for VW to 0.69 for PH. In the TC population, it ranged from 0.31 for VW to 0.70 for RNPE and in the TM population, it ranged from 0.38 for VW to 0.72 for EH. Generally, the proportion of phenotypic variance explained by the additive effects was greater than that explained by the additive-by-additive effects for all 10 traits. We also found that the proportion of variance explained by the additive-by-additive effects for the traits RNPE, KT, KW, KL, and GY was larger in the RIL population than the corresponding proportion in the TC and TM populations (**Supplementary Table 4**), illustrating that further studies are needed to understand the non-additive genetic architecture of these traits.

We also mapped QTL for the 10 traits in the RIL, TC, and TM populations, respectively (**Figure 3A** and **Supplementary Table 5**). In the RIL population, a total of 16 QTL were identified on eight chromosomes and five superior alleles were

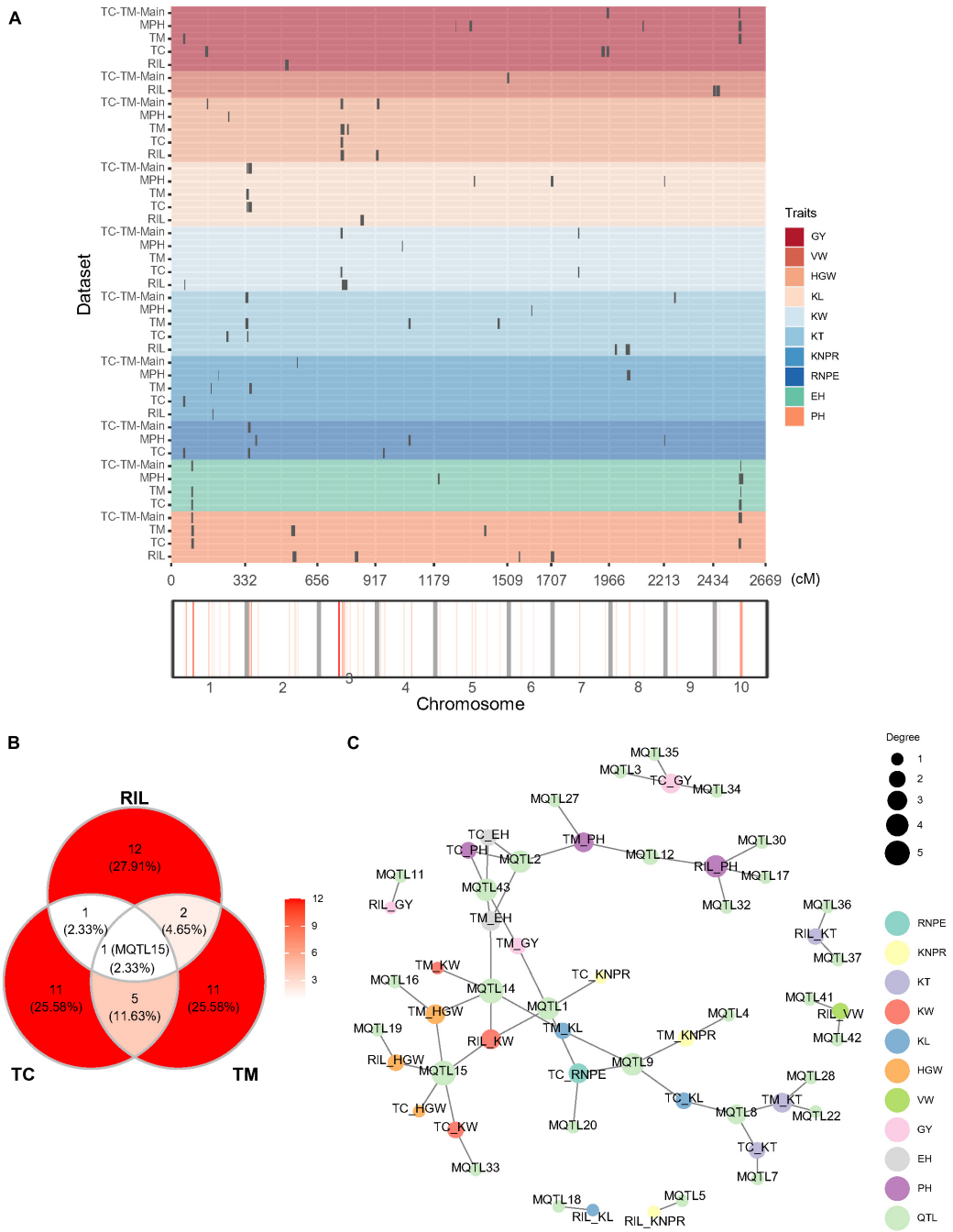

FIGURE 3 | Quantitative trait loci (QTL) distribution and pleiotropic QTL detected in the recombinant inbred line population developed by Ye478 × Qi319 (RIL), Chang7-2 × RIL (TC), and Mo17 × RIL (TM) populations. **(A)** QTL distribution and hotspots in the whole genome shown for RIL, TC, and TM populations. TC-TM-Main is the mapping results for the additive and dominance effects of QTL from the pooled population of TC and TM. MPH represents the result of dominance QTL mapping for MPH. **(B)** Venn diagram showing the numbers of overlapping QTL between the RIL, TC, and TM populations. **(C)** Trait-QTL network for the 10 traits and QTL identified in the RIL, TC, and TM populations. The connections between traits and QTL are linked if a QTL was identified for the respective trait. PH, plant height; EH, ear height; RNPE, row number per ear; KNPR, kernel number per row; KT, kernel thickness; KW, kernel width; KL, kernel length; VW, volume weight; HGW, hundred grain weight; GY, grain yield per plant.

from the Ye478 parental line over the 10 traits. In the TC population, a total of 18 QTL were identified, among which 10 superior alleles came from the Ye478 parent. 19 QTL were identified in the TM population, among which eight superior alleles originated from the Ye478 parent. Three common QTL were jointly identified in the RIL and TM populations, two

QTL were shared by the RIL and TC populations, and six QTL were jointly detected in the TC and TM populations (**Figure 3B**). One QTL (MQTL15) located in the interval from 162.93 to 172.23 Mb on chromosome 3 was shared among all three populations. This QTL was associated with KW in the RIL and the TC population and with HGW in all three populations (**Figure 3C**). A few other QTL also showed pleiotropic effects. For example, MQTL9 located between 1.71 and 4.67 Mb on chromosome 2 was associated with RNPE and KL in the TC population and with KNPR and KL in the TM population.

In the above QTL mapping results, low overlapping ratios among TC, TM and RIL populations were observed. In addition, in phenotype, the top 10 lines in the TC population did not match those identified in the TM population or vice versa (**Supplementary Figures 3A,B**). The two genetic phenomenons suggested that non-additive effects were important for hybrid performance. In this case, it is interesting to further dissect the hybrid performance to mine dominance and epistatic QTL.

Multiple Variance Components Dissection and Main Effect Quantitative Trait Loci Mapping for Hybrid Performance and Midparent Heterosis in the TC-TM Population

We dissected the contribution of all five variance components (additive, dominance, and three epistatic polygenic variances) by Bayesian generalized linear regression (Pérez and De Los Campos, 2014) based on the hybrid performance and MPH in the TC-TM population. The results for the hybrid performance showed that additive-by-additive was the most important polygenic effect for the traits PH, EH, and KT, additive-by-dominance was predominant for VW and dominance was the most important polygenic effect for the remaining traits (**Supplementary Table 6**). For the analysis of MPH, the additive-by-dominance variance contributed the most for traits KT, KW, VW, and HGW, while the dominance variance contributed the most for the other six traits (**Supplementary Table 7**). Different proportions of dominant variances among 10 traits showed the complexity of heterosis.

We implemented a mixed model to test the main (additive and dominance) effects of a specific marker for both the hybrid performance and the MPH for all traits in the pooled TC-TM population. A total of 21 main effect QTL were identified for the 10 traits for hybrid performance (**Supplementary Table 5** and **Figure 4A**). Among them, one had a significant dominance effect for KNPR and was located in the interval 210.29–211.57 Mb on chromosome 2 (**Figure 4B** and **Supplementary Figure 4**). For the other 20 QTL, the additive and dominance effects were confounded due to the fact that there were only two genotypes per locus (**Supplementary Figure 4**). Moreover, a total of 17 dominance QTL were detected for MPH for the 10 traits (**Figure 4C** and **Supplementary Table 5**). Interestingly, only two detected QTL were in common between MPH and hybrid performance (**Figure 4A**). The pleiotropic QTL MQTL43 located in the interval around 80.08–112.87 Mb on chromosome 10 was associated with EH and GY in the MPH dataset and with PH, EH and GY in the TC-TM-Main dataset (**Supplementary Table 5** and **Figure 4C**). The lack of common QTL between MPH and hybrid performance implies that the two phenomena might have different genetic architectures, consistent with the results of the variance component analysis.

Epistasis Plays an Important Role in Hybrid Performance

For hybrid performance in the TC-TM population, we scanned the entire genome to identify significant epistasis loci for the 10 traits and 197, 176, 131, and 112 significant epistatic pairs of loci were identified for additive-by-additive, additive-by-dominance, dominance-by-additive and dominance-by-dominance effects, respectively (**Supplementary Table 8**). The number of significant locus pairs varied across traits and the proportion of explained variance of an epistatic interaction ranged from 3.46 to 4.52%. For grain yield, only one significant additive-by-dominance QTL were detected. We observed the phenomenon of a continuous region interacting with another locus in the genome. For example, for additive-by-additive mapping, the interaction between a cluster of adjacent SNPs on chromosome 8 (Chr8_180048590, Chr8_180913576, Chr8_181023046, and Chr8_180032314) and a locus on chromosome 6 (Chr6_166754537) was significantly associated with PH (**Supplementary Table 8**).

EH had a more simple genetic architecture compared to GY and the variation of MPH for EH was also higher (**Supplementary Table 1**). We therefore used the trait EH as an example to investigate the epistatic effects in the RIL and the two hybrid populations. In the RIL population, no QTL was identified (**Figure 5A**). However, in the pooled TC-TM population, two main effect QTL for hybrid performance were identified on chromosomes 1 (*TC-TM-Main-qEH1* represented by the peak SNP Chr1_131115160) and 10 (*TC-TM-Main-qEH10* represented by the peak SNP Chr10_91890676) (**Figure 5B**). For MPH, however, only the *MPH-qEH10* QTL had a significant dominance effect as well as several additional small-effect QTL (**Figure 5C**). We further tested the additive-by-additive interactions between *TC-TM-Main-qEH1* and all other SNPs (4,140 in total). None of the tested effects were significant in the RIL population (**Figure 5D**). However, several significant interactions were identified in the pooled TC-TM population (**Figure 5E**). Further analysis confirmed the interaction between the two loci *TC-TM-Main-qEH1* and *TC-TM-Main-qEH10* in the pooled TC-TM population (**Figure 5F**). The specific type of epistatic effect between the two loci in the TC-TM population could not be determined because there were only two different genotypes at each locus. However, as we observed that the additive-by-additive effect was not significant between these two loci in the RIL population, we concluded that it is likely the additive-by-dominance or dominance-by-dominance effects that led to the detection

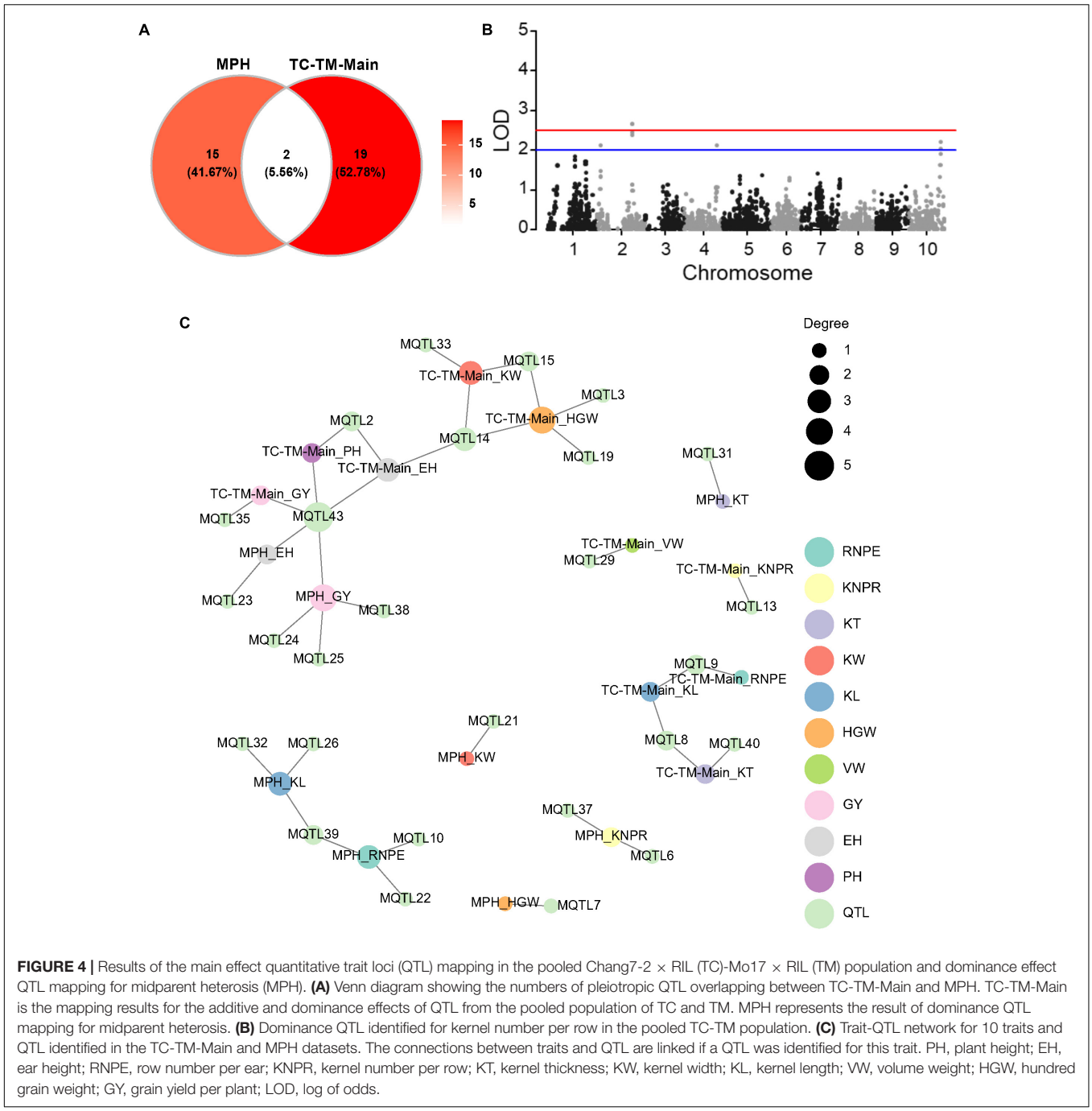

FIGURE 4 | Results of the main effect quantitative trait loci (QTL) mapping in the pooled Chang7-2 × RIL (TC)-Mo17 × RIL (TM) population and dominance effect QTL mapping for midparent heterosis (MPH). **(A)** Venn diagram showing the numbers of pleiotropic QTL overlapping between TC-TM-Main and MPH. TC-TM-Main is the mapping results for the additive and dominance effects of QTL from the pooled population of TC and TM. MPH represents the result of dominance QTL mapping for midparent heterosis. **(B)** Dominance QTL identified for kernel number per row in the pooled TC-TM population. **(C)** Trait-QTL network for 10 traits and QTL identified in the TC-TM-Main and MPH datasets. The connections between traits and QTL are linked if a QTL was identified for this trait. PH, plant height; EH, ear height; RNPE, row number per ear; KNPR, kernel number per row; KT, kernel thickness; KW, kernel width; KL, kernel length; VW, volume weight; HGW, hundred grain weight; GY, grain yield per plant; LOD, log of odds.

of this epistatic QTL in the hybrid population but not in the RIL population.

Correlation Between the Number of Favorable Quantitative Trait Loci and Hybrid Performance

We chose a slightly lower significance threshold of LOD = 2.0 to obtain more loci for this analysis, which yielded four and six significant QTL for GY in the TC and TM population, respectively. If the performance of heterozygous genotypes was better than that of homozygous genotypes at one QTL, it was called a heterozygous favorable QTL; otherwise, it was called a homozygous favorable QTL. The correlations between the number of favorable QTL and the hybrid performance were calculated for all 10 traits (**Supplementary Table 9**). The correlations between the hybrid performance and the number of favorable homozygous QTL (r_1), the number of favorable heterozygous QTL (r_2) and the total number of favorable QTL (r_3) varied across traits, but were significant for most of traits in both the TC and TM populations.

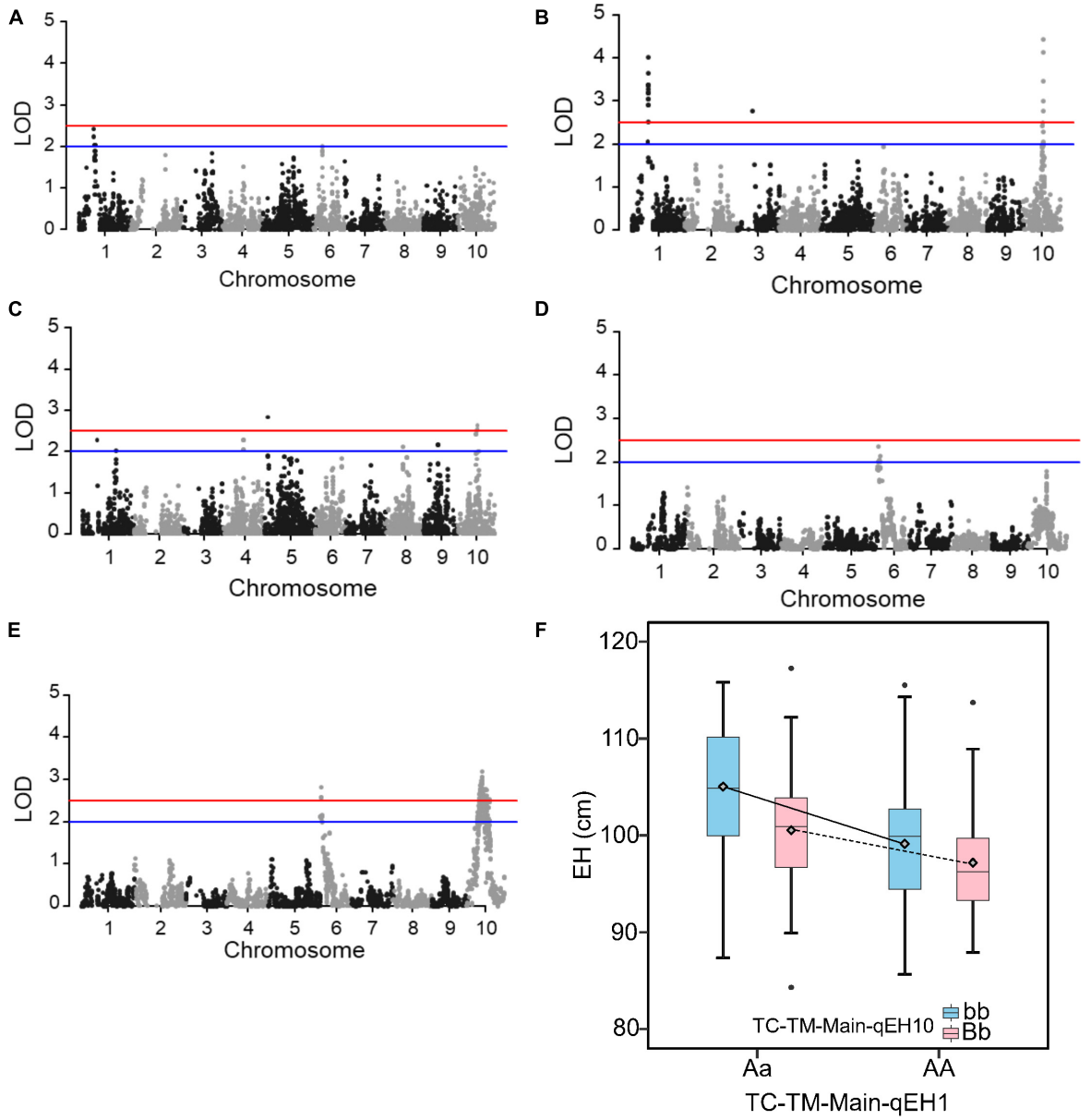

FIGURE 5 | Quantitative trait loci (QTL) mapping results for the trait ear height. **(A)** QTL mapping in the recombinant inbred line population developed by Ye478 × Qi319 (RIL) population. **(B)** In the pooled Chang7-2 × RIL (TC)-Mo17 × RIL (TM) population, and **(C)** dominance QTL mapping for midparent heterosis. **(D)** Test for epistasis between the QTL *TC-TM-Main-qEH1* (peak single nucleotide polymorphisms is Chr1_131115160) on chromosome 1 and the other 4,140 markers in the RIL population. **(E)** Test for epistasis between the QTL *TC-TM-Main-qEH1* (peak single nucleotide polymorphisms is Chr1_131115160) on chromosome 1 and the other 4,140 markers in the pooled TC-TM population. The red horizontal line indicates the significance threshold used for QTL detection and the blue line is the threshold to identify the loci for the favorable QTL analysis. **(F)** The interactions between different genotypes of QTL *TC-TM-Main-qEH1* (Chr1_116118501; Aa, AA) and different genotypes of QTL *TC-TM-Main-qEH10* (Chr10_91890676; bb, Bb). The diamond indicates the mean value of different genotypes. LOD, log of odds.

In the TC population, three of the four QTL for GY were heterozygous favorable QTL and r_1, r_2, and r_3 were 0.16 (**Supplementary Table 9**), 0.43 (**Figure 6A**) and 0.41 (**Figure 6B**), respectively. In the TM population, only two of the six detected QTL were heterozygous favorable QTL and r_1, r_2, and r_3 were 0.42 (**Figure 6C**), 0.25 (**Figure 6D**), and 0.47 (**Figure 6E**), respectively. These results illustrate that superior hybrids can be selected by combining favorable alleles at significant loci.

Genomic Selection Accuracy in Different Breeding Schemes

For genomic selection within populations, the prediction accuracy ranged from 0.70 for RNPE to 0.40 for VW in

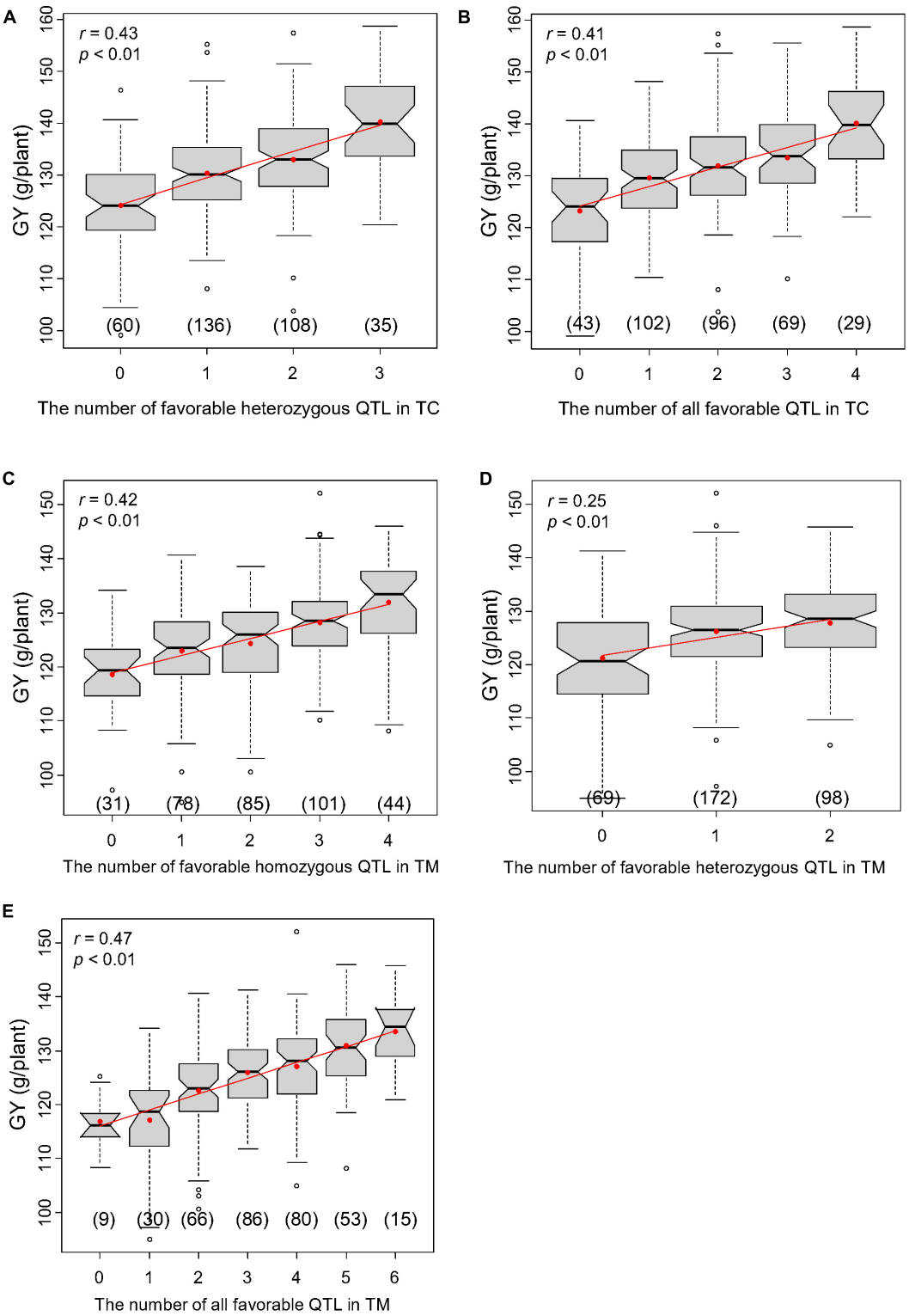

FIGURE 6 | Correlations between the number of favorable quantitative trait loci (QTL) and grain yield per plant (GY). **(A)** Correlation between the number of all favorable QTL and GY in Chang7-2 × RIL (TC) population and **(B)** between the number of favorable heterozygous QTL and GY in TC population. **(C)** Correlation between the number of all favorable QTL and GY in Mo17 × RIL (TM) population and **(D)** between the number of favorable heterozygous QTL and GY in TM population. **(E)** Correlation between the number of favorable homozygous QTL and GY in TM population.

TC and ranged from 0.63 for KT to 0.51 for GY in TM population (**Figure 7A**). Generally, traits with a low heritability usually had a low prediction accuracy, like VW in TC and GY in TM population.

The three cross-validation schemes were shown in **Supplementary Figure 1**. For strategy 1 (CV1), when using the TC population to predict the TM population (TC_TM), the prediction accuracy ranged from 0.305 for GY to 0.699 for EH (**Figure 7B**). Conversely, the prediction accuracy ranged from 0.287 for GY to 0.73 for EH when the TM population was used to predict the TC population (TM_TC). When the significant QTL identified in the training population were included as fixed effects in the prediction model, this did not result in an improvement of the prediction accuracy for most of the traits. Only a few traits, e.g., EH and KL, showed a slight improvement. For some traits, e.g., KNPR and KT, the prediction accuracy even decreased.

For the second cross-validation strategy (CV2), the lowest prediction accuracy was 0.49 for VW and the highest prediction accuracy was 0.90 for RNPE. For the third cross-validation strategy (CV3), the lowest prediction accuracy was 0.53 for GY and the highest prediction accuracy was 0.92 for KT and RNPE (**Figure 7C**). The results also showed that the prediction accuracy of CV3 was higher than within population scheme and CV1, CV2, regardless of whether GS or wGS was applied. The wGS taking potential QTL as fixed had a higher prediction accuracy than GS in both CV2 and CV3 (**Figure 7C**).

DISCUSSION

Hundred Grain Weight and Kernel Number per Row Significantly Contribute to the Variation of Grain Yield

Grain yield is a complex trait, affected by many genetic and non-genetic factors. The three traits that were found to mainly contribute to GY are HGW, RNPE, and KNPR. In earlier studies, the focus has been placed on correlation analysis between traits. In general, moderate to high correlations were observed between GY and many other traits (Cui et al., 2016; Li et al., 2020, 2021). However, it is difficult to determine which trait contributes the most to grain yield. Path analysis is an alternative approach that allows examining the relative importance of a component trait to the variation of the target trait. Our results revealed that regardless of whether the population was the RIL population or the hybrid population, HGW and KNPR had the highest path coefficients thus, contributed the most to the variation of GY (**Figures 2B–D**). Consequently, HGW and KNPR are promising indirect traits to improve GY in hybrid breeding in maize.

Identification of Quantitative Trait Loci Using an Additive Genetic Model

Previous studies in the underlying populations (RIL, TC, and TM) focused on additive genetic models, where the genotypes and phenotypes of the RIL population were used to map QTL by treating the GCA as traits (Zhou et al., 2018; Lu et al., 2020). In this study, the genotypes and phenotypes of two hybrid populations were used to detect significant main (additive and dominance) effects and epistatic QTL. Only two common QTL were identified between the RIL and the TC hybrid population, and three common QTL were identified between the RIL and the TM hybrid population (**Figure 3B**). These results indicated that the non-additive genetic effect played an important role, which meant that a line with a moderate value of GY can still yield a high GY when crossing with testers (**Supplementary Figure 3**). And an additive model was not enough to explain heterosis.

Non-additive Polygenic Effects Play an Important Role in Hybrid Performance

The proportion of phenotypic variance explained by the additive-by-additive effects in the RIL population was higher for most traits than that in the TC and TM populations (**Supplementary Table 4**). And the unparallel relationship between RILs and hybrid populations (**Supplementary Figure 3**) inspired further exploration of the genetic basis in the combined TC-TM hybrid population using a model integrating non-additive polygenic effects. Based on the estimated variance components, we conclude that the prominent gene action varies across traits (**Supplementary Table 6**). For example, the prominent variance was the additive-by-additive component for the trait PH, but for GY the prominent variance was the dominance component. This presents a considerable challenge for selecting elite single-cross hybrids and for uncovering the importance of non-additive genetic effects because additive variance is not a prominent factor to control the variation of any of the investigated traits.

Identification of Quantitative Trait Loci Using a Non-additive Model

For genetic dissection of the hybrid performance, Xu (2013) proposed a new mixed model method for QTL mapping by incorporating multiple polygenic covariance structures, which consist of the additive, dominance, and epistatic variance components. In theory, each particular effect could be tested in a model by controlling all other genetic effects as background. Similarly, a quantitative genetic framework was proposed for the genetic dissection of MPH (Jiang et al., 2017). The above two linear mixed models are very similar in the polygenic background control. The main difference is the response variable, in the former it is the performance of the hybrid (Xu, 2013) and in the latter the MPH (Jiang et al., 2017). It should be noted that the hybrid performance could not be simply replaced by MPH with just the removal of the additive effect from the linear mixed model (Jiang et al., 2017). In our study, the GBS technology only covered about 0.07-fold of the genome in our populations. The two tester lines had the same genotype at 95% of the loci, which resulted in the pooled TC-TM population having just two genotypes at 95% of loci. Two genotypes per locus mimick a backcross population so that the additive effects are confounded with the dominance effects, which explains why the TC-TM-Add model was nearly the same as the TC-TM-Dom model (with one exception for a dominance QTL) (**Supplementary Figure 4**). Regardless of the high similarity between the hybrid populations and the hypothetical BC population, the hybrid

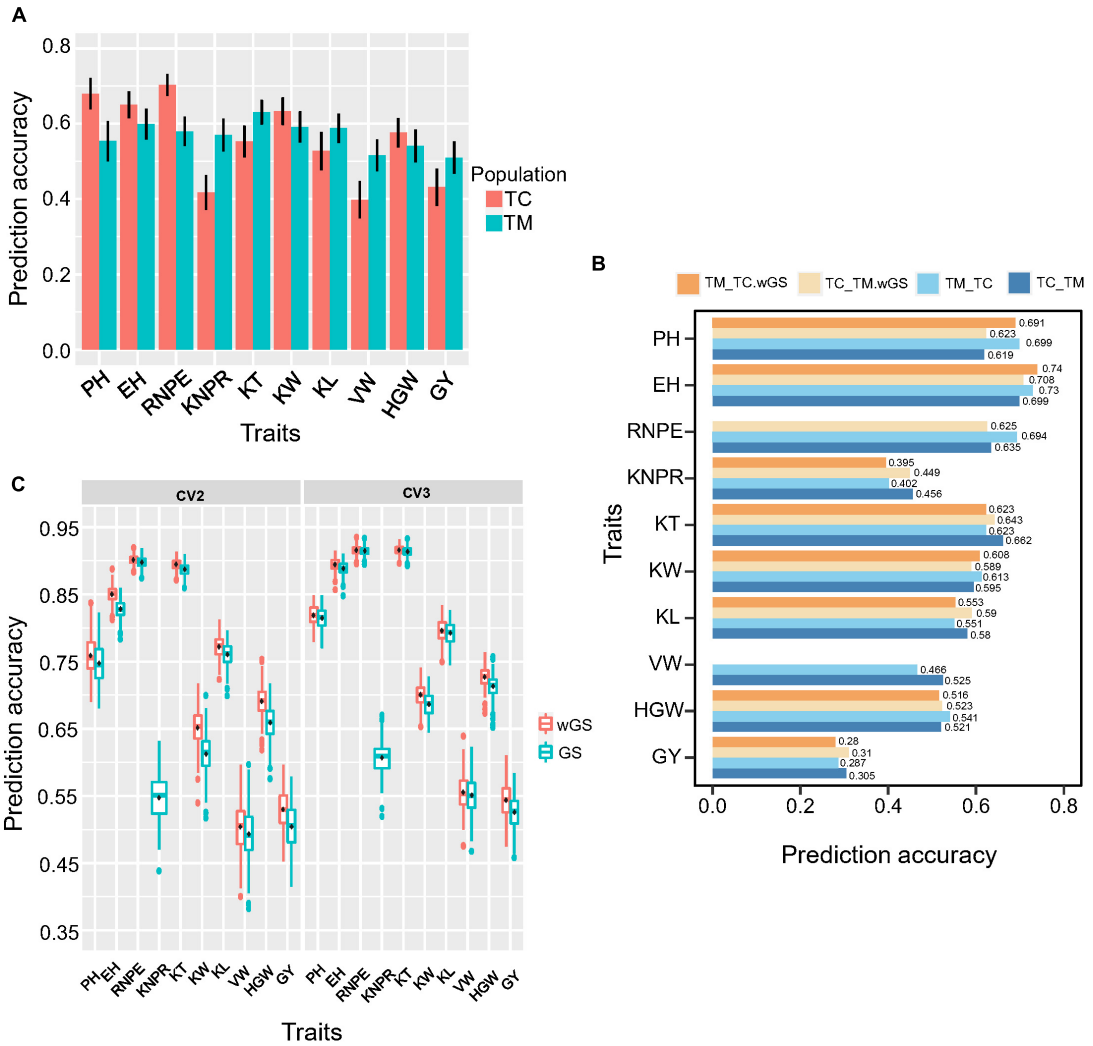

FIGURE 7 | Genomic prediction accuracy of different cross-validation strategies. **(A)** Prediction accuracy within Chang7-2 × RIL (TC) and Mo17 × RIL (TM) population. RIL, the recombinant inbred line population developed by Ye478 × Qi319. **(B)** Prediction accuracy of 10 traits of cross-validation strategy 1 (CV1). TC_TM represents TM predicted by the TC population; TM_TC represents TC predicted by the TM population. **(C)** Prediction accuracy of 10 traits of cross-validation strategy 2 (CV2) and cross-validation strategy 3 (CV3). wGS is a weighted genomic selection that incorporates the peak single nucleotide polymorphisms (SNP) of potential target quantitative trait loci as fixed effects. PH, plant height; EH, ear height; RNPE, row number per ear; KNPR, kernel number per row; KT, kernel thickness; KW, kernel width; KL, kernel length; VW, volume weight; HGW, hundred grain weight; GY, grain yield per plant.

populations still had advantages. In a dominance test for MPH, 17 significant dominance loci were detected (**Figure 4A**) and for hybrid performance, a set of significant epistatic loci was identified (**Supplementary Table 8**).

Improve Prediction Accuracy by Integrating Functional Markers

In genomic selection models like GBLUP or rrBLUP, all SNPs were treated equally or had the same distribution when treated as random. Actually, significant QTL contributed more to the variation of traits. In such a case, significant SNPs should be treated differently. In this study, those SNPs were included in the fixed effect in the GS model to explore whether prediction accuracy could be improved. For cross-validation scheme 1, just two of 10 traits, namely EH and KL showed slight improvement. We guess that TC and TM populations had different genetic backgrounds (Li et al., 2019) and those QTL made different effect within two different populations. So the QTL effect was estimated biasedly when only one tester population as training population. But for CV2 and CV3, the population used for QTL mapping consisted of lines from both TC and TM, in this case, improvements were observed for all traits harboring QTL (**Figure 7C**). By comparison, it showed that CV3 scheme had superiority over the within population scheme, CV1, and CV2, which inspired us that when there were some known functional QTL in a target population, a strategy treating known QTL as fixed effects with CV3 design was a better choice for genomic prediction.

Relationship Between Midparent Heterosis and Hybrid Performance

Hybrid performance is the phenotypic value of the hybrid, which is the sum of the midparent heterosis and the midparent value. Hybrid performance is controlled by the additive, the dominance and all four epistatic polygenic effects, whereas MPH is not affected by the additive effect because the additive effect does not contribute to heterosis (Jiang et al., 2017). In this study, we confirmed the different genetic architecture of hybrid performance and MPH as both had only two QTL in common (**Figure 4A**), which is consistent with a previous study (Hua et al., 2003). Furthermore, the variance component ratios were also different between hybrid performance and MPH (**Supplementary Tables 6**, **7**). In a wheat study, the midparent value showed a negative correlation with MPH but was positively correlated with the hybrid performance (Boeven et al., 2020). In our study, we observed the correlation between the hybrid performance and MPH was 0.77 ($p < 0.01$) (**Supplementary Figure 5A**), and a positive correlation of 0.23 ($p < 0.01$) between the hybrid performance and midparent value (**Supplementary Figure 5B**), while a negative correlation between midparent value and MPH for grain yield of −0.45 ($p < 0.01$) (**Supplementary Figure 5C**). And the path analysis also highlighted the superior contribution of MPH to hybrid performance in hybrid population (**Supplementary Figure 5D**). In plant hybrid breeding, we aim to select a single-cross hybrid with both high MPH and midparent value, but these seem to be contradictory goals considering the negative correlation. Consequently, hybrid breeding must balance the two and target hybrid performance to achieve high performing hybrids.

Mechanisms of Midparent Heterosis and Hybrid Performance

Although dominant and additive effects couldn't be separated and dominant degree couldn't be estimated in this study, multiple variance components dissection provided possibility to assess the mechanism of heterosis and hybrid performance. Results showed the dominance contributed the highest proportion for MPH of most traits, especially for GY and KNPR (**Supplementary Table 7**). However, it was found that the epistasis (sum of additive-by-additive, additive-by-dominance, and dominance-by-dominance) contributes the highest proportion to hybrid performance of GY, PH, EH, and KNPR (**Supplementary Table 6**). The results were similar to a previous report in maize (Tang et al., 2010). A series of linear mixed models incorporating multiple polygenic covariance structures together with NCII population provide possibility to explore the genetic factors and mechanism of heterosis.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

WL, ML, and SX designed the study. ZZ collected the phenotypic data. DL, XiaL, GL, JL, and HW performed data analysis. DL drafted the manuscript. DL, XiaL, YJ, SC, TW, JR, XinL, SX, ML, and WL revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Key Research and Development Program of China (grant numbers: 2016YFD0101201) and Science and Technology Innovation Team of Maize Modern Seed Industry in Hebei (grant numbers: 21326319D).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.774478/full#supplementary-material

Supplementary Figure 1 | Three different cross-validation schemes. **(A)** Cross-validation strategy 1. **(B)** Cross-validation strategy 2. **(C)** Cross-validation strategy 3. RILs, the recombinant inbred lines developed by Ye478 × Qi319.

Supplementary Figure 2 | Collinearity between the genetic and physical maps.

Supplementary Figure 3 | Parallel maps of grain yield per plant (GY) in the recombinant inbred line population developed by Ye478 × Qi319 (RIL), Chang7-2 × RIL (TC), and Mo17 × RIL (TM) populations. **(A)** GY was ranked based on the TC population. **(B)** GY was ranked based on the TM population.

Supplementary Figure 4 | Venn diagram showing the numbers of pleiotropic quantitative trait loci (QTL) overlapping among TC-TM-Add, TC-TM-Dom and MPH. TC-TM-Add represents the mapping results for the additive effects in the pooled population of Chang7-2 × RIL (TC) and Mo17 × RIL (TM). RIL, the recombinant inbred line population developed by Ye478 × Qi319; TC-TM-Dom represents the mapping results for the dominance effects in the pooled population of TC and TM; MPH represents the result of dominance QTL mapping for midparent heterosis.

Supplementary Figure 5 | Correlations and path coefficients among hybrid performance, midparent heterosis (MPH) and midparent value. **(A)** Correlation between MPH and hybrid performance. **(B)** Correlation between midparent value and hybrid performance. **(C)** Correlation between midparent value and MPH. **(D)** The path coefficients among hybrid performance, MPH and midparent value.

Supplementary Table 1 | Summary statistics for 10 traits for midparent heterosis in the Chang7-2 × RIL (TC), and Mo17 × RIL (TM) populations.

Supplementary Table 2 | Variance of general combining ability (GCA) and specific combining ability (SCA) and their interaction with the environment.

Supplementary Table 3 | Summary statistics for the genetic distances across 10 linkage groups of the maize genome.

Supplementary Table 4 | Variance components and proportion of the phenotypic variance contributed by each variance component in the recombinant inbred line population developed by Ye478 × Qi319 (RIL), Chang7-2 × RIL (TC), and Mo17 × RIL (TM) populations, respectively.

Supplementary Table 5 | QTL mapping results for 10 traits in the recombinant inbred line population developed by Ye478 × Qi319 (RIL), Chang7-2 × RIL (TC), Mo17 × RIL (TM) populations, the pooled TC-TM population and heterosis dataset.

Supplementary Table 6 | Variance components and proportion of the phenotypic variance contributed by each variance component in the pooled Chang7-2 × RIL (TC)-Mo17 × RIL (TM) population.

Supplementary Table 7 | Variance components and proportion of the phenotypic variance contributed by each variance component for midparent heterosis.

Supplementary Table 8 | Significant epistatic paired loci for 10 traits identified from the pooled Chang7-2 × RIL (TC)-Mo17 × RIL (TM) population.

Supplementary Table 9 | Correlations between the number of harbored favorable quantitative trait loci (QTL) and hybrid performance for 10 traits in Chang7-2 × RIL (TC) and Mo17 × RIL (TM) population, respectively.

REFERENCES

Albrecht, T., Auinger, H. J., Wimmer, V., Ogutu, J. O., Knaak, C., Ouzunova, M., et al. (2014). Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. *Theor. Appl. Genet.* 127, 1375–1386. doi: 10.1007/s00122-014-2305-z

Bernal-Vasquez, A. M., Utz, H. F., and Piepho, H. P. (2016). Outlier detection methods for generalized lattices: a case study on the transition from ANOVA to REML. *Theor. Appl. Genet.* 129, 787–804. doi: 10.1007/s00122-016-2666-6

Bernardo, R. (2014). Genomewide selection when major genes are known. *Crop Sci.* 54, 68–75. doi: 10.2135/cropsci2013.05.0315

Birchler, J. A., Yao, H., and Chudalayandi, S. (2006). Unraveling the genetic basis of hybrid vigor. *Proc. Natl. Acad. Sci. U.S.A.* 103, 12957–12958. doi: 10.1073/pnas.0605627103

Boeven, P. H. G., Zhao, Y., Thorwarth, P., Liu, F., Maurer, H. P., Gils, M., et al. (2020). Negative dominance and dominance-by-dominance epistatic effects reduce grain-yield heterosis in wide crosses in wheat. *Sci. Adv.* 6:eaay4897. doi: 10.1126/sciadv.aay4897

Broman, K. W. (2001). Review of statistical methods for QTL mapping in experimental crosses. *Lab. Anim.* 30, 44–52. doi: 10.1038/5000133

Broman, K. W., Wu, H., Sen, Ś, and Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19, 889–890. doi: 10.1093/bioinformatics/btg112

Browning, B. L., and Browning, S. R. (2016). Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* 98, 116–126. doi: 10.1016/j.ajhg.2015.11.020

Bruce, A. B. (1910). The Mendelian theory of heredity and the augmentation of vigor. *Science* 32, 627–628.

Bu, S. H., Xinwang, Z., Yi, C., Wen, J., Jinxing, T., and Zhang, Y. M. (2015). Interacted QTL mapping in partial NCII design provides evidences for breeding by design. *PLoS One* 10:e0121034. doi: 10.1371/journal.pone.0121034

Churchill, G. A., and Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* 138, 963–971. doi: 10.1093/genetics/138.3.963

Cui, Y., Li, R., Li, G., Zhang, F., Zhu, T., Zhang, Q., et al. (2020). Hybrid breeding of rice via genomic selection. *Plant Biotechnol. J.* 18, 57–67. doi: 10.1111/pbi.13170

Cui, Z., Luo, J., Qi, C., Ruan, Y., Li, J., Zhang, A., et al. (2016). Genome-wide association study (GWAS) reveals the genetic architecture of four husk traits in maize. *BMC Genomics* 17:946. doi: 10.1186/s12864-016-3229-6

East, E. M. (1936). Heterosis. *Genetics* 21, 375–397.

Falconer, D. S., and Mackay, T. F. C. (1996). *An Introduction to Quantitative Genetics*, 4th Edn. London: Addison Wesley Longman.

Garcia, A. A. F., Wang, S., Melchingerand, A. E., and Zeng, Z. B. (2008). Quantitative trait loci mapping and the genetic basis of heterosis in maize and rice. *Genetics* 180, 1707–1724. doi: 10.1534/genetics.107.082867

Garin, V., Wimmer, V., Mezmouk, S., Malosetti, M., and van Eeuwijk, F. (2017). How do the type of QTL effect and the form of the residual term influence QTL detection in multi-parent populations? A case study in the maize EU-NAM population. *Theor. Appl. Genet.* 130, 1753–1764. doi: 10.1007/s00122-017-2923-3

Gilmour, A. R., Gogel, B. J., Cullis, B. R., and Thompson, R. (2009). *ASReml User Guide Release 3.0*. Hemel Hempstead: VSN International Ltd.

Guo, T., Yang, N., Tong, H., Pan, Q., Yang, X., Tang, J., et al. (2014). Genetic basis of grain yield heterosis in an "immortalized F2" maize population. *Theor. Appl. Genet.* 127, 2149–2158. doi: 10.1007/s00122-014-2368-x

Hochholdinger, F., and Baldauf, J. A. (2018). Heterosis in plants. *Curr. Biol.* 28, R1089–R1092. doi: 10.1016/j.cub.2018.06.041

Hua, J., Xing, Y., Wu, W., Xu, C., Sun, X., Yu, S., et al. (2003). Single-locus heterotic effects and dominance by dominance interactions can adequately explain the genetic basis of heterosis in an elite rice hybrid. *Proc. Natl. Acad. Sci. U.S.A.* 100, 2574–2579. doi: 10.1073/pnas.0437907100

Hyun, M. K., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., et al. (2008). Efficient control of population structure in model organism association mapping. *Genetics* 178, 1709–1723. doi: 10.1534/genetics.107.080101

Jiang, Y., Schmidt, R. H., Zhao, Y., and Reif, J. C. (2017). A quantitative genetic framework highlights the role of epistatic effects for grain-yield heterosis in bread wheat. *Nat. Genet.* 49, 1741–1746. doi: 10.1038/ng.3974

Jones, D. F. (1917). Dominance of linked factors as a means of accounting for heterosis. *Genetics* 3, 310–312. doi: 10.1073/pnas.3.4.310

Li, D., Chen, Z., Wang, M., Leiser, W. L., Weiß, T. M., Zhao, Z., et al. (2021). Dissecting the phenotypic response of maize to low phosphorus soils by field screening of a large diversity panel. *Euphytica* 217, 1–12. doi: 10.1007/s10681-020-02727-2

Li, D., Wang, P., Gu, R., Fu, J., Xu, Z., Lyle, D., et al. (2019). Genetic relatedness and the ratio of subpopulation-common alleles are related in genomic prediction across structured subpopulations in maize. *Plant Breed.* 138, 802–809. doi: 10.1111/pbr.12717

Li, G., Dong, Y., Zhao, Y., Tian, X., Würschum, T., Xue, J., et al. (2020). Genome-wide prediction in a hybrid maize population adapted to Northwest China. *Crop J.* 8, 830–842. doi: 10.1016/j.cj.2020.04.006

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993. doi: 10.1093/bioinformatics/btr509

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

Lippman, Z. B., and Zamir, D. (2007). Heterosis: revisiting the magic. *Trends Genet.* 23, 60–66. doi: 10.1016/j.tig.2006.12.006

Liu, N., Du, Y., Warburton, M. L., Xiao, Y., and Yan, J. (2021). Phenotypic plasticity contributes to maize adaptation and heterosis. *Mol. Biol. Evol.* 38, 1262–1275. doi: 10.1093/molbev/msaa283

Lu, X., Zhou, Z., Yuan, Z., Zhang, C., Hao, Z., Wang, Z., et al. (2020). Genetic dissection of the general combining ability of yield-related traits in maize. *Front. Plant Sci.* 11:788. doi: 10.3389/fpls.2020.00788

Ma, X., Shao, Y., Tian, L., Flasch, D. A., Mulder, H. L., Edmonson, M. N., et al. (2019). Analysis of error profiles in deep next-generation sequencing data. *Genome Biol.* 20, 1–15. doi: 10.1186/s13059-019-1659-6

McCouch, S. R., Cho, Y. G., Yano, M., Paul, E., Blinrub, M., Morishima, H., et al. (1997). Rice report on QTL nomenclature. *Rice Genet. Newsl.* 14, 11–13.

Melchinger, A. E., Utz, H. F., Piepho, H. P., Zeng, Z. B., and Schön, C. C. (2007b). The role of epistasis in the manifestation of heterosis: a systems-oriented approach. *Genetics* 177, 1815–1825. doi: 10.1534/genetics.107.077537

Melchinger, A. E., Piepho, H. P., Utz, H. F., Muminović, J., Wegenast, T., Törjék, O., et al. (2007a). Genetic basis of heterosis for growth-related traits in *Arabidopsis* investigated by testcross progenies of near-isogenic lines reveals a significant role of epistasis. *Genetics* 177, 1827–1837. doi: 10.1534/genetics.107.080564

Meng, L., Li, H., Zhang, L., and Wang, J. (2015). QTL IciMapping: integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations. *Crop J.* 3, 269–283. doi: 10.1016/j.cj.2015.01.001

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829.

Pérez, P., and De Los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics* 198, 483–495. doi: 10.1534/genetics.114.164442

Powers, D. L. (1944). An expansion of Jones's theory for the explanation of heterosis. *Am. Nat.* 78, 275–280. doi: 10.1086/281199

Reif, J. C., Gumpert, F. M., Fischer, S., and Melchinger, A. E. (2007). Impact of interpopulation divergence on additive and dominance variance in hybrid populations. *Genetics* 176, 1931–1934. doi: 10.1534/genetics.107.074146

Reif, J. C., Kusterer, B., Piepho, H. P., Meyer, R. C., Altmann, T., Schön, C. C., et al. (2009). Unraveling epistasis with triple testcross progenies of near-isogenic lines. *Genetics* 181, 247–257. doi: 10.1534/genetics.108.093047

Stuber, C. W., Lincoln, S. E., Wolff, D. W., Helentjaris, T., and Lander, E. S. (1992). Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. *Genetics* 132, 823–839. doi: 10.1093/genetics/132.3.823

Su, G., Christensen, O. F., Ostersen, T., Henryon, M., and Lund, M. S. (2012). Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. *PLoS One* 7:e45293. doi: 10.1371/journal.pone.0045293

Tang, J., Yan, J., Ma, X., Teng, W., Wu, W., Dai, J., et al. (2010). Dissection of the genetic basis of heterosis in an elite maize hybrid by QTL mapping in an immortalized F2 population. *Theor. Appl. Genet.* 120, 333–340. doi: 10.1007/s00122-009-1213-0

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

Utz, H. F., Melchinger, A. E., and Schön, C. C. (2000). Bias and sampling error of the estimated proportion of genotypic variance explained by quantitative trait loci determined from experimental data in maize using cross validation and validation with independent samples. *Genetics* 154, 1839–1849.

Wei, X., Lu, X., Zhang, Z., and Xu, M. (2016). Genetic analysis of heterosis for maize grain yield and its components in a set of SSSL testcross populations. *Euphytica* 210, 181–193. doi: 10.1007/s10681-016-1695-1

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag.

Würschum, T., Leiser, W. L., Langer, S. M., Tucker, M. R., and Longin, C. F. H. (2018). Phenotypic and genetic analysis of spike and kernel characteristics in wheat reveals long-term genetic trends of grain yield components. *Theor. Appl. Genet.* 131, 2071–2084. doi: 10.1007/s00122-018-3133-3

Xie, W., Feng, Q., Yu, H., Huang, X., Zhao, Q., Xing, Y., et al. (2010). Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 107, 10578–10583. doi: 10.1073/pnas.1005931107

Xu, S. (2013). Mapping quantitative trait loci by controlling polygenic background effects. *Genetics* 195, 1209–1222. doi: 10.1534/genetics.113.157032

Xu, S., Zhu, D., and Zhang, Q. (2014). Predicting hybrid performance in rice using genomic best linear unbiased prediction. *Proc. Natl. Acad. Sci. U.S.A.* 111, 12456–12461. doi: 10.1073/pnas.1413750111

Yi, Q., Liu, Y., Hou, X., Zhang, X., Li, H., Zhang, J., et al. (2019). Genetic dissection of yield-related traits and mid-parent heterosis for those traits in maize (*Zea mays* L.). *BMC Plant Biol.* 19:392. doi: 10.1186/s12870-019-2009-2

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38, 203–208. doi: 10.1038/ng1702

Zhao, Y., Mette, M. F., and Reif, J. C. (2015b). Genomic selection in hybrid breeding. *Plant Breed.* 134, 1–10. doi: 10.1111/pbr.12231

Zhao, Y., Li, Z., Liu, G., Jiang, Y., Maurer, H. P., Würschum, T., et al. (2015a). Genome-based establishment of a high-yielding heterotic pattern for hybrid wheat breeding. *Proc. Natl. Acad. Sci. U.S.A.* 112, 15624–15629. doi: 10.1073/pnas.1514547112

Zhao, Y., Zeng, J., Fernando, R., and Reif, J. C. (2013). Genomic prediction of hybrid wheat performance. *Crop Sci.* 53, 802–810. doi: 10.2135/cropsci2012.08.0463

Zhou, Z., Zhang, C., Lu, X., Wang, L., Hao, Z., Li, M., et al. (2018). Dissecting the genetic basis underlying combining ability of plant height related traits in maize. *Front. Plant Sci.* 9:1117. doi: 10.3389/fpls.2018.01117

Zhou, Z., Zhang, C., Zhou, Y., Hao, Z., Wang, Z., Zeng, X., et al. (2016). Genetic dissection of maize plant architecture with an ultra-high density bin map based on recombinant inbred lines. *BMC Genomics* 17:178. doi: 10.1186/s12864-016-2555-z

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 03 January 2022
doi: 10.3389/fpls.2021.754439

Dissecting the Genetics of Early Vigour to Design Drought-Adapted Wheat

Stjepan Vukasovic[1], Samir Alahmad[2], Jack Christopher[3], Rod J. Snowdon[1], Andreas Stahl[4] and Lee T. Hickey[2]*

[1] Department of Plant Breeding, IFZ Research Centre for Biosystems, Land Use and Nutrition, Justus Liebig University Giessen, Giessen, Germany, [2] Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, Brisbane, QLD, Australia, [3] Leslie Research Facility, Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, Toowoomba, QLD, Australia, [4] Federal Research Centre for Cultivated Plants, Institute for Resistance Research and Stress Tolerance, Julius Kühn-Institute, Quedlinburg, Germany

Due to the climate change and an increased frequency of drought, it is of enormous importance to identify and to develop traits that result in adaptation and in improvement of crop yield stability in drought-prone regions with low rainfall. Early vigour, defined as the rapid development of leaf area in early developmental stages, is reported to contribute to stronger plant vitality, which, in turn, can enhance resilience to erratic drought periods. Furthermore, early vigour improves weed competitiveness and nutrient uptake. Here, two sets of a multi-reference nested association mapping (MR-NAM) population of bread wheat (*Triticum aestivum* ssp. *aestivum* L.) were used to investigate early vigour in a rain-fed field environment for 3 years, and additionally assessed under controlled conditions in a greenhouse experiment. The normalised difference vegetation index (NDVI) calculated from red/infrared light reflectance was used to quantify early vigour in the field, revealing a correlation ($p < 0.05$; $r = 0.39$) between the spectral measurement and the length of the second leaf. Under controlled environmental conditions, the measured projected leaf area, using a green-pixel counter, was also correlated to the leaf area of the second leaf ($p < 0.05$; $r = 0.38$), as well as to the recorded biomass ($p < 0.01$; $r = 0.71$). Subsequently, genetic determination of early vigour was tested by conducting a genome-wide association study (GWAS) for the proxy traits, revealing 42 markers associated with vegetation index and two markers associated with projected leaf area. There are several quantitative trait loci that are collocated with loci for plant developmental traits including plant height on chromosome 2D ($\log_{10}(P) = 3.19$; PVE = 0.035), coleoptile length on chromosome 1B ($-\log_{10}(P) = 3.24$; PVE = 0.112), as well as stay-green and vernalisation on chromosome 5A ($-\log_{10}(P) = 3.14$; PVE = 0.115).

Keywords: *Triticum aestivum*, normalised difference vegetation index, NDVI, nested association mapping, genome-wide association study, GWAS

OPEN ACCESS

Edited by:
Rodomiro Ortiz, Swedish University of Agricultural Sciences, Sweden

Reviewed by:
Greg Rebetzke, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia
Francisco Pinera, International Maize and Wheat Improvement Center, Mexico
David Anthony Van Sanford, University of Kentucky, United States

****Correspondence:***
Stjepan Vukasovic
Stjepan.Vukasovic@agrar.uni-giessen.de

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *06 August 2021*
Accepted: *01 December 2021*
Published: *03 January 2022*

Citation:
Vukasovic S, Alahmad S, Christopher J, Snowdon RJ, Stahl A and Hickey LT (2022) Dissecting the Genetics of Early Vigour to Design Drought-Adapted Wheat. Front. Plant Sci. 12:754439. doi: 10.3389/fpls.2021.754439

Abbreviations: Abbreviations; CoV, coefficient of variation; DAS, days after sowing; EV, early vigour; GH, greenhouse; GRAMMAR, genome-wide rapid association using mixed model and regression; GWAS, genome-wide association study; L1, leaf one; L2, leaf two; LMM, linear mixed model; MAF, minor allele frequency; Ma-NAM, Mace derived nested association mapping population; MR-NAM, multi-parent nested association mapping population; N, nitrogen; NDVI, normalised difference vegetation index; PC, principal component; PLA, projected leaf area (measured using the green pixel counter); Sc-NAM, scout derived nested association mapping population; SD, standard deviation; Su-NAM, Suntop derived nested association mapping population; TLA, total leaf area (= calculated area of L1 + L2); Var, variance.

INTRODUCTION

Global climate change is considered one of the biggest and most complex challenges the mankind has faced. One effect that has been observed since 1970, which leads to severe yield losses, is the increased occurrence of erratic drought phenomena (Liu et al., 2019). Specifically, the agricultural sector is facing serious challenges since drought-stress is considered the most limiting factor in rain-fed cropping systems (Hu and Xiong, 2014). Based on calculations of the Intergovernmental Panel on Climate Change (IPCC), it is predicted that the global mean surface temperature will rise 2°C more in the 20-year period from 2046 to 2065, than in the comparable period between 1986 and 2005, with a total increase of 4.8°C by 2100 (IPCC, 2015). As a result, more frequent heat and drought events are to be expected and to be classified as a major threat to the primary production sector in general and the wheat production in particular (Steinfort et al., 2017). Based on crop modelling scenarios, it is predicted that global wheat production will fall by 6% per 1°C temperature increase (Asseng et al., 2013). In particular, the Australian wheat production region is expected to experience a strong increase in drought and heat events, with a yield decrease of up to 20% projected from a median temperature increase of 2°C (Asseng et al., 2015). The severe impact of strong drought events has already been observed during the "Millennium Drought" between 2001 and 2009, where major reductions in production were recorded. In southern Australia in particular, production was severely decreased due to the impact of drought during this phase (Dijk et al., 2013). Large areas of the southern and western wheat-cropping regions in Australia have a Mediterranean climate, which is defined by terminal droughts (Siddique et al., 1990; Whan et al., 1991; Botwright et al., 2002; Rebetzke et al., 2008; Sadras and Dreccer, 2015; Rebetzke et al., 2017).

Consequently, there has been a growing focus on commercial and public breeding programs to identify traits associated with water-use efficiency to increase the yield potential under water-limited conditions (Lopes and Reynolds, 2012; Passioura, 2012). Potential traits of interest include long coleoptiles, which enable deeper sowing in the soil profile and improved access to water reservoirs underneath a dry surface soil, as well as reduced tillering to lessen unnecessarily metabolism into the non-fertile emerging tillers (Richards et al., 2010). Another is early vigour (EV), defined as the rapid production of leaf area during the early development phase of the plant (López-Castañeda et al., 1996). The primary advantage of EV is the increased biomass production early in the season and the rapid closure of the canopy, which can reduce evaporation of soil water, which then increases water availability (López-Castañeda and Richards, 1994; Condon et al., 2004). Early canopy closure also leads to minimized solar radiation on the soil and to an enhanced competitiveness of the crop against weeds (Dingkuhn et al., 1999; Coleman et al., 2001; Lemerle et al., 2001; Bertholsson, 2005). Mediterranean growing areas are water-limited environments which are characterized by experiencing late seasons droughts. According to Richards et al. (1987) and López-Castañeda et al. (1995), EV offers great potential for increasing water-use efficiency in such drought-prone regions. Additional advantages that may be associated with EV include larger uptake of essential plant nutrients, superior tolerance to aluminium stress, as well as improved yield under high temperatures and elevated atmospheric CO_2 concentration (Coleman et al., 2001; Lemerle et al., 2001; Liao et al., 2004; Bertholsson, 2005; Ludwig and Asseng, 2010; Valle and Calderini, 2010; Ryan et al., 2015). Previous studies in wheat have already identified various physiological traits associated with EV, such as the embryo size (Moore and Rebetzke, 2015), coleoptile length (Clarke et al., 1991; Rebetzke et al., 2007), tiller size, and leaf characteristics (Rebetzke and Richards, 1999; Rebetzke et al., 2007, 2017).

Several studies have highlighted the beneficial effect of increased EV on yield performance in specific environments. Nevertheless, due to insufficient knowledge about genetic variation and lack of information on the economic value of the trait, EV has only been introduced into breeding programs to a limited degree (Rebetzke et al., 2017). However, Botwright et al. (2002) demonstrated the positive effect of EV on yield performance in medium and low rainfall regions in combination with favourable soil conditions, such as sandy soils. Early vigour has significant implications for water demand. For example, a slight increase in leaf area growth during the vegetative growth stages can lead to an increase in biomass, transpiration area, and water use (Asseng and van Herwaarden, 2003). This can result in rapid depletion of soil water prior to anthesis, which may have a negative impact on grain yield during flowering time and grain filling (Richards and Townley-Smith, 1987). According to Turner and Nicolas (1998), strong vigorous genotypes have deeper and greater water uptake compared to less vigorous genotypes and are subsequently considered to be advantageous in low rainfall environments, including Mediterranean growing regions. However, despite a clear genetic effect on EV, its interaction with the environment, soil type, and available fertilizer is substantial. For example, EV may adversely affect yield if managed unfavourably. Therefore, ensuring that the crop is provided with sufficient nitrogen (N) is essential in order to prevent premature N deficiency due to excessive biomass production (Asseng and van Herwaarden, 2003).

To determine the impact of EV on yield, a better understanding of the physiological mechanisms of the trait along with its genetic control is required. In addition, efficient and low-cost phenotyping procedures are needed to assess EV in wheat. The aim of this study was (i) to investigate the physiological characteristics that drive EV in wheat in the field and under controlled conditions, (ii) to assess the efficiency of phenotyping methodologies under greenhouse and field conditions, and (iii) to identify genomic regions influencing EV in wheat.

MATERIALS AND METHODS

Plant Material

To evaluate EV in the field, a set of 685 spring wheat genotypes (further referred to as Set 1) was randomly selected from a multi-reference nested association mapping population (MR-NAM).

The MR-NAM population was developed based on 11 diverse founder lines which were crossed with the commercially used wheat varieties Suntop (AGT), Scout (LPB), and Mace (AGT), then consequently adapted to the environmental conditions of the western, northern, and southern cropping regions of Australia, respectively (Richard et al., 2015). The founder lines were selected according to key traits, such as drought adaptation and stay-green (e.g., Dharwar Dry, Drysdale), root architecture traits (e.g., Seri) or adaptation to nematodes, and disease resistance (e.g., Wylie, Gregory) (Richard et al., 2015; Christopher et al., 2021). After crossing the founder lines with the three parental lines, using an incomplete crossing scheme, the 15 F_1 lines were generated. Subsequently, these 15 F_1 lines were used for population development through inbreeding, which produces1474 F_4-derived lines and were then segmented into 15 genetically diverse families. These 15 families comprised four Mace-derived families, forming a conventional NAM population which was denoted as the Mace-NAM (Ma-NAM) component of the MR-NAM, five Scout-derived families (Sc-NAM), and six Suntop-derived families (Su-NAM). The NAM population was genotyped using the DArT-seq genotype-by-sequencing platform, producing over 25,000 polymorphic markers (Richard et al., 2015). The first set (SET 1) was tested in experimental years, 2015 and 2016, respectively. In 2017, a randomly selected subset comprising 210 lines (referred to as Set 2) was selected and was tested in the field in a greenhouse (GH) environment. In order to provide a more detailed information of the physiological characteristics, a core set was formed within Set 2, which was intensively investigated in the field and in the GH experiment.

Experimental Design

Evaluating Early Vigour Under Field Conditions

Field trials were conducted over 3 years from 2015 to 2017. All field trials were carried out under rain-fed conditions at the Hermitage Research Facility (HRF), Warwick, Queensland, Australia (28.21°S, 152.10°E, 480 m above sea level). The HRF site is characterized by alkaline, cracking, and heavy clay soils with high water-holding capacity. Cropping season is from May to October, with an average rainfall of 211 mm and an average temperature of 14°C. Further information regarding the environmental conditions is given in **Table 1**. Sites were sown with yield plots, each plot measuring 2 m × 6 m, containing 7 rows at 25 cm spacing, with a target crop density of 100 plants/m^2. To precisely reach the target crop density, thousand seed weight and germination rate were determined for seed of each genotype, while the sowing rates were calculated for each genotype. To avoid any artefacts in seed size, the seeds of the lines used in each trial were sourced from a common site and the year of the seed propagation. Seeds were generated from fully irrigated seed-increase rows sown at 0.5 m row spacing and fertilizer was applied to provide for the non-limiting conditions for both water and nutrients. Furthermore, diseases and weeds were controlled as necessary. This allowed for the full potential seed size of each genotype to be expressed during seed production. In all trials, no specific selection for seed size were performed. In each year the trials received 120 kg/ha^{-1} of urea prior to sowing, and 40 kg/ha^{-1} of Starter Z® (Incitec Pivot Fertilisers, Southbank, VIC, Australia; 10.5% N, 19.5% P, 2.2% S, 2.2% Zn) was applied at sowing. Plant protection measures were applied as necessary. All field trials were designed as a partially replicated (p-rep) block design with percentage of partial replication of 34, 29, and 62% in 2015, 2016, and 2017, respectively.

Across the field trials, the normalised difference vegetation index (NDVI) was used as a quantitative measure for EV. The NDVI measurements were collected using a hand-held NTech Greenseeker® model 505, manufactured by NTech Industries, Ukiah, CA, United States. By attaching the device to the body by using a harness, the measurement could be carried out constantly at a height of 1 m. The NDVI has been reported as a very useful index for studying the dynamics of canopy development and of senescence patterns of wheat (Lopes and Reynolds, 2012; Christopher et al., 2014). This vegetation index has also been previously used to evaluate EV in wheat (Li et al., 2014). In this study, NDVI measurements were collected at 29 days after sowing (DAS) for all field trials.

In 2017, detailed measures were captured to fully understand the physiological properties. A core set consisting of 30 genotypes was compiled from Set 2 and was intensively examined at the time of the NDVI measurement. The core set included all founder and reference lines from the MR-NAM population, along with selected good-performing lines from other unpublished experiments. From each of the 30 genotypes, ten plants per plot were randomly sampled for each genotype, while the leaf characteristics were recorded by measuring the length and width of the first (L1) and the second leaf (L2). Subsequently, the approximate leaf area for L1 and L2 was calculated by multiplying the measured length by the width. The sum of the calculated leaf areas of L1 and L2 was then expressed as the total leaf area (TLA).

Evaluating Early Vigour Under Controlled Conditions

The GH experiment was conducted using 210 genotypes from Set 2. The greenhouse chamber temperature was set on 22°C during daytime and 17°C during night time, while the lighting conditions were set to provide a 12-h photoperiod. Furthermore,

TABLE 1 | Details for environmental conditions for wheat trials subjected to analyses in this study.

Trial	Location	State	Sowing date	CIR [mm]	PAWC [mm]	AvgT [°C]	RAD (MJ m^{-2})
NAM 2015	HRF, Warwick	QLD	11.06.2015	103.4	329.6	13.1	2,318
NAM 2016	HRF, Warwick	QLD	22.07.2016	303.1	212.1	14.9	2,430
NAM 2017	HRF, Warwick	QLD	27.05.2017	112.4	286.47	14.2	2,276

Shon are trials name and year (Trial), location and state, sowing date, cumulative in-crop rainfall in mm (CIR), plant available water capacity of the soil in mm (PAWC), daily average temperature from sowing to maturity in °C (AvgT), and cumulative radiation from sowing to maturity in MJ m^{-2} (RAD).

the plants were irrigated on a daily basis. A single plant was grown per pot using 250 ml pots with 70 mm diameter. The potting media had a pH of 5.5–6.5 and was a composition of 70% pine bark (0–5 mm) and 30% coco peat, as well as fertilizers. Furthermore, Osmocote® (ICL SF, Sydney, NSW, Australia), containing 19.4% N, 16% P, and 5% K, was added to the potting media to guarantee sufficient nutrients during the experiment. The trial was designed as a fully replicated, randomized, and complete block design with six replications per genotype.

To measure leaf dimensions on a larger scale within the GH experiment setup, this study evaluated the extent of the image analysis methods that will be suitable for this purpose. For this purpose, a green pixel counter was programmed using the MATLAB® (MathWorks, Inc., Natick, MA, United States)

FIGURE 1 | Mode of operation of the green pixel counter, analysis takes place in two steps. First capture of the green pixels in the image as indicated by the red outlined area (left). Second, removal of all non-green pixels (right).

TABLE 2 | Descriptive statistics for collected field data for Set 1 (685 lines) and the subsets of lines Set 2 (210 lines) and core set (30 lines) captured at 21 and 29 days after sowing (DAS).

Set	Parameter	Unit	DAS	Year	Descriptive statistics				
					Mean	Min	Max	SD	CoV
Set 1	NDVI		29	2015	0.343	0.26	0.449	0.17	0.5
11	NDVI		29	2016	0.259	0.14	0.447	0.2	0.77
Set 2	NDVI		29	2015	0.341	0.26	0.456	0.03	0.09
12	NDVI		29	2016	0.339	0.27	0.461	0.04	0.11
13	NDVI		21	2017	0.325	0.22	0.394	0.04	0.12
14	NDVI		29	2017	0.264	0.19	0.416	0.03	0.13
Core Set	NDVI		29	2015	0.344	0.28	0.398	0.03	0.08
15	NDVI		29	2016	0.276	0.19	0.403	0.06	0.22
16	NDVI		21	2017	0.245	0.22	0.317	0.03	0.11
17	NDVI		29	2017	0.321	0.24	0.391	0.04	0.13
18	TLA	[cm^2]	21	2017	6.973	5.7	9.609	1.08	0.16
19	Leaf area L1	[cm^2]	21	2017	3.569	2.67	4.658	0.48	0.13
20	Leaf area L2	[cm^2]	21	2017	3.404	1.82	5.904	0.98	0.29
21	Length L1	[cm]	21	2017	10.83	8.59	13.12	1.11	0.1
22	Width L1	[cm]	21	2017	0.33	0.3	0.39	0.03	0.08
23	Length L2	[cm]	21	2017	10.68	8.15	16.96	1.88	0.18
24	Width L2	[cm]	21	2017	0.315	0.22	0.41	0.05	0.15
25	TLA	[cm^2]	29	2017	9.971	6.66	12.78	1.46	0.15
26	Leaf area L1	[cm^2]	29	2017	4.221	2.75	5.62	0.63	0.15
27	Leaf area L2	[cm^2]	29	2017	5.75	3	7.793	1.04	0.18
28	Length L1	[cm]	29	2017	12.04	9.01	13.85	1.04	0.09
29	Width L1	[cm]	29	2017	0.349	0.3	0.42	0.03	0.09
30	Length L2	[cm]	29	2017	15.49	10.3	18.33	1.6	0.1
	Width L2	[cm]	29	2017	0.369	0.29	0.47	0.04	0.12

Shown are the set of lines examined (Set), parameter examined, unit (NDVI is an index without units), time of measurement in days after sowing (DAS), year of experiment as well as for each set examined, mean value, minimum (Min), maximum (Max), variance (Var), standard deviation (SD), and coefficient of variation (CoV).

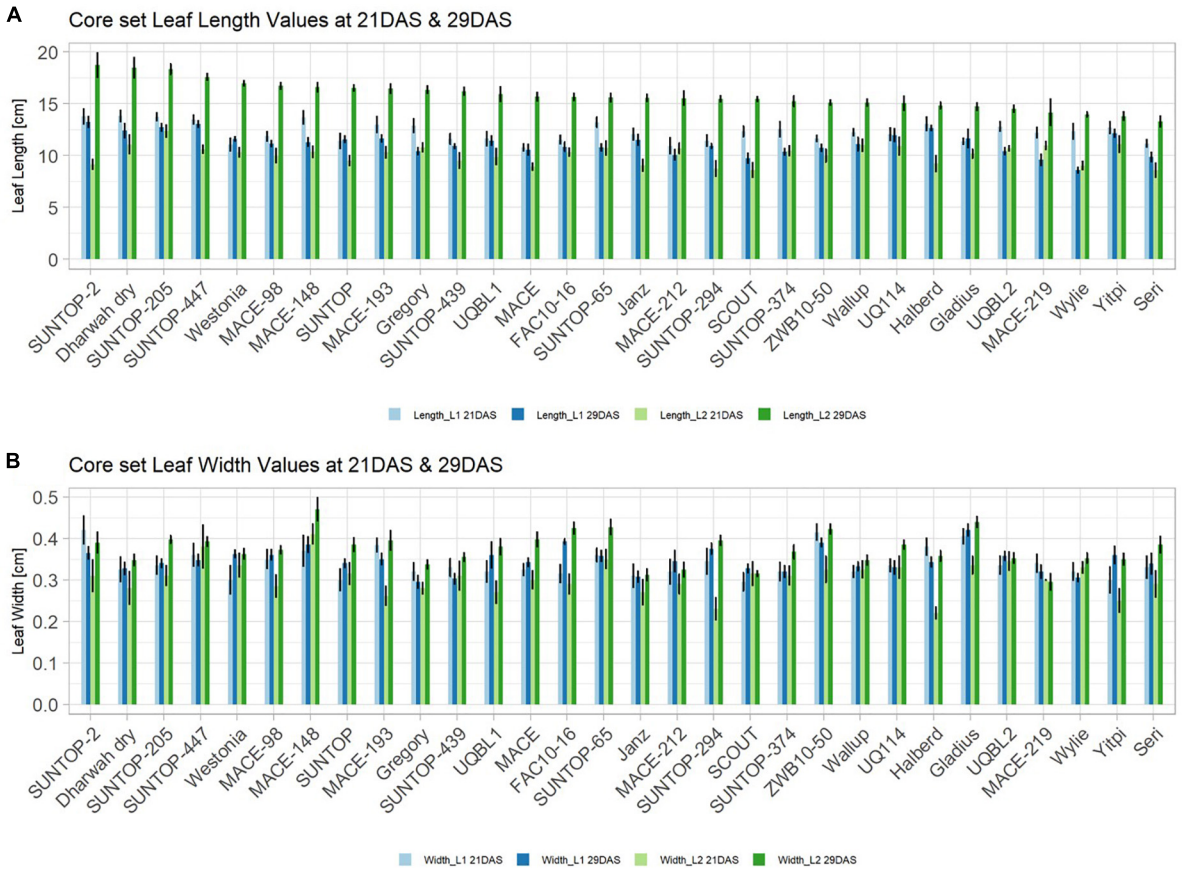

FIGURE 2 | Core set leaf parameters recorded under field conditions. **(A)** Showing leaf length values captured at 21 DAS and 29 DAS and **(B)** showing leaf width values captured at 21 DAS and 29 DAS. Error bars represent standard error.

programming language. The pixel counter identifies green pixels in an image to calculate a projected leaf area (PLA) and is a low-cost approach for image analysis (**Figure 1**). The measured PLA, using the green pixel counter, is analogous to the TLA, estimated as the sum of the areas of L1 and L2, as calculated from the length and width of each. With the default settings, images were taken using a Canon Eos 750D® camera. To reduce variation between images, the camera was mounted on a tripod at 80 cm distance from a platform on which every pot was placed. To avoid interference with other pixels or light sources, the image was taken in a closed room with consistent lighting conditions. In addition, a black background was placed behind the plant to exclude any other colour pigments from the picture, reducing them only to black and green pigments. The image analysis was conducted two times: first at 17 DAS and second at 21 DAS. At 17 DAS, images were captured for genotypes in Set 2. At 21 DAS all replications of the core set of Set 2 were imaged, as well as recorded manually. Again, the length and width of L1 and L2 were measured.

Statistical Analyses of Phenotype Data

For the calculation of the best linear unbiased estimators (BLUEs), the linear mixed model (LMM) described in Eq. 1 was used. For the calculation of the LMM, the R-language-based packages lme4[1] combined with lsmeans[2] were used.

$$P_{ijkl} = \mu + g_i + W_k + C_j + R_l + e_{ijkl} \quad (1)$$

where P_{ijkl} is the phenotypic value of the i^{th} genotype, in the k^{th} replication, μ stands for the overall mean, g_i describes the fixed effects of the i^{th} genotype. The random effects are W_k, which is the k^{th} replication, C_j, which represents the j^{th} column, and R_l representing the l^{th} row. The error term is represented by e_{ijkl}.

Pearson correlation matrices were created using the R package psych. Principal component analysis was conducted using the R package stats. For the Pearson correlation, as well as for the principal component analysis, we used the BLUEs described in Eq. 1.

Association Mapping

Association mapping was performed using the R package GenABEL (Aulchenko et al., 2007). Genome-wide rapid

[1]http://CRAN.R-project.org/package=lme4

[2]https://cran.r-project.org/web/packages/lsmeans/index.html

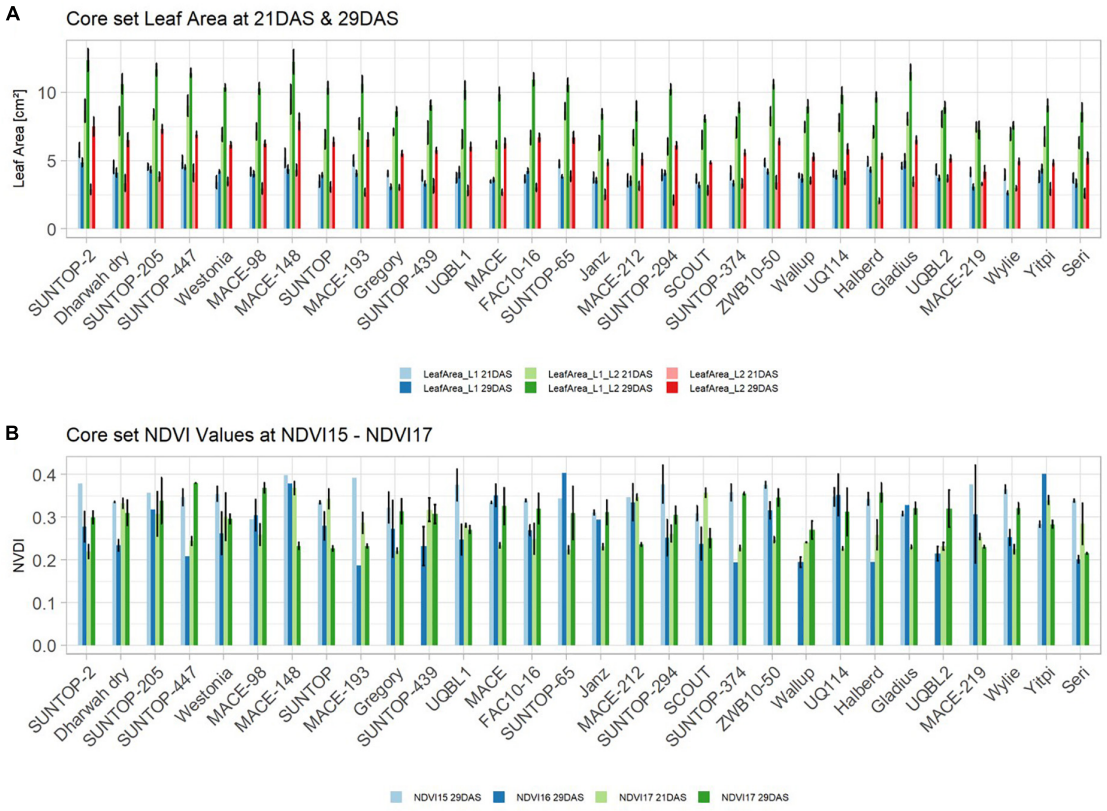

FIGURE 3 | (A) Core set leaf area parameters showing leaf area of leaf 1 (L1), leaf 2 (L2) as well as total leaf area (TLA) measured at 21 DAS and 29 DAS, respectively. **(B)** Core set NDVI parameters from NDVI15, NDVI16, and NDVI17 at 21DAS as well as NDVI17 at 29DAS. Error bars represent standard error.

association studies, using the mixed model and regression (GRAMMAR) method, initially estimate the residuals from the LMM on the assumption that the SNPs have no effect (null model). Subsequently, GRAMMAR then treats the residuals as phenotypes for further genome-wide analysis, using a standard linear mixed model (Zhou and Stephens, 2012). A total of 685 lines were genotyped using the presence/absence Diversity Arrays Technology genotyping-by-sequencing (SillicoDArTs^{TM}) platform. By applying zero mismatches and gaps, as well as a stringent alignment using BLASTN (Altschul et al., 1990), we were able to uniquely anchor 15,146 SNPs to a single position of the wheat Chinese Spring reference genome (RefSeq v1.0). The allelic association was calculated for NDVI, after accounting for population structure by implementing the first principal component, as well as the genome-wide kinship matrix for the genotypic trait values of NDVI and PLA. Prior to analysis, markers with more than 10% missing data or minor allele frequency (MAF) less than 5% were excluded from the analysis. A total number of 9,432 high-quality and polymorphic markers remained for the analysis. The cut-off value for markers being identified as significantly associated with the trait was set at the arbitrary threshold of $-\log_{10}(P) > 3$ which corresponds to a p cut-off of 0.001 significance level. The phenotypic variation explained by a given QTL (PVE) was calculated separately, according to Shim et al. (2015).

This study estimated broad sense (H^2) and narrow sense (h^2) heritability by using the R package sommer (Covarrubias-Pazaran, 2016). A marker-based approach to estimate σ^2_A, σ^2_D, by calculating additive and dominance relationship matrices, was applied. The models to estimate h^2 and H^2 are given in Eqs. 2, 3, respectively.

$$h^2 = \frac{\sigma^2_A}{\sigma^2_P} \quad (2)$$

$$H^2 = \frac{\sigma^2_A + \sigma^2_D}{\sigma^2_P} \quad (3)$$

With σ^2_A as the additive genetic variance, σ^2_D as the dominance genetic variance, and σ^2_P as the phenotypic variance.

RESULTS

Phenotypic Characterisation of Early Vigour

Expression of Early Vigour in the Field

Basic descriptive statistical indicators (minimum, maximum, and mean), variation (Var), standard deviation (SD), coefficient of

variation (CoV) for NDVI, and leaf parameters of Set 1, Set 2, and the core set are all given in **Table 2**. For Set 1, the largest mean NDVI values were observed in 2015. For Set 2 and the core set, the largest mean values were reached in 2016, while in Set 2, the overall maximum mean NDVI was recorded in 2015. All leaf parameters which were measured in the core set showed larger values at 29 DAS compared to those recorded at 21 DAS (**Table 2**). Within the core set the largest values for L1 were reached by

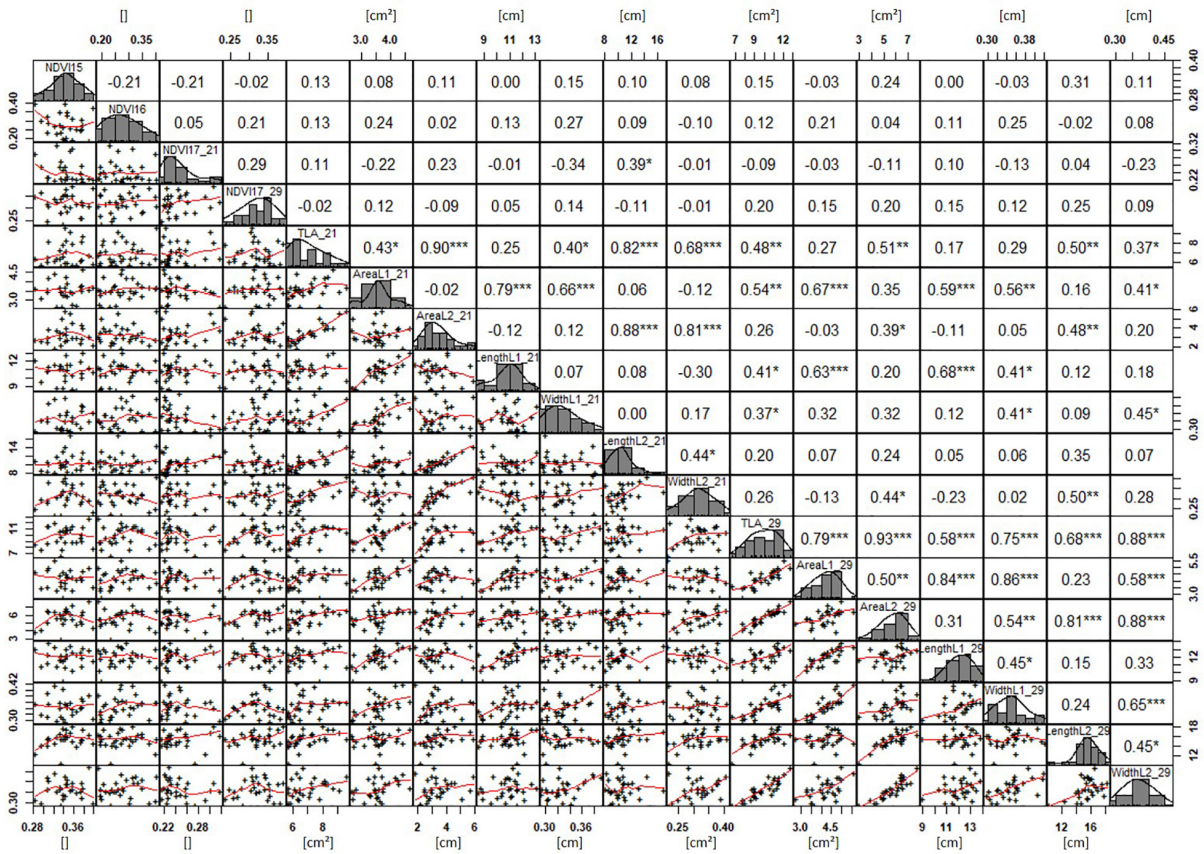

FIGURE 4 | NDVI and leaf parameters captured for the core set of 30 lines in field trials from 2015 to 2017. For each trait, the population distribution is displayed in the centre diagonal. The upper right shows the Pearson correlation coefficient for each trait combination. The lower left half shows the scatter plot with fitted line for each trait combination. * significant at $p < 0.05$; ** significant at $p < 0.005$; *** significant at $p < 0.001$.

TABLE 3 | Descriptive statistics for collected GH data for core set (30 lines) captured at 17 and 21 days after sowing (DAS).

Set	Parameter		DAS	Year	Descriptive statistics					
					Mean	Min	Max	Var	SD	CoV
Core Set	Seed weight	[g]	–	2017	0.095	0.068	0.121	0.000	0.012	0.129
31	PLA	[cm^2]	17 DAS	2017	13.895	8.236	19.681	7.740	2.913	0.210
32	PLA	[cm^2]	21 DAS	2017	14.795	10.635	20.348	4.350	2.158	0.146
33	Biomass	[g]	21 DAS	2017	0.067	0.010	0.108	0.001	0.024	0.364
34	TLA	[cm^2]	21 DAS	2017	11.254	6.915	16.493	5.659	2.518	0.224
35	Leaf area L1	[cm^2]	21 DAS	2017	3.707	1.983	5.125	0.714	0.904	0.244
36	Leaf area L2	[cm^2]	21 DAS	2017	7.548	4.600	12.203	2.933	1.799	0.238
37	Length L1	[cm]	21 DAS	2017	9.943	6.750	12.100	1.239	1.273	0.128
38	Width L1	[cm]	21 DAS	2017	0.366	0.267	0.483	0.003	0.060	0.164
39	Length L2	[cm]	21 DAS	2017	16.445	12.500	21.700	3.487	2.034	0.124
	Width L2	[cm]	21 DAS	2017	0.452	0.358	0.583	0.004	0.063	0.138

Shown are the set of lines examined (Set), parameter examined, unit, time of measurement (DAS), year of experiment as well as mean value, minimum (Min), maximum (Max), variance (Var), standard deviation (SD), and coefficient of variation (CoV).

Dharwah dry at 21 DAS and SUNTOP-2 at 29 DAS (**Figure 2A**). Furthermore, a significant ($p < 0.05$) difference between the genotypes was observed at 21 DAS and 29 DAS, with a smaller phenotypical variation at 21 DAS compared to the measurement at 29 DAS (**Supplementary Table 1**). For leaf length L2, similar results were observed with significant ($p < 0.05$) phenotypical variation at both time points, with a smaller variation at 21 DAS compared to 29 DAS (**Supplementary Table 1**). The largest leaf length L2 values were reached by SUNTOP-205 at 21 DAS and SUNTOP-2 at 29 DAS, respectively (**Figure 2B**). Regarding the leaf length of L1, the largest values were reached by SUNTOP-2 at 21 DAS and by Gladius at 29 DAS. For width L1, only at 29 DAS, a significant ($p < 0.05$) phenotypical variation could be observed (**Supplementary Table 1**). By comparing leaf area values calculated within the core set, it was observed that SUNTOP-2 showed the largest values for L1 at 21 DAS and 29 DAS, as well as for TLA at 21 DAS. For leaf area, L2 at 21 DAS, 29 DAS and TLA at 29 DAS, MACE-148 showed the largest values (**Figure 3**). For all leaf area traits which were recorded at 21 DAS, the phenotypical variation is smaller compared to the phenotypical variation at 29 DAS (**Supplementary Table 1**). In order to examine the impact of each leaf characteristic on the TLA, Pearson correlation coefficients for each factor were estimated (**Figure 4**). Person correlation analysis revealed a larger significant ($p < 0.001$) correlation between TLA at 21 DAS and leaf area L2 at 21 DAS ($r = 0.89$), than between TLA and leaf area L1 at 21 DAS ($p < 0.05$ $r = 0.42$). A similar correlation can be observed for L1 and L2 at 21 DAS, and L2 parameters at 21 DAS. Notably, a significant ($p < 0.05$) correlation was observed between NDVI 17 21 DAS and leaf length L2 at 21 DAS ($r = 0.39$). Furthermore, the relationship between the leaf area of L1 and L2 at 29 DAS and leaf parameters at 29 DAS shows similarities to the calculation made at 21 DAS. Leaf area L1 at 29 DAS shows stronger correlations with leaf TLA at 29 DAS ($r = 0.79$), compared to 21 DAS. However, the L2 area at 29 DAS ($r = 0.93$) shows a stronger correlation to TLA at 29 DAS. Interestingly, the correlation between area L1 and L2 at 21 DAS, and the L1 and L2 at 29 DAS ($r = 0.47$) is slightly smaller than the correlation with L2 area at 29 DAS ($r = 0.51$) and leaf length L2 at 29 DAS.

Expression of Early Vigour Under Controlled Conditions

In order to understand how EV can be determined under controlled conditions, this study conducted a trial testing Set 2 as well as the core set within a GH environment. Basic statistical indicators are given in **Table 3**. **Figure 5A** reveals that the leaf

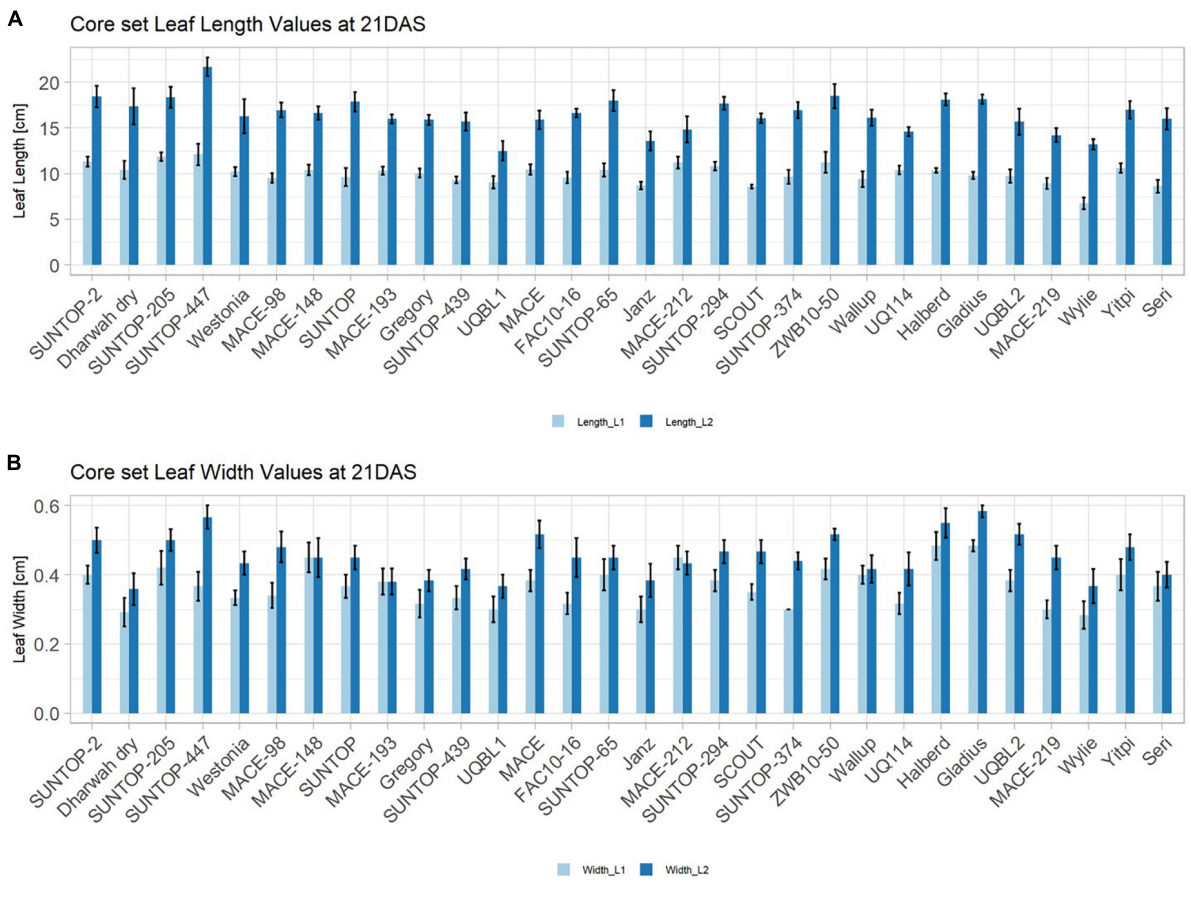

FIGURE 5 | Core set leaf parameters recorded under controlled conditions. **(A)** Showing leaf length values captured at 21 DAS and **(B)** showing leaf width values captured at 21 DAS. Error bars represent standard error.

length and leaf width of L2 exceeds L1 for almost all genotypes at 21 DAS. The only exceptions are the leaf widths of MACE-148 and of MACE-193, where L1 and L2 had similar values (**Figure 5B**). The largest leaf length values for L1 and L2 were reached by SUNTOP-447 (**Figure 5A**). As in field conditions, Gladius reached the largest values for leaf width L1 as well as for leaf width L2 (**Figure 5B**). Furthermore, for all the measured leaf parameters, significant ($p < 0.05$) differences can be observed. However, the recorded phenotypical differences show a lower variation compared to the field conditions (**Supplementary Table 2**). The calculated leaf area showed a significant ($p < 0.05$) phenotypical difference, particularly the L2, as well as the L1&L2 values (**Supplementary Table 2**). The SUNTOP-447 exhibited the largest values for leaf length, leaf area L2, and L1&L2, respectively. Maximum leaf area and L1 were exhibited by MACE-212 (**Figure 6A**). Only around two-thirds of the core set showed increased PLA at 21, compared to 17 DAS (**Figure 6B**). Pearson coefficients of correlation showed a significant positive relationship of the collected dry matter with the PLA 17 DAS ($p < 0.05$; $r = 0.46$) and with PLA 21 DAS ($p < 0.001$; $r = 0.71$). Furthermore, the PLA at 21 DAS shows a positive correlation to leaf width L2 ($r = 0.34$), as well as to area L2 ($r = 0.38$) and TLA ($r = 0.39$). As under field conditions, the TLA of L1 and L2 is more affected by area L2 ($r = 0.96$). However, area L1 ($r = 0.85$) seems to have a greater impact on TLA in the GH compared to the field. Furthermore, it is noteworthy that dry mass measured at harvest at 21 DAS shows a positive correlation to all captured leaf characteristics. Moreover, seed weight (SW) showed no significant correlation to any other trait (**Figure 7**).

Components of Early Vigour

To determine which physiological parameters of the leaf had the greatest influence on EV, a principal component (PC) analysis was performed. The analysis was conducted for the core set data and included parameters such as the leaf measurements at 21 DAS and at 29 DAS in the field, or at 21 DAS in the GH, as well as dry matter content, grain weight, and the respective NDVI values from 2015 to 2017 (**Figure 8**). For the field data, the NDVI in 2015 (NDVI 15) was largely associated with parameters connected to L2 at 21 DAS, and to leaf length and area of L2 at 29 DAS. The NDVI 16 and NDVI 17 at 29 DAS showed a strong association with area L1 at 21 DAS and area L1 at 29 DAS. In this regard, it became apparent that several parameters recorded at 29 DAS, such as area L1 and L2 and leaf width L2, were strongly correlated as well. Interestingly, for NDVI at 21 DAS, no association to any physiological parameters of the leaf was observed. In the GH, it could be observed that area L1 and L2 were slightly and closely more associated with area L2 than with area L1. Furthermore,

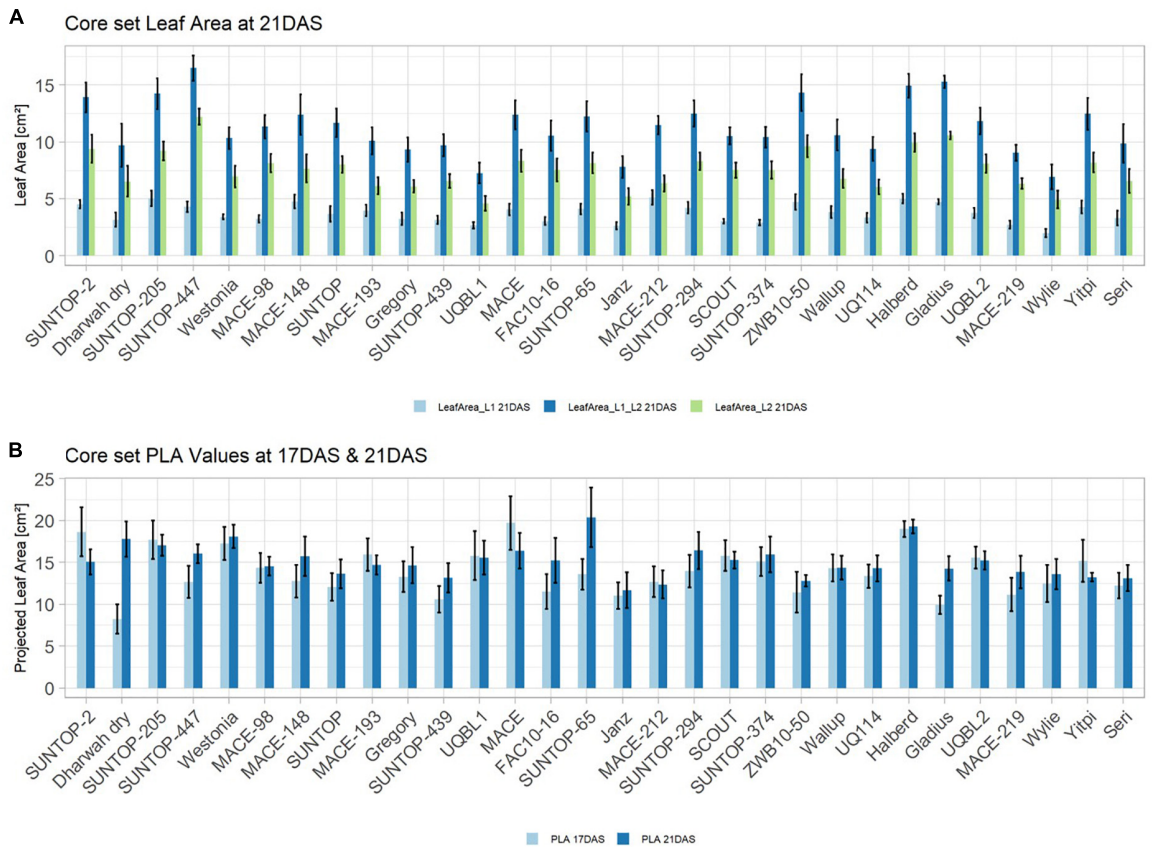

FIGURE 6 | (A) Core set leaf area parameters showing leaf area of leaf 1 (L1), leaf 2 (L2) as well as total leaf area (TLA) measured at 21 DAS in greenhouse experiment in 2017. **(B)** Core set projected leaf parameters (PLA) recorded at 17DAS and 21DAS. Error bars represent standard error.

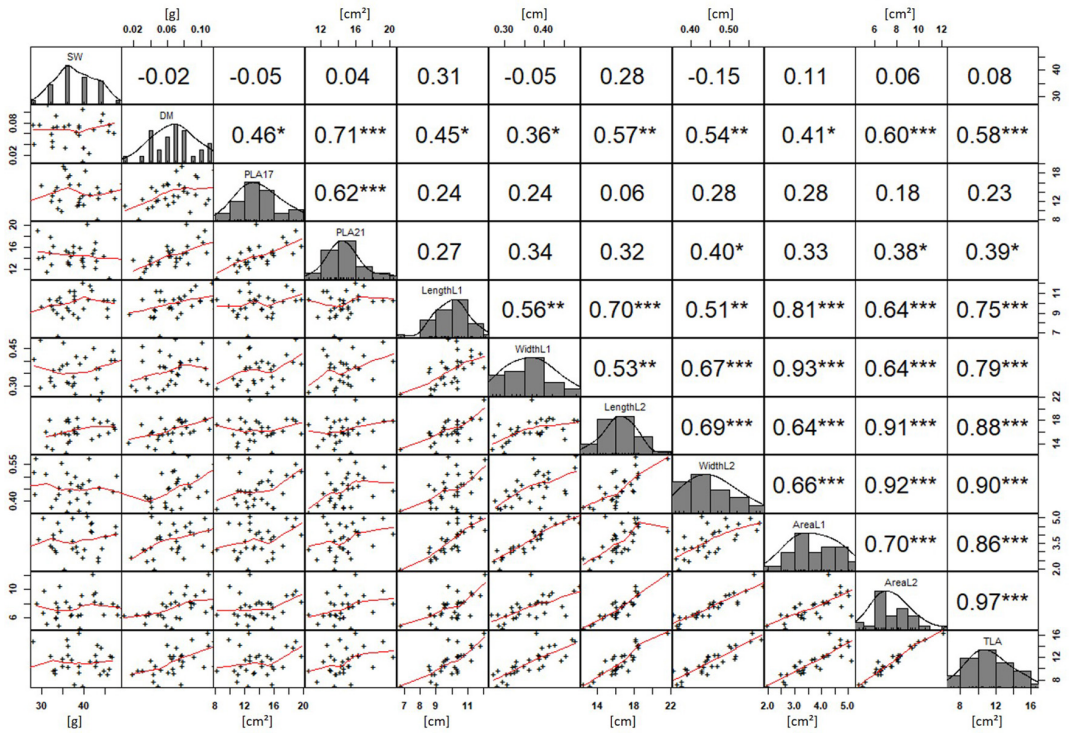

FIGURE 7 | Projected leaf area (PLA) by green pixel count method and leaf parameters captured for core set in greenhouse experiment. For each trait, the population distribution is displayed in the centre diagonal. The upper right shows the Pearson correlation coefficient for each trait combination. The lower left half shows the scatter plot with fitted line for each trait combination. * significant at $p < 0.05$; ** significant at $p < 0.005$; *** significant at $p < 0.001$.

area L2 appeared to be in a stronger association to leaf width L2 than to leaf length L2. The PLA at 17 DAS and PLA at 29 DAS showed a strong association to each other. However, apart from dry mass, no significant associations to any of the leaf parameters were observed.

Identifying Genetic Determinants of Early Vigour

Genetic analysis was performed with NDVI values for the 685 genotypes from Set 1 in 2015 (NDVI'15) and in 2016 (NDVI'16), along with NDVI values from 2015 to 2017, and the PLA data from the GH trial in 2017 in the 221 genotypes of Set 2. Several SNP markers exceeded the arbitrary threshold of association ($-\log_{10}(P) = 3$) for NDVI and PLA across all years in the genome-wide association study. A total of 41 QTL were associated [$-\log_{10}(P) = 3$] with either NDVI or PLA. The majority of trait-associated SNP markers (21) were associated with NDVI at 21 DAS in 2017 in Set 2 (**Figure 9**). Chromosome positions of all identified marker-trait associations for NDVI and PLA in the different environments are summarized in **Table 4**. All the identified QTL PVE was low and was ranged between 0.024 to 1.35%. The five most significant QTL, *QSG.qwr-3B.1*, *QSG.qwr-2A.3*, *QSG.qwr-3D.1*, *QSG.qwr-1A.1*, and *QSG.qwr-5B.2* accounted for 0.4% of the variation (**Table 4**). Most QTL effects were small and few, as QTL were detected in multiple environments or across traits, reflecting the genetic complexity and the strong environmental dependency of EV traits. This finding is consistent with several other studies that have also identified multiple QTL for several EV related traits located on different chromosomes (**Table 5**). In accordance with this observation, the narrow-sense heritability of both NDVI in Set 1 was found to be moderate with $h^2 = 0.22$–0.28, and small for NDVI and PLA in Set 2 with $h^2 = 0.08$–0.04 (**Table 6**).

DISCUSSION

The first objective of this study was to provide information on physiological components that contribute to EV of wheat, both in the field and under controlled conditions. Several studies observed a strong contribution of leaf width to a specific leaf area, and consequently to EV (Rebetzke and Richards, 1999; Richards and Lukacs, 2002; Maydup et al., 2012). Our results showed that the variation in EV is strongly associated with the leaf length of both L1 and L2 in the field, as well as in the GH. For both leaves (L1 and L2), leaf length is the main contributor to a larger leaf area development, and, consequently, to an increased EV. Furthermore, leaf length of L2 showed a slightly greater impact on the area of L1 and L2 with an advancing development. This finding agrees with other studies reporting a significant positive correlation between the area of L2 and the leaf length (Nursinow et al., 2011; Boden et al., 2014; Moore and Rebetzke, 2015; Duan et al., 2016). The data revealed faster development of L2 in the

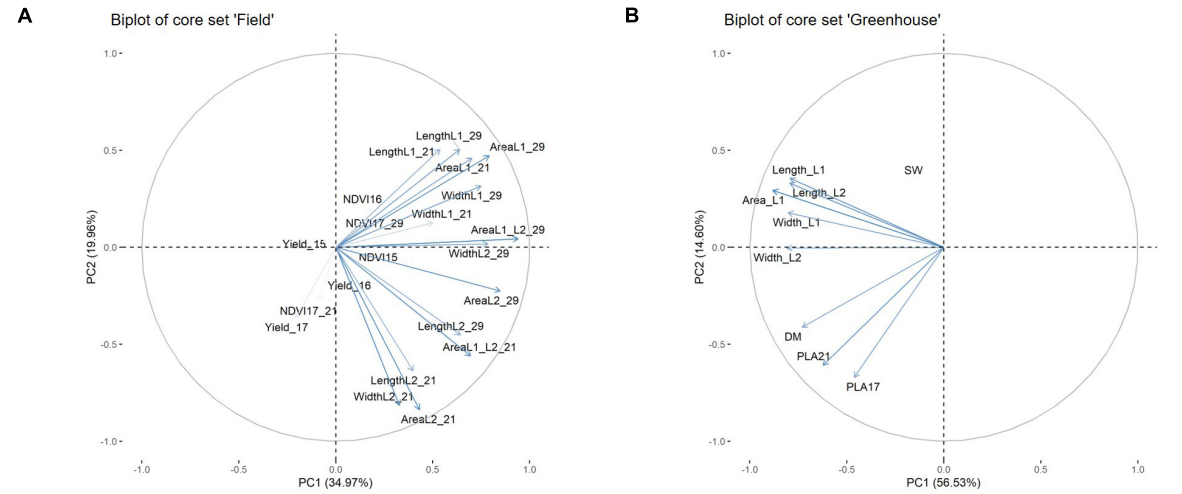

FIGURE 8 | Biplot from principal component analysis for core set field **(A)**, core set greenhouse **(B)**.

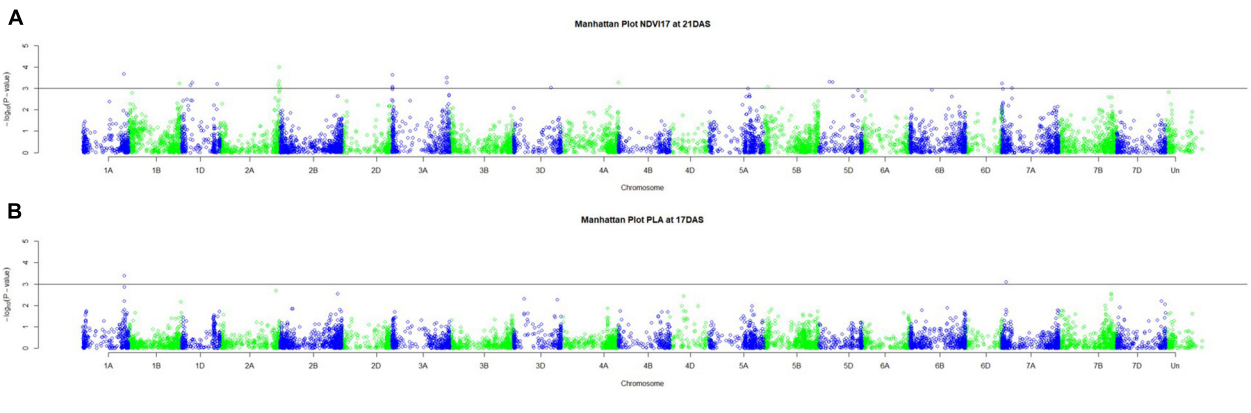

FIGURE 9 | Manhattan plots for **(A)** NDVI17 at 21DAS and **(B)** PLA at 17DAS. The horizontal line at $-\log_{10}(P) = 3$ indicates the chosen probability threshold to indicate a significant association with the trait.

GH compared to the field studies. This was also observed by Rebetzke et al. (2007) and leads to the conclusion that EV can be recorded at earlier stages under controlled conditions than in the field. Previous studies have reported that embryo size is a highly heritable trait that is strongly associated with leaf area (López-Castañeda et al., 1996; Rebetzke and Richards, 1999; Aparicio et al., 2002; Moore and Rebetzke, 2015). Since it has been established that embryo size increases with seed weight in wheat (Moore and Rebetzke, 2015) and barley (López-Castañeda et al., 1996), this study used seed weight to indirectly evaluate the impact of embryo size on EV. In the GH experiment, it was not possible to demonstrate the positive impact of embryo size on EV by using seed weight, since no significant correlation with any measured leaf parameter was observed. Nonetheless, several studies reported a positive effect of the embryo size on early vigour in wheat (Rebetzke and Richards, 1999; Richards and Lukacs, 2002; Moore and Rebetzke, 2015). In other studies, however, the total variation in EV could not be exclusively explained by the considering seed weight (Maydup et al., 2012), and the seed density has also been suggested as a potentially more decisive factor in determining EV (Ball et al., 2011). This partially explains why no correlation between seed weight and EV parameters could be found in the present study. In the GH experiment, biomass was recorded and was exhibited as a significant correlation with the area of L2, as well as area of L1 and L2, suggesting an increased above-ground biomass for genotypes with greater EV. Richard et al. (2019) reported that lines with increased above-ground biomass were strongly associated with an increased grain yield.

The second objective of this study was to evaluate high-throughput methods to precisely record physiological leaf characteristics in field and greenhouse trials. Mullan and Reynolds (2010) reported a significant correlation between NDVI, leaf area, and biomass. These PCA results from the current study tended to confirm that NDVI and the leaf area are related. Although the biomass in the field was not separately measured, a significant positive correlation between biomass and leaf area in the GH experiment suggested a putative relationship. In the

TABLE 4 | QTL identified with significant association to phenotypic traits (*P*-value > −log10 (*P*) = 3).

QTL name	Chr	Position [bp]	Trait	DAS	Year	*P*-value	PVE	No. of reported genes in ± 20kb	Gene-ID	Start position [bp]	End position [bp]
QSG.qwr-1A.1	1A	530870345	NDVI	21	2017	3.68	0.135	1	TraesCS1A02G341200	530883540	530884786
QSG.qwr-1A.2	1A	537895686	PLA	17	2017	3.39	0.125	–	NA	NA	NA
QSG.qwr-2A.1	2A	758570584	NDVI	21	2017	3.17	0.114	3	TraesCS2A02G552800	758553979	758559790
									TraesCS2A02G552900	758583247	758587695
									TraesCS2A02G553000	758588329	758589199
QSG.qwr-2A.2	2A	768989573	NDVI	21	2017	3.34	0.120	3	TraesCS2A02G573200	768981429	768983083
									TraesCS2A02G573300	768996829	768998744
									TraesCS2A02G573400	769002934	769004935
QSG.qwr-2A.3	2A	769344238	NDVI	21	2017	4	0.152	1	TraesCS2A02G574300	769340976	769344226
QSG.qwr-2A.4	2A	775175704	NDVI	21	2017	3.06	0.113	1	TraesCS2A02G583000	775165226	775166200
QSG.qwr-3A.1	3A	8267133	NDVI	21	2017	3.08	0.107	2	TraesCS3A02G008800	8253788	8255463
									TraesCS3A02G008900	8258758	8260510
QSG.qwr-3A.2	3A	9212918	NDVI	21	2017	3.64	0.134	1	TraesCS3A02G011400	9210008	9210976
QSG.qwr-3A.3	3A	9463808	NDVI	21	2017	3.03	0.107	3	TraesCS3A02G011800	9444895	9445740
									TraesCS3A02G011900	9473058	9473828
									TraesCS3A02G012000	9480858	9481594
QSG.qwr-3A.4	3A	711365648	NDVI	21	2017	3.5	0.122	1	TraesCS3A02G480500	711380680	711385803
	3A	711366350	NDVI	21	2017	3.27		–	NA	NA	NA
QSG.qwr-4A.1	4A	744471786	NDVI	21	2017	3.28	0.119	1	TraesCS4A02G499600	744456472	744458084
QSG.qwr-5A.1	5A	526386152	NDVI	29	2015	3.14	0.115	5	ENSRNA050011196	526378014	526378087
									ENSRNA050021751	526384607	526384709
									TraesCS5A02G315700	526393365	526396850
									TraesCS5A02G315800	526396904	526400499
									TraesCS5A02G315900	526400750	526402134
QSG.qwr-5A.2	5A	529810097	NDVI	29	2015	3.54	0.025	–	NA	NA	NA
QSG.qwr-5A.3	5A	592515889	NDVI	29	2015	3.07	0.030	1	TraesCS5A02G398700	592518737	592520426
QSG.qwr-6A.1	6A	5480626	NDVI	29	2015	3.14		1	TraesCS6A02G011800	5473439	5480862
QSG.qwr-7A.1	7A	263934	NDVI	21	2017	3.24	0.112	2	TraesCS7A02G000200	242359	246258
									TraesCS7A02G000300	250565	253987
QSG.qwr-7A.2	7A	51498553	PLA	17	2017	3.1	0.112	–	NA	NA	NA
QSG.qwr-7A.3	7A	130596094	NDVI	21	2017	3.03	0.106	1	TraesCS7A02G177400	130608975	130613414
QSG.qwr-7A.4	7A	694049656	NDVI	29	2016	3.63	0.040	2	TraesCS7A02G506800	694053373	694053943
									TraesCS7A02G506900	694060394	694067693
QSG.qwr-1B.1	1B	37762680	NDVI	29	2016	3.12	0.037	–	NA	NA	NA
QSG.qwr-1B.2	1B	84949441	NDVI	29	2016	3.47	0.039	–	NA	NA	NA
QSG.qwr-1B.3	1B	658902240	NDVI	21	2017	3.24	0.113	3	TraesCS1B02G434300	658908133	658910735
									TraesCS1B02G434400	658911233	658914898
									TraesCS1B02G434500	658915051	658919932
QSG.qwr-3B.1	3B	22137315	NDVI	29	2016	4.24	0.051	1	TraesCS3B02G042800	22129256	22130588
QSG.qwr-5B.1	5B	52306024	NDVI	21	2017	3.06	0.114	1	TraesCS5B02G046800	52317367	52318139
QSG.qwr-5B.2	5B	162778081	NDVI	29	2015	3.64	0.030	1	TraesCS5B02G110800	162771499	162776936
QSG.qwr-5B.3	5B	576298790	NDVI	29	2015	3	0.024	–	NA	NA	NA
QSG.qwr-7B.1	7B	748033995	NDVI	29	2015	3.27	0.026	1	TraesCS7B02G497400	748012557	748015135
QSG.qwr-1D.1	1D	52680610	NDVI	29	2016	3.2	0.035	1	TraesCS1D02G072300	52658816	52664518
QSG.qwr-1D.2	1D	111980607	NDVI	21	2017	3.15	0.112	1	TraesCS1D02G116200	111967743	111980590
QSG.qwr-1D.3	1D	134460288	NDVI	21	2017	3.28	0.120	–	NA	NA	NA
QSG.qwr-1D.4	1D	461051177	NDVI	21	2017	3.21	0.112	1	TraesCS1D02G389300	461052237	461057083
QSG.qwr-2D.1	2D	19551003	NDVI	29	2016	3.19	0.035	3	TraesCS2D02G051300	19543541	19544843
									TraesCS2D02G051400	19555735	19557093
									TraesCS2D02G051500	19561151	19563368
QSG.qwr-2D.2	2D	69503900	NDVI	29	2017	3.05	0.100	2	TraesCS2D02G120100	69502005	69504329
									TraesCS2D02G120200	69504557	69510365
QSG.qwr-2D.3	2D	362486328	NDVI	29	2017	3.11	0.108	–	NA	NA	NA
QSG.qwr-3D.1	3D	325183855	NDVI	29	2015	3.76	0.032	–	NA	NA	NA
QSG.qwr-3D.2	3D	481926113	NDVI	21	2017	3.04	0.113	1	TraesCS3D02G368700	481908948	481926630
QSG.qwr-5D.1	5D	138341788	NDVI	21	2017	3.31	0.115	–	NA	NA	NA
QSG.qwr-5D.2	5D	179786153	NDVI	21	2017	3.3	0.115	1	TraesCS5D02G123000	179762666	179778583
QSG.qwr-6D.1	6D	50627689	NDVI	29	2015	3.35	0.027	–	NA	NA	NA
QSG.qwr-7D.1	7D	533491186	NDVI	29	2015	3.11	0.026	–	NA	NA	NA

Shown are QTL identifier, chromosome location (Chr), Positions refer to physical positions [bp] on reference genome Chinese Spring (RefSeq v1.0), trait, time of measurement (DAS), year of measurement (Year), probability of association with trait by chance [−log10 (p-value)], phenotypic variation explained by a given QTL (PVE), number of genes with 20 kb of the QTL SNP location (No. of genes ± 20 kb), name of genes reported within the 20 kb radius of the QTL SNP (Gene-ID), start and end position of reported gene in [bp] (Start Position [bp]/ End Position [bp]).

TABLE 5 | Summary of early vigour related QTL reported in previous publications.

Trait	QTL/marker name	Chr	Publication
Coleoptile length	QClp.ipk-1A	1A	Landjeva et al., 2008
Coleoptile length	QClp.ipk-1B	1B	
Coleoptile length	ksuG9c	1A	Rebetzke et al., 2007
Coleoptile length	Stm55ltgag	2D	
Coleoptile length	psr426	5A	
Coleoptile length	psr326b	5D	
Embryo size	gwm18	1B	Moore and Rebetzke, 2015
EV, canopy temperature	41	3B	Bennet et al., 2012
Ground cover	QGCw.caas-1A.1	1A	Li et al., 2014
Ground cover	QGCw-caas-1D	1D	
Ground cover	QGCs-caas-2A.2	2A	
Ground cover	QGCs.caas-3B.1	3B	
Ground cover	QGCw-caas-5B	5B	
Ground cover	QGCw.caas-5B	5B	
Ground cover	QGCwcaas-5D	5D	
Ground cover	QGCscaas-6A	6A	
Ground cover	QGCs-caas-6A	6A	
Leaf length	gwm261	2D	Moore and Rebetzke, 2015
Leaf length	cdo669b	4B	
Leaf length	E36/M60-210-P1	2D	Steege et al., 2005
Leaf length	E48/M48-217-P2	5D	
Leaf length	E48/M60-225-P1	6D	
Leaf length	E45/M52-150-P1	7D	
Leaf width	wmc190	2D	Moore and Rebetzke, 2015
Leaf width	wmc289	5B	
Leaf width	E45/M52-274-P1	1D	Steege et al., 2005
Leaf width	Xgwm458	1D	
Leaf width	E42/M51-482-P2	2D	
Leaf width	Xgwm165	4D	
Leaf width	E42/M52-241-P1	5D	
Leaf width	E51/M52-189-P1	7D	
NDVI	QNDVIs-caas-3A	3A	Li et al., 2014
NDVI	QNDVIw-caas-6D	6D	
NDVI	QYld.aww-1B.2	1B	Tura et al., 2020
NDVI	QTgw.aww-1B	1B	
Relative growth rate	QRgr.saas-5A	5A	Li et al., 2017
Root dry weight	QRdw.saas-5A	5A	
Root length	QRlp.ipk-1A	1A	Landjeva et al., 2008
Root length	QRlp.ipk-7D	7D	
Shoot biomass	Rht-B1	4B	Ryan et al., 2015
Shoot biomass	Rht-D1	4D	
Shoot biomass	wmc525	7A	
Shoot fresh weight	QSfw.saas-5A	5A	Li et al., 2017
Shoot dry weight	QSdw.saas-5A	5A	
Total leaf area	QTla.saas-5A	5A	

Reported trait (Trait), name of the respective QTL or marker that is associated with the trait (QTL/Marker Name), chromosome on which the QTL or marker is located (Chr), cited publication which reported QTL (Publication).

TABLE 6 | Narrow-sense (h2) and broad-sense heritability for NDVI and PLA recorded in set 1 (685 lines) and set 2 (210 lines).

Set	Trait	DAS	Year	Heritability	
				h^2	H^2
Set 1	NDVI	29	2015	0.22	0.28
	NDVI	29	2016	0.28	0.38
Set 2	NDVI	29	2015	0.05	0.06
	NDVI	29	2016	0.04	0.05
	NDVI	21	2017	0.07	0.08
	NDVI	29	2017	0.06	0.07
	PLA	17	2017	0.04	0.04

Trait, time of measurement (DAS), year of measurement (Year), and narrow sense heritability (h^2) as well as broad sense heritability (H^2).

correlation analysis, only one significant correlation between NDVI and leaf parameters could be established, which was NDVI17 at 21 DAS and leaf length L2 at 21 DAS. This confirms that leaf length L2 contributes more to an increased EV compared to leaf dimensions of L1. This is also supported by other studies (López-Castañeda et al., 1996; Richards and Lukacs, 2002; Duan et al., 2016). In terms of high-throughput phenotyping methods for EV in GH environments, this study tested the ability of a green pixel counter as a low-cost method. Since embryo size strongly affects EV, the method is not able to explain the trait completely. However, the results for PLA calculated by the green pixel counter indicate great potential to measure EV under controlled conditions. In particular, the leaf parameters of L2 showed a significant correlation with PLA, as well as with the biomass. Furthermore, the green pixel counter was successfully used to measure coleoptile tiller length, a trait which strongly affects EV and is also highly correlated to embryo size.

The third aim of this study was to achieve a better understanding of EV genetics in wheat. The GWAS for Set 1 and Set 2 revealed 41 SNP markers for NDVI and for PLA, which were linked to 60 protein-coding regions across 17 chromosomes. Consistent with previous studies, the present study shows that EV in wheat is a quantitative trait with numerous QTL located across several chromosomes (Li et al., 2014; Moore and Rebetzke, 2015; Boudiar et al., 2016). Numerous studies have reported the effect of dwarfing genes on coleoptile length (Rebetzke et al., 2007; Li et al., 2017), coleoptile width (Rebetzke et al., 2014), and leaf epidermal cell dimension (Botwright et al., 2005). In most of these studies, the influence of dwarfing genes was highlighted. The gibberellic acid (GA)-insensitive dwarfing genes *Rht B1b on* chromosome 4B and *Rht D1db* on chromosome 4D have been reported to reduce coleoptile length and, consequently, EV, since they decrease epidermal cell length in leaf tissue (Ellis et al., 2004; Rebetzke et al., 2007; Yu and Bai, 2010; Li et al., 2017; Rebetzke et al., 2017). In our study, no SNP markers were detected on either one of the 4B or 4D chromosomes. This suggests that NDVI and PLA are the less affected traits by the presence of *Rht B1b* and of *Rht D1b*. Comparable results were reported in Li et al. (2014), were none of the parental lines that carried *Rht-B1b*, and only one parental line contained *Rht-D1b*, while no QTL was identified on chromosome 4D. However, a significant correlation could be observed between *Rht-D1* and NDVI and the ground cover in certain environments. The *Rht 8*, on the short arm of chromosome 2D, is a GA-responsive dwarfing gene reported to have a secondary effect of reducing epidermal cell length. Hence, it is more appropriate for achieving

good canopy cover in combination with a semi-dwarf growth habit (Botwright et al., 2005). Chai et al. (2019) reported the WRKY transcription factor *TraesCS2D01G051500* as a possible candidate gene for *Rht8*. We identified QTL *QSG.qwr-2D.1* in the vicinity of *TraesCS2D01G051500*, and found this QTL to be significantly associated with NDVI17 at 29 DAS. Several studies have identified QTL associated with coleoptile length on chromosome 1B, including markers *XpGTG-mTCGA294* (Yu and Bai, 2010) and *wsnp_CAP11_c2596_1325540* (Ma et al., 2020). This study confirms that chromosome 1B is a region of interest for EV, since three significant QTL were detected on this chromosome. That applies particularly to *QSG.qwr-1B*, which is located at the same region on the long arm region of chromosome 1B as *wsnp_CAP11_c2596_1325540*, as reported in Ma et al. (2020). Another region of interest is chromosome 5A, which is also considered as the most important chromosome for stay-green traits (Shi et al., 2017; Liu et al., 2019), including the isopentenyl transferase gene (Gan and Amasino, 1995). Furthermore, chromosome 5A harbours several major developmental genes, such as the vernalisation gene *Vrn1*, frost resistance gene *Fr1,* as well as genes for ear emergence time and for the plant height (Sutka and Snape, 1989; Kato et al., 1999; Galiba et al., 1995).

The results of the current phenotypic investigation extended our insights into the EV trait in wheat. The key characteristics of EV and the relationship with other traits, such as biomass, were successfully identified. In addition, this study presents effective methods that can be used to detect EV in the field, as well as under controlled conditions. In particular, the connection between the leaf length parameters and the NDVI highlights the great potential of NDVI, especially if given recent advances in unmanned aerial vehicle or drone phenotyping platforms (Shi et al., 2016). Nevertheless, potential interactions due to environmental factors must be clarified by practical crop management for a better understanding, since factors such as sowing time, sowing depth, sowing rate, and row spacing may also influence EV. In particular, the interaction and value of EV in specific target environments must be clarified. Furthermore, it has to be considered that yield predictions based on EV can be very challenging. Since the trait is recorded at a very early developmental stage, its relationship to yield performance can subsequently be influenced by a multitude of complex environmental factors. For example, abiotic stress factors have a particularly decisive influence on yield and its yield components, especially during key developmental stages, such as tillering and flowering. Nevertheless, EV is essential for good crop establishment and, therefore, can impact yield even at this early stage. Hence, we suggest incorporating EV measurements into experiments by using the NDVI data in performance evaluations, such as stay-green trials. Our genomic analysis has identified QTL that is associated with EV, which are co-located or are closely linked to key genes controlling the plant development, such as plant height, coleoptile length, stay-green, and vernalisation. The results support the theory that EV is a trait regulated by pleiotropic genes. These findings may help identify the key drivers and determine potential trade-offs with important agronomic traits. Given that the trait is underpinned by many QTL with small effects, marker-assisted selection or gene-based approaches are likely to be challenging; however, genomic prediction approaches provide a suitable option for future breeding.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

JC and LTH conceived the study. SV, SA, and JC performed the experiments. SV and AS analysed the phenotypic data. SV wrote the manuscript with further input from AS, RJS, JC, SA, and LTH. All authors contributed to the article and approved the submitted version.

FUNDING

JC and LTH received funding from the Grains Research and Development Corporation of Australia (project code UQ00068). LTH was supported by an ARC Early Career Discovery Research Award (project code DE170101296). SV received a PROMOS scholarship funded by the German Academic Exchange Service. Further, the authors acknowledge funding from the German Federal Ministry of Food and Agriculture (BMEL grant 28-1-B2.014-16/WinEffizient).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.754439/full#supplementary-material

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Aparicio, N., Villegas, D., Araus, J., Blanco, R., and Royo, C. (2002). Seedling development and biomass as affected by seed size and morphology in durum wheat. *J. Agric. Sci.* 139, 143–150. doi: 10.1017/s0021859602002411

Asseng, S., and van Herwaarden, A. F. (2003). Analysis of the benefits to wheat yield from assimilates stored prior to grain filling in a range of environments. *Plant Soil* 256, 217–229. doi: 10.1023/A:1026231904221

Asseng, S., Ewert, F., Martre, P., Rotter, R. P., Lobell, D. B., Cammarano, D., et al. (2015). Rising temperatures reduce global wheat production. *Nat. Clim. Change* 4, 143–147.

Asseng, S., Ewert, F., Rosenzweig, C., Jones, J. W., Hatfield, J. L., Ruane, A. C., et al. (2013). Uncertainty in simulating wheat yields under climate change. *Nat. Clim. Change* 3, 827–832. doi: 10.1038/nclimate1916

Aulchenko, Y. S., de Koning, D.-J., and Haley, C. (2007). Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* 177, 577–585. doi: 10.1534/genetics.107.075614

Ball, B., Meharry, D., Acuña, T. L. B., Sharma, D. L., Hamza, M., and Wade, L. J. (2011). Increases in seed density can improve plant stand and increase seedling vigour from small seeds of wheat (*Triticum aestivum*). *Ex. Agric.* 47, 445–457. doi: 10.1017/S0014479710001006

Bennet, D., Reynolds, M., Mullan, D., Izanloo, A., Kuchel, H., Langridge, P., et al. (2012). Detection of two major grain yield QTL in bread wheat (*Triticum aestivum* L.) under heat, drought and high yield potential environments. *Theor. Appl. Genet.* 125, 1473–1485. doi: 10.1007/s00122-012-1927-2

Bertholsson, N. O. (2005). Early vigour and allelopathy – two useful traits for enhanced barley and wheat competitiveness against weeds. *Weed Res.* 45, 94–102. doi: 10.1111/j.1365-3180.2004.00442.x

Boden, S., Weiss, D., Ross, J., Davies, N., Trevaskis, B., Chandler, P., et al. (2014). Early flowering3 regulates flowering in spring barley by mediating gibberellin production and flowering locus t expression. *Plant Cell* 26, 1557–1569. doi: 10.1105/tpc.114.123794

Botwright, T. L., Condon, A. G., Rebetzke, G. J., and Richards, R. A. (2002). Field evaluation of early vigour for genetic improvement of grain yield in wheat. *Aust. J. Agric. Res.* 53, 1137–1145. doi: 10.1071/AR02007

Botwright, T. L., Rebetzke, G. J., Condon, A. G., and Richards, R. A. (2005). Influence of the gibberellin-sensitive Rht8 dwarfing gene on leaf epidermal cell dimensions and early vigour in wheat (*Triticum aestivum* L.). *Ann. Bot.* 95, 631–639. doi: 10.1093/aob/mci069

Boudiar, R., Casas, A. M., Cantalapiedra, C. P., Gracia, M. P., and Igartua, E. (2016). Identification of quantitative trait loci for agronomic traits contributed by a barley (*Hordeum vulgare*) Mediterranean landrace. *Crop Pasture Sci.* 67, 37–46. doi: 10.1071/CP15149

Chai, L., Chen, Z., Bian, R., Zhai, H., Cheng, X., Peng, H., et al. (2019). Dissection of two quantitative trait loci with pleiotropic effects on plant height and spike length linked in coupling phase on the short arm of chromosome 2D of common wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* 132:3225. doi: 10.1007/s00122-019-03318-z

Christopher, J. T., Veyradier, M., Borrell, A. K., Harvey, G., Fletcher, S., and Chenu, K. (2014). Phenotyping novel stay-green traits to capture genetic variation in senescence dynamics. *Funct. Plant Biol.* 41, 1035–1048. doi: 10.1071/FP14052

Christopher, M., Paccapelo, V., Kelly, A., Macdonald, B., Hickey, L., Richard, C., et al. (2021). QTL identified for stay-green in a multi-reference nested association mapping population of wheat exhibit context dependent expression and parent-specific alleles. *Field Crops Res.* 270:108181. doi: 10.1016/j.fcr.2021.108181

Clarke, J. M., Richards, R. A., and Condon, A. G. (1991). Effect of drought stress on residual transpiration and its relationship with water use of wheat. *Can. J. Plant Sci.* 71, 695–702. doi: 10.4141/cjps91-102

Coleman, R. K., Gill, G. S., and Rebetzke, G. J. (2001). Identification of quantitative trait loci for traits conferring weed competitiveness in wheat (*Triticum aestivum* L.). *Aust. J. Agric. Res.* 52:e0228775. doi: 10.1071/AR01055

Condon, A. G., Richards, R. A., Rebetzke, G. J., and Farquhar, G. D. (2004). Breeding for high water-use efficiency. *J. Exp. Bot.* 55:407. doi: 10.1093/jxb/erh277

Covarrubias-Pazaran, G. (2016). Genome-assisted prediction of quantitative traits using the r package sommer. *PLoS One* 11:e0156744. doi: 10.1371/journal.pone.0156744

Dijk, A. I., Beck, H. E., Crosbie, R. S., Jeu, R. A., Liu, Y. Y., Podger, G. M., et al. (2013). The millennium drought in southeast Australia (2001–2009): natural and human causes and implications for water resources, ecosystems, economy, and society. *Water Resour. Res.* 49, 1040–1057. doi: 10.1002/wrcr.20123

Dingkuhn, M., Johnson, D. E., Sow, A., and Audebert, A. Y. (1999). Relationships between upland rice canopy characteristics and weed competitiveness. *Field Crops Res.* 61, 79–95. doi: 10.1016/S0378-4290(98)00152-X

Duan, T., Chapman, S. C., Holland, E., Rebetzke, G. J., Guo, Y., and Zheng, B. (2016). Dynamic quantification of canopy structure to characterize early plant vigour in wheat genotypes. *J. Exp. Bot.* 67, 4523–4534. doi: 10.1093/jxb/erw227

Ellis, M. H., Rebetzke, G. J., Chandler, P., Bonnett, D., Spielmeyer, W., and Richards, R. A. (2004). The effect of different height reducing genes on the early growth of wheat. *Funct. Plant Biol.* 31, 583–589. doi: 10.1071/FP03207

Galiba, G., Quarrie, S. A., Sutka, J., Morgounov, A., and Snape, J. W. (1995). RFLP mapping of the vernalization (Vrn1) and frost resistance (Fr1) genes on chromosome 5A of wheat. *Theor. Appl. Genet.* 90, 1174–1179. doi: 10.1007/BF00222940

Gan, S., and Amasino, R. M. (1995). Inhibition of leaf senescence by autoregulated production of cytokinin. *Science* 270:5244. doi: 10.1126/science.270.5244.1986

Hu, H., and Xiong, L. (2014). Genetic engineering and breeding of drought-resistant crops. *Annu. Rev. Plant Biol.* 65, 715–741. doi: 10.1146/annurev-arplant-050213-040000

IPCC (2015). *Climate Change 2014: Synthesis report*. Geneva: IPCC.

Kato, K., Miura, H., and Sawada, S. (1999). QTL mapping of genes controlling ear emergence time and plant height on chromosome 5A of wheat. *Theor. Appl. Genet.* 98, 472–477.

Landjeva, S., Neumann, K., Lohwasser, U., and Börner, A. (2008). Molecular mapping of genomic regions associated with wheat seedling growth under osmotic stress. *Biol. Plant.* 52, 259–266. doi: 10.1007/s10535-008-0056-x

Lemerle, D., Gill, G. S., Murphy, C. E., Walker, S. R., Cousens, R. D., Mokhtari, S., et al. (2001). Genetic improvement and agronomy for enhanced wheat competitiveness with weeds. *Aust. J. Agric. Res.* 52, 527–548. doi: 10.1071/AR00056

Li, G., Bai, G., Carver, B. F., Elliott, N. C., Bennett, R. S., Wu, Y., et al. (2017). Genome-wide association study reveals genetic architecture of coleoptile length in wheat. *Theor. Appl. Genet.* 130, 391–401. doi: 10.1007/s00122-016-2820-1

Li, X.-M., Chen, X.-M., Xiao, Y.-G., Xia, X.-C., Wang, D.-S., He, Z.-H., et al. (2014). Identification of QTLs for seedling vigor in winter wheat. *Euphytica* 198, 199–209. doi: 10.1007/s10681-014-1092-6

Liao, M., Fillery, I. R. P., and Palta, J. A. (2004). Early vigorous growth is a major factor influencing nitrogen uptake in wheat. *Funct. Plant Biol.* 31, 121–129. doi: 10.1071/FP03060

Liu, C., Sukumaran, S., Claverie, E., Sansaloni, C., Dreisigacker, S., and Reynolds, M. (2019). Genetic dissection of heat and drought stress QTLs in phenology-controlled synthetic-derived recombinant inbred lines in spring wheat. *Mol. Breed.* 39:34. doi: 10.1007/s11032-019-0938-y

Lopes, M. S., and Reynolds, M. P. (2012). Stay-green in spring wheat can be determined by spectral reflectance measurements (normalized difference vegetation index) independently from phenology. *J. Exp. Bot.* 63, 3789–3798. doi: 10.1093/jxb/ers071

López-Castañeda, C., and Richards, R. A. (1994). Variation in temperate cereals in rainfed environments III. Water use and water-use efficiency. *Field Crops Res.* 39, 85–98. doi: 10.1016/0378-4290(94)90011-6

López-Castañeda, C., Richards, R. A., Farquhar, G. D., and Williamson, R. E. (1996). Seed and seedling characteristics contributing to variation in early vigor among temperate cereals. *Crop Sci.* 36, 1257–1266. doi: 10.2135/cropsci1996.0011183X003600050031x

López-Castañeda, C. L., Richards, R. A., and Farquhar, G. D. (1995). Variation in Early Vigor between Wheat and Barley. *Crop Sci.* 36, 1257–1266.

Ludwig, F., and Asseng, S. (2010). Potential benefits of early vigor and changes in phenology in wheat to adapt to warmer and drier climates. *Agric. Syst.* 103, 127–136. doi: 10.1016/j.agsy.2009.11.001

Ma, J., Lin, Y., Tang, S., Duan, S., Wang, Q., Wu, F., et al. (2020). A genome-wide association study of coleoptile length in different chinese wheat landraces. *Front. Plant Sci.* 11:677. doi: 10.3389/fpls.2020.00677

Maydup, M. L., Graciano, C., Guiamet, J. J., and Tambussi, E. A. (2012). Analysis of early vigour in twenty modern cultivars of bread wheat (*Triticum aestivum* L.). *Crop Pasture Sci.* 63, 987–996. doi: 10.1071/CP12169

Moore, C., and Rebetzke, G. (2015). Genomic regions for embryo size and early vigour in multiple wheat (*Triticum aestivum* L.) populations. *Agronomy* 5, 152–179. doi: 10.3390/agronomy5020152

Mullan, D. J., and Reynolds, M. P. (2010). Quantifying genetic effects of ground cover on soil water evaporation using digital imaging. *Funct. Plant Biol.* 37, 703–712. doi: 10.1071/FP09277

Nursinow, D., Helfer, A., Hamilton, E., King, J., Imaizumi, T., Schultz, T., et al. (2011). The ELF4–ELF3–LUX complex links the circadian clock to diurnal control of hypocotyl growth. *Nature* 475, 398–402. doi: 10.1038/nature10182

Passioura, J. B. (2012). Phenotyping for drought tolerance in grain crops: when is it useful to breeders? *Funct. Plant Biol.* 39, 851–859. doi: 10.1071/FP12079

Rebetzke, G. J., and Richards, R. A. (1999). Genetic improvement of early vigour in wheat. *Aust. J. Agric. Res.* 50, 291–301. doi: 10.1071/A98125

Rebetzke, G. J., Ellis, M. H., Bonnett, D. G., and Richards, R. A. (2007). Molecular mapping of genes for *Coleoptile* growth in bread wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* 114, 1173–1183. doi: 10.1007/s00122-007-0509-1

Rebetzke, G. J., López-Castañeda, C., Acuña, T. L. B., Condon, A. G., and Richards, R. A. (2008). Inheritance of coleoptile tiller appearance and size in wheat. *Aust. J. Agric. Res.* 59, 863–873. doi: 10.1071/AR07397

Rebetzke, G. J., Richards, R. A., and Holland, J. B. (2017). Population extremes for assessing trait value and correlated response of genetically complex traits. *Field Crops Res.* 201, 122–132. doi: 10.1016/j.fcr.2016.10.019

Rebetzke, G. J., Verbyla, A. P., Verbyla, K. L., Morell, M. K., and Cavanagh, C. R. (2014). Use of a large multiparent wheat mapping population in genomic dissection of coleoptile and seedling growth. *Plant Biotechnol. J.* 12, 219–230. doi: 10.1111/pbi.12130

Richard, A. R., Colin, R. C., and Penny, R. (2019). Selection for erect canopy architecture can increase yield and biomass of spring wheat. *Field Crops Res.* 244, 107649. doi: 10.1016/j.fcr.2019.107649

Richard, C. A., Hickey, L. T., Fletcher, S., Jennings, R., Chenu, K., and Christopher, J. T. (2015). High-throughput phenotyping of seminal root traits in wheat. *Plant Method.* 11:13. doi: 10.1186/s13007-015-0055-9

Richards, R. A., Dennett, C. W., Qualset, C. O., Epstein, E., Norlyn, J. D., and Winslow, M. D. (1987). Variation in yield of grain and biomass in wheat, barley, and triticale in a salt-affected field. *Field Crops Res.* 15, 277–287. doi: 10.1016/0378-4290(87)90017-7

Richards, R. A., and Lukacs, Z. (2002). Seedling vigour in wheat – sources of variation for genetic and agronomic improvement. *Aust. J. Agric. Res.* 53, 41–50. doi: 10.1071/AR00147

Richards, R. A., and Townley-Smith, T. F. (1987). Variation in leaf area development and its effect on water use, yield and harvest index of droughted wheat. *Aust. J. Agric. Res.* 38, 983–992. doi: 10.1071/AR9870983

Richards, R. A., Rebetzke, G. J., Watt, M., Condon, A. G., Spielmeyer, W., and Dolferus, R. (2010). Breeding for improved water productivity in temperate cereals: phenotyping, quantitative trait loci, markers and the selection environment. *Funct. Plant Biol.* 37, 85–97. doi: 10.1071/FP09219

Ryan, P. R., Liao, M., Delhaize, E., Rebetzke, G. J., Weligama, C., Spielmeyer, W., et al. (2015). Early vigour improves phosphate uptake in wheat. *J. Exp. Bot.* 66, 7089–7100. doi: 10.1093/jxb/erv403

Sadras, V., and Dreccer, M. F. (2015). Adaptation of wheat, barley, canola, field pea and chickpea to the thermal environments of Australia. *Crop Pasture Sci.* 66, 1137–1150. doi: 10.1071/CP15129

Shi, S., Azam, F. I., Li, H., Chang, X., Li, B., and Jing, R. (2017). Mapping QTL for stay-green and agronomic traits in wheat under diverse water regimes. *Euphytica* 213:246. doi: 10.1007/s10681-017-2002-5

Shi, Y., Thomasson, J. A., Murray, S. C., Pugh, N. A., Rooney, W. L., Shafian, S., et al. (2016). Unmanned aerial vehicles for high-throughput phenotyping and agronomic research. *PLoS One* 11:e0159781. doi: 10.1371/journal.pone.0159781

Shim, H., Chasman, D. I., Smith, J. D., Mora, S., Ridker, P. M., Nickerson, D. A., et al. (2015). A multivariate genome-wide association analysis of 10 LDL subfractions, and their response to statin treatment, in 1868 Caucasians. *PLoS One* 10:e0120758. doi: 10.1371/journal.pone.0120758

Siddique, K. H. M., Tennant, D., Perry, M. W., and Belford, R. K. (1990). Water use and water use efficiency of old and modern wheat cultivars in a mediterranean-type environment. *Aust. J. Agric. Res.* 41, 431–447. doi: 10.1071/ar9900431

Steege, M. W., den Ouden, F. M., Lambers, H., Stam, P., and Peeters, A. J. (2005). Genetic and physiological architecture of early vigor in *Aegilops tauschii*, the D-genome donor of hexaploid wheat. A quantitative trait loci analysis. *Plant Physiol.* 139, 1078–1094. doi: 10.1104/pp.105.063263

Steinfort, U., Trevaskis, B., Fukai, S., Bell, K. L., and Dreccer, M. F. (2017). Vernalisation and photoperiod sensitivity in wheat: Impact on canopy development and yield components. *Field Crops Res.* 201, 108–121. doi: 10.1016/j.fcr.2016.10.012

Sutka, J., and Snape, J. W. (1989). Location of a gene for frost resistance on chromosome 5A of wheat. *Euphytica* 42, 41–42. doi: 10.1007/BF00042613

Tura, H., Edwards, J., Gahlaut, V., Garcia, M., Sznajder, B., Baumann, U., et al. (2020). QTL analysis and fine mapping of a QTL for yield-related traits in wheat grown in dry and hot environments. *Theor. Appl. Genet* 133, 239–257. doi: 10.1007/s00122-019-03454-6

Turner, N. C., and Nicolas, M. E. (1998). "Early vigour : a yield-positive characteristic for wheat in drought-prone mediterranean-type environment," in *Crop Improvement for Stress Tolerance*, eds R. K. Behl, D. P. Singh, and G. P. Lodhi (New Delhi: CCSHAU, Hisar and MMB), 47–62.

Valle, S. R., and Calderini, D. F. (2010). Phyllochron and tillering of wheat in response to soil aluminum toxicity and phosphorus deficiency. *Crop Pasture Sci.* 61, 863–872. doi: 10.1071/CP09310

Whan, B. R., Carlton, G. P., and Anderson, W. K. (1991). Potential for increasing early vigour and total biomass in spring wheat. I. Identification of genetic improvements. *Aust. J. Agric. Res* 42, 541–553. doi: 10.1071/AR9910347

Yu, J.-B., and Bai, G.-H. (2010). Mapping quantitative trait loci for long coleoptile in chinese wheat landrace wangshuibai. *Crop Sci.* 50, 43–50. doi: 10.2135/cropsci2009.02.0065

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44, 821–824. doi: 10.1038/ng.2310

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 13 January 2022
doi: 10.3389/fpls.2021.727835

Unraveling Heat Tolerance in Upland Cotton (*Gossypium hirsutum* L.) Using Univariate and Multivariate Analysis

Muhammad Mubashar Zafar[1,2†], Xue Jia[1†], Amir Shakeel[3], Zareen Sarfraz[2], Abdul Manan[3], Ali Imran[3], Huijuan Mo[1,2], Arfan Ali[4], Yuan Youlu[1,2], Abdul Razzaq[2,5], Muhammad Shahid Iqbal[1,2,6*] and Maozhi Ren[1,2*]*

[1] Zhengzhou Research Base, State Key Laboratory of Cotton Biology, School of Agricultural Sciences, Zhengzhou University, Zhengzhou, China, [2] Institute of Cotton Research, Chinese Academy of Agricultural Sciences, Anyang, China, [3] Department of Plant Breeding and Genetics, University of Agriculture, Faisalabad, Pakistan, [4] FB Genetics, Four Brothers Group, Lahore, Pakistan, [5] Institute of Molecular Biology and Biotechnology, The University of Lahore, Lahore, Pakistan, [6] Cotton Research Station, Ayub Agricultural Research Institute, Faisalabad, Pakistan

OPEN ACCESS

Edited by:
Suchismita Mondal,
International Maize and Wheat Improvement Center, Mexico

Reviewed by:
Alireza Pour-Aboughadareh,
Seed and Plant Improvement Institute, Iran
Huiying Li,
Chinese Academy of Sciences (CAS), China

***Correspondence:**
Abdul Razzaq
biolformanite@gmail.com
Muhammad Shahid Iqbal
shahidkooria@gmail.com
Maozhi Ren
renmaozhi01@caas.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: 19 June 2021
Accepted: 10 November 2021
Published: 13 January 2022

Citation:
Zafar MM, Jia X, Shakeel A, Sarfraz Z, Manan A, Imran A, Mo H, Ali A, Youlu Y, Razzaq A, Iqbal MS and Ren M (2022) Unraveling Heat Tolerance in Upland Cotton (Gossypium hirsutum L.) Using Univariate and Multivariate Analysis. Front. Plant Sci. 12:727835. doi: 10.3389/fpls.2021.727835

The ever-changing global environment currently includes an increasing ambient temperature that can be a devastating stress for organisms. Plants, being sessile, are adversely affected by heat stress in their physiology, development, growth, and ultimately yield. Since little is known about the response of biochemical traits to high-temperature ambiance, we evaluated eight parental lines (five lines and three testers) and their 15 F_1 hybrids under normal and high-temperature stress to assess the impact of these conditions over 2 consecutive years. The research was performed under a triplicate randomized complete block design including a split-plot arrangement. Data were recorded for agronomic, biochemical, and fiber quality traits. Mean values of agronomic traits were significantly reduced under heat stress conditions, while hydrogen peroxide, peroxidase, total soluble protein, superoxide dismutase, catalase (CAT), carotenoids, and fiber strength displayed higher mean values under heat stress conditions. Under both conditions, high genetic advance and high heritability were observed for seed cotton yield (SCY), CAT, micronaire value, plant height, and chlorophyll-a and b content, indicating that an additive type of gene action controls these traits under both the conditions. For more insights into variation, Pearson correlation analysis and principal component analysis (PCA) were performed. Significant positive associations were observed among agronomic, biochemical, and fiber quality-related traits. The multivariate analyses involving hierarchical clustering and PCA classified the 23 experimental genotypes into four groups under normal and high-temperature stress conditions. Under both conditions, the F_1 hybrid genotype FB-SHAHEEN × JSQ WHITE GOLD followed by Ghuari-1, CCRI-24, Eagle-2 × FB-Falcon, Ghuari-1 × JSQ White Gold, and Eagle-2 exhibited better performance in response to high-temperature stress regarding the agronomic and fiber quality-related traits. The mentioned genotypes could be utilized in future cotton breeding programs to enhance heat tolerance and improve cotton yield and productivity through resistance to environmental stressors.

Keywords: high-temperature stress, upland cotton (*Gossypium hirsutum* L.), principal component analysis (PCA), heritability, *Gossypium*

INTRODUCTION

Cotton is the world's most important natural fiber and oil crop (Patel et al., 2021; Salimath et al., 2021). Climate change in recent decades has resulted in outbreaks of biotic and abiotic stressors that negatively affect plant yield and quality. Among abiotic stressors, heat stress is one of the most detrimental constraints, limiting cotton production by disturbing its normal growth, physiological, and developmental processes. Pakistan ranks fifth among top cotton-producing countries after India, China, United States, and Brazil (Khan et al., 2020), even though the nations cotton zone average temperature ranks highest among cotton-growing areas worldwide (Saleem et al., 2021). In Pakistan, the average temperature of the cotton-growing belt remains at 37°C/25°C (day/night) as compared to the United States 30°C /24°C, China's 29°C/18°C, and India's 34°C/21°C (Saleem et al., 2021). During the early growth period (May–June), the temperature remains as high as 40–45°C, and at times reaches 50°C.

Like most crop plants, cotton is susceptible to heat stress, especially during the developmental (Zahid et al., 2016) and reproductive phases (Salman et al., 2019). The most notable effects include flower shedding at the flowering phase, leading to stunted growth and reduced boll weight, and ultimately lower yields (Xu et al., 2020). At the peak of the reproductive phase, exposure to heat stress very often results in reduced seed cotton yield (SCY); whereas slightly lower temperatures at this time are more favorable and produce a better yield (Sarwar et al., 2017). Previous studies report some of the optimum temperatures for various growth and developmental stages of the cotton crop: for cottonseed germination 12°C, root development 30°C, and seedling development for boll development 25.5–29.5°C (Conaty et al., 2012; Lokhande and Reddy, 2014). A temperature range of 32–40°C usually negatively impacts root development, and when the temperature rises to 36°C, stomatal conductance decreases. The weight of the boll reduces as the temperature rise from 25.5 to 29.5°C (Conaty et al., 2012; Lokhande and Reddy, 2014). Singh et al., 2007 reported that each 1°C rise of temperature in the field reduces the SCY by 110 kg ha^{-1}. Pollen tube germination, growth, and elongation are adversely affected by a temperature increase from 28 to 30°C. At 28°C, optimum pollen germination occurs (Burke et al., 2004) and the germination rate decreases as the temperature rises sharply from 28 to 37°C. Therefore, high-temperature stress reduces the germination rate, plant growth, photosynthetic rate, fruiting branches, membrane integrity, boll weight, and increases boll abscission, all of which lead to lower yield (Salman et al., 2019). Heat stress is also related to reduced boll size and the number of seeds per boll that limit fertilization efficiency (Pettigrew, 2008). In Pakistan, the average boll weight is 2–3 g, which is lower than in other countries (Saleem et al., 2021). High-temperature stress decreases the chlorophyll content, ultimately reducing the photosynthetic rate and translocation of assimilates to reproductive organs and thus, increases senescence (Rafiq et al., 2013; Dabbert and Gore, 2014). High-temperature stress also distorts stomatal movement and stomatal conductance 28–30°C resulting in poor gaseous exchange, ultimately effecting photosynthesis and ultimately the productivity (Conaty et al., 2012).

This high temperature abiotic stress affects plant antioxidant activities, leading to decreased SCY (Kamal et al., 2017). Under high-temperature stress, oxidative stress is induced to generate reactive oxygen species (ROS) (Kocsy et al., 2004). Under heat stress, ROS such as hydrogen peroxide (H_2O_2), superoxide radical (O_2^-), singlet oxygen ($^1O_2$), and hydroxyl radicals are produced in higher amounts (Choudhury et al., 2013). Their increased production may irreversibly damage plant cells through the oxidation of different cellular compartments such as chloroplasts, peroxisomes, and mitochondria (Roychoudhury et al., 2012). In plants, tolerance to oxidative damage is directly correlated with the production of antioxidant enzymes (Almeselmani et al., 2009). To scavenge the damaging ROS, the cotton plant activates the production of detoxifying enzymes including superoxide dismutase (SOD), peroxidases (POD), catalase (CAT), and non-enzymatic antioxidants including carotenoids, flavonoids, ascorbate, and tocopherols (Suzuki et al., 2012). Since high-temperature stress is a detrimental factor in cotton production, the development of heat-tolerant germplasm is a prominent objective of today's cotton breeders, who aim to achieve higher yields under heat stress (Teixeira et al., 2013).

Numerous approaches can be adopted to face and overcome high-temperature abiotic stress in cotton. Scientists consider the development of heat-tolerant germplasm as reliable, long-lasting, cost-effective, and the best possible solution for combating this stress. Through conventional breeding, a significant level of tolerance in cotton can be achieved. Moreover, the global mean temperature is constantly increasing, urging cotton breeders to search for hidden potential genotypes from the existing germplasm *via* effective screening approaches based on particular morphological, biochemical, and physiological traits (Salman et al., 2019). The selection of cotton genotypes tolerant to heat stress is a prerequisite for cotton breeding improvement programs.

Understanding multivariate statistics necessitates an understanding of high-dimensional geometry and a conceptualization of linear algebra (Stewart and Thomas, 2008). Unlike univariate and bivariate models, multivariate data addresses several issues simultaneously. Multivariate analysis can provide many options to test and summarize the power of linear relationships across multiple variables (Timm, 2002). For example, in correlation tests, parametric and non-parametric options are present while using this technique. Plant breeders can use multivariate analysis to understand differences across variables and their possible associations (Dhamayanthi et al., 2018). For using standard least-square fit, several reports state that environmental indicators may significantly correlate with quantitative traits, including crop yield (Zhang et al., 2010; Sellam and Poovammal, 2016; Zhou et al., 2021). Hence, simple ANOVA functions are usually inefficient for describing the effect of environmental indicators with the desired productivity and quality owing to higher complexity in a different set of variables (Hoaglin and Welsch, 1978; Goos and Meintrup, 2016). Principal component analysis (PCA) and cluster analysis

have received more attention owing to impact during recent decades (Mohammadi and Prasanna, 2003). PCA can efficiently analyze the interaction among traits and the performance of genotypes and efficiently dissect trait association (Aslam et al., 2017). Cluster analysis groups have rows together that share similar values across several variables and are considered as an exploratory technique that helps to understand clumping structures in data (Rathinavel, 2018). Correlation analysis is mainly used to understand the degree of relationship and its nature across traits. It can deal with the basic notion of an association across various traits, which can help in the selection of genotypes with the desired combination of traits (Ghafoor et al., 2013).

Multivariate analyses comprise highly acceptable and precise methods and techniques to explore the potential genetic variation existing in the available germplasm (Malik et al., 2014). It can also exploit prospective genetic associations and patterns of variability within the studied germplasm. Worldwide, multivariate analyses are used to study a range of crops, especially maize, wheat, cotton, and sorghum (Ali et al., 2011; Ajmal et al., 2013; Jarwar et al., 2019). Hence, the current work was designed to evaluate suites of biochemical, morphological, and agronomic traits of available potential cotton existing genotypes *via* univariate and multivariate analyses under heat stress with the aim of developing new heat resilient cultivars. Such research would yield basic information regarding high-temperature resilience potential in existing genotypes that may prove valuable and advantageous for developing heat-tolerant cotton cultivars in cotton breeding improvement programs across heat-stricken climatic regions.

MATERIALS AND METHODS

Plant Materials and Experimental Layout

During the normal cotton growing season of November 2017, 50 cotton genotypes were screened against heat stress based on SCY under field conditions at FB Genetics, Four Brothers Group, Pakistan. The experiment was performed in the field under two treatments, i.e., normal and heat stress (5–6°C above normal for 12 days at 50% flowering) following a split-plot arrangement under a randomized complete block design (RCBD) with two replications. The plant × plant and row × row distance was 30 and 75 cm, respectively, with a 6 m row length for each genotype under each replication. The temperature was raised by constructing a tunnel using polythene sheets for 12 days at 50% flowering and it was removed at night. After screening, eight cotton genotypes were selected as parents regarding their SCY under heat stress. The selected parental material was crossed in line × tester mating fashion in the following season to obtain the subsequent F_1 hybrids. The five heat-tolerant genotypes were kept as female lines, namely, Ghuari-1, Badar-1, Eagle-2, CCRI-24, and Fb-Shaheen. The three sensitive genotypes were kept as male testers: Fb-Falcon, Fb-Smart 1, and JSQ White Gold.

The 15 F_1 hybrids and their eight parents are listed in **Table 1**, and were planted at the field research area in the normal cotton growing season under normal and high-temperature stress conditions for 2 consecutive years, 2018 and 2019. Experimental layout was performed in a split-plot arrangement under RCBD. The plant × plant and row × row distance was kept as 30 and 75 cm, respectively, with a 6 m row length for each genotype under each replication. The seed of the selected genotypes was manually sown (dibble method) on furrows in June. The crop was harvested in October each year. The R × R and P × P distance was 75 and 30 cm, respectively. The experiment was laid out in a triplicated manner under RCBD following a split-plot arrangement. All the culture and agronomic practices were performed following local recommendations across crop-growing seasons over 2 years.

Imposition of High-Temperature Stress

During September, when all genotypes were at the 50% flowering stage, heat stress was implemented for 12 days during both the years. The covering of the polythene tunnel enhanced the temperature (5–6°C) during the daytime and the tunnel was removed during the night (Muhammad et al., 2018). The minimum and maximum temperatures inside the tunnel were continuously recorded (Both et al., 2015) throughout the crop growing season (**Supplementary Table 1A**). After the implementation of high-temperature stress, data were collected regarding biochemical characters. The maximum and minimum temperature ranges recorded during the crop growing season are given in **Supplementary Table 1B**.

Biochemical Traits

For the determination of biochemical traits, leaf samples were collected from the experimental genotypes after the imposition of high-temperature treatment for 12 days. The quantification of H_2O_2, CAT, peroxidase (POD), total soluble proteins (TSP), chlorophyll contents (Chl), and carotenoids (Car) in the leaves was performed to assess the effect of stress on biochemical attributes of the plants. For this purpose, the fourth fully expanded top leaf was considered for sampling from each genotype for biochemical analyses following the sampling method used by Song et al. (2014). Enzyme extraction was conducted on 0.5 g of cotton leaf samples. The leaves were cut with the help of a leaf pincher and then crushed and ground with 1–2 mL of chilled potassium phosphate buffer (pH 7.8). The prepared mixture was then centrifuged for 5 min at 1,400 rpm. Residues were discarded and the supernatant was collected for the determination of biochemical attributes *via* UV spectrophotometer at different wavelengths (Sarwar et al., 2019).

Hydrogen Peroxide (μmol/g-FW)

For the determination of H_2O_2, the Velikova protocol was followed (Velikova et al., 2000). Fresh leaf tissues (0.5 g) were blended by using trichloroacetic acid (TCA, 5 mL of 0.1% (w/v) solution) and then centrifuged at 12,000 rpm for 12 min. The supernatant was collected in a volume of 0.5 mL, and then 0.5 mL of phosphate buffer (pH 7.0) and 1 mL of potassium iodide were added. At the 390 nm wavelength of the UV spectrophotometer, the absorbance capacity of each sample was recorded.

TABLE 1 | List of lines, testers, and their cross combinations used in the experiment.

Lines	Testers	Crosses	Crosses	Crosses
Ghuari-1	Fb-Falcon	Ghuari-1 × Fb-Falcon	Badar-1 × Js White Gold	CCRI-24 × Fb-Smart1
Badar-1	Fb-Smart 1	Ghuari-1 × Fb-Smart1	Eagle-2 × Fb-Falcon	CCRI-24 × JSQ White Gold
Eagle-2	JSQ White Gold	Ghuari-1 × JSQ White Gold	Eagle-2 × Fb-Smart1	Fb-Shaheen × Fb-Falcon
CCRI-24		Badar-1 × Fb-Falcon	Eagle-2 × JSQ White Gold	Fb-Shaheen × Fb-Smart1
Fb-Shaheen		Badar-1 × Fb-Smart1	CCRI-24 × Fb-Falcon	Fb-Shaheen × JSQ White Gold

Catalase (U/mg Protein)

Enzyme extract (0.1 mL) was mixed with 3 mL of the reaction mixture, containing 5.9 mM H_2O_2 and 50 mM potassium phosphate buffer (7.0 pH). CAT activity was recorded at the wavelength of 240 nm (Liu et al., 2009) using a spectrophotometer.

Peroxidase (U/mg Protein)

The POD solution contained 50 mM phosphate buffer (pH = 5), 40 mM H_2O_2, 20 mM guaiacol, and 0.1 mL of enzyme extract according to Liu's protocol, after certain amendments (Liu et al., 2009). At 470 nm, absorbance changes were recorded by the spectrophotometer.

Total Soluble Proteins (mg/g-FW)

The Bradford reagent method was used for the determination of protein content. Aliquots of 100 μL of the sample were blended with 5 mL of Bradford reagent. At 595 nm wavelength, the absorbance was recorded (Bradford, 1976) using a spectrophotometer.

Chlorophyll Content and Carotenoids Assay

The Arnon method (Arnon, 1949) with specific alterations measured Chl a and b contents and carotenoid pigments. A volume of 8–10 mL of 80% acetone (v/v) was used for crushing a 0.50 g sample of the cotton leaf. Filter paper was used to obtain a homogenized solution. A spectrophotometer was employed to record the absorbance of the final solution at 645 and 663 nm wavelengths. Chl a and b contents and Car were estimated by using the following formulas.

$$\text{Chlorophyll a}\left(\frac{\mu g}{g}\text{FW}\right) = [12.7\,(\text{OD }663) - 2.69\,(\text{OD }645)] \times \frac{v}{1000 \times w}$$

$$\text{Chlorophyll b}\left(\frac{\mu g}{g}\text{FW}\right) = [22.9\,(\text{OD }665) - 4.48\,(\text{OD }663)] \times \frac{V}{1000 \times w}$$

$$\text{Carotenoids}\left(\frac{\mu g}{g}\text{FW}\right) = \frac{A^{car}}{Em} \times 1000$$

$$A^{car} = \text{OD }480 + 0.114(\text{OD }663) - 0.638(\text{OD }645)$$

where,

W = weight of leaf sample, V = volume of sample, Em = 2,500

Yield and Fiber Quality Traits

At crop agronomic maturity, data from five plants of each genotype regarding yield-related traits were recorded. The yield-related traits included plant height (PH), the number of bolls (TNB), boll weight (BW), SCY, and lint percentage (lint%). A representative sample from seed cotton obtained from the experimental genotypes was taken and weighed. The ginning of seed cotton samples was accomplished with the help of a single roller ginning machine (Testex, Model: TB510C) to separate seed and lint, and the ginning outturn was estimated by dividing the weight of lint in a sample by the seed cotton weight of the sample, expressed in percentage. Lint was further subjected to fiber quality analysis for the estimation of fiber strength (g/tex) (STR), short fiber (SF), micronaire value (MIC), reflectance (%) (RD), upper half mean length (mm) (UHML), and uniformity index ratio (%) (UI) with a high-volume instrument (HVI-900, USTER, United States), following ASTM protocol, publication D5867-05 for HVI analysis (ASTM, 2005).

Statistical Analysis

The preliminary screening data comprising of 50 upland cotton accessions based on yield performance across heat and normal conditions were subjected to a linear mixed model using ANOVA, followed by the construction of an ANOM-decision chart to graphically represent the genotypic behavior for selection of parents for crossing through analysis of mean methods as described by Nelson et al. (2005). These analyses were performed using default and standardization options with SAS-JMP Pro 16 (SAS Institute Inc., Cary, NC, United States, 1989–2021). In this method, if a single or group of genotypes plotted statistics fell outside of the decision limits, then the test indicated a statistical difference between that group's statistic and the overall average of the statistic for genotypes/groups (Yiğit and Mendeş, 2017). Based on the ANOM-decision charts, genotypes performing better across both normal and stress conditions having significant differences from means and falling above the upper decision level (UDL) can be considered as tolerant, whereas the genotypes having significant reduction of yield having a significant difference from means below the lower decision level (LDL) in stress treatment can be considered as susceptible.

Data collected from evaluation of parents and F_1s across normal and heat stress for agronomic, biochemical, and fiber traits has been subjected to analysis of variance (Steel et al., 1997) to estimate genetic variability among parents and their

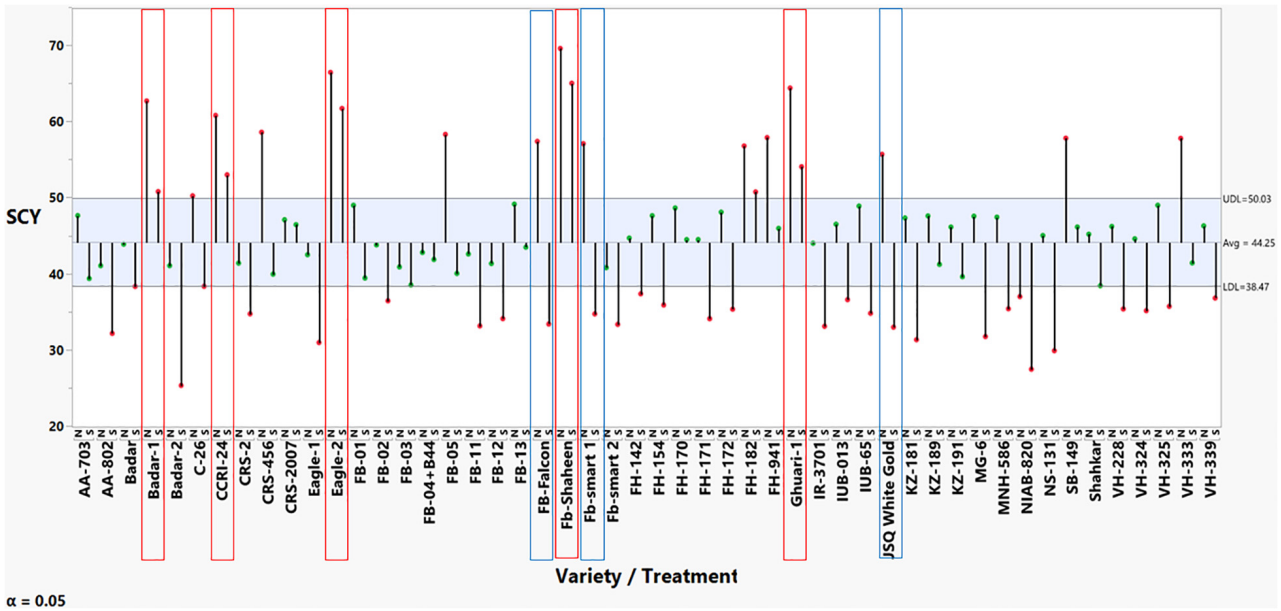

FIGURE 1 | ANOM-decision chart with decision limits 38.47–50.03 for seed cotton yields across normal and heat stress (alpha > 0.05%). It provides a graphical test for simultaneously comparing the mean performance of these 50 cotton genotypes across normal and heat stress. Red-colored heads represent significant deviation from the mean, either above upper decision level (UDL) or below lower decision level (LDL).

subsequent hybrids. Means and standard errors were calculated and used throughout all the data sets (Gomez and Gomez, 1984). The years were then pooled to obtain mean values for further analyses. The method proposed by Singh and Chaudhary (1985) was used to calculate broad-sense heritability (H^2b). For the categorization of H^2b, a method based on a range of values was used, as follows: low H^2b had a < 30% value, medium H^2b values range between 30 and 70%, and high H^2b values above 70%. The H^2b was classified following the procedure given by Johnson (Johnson et al., 1955). Genetic advance percentage (GAM) is calculated by the method proposed by Poehlman and Sleper (1995) under a 20% selection intensity. For multivariate analyses, average data was taken for replications. Subsequently, means were subjected to multivariate analyses, including correlation matrix (Pearson correlation) and PCA (Correlation based), two-way cluster analysis (hierarchical clustering), and construction of a distance-based tree using Ward's method; all these analyses were performed using default analyses and standardization options with SAS-JMP Pro 16 (SAS Institute Inc., Cary, NC, United States, 1989–2021).

To validate the results and also to determine the performance of genotypes with a high SCY and good fiber quality across normal and heat stress, 3D scatterplots were constructed based on stress tolerance indices, including mean performance (MP), geometric mean performance (GMP), and stress tolerance index (STI) for normal and heat stress with the help of the freely available online software package iPASTIC developed by Pour-Aboughadareh et al. (2019). To rank and identify the best genotypes having stable and better yields across both conditions, the representative trait was used according to the method given by Ketata et al. (1989). Based on this approach, the average sum of rank (ASR) corresponding to all variables/indices was used as an indicator for selecting the best genotypes. According to this procedure, the lowest rank was assigned to the genotype having the best performance for the corresponding variable; hence, genotypes with the lowest value for ASR and lowest values for standard deviation were denoted as the best ones.

RESULTS

The results from the preliminary screening experiment across normal and heat stress conditions, represented in **Supplementary Table 2**, revealed non-significant results for replications, whereas significant effect estimates were found for genotypes, treatment, and genotype treatment interaction for SCY. The results for mean comparisons through the analysis of mean methods (ANOM)-decision chart were constructed to represent the genotypic behavior for selection based on the ANOM. Based on the results shown in **Figure 1**, five genotypes were declared as tolerant since their means fell above the UDL across both normal and stress conditions without significantly decreasing SCY due to imposed stress. These five heat-tolerant cotton lines/genotypes were used as female lines for crossing: Ghuari-1, Badar-1, Eagle-2, CCRI-24, and Fb-Shaheen. We also selected three heat susceptible genotypes because they produced a significant decline in SCY across heat-stress conditions compared with a higher yield performance in normal conditions. These genotypes had a significantly higher yield in normal conditions, i.e., above the UDL, whereas they had significantly reduced yield performance in heat stress, i.e., below the LDL. These genotypes were: Fb-Falcon, Fb-Smart 1, and JSQ White

TABLE 2 | Summary of mean square values regarding the influence of high-temperature stress on different traits of the studied cotton genotypes of parents and their F_1 hybrids across the 2 experimental years.

Traits	Genotypes		Heat Stress		Genotypes × Heat Stress		Heat stress × Year	Genotypes × Year	Heat stress × Year × genotype
	1st year	2nd year	1st year	2nd year	1st year	2nd year			
Plant Height	137.68**	151.7**	487.600	619.8	27.789**	30.0**	3.96	2.18	1.37
Number of bolls	36.650**	37.90**	877.93*	838.2*	14.092**	10.53**	0.23	1.61	1.26
Boll weight	0.2247**	0.184**	0.0088	0.182	0.0824**	0.04817	0.05	0.01	0.03
Seed cotton yield	249.63**	237.4**	1808.5*	2329.0*	57.09**	60.01**	16.44	5.44	7.33
Ginning out-turn	65.164 **	51.28**	67.151*	103.5	13.489	18.16**	1.962	1.552	1.59
Hydrogen peroxide	29.950**	23.57**	326.4**	611.5**	7.286**	3.691**	22.1**	1.278	1.8 *
Catalase	2,261**	2,123**	7,143**	7,178**	598**	556**	2.0	5.0	4.0
Peroxidase	542.3**	446.5**	9,675**	90,382.*	321.5**	323.5**	54.3**	9.45*	5.32
Super-oxidase dismutase	145.4**	131.2**	38,581**	40,782.*	120.0**	87.3**	15.2	15.6*	21.0**
Total soluble protein	24.11**	18.4**	3,009**	2,978.8*	14.47**	16.6**	0.04	2.41**	2.47**
Chlorophyll contents a	0.094**	0.10**	6.723**	8.52**	0.055**	0.04**	0.05**	0.00	0.00
b	0.011**	0.01**	0.431**	0.59**	0.005**	0.004**	0.006**	5.01	7.94*
Carotenoids	0.011**	0.010**	4.248*	4.019**	0.005**	0.003**	0.001	0.000	0.00
Short fiber	0.900**	0.86**	0.960	1.86	1.524**	1.46**	0.074	0.02	0.079*
Fiber strength	22.62**	30.8**	29.50	3.25	13.04**	20.4**	26.18**	8.26 **	11.7**
MIC value	1.351**	1.06**	0.841**	0.26**	0.262**	0.19**	1.02**	0.10*	0.10*
Reflectance	10.44**	14.4**	225.3**	186.5*	9.09**	11.7**	0.918	2.759	2.214
Uniformity index	6.96**	22.07**	102.6*	53.5**	14.28**	16.62**	3.962	10.9**	11.0**
Upper half mean length	9.54**	11.4**	40.445	16.5	4.85**	7.6**	2.63	2.54	2.92

**Significance (α = 0.05), **highly Significant (α = 0.01).*

TABLE 3 | Genetic components of variability, genetic advance percentage means, and heritability (broad sense) estimate studied traits across normal (N) and heat stress (HT) conditions for pooled data across the years 2018 and 2019.

SOV		Max	Mini	Mean	CV%	GCV	PCV	H2b%	GAM
Plant height (cm)	N	99.80	66.00	83.68	3.86	7.33	8.29	78.29	13.37
	HT	98.80	68.20	78.49	3.45	8.49	9.16	85.85	16.20
Number of bolls	N	29.80	12.20	20.63	11.15	19.04	22.06	74.46	33.84
	HT	20.60	7.20	14.59	11.87	14.78	18.95	60.80	23.74
Boll weight (g)	N	3.75	2.31	3.01	5.46	7.18	9.02	63.32	11.77
	HT	3.74	2.04	2.92	6.36	6.77	9.29	53.17	10.17
Seed cotton yield (g)	N	69.20	29.60	50.31	5.29	15.52	16.40	89.59	30.26
	HT	69.63	26.41	40.25	8.11	22.06	23.51	88.09	42.66
Ginning out turn%	N	59.87	39.82	50.55	5.22	7.90	9.47	69.60	13.58
	HT	59.57	39.41	48.57	5.17	7.09	8.77	65.27	11.79
Hydrogen peroxide (μmol/g)	N	15.10	2.30	6.63	23.69	38.80	45.46	72.85	68.22
	HT	18.58	3.36	11.78	14.21	17.78	22.76	61.05	28.62
Catalase (U mg^{-1} protein)	N	112.00	15.00	66.54	4.82	40.36	40.65	98.59	82.56
	HT	314.00	207.10	243.21	2.24	10.06	10.30	95.26	20.22
Peroxidase (U mg^{-1} protein)	N	90.00	12.40	42.62	7.29	43.30	43.91	97.25	87.96
	HT	123.60	89.90	105.29	2.55	5.69	6.24	83.33	10.71
Superoxide dismutase (U mg^{-1} protein)	N	71.30	36.30	55.90	5.72	13.85	14.98	85.44	26.36
	HT	119.10	79.60	98.01	3.96	6.18	7.34	70.84	10.71
Carotenoids	N	0.34	0.05	0.21	12.46	26.85	29.60	82.29	50.17
(mg g^{-1} FW)	HT	0.80	0.47	0.63	6.05	8.84	10.71	68.14	15.04
Reflectance	N	84.12	69.40	75.38	3.28	3.46	4.77	52.76	5.18
	HT	76.95	68.45	72.53	1.61	2.20	2.72	64.97	3.65
The upper half mean length (mm)	N	33.14	21.92	26.89	6.82	7.40	10.07	54.09	11.22
	HT	30.01	20.92	26.04	5.44	6.50	8.48	58.81	10.28
Uniformity index %	N	97.36	80.14	86.20	3.09	3.78	4.88	59.88	6.02
	HT	87.80	80.14	84.67	2.06	2.26	3.06	54.63	3.44
Short fiber contents (%)	N	10.20	5.40	8.13	5.87	10.50	12.03	76.17	18.87
	HT	9.50	7.20	7.85	3.44	6.80	7.62	79.63	12.50
MIC (units)	N	6.90	4.00	5.12	5.25	12.01	13.11	83.94	22.67
	HT	5.90	3.80	5.23	4.06	8.50	9.42	81.43	15.80
Total soluble protein contents (mg g^{-1} FW)	N	9.42	1.18	4.65	21.92	20.17	29.79	45.86	28.15
	HT	29.33	9.30	16.03	10.85	23.85	26.20	82.85	44.71
Chlorophyll a contents (mg g^{-1} FW)	N	1.75	0.59	1.12	6.43	21.99	22.91	92.13	43.48
	HT	0.80	0.22	0.51	10.58	21.56	24.02	80.60	39.88
Chlorophyll b contents (mg g^{-1} FW)	N	0.50	0.11	0.37	7.53	20.04	21.41	87.63	38.65
	HT	0.29	0.09	0.21	12.60	20.23	23.83	72.05	35.38
Fiber strength (g/tex)	N	39.80	20.03	30.54	8.74	13.14	15.78	69.34	22.53
	HT	39.80	25.10	32.05	5.70	7.53	9.44	63.62	12.37

SOV, Source of variation); CV%, coefficient of variation; GCV%, genotypic coefficient of variation; PCV%, phenotypic coefficient of variance); H^2b, broad sense heritability; GAM, genetic advance per percent means.

Gold, and were selected to be used as testers (male parent) in crossing (**Figure 1**).

Analysis of variance showed significant differences among parents and F_1 hybrids under heat stress during both years; this pointed toward the existence of genetic divergence (**Table 2**). The Genotypes × Treatment interaction for all traits was highly significant, suggesting that all parents and hybrids behaved differently under heat stress (**Table 2**). The analysis of variance for heat stress × year showed non-significant interactions for all traits except H_2O_2, POD, Chl a and b contents, STR, and MIC. The genotypes × year interaction showed non-significant interactions for all traits except SOD, TSP, Chl-a content, STR, MIC, and UI. The heat stress × year × genotypes interaction showed non-significant interactions for all traits except H_2O_2, SOD, TSP, Chl-b content, SF, STR, MIC, and UI (**Table 2**). Overall, heat stress negatively affected all agronomic and yield-related traits in all-cotton genotypes during both years. The following traits: PH, TNB, BW, SCY, Lint%, Chl a an b, SF, RD, and UI were reduced in all genotypes under high-temperature stress (**Table 3**). The mean values for H_2O_2, TSP, STR, CAT, SOD, POD, and Car increased under heat stress whereas the mean values for Chl a and b decreased (**Table 3**).

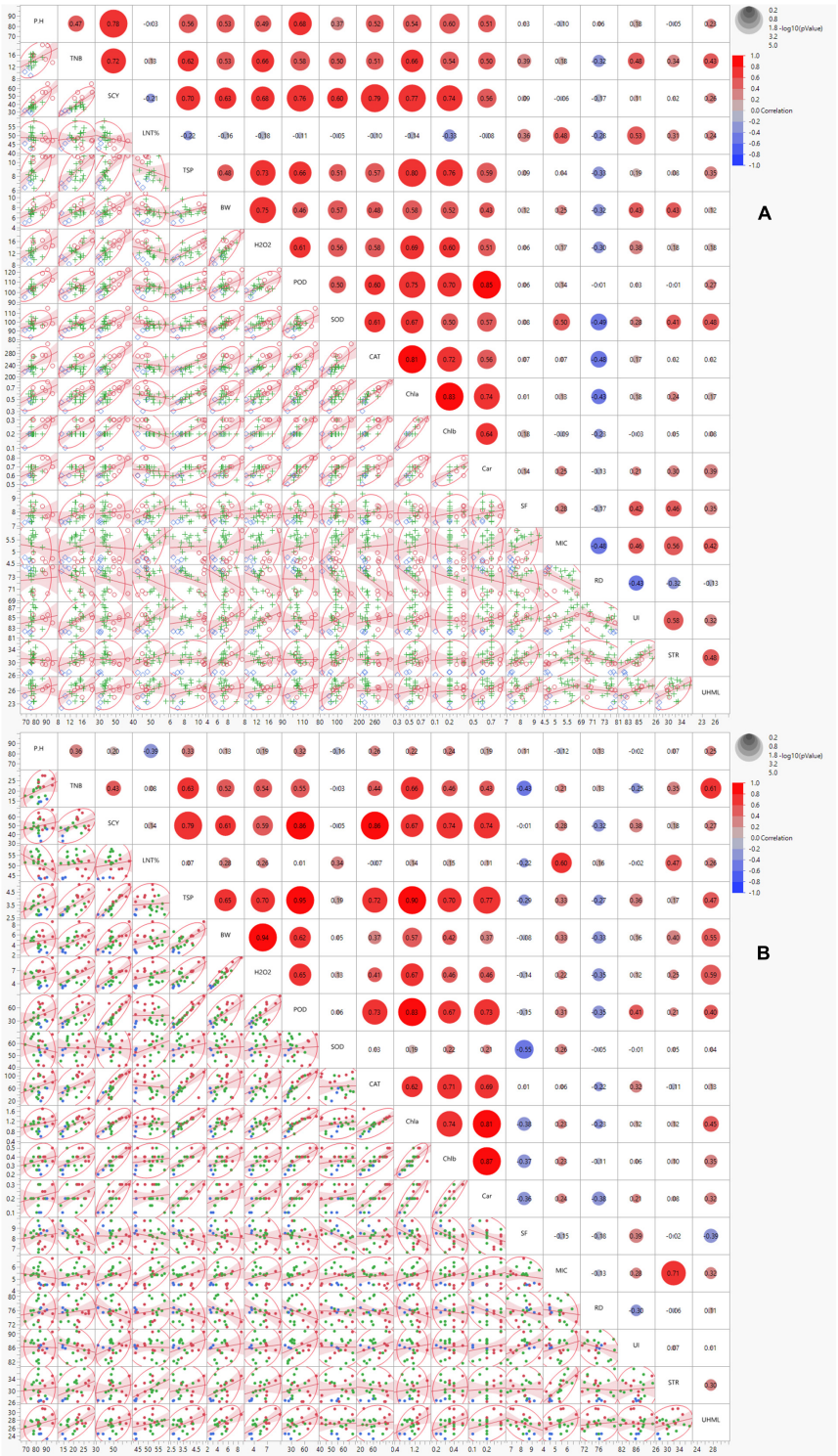

FIGURE 2 | Scatterplot correlation matrix of the 19 ionic yield and fiber-related traits of 23 cotton genotypes grown under normal (left) and high-temperature stress (right) conditions. In the upper panel, red and blue circles indicated positive and negative correlations, respectively, with increasing color intensity reflecting a higher coefficient. The lower panel indicates the bivariate density distributions with ellipses between each pair of traits and trendline of the correlated traits. PH = plant height (cm), TNB = number of bolls, BW = boll weight (g), SCY = seed cotton yield (g), SF = short fiber contents (%), STR = fiber strength (g/tex), UHML = upper half mean length (mm), MIC = micronaire value (unit), RD = reflectance, UI = uniformity index (%), H_2O_2 = hydrogen peroxide (μmol/g), CAT = catalase (U mg^{-1} protein), POD = peroxidase (U mg^{-1} protein), SOD = superoxide dismutase (U mg^{-1} protein), TSP = total soluble protein (mg g^{-1} FW), Chl a and b = chlorophyll contents **(A,B)** (mg g^{-1} FW), Caro = carotenoid (mg g^{-1} FW).

Genetic Components of Various Characters Under Normal and Heat Stress Conditions

The mean values for all traits under normal and heat stress conditions were estimated. Based on these mean values following traits: PH (83.68), TNB (20.63), BW (3.01), UHML (26.89), SCY (50.31), Lint% (50.55), RD (75.38), UI (86.20), MIC (5.12), Chl-a contents (1.12), and Chl-b contents (0.37) exhibited higher mean values under normal conditions. In contrast, H_2O_2 (11.78), POD (105.29), TSP (16.03), SOD (98.01), Car (0.63), CAT (243.21), and STR (32.05) displayed higher mean values under high-temperature conditions (**Table 2**). The coefficient of variation was also computed to determine the precision of the experiment. A lower value of coefficient of variation (CV%) indicated a precise and accurate experiment. The following traits: PH, BW, Lint%, CAT, POD, SOD, SF, MIC, RD, UHML, UI, MIC, and STR had lower coefficient of variation (CV%) values under normal and stress conditions. Under both conditions, the traits TNB, SCY, Car, TSP, Chl a and b, and H_2O_2 had moderate to high coefficient of variation (CV%) values (**Table 3**). The genotypic coefficient of variation (GCV) was observed to be slightly lower than the phenotypic coefficient of variation (PCV) for all the studied traits under both conditions, indicating that the environment was the least influential on these traits. An increasing H^2b and genetic advance mean per percent (GAM) was observed in CAT, MIC, SCY, Chl a and b contents, and PH under both conditions. Moderate H^2b and GAM were exhibited by TNB, H_2O_2, Car, SF, and STR under normal and stress conditions (**Table 3**).

Correlation Analysis

Correlation analysis was performed to estimate the relationship among studied traits under normal and high-temperature stress conditions. The morphological trait, SCY, revealed significant positive associations with TSP, BW, POD, CAT, and Chl a and b under both the conditions. The biochemical trait, TSP, exhibited a significant positive association with H_2O_2, POD, and Chl a and b under both the conditions. BW displayed a highly significant positive association with H_2O_2. The biochemical traits showed a significant positive correlation with POD, CAT, Chl a and b, and Car under both conditions (**Figure 2**). The remainder of the correlations were inconsistently significant under both conditions, and some were insignificant or negatively correlated among themselves under normal and stress conditions. The correlation of PH with most traits in normal conditions was significantly positive, reducing TNB, SCY, TSP, BW, H_2O_2, POD, CAT, Chl a, and b, and Car under control stress conditions. The trait of SOD also reduced its significantly positive correlations with POD, H_2O_2, BW, TSP, CAT, Chl a and b, Car, MIC, STR, and UHML from normal to stress conditions. However, with SCY, TNB, PH, SF, and UI the positive correlations of SOD changed to negative ones under heat stress conditions. The negative correlation of SOD with Lint% changed to a positive correlation under heat stress. The significant positive correlation of UI with STR changed to a non-significant level under the stress condition (**Figure 2**).

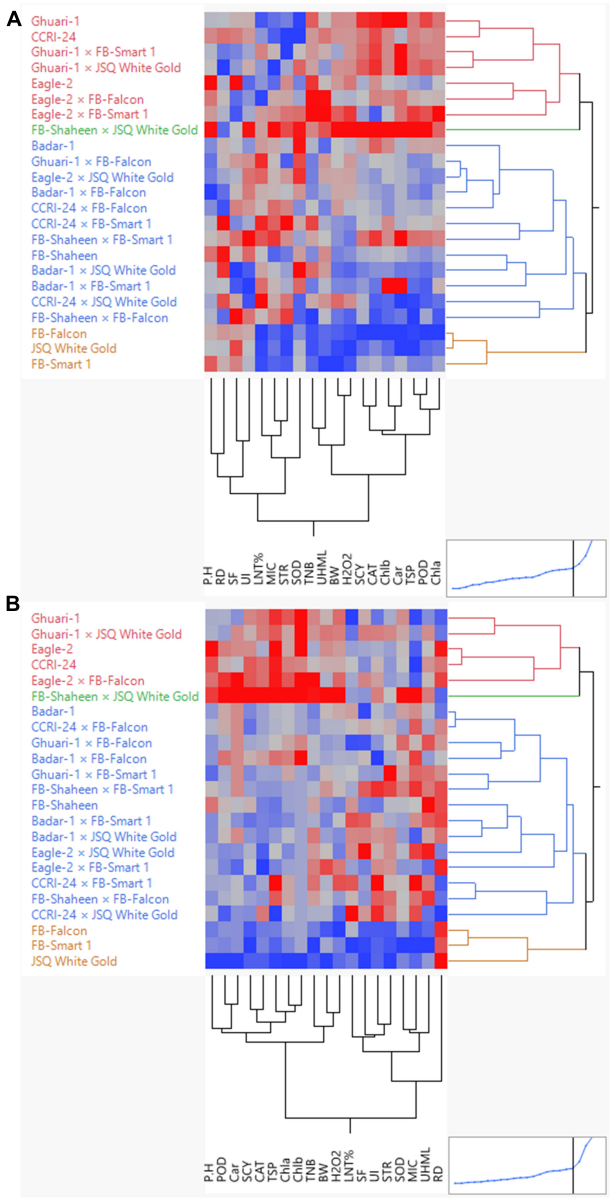

FIGURE 3 | Hierarchical clustering of 23 cotton genotypes for biochemical, yield, and fiber-related traits under normal **(A)** and high-temperature stress **(B)** conditions.

Cluster Analysis

Agglomerative hierarchical clustering (AHC) analysis was performed for the estimation of the degree of dissimilarity among experimental genotypes based on morphological, physiological, and biochemical traits measured under normal and high-temperature stress conditions. The cluster tree was shaped using the agglomerative hierarchical approach based on a "bottom-up" technique. The technique uses every single observation at the initial level as an individual cluster. These individual observations move forward to the next level, forming a hierarchy after pairing up successively until the final distinct cluster. The Euclidean

distance method was used to calculate the distances between genotype pairs. Subsequently, all the genotypes were clustered together to create a full-fledged dendrogram *via* operating Ward's method. A two-way clustering technique was utilized through AHC to build a two-way cluster diagram.

This analysis divided all 23 experimental genotypes into four groups under both normal and heat stress conditions. Under normal conditions, Group-1, Group-2, Group-3, and Group-4 enclosed seven, one, 12, and three genotypes, respectively (**Figure 3**). Under high-temperature stress, these 23 genotypes were clustered again into Group-I, Group-II, and Group-III, comprising five, one, 14, and three genotypes, respectively (**Figure 3**). Different colors represented different clusters. Based on the performance of genotypes, which is depicted in the diagram through a color gradient from red to blue (highest to lowest) obtained from clustering, and under both under normal and heat stress conditions, the following genotypes performed well: FB-SHAHEEN × JSQ WHITE GOLD, CCRI-24, Ghauri-1, Eagle-2 × FB-Falcon, Ghuari-1 × JSQ White Gold, and Eagle-2 regarding agronomic, biochemical, and fiber quality attributes.

Principal Component Analysis

Principal component analysis is a multivariate statistical approach to studying and simplifying complicated and huge datasets. Based on the correlation among studied characters and extracted clusters, the variation patterns in cotton genotypes were also investigated using PCA to assess the genetic diversity of the genotypes and their relationship with the studied traits. Under both conditions, the total variation was divided into 19 principal components (PCs), out of which the first four PCs displayed > 1 eigenvalue. In contrast, the remaining PCs exhibited lower eigenvalues (**Figure 4**). The first four PCs contributed 79.56% to total variability among the cotton genotypes evaluated for various ionic, yield, and fiber quality traits under both conditions. While the remainder of all PCs shared 20.44% of the total variability under both conditions. PC-1 shared 44.3%, PC-2 exhibited 17%, PC-3 revealed 10.8%, and PC-4 displayed 7.46% of total variability among the genotypes for the studied characters. PC-1 contributed the most cumulative variability to the treatment, followed by PC-2, PC-3, and PC-4 (**Figure 4**).

The summary biplot of studied traits along with their magnitudes of variation is displayed in **Figure 4**. All the genotypes under normal and stress conditions were distributed inside the correlation eclipse between the first two PCs (**Figure 4**, left). A relative distance of variables from the origin of PC-1 and PC-2 revealed a contribution of each variable to total variation for the accessions studied. It covered the plot from start to end and provided information about the diversity present among the genotypes. The second summary biplot in **Figure 4** (right) between PC-1 and PC-2 explained 61.3% of the total variation. It reveals that most biochemical traits and a few others between the two PCs were positively correlated with each other: namely Car, CAT, SOD, TSP, POD, H2O2, STR, and BW. The length of vectors originating from the center is a depiction of the correlation amount among traits. These were validation of the correlations mentioned above among the studied traits under both conditions. The TNB, SCY, BW, UHML, Chl a and b, Car, H_2O_2, POD, TSP, and SOD had long vectors and revealed higher variation, whereas lint%, PH, MIC, STR, and UI exhibited the least variability. The SF, UI, and RD did not follow a desirable direction. PCA results displayed clear discrimination among all studied genotypes across normal and high-temperature stress conditions among the four PCs contributing to the maximum. The elaborated distribution details of studied traits under both normal and stress conditions among the four PCs are displayed in a scatterplot matrix in **Figure 5**. In this biplot, Car, CAT, SOD, TSP, POD, H_2O_2, STR, and BW show a positive correlation between PC2 and PC3 biplot, validating the correlation results.

The traits of POD, SOD, and H_2O_2 were closed and positively correlated in the biplot of PC-1 and PC-3. The biplot of PC-4 had the least variation compared with PC-1, with PC-2, and PC-3 being independent. In this biplot, SF and UI were more discriminating traits and had a strong positive correlation. The biplot of PC-1 with PC-4 contributed lower variation as compared with PC-2 and PC-3. In this biplot, RD, UHML, PH, and TNB lay close to each other and exhibited positive associations among themselves.

Stress Tolerance Indices

We have estimated stress tolerance indices (STI) based on mean performance (MP), geometric mean performance (GMP), and STI for test genotypes, considering yield as the most critical indicator for screening regarding heat-tolerant genotypes. In this method, genotypes were ranked according to their MP, GMP, and STI. For MP and GMP genotypes were with higher values, whereas STI ≥ 1 have been considered for heat tolerance (Fernandez, 1992). Out of the 23 studied genotypes, five accessions had an STI value ≥ 1, and the highest values for STI were recorded for FB-Shaheen × JSQ White Gold (1.82), Ghuari-1 (1.27), CCRI-24 (1.26), Eagle-2 × FB-Falcon (1.11), Ghuari-1 × JSQ White Gold (1.04), and Eagle-2 (0.91) (**Supplementary Table 3**). According to the theory proposed by Fernandez (1992), a 3D scatterplot plot was constructed to categorize 23 test genotypes of upland cotton, including lines and their F_1 hybrids; four groups were observed (**Figure 6**). The genotypes categorized as Group A had a relatively consistent performance across normal temperature and heat stress. Group B included accessions with higher performance through normal conditions; as far as group C was concerned, it comprised genotypes having high performance across the stress. In contrast, group D had genotypes with lower performance across both conditions (**Supplementary Table 3**). Based on cluster analysis, PCA, and STI, the genotypes FB-Shaheen × JSQ White Gold, CCRI-24, Ghuari-1, Eagle-2 × FB-Falcon, Ghuari-1 × JSQ White Gold, and Eagle-2 demonstrated superior performance under both conditions, and thus were identified as heat stress-tolerant (**Figure 6** and **Supplementary Table 3**).

DISCUSSION

Of all abiotic stressors, high-temperature stress is a major constraint in improving cotton yield and production, affecting numerous attributes and physiological and metabolic processes

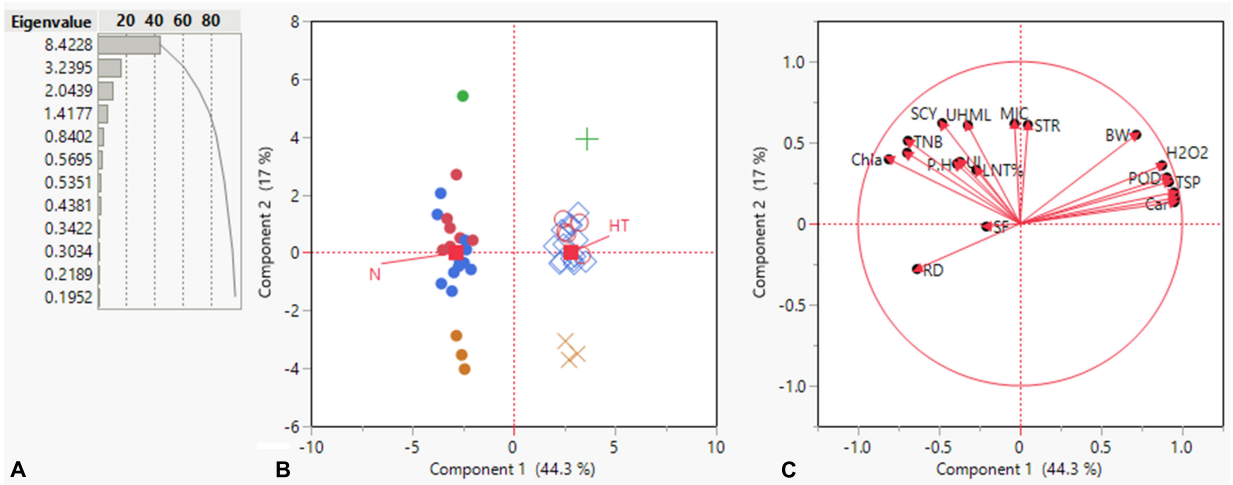

FIGURE 4 | Summary of **(A)** bar chart displaying eigenvalue and variation percentage contribution by all principal components (PCs), **(B)** a biplot between PC1 and PC2 displaying the distribution of 23 cotton genotypes under normal and high-temperature stress conditions **(C)** contribution of various traits in variation for genotypes under normal and high-temperature stress conditions.

(Snider et al., 2011; Xu et al., 2020). The development of high-yielding cotton cultivars with high-temperature resilience are needed to endure the warming global climate. To date, enormous efforts have been made to develop heat-tolerant cotton genotypes. Plant breeders' first choice always relies on the available genetic diversity of various desirable characters among existing germplasm (Pour-Aboughadareh et al., 2018; Majeed et al., 2019). Up-to-date information regarding genetic variability and heritability is necessary to enhance breeding programs in order to develop heat-tolerant cotton cultivars (Tang et al., 1996).

Five lines and three testers were crossed in Line × Tester fashion (5 × 3), and 15 F_1 hybrids were obtained subsequently. Using CV for the studied traits, the variation assists with enhancing crop yields by assembling beneficial genes from genetically divergent genotypes. CV also assists in depiction of the precision regarding the experiment conducted. Genetic variation is highly prone to fluctuations that take place in a plants' environment. As the genome of a plant tries to adapt according to the vagaries of its environment, internal modifications occur to produce desirable, modified, and flexible phenotypes. Those traits exhibiting high GCV and PCV with low adverse environmental effects are advantageous for selection. Cultivars with such characters should be selected to develop desirable and adaptable genotypes (Kaleri et al., 2016; Chaudhari et al., 2017). In this work, phenotypic variance was higher than the genotypic variance for all the studied characters. The GCV was slightly lower than PCV showing the lower environmental variance, which indicates that these characters were less affected by the environment (Singh et al., 2013; Ahmadi et al., 2016). Our findings are also supported by previous reports (Khan et al., 2014).

Genetic improvement of crop plants relies on the magnitude of heritability of economic traits (Ma-Teresa et al., 1994; Ahmadi et al., 2016). Traits with high heritability and genetic advance express their features by being transmitted to the next generation in higher percentages. A high H^2b, coupled with high GAM, may contribute to genetic gain owing to the selection process. Such a trend was observed in this work with CAT, MIC, SCY, PH, and Chl a and b content, indicating additive gene action. The mentioned traits can prove helpful in the selection of genotypes at early stages to be used further in improvement-based breeding programs. Interestingly, similar results were observed in earlier studies (Nawaz et al., 2019; Singh et al., 2019; Bhatti et al., 2020). The following traits: TNB, H_2O_2, Car, SF, and STR, had moderate H^2b and GAM under both normal and stress conditions (Adhikari et al., 2018). Higher PCV, GCV, H^2b, and GAM favor stabilized selection regarding the accumulation of alleles owing to the predominance of additive genes (Jamil et al., 2020). Some studies also suggest that heat tolerance is heritable (Snider et al., 2010, 2011).

Previous findings on high mean values are incongruent with our findings for PH (Majeed et al., 2019), Chl content (Van Der Westhuizen et al., 2020), Lint% (Azhar et al., 2009), BW (Snider et al., 2011), fiber quality traits (Snider et al., 2009), and antioxidant enzymes (Gür et al., 2010; Kamal et al., 2017; Majeed et al., 2019). In this work, as temperature elevated, H_2O_2 production was observed to increase; however, owing to the scavenging activity of CAT and POD, its damaging impacts were prevented (Li et al., 2007; Sekmen et al., 2014). CAT and POD convert the toxic H_2O_2 into water and oxygen (Farooq et al., 2018). An increase in CAT, TSP, and POD contents is generally observed in high-yielding cultivars as these are actively involved in scavenging H_2O_2 to maintain its optimum level (Gosavi et al., 2014; Hussain et al., 2021). The genotypes that showed higher CAT, POD, and TSP values were optimal for H_2O_2 level and were declared as heat-tolerant genotypes. Similar results have been observed previously in cotton (Gür et al., 2010), wheat (Sairam et al., 2000), chickpea (Kaushal et al., 2011), and moth

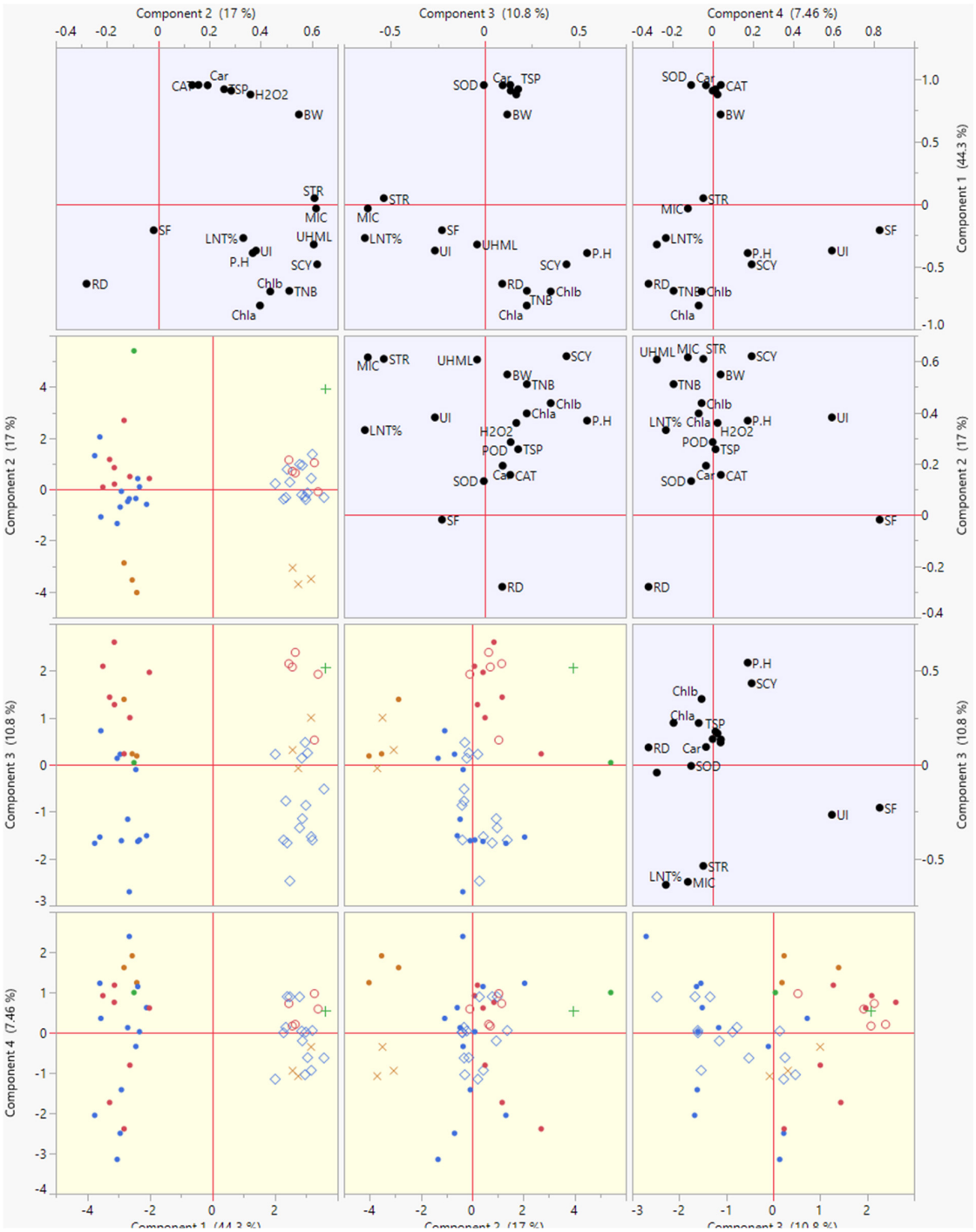

FIGURE 5 | Scatterplot of PC1, PC2, PC3, and PC4 displaying their contribution to the total variability of genotypes based on studied agronomic, biochemical, and fiber-related traits under normal and high temperature stress conditions.

bean (Harsh et al., 2016). Carotenoids also increase under stress conditions as they are mainly involved in safeguarding singlet oxygen (McElroy and Kopsell, 2009).

A correlation matrix is used to study the dependency of variables upon each other for better phenotypes to give improved yields (Li and Ji, 2005; Pour-Aboughadareh et al., 2021). The positively associated traits TSP, BW, POD, CAT, Chl a and b, and Car were in line with earlier reports in cotton (Wan et al., 2007). SCY exhibited a positive relationship with TSP, BW, POD, CAT, and Chl a and b content. BW showed a higher positive

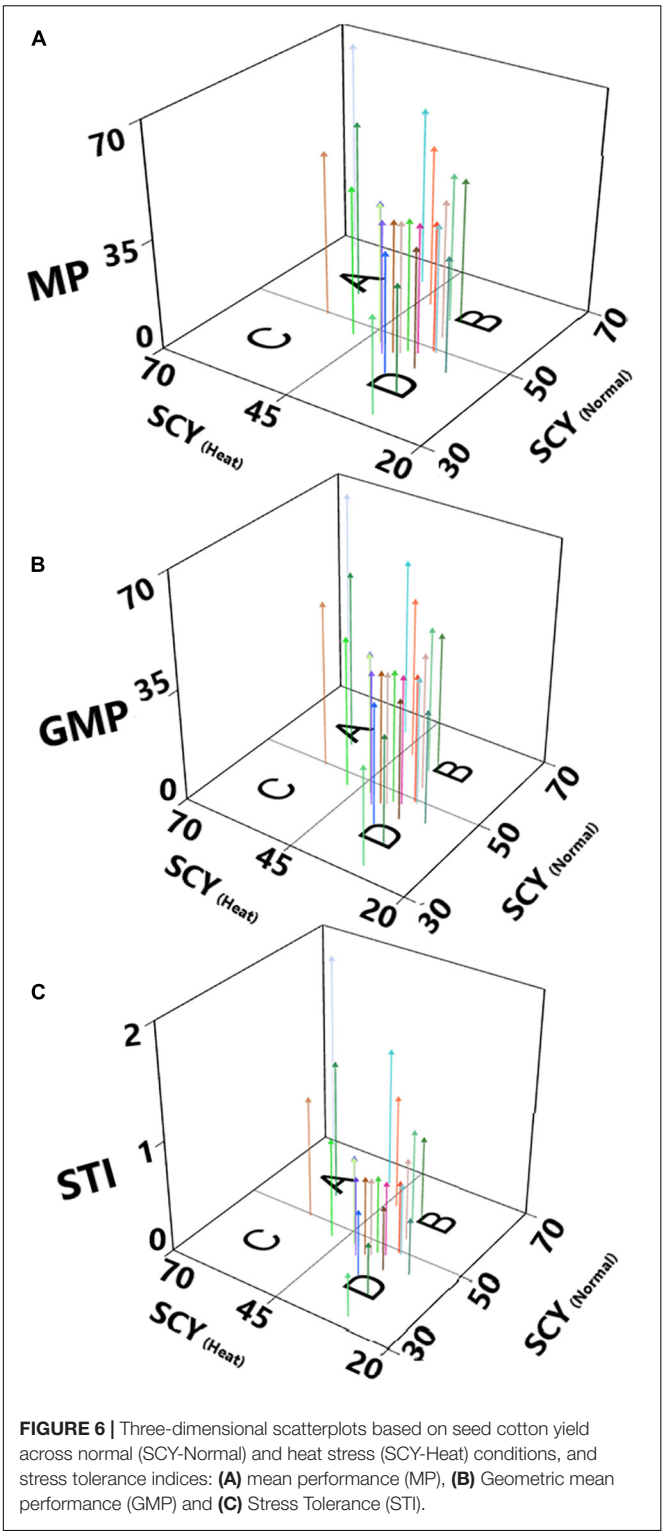

FIGURE 6 | Three-dimensional scatterplots based on seed cotton yield across normal (SCY-Normal) and heat stress (SCY-Heat) conditions, and stress tolerance indices: **(A)** mean performance (MP), **(B)** Geometric mean performance (GMP) and **(C)** Stress Tolerance (STI).

relationship with H_2O_2. Similar positive correlations among traits were also reported in earlier studies (Song et al., 2015; Majeed et al., 2019; Mangi et al., 2021).

In some cotton cultivar leaves, the antioxidant enzymes become upregulated under heat stress but remain unable to safeguard cells from oxidative injury (Snider et al., 2009). This study represented the F_1 hybrid genotype FB-SHAHEEN × JSQ WHITE GOLD with high SCY, BW, TNB, CAT, SOD, POD, and Chl content under both normal and stress conditions. This F_1 hybrid genotype was also observed to be superior in terms of fiber quality traits. The other F_1 hybrids, CCRI-24 × JSQ WHITE GOLD, and EAGLE-2 × JSQ WHITE GOLD showed maximum lint%, whereas the minimum lint% was recorded for GHUARI-1 × FB-FALCON under both conditions. The parental genotypes Eagle-2 and CCRI-24 were superior in yield and fiber quality parameters under both conditions.

To select the best genotypes for agronomic, fiber-related, and biochemical traits, their discrimination from remaining low-performing ones was attained using hierarchical cluster analysis, indicating that they be utilized further in breeding programs (Chunthaburee et al., 2016). The genotypes were clustered into four distinct groups. Group-1 and Group-2 included superior performing genotypes under normal and stress conditions, discriminating them as heat tolerant. Similarly, the PCA analyses revealed the first four PCs as significant contributors to the total variation covering 79.56% toward biochemical, fiber-related, and agronomic traits. These results affirmed the differences among genotypes regarding studied traits under normal and stress conditions, which can prove helpful for their utilization in future breeding programs regarding the improvement of heat tolerance of cotton cultivars. These efficient statistical techniques are employed for the discrimination of genotypes for their diversity evaluation. The results of PCA in the current study are congruent with previous findings on cotton genotypes by other researchers (Saeed et al., 2015; Shabbir et al., 2016; Jamil et al., 2020). Out of the first four PCs the maximum contribution to the total variation residing in the experimental germplasm was from PC1 and PC2, which is in line with earlier reports related to PCA (Amna et al., 2013; Isong et al., 2017). The traits Car, CAT, SOD, TSP, POD, H2O2, STR, and BW, contributed to the first two PCs under both conditions (Javed et al., 2017). Thus, multivariate analyses are a rich source of efficiency, precision, and accuracy regarding the outcomes obtained from experimental studies.

Among the various stress tolerance indices, MP, GMP, and STI have been extensively used in various studies and are suitable selection criteria, as these parameters enable us to identify individuals with high performance regarding stress-tolerance potential under unfavorable conditions (Pour-Aboughadareh et al., 2017). In the same way, many scientists have used these indices in several crops to enable them to assess stress-tolerant genotypes for further utilization in stress breeding programs. These indices have successfully helped to discriminate the genotypes as they revealed a minimal reduction in yield in response to a stress condition, compared with the other studied genotypes. These outcomes align with the findings of other research where these indices distinguished tolerant genotypes from sensitive genotypes (Naghavi et al., 2013; Khalili et al., 2014, 2016, 2018; Etminan et al., 2019; Noorka et al., 2019). Furthermore, the grouping of genotypes for high-temperature tolerance made through these indices is almost the same as we obtained from results of hierarchical clustering and PCA, thus validating the high reliability of the methods used. Hence, tolerant accessions based on STI, AHC, and

PCA results could be grown across higher temperature regions with limited penalties to their growth. The F_1 hybrid FB-SHAHEEN × JSQ WHITE GOLD followed by Ghuari-1, CCRI-24, Eagle-2 × FB-Falcon, Ghuari-1 × JSQ White Gold, and Eagle-2 were identified as more heat tolerant as compared with the remaining experimental genotypes.

Several previous studies have documented that species with higher heat tolerance show an increasing trend in antioxidant enzyme activity in response to high-temperature stress, but susceptible species fail to do so. Thus, the evidence accumulated from current data indicates that intrinsic antioxidant resistance mechanisms of plants may exhibit a strategy for the enhancement of tolerance against heat stress. However, to perform selection efficiently for genetically transformed heat-tolerant plants, the effects of underlying mechanisms under heat stress on plant morphology, physiology, growth, and antioxidative responses must first be identified.

CONCLUSION

The continuously warming global climate drives plant genotypes to adapt through the modification of specific phenotypes. With this scenario of escalating temperature, the development of cultivars that may endure abrupt fluctuations without adversely affecting yield is necessary. The first solution is to screen the available cotton germplasm for its potential against high-temperature stress. Most plants exhibit high antioxidant enzyme activities as an important step involving the heat tolerance mechanism. This work identified that the F_1 hybrid genotype: FB-SHAHEEN × JSQ WHITE GOLD, followed by Ghuari-1, CCRI-24, Eagle-2 × FB-Falcon, Ghuari-1 × JSQ White Gold, and Eagle-2 were the best performers under stress and normal conditions as they were not adversely affected. The adverse effects of heat stress usually include disruption of routine morphological, physiological, biochemical, and fiber characters in cotton and ultimately affect yield. Potential genotypes can be efficiently employed in future cotton breeding programs to improve cotton crop yield and productivity by enhancing their heat tolerance to withstand the changing climate.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

MZ: experimentation, data collection, and drafting the manuscript. XJ and HM: visualization, validation, review, and editing manuscript. AS: conceptualization, resources, supervision, experimentation, review, and editing. ZS: formal analysis, visualization, validation, review, and editing. AM, AI, and AR: experimentation, data acquisition, review, and editing. AA: data acquisition, experimentation, review, and editing. YY: resources, visualization, validation, review, and editing. MI: formal analysis, software, visualization, validation, review, and editing. MR: conceptualization, funding, supervision, validation, review, and editing. All authors have reviewed the manuscript critically and approved the final draft for publication in Frontiers in Plant Science.

FUNDING

This work was supported by the Genetically Modified Organisms Breeding Major Project of China (2019ZX08010004–004) and China's National Natural Science Foundation (31901579).

ACKNOWLEDGMENTS

We are thankful to the Four Brothers Group for providing us with the facilities and materials for the experiment.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.727835/full#supplementary-material

Supplementary Table 1A | Temperature recorded in the tunnel during experiments.

Supplementary Table 1B | Weather data during the crop seasons of 2018 and 2019.

Supplementary Table 2 | Linear mixed-effects model ANOVA for seed cotton yield of 50 cotton genotypes across normal and heat stress conditions.

Supplementary Table 3 | Yield performance of 23 cotton accessions under normal and heat stress conditions, and tolerance and susceptibility indices.

REFERENCES

Adhikari, B. N., Joshi, B. P., Shrestha, J., and Bhatta, N. R. (2018). Genetic variability, heritability, genetic advance and correlation among yield and yield components of rice (Oriza sativa L.). *J. Agric. Nat. Resour.* 1, 14 9–160. doi: 10.3126/janr.v1i1.22230

Ahmadi, J., Vaezi, B., and Pour-Aboughadareh, A. (2016). Analysis of variability, heritability, and interrelationships among grain yield and related characters in barley advanced lines. *Genetika* 48, 73–85. doi: 10.2298/GENSR1601073A

Ajmal, S. U., Minhas, N. M., Hamdani, A., Shakir, A., Zubair, M., and Ahmad, Z. (2013). Multivariate analysis of genetic divergence in wheat (Triticum aestivum) germplasm. *Pak. J. Bot.* 45, 1643–1648.

Ali, M. A., Jabran, K., Awan, S., Abbas, A., Zulkiffal, M., Acet, T., et al. (2011). Morpho-physiological diversity and its implications for improving drought tolerance in grain sorghum at different growth stages. *Austr. J. Crop Sci.* 5:311.

Almeselmani, M., Deshmukh, P., and Sairam, R. (2009). High temperature stress tolerance in wheat genotypes: role of antioxidant defence enzymes. *Acta Agron. Hungari.* 57, 1–14. doi: 10.1556/AAgr.57.2009.1.1

Amna, N., Jehanzeb, F., Abid, M., Muhammad, S., and Muhammad, R. (2013). Estimation of genetic diversity for CLCuV, earliness and fiber quality traits using various statistical procedures in different crosses of Gossypium hirsutum L. *Âåñòíèê Àãðàðíîé Íàóêè* 43, 2–9.

Arnon, D. I. (1949). Copper enzymes in isolated chloroplasts. Polyphenoloxidase in Beta vulgaris. *Plant Physiol.* 24:1. doi: 10.1104/pp.24.1.1

Aslam, M., Maqbool, M. A., Zaman, Q. U., Shahid, M., Akhtar, M. A., and Rana, A. S. (2017). Comparison of different tolerance indices and PCA biplot analysis for assessment of salinity tolerance in lentil (Lens culinaris) genotypes. *Int. J. Agric. Biol.* 19, 470-478. doi: 10.17957/IJAB/15.0308

ASTM (2005). *ASTM D5867-05, Standard Test Methods for Measurement of Physical Properties of Cotton Fibers by High Volume Instruments*. West Conshohocken: ASTM International.

Azhar, F., Ali, Z., Akhtar, M., Khan, A., and Trethowan, R. (2009). Genetic variability of heat tolerance, and its effect on yield and fibre quality traits in upland cotton (Gossypium hirsutum L.). *Plant Breed.* 128, 356–362. doi: 10.1111/j.1439-0523.2008.01574.x

Bhatti, M. H., Yousaf, M. I., Ghani, A., Arshad, M., and Shehzad, A. A. (2020). Assessment of genetic variability and traits association in upland cotton (Gossypium hirsutum L.). *Int. J. Bot. studies.* 5, 148–151.

Both, A.-J., Benjamin, L., Franklin, J., Holroyd, G., Incoll, L. D., Lefsrud, M. G., et al. (2015). Guidelines for measuring and reporting environmental parameters for experiments in greenhouses. *Plant Methods* 11:43.

Bradford, M. M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* 72, 248–254. doi: 10.1016/0003-2697(76)90527-3

Burke, J. J., Velten, J., and Oliver, M. J. (2004). In vitro analysis of cotton pollen germination. *Agron. J.* 96, 359–368. doi: 10.2134/agronj2004.3590

Chaudhari, M., Faldu, G., and Ramani, H. J. A. I. B. (2017). Genetic variability, Correlation and Path coefficient analysis in cotton (*Gossypium hirsutum* L.). *Adv. Biores.* 8, 226–233.

Choudhury, S., Panda, P., Sahoo, L., and Panda, S. K. (2013). Reactive oxygen species signaling in plants under abiotic stress. *Plant Signal. Behav.* 8:e23681. doi: 10.4161/psb.23681

Chunthaburee, S., Dongsansuk, A., Sanitchon, J., Pattanagul, W., and Theerakulpisut, P. (2016). Physiological and biochemical parameters for evaluation and clustering of rice cultivars differing in salt tolerance at seedling stage. *Saudi J. Biol. Sci.* 23, 467–477. doi: 10.1016/j.sjbs.2015.05.013

Conaty, W., Burke, J., Mahan, J., Neilsen, J., and Sutton, B. (2012). Determining the optimum plant temperature of cotton physiology and yield to improve plant-based irrigation scheduling. *Crop Sci.* 52, 1828–1836. doi: 10.2135/cropsci2011.11.0581

Dabbert, T., and Gore, M. A. (2014). Challenges and perspectives on improving heat and drought stress resilience in cotton. *J. Cotton Sci.* 18, 393–409.

Dhamayanthi, K., Manivannan, A., and Saravanan, M. (2018). Evaluation of new germplasm of Egyptian cotton (G. barbadense) through multivariate genetic component analysis. *Electr. J. Plant Breed.* 9, 1348–1354. doi: 10.5958/0975-928X.2018.00168.0

Etminan, A., Pour-Aboughadareh, A., Mohammadi, R., Shooshtari, L., Yousefiazarkhanian, M., and Moradkhani, H. (2019). Determining the best drought tolerance indices using artificial neural network (ANN): insight into application of intelligent agriculture in agronomy and plant breeding. *Cereal Res. Commun.* 47, 170–181. doi: 10.1556/0806.46.2018.057

Farooq, M. A., Shakeel, A., Atif, R. M., and Saleem, M. F. (2018). Genetic Variability Studies for Salinity Tolerance in Gossypium hirsutum. *Int. J. Agric. Biol.* 20, 2871–2878.

Fernandez, G. C. (1992). "Effective selection criteria for assessing plant stress tolerance," in *Proceeding of the International Symposium on Adaptation of Vegetables and other Food Crops in Temperature and Water Stress, Aug. 13-16, 1992,* Taiwan, 257–270.

Ghafoor, G., Hassan, G., Ahmad, I., Khan, S. N., and Suliman, S. (2013). Correlation analysis for different parameters of F2 bread wheat population. *Pure Appl. Biol.* 2:28. doi: 10.19045/bspab.2013.21005

Gomez, K. A., and Gomez, A. A. (1984). *Statistical Procedures for Agricultural Research*. United States: John Wiley & Sons.

Goos, P., and Meintrup, D. (2016). *Statistics with JMP: hypothesis Tests, ANOVA and Regression*. United States: John Wiley & Sons.

Gosavi, G., Jadhav, A., Kale, A., Gadakh, S., Pawar, B., and Chimote, V. (2014). *Effect of Heat Stress on Proline, Chlorophyll Content, Heat Shock Proteins and Antioxidant Enzyme Activity in Sorghum (Sorghum bicolor) at Seedlings Stage*. India: NISCAIR-CSIR.

Gür, A., Demirel, U., Özden, M., Kahraman, A., and Çopur, O. (2010). Diurnal gradual heat stress affects antioxidant enzymes, proline accumulation and some physiological components in cotton (Gossypium hirsutum L.). *Afr. J. Biotechnol.* 9, 1008–1015. doi: 10.5897/AJB09.1590

Harsh, A., Sharma, Y., Joshi, U., Rampuria, S., Singh, G., Kumar, S., et al. (2016). Effect of short-term heat stress on total sugars, proline and some antioxidant enzymes in moth bean (Vigna aconitifolia). *Ann. Agric. Sci.* 61, 57–64. doi: 10.1016/j.aoas.2016.02.001

Hoaglin, D. C., and Welsch, R. E. (1978). The hat matrix in regression and ANOVA. *Am. Statist.* 32, 17–22. doi: 10.1080/00031305.1978.10479237

Hussain, R., Ayyub, C. M., Shaheen, M. R., Rashid, S., Nafees, M., Ali, S., et al. (2021). Regulation of Osmotic Balance and Increased Antioxidant Activities under Heat Stress in Abelmoschus esculentus L. Triggered by Exogenous Proline Applications. *Agronomy* 11:685. doi: 10.3390/agronomy11040685

Isong, A., Balu, P. A., and Ramakrishnan, P. (2017). Association and principal component analysis of yield and its components in cultivated cotton. *Electr. J. Plant Breed.* 8, 857–864. doi: 10.5958/0975-928X.2017.00140.5

Jamil, A., Khan, S. J., and Ullah, K. (2020). Genetic diversity for cell membrane thermostability, yield and quality attributes in cotton (Gossypium hirsutum L.). *Genet. Resour. Crop Evol.* 67, 1405–1414. doi: 10.1007/s10722-020-00911-w

Jarwar, A. H., Wang, X., Iqbal, M. S., Sarfraz, Z., Wang, L., Ma, Q., et al. (2019). Genetic divergence on the basis of principal component, correlation and cluster analysis of yield and quality traits in cotton cultivars. *Pak. J. Bot.* 51, 1143–1148. doi: 10.30848/PJB2019-3(38)

Javed, M., Hussain, S., and Baber, M. (2017). Assessment of genetic diversity of cotton genotypes for various economic traits against cotton leaf curl disease (CLCuD). *Genet. Mol. Res* 16, 1–12. doi: 10.4238/gmr16019446

Johnson, H. W., Robinson, H., and Comstock, R. (1955). Estimates of genetic and environmental variability in soybeans 1. *Agron. J.* 47, 314–318. doi: 10.2134/agronj1955.00021962004700070009x

Kaleri, A. A., Baloch, A. W., Baloch, M., Wahocho, N. A., Abro, T. F., Jogi, Q., et al. (2016). Heritability and correlation analysis in Bt and non-Bt cotton (Gossypium hirsutum L.) genotypes. *Pure Appl. Biol.* 5:1. doi: 10.19045/bspab.2016.50114

Kamal, M., Saleem, M., Shahid, M., Awais, M., Khan, H., and Ahmed, K. (2017). Ascorbic acid triggered physiochemical transformations at different phenological stages of heat-stressed Bt cotton. *J. Agron. Crop Sci.* 203, 323–331. doi: 10.1111/jac.12211

Kaushal, N., Gupta, K., Bhandhari, K., Kumar, S., Thakur, P., and Nayyar, H. (2011). Proline induces heat tolerance in chickpea (Cicer arietinum L.) plants by protecting vital enzymes of carbon and antioxidative metabolism. *Physiol. Mol. Biol. Plants* 17, 203–213. doi: 10.1007/s12298-011-0078-2

Ketata, H., Yau, S., and Nachit, M. (1989). "Relative consistency performance across environments," in *International symposium on physiology and breeding of winter cereals for stressed mediterranean environments*. (Paris: INRA).

Khalili, M., Alireza, P.-A., Naghavi, M. R., and Mohammad-Amini, E. (2014). Evaluation of drought tolerance in safflower genotypes based on drought tolerance indices. *Notul. Bot. Horti Agrobotan. Cluj-Napoca* 42, 214–218. doi: 10.15835/nbha4219331

Khalili, M., Pour-Aboughadareh, A., and Naghavi, M. R. (2016). Assessment of drought tolerance in barley: integrated selection criterion and drought tolerance indices. *Environ. Exp. Biol.* 14, 33–41. doi: 10.22364/eeb.14.06

Khalili, M., Zhang, X., Polycarpou, M. M., Parisini, T., and Cao, Y. (2018). Distributed adaptive fault-tolerant control of uncertain multi-agent systems. *Automatica* 87, 142–151. doi: 10.1016/j.automatica.2017.09.002

Khan, M. A., Wahid, A., Ahmad, M., Tahir, M. T., Ahmed, M., Ahmad, S., et al. (2020). World cotton production and consumption: an overview. *Cotton Prod. Uses* 2020, 1–7. doi: 10.1007/978-981-15-1472-2_1

Khan, N., Azhar, F. M., Khan, A., and Ahmad, R. (2014). Measurement of canopy temperature for heat tolerance in upland cotton: variability and its genetic basis. *Pak. J. Agri. Sci* 51, 359–365.

Kocsy, G., Szalai, G., Sutka, J., Páldi, E., and Galiba, G. (2004). Heat tolerance together with heat stress-induced changes in glutathione and

hydroxymethylglutathione levels is affected by chromosome 5A of wheat. *Plant Sci.* 166, 451–458. doi: 10.1016/j.plantsci.2003.10.011

Li, H. B., Qin, Y. M., Pang, Y., Song, W. Q., Mei, W. Q., and Zhu, Y. X. (2007). A cotton ascorbate peroxidase is involved in hydrogen peroxide homeostasis during fibre cell development. *New Phytol.* 175, 462–471. doi: 10.1111/j.1469-8137.2007.02120.x

Li, J., and Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* 95, 221–227. doi: 10.1038/sj.hdy.6800717

Liu, D., Zou, J., Meng, Q., Zou, J., and Jiang, W. (2009). Uptake and accumulation and oxidative stress in garlic (Allium sativum L.) under lead phytotoxicity. *Ecotoxicology* 18, 134–143. doi: 10.1007/s10646-008-0266-1

Lokhande, S., and Reddy, K. R. (2014). Quantifying temperature effects on cotton reproductive efficiency and fiber quality. *Agron. J.* 106, 1275–1282. doi: 10.2134/agronj13.0531

Majeed, S., Malik, T. A., Rana, I. A., and Azhar, M. T. (2019). Antioxidant and physiological responses of upland cotton accessions grown under high-temperature regimes. *Iran. J. Sci. Technol. Transac. A Sci.* 43, 2759–2768. doi: 10.1007/s40995-019-00781-7

Malik, R., Sharma, H., Sharma, I., Kundu, S., Verma, A., Sheoran, S., et al. (2014). Genetic diversity of agro-morphological characters in Indian wheat varieties using GT biplot. *Austr. J. Crop Sci.* 8:1266.

Mangi, N., Nazir, M. F., Wang, X., Iqbal, M. S., Sarfraz, Z., Jatoi, G. H., et al. (2021). Dissecting Source-Sink Relationship of Subtending Leaf for Yield and Fiber Quality Attributes in Upland Cotton (*Gossypium hirsutum* L.). *Plants* 10:1147. doi: 10.3390/plants10061147

Ma-Teresa, L., Gerpcio-Sta, C., and Enrique, C. (1994). Heritability estimates of some root characters in sweetpotato. *Philipp. J. Crop Sci.* 19, 27–32.

McElroy, J. S., and Kopsell, D. A. (2009). Physiological role of carotenoids and other antioxidants in plants and application to turfgrass stress management. *New Zeal. J. Crop Hortic. Sci.* 37, 327–333. doi: 10.1080/01140671.2009.9687587

Mohammadi, S. A., and Prasanna, B. (2003). Analysis of genetic diversity in crop plants—salient statistical tools and considerations. *Crop Sci.* 43, 1235–1248. doi: 10.2135/cropsci2003.1235

Muhammad, A., Amir, S., Khan, T. M., and Irfan, A. (2018). Genetic basis of variation for high temperature tolerance in upland cotton. *Int. J. Agric. Biol.* 20, 2637–2646.

Naghavi, M. R., Aboughadareh, A. P., and Khalili, M. (2013). Evaluation of drought tolerance indices for screening some of corn (*Zea mays* L.) cultivars under environmental conditions. *Not. Sci. Biol.* 5, 388–393. doi: 10.15835/nsb539049

Nawaz, B., Naeem, M., Malik, T. A., Muhae-Ud-Din, G., Ahmad, Q., and Sattar, S. (2019). Estimation of Gene Action, Heritability and Pattern of Association among Different Yield Related Traits in Upland Cotton. *Int. Multidiscip. Res. J.* 9, 13–19. doi: 10.29329/ijiaar.2019.188.4

Nelson, P. R., Wludyka, P. S., and Copeland, K. A. (2005). *The Analysis of Means: a Graphical Method for Comparing Means, Rates, and Proportions*. Philadelphia: SIAM. doi: 10.1137/1.9780898718362

Noorka, I. R., Iqbal, M. S., ÖztÃijrk, M., Shahid, M. R., and Khaliq, I. (2019). *Cotton, White Gold of Pakistan: An Efficient Technique for Bumper Crop Production.* Boca Raton, FL: Apple Academic Press, Inc, 87–96.

Patel, J., Lubbers, E., Kothari, N., Koebernick, J., and Chee, P. (2021). Genetics and Genomics of Cottonseed Oil. *Oil Crop Genomics* 2021, 53–74. doi: 10.1007/978-3-030-70420-9_3

Pettigrew, W. (2008). The effect of higher temperatures on cotton lint yield production and fiber quality. *Crop Sci.* 48, 278–285. doi: 10.2135/cropsci2007.05.0261

Poehlman, J., and Sleper, D. (1995). *Breeding Field Crops* 4th Edn. New Dehli: Panima Publishing Corporation.

Pour-Aboughadareh, A., Ahmadi, J., Mehrabi, A. A., Etminan, A., Moghaddam, M., and Siddique, K. H. (2017). Physiological responses to drought stress in wild relatives of wheat: implications for wheat improvement. *Acta Physiol. Plant.* 39:106.

Pour-Aboughadareh, A., Ahmadi, J., Mehrabi, A. A., Etminan, A., and Moghaddam, M. (2018). Insight into the genetic variability analysis and relationships among some Aegilops and Triticum species, as genome progenitors of bread wheat, using SCoT markers. *Plant Biosyst. Int. J. Deal. all Aspects Plant Biol.* 152, 694–703. doi: 10.1080/11263504.2017.1320311

Pour-Aboughadareh, A., Sanjani, S., Nikkhah-Chamanabad, H., Mehrvar, M. R., Asadi, A., and Amini, A. (2021). Identification of salt-tolerant barley genotypes using multiple-traits index and yield performance at the early growth and maturity stages. *Bull. Natl. Res. Centre* 45:117. doi: 10.1186/s42269-021-00576-0

Pour-Aboughadareh, A., Yousefian, M., Moradkhani, H., Moghaddam, V. M., Poczai, P., and Siddique, K. H. (2019). iPASTIC: an online toolkit to estimate plant abiotic stress indices. *Appl. Plant Sci.* 7:e11278. doi: 10.1002/aps3.11278

Rafiq, A., Iqbal, M. S., Ibrar, D., Mahmood, T., Naveed, M. S., and Naeem, M. K. (2013). A review on heat stress response in different genotypes of tomato crop (Solanum lycopersicon L.). *Int. J. Mod. Agri.* 2, 64–71.

Rathinavel, K. (2018). Principal Component Analysis with Quantitative Traits in Extant Cotton Varieties (Gossypium Hirsutum L.) and Parental Lines for Diversity. *Curr. Agric. Rese. J.* 6, 54–64. doi: 10.12944/CARJ.6.1.07

Roychoudhury, A., Basu, S., and Sengupta, D. N. (2012). Antioxidants and stress-related metabolites in the seedlings of two indica rice varieties exposed to cadmium chloride toxicity. *Acta Physiol. Plantar.* 34, 835–847. doi: 10.1007/s11738-011-0881-y

Saeed, F., Shabbir, R. H., Farooq, J., Riaz, M., and Mahmood, K. (2015). Genetic Diversity Analysis for Earliness, Fiber Quality and Cotton Leaf Curl Virus in *Gossypium hirsutum* L. Accessions. *Cotton Genomics Genet.*6, 1–7.

Sairam, R., Srivastava, G., and Saxena, D. (2000). Increased antioxidant activity under elevated temperatures: a mechanism of heat stress tolerance in wheat genotypes. *Biol. Plantar.* 43, 245–251. doi: 10.1023/A:1002756311146

Saleem, M. A., Malik, W., Qayyum, A., Ul-Allah, S., Ahmad, M. Q., Afzal, H., et al. (2021). Impact of heat stress responsive factors on growth and physiology of cotton (Gossypium hirsutum L.). *Mol. Biol. Rep.* 48, 1069–1079. doi: 10.1007/s11033-021-06217-z

Salimath, S. S., Romsdahl, T. B., Konda, A. R., Zhang, W., Cahoon, E. B., Dowd, M. K., et al. (2021). Production of tocotrienols in seeds of cotton (Gossypium hirsutum L.) enhances oxidative stability and offers nutraceutical potential. *Plant Biotechnol. J.* 19, 1268–1282. doi: 10.1111/pbi.13557

Salman, M., Majeed, S., Rana, I. A., Atif, R. M., and Azhar, M. T. (2019). *Novel Breeding and Biotechnological Approaches to Mitigate the Effects of Heat Stress on Cotton*. Germany: Springer. 251–277. doi: 10.1007/978-3-030-21687-0_11

Sarwar, M., Saleem, M., Najeeb, U., Shakeel, A., Ali, S., and Bilal, M. (2017). Hydrogen peroxide reduces heat-induced yield losses in cotton (G ossypium hirsutum L.) by protecting cellular membrane damage. *J. Agron. Crop Sci.* 203, 429–441. doi: 10.1111/jac.12203

Sarwar, M., Saleem, M. F., Ullah, N., Ali, S., Rizwan, M., Shahid, M. R., et al. (2019). Role of mineral nutrition in alleviation of heat stress in cotton plants grown in glasshouse and field conditions. *Sci. Rep.* 9:13022.

Sekmen, A. H., Ozgur, R., Uzilday, B., and Turkan, I. (2014). Reactive oxygen species scavenging capacities of cotton (Gossypium hirsutum) cultivars under combined drought and heat induced oxidative stress. *Environ. Exp. Bot.* 99, 141–149. doi: 10.1016/j.envexpbot.2013.11.010

Sellam, V., and Poovammal, E. (2016). Prediction of crop yield using regression analysis. *Ind. J. Sci. Technol.* 9, 1–5. doi: 10.17485/ijst/2016/v9i38/91714

Shabbir, R. H., Bashir, Q. A., Shakeel, A., Khan, M. M., Farooq, J., Fiaz, S., et al. (2016). Genetic divergence assessment in upland cotton (Gossypium hirsutum L.) using various statistical tools. *J. Global Innov. Agric. Soc. Sci.* 4, 62–69. doi: 10.22194/JGIASS/4.2.744

Singh, R. P., Prasad, P. V., Sunita, K., Giri, S., and Reddy, K. R. (2007). Influence of high temperature and breeding for heat tolerance in cotton: a review. *Adv. Agron.* 93, 313–385.

Singh, D., Gill, J., Gumber, R., Singh, R., and Singh, S. (2013). Yield and fibre quality associated with cotton leaf curl disease of Bt-cotton in Punjab. *J. Environ. Biol.* 34:113.

Singh, M., Singh, V., Yadav, G., and Kumar, P. (2019). Studies on variability, heritability (narrow sense) and genetic advance analysis for growth, yield and quality traits in pumpkin (Cucurbita moschata Duch. ex. Poir). *J. Pharmacogn. Phytochem.* 8, 3621–3624. doi: 10.20546/ijcmas.2019.807.120

Singh, R., and Chaudhary, B. (1985). *Biometrical Methods in Quantitative Genetic Analysis*. India: Kalyani Publishers.

Snider, J., Oosterhuis, D., and Kawakami, E. (2011). Mechanisms of reproductive thermotolerance in Gossypium hirsutum: the effect of genotype and exogenous

calcium application. *J. Agron. Crop Sci.* 197, 228–236. doi: 10.1111/j.1439-037X.2010.00457.x

Snider, J. L., Oosterhuis, D. M., and Kawakami, E. M. (2010). Genotypic differences in thermotolerance are dependent upon prestress capacity for antioxidant protection of the photosynthetic apparatus in Gossypium hirsutum. *Physiol. Plantar.* 138, 268–277. doi: 10.1111/j.1399-3054.2009.01325.x

Snider, J. L., Oosterhuis, D. M., Skulman, B. W., and Kawakami, E. M. (2009). Heat stress-induced limitations to reproductive success in Gossypium hirsutum. *Physiol. Plantar.* 137, 125–138. doi: 10.1111/j.1399-3054.2009.01266.x

Song, M., Fan, S., Pang, C., Wei, H., Liu, J., and Yu, S. (2015). Genetic analysis of yield and yield-related traits in short-season cotton (Gossypium hirsutum L.). *Euphytica* 204, 135–147. doi: 10.1007/s10681-014-1348-1

Song, M., Fan, S., Pang, C., Wei, H., and Yu, S. (2014). Genetic analysis of the antioxidant enzymes, methane dicarboxylic aldehyde (MDA) and chlorophyll content in leaves of the short season cotton (Gossypium hirsutum L.). *Euphytica* 198, 153–162. doi: 10.1007/s10681-014-1100-x

Steel, R. G., Torrie, J. H., and Dickey, D. A. (1997). *Principles and Procedures of Statistics: a Biological Approach*. United States: McGraw-Hill.

Stewart, S., and Thomas, M. O. (2008). "Student learning of basis in linear algebra" in *Proceedings of the Joint Conference of PME*. (New Zealand: The University of Auckland). 281–288.

Suzuki, N., Koussevitzky, S., Mittler, R., and Miller, G. (2012). ROS and redox signalling in the response of plants to abiotic stress. *Plant Cell Environ.* 35, 259–270. doi: 10.1111/j.1365-3040.2011.02336.x

Tang, B., Jenkins, J., Watson, C., Mccarty, J., and Creech, R. (1996). Evaluation of genetic variances, heritabilities, and correlations for yield and fiber traits among cotton F 2 hybrid populations. *Euphytica* 91, 315–322. doi: 10.1007/BF00033093

Teixeira, E. I., Fischer, G., Van Velthuizen, H., Walter, C., and Ewert, F. (2013). Global hot-spots of heat stress on agricultural crops due to climate change. *Agric. For. Meteorol.* 170, 206–215. doi: 10.1016/j.agrformet.2011.09.002

Timm, N. H. (2002). *Applied Multivariate Analysis*. Germany: Springer.

Van Der Westhuizen, M., Oosterhuis, D., Berner, J., and Boogaers, N. (2020). Chlorophyll a fluorescence as an indicator of heat stress in cotton (Gossypium hirsutum L.). *South Afr. J. Plant Soil* 37, 116–119. doi: 10.1080/02571862.2019.1665721

Velikova, V., Yordanov, I., and Edreva, A. (2000). Oxidative stress and some antioxidant systems in acid rain-treated bean plants: protective role of exogenous polyamines. *Plant Sci.* 151, 59–66. doi: 10.1016/S0168-9452(99)00197-1

Wan, Q., Zhang, Z., Hu, M., Chen, L., Liu, D., Chen, X., et al. (2007). T 1 locus in cotton is the candidate gene affecting lint percentage, fiber quality and spiny bollworm (Earias spp.) resistance. *Euphytica* 158, 241–247. doi: 10.1007/s10681-007-9446-y

Xu, W., Zhou, Z., Zhan, D., Zhao, W., Meng, Y., Chen, B., et al. (2020). The difference in the formation of thermotolerance of two cotton cultivars with different heat tolerance. *Arch. Agron. Soil Sci.* 66, 58–69. doi: 10.1080/03650340.2019.1593967

Yiğit, S., and Mendeş, M. (2017). ANOM technique for evaluating practical significance of observed difference among treatment groups. *Int. J. Agric. Sci. Res.* 6, 1–7.

Zahid, K. R., Ali, F., Shah, F., Younas, M., Shah, T., Shahwar, D., et al. (2016). Response and tolerance mechanism of cotton Gossypium hirsutum L. to elevated temperature stress: a review. *Front. Plant Sci.* 7:937. doi: 10.3389/fpls.2016.00937

Zhang, L., Lei, L., and Yan, D. (2010). "Comparison of two regression models for predicting crop yield" in *2010 IEEE International Geoscience and Remote Sensing Symposium*. (United States: IEEE). 1521–1524. doi: 10.1109/IGARSS.2010.5652764

Zhou, J., Zhou, J., Ye, H., Ali, M. L., Chen, P., and Nguyen, H. T. (2021). Yield estimation of soybean breeding lines under drought stress using unmanned aerial vehicle-based imagery and convolutional neural network. *Biosyst. Eng.* 204, 90–103. doi: 10.1016/j.biosystemseng.2021.01.017

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Plant Science

METHODS
published: 21 January 2022
doi: 10.3389/fpls.2021.791859

Breeding Schemes: What Are They, How to Formalize Them, and How to Improve Them?

Giovanny Covarrubias-Pazaran[1,2], Zelalem Gebeyehu[2], Dorcus Gemenet[1,3], Christian Werner[1,3], Marlee Labroo[1,3], Solomon Sirak[1], Peter Coaldrake[1], Ismail Rabbi[4], Siraj Ismail Kayondo[4], Elizabeth Parkes[4], Edward Kanju[4], Edwige Gaby Nkouaya Mbanjo[4], Afolabi Agbona[4], Peter Kulakow[4], Michael Quinn[1,3] and Jan Debaene[1,3]*

[1] Excellence in Breeding Platform, Consultative Group on International Agricultural Research, Texcoco, Mexico, [2] Independent Researcher, Addis Ababa, Ethiopia, [3] International Maize and Wheat Improvement Center (CIMMYT), Texcoco, Mexico, [4] International Institute for Tropical Agriculture (IITA), Ibadan, Nigeria

OPEN ACCESS

Edited by:
Diego Rubiales, Institute for Sustainable Agriculture, Spanish National Research Council (CSIC), Spain

Reviewed by:
Maryke T. Labuschagne, University of the Free State, South Africa
Valheria Castiblanco, International Center for Tropical Agriculture (CIAT), Colombia

***Correspondence:**
Giovanny Covarrubias-Pazaran
g.covarrubias@cgiar.org;
covaruberpaz@gmail.com

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *09 October 2021*
Accepted: *10 December 2021*
Published: *21 January 2022*

Citation:
Covarrubias-Pazaran G, Gebeyehu Z, Gemenet D, Werner C, Labroo M, Sirak S, Coaldrake P, Rabbi I, Kayondo SI, Parkes E, Kanju E, Mbanjo EGN, Agbona A, Kulakow P, Quinn M and Debaene J (2022) Breeding Schemes: What Are They, How to Formalize Them, and How to Improve Them? Front. Plant Sci. 12:791859. doi: 10.3389/fpls.2021.791859

Formalized breeding schemes are a key component of breeding program design and a gateway to conducting plant breeding as a quantitative process. Unfortunately, breeding schemes are rarely defined, expressed in a quantifiable format, or stored in a database. Furthermore, the continuous review and improvement of breeding schemes is not routinely conducted in many breeding programs. Given the rapid development of novel breeding methodologies, it is important to adopt a philosophy of continuous improvement regarding breeding scheme design. Here, we discuss terms and definitions that are relevant to formalizing breeding pipelines, market segments and breeding schemes, and we present a software tool, Breeding Pipeline Manager, that can be used to formalize and continuously improve breeding schemes. In addition, we detail the use of continuous improvement methods and tools such as genetic simulation through a case study in the International Institute of Tropical Agriculture (IITA) Cassava east-Africa pipeline. We successfully deploy these tools and methods to optimize the program size as well as allocation of resources to the number of parents used, number of crosses made, and number of progeny produced. We propose a structured approach to improve breeding schemes which will help to sustain the rates of response to selection and help to deliver better products to farmers and consumers.

Keywords: breeding scheme, breeding pipeline, market segment, product profile, continuous improvement, genetic simulation

INTRODUCTION

A breeding program is the sum of breeding pipelines to achieve breeding targets for a set of market/target segments[1] Only after rigorous market and social studies have been carried out and an impactful pipeline investment case is presented to the leadership of an organization/institution, a breeding pipeline is created to carry out trait discovery, population improvement, product development, introgression efforts, seed dissemination/commercialization or a combination of one or several of these (tiers). Any pipeline should have a clear deliverable/product to be handed at the end of the pipeline and a clear customer (another pipeline lead, another organization, etc.). A market segment is defined by the target population of environments in which the final

[1]Breeding Pipeline (2021). *Breeding Pipeline: Scope and Approach*. Available online at: https://globalrust.org/dggw/breeding-pipeline

product is grown, as well as descriptions of the target clients and product traits that are valued for production and consumption by farmers and end-users. Products to be placed in a market segment are described through product profiles/concepts; detailed descriptions of the traits and their thresholds (or range of values) to be found in the desired product or variety (sometimes based on current variety in the market) that aims to increase the likelihood of acceptance in the market. A breeding pipeline within a program may target one or more market segments and the associated product profiles using one or more breeding schemes. Breeding schemes are a collection of crossing, evaluation, and selection (CES) tasks and decisions which vary across breeding stages (e.g., in the crossing block vs. advanced yield testing in plants) and ultimately define a breeding strategy (Henryon et al., 2014; Yabe et al., 2017; Cobb et al., 2019; Pook et al., 2020; Gaynor et al., 2021; **Figure 1**).

Because CES decisions are numerous in a breeding program, breeding schemes can be difficult to describe succinctly and consistently, especially in the context of particular modes of crop reproduction and emerging breeding technologies (Yabe et al., 2017). Breeding leads or other experts typically visualize CES tasks and decisions as illustrative flow charts or tables. Unfortunately, some may not contain all information necessary to reproduce the breeding scheme in other places and may not fully visualize the resource allocation at different stages. Examples of these decisions, which may happen once or repeatedly at different stages of the breeding scheme, are:

- Crossing decisions: number of parents, number of crosses, number of progeny, type of cross, and mate allocation method, etc.
- Evaluation decisions: number of locations, replication level within and among locations, number of checks, experimental design, and plot sizes, etc.
- Selection decisions: percentage of individuals selected (selection intensity), the selection method (e.g., culling, index, tandem), and the selection unit, etc.

Another layer of complexity in communicating breeding schemes is that the number of stages in a scheme depends on the biology of the species, the multiplication ratio, the evaluation steps required to identify new parents, and the complexity of the market segment and product profile(s) for the desired final product (Henryon et al., 2014). Most breeding programs have a crossing stage to recombine elite parents, stages to multiply progeny and/or generate progeny derivatives such as testcrosses or inbred derivatives (e.g., lines), and multiple stages to test progeny derivatives for their potential as new parents or products. This stage-gate process in breeding programs is repeated cyclically, generating a recurrent selection scheme which, if effective, increases the population mean for the set of traits of interest (Allard, 1999; Chao and Ishii, 2005; Cooper, 2008). Additionally, programs do not wait until a cycle of the stage-gate process is completed to restart the process, and instead run several generations in parallel. A set of genotypes at a given stage within a given cycle is commonly referred to as a cohort or a selection stream (**Figure 2**). Generations may be discrete or overlapping depending on whether the parents of the cohort genotypes are selected from a single unique cohort or from multiple cohorts. Overlapping generations are more common and lead to more blurred genetic boundaries between cohorts, as cohorts tend to be more related with overlapping generations compared to discrete in absence of inbreeding control (Meuwissen and Sonesson, 1998). In summary, formalized breeding schemes are necessary to clarify the structure of breeding program pipelines.

Despite the inherent complexity of CES decisions, in some organizations breeding schemes are rarely shared formally or presented in writing. It is common for breeding leads to inherit a breeding program and its scheme from their predecessor. Usually, the predecessor transfers the breeding scheme verbally and practically rather than providing a quantitative description of the scheme in a formal document or software. This requires overlap between breeders and on-site presence of the predecessor, potentially for years at a time. Information about the breeding scheme may also be spread among several staff members within the program, interspersed in various publications, or buried in personal notes or presentations. Unfortunately, this method of transferring breeding schemes has led to the total loss of information (and even germplasm) of many breeding programs that have disappeared in the last century (Baenziger, 2006; Gepts and Hancock, 2006; Morris et al., 2006). Improved transferring methods could allow increased interoperability among breeders and better preservation of pedigrees, data, and germplasm.

In addition, codified, systematic documentation of breeding schemes could spur continuous improvement and lead to increased genetic gain and varietal turnover (EiB[2]). As suggested by Bernardo (2002), plant breeding programs should be managed as formal industrial processes that allow better breeding methods to be adopted as they become available to ensure sustainable, steady production of high-quality products. Industrial processes require a clear flow of subprocesses (tasks and decisions) and development of standard operating procedures (SOPs) that ensure minimization of production errors. Several methodologies, such as SixSigma and LEAN among others, were proposed in the 20th century to manage and continuously improve different components of industrial processes in the automotive, communications, and robotics industries (Bhuiyan and Baghel, 2005; Schroeder et al., 2008). Project management tools used in these methodologies, together with modern mathematical and computational tools like simulation and optimization, could easily be extrapolated to draft, formalize, manage, and improve breeding schemes successfully, in contrast to the artisanal approach to breeding common during the 20th century.

Improving a complex process like a breeding program requires understanding of how each process-related decision affects the outcome (e.g., genetic gain or probability of releasing a new product) and how varying these decisions affect the outcome. Given the cost and time associated with piloting new methods or ways to run this complex process, the use of simulations

[2]https://www.cgiar.org/wp/wp-content/uploads/2018/05/SC6-04_Multi-Funder-Breeding-Initiative-update.pdf

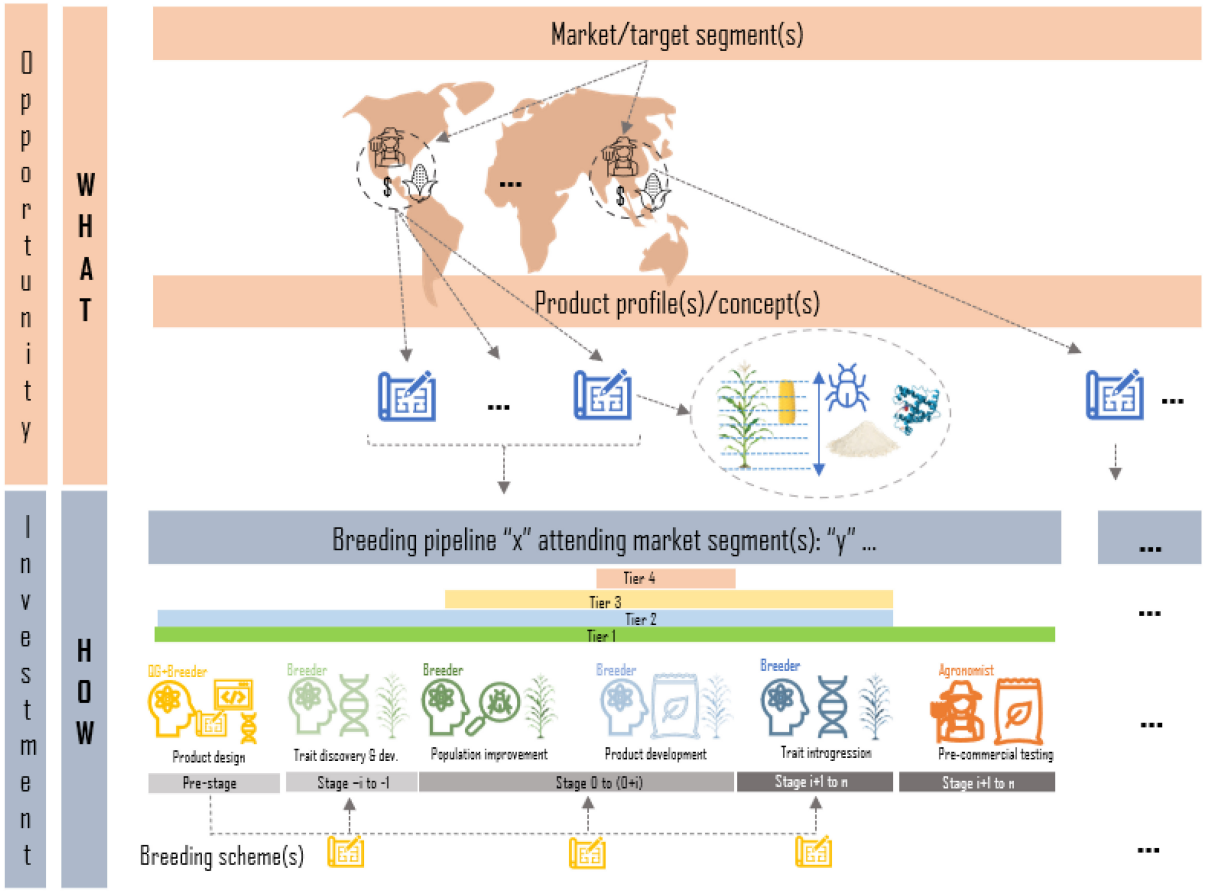

FIGURE 1 | Graphical representation of the relationship between target segments and pipelines of a breeding program. Market segments defined by agroclimatic regions, clients and product features are formalized in product profiles; trait descriptions of desired products to irrupt in the market. Breeding programs are represented as the sum of breeding pipelines focused on one or more of the following tasks: product design, trait discovery, population improvement, product development, trait introgression, or product dissemination. Each breeding pipeline may have one or more breeding schemes (strategies) to attend the market segments and the associated product profiles.

has a particular relevance to the design of breeding schemes (Li et al., 2012; Murray and Atlin, 2017; Yabe et al., 2017). Stochastic simulations of whole breeding programs rarely have been used to improve performance of breeding programs due to lack of computational and software resources in past decades (Gaynor et al., 2021). Currently, simulation technology is available and practical, and it should be incorporated into breeding scheme improvement efforts.

Here, we propose a process to formalize and improve the breeding schemes. In addition, we introduce a publicly available software tool, Breeding Pipeline Manager (BPM), which has capabilities to quantitatively document and record breeding schemes as well as the market segments and product profiles they target in a standardized yet customizable way. The BPM module can be added to any compatible enterprise breeding system (database) to link the phenotypic data to clear targets, pipelines and breeding schemes (Gao et al., 2020). In addition, we discuss the use of classical continuous improvement tools combined with state-of-the-art simulation and mathematical tools to continuously improve breeding schemes. We conclude by providing a case study of the use of these tools and methods in the improvement of the International Institute of Tropical Agriculture (IITA)-cassava breeding scheme. We expect that this framework will assist plant breeding professionals in conducting breeding as a systematic process and to help establish continuity and prevent inconsistency in breeding programs. Furthermore, we expect that formalizing breeding schemes will increase their rates of response to selection (i.e., genetic gain) by motivating critical examination of the schemes used and their opportunities for improvement.

MATERIALS AND METHODS

Enabling Methodology and Software to Formalize Breeding Targets and Schemes

We applied the continuous improvement methodology known as six-sigma to approach breeding as a process and enable

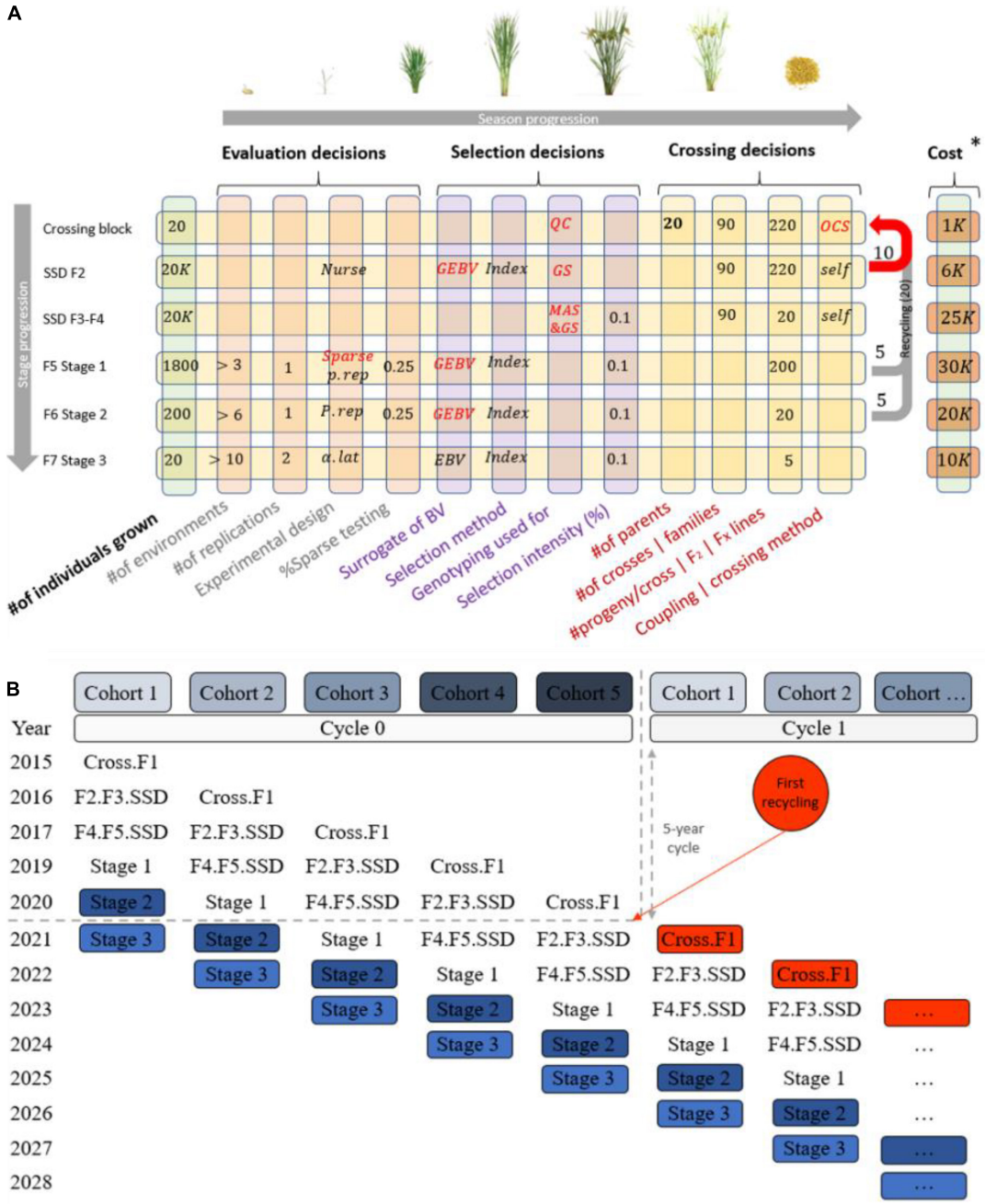

FIGURE 2 | Typical structure of a breeding pipeline. **(A)** Breeding scheme illustrating crossing, evaluation, and selection decisions (columns), which are made once or multiple times across the stage gate process (rows) in a cyclical fashion to achieve genetic gain. **(B)** A graphical representation of cohorts (parallel cycles of breeding) and how they overlap when recycling occurs and parents are taking from multiples stages.

breeding scheme improvements. Six-sigma is a five-step method: *define*, *measure*, *analyze*, *improve*, and *control* [DMAIC, (də.′meɪ.ɪk)]. The steps are undertaken iteratively to create a cyclical method for continuous improvements. Six-sigma was originally proposed by Bill Smith at Motorola in 1986 to improve industrial processes, in which 99.99966% of all opportunities to produce some features of a part are statistically expected to be free of defects (Tennant, 2017). In the case of plant breeding, this would imply producing better varieties than the ones existing in the market and steady genetic gains with higher probability. The six-sigma method reflects the scientific method, but it is used for process management rather than hypothesis testing. To increase the efficiency and ease of use of six sigma tools (e.g., value stream mapping, correlation analysis, etc.) by breeding teams, we developed a software named Breeding Pipeline Manager

(BPM) to document, describe, and visualize market segments, product profiles and breeding schemes. BPM is available at http://bpm.excellenceinbreeding.org? In collaboration with multiple breeding programs for a wide range of crops within the CGIAR, such as line, hybrid, and clonally propagated species, we identified breeding decisions which fell into three categories: crossing, evaluation and selection decisions. We then summarized breeding schemes and decisions as a table containing the breeding stages in rows (e.g., seedling nursery, stage 1 yield testing, etc.) and the CES tasks and decisions in columns. BPM provides a graphical user interface for capturing breeding schemes in a standard format. BPM also allows users to create visualizations (flowcharts) of their breeding scheme. In addition, the BPM allows market segment and product profile definition and users can link breeding schemes to market segments.

The BPM back-end was developed in Node JS, an open-source, cross-platform, and scalable JavaScript runtime environment. The front-end graphical user interface was developed in React JS. The source code is available at https://gitlab.com/excellenceinbreeding/module2. The platform leverages NodeJS asynchronous technology to perform intensive calculations without affecting the performance of other functionalities of the system. In addition, the platform uses Docker containerization technology for continuous development and integration (Merkel, 2014; Boettiger, 2015). This will not only enable automation of the deployment process but also horizontal scaling on any cloud infrastructure (depending on traffic). An online manual showing the details of the available features of the software and their use can be found once connected in the tool under the question mark bar at the bottom menu.

Application of Continuous Improvement in Breeding Programs: IITA-Cassava Example

We selected the CGIAR IITA-Cassava program to showcase the importance of using enabling tools and simulations to continuously optimize breeding programs. The IITA cassava program is situated in different parts of sub-Saharan Africa to serve region-specific challenges and market segments. The IITA-Cassava east-Africa pipeline, situated in Uganda, was chosen to showcase the use of SixSigma and the BPM tool to improve their breeding scheme. The five six-sigma steps were applied as follows in the cassava program. The *problem was defined* as lower than achievable genetic gain for traits of interest under the current scheme. The breeding targets and scheme were *measured* (documented) by capturing all CES decisions across all stages using the BPM tool (as described in the next section) through several interactions with the breeders. The analysis of the measured decisions and the genetic gain indicators revealed many possible improvements. We first chose to *analyze* the crossing decisions in the breeding scheme to identify possible improvements. The number of parents (nParents), number of crosses (nCrosses), number of progeny per cross (nProgeny), and recycling strategy were prioritized for evaluation *via* stochastic simulation in AlphaSimR (Gaynor et al., 2021). We proposed an *improvement* plan based on the close-to-optimal number of parents, number of crosses, number of progeny, and recycling strategy identified *via* simulation. The improvement plan used the A3 format (referring to the size of an A3 sheet that describes a project briefly) common in project management (Anderson et al., 2011). We then *controlled* the improvement by monitoring how key performance indicators (a set of quantifiable measurements used to gauge an institution's overall long-term performance) stated in the improvement plan changed as the improvements proceeded.

Stochastic Simulation to Improve Crossing Decisions in IITA-Cassava East-Africa Pipeline

Current and Alternative Programs

As a clonally propagated crop, cassava breeders currently have adopted a four-stage evaluation strategy in addition to the crossing block stage and the seedling nursery stage where planting material is multiplied. These evaluation stages include stage 1 (clonal evaluation; CE), stage 2 (preliminary yield trial; PYT), stage 3 (advanced yield trial; AYT), and stage 4 (uniform yield trial; UYT; **Table 1**). The summary of the advancement decisions across the different stages in the current (baseline) pipeline that was to be improved is as shown in **Table 1**. The pipeline began with only four parents selected to have the target traits for the target markets. From the four parents, 12 crosses were made, each with 136 progeny, thus resulting in 1,632 individuals. All 1,632 were multiplied in the seedling nursery and then evaluated at stage 1 in one environment and one replication per environment. Based on performance at stage 1 testing, 120 individuals were selected and advanced to stage 2 testing in two environments and two replications per environment. From stage 2 evaluation, 64 individuals were selected and advanced to stage 3 testing in two environments and three replications per environment. Finally, 24 individuals were selected and advanced to stage 4 testing in two environments and three replications per environment. Recycling of parents was planned to occur at PYT and UYT. This information was input into BPM and the scheme was simulated to address specific questions related to crossing decisions as prioritized by the breeding team. The program was interested in knowing if the use of four parents was adequate to sustain genetic gain. The program also inquired how to improve their recycling strategy, particularly from which

TABLE 1 | Summary of IITA-Cassava east-Africa pipeline numbers handled by stage.

Stage	Year	nParents	nCrosses	nProgeny/ cross	nIndividuals	% Selected
Crossing block	1	4	12	136	1,632	–
Seedling nursery	1	–	12	136	1,632	100
Stage 1 (CE)	2	–	12	136	1,632	100
Stage 2 (PYT)	3	–	–	–	120	7.35*
Stage 3 (AYT)	4	–	–	–	64	53.3*
Stage 4 (UYT)	5-6	–	–	–	24	37.5

**Stages where the recycling occurs to form the new crossing block. Recycling from the combined PYT and AYT leads to an average cycle time of 3.5 years.*

stage to recycle and whether to recycle from multiple stages. Here, we share the results for improving these decisions among many others that are currently being improved. It should be noted that the IITA pipelines in other regions, particularly for West Africa, use a greater number of parents (∼100) in their crossing block and therefore were not subject to this improvement. The simulation exercise is expected to find an optimal number of parents between these two extremes and useful for the East-Africa pipeline and develop some high-level guidelines for the test of the IITA-cassava pipelines.

Simulation Parameters: Treatments

To keep the resources constant with the baseline, we restricted the number of individuals (nIndividuals) at the F1 stage to 1,632 in all experimental simulation treatments. We then developed a grid to evaluate different numbers of parents in the crossing block using the following possible numbers of parents (nParents): 4, 8, 16, 32, and 64. The number of possible crosses for each level of number of parents was constrained to a maximum of nParents * (nParents − 1)/2, which is equivalent to all possible combinations of parents or a half-diallel, while considering the initial restriction that the number of individual progeny (nIndividuals) must be equal to 1,632. This resulted in the following possible numbers of crosses: 6, 12, 24, 48, 96, 204, 408, or 816. To keep the number of individual progeny constant at 1,632, the number of progeny per cross was set to 272, 136, 68, 34, 17, 8, 4, 2 for numbers of crosses equal to 6, 12, 24, 48, 96, 204, 408, and 816 respectively. The number of individual progeny is always equal to the number of crosses multiplied by the number of progeny per cross.

As such, a total of 24 simulation treatments were defined (**Table 2**). To identify the optimal number of parents, number of crosses, number of progeny per cross, and the best recycling strategy, a stochastic genetic simulation was conducted in the R package AlphaSimR (Gaynor et al., 2021).

Simulation Parameters: Genome and Evaluation

Burn-In Genome Sequence

For each replicate, a genome consisting of 18 chromosome pairs was simulated for the hypothetical plant species similar to cassava. These chromosomes were assigned a genetic length of 1.43 Morgans and a physical length of 8×10^8 base pairs. Sequences for each chromosome were generated using the Markovian coalescent simulator (MaCS; Chen et al., 2009) implemented in AlphaSimR (). Recombination rate was inferred from genome size (i.e., 1.43 Morgans/ 8×10^8 base pairs = 1.8×10^{-9} per base pair), and mutation rate was set to 2×10^{-9} per base pair. Effective population size was set to 30 to mimic an evolution history of natural and artificial selection.

TABLE 2 | Summary of factor values combined for number of parents, number of crosses, and number of progeny per cross to produce a total of 1,632 progeny.

Number of Parents	Number of Crosses	Number of Progeny per cross
4	6	2
8	12	4
16	24	8
. . .*	. . .*	. . .*
64	816	272

. . . indicates the numbers duplicate until reaching the final numbers in the row. All treatment combinations going beyond the 1,632 progeny were not run. This allowed comparison of these factors' influence on genetic gain at a fixed program size.*

Burn-In Founder Genotypes

Simulated genome sequences were used to produce 4 founder non-inbred individuals. These founder individuals served as the initial parents in the burn-in phase. Sites segregating in the founders' sequences were randomly selected to serve as 100 quantitative trait nucleotides (QTN) per chromosome (1,800 total).

Burn-In Phenotypes

A single highly complex trait representing an index of tuber yield, dry matter, cassava mosaic disease, total carotenoids and sprouting was simulated for all founders. The genetic value of this trait was determined by summing its QTN allelic effects. To model genotype-by-environment (GxE) interaction, allele effects depends on the value of an environmental effect which changes over years. For a given year, the allele effects followed this formula:

$$a_i\left(w_j\right) \;=\; b_i + m_i w_j,$$

where a_i is the allele effect for QTN i, w_j is the environmental effect for year j, b_i is the QTN intercept and m_i is the QTN slope on the environmental effect. The slope, intercept, and environmental effects were sampled from the following normal distributions. This equation is equivalent to Finlay–Wilkinson regression. Details on the full formulation of genotype by environment interaction simulation features enabled in AlphaSimR can be found in Gaynor (2021). In the case of the cassava program, a variance component for genotype by year (σ^2_{GxY} 2) and genotype by location (σ^2_{GxL} 1) interactions were defined and summed to produce the genotype by year by location (σ^2_{GxYxL} 3) interaction variance components used in the addTraitAG() function in the varGxE argument in AlphaSimR for a trait with additive gene action and GxE interaction. Main genotype variance component was assumed equal to 1 (σ^2_G 1). The genetic values of each non-inbred individual were used to produce phenotypic values by adding random noise sampled from a normal distribution with mean 0. The variance of the random error was varied according to the three stages of evaluation in the breeding program based on the plot size and number of replications per entry currently used according to the different experimental designs used at the different stages (augmented design at stage 1 and randomized completely blocked design in posterior stages). The values for these error variances were set to achieve levels of plot heritability reported by the cassava program currently estimated at the different stages.

In order to simulate the multi-environment testing common in breeding programs, the variance components for genotype by year and genotype by locations were used to simulate a matrix

TABLE 3 | Summary of simulation features for the genome and phenotypes.

Simulation features		
Burn-in	Genome sequence	100,000 generations of evolution
		18 chromosome pairs
		1.43 Morgans per Chromosome
		8×10^8 base pairs per chromosome
		2×10^{-9} mutation rate
	Founder genotypes	4 non-inbred founders
		1,800 QTN (additive GxE effects)
		Normally distributed QTN effects
		σ^2_{GxY} 2, σ^2_{GxL} 1, σ^2_{GxYxL} 3, σ^2_G 1
	Recent breeding	20 years of modern breeding
		Non-inbred cloned individuals
		Conventional breeding
Evaluation	Future breeding	20–60 years of breeding
		Testing alternative allocation of resources
		Equal cost programs
		Conventional breeding

of possible slopes for the environmental covariate used by the setPheno() function in the *p*-value argument (years in rows and locations in columns). The values were sampled depending on the year and number of locations used for a given stage to approximate the GxE. A summary of simulation features for the genome and phenotypes can be found in **Table 3**.

Population means and standard errors at Stage 1 of yield evaluation across the 20–60 years of breeding for the treatments described previously were computed using the dplyr library in R (Wickham et al., 2021), and plotted using the ggplot2 library in R (Wickham, 2011). One hundred replicates were run for each simulation treatment.

RESULTS AND DISCUSSION

Adapting Continuous Improvement Tools and Concepts in the Improvement of Breeding Schemes

Following the paradigm of approaching breeding as an industrial process (Bernardo, 2002), we adapted the six-sigma methodology to continuously improve the breeding schemes of breeding programs (**Figure 3**). Under this framework, we follow the project methodology Plan-Do-Study-Act inspired by William Edwards Deming named DMAIC, an acronym standing for Define, Measure, Analyze, Improve and Control steps that

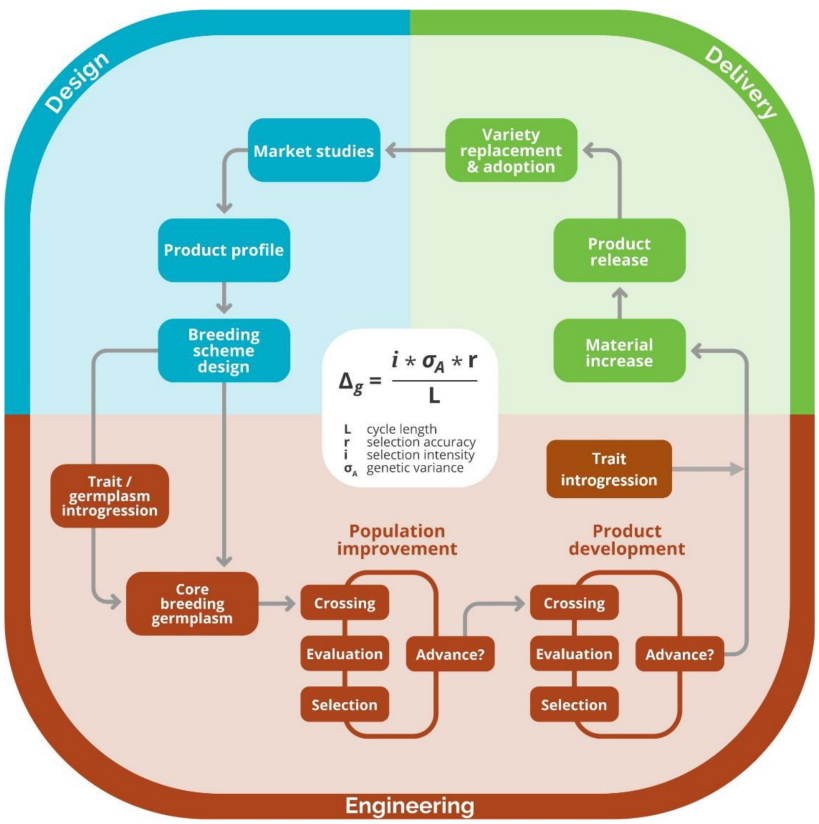

FIGURE 3 | Graphical representation of breeding as a process. The design component of the breeding process, which includes activities such as defining market segments, product profiles, and breeding schemes, is shown in blue. The engineering component of the process, where crossing, evaluation, and selection activities for product development and population improvement are made, is shown in red. The delivery component of the breeding process, where activities like material increase and registration, occur are shown in green. Image taken with permission from Covarrubias-Pazaran (2020).

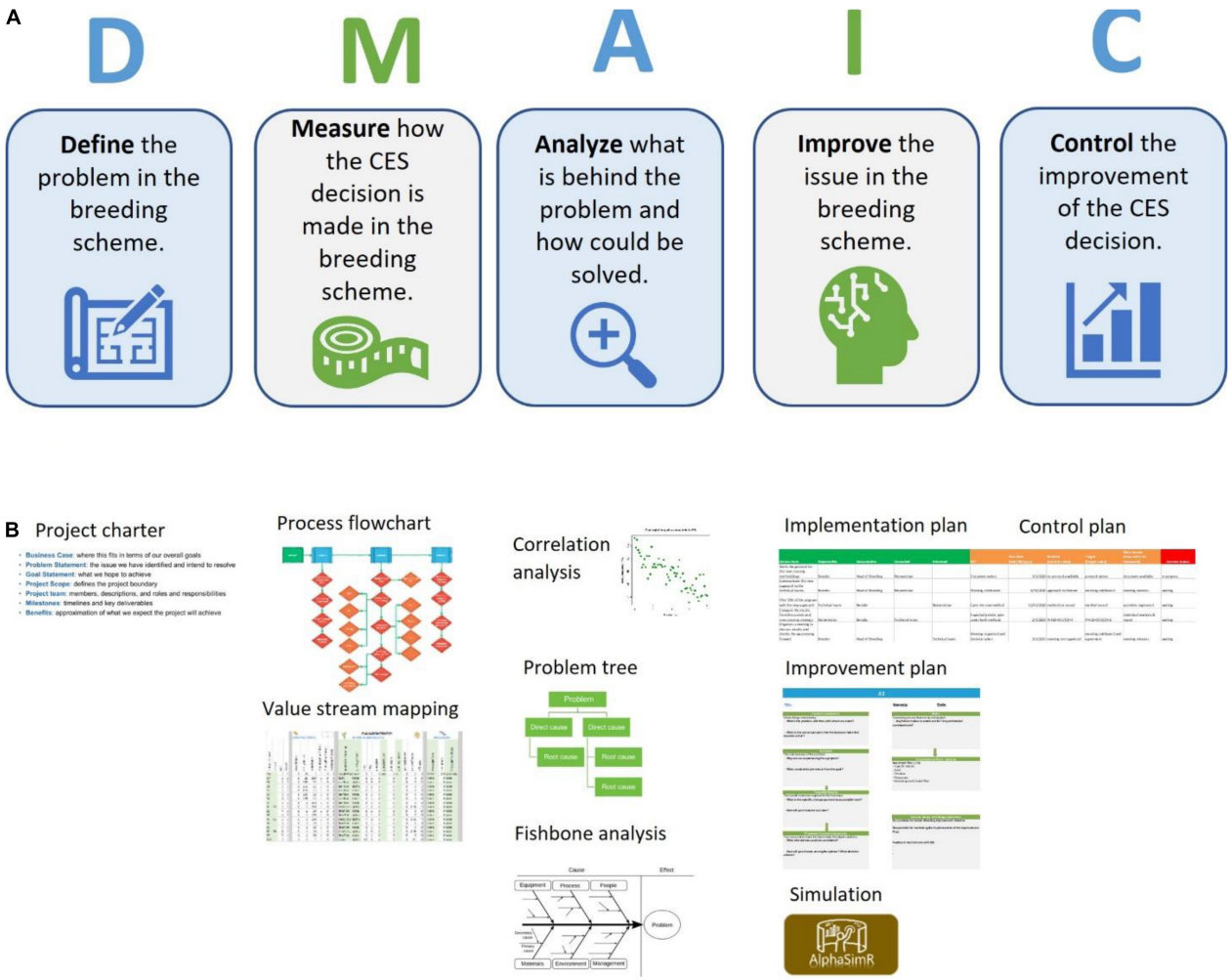

FIGURE 4 | Graphical representation of the six-sigma process applied to the continuous improvement of breeding schemes (strategies). **(A)** Description of the DMAIC steps. **(B)** Different tools to support continuous improvement of crossing, evaluation, and selection (CES) decisions in breeding schemes.

are cyclically repeated to reflect the continuous or cyclical approach (Aguayo, 1991; **Figure 4**). To demonstrate the use of the continuous improvement methodologies to optimize breeding processes leveraging from measuring tools like the breeding pipeline manager (BPM) and stochastic simulation, we engaged in discussions with the IITA Cassava program. First, the cassava team registered their breeding pipelines and the market segments targeted per pipeline. We found the IITA-cassava program to be composed of five breeding pipelines and on average tackling six market segments. The market segments and accompanying pipelines are stratified by a combination of regional consumption preferences and prevailing biotic and abiotic stresses. For example, most of the produced cassava in West Africa goes to processed (granulated and paste) products while in east Africa, the predominant preference is for fresh consumption with minimal processing (boiling, roasting and flour from dried roots). Subsequently, we focused on the IITA-Cassava east-Africa pipeline targeting market segments listed by breeders and generally described as fresh market and high-quality flour. Even though we proposed six-sigma for improving breeding schemes, the reader should keep in mind that continuous improvement applies to all components of the breeding process including the management roles which are responsible of the encouraging and incentivizing improvements.

Defining a Problem

The step of *defining* the problem was adapted to breeding scheme improvement by defining the problem as a suboptimal rate of response to selection (genetic gain) but pointing to one of the many crossing-evaluation-selection (CES) tasks and decisions at a given stage as the possible root of the problem. We found tools such as Project charter useful to define the problem (McKeever, 2006). We proceeded to use the project charter to *define* or state the problem in the IITA-Cassava east-Africa pipeline as having "potential for greater response to selection without increased expenditures." Details in the business case, goal statement, timeline, scope, and

TABLE 4 | Project charter applied to the IITA-Cassava east Africa program.

Problem statement The rate of genetic gain in the IITA-Cassava east-Africa breeding program is less than or equal to 1% per year for traits of interest, and the rate of variety turnover is lower than possible.	**Business case** By optimizing the breeding schemes using quantitative genetics principles, we can increase the response to selection per dollar invested per unit time.
Goal statement Reduce cycle time to the biological limit, optimize the trade-off between selection intensity and accuracy, manage the genetic variance, while constraining possible alternatives to similar level of resources.	**Team members** Cassava head of breeding Cassava breeders Quantitative Geneticist
Scope Crossing, evaluation and selection (CES) decisions included in the breeding scheme.	**Timeline** One to two CES tasks and decisions improved per year.

TABLE 5 | Features defining a market segment.

Client features	Environment features	Product features
Geographical region	Temperature	Mode of reproduction
Income	Humidity	Maturity
Education	Vegetation	Color
Farm size	Water availability	Shape
	Soil fertility	Biofortification
	Altitude	End use
	Soil pH	
	Production system	
	Prevailing biotic stresses	

The features of the client being served, the features of the target population of environments (TPE), and the final product characteristics are displayed. These three sets of features define a market segment in the breeding pipeline manager (BPM) tool.

team members can be found in **Table 4**. Unfortunately, we found that estimates of realized genetic gain were not available in the program to justify the definition of the problem. However, given the lack of an efficient recurrent selection strategy, we assumed the definition of the problem to be relevant to the program.

Software Development to Measure/Document Breeding Programs

To facilitate the *measuring* step of the continuous improvement approach proposed, in which CES decisions are recorded for further *analysis* (Bhuiyan and Baghel, 2005), we developed the BPM software. The breeding pipeline manager tool (BPM) is equipped with a module to define breeding pipelines as the sum of efforts to deliver a product. Breeding pipeline definition is the highest-level unit of information clustering in the BPM tool. The pipeline can be linked to market segments defined by the user. The market segment is defined in the BPM tool as the sum of the client, the target population of environments (TPE), and final product characteristics displayed in **Table 5**. These aim to capture the characteristics that can make breeding a more targeted effort according to Ragot et al. (2018) (**Figure 5A**). The BPM module can be incorporated to any enterprise breeding system (database) to properly link the generated phenotypic data to clear target segments and pipelines. Market segments for the cassava pipeline were captured using the BPM tool and are shown in **Figure 5A**. The major focus is on lowland high-rainfall, late maturity, long, hard cassava for fresh and flour consumption.

On top of defining the market segments, breeding programs must describe specifics of the product to be released in the market. Here, the concept of product profile (sometimes referred to as product concept) applies (Ragot et al., 2018; Carey et al., 2021). The existence of these profiles can make the difference between success and failure (Carey et al., 2021; Mwanga et al., 2021). The BPM tool has a module to define product profiles and link them to specific market segments, and the cassava breeders used the tool to formalize such profiles (**Figure 5B**). One of the product profiles for example is focused on achieving defined levels of fresh yield, plant height, dry matter and cassava mosaic disease resistance.

Part of the design of breeding pipelines is the creation of a blueprint or a breeding scheme that will allow the breeder to achieve the product profile for the market segment while maximizing the genetic gain of the breeding population per dollar invested (Henryon et al., 2014). The *blueprint* should specify all the crossing, evaluation and selection tasks and decisions occurring at the different stages (e.g., recombination, multiplication and testing stages) for the purposes of population improvement and product development. Most breeding programs have these two purposes coupled in a way that advancement decisions influence the recycling decisions. Others have proposed and shown that decoupling the population improvement from product development by moving the recycling decision to very early stages (e.g., F2, nursery or multiplication stages) will increase the rates of genetic gain. Better products can be expected when the product development process is regarded as separate from a rapid cycling population improvement strategy (Gaynor et al., 2017).

Crossing, evaluation and selection (CES) decisions comprising the breeding scheme can and should be recorded at the highest level of detail and safeguarded for the benefit of the breeding organization in case of any adverse circumstances. In **Table 6** we show the CES decisions that should be considered to capture the level of resolution and detail necessary to avoid loss of valuable information; these can be recorded by the BPM tool in the breeding scheme module. The software allows for breeding pipelines to manage multiple breeding schemes, as may happen when a program has a principal breeding scheme, but one or more parallel experimental breeding schemes, to accelerate genetic gains.

Measuring the Process

The *measuring* step of the six-sigma process was adapted by recording numerically and categorically all the different CES decisions across the various stages directly impacting genetic gain (e.g., number of parents, # crosses, # progeny, coupling method, etc.). The lack of available tools to measure/record breeding schemes was the motivation to develop the BPM

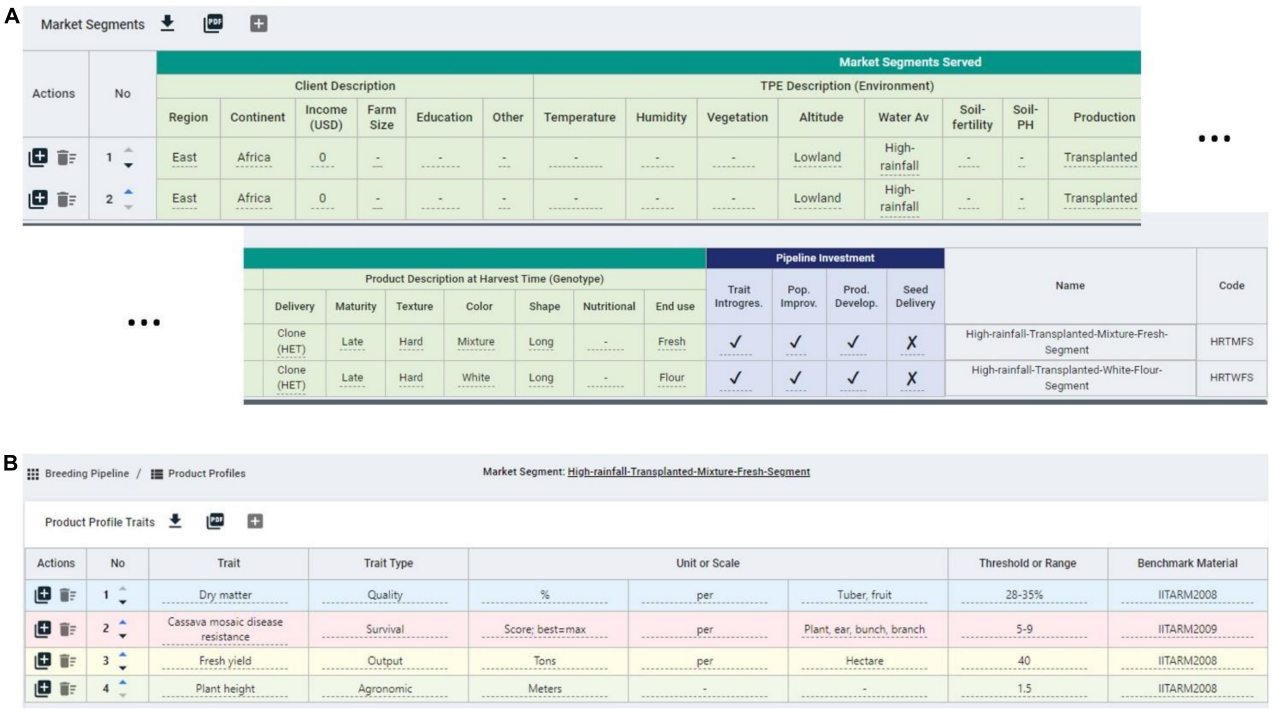

FIGURE 5 | Snapshot of some market segments and product profiles defined in the breeding pipeline manager (BPM) tool for the IITA-cassava east-Africa pipeline. **(A)** Two market segments for a cassava program with defined client, environment and product features, mainly distinguished by the end use. **(B)** Example product profile for a cassava market segment featuring quality, survival, output, and agronomic traits.

TABLE 6 | Examples of crossing, evaluation and selection decision recorded by the BPM tool across the different stages of the breeding program, defining the breeding strategy.

Evaluation	Selection	Crossing
Plant portion harvested in the previous season to be planted in the current season (e.g., seed, tuber, cutting)	Surrogate of merit (e.g., BLUE, BLUP, GBLUP) per phenotyped trait	Crossing or multiplication unit (e.g., family, individual)
Cultivation method of the plant portion (e.g., pot, plot, petri dish)	Number of locations per phenotyped trait	Crossing or multiplication method (e.g., 2-way cross, 3-way cross)
Experimental design	Selection method (e.g., visual, culling, index)	Parent coupling method (e.g. random mating, optimum contribution)
Total number of locations	Method to model genotype x environment interaction	Number of potential female parents
Replications per location	Method to model spatial adjustment	Number of potential male parents
Plot width and units (e.g., 1 m^2)	Selection intensities for different selection units (e.g., families, lines, female parents)	Total number of crosses or total number of unique materials to multiply
Plot length and units (e.g., 1 m^2)	Recycling unit	Number of progeny per cross or number of clones multiplied
Sparse testing percentage	Recycling generation	Molecular technology
Sparse testing bridging method	Number of selection units recycled	Number of molecular marker sites
Number of checks		Purpose of molecular technology (QC, GS, etc.)
Percentage of check plots		Population used in genomic selection as the training (prediction) set

tool presented above, although the BPM tool is inspired by the Value Stream Mapping approaches commonly used in process management (Singh et al., 2011). We *measured* or recorded the breeding scheme of the IITA-Cassava program using the BPM tool to capture all crossing, evaluation and selection decisions across the different breeding stages and a portion can be observed in **Figure 6**. We captured seven stages (crossing block, multiplication and five stages of yield and agronomic evaluation) across 52 different CES decisions for the East-Africa cassava pipeline that informed the *analysis* step to identify areas for improvement (**Figures 4**, **5**). These decisions comprise the crossing, evaluation and

A

Actions	No.	STARTING POINTS										EVALUATION STRATEGY													
		Stage Name	Generation Material Planted	Derived From	Year	Season	# of_HSfamilies	# of_F1 or cross	# of_Fx(x>1)ind	# of_sublines	Total Genotypes Planted	Plant Portion To Be Planted	Number Of Plant Portions	Plant Portion Planted In	Experimental Design	Total # of Locations	# of Replications Per Loc	Plot Width Units	Plot Width	Plot Length Units	Plot Length	Sparse Test Purpose	Sparse Test Bridge Method	Sparse Testing %(0-1)	% of Check Plots (0-1)	# Of Unique Checks (0 to n)	Evaluation Season Cost
	1	CB	PAR	F1	1	1	0	0	4	0	4	plant	4	fieldSpot	none	1	1	plant	1	plant	1	-	-	0	0	0	$1,000
	2	SeedProd	F1	PAR	1	2	0	12	136	0	1,632	seed	1,632	pot	none	1	1	plant	1	plant	1	-	-	0	0	0	$1,163
	3	STG1	F1	F1	2	1	0	0	1,632	0	1,632	cutting	5	row	augm	1	1	m	1	m	5	-	-	0	0.05	5	$2,632
	4	STG2	F1	F1	3	1	0	0	120	0	120	cutting	60	smallPlot	rcbd	2	2	m	3	m	5	-	-	0	0.06	2	$2,240
	5	STG3	F1	F1	4	1	0	0	64	0	64	cutting	252	bigPlot	rcbd	2	3	m	6	m	7	-	-	0	0.1	2	$2,960
	6	STG4.1	F1	F1	5	-	0	0	24	0	24	cutting	252	bigPlot	rcbd	2	3	m	6	m	7	-	-	0	0.1	2	$2,360
	7	STG4.2	F1	F1	6	-	0	0	15	0	15	cutting	378	bigPlot	rcbd	3	3	m	6	m	7	-	-	0	0.1	2	$3,225

B

FIGURE 6 | Graphical representation of BPM capabilities to record and display breeding schemes. **(A)** The evaluation decision across stages of an IITA-Cassava breeding scheme mapped in the breeding pipeline manager (BPM) tool are displayed in columns, and sequential stages are displayed in rows. **(B)** The capability to draw flowcharts with the available information in the breeding schemes is displayed.

selection strategies for both population improvement and product development.

Analyzing the Problem

The *analysis* step was adapted to breeding scheme improvement by replacing approaches such as correlation analysis. Correlation analysis is a method that links a response variable or key performance indicator (KPI) to another variable in the production process to understand relationships that could indicate the part of the process that needs to be refined. We instead conducted an analysis based on known quantitative genetic relationships between the various CES decisions and genetic gain (e.g., program size affects genetic gain depending on how effectively genetic variance is utilized and also linked to selection intensity). Additional tools like Fishbone (diagram to articulate the root causes of the problem) are not discouraged but we limited this exercise to one-to-one meetings with the breeding team to discuss the possible gaps while analyzing the current scheme together in the light of quantitative genetic principles (Ishii and Lee, 1996). We initially found

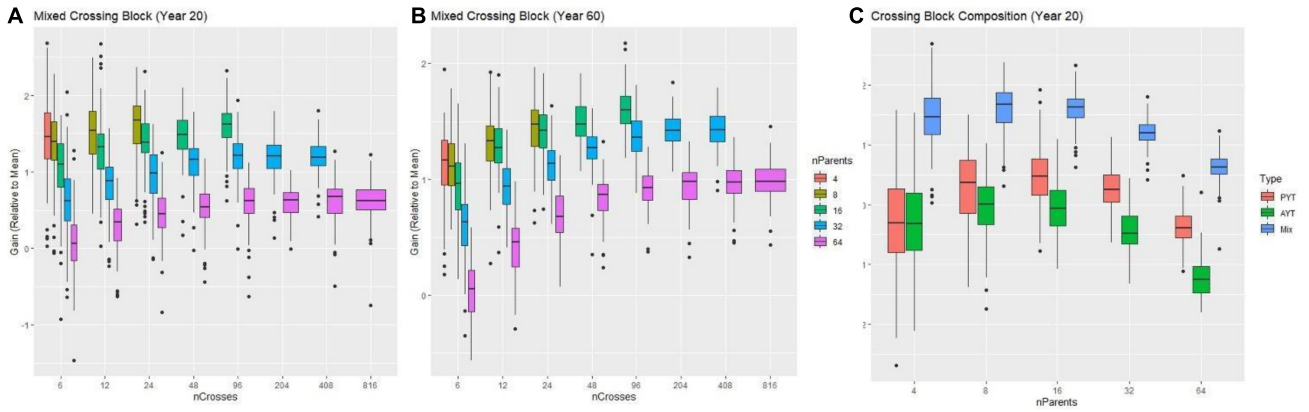

FIGURE 7 | Results from simulations comparing different numbers of parents and crosses combinations subject to the constraint of ∼1,600 individuals manageable for a time-horizon of 20 and 60 years of breeding. In panels **(A,B)** the genetic gain (relative to the mean) measured in advanced yield trial (AYT) individuals is shown (y-axis) as a function of a different number of crosses (x-axis) for two breeding-time horizons (20 and 60 years) for different number of parents (colored boxes) is shown. In panel **(C)** the genetic gain (relative to the mean) measured in advanced yield trial (AYT) individuals is shown (y-axis) as a function of a different number of parents (x-axis) for the breeding-time horizon of 20 years comparing different compositions of the crossing block are shown. The red, green and blue boxes represent crossing blocks composed by recycling AYT, PYT, or a mix of PYT and AYT individuals respectively.

several possible areas of improvement, including the small size of the program, the experimental design used at yield and agronomic evaluation stages, the coverage of the target population of environments (TPE), the opportunity to use molecular information, the potential improvement of analytical methods for genetic evaluation, a possibility to select the best families at earlier stages, the possibility to reduce the cycle length, and other decisions such as an improved crossing plan. Since it is well-known from classic quantitative genetics theory that using resources properly to maximize the genetic variance observed among and within families can maximize response to selection (Lynch and Walsh, 1998; Hallauer et al., 2010), we chose to optimize the decisions of number of parents, number of crosses and number of progeny per cross given the low number of parents used by the program in the crossing block and very likely limiting the rate or sustainability of genetic gains. Although we first focused on improving the resource allocation for the number of parents, crosses and progeny, the reader should remember that as a continuous improvement process, the other areas of opportunity identified should also be improved right after or at the same time depending on the resources available. This is just an example of how to implement breeding scheme improvement.

Using Simulation to Optimize the Process

Prior to recommending an *improvement* plan, we used genetic simulation (Gaynor et al., 2021) to identify optimal use of resources (plots available) by defining a grid of possible treatments that contained different combinations of number of parents, crosses and progeny subject to the constraint of 1,632 individuals at the F1 stage assuming other factors constant (e.g., properly resourced, properly tested, etc.). Regarding recycling strategy, using overlapping cohorts to recycle (i.e., a mixed crossing block composed half of parents from the PYT and half of parents from the AYT) lead to higher genetic gain regardless of the number of parents (**Figure 7A**). Based on this observation, we evaluated the effect on genetic gain of the number of parents, crosses and progeny while recycling from the mixed PYT and AYT.

For the single complex trait which represented an index of multiple traits, the decision of the number of parents provided the greatest opportunity to increase genetic gain. An excessive number of parents -here, more than 30- always resulted in decreased genetic gain compared to use of fewer than 30 parents at both the 20 and 60-year time horizons (**Figure 7C**). At the 20-year time horizon, the optimal number of parents was ∼8–16. However, at the 60-year time, the optimal number of parents increased to between 16 and 32.

Increasing the number of crosses generally increased gain, but with diminishing returns to additional crosses at a given number of parents. At low numbers of parents, not enough possible unique crosses were available to take advantage of gains possible by increasing the number of crosses. Interestingly, the optimal number of crosses also differed in the short (20-year) and long (60-year) terms. At the 20-year timepoint, schemes with fewer crosses and more progeny per cross tended to have higher gain across numbers of parents, but at the 60-year timepoint schemes with relatively more crosses and fewer progeny had higher gain. However, even at the 60-year timepoint, the optimal number of crosses was much less than the possible half-diallel of unique crosses.

Given the genetic parameters specified for the cassava program, the use of ∼8–16 parents, ∼24 crosses and ∼68 progeny per cross in each crossing block per year was the optimal distribution to maximize genetic gain at the 20-year time horizon (**Figure 7A**). At the 60-year time horizon, the optimal distribution was 16–32 parents, 60 crosses, and ∼30 progeny per cross (**Figure 7B**). To consider both short- and long-term interests of the breeding program, we chose to recommend use

of 15–30 parents recycled from the combined PYT and AYT stages with 40 crosses and 40 progeny, given the constraint of the program to handle ∼1,632 materials to start.

Improving the Process

The *improvement* step in the six-sigma method was adapted to breeding scheme improvement by using management tools like the A3 format to reflect the current and future state of the CES decision (subprocess) together with an action plan laying with detail the actions required to achieve the future state (**Supplementary File 1**; Anderson et al., 2011). We included a RACI chart (responsible, accountable, consulted and informed people in the improvement plan) to formalize the process to achieve the desired improvement. Is important to notice that a RACI chart can and should be employed during the management of the different tasks of the breeding process and not only for the continuous improvement of breeding schemes. We propose that the future state and actions included in the improvement plan should be guided by sound quantitative genetics principles and recommendations coming from state-of-the-art tools, such as evaluation of new strategies by genetic simulation (Mi et al., 2014; Gaynor et al., 2017; Pook et al., 2020, 2021). We expect that results obtained through simulation can identify close-to-optimal solutions and changes to the breeding CES tasks and decisions.

Based on the simulation findings, a meeting with the IITA-cassava breeding team was held to discuss the optimal scenarios revealed by simulations and the next steps. The recommendation to use between 15 and 30 parents in the crossing block depending on the target breeding-time-horizon was and to use a mixed crossing block of parents from both the PYT and the AYT was accepted by the team. The improvement plan developed by the IITA-cassava program included detailing the current and future state can be found in the **Supplementary File 1**. The improvement plan developed included actions like team agreement on the modification of the number of parents, number of crosses and number of progeny per cross used in the crossing block, the development of a new SOPs for the crossing block stage, training the technicians to execute the new SOPs, monitoring the genetic gain across years to confirm the positive change, among others.

Controlling the Improvement Process

The *control* step was adapted to breeding program improvement by adding a monitoring section to the improvement plan that keeps track of the progress of the action plan through the inclusion of key performance indicators (KPIs), deadlines, and risks, as it does in other industrial processes. To monitor or control the progress of the improvement plan in the IITA-cassava, deadlines and key performance indicators for the different actions were defined and monitored to ensure that changes occur. Once the new process was adopted, we moved to the next possible crossing evaluation or selection decision that could be causing low rates of genetic gain. This process is still undergoing together with other improvements identified.

CONCLUSION

There is tremendous potential of systematizing breeding as an industrial process and enabling continuous improvement methodologies (e.g., six-sigma) to the different crossing, evaluation, selection decisions and other parts of the breeding process. Successful implementation of these methodologies has potential to increase the rate of genetic gain and delivery of better products in breeding programs. To guarantee such improvements in genetic gain, the recommended changes must be near-optimal or at least better than the current strategy. We propose the use of genetic simulation to identify these solutions to guide the continuous improvement steps. The work with the IITA-cassava program resulted in improved resource allocations and adjustments to the proper number of parents to sustain gains for the breeding time horizon of interest. These and other improvements achieved through the same approach in other CES decisions are ongoing. We expect that this generalized framework will assist plant breeding professionals in transitioning toward conducting breeding as an industrial process, help prevent discontinuity and inconsistency in breeding pipelines and their schemes and implement a culture of continuous improvement in all areas of their breeding programs.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

GC-P, PC, JD, and MQ conceived the study. GC-P, ZG, and SS developed the software. DG, CW, and ML performed the simulations and applied the continuous improvement methodologies. IR, SK, EP, EK, EM, AA, and PK produced the data and ran the programs used for the study. All authors contributed to writing the manuscript and agreed to be accountable for the content of the work.

FUNDING

The Excellence in Breeding Platform received funding from the Bill and Melinda Gates Foundation (BMGF) grant number OPP1177070.

ACKNOWLEDGMENTS

We would like to thank the different breeding teams from the CGIAR that helped to improve the concepts and tools presented

while working to formalize their targets and breeding schemes. We would also like to thank Sam Storr for support in creating the figures. We would also further like to thank the cassava program for their assistance in testing the methodology for breeding scheme improvement in their programs.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.791859/full#supplementary-material

REFERENCES

Aguayo, R. (1991). *Dr. Deming: The American Who Taught the Japanese About Quality*. New York, NY: Simon and Schuster.

Allard, R. W. (1999). *Principles of Plant Breeding*. Hoboken, NJ: John Wiley & Sons.

Anderson, J. S., Morgan, J. N., and Williams, S. K. (2011). Using Toyota's A3 thinking for analyzing MBA business cases. *Decis. Sci. J. Innov. Educ.* 9, 275–285. doi: 10.1111/j.1540-4609.2011.00308.x

Baenziger, P. S. (2006). Plant breeding training in the US. *Hortscience* 41, 40–44. doi: 10.21273/HORTSCI.41.1.40

Bernardo, R. (2002). *Breeding for Quantitative Traits in Plants*. Woodbury, NJ: Stemma press.

Bhuiyan, N., and Baghel, A. (2005). An overview of continuous improvement: from the past to the present. *Manag. Decis.* 43, 761–771. doi: 10.1108/00251740510597761

Boettiger, C. (2015). An introduction to Docker for reproducible research. *ACM SIGOPS Oper. Syst. Rev.* 49, 71–79. doi: 10.1145/2723872.2723882

Carey, E. E., Ssali, R., and Low, J. W. (2021). Review of knowledge to guide product development and breeding for sweetpotato frying quality in West Africa. *Int. J. Food Sci. Technol.* 56, 1410–1418. doi: 10.1111/ijfs.14934

Chao, L. P., and Ishii, K. (2005). "Design process error-proofing: benchmarking gate and phased review life-cycle models," in *Proceedings of the ASME 2005 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, (New York, NY: American Society of Mechanical Engineers Digital Collection), 301–310. doi: 10.1115/DETC2005-84235

Chen, G. K., Marjoram, P., and Wall, J. D. (2009). Fast and flexible simulation of DNA sequence data. *Genome Res.* 19, 136–142. doi: 10.1101/gr.083634.108

Cobb, J. N., Juma, R. U., Biswas, P. S., Arbelaez, J. D., Rutkoski, J., Atlin, G., et al. (2019). Enhancing the rate of genetic gain in public-sector plant breeding programs: lessons from the breeder's equation. *Theor. Appl. Genet.* 132, 627–645. doi: 10.1007/s00122-019-03317-0

Cooper, R. G. (2008). Perspective: the stage-gate® idea-to-launch process—update, what's new, and nexgen systems. *J. Prod. Innov. Manag.* 25, 213–232. doi: 10.1111/j.1540-5885.2008.00296.x

Covarrubias-Pazaran, G. (2020). *Guidelines for Germplasm and Trait Introgression. Online Manual*. Available online at: https://excellenceinbreeding.org/sites/default/files/manual/EiB-M2_Germplasm%20%20%20trait%20introgression_01-06-20.pdf (accessed June 1, 2021).

Gao, S. Y., Hagen, T. J., Robbins, K., Jones, E., Karkkainen, M., Dreher, K. A., et al. (2020). "Transforming breeding through enterprise breeding system and analytics," in *Proceedings of the Plant and Animal Genome*, Pag.

Gaynor, R. C., Gorjanc, G., Bentley, A. R., Ober, E. S., Howell, P., Jackson, R., et al. (2017). A two-part strategy for using genomic selection to develop inbred lines. *Crop Sci.* 57, 2372–2386. doi: 10.2135/cropsci2016.09.0742

Gaynor, R. C., Gorjanc, G., and Hickey, J. M. (2021). AlphaSimR: an R package for breeding program simulations. *G3* 11:jkaa017. doi: 10.1093/g3journal/jkaa017

Gaynor, R. C. (2021). *Traits in AlphaSimR*. Available online at: https://cran.r-project.org/web/packages/AlphaSimR/index.html (accessed June 1, 2021).

Gepts, P., and Hancock, J. (2006). The future of plant breeding. *Crop Sci.* 46, 1630–1634. doi: 10.2135/cropsci2005-12-0497op

Hallauer, A. R., Carena, M. J., and Miranda Filho, J. D. (2010). *Quantitative Genetics in Maize Breeding*, Vol. 6. Berlin: Springer Science & Business Media. doi: 10.1007/978-1-4419-0766-0_12

Henryon, M., Berg, P., and Sørensen, A. C. (2014). Animal-breeding schemes using genomic information need breeding plans designed to maximise long-term genetic gains. *Livest. Sci.* 166, 38–47. doi: 10.1016/j.livsci.2014.06.016

Ishii, K., and Lee, B. (1996). "Reverse fishbone diagram: a tool in aid of design for product retirement," in *Proceedings of the ASME Design Engineering Technical Conferences and Computers in Engineering Conference*, (Stanford, CA: Stanford University). doi: 10.1115/96-DETC/DFM-1272

Li, X., Zhu, C., Wang, J., and Yu, J. (2012). Computer simulation in plant breeding. *Adv. Agron.* 116, 219–264. doi: 10.1016/B978-0-12-394277-7.00006-3

Lynch, M., and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sunderland, MA: Sinauer Associates.

McKeever, C. (2006). The project charter–blueprint for success. *Crosstalk* 2006:19.

Merkel, D. (2014). Docker: lightweight linux containers for consistent development and deployment. *Linux J.* 2014:2.

Meuwissen, T. H. E., and Sonesson, A. K. (1998). Maximizing the response of selection with a predefined rate of inbreeding: overlapping generations. *J. Anim. Sci.* 76, 2575–2583. doi: 10.2527/1998.76102575x

Mi, X., Utz, H. F., Technow, F., and Melchinger, A. E. (2014). Optimizing resource allocation for multistage selection in plant breeding with R package Selectiongain. *Crop Sci.* 54, 1413–1418. doi: 10.2135/cropsci2013.10.0699

Morris, M., Edmeades, G., and Pehu, E. (2006). The global need for plant breeding capacity: what roles for the public and private sectors? *Hortscience* 41, 30–39. doi: 10.21273/HORTSCI.41.1.30

Murray, S. C., and Atlin, G. N. (2017). "Symposium–training plant breeders to design and manage 21st century cultivar development pipelines," in *Proceedings of the ASA, CSSA and SSSA International Annual (2017)*, (Madison, WI: ASA-CSSA-SSSA).

Mwanga, R. O., Mayanja, S., Swanckaert, J., Nakitto, M., Zum Felde, T., Grüneberg, W., et al. (2021). Development of a food product profile for boiled and steamed sweetpotato in Uganda for effective breeding. *Int. J. Food Sci. Technol.* 56, 1385–1398. doi: 10.1111/ijfs.14792

Pook, T., Schlather, M., and Simianer, H. (2020). MoBPS-modular breeding program simulator. *G3* 10, 1915–1918. doi: 10.1534/g3.120.401193

Pook, T., Büttgen, L., Ganesan, A., Ha, N. T., and Simianer, H. (2021). MoBPSweb: a web-based framework to simulate and compare breeding programs. *G3* 11:jkab023. doi: 10.1093/g3journal/jkab023

Ragot, M., Bonierbale, M. W., and Weltzien, E. (2018). "From market demand to breeding decisions: a framework," in *Paper Presented at the CGIAR Gender and Breeding Initiative*, Lima.

Schroeder, R. G., Linderman, K., Liedtke, C., and Choo, A. S. (2008). Six sigma: definition and underlying theory. *J. Operat. Manag.* 26, 536–554. doi: 10.1016/j.jom.2007.06.007

Singh, B., Garg, S. K., and Sharma, S. K. (2011). Value stream mapping: literature review and implications for Indian industry. *Int. J. Adv. Manuf. Technol.* 53, 799–809. doi: 10.1007/s00170-010-2860-7

Tennant, G. (2017). *Six Sigma: SPC and TQM in Manufacturing and Services*. Milton Park: Routledge. doi: 10.4324/9781315243023

Wickham, H., François, R., Henry, L., and Müller, K. (2021). *dplyr: A Grammar of Data Manipulation. R Package Version 1.0.7*. Available online at: https://CRAN.R-project.org/package=dplyr (accessed May 1, 2021).

Wickham, H. (2011). ggplot2. *Wiley Interdiscip. Rev.* 3, 180–185. doi: 10.1002/wics.147

Yabe, S., Iwata, H., and Jannink, J. L. (2017). A simple package to script and simulate breeding schemes: the breeding scheme language. *Crop Sci.* 57, 1347–1354. doi: 10.2135/cropsci2016.06.0538

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 27 January 2022
doi: 10.3389/fpls.2021.808001

Dissecting the Root Phenotypic and Genotypic Variability of the Iowa Mung Bean Diversity Panel

*Kevin O. Chiteri[1], Talukder Zaki Jubery[2], Somak Dutta[3], Baskar Ganapathysubramanian[2], Steven Cannon[1,4] and Arti Singh[1]**

[1] Department of Agronomy, Iowa State University, Ames, IA, United States, [2] Department of Mechanical Engineering, Iowa State University, Ames, IA, United States, [3] Department of Statistics, Iowa State University, Ames, IA, United States, [4] USDA—Agricultural Research Service, Corn Insects and Crop Genetics Research Unit, Ames, IA, United States

Mung bean [*Vigna radiata* (L.) Wilczek] is a drought-tolerant, short-duration crop, and a rich source of protein and other valuable minerals, vitamins, and antioxidants. The main objectives of this research were (1) to study the root traits related with the phenotypic and genetic diversity of 375 mung bean genotypes of the Iowa (IA) diversity panel and (2) to conduct genome-wide association studies of root-related traits using the Automated Root Image Analysis (ARIA) software. We collected over 9,000 digital images at three-time points (days 12, 15, and 18 after germination). A broad sense heritability for days 15 (0.22–0.73) and 18 (0.23–0.87) was higher than that for day 12 (0.24–0.51). We also reported root ideotype classification, i.e., PI425425 (India), PI425045 (Philippines), PI425551 (Korea), PI264686 (Philippines), and PI425085 (Sri Lanka) that emerged as the top five in the topsoil foraging category, while PI425594 (unknown origin), PI425599 (Thailand), PI425610 (Afghanistan), PI425485 (India), and AVMU0201 (Taiwan) were top five in the drought-tolerant and nutrient uptake "steep, cheap, and deep" ideotype. We identified promising genotypes that can help diversify the gene pool of mung bean breeding stocks and will be useful for further field testing. Using association studies, we identified markers showing significant associations with the lateral root angle (LRA) on chromosomes 2, 6, 7, and 11, length distribution (LED) on chromosome 8, and total root length-growth rate (TRL_GR), volume (VOL), and total dry weight (TDW) on chromosomes 3 and 5. We discussed genes that are potential candidates from these regions. We reported beta-galactosidase 3 associated with the LRA, which has previously been implicated in the adventitious root development *via* transcriptomic studies in mung bean. Results from this work on the phenotypic characterization, root-based ideotype categories, and significant molecular markers associated with important traits will be useful for the marker-assisted selection and mung bean improvement through breeding.

Keywords: root system architecture, GWAS, high throughput phenotyping, phenomics, pulses

OPEN ACCESS

Edited by:
Rodomiro Ortiz, Swedish University of Agricultural Sciences, Sweden

Reviewed by:
Santosh Kumar Gupta, National Institute of Plant Genome Research (NIPGR), India
Karl Peter Pauls, University of Guelph, Canada

****Correspondence:***
Arti Singh
arti@iastate.edu

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *02 November 2021*
Accepted: *06 December 2021*
Published: *27 January 2022*

Citation:
Chiteri KO, Jubery TZ, Dutta S, Ganapathysubramanian B, Cannon S and Singh A (2022) Dissecting the Root Phenotypic and Genotypic Variability of the Iowa Mung Bean Diversity Panel. Front. Plant Sci. 12:808001. doi: 10.3389/fpls.2021.808001

INTRODUCTION

There is an increasing demand, particularly in Western cultures, for plant-based protein sources, including analogs of meat, egg, and dairy (Wild et al., 2014; Joshi and Kumar, 2016; Niva et al., 2017; Aschemann-Witzel et al., 2020). Numerous factors influence this change, including social, political, environmental, ethical, health-focused, technological, and economical (Vinnari, 2008; Markiewicz, 2010). Pulses such as lentils (*Lens culinaris*), horse beans (*Dolichos* spp.), lupins (*Lupinus albus* L.), common beans (*Phaseolus vulgaris*), chickpea (*Cicer arietinum*), field peas (*Pisum sativum*), cowpeas (*Vigna unguiculata*), fava beans (*Vicia faba*), mung bean [*Vigna radiata* (L.) Wilczek], urd beans (*Vigna mungo*), and food-grade soybeans (*Glycine max*) have consistently been used as protein sources in the global south (Niva et al., 2017). The plant protein demand has been fueled by the sustainable production of pulses coupled with their health benefits and the production of meat analogs (Wild et al., 2014; Niva et al., 2017). Residents of sub-Saharan Africa, the Caribbeans, and South America consume more than 10 kg/capita/year of pulses, compared to 3 kg/capita/year among Western cultures (Akibode and Maredia, 2012).

Mung bean (*V. radiata* L. Wilczek), initially domesticated in India, is now cultivated in over 7 million hectares worldwide (Nair et al., 2019; Aski et al., 2021). Mung bean is a short-duration crop, usually between 60 and 90 days from planting to harvest (Sandhu and Singh, 2021). Mung beans, being relatively heat- and drought-tolerant, may be helpful in the agricultural adaptation to climate change (Pataczek et al., 2018; Wang et al., 2018). Mung bean is easily digestible, with seed composition of 22–28% protein, 1–1.5% fat, and 60–65% carbohydrates, as well as minerals, vitamins, and antioxidants (Jahan et al., 2020; Aski et al., 2021; Sandhu and Singh, 2021). Mung beans are consumed as whole grain, sprouted gram, split dhal, and mung bean flour in Indian dishes (Nair et al., 2019; Aski et al., 2021). The complementation of mung beans with cereals provides a balanced intake of required nutrients. In Western cultures, mung beans are consumed mostly as sprouts and, more recently, as processed products such as meat substitutes, egg substitutes, chips, no-nut butter, and pasta (Sandhu and Singh, 2021). The demand for plant-based protein in the United States has led to enhancing existing and establishment of breeding programs at United States institutions, including at Iowa State University (Sandhu and Singh, 2021). However, due to limited breeding efforts in North America, there is a knowledge gap in agronomic trait diversity including root traits that are emerging as important area of research and breeding efforts (Lynch, 2007; White et al., 2013; Burridge et al., 2017).

Root system architecture (RSA) can be defined as the morphology of the root system at many scales, both global (i.e., the entire root system) and local (i.e., primary and lateral root levels), as well as the spatial variability of the morphology (Hodge et al., 2009; Rogers and Benfey, 2015; Lobet et al., 2019; Aski et al., 2021). The morphometric traits include the number, length, volume, mass, shape, angle, depth, etc. The spatio-temporal variation seen in RSA of different plants reflects the phenotypic plasticity and the genotype × environment interaction (Rogers and Benfey, 2015; Lobet et al., 2019). Roots have a great impact on yield and plant fitness by providing plants with the structural stability, nutrient foraging, plant-microbe interactions, preventing soil erosion, aeration, and water extraction (Hodge et al., 2009; Rogers and Benfey, 2015).

The desired root phenotypes by plant breeders will be ones that enhances plant adaptation to the edaphic stress while maintaining or increasing yields, for example, deeper and proliferating roots are desired during water-deficient stresses in the changing climate (Gaur et al., 2008; Aski et al., 2021). Lynch and Brown (2001) coined the term "topsoil foraging ideotype," which is characterized by proliferation of lateral roots, long root hairs, association with mycorrhizal fungi, and suited to uptake of the immobile phosphorus mineral from the topsoil stratum (White et al., 2013). The "steep, cheap, and deep" ideotype (Lynch, 2013) optimizes on the uptake of water and the soluble nitrogen in the soil minimizing leaching. The "steep, cheap, deep" is characterized by thick and long primary roots, high affinity for N by epidermal cells, and the high concentration of cortical aerenchyma cells (White et al., 2013). Falk et al. (2020b) used the term "informative root" (iRoot) category to capture the biological significance of the captured root traits as would simulate field conditions. They reported that the topsoil foraging had a faster total root length-growth rate (TRL_GR), wider (WID), and a large TRL upper root ratio (TRL_{Upper}). The steep, cheap, and deep ideotype contained a deeper primary root length (PRL), faster TRL_GR, steep lateral root angles (LRA), and lower solidity traits (SOL2). These works have been possible due to the use of computer vision and machine learning in extraction of complex traits.

Advances in computer vision, machine learning, and high-throughput phenotyping (HTP) technologies, coupled with efficient statistical methods and collaborative research, have opened the way for more research to be carried out in plants as reviewed in Singh et al. (2016, 2018), Atkinson et al. (2019), Ghosal et al. (2019), Parmley et al. (2019), and Singh A. et al. (2021). The use of these technologies has been implemented in the collection of agronomic and yield estimation traits (Riera et al., 2021), detection of abiotic and biotic stress (Naik et al., 2017; Zhang et al., 2017; Nagasubramanian et al., 2018, 2019), and monitoring plant health (Ghosal et al., 2018). However, as shown in Falk et al. (2020a), computer vision and machine learning-based methods are essential to advance the root phenotyping and large-scale studies (Singh et al., 2016, 2018; Singh D. P. et al., 2021). Root phenotyping is classified depending on where it is carried out, i.e., in controlled environments or in the field, destructive or nondestructive, and whether the HTP uses 2-dimensional (2D) or 3-dimensional (3D) to capture the traits of interest (see reviews, Atkinson et al., 2019; Singh A. K. et al., 2021).

Previous methods developed for extracting roots in the field include destructive methods such as "shovelomics" (Trachsel et al., 2011), the use of soil cores (Wasson et al., 2016), and nondestructive methods such as electrical resistance tomography (Srayeddin and Doussan, 2009), electromagnetic inductance (Shanahan et al., 2015), and ground penetrating radar (Liu et al., 2018). Soil opacity is still a limiting factor to access roots

in most field experiments. Controlled environmental methods include the use of rhizotrons, which utilize soil (Rellán-Álvarez et al., 2015), nonsoil methods such as hydroponics (Aski et al., 2021), transparent artificial growth media (Ma et al., 2019), and growth pouches (Tan and Nopamornbodi, 1979). The high-throughput nature of acquiring 2D root images from controlled environments necessitated the development of the image analysis software to extract the traits (Atkinson et al., 2019). Commercially available software includes WinRhizo (Regent Instruments, Quebec, Canada). The open-source software available for use includes SmartRoot (Lobet et al., 2011), RootNav (Pound et al., 2013), GiaRoots (Galkovskyi et al., 2012), DART (Le Bot et al., 2010), Ez-Rhizo (Armengaud, 2009), DIRT (Das et al., 2015), ARIA (Pace et al., 2014; Falk et al., 2020a), RhizoVision (Seethepalli et al., 2020), MyRoot (Betegón-Putze et al., 2019), and IJ_Rhizo (Pierret et al., 2013). A combination of the methods above has been used to study the roots of a variety of plants under various conditions. Species of plant roots studied include common bean (Bonser et al., 1996), maize (Hund et al., 2009; Zheng et al., 2020), wheat (Atkinson et al., 2015), pearl millet (Passot et al., 2016), soybean (Falk et al., 2020a), and canola (Gioia et al., 2016). In a recent study, Aski et al. (2021) utilized modified hydroponics to study the RSA phenotypic diversity of the mung bean mini-core collection at the World Vegetable Center (formerly known as Asian Vegetable Development and Research Center [AVRDC]) (Schafleitner et al., 2015). As this software is capable of generating the useful data on multiple phenotypic root traits, these also lend themselves to genetic studies.

Genome-wide association studies (GWAS) is a statistical tool that uses historical recombination events to uncover the significant genotypic variation associated with the phenotypic variation for the trait of interest (Huang and Han, 2014; Tibbs Cortes et al., 2021). GWAS has been extensively used to investigate important agronomic traits such as plant height, days to flower, yield, nutrient content, flood and drought tolerance, and insect and pest resistance in maize (Yang et al., 2014), soybean (Zhang et al., 2015; Fang et al., 2017), common bean (Kamfwa et al., 2015), mung bean (Sandhu and Singh, 2021), and rice (Huang et al., 2010) among others.

The current study was conducted with the objectives to (1) study the diversity of the RSA trait in the Iowa (IA) mung bean panel, (2) contextualize these RSA traits with root-based ideotypes, and (3) conduct GWAS on RSA traits and identify candidate genes for these associations.

MATERIALS AND METHODS

Plant Materials

A total of 376 accessions were used in this study. A total of 372 Plant Introductions (PI) were filtered from the 482 IA mung bean panel (Sandhu and Singh, 2021) using the identity-by-state method in SNPRelate and genetic distance of Nei (Zheng et al., 2012). PIs that were common among the two methods were dropped. The 482 PIs were a part of the over 3,000 mung bean accessions obtained from the United States Department of Agriculture-Germplasm Resources Information Network (USDA-GRIN), in Griffin Georgia that were able to flower and form pods in IA conditions. Three Asian Vegetable mung bean (AVMU) lines, namely, AVMU001, AVMU0201, and AVMU9701, were included as checks, since they are improved cultivars from the WVC, formerly AVRDC (Fernandez and Shanmugasundaram, 1988).

Experimental Design and Germination Protocol

This study used a randomized incomplete block design, with each growth chamber serving as a replicate, for a total of eight replications for the experiment. Two growth chambers were used for an increased throughput. Each chamber had four blocks. Each block contained six complete and two incomplete sub-blocks. Each complete sub-block held twelve genotypes, while each incomplete sub-block held eleven genotypes. The genotypes were randomized within each block and sub-blocks. Randomization was generated using the R package blocksdesign (Edmondson and Edmondson, 2021). The procedures described in Falk et al. (2020a) were followed with little modification in the experimental design (**Figure 1**). First, ten seeds of each genotype were equally spaced near the top (∼1″) of a 9″ × 12″ germination paper. The paper was rolled into germination rolls. All the germination rolls for each sub-block were rubber banded and labeled with a tag. Once all the 376 were planted, water was filled halfway in the rectangular bucket, and the rolls transferred to a Conviron growth chamber (Controlled Environments Ltd., Winnipeg, Canada) set at 25°C for 16 h of light and 20°C for 8 h darkness. The lighting was set to 276–280 μmol/s/m^2 and constantly monitored using the LI-250A photometer (Li-Cor Biosci-sciences, Lincoln, NE, United States). On the 5th day of germination, a representative sample for each genotype was picked and carefully placed onto the 12″ × 18″ blue germination paper (Anchor Paper, Minneapolis, MN, United States). Labeled bar-coded tags are stapled onto the 1″ folded top of the blue paper. A 12″ × 16″ brown blotting paper (Anchor Paper, Minneapolis, MN, United States) was carefully placed on top of the blue paper. Two full blue papers are clipped together using binder clips and placed in the rack on the plot number in the chamber. Chamber conditions were monitored daily.

Imaging, Image Processing, and Trait Extraction

A high-throughput imaging station was set up similar to the one reported by Falk et al. (2020a). Images were captured using a Canon T5i digital SLR camera (lens: EF-S 18–55 mm f/3.5–5.6 IS II) (Canon USA, Inc., Melville, NY, United States). The setup allowed for the automated renaming of images captured using the SmartShooter software (Hart, 2021; **Figure 1E**). The seedlings were imaged on days 12, 15, and 18 without moving roots. Exceptions to moving were on day 12 where some secondary roots did not emerge from the fold and on day 18, when some of the roots of the genotypes were overgrowing the length of the blue paper. The days to image and the spacing were determined by a preliminary study. On the 18th day, the seedlings were cut at the junction between the shoot and camera-visible root section.

FIGURE 1 | Workflow from seed to phenotyping roots, **(A)** germination roll, **(B)** germination rolls banded by sub-block, **(C)** seedlings in the four blocks inside the growth chamber, **(D)** genotypes pooled out for imaging, **(E)** imaging platform, **(F)** captured root, and **(G)** preprocessed root ready for trait extraction.

The root and shoot of each genotype were placed in small brown bags. The wet-cut seedlings were dried in growth chambers set at 34°C/24 h for 2 days, with the light set to 276–280 μmol/s/m^2 and stored for weighing. Each root and shoot of the genotype were measured using the Ohaus portable weighing balance (Ohaus Corporation, NJ, United States).

More than 9,000 (376 genotypes × 8 reps × 3 time points) images were collected from the whole experiment. The images were first rotated manually to portrait orientation to enable consistent preprocessing. Images with no germinated seed, herein referred to as blank, were excluded from processing. JPEGCrops (2021), an open-source software, was used to auto crop all the images in a batch by cutting off the top part with the labeled bar-coded tag. The images were then converted into black/white images by thresholding (heuristically using red, green, and blue, LAB, or Hue, Saturation, and Value color spaces). This was carried out using the image processing step and followed by the trait extraction step within the improved Automatic Root Image Analysis (ARIA) 2.0 tool (Falk et al., 2020a). Different color spaces were used due to the variations in the images caused by unequal lightning and water spots. The ARIA 2.0 tool runs on Matlab (2020a). Traits extracted in ARIA are shown in the **Supplementary Materials** (**Supplementary Table 1**).

Statistical Analysis

All the analyses were carried out using the R statistical software (R Core Team, 2021). A separate code was written for the extraction of median LRA. Outliers were filtered out using the Tukey's box plot method (Hoaglin, 2003). A soybean genotype previously included was dropped due to clear visual differences with mung beans. The preprocessing steps above left 8,611 observations represent 375 genotypes for the analysis. Most of the reported analysis is from day 15 data with references and comparisons to days 12 and 18. Day 15 was chosen as a good representation of root growth between day 12 and day 18. A subset of ARIA traits was used for the analysis (**Table 1**). They were informed by traits used in a similar study by Aski et al. (2021) and traits important for the iRoot categories: topsoil foraging, and steep, cheap, and deep ideotypes described by Falk et al. (2020b). A mixed linear model (Eq. 1) was used to extract the best linear unbiased predictors (BLUPs) for each trait per genotype. All model variables were considered a random effect except chamber, which was a fixed effect. The model was run within the H2Cal function from the inti package (Lozano-Isla, 2021), which utilized the unbalanced data (Cullis et al., 2006; Piepho and Möhring, 2007; Schmidt et al., 2019). Broad sense heritability (H) was calculated using Eq. 2 (Cullis et al., 2006). Pearson correlations were used to draw correlations among the root traits.

$$Y_{ijkl} = \mu + \text{Chamber}_i + (1|\text{Chamber}:\text{Block})_{ij} + (1|\text{Chamber}:\text{Block}:\text{Sub}-\text{block})_{ijk} + (1|\text{Genotype})_l + e_{ijkl} \quad (1)$$

Where μ is the overall population mean, Y_{ijkl} is the phenotypic trait, Chamber_i is the fixed effect of the ith growth chamber $(1|\text{Chamber:Block})_{ij}$ is the random interaction effect between the ith chamber and the jth block $(1|\text{Chamber:Block:Sub-block})_{ijk}$ is the three way random interaction effect between the ith chamber, jth block and kth sub-block $(1|\text{Genotype})_l$ is the random effect of

TABLE 1 | Subset of traits of mung bean root architecture extracted from the Automated Root Image Analysis (ARIA) software and used for the analysis, clustering, and iRoot category.

Trait name	Symbol	Unit	Trait description
Total root length	TRL	cm	Cumulative length of all the roots in centimeters
Primary root length	PRL*	cm	Length of the primary root in centimeters
TRLUpper	TRLUpper*	cm	Total root length of the upper one third
Depth	DEP	cm	The maximum vertical distance reached by the root system
Width	WID*	cm	The maximum horizontal width of the whole RSA
Diameter	DIA	cm	Diameter of the primary root
Lateral root branches	LRB	Count	Number of lateral root branches
Network area	NWA	Count	The number of pixels that are connected in the skeletonized image
Convex area	CVA	cm^2	The area of the convex hull that encloses the entire root image
RhizoArea	RHZO	cm^2	Length of 2 mm surrounding the TRL
Primary root surface area	PRA	cm^2	Surface area of the primary root
Volume	VOL	cm^3	Volume of the primary root
Lateral root angles	LRA*	Angle	Root angles along the extent of all lateral roots
Solidity	SOL*	Ratio	The fraction equal to the network area divided by the convex area
Length distribution	LED	Ratio	TRL_{Upper}/TRL_{Lower}
Total root length-growth rate	TRL_GR*	cm/day	(TRL_day 15 – TRL_day 12)/3

**Traits used for iRoot ideotypes. RSA, root system architecture.*

the lth genotype, and e_{ijkl} is the random error term following the $N(0,q^2e)$.

$$H^2_{\text{Cullis}} = 1 - \bar{V}_{\Delta}^{BLUP}/2\delta_g^2 \quad (2)$$

Where δ_g^2 is the genotypic variance and $\bar{V}_{\Delta}^{BLUP}$is the mean variance of the difference of two genotypic BLUPs for the genotypic effect (Schmidt et al., 2019).

Root Ideotypes, Phenotypic, and Genotypic Diversity

The iRoots were formed by first ranking the genotypes under each trait, getting the sum of the ranks and then ranking the sums for each category. For topsoil foraging, the genotypes were ranked individually under the TRL_GR, WID, and TRL_{Upper}. The sum of the ranks was ranked, and this yielded to the final ranking of each genotype. A similar approach was used for the "steep, cheap, and deep" ideotype using the TRL_GR, steep LRA, and SOL2.

The principal component analysis (PCA) and hierarchical clustering were used in both the phenotypic and genotypic clustering of the genotypes using the Euclidean distance matrix. The base R function hclust with methods "complete" and "prcomp" was used. The package factoextra (Kassambara and Mundt, 2020) was used to determine the optimum number of clusters to be used by comparing 30 different indices. The clusters were related to the country of origin. Heat maps were developed according to the trait performance and iRoot category ranking using the Complex Heatmap package (Gu et al., 2016). Phenotypic and genotypic dendrograms were made using the dendextend (Galili, 2015) and circlize (Gu et al., 2014) packages. The pairwise fixation index (Fst) was calculated between the two genotypic clusters using the function genet.dist (method = "WC84") within the ade4 package (Dray Stéphane, 2007). Fst is an indication of the amount of differentiation within subpopulations, with low Fst indicating high gene flow (low genetic diversity) (Wright, 1965).

Genome-Wide Association Analysis

In total, 26,550 SNPs (marker data) were obtained using genotype-by-sequencing and preprocessed earlier by Sandhu and Singh (2021). Sites with >15% missing data and minor allele frequency > 0.01 were filtered out. GWAS was carried out using BLUPs on all the trait data. Associations were conducted using the Trait Analysis by aSSociation, Evolution, and Linkage (TASSEL, Bradbury et al., 2007) software using a linear mixed model (Yu et al., 2006). Both the kinship matrix and PCA were generated in TASSEL controlling for population structure. Bonferroni correction with p-value = 0.05 was used to control for false positives and declare significant associations (Kuo, 2017). Manhattan plots for visualizing the associations were carried out in R using the CMplot library in the rMVP package (Yin et al., 2021). Authors also used a newly developed computational framework, selection of variables with embedded screening (SVEN), a Bayesian based model to run GWAS (Li et al., 2020). The identification of candidate genes was carried out by locating the significant SNP on the sequenced and annotated mung bean genome using the "genome data viewer" tool at the National Center for Biotechnology Information (NCBI; Kang et al., 2014).

RESULTS

Descriptive Statistics, Correlation, and Heritabilities

We observed the significant phenotypic variability for root traits. The coefficient of variation (CV) ranged from 2 to 19% and standard deviation (SD) from 0.01 to 628.67 (different units of measurements for traits). Most of the traits had low SD, i.e., <10

except for TRL, VOL, TRL_{Upper}, and CVA that had SD < 100, while RHZO had an SD > 500. TRL, TRL_{Upper}, CVA, WID, NWA, RHZO, and TRL_GR had 10% < CV < 20%, while the rest of the traits had CV < 10% (**Table 2**). Day 12 CV ranged from 0 to 22% while for day 18 was 0–28%. Dry matter weight measurements, including shoot dry weight (SDW), root dry weight (RDW), and total dry weight (TDW) had CV 24, 28, and 26%, respectively, at day 18 (**Supplementary Table 2**).

The correlation between the root traits varied. TRL_{Upper} was highly correlated with WID, CVA, and TRL_GR. NWA was highly correlated with WID, CVA, TRL_GR, TRL_{Upper}, RHIZO, and TRL. LRA had the lowest correlation with the other traits. There was no correlation between LRA and VOL, and DIA (**Figure 2**). Correlation on day 12 was high. Negative correlations were observed at day 18 with SOL2 being negatively correlated to most traits and LRA negatively correlated with root shoot ratio (RSR) (image not shown).

Broad sense H ranged from 0.22 to 0.73. LRA and WID had the lowest and highest H at 0.22 and 0.73, respectively. DIA, VOL, surface area, LRB, and LRA had H < 0.5, while TRL, PRL, LED, TRL_{Upper}, CVA, DEP, WID, NWA, RHZO, SOL2, and TRL_GR had H > 0.5 (**Table 2** and **Figure 3**). H was high at days 15 and 18 and low on days 12 for most of the traits. Day 12 H ranged from 0.24 to 0.51, with TRL_{Upper} having the highest H. Day 18 H ranged from 0.23 to 0.87, with dry weight traits (i.e., SDW, RDW, and TDW) showing higher levels at 0.84, 0.87, and 087, respectively (**Supplementary Table 2**).

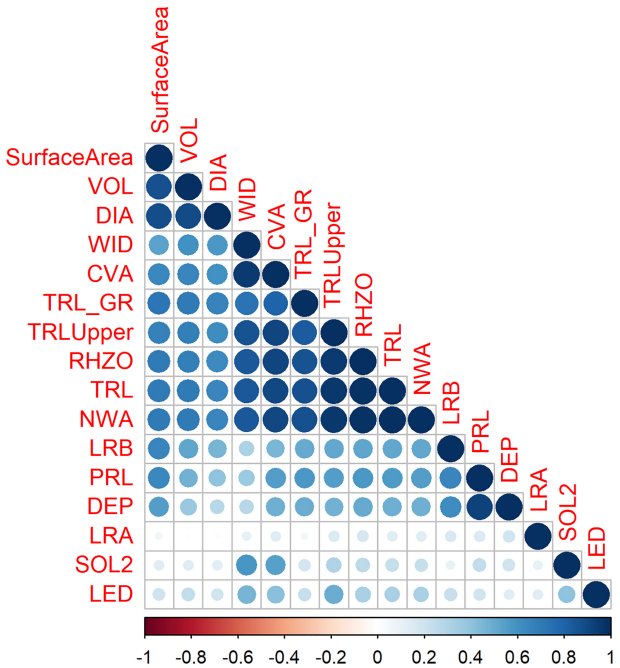

FIGURE 2 | Correlations among the root traits at day 15 using the Iowa (IA) mung bean panel. Experiments were conducted in controlled environment chambers and included eight replications.

Root Ideotypes

We described two root ideotypes, namely, topsoil foraging and "steep, cheap, and deep." PI425425 (India), PI425045 (Philippines), PI425551 (Korea), PI264686 (Philippines), and PI425085 (Sri Lanka) emerged as top five in the topsoil foraging category. PI425594 (unknown origin), PI425599 (Thailand), PI425610 (Afghanistan), PI425485 (India), and AVMU0201 (Taiwan) were top five in the "steep, cheap, and deep" ideotype (**Table 3** and **Figure 4**). For day 18, the PI425551 (Korea), PI264686 (Philippines), PI426026 (Thailand), PI425085 (Sri Lanka), and PI426042 (Australia) were the top five in the topsoil foraging category. In the "steep, cheap, and deep" ideotype, PI264686 (Philippines), PI425551 (Korea), PI363514 (India), and PI425599 (Thailand) were the top four (**Supplementary Table 3**). No iRoot categories were created on day 12 since TRL_GR could not be calculated.

TABLE 2 | Descriptive statistics and broad sense heritability of a subset of root traits from day 15 of the Iowa (IA) mung bean genotypes estimated from eight replications.

Trait	Mean	Median	Min	Max	SD	CV (%)	H
TRL	230.22	225.56	159.22	325.48	35.1	15	0.66
PRL	42.72	42.65	36.48	48.06	1.7	4	0.54
LED	2.05	2.04	1.53	2.6	0.19	9	0.51
DIA	0.24	0.24	0.23	0.26	0.01	3	0.31
VOL	261.41	260.76	223.14	316.38	17.76	7	0.29
Surface area	31.42	31.45	28.9	34.76	1.06	3	0.24
TRL_{Upper}	150.85	149.35	99.82	216.32	22.63	15	0.64
CVA	412.66	407.67	258.73	567.28	65.28	16	0.66
DEP	37.74	37.7	33.47	40.66	1.16	3	0.53
WID	18.68	18.4	11.69	25.47	2.88	15	0.73
NWA	2.82	2.77	1.98	3.95	0.41	15	0.64
LRB	137.67	137.82	124.39	152.1	4.28	3	0.32
RHZO	4651.19	4588.15	3272.6	6280.68	628.27	14	0.63
SOL2	140.92	140.94	107.82	164.14	9.6	7	0.57
LRA	50.23	50.18	46.64	53.36	1.11	2	0.22
TRL_GR	24.64	23.88	15.76	39.69	4.64	19	0.68

*Full trait descriptions are in **Table 1**. SD, standard deviation; CV, coefficient of variation; H, broad sense heritability.*

Phenotypic and Genotypic Clusters

Three distinct phenotypic clusters were observed using the root trait data, while two clusters were observed from the SNP data of the genotypes (**Figure 5**). Phenotypic clusters 1, 2, and 3 had 69, 163, and 135 genotypes, respectively. Genotypic clusters 1 and 2 had 48 and 319 genotypes, respectively. India had the highest number of genotypes in both genotypic clusters 1 (37) and 2 (197). The United Kingdom had 13 genotypes in genotypic cluster 2, while the rest of the countries had less than 10 genotypes in each cluster. The United Kingdom and United States had no genotypes in genotypic cluster 1 (**Supplementary Figure 1**). Similarly in the phenotypic clusters 1, 2, and 3, India led with 17, 94, and 123 genotypes. The rest of the countries had less than ten genotypes (**Supplementary Table 3**). On day 18, there were two phenotypic clusters and two genotypic clusters. Phenotypic clusters 1 and 2 had 250 and 117 genotypes, respectively.

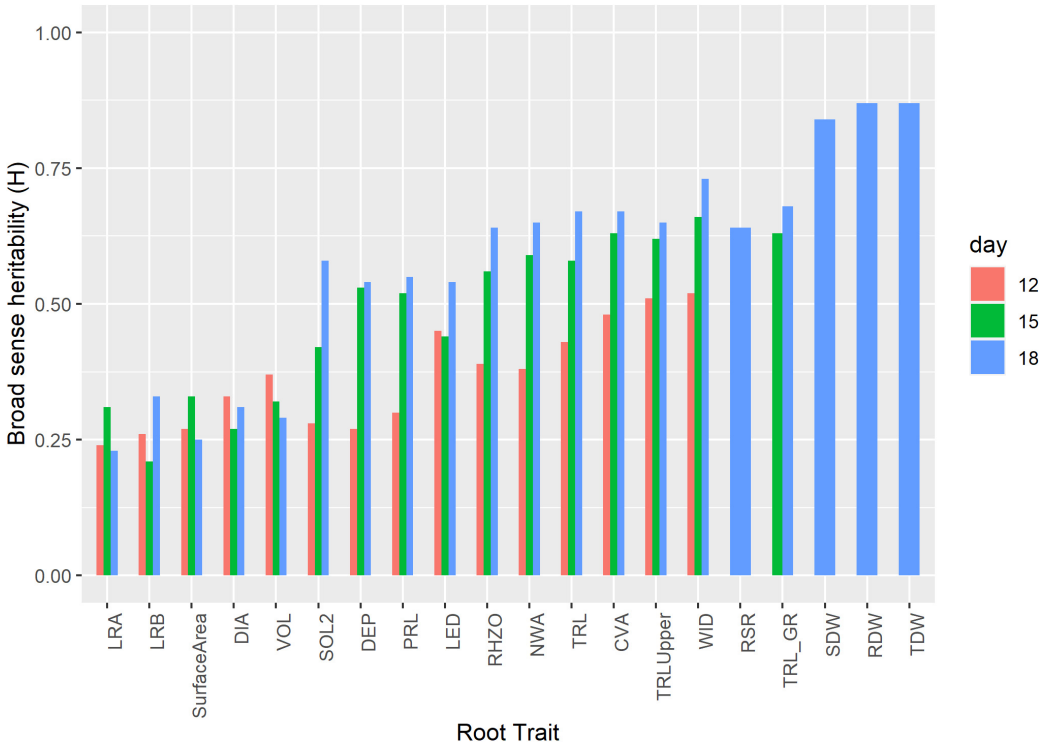

FIGURE 3 | Broad sense heritability (H) of select root related traits at days 12, 15, and 18. Experiments were carried out with the IA mung bean panel with genotypes grown in growth chambers.

Genotypic clusters had similar composition as from day 15. In the phenotypic cluster, India had 132 and 102 genotypes in clusters 1 and 2. The rest had less than 10 genotypes (**Supplementary Figure 2** and **Supplementary Table 4**).

PC1 and PC2 explained 7.6 and 3.9% of the total genotypic variation in the IA mung bean GWAS panel (**Figure 6**). The PCs were not able to discern any distinct subpopulations. Superimposition of iRoot ranking on the PCs showed that genotypes from India dominated both in the "steep, cheap, and deep" and topsoil foraging (**Figures 6C,D**). For day 18, the top genotypes in the "steep, cheap, and deep" category are mostly from India, while in the topsoil foraging, they are mostly from the other countries, Australia, the United Kingdom, and others with few from India (**Supplementary Figure 3**). The complex heat map showed the patterns and correlations among the genotypic clusters, iRoot type rank, and root trait performance used in clustering (**Figure 7**). Most of the traits in genotypic cluster 2 had a better ranking in the topsoil foraging, while cluster 1 contained mostly the worst ranked in the same category. Genotypes were evenly distributed in ranking among the genotypic clusters 1 and 2 in the "steep, cheap, and deep" iRoot category. Some of the best genotypes for the traits, including TRL_{Upper}, RHZO, NWA, WID, and CVA, were in genotypic cluster 2, while cluster 1 was dominated by low values in the above traits. LRA, SOL2, LED, LRB, PRL, and DEP looked evenly distributed within genotypic clusters 1 and 2 (**Figure 7**). The pairwise Fst was 0.05.

TABLE 3 | Top five genotypes by iRoot rank categories from day 15 image analysis.

Topsoil foraging	Country	"Steep, cheap, and deep"	Country
PI425425	India	PI425594	Unknown origin
PI425045	Philippines	PI425599	Thailand
PI425551	Korea	PI425610	Afghanistan
PI264686	Philippines	PI425485	India
PI425085	Sri Lanka	AVMU0201	Taiwan

Genome-Wide Association Studies and Candidate Genes

Association studies revealed significant SNPs for traits on different days. Day 12 LRA had seven significant SNPs. Day 15 LED had one significant SNP. On day 18, TRL_GR, TDW, and volume each had one significant SNP, while LED had two significant SNPs (**Figure 8**). Out of the seven SNPs for day 12 LRA, the first three had no mapping on the mung bean genome with no gene ID, genomic context, and gene description. On day 18, significant marker 8_10447903 for LED is an uncharacterized gene. Significant SNPs were found for the same trait LED, for days 15 and 18, marker 8_11481602 and marker 8_10447903, respectively. A summary of the significant SNP associations from the TASSEL software is presented in **Table 4**.

Day 12 SNP markers 2_2226549, 2_19972687, 7_19972687, and 11_7608353 were associated with LRA. Marker 2_2226549

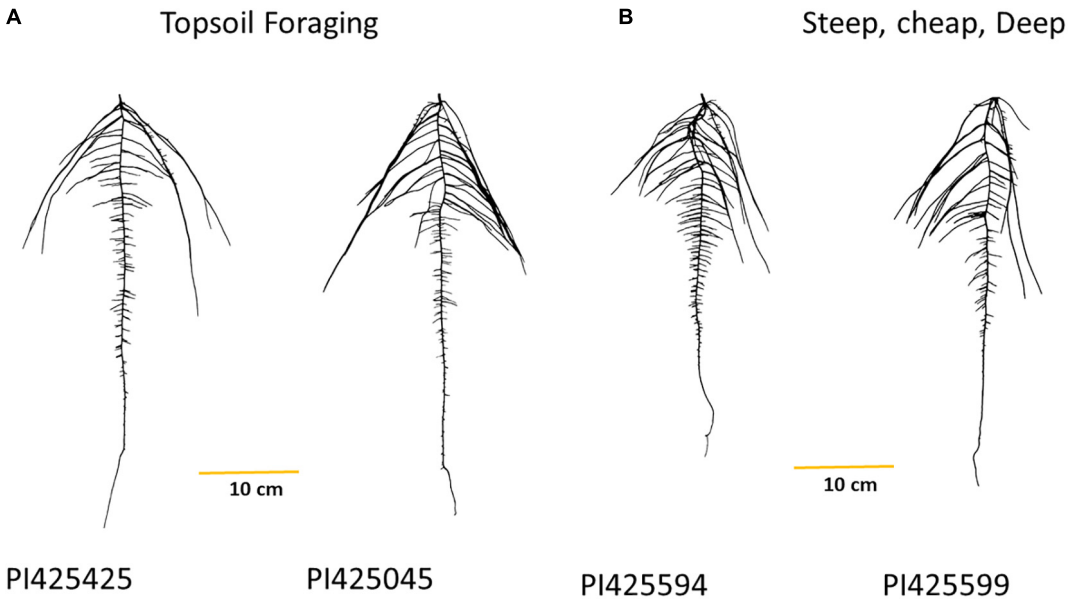

FIGURE 4 | Mung bean iRoot ideotypes, **(A)** top two genotypes in the topsoil foraging and **(B)** top two genotypes in the steep, cheap, and deep after ranking the genotypes in the IA mung bean panel.

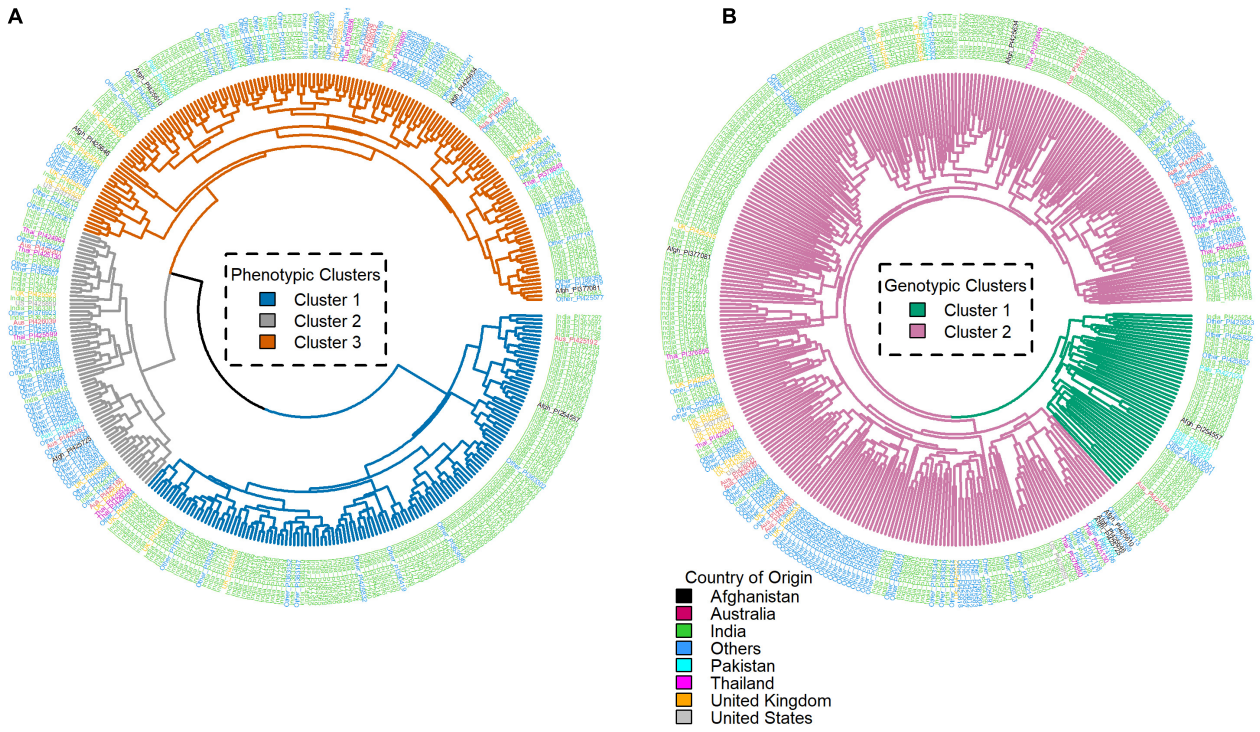

FIGURE 5 | Day 15 phenotypic **(A)** and genotypic **(B)** clusters of the root traits and single-nucleotide polymorphism (SNP) data, respectively, generated using hierarchical clustering of the core traits for all the genotypes. The labels represent the Plant Introductions (PI) and country of origin.

is located within an exon for a gene described as lignin forming anionic peroxidase (LOC106755829). Marker 2_19972687 is located within an exon encoding a gene (-)-germacrene D synthase-like (LOC106753988). Marker 7_19972687 is located within an exon of the beta-galactosidase 3 gene (LOC106768494). Marker 11_7608353 associated with LRA also located within an exon for a gene described as protein FAR1-RELATED SEQUENCE 5 (LOC106776541). The same significant SNP

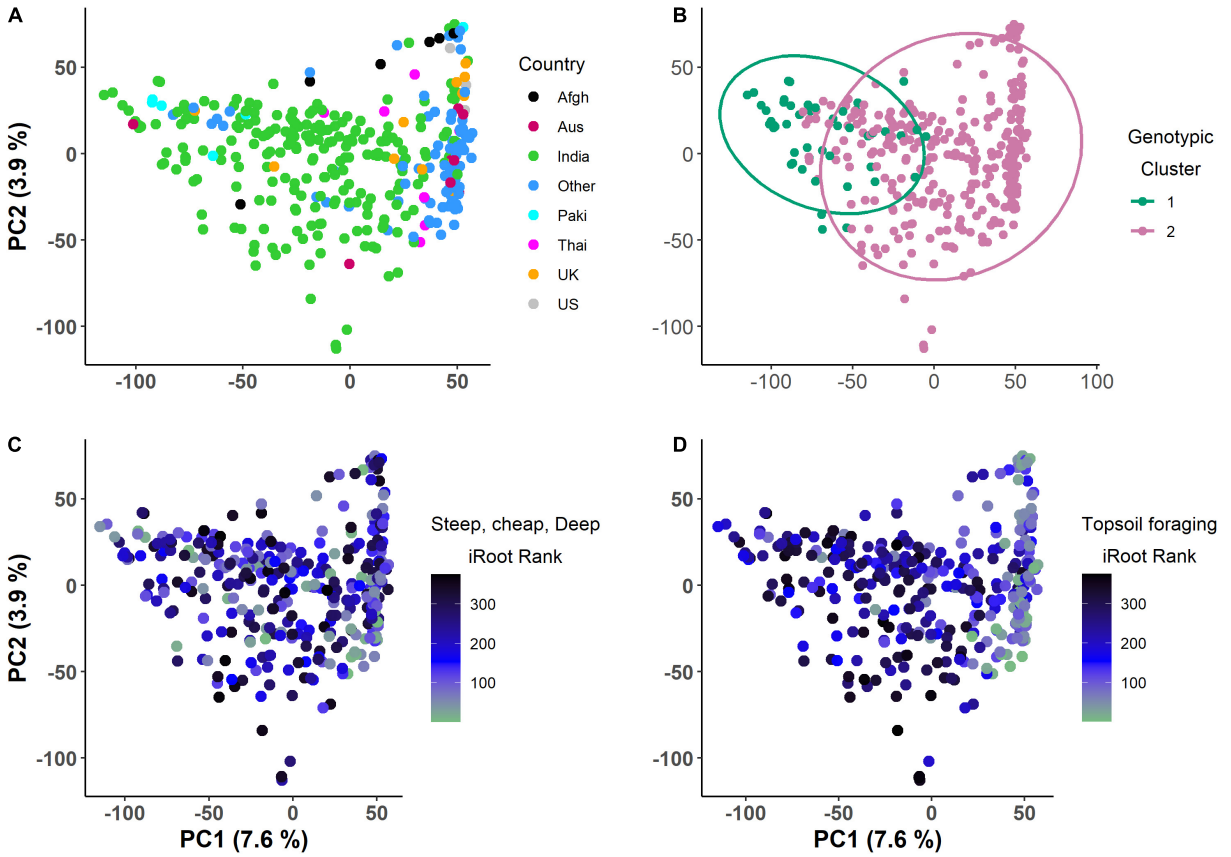

FIGURE 6 | Day 15 principal component analysis of the genotypes for the IA mung bean panel, **(A)** colored by country of origin, **(B)** colored by genotypic clusters, **(C)** color gradient showing ranking in the steep, cheap, and deep iRoot category, and **(D)** color gradient showing ranking in the topsoil foraging iRoot category (the lower the rank, the better the genotype).

marker 8_11481602 associated with LED from days 15 and 18 was found within the exon of a monodehydroascorbate reductase gene (LOC106772343). Day 18 SNP marker 8_10447903 was found within an intron for an uncharacterized gene but close to the LOC106772343 gene. Day 18 SNP marker 3_10004492 associated with TRL_GR is located within the exon for a gene coding for mannose-1-phosphate guanylyltransferase 1 (LOC106757974). Day 18 SNP maker 5_35265704 associated with TDW is located in the exon for a gene described as putative dehydration-responsive element-binding protein 2H (LOC106760865).

Selection of variables with embedded screening resulted in several significant markers for most of the traits across the 3 days (**Figure 9**, **Supplementary Figures 4–6**, and **Supplementary Table 5**). Two markers for LED (8_44518003) and TDW (5_35265704) from day 18 did overlap with TASSEL results. Marker 8_44518003 is an exon within the gene encoding monodehydroascorbate reductase (LOC106772343), while marker 5_35265704 was found within the gene encoding putative dehydration-responsive element-binding protein 2H (LOC106760865). Day 18 marker for TRL_{Upper} (2_22583526) was found within an exon in the gene encoding coilin (LOC106756657), while marker for DEP (5_23119832) was found in an exon within the gene encoding expansin-A11 (LOC106761944).

DISCUSSION

Controlled environments have been successfully used to study organisms out of their *in situ* environments (Crop Science Controlled Environment Research Guidelines, 2021). Plants in controlled environments may be exposed to similar conditions as would be in the field to help better achieve the objectives under study (Tibbitts and Langhans, 1993). There have been successful results for measuring various above-ground phenotypes in controlled environments, but below-ground phenotypes pose additional challenges (White et al., 2013). While studies in controlled environments do not imitate what *in situ* root environments look like, they are helpful in *a priori* screening of genotypes to minimize the heavy below-ground phenotyping work required in the field (Lynch and Brown, 2012; Li R. et al., 2015; Ye et al., 2018).

Mung beans are mostly grown on residual moisture after primary crops in most of Southeast Asia (Poehlman and Milton, 1991; Aski et al., 2021). In the Western world, mung beans

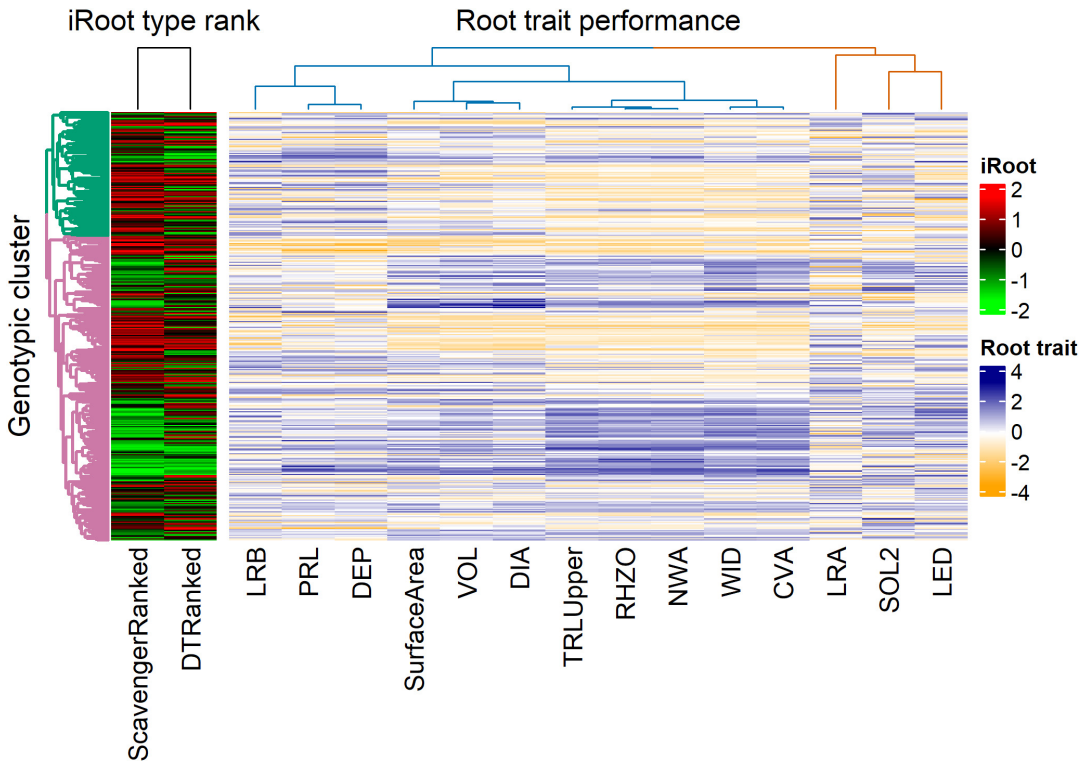

FIGURE 7 | Heat map revealing patterns between the genotypic clusters, iRoot category, and root traits performance for the genotypes in the IA mung bean panel. The dendrogram on the y-axis represents the genotypic clusters (green = 1 and pink = 2). The dendrograms on the top represent the iRoot category and root trait performance.

planted in the summers depend highly on the moisture residue, often following a wet, cold spring. In IA, mung beans are planted around the first week of June, capitalizing on the intense solar radiation for rapid growth (Sandhu and Singh, 2021). This would explain why mung beans, like other legume species, would be ideal with the "steep, cheap, and deep" root ideotypes, to chase the water and the soluble nitrogen before the establishment of root nodules for atmospheric nitrogen fixation. Schneider et al. (2021) showed that steep root angles improved nitrogen uptake *in silico* in maize. Using the OpenSimRoot model, an 11% increase in nitrogen uptake and a 4% increase in plant biomass were predicted at 40 days of growth (Schneider et al., 2021).

Lynch and Brown (2012) showed that common bean genotypes with wide basal root angles were superior in phosphorus (P) acquisition, while the ones with narrow basal root angles were superior in water acquisition during drought conditions. A recent study looked at the P efficiency of mung bean root morphology traits in low and optimum conditions (Reddy et al., 2020), a trait associated with topsoil foraging. They found Indian improved cultivars would be better with regards to P foraging. We identified the top genotypes, including PI425425 (India), PI425045 (Philippines), PI425551 (Korea), PI264686 (Philippines), and PI425085 (Sri Lanka), in the topsoil foraging (**Table 3**). Our hypothesis is that this represents the improved germplasm developed in India or after migration from India, while some of the lower ranks are still landraces or wild relatives, but this will need to be evaluated further in field conditions. For example, AVMU0201 is from Taiwan, the World Vegetable Center (Brassica, 2014) (formerly AVRDC), which has been breeding mung beans since the 1970s. Accessions PI425045 and PI264686 are from the Philippines, which also hosts a duplicate mung bean collection at the University of the Philippines, Los Banos (Poehlman and Milton, 1991).

We reported a wide variability of the root trait phenotype in the IA mung bean panel during the early stages of development (**Table 2**). Indian genotypes represented 24, 56, and 89% in the phenotypic clusters 1, 2, and 3 with an overall presence of 67% (**Supplementary Figure 4**). The high H for traits on days 15 and 18 could be due to better capture of the traits by ARIA unlike day 12 as it might be too early for trait development and differentiation. The high correlation could also be explained by the fact that young plants are utilizing all nutrients for the vegetative growth. These conclusions cannot be assumed to represent the rest of the developmental stages of mung bean plants, prompting the need for further studies. Similar observations were made in soybean (Falk et al., 2020b).

Genetic variability is one of the most important factors in a breeder's toolbox (Cobb et al., 2019). Indian genotypes were 76 and 67% in the genotypic clusters 1 and 2 with an overall presence of 63% (**Supplementary Figure 4**). The lack of clear subpopulations as indicated by the PCs shows the homogeneity within the mungbean accessions (**Figure 6**). Previous studies

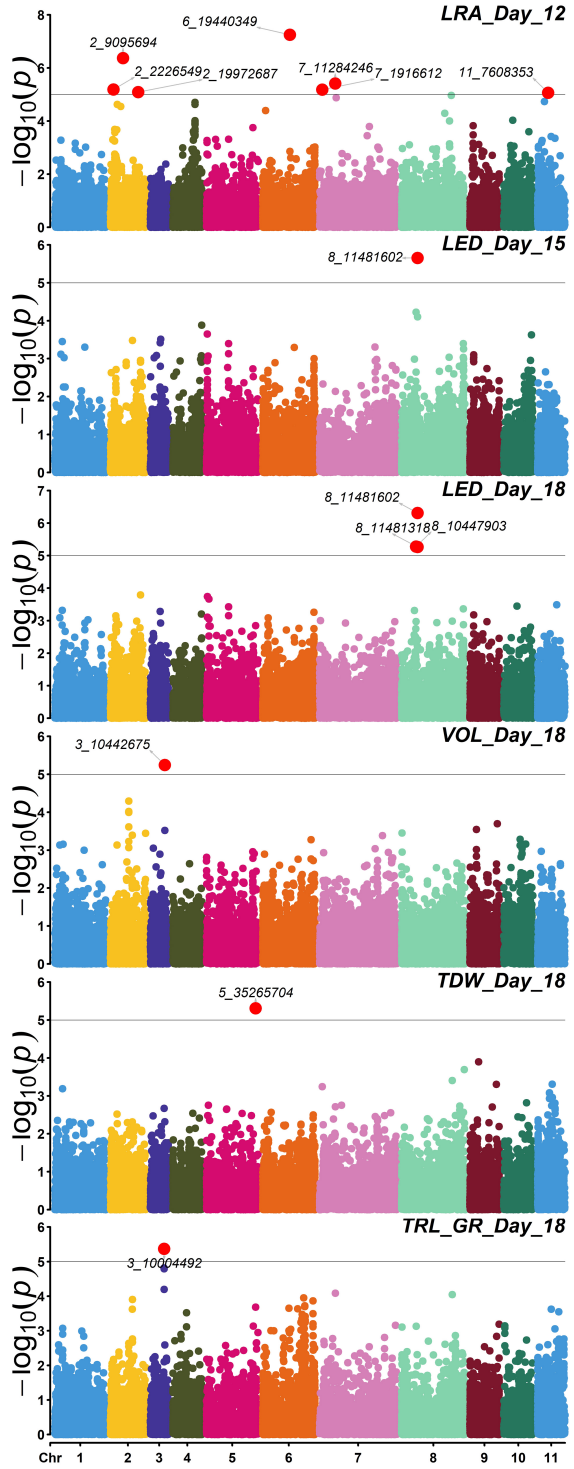

FIGURE 8 | Manhattan plots of –log10(*p*) vs chromosomes of SNP markers associated with the mung bean traits lateral root angles (LRA) (day 12), length distribution (LED) (day 15), and volume (VOL), LED, total dry weight (TDW), total root length-growth rate (TRL_GR) for day 18. The threshold line is the Bonferroni correction at $p = 0.05$ [at –log10(*p*) = 5]. Significant SNPs are highlighted in red and annotated with the marker name. Trait—SNP associations performed using the Trait Analysis by aSSociation, Evolution, and Linkage (TASSEL) software.

have shown similar results using simple sequence repeat markers in the Indonesian germplasm (Lestari et al., 2014) and the USDA germplasm (Wang et al., 2018). In other studies, the STRUCTURE (Pritchard et al., 2000) analysis showed between 3 and 6 subpopulations although no clear pattern was seen according to their geographical origins (Lestari et al., 2014; Wang et al., 2018; Sandhu and Singh, 2021). A similar observation, attributed to population admixture, was shown in common bean (Burle et al., 2010). The low Fst of 0.05 shows a high gene flow or low differentiation between the two genotypic clusters. The low Fst in this study confirms similar earlier reports within the USDA collections (Wang et al., 2018) and within the Indonesian germplasm (Lestari et al., 2014). Overall, results indicate a narrow genetic diversity in mung bean (Poehlman and Milton, 1991; Singh D. P. et al., 2021). Early breeders had to opt for mutation breeding to increase the genetic diversity. The narrow genetic base can be explained by the self-pollinated nature, very low cross-pollination frequencies, and poor hybridization of mung bean with other *Vigna* species (Poehlman and Milton, 1991; Singh D. P. et al., 2021). The narrow genetic diversity within the IA panel seems to reflect the fact that most of the accessions were collected on the Indian subcontinent, where mung bean was domesticated (Fuller, 2007). Our results support the idea that, in pulses, the lack of genetic diversity is due in part to the continuous use of a few genotypes as parents in the population development (Kumar et al., 2011). This shows the urgency of breeding efforts to diversify the genetic basis.

Adaptive roots to biotic and abiotic stresses will play a key role in bridging the yield gap in crop plants in the changing climate. A solid understanding of the genetic and environmental factors impacting the RSA will be important to the breeding of stable cultivars (see review, Lynch, 2007; Koevoets et al., 2016). RSA traits associated with response to abiotic stresses, including nutrient deficiency, drought tolerance, salinity, flooding, and temperature, and the underlying candidate genes have previously been studied (see review, Koevoets et al., 2016). Narrow LRA, high LED, and increased LRB were highly correlated to high P accumulation in *Arabidopsis* (Gruber et al., 2013), maize (Zhu et al., 2005), and common bean (Bonser et al., 1996). Auxins and strigolactones are key regulators in root and shoot development. An auxin receptor TRANSPORT INHIBITOR RESPONSE1 (TIR1) was shown to be responsible for the change in LRB as a response to low P levels (Pérez-Torres et al., 2008). Reduced LRB and increased PRL are characteristics of the "steep, cheap, and deep" ideotype, where the plant increases resource allocation to chase water and the mobile N in the deeper soil as evidenced in *Arabidopsis* and maize (Lynch, 2013). The nitrate transporters NRT1.1 and NRT2.1 were identified for the reduced LRB and increased PRL (Linkohr et al., 2002). The extended root system in *Arabidopsis* (Yu et al., 2008), rice, cotton, and poplar (Yu et al., 2013) was attributed to HD-ZIP transcription factor (HDG11) which promotes cell elongation by up-regulating cell wall loosening proteins hence important for drought tolerance.

In the current study, several putative candidate genes were identified for root traits associated with genes involved in the plant growth and development and stress tolerance response (**Table 4**). Lagrimini et al. (1997) proposed that

TABLE 4 | Significant single-nucleotide polymorphisms (SNPs) for results of association studies for traits across days 12, 15, and 18 as run in the Trait Analysis by aSSociation, Evolution, and Linkage (TASSEL) software.

Day	Trait	Marker	Chr	Pos	p	Add_effect	MarkerR2	Gene ID	Genomic context	Gene description
12	LRA	6_19440349	6	19440349	5.65E-08	−1.53E+00	0.10564	None	None	None
	LRA	2_9095694	2	9095694	4.26E-07	1.44515	0.08585	None	None	None
	LRA	7_11284246	7	11284246	3.86E-06	NaN	0.06085	None	None	None
	LRA	2_2226549	2	2226549	6.50E-06	−1.10E+00	0.06824	LOC106755829	Exon	Lignin-forming anionic peroxidase-like
	LRA	7_1916612	7	1916612	6.55E-06	−1.12E+00	0.06827	LOC106768494	Exon	Beta-galactosidase 3
	LRA	2_19972687	2	19972687	8.13E-06	1.54662	0.07675	LOC106753988	Exon	(–)-Germacrene D synthase-like
	LRA	11_7608353	11	7608353	8.66E-06	1.06558	0.06673	LOC106776541	Exon	Protein FAR1-RELATED SEQUENCE 5
15	LED	8_11481602	8	11481602	2.22E-06	0.09192	0.07531	LOC106772343	Exon	Monodehydroascorbate reductase
18	TRL_GR	3_10004492	3	10004492	4.26E-06	−2.48E+00	0.06752	LOC106757974	Exon	Mannose-1-phosphate guanylyltransferase 1
	LED	8_11481602	8	11481602	4.86E-07	0.1291	0.08614	LOC106772343	Exon	Monodehydroascorbate reductase
	LED	8_10447903	8	10447903	5.27E-06	−1.14E-01	0.07111	LOC106771882	Intron	Uncharacterized LOC106771882
	TDW	5_35265704	5	35265704	4.92E-06	0.00902	0.06543	LOC106760865	Exon	Putative dehydration-responsive element-binding protein 2H (DREB2)

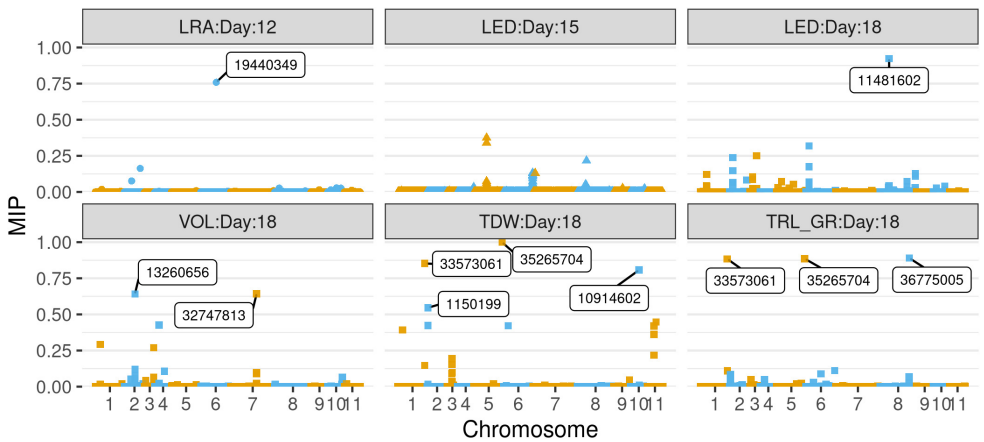

FIGURE 9 | Selection of variables with embedded screening (SVEN) plots of marginal inclusion probability (MIP) vs chromosomes of SNP markers associated with the mung bean traits LRA (day 12), LED (day 15), and LED, VOL, TDW, TRL_GR for day 18. Significant SNPs are boxed with the marker name.

anionic peroxidases, associated with LOC106755829 (day 12 LRA), play a role in plant host defense using a transformed tobacco (*Nicotiana tabacum* L.) plant. They have also been identified as major enzymes in cell wall lignification and found in large quantities in the xylem tissue (Sasaki et al., 2007). (-)- Germacrene D synthase, associated with LOC106753988 (day 12 LRA), is a member of sesquiterpene synthases family of plant proteins that have the capability of converting a precursor molecule farnesyl diphosphate into many sesquiterpene isoforms (Picaud et al., 2006). (-)- Germacrene D synthase catalyzes the formation (-)- Germacrene D, which is known to have strong effects on insects. Beta-galactosidase 3 is associated with Loci LOC106768494 (day 12 LRA), which has been implicated in adventitious root development *via* transcriptomic studies in mung bean (Li S.-W. et al., 2015). In rice, beta-galactosidase 1 and 2 were found to be highly expressed in the root and shoot seedlings, with less expression in flowers and immature seeds. Beta-galactosidases are important in the breakdown of molecular complexes (carbohydrates, glycolipids, and glycoproteins) that contain galactose (Chantarangsee et al., 2007). Beta-galactosidases would be important in the supply of the required energy from storage reserves during the rapid growth phase. The Far-related sequence (FRS) family, associated with LOC106776541 (day 12 LRA), is conserved among plants. Genes in this family are involved in multiple cellular processes (Lin Ma-FAR1). For example, *Arabidopsis* (*Arabidopsis thaliana*)

mutants of *fhy3* were less sensitive to both osmotic and salinity stress while also reducing the ABA-dependent inhibition of seedling root elongation, seedling greening, and germination (Tang et al., 2013; Ma and Li, 2018).

Transcripts of the gene encoding monodehydroascorbate reductase associated with LOC106772343 (day 15 and 18 LED), an antioxidant enzyme, were significantly reduced in the root elongation zone when roots for tall fescue (*Festuca arundinaceaSchreb.cv.* "K-31") when exposed to water stress (Xu et al., 2015). Water stress is associated with high concentration of reactive oxygen species. A mutation in the *Arabidopsis CYT1* gene encoding mannose-1-phosphate guanylyltransferase 1 associated with LOC106757974 (day 18 TRL_GR) showed deficiency in the cell wall after depletion of GDP mannose. The mutants exhibited radial swelling and accumulation of callose at the root tip. The functional analysis revealed mannose-1-phosphate guanylyltransferase 1 is involved in N-glycosylation during the cellulose synthesis (Lukowitz et al., 2001). An orthologous gene (DREB1A/CBF3 and DREB2A) associated with LOC106760865 (day 18 TDW), in *Arabidopsis*, encodes transcription factors that are involved in activating downstream genes involved in drought and cold stress (Sakuma et al., 2006). In another study, DREB2A proteins were found to increase the stress tolerance by modulating root architecture traits like the lateral root number and root length (Shukla et al., 2006; Agarwal et al., 2010).

Selection of variables with embedded screening loci associated with LOC106756657 (day 18 TRLUpper) and LOC106761944 (day 18 DEP) were associated with the adventitious root development in mung bean like the TASSEL results (Li S.-W. et al., 2015). Coilin is important in the formation of Cajal bodies, which are mostly associated with RNA processes. Kanno et al. (2016) suggest that coilin may be acting in multiple levels fine tuning expression of some genes important for environmental adaptation. Expansins are proteins involved with cell wall loosening and modification, partly mediated by the pH expansion of the cell wall during plant growth (Lee et al., 2001). In rice, Zhiming et al. (2011) identified a gene encoding *EXPA17* that was important for the root hair growth, which requires intensive cell wall modification.

The high H among the dry weight measurements can be used in the selection of parents with the root to shoot ratio (RSR) previously used as a measure of the photosynthetic materials allocations (**Figure 3**). During a low supply of water, nitrogen, and phosphorus in the soil, more resources are allocated to roots relative to shoots (Xu et al., 2015; Lynch et al., 2021). Within legumes QTLs for fibrous rooting/surface area (Abdel-Haleem et al., 2011), root length (Prince et al., 2015), lateral root number, and root thickness (Manavalan et al., 2015; Prince et al., 2019) in soybean have been mapped. In cowpea, QTLs for basal root angle, root diameter, median width, and width accumulation were reported (Burridge et al., 2017). In pea, root length QTL (Fondevilla et al., 2010) and in common bean basal root angle QTL have been identified. Root length density, root surface area, RDW ratio, and root depth in chickpea have been mapped (Jaganathan et al., 2015). In cereals for, maize and sorghum associations with area, convex hull area, median width, maximum width, width-profile angle, and adjusted depth were identified (Zheng et al., 2020), deep root mass, and the number of deep roots in rice (Courtois et al., 2013) and PRL, RDW in wheat (Sanguineti et al., 2007).

Our study has elucidated the phenotypic and genotypic variability for the root traits in the 375 genotypes in the IA mung bean panel. We identified candidate genotypes that can now be advanced to the greenhouse or field for further testing, especially for the root ideotypes. If their trait response and expression can be confirmed, these can be utilized as parents in the breeding program. Using GWAS, we identified significant markers associated with several RSA traits. Taken together, the ideotypes after field evaluation and significant markers can be utilized as tools for marker-assisted selection and crop improvement in mung bean breeding programs.

DATA AVAILABILITY STATEMENT

A subset of the dataset and the R scripts used in the analysis can be found in the github account mungbeanpaper, https://github.com/yalek/mungbeanpaper.git. Raw data is provided in the **Supplementary Material**. Marker dataset can be found at Dryad Data, doi: 10.5061/dryad.wdbrv15mb. Images can be provided upon request.

AUTHOR CONTRIBUTIONS

KC and AS conceptualized and designed the experiments. KC performed data collection with the assistance of student workers, did the image preprocessing, data analysis, and wrote the first draft manuscript with feedback from AS and SC on GWAS, and SD and AS on phenotypic analysis. TJ helped with image preprocessing. BG provided feedback on root imaging software. AS provided overall leadership on the project. All authors revised and approved the submitted manuscript.

FUNDING

This material is based upon work supported by the National Science Foundation under Predictive Plant Phenomics (P3) Grant No. DGE-1545453 (KC), the USDA National Institute of Food and Agriculture (NIFA) Food and Agriculture Cyberinformatics Tools (FACT) (award 2019-67021-29938) (AS and BG), RF Baker Funding (AS), USDA-NIFA Hatch project (IOW03717) (SD), USDA-CRIS (IOW04714) project (AS), USDA Agricultural Research Service, project 5030-21000-069-00D (SC) and AI Institute for Resilient Agriculture (AIIRA), supported by the NSF and USDA award #2021-67021-35329 to BG and AS.

ACKNOWLEDGMENTS

We are grateful to Asheesh K. Singh for his infrastructural support on the project. We are grateful to Ken Moore for initial ideas on the experimental designs. We would like to thank the undergraduate students Grace Heck, Megan Besch, Michael Cook, Erin Stichter, and Melinda Zubrod for their unwavering support of this project. We also thank all the graduate students

Joscif Raigne, Clayton Carley, Liza Van der Laan, Matthew Carroll, Ashlyn Rairdin, and Sarah Jones for their immense help in data collection and discussions. We also express gratitude to Jennifer Hicks for helping with procuring materials for the project. We are grateful to Aaron Brand for making sure the growth chambers ran smoothly. The findings and conclusions in this publication are those of the author(s) and should not be construed to represent any official USDA or US Government determination or policy. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture. USDA is an equal opportunity provider and employer.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.808001/full#supplementary-material

REFERENCES

Abdel-Haleem, H., Lee, G.-J., and Boerma, R. H. (2011). Identification of QTL for increased fibrous roots in soybean. *Theor. Appl. Genet.* 122, 935–946. doi: 10.1007/s00122-010-1500-9

Agarwal, P., Agarwal, P. K., Joshi, A. J., Sopory, S. K., and Reddy, M. K. (2010). Overexpression of PgDREB2A transcription factor enhances abiotic stress tolerance and activates downstream stress-responsive genes. *Mol. Biol. Rep.* 37, 1125–1135. doi: 10.1007/s11033-009-9885-8

Akibode, S., and Maredia, M. (2012). *Global and Regional Trends in Production, Trade and Consumption of Food Legume Crops*. Available online at: https://ageconsearch.umn.edu/record/136293 (accessed October 11, 2021).

Armengaud, P. (2009). EZ-Rhizo software: the gateway to root architecture analysis. *Plant Signal. Behav.* 4, 139–141. doi: 10.4161/psb.4.2.7763

Aschemann-Witzel, J., Gantriis, R. F., Fraga, P., and Perez-Cueto, F. J. A. (2020). Plant-based food and protein trend from a business perspective: markets, consumers, and the challenges and opportunities in the future. *Crit. Rev. Food Sci. Nutr.* 61, 3119–3128. doi: 10.1080/10408398.2020.1793730

Aski, M. S., Rai, N., Reddy, V. R. P., Gayacharan, Dikshit, H. K., Mishra, G. P., et al. (2021). Assessment of root phenotypes in mungbean mini-core collection (MMC) from the World Vegetable Center (AVRDC) Taiwan. *PLoS One* 16:e0247810. doi: 10.1371/journal.pone.0247810

Atkinson, J. A., Pound, M. P., Bennett, M. J., and Wells, D. M. (2019). Uncovering the hidden half of plants using new advances in root phenotyping. *Curr. Opin. Biotechnol.* 55, 1–8. doi: 10.1016/j.copbio.2018.06.002

Atkinson, J. A., Wingen, L. U., Griffiths, M., Pound, M. P., Gaju, O., Foulkes, M. J., et al. (2015). Phenotyping pipeline reveals major seedling root growth QTL in hexaploid wheat. *J. Exp. Bot.* 66, 2283–2292. doi: 10.1093/jxb/erv006

Betegón-Putze, I., González, A., Sevillano, X., Blasco-Escámez, D., and Caño-Delgado, A. I. (2019). MyROOT: a method and software for the semiautomatic measurement of primary root length in Arabidopsis seedlings. *Plant J.* 98, 1145–1156. doi: 10.1111/tpj.14297

Bonser, A. M., Lynch, J., and Snapp, S. (1996). Effect of phosphorus deficiency on growth angle of basal roots in Phaseolus vulgaris. *New Phytol.* 132, 281–288. doi: 10.1111/j.1469-8137.1996.tb01847.x

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23, 2633–2635. doi: 10.1093/bioinformatics/btm308

Brassica (2014). *Home - World Vegetable Center*. Available online at: https://avrdc.org/ (accessed October 09, 2021).

Burle, M. L., Fonseca, J. R., Kami, J. A., and Gepts, P. (2010). Microsatellite diversity and genetic structure among common bean (*Phaseolus vulgaris* L.) landraces in Brazil, a secondary center of diversity. *Theor. Appl. Genet.* 121, 801–813. doi: 10.1007/s00122-010-1350-5

Burridge, J. D., Schneider, H. M., Huynh, B.-L., Roberts, P. A., Bucksch, A., and Lynch, J. P. (2017). Genome-wide association mapping and agronomic impact of cowpea root architecture. *Theor. Appl. Genet.* 130, 419–431. doi: 10.1007/s00122-016-2823-y

Chantarangsee, M., Tanthanuch, W., Fujimura, T., Fry, S. C., and Ketudat Cairns, J. (2007). Molecular characterization of β-galactosidases from germinating rice (*Oryza sativa*). *Plant Sci.* 173, 118–134. doi: 10.1016/j.plantsci.2007.04.009

Cobb, J. N., Juma, R. U., Biswas, P. S., Arbelaez, J. D., Rutkoski, J., Atlin, G., et al. (2019). Enhancing the rate of genetic gain in public-sector plant breeding programs: lessons from the breeder's equation. *Theor. Appl. Genet.* 132, 627–645. doi: 10.1007/s00122-019-03317-0

Courtois, B., Audebert, A., Dardou, A., Roques, S., Ghneim-Herrera, T., Droc, G., e1t al. (2013). Genome-wide association mapping of root traits in a japonica rice panel. *PLoS One* 8:e78037. doi: 10.1371/journal.pone.0078037

Crop Science Controlled Environment Research Guidelines (2021). Available online at: https://www.crops.org/publications/journals/author-resources/cs-instructions/controlled-environment-research/ (accessed October 6, 2021).

Cullis, B. R., Smith, A. B., and Coombes, N. E. (2006). On the design of early generation variety trials with correlated data. *J. Agric. Biol. Environ. Stat.* 11, 381–393.

Das, A., Schneider, H., Burridge, J., Ascanio, A. K. M., Wojciechowski, T., Topp, C. N., et al. (2015). Digital imaging of root traits (DIRT): a high-throughput computing and collaboration platform for field-based root phenomics. *Plant Methods* 11:51. doi: 10.1186/s13007-015-0093-3

Dray Stéphane, D. A.-B. (2007). The ade4 Package: implementing the duality diagram for ecologists. *J. Stat. Softw.* 22, 1–20.

Edmondson, R. N., and Edmondson, M. R. (2021). *Package 'blocksdesign'*. Available online at: https://cran.r-project.org (accessed October 27, 2021).

Falk, K. G., Jubery, T. Z., O'Rourke, J. A., Singh, A., Sarkar, S., Ganapathysubramanian, B., et al. (2020b). Soybean root system architecture trait study through genotypic, phenotypic, and shape-based clusters. *Plant Phenomics* 2020:1925495. doi: 10.34133/2020/1925495

Falk, K. G., Jubery, T. Z., Mirnezami, S. V., Parmley, K. A., Sarkar, S., Singh, A., et al. (2020a). Computer vision and machine learning enabled soybean root phenotyping pipeline. *Plant Methods* 16:5. doi: 10.1186/s13007-019-0550-5

Fang, C., Ma, Y., Wu, S., Liu, Z., Wang, Z., Yang, R., et al. (2017). Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. *Genome Biol.* 18:161. doi: 10.1186/s13059-017-1289-9

Fernandez, G. C. J., and Shanmugasundaram, S. (1988). "The AVRDC mungbean improvement program: the past, present and future," in *Proceedings of the 1988 Second International Symposium held at Bangkok*, Bangkok, 58–70.

Fondevilla, S., Fernández-Aparicio, M., Satovic, Z., Emeran, A. A., Torres, A. M., Moreno, M. T., et al. (2010). Identification of quantitative trait loci for specific mechanisms of resistance to Orobanche crenata Forsk. in pea (*Pisum sativum* L.). *Mol. Breed.* 25, 259–272. doi: 10.1007/s11032-009-9330-7

Fuller, D. Q. (2007). Contrasting patterns in crop domestication and domestication rates: recent archaeobotanical insights from the Old World. *Ann. Bot.* 100, 903–924. doi: 10.1093/aob/mcm048

Galili, T. (2015). dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* 31, 3718–3720. doi: 10.1093/bioinformatics/btv428

Galkovskyi, T., Mileyko, Y., Bucksch, A., Moore, B., Symonova, O., Price, C. A., et al. (2012). GiA Roots: software for the high throughput analysis of plant root system architecture. *BMC Plant Biol.* 12:116. doi: 10.1186/1471-2229-12-116

Gaur, P. M., Krishnamurthy, L., and Kashiwagi, J. (2008). Improving Drought-Avoidance Root Traits in Chickpea (*Cicer arietinum* L.) -Current Status of Research at ICRISAT. *Plant Prod. Sci.* 11, 3–11.

Ghosal, S., Blystone, D., Singh, A. K., Ganapathysubramanian, B., Singh, A., and Sarkar, S. (2018). An explainable deep machine vision framework for plant stress phenotyping. *Proc. Natl. Acad. Sci. U.S.A.* 115, 4613–4618. doi: 10.1073/pnas.1716999115

Ghosal, S., Zheng, B., Chapman, S. C., Potgieter, A. B., Jordan, D. R., Wang, X., et al. (2019). A weakly supervised deep learning framework for sorghum head detection and counting. *Plant Phenomics* 2019, 1525874. doi: 10.34133/2019/1525874

Gioia, T., Galinski, A., Lenz, H., Müller, C., Lentz, J., Heinz, K., et al. (2016). GrowScreen-PaGe, a non-invasive, high-throughput phenotyping system based on germination paper to quantify crop phenotypic diversity and plasticity of root traits under varying nutrient supply. *Funct. Plant Biol.* 44, 76–93. doi: 10.1071/FP16128

Gruber, B. D., Giehl, R. F. H., Friedel, S., and von Wirén, N. (2013). Plasticity of the Arabidopsis root system under nutrient deficiencies. *Plant Physiol.* 163, 161–179. doi: 10.1104/pp.113.218453

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* 32, 2847–2849. doi: 10.1093/bioinformatics/btw313

Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). circlize Implements and enhances circular visualization in R. *Bioinformatics* 30, 2811–2812. doi: 10.1093/bioinformatics/btu393

Hart, F. (2021). *Smart Shooter Photography Software*. Available on: https://kuvacode.com/smartshooter3 (accessed September 13, 2021).

Hoaglin, D. C. (2003). John W. Tukey and data analysis. *Stat. Sci.* 18, 311–318.

Hodge, A., Berta, G., Doussan, C., Merchan, F., and Crespi, M. (2009). Plant root growth, architecture and function. *Plant Soil* 321, 153–187.

Huang, X., and Han, B. (2014). Natural variations and genome-wide association studies in crop plants. *Annu. Rev. Plant Biol.* 65, 531–551. doi: 10.1146/annurev-arplant-050213-035715

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., et al. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* 42, 961–967. doi: 10.1038/ng.695

Hund, A., Trachsel, S., and Stamp, P. (2009). Growth of axile and lateral roots of maize: I development of a phenotying platform. *Plant Soil* 325, 335–349. doi: 10.1007/s11104-009-9984-2

Jaganathan, D., Thudi, M., Kale, S., Azam, S., Roorkiwal, M., Gaur, P. M., et al. (2015). Genotyping-by-sequencing based intra-specific genetic map refines a "QTL-hotspot" region for drought tolerance in chickpea. *Mol. Genet. Genomics* 290, 559–571. doi: 10.1007/s00438-014-0932-3

Jahan, I., Rahman, M. M., Tuzzohora, M. F., Hossain, M. A., Begum, S. N., Burritt, D. J., et al. (2020). Phenotyping of mungbean (*Vigna radiata* L.) genotypes against salt stress and assessment of variability for yield and yield attributing traits. *J. Plant Stress Physiol.* 6, 7–17.

Joshi, V., and Kumar, S. (2016). *Meat analogues: plant based alternatives to meat products- a review*. *Int. J. Food Ferment. Technol.* 5, 107–119. doi: 10.5958/2277-9396.2016.00001.5

JPEGCrops (2021). Available online at: http://ekot.dk/programmer/JPEGCrops/ (accessed September 13, 2021).

Kamfwa, K., Cichy, K. A., and Kelly, J. D. (2015). Genome-wide association study of agronomic traits in common bean. *Plant Genome* 8:elantgenome2014.09.0059.

Kanno, T., Lin, W. D., Fu, J. L., Wu, M. T., Yang, H. W., Lin, S. S., et al. (2016). Identification of coilin mutants in a screen for enhanced expression of an alternatively spliced GFP reporter gene in *Arabidopsis thaliana*. Genetics 203, 1709–1720. doi: 10.1534/genetics.116.190751

Kassambara, A., and Mundt, F. (2020). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. Available online at: https://CRAN.R-project.org/package=factoextra (accessed September 27, 2021).

Kang, Y. J., Kim, S. K., Kim, M. Y., Lestari, P., Kim, K. H., Ha, B.-K., et al. (2014). Genome sequence of mungbean and insights into evolution within *Vigna* species. *Nat. Commun.* 5:5443. doi: 10.1038/ncomms6443

Koevoets, I. T., Venema, J. H., Elzenga, J. T. M., and Testerink, C. (2016). Roots withstanding their environment: exploiting root system architecture responses to abiotic stress to improve crop tolerance. *Front. Plant Sci.* 7:1335. doi: 10.3389/fpls.2016.01335

Kumar, J., Choudhary, A. K., Solanki, R. K., and Pratap, A. (2011). Towards marker-assisted selection in pulses: a review. *Plant Breed.* 130, 297–313. doi: 10.1007/s00299-017-2127-y

Kuo, K. H. M. (2017). Multiple testing in the context of gene discovery in sickle cell disease using genome-wide association studies. *Genomics Insights* 10:1178631017721178. doi: 10.1177/1178631017721178

Lagrimini, L. M., Gingas, V., Finger, F., Rothstein, S., and Liu, T. (1997). Characterization of antisense transformed plants deficient in the tobacco anionic peroxidase. *Plant Physiol.* 114, 1187–1196. doi: 10.1104/pp.114.4.1187

Le Bot, J., Serra, V., Fabre, J., Draye, X., Adamowicz, S., and Pagès, L. (2010). DART: a software to analyse root system architecture and development from captured images. *Plant Soil* 326, 261–273.

Lee, Y., Choi, D., and Kende, H. (2001). Expansins: ever-expanding numbers and functions. *Curr. Opin. Plant Biol.* 4, 527–532. doi: 10.1016/s1369-5266(00)00211-9

Lestari, P., Kim, S. K., Reflinur, Kang, Y. J., Dewi, N., and Lee, S.-H. (2014). Genetic diversity of mungbean (*Vigna radiata* L.) germplasm in Indonesia. *Plant Genet. Resour.* .12, S91–S94.

Li, D., Dutta, S., and Roy, V. (2020). Model Based Screening Embedded Bayesian Variable Selection for Ultra-high Dimensional Settings. *arXiv* [Preprint]. Available online at: http://arxiv.org/abs/2006.07561 (accessed September 27, 2021).

Li, R., Zeng, Y., Xu, J., Wang, Q., Wu, F., Cao, M., et al. (2015). Genetic variation for maize root architecture in response to drought stress at the seedling stage. *Breed. Sci.* 65, 298–307. doi: 10.1270/jsbbs.65.298

Li, S.-W., Shi, R.-F., and Leng, Y. (2015). *De novo* characterization of the mung bean transcriptome and transcriptomic analysis of adventitious rooting in seedlings using RNA-Seq. *PLoS One* 10:e0132969. doi: 10.1371/journal.pone.0132969

Linkohr, B. I., Williamson, L. C., Fitter, A. H., and Leyser, H. M. O. (2002). Nitrate and phosphate availability and distribution have different effects on root system architecture of Arabidopsis. *Plant J.* 29, 751–760. doi: 10.1046/j.1365-313x.2002.01251.x

Liu, X., Dong, X., Xue, Q., Leskovar, D. I., Jifon, J., Butnor, J. R., et al. (2018). Ground penetrating radar (GPR) detects fine roots of agricultural crops in the field. *Plant Soil* 423, 517–531. doi: 10.1007/s11104-017-3531-3

Lobet, G., Paez-Garcia, A., Schneider, H., Junker, A., Atkinson, J. A., and Tracy, S. (2019). Demystifying roots: a need for clarification and extended concepts in root phenotyping. *Plant Sci.* 282, 11–13. doi: 10.1016/j.plantsci.2018.09.015

Lobet, G., Pagès, L., and Draye, X. (2011). A novel image-analysis toolbox enabling quantitative analysis of root system architecture. *Plant Physiol.* 157, 29–39. doi: 10.1104/pp.111.179895

Lozano-Isla, F. (2021). *inti: Tools and Statistical Procedures in Plant Science*. Available online at: https://cran.r-project.org/web/packages/inti/citation.html (accessed September 13, 2021).

Lukowitz, W., Nickle, T. C., Meinke, D. W., Last, R. L., Conklin, P. L., and Somerville, C. R. (2001). *Arabidopsis cyt1* mutants are deficient in a mannose-1-phosphate guanylyltransferase and point to a requirement of N-linked glycosylation for cellulose biosynthesis. *Proc. Natl. Acad. Sci. U.S.A.* 98, 2262–2267. doi: 0.1073/pnas.051625798

Lynch, J. P. (2007). Roots of the second green revolution. *Aust. J. Bot.* 55, 493–512.

Lynch, J. P. (2013). Steep, cheap and deep: an ideotype to optimize water and N acquisition by maize root systems. *Ann. Bot.* 112, 347–357. doi: 10.1093/aob/mcs293

Lynch, J. P., and Brown, K. M. (2001). Topsoil foraging – an architectural adaptation of plants to low phosphorus availability. *Plant Soil* 237, 225–237.

Lynch, J. P., and Brown, K. M. (2012). New roots for agriculture: exploiting the root phenome. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367, 1598–1604. doi: 10.1098/rstb.2011.0243

Lynch, J. P., Strock, C. F., Schneider, H. M., Sidhu, J. S., Ajmera, I., Galindo-Castañeda, T., et al. (2021). Root anatomy and soil resource capture. *Plant Soil* 466, 21–63.

Ma, L., and Li, G. (2018). FAR1-RELATED SEQUENCE (FRS) and FRS-RELATED FACTOR (FRF) family proteins in *Arabidopsis* growth and development. *Front. Plant Sci.* 9:692. doi: 10.3389/fpls.2018.00692

Ma, L., Shi, Y., Siemianowski, O., Yuan, B., Egner, T. K., Mirnezami, S. V., et al. (2019). Hydrogel-based transparent soils for root phenotyping *in vivo*. *Proc. Natl. Acad. Sci. U.S.A.* 116, 11063–11068. doi: 10.1073/pnas.1820334116

Manavalan, L. P., Prince, S. J., Musket, T. A., Chaky, J., Deshmukh, R., Vuong, T. D., et al. (2015). Identification of novel QTL governing root architectural traits in an interspecific soybean population. *PLoS One* 10:e0120490. doi: 10.1371/journal.pone.0120490

Markiewicz, K. (2010). *The Economics of Meeting Future Protein Demand*. Wageningen: Wageningen University.

Nagasubramanian, K., Jones, S., Sarkar, S., Singh, A. K., Singh, A., and Ganapathysubramanian, B. (2018). Hyperspectral band selection using genetic algorithm and support vector machines for early identification of charcoal rot disease in soybean stems. *Plant Methods* 14:86. doi: 10.1186/s13007-018-0349-9

Nagasubramanian, K., Jones, S., Singh, A. K., Sarkar, S., Singh, A., and Ganapathysubramanian, B. (2019). Plant disease identification using explainable 3D deep learning on hyperspectral images. *Plant Methods* 15:98. doi: 10.1186/s13007-019-0479-8

Naik, H. S., Zhang, J., Lofquist, A., Assefa, T., Sarkar, S., Ackerman, D., et al. (2017). A real-time phenotyping framework using machine learning for plant stress severity rating in soybean. *Plant Methods* 13:23. doi: 10.1186/s13007-017-0173-7

Nair, R. M., Pandey, A. K., War, A. R., Hanumantharao, B., Shwe, T., Alam, A., et al. (2019). Biotic and abiotic constraints in mungbean production-progress in genetic improvement. *Front. Plant Sci.* 10:1340. doi: 10.3389/fpls.2019.01340

Niva, M., Vainio, A., and Jallinoja, P. (2017). "10 - Barriers to increasing plant protein consumption in western populations," in *Vegetarian and Plant-Based Diets in Health and Disease Prevention* (bll 157–171), ed. F. Mariotti (Cambridge, MA: Academic Press).

Pace, J., Lee, N., Naik, H. S., Ganapathysubramanian, B., and Lübberstedt, T. (2014). Analysis of maize (*Zea mays* L.) seedling roots with the high-throughput image analysis tool ARIA (Automatic Root Image Analysis). *PLoS One* 9:e108255. doi: 10.1371/journal.pone.0108255

Parmley, K. A., Higgins, R. H., Ganapathysubramanian, B., Sarkar, S., and Singh, A. K. (2019). Machine Learning Approach for Prescriptive Plant Breeding. *Sci. Rep.* 9:17132. doi: 10.1038/s41598-019-53451-4

Passot, S., Gnacko, F., Moukouanga, D., Lucas, M., Guyomarc'h, S., Ortega, B. M., et al. (2016). Characterization of pearl millet root architecture and anatomy reveals three types of lateral roots. *Front. Plant Sci.* 7:829. doi: 10.3389/fpls.2016.00829

Pataczek, L., Zahir, Z. A., Ahmad, M., Rani, S., Nair, R., Schafleitner, R., et al. (2018). Beans with Benefits—The Role of Mungbean (*Vigna radiate*) in a Changing Environment. *Am. J. Plant Sci.* 09, 1577–1600. doi: 10.4236/ajps.2018.97115

Pérez-Torres, C.-A., López-Bucio, J., Cruz-Ramírez, A., Ibarra-Laclette, E., Dharmasiri, S., Estelle, M., et al. (2008). Phosphate availability alters lateral root development in Arabidopsis by modulating auxin sensitivity via a mechanism involving the TIR1 auxin receptor. *Plant Cell* 20, 3258–3272. doi: 10.1105/tpc.108.058719

Picaud, S., Olsson, M. E., Brodelius, M., and Brodelius, P. E. (2006). Cloning, expression, purification and characterization of recombinant (+)-germacrene D synthase from Zingiber officinale. *Arch. Biochem. Biophys.* 452, 17–28. doi: 10.1016/j.abb.2006.06.007

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

Pierret, A., Gonkhamdee, S., Jourdan, C., and Maeght, J.-L. (2013). IJ_Rhizo: an open-source software to measure scanned images of root samples. *Plant Soil* 373, 531–539. doi: 10.1093/aobpla/plab056

Poehlman, and Milton, J. (1991). *The Mungbean*. New Delhi: Oxford & IBH Pub.

Pound, M. P., French, A. P., Atkinson, J. A., Wells, D. M., Bennett, M. J., and Pridmore, T. (2013). RootNav: navigating images of complex root architectures. *Plant Physiol.* 162, 1802–1814. doi: 10.1104/pp.113.221531

Prince, S. J., Song, L., Qiu, D., Maldonado Dos, Santos, J. V., Chai, C., et al. (2015). Genetic variants in root architecture-related genes in a *Glycine soja* accession, a potential resource to improve cultivated soybean. *BMC Genomics* 16:132. doi: 10.1186/s12864-015-1334-6

Prince, S. J., Valliyodan, B., Ye, H., Yang, M., Tai, S., Hu, W., et al. (2019). Understanding genetic control of root system architecture in soybean: insights into the genetic basis of lateral root number. *Plant Cell Environ.* 42, 212–229. doi: 10.1111/pce.13333

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959. doi: 10.1093/genetics/155.2.945

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Reddy, V. R. P., Aski, M. S., Mishra, G. P., Dikshit, H. K., Singh, A., Pandey, R., et al. (2020). Genetic variation for root architectural traits in response to phosphorus deficiency in mungbean at the seedling stage. *PLoS One* 15:e0221008. doi: 10.1371/journal.pone.0221008

Rellán-Álvarez, R., Lobet, G., Lindner, H., Pradier, P.-L., Sebastian, J., Yee, M.-C., et al. (2015). GLO-Roots: an imaging platform enabling multidimensional characterization of soil-grown root systems. *eLife* 4:e07597. doi: 10.7554/eLife.07597

Riera, L. G., Carroll, M. E., Zhang, Z., Shook, J. M., Ghosal, S., Gao, T., et al. (2021). Deep multiview image fusion for soybean yield estimation in breeding applications. *Plant Phenomics* 2021:9846470. doi: 10.34133/2021/9846470

Rogers, E. D., and Benfey, P. N. (2015). Regulation of plant root system architecture: implications for crop advancement. *Curr. Opin. Biotechnol.* 32, 93–98. doi: 10.1016/j.copbio.2014.11.015

Sakuma, Y., Maruyama, K., Osakabe, Y., Qin, F., Seki, M., Shinozaki, K., et al. (2006). Functional analysis of an *Arabidopsis* transcription factor, DREB2A, involved in drought-responsive gene expression. *Plant Cell* 18, 1292–1309. doi: 10.1105/tpc.105.035881

Sandhu, K., and Singh, A. (2021). Strategies for the utilization of the USDA mung bean germplasm collection for breeding outcomes. *Crop Sci.* 61, 422–442. doi: 10.1002/csc2.20322

Sanguineti, M. C., Li, S., Maccaferri, M., Corneti, S., Rotondo, F., Chiari, T., et al. (2007). Genetic dissection of seminal root architecture in elite durum wheat germplasm. *Ann. Appl. Biol.* 151, 291–305. doi: 10.1111/j.1744-7348.2007.00198.x

Sasaki, S., Shimizu, M., Wariishi, H., Tsutsumi, Y., and Kondo, R. (2007). Transcriptional and translational analyses of poplar anionic peroxidase isoenzymes. *J. Wood Sci.* 53, 427–435. doi: 10.1007/s10086-007-0888-6

Schafleitner, R., Nair, R. M., Rathore, A., Wang, Y.-W., Lin, C.-Y., Chu, S.-H., et al. (2015). The AVRDC - The World Vegetable Center mungbean (*Vigna radiata*) core and mini core collections. *BMC Genomics* 16:344. doi: 10.1186/s12864-015-1556-7

Schmidt, P., Hartung, J., Rath, J., and Piepho, H.-P. (2019). Estimating broad-sense heritability with unbalanced data from agricultural cultivar trials. *Crop Sci.* 59, 525–536.

Schneider, H. M., Lor, V. S. N., Hanlon, M. T., Perkins, A., Kaeppler, S. M., Borkar, A. N., et al. (2021). Root angle in maize influences nitrogen capture and is regulated by calcineurin B-like protein (CBL)-interacting serine/threonine-protein kinase 15 (ZmCIPK15). *Plant Cell Environ.* doi: 10.1111/pce.14135 [Epub ahead of print].

Seethepalli, A., Guo, H., Liu, X., Griffiths, M., Almtarfi, H., Li, Z., et al. (2020). RhizoVision crown: an integrated hardware and software platform for root crown Phenotyping. *Plant Phenomics* 2020:3074916. doi: 10.34133/2020/3074916

Shanahan, P. W., Binley, A., Whalley, W. R., and Watts, C. W. (2015). The use of electromagnetic induction to monitor changes in soil moisture profiles beneath different wheat genotypes. *Soil Sci. Soc. Am. J.* 79, 459–466. doi: 10.2136/sssaj2014.09.0360

Shukla, R. K., Raha, S., Tripathi, V., and Chattopadhyay, D. (2006). Expression of CAP2, an APETALA2-family transcription factor from chickpea, enhances growth and tolerance to dehydration and salt stress in transgenic tobacco. *Plant Physiol.* 142, 113–123. doi: 10.1104/pp.106.081752

Singh, A., Ganapathysubramanian, B., Singh, A. K., and Sarkar, S. (2016). Machine learning for high-throughput stress phenotyping in plants. *Trends Plant Sci.* 21, 110–124. doi: 10.1016/j.tplants.2015.10.015

Singh, A., Jones, S., Ganapathysubramanian, B., Sarkar, S., Mueller, D., Sandhu, K., et al. (2021). Challenges and opportunities in machine-augmented plant stress phenotyping. *Trends Plant Sci.* 26, 53–69. doi: 10.1016/j.tplants.2020.07.010

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

Singh, A. K., Singh, A., Sarkar, S., Ganapathysubramanian, B., Schapaugh, W., Miguez, F. E., et al. (2021). "High-Throughput Phenotyping in Soybean," in *High-Throughput Crop Phenotyping* (bll 129–163), eds J. Zhou and H. T. Nguyen (Cham: Springer International Publishing).

Singh, D. P., Singh, A. K., and Singh, A. (2021). "Chapter 25 - Breeding of crop ideotypes," in *Plant Breeding and Cultivar Development* (bll 497–516), eds D. P. Singh, A. K. Singh, and A. Singh (Cambridge, MA: Academic Press).

Srayeddin, I., and Doussan, C. (2009). Estimation of the spatial variability of root water uptake of maize and sorghum at the field scale by electrical resistivity tomography. *Plant Soil* 319, 185–207.

Tan, K. H., and Nopamornbodi, V. (1979). Effect of different levels of humic acids on nutrient content and growth of corn (*Zea mays* L.). *Plant Soil* 51, 283–287.

Tang, W., Ji, Q., Huang, Y., Jiang, Z., Bao, M., Wang, H., et al. (2013). FAR-RED ELONGATED HYPOCOTYL3 and FAR-RED IMPAIRED RESPONSE1 transcription factors integrate light and abscisic acid signaling in Arabidopsis. *Plant Physiol.* 163, 857–866. doi: 10.1104/pp.113.224386

Tibbitts, T. W., and Langhans, R. W. (1993). "Controlled-environment studies," in *Photosynthesis and Production in a Changing Environment: A Field and Laboratory Manual* (bll 65–78), eds D. O. Hall, J. M. O. Scurlock, H. R. Bolhàr-Nordenkampf, R. C. Leegood, and S. P. Long (Dordrecht: Springer).

Tibbs Cortes, L., Zhang, Z., and Yu, J. (2021). Status and prospects of genome-wide association studies in plants. *Plant Genome* 14:e20077. doi: 10.1002/tpg2.20077

Trachsel, S., Kaeppler, S. M., Brown, K. M., and Lynch, J. P. (2011). Shovelomics: high throughput phenotyping of maize (*Zea mays* L.) root architecture in the field. *Plant Soil* 341, 75–87.

Vinnari, M. (2008). The future of meat consumption — Expert views from Finland. *Technol. Forecast. Soc. Change* 75, 893–904.

Wang, L., Bai, P., Yuan, X., Chen, H., Wang, S., Chen, X., et al. (2018). Genetic diversity assessment of a set of introduced mung bean accessions (*Vigna radiata* L.). *Crop J.* 6, 207–213. doi: 10.1016/j.cj.2017.08.004

Wasson, A., Bischof, L., Zwart, A., and Watt, M. (2016). A portable fluorescence spectroscopy imaging system for automated root phenotyping in soil cores in the field. *J. Exp. Bot.* 67, 1033–1043. doi: 10.1093/jxb/erv570

White, P. J., George, T. S., Gregory, P. J., Bengough, A. G., Hallett, P. D., and McKenzie, B. M. (2013). Matching roots to their environment. *Ann. Bot.* 112, 207–222. doi: 10.1093/aob/mct123

Wild, F., Czerny, M., Janssen, A. M., Kole, A. P. W., and Domig, K. J. (2014). The evolution of a plant-based alternative to meat: from niche markets to widely accepted meat alternatives. *Agro Food Ind. Hi Tech* 25, 45–49.

Wright, S. (1965). The interpretation of population structure by f-statistics with special regard to systems of mating. *Evolution* 19, 395–420.

Xu, W., Cui, K., Xu, A., Nie, L., Huang, J., and Peng, S. (2015). Drought stress condition increases root to shoot ratio via alteration of carbohydrate partitioning and enzymatic activity in rice seedlings. *Acta Physiol. Plant.* 37:9.

Yang, N., Lu, Y., Yang, X., Huang, J., Zhou, Y., Ali, F., et al. (2014). Genome wide association studies using a new nonparametric model reveal the genetic architecture of 17 agronomic traits in an enlarged maize association panel. *PLoS Genet.* 10:e1004573. doi: 10.1371/journal.pgen.1004573

Ye, H., Roorkiwal, M., Valliyodan, B., Zhou, L., Chen, P., Varshney, R. K., et al. (2018). Genetic diversity of root system architecture in response to drought stress in grain legumes. *J. Exp. Bot.* 69, 3267–3277. doi: 10.1093/jxb/ery082

Yin, L., Zhang, H., Tang, Z., Xu, J., Yin, D., Zhang, Z., et al. (2021). rMVP: a memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. *Genomics Proteomics Bioinformatics.* doi: 10.1016/j.gpb.2020.10.007 [Epub ahead of print].

Yu, H., Chen, X., Hong, Y.-Y., Wang, Y., Xu, P., Ke, S.-D., et al. (2008). Activated expression of an Arabidopsis HD-START protein confers drought tolerance with improved root system and reduced stomatal density. *Plant Cell* 20, 1134–1151. doi: 10.1105/tpc.108.058263

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38, 203–208. doi: 10.1038/ng1702

Yu, L., Chen, X., Wang, Z., Wang, S., Wang, Y., Zhu, Q., et al. (2013). Arabidopsis enhanced drought tolerance1/HOMEODOMAIN GLABROUS11 confers drought tolerance in transgenic rice without yield penalty. *Plant Physiol.* 162, 1378–1391. doi: 10.1104/pp.113.217596

Zhang, J., Naik, H. S., Assefa, T., Sarkar, S., Reddy, R. V. C., Singh, A., et al. (2017). Computer vision and machine learning for robust phenotyping in genome-wide studies. *Sci. Rep.* 7:44048. doi: 10.1038/srep44048

Zhang, J., Song, Q., Cregan, P. B., Nelson, R. L., Wang, X., Wu, J., et al. (2015). Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (*Glycine max*) germplasm. *BMC Genomics* 16:217. doi: 10.1186/s12864-015-1441-4

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., and Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28, 3326–3328. doi: 10.1093/bioinformatics/bts606

Zheng, Z., Hey, S., Jubery, T., Liu, H., Yang, Y., Coffey, L., et al. (2020). Shared Genetic Control of Root System Architecture between Zea mays and Sorghum bicolor. *Plant Physiol.* 182, 977–991. doi: 10.1104/pp.19.00752

Zhiming, Y., Bo, K., Xiaowei, H., Shaolei, L., Youhuang, B., Wona, D., et al. (2011). Root hair-specific expansins modulate root hair elongation in rice. *Plant J.* 66, 725–734. doi: 10.1111/j.1365-313X.2011.04533.x

Zhu, J., Kaeppler, S. M., and Lynch, J. P. (2005). Topsoil foraging and phosphorus acquisition efficiency in maize (*Zea mays*). *Funct. Plant Biol.* 32, 749–762. doi: 10.1071/FP05005

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Plant Science

METHODS
published: 29 April 2022
doi: 10.3389/fpls.2022.800161

NeuralLasso: Neural Networks Meet Lasso in Genomic Prediction

Boby Mathew[1,2], *Andreas Hauptmann*[3,4], *Jens Léon*[2] *and Mikko J. Sillanpää*[3]*

[1] Bayer CropScience, Monheim am Rhein, Germany, [2] Institute of Crop Science and Resource Conservation, University of Bonn, Bonn, Germany, [3] Research Unit of Mathematical Sciences, University of Oulu, Oulu, Finland, [4] Department of Computer Science, University College London, London, United Kingdom

Prediction of complex traits based on genome-wide marker information is of central importance for both animal and plant breeding. Numerous models have been proposed for the prediction of complex traits and still considerable effort has been given to improve the prediction accuracy of these models, because various genetics factors like additive, dominance and epistasis effects can influence of the prediction accuracy of such models. Recently machine learning (ML) methods have been widely applied for prediction in both animal and plant breeding programs. In this study, we propose a new algorithm for genomic prediction which is based on neural networks, but incorporates classical elements of LASSO. Our new method is able to account for the local epistasis (higher order interaction between the neighboring markers) in the prediction. We compare the prediction accuracy of our new method with the most commonly used prediction methods, such as BayesA, BayesB, Bayesian Lasso (BL), genomic BLUP and Elastic Net (EN) using the heterogenous stock mouse and rice field data sets.

Keywords: neural networks, LASSO, local epistasis, genomic selection, whole genome prediction

OPEN ACCESS

Edited by:
Rodomiro Ortiz, Swedish University of Agricultural Sciences, Sweden

Reviewed by:
Hao Cheng, University of California, Davis, United States
Osval Antonio Montesinos-López, Universidad de Colima, Mexico

***Correspondence:**
Mikko J. Sillanpää
mikko.sillanpaa@oulu.fi

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *22 October 2021*
Accepted: *18 March 2022*
Published: *29 April 2022*

Citation:
Mathew B, Hauptmann A, Léon J and Sillanpää MJ (2022) NeuralLasso: Neural Networks Meet Lasso in Genomic Prediction. Front. Plant Sci. 13:800161. doi: 10.3389/fpls.2022.800161

INTRODUCTION

The introduction of Genomic Selection (GS) (Meuwissen et al., 2001) along with the availability of low cost genotyping platforms has resulted in a major paradigm shift in both animal and plant breeding. Since then, GS has been successfully applied for efficient selection and accelerating the breeding process in various breeding programs (Spindel et al., 2015; Garner et al., 2016; Hickey et al., 2017; Voss-Fels et al., 2019). Even though GS has now been widely implemented in practice, still considerable effort has been given to improve the prediction accuracy in GS beyond the current limits. Various factors can affect the prediction accuracy in GS including marker density, heritability of the trait, population size, constitution of the learning population and the statistical model used to predict the genomic breeding values (Meuwissen, 2009; Liu et al., 2018; Norman et al., 2018). Recently many studies tried to incorporated the transcriptome data (Li et al., 2019; Azodi et al., 2020) into genomic prediction models, in order to improve the prediction accuracy in GS.

The genomic prediction models can be divided roughly into two classes: (1) genomic best linear unbiased prediction (GBLUP) based on linear mixed models and (2) the whole-genome regression (WGR) based on multilocus regression models. In the first approach, the genetic background of the trait is assumed to be polygenic while in the latter, more oligogenic genetic background is assumed. Again in the first, molecular markers are used to construct the genomic relationship matrix while in the latter, molecular markers represent considered set of regression variables in the model. However, note that WGR model can be written also as the GBLUP model with a

trait-specific relationship matrix having own variance component for each SNP in the diagonal (Zhang et al., 2010; Piepho et al., 2012; Resende et al., 2012; Shen et al., 2013).

Epistasis (genetic interaction) is one of the major reason for the non-linearity in the genotype-phenotype relationship and considerable efforts have been given to model epistasis in genomic prediction models (Hu et al., 2011; Wittenburg et al., 2011; Wang et al., 2012; Jiang and Reif, 2015). Recently, many studies even pointed out the importance of local epistasis (interactions that span short segments of the genome) (Wei et al., 2014; Akdemir and Jannink, 2015; Akdemir et al., 2017; He et al., 2017; Liang et al., 2020). Although it is well known that epistasis (both local and global) interactions contribute to many complex traits (Taylor and Ehrenreich, 2014, 2015; Albert and Kruglyak, 2015), most of the genomic prediction models account for the pair-wise interactions due to the computational complexity of screening through all possible combinations.

Most of the WGR models used in GS are based on linear regression procedure and have been successfully adopted to predict complex phenotype in plant and animal breeding programs (Meuwissen et al., 2001; Park and Casella, 2008; Mathew et al., 2019). Nonlinear extensions of these methods with dominance and epistasis has been also considered (Nishio and Satoh, 2014; Jiang and Reif, 2015; Varona et al., 2018; Olatoye et al., 2019). However, recent development in the field of machine learning enable us to use profound nonlinear methods for the prediction of complex traits in breeding. Among the machine learning methods, deep learning (DL) methods received much attention due to their outstanding prediction properties (LeCun et al., 2015). Although improved accuracy can be questioned, many recent studies successfully applied deep learning for various genomic problems (Uppu et al., 2016; Bellot et al., 2018; Montesinos-López et al., 2018, 2019; Crossa et al., 2019; Liu et al., 2019; Pérez-Enciso and Zingaretti, 2019).

Often these learning methods are applied in a black-box manner and standard architectures that worked well in disciplines like natural language processing and computer vision are transferred to genomic prediction. Even though results are encouraging, interpretability remains an issue (Waldmann, 2018). However, as an exception, there is a study presenting an interpretable neural network model (see Zhao et al., 2021). Also, in this study we propose to design a domain specific learning system that is motivated by neural networks, but incorporates classical elements of lasso. The resulting algorithm is termed NeuralLasso, that is capable of incorporating higher order nonlinear interactions between contributing markers in the local neighborhood. Unlike the method of Zhao et al. (2021), our non-Bayesian approach is focusing on modeling high-order local interactions. In the terminology of neural networks, predictions are performed in a single layer and ℓ_1 sparsity on the learned parameters is incorporated, hence the relation to classical lasso models. We compare the prediction accuracy of NeuralLasso with the most commonly used GP methods such as BayesA, BayesB, BL, GBLUP, and EN using the mouse and rice data sets.

MODELS AND METHODS

Whole Genome Regression Model

Let us consider a standard genomic prediction model

$$y = X\beta + Zw + \epsilon. \tag{1}$$

Here, y is a vector of observed phenotypes for n lines, β contains the fixed effects, X represents the incidence matrix for the fixed effects, $Z = Z_{i1}, Z_{i2}, ...Z_{ip}$ is the $n \times p$ (p is the number of markers) matrix for the genotypes coded as 0,1,2, $w = (w_1, w_2....w_p)$ is a column vector of marker effects and ϵ corresponds to the residual, following a normal distribution as $\epsilon \sim N(0, I\sigma_e^2)$. For simplicity, here we assume no fixed effects other than overall mean (note that it is possible to pre-correct fixed effects away from the phenotype before neural network analysis).

The number of markers usually exceeds the number of observations in genomic prediction problems and regularization is applied in order to obtain solution to Equation (1). A regularized regression function can be formulated as

$$\hat{\beta}, \hat{w} = \underset{\beta, w}{\operatorname{argmin}} \left[\sum_{i=1}^{n} (y_i - X\beta - \sum_{j=1}^{p} Z_{ij} w_j)^2 + P(\lambda, w) \right]. \tag{2}$$

Here, the function $P(\lambda, w)$ is the penalty function with regularization parameter $\lambda \geq 0$. Least absolute shrinkage and selection operator (lasso) (Tibshirani, 1996) based on the penalty term called ℓ_1-norm, which is the sum of the absolute coefficients and Ridge Regression based on the ℓ_2-norm penalty which the sum of squared coefficients are the most commonly used regularized regression methods. The EN method which is a compromise between lasso and ridge regression penalties can be represented as:

$$\hat{\beta}, \hat{w} = \underset{\beta, w}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - X\beta - \sum_{j=1}^{p} Z_{ij} w_j)^2 + \lambda \sum_{j=1}^{p} \left[\frac{1}{2}(1-\alpha) w_j^2 + \alpha |w_j| \right] \right\} \tag{3}$$

where $0 \leq \alpha \leq 1$ is the penalty weight. The EN penalty is controlled by α and when $\alpha = 1$ EN is identical to lasso, whereas EN is equivalent to ridge regression when $\alpha = 0$.

NeuralLasso

We design our model based on the underlying Equation (1). That is, given the genotypes in Z, instead of finding the matrix w we seek to find a parametrizable (nonlinear) mapping Λ_θ, with parameters θ, such that

$$\Lambda_\theta(Z) = y. \tag{4}$$

The question is how to construct such a mapping Λ_θ and what do the parameters of θ represent. In the following we aim to derive a model, that is motivated by neural networks, but follows the classical architecture of lasso as presented in Equation (3). For this purpose, we will shortly review classic neural network architectures.

Background on Neural Networks

The underlying premise of a neural network is to combine affine linear mappings and pointwise nonlinearities to construct a nonlinear mapping in a repeating multi-layered fashion (LeCun et al., 2015; Schmidhuber, 2015; Goodfellow et al., 2016). In its most general form we can write the main building block of a neural network for an input $z \in \mathbb{R}^n$ (genotypes) and output $y \in \mathbb{R}^m$ (phenotype) as

$$y = \varphi(Cz + b), \tag{5}$$

where $C \in \mathbb{R}^{m \times n}$ is a linear transformation matrix, $b \in \mathbb{R}^m$ an additive affine component, the so-called bias, and finally $\varphi(\cdot)$ a point-wise acting nonlinear function. A popular choice for this nonlinear function is given by the rectified linear unit $\mathrm{ReLU}(x) := \max(x, 0)$. A multi-layered neural network would be now given as a repeated composition the blocks in Equation (5), where each block is called one layer. Nevertheless, in this work we concentrate on so-called shallow networks that consist only of one layer. The specific network architecture is now defined by the structure of the affine linear transformation C. The obvious choice of a dense matrix C is called a fully connected layer, as each data point in the input vector is related with each point in the output vector. Such a fully connected layer learns specific weights for each location in the input and hence is locally varying. Another option for the choice of linear mapping would be given by convolutions, if represented as matrices this would result in a sparse representation. Sparse representations are desirable, as they can be implemented efficiently and reduce the amount of parameters significantly. Nevertheless, the choice of convolutions as linear transformation is not optimal in our setting, as these are translationally invariant and hence do not encode any locality. In the following, we aim to design a transformation that is sparse, but does also encode locality to combine the strength of both.

Formulating NeuralLasso

The first important part is to define the underlying transformation given as the matrix C for our proposed model is based on the requirement to encode locality, while taking neighborhood relationships into account. For this purpose, we follow (Arridge and Hauptmann, 2019) and define a sparse subdiagonal matrix $C \in \mathbb{R}^{p \times p}$, where p is the number of markers, and a neighborhood of size N, such that the main diagonal and the N subdiagonals below and above are non-zero. That is for $N = 0$ we simply have a diagonal matrix and for $N = 1$ we have a tridiagonal matrix such as

$$C = \begin{pmatrix} c_{0,1} & c_{1,1} & & & & \\ c_{-1,2} & c_{0,2} & c_{1,2} & & & \\ & \ddots & \ddots & \ddots & & \\ & & c_{-1,n-1} & c_{0,n-1} & c_{1,n-1} & \\ & & & c_{-1,n} & c_{0,n} & \end{pmatrix}. \tag{6}$$

Given the matrix C we could formulate a lasso problem that takes interactions in the local neighborhood into account by minimizing

$$\hat{C} = \arg\min_C \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} (CZ_i^T)_j \right)^2 + \lambda \sum_{j=-N}^{N} \sum_{i=1}^{p} |c_{j,i}|. \tag{7}$$

Note, that for $N = 0$ no neighborhood relation is taken into account and the model reduces to the basic lasso scheme similar to Equation (3). As the above model in Equation (7) only considers linear interactions in the local neighborhood, we want to combine this sparse subdiagonal matrix with classical elements of neural networks, *i.e.*, nonlinear activation functions and additional bias vectors to allow for nonlinear interactions, as outlined previously.

The Proposed Model for Local Epistatic Interactions

We will now consider the building block of a neural network as in Equation (5) for one layer, but consider multiplication with the subdiagonals of C separately to introduce nonlinear effects between neighboring loci. In the following, we will fix the neighborhood to $N = 2$, that is a neighborhood window of 5 loci. We will model the nonlinear interaction by a maximum thresholding using ReLU for the 3 central loci and no nonlinearity for the outer two loci. This way we enforce an interaction effect of the 5-neighborhood. Given the (sub)diagonal vectors $c_i \in \mathbb{R}^p$ for $i = -2, \ldots, 2$ the non-linear parametrized model can be formulated as

$$\Lambda_C(Z_j) = \sum_{n=-2}^{2} \sum_{i=1}^{p} \varphi_i(c_{n,i} z_{i+n} + b_{n,i}), \tag{8}$$

where $z_i = 0$ for $i < 1$ or $i > p$, and $\varphi_i(x) = \mathrm{ReLU}(x) = \max(x, 0)$ for $i = -1, 0, 1$ and $\varphi_i(x) = x$ otherwise. That is, if we write all terms down we get

$$\begin{aligned} \Lambda_C(Z_j) = \sum_{i=1}^{p} & (c_{-2,i} z_{i-2} + b_{-2,i}) + (c_{2,i} z_{i+2} + b_{2,i}) \\ & + \mathrm{ReLU}(c_{-1,i} z_{i-1} + b_{-1,i}) + \mathrm{ReLU}(c_{0,i} z_i + b_{0,i}) \\ & + \mathrm{ReLU}(c_{1,i} z_{i+1} + b_{1,i}). \end{aligned} \tag{9}$$

The resulting NeuralLasso then formulates as

$$\{\hat{C}, \hat{b}\} = \arg\min_{\{C,b\}} \sum_{j=1}^{n} \left[y_j - \Lambda_C(Z_j) \right]^2 + \lambda \sum_{n=-2}^{2} \sum_{i=1}^{p} \left(|c_{n,i}| + |b_{n,i}| \right). \tag{10}$$

The parameters $\hat{C}$ and $\hat{b}$ can then be found by any suitable optimisation algorithm. ReLU functions were chosen, here, because of their ability to keep some of the linearity and introducing nonlinearity only by thresholding. Note that if all the ReLU activations are changed to linear functions then the model reduces to a sparse perceptron with biases (*i.e.*, a single-layer neural network), which will be an overparametrized version of the lasso approach. We will shortly discuss our implementation in the next section.

For the final estimation, we are only left with estimating the penalty weight λ as in the classic lasso model. This can be

achieved in a similar manner as used by Waldmann et al. (2019), here, we use a slightly modified bisection method and a single set of training data *i.e.*, one realization of training and validation split. We then initialize a starting interval $[a, b]$ for λ, chosen based on prior knowledge for the range of λ. We then compute the correlation coefficient for $\lambda = a, b$, *i.e.*, the end points of the interval $[a, b]$, and the mid point $\lambda = a + b/2$. Then, we identify the value for λ with the largest correlation coefficient. If it is one of the end points, we shift the interval around the end point, which becomes the new mid point. If, otherwise, the mid point has the highest correlation value, we keep the mid point, but halve the interval size. We then repeat the process for the new subinterval and compute the correlation for either one new point, if shifted, or two, if halved.

We note that for simplicity we have made here certain fixed choices and formulated NeuralLasso only for univariate continuous outcomes using fixed neighborhood size of 5, with ReLU as activation function and one layer. However, note that these are not arbitrary choices. As was stated earlier, ReLU was employed for its beneficial property of including linear functions as special case, if appropriate biases are learned. Some choices were found based on experimenting (*e.g.*, neighborhood size of 5 provided good predictive performance) and some of the choices (use of ReLU and linear activation functions) are discussed more in the discussion section. We refer to the **Appendix A** for more a general formulation of NeuralLasso using variable neighborhood size and activation functions.

Example Analysis

In order to compare the prediction accuracy of different methods, we analyzed the rice field data which is publicly available at http://www.ricediversity.org/data/ and a heterogeneous stock mouse population [see Valdar et al. (2006) for more details]. We selected traits in these data sets, which cover many levels of heritabilities (ranging from 0.25 to 0.75) and arguably many different genetic architectures.

Rice field data: The rice data set consists of 413 diverse accessions of *O. sativa* collected from 82 different countries (Zhao et al., 2011). The accessions were genotyped with single nucleotide polymorphism (SNP) markers and 33,569 SNPs were available for the analysis after excluding markers with minor allele frequency (MAF)> 0.05, duplicated markers and missing values > 20%. In this study, we analyzed the traits flowering time (FT) (in three different locations) and amylose content (AMY). The trait FT was measured in three different locations, the first location (ARK) was in Stuttgart, Arkansas, USA, the second one in Aberdeen (ABR) and the third location was Faridpur (FAD), Bangladesh (see Zhao et al., 2011 for more details). Out of the 413 lines, phenotypic informations were available for 371 lines in all three environments with the trait FT and 393 lines for the trait AMY. Genetic architecture underlying the trait (some traits are affected by many genes and some are by only few number of genes) is often play an important role in the prediction accuracy of different statistical methods. Thus we decided to consider two traits (FT and AMY) with different genetic architecture in this study. The narrow-sense SNP-heritabilities (h^2) of the traits were 0.50, 0.70, 0.50, and 0.26 for the phenotypes AMY, ARK, ABR, and FAD, respectively. Here, h^2 were estimated as: $h^2 = \sigma_g^2/(\sigma_g^2 + \sigma_e^2)$, where σ_g^2 and σ_e^2 are the genomic and residual variances, respectively. The variance components were estimated using GBLUP method.

Heterogeneous stock mouse data: The mice data (see Valdar et al., 2006) consists of 1940 individuals with 10345 biallelic SNP markers after excluding markers with minor allele frequency (MAF) = 0.05 and missing values = 20%. In this study, we analyzed the trait "body weight," which was measured at the age of 6 weeks. The narrow-sense SNP-heritability (h^2) of the trait "body weight" was 0.58.

Results: To demonstrate the superiority of our new approach, we compared the prediction accuracy (Pearson correlation coefficient between the observed and predicted phenotypes) of NeuralLasso with the most commonly used GP methods using the rice data set. The GP methods we are considering here are the GBLUP (Meuwissen et al., 2001), least absolute shrinkage and selection operator (lasso) (Tibshirani, 1996) and elastic net (EN) (Hoerl and Kennard, 1970). Also, Bayesian WGR models we choosed to consider here are the BL (Park and Casella, 2008), BayesA and BayesB (Meuwissen et al., 2001). Predictive abilities of BayesA, BayesB and BL were estimated using the R-package BGLR (Pérez and de los Campos, 2014). Whereas the predictive abilities of GBLUP and EN were estimated using the R-packages rrBLUP (Endelman, 2011) and glmnet (Simon et al., 2011), respectively. To estimate the predictive accuracy of NeuralLasso, the model was implemented in Python with TensorFlow and the scripts used in this study will be publicly available at: https://github.com/asHauptmann/NeuralLasso. Optimization was performed with the Adam algorithm and a cosine decay from 10^{-3} to 10^{-5} with 3,000 iterations, as batch size we used the full sample size.

In order to compare the prediction accuracies we used five-fold cross-validation (CV), for that we used 80% of the data as the training set and the remaining 20% as the validation set. To remove the influence of random partitions on the accuracy, we repeated the cross-validation procedure 50 times and took the mean value. Additionally, we also used the same training and validation sets with the different GP methods. In the analysis using BGLR, we used the default priors and considered 10,000 Markov Chain Monte Carlo iterations with a burn-in period of 3,000 iterations. For the EN estimation using glmnet, we set α to 0.33 in Equation (3) based on cross validation.

In **Table 1**, we can see the prediction accuracies of different methods in four traits (ARK, ABR, FAD, AMY) of rice data set, as well as trait body weight of mice data set. These traits together cover many levels of SNP-heritabilities. In traits ARK, ABR, and AMY of rice and trait body weight of mice, NeuralLasso seems to slightly outperform all the other methods, suggesting some role of local interactions in the genetic architecture of the trait. In trait FAD, the superior performance is much smaller and performance is practically the same with the GBLUP. This is likely due to much smaller SNP-heritability of the FAD than the other traits (see also **Figure 1** which shows ordering of the methods in their prediction

TABLE 1 | Mean prediction accuracy based on 50 CV replicates using different approaches for the traits with rice (ARK, ABR, FAD, AMY) and mice (WEIGHT) data sets are shown along with the corresponding heritability (h^2) estimate for the trait.

	GBLUP	BayesA	BayesB	BL	ElasticNet	NeuralLasso	h^2
Rice							
ARK	0.664	0.666 (+0.30)	0.662 (−0.30)	0.665 (+0.15)	0.613 (−7.68)	0.672 (+1.20)	0.70
ABR	0.568	0.579 (+1.93)	0.565 (−0.52)	0.562 (−1.05)	0.546 (−3.87)	0.589 (+3.69)	0.50
FAD	0.473	0.477 (+0.84)	0.477 (+0.84)	0.474 (+0.21)	0.416 (−12.05)	0.478 (+1.05)	0.26
AMY	0.447	0.45 (+0.67)	0.451 (+0.89)	0.442 (−1.11)	0.419 (−6.26)	0.463 (+3.58)	0.50
Mice							
WEIGHT	0.512	0.525 (+2.53)	0.521 (+1.75)	0.527 (+2.92)	0.503 (−1.75)	0.532 (+3.90)	0.58

Additionally, the percentage difference in prediction accuracy compared to the commonly used GBLUP estimation method is provided in the bracket.

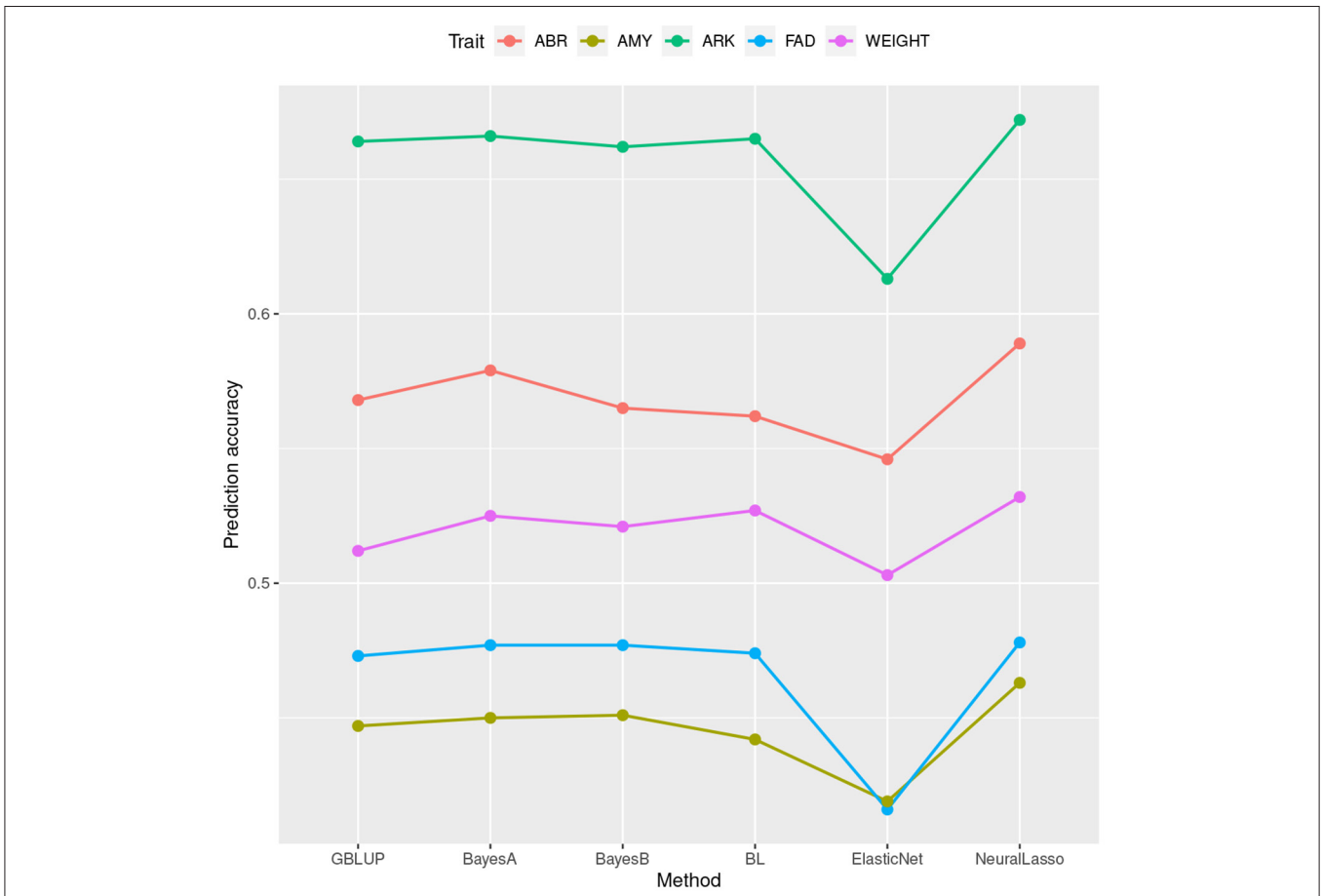

FIGURE 1 | Mean prediction accuracy calculated based on 50 cross validations for different traits from the rice and mice data sets plotted against the corresponding estimation methods.

accuracies). Superiority of NeuralLasso method becomes clear also from here.

DISCUSSION

In this study, we have presented a shallow neural network method which takes into account higher order local epistatic interactions in each marker's neighborhood. In recent years, machine learning methods including deep learning (DL) methods have been widely considered for GP, however neural network methods perform similarly or worst to the classical linear methods (Azodi et al., 2019; Zingaretti et al., 2020; Montesinos-López et al., 2021). In this study with the tested cases, our proposed method seems to improve the prediction accuracy slightly over traditional methods. We believe that the accuracy of NeuralLasso will depend on the complexity of the trait. Unlike the traditional genomic prediction models which are able to account for the two-loci genome-wide interactions, NeuralLasso account for only the additive and higher order local epistatic genetic effects. Thus there is a reduced chance that the local epistatic genetic effect will

disappear due to recombination and will be passed on to several generations (Akdemir and Jannink, 2015).

Even though, deep neural networks have been popular so far, size of learning data need to be large in many cases. We believe that more shallow networks like the one presented here may turn out to be useful and important in the future due to their more limited learning data size requirements.

In fact, our proposed model is not a single-layer neural network, *i.e.*, a perceptron. In the classic perceptron the nonlinearity is applied after summation, in our case the nonlinearity is applied before to allow for nonlinear interactions. On the other hand, one could say it is only one layer, but with several channels, for each member of the neighborhood, that are combined nonlinearly. In summary, this is why we say the model is motivated by neural networks, but does not clearly fit in the classic notion of a neural network. Finally, that is why we also do not describe our model as a neural network, but as NeuralLasso, motivated by the design of neural networks.

We also tested the performance of NeuralLasso when changing all non-linear ReLU functions to linear ones (results not shown). In those experiments, the prediction accuracies of NeuralLasso method clearly dropped down in the rice data set but stayed at about the same level in the mice data [when ReLU functions in Equation (9) were replaced by linear functions]. This is well in line with what one expects to see in rice data (high level of epistasis) and in mice data (small or no level of epistasis). Therefore, the latter experiment arguably means that our NeuralLasso may also be capable of taking into account some other context-specific effects than only epistasis, because its predictive performance was so high in mice data set.

In this study, we only considered small genomic region, however, NeuralLasso can be adjusted to account for higher order genetic interactions in larger genomic region of interest, chromosome-wise or whole genome scale. Although this might be computationally challenging, it will be interesting to see if this turns out to be important in the future. In order to reduce the computational burden, one can also first perform a genome-wide association study (GWAS) and only account the regions of interest (*e.g.,* candidate gene regions) in NeuralLasso.

As in all genomic predictions, not any single statistical method is clearly superior in their prediction accuracy for all traits, but their performance depends on factors such as genetic architecture and heritability of the trait. However, NeuralLasso performance was found here to be promising and it is worth of considering in the future.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: details are provided in the article.

AUTHOR CONTRIBUTIONS

BM, AH, JL, and MS: writing—review and editing and conceptualization. BM and AH: writing—original draft and formal analysis. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We are grateful for the editor and two reviewers for their comments which significantly helped us to improve presentation of this work. AH acknowledges funding from the Academy of Finland projects 338408 and 336796. The codes used in this study are made available at: https://github.com/asHauptmann/NeuralLasso.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.800161/full#supplementary-material

REFERENCES

Akdemir, D., Jannink, J.-L., and Isidro-Sánchez, J. (2017). Locally epistatic models for genome-wide prediction and association by importance sampling. *Genet. Select. Evol.* 49, 1–14. doi: 10.1186/s12711-017-0348-8

Akdemir, D., and Jannink, J. L. (2015). Locally epistatic genomic relationship matrices for genomic association and prediction. *Genetics* 199, 857–871. doi: 10.1534/genetics.114.173658

Albert, F. W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* 16, 197–212. doi: 10.1038/nrg3891

Arridge, S., and Hauptmann, A. (2019). Networks for nonlinear diffusion problems in imaging. *J. Math. Imag. Vis.* 62, 1–17. doi: 10.1007/s10851-019-00901-3

Azodi, C. B., Bolger, E., McCarren, A., Roantree, M., de Los Campos, G., and Shiu, S.-H. (2019). Benchmarking parametric and machine learning models for genomic prediction of complex traits. *G3 Gen. Gen. Genet.* 9, 3691–3702. doi: 10.1534/g3.119.400498

Azodi, C. B., Pardo, J., VanBuren, R., de los Campos, G., and Shiu, S.-H. (2020). Transcriptome-based prediction of complex traits in maize. *Plant Cell* 32, 139–151. doi: 10.1105/tpc.19.00332

Bellot, P., de los Campos, G., and Pérez-Enciso, M. (2018). Can deep learning improve genomic prediction of complex human traits? *Genetics* 210, 809–819. doi: 10.1534/genetics.118.301298

Crossa, J., Martini, J. W., Gianola, D., Pérez-Rodríguez, P., Jarquin, D., Juliana, P., et al. (2019). Deep kernel and deep learning for genome-based prediction of single traits in multienvironment breeding trials. *Front. Genet.* 10, 1168. doi: 10.3389/fgene.2019.01168

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Gen.* 4, 250–255. doi: 10.3835/plantgenome2011.08.0024

Garner, J., Douglas, M., Williams, S. O., Wales, W., Marett, L., Nguyen, T., et al. (2016). Genomic selection improves heat tolerance in dairy cattle. *Sci. Rep.* 6, 34114. doi: 10.1038/srep34114

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep Learning*. Cambridge, MA: MIT Press.

He, S., Reif, J. C., Korzun, V., Bothe, R., Ebmeyer, E., and Jiang, Y. (2017). Genome-wide mapping and prediction suggests presence of local epistasis in a vast elite winter wheat populations adapted to Central Europe. *Theor. Appl. Genet.* 130, 635–647. doi: 10.1007/s00122-016-2840-x

Hickey, J. M., Chiurugwi, T., Mackay, I., Powell, W., Eggen, A., Kilian, A., et al. (2017). Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nat. Genet.* 49, 1297. doi: 10.1038/ng.3920

Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12, 55–67.

Hu, Z., Li, Y., Song, X., Han, Y., Cai, X., Xu, S., et al. (2011). Genomic value prediction for quantitative traits under the epistatic model. *BMC Genet.* 12, 1–11. doi: 10.1186/1471-2156-12-15

Jiang, Y., and Reif, J. C. (2015). Modeling epistasis in genomic selection. *Genetics* 201, 759–768. doi: 10.1534/genetics.115.177907

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Li, Z., Gao, N., Simianer, H., and Martini, J. W. (2019). Integrating gene expression data into genomic prediction. *Front. Genet.* 10, 126. doi: 10.3389/fgene.2019.00126

Liang, Z., Tan, C., Prakapenka, D., Ma, L., and Da, Y. (2020). Haplotype analysis of genomic prediction using structural and functional genomic information for seven human phenotypes. *Front. Genet.* 11, 1461. doi: 10.3389/fgene.2020.588907

Liu, X., Wang, H., Wang, H., Guo, Z., Xu, X., Liu, J., et al. (2018). Factors affecting genomic selection revealed by empirical evidence in maize. *Crop J.* 6, 341–352. doi: 10.1016/j.cj.2018.03.005

Liu, Y., Wang, D., He, F., Wang, J., Joshi, T., and Xu, D. (2019). Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. *Front. Genet.* 10, 1091. doi: 10.3389/fgene.2019.01091

Mathew, B., Sillanpää, M. J., and Léon, J. (2019). "Advances in crop breeding techniques in cereal crops," in *Advances in Statistical Methods To Handle Large Data Sets for GWAS in Crop Breeding*, eds F. Ordon and W. Friedt (London: Burleigh Dodds Science Publishing), 437–450. doi: 10.19103/AS.2019.0051.20

Meuwissen, T. H. (2009). Accuracy of breeding values of'unrelated'individuals predicted by dense SNP genotyping. *Genet. Select. Evol.* 41, 35. doi: 10.1186/1297-9686-41-35

Meuwissen, T. H., Hayes, B., and Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. doi: 10.1093/genetics/157.4.1819

Montesinos-López, O. A., Martín-Vallejo, J., Crossa, J., Gianola, D., Hernández-Suárez, C. M., Montesinos-López, A., et al. (2019). New deep learning genomic-based prediction model for multiple traits with binary, ordinal, and continuous phenotypes. *G3 Gen. Gen. Genet.* 9, 1545–1556. doi: 10.1534/g3.119.300585

Montesinos-López, O. A., Montesinos-López, A., Crossa, J., Gianola, D., Hernández-Suárez, C. M., and Martín-Vallejo, J. (2018). Multi-trait, multi-environment deep learning modeling for genomic-enabled prediction of plant traits. *G3 Gen. Gen. Genet.* 8, 3829–3840. doi: 10.1534/g3.118.200728

Montesinos-López, O. A., Montesinos-López, A., Pérez-Rodríguez, P., Barrón-López, J. A., Martini, J. W., Fajardo-Flores, S. B., Gaytan-Lugo, L. S., Santana-Mancilla, P. C., and Crossa, J. (2021). A review of deep learning applications for genomic selection. *BMC Gen.* 22, 1–23. doi: 10.1186/s12864-020-07319-x

Nishio, M., and Satoh, M. (2014). Including Dominance Effects in the Genomic BLUP Method for Genomic Evaluation. *PloS ONE* 9, e85792. doi: 10.1371/journal.pone.0085792

Norman, A., Taylor, J., Edwards, J., and Kuchel, H. (2018). Optimising genomic selection in wheat: effect of marker density, population size and population structure on prediction accuracy. *G3 Gen. Gen. Genet.* 8, 2889–2899. doi: 10.1534/g3.118.200311

Olatoye, M. O., Hu, Z., and Aikpokpodion, P. O. (2019). Epistasis detection and modeling for genomic selection in cowpea (*Vigna unguiculata* L. Walp.). *Front. Genet.* 10, 677. doi: 10.3389/fgene.2019.00677

Park, T., and Casella, G. (2008). The Bayesian Lasso. *J. Am. Stat. Assoc.* 103, 681–686. doi: 10.1198/016214508000000337

Pérez, P., and de los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics* 198, 483–495. doi: 10.1534/genetics.114.164442

Pérez-Enciso, M., and Zingaretti, L. M. (2019). A guide on deep learning for complex trait genomic prediction. *Genes* 10, 553. doi: 10.3390/genes10070553

Piepho, H., Ogutu, J., Schulz-Streeck, T., Estaghvirou, B., Gordillo, A., and Technow, F. (2012). Efficient computation of ridge-regression best linear unbiased prediction in genomic selection in plant breeding. *Crop Sci.* 52, 1093–1104. doi: 10.2135/cropsci2011.11.0592

Resende M. Jr., Muñoz, P., Resende, M. D., Garrick, D. J., Fernando, R. L., Davis, J. M., et al. (2012). Accuracy of genomic selection methods in a standard data set of loblolly pine (pinus taeda l.). *Genetics* 190, 1503–1510. doi: 10.1534/genetics.111.137026

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Netw.* 61, 85–117. doi: 10.1016/j.neunet.2014.09.003

Shen, X., Alam, M., Fikse, F., and Rönnegård, L. (2013). A novel generalized ridge regression method for quantitative genetics. *Genetics* 193, 1255–1268. doi: 10.1534/genetics.112.146720

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* 39, 1–13. doi: 10.18637/jss.v039.i05

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redona, E., et al. (2015). Genomic selection and association mapping in rice ((*Oryza sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS Genet.* 11, e1004982. doi: 10.1371/journal.pgen.1004982

Taylor, M. B., and Ehrenreich, I. M. (2014). Genetic interactions involving five or more genes contribute to a complex trait in yeast. *PLoS Genet.* 10, e1004324. doi: 10.1371/journal.pgen.1004324

Taylor, M. B., and Ehrenreich, I. M. (2015). Higher-order genetic interactions and their contribution to complex traits. *Trends Genet.* 31, 34–40. doi: 10.1016/j.tig.2014.09.001

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Series B (Methodol.)* 58, 267–288.

Uppu, S., Krishna, A., and Gopalan, R. P. (2016). A deep learning approach to detect SNP interactions. *J. Softw.* 11, 965–975. doi: 10.17706/jsw.11.10.960-975

Valdar, W., Solberg, L. C., Gauguier, D., Burnett, S., Klenerman, P., Cookson, W. O., et al. (2006). Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat. Genet.* 38, 879–887. doi: 10.1038/ng1840

Varona, L., Legarra, A., Toro, M. A., and Vitezica, Z. G. (2018). Non-additive effects in genomic selection. *Front. Genet.* 9, 78. doi: 10.3389/fgene.2018.00078

Voss-Fels, K. P., Cooper, M., and Hayes, B. J. (2019). Accelerating crop genetic gains with genomic selection. *Theor. Appl. Genet.* 132, 669–686. doi: 10.1007/s00122-018-3270-8

Waldmann, P. (2018). Approximate Bayesian neural networks in genomic prediction. *Genet. Select. Evol.* 50, 70. doi: 10.1186/s12711-018-0439-1

Waldmann, P., Ferenčaković, M., Mészáros, G., Khayatzadeh, N., Curik, I., and Sölkner, J. (2019). AUTALASSO: an automatic adaptive LASSO for genome-wide prediction. *BMC Bioinformat.* 20, 1–10. doi: 10.1186/s12859-019-2743-3

Wang, D., El-Basyoni, I. S., Baenziger, P. S., Crossa, J., Eskridge, K., and Dweikat, I. (2012). Prediction of genetic values of quantitative traits with epistatic effects in plant breeding populations. *Heredity* 109, 313–319. doi: 10.1038/hdy.2012.44

Wei, W.-H., Hemani, G., and Haley, C. S. (2014). Detecting epistasis in human complex traits. *Nat. Rev. Genet.* 15, 722–733. doi: 10.1038/nrg3747

Wittenburg, D., Melzer, N., and Reinsch, N. (2011). Including non-additive genetic effects in Bayesian methods for the prediction of genetic values based on genome-wide markers. *BMC Genet.* 12, 1–14. doi: 10.1186/1471-2156-12-74

Zhang, Z., Liu, J., Ding, X., Bijma, P., de Koning, D.-J., and Zhang, Q. (2010). Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. *PloS ONE* 5, e12648. doi: 10.1371/journal.pone.0012648

Zhao, K., Tung, C.-W., Eizenga, G. C., Wright, M. H., Ali, M. L., Price, A. H., et al. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. *Nat. Commun.* 2, 467. doi: 10.1038/ncomms1467

Zhao, T., Fernando, R., and Cheng, H. (2021). Interpretable artificial neural networks incorporating Bayesian alphabet models for genome-wide prediction and association studies. *G3 Gen. Gen. Genet.* 11, jkab228. doi: 10.1093/g3journal/jkab228

Zingaretti, L. M., Gezan, S. A., Ferrão, L. F. V., Osorio, L. F., Monfort, A., Muñoz, P. R., et al. (2020). Exploring deep learning for complex trait genomic prediction in polyploid outcrossing species. *Front. Plant Sci.* 11, 25. doi: 10.3389/fpls.2020.00025

Conflict of Interest: BM was employed by company Bayer CropScience.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict ofinterest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in

this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership