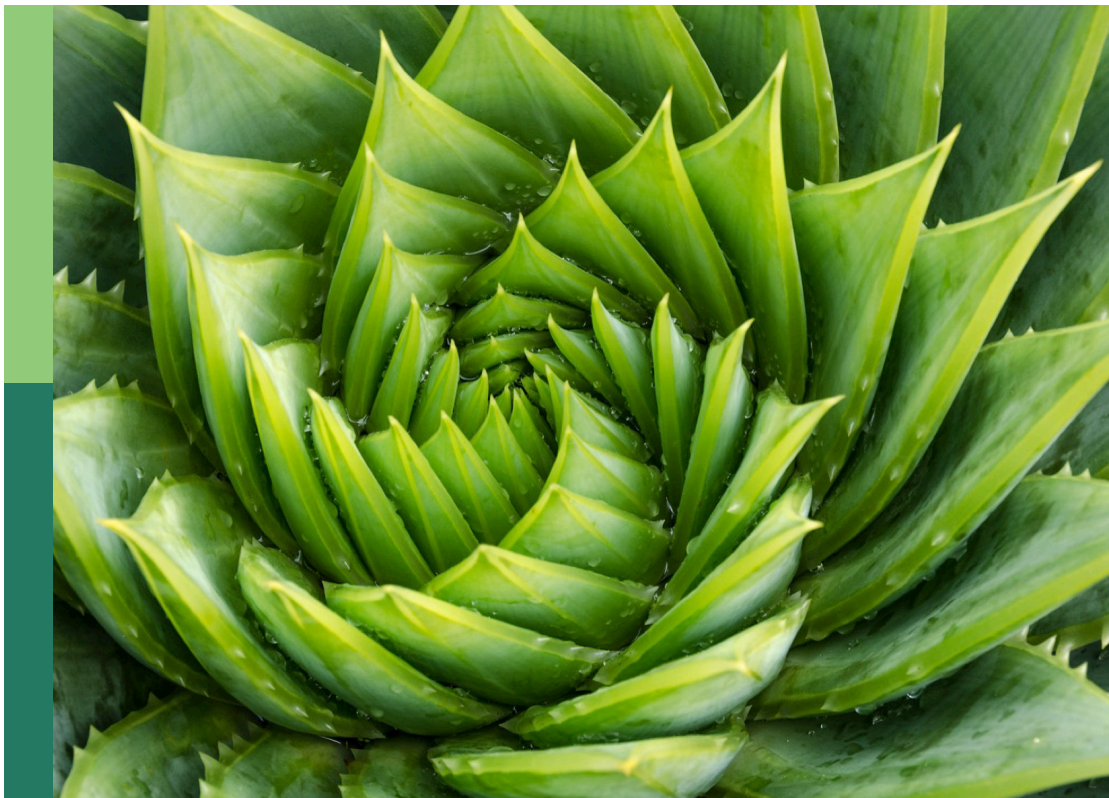
frontiers | Research Topics

Advances in statistical methods for the genetic dissection of complex traits in plants

Edited by
Yuan-Ming Zhang, Zhenyu Jia and Shang-Qian Xie

Coordinated by
Jia Wen, Ya-Wen Zhang and Shibo Wang

Published in
Frontiers in Plant Science



ISSN 1664-8714
ISBN 978-2-8325-4369-6
DOI 10.3389/978-2-8325-4369-6

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

Advances in statistical methods for the genetic dissection of complex traits in plants

Topic editors
Yuan-Ming Zhang — Huazhong Agricultural University, China
Zhenyu Jia — University of California, Riverside, United States
Shang-Qian Xie — University of Idaho, United States

Topic coordinators
Jia Wen — University of North Carolina at Chapel Hill, United States
Ya-Wen Zhang — Huazhong Agricultural University, China
Shibo Wang — University of California, Riverside, United States

Citation
Zhang, Y.-M., Jia, Z., Xie, S.-Q., Wen, J., Zhang, Y.-W., Wang, S., eds. (2024). *Advances in statistical methods for the genetic dissection of complex traits in plants*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4369-6

Table of contents

frontiers | Frontiers in Plant Science

TYPE Editorial
PUBLISHED 15 January 2024
DOI 10.3389/fpls.2024.1357564

OPEN ACCESS

EDITED AND REVIEWED BY
Roger Deal,
Emory University, United States

*CORRESPONDENCE
Yuan-Ming Zhang
soyzhang@mail.hzau.edu.cn

RECEIVED 18 December 2023
ACCEPTED 02 January 2024
PUBLISHED 15 January 2024

CITATION
Zhang Y-M, Jia Z, Xie S-Q, Wen J, Wang S and Zhang Y-W (2024) Editorial: Advances in statistical methods for the genetic dissection of complex traits in plants.
Front. Plant Sci. 15:1357564.
doi: 10.3389/fpls.2024.1357564

Editorial: Advances in statistical methods for the genetic dissection of complex traits in plants

Yuan-Ming Zhang[1]*, Zhenyu Jia[2], Shang-Qian Xie[3], Jia Wen[4], Shibo Wang[2] and Ya-Wen Zhang[5]

[1]College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, China, [2]Department of Botany and Plant Sciences, University of California, Riverside, Riverside, CA, United States, [3]Department of Animal, Veterinary & Food Sciences, University of Idaho, Moscow, ID, United States, [4]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States, [5]International Genome Center, Jiangsu University, Zhenjiang, China

KEYWORDS

multi-omics profiling, plants, 3VmrMLM, genome-wide association study, complex traits

Editorial on the Research Topic
Advances in statistical methods for the genetic dissection of complex traits in plants

1 Multi-locus genome-wide association study methods

In real data analysis, most commonly used genome-wide association study (GWAS) methods often miss some important loci and trait heritability. To address these challenges, Li et al. (2022a) established an innovative method named 3VmrMLM based on a compressed variance component mixed model. In 3VmrMLM, all the effects in quantitative trait nucleotide (QTN), QTN-by-environment interaction (QEI), and QTN-by-QTN interaction (QQI) detection are compressed into an effect-related vector, while all polygenic backgrounds are compressed into a vector-related polygenic background. This method is especially well suited for species with a high proportion of heterozygous genotypes, such as human, forests, chrysanthemums, and grasslands.

Can 3VmrMLM replace existing methods? The answer is no, despite 3VmrMLM demonstrating superiority over existing methods. For the detection of loci dominated by additive effects, existing methods remain appropriate, as observed in rice, wheat, and soybean. Since GWAS is based on linkage disequilibrium from historical recombination, there is complementarity between methods (Zhang et al., 2019). However, existing methods face challenges in detecting dominant effects and small allele substitution effects (Zhang et al., 2023).

When analyzing real data, the inflation factor or quantile–quantile plot serves as a common metric to assess method performance. However, this is not crucial for our mrMLM and 3VmrMLM methods (Zhang et al., 2020; Li et al., 2022a), because their genome-wide scanning aims to select potentially associated markers rather than identify

significant loci. A method is considered effective when it mines some importantly known and candidate genes around these loci, supported by strong evidence, as seen in 3VmrMLM. These identified loci may be used for genomic selection (Su et al., 2024), while more associated known and candidate genes can be mined and highlighted in the Manhattan plot.

This Research Topic contains three articles focusing on methodological studies and comparisons. Yang et al. proposed the MTOTC method to transform hierarchical data of ordinal traits into continuous phenotypes, which were then analyzed by multi-locus methods. This showed that the combination of MTOTC with any multi-locus method detects more QTNs. To identify QQI via the IIIVmrMLM software (Li et al., 2022b), Han et al. performed Levene's test to obtain the top 5,000 loci for each trait, and these loci were used to detect QTNs and QQIs associated with 11 flowering time-related traits in 199 Arabidopsis accessions with 216,130 markers. Around 89 QTNs and 130 QQIs, 34 identified genes were reported in previous studies, while 20 candidate genes were predicted; in particular, *AT1G12990* and *AT1G09950* around QQIs may have an interaction effect on flowering time. In addition, He et al. measured five free amino acid levels in 448 rice accessions across two environments, used nine GWAS methods to perform association analysis between phenotypes and 4,325,832 SNPs, and identified 88 stable QTLs, demonstrating the advantages of 3VmrMLM, including the most common QTNs, the highest LOD score, and the highest proportion of all stable QTLs.

2 The applications of new multi-locus GWAS methodologies in the genetic dissection of complex traits

Yield is one of the paramount breeding objectives, with nine articles in the Research Topic focusing on identifying QTNs and/or QEIs for yield-related traits. Zhang et al. used 3VmrMLM to re-associate 44,000 SNPs with eight yield-related traits from 413 rice accessions across three environments. They identified 87 known genes around QTNs and QEIs, including *OsMADS5* and *FZP*. Differential expression, functional enrichment, and haplotype analysis revealed the association of *LOC_Os04g53210* and *LOC_Os07g42440* with yield, while *LOC_Os04g53210* around a QEI potentially influenced flowering time. Zhao et al. employed 3VmrMLM to perform association analysis between three measured grain size traits of 159 rice accessions in two environments and 2,017,495 SNPs, identifying 393 QTNs and 8 QEIs. They found 22 genes around QTNs and 2 genes around QEIs to be genuinely associated with these traits. Additionally, 14 candidate genes were significant in differential expression, GO annotation, and haplotype analysis. Moreover, in a joint analysis of main crop and ratoon rice, 4 known genes, 8 additional candidate genes, and 2 candidate gene-by-environment interactions (GEIs) were identified as responsible for grain size-related traits.

Shu et al. evaluated plant height (PH) and ear height (EH) in 203 maize inbred lines at five locations and used 3VmrMLM to perform association analysis between phenotypes and 73,174 SNPs. They detected 23 significant QEIs and 53 corn belt-specific QTNs for the two traits. Transcriptomic and haplotype analysis highlighted the EH-related QEI S10_135 and the PH-related QEI S4_4, as well as corn belt-specific QTNs (S10_4 and S7_1), showcasing the power of 3VmrMLM in QEI discovery. Sun et al. measured the tassel branch number (TBN) of 190 F_2 individuals and $F_{2:3}$ families, using four methods to associate the phenotypes with 4,136 SNPs. They identified 13 QTLs and 22 QTNs, including large-effect QTLs qTBN6.06-1 and qTBN6.06-2 on chromosome 6. RNA-seq analysis revealed 5 candidate genes associated with TBN. Wen et al. identified 76 QTNs and 73 QEIs for three yield-related traits in 300 tropical and subtropical maize lines with 332,641 SNPs under well-watered, drought, and heat-stress conditions. They reported 34 genes from previous studies, confirming genes associated with drought tolerance (*ereb53* and *thx12*) and heat stress (*hsftf27* and *myb60*). Differential expression, tissue-specific expression, and haplotype analysis confirmed 24 candidate genes, while three yield GEIs (*GRMZM2G064159*, *GRMZM2G146192*, and *GRMZM2G114789*) were predicted.

Feng et al. measured the boll weight (BW) of 290 cotton accessions in nine environments and used GEMMA to perform association analysis between the phenotypes and 25,169 SNPs and 2,315 InDels, identifying two major QTLs on chromosomes A08 and D13. *Ghir_A08G009110* and *Ghir_D13G023010* were confirmed by both transcript-level and differential expression analysis between high- and low-BW accessions during the ovule development stage. Liu et al. measured three seed size-related traits in 196 mung bean accessions across two environments and used four methods to perform association analysis between the phenotypes and 3,607,508 SNPs. *VrKIX8*, *VrPAT14*, *VrEmp24/25*, *VrIAR1*, *VrBEE3*, *VrSUC4*, and *Vrflo2* around QTNs were homologous to known seed development genes in rice and *Arabidopsis thaliana* and further verified by differential expression and RT-qPCR analysis. *VrFATB*, *VrGSO1*, *VrLACS2*, and *VrPAT14* around QEIs were homologous to known seed development genes in *A. thaliana*. Hong et al. measured two epicotyl length traits in 951 soybean accessions over two years and used 3VmrMLM to perform association analysis between phenotypes and 1,639,846 SNPs, identifying 180 QTNs and QEIs. Based on transcript abundance, GO enrichment, and haplotype analysis, 10 candidate genes were predicted to be involved in the process of seed germination and seedling development, and it was found that *Glyma.04G122400* and *Glyma.18G183600* may affect epicotyl length elongation. Han et al. measured the flowering time (FT) of 490 *Brassica napus* accessions in eight environments and used 3VmrMLM to perform association analysis between the phenotypes and 11,700,689 SNPs, identifying 19 stable QTNs and 32 QEIs for FT and 10 QTNs for FT-related climatic indices. A total of 12 and 14 differentially expressed genes were found to be candidate genes for stable QTNs and QEIs, respectively, while five DEGs were found to be candidate genes for the indices. *BnaFLCs*, *BnaFTs*, *BnaA02.VIN3*, and *BnaC09.PRR7* were further validated.

With the improvement in people's living standards, crop quality traits are becoming increasingly important. Yu et al. measured four seed tocopherol content traits of 175 soybean accessions in three environments, used six methods to perform association analysis between the phenotypes and 23,149 SNPs, identifying 101 QTNs in single-environment analysis and 57 QTNs and 13 QEIs in multi-environment analysis. A total of 11 candidate genes residing in eight novel QTLs were confirmed using haplotype, RNA-Seq, and RT-qPCR analysis. Zheng et al. evaluated three cooking quality traits in 345 rice accessions over two years and used seven multi-locus methods to perform association analysis between phenotypes and 193,582 SNPs, identifying 144 QTNs and 21 QEIs. Based on analyses of mutation type, gene ontology classification, haplotype, and expression pattern, *OsSSIIIb*, *OsMT2b*, *wx*, *OsSSIIa*, and *OsSSIIIa*, which are related to starch synthesis and endosperm development, were found to influence grain expansion after cooking. Azam et al. measured the seed isoflavone accumulation of 1551 soybean accessions in five environments, used cMLM to perform association analysis between the phenotypes and 6,149,599 SNPs, and revealed that the allelic variation of *Glyma.11G108100* significantly influenced isoflavone accumulation.

Resistance, a key trait affecting crop yield, is the focus of two articles in this Research Topic. Kou et al. measured the pre-harvest sprouting of 629 Chinese wheat varieties in two environments, and they used the mrMLM and IIIVmrMLM software to perform association analysis between the phenotypes and 314,548 SNPs, identifying 22 stable QTNs for PHS resistance, such as AX-95124645 ($r^2 \geq 36\%$). Importantly, all white-grained varieties with the QSS.TAF9-3DTT haplotype showed resistance to spike sprouting. Around this locus, *TraesCS3D01G466100* and *TraesCS3D01G468500* were differentially expressed and found by GO annotation to be related to pre-harvest sprouting resistance. He et al. evaluated Pasmo resistance in 445 flax accessions over 5 years and used four methods to perform association analysis between phenotypes and 246,035 SNPs, identifying 132 tag QTNs and 50 QEIs. A total of 37 and 9 resistance gene analogs were considered potential candidates for QTNs and QEIs, respectively.

In addition, Wu et al. evaluated eight traits of 226 sunflower inbred lines under control and drought stress conditions and used three methods to perform association analysis between these phenotypes and 94,162 SNPs. Among the 118 genes around 80 QTNs, 14 candidate genes were validated by RNA-seq and RT-qPCR analysis, and *LOC110885273*, *LOC110872899*, *LOC110891369*, and *LOC110920644* were found to be abscisic acid-related protein kinases and transcription factors.

3 Future perspectives

To effectively identify QEIs across diverse environments and QQIs across numerous markers, it is imperative to devise new algorithms tailored to sample size, computational speed, and minimal memory requirements to meet the needs of human large data analysis. As the field advances, the genetic model for quantitative traits may transition from the classic Fisher genetic model to a more comprehensive framework through the integration of artificial intelligence. We anticipate that our compressed variance component mixed model will emerge as a pivotal tool in the genetic analysis of complex traits and multi-omics data in the future.

Author contributions

Y-MZ: Writing – original draft, Writing – review & editing. ZJ: Writing – review & editing. S-QX: Writing – review & editing. JW: Writing – review & editing. SW: Writing – review & editing. Y-WZ: Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The work was supported by the National Natural Science Foundation of China (32070557; 32270673; 32060149).

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

References

Li, M., Zhang, Y. W., Xiang, Y., Liu, M. H., and Zhang, Y. M. (2022b). IIIVmrMLM: The R and C++ tools associated with 3VmrMLM, a comprehensive GWAS method for dissecting quantitative traits. *Mol. Plant* 15 (8), 1251–1253. doi: 10.1016/j.molp.2022.06.002

Li, M., Zhang, Y. W., Zhang, Z. C., Xiang, Y., Liu, M. H., Zhou, Y. H., et al. (2022a). A compressed variance component mixed model for detecting QTNs and QTN-by-environment and QTN-by-QTN interactions in genome-wide association studies. *Mol. Plant* 15, 630–650. doi: 10.1016/j.molp.2022.02.012

Su, J. S., Lu, Z. W., Zeng, J. W., Zhang, X. F., Yang, X. W., Wang, S. Y., et al. (2024). Multi–locus genome–wide association study and genomic prediction for flowering time in chrysanthemum. *Planta* 259, 13. doi: 10.1007/s00425-023-04297-8

Zhang, Y. M., Jia, Z., and Dunwell, J. M. (2019). Editorial: The applications of new multi-locus GWAS methodologies in the genetic dissection of complex traits. *Front. Plant Sci.* 10, 100. doi: 10.3389/fpls.2019.00100

Zhang, Y. M., Jia, Z., and Dunwell, J. M. (2023). Editorial: The applications of new multilocus GWAS methodologies in the genetic dissection of complex traits, volume II. *Front. Plant Sci.* 14, 1340767. doi: 10.3389/fpls.2023.1340767

Zhang, Y. W., Tamba, C. L., Wen, Y. J., Li, P., Ren, W. L., Ni, Y. L., et al. (2020). mrMLM v4.0.2: An R platform for multi-locus genome-wide association studies. *Genom. Proteom. Bioinf.* 18, 481–487. doi: 10.1016/j.gpb.2020.06.006

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 03 May 2022
doi: 10.3389/fpls.2022.847435

Combined GWAS and Transcriptome Analyses Provide New Insights Into the Response Mechanisms of Sunflower Against Drought Stress

Yang Wu[1], *Huimin Shi*[1], *Haifeng Yu*[2], *Yu Ma*[2], *Haibo Hu*[1], *Zhigang Han*[2], *Yonghu Zhang*[2], *Zilong Zhen*[1], *Liuxi Yi*[1]* *and Jianhua Hou*[1]*

[1] College of Agricultural, Inner Mongolia Agricultural University, Hohhot, China, [2] Institute of Crop Breeding and Cultivation, Inner Mongolia Academy of Agricultural and Husbandry Sciences, Hohhot, China

OPEN ACCESS

Edited by:
Dirk Walther,
Max Planck Institute of Molecular Plant Physiology, Germany

Reviewed by:
Samar Gamal Thabet,
Fayoum University, Egypt
Frank Maulana,
Louisiana State University Agricultural Center, United States
Zohreh Hajibarat,
Shahid Beheshti University, Iran

***Correspondence:**
Jianhua Hou
houjh@imau.edu.cn
Liuxi Yi
yiliuxivip@163.com

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

Received: 02 January 2022
Accepted: 31 March 2022
Published: 03 May 2022

Citation:
Wu Y, Shi H, Yu H, Ma Y, Hu H, Han Z, Zhang Y, Zhen Z, Yi L and Hou J (2022) Combined GWAS and Transcriptome Analyses Provide New Insights Into the Response Mechanisms of Sunflower Against Drought Stress. Front. Plant Sci. 13:847435. doi: 10.3389/fpls.2022.847435

Sunflower is one of the most important oil crops in the world, and drought stress can severely limit its production and quality. To understand the underlying mechanism of drought tolerance, and identify candidate genes for drought tolerance breeding, we conducted a combined genome-wide association studies (GWAS) and RNA-seq analysis. A total of 226 sunflower inbred lines were collected from different regions of China and other countries. Eight phenotypic traits were evaluated under control and drought stress conditions. Genotyping was performed using a Specific-Locus Amplified Fragment Sequencing (SLAF-seq) approach. A total of 934.08 M paired-end reads were generated, with an average Q30 of 91.97%. Based on the 243,291 polymorphic SLAF tags, a total of 94,162 high-quality SNPs were identified. Subsequent analysis of linkage disequilibrium (LD) and population structure in the 226 accessions was carried out based on the 94,162 high-quality SNPs. The average LD decay across the genome was 20 kb. Admixture analysis indicated that the entire population most likely originated from 11 ancestors. GWAS was performed using three methods (MLM, FarmCPU, and BLINK) simultaneously. A total of 80 SNPs showed significant associations with the 8 traits ($p < 1.062 \times 10^{-6}$). Next, a total of 118 candidate genes were found. To obtain more reliable candidate genes, RNA-seq analysis was subsequently performed. An inbred line with the highest drought tolerance was selected according to phenotypic traits. RNA was extracted from leaves at 0, 7, and 14 days of drought treatment. A total of 18,922 differentially expressed genes were obtained. Gene ontology and Kyoto Encyclopedia of Genes and Genomes analysis showed up-regulated genes were mainly enriched in the branched-chain amino acid catabolic process, while the down-regulated genes were mainly enriched in the photosynthesis-related process. Six DEGs were randomly selected from all DEGs for validation; these genes showed similar patterns in RNA-seq and RT-qPCR analysis, with a correlation coefficient of 0.8167. Through the integration of the genome-wide association study and the RNA-sequencing, 14 candidate genes were identified. Four of them (LOC110885273, LOC110872899, LOC110891369, LOC110920644) were abscisic acid related protein kinases and transcription factors.

These genes may play an important role in sunflower drought response and will be used for further study. Our findings provide new insights into the response mechanisms of sunflowers against drought stress and contribute to further genetic breeding.

Keywords: sunflower, drought stress, genome-wide association studies (GWAS), RNA-seq, single-nucleotide polymorphisms (SNPs), specific-locus amplified fragment sequencing (SLAF-seq)

INTRODUCTION

Sunflower (*Helianthus annuus. L*) belongs to the Compositae family (Schilling and Heiser, 1981), and is native to North America (Schilling and Heiser, 1981). As one of the major oilseed crops in the world, sunflower is considered an important source of high-quality oil and dietary fiber for human health (Khan et al., 2015). The world harvested area of sunflower seed has increased by 20% (from 23.07 million hectares to 27.87 million hectares), and the production has increased by more than 50% (from 31.45 million tons to 50.23 million tons) from 2010 to 2020 (FAO, 2021). China is the sixth-largest sunflower-producing country in the world. The main production areas of sunflowers in China are in the northwest region, such as Inner Mongolia Autonomous Region and Xinjiang Uygur autonomous region. The sunflower is an important economic source for local farmers, and the status of sunflower production directly affects farmers' living standards.

The global average temperature has risen by about 0.85°C from the year 1880 to 2012 (Adopted, 2014), resulting in a series of extreme weather events, such as heavy rains, flooding, drought, and desertification. Among them, drought is the most serious abiotic stress limiting global agricultural production (Wilhite and Buchanan-Smith, 2005). A persistent drought can cause a large number of deaths and force large-scale migration, while severe droughts can even impact human civilization (Ault, 2020). With the continued climate change and population growth, drought may pose a serious threat to global and regional food security in the coming decades (Riddell et al., 2018). Due to the strong root system, the sunflower was considered to be relatively tolerant to water stress. They are often seeded on beds and ridges with poor moisture conditions where many other crops are unable to survive (Hussain et al., 2018). As a result, it is more susceptible to drought stress leading to yield reduction (Pasda and Diepenbrock, 1990; Adeleke and Babalola, 2020; Grasso et al., 2020). Studies have shown that drought stress in sunflower seedlings can lead to severe yield loss (Mwale et al., 2003; Rauf and Ahmad Sadaqat, 2008).

The sunflower drought stress response behavior involves a series of changes in morphological, physiological, and molecular levels. The drought stress negatively influenced seed germination and seedling emergence at the germination stage (Kaya et al., 2006). Drought stress at the vegetative stage reduces plant height (PH), leaf surface area (LSA), and biomass production while causing pollen sterility at the reproductive stage (Turhan and Baser, 2004; Hussain et al., 2008). From a physiological perspective, drought affects the uptake of water and nutrition, leads to a reduction of relative water content (RWC), and the turgor of cells (Hussain et al., 2008, 2016; Ibrahim et al., 2016). Plants respond to drought stress by reducing water evaporation through stomatal closure. As a result, it also reduces the photosynthetic rate (Flexas et al., 2004). The decreased photosynthesis rate leads to a decrease in CO_2 fixation, which affects the regeneration of the final acceptor of the electron transport chain ($NADP^+$). The leaked electrons flow to O_2 to produce reactive oxygen species (ROS) (Flexas et al., 2004). ROS cause oxidation of membrane lipids, resulting in decreased cell membrane stability. The decrease in cell membrane permeability results in the accumulation of the relative electrical conductivity (REC) and malondialdehyde (MDA) (Gunes et al., 2008). From the molecular level, plants involve a series of pathways for signal perception, transduction, gene expression, and other stress metabolites to accommodate drought. Drought-induced genes can mainly be classified into two groups. The first group constitutes genes whose products directly function in tolerance to stress, such as LEA proteins, osmolytes, proline (Pro), CAT, POD. Another group includes genes playing a role in signal transduction as well as the regulation of gene expression including various transcription factors (TF), protein kinases (PK), and transcriptional regulators (TR) (Lata et al., 2015).

Some agronomic measures can mitigate the damage of drought impact on plants, such as exogenous applications of plant hormones, osmotic regulators, and mineral nutrients (Salami and Saadat, 2013; Rabert et al., 2014). However, these changes are not heritable, and need additional labor, capital, and technology investment. Coping with drought through the breeding approach is usually the most effective and economical strategy. The genetic modification within the plant is heritable. Once a gene is introduced into a breeding material, it will be a permanent source of drought tolerance (Rauf, 2008). Drought tolerance in plants is a complex quantitative trait involving many micro-effective genes (Blum, 2011). Molecular-based plant drought resistance breeding is a hot spot in recent years (Wang and Qin, 2017). Previous studies on the molecular mechanism of sunflower drought resistance were mostly based on linkage analysis (Kiani et al., 2007; Poormohammad Kiani et al., 2009; Haddadi et al., 2011). However, the linkage analysis population was on two parents with significantly different phenotypes and the recombinant inbred lines (RILs). Only genes in RILs that show a significant difference between parental lines could be detected.

Genome wide association study (GWAS) is an observational study to detect associations between genetic variants and traits in individuals (Togninalli et al., 2018). Compared to linkage analysis, GWAS uses a natural population, which eliminates the need to construct a population. Therefore, the time consumption is greatly reduced. The use of natural populations

allows GWAS to simultaneously detect many natural allelic variations (Ma et al., 2018). In addition, the natural population contains all the historical recombination information and thus provide relatively higher detection accuracy than bi-parental populations (Kofsky et al., 2020). GWAS has been widely used in plant drought research, such as wheat (*Triticum aestivum L.*), cotton (*Gossypium herbaceum L.*), rice (*Oryza sativa L.*), and potato (*Solanum tuberosum L.*) (Ma et al., 2016; Mwadzingeni et al., 2017; Hou et al., 2018; Tagliotti et al., 2021). RNA-sequencing (RNA-Seq) is another attractive omics tool to identify differentially expressed genes (DEGs) under different conditions. Further analysis can provide insight into the changes in the DEGs expression level, important biological processes, and pathways (Zhang et al., 2017). Combining GWAS with RNA-seq can decrease the higher false-positive rate (FDR) inherent in GWAS analysis, and improve the accuracy of gene selection (Xie et al., 2019; Wang et al., 2022). However, to our knowledge, there are no relevant studies on sunflowers.

Molecular marker-based genotyping is an important step in GWAS analysis. Most traditional molecular markers were based on sequence length polymorphism. However, it could not be used for large-scale genotyping due to low throughput (Sun et al., 2013b). Whole gene sequencing technology is restricted in its use for non-model organisms due to population size and price (Muir et al., 2016). One strategy to reduce the sequencing cost was to reduce representation libraries (RRL). Specific length amplified fragment sequencing (SLAF) is one of the representative techniques, which uses specific enzymes to digest the genomes, and select a given size range of restriction fragments based on personalized research purposes (Sun et al., 2013b). This approach maintains the marker density while reducing the volume of sequencing, lowering the cost.

In this study, we performed a GWAS analysis of 226 sunflower varieties based on SLAF-seq. Then, a drought-tolerant accession was selected for RNA-seq analysis. Several important candidate genes were obtained using a combined analysis. Our research objectives were to (1) investigate the phenotypic variations among accessions under different water conditions; (2) develop new drought-related SNPs and identify genetic variants; (3) understand gene expression patterns under different drought stress time points, and reveal important biological processes and pathways; (4) obtain important genes associated with drought tolerance.

MATERIALS AND METHODS

Plant Materials and Growth Condition

A total of 226 sunflower inbred lines were collected from different countries (Australia, U.S.A., and France) and different provinces in China (Inner Mongolia, Ningxia, Xinjiang, Liaoning, Jilin). Seventy-three of them were provided by the Inner Mongolia Academy of Agriculture and Animal Husbandry, and 153 were kept in our laboratory. The experiment was conducted in the summer of 2019 at the Inner Mongolia Agricultural University, China (111.71, 40.82, 1,000 m above sea level). Seeds with fully mature, healthy, and uniform sizes were sorted for drought-stress experiments. After sterilization with 0.2% (w/v) mercuric chloride ($HgCl_2$), all seeds were rinsed several times with distilled water and soaked in deionized water for 24 h. Then the seeds were sown in plastic flowerpots (25 × 19 × 16 cm) filled with 3 kg soil (sandy soil and organic humus in a ratio of 2:1). Each pot was planted with 10 seeds and each accession had 6 pots. To avoid interference from natural rainfall and other factors, all pots were placed in a greenhouse (light/dark cycles: 14 h/10 h; 28/22°C; 45 ± 5% relative humidity) without water and nutritional limitation.

Experimental Design and Drought Treatments

When seedlings grew to the stage of three leaves, six pots of each accession were randomly and equally divided into two groups. Each group contained three pots as three biological replicates. The different watering regime was imposed on these two groups. One group continued to irrigate with sufficient water, and maintain the soil moisture content of 30 ± 2% as a control group (WW). Another group kept the soil moisture content to 10 ± 2% as a treatment group (DS). The soil moisture content of each pot was determined at 9 a.m. every day using the weight method described by Soni and Abdin (2017) and supplemented with water according to the target soil moisture content.

Phenotypic Evaluation and Statistical Analysis

The experiment lasted for 15 days, then 5 plants were randomly selected from each pot for phenotypic evaluation. Plant height (PH) was measured directly with a ruler. Leaf surface area (LSA) was calculated by the leaf area co-efficient method (Alza and Fernandez-Martinez, 1997). Root shoot ratio (RSR) was measured by the gravimetric method. Total root length (RL), root volume (RV), and root surface area (RSA) were measured with an LA-S root scanner (Wanshen Testing Technology Co., Ltd., Hangzhou, China). The relative water content (RWC) was detected using the saturate water method by Galmes et al. (2011). The chlorophyll concentration was assessed using a SPAD chlorophyll meter (TYS-A, TOP Instrument Co., Ltd., Hangzhou, China).

Data were analyzed using SPSS software (SPSS for Windows, V20.0.0, SPSS, Chicago, Illinois). Normality distribution was preliminarily assessed by a one-sample Kolmogorov-Smirnov's goodness-to-fit test (K-S test). For statistical differences between WW and DS growth condition, the Student *t*-test (normal distribution) and Wilcoxon signed-rank test (non-normal distribution) was used. Spearman non-parametric correlations were used to determine the correlation coefficient and statistical significance. Corrplot and Pheatmap R package were used to visualize the correlation.

Genomic DNA Extraction and Restriction Enzyme Selection

Total genomic DNA was extracted from 100 mg of fresh leaves by the CTAB method with a plant genomic DNA kit DP305 (Tiangen Biotech, China). To ensure it met the requirements for SLAF-seq (concentration ≥ 20 ng/μl; volume ≥ 30/μl), the concentration and quality of DNA were determined

using a Nanodrop 2000 spectrophotometer (Thermo Scientific, Waltham, MA, USA).

The SLAF-seq technique requires breaking the genome into small fragments using restriction enzymes. Then selecting restriction fragments of a specific length range (defined as SLAF-seq) for sequencing. To evaluate the number of target fragments produced *via* different combinations of restriction enzymes, a *in silicon* pre-experiment for enzyme selection was conducted. The criteria for enzyme selection were as follows: (1) the proportion of restriction fragments located in repetitive sequences is as low as possible; (2) The restriction fragments are distributed evenly on the genome as far as possible; (3) Consistency between the length of restriction fragments and the specific experimental system (Davey et al., 2013); (4) The number of restriction fragments with lengths 364–464 pb (SLAF tags in sunflower) should exceed 300,000.

SLAF Library Construction and High Throughput Sequencing

The SLAF library construction and high-throughput sequencing were performed as described by Sun et al. (2013b). After a series of polymerase chain reactions (PCR), adapter ligation reactions, and agarose gel purification, the SLAF-tags were isolated and subjected to PCR amplification following the guide of Illumine sample preparation. The paired-end sequencing was performed on an Illumina HiSeq 2500 platform (Illumina Inc., San Diego, CA, USA) at Beijing Biomarker Technologies Corporation (Beijing, China). Sequencing quality was estimated by measuring the guanine-cytosine (GC) content and Q30 ratio. A *Q* value of 30 represents a 0.1% error probability and 99.9% confidence level. Reads with >90% identity were clustered into a single SLAF-tag using BLAT software, and SLAF-tags with a sequence that varied across samples were defined as polymorphic SLAF tags (Zhang et al., 2018). To test the accuracy of the restriction enzyme digestion protocol, we used the genome of *Oryza sativa ssp. japonica* as a control (374.30 Mb, http://rapdb.dna.affrc.go.jp/).

SNP Genotyping and Linkage Disequilibrium Analysis

All reads were processed for quality control and filtered using Seqtk (https://github.com/lh3/seqtk) software. High-quality paired-end reads were aligned to the reference genome (https://ftp.ncbi.nlm.nih.gov/genomes/all/annotation_releases/4232/100/GCF_002127325.1_HanXRQr1.0/) using Burrows-Wheeler Aligner (BWA) software (Li and Durbin, 2009). SNP calling was conducted using the HaplotypeCaller function of Genome Analysis Toolkit (GATK) (McKenna et al., 2010). The VCF files obtained by GATK were converted to PLINK files using VCFtools (v0.1.16) (Danecek et al., 2011). SNPs with an integrity ratio of <0.8 and MAF <0.05 were filtered out *via* PLINK software (v1.90b6.21) (Purcell et al., 2007). Linkage disequilibrium (LD) was estimated by measuring the squared allele frequency correlations (r^2) (VanLiere and Rosenberg, 2008) between pairs of SNPs *via* PLINK software, with $r^2 = 1$ indicating complete LD, and $r^2 = 0$ indicating absent LD. LD decay extent was defined as the physical genomic distance at which the r^2 decreased to half of its maximum value. PopLDdecay software (Zhang et al., 2019) was used to visualize the mean r^2 of all chromosomes within the 100 kb region.

Population Structure Analysis

Based on the filtered SNPs, population analysis, phylogeny analysis, and principal component analysis (PCA) were performed in turns. Admixture software v1.3.0 (Alexander et al., 2009) was used to analyze the population structure. The number of underlying population groups *K* was predefined as 1–13 using the maximum likelihood estimation approach. The cross-validation errors (CV) for each *K* value were calculated. The *K* value with the lowest CV error was selected as the optimal number of populations. The Pophelper R package was used to make multiline plots (Francis, 2017). The genetic distances were calculated using VCF2Dis-1.45 (https://github.com/BGI-shenzhen/VCF2Dis). The FastME (v 2.0) software (Lefort et al., 2015) was used to convert the mat file obtained in the previous step into a distance matrix file (*nwk). The phylogenetic trees were constructed using the neighbor-joining method in the iTOL server (https://itol.embl.de/) (Letunic and Bork, 2021). PCA was performed using PLINK software by the –pca function. The first three components were used to plot the PCA *via* the rgl (v. 0.107.14) R package (Adler et al., 2003).

Genomic-Wide Association Study

The GWAS analysis was conducted using three methods: mixed linear model (MLM), Fixed and random model Circulating Probability Unificatin (FarmCPU), and Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK) in GAPIT R package (Lipka et al., 2012). The phenotypic data of each accession was represented using two indices: stress tolerance index (STI) (Fernandez, 1992), and stress susceptibility index (SSI) (Fischer and Maurer, 1978). These were calculated as follows:

$$STI = \frac{Y_{si} \times Y_{pi}}{\overline{Y}_{pi}^{\,2}}$$

$$SSI = \frac{1 - \frac{Y_{si}}{Y_{pi}}}{1 - \frac{\overline{Y}_{si}}{\overline{Y}_{pi}}}$$

where Y_{si} = performance of a genotype under stress; Y_{pi} = performance of the same genotype under control conditions; $\overline{Y}_{\mathrm{si}}$ = mean $\mathrm{Y_{si}}$ of all genotypes, $\overline{Y}_{\mathrm{pi}}$ = mean $\mathrm{Y_{pi}}$ of all genotypes.

The first three principal components were used as covariates. The GAPIT uses genotype data to automatically generate kinship matrix and calculate population structure according to the needs of different methods. For the identification of true marker-trait association, the significant *p*-value was set as $p < 1.062 \times 10^{-6}$ ($p = 0.1/n$; n = total markers used, which is roughly a Bonferroni correction, corresponding to $-\log_{10}(p) = 5.97$, blue line in the Manhattan plots) (Zhou et al., 2017). The Manhattan plot was used to show the correlation between SNP and phenotypic traits.

The Quantile-quantile (Q-Q) plot was used to display the level of difference between observed and predicted values. Both the Manhattan plots and Q-Q plots were constructed using CMplot R package (Yin, 2018).

GWAS Candidate Gene Search and Combined Analysis

The region of GWAS candidate genes was defined by the average LD decay distance. Genes located within 20 kb flanking regions on either side of the significantly associated SNPs were considered as candidate genes. Function annotations were conducted using the Eggnog (Huerta-Cepas et al., 2019) and Pfam (Bateman et al., 2004) software. The blast software was used to search for *Arabidopsis thaliana* genes homologous to candidate genes in the TAIR database (https://www.arabidopsis.org). Transcription factors (TF), protein kinase (PK), and transcriptional regulators (TR) were identified using iTAK software (Zheng et al., 2016).

Material Screening and RNA-Sequencing

To reveal important biological processes and significant pathways involved in sunflower drought-response, and narrow down the candidate genes, RNA-seq was conducted. We screened the 226 GWAS accessions based on phenotypic evaluation results. A comprehensive drought tolerance coefficient value (*D*-value) was used to evaluate the drought tolerance of all accessions (Li et al., 2015). The *D*-value integrated the results of multi-traits measured under two watering regimes and can represent the comprehensive drought tolerance of an accession. Finally, an inbred line with the highest *D*-value was selected and named "K58" (Zilong et al., 2021).

The drought stress experiment was the same as GWAS. Young leaves were sampled at 0, 7, and 14 days after drought treatment. Total mRNA was isolated using the RNA prep pure plant kit DP411 (Tiangen Biotech, China) according to the instruction manual. A total of 1 μg RNA per sample was used for cDNA library construction. Sequencing libraries were generated using NEBNext UltraTM RNA Library Prep Kit for Illumina (NEB, USA) following the manufacturer's recommendations. The quality of libraries was assessed through the Agilent Bioanalyzer 2100 system. After the quality test, all samples were sequenced in the Illumina Novaseq 6000 system, and 150-bp paired-end sequences were obtained (raw reads). Clean reads were obtained by eliminating reads containing ploy-N, reads containing adapter and low-quality reads from raw reads. The Q30, GC content of clean reads were calculated simultaneously.

Analysis of Differentially Expressed Genes

Differentially expressed genes analysis was conducted using the HISAT2-Stringtie(merge)-DESeq2 pipeline. High-quality clean reads were aligned to the reference genome using the Hisat2 software (version 2.2.1) (Kim et al., 2015) with default parameters. In the gene count step, we used a "Transcript merge mode" *via* StringTie software (Pertea et al., 2015). Briefly, the alignment files (*.BAM) of each sample was converted to GTF file using StringTie software. Then all the GTF files were merged into one single file containing a non-redundant set of transcripts. This file was then used as a reference to recalculate the count for each gene. With this model, novel genes/transcripts can be identified that differ from the reference genome.

A python script [prepDE.py (https://ccb.jhu.edu/software/stringtie/dl/prepDE.py)] was used to generate a gene count matrix from the GTF file of each sample. Normalization and differential expression analysis were performed using DESeq2 R packages (Love et al., 2014). By default, DESeq2 computes a Benjamini-Hochberg adjusted *p*-value (P_{adj}) to control the false discovery rate (FDR) (Anders and Huber, 2012). The "Fold Changes" of a gene is the FPKM ratio at day 7 (or 14) to that at day 0. For comparison purposes, we take the logarithm of the fold change and calculate the absolute value ($|\log_2(\text{Fold Changes})|$). The $|\log2(\text{Fold Changes})|$ of a gene equal to 1 means that the expression level of this gene has doubled or halved. Genes with $P_{adj} \leq 0.01$ and $|\log_2(\text{Fold Changes})| \geq 1$ was considered as DEG.

Enrichment Analyses of Gene Ontology and KEGG Pathways

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses were performed to reveal the biological functions and pathways of DEGs. The sequence file of each gene was input into Eggnog software (version 2.0.1) to obtain gene annotation (Huerta-Cepas et al., 2019). GO and KEGG analysis was conducted using the ClusterProfiler (version 4.0.0) R package (Yu et al., 2012). Only GO-terms or KEGG pathways with p-value < 0.05 were screened for subsequent analysis. The REVIGO program (http://revigo.irb.hr/) was used to remove redundant GO-terms (Supek et al., 2011).

RT-qPCR Validation

To validate RNA-seq results, reverse transcription quantitative PCR (RT-qPCR) was conducted on 6 randomly selected DEGs with three technical replicates. Experimental samples are the same as for RNA-seq. Reverse transcription was conducted using Biomarker Script II 1st Strand cDNA Synthesis Kit (Biomarker Technologies, Beijing, China) with Oligo $d(T)_{23}$ VN as a primer, and qPCR reactions were performed with Biomarker 2X SYBR Green Fast qPCR Mix (Biomarker Technologies, Beijing, China) on the FTC-3000 qPCR system (Funglyn Biotech Inc., Toronto, ON, Canada). Gene expression levels were calculated using the method of $2^{-\Delta\Delta Ct}$ according to Livak and Schmittgen (Livak and Schmittgen, 2001), and standard deviation was calculated among three biological replicates. The 18S rRNA gene was used as the endogenous control (Ebrahimi Khaksefidi et al., 2015).

Combined Analysis of GWAS and RNA-Seq

To reduce the number of candidate genes, we conducted a combined analysis. The two gene sets obtained by GWAS and RNA-seq were subjected to the

intersection operation. Genes within the intersection were considered to be important genes and were investigated in depth.

RESULTS

Phenotypic Variation Among Accessions

Drought stress led to different degrees of changes in all phenotypic traits (**Figure 1**; **Table 1**). Drought stress inhibited plant height (PH). Mean PH was 31.37 cm (ranged from 15.07 to 56.10 cm) at WW condition, whereas it was 22.23 cm (ranged from 6.4 to 38.55 cm) under DS conditions. Over 90% of the accessions (208/226) had a decrease in PH under drought stress.

Mean leaf surface area (LSA) was 46.34 cm^3 (ranged from 4.62 to 143.62 cm^3) for the WW condition compared with 24.21 cm^3 (ranged from 3.73 to 65.36 cm^3) for the DS condition. Over 88% (200/226) of the accessions had a decrease in LSA under drought stress.

The root-shoot ratio (RSR) increased slightly under the DS condition compared with in WW condition. Mean RSR was 0.16 (ranged from 0.05 to 0.79) under DS condition, whereas it was 0.12 (ranged from 0.02 to 0.62) under WW condition, with 71.7% (162/226) of the accessions showing an increased RSR under DS conditions. Notably, drought stress significantly increased three root-related traits, the average root length (RL), root volume (RV), and root surface area (RSA) increased by 44.1, 131, and 76.4% under DS condition compared with plants under WW condition. Among the 226 accessions, 77.4% (175/226), 83.2% (188/226), 83.2% (188/226) of them showed an increased RL, RV, and RSA under drought conditions, respectively. Drought stress has relatively little effect on the relative water content (RWC) of sunflower leaves, and the mean value was reduced from 0.74% under WW condition to 0.69% under the DS condition, with a reduction rate of 7.3%. Among 226 sunflower plants, 83.6% (189/226) had lower RWC under the DS condition. Similarly, the SPAD value was also decreased slightly in DS compared to WW, with a reduction rate of 5.7%. Mean values were 31.08 (ranged from 22.1 to 39.77) and 29.31 (ranged from 18.6 to 38.67) under WW and DS, respectively, and 72.6% (164/226) accessions showed a decreased SPAD value under DS condition.

The coefficient of variation (CV) was used to describe the variance within accessions. In this study, the CV of some traits was very high, the average CV among all traits were 40.36%, varying from 11.94 to 71.86%. It shows that our experiment materials have strong heterogeneity. RSR had the highest CV values (61.42–65.49%) while the SPAD value showed the lowest CV values (11.94–14.09%) (**Supplementary Table 1**).

The correlation between the same indicator under different conditions is shown in **Supplementary Figure 1**. The correlation coefficients of LSA and SPAD were higher than 0.6 in the WW vs. DS, while the correlation coefficients of RSA, RL, and RSR were all lower than 0.1. The correlation between different indicators under the same condition is shown in **Figure 2**. The three root-related indexes (RL, RV, and RSA) showed positive correlation under both WW and DS growth conditions. Under DS conditions, RV was positively correlated with RSA (spearman Cor. = 0.776). whereas negatively correlated with PH (spearman Cor. = −0.59). Under WW conditions, LSA is positively correlated with SPAD with a spearman correlation coefficient of 0.61.

SLAF-Sequencing, Genotyping, and Linkage Disequilibrium

Enzyme digestion efficiency is an important indicator of SLAF-seq quality. According to the results of the pre-experiment, Hae III was selected to digest the genomic DNA. The enzyme digestion efficiency of control genome *Oryza sativa* ssp. *japonica* was 94.12%, indicating the enzyme digestion reaction was normal. A total of 934.08 MB paired-end reads were obtained, with an average Q30 of 91.97% (89.04–93.44%) and a GC content of 43.67% (42.13–45.56%) (**Supplementary Table 2**). The mapping rate and the proper mapped rate were 98.20 and 90.96%, respectively (**Supplementary Table 3**).

A total of 565,668 SLAF tags were obtained, 243,291 of them were polymorphic SLAF tags. These SLAF-tags were evenly distributed on 17 chromosomes (**Figure 3**; **Supplementary Table 4**). SLAF tags on chromosome 13 had the highest polymorphic rate (48.25%), while chromosome 12 had the lowest polymorphic rate (38.85%). A total of 2,124,143 population SNP markers were developed *via* GATK software (**Supplementary Table 5**; **Figure 4**). After quality control, 94,162 high-quality SNPs were obtained for subsequent analysis (**Supplementary Table 6**; **Figure 5**). Chromosome 10 harbored the highest proportion of SNPs (8.68%, 8,173 of 94,162), while the shortest chromosome 6 contained the lowest proportion of SNPs (3.08%, 2,898 of 94,162). There were 31.37 SNP per 1 MB on average across 17 chromosomes. Chromosome 10 had the highest SNPs/Mb ratio (47.68 SNPs per Mb), while chromosome 6 had the lowest SNPs/Mb ratio (19.56 SNPs per Mb) (**Supplementary Table 6**). LD was estimated as the r^2 value, r^2 ranged from 0.135 on chromosome 6 to 0.218 on chromosome 10, with an average of 0.174, revealing differences in the level of LD among chromosomes (**Supplementary Table 7**). The average distance of LD decay was about 20 kb (**Figure 6**).

Genetic Diversity and Population Structure

Divergence of the 226 accessions during evolution was the major factor leading to high rates of false positive errors in GWAS analysis (Yu and Buckler, 2006). The admixture software was used to analyze the population structure, and the CV for K = 1–13 was examined. The results showed that when K = 11, the CV dropped to the lowest value (0.659), suggesting the entire population most likely originated from 11 ancestors (**Figures 7**, **8A**). The phylogenetic tree has divided the accessions into 7 main clusters with identical tree topologies (**Figure 8B**). PCA analysis revealed that all the 11 principal components had eigenvalues of over 1, and the first 8 principal components can explain 85.73% of the total variance. The first three principal components PC1 (with variance explain 15.71%), PC2 (with variance explain 13.55%), and PC3 (with variance explain 11.77%) were displayed in **Figure 8C**. All these results showed that our experimental materials are highly heterogeneous and is ideal for GWAS analysis.

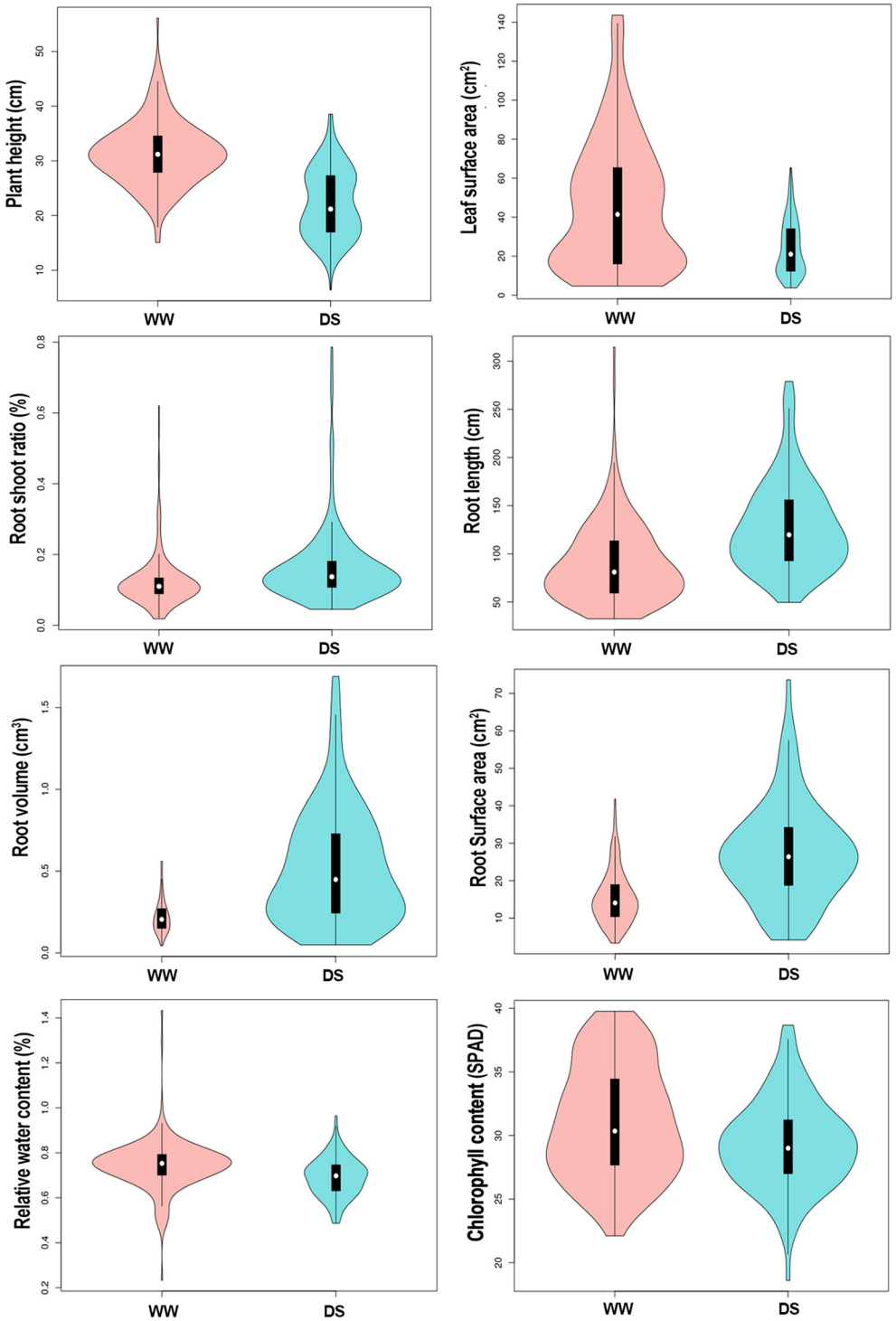


FIGURE 1 | Vioplot visualizing the 8 physiological traits of sunflower in response to different water treatments. *Y*-axis represent the density distribution of all 226 samples. WW, well-water growth condition; DS, drought-stress growth condition.

Genome-Wide Association Analysis

The GWAS was performed on 8 traits using 3 methods (MLM, FarmCPU, BLINK). A total of 80 SNPs were detected under the significance threshold of $p < 1.062 \times 10^{-6}$. Among them, 59 were obtained by STI, and 22 were obtained by SSI, and there was only one common SNP between the two indicators

TABLE 1 | Descriptive statistics values for traits of 226 sunflowers under drought stress.

Traits	Trt.	Min.	Max.	Mean	SD.	CV. (%)	Skewness	Kurtosis
Plant height	WW	15.07	56.10	31.37	5.80	18.48	0.50	1.56
	DS	6.40	38.55	22.23	6.16	27.72	0.25	−0.65
Leaf surface area	WW	4.62	143.62	46.34	33.30	71.85	0.91	0.30
	DS	3.73	65.36	24.21	14.03	57.93	0.71	−0.31
Root shoot ratio	WW	0.02	0.62	0.12	0.08	61.42	3.24	14.85
	DS	0.05	0.79	0.16	0.11	65.49	3.21	13.05
Root length	WW	32.56	314.68	89.47	40.83	45.63	1.63	5.11
	DS	49.66	279.06	128.90	46.76	36.28	0.96	0.86
Root volume	WW	0.04	0.56	0.22	0.11	46.99	1.21	1.53
	DS	0.05	1.69	0.52	0.35	67.17	1.01	0.88
Root surface area	WW	3.33	41.79	15.50	7.04	45.43	1.00	1.03
	DS	4.22	73.59	27.34	13.19	48.23	0.74	0.88
Relative water content	WW	0.23	1.43	0.74	0.11	14.59	0.84	11.26
	DS	0.49	0.96	0.69	0.09	12.54	0.10	0.19
SPAD	WW	22.10	39.77	31.08	4.38	14.09	0.19	−0.97
	DS	18.60	38.67	29.31	3.50	11.93	0.18	0.42

Trt., Treatment; Min., Minimum; Max., Maximum; SD, Standard deviation; CV, Coefficient of variance; CK, Well water condition; DS, Drought stress condition.

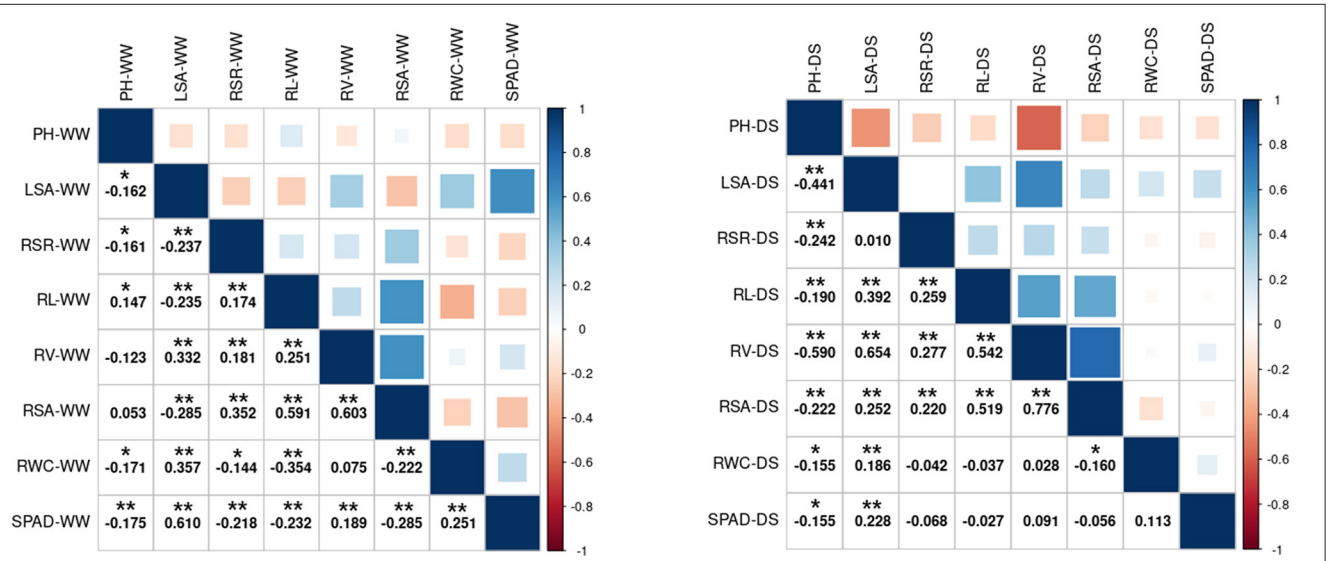


FIGURE 2 | Spearman's correlation analysis between the 8 drought-related traits under two water condition. Left: Under WW growth condition. Right: Under DS growth condition. WW, well-water growth condition; DS, drought-stress growth condition; PH, plant height; LSA, leaf surface area; RSR, Root shoot ratio; RL, Root length; RV, Root volume; RSA, Root surface area; RWC, Relative water content; SPAD, SPAD value. * and **Significant at the 0.05 and 0.01 probability levels between genotypes, respectively.

(**Supplementary Figures 2**, **3**; **Supplementary Table 8**). A total of 19, 44, and 33 SNPs were discovered by MLM, FarmCPU, and BLINK methods, respectively. For 8 phenotypic traits, LSA detected the most associated SNPs (27), followed by RWC detected 13, SPAD, RSR, and PH detected 12, 11, and 11, respectively. RL, RV, and RSA were detected 2, 4, and 3 SNPs, respectively. A total of 118 genes were found within the 20 kb of 80 significant SNPs, 85 of them were protein-coding genes (**Supplementary Table 9**).

RNA-Sequencing and Expression Analysis

A total of 70 Gb clean data were obtained after filtering and quality control. The Q30 of each library ranged from 93.57 to 94.97%, and the GC content ranged from 44.86 to 45.68% (**Supplementary Table 10**). A total of 18,922 DEGs were obtained (**Supplementary Table 11**), 6,698 of them were newly discovered. In general, there were more DEGs under 14 days of drought stress compared with the 7 days, and down-regulated DEGs were more than

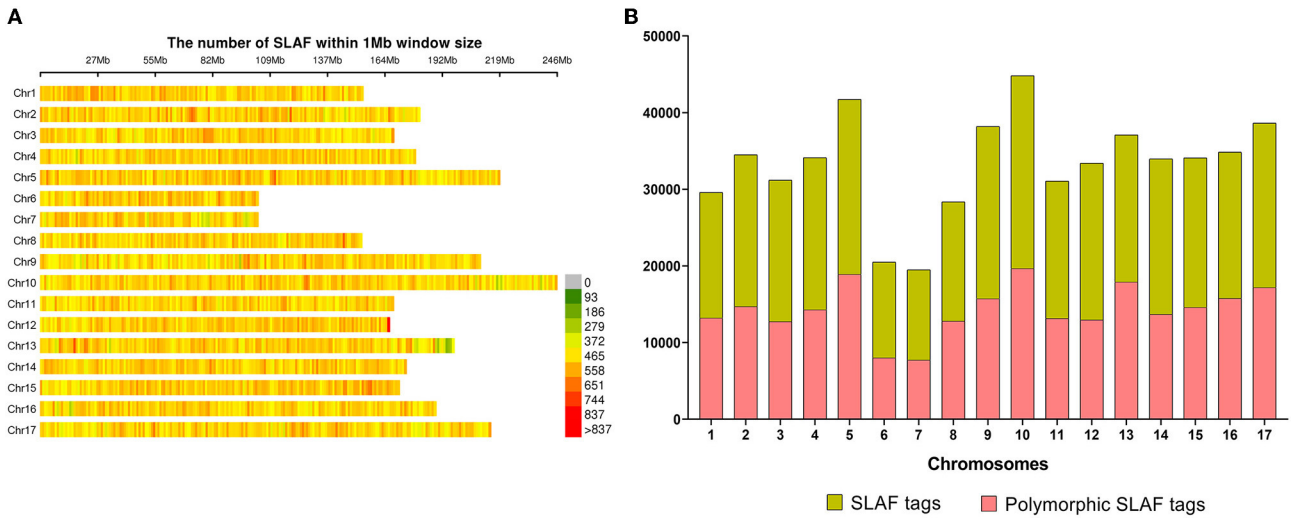


FIGURE 3 | Specific length amplified fragments (SLAF) tags distribution. **(A)** Distribution of all 565,668 SLAF tags on sunflower genome based on 226 accessions. The colors indicate the number of SLAF tags within a 1 Mb window. **(B)** The number of SLAF tags and polymorphic SLAF tags on each chromosome.

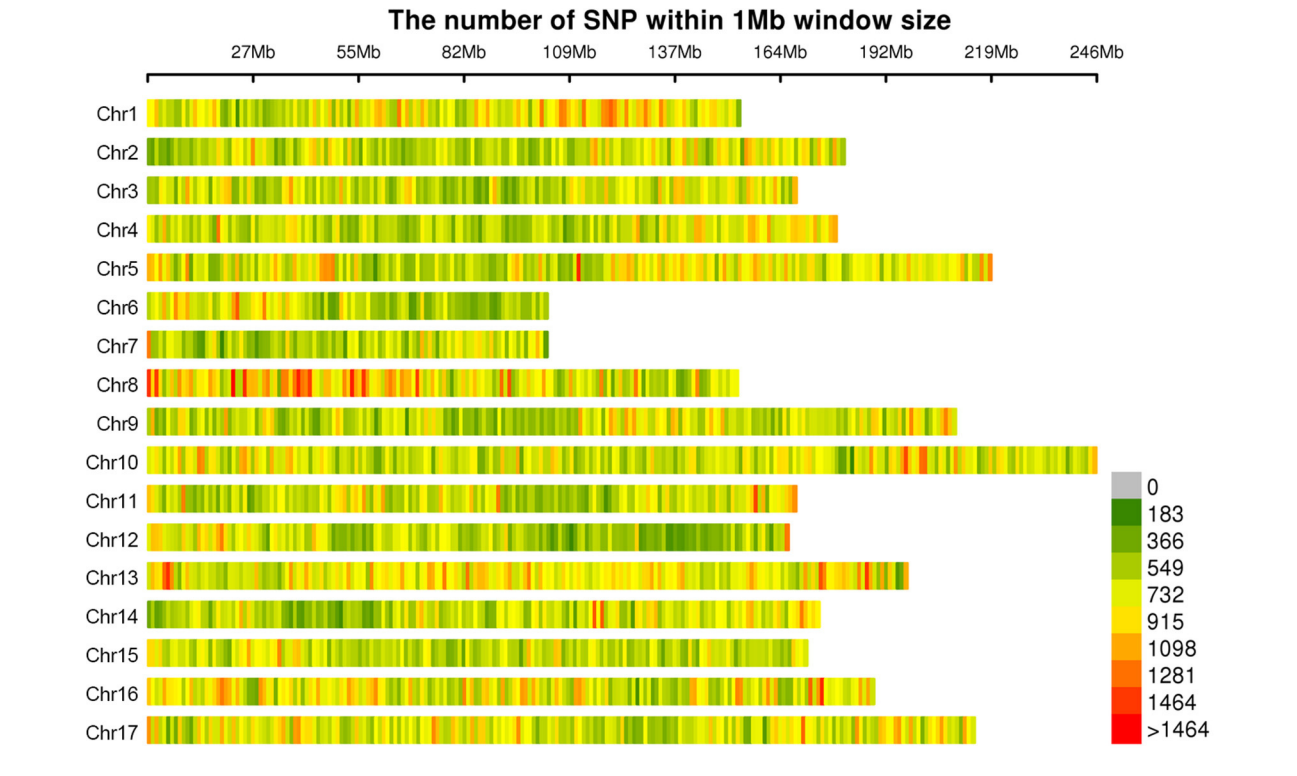


FIGURE 4 | Distribution of all 2,124,143 single nucleotide polymorphisms (SNPs) on sunflower genome. The colors indicate the number of SNPs within a 1 Mb window.

up-regulated DEGs (**Figure 9**). From day-7 to day-14, the up-regulated DEGs were increasing from 3,848 to 7,174, whereas the down-regulated DEGs were increasing from 5,201 to 8,521, respectively.

Enrichment Analysis

GO Analysis

The up-regulated genes were enriched in 46, 90 GO-terms at 7, 14 days. On day-7, the most significant GO-terms were

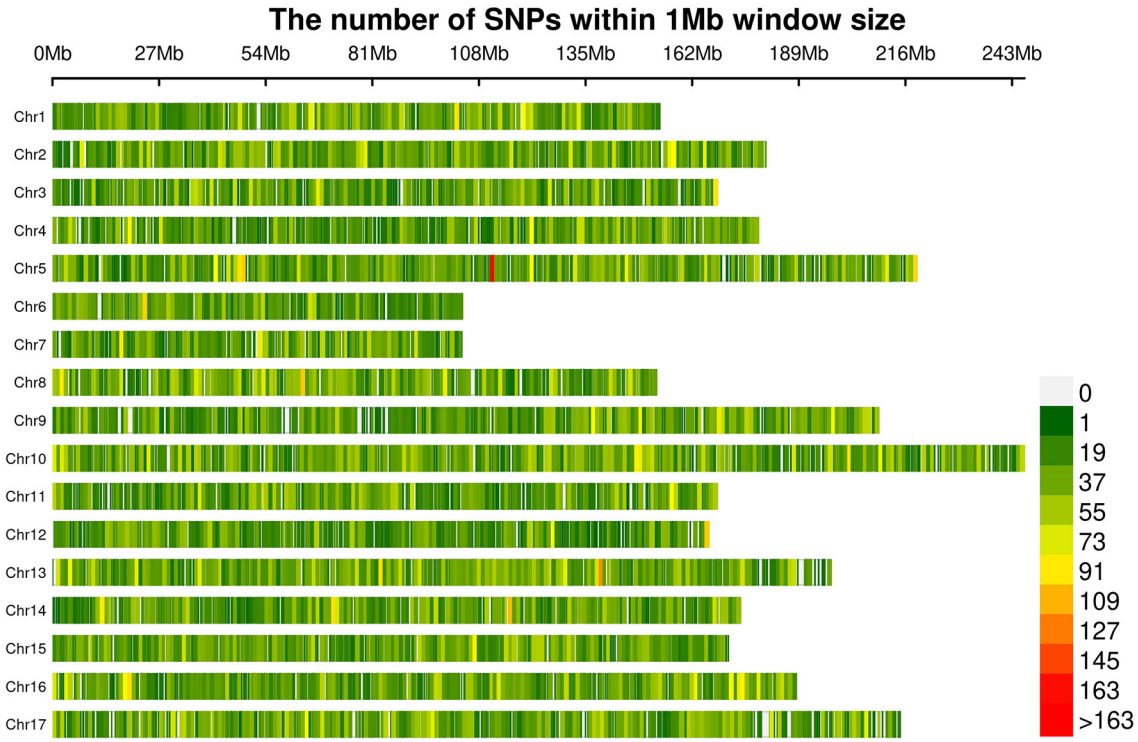


FIGURE 5 | Distribution of filtered SNPs among the 17 chromosomes. The colors indicate the number of SNPs within a 1 Mb window.

cellular amino acid catabolic process (GO:0009063), branched-chain amino acid catabolic process (GO:0009083), and seed maturation (GO:0010431). On day-14, the most significant GO-terms were leaf senescence (GO:0010150), aging (GO:0007568), and carboxylic acid catabolic process (GO:0046395). For down-regulated genes, there were 127, and 199 GO-terms enriched at 7, 14 days. On day-7, the most significant GO-terms were cellular polysaccharide metabolic process (GO:0044264), cell wall biogenesis (GO:0042546), and photosynthesis, light reaction (GO:0019684); At day-14, the most significant GO-terms were photosynthesis (GO:0015979), photosynthesis, light reaction (GO:0019684), and plastid organization (GO:0009657) (**Supplementary Figure 4**).

KEGG Analysis

Up-regulated genes were enriched in 13 and 48 significant KEGG pathways at 7 and 14 days. On day-7, the most significant pathways were Valine, leucine and isoleucine degradation, MAPK signaling pathway—plant, and FoxO signaling pathway; On day-14, the most significant pathways were valine, leucine, and isoleucine degradation, MAPK signaling pathway—plant, and longevity regulating pathway. For down-regulated genes, there were 36, 48 significant KEGG pathways enriched at 7, 14 days. On day-7 and day-14, the most significant KEGG pathways were both related to photosynthesis, such as photosynthesis proteins (BR:ko00194), photosynthesis-antenna proteins, and photosynthesis (**Supplementary Figure 5**).

RT-qPCR Validation

To validate the accuracy of RNA-seq, RT-qPCR was performed. Six genes were randomly selected from all DEGs. The primer sequence was shown in **Supplementary Table 12**. Correlation analysis showed that RNA-seq was closely related to RT-qPCR results. The correlation coefficient (R^2) was 0.8167, endorsing our RNA-seq data were reliable (**Figure 10**; **Supplementary Figure 6**).

Candidate Genes Identification

By integrating the results of GWAS and RNA-seq analysis, a total of 18 common genes were obtained, 14 of them were protein-coding genes (**Table 2**; **Figure 11**). These genes are distributed on chromosomes 4, 5, 8, 9, 10, 11, 12, 13, 16, and 17. Two genes are associated with both LSA and PH. One gene is associated with both LSA and SPAD. Their details are as follows.

Candidate Genes Associated With Plant Height

There were 2 candidate genes that were screened using combined analysis. Both of them were located on chromosome 13. The LOC110899235 gene encoding "inosine-uridine preferred nuclear hydrate" is homologous to the AT5G18860.2 gene in *Arabidopsis thaliana*. Another LOC110899238 gene encoding "ABC transporter c family member 3-like" is homologous to the AT3G13080.1 gene in *Arabidopsis thaliana*. Both two genes were down-regulated with the extension of drought stress time in K58.

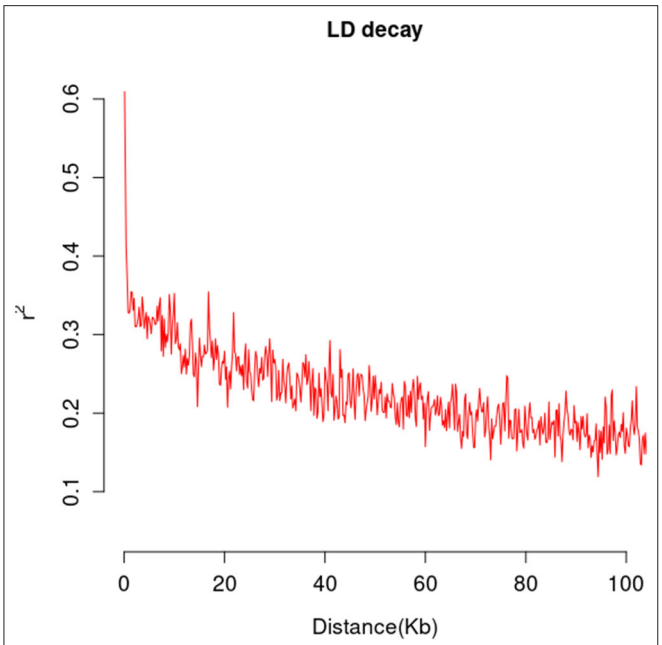


FIGURE 6 | The mean LD decay rate was estimated by the squared allele frequency correlations (r^2) using all pairs of SNPs located within 100 kb of physical distance in genomic regions in a population of 226 sunflower accessions.

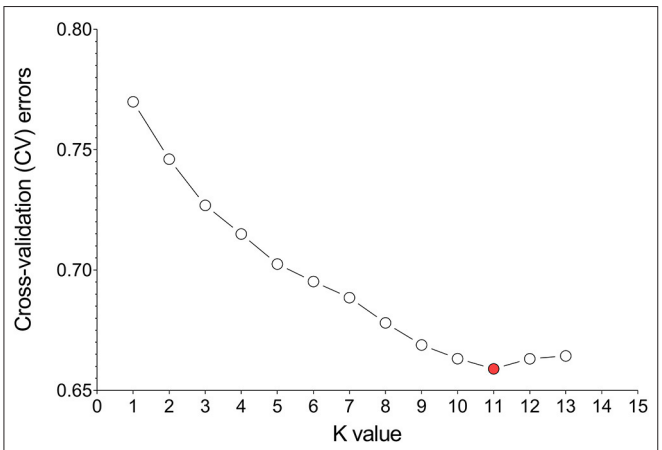


FIGURE 7 | Estimation of cross-validation (CV) errors for K values ranging from 2 to 13. The CV errors declined rapidly from K = 2 ∼ 11 and reached the lowest value at K = 11.

Candidate Genes Associated With Leaf Surface Area

There were 8 common candidate genes associated with LSA, 2 of which were also associated with PH. The function of the gene LOC10936334 located on chromosome 4 was annotated as "Jacalin-like lectin domain", which is homologous to the AT1G73040.1 gene in *Arabidopsis thaliana*, and its expression level continues to decrease under drought stress in K58. Gene LOC110941963 located on chromosome 5 was annotated as "microtubule-associated protein RP EB family member", which was homologous to the AT3G47690.1 gene in *Arabidopsis thaliana*. This gene was down-regulated after 14 days of drought stress in K58. At 19.52 kb upstream of an SNP (S10_123892851) on chromosome 10, a gene (LOC110885273) encoding "Serine threonine-protein kinase" was identified. It is worth noting that the gene was also associated with SPAD. This gene belongs to the protein kinase family of RLK-Pelle_SD-2b, and is homologous to the *Arabidopsis* AT4G32300.1 gene. RNA-seq showed it was down-regulated with the extension of drought stress in K58. Gene LOC110894816 encoding "Equilibrative nucleotide transporter" were down-regulated at 7, 14 days in K58, which is homologous to AT1G70330.1 in *Arabidopsis thaliana*. Gene LOC110920644 belongs to the PLATZ transcription factor family. It was up-regulated at 7 days and down-regulated at 14 days of drought stress in K58. Gene LOC110891369 encoding "receptor-like protein kinase" was sharply up-regulated at 14 days. This protein kinase belongs to the RLK-Pelle_SD-2b RLK-Pelle_CrRLK1L-1 protein kinase family.

Candidate Genes Associated With Root-Shoot Ratio

There were 2 candidate genes obtained by combined analysis. One gene LOC110937937 encoding "Component of the peroxisomal and mitochondrial division machineries" was up-regulated at 14 days post drought stress, another gene LOC110915715 encoding "Protein of unknown function (DUF1666)" were continuously down-regulated with the drought stress.

Candidate Genes Associated With Three Root Related Traits

Notably, there are relatively fewer SNPs related to three root traits (RL, RV, and RSA). No genes were found within the 20 kb region of RL associated SNPs. The combined analysis identified 2 genes associated with RV and 1 gene associated with RSA. For RV, gene LOC110877324 on chromosome 9 was annotated as "Belongs to the UDP-glycosyl transferase family", which was down-regulated in K58 after 14 days of drought stress. Another gene (LOC110917707) located on chromosome 16 was annotated as "domain presence in VPS-27, Hrs and Stam", which was up-regulated in K58 after 14 days of drought stress. These two genes are homologous to the AT2G18570.1 gene and AT2G38410.1 gene in *Arabidopsis thaliana,* respectively.

For RSA, gene LOC110872899 was located on chromosome 8, and annotated as "Inactive leucine-rich repeat receptor-like serine threonine-protein kinase". This gene is homologous to the *Arabidopsis* AT1G10850.1 gene. It was slightly up-regulated in K58 at 7 days and then sharply down-regulated at 14 days of drought stress.

Candidate Genes Associated With Relative Water Content

LOC110941862 is the unique gene screened by the combined analysis. This gene encodes the "Topless-related protein", which is homologous to the AT1G15750.3 gene in *Arabidopsis thaliana*. RNA-seq results showed that this gene was continuously down-regulated in K58 under drought stress.

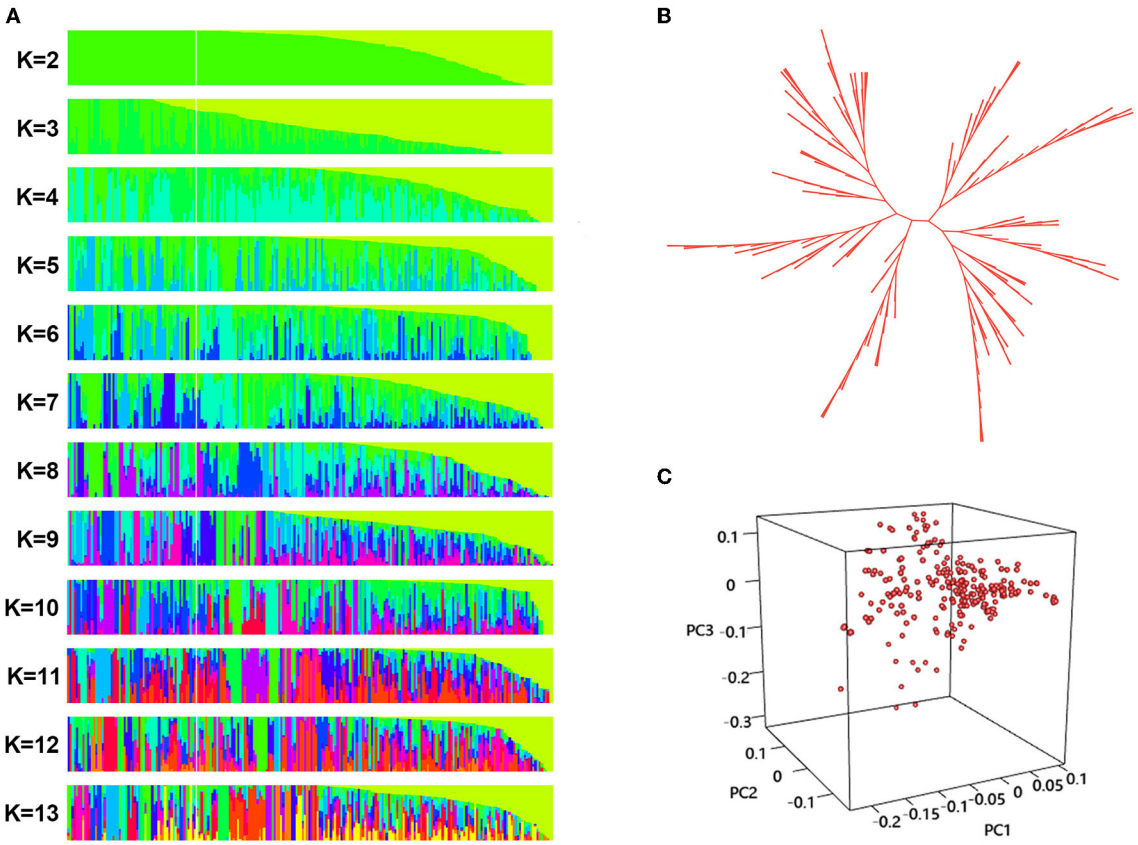


FIGURE 8 | Population structure analysis phylogenetic tree construction, and principal component analysis (PCA) of the 226 sunflower accessions. **(A)** Population structure of sunflower accessions estimated by ADMIXTURE, each row represents a given number of clusters (*K*, *K* = 2–13), each vertical column represents one individual and each colored segment in each column represents the percentage of the individual in the population. **(B)** The unrooted neighbor-joining three a of 226 sunflower accessions. Each branch indicates a sample, and the length of the branches represents the genetic distance, **(C)** PCA scatter plots shows the distribution of 226 sunflower accessions defined by the eigenvectors of the first three principal components (PC). The three axes represent PC1, PC2, and PC3 respectively. Each dot represents a sample.

DISCUSSION

Global climate change threatens crop production worldwide. Plants adopt diverse strategies to combat drought stress such as reducing the stomatal conductance, decreased photosynthetic rate, accumulation of different osmoprotectants, activation of stress-responsive genes and transcription factors, etc. (Farooq et al., 2009; Kaur and Asthir, 2017). Drought resistance is a complex quantitative trait. One difficulty in drought-tolerant genetic breeding is the unequivocal evaluation of plant response to soil-water deficits (Pereyra-Irujo et al., 2007). Based on the previous research, we evaluated 8 phenotypic traits among 226 accessions under WW and DS conditions. Compared to the WW condition, the average PH, LSA, RWC, and SPAD value were decreased, while RSR and three root related traits (RL, RV, RSA) were increased under the DS condition.

It has long been known that drought stress at the vegetative stage impedes phenotypic traits like PH, LSA, whereas an increase in RL at the expense of above-ground dry matter occurs resulting in higher RSR (Petcu et al., 2001; Hussain et al., 2010; Javaid et al., 2015). In our results, the change trends of mean PH, RL, RSR, and LSA were consistent with previous studies. However, the mean RV increased under drought, which was not consistent with a previous study. Geetha et al. (2012) found that the RV decreased by 40.2% under drought stress

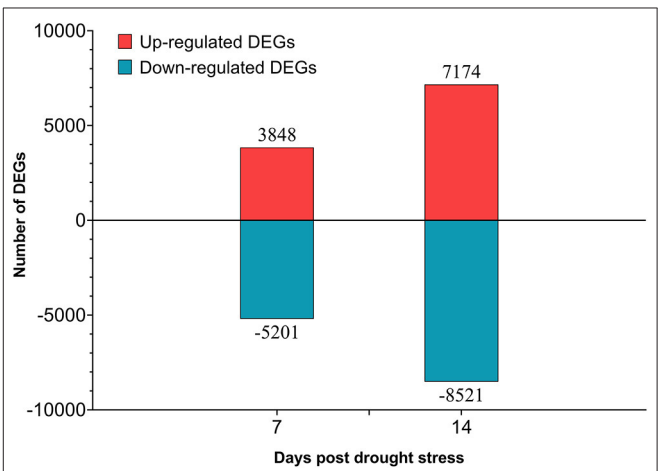


FIGURE 9 | Number of differentially expressed genes (DEGs) in different drought stress time.

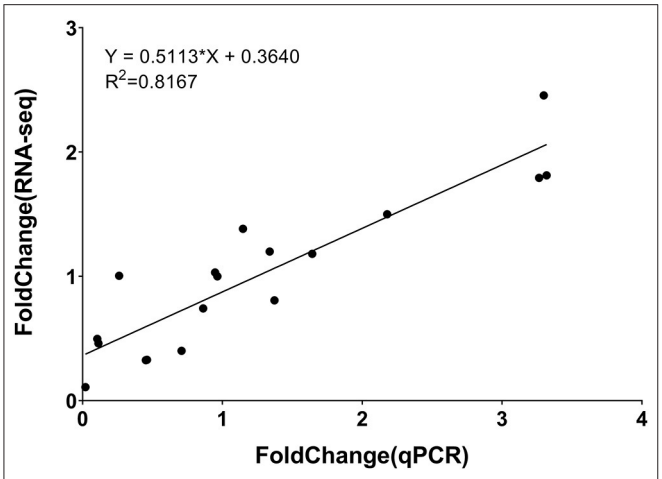


FIGURE 10 | Correlation between results of RT-qPCR and RNA-seq for select DEGs.

among 29 sunflower varieties, while we found 83% of accessions have an increase in RV. This may be due to differences in the genotypes of the study materials. Different genotypes of plants have different adaptability to drought stress (Petcu et al., 2001). Even in the most consistent trend of variation in PH (92% decreased under drought stress), there were still 16 accessions increased under drought stress. These specific materials may include important drought-tolerance genes and will be good sources for our drought tolerance molecular breeding. In some previous studies, the relationship between SPAD and chlorophyll content per unit leaf area is fitted as linear regression. SPAD value is often used to represent chlorophyll content (Costa et al., 2001; Martínez and Guiamet, 2004). Our results show that under WW growth conditions, SPAD value is positively correlated with LSA. It demonstrates that a larger LSA has more chlorophyll, which increases the photosynthetic rate (Espina et al., 2018). The correlation coefficients of LSA and SPAD in WW vs. DS conditions were higher than 0.6, indicating that drought affects these two traits more by environment than by genotype. The correlation coefficients of RSA, RL and RL were very low, indicating that they were more influenced by genotype.

Studies have shown that the genetic relatedness of the mapping population can increase the false positive risk of GWAS results (Ali et al., 2020). A population with enough genotype and trait diversity is considered to be the expected GWAS population (Flint-Garcia et al., 2005). In this study, the population panel consisting of 226 accessions were collected from different ecological regions. Three population structure analysis methods (admixture, phylogenetic, and PCA) were conducted. Results showed that 226 sunflower materials had large genetic differences and were an ideal GWAS population. Linkage disequilibrium (LD) is the basis of GWAS (Ali et al., 2020). When LD declines rapidly with distance, LD mapping is potentially very precise (Gaut and Long, 2003). Since our materials have high genetic variability, the LD-decay distance is about 20 kb. Overall patterns of LD decay show chromosome specificity. Chr10 showed the highest LD value, followed by Chr7, Chr5, Chr13, and Chr17. This result is consistent with a previous study conducted by Filippi et al. (2020). They have reported different patterns of LD across chromosomes, with Chr10, Chr17, Chr5, and Chr2 showing the highest LD. The extended LD in Chr10 and Chr5 were also reported by other researchers (Cadic et al., 2013; Mandel et al., 2013). Owens et al. showed that the extended LD on Chr10 could be the result of the wild introgression in the fertility restoring male lines (Owens et al., 2019).

GWAS methods have evolved over years. Several new methods are being developed to improve the statistical power and reduce the computational time. FarmCPU uses a set of markers associated with a casual gene as a co-factor instead of kinship to avoid overfitting and eliminate confounding between kinship and testing markers iteratively (Liu et al., 2016). More recently, along with improvements in statistical power and reduction in computing time compared to FarmCPU, the new method called BLINK is set to eliminate FarmCPU requirement that quantitative trait nucleotides (QTNs) are evenly distributed in the genome (Huang et al., 2019). In the present study, we used 3 methods simultaneously to conduct GWAS. The FarmCPU method detected 44 SNPs, the BLINK method detected 33 SNPs, and the MLM method detected the lowest of 19 SNPs, respectively. There were 12 SNPs found simultaneously by FarmCPU and BLINK method, and only 3 common SNPs were found by 3 methods. Most SNPs were only found in one method. Therefore, it may be prudent to use multiple methods to conduct a GWAS survey (Nida et al., 2021).

STI and SSI are two commonly used evaluation indexes in the study of plant abiotic stress. According to the research of Mehdi GHAFFARI, STI is more efficient for identifying drought-resistant lines, and SSI is more efficient for identifying drought-sensitive lines (Ghaffari et al., 2012). Applying both indicators simultaneously could provide a complete and accurate assessment of drought tolerance. Strangely, the calculation methods of STI in different articles are inconsistent (Sukumaran et al., 2018; Khanzada et al., 2020; Chaurasia et al., 2021). In the present study, we carefully chose a scientific STI calculation method for GWAS analysis. A total of 80 significant SNP markers associated with 8 phenotypic traits were detected, 22 of them were detected using SSI, and 59 of them were detected using STI, only one common SNP was detected by both of the two indexes.

To further understand the biological processes, pathways, and gene expression patterns in sunflowers under drought stress, we conducted an RNA-seq analysis. Based on the phenotypic traits, a drought-tolerant plant was selected from the GWAS population. We sampled the leaves at 0, 7, and 14 days after drought stress. A total of 18,922 differentially expressed genes were obtained.

There was a noticeable consistency between the results of GO and KEGG analysis. For example, up-regulated genes were enriched in GO-terms such as cellular amino acid catabolic process (GO:0009063), branched-chain amino acid catabolic process (GO:0009083), while KEGG analysis showed "Valine, leucine and isoleucine degradation" was the most significant pathway. Down-regulated genes were enriched in photosynthesis (GO:0015979), photosynthesis, light reaction (GO:0019684) according to GO analysis, while KEGG analysis showed down-regulated genes enriched in pathways such as

TABLE 2 | Detail information of 14 genes obtained by combine-analysis of GWAS and RNA-seq.

Traits	Gene name	Chromosome	Gene_start	Gene_end	Description	iTak	Families
(PH/LSA)-STI	LOC110899235	13	138759411	138769673	Inosine-uridine preferring nucleoside hydrolase		
	LOC110899238	13	138795923	138801286	ABC transporter C family member 3-like		
(LSA/SPAD)-(SSI/STI)	LOC110885273	10	123870286	123873341	Serine threonine-protein kinase	PK	RLK-Pelle_SD-2b
LSA-SSI	LOC110894816	12	55570908	55573165	Equilibrative nucleotide transporter		
	LOC110936334	4	60470023	60472767	Jacalin-like lectin domain		
	LOC110941963	5	200316650	200319863	Microtubule-associated protein RP EB family member		
	LOC110891369	11	160106964	160111888	Receptor-like protein kinase	PK	RLK-Pelle_CrRLK1L-1
	LOC110920644	17	8881157	8883452	PLATZ transcription factor	TF	PLATZ
RSR-SSI	LOC110937937	4	169924932	169927583	Component of the peroxisomal and mitochondrial division machineries. Plays a role in promoting the fission of mitochondria and peroxisomes		
	LOC110915715	16	39633096	39638580	Protein of unknown function (DUF1666)		
RV-STI	LOC110877324	9	29925998	29927713	Belongs to the UDP-glycosyltransferase family		
	LOC110917707	16	74795170	74799439	Domain present in VPS-27, Hrs and STAM		
RSA-STI	LOC110872899	8	68366663	68376805	Inactive leucine-rich repeat receptor-like serine threonine-protein kinase	PK	RLK-Pelle_LRR-III
RWC-SSI	LOC110941862	5	195719872	195730082	Topless-related protein		

The content in brackets indicates simultaneous, for example, (PH/LSA)-STI, indicating that this gene is recognized by both PH-STI and LSA-STI.

Photosynthesis proteins (BR:ko00194), Photosynthesis—antenna proteins, Photosynthesis. The branched-chain amino acids (BCAAs), including isoleucine, leucine, and valine, are essential for plants (Binder et al., 2007). Pires et al. (2016) results highlight that catabolism of BCAA appears to play an important role in the mechanism of tolerance to short-term drought, most likely by delaying the onset of stress. Our results also proved that the degradation of BCAA may be an important mechanism of sunflower drought resistance. Abiotic stress damage the thylakoid membrane, disturb its functions, and ultimately decrease photosynthesis. Down-regulated expression of photosynthesis-related genes under drought stress has been reported in several plants, such as *Arabidopsis* (Bechtold et al., 2016; Bouzid et al., 2019), wheat (Derakhshani et al., 2020), and grapevines (Franck et al., 2020). In a previous study, Escalante et al. found a down-regulation of photosynthesis-related genes in the aerial part of sunflowers (Moschen et al., 2017). However, another study revealed that the expression levels of photosynthesis-related genes were increased under drought stress in sunflowers (Escalante et al., 2020). This difference may be caused by differences in drought intensity and genotype, and our results were identical with the former.

With the development of high-throughput technologies, omics research is also undergoing a shift from a single-omics to a large-scale multi-omics approach (Liu et al., 2020). Through the multi-omics approach, researchers can obtain a deeper understanding of the fundamental biological processes, a more accurate prediction of the response variable, and gain further insight into mechanistic aspects of the system (Cavill et al., 2015). By integrating the transcriptome and metabolome, Sebastián Moschen et al. (2017) gained a deeper insight into the sunflower drought-response mechanism. The integration of genomic and transcriptomic analysis has also been reported in many recent studies. This approach can be used as an effective way to identify candidate genes. For example, eight salt stress-related candidate genes were identified by a combination of GWAS analysis and transcriptome analysis in Alfalfa (*Medicago sativa* L.) (He et al., 2021). Seven candidate genes for seminal root length in maize (*Zea mays* L.) were identified by integrating the results of the GWAS, the common DEGs, and the co-expression network analysis (Guo et al., 2020). Using a combined analysis, we identified 18 common genes.

The total genes in the sunflower reference genome were 81,496, and we found 18,922 DEGs *via* RNA-seq. According to this proportion, we should find at least 29 DEGs among the 118

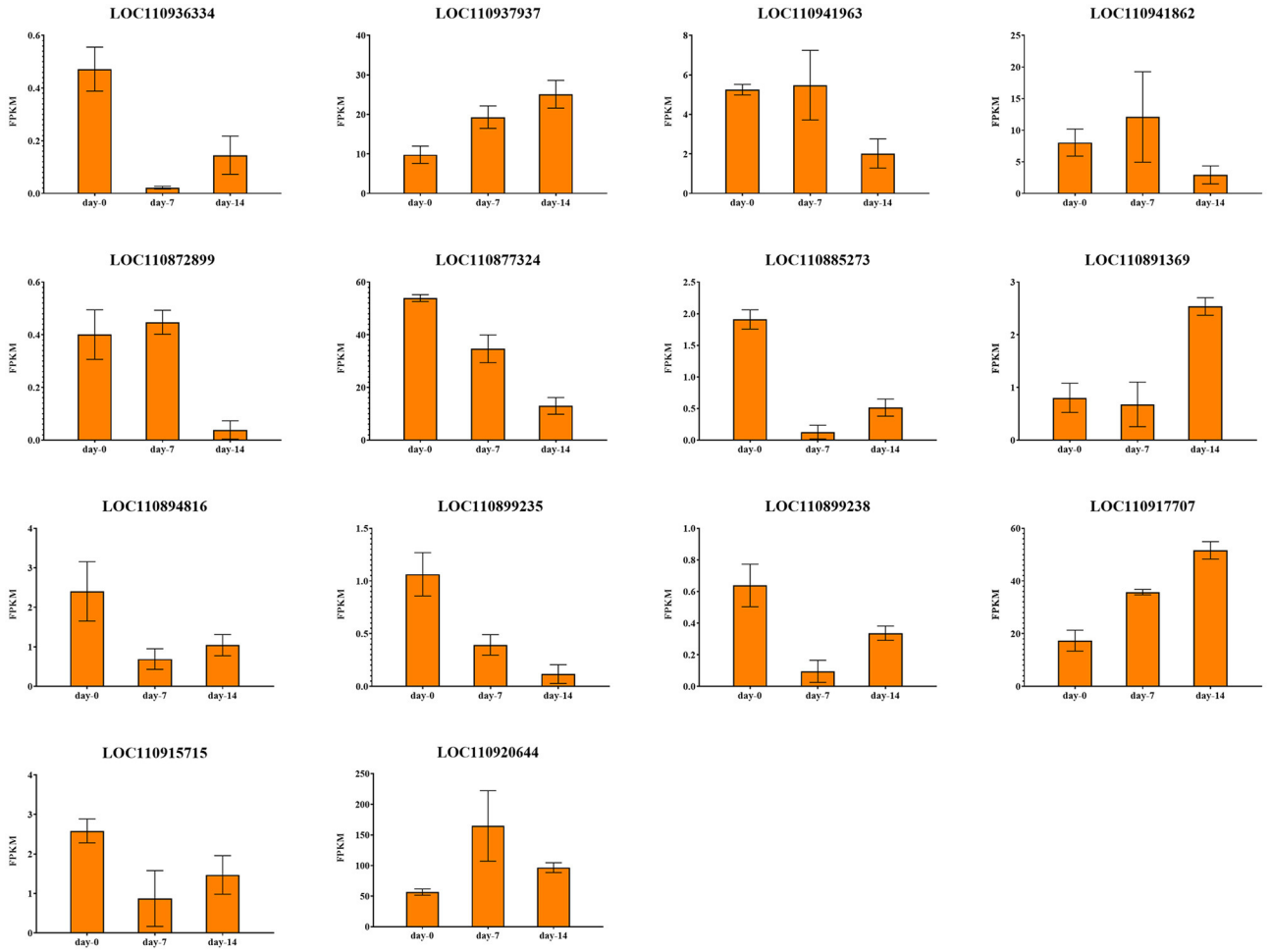


FIGURE 11 | Expression profiles of 14 drought response candidate genes.

genes of GWAS. However, the number of common genes that we have found was relatively small (18). This is because among the 18,922 DEGs, only 12,124 of them exist in the reference genome and the rest are novel genes. A subsequent chi-square test using this number found no significant difference between the two proportions ($P = 0.908$). Nonetheless, the proportion of overlapped genes was still lower than we expected. The reason we speculate is that GWAS candidate genes are mainly regulatory genes that act in all accessions. A slight regulation of expression level under drought stress, which did not reach the threshold of significant difference, can affect the physiological processes in plants, whereas the DEGs of RNA-seq are mainly a series of drought-responsive functional genes that are regulated in K58 under drought stress. The difference in the class and function of the genes from these two gene sets results in a low percentage of overlapping genes. Of course, this needs further confirmation.

Among these 18 genes, 14 are protein-coding genes, of which 3 are encoding PK and 1 encodes TF. These genes may play an important role in drought response in sunflowers. The LOC110885273 gene encodes G-type lectin S-receptor-like serine/threonine-protein kinase (LecRLKs). The protein kinase is involved in plant responses to biotic and abiotic stresses (Bonaventure, 2011; Singh et al., 2012; Zhao et al., 2016). Overexpression of G-type LecRLKs enhances the drought tolerance of *Arabidopsis thaliana* (Sun et al., 2013a), which may be achieved by controlling stomata size through interaction with abscisic acid (ABA) (Arnaud et al., 2012). Pan et al. (2020) identified a LecRLKs gene OsESG1 in rice and found it could be induced by treating with PEG, NaCl, and ABA. However, we found the LOC110885273 gene was down-regulated under drought stress, which may lead to the decrease of SPAD value under drought stress.

The receptor like kinase (RLKs) family has been defined as the most abundant gene family in *Arabidopsis*. Leucine rich repeat-RLKs (LRRRLKs) are the largest group of receptor-kinases in Arabidopsis, which is widely involved in responses to various biotic and abiotic stresses (Diévart and Clark, 2003; Lehti-Shiu et al., 2009). Osakabe et al. (2005) found that an LRRRLKs gene (RPK1) is involved in the early steps in the

ABA signaling pathway through a gene knock-out experiment. The overexpression of receptor-like kinase rich in the Leucine Repetition gene improves the *Arabidopsis thaliana* drought resistance (Xing et al., 2011). Receptor-like cytoplasmic kinase GUDK and OsSIK1 were shown to enhance drought tolerance in rice (Ouyang et al., 2010; Harb et al., 2020). In the present study, a down-regulated LRRRLKs gene LOC110872899 was identified, which is located at chromosome 8, and associated with RSA, maybe the mechanism of this gene in sunflower drought tolerance response is different. Another receptor-like protein kinase gene LOC110891369 was up-regulated at 14-days of drought stress in K58, which belongs to the family of RLK-Pelle_CrRLK1L-1, and is associated with LSA.

PLATZ transcription factors play important roles in plant growth, development, and biotic and abiotic stress responses. Liu et al. (2021) reveal that PLATZ4 interacts with AITR6 to increase ABA sensitivity and drought tolerance in *Arabidopsis* by regulating the expression of different genes. Zenda et al. (2019) identified a PLATZ gene (Zm00001d051511) in maize. It was up-regulated in tolerant line YE8112, whilst down-regulated in drought-sensitive line after drought stress. This result indicated the TF genes could be the key contributors to drought stress tolerance in the drought-tolerant maize inbred line. This different expression pattern was also proved in Ray's research on rice (Ray et al., 2011), PLATZ (LOC_Os10g42410) gene was down-regulated in panicle, while up-regulated in vegetative tissues under drought stress. Even in the same tissue at the same time, it was found that the expression levels of two PLAZT genes were up-regulated and down-regulated, respectively, which indicated the complexity of drought stress regulation. In this study, a PLAZT gene LOC110920644, which is related with LSA, was up-regulated at the early stage in K58 under drought stress.

ABA is an important hormone for plant drought response (Zotova et al., 2018). The cell ABA level increases under drought stress, leading to stomatal closure and active several stress-responsive genes (Cutler et al., 2010). Drought stress increased ABA levels in sunflowers have been reported (Robertson et al., 1985). In this study, the functions of the four TF/PK genes are all related to ABA, indicating the important role of the ABA-dependent process in the drought response of sunflowers.

CONCLUSION

Sunflower is one of the most important oil crops in the world, which is often grown as a rainfed crop. Water limitation at the seedling stage can severely reduce stand establishment and negatively impact yields. However, the molecular mechanism underlying drought resistance is still not fully understood. In this study, we used SLAF-seq to perform GWAS for 8 important phenotypic traits in 226 sunflower inbred lines. Using three methods (i.e., MLM, FarmCPU, and BLINK) for sunflower grown in two conditions (i.e., well-water and drought stress), we identified a total of 80 SNP displaying a significant association ($p < 1.062 \times 10^{-6}$). Candidate genes were searched in the 20 kb up/down-stream of each SNP. There were 85 protein-coding candidate genes possibly related to the 8 important phenotypic traits. Next, we conducted an RNA-seq based on a drought-tolerance inbred line (K58). A total of 18,922 DEGs were identified on 7 and 14 days after drought treatment. Up-regulated genes were mainly enriched in BCAA catabolic process, while down-regulated genes were mainly enriched in the photosynthesis process. Using a combined analysis, we found 14 common genes between GWAS and RNA-seq, three of them were PK genes, and one of them was TF gene. LOC110885273 was associated with LSA and SPAD, belongs to the RLK-Pelle_SD-2b protein kinase family. LOC110872899 belongs to the RLK-Pelle_LRR-III protein kinase family and is associated with RSA. LOC110891369 belongs to the RLK-Pelle_CrRLK1L-1 protein kinase family and is associated with LSA. The PLAZT gene LOC110920644 is related to LSA, and belongs to PLAZT TF family. Through functional analysis, there are 4 genes involving the ABA-dependent drought response pathway of plants.

The integrative analysis of omics data is a promising approach to identify candidate genes for complex traits. This study is the first attempt to combine GWAS and RNA-seq to explore the genetic mechanism of sunflower drought tolerance to our knowledge. We will further validate the functions of these genes, possibly by overexpression or by CRISPER/Cas genome editing. Our research reveals the phenotypic and molecular mechanisms of drought response in sunflowers. The results will be useful for the genetic enhancement of drought-resistant sunflowers.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available. This data can be found here: https://www.ncbi.nlm.nih.gov/search/all/?term=PRJNA797473.

AUTHOR CONTRIBUTIONS

JH and LY: conceptualization. LY, YM, and YW: methodology. YW: software, writing—original draft preparation, and visualization. JH: validation, project administration, and funding acquisition. YW, ZZ, and HS: formal analysis. YM, ZZ, HH, and ZH: investigation. HY: resources. YM and ZZ: data curation. LY, YZ, and HY: writing—review and editing. LY: supervision. All authors have read and agreed to the published version of the manuscript.

FUNDING

This work was financially supported by grants from the National Natural Science Foundation of China (Nos. 32160450 and 31760396).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.847435/full#supplementary-material

REFERENCES

Adeleke, B. S., and Babalola, O. O. (2020). Oilseed crop sunflower (*Helianthus annuus*) as a source of food: nutritional and health benefits. *Food Sci. Nutr.* 8, 4666–4684. doi: 10.1002/fsn3.1783

Adler, D., Nenadic, O., and Zucchini, W. (2003). "Rgl: a r-library for 3d visualization with opengl," in *Proceedings of the 35th Symposium of the Interface: Computing Science and Statistics. Salt Lake City*, 1–11.

Adopted, I. (2014). *Climate Change 2014 Synthesis Report*. Geneva: IPCC.

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi: 10.1101/gr.094052.109

Ali, N., Li, D., Eltahawy, M. S., Abdulmajid, D., Bux, L., Liu, E., et al. (2020). Mining of favorable alleles for seed reserve utilization efficiency in Oryza sativa by means of association mapping. *BMC Genetics* 21, 1–15. doi: 10.1186/s12863-020-0811-3

Alza, J., and Fernandez-Martinez, J. (1997). Genetic analysis of yield and related traits in sunflower (*Helianthus annuus* L.) in dryland and irrigated environments. *Euphytica* 95, 243–251. doi: 10.1023/A:1003056500991

Anders, S., and Huber, W. (2012). *Differential Expression of RNA-Seq Data at the Gene Level–the DESeq Package*. Heidelberg: European Molecular Biology Laboratory (EMBL), 10.

Arnaud, D., Desclos-Theveniau, M., and Zimmerli, L. (2012). Disease resistance to Pectobacterium carotovorum is negatively modulated by the Arabidopsis Lectin Receptor Kinase LecRK-V. 5. *Plant Signal. Behav.* 7, 1070–1072. doi: 10.4161/psb.21013

Ault, T. R. (2020). On the essentials of drought in a changing climate. *Science* 368, 256–260. doi: 10.1126/science.aaz5492

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., et al. (2004). The Pfam protein families database. *Nucleic Acids Res.* 32(Suppl. 1), D138–D141. doi: 10.1093/nar/gkh121

Bechtold, U., Penfold, C. A., Jenkins, D. J., Legaie, R., Moore, J. D., Lawson, T., et al. (2016). Time-series transcriptomics reveals that AGAMOUS-LIKE22 affects primary metabolism and developmental processes in drought-stressed Arabidopsis. *Plant Cell* 28, 345–366. doi: 10.1105/tpc.15.00910

Binder, S., Knill, T., and Schuster, J. (2007). Branched-chain amino acid metabolism in higher plants. *Physiol. Plantarum* 129, 68–78. doi: 10.1111/j.1399-3054.2006.00800.x

Blum, A. (2011). Drought resistance-is it really a complex trait? *Func. Plant Biol.* 38, 753–757. doi: 10.1071/FP11101

Bonaventure, G. (2011). The Nicotiana attenuata LECTIN RECEPTOR KINASE 1 is involved in the perception of insect feeding. *Plant Signal. Behav.* 6, 2060–2063. doi: 10.4161/psb.6.12.18324

Bouzid, M., He, F., Schmitz, G., Häusler, R., Weber, A., Mettler-Altmann, T., et al. (2019). Arabidopsis species deploy distinct strategies to cope with drought stress. *Ann. Botany* 124, 27–40. doi: 10.1093/aob/mcy237

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

Cavill, R., Jennen, D., Kleinjans, J., and Briedé, J. J. (2015). Transcriptomic and metabolomic data integration. *Briefings Bioinform.* 17, 891–901. doi: 10.1093/bib/bbv090

Chaurasia, S., Singh, A. K., Kumar, A., Songachan, L., Yadav, M. C., Kumar, S., et al. (2021). Genome-wide association mapping reveals key genomic regions for physiological and yield-related traits under salinity stress in wheat (*Triticum aestivum* L.). *Genomics* 113, 3198–3215. doi: 10.1016/j.ygeno.2021.07.014

Costa, C., Dwyer, L. M., Dutilleul, P., Stewart, D. W., Ma, B. L., and Smith, D. L. (2001). Inter-relationships of applied nitrogen, SPAD, and yield of leafy and non-leafy maize genotypes. *J. Plant Nutr.* 24, 1173–1194. doi: 10.1081/PLN-100106974

Cutler, S. R., Rodriguez, P. L., Finkelstein, R. R., and Abrams, S. R. (2010). Abscisic acid: emergence of a core signaling network. *Ann. Rev. Plant Biol.* 61, 651–679. doi: 10.1146/annurev-arplant-042809-112122

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Davey, J. W., Cezard, T., Fuentes-Utrilla, P., Eland, C., Gharbi, K., and Blaxter, M. L. (2013). Special features of RAD Sequencing data: implications for genotyping. *Mol. Ecol.* 22, 3151–3164. doi: 10.1111/mec.12084

Derakhshani, B., Ayalew, H., Mishina, K., Tanaka, T., Kawahara, Y., Jafary, H., et al. (2020). Comparative analysis of root transcriptome reveals candidate genes and expression divergence of homoeologous genes in response to water stress in wheat. *Plants.* 9, 596. doi: 10.3390/plants9050596

Diévart, A., and Clark, S. E. (2003). Using mutant alleles to determine the structure and function of leucine-rich repeat receptor-like kinases. *Curr. Opin. Plant Biol.* 6, 507–516. doi: 10.1016/S1369-5266(03)00089-X

Ebrahimi Khaksefidi, R., Mirlohi, S., Khalaji, F., Fakhari, Z., Shiran, B., Fallahi, H., et al. (2015). Differential expression of seven conserved microRNAs in response to abiotic stress and their regulatory network in Helianthus annuus. *Front. Plant Sci.* 6, 741. doi: 10.3389/fpls.2015.00741

Escalante, M., Vigliocco, A., Moschen, S., Fernández, P., Heinz, R., Garcia-Garcia, F., et al. (2020). Transcriptomic analysis reveals a differential gene expression profile between two sunflower inbred lines with different ability to tolerate water stress. *Plant Mol. Biol. Rep.* 38, 1–16. doi: 10.1007/s11105-020-01192-4

Espina, M. J., Ahmed, C. M. S., Bernardini, A., Adeleke, E., Yadegari, Z., Arelli, P., et al. (2018). Development and phenotypic screening of an ethyl methane sulfonate mutant population in soybean. *Front. Plant Sci.* 9, 394. doi: 10.3389/fpls.2018.00394

FAO (2021). *FAO Statistical Yearbook – World Food and Agriculture*. FAO.

Farooq, M., Wahid, A., Kobayashi, N., Fujita, D., and Basra, S. (2009). Plant drought stress: effects, mechanisms and management. *Sustain. Agric.* 29, 153–188. doi: 10.1007/978-90-481-2666-8_12

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.* Shanhua (1992), 257–270.

Filippi, C. V., Merino, G. A., Montecchia, J. F., Aguirre, N. C., Rivarola, M., Naamati, G., et al. (2020). Genetic diversity, population structure and linkage disequilibrium assessment among international sunflower breeding collections. *Genes* 11, 283. doi: 10.3390/genes11030283

Fischer, R., and Maurer, R. (1978). Drought resistance in spring wheat cultivars. I. Grain yield responses. *Austr. J. Agric. Res.* 29, 897–912. doi: 10.1071/AR9780897

Flexas, J., Bota, J., Loreto, F., Cornic, G., and Sharkey, T. (2004). Diffusive and metabolic limitations to photosynthesis under drought and salinity in C3 plants. *Plant Biol.* 6, 269–279. doi: 10.1055/s-2004-820867

Flint-Garcia, S. A., Thuillet, A. C., Yu, J., Pressoir, G., Romero, S. M., Mitchell, S. E., et al. (2005). Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J.* 44, 1054–1064. doi: 10.1111/j.1365-313X.2005.02591.x

Francis, R. M. (2017). pophelper: an R package and web app to analyse and visualize population structure. *Mol. Ecol. Res.* 17, 27–32. doi: 10.1111/1755-0998.12509

Franck, N., Zamorano, D., Wallberg, B., Hardy, C., Ahumada, M., Rivera, N., et al. (2020). Contrasting grapevines grafted into naturalized rootstock suggest scion-driven transcriptomic changes in response to water deficit. *Sci. Horticult.* 262, 109031. doi: 10.1016/j.scienta.2019.109031

Galmes, J., Conesa, M. A., Ochogavía, J. M., Perdomo, J. A., Francis, D. M., Ribas-Carb,ó, M., et al. (2011). Physiological and morphological adaptations in relation to water use efficiency in Mediterranean accessions of *Solanum lycopersicum*. *Plant Cell. Environ.* 34, 245–260. doi: 10.1111/j.1365-3040.2010.02239.x

Gaut, B. S., and Long, A. D. (2003). The lowdown on linkage disequilibrium. *Plant Cell.* 15, 1502–1506. doi: 10.1105/tpc.150730

Geetha, A., Suresh, J., and Saidaiah, P. (2012). Study on response of sunflower (*Helianthus annuus* L.) genotypes for root and yield characters under water stress. *Curr. Biot.* 6, 32–41. doi: 10.1111/j.1365-313X.2003.01987.x

Ghaffari, M., Toorchi, M., Valizadeh, M., and Shakiba, M. R. (2012). Morpho-physiological screening of sunflower inbred lines under drought stress condition. *Turkish J. Field Crop.* 17, 185–190. doi: 10.2298/GENSR1203701Z

Grasso, S., Pintado, T., Pérez-Jiménez, J., Ruiz-Capillas, C., and Herrero, A. M. (2020). Potential of a sunflower seed by-product as animal fat replacer in healthier frankfurters. *Foods* 9, 445. doi: 10.3390/foods9040445

Gunes, A., Pilbeam, D. J., Inal, A., and Coban, S. (2008). Influence of silicon on sunflower cultivars under drought stress, I: growth, antioxidant mechanisms, and lipid peroxidation. *Commun. Soil Sci. Plant Anal.* 39, 1885–1903. doi: 10.1080/00103620802134651

Guo, J., Li, C., Zhang, X., Li, Y., Zhang, D., Shi, Y., et al. (2020). Transcriptome and GWAS analyses reveal candidate gene for seminal root length of maize seedlings under drought stress. *Plant Sci.* 292, 110380. doi: 10.1016/j.plantsci.2019.110380

Haddadi, P., Yazdi-Samadi, B., Naghavi, M. R., Kalantari, A., Maury, P., and Sarrafi, A. (2011). QTL analysis of agronomic traits in recombinant inbred lines of sunflower under partial irrigation. *Plant Biotechnol. Rep.* 5, 135–146. doi: 10.1007/s11816-011-0164-5

Harb, A., Simpson, C., Guo, W., Govindan, G., Kakani, V. G., and Sunkar, R. (2020). The effect of drought on transcriptome and hormonal profiles in barley genotypes with contrasting drought tolerance. *Front. Plant Sci.* 11, 618491. doi: 10.3389/fpls.2020.618491

He, F., Wei, C., Zhang, Y., Long, R., Li, M., Wang, Z., et al. (2021). Genome-wide association analysis coupled with transcriptome analysis reveals candidate genes related to salt stress in alfalfa (*Medicago sativa* L.). *Front. Plant Sci.* 12, 826584–826584. doi: 10.3389/fpls.2021.826584

Hou, S., Zhu, G., Li, Y., Li, W., Fu, J., Niu, E., et al. (2018). Genome-wide association studies reveal genetic variation and candidate genes of drought stress related traits in cotton (*Gossypium hirsutum* L.). *Front. Plant Sci.* 9, 1276. doi: 10.3389/fpls.2018.01276

Huang, M., Liu, X., Zhou, Y., Summers, R. M., and Zhang, Z. (2019). BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. *Gigascience*. 8, giy154. doi: 10.1093/gigascience/giy154

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 47, D309–D314. doi: 10.1093/nar/gky1085

Hussain, M., Farooq, S., Hasan, W., Ul-Allah, S., Tanveer, M., Farooq, M., et al. (2018). Drought stress in sunflower: physiological effects and its management through breeding and agronomic alternatives. *Agric. Water Manag.* 201, 152–166. doi: 10.1016/j.agwat.2018.01.028

Hussain, M., Malik, M., Farooq, M., Ashraf, M., and Cheema, M. (2008). Improving drought tolerance by exogenous application of glycinebetaine and salicylic acid in sunflower. *J. Agron. Crop. Sci.* 194, 193–199. doi: 10.1111/j.1439-037X.2008.00305.x

Hussain, R. A., Ahmad, R., Nawaz, F., Ashraf, M. Y., and Waraich, E. A. (2016). Foliar NK application mitigates drought effects in sunflower (*Helianthus annuus* L.). *Acta Physiol. Plant.* 38, 1–14. doi: 10.1007/s11738-016-2104-z

Hussain, S., Saleem, M., Ashraf, M., Cheema, M., and Haq, M. (2010). Abscisic acid, a stress hormone helps in improving water relations and yield of sunflower (*Helianthus annuus* L.) hybrids under drought. *Pakis. J. Bot.* 42, 2177–2189. doi: 10.3417/2008072

Ibrahim, M., Faisal, A., and Shehata, S. (2016). Calcium chloride alleviates water stress in sunflower plants through modifying some physio-biochemical parameters. *American-Eurasian J. Agric. Environ. Sci.* 16, 677–693. doi: 10.5829/idosi.aejaes.2016.16.4.12907

Javaid, T., Bibi, A., Sadaqat, H., and Javed, S. (2015). Screening of sunflower (*Helianthus annuus* L.) hybrids for drought tolerance at seedling stage. *Int. J. Plant Sci. Ecol.* 1, 6–16.

Kaur, G., and Asthir, B. (2017). Molecular responses to drought stress in plants. *Biol. Plantarum* 61, 201–209. doi: 10.1007/s10535-016-0700-9

Kaya, M. D., Okçu, G., Atak, M., Cikili, Y., and Kolsarici, Ö. (2006). Seed treatments to overcome salt and drought stress during germination in sunflower (*Helianthus annuus* L.). *Eur. J. Agron.* 24, 291–295. doi: 10.1016/j.eja.2005.08.001

Khan, S., Choudhary, S., Pandey, A., Khan, M. K., and Thomas, G. (2015). Sunflower oil: efficient oil source for human consumption. *Emerg. Life Sci. Res.* 1, 1–3. Available online at: https://www.emergentresearch.org/uploads/38/1768_pdf.pdf

Khanzada, H., Wassan, G. M., He, H., Mason, A. S., Keerio, A. A., Khanzada, S., et al. (2020). Differentially evolved drought stress indices determine the genetic variation of Brassica napus at seedling traits by genome-wide association mapping. *J. Adv. Res.* 24, 447–461. doi: 10.1016/j.jare.2020.05.019

Kiani, S. P., Talia, P., Maury, P., Grieu, P., Heinz, R., Perrault, A., et al. (2007). Genetic analysis of plant water status and osmotic adjustment in recombinant inbred lines of sunflower under two water treatments. *Plant Sci.* 172, 773–787. doi: 10.1016/j.plantsci.2006.12.007

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi: 10.1038/nmeth.3317

Kofsky, J., Zhang, H., and Song, B.-H. (2020). Genetic architecture of early vigor traits in wild soybean. *Int. J. Mol. Sci.* 21, 3105. doi: 10.3390/ijms21093105

Lata, C., Muthamilarasan, M., and Prasad, M. (2015). "Drought Stress Responses and Signal Transduction in Plants," in *Elucidation of Abiotic Stress Signaling in Plants: Functional Genomics Perspectives, Vol. 2*, ed. G. K. Pandey (New York, NY: Springer New York), 195–225.

Lefort, V., Desper, R., and Gascuel, O. (2015). FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol. Biol. Evol.* 32, 2798–2800. doi: 10.1093/molbev/msv150

Lehti-Shiu, M. D., Zou, C., Hanada, K., and Shiu, S.-H. (2009). Evolutionary history and stress regulation of plant receptor-like kinase/pelle genes. *Plant Physiol.* 150, 12–26. doi: 10.1104/pp.108.134353

Letunic, I., and Bork, P. (2021). Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* 49, W293–W296. doi: 10.1093/nar/gkab301

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

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

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, M., Wang, C., Ji, Z., Zhang, L., Li, C., Huang, J., et al. (2021). Regulation of drought tolerance in Arabidopsis involves PLATZ4-mediated transcriptional suppression of PIP2. *bioRxiv*. doi: 10.1101/2021.10.06.463369

Liu, S.-H., Shen, P.-C., Chen, C.-Y., Hsu, A.-N., Cho, Y.-C., Lai, Y.-L., et al. (2020). DriverDBv3: a multi-omics database for cancer driver gene research. *Nucleic Acids Res.* 48, D863–D870. doi: 10.1093/nar/gkz964

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

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2– ΔΔCT method. *Methods* 25, 402–408. doi: 10.1006/meth.2001.1262

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550. doi: 10.1186/s13059-014-0550-8

Ma, J., Geng, Y., Pei, W., Wu, M., Li, X., Liu, G., et al. (2018). Genetic variation of dynamic fiber elongation and developmental quantitative trait locus mapping of fiber length in upland cotton (Gossypium hirsutum L.). *BMC Genomics* 19, 882–882. doi: 10.1186/s12864-018-5309-2

Ma, X., Feng, F., Wei, H., Mei, H., Xu, K., Chen, S., et al. (2016). Genome-wide association study for plant height and grain yield in rice under contrasting moisture regimes. *Front. Plant Sci.* 7, 1801. doi: 10.3389/fpls.2016.01801

Mandel, J. R., Nambeesan, S., Bowers, J. E., Marek, L. F., Ebert, D., Rieseberg, L. H., et al. (2013). Association mapping and the genomic consequences of selection in sunflower. *PLoS Genetics* 9, e1003378. doi: 10.1371/journal.pgen.1003378

Martínez, D., and Guiamet, J. (2004). Distortion of the SPAD 502 chlorophyll meter readings by changes in irradiance and leaf water status. *Agronomie* 24, 41–46. doi: 10.1051/agro:2003060

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Moschen, S., Di Rienzo, J. A., Higgins, J., Tohge, T., Watanabe, M., González, S., et al. (2017). Integration of transcriptomic and metabolic data reveals hub transcription factors involved in drought stress response in sunflower (*Helianthus annuus* L.). *Plant Mol. Biol.* 94, 549–564. doi: 10.1007/s11103-017-0625-5

Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D. J., Salichos, L., et al. (2016). The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol.* 17, 53. doi: 10.1186/s13059-016-0917-0

Mwadzingeni, L., Shimelis, H., Rees, D. J. G., and Tsilo, T. J. (2017). Genome-wide association analysis of agronomic traits in wheat under drought-stressed and non-stressed conditions. *PloS ONE* 12, e0171692. doi: 10.1371/journal.pone.0171692

Mwale, S., Hamusimbi, C., and Mwansa, K. (2003). Germination, emergence and growth of sunflower (*Helianthus annuus* L.) in response to osmotic seed priming. *Seed Sci. Technol.* 31, 199–206. doi: 10.15258/sst.2003.31.1.21

Nida, H., Girma, G., Mekonen, M., Tirfessa, A., Seyoum, A., Bejiga, T., et al. (2021). Genome-wide association analysis reveals seed protein loci as determinants of variations in grain mold resistance in sorghum. *Theor. Appl. Genetics* 134, 1167–1184. doi: 10.1007/s00122-020-03762-2

Osakabe, Y., Maruyama, K., Seki, M., Satou, M., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2005). Leucine-rich repeat receptor-like kinase1 is a key membrane-bound regulator of abscisic acid early signaling in arabidopsis. *Plant Cell* 17, 1105–1119. doi: 10.1105/tpc.104.027474

Ouyang, S. Q., Liu, Y. F., Liu, P., Lei, G., He, S. J., Ma, B., et al. (2010). Receptor-like kinase OsSIK1 improves drought and salt stress tolerance in rice (*Oryza sativa*) plants. *Plant J.* 62, 316–329. doi: 10.1111/j.1365-313X.2010.04146.x

Owens, G. L., Baute, G. J., Hubner, S., and Rieseberg, L. H. (2019). Genomic sequence and copy number evolution during hybrid crop development in sunflowers. *Evolut. Appl.* 12, 54–65. doi: 10.1111/eva.12603

Pan, J., Li, Z., Wang, Q., Yang, L., Yao, F., and Liu, W. (2020). An S-domain receptor-like kinase, OsESG1, regulates early crown root development and drought resistance in rice. *Plant Sci.* 290, 110318. doi: 10.1016/j.plantsci.2019.110318

Pasda, G., and Diepenbrock, W. (1990). The physiological yield analysis of sunflower (*Helianthus annuus* L.). Part II. Climatic factors. *Fett Wissenschaft Technol.* 93, 155–168. doi: 10.1002/lipi.19910930501

Pereyra-Irujo, G. A., Velázquez, L., Granier, C., and Aguirrezabal, L. A. (2007). A method for drought tolerance screening in sunflower. *Plant Breed.* 126, 445–448. doi: 10.1111/j.1439-0523.2007.01375.x

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

Petcu, E., Arsintescu, A., and Stanciu, D. (2001). The effect of hydric stress on some characteristics of sunflower plants. *Romanian Agric. Res.* 16, 15–22.

Pires, M. V., Pereira Júnior, A. A., Medeiros, D. B., Daloso, D. M., Pham, P. A., Barros, K. A., et al. (2016). The influence of alternative pathways of respiration that utilize branched-chain amino acids following water shortage in Arabidopsis. *Plant Cell Environ.* 39, 1304–1319. doi: 10.1111/pce.12682

Poormohammad Kiani, S., Maury, P., Nouri, L., Ykhlef, N., Grieu, P., and Sarrafi, A. (2009). QTL analysis of yield-related traits in sunflower under different water treatments. *Plant Breed.* 128, 363–373. doi: 10.1111/j.1439-0523.2009.01628.x

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Human Genet.* 81, 559–575. doi: 10.1086/519795

Rabert, G. A., Manivannan, P., Somasundaram, R., and Panneerselvam, R. (2014). Triazole compounds alters the antioxidant and osmoprotectant status in drought stressed *Helianthus annuus* L. plants. *Emir. J. Food Agric*. 26, 265–276. doi: 10.9755/ejfa.v26i3.17385

Rauf, S. (2008). Breeding sunflower (*Helianthus annuus* L.) for drought tolerance. *Commun. Biomet. Crop Sci.* 3, 29–44. doi: 10.3923/pjbs.1999.846.848

Rauf, S., and Ahmad Sadaqat, H. (2008). Effect of osmotic adjustment on root length and dry matter partitioning in sunflower (*Helianthus annuus* L.) under drought stress. *Acta Agric. Scand. Section B–Soil Plant Sci.* 58, 252–260. doi: 10.1080/09064710701628958

Ray, S., Dansana, P. K., Giri, J., Deveshwar, P., Arora, R., Agarwal, P., et al. (2011). Modulation of transcription factor and metabolic pathway genes in response to water-deficit stress in rice. *Funct. Integr. Genomics* 11, 157–178. doi: 10.1007/s10142-010-0187-y

Riddell, E. A., Odom, J. P., Damm, J. D., and Sears, M. W. (2018). Plasticity reveals hidden resistance to extinction under climate change in the global hotspot of salamander diversity. *Sci. Adv.* 4, eaar5471. doi: 10.1126/sciadv.aar5471

Robertson, J. M., Pharis, R. P., Huang, Y. Y., Reid, D. M., and Yeung, E. C. (1985). Drought-induced increases in abscisic acid levels in the root apex of sunflower 1. *Plant Physiol.* 79, 1086–1089. doi: 10.1104/pp.79.4.1086

Salami, M., and Saadat, S. (2013). Study of potassium and nitrogen fertilizer levels on the yield of sugar beet in jolge cultivar. *J. Novel Appl. Sci.* 2, 94–100.

Schilling, E. E., and Heiser, C. B. (1981). Infrageneric classification of Helianthus (Compositae). *Taxon* 30, 393–403. doi: 10.2307/1220139

Singh, P., Kuo, Y.-C., Mishra, S., Tsai, C.-H., Chien, C.-C., Chen, C.-W., et al. (2012). The lectin receptor kinase-VI. 2 is required for priming and positively regulates Arabidopsis pattern-triggered immunity. *Plant Cell* 24, 1256–1270. doi: 10.1105/tpc.112.095778

Soni, P., and Abdin, M. Z. (2017). Water deficit-induced oxidative stress affects artemisinin content and expression of proline metabolic genes in *Artemisia annua* L. *FEBS Open Bio.* 7, 367–381. doi: 10.1002/2211-5463.12184

Sukumaran, S., Reynolds, M. P., and Sansaloni, C. (2018). Genome-wide association analyses identify QTL hotspots for yield and component traits in durum wheat grown under yield potential, drought, and heat stress environments. *Front. Plant Sci.* 9, 81. doi: 10.3389/fpls.2018.00081

Sun, X., Liu, D., Zhang, X., Li, W., Liu, H., Hong, W., et al. (2013b). SLAF-seq: an efficient method of large-scale de novo SNP discovery and genotyping using high-throughput sequencing. *PLoS ONE* 8, e58700. doi: 10.1371/journal.pone.0058700

Sun, X.-L., Yu, Q.-Y., Tang, L.-L., Ji, W., Bai, X., Cai, H., et al. (2013a). GsSRK, a G-type lectin S-receptor-like serine/threonine protein kinase, is a positive regulator of plant tolerance to salt stress. *J. Plant Physiol.* 170, 505–515. doi: 10.1016/j.jplph.2012.11.017

Supek, F., and Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* 6, e21800. doi: 10.1371/journal.pone.0021800

Tagliotti, M. E., Deperi, S. I., Bedogni, M. C., and Huarte, M. A. (2021). Genome-wide association analysis of agronomical and physiological traits linked to drought tolerance in a diverse potatoes (*Solanum tuberosum*) panel. *Plant Breed*. 140, 654–664. doi: 10.1111/pbr.12938

Togninalli, M., Roqueiro, D., Investigators, C. O., and Borgwardt, K. M. (2018). Accurate and adaptive imputation of summary statistics in mixed-ethnicity cohorts. *Bioinformatics* 34, i687–i696. doi: 10.1093/bioinformatics/bty596

Turhan, H., and Baser, I. (2004). *In vitro* and *in vivo* water stress in sunflower (*Helianthus Annuus* L.)/Estrés Hídrico En Girasol (*Helianthus annuus* L.) En Las Condiciones *in vitro* E *in vivo*/stress D'eau Du Tournesol (*Helianthus annus* L.) Dans Les Conditions *in vitro* Et *in vivo*. *Helia* 27, 227–236. doi: 10.2298/HEL0440227T

VanLiere, J. M., and Rosenberg, N. A. (2008). Mathematical properties of the r2 measure of linkage disequilibrium. *Theor. Popul. Biol.* 74, 130–137. doi: 10.1016/j.tpb.2008.05.006

Wang, H., and Qin, F. (2017). Genome-wide association study reveals natural variations contributing to drought resistance in crops. *Front. Plant Sci.* 8, 1110. doi: 10.3389/fpls.2017.01110

Wang, L., Liu, Y., Gao, L., Yang, X., Zhang, X., Xie, S., et al. (2022). Identification of candidate forage yield genes in sorghum (*Sorghum bicolor* L.) using integrated genome-wide association studies and RNA-Seq. *Front. Plant Sci.* 12, 433. doi: 10.3389/fpls.2021.788433

Wilhite, D. A., and Buchanan-Smith, M. (2005). Drought as hazard: understanding the natural and social context. *Drought Water Crises Sci. Technol. Manag. Issues* 3, 29. doi: 10.1201/9781420028386.pt1

Xie, D., Dai, Z., Yang, Z., Tang, Q., Deng, C., Xu, Y., et al. (2019). Combined genome-wide association analysis and transcriptome sequencing to identify candidate genes for flax seed fatty acid metabolism. *Plant Sci.* 286, 98–107. doi: 10.1016/j.plantsci.2019.06.004

Xing, H. T., Guo, P., Xia, X. L., and Yin, W. L. (2011). PdERECTA, a leucine-rich repeat receptor-like kinase of poplar, confers enhanced water use efficiency in Arabidopsis. *Planta* 234, 229–241. doi: 10.1007/s00425-011-1389-9

Yin, L. (2018). *CMplot: Circle Manhattan Plot*. San Francisco, CA: GitHub, Inc.

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics J. Integ. Biol.* 16, 284–287. doi: 10.1089/omi.2011.0118

Yu, J., and Buckler, E. S. (2006). Genetic association mapping and genome organization of maize. *Curr. Opin. Biotechnol.* 17, 155–160. doi: 10.1016/j.copbio.2006.02.003

Zenda, T., Liu, S., Wang, X., Liu, G., Jin, H., Dong, A., et al. (2019). Key maize drought-responsive genes and pathways revealed by comparative transcriptome and physiological analyses of contrasting inbred lines. *Int. J. Mol. Sci.* 20, 1268. doi: 10.3390/ijms20061268

Zhang, C., Dong, S.-S., Xu, J.-Y., He, W.-M., and Yang, T.-L. (2019). PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* 35, 1786–1788. doi: 10.1093/bioinformatics/bty875

Zhang, T., Zhang, X., Han, K., Zhang, G., Wang, J., Xie, K., et al. (2017). Analysis of long noncoding RNA and mRNA using RNA sequencing during the differentiation of intramuscular preadipocytes in chicken. *PLoS ONE* 12, e0172389. doi: 10.1371/journal.pone.0172389

Zhang, X., Wang, G., Chen, B., Du, H., Zhang, F., Zhang, H., et al. (2018). Candidate genes for first flower node identified in pepper using combined SLAF-seq and BSA. *PLoS ONE* 13, e0194071. doi: 10.1371/journal.pone.0194071

Zhao, W., Liu, Y.-W., Zhou, J.-M., Zhao, S.-P., Zhang, X.-H., and Min, D.-H. (2016). Genome-wide analysis of the lectin receptor-like kinase family in foxtail millet (*Setaria italica* L.). *Plant Cell Tissue Organ Cult.* 127, 335–346. doi: 10.1007/s11240-016-1053-y

Zheng, Y., Jiao, C., Sun, H., Rosli, H., Pombo, M. A., Zhang, P., et al. (2016). iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Mol. Plant* 9, 1667–1670. doi: 10.1016/j.molp.2016.09.014

Zhou, Q., Zhou, C., Zheng, W., Mason, A. S., Fan, S., Wu, C., et al. (2017). Genome-wide SNP markers based on SLAF-seq uncover breeding traces in rapeseed (*Brassica napus* L.). *Front. Plant Sci.* 8, 648. doi: 10.3389/fpls.2017.00648

Zilong, Z., Jianhua, H., Liuxi, Y., HaiFeng, Y., Huimin, S., Jinglin, W., et al. (2021). Drought resistance identification and drought resistance index screening of sunflower germplasm resources at seedling stage. *Agric. Res. Arid Area*. 39, 228–238. doi: 10.7606/j.issn.1000-7601.2021.04.29

Zotova, L., Kurishbayev, A., Jatayev, S., Khassanova, G., Zhubatkanov, A., Serikbay, D., et al. (2018). Genes encoding transcription factors TaDREB5 and TaNFYC-A7 are differentially expressed in leaves of bread wheat in response to drought, dehydration and ABA. *Front. Plant Sci.* 9, 1441. doi: 10.3389/fpls.2018.01441

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

ORIGINAL RESEARCH
published: 13 June 2022
doi: 10.3389/fpls.2022.929168

Detection of Stable Elite Haplotypes and Potential Candidate Genes of Boll Weight Across Multiple Environments via GWAS in Upland Cotton

Zhen Feng[1,2†], Libei Li[1,2†], Minqiang Tang[3†], Qibao Liu[1], Zihan Ji[1,2], Dongli Sun[1,2], Guodong Liu[4], Shuqi Zhao[5], Chenjue Huang[1,2], Yanan Zhang[1,2], Guizhi Zhang[4] and Shuxun Yu[1,2*]*

[1] College of Advanced Agriculture Sciences, Zhejiang A&F University, Hangzhou, China, [2] The Key Laboratory for Quality Improvement of Agricultural Products of Zhejiang Province, Zhejiang A&F University, Hangzhou, China, [3] Key Laboratory of Genetics and Germplasm Innovation of Tropical Special Forest Trees and Ornamental Plants (Ministry of Education), College of Forestry, Hainan University, Haikou, China, [4] Institute of Industrial Crops, Shandong Academy of Agricultural Sciences, Jinan, China, [5] Huanggang Academy of Agricultural Sciences, Huanggang, China

Boll weight (BW) is a key determinant of yield component traits in cotton, and understanding the genetic mechanism of BW could contribute to the progress of cotton fiber yield. Although many yield-related quantitative trait loci (QTLs) responsible for BW have been determined, knowledge of the genes controlling cotton yield remains limited. Here, association mapping based on 25,169 single-nucleotide polymorphisms (SNPs) and 2,315 insertions/deletions (InDels) was conducted to identify high-quality QTLs responsible for BW in a global collection of 290 diverse accessions, and BW was measured in nine different environments. A total of 19 significant markers were detected, and 225 candidate genes within a 400 kb region (± 200 kb surrounding each locus) were predicted. Of them, two major QTLs with highly phenotypic variation explanation on chromosomes A08 and D13 were identified among multiple environments. Furthermore, we found that two novel candidate genes (*Ghir_A08G009110* and *Ghir_D13G023010*) were associated with BW and that *Ghir_D13G023010* was involved in artificial selection during cotton breeding by population genetic analysis. The transcription level analyses showed that these two genes were significantly differentially expressed between high-BW accession and low-BW accession during the ovule development stage. Thus, these results reveal valuable information for clarifying the genetic basics of the control of BW, which are useful for increasing yield by molecular marker-assisted selection (MAS) breeding in cotton.

Keywords: SNP, boll weight, association mapping, candidate genes, MAS

OPEN ACCESS

Edited by:
Mallikarjuna Swamy, International Rice Research Institute (IRRI), Philippines

Reviewed by:
Hantao Wang, Cotton Research Institute (CAAS), China
Wenfeng Pei, Cotton Research Institute (CAAS), China
Navraj Sarao, Punjab Agricultural University, India
HongGe Li, Cotton Research Institute (CAAS), China

***Correspondence:**
Shuxun Yu
yushuxun@zafu.edu.cn
Libei Li
libeili@zafu.edu.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: *26 April 2022*
Accepted: *17 May 2022*
Published: *13 June 2022*

Citation:
Feng Z, Li L, Tang M, Liu Q, Ji Z, Sun D, Liu G, Zhao S, Huang C, Zhang Y, Zhang G and Yu S (2022) Detection of Stable Elite Haplotypes and Potential Candidate Genes of Boll Weight Across Multiple Environments via GWAS in Upland Cotton. Front. Plant Sci. 13:929168. doi: 10.3389/fpls.2022.929168

INTRODUCTION

Cotton has an ancient history of cultivation dating back seven thousand years or more according to the oldest archeological evidence, which was found in Pakistan (Rajpal et al., 2016). Subsequently, the invention of the cotton gin in the late 18th century caused massive growth in cotton production, and cotton gradually became an important cash crop (Sunilkumar et al., 2006). Previous studies have suggested that allotetraploids emerged approximately 1.5 million years ago (MYA) through a single allopolyploidization event in a propagule resembling diploid cotton (*Gossypium herbaceum* L.) that dispersed across the Atlantic Ocean from Africa to the New World and subsequently hybridized with a resembling diploid cotton (*Gossypium raimondii*) and produced upland cotton after long-term evolution (Wendel, 1989; Sunilkumar et al., 2006; Liu et al., 2015). Currently, upland cotton has become a predominant cotton species in global cotton commerce, with ∼ 27 million metric tons produced per year. In addition, it also provides natural fiber for the textile industry, which has high yield and wider adaptation (Chen et al., 2007). In recent years, due to population growth, climate change, and the challenges associated with maintaining the grain-cotton balance in farmlands, the cotton planting area has decreased. Therefore, the urgent need to increase cotton production is particularly important.

The application of quantitative trait locus linkages or QTL-related molecular markers of target traits by MAS can prevent environmental interference and improve breeding efficiency (Yin et al., 2003). The study of QTLs in cotton has focused mainly on yield and fiber quality component traits (Said et al., 2015). Cotton yield component traits include fruit branch number (FBN), lint percentage (LP), boll number per plant (BN), boll weight (BW), and seed index (SI), which were controlled by QTLs and environmental factors. Among these traits, BW is more stably inherited and has relatively high heritability (Fan et al., 2018; Liu et al., 2018; Zhang et al., 2019b; Gu et al., 2020; Zhu et al., 2021). In the past three decades, BW has been widely used for quantitative genetics studies, and a great number of studies have been conducted to identify genetic locus for BW distributed on almost all chromosomes via classic linkage maps and genome-wide association studies (GWAS) using cotton panels; over 170 QTLs for BW have been discovered (Said et al., 2015; Liu et al., 2018; Wang et al., 2019b; Zhu et al., 2021). By using F_2 and $F_{2:3}$ populations derived from an upland cotton intraspecific cross (Simian3 x TM-1), several yield-related QTLs were identified by simple sequence repeat (SSR) and random amplified polymorphic DNA (RAPD) markers, and common QTLs explaining 15.6% of the phenotypic variation (PV) were identified for BW and 100-seed weight on chromosome A09 (Yin et al., 2002). Wang et al. (2015) constructed a linkage map, which included 178 loci spanning 2016.44 cM, and a total of 19 QTLs for BW were detected on seven chromosomes; two QTLs were identified in more than two environments. In addition, a previous study involving 356 cotton accessions identified four favorable alleles for BW by a GWAS panel (Mei et al., 2013). The elucidation of the genetic architecture of BW can provide strong theoretical support for breeders to increase cotton production. However, there still exists inadequacy in previous research, such as the use of low-density linkage maps constructed based on traditional molecular markers, incomplete genetic information of the reference genome, and rough resolution of the mapping interval, resulting in candidate genes that could not be directly identified. SNP markers could be more effectively to explore the genetic structure in important agronomic traits in biparental map-based cloning and association analysis based on their highly polymorphism, wide distribution, and low research costs (Van Tassell et al., 2008; Ganal et al., 2009). Along with the reduction in high-throughput sequencing costs, a great quantity of SNP markers has been extensive development (Michael et al., 2018; Sun et al., 2020), leading to more candidate genes can be identified by QTL mapping and GWAS through SNP markers (Zhou et al., 2020; Li et al., 2021). In recent years, candidate genes for yield component traits in cotton have been wide-ranging explored in genetic studies with SNP markers rather than traditional molecular markers. For example, Zhang et al. (2016) constructed a high-density genetic map containing 5,521 SNP markers developed with a recombinant inbred line (RIL) population in 11 environments, and 344 candidate genes for BW were annotated. In addition, Fang et al. (2017) employed whole-genome resequencing using 1,871,401 high-quality SNP markers in 258 diverse accessions and discovered that the candidate gene *Gh_D08G0312* may be a key gene determining cotton yield. Moreover, two candidate genes associated with lint percentage were uncovered using 276 upland cotton accessions with 10,660 SNPs in multiple environments; these genes were highly expressed during ovule and fiber development, indicating that they may play important roles in influencing LP (Song et al., 2019). Although QTLs for yield component traits have been extensively explored in upland cotton, compared to those in important crops such as rice and maize, few candidate genes have been identified.

For this study, to gain better insight into the genetic basics of BW, specific locus amplified fragment sequencing (SLAF-seq) was taken as for whole-genome identification of SNPs and InDels in a natural population. PV for BW in nine environments was evaluated across four representative agroecological regions. In addition, several QTLs and candidate genes were further identified by a GWAS. This study provides information regarding a valuable cotton germplasm potentially useful for MAS in cotton breeding practice for raising yield in upland cotton.

MATERIALS AND METHODS

GWAS Population and Field Experiments

A total of 290 elite upland cotton accessions were obtained from CRICAAS (http://www.cricaas.com.cn/). Among these accessions, 263 (90.7%) representative cultivars were collected from four major cotton production regions of China: Northern-Specific Early-Maturity region (NSER), Yellow River region (YRR), Yangtze River region (YZRR), and Northwest Inland region (NIR). The remaining 27 (9.3%) cultivars were introduced from six different countries (USA, Azerbaijan, Israel, Kyrgyzstan, Tajikistan, and Uzbekistan). Complete GWAS population material of each accession is shown in **Supplementary Table S1**.

A natural population of 290 upland cotton accessions was planted at Anyang (36 08'N, 114 48'E) in three consecutive years (2014, 2015, and 2016) (E1: Anyang-2014, E2: Anyang-2015, and E3: Anyang-2016); Shihezi (44 31'N, 86 01'E) in three consecutive years (2014, 2015, and 2016) (E4: Shihezi-2014, E5: Shihezi-2015, and E6: Shihezi-2016); Huanggang (30 57'N, 114 92'E) in 2 years (2016 and 2021) (E7: Huanggang-2016 and E8: Huanggang-2021); and Sanya (18 36'N, 109 17'E) from 2020 to 2021 (E9: Sanya-2020-2021). Each environment was conducted with a randomized complete block for three replications.

Phenotyping and Statistical Analysis of BW

In total, 20 mature cotton bolls were randomly harvested from the middle branches and dried under sunlight for 2 days in each line. The phenotypic data from all the environments were analyzed with the base packages of R software (version: 3.5.0), and the correlation analysis results were exhibited with the "corrplot" (Wei et al., 2017). The broad-sense heritability (H^2) of BW progressed with the "sommer" (Covarrubias-Pazaran, 2016). In addition, the BLUP value of boll weight in the nine environments for the GWAS analyses was conducted by the "lme4" (Bates et al., 2014).

Genome Sequencing and Variation Detection

We collected young leaves at seedling stage of each line for genotyping. The SLAF-seq libraries were constructed for each accession based on the restriction enzymes *Rsa* I and *Hae* III (New England Biolabs, NEB). All accessions were genotyped with the Illumina HiSeq2500 platform. The detailed protocols used for library preparation and sequencing using the SLAF strategy have been described previously (Li et al., 2017). The quality control process was employed by Trimmomatic (version: 0.32) (Bolger et al., 2014), and then, the filter reads were aligned to reference genomes of the three upland cotton accessions ("TM-1," "CRI24," and "NDM8") by using BWA (version: 0.7.17) (Li and Durbin, 2009; Yu et al., 2021). The high-quality SNPs and InDels were detected using Genome Analysis Toolkit software (version: 3.8) (McKenna et al., 2010).

GWAS and Genetic Diversity Analysis

For GWAS analysis, we first filtered the SNPs and InDels with a minor allele frequency (MAF) less than 0.05 and a missing rate greater than 80%. Second, population structure was calculated as the covariate to reduce false positives (**Supplementary Figure S1**). Finally, the linear mixed mode in GEMMA (version: 0.98.3) (Zhou and Stephens, 2012) was used for discovering the significant locus by high-quality markers and BW values from each individual environment. The $-\log_{10}(P)$ value was 4.43, which was used as 1/n (n = total number of SNPs and InDels in the GWAS panel) according to the Bonferroni-corrected method. The phenotypic variation explained (PVE) of each marker was calculated by the formula as follows: $\mathrm{PVE} = [2\beta^2 \times \mathrm{MAF} \times (1 - \mathrm{MAF})] / [2\beta^2 \times \mathrm{MAF} \times (1 - \mathrm{MAF}) + ((\mathrm{se}(\beta))^2 \times 2 \times \mathrm{N} \times \mathrm{MAF} \times (1–\mathrm{MAF}))]$, where β and MAF were obtained by the GEMMA software, and N represented the sample size according to previous reports (Shim et al., 2015). The R package "qqman" was used to generate Manhattan plots (Turner, 2014). The 290 accessions were split into three populations based on the release years, including cultivars released before the 1980s, cultivars bred within the 1980s−2000s, and cultivars bred after the 2000s; VCFtools (version: 0.1.16) was used to estimate nucleotide diversity (π) (Danecek et al., 2011) in the three populations. LD block analysis was conducted with the "LDheatmap" (Shin et al., 2006) to find existing LD blocks.

Haplotype Analysis and Candidate Gene Identification

Haplotype analysis of associated markers on chromosomes A08 and D13 was conducted based on the phenotypic values and genotype data, and box plots were created using the R package "ggplot2" (Wickham, 2011). Candidate BW-related genes were identified and annotated on the basis of the "TM-1" genome released from COTTONGENE (https://www.cottongen.org/), which was in the upstream and downstream of 200 kb regions by significant markers according to previous reports (Su et al., 2018; Wang et al., 2019a). GO enrichment was performed on the agriGO to identify the enriched pathways by using default parameters (Tian et al., 2017).

Gene Expression Level Analysis

The expression patterns in *G. hirsutum* L. "TM-1" and "CRI12" at the ovule development stage (10 days post-anthesis (DPA), 20 DPA, 30 DPA, and 40 DPA) were analyzed using the published RNA-seq dataset PRJNA248163 (Fang et al., 2017). The TPM values were determined using GFOLD software (version: 1.1.4) (Feng et al., 2012). We further performed qRT-PCR analysis. All gene-specific primers used in this study were designed using Primer3 (version: 0.4.0); they are listed in **Supplementary Table S2**. Seeds of upland cotton (G. *hirsutum* cv. "TM-1" and "CRI16") were planted at Zhejiang A&F University in Hangzhou. Flowers were tagged on the day of anthesis. We collected bolls at 0, 5, 15, 20, and 25 DPA, and then, the young seeds with fibers were stripped of hulls, frozen in liquid nitrogen, and stored at -80^{circ}C. Total RNA was extracted from the frozen 0, 5, 15, 20, and 25 DPA fibers and ovule using the MolPure® Plant Plus RNA Kit (Yeasen, Shanghai, China), and cDNA was synthesized using the MonScript™ RTIII Super Mix with dsDNase (Monad, Shanghai, China). Then, real-time PCR was performed to identify transcript levels using LightCycler 480 II PCR System (Mannheim, Germany) and MonAmp™ ChemoHS qPCR Mix (Monad, Shanghai, China). The $2^{-\Delta\Delta\mathrm{CT}}$ method was applied to analyze the gene transcript abundance with three biological replicates (Livak and Schmittgen, 2001). Data visualization for qRT-PCR and RNA-seq was performed using custom R scripts.

RESULTS

Detection of SNPs and InDels in Cotton Genome

A total of 290 cotton accessions (**Supplementary Table S1**) were selected from a wide global distribution, spanning over

100 years of cotton breeding, and genotyped using the SLAF-seq approach (**Figure 1**). To identify high-quality SNPs and InDels, we compared the mapping rates across seven high-quality published reference genomes from multiple research communities (Yu et al., 2014; Hu et al., 2019; Wang et al., 2019a; Yang et al., 2019; Chen et al., 2020; Huang et al., 2020; Ma et al., 2021). The number of SLAF reads with mapping rates ranging from 98.62 to 98.93% revealed no evidence of a significant difference, while HAU_v1 showed the largest number of high-quality SNPs and InDels (**Supplementary Table S3**). Thus, we selected HAU_v1 as a reference for further GWAS. A final set of 25,169 SNPs and 2,315 InDels were obtained with a MAF greater than 0.05 and missing data less than 20% in GWAS population. The mean marker density was one per 80.3 kb in the At subgenomes and one per 81.8 kb in the Dt subgenomes. Moreover, chromosome A06 possesses the highest number of markers (3,003 SNPs and 178 InDels), followed by chromosome A08 (2,827 SNPs and 189 InDels), and the smallest number of markers was observed on chromosome D03 (403 SNPs and 58 InDels) (**Supplementary Figure S2**).

PV of BW

The BW of 290 upland cotton accessions in nine environments followed an approximately normal distribution according to Shapiro–Wilk tests (**Table 1**). The frequency distributions of BW in the natural population are summarized in **Figure 2A**. The lowest average BW was 3.08 g in E7, and the highest average BW was 8.21 g in E6, with an average variation from 4.16 $\pm$ 0.44 to 6.48 $\pm$ 0.57 across the nine environments, suggesting extensive PV in the association panel (**Table 1**). The correlation analysis for BW exhibited relatively high positive correlations between environments ($P < 0.001$), with Pearson's correlation coefficients ranging from 0.26 to 0.75 (**Figure 2B**). On the contrary, a two-way ANOVA showed that genotypic variance (G) and the genotype-by-environment variance (G $\times$ E) had significant effects on BW ($P < 0.001$). This finding confirmed that a large number of genetic variations existed in the natural population. The H^2 for BW was calculated as 69.65%, indicating that BW was mainly affected by the genotype, which was suitable for making further efforts association analysis (**Supplementary Table S4**).

GWAS of BW in Upland Cotton

A GWAS of boll weight was performed with a linear mixed model (LMM) (**Figures 3A,B** and **Supplementary Figures S3, S4**). In total, 19 significant elite alleles with 16 SNPs and three InDels were identified on six chromosomes (A06, A07, A08, D01, D07, and D13) across nine individual environments and BW-BLUP values. Each allele explained 5.58 to 10.95% of the PV, and the -$\log_{10}(P)$ values ranged from 4.53 to 6.13 (**Table 2**). A total of six loci were identified in at least two environments, and two major QTLs flanked by four alleles (rsA08_30171616, rsD13_60955253, rsD13_60955261, and rsD13_60955462) were further associated with BW-BLUP values (**Table 2**). Among them, one QTL significantly associated with a SNP (-$\log_{10}(P) =$ 5.04) on chromosome A08 explained 9.38% of the PV. Notably, another major QTL region on chromosome D13 (60,820,223–60,955,462) was stably detected in six environments, and the BW-BLUP values were based on two SNPs and an InDel. The PV explained and -$\log_{10}(P)$ values ranged from 10.32 to 10.95% and 6.06 to 6.13, respectively.

Analysis of Candidate Genes Associated With BW

Potential candidate genes linked to 19 significant BW-associated markers were extracted based on the "TM-1" reference genome (Wang et al., 2019a). A total of 225 candidate genes were identified for BW, with most genes distributed on chromosome D13 and only one candidate gene located on chromosome A08 within the 400 kb genome region (**Supplementary Table S5**). Then, we identified orthologs for 225 candidate genes based on sequence similarity analysis by comparing the candidate genes to the *Arabidopsis thaliana* reference genome, which included 215 annotated genes and 10 novel genes (**Supplementary Table S5**). Furthermore, the expression levels of the 225 genes exhibited extensive variation among different cotton tissues representing vegetative growth processes, ovule developmental stages, and the primary fiber developmental stages of initiation, elongation, and secondary wall biosynthesis. The expression patterns of candidate genes were categorized into three groups, referred to here as lineages I, II, and III, based on similarities among the expression profiles (**Figure 3C**). Gene Ontology (GO) analysis found that a large proportion of genes (33.22%) had unknown functions, but most of the candidate genes were involved in metabolic processes (42.68%), catalytic activity (38.85%), cellular processes (38.22%), or single-organism processes (24.20%) (**Figure 3D**). For example, *Ghir_D13G021550* (*PLA2-BETA*) has been reported to be involved in pollen development, germination, and stomatal opening in response to light (Kim et al., 2011). Orthologs of *Ghir_A07G004250* (*AT4G32280.1*) have been reported to be involved in the regulation of indoleacetic acid (IAA) signaling (Shimizu et al., 2016) and have ovule-specific expression at 0 DPA and 1 DPA (**Supplementary Figure S5**). In addition, six genes in the Dt subgenome (*Ghir_D01G001790*, *Ghir_D13G021810*, *Ghir_D13G022780*, *Ghir_D13G023170*, *Ghir_D13G023060*, and *Ghir_D13G023090*) were shown to be involved in response to stimulus, which is consistent with previous reports (Liu et al., 2012; Su et al., 2020). In addition, some genes were involved in cellular component organization, organelle part, biological regulation, and cell part, with proportions ranging from 3.18 to 13.38% (**Figure 3C**). Specifically, *Ghir_D13G023010* (*RHIP1*) encodes a protein predicted to have a three-stranded helical structure, which has been previously shown to modulate early seedling development in *Arabidopsis* (Huang et al., 2015).

Two Candidate Genes Pleiotropically Increase BW in Cotton Accessions

Previous studies have indicated that QTLs for BW were widely distributed on all the chromosomes of cotton, but few QTLs mapped to chromosome A08 (Said et al., 2015; Li et al., 2016; Zhang et al., 2016). In this study, a novel QTL with a significant SNP (rsA08_30171616) on chromosome A08 exhibited the

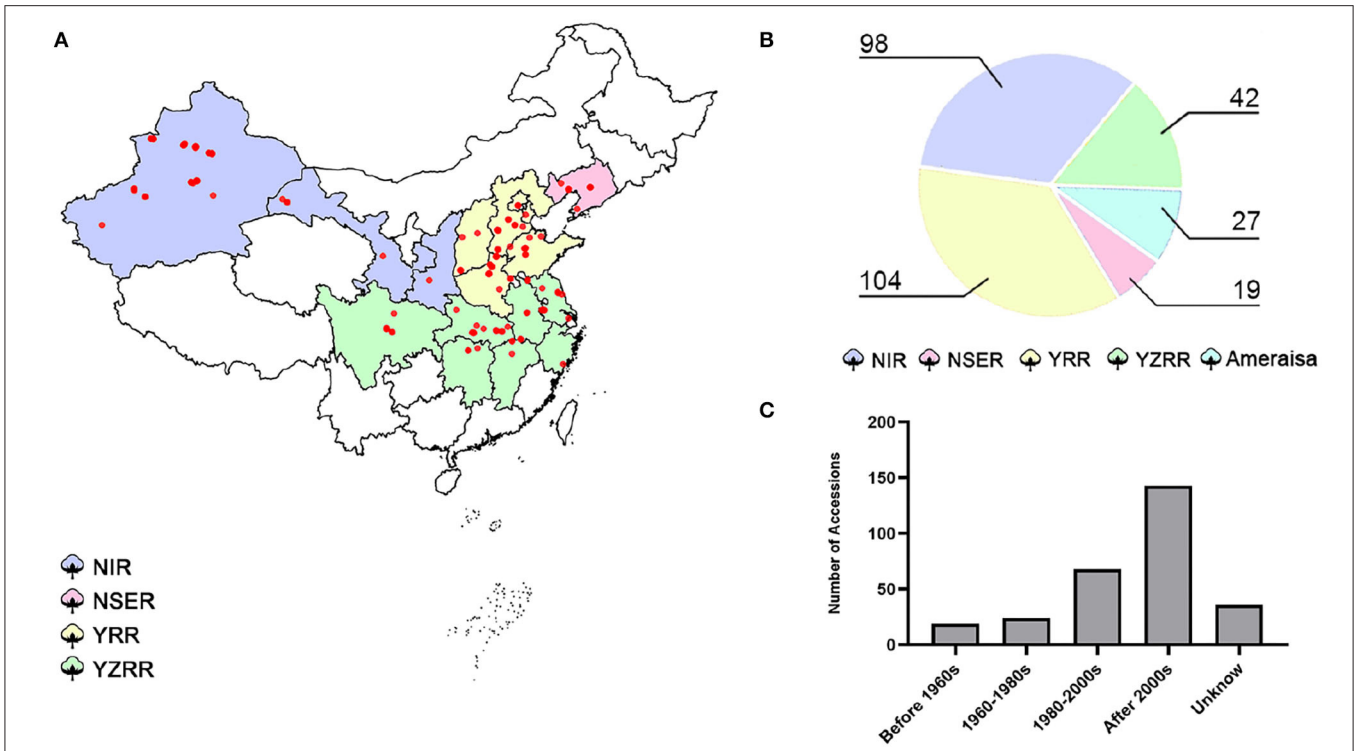


FIGURE 1 | Map of the 290 cotton accessions. **(A)** Geographic distribution of the natural population; each accession is represented by a dot. **(B)** Pie chart of the proportions of diverse cotton-growing areas in 290 accessions. NIR: Northwest Inland region in China; NSER: Northern-Specific Early-Maturity region; YRR: Yellow River region; YZRR: Yangtze River region; and Amerasian: 27 accessions primarily introduced from six different countries (USA, Azerbaijan, Israel, Kyrgyzstan, Tajikistan, and Uzbekistan). **(C)** Breeding stage distribution of the GWAS panel; Unknown: accessions that were not found among the pedigrees.

TABLE 1 | Phenotypic variation of BW in the natural populations.

Environment	Minimum	Maximum	Mean	SD	Shapiro–Wilk *P* value
E1 (Anyang-2014)	3.43	7.61	5.44	0.73	0.86
E2 (Anyang-2015)	3.30	6.93	5.35	0.60	0.65
E3 (Anyang-2016)	3.42	7.81	5.72	0.63	0.05
E4 (Shihezi-2014)	3.87	6.94	5.55	0.45	0.00
E5 (Shihezi-2015)	4.03	7.15	5.57	0.47	0.09
E6 (Shihezi-2016)	3.97	8.21	6.48	0.57	0.01
E7 (Huanggang-2016)	3.08	5.38	4.16	0.44	0.40
E8 (Huanggang-2021)	3.22	6.48	4.76	0.57	0.63
E8 (Sanya-2020-2021)	3.24	7.07	5.11	0.58	0.13

strongest association with BW, explaining 9.38% of the PV in two environments and the BW-BLUP (**Figure 4A**). This SNP has two haplotypes AA and GG, which led to the accessions carrying the GG haplotype having a significantly lower BW than those carrying the AA haplotype in nine environments ($P < 0.05$) (**Figure 4B**). In addition, to gain insight into the geographic distribution of the favorable haplotype (AA) for rsA08_30171616, the 290 cotton accessions were divided into five groups: NIR, NSER, YRR, YZRR, and Amerasian. NIR and YRR had a high proportion of the lines (**Figure 1B**) and showed an extraordinarily low AA frequency (**Figure 4C**), while the lines obtained from YZRR and Amerasian had a relatively high frequency of the favorable haplotype (>20%). We further performed an LD analysis of the significant SNP rsA08_30171616, and only one gene, *Ghir_A08G009110,* in the LD block was found in this region (**Figure 4A**). The quantitative reverse-transcription PCR (qRT-PCR) analysis and RNA-seq data showed that *Ghir_A08G009110* had higher expression levels in "TM-1" (BW = 6.18 ±0.83 g) carrying the AA allele than in "CRI12" (BW = 5.28 ±0.59 g) and "CRI16" (BW = 5.08 ±0.97 g) with GG allele during ovule development stage (**Figures 4D,E**). Through the above empirical results, we inferred

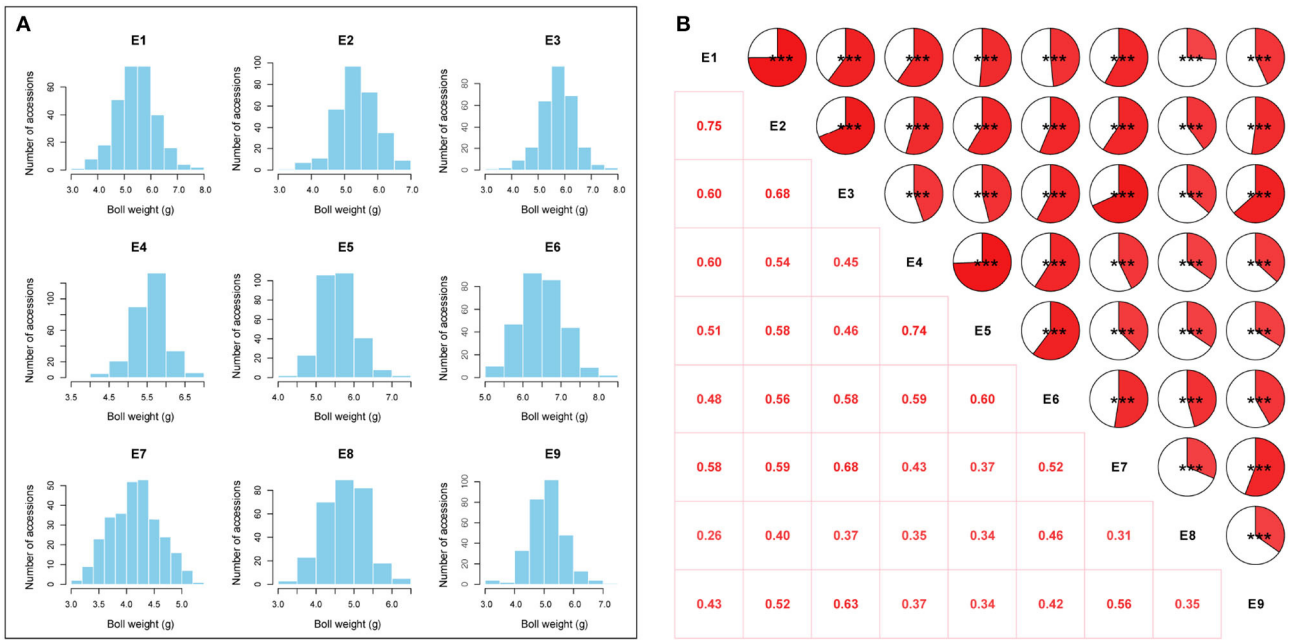


FIGURE 2 | Phenotypic variation analysis of boll weight. **(A)** Distributions of the mean values for boll weight in nine environments (E1: Anyang-2014, E2: Anyang-2015, E3: Anyang-2016, E4: Shihezi-2014, E5: Shihezi-2015, E6: Shihezi-2016, E7: Huanggang-2016, E8: Huanggang-2021, and E9: Sanya-2020-2021). **(B)** Correlation analysis of boll weight in nine environments (***$P < 0.001$, **$P < 0.01$, and *$P < 0.05$).

that *Ghir_A08G009110* on chromosome A08 has potential role responsible for improving BW and may be beneficial to cotton breeding.

We then focused on a stable QTL on chromosome D13 (**Figure 5A**). Two SNPs and one InDel in this interval were stably associated with BW in six environments and with BW-BLUP, which could explain the relatively high PV from 10.32 to 10.95% (**Table 2**). Notably, three genes (*Ghir_D13G023000*, *Ghir_D13G023010,* and *Ghir_D13G023020*) were observed and tightly linked within the candidate region (**Figure 5B**). Furthermore, we found that the genetic diversity of this interval decreased with the breeding period; cotton cultivars released before the 1980s were dramatically more diverse than the cultivars bred in the 1980–2000s, and the cultivars bred after the 2000s showed the lowest diversity. These three elite alleles generated two haplotypes (HapA and HapB) in this LD block. Among them, rsD13_60955462 was located in the 3' UTR of *Ghir_D13G023010*. Varieties carrying HapB exhibited a higher average BW than those carrying HapA (**Figure 5C**). The RNA-seq data showed that *Ghir_D13G023010* had higher expression abundance level in the low-BW variety "CRI12" than in the high-BW variety "TM-1" compared with the other two genes during ovule development from 10 to 40 DPA (**Figure 5D**). The qRT-PCR analysis also showed that *Ghir_D13G023010* had higher expression levels in low-BW variety "CRI16" than in the high-BW variety "TM-1" during ovule development (**Supplementary Figure S6**). Thus, we inferred that *Ghir_D13G023010* is a novel gene that influences BW in cotton by negative regulation.

DISCUSSION

Accurate Identification of SNPs and InDels

GWAS has become a commonly used method to identify elite allelic variation and candidate genes for important agronomic traits in cotton breeding and improvement (Fang et al., 2017; Wang et al., 2017; Ma et al., 2018). However, accurate genome sequence information enables the exploration and utilization of key genes that control important agronomic traits. It has been over 10 years since the first cotton genome sequence was published (Paterson et al., 2012; Wang et al., 2012). Since then, the number of cotton genomes sequenced has increased continually via multiple research studies due to the improvement in sequencing technologies in terms of cost, accuracy, and speed. The high rate at which genome sequences are becoming available is due to the development of next-generation sequencing (NGS), third-generation sequencing (TGS), and chromosome-scale scaffolding tools (Bio-Nano and Hi-C), with contig N50 values ranging from 0.11 Mb to 13.15 Mb in multiple upland cotton accessions ("TM-1," "NDM8," and "CRI24") (Yu et al., 2014). A previous study demonstrated that the development of different reference-quality genomes could facilitate the investigation of novel variation and found new genes that were not discovered in previous SNP/InDel-based association analyses for important agronomic traits. For example, in maize, Tao et al. (2019) uncovered a novel causal mutation with an 8.9-kb insertion of a grain-size QTL (qHKW1) in an RIL population with the assistance of the newly assembled "SK" genome (Tao et al., 2019). In this study, to obtain accurate genetic markers, we employed a reference genome with a contig N50 greater than 100 kb

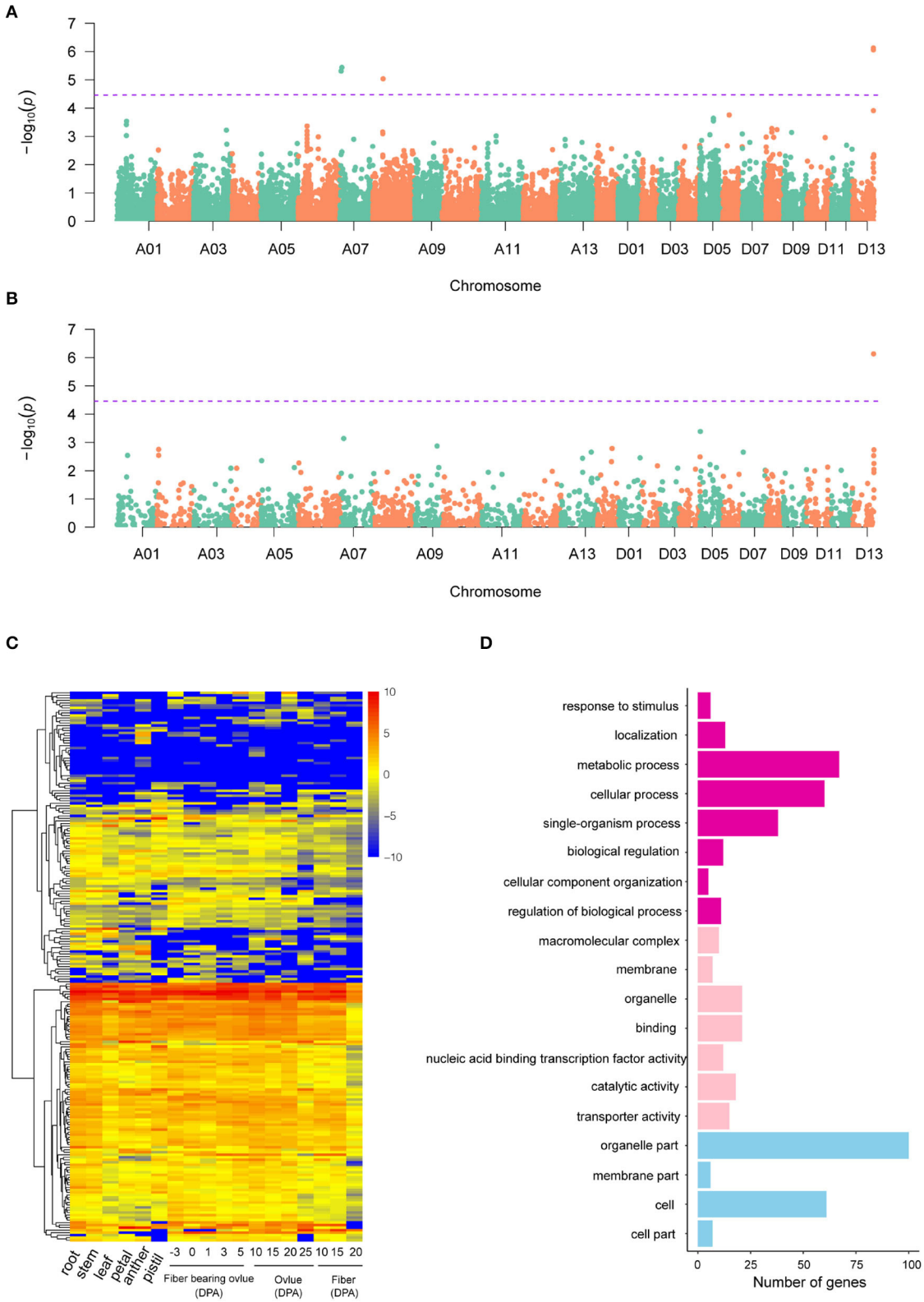


FIGURE 3 | GWAS results of SNP and InDel markers and candidate gene analysis. **(A,B)** Manhattan plots of BW-BLUP for SNPs and InDels, respectively; significant BW-associated markers are distinguished by purple lines. **(C)** Heatmap of candidate gene expression patterns in 18 cotton tissues. **(D)** GO analysis of candidate genes associated with boll weight. The chart of purple, pink, and blue represented biological process, molecular function, and cellular component, respectively.

TABLE 2 | List of significant markers (SNPs and InDels) associated with boll weight.

Marker	Marker type	Chromosome	Position	Major allele	Minor allele	*P* value	R^2	Environment
rsGhir_A06_26390257	SNP	A06	26,390,257	T	G	6.84E-06	6.67	E2
rsGhir_A06_26390265	SNP	A06	26,390,265	G	A	2.07E-05	5.92	E2
rsGhir_A06_26390284	SNP	A06	26,390,284	G	A	2.43E-05	6.04	E2
rsGhir_A06_26390468	SNP	A06	26,390,468	A	C	2.97E-05	5.58	E2
rsGhir_A06_26390491	SNP	A06	26,390,491	A	G	2.23E-05	6.09	E2
rsGhir_A06_32168831	Indel	A06	32,168,831	G	GT	2.14E-05	6.81	E8
rsGhir_A07_4798628	SNP	A07	4,798,628	G	A	4.91E-06	8.36	E7, BLUP
rsGhir_A07_6937342	SNP	A07	6,937,342	C	T	3.66E-06	8.57	BLUP
rsGhir_A07_6937395	SNP	A07	6,937,395	C	T	3.66E-06	8.38	BLUP
rsGhir_A07_9574709	SNP	A07	9,574,709	C	G	2.94E-05	7.69	E4, E5
rsGhir_A08_30171616	SNP	A08	30,171,616	A	G	9.20E-06	9.38	E6, E9, BLUP
rsGhir_D01_1229290	SNP	D01	1,229,290	A	G	8.35E-07	9.25	E8
rsGhir_D01_1229442	SNP	D01	1,229,442	T	C	1.94E-06	9.01	E8
rsGhir_D07_19492198	SNP	D07	19,492,198	G	A	2.95E-05	6.71	E1
rsGhir_D13_59526001	SNP	D13	59,526,001	G	C	2.41E-05	6.19	E4, E5
rsGhir_D13_60955253	Indel	D13	60,955,253	A	AT	7.40E-07	10.95	E1, E2, E3, E4, E5, E6, BLUP
rsGhir_D13_60955261	SNP	D13	60,955,261	G	T	7.51E-07	10.88	E1, E2, E3, E4, E5, E6, BLUP
rsGhir_D13_60955462	SNP	D13	60,955,462	A	G	8.73E-07	10.32	E1, E2, E3, E4, E5, E6, BLUP
rsGhir_D13_62059670	Indel	D13	62,059,670	GC	G	4.99E-06	6.06	E3

for SNP and InDel calling. Although there was no significant difference in mapping rate, the genome version of HAU_v1 had more high-quality SNP and InDel markers. This genome provided a genetic basis for us to find a novel BW-associated locus. It is worth noting that 73.68% of associated BW loci could be detected via the comparison of multiple genomes. Five loci (rsGhir_A06_26390257, rsGhir_A06_26390265, rsGhir_A06_26390284, rsGhir_A06_26390468, and rsGhir_A06_26390491) on chromosome A06 are unique to HUA and are likely due to the diversity within the species and the quality of the reference genome. Therefore, the development of multiple reference genomes would enable the integration of these resources into high-quality pangenomes and will provide a better understanding of genetic diversity and a comprehensive guiding principle for the further exploration and utilization of this diversity for cotton improvement.

Comparison of GWAS Results With Previously Reported Results

BW is an important determinant of yield and profitability in cotton and is controlled by multiple genes. Indeed, cotton breeding has constantly focused on the improvement of BW. Thus far, most QTLs for BW have been identified based on linkage analysis in the CottonQTLdb by using traditional molecular markers (Said et al., 2015). In addition, due to the limitation of traditional markers with lower levels of polymorphism and distribution density, it is difficult to attain sufficient resolution for fine map-based cloning and direct identification of candidate genes. GWAS has become a popular and powerful method to detect variants associated with major agricultural traits (Su et al., 2016, 2018; Fang et al., 2017; Wan et al., 2017; Ma et al., 2018; Zhang et al., 2019a). However, few studies have dissected the genetic basis of BW in cotton via GWAS in combination with high-throughput SNPs and diverse accessions across multiple environments in recent years, and even fewer candidate genes have been reported. In this study, 290 upland cotton accessions that were widely collected worldwide were used to conduct GWASs using high-throughput SNPs and diverse environments over multiple years. In total, 19 significant loci were identified among six different cotton chromosomes (**Table 2**), including 16 SNPs and three InDels. The identification of cotton varieties with stable yield and wide adaptation across a range of environments is one of the important objectives of modern cotton breeding programs in China. Although BW has relatively high heritability (69.65%), still lower than other agronomic traits in cotton, including oil content (96.6%) (Zhao et al., 2019), fiber length (81%) (Zhang et al., 2019a), flowering time (79%) (Li et al., 2021), and resulting, only a few stable QTLs were identified in 19 significant loci. This indicates that the remaining QTLs are affected by environment or genotype-by-environment. Meanwhile, phenotypic variation analysis found the BW of cotton grown in Huanggang is lower than that in Shihezi and Anyang. It is mainly caused by the high temperature in summer and the excessive rainfall in the later stage of cotton growth at the Yangtze River basin, leading to the correlation coefficient of E7 and E8 with other environments (E1–E6, E9) being low. Furthermore, although the SNPs obtained by SLAF-seq technology can well cover the whole genome of cotton, it must be admitted that there are indeed fewer stable QTLs than those obtained based on resequencing of GWAS. Therefore, we could employ resequencing for GWAS analysis in further to obtain more reliable QTLs for BW. To screen QTLs with high precision, high stability, and small confidence intervals for MAS and

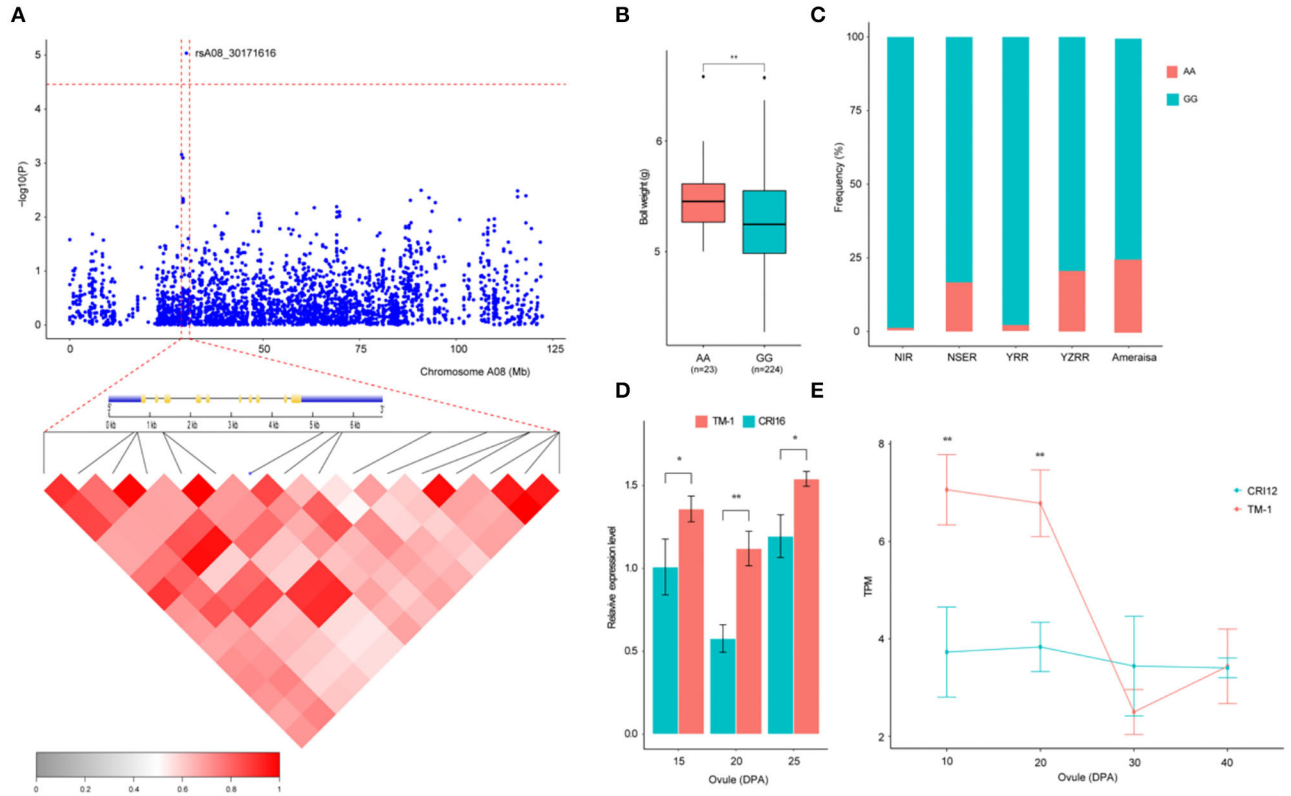


FIGURE 4 | Variation analysis of the boll weight-related gene *Ghir_A08G009110* on candidate region. **(A)** Local Manhattan plots for BW-related genes on chromosome A08 and LD heatmap for the candidate region within the peak region of rsA08_30171616, including the exon–intron structure of *Ghir_A08G009110*. **(B)** Box plots for BW between the two haplotypes mentioned above (** $P < 0.01$, * $P < 0.05$). **(C)** Differentiation of the genetic diversity distribution of the favorable haplotype for rsA08_30171616 in five geographic areas. **(D)** Expression level analysis of *Ghir_A08G009110* between "TM-1" (red) and "CRI16" (green) during ovule developmental stages (15, 20, and 25 DPA) by qRT-PCR (** $P < 0.01$, * $P < 0.05$). **(E)** Expression abundance analysis of *Ghir_A08G009110* between "TM-1" (red) and "CRI12" (green) during ovule developmental stages (10, 20, 30, and 40 DPA) by RNA-seq (** $P < 0.01$, * $P < 0.05$).

gene cloning, we further compared our results with published studies based on SNP and SSR markers (Said et al., 2015). Eleven reliable and significant markers located on chromosomes A07, D01, D07, and D13 were reported in previous studies. Three SNPs (rsGhir_A07_6937342, rsGhir_A07_6937395, and rsGhir_A07_9574709) on chromosome A07 overlapped with the region i49554Gh, which was named *qGhLP-c7* by Huang et al. (2017). rsGhir_D01_1229290, rsGhir_D01_1229442, rsGhir_D07_19492198, and rsGhir_D13_59526001 on chromosomes D01, D07, and D13 were mapped to regions adjacent to TM47842_TM47844, TM64105, and TM82005, respectively, as reported by Zhu et al. (2021). Most importantly, we also discovered a major QTL that was detected in multiple environments and with multiple BW-BLUP values and that could explain more than 10% of the observed PV. Furthermore, this region also overlapped with TM82122, as described by Liu et al. (2018), and narrowed the candidate region to 60.82–60.95 Mb on chromosome D13 containing three candidate genes. To date, few QTLs for BW on chromosome A08 have been identified in previous studies. Interestingly, a tightly linked region flanked by rsGhir_A08_30171616 on chromosome A08 was detected in two environments and with BW-BLUP values. This region contained only one gene (*Ghir_A08G009110*), which was not reported to control the boll weight of cotton in previous studies. Thus, these stable QTLs that are responsible for BW may have a significant effect on further yield improvement in cotton with appropriate BW.

Candidate Genes Related to BW

It is known that BW is a complex quantitative trait controlled by many genes. Here, based on the association analysis, candidate gene expression analysis, and genetic diversity analysis of BW in 290 diverse cultivated upland cotton accessions, *Ghir_A08G009110* and *Ghir_D13G023010* on chromosomes A08 and D13, respectively, were identified as candidate genes for QTLs controlling BW in a natural population. Interestingly, *Ghir_A08G009110,* a unique candidate gene within the strong LD region 200 kb upstream and downstream of rsA08_30171616, encodes a protein containing ankyrin and DHHC-CRD domains in *A. thaliana* and is involved in root hair cell growth (Wan et al., 2017). We also discovered that the candidate gene *Ghir_A08G009110* in this region was highly expressed during the early stage of ovule development in the high-BW variety (**Figures 4D,E**). In addition, *Ghir_A08G009110* showed excellent

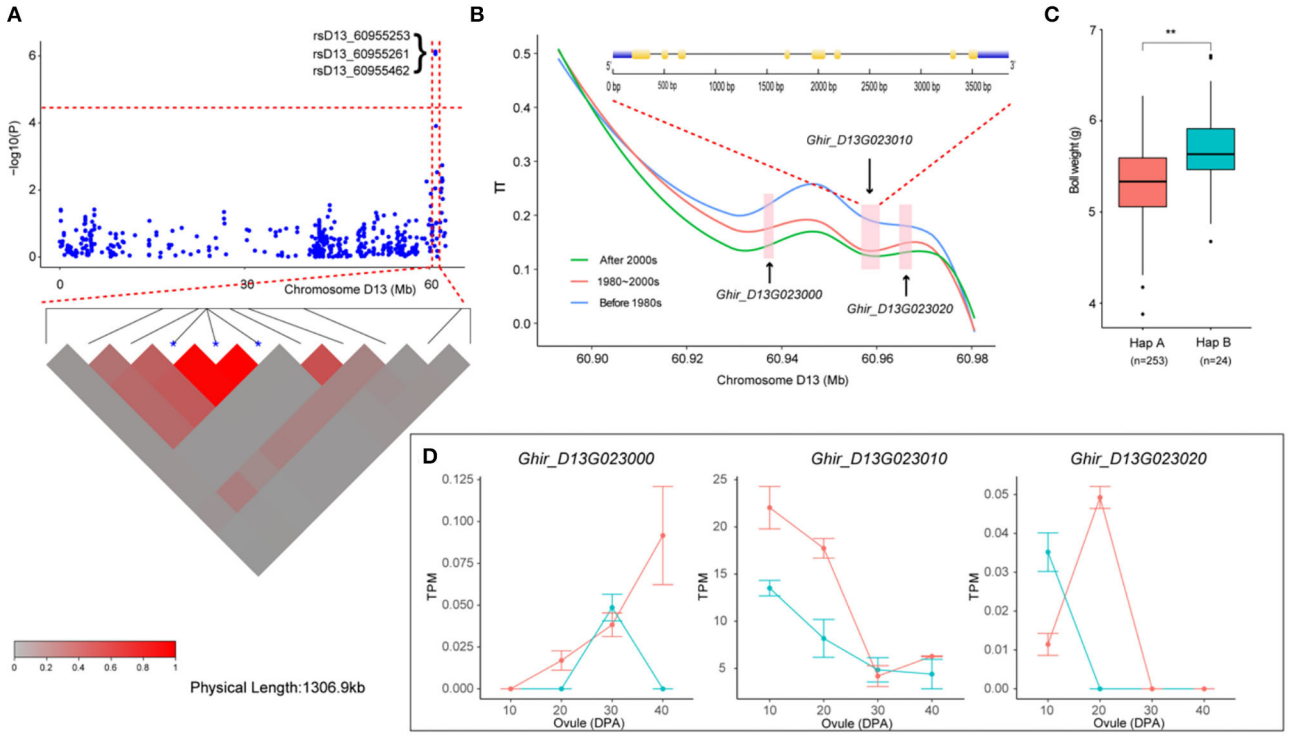


FIGURE 5 | Variation analysis of the boll weight-related gene *Ghir_D13G023010* on candidate region. **(A)** Local Manhattan plots for BW-related genes on chromosome D13 and LD heatmap for the candidate region within the peak region of rsD13_60955253, rsD13_60955261, and rsD13_60955462. **(B)** Genetic diversity across the three populations and exon–intron structure of *Ghir_D13G023010*. **(C)** Box plots for BW of the two haplotypes mentioned above (** $P < 0.01$, * $P < 0.05$). **(D)** Expression abundance analysis of *Ghir_D13G023010* between "TM-1" (green) and "CRI12" (red) during ovule developmental stages (0, 10, 20, 30, and 40 DPA) by RNA-seq.

potential for improving cotton yield and was not associated with other important agronomic traits in a previous QTL analysis (Said et al., 2015). Therefore, it is reasonable to postulate that *Ghir_A08G009110* is a new candidate gene for influencing BW in cotton. However, cotton accessions with rsA08_30171616-A had a much higher allele frequency than those with the potential superior alleles for *Ghir_A08G009110* in NESR and Amerasian, including accessions with a higher genomic proportion of some early core accessions. YRR and NIR, which contained mostly modern accessions, had a lower proportion of superior alleles for *Ghir_A08G009110* (rsA08_30171616-G). Thus, it is possible that the locus rsA08_30171616-A associated with excellent BW was screened out during the breeding process, so it is necessary to use rsA08_30171616-A as a tagging SNP in MAS of cotton lines to further improve yield.

Seed weight is also selected for during crop domestication, and understanding the genetic and molecular mechanisms controlling seed size has become an important research topic in plant science (Lin et al., 2014). Cotton is the largest economically important crop in the world, and breeders have expended a great deal of effort in improving the yield of cotton during long-term selection. Recently, *Ghir_D03G011310* was considered a candidate gene underlying the natural variation in cotton that controls early maturity in a natural population during long-term artificial selection, as stated in our previous report (Li et al., 2021). Furthermore, Wang et al. (2017) found many genes involved in the domestication of white fiber. However, the genes underlying the natural variation in cotton BW are still largely unknown. Here, we compared the genetic diversity of the region from 60.91 to 60.97 Mb on chromosome D13 containing *Ghir_D13G023010* in different breeding periods, and it was found that cultivars bred after the 2000s had lower genetic diversity than cultivars released before the 1980s and cultivars released in the 1980s−2000s. This result implied that with the continuous increase in cotton yield during the breeding process, this region is associated with artificial selection and with the increase in the BW of cotton. In addition, *Ghir_D13G023010* was the only *RHIP1* homolog in the cotton genome and was the best match with *Ghir_D13G023010* in the *Arabidopsis* genome. *RHIP1* is an uncharacterized conserved protein that participates in sugar signaling and plays significant role in negatively regulating seeding development (Huang et al., 2015). In particular, *Ghir_D13G023010* has highly expression abundance in the low-BW variety than in the high-BW variety (**Figure 5D** and **Supplementary Figure S6**). From the above results, we inferred that *Ghir_A08G009110* and *Ghir_D13G023010* were major candidate genes that may play an important role in influencing cotton boll weight.

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

SY and ZF supervised the study and were involved in writing—reviewing and editing. LL was involved in funding acquisition, investigation, visualization, and writing—original draft. MT conceptualized the study and was involved in data curation and formal analysis. QL and CH designed software and data curation. DS validated the study. ZJ, GL, SZ, and GZ investigated the study. YZ designed software and visualized the study. All authors contributed to the article and approved the submitted version.

FUNDING

This research was sponsored by the Program for Research and Development of Zhejiang A&F University (2021LFR005).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.929168/full#supplementary-material

REFERENCES

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Chen, Z. J., Scheffler, B. E., Dennis, E., Triplett, B. A., Zhang, T., Guo, W., et al. (2007). Toward sequencing cotton (*Gossypium*) genomes. *Plant Physiol.* 145, 1303–1310. doi: 10.1104/pp.107.107672

Chen, Z. J., Sreedasyam, A., Ando, A., Song, Q., De Santiago, L. M., Hulse-Kemp, A. M., et al. (2020). Genomic diversifications of five Gossypium allopolyploid species and their impact on cotton improvement. *Nat. Genet.* 52, 525–533. doi: 10.1038/s41588-020-0614-5

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

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., Depristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics.* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Fan, L., Wang, L., Wang, X., Zhang, H., Zhu, Y., Guo, J., et al. (2018). A high-density genetic map of extra-long staple cotton (*Gossypium barbadense*) constructed using genotyping-by-sequencing based single nucleotide polymorphic markers and identification of fiber traits-related QTL in a recombinant inbred line population. *BMC Genomics.* 19, 489. doi: 10.1186/s12864-018-4890-8

Fang, L., Wang, Q., Hu, Y., Jia, Y., Chen, J., Liu, B., et al. (2017). Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. *Nat. Genet.* 49, 1089–1098. doi: 10.1038/ng.3887

Feng, J., Meyer, C. A., Wang, Q., Liu, J. S., Shirley Liu, X., and Zhang, Y. (2012). GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. *Bioinformatics.* 28, 2782–2788. doi: 10.1093/bioinformatics/bts515

Ganal, M. W., Altmann, T., and Roder, M. S. (2009). SNP identification in crop plants. *Curr. Opin. Plant Biol.* 12, 211–217. doi: 10.1016/j.pbi.2008.12.009

Gu, Q., Ke, H., Liu, Z., Lv, X., Sun, Z., Zhang, M., et al. (2020). A high-density genetic map and multiple environmental tests reveal novel quantitative trait loci and candidate genes for fibre quality and yield in cotton. *Theor. Appl. Genet.* 133, 3395–3408. doi: 10.1007/s00122-020-03676-z

Hu, Y., Chen, J., Fang, L., Zhang, Z., Ma, W., Niu, Y., et al. (2019). Gossypium barbadense and Gossypium hirsutum genomes provide insights into the origin and evolution of allotetraploid cotton. *Nat. Genet.* 51, 739–748. doi: 10.1038/s41588-019-0371-5

Huang, C., Nie, X., Shen, C., You, C., Li, W., Zhao, W., et al. (2017). Population structure and genetic basis of the agronomic traits of upland cotton in China revealed by a genome-wide association study using high-density SNPs. *Plant Biotechnol. J.* 15, 1374–1386. doi: 10.1111/pbi.12722

Huang, G., Wu, Z., Percy, R. G., Bai, M., Li, Y., Frelichowski, J. E., et al. (2020). Genome sequence of Gossypium herbaceum and genome updates of Gossypium arboreum and Gossypium hirsutum provide insights into cotton A-genome evolution. *Nat. Genet.* 52, 516–524. doi: 10.1038/s41588-020-0607-4

Huang, J. P., Tunc-Ozdemir, M., Chang, Y., and Jones, A. M. (2015). Cooperative control between AtRGS1 and AtHXK1 in a WD40-repeat protein pathway in *Arabidopsis thaliana*. *Front. Plant Sci.* 6, 851. doi: 10.3389/fpls.2015.00851

Kim, H. J., Ok, S. H., Bahn, S. C., Jang, J., Oh, S. A., Park, S. K., et al. (2011). Endoplasmic reticulum- and Golgi-localized phospholipase A2 plays critical roles in *Arabidopsis* pollen development and germination. *Plant Cell.* 23, 94–110. doi: 10.1105/tpc.110.074799

Li, C., Dong, Y., Zhao, T., Li, L., Li, C., Yu, E., et al. (2016). Genome-wide SNP linkage mapping and QTL analysis for fiber quality and yield traits in the upland cotton recombinant inbred lines population. *Front. Plant Sci.* 7, 1356. doi: 10.3389/fpls.2016.01356

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

Li, L., Zhang, C., Huang, J., Liu, Q., Wei, H., Wang, H., et al. (2021). Genomic analyses reveal the genetic basis of early maturity and identification of loci and candidate genes in upland cotton (*Gossypium hirsutum* L.). *Plant Biotechnol. J.* 19, 109–123. doi: 10.1111/pbi.13446

Li, L., Zhao, S., Su, J., Fan, S., Pang, C., Wei, H., et al. (2017). High-density genetic linkage map construction by F2 populations and QTL analysis of early-maturity traits in upland cotton (*Gossypium hirsutum* L.). *PLoS ONE.* 12, e0182918. doi: 10.1371/journal.pone.0182918

Lin, T., Zhu, G., Zhang, J., Xu, X., Yu, Q., Zheng, Z., et al. (2014). Genomic analyses provide insights into the history of tomato breeding. *Nat. Genet.* 46, 1220–1226. doi: 10.1038/ng.3117

Liu, K., Sun, J., Yao, L., and Yuan, Y. (2012). Transcriptome analysis reveals critical genes and key pathways for early cotton fiber elongation in Ligon lintless-1 mutant. *Genomics.* 100, 42–50. doi: 10.1016/j.ygeno.2012.04.007

Liu, R., Gong, J., Xiao, X., Zhang, Z., Li, J., Liu, A., et al. (2018). GWAS analysis and QTL identification of fiber quality traits and yield components in upland cotton using enriched high-density SNP markers. *Front. Plant Sci.* 9, 1067. doi: 10.3389/fpls.2018.01067

Liu, X., Zhao, B., Zheng, H. J., Hu, Y., Lu, G., Yang, C. Q., et al. (2015). *Gossypium barbadense* genome sequence provides insight into the evolution of extra-long staple fiber and specialized metabolites. *Sci. Rep.* 5, 14139. doi: 10.1038/srep14139

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods.* 25, 402–408. doi: 10.1006/meth.2001.1262

Ma, Z., He, S., Wang, X., Sun, J., Zhang, Y., Zhang, G., et al. (2018). Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. *Nat. Genet.* 50, 803–813. doi: 10.1038/s41588-018-0119-7

Ma, Z., Zhang, Y., Wu, L., Zhang, G., Sun, Z., Li, Z., et al. (2021). High-quality genome assembly and resequencing of modern cotton cultivars provide resources for crop improvement. *Nat. Genet.* 53, 1385–1391. doi: 10.1038/s41588-021-00910-2

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

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.* 8, e82193. doi: 10.1371/journal.pone.0082193

Michael, T. P., Jupe, F., Bemm, F., Motley, S. T., Sandoval, J. P., Lanz, C., et al. (2018). High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. *Nat. Commun.* 9, 541. doi: 10.1038/s41467-018-03016-2

Paterson, A. H., Wendel, J. F., Gundlach, H., Guo, H., Jenkins, J., Jin, D., et al. (2012). Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature.* 492, 423–427. doi: 10.1038/nature11798

Rajpal, V. R., Rao, S. R., and Raina, S. (2016). *Gene Pool Diversity and Crop Improvement.* Cham: Springer. doi: 10.1007/978-3-319-27096-8

Said, J. I., Knapka, J. A., Song, M., and Zhang, J. (2015). Cotton QTLdb: a cotton QTL database for QTL analysis, visualization, and comparison between *Gossypium hirsutum* and G. hirsutum x G. barbadense populations. *Mol. Genet. Genomics.* 290, 1615–1625. doi: 10.1007/s00438-015-1021-y

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

Shimizu, H., Torii, K., Araki, T., and Endo, M. (2016). Importance of epidermal clocks for regulation of hypocotyl elongation through PIF4 and IAA29. *Plant Signal. Behav.* 11, e1143999. doi: 10.1080/15592324.2016.1143999

Shin, J. H., Blay, S., Graham, J., and Mcneney, B. (2006). LDheatmap: an R Function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *J. Stat. Softw.* 16, 1–10. doi: 10.18637/jss.v016.c03

Song, C., Li, W., Pei, X., Liu, Y., Ren, Z., He, K., et al. (2019). Dissection of the genetic variation and candidate genes of lint percentage by a genome-wide association study in upland cotton. *Theor. Appl. Genet.* 132, 1991–2002. doi: 10.1007/s00122-019-03333-0

Su, J., Li, L., Zhang, C., Wang, C., Gu, L., Wang, H., et al. (2018). Genome-wide association study identified genetic variations and candidate genes for plant architecture component traits in Chinese upland cotton. *Theor. Appl. Genet.* 131, 1299–1314. doi: 10.1007/s00122-018-3079-5

Su, J., Pang, C., Wei, H., Li, L., Liang, B., Wang, C., et al. (2016). Identification of favorable SNP alleles and candidate genes for traits related to early maturity via GWAS in upland cotton. *BMC Genomics.* 17, 687. doi: 10.1186/s12864-016-2875-z

Su, X., Zhu, G., Song, X., Xu, H., Li, W., Ning, X., et al. (2020). Genome-wide association analysis reveals loci and candidate genes involved in fiber quality traits in sea island cotton (*Gossypium barbadense*). *BMC Plant Biol.* 20, 289. doi: 10.1186/s12870-020-02502-4

Sun, C., Dong, Z., Zhao, L., Ren, Y., Zhang, N., and Chen, F. (2020). The Wheat 660K SNP array demonstrates great potential for marker-assisted selection in polyploid wheat. *Plant Biotechnol. J.* 18, 1354–1360. doi: 10.1111/pbi.13361

Sunilkumar, G., Campbell, L. M., Puckhaber, L., Stipanovic, R. D., and Rathore, K. S. (2006). Engineering cottonseed for use in human nutrition by tissue-specific reduction of toxic gossypol. *Proc. Natl. Acad. Sci. U. S. A.* 103, 18054–18059. doi: 10.1073/pnas.0605389103

Tao, Y., Jordan, D. R., and Mace, E. S. (2019). Crop genomics goes beyond a single reference genome. *Trends Plant Sci.* 24, 1072–1074. doi: 10.1016/j.tplants.2019.10.001

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., et al. (2017). agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Res.* 45, W122–W129. doi: 10.1093/nar/gkx382

Turner, S. D. (2014). qqman: an R package for visualizing GWAS results using QQ and manhattan plots. *J. Open Source Softw.* 3, 731. doi: 10.21105/joss.00731

Van Tassell, C. P., Smith, T. P., Matukumalli, L. K., Taylor, J. F., Schnabel, R. D., Lawley, C. T., et al. (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat. Methods.* 5, 247–252. doi: 10.1038/nmeth.1185

Wan, Z. Y., Chai, S., Ge, F. R., Feng, Q. N., Zhang, Y., and Li, S. (2017). Arabidopsis PROTEIN S-ACYL TRANSFERASE4 mediates root hair growth. *Plant J.* 90, 249–260. doi: 10.1111/tpj.13484

Wang, H., Huang, C., Guo, H., Li, X., Zhao, W., Dai, B., et al. (2015). QTL mapping for fiber and yield traits in upland cotton under multiple environments. *PLoS ONE.* 10, e0130742. doi: 10.1371/journal.pone.0130742

Wang, K., Wang, Z., Li, F., Ye, W., Wang, J., Song, G., et al. (2012). The draft genome of a diploid cotton *Gossypium raimondii*. *Nat. Genet.* 44, 1098–1103. doi: 10.1038/ng.2371

Wang, M., Tu, L., Lin, M., Lin, Z., Wang, P., Yang, Q., et al. (2017). Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. *Nat. Genet.* 49, 579–587. doi: 10.1038/ng.3807

Wang, M., Tu, L., Yuan, D., Zhu, S.hen, C., Li, J., Liu, F., et al. (2019a). Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. *Nat. Genet.* 51, 224–229. doi: 10.1038/s41588-018-0282-x

Wang, Y., Li, G., Guo, X., Sun, R., Dong, T., Yang, Q., et al. (2019b). Dissecting the genetic architecture of seed-cotton and lint yields in Upland cotton using genome-wide association mapping. *Breed. Sci.* 69, 611–620. doi: 10.1270/jsbbs.19057

Wei, T., Simko, V., Levy, M., Xie, Y., Jin, Y., and Zemla, J. (2017). Package 'corrplot'. *Statistician* 56, e24. doi: 10.1002/mus.25583

Wendel, J. F. (1989). New World tetraploid cottons contain Old World cytoplasm. *Proc. Natl. Acad. Sci. U. S. A.* 86, 4132–4136. doi: 10.1073/pnas.86.11.4132

Wickham, H. (2011). ggplot2. *WIREs Comput. Stat.* 3, 180–185. doi: 10.1002/wics.147

Yang, Z., Ge, X., Yang, Z., Qin, W., Sun, G., Wang, Z., et al. (2019). Extensive intraspecific gene order and gene structural variations in upland cotton cultivars. *Nat. Commun.* 10, 1–13. doi: 10.1038/s41467-019-10820-x

Yin, J. M., Wu, Y. T., Zhang, J., Zhang, T. Z., Guo, W. Z., and Zhu, X. F. (2002). Tagging and mapping of QTLs controlling lint yield and yield components in upland cotton (*Gossypium hirsutum* L.) using SSR and RAPD markers. *Sheng Wu Gong Cheng Xue Bao.* 18, 162–166. doi: 10.3321/j.issn:1000-3061.2002.02.007

Yin, X., Stam, P., Kropff, M. J., and Schapendonk, A. H. C. M. (2003). Crop modeling, QTL mapping, and their complementary role in plant breeding. *Agron. J.* 95, 90. doi: 10.2134/agronj2003.0090

Yu, J., Jung, S., Cheng, C. H., Ficklin, S. P., Lee, T., Zheng, P., et al. (2014). CottonGen: a genomics, genetics and breeding database for cotton research. *Nucleic Acids Res.* 42, D1229–D1236. doi: 10.1093/nar/gkt1064

Yu, J., Jung, S., Cheng, C. H., Lee, T., Zheng, P., Buble, K., et al. (2021). CottonGen: the community database for cotton genomics, genetics, and breeding research. *Plants (Basel).* 10, 2805. doi: 10.3390/plants10122805

Zhang, C., Li, L., Liu, Q., Gu, L., Huang, J., Wei, H., et al. (2019a). Identification of loci and candidate genes responsible for fiber length in upland cotton (*Gossypium hirsutum* L.) via association mapping and linkage analyses. *Front. Plant Sci.* 10, 53. doi: 10.3389/fpls.2019.00053

Zhang, K., Kuraparthy, V., Fang, H., Zhu, L., Sood, S., and Jones, D. C. (2019b). High-density linkage map construction and QTL analyses for fiber quality, yield and morphological traits using CottonSNP63K array in upland cotton (*Gossypium hirsutum* L.). *BMC Genomics.* 20, 889. doi: 10.1186/s12864-019-6214-z

Zhang, Z., Shang, H., Shi, Y., Huang, L., Li, J., Ge, Q., et al. (2016). Construction of a high-density genetic map by specific locus amplified fragment sequencing (SLAF-seq) and its application to Quantitative Trait Loci (QTL) analysis for boll weight in upland cotton (*Gossypium hirsutum.*). *BMC Plant Biol.* 16, 79. doi: 10.1186/s12870-016-0741-4

Zhao, W., Kong, X., Yang, Y., Nie, X., and Lin, Z. (2019). Association mapping seed kernel oil content in upland cotton using genome-wide SSRs and SNPs. *Mol Breeding.* 39, 1–11. doi: 10.1007/s11032-019-1007-2

Zhou, Q., Tang, D., Huang, W., Yang, Z., Zhang, Y., Hamilton, J. P., et al. (2020). Haplotype-resolved genome analyses of a heterozygous diploid potato. *Nat. Genet.* 52, 1018–1023. doi: 10.1038/s41588-020-0699-x

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

Zhu, G., Hou, S., Song, X., Wang, X., Wang, W., Chen, Q., et al. (2021). Genome-wide association analysis reveals quantitative trait loci and candidate genes involved in yield components under multiple field

environments in cotton (*Gossypium hirsutum*). *BMC Plant Biol.* 21, 250. doi: 10.1186/s12870-021-03009-2

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

TYPE Original Research
PUBLISHED 13 October 2022
DOI 10.3389/fpls.2022.997988

OPEN ACCESS

EDITED BY
Zhenyu Jia,
University of California, Riverside, United States

REVIEWED BY
Jia Wen,
The University of North Carolina at Chapel Hill, United States
Youlu Yuan,
Cotton Research Institute (CAAS), China

*CORRESPONDENCE
Xin Chen
cx@jaas.ac.cn
Xingxing Yuan
yxx@jaas.ac.cn

SPECIALTY SECTION
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

RECEIVED 19 July 2022
ACCEPTED 15 August 2022
PUBLISHED 13 October 2022

CITATION
Liu J, Lin Y, Chen J, Yan Q, Xue C, Wu R, Chen X and Yuan X (2022) Genome-wide association studies provide genetic insights into natural variation of seed-size-related traits in mungbean.
Front. Plant Sci. 13:997988.
doi: 10.3389/fpls.2022.997988

Genome-wide association studies provide genetic insights into natural variation of seed-size-related traits in mungbean

Jinyang Liu, Yun Lin, Jingbin Chen, Qiang Yan, Chenchen Xue, Ranran Wu, Xin Chen* and Xingxing Yuan*

Institute of Industrial Crops, Jiangsu Academy of Agricultural Sciences/Jiangsu Key Laboratory for Horticultural Crop Genetic Improvement, Nanjing, China

Although mungbean (*Vigna radiata* (L.) R. Wilczek) is an important legume crop, its seed yield is relatively low. To address this issue, here 196 accessions with 3,607,508 SNP markers were used to identify quantitative trait nucleotides (QTNs), QTN-by-environment interactions (QEIs), and their candidate genes for seed length (SL), seed width, and 100-seed weight (HSW) in two environments. As a result, 98 QTNs and 20 QEIs were identified using 3VmrMLM, while 95, >10,000, and 15 QTNs were identified using EMMAX, GEMMA, and CMLM, respectively. Among 809 genes around these QTNs, 12 were homologous to known seed-development genes in rice and *Arabidopsis thaliana*, in which 10, 2, 1, and 0 genes were found, respectively, by the above four methods to be associated with the three traits, such as *VrEmp24/25* for SL and *VrKIX8* for HSW. Eight of the 12 genes were significantly differentially expressed between two large-seed and two small-seed accessions, and *VrKIX8*, *VrPAT14*, *VrEmp24/25*, *VrIAR1*, *VrBEE3*, *VrSUC4*, and *Vrflo2* were further verified by RT-qPCR. Among 65 genes around these QEIs, *VrFATB*, *VrGSO1*, *VrLACS2*, and *VrPAT14* were homologous to known seed-development genes in *A. thaliana*, although new experiments are necessary to explore these novel GEI-trait associations. In addition, 54 genes were identified in comparative genomics analysis to be associated with seed development pathway, in which *VrKIX8*, *VrABA2*, *VrABI5*, *VrSHB1*, and *VrIKU2* were also identified in genome-wide association studies. This result provided a reliable approach for identifying seed-size-related genes in mungbean and a solid foundation for further molecular biology research on seed-size-related genes.

KEYWORDS

multiple genome-wide association studies, QTN-by-environment interactions, *VrEmp24/25*, multi-omics analysis, RT-qPCR

Background

Mungbean (*Vigna radiata* (L.) R. Wilczek) is a basic source of protein and carbohydrate, as it contains approximately 20% protein and 75% carbohydrate, and is a traditional and important legume in Asia (Somta et al., 2007). Due to its short life cycle (60–75 days), relative drought tolerance, and the ability to restore atmospheric nitrogen in association with *Rhizobium*/*Bradyrhizobium* bacteria, mungbean plays a crucial role in cropping systems and soil improvement (Somta et al., 2007; Alam et al., 2014).

The crop is generally grown as a cash crop in cereal-based farming systems. However, the major constraint in mungbean production is low seed yield. The average seed yield of mungbean is only approximately 700 kg per ha (Islam et al., 2015). Therefore, improving seed yield is the main goal in mungbean breeding. Understanding the genetic basis underlying seed-size-related traits is critical for the genetic improvement of mungbeans. In mungbeans, the ideotype of high-yielding cultivars are generally characterized by a large seed size, a short and synchronous maturity, a low sensitivity or insensitivity to day length, and the resistances to insects and disease (Fernandez et al., 1988). However, the knowledge on genes related to seed size has been limited. Moreover, the genes involved in the pathway of seed developments are not yet fully known.

Seed weight is the most important yield component and directly proportional to seed yield per plant in mungbean. To date, there have been seven studies of QTLs for seed weight in mungbean. Most of these studies are based on bi-parental segregation populations derived from interspecific crosses between cultivated and wild (*V. radiata* var. *sublobata*) mungbeans, and only two studies have evaluated seed size in more than one environments. The number of QTLs identified in those studies ranged from 3 to 11. Humphry et al. (2010) reported 11 loci for seed weight using SSR-marks, and Mei et al. (2009) identified a major QTL associated with both bruchid resistance and seed mass. Nonetheless, no candidate gene was identified for this trait.

Although many genes for seed weight have been reported in *Arabidopsis* (Plackett et al., 2012; Ge et al., 2016; Lu et al., 2016; Cheng et al., 2018; Zhang et al., 2020), soybeans, and rice (Luo et al., 2013; Ge et al., 2016; Liu et al., 2020a; Hao et al., 2021; Nguyen et al., 2021), few genes were reported in mungbean. In *Arabidopsis*, *FATB* (Bonaventure et al., 2003) was involved in the synthesis of short-chain fatty acids and influenced seed development. Although *GA20OX* regulated *Arabidopsis* in late floral development (Plackett et al., 2012), the overexpression of *GmGA20OX* in *Arabidopsis* enhanced seed size and weight. *KIX8* controlled seed size in *Arabidopsis* and soybeans (Liu et al., 2020a; Nguyen et al., 2021). *BES1* suppressed the cell elongation and increased seed size in legume species (Ge et al., 2016). *ERG2* promoted early seed development and influenced the length of mature siliques (Cheng et al., 2018). In soybeans, *GA20OX* (Lu et al., 2016), *GmFAD3* (Singh et al., 2011), *GmLEC2* (Manan et al., 2017), *GmPDAT* (Liu et al., 2020c), *GmKIX8-1* (Nguyen et al., 2021), and *GmGA3ox1* (Hu et al., 2022) were found to influence seed size by regulating lipid accumulation or increasing cell proliferation. In rice, *D1* (Sun et al., 2018), *D2* (Fang et al., 2016), *flo2* (She et al., 2010), *GS3* (Sun et al., 2018), *OsBZR1* (Liu et al., 2021), *GW2* (Hao et al., 2021), *D11* (Wu et al., 2016), and *OsHT* (Guo et al., 2020) were found to control seed weight by regulating rice grain size or starch quality.

Knowledge regarding seed development pathway is also a valuable source for transgenic strategies to improve crop production. As reported, there are several signaling pathways that control seed size, including the G-protein signaling, ubiquitin proteasome pathways, mitogen-activated protein kinase (MAPK) signaling, auxin pathways, and some transcriptional regulators (Li et al., 2019). In *Arabidopsis*, *GPA1*, *AGB*, and *AGG3* were involved in G-protein-signaling pathways. *DA1*, *DA2*, *SOD2*, *UBP15*, *EOD1*, and *SAMBA* were involved in ubiquitin proteasome pathways. In addition, *ABA2*, *ABI5*, *SHB1*, *MINI3*, *IKU2*, and *CKX* were involved in the HAIKU (IKU) pathway. Additional genes were found to be related to seed size developments, but their pathways are uncertain, such as *KIX8*, *BES1*, *MES1*, and *KLU* (Orozco-Arroyo et al., 2015; Li et al., 2019). However, some reports have been focused on genetic foundation and molecular mechanism of seed developments in mungbean.

Genome-wide association studies (GWASs), along with multi-omics analysis, have been frequently used to mine candidate genes for most important agronomic traits in crops. Integrating GWAS with comparative genomics, transcriptome analysis, and molecular experiments, genes have been identified to be associated with complex traits (Liu et al., 2020c). For example, Gong et al. (2022) conducted a GWAS with high-quality single nucleotide polymorphism (SNP) data and seed-size traits, and found that Cla97C05G104360 and Cla97C05G104380, which are involved in abscisic acid metabolism, played important role in regulating the seed size in watermelon. Duan et al. (2022) identified *GmST05* to be associated with soybean seed size through the GWAS of 1800 soybean germplasm resources, and *GmST05* differed significantly at the transcriptional level. Liu et al., 2022a,c used GWASs and biological experiments to identify a pleiotropic gene *GmPDAT* for seed size- and oil-related traits in

Abbreviations: GWAS, genome-wide association study; HSW, 100-seed weight; FPKM, Fragments Reads Per Kilobases per Million reads; PPI, protein–protein interaction; RNA-seq, RNA sequencing; QEIs, QTN-by-environment interactions; GEMMA, genome-wide efficient mixed-model association; CMLMs, compressed mixed linear models; EMMAX, efficient mixed-model association expedited; KIX8, KINASE-INDUCIBLE DOMAIN INTERACTING8; Emp24/25, emp24/gp25L/p24 family; QTNs, quantitative trait nucleotides; SNP, single nucleotide polymorphism; SW, seed width; SL, seed length.

soybean, and a salt-stress-tolerance gene *VrFRO8* in mungbean. Nonetheless, the related genes responsible for seed-size-related traits remained unknown in mungbean.

To address the above issues, 196 mungbean accessions with 3,607,508 SNP markers were used to conduct GWAS for seed length (SL), seed width (SW), 100-seed weight (HSW) using 3VmrMLM (Li et al., 2022b), efficient mixed-model association expedited (EMMAX) (Kang et al., 2010), genome-wide efficient mixed-model association (GEMMA) (Zhou and Stephens, 2012), and compressed mixed linear model (CMLM) (Zhang et al., 2010) methods. Candidate genes around quantitative trait nucleotides (QTNs) and QTN-by-environment interactions (QEIs) for the three traits were predicted by transcriptomics and comparative genomics. Key candidate genes were verified by RT-PCR analysis. Moreover, genes in seed-development-regulation pathway were also mined by comparative genomics. It should be noted that *VrEmp24/25* and *VrKIX8* were found to be associated with SL and HSW, and a major gene *VrPAT14* (LOD = 61.95, r^2 = 5.80%) was identified in QEI detection *via* 3VmrMLM.

Materials and methods

Plant materials and treatments

A diverse set of 196 mungbean accessions including 20 wild and 176 cultivated accessions from 23 countries, were used in this study (**Supplementary Data Set 1**). All the accessions were planted in a randomized complete block design with two replicates in an experimental field of Kasetsart University, Kamphaeng Saen Campus, Nakhon Pathom, Thailand in 2018 and 2020. In each replicate, each accession was planted in a single row 2.5 m long with 12.5 cm intra-row spacing (ca. 20 plants/row) and 50 cm inter-row spacing. Cultural practices were performed according to Park (1978). SW (mm), SL (mm), and HSW (g) were measured. At maturity. The SL and SW traits for each accession were averaged based on 20 seeds and 100SW for each accession was averaged based on three replicates.

Whole-genome resequencing

The young leaves of the above 196 mungbean accessions were collected 1 week after planting. The DNA was extracted in 2018, using the CTAB method (Smith et al., 2005). Short reads sequenced by an Illumina HiSeq 4000 platform (Illumina, San Diego, CA, United States), and mapped to scaffolds using Burrows-Wheeler-Alignment Tool (BWA) (Version 0.7.15)[1] (Li and Durbin, 2009). Genome Analysis Toolkit (GATK) was used to select SNP and indel[2] (McKenna et al., 2010). Sulv 1 genome was selected as the reference genome in the GATK analysis (Yan et al., 2020). High-quality SNPs and Indel variations were obtained as the following steps. (a) Retaining concordant sites both identified by GATK and VCFtools were retained (Danecek et al., 2011). (b) Filtering out SNP with quality value below 30, removing SNPs with an average coverage depth < 8× and with minor allele frequency (MAF) less than 5%. (c) Deleting insertions and deletions (InDels) with length less than10 bp were deleted. A total of 3,607,508 SNPs were identified.

As described in Liu et al. (2022a), the number of subpopulations was five ($K = 5$), and the population structure (Q matrix) was calculated using ADMIXTURE software (version is 1.3.0).[3] The K matrix was calculated using the above CMLM (GAPIT version 3),[4] EMMAX (GAPIT),[5] GEMMA (Version 0.94.1)[6], and 3VmrMLM programs (IIIVmrMLM)[7] (**Supplementary Data Set 2**; Li et al., 2022a).

Genome-wide association study for seed width, seed length, and 100-seed weight

Only the SNPs with MAF ≥ 0.05 and missing rate < 10% were used in GWAS (Pongpanich et al., 2010). The lines with more than 95% missing for trait were filtered out (Liaw and Wiener, 2002). SW, SL, and HSW, and the above SNP markers in 196 mungbean accessions were used to conduct GWAS using four different methods, including 3VmrMLM (Li et al., 2022b) *via* software IIIVmrMLM (Li et al., 2022a), EMMAX (Kang et al., 2010), GEMMA (Zhou and Stephens, 2012), and CMLM (Zhang et al., 2010). The probability threshold for significant QTNs was set at $1/m$ = 2.77e-07 (m = 3,607,508) for all the four methods (Xu et al., 2018; Zhang Y. M. et al., 2019; Zhang Y. M. et al., 2019), and the LOD score threshold for suggested QTNs was set at LOD ≥ 3.0 for 3VmrMLM (Li et al., 2022b). Heatmaps of the linkage disequilibrium was generated by LDheatmap package (Shin et al., 2006), haplotype analysis was conducted by LDheatmap package (Barrett et al., 2005). The averages for those traits measured in 2018 and 2020 were used in GWAS.

Candidate gene identification

Candidate genes for salt tolerance were mined in the follow steps. (a) All the genes between the 30 Kb around regions for each of the significantly QTN were mined, where the LD-value was about 20 Kb in mungbean, (b)

1 http://bio-bwa.sourceforge.net/bwa.shtml

2 https://gitee.com/mirrors/GATK

3 http://dalexander.github.io/admixture/download.html

4 http://zzlab.net/GAPIT

5 http://csg.sph.umich.edu//kang/emmax/download/index.html

6 https://github.com/genetics-statistics/GEMMA

7 https://github.com/YuanmingZhang65/IIIVmrMLM

mined the *Arabidopsis*, rice and soybean homologous genes of those candidate genes, which were reported related to seed developments, seed production, phytohormone signaling pathways and carbohydrate metabolism pathways, etc. (Li et al., 2019), as the candidate genes. (c) The selected genes showing different expression between two groups of mungbean accessions contrasting in seed size (large seed vs. small seed) (see below) were considered as candidate genes.

Differentially expressed gene based on RNA-sequenced data

Two large-seeded accessions [G141 and G143; 19.32 ± 7.09 (g)] and two small-seeded accessions [G169 and G171; 11.58 ± 5.93 (g)] were selected for RNA sequencing (RNA-seq) analysis. Data in seed set were collected at three seed development stages (10, 15, and 25 DAF) for RNA extraction in 2021. Total RNA was extracted using RNAprep Pure Plant Kit (DP441) according to the manufacturer's instructions. 1 μg high-quality RNA samples (OD260/280 = 1.8∼2.2; OD260/230 ≥ 2.0; RIN ≥ 6.5; 28S:18S ≥ 1.0 and >10 μg) were used to construct the sequencing library (G9691B, Agilent). The RNA were analyzed in an Illumina Novaseq Sequencer. Raw reads were cleaned by trimmomatic[8] (Bolger et al., 2014), and clean reads were mapped to reference sequences using Hisat2 (Pertea et al., 2016). The gene expression level was calculated by using RPKM method by Subread package (Mortazavi et al., 2008).

In the key candidate gene identification, the extracted RNA in two large-seeded accessions at 10 and 25 DAF were treated with RNase-free DNase I (Promega, Madison, WI, United States). After reverse transcription, the cDNA was used as a template for RT-qPCR using the Takara Bio TB Green Premix Ex Taq (Tli RNase H Plus). The detail progress was described by Liu et al. (2022b). Reactions were run on a Bio-Rad CFX96 system. *EVM0007380* (homologous of *At3g18780*) was used as the CK in this experiment. Primers were designed by NCBI and tested by RCR of tubulin. The *t*-test was adopted in the hypothesis testing, $P < 0.05$, $P < 0.01$, and $P < 0.001$ indicated significant probability levels at 0.05, 0.01, and 0.001, respectively. Information of the primers used is presented in **Supplementary Table 1**.

Protein–protein interaction

The protein–protein interactions (PPIs) were detected used the online tools STRING[9] (Jensen et al., 2009). The mungbean (*V. radiata* (L.) R. Wilczek) protein database was used as the protein library.

Results

Phenotypic variation for mungbean seed-size-related traits

100-seed weight, SW, and SL in 196 mungbean accessions were measured in 2018 and 2020. The average-plus-standard deviations for the three traits across the 2 years were 5.05 ± 1.91 (g), 3.48 ± 0.51 (mm), and 4.64 ± 0.99 (mm), respectively, and their average coefficients of variation (CV) across the 2 years were 38.5, 14.5, and 16.5 (%), respectively (**Supplementary Table 2**). Although the trends for those traits in the 2 years were similar (**Figures 1A–C**), HSW (38.5%) had much larger phenotypic variation than SW (14.5%) and SL (16.5%), indicating their large phenotypic variation and typical quantitative traits. In general, the wild mungbeans showed low seed weight (1.68 ± 0.61) as well as short SW (2.45 ± 0.401) and SL (3.12 ± 0.43), while the cultivated mungbeans had high seed weights (5.29 ± 1.68) as well as long SW (3.56 ± 0.41) and SL (4.76 ± 0.92) (**Supplementary Table 2**). Moreover, significant difference for each trait between the 2 years was observed ($P < 0.001$), and these traits had significant correlations with each other ($r > 0.87$, $P < 0.001$) (**Figure 1D**), indicating the existence of common QTNs among these traits (Liu et al., 2020b).

Genome-wide association studies for seed-size-related traits in mungbean

Detection of main-effect quantitative trait nucleotides for seed-size-related traits in each environment

After removing the SNPs with an average coverage depth < 8× and with a MAF less than 5%, we identified more than 3.6 million SNP markers. In the single-environment analysis, the phenotypic observations for each trait in 196 accessions measured in 2018 and 2020 were used to associate with 3,607,508 SNPs using 3VmrMLM, EMMAX, GEMMA, and CMLM under the situations of five subpopulations and polygenic background control (kinship matrix) (**Supplementary Data Set 3**). As more than 10,000 QTNs were identified by GEMMA for HSW in 2018, the relevant results were not used in the subsequent analysis. As a result, 208 significant QTNs were identified for the above traits. Thirteen significant QTNs were simultaneously identified in two environments by two GWAS methods (**Supplementary Table 3**; **Supplementary Data Set 4**), some significant QTNs are presented in **Figure 2**. For example, Chr10-25206533-25223155

8 http://www.usadellab.org/cms/index.php?page=trimmomatic

9 https://string-db.org//

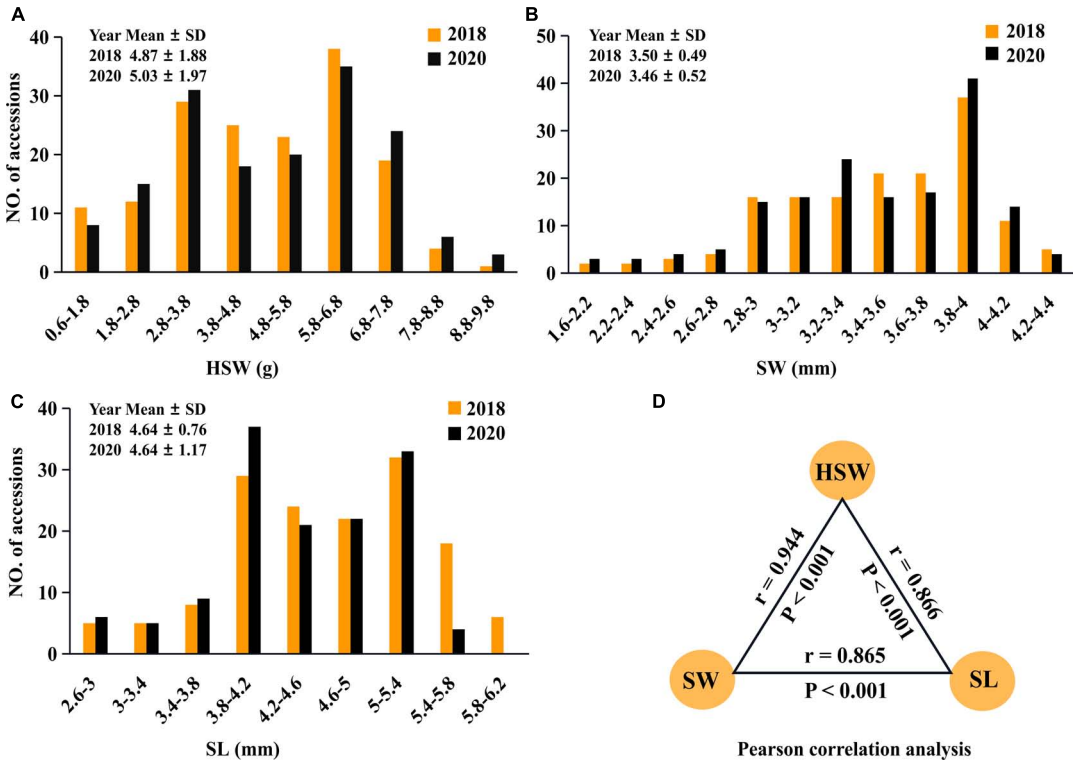


FIGURE 1
The frequency distributions of seed-size-related traits. Frequency distributions of HSW **(A)** (g), SL **(B)** (mm), and SW **(C)** (mm) in 196 mungbean accessions, which were measured in 2018 (brown bar) and 2020 (black bar). SD, standard deviation. The associations of HSW with SW and SL, the average dates of those traits measured in 2018 and 2020 were used in the partial correlation analysis **(D)**.

(LOD = 15.40∼37.89, *P* = 3.16E-08∼5.15E-09) was detected in 2018 and 2020 by MLM, EMMAX, and 3VmrMLM to be associated with HSW, SW, and SL (Table 1; Figures 2A–F), and the Q-Q plot in the Supplementary Figures 1A–D, which was corresponding to the GWAS results in Figure 2, except 3VmrMLM. And Chr1-71543546 (LOD = 7.70∼12.44) was detected in 2018 and 2020 by 3VmrMLM to be associated with SW (Supplementary Table 3). These QTNs were distributed on chromosomes 1–4, and 10 (≥20 QTNs for each chromosome) and had a 1.15% average proportion of their total phenotypic variation explained by each QTN, and there were 47, 115, and 46 QTNs, respectively, for HSW, SL, and SW (Supplementary Data Set 4).

Detection of quantitative trait nucleotides for seed-size-related traits in multiple environments

To detect more stable QTNs, three seed-size-related traits of 196 mungbean accessions measured in 2018 and 2020 were used to associate with 3607508 SNP markers using two-environment 3VmrMLM joint analysis. As a result, 32, 33, and 18 significant QTNs were identified for HSW, SL, and SW, respectively (Supplementary Table 3), and had a 1.08% average proportion of total phenotypic variation explained by each QTN. Moreover, eight significant QTNs were identified (Supplementary Table 4). For example, Chr1-8161305-8347626 (LOD = 24.09∼36.33) and Chr10-25222572-25223133 loci (LOD = 29.75∼37.89) were detected to be associated with HSW and SL, respectively (Supplementary Tables 3, 4).

Based on all the above main-effect QTNs in single- and multiple-environment analysis, five stable QTNs across various methods and/or two environments were found (Supplementary Table 5), including Chr1-8161305-8347626 (LOD = 24.09∼36.33), Chr2-12602704 (LOD = 17.71∼38.08), Chr4-10069367 (LOD = 17.72∼34.19), Chr5-10834954 (LOD = 9.53∼30.03), and Chr10-Chr10-25222572-25223133 (LOD = 29.75∼37.89), especially, Chr1-8161305-8347626 and Chr10-25222572-25223133 were simultaneously identified across methods and two environments.

Detection of quantitative trait nucleotide-by-environment interactions for seed-size-related traits in multiple environments

All the above datasets in GWAS were used to detect QEIs using 3VmrMLM. As a result, 5, 10, and 5 significant QEIs were

TABLE 1 Eight key candidate genes derived from genome-wide association studies for seed-related traits.

Trait	Genome-wide association studies					Comparative genomics				Function	Reference
	Chromosome	Position (bp)	LOD score or P_1-value	r^2 (%)	Method	Candidate genes	P_2-value	$\log_2$FC	Arabidopsis homologs		
Single_env: Detection of main-effect QTNs for seed size-related traits											
2018-HSW	1	52015258	21.84	0.81	3VmrMLM	EVM0016442/IAR1	0.05*	0.39	AT1G68100	IAA-alanine resistance protein 1	Rampey et al., 2013
	4	36876485	35.25	1.3	3VmrMLM	EVM0019602/flo2	0.02*	1.09	AT4G36920	Seed development	She et al., 2010
	11	3018112	25.95	2.62	3VmrMLM	EVM0010067/ABA2	0.18	0.21	AT1G52340	Seed maturation	Chauffour et al., 2019
2020-HSW	1	8177726	28.09	1.31	3VmrMLM	EVM0032114/KIX8	0.03*	0.49	AT3G24150	Seed development	Li et al., 2019
	4	7755858	19.6	1.03	3VmrMLM	EVM0015332/SUC4	0.02*	0.29	AT1G09960	Sucrose transport protein SUC4	Xu and Liesche, 2021
	10	25206533	15.41	0.59	3VmrMLM	EVM0015812/Emp24	0.02*	0.67	AT1G26690	Emp24 family protein	Ren et al., 2019
2018-SW	1	71543546	12.44	1.65	3VmrMLM	EVM0002784/BEE3	0.01*	1.24	AT1G73830	Seed development	Moreno et al., 2018
2020-SW	1	30724948	29.81	1.74	3VmrMLM	EVM0033315/SHB1	0.15	0.04	AT4G25350	Seed development	Zhang H. et al., 2017
	1	71543546	7.70	0.57	3VmrMLM	EVM0002784/BEE3	0.01*	1.24	AT1G73830	Seed development	Moreno et al., 2018
	6	13463604	12.93	0.55	3VmrMLM	EVM0028931/ZIP6	0.02*	−0.85	AT2G30080	Seed development	Lee et al., 2021
	9	24007163	61.96	5.8	3VmrMLM	EVM0027211/PAT14	0.03*	1.19	AT3G60800	Leaf senescence	Zhao et al., 2016
2018-SL	3	34837582	3.24E-08	NA	EMMAX	EVM0028440/ABI5	0.19	0.25	AT2G36270	ABSCISIC ACID-INSENSITIVE 5 isoform X4	Lynch et al., 2022
	6	1650897	1.92E-08	NA	EMMAX	EVM0030447/IKU2	0.43	0.78	AT3G19700	Embryo development	Xiao et al., 2016
	10	25223155	5.15E-09	0.992	CMLM	EVM0015812/Emp24	0.01*	0.67	AT1G26690	Emp24 family protein	Ren et al., 2019
	10	25222572	1.91E-06	0.515	CMLM	EVM0015812/Emp24	0.01*	0.67	AT1G26690	Emp24 family protein	Ren et al., 2019
	10	25223133	9.34E-09	2.264	CMLM	EVM0015812/Emp24	0.01*	0.67	AT1G26690	Emp24 family protein	Ren et al., 2019
	10	25223155	3.16E-08	3.411	CMLM	EVM0015812/Emp24	0.01*	0.67	AT1G26690	Emp24 family protein	Ren et al., 2019
	10	25223133	9.34E-09	NA	EMMAX	EVM0015812/Emp24	0.01*	0.67	AT1G26690	Emp24 family protein	Ren et al., 2019
Multi_env: Detection of main-effect QTNs for seed size-related traits											
HSW	1	8161305	36.33	0.8	3VmrMLM	EVM0032114/KIX8	0.03*	0.50	AT3G24150	Seed development	Li et al., 2019
	1	52015258	13.52	0.12	3VmrMLM	EVM0016442/IAR1	0.06	0.39	AT1G68100	IAA-alanine resistance protein 1	Rampey et al., 2013
	4	7755858	28.43	0.66	3VmrMLM	EVM0015332/SUC4	0.02*	0.30	AT1G09960	Sucrose transport protein SUC4	Xu and Liesche, 2021
	4	36876485	71.71	0.95	3VmrMLM	EVM0019602/flo2	0.02*	1.09	AT4G36920	Seed development	She et al., 2010
	10	25222572	37.89	0.67	3VmrMLM	EVM0015812/Emp24	0.01*	0.67	AT1G26690	Emp24 family protein	Ren et al., 2019
SL	1	8347626	24.09	0.35	3VmrMLM	EVM0032114/KIX8	0.03*	0.50	AT3G24150	Seed development	Li et al., 2019
	4	19559337	16.8	0.32	3VmrMLM	EVM0022984/flo2	NA	NA	Os04g0645100	Seed development	She et al., 2010
	10	25223133	29.75	0.64	3VmrMLM	EVM0015812/Emp24	0.01*	0.67	AT1G26690	Emp24 family protein	Ren et al., 2019
SW	6	13463604	27.54	1.62	3VmrMLM	EVM0028931/ZIP6	0.02*	−0.85	AT2G30080	Seed development	Lee et al., 2021

The P_1-values were calculated by CMLM, EMMA, and 3VmrMLM, The P_2-values were calculated using paired t-test from the average FPKM values at three stages between two high seed weight (n_1 = 2) and tow seed weight (n_2 = 2) mungbeans, and their significances were marked by * (0.05 level); FC and NA represent fold change and no expression, respectively.

10.3389/fpls.2022.997988

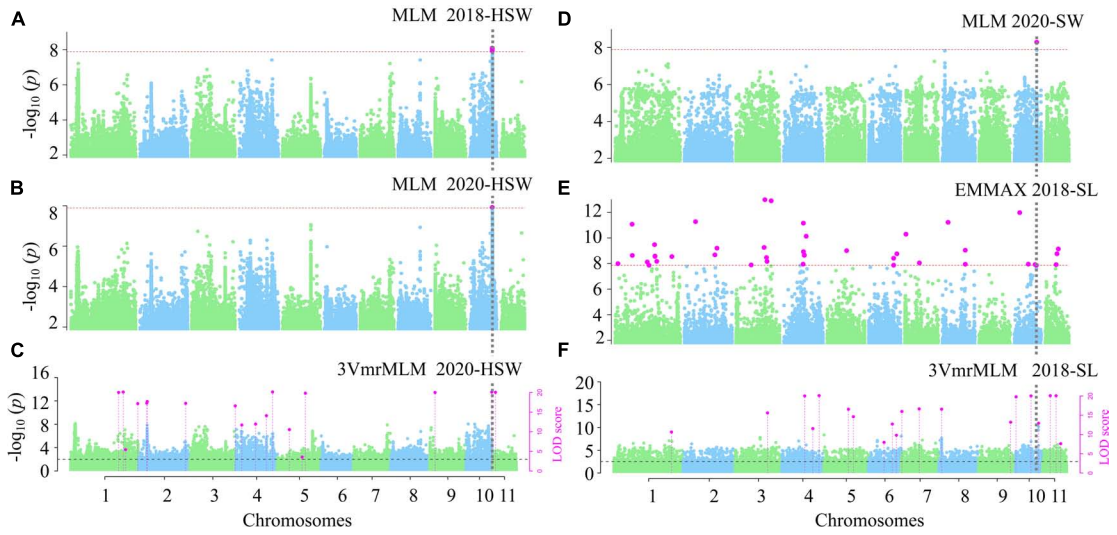


FIGURE 2
Manhattan plots for the GWAS for seed-yield-related traits. GWAS for HSW **(A–C)**, SW **(D)** and SL **(E,F)**. Significant QTN in phenotypic GWAS was set at *P*-value ≤ 0.05/m = 1.39e-08 (*m* = 3607508), ≤2.77e-09 for CMLM and EMMAX **(A,B,D,E)**; and LOD ≥ 3.0 for the 3VmrMLM as the significant QTN, and all the critical values were marked by horizontal lines, *Y*-axis on the left side reports −log10 *P*-values of SNP, while *Y*-axis on the right side reports LOD scores, and LOD scores are shown in points with straight lines.

found to be associated with HSW, SL, and SW, respectively (**Supplementary Figure 2**; **Table 2**). Among these QEIs, 5 had zero dominant-by-environment interaction effects, and 7 had zero additive-by-environment interaction effects. For example, the two loci Chr4-26262890 and Chr4-31677341 for HSW had only additive-by-environment interaction effects of 0.12 (**Supplementary Figures 2A–C**, LOD = 12.70; r^2 = 0.26) and 0.08 (**Supplementary Figures 2A–C**, LOD = 12.65; r^2 = 0.27), respectively.

The two loci Chr1-155976 and Chr1-3598291 for HSW had only dominant-by-environment interaction effects of −0.61 (LOD = 12.73; r^2 = 0.25) and 0.44 (LOD = 13.25; r^2 = 0.27), respectively. Among the 20 QEIs, the loci Chr4-5255551 and Chr7-16074671 had inconsistent directions between additive- and dominant-by-environment interaction effects.

In addition, among these QEIs, the QEI locus Chr9-24007163 for SW had large effect, and r^2 was 5.8% (**Supplementary Figure 2B**, LOD = 61.95). The additive and dominant effects in environment 1 were −0.14 and −0.098, respectively.

Candidate genes for seed-size-related traits

A total of 6912 DEGs were identified between two high-seed-weight and low-seed-weight mungbeans (FDR ≤ 0.05) (**Supplementary Figures 3A,B**; **Supplementary Data Set 6**). These DEGs were intersected with 809 genes around significant QTNs for HSW, SL, and SW (**Supplementary Tables 3**, **4**; Supplementary Data Sets 4, 5). As a result, 53 out of 809 genes were differentially expressed ($P \leq 0.05$, $Log_2FC \geq 0.5$). Using comparative genomics analysis, 12 out of 53 DEGs were homologous to previously reported seed development related genes in rice and *Arabidopsis thaliana*, in which *KIX8*, *PAT14*, *Emp24/25*, *IAR1*, *BEE3*, *SUC4*, *flo2*, and *Zip6* had been confirmed *via* functional analysis in rice and *A. thaliana* (**Table 1**), such as *VrKIX8* (LOD = 24.09∼36.33), *VrEmp24/25* (LOD = 15.40∼37.89, *P* = 3.16E-08∼5.15E-09), *VrPAT14* (LOD = 61.96), and *VrZIP6* (LOD = 27.54). Among the eight genes, *VrKIX8*, *VrEmp24/25*, *VrIAR1*, *VrBEE3*, *VrSUC4*, and *Vrflo2* were significantly upregulated in high-HSW accessions, *VrPAT14* was significantly downregulated, and *VrZIP6* had no significant difference (**Figure 3A**), as compared to those in low-HSW accessions using the transcriptome data at 10, 15, and 25 DAF (**Supplementary Data Set 4**). We conducted RT-qPCR analysis to further confirm the eight key candidate genes. The results showed that seven genes were confirmed, except *VrZIP6*, a transcription factor related to seed development. All the seven genes had higher expression levels in the early stage of seed development (10 DAF) than in the late maturation stage of seed development (25 DAF) (**Figure 3B**; **Supplementary Data Set 7**), indicating their essential roles at early stage of seed development.

Using the same approach described above, among 65 genes around 20 QEIs, four were homologous to previously reported seed development related genes in rice and *A. thaliana* (**Table 2**), although new experiments are necessary to explore these novel GEI-trait associations. The four genes were described as below. *VrFATB* was linked to the locus Chr4-30176682 (**Supplementary Figure 2A**). As described in Bonaventure et al. (2003) and Sun et al. (2014), *FATB* is

TABLE 2 Twenty significant QTN-by-environment interactions for seed-size-related traits under multi-environments.

Trait	3VmrMLM						Candidate genes		*P*-value	log_2FC	Arabidopsis homologs	Function	References
	Chr	Position (bp)	LOD (QE)	Add × Env1	Dom × Env1	r^2 (%)							
HSW	1	25048694	7.99	0.08		0.18	EVM0010707; EVM0020394	EVM0010707	0.11	0.05	NA	NA	
	3	5498494	14.34	0.11		0.33	EVM0013436; EVM0027482; EVM002290	EVM0013436	0.21	1.53	AT3G61060	F-box protein PP2-A13	
	4	30176682	15.23	0.12		0.38	EVM0013210	EVM0013210/ FATB	0.09	0.50	AT1G08510	FATB	Bonaventure et al., 2003; Sun et al., 2014
	4	42563100	6.50	0.08		0.15	EVM0019039; EVM0011516	EVM0019039/ GSO1	0.09	0.91	AT4G20140	Seed development	Creff et al., 2019
	5	8962133	10.49	0.09		0.23	EVM0027740; EVM0007126	EVM0007126	0.05	−4.53	AT1G21450	Seed development	
SL	1	155976	12.73	0.00	−0.61	0.25	EVM0006618; EVM0002787; EVM0025368; EVM0002245; EVM0007007	EVM0006618	0.00	0.43	AT3G59910	Ankyrin repeat protein SKIP35 isoform X1	
	1	35982911	13.25	0.00	0.44	0.27	EVM0014255	EVM0014255	NA	NA	AT3G26570	Inorganic phosphate transporter 2-1, chloroplastic	
	4	22723706	12.93	−0.01	−0.61	0.26	EVM0015688	EVM0015688	0.03	0.07	AT5G50920	Chaperone protein ClpC, chloroplastic	
	4	26262890	12.70	0.00	−0.43	0.26	EVM0003123; EVM0001918	EVM0003123	NA	NA	NA	Citrate-binding protein-like	
	4	31677341	12.65	0.00	−0.61	0.27	EVM0009176; EVM0033509; EVM0023714; EVM0033630; EVM0032994	EVM0033630	0.03	NA	AT3G57520	Probable galactinol–sucrose galactosyltransferase 2 isoform X2	
	4	40101763	13.31	−0.01	−0.61	0.29	EVM0000524; EVM0025504	EVM0000524	0.21	NA	AT4G33140	Uncharacterized protein	
	7	16074671	12.90	0.01	−0.61	0.25	EVM0007632; EVM0003451; EVM0005587; EVM0017922; EVM0009325	EVM0007632	0.14	0.66	AT5G10330	Histidinol-phosphate aminotransferase, chloroplastic	
	7	28608053	12.99	−0.01	−0.61	0.27	EVM0025691; EVM0014665	EVM0025691	NA	NA	AT2G34930	Hypothetical protein	

(Continued)

10.3389/fpls.2022.997988

TABLE 2 (Continued)

Trait	3VmrMLM						Candidate genes		*P*-value	log_2FC	Arabidopsis homologs	Function	References
	Chr	Position (bp)	LOD (QE)	Add × Env1	Dom × Env1	r^2 (%)							
	8	32848165	12.70	0.00	−0.61	0.26	EVM0033747; EVM0012210; EVM0020228; EVM0006042; EVM0026839; EVM0012261; EVM0001209; EVM0016212; EVM0027531; EVM0030105; EVM0021224; EVM0011572	EVM0012210/ LACS2	0.03	−2.53	AT1G49430	Long chain acyl-CoA synthetase 2 isoform X1	Schnurr et al., 2004; Bai et al., 2022
	11	24829262	12.65	0.00	−0.61	0.25	EVM0006035; EVM0003000; EVM0020076; EVM0004982	EVM0020076	0.03	0.22	AT1G59870	ABC transporter G family member 36	
SW	2	29996834	9.66	0.02	0.26	0.62	EVM0004520; EVM0005114	EVM0004520	0.09	1.02	AT3G09300	Oxysterol-binding Protein-related protein 3B	
	4	5255551	7.38	0.02	−0.12	0.48	EVM0010724; EVM0028229	EVM0010724	0.11	NA	AT1G80550	Pentatricopeptide repeat-containing protein	
	4	19640302	16.41	0.00	−0.39	1.17	NA	NA	NA	NA	NA		
	7	18410421	9.28	−0.03	−0.20	0.61	EVM0022194; EVM0018119; EVM0020361; EVM0025547	EVM0022194	0.08	0.47	AT1G68690	Proline-rich receptor-like protein kinase PERK9	
	9	24007163	61.96	−0.14	−0.10	5.80	EVM0027211; EVM0026090; EVM0028888; EVM0024624; EVM0026781; EVM0029904; EVM0012085; EVM0004220	EVM0027211/ PAT14	0.03	1.19	AT3G60800	Leaf senescence	Zhao et al., 2016

The *P*-values were calculated using paired *t*-test from the average RPKM values at three stages between two high seed weight (n_1 = 2) and tow seed weight (n_2 = 2) mungbeans, and their significances were marked by * (0.05 level); FC and NA represent fold change and no expression, respectively.

10.3389/fpls.2022.997988

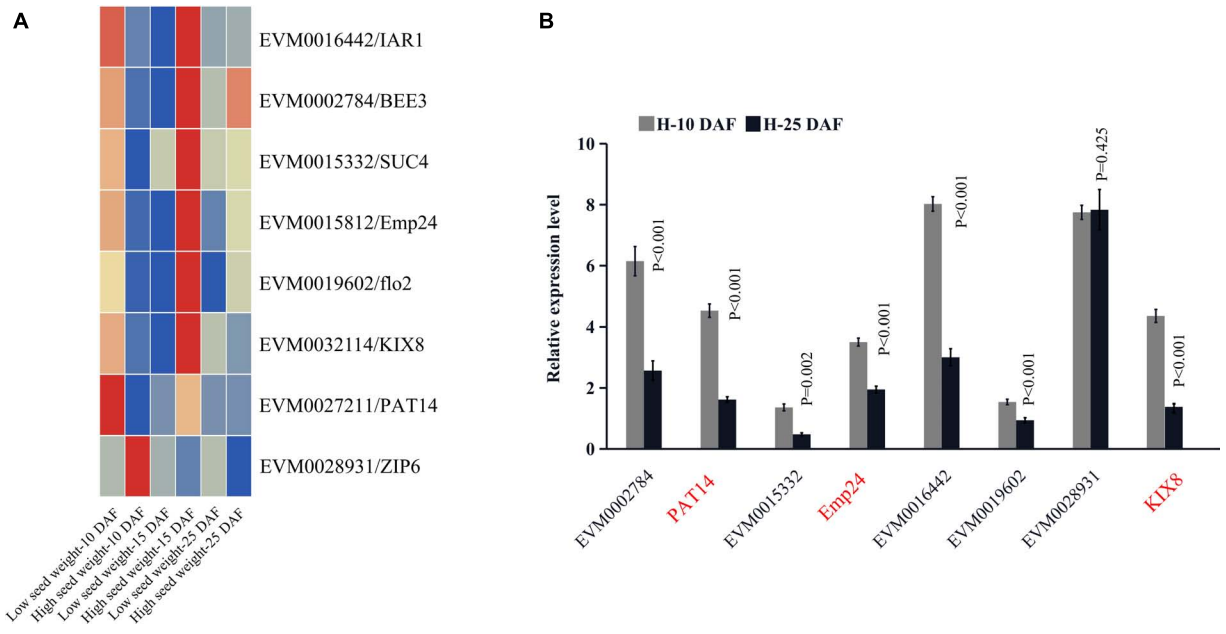


FIGURE 3
The expression of eight key candidate genes. The expression profiling of eight key candidate genes significantly associated with seed-size-related traits. The expression profiling of eight key candidate genes between two high-seed-weight and two low-seed-weight mungbeans **(A)**. Real-time PCR analysis of the eight key candidate genes; the *t*-test was used to test the significant differences of genes expression between two high-seed-weight mungbeans at 10 DAF and 25 DAF **(B)**. DAF, days after flowering.

a major determinant of saturated fatty-acid synthesis, and increases *FATB* activity at low temperature during seedling establishment caused high saturated fatty-acid content in plant. *VrGSO1* was linked to the locus Chr4-42563100 (**Supplementary Figure 2A**). As observed in Creff et al. (2019), *GSO1* was a stress signal-pathway-related gene, and stress-associated *MPK6* protein acted downstream of *GSO1* in developing embryo. *VrPAT14* was linked to the locus Chr9-24007163 (**Supplementary Figure 2B**). In Zhao et al. (2016), *PAT14* was involved with NPR1-dependent salicylic-acid signaling. *VrLACS2* was linked to the locus Chr8-32848165 (**Supplementary Figure 2C**), in which *VrLACS2* was essential for normal cuticle development in *Arabidopsis* (Schnurr et al., 2004) and *CrLACS2* suppression resulted in 50% less oil, yet with a higher amount of chloroplast lipids under N-deprivation (Bai et al., 2022).

Haplotype analysis of the main candidate genes

Two DEGs, *VrEmp24/25* and *VrKIX8*, were detected in the single- and multi-environment analyses (**Figures 4A,B**), and verified by RT-qPCR. Their haplotypic analyses were described as below.

In the haplotype analysis of *VrEmp24/25*, five SNP markers were found to be within *VrEmp24/25* and the promoter region (**Supplementary Data Set 8**), and the two SNP markers in *VrEmp24/25* were used to consist of three haplotypes (**Figure 4D**). Among the three haplotypes, hap 1 (5.17 g) had significantly higher HSW than hap 2 (1.58 g) and hap 3 (4.50 g; P = 2.11E-29) (**Supplementary Table 7**). Thus, hap 1 is elite haplotype. And the elite haplotypes TT made up more than 90.9% (160/176) in the cultivated mungbeans. *VrEmp24/25* with elite haplotype frequencies less than 45% in wild mungbeans (**Supplementary Table 7**; **Figure 4**) can be exploited for the improvement of mungbean cultivars.

Around the significant QTN Chr1-8161305-8347626 (**Figure 5A**; **Supplementary Data Set 8**), eight genes were found distributed in the region (**Figure 5B**). And six polymorphic loci, i.e., Chr1_8243935, Chr1_8243938, Chr1_8243939, Chr1_8243940, Chr1_8243945, and Chr1_8244001 were found in *VrKIX8* and the promoter region. All the six SNP were used to conduct the haplotype analysis (**Figure 5C**). Among the three haplotypes, hap 1 (5.09 g) had significantly higher HSW than hap 2 (4.56 g), hap 3 (3.47 g), and hap 4 (3.86 g) (**Supplementary Table 7**). Thus, hap 1 is elite haplotype. The elite haplotypes ATCGAA made up more than 73.2% (129/176) in the cultivated mungbeans, while the haplotype frequencies of CGAGT and CTAGGA were more than 25% (5/20) in wild mungbeans. Though Chr1_8243945 and Chr1_8244001 were located within the 5′ UTR of *VrKIX8*, and the amino acid sequence had not changed between cultivated mungbeans and wild mungbeans (**Figure 5D**). The SNP in 5′ UTRs could influence the translation efficiency of *VrKIX8* (Evfratov et al., 2017). The HSW in hap 1 (5.16 g) was significantly higher than that in hap 2 to hap 4 (3.50–4.66 g; P = 1.19E-21).

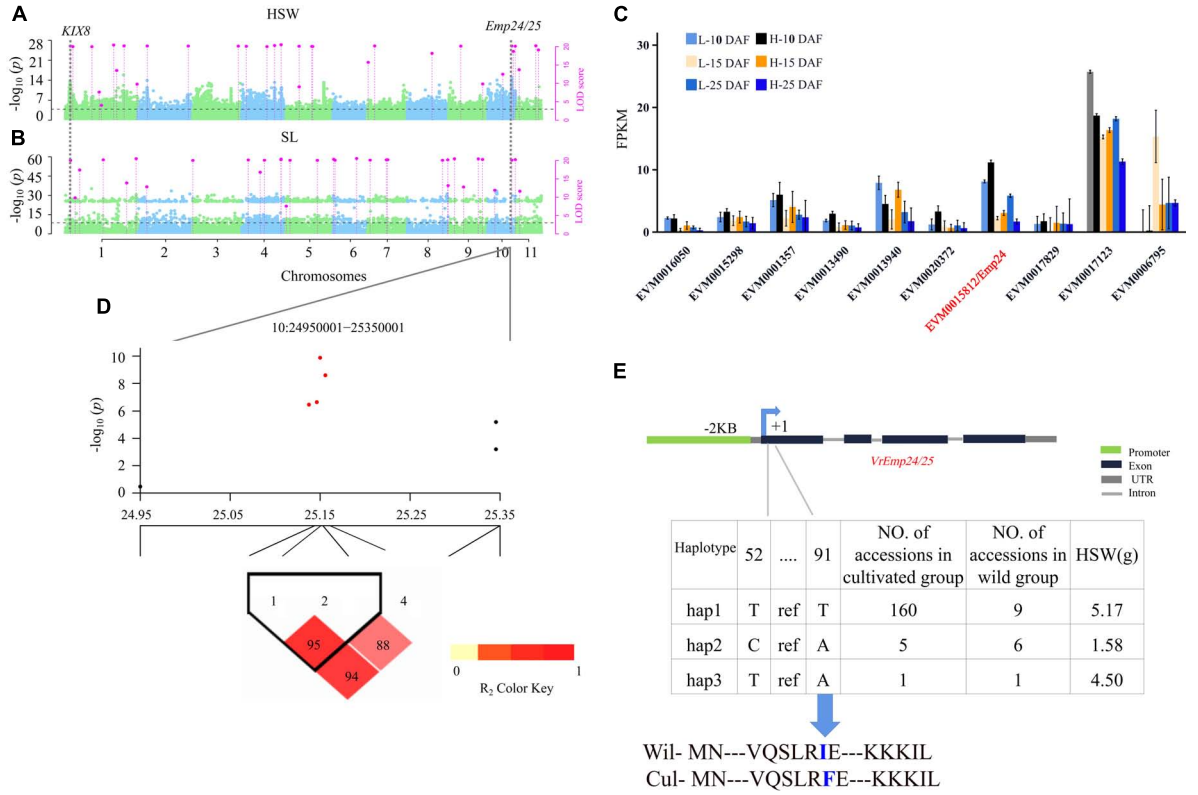


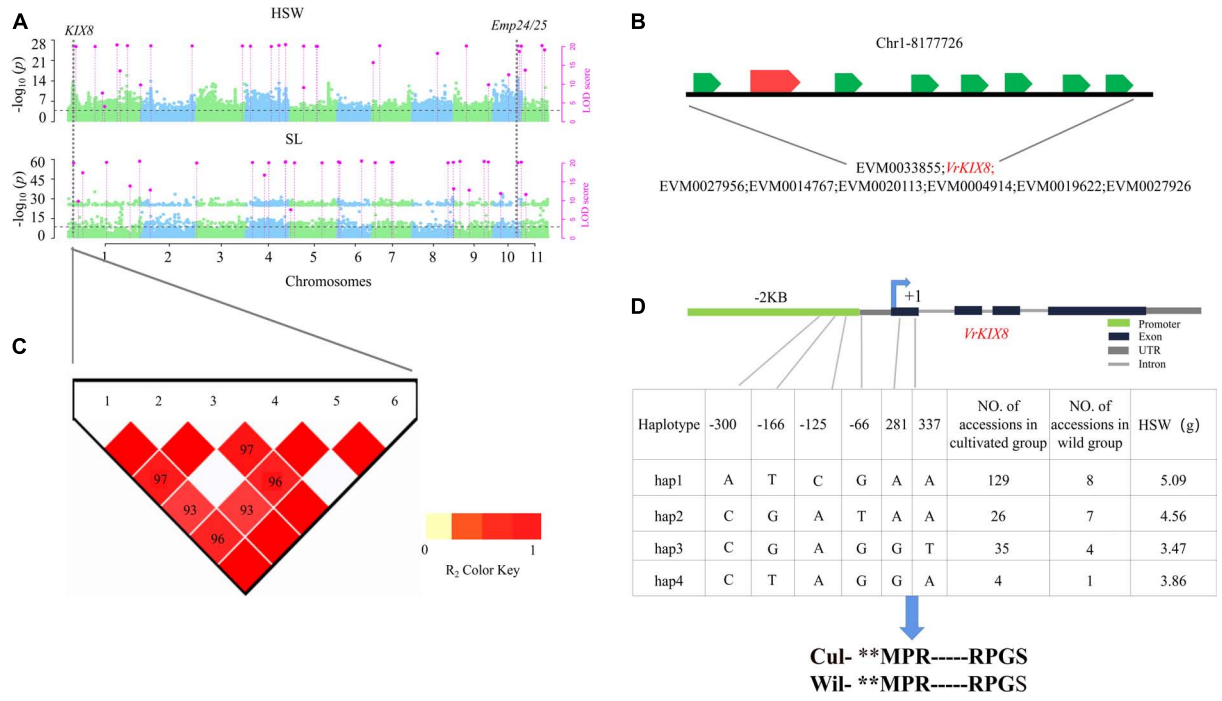
FIGURE 4
Genetic analysis of *VrEmp24/25*. Local Manhattan plots for HSW under multi-environments. LOD ≥ 3.0 for the 3VmrMLM as the significant QTN **(A,B)**. The expression profiling of 10 candidate genes for HSW identified at 30 Kb around Chr10-25222572-25223133 loci in the seed between two high-seed-weight and two low-seed-weight mungbeans **(C)**. LD heatmaps surrounding Chr10-25222572-25223133 loci **(D)**. Haplotype analysis of *VrEmp24/25* **(E)**, the thirtieth amino acid of *VrEmp24/25* changed from ATT (Ile, I) to TTT (Phe, F). DAF, days after flowering. Wil, the wild accessions. Cul, the cultivated accessions.

Based on these results, we deduced that these two SNP and six SNP cause the difference expression of the *VrEmp24/25* and *VrKIX8* gene, respectively. The discovery of *VrEmp24/25* and *VrKIX8* two domestication/improvement genes can accelerate breeding selections and facilitate ideal crop designs.

Expression patterns of seed development pathway genes in mungbean

As seed development pathway genes were largely unknown in mungbean, we mined seed development pathway genes by comparative genomics and transcriptomics analysis. As a result, 54 genes in seed-development pathway were identified in this study (Figure 6; Supplementary Data Set 9), such as two *GPA1*, one *AGB*, and one *AGG3*. In the ubiquitin proteasome pathways, two *DA1*, one *DA2*, one *SOD2*, one *EOD1*, and one *UBP15* rather than *SAMBA* were identified. In the auxin pathways, two *ABA2*, one *ABI5*, three *SHB1*, five *IKU2*, and three *CKX2* rather than IKU1 and MINI3 were identified (Figure 6A). Five transcription factors including three *BES*1, and two *SOD7* were identified. Moreover, 16 genes for seed size developments were found to be with uncertain pathways, including three *KIX8*, five *MES1*, and one *KLU* (Figure 6A; Supplementary Data Set 9). Among the 54 genes, 13 genes were significantly differentially expressed (*P*-value < 0.05, *t*-test) between two low-seed-weight (nos. G169 and G171) and two high-seed-weight (no. G141 and G143) accessions in the 196 mungbean accessions using the transcriptome data at 10, 15, and 25 DAF (Figure 6B; Supplementary Data Set 8). Moreover, almost 90% of the 54 genes (48/54) had higher expressions in the early stage of seed development (10 and 15 DAF) than in the late maturation stage (25 DAF), including *VrKIX8* (EVM0032114), which was commonly identified in the GWAS by 3VmrMLM for HSW and SL. And EVM0010067/*VrABA2*, EVM0033315/*VrSHB1*, EVM0028440/*VrABI5*, and EVM0030447/*VrIKU2* were also identified in the GWAS by 3VmrMLM, within 100 Kb region of significant QTNs (Table 1).

We also did the PPI analysis among the seed development pathway genes, and found five pairs of PPIs were larger than the medium confidence value of 0.40 (Supplementary Table 7), indicating the existence of significant PPIs, i.e., EVM0013794.1 (VrAGG3) and EVM0006667.1 (VrDA2)



Haplotype	-300	-166	-125	-66	281	337	NO. of accessions in cultivated group	NO. of accessions in wild group	HSW (g)
hap1	A	T	C	G	A	A	129	8	5.09
hap2	C	G	A	T	A	A	26	7	4.56
hap3	C	G	A	G	G	T	35	4	3.47
hap4	C	T	A	G	G	A	4	1	3.86

FIGURE 5
Genetic analysis of *VrKIX8*. Local Manhattan plots for HSW in multi-environments. LOD ≥ 3.0 for the 3VmrMLM as the significant QTN **(A)**. LD heatmaps surrounding Chr1-8161305-8347626 loci **(B)**. Genes around the significant QTN region, shown at the bottom **(C)**. Haplotype analysis of *VrKIX8* **(D)**. Wil, the wild accessions. Cul, the cultivated accessions. symbol "**" means omit the same sequence part.

(0.478), EVM0033720.1 (VrAGB) and EV944.1 (VrGPA1-1) (0.995), as well as EVM0033720.1 (VrAGB) and EVM0015092.1 (VrGPA1-2) (0.995).

Discussion

The high-yield and efficiency breeding progress of mungbeans have been limited by the lack of ideal yield-related genes. At present, few QTNs or QTLs of yield-related traits in mungbeans have been reported (Kang et al., 2014). This study provided a genetic analysis of seed-size-related traits in mungbeans, to improve the accuracy of significant QTNs, we used multiple genome-wide M0017 association studies combined with multi-omics analysis to mine candidate genes associated with yield-related traits. Firstly, a total of 98 QTNs and 20 QEIs were identified using 3VmrMLM, while 95 and 15 QTNs were identified using EMMAX, and CMLM, respectively. Then, in the identification of candidate genes, 12 key candidate genes were mined, and seven of them including *VrKIX8*, *VrEmp24/25*, and *VrPAT14* were evidenced by transcriptome analysis and RT-qPCR analysis. Lastly, through haplotype analysis, the thirtieth amino acid of *VrEmp24/25* in the elite haplotype was changed from Ile to Phe. And there were six SNP in the promoter and 5′ UTRs of *VrKIX8*, however, the amino acid sequence of *VrKIX8* in the elite haplotype was not changed. The results provided the theoretical basis for both the functional identification of seed-size-related genes and for quality improvements in mungbean breeding.

Multiple genome-wide association studies methods combined with multi-omics analysis in mining candidate genes

In the GWAS, how to identify candidate genes around significant QTNs has been a challenge. Liu et al. (2020c), Zhang et al. (2021), and Gong et al. (2022) selected the 100-kb interval upstream and downstream of the significant QTN as the candidate interval in watermelon and soybeans. Usually, the interval has been chosen according to the LD decay values.

In order to determine stable QTNs and key candidate genes for seed-size-related traits, we adopted the following analyses. Firstly, we used CMLM, EMMAX, GEMMA, and 3VmrMLM to identify stable QTNs, as a result, five stable QTNs for seed-size-related traits were detected in single- and multiple-environments (**Supplementary Table 5**), i.e., Chr1-8161305-8347626 (LOD = 24.09∼36.33), and Chr10-25222572-25223133 loci (LOD = 29.75∼37.89).

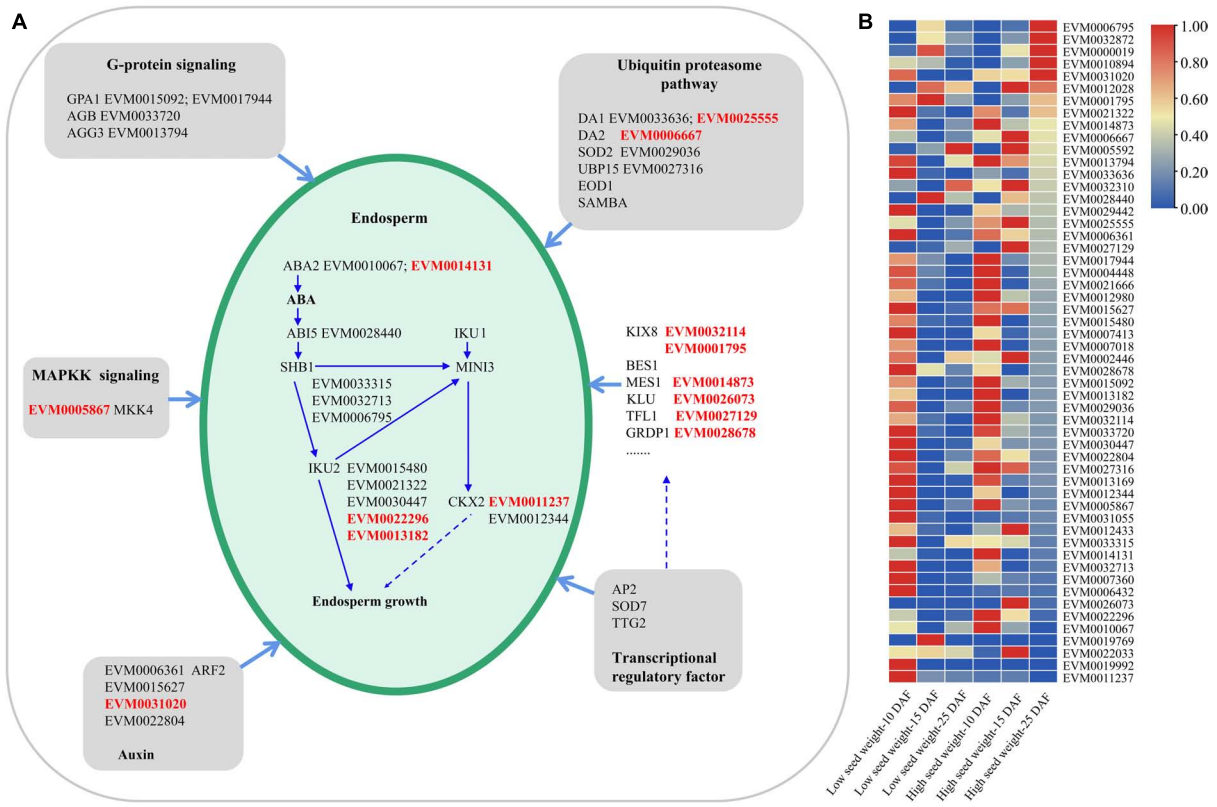


FIGURE 6
The seed development pathway in mungbean **(A)** and the expression profiling of 54 candidate genes predicted by comparative genomics, identified in this study **(B)**. GPA1, G PROTEIN ALPHA SUBUNIT 1; AGB, heterotrimeric G-protein beta subunit; AGG3, heterotrimeric G-protein gamma-subunit; DA1/DA2, a ubiquitin-activated peptides; SOD2, chloroplastic copper/zinc superoxide dismutase CSD2; EOD1, E3 ubiquitin ligase; UBP15, ubiquitin-specific protease; SAMBA, plant-specific negative regulator of the APC/C complex; MKK4, mitogen-activated map kinase; ABA2, ABA DEFICIENT 2; ABI5, ABA INSENSITIVE 5; SHB1, a nuclear and cytosolic protein; IKU1/IKU2, leucine rich repeat (LRR) kinase; CKX2, CYTOKININ OXIDASE 2; MINI3, MINISEED 3; ARF2, auxin response factor; AP2, AP2/EREBP (ethylene-responsive element-binding protein) class of transcription factors (**Figure 5A**). We also identified five transcription factors, three BES1 and two SOD7, transcription repressor involved in regulation of inflorescence architecture; KIX8, KINASE-INDUCIBLE DOMAIN INTERACTING8; MES1, METHYL ESTERASE 1; KLU, cytochrome P450 CYP78A5 monooxygenase. DAF, days after flowering; ABA, abscisic acid; DAF, days after flowering.

Second, in the identification of candidate genes, we conducted issue expression analysis, and comparative genomics analysis. 53 out of the 809 candidate genes were significantly differentially expressed between high and low HSW accessions ($P \leq 0.05$, $Log_2FC \geq 0.5$). Among the 53 DEGs, *Arabidopsis* homologous genes of the 12 key candidate genes had certain molecular functions. Notably, 10 of those genes were identified by 3VmrMLM (**Table 1**). Seven key candidate genes (*VrKIX8*, *VrEmp24/25*, *VrIAR1*, *VrBEE3*, *VrSUC4*, *VrPAT14*, and *Vrflo2*) were significantly differentially expressed between the low-seed-weight and high-seed-weight accessions, and further verified by RT-qPCR analysis (**Table 1**; **Figure 4**). *VrKIX8* (Chr1-8161305-8347626) and *VrEmp24/25* (Chr10-25222572-25223133) may be main genes in controlling seed-size-related traits.

Notably, 3VmrMLM showed more powerful ability in the detection of significant QTN than GEMMA, EMMAX, and CMLM, as it found more differentially expressed key candidate genes than other methods. The combination of 3VmrMLM and multi-omics analysis in the genetic analysis of complex traits was helpful.

Genome-wide association study provided potential genes *VrEmp24/25* and *VrKIX8* for mungbean seed-size-related traits

VrEmp24/25 was an important seed-size traits related gene, the evidence was as below: Firstly, Chr10-25206533-25223155 locus for seed size traits was detected in 2018 and 2020 by CMLM, EMMAX, and 3VmrMLM (**Figure 2**), and there were 10 genes in its interval (**Figure 4C**). Secondly, among the 10 genes, only *VrEmp24/25* (EVM0015812) ($P = 0.014$, $Log_2FC = 0.67$) had deferentially expressed across different phenotype accessions (**Figure 4C**; **Supplementary Data Set 4**).

Besides, in maize, the loss function of *EMP24* and *Emp25* would impair embryo and endosperm development (Xiu et al., 2020). *EMP24* was required for the splicing of nad4 (Ren et al., 2019), and the lack of either Nad4 or Nad5 blocked the assembly of complex I holoenzyme in *Arabidopsis* (Ligas et al., 2019). The loss of the steady-state level of mitochondrial nad5 mature mRNA blocked the assembly of complex I and caused an arrest in endosperm development (Zhang Y. F. et al., 2017). Lastly, the elite haplotypes of *VrEmp24/25* (TT) made up the main proportion of more than 90.9% in cultivated mungbeans, 45% in wild mungbeans (**Figure 4E**). The HSW in hap 1 haplotypes accessions was significantly higher than that in hap 2 and hap 3 (P = 2.11E-29). It was reported that a single amino acid completely prevented the appearance of the enzyme in the medium, and we inferred that the related variation could lead to the change in enzyme activity (East et al., 1990; Alfson et al., 2018).

There have four evidences to take *VrKIX8* as another important seed-size trait gene. Firstly, *VrKIX8* associated with Chr1-8161305-8347626 (LOD = 24.09∼36.33) for HSW and SL were detected in multi-environment by 3VmrMLM (**Figure 5A**; **Supplementary Table 5**). Secondly, *VrKIX8* (LOD = 24.09∼36.33) had significantly differentially expressed between high- and low-HSW accessions (**Figure 3A**). Then, in *Arabidopsis*, the disruption of *KIX8/9* and *PPD1/2* could cause large seeds due to increased cell proliferation and cell elongation in the integuments (Liu et al., 2020a). In soybeans, the loss of the function *GmKIX8-1* showed a significant increase in the size of seeds and leaves. In addition, the increase in organ size was due to the increased cell proliferation, rather than cell expansion. *GmKIX8-1* showed negatively regulated cell proliferation in plants (Nguyen et al., 2021). Lastly, the elite haplotypes of *VrKIX8* (ATCGAA) made up the main proportion of more than 73% in cultivated mungbeans, 40% in wild mungbeans. Moreover, there are four SNPs in the promoter and of *VrKIX8*, and two SNPs in the CDS region, however the amino acid sequence did not change between the elite haplotypes and the other haplotypes (**Figure 5C**). The HSW in hap 1 haplotypes accessions was higher than that in hap 2 to hap 4 (P = 1.19E-21). We supposed that the mutations may have influenced the translation efficiency of *VrKIX8* and caused low expression in cultivated accessions during mungbean domestication.

Genes participate in seed development progress

The genes controlling seed development progress in mungbean are largely unknown (Ha et al., 2021). In this study, we identified fifty-four candidate genes in the seed-development pathways, i.e., *aba2* (Cheng et al., 2014; Chauffour et al., 2019), ABI5 (Lynch et al., 2022), *SHB1*, MINI3, and IKU2 (Garcia et al., 2003; Xiao et al., 2016; Zhang H. et al., 2017), mutants of those genes induced abnormal seed development in *Arabidopsis*. And, five genes were also commonly identified *via* GWAS (**Table 1**). Those five genes (*VrKIX8*, *VrABA2*, *VrSHB1*, *VrABI5*, and *VrIKU2*) are more likely to be reliable, especially for *VrKIX8*, as described above.

We also analyze the possible correlation between the main seed development pathways. Among the 54 genes, five genes (*VrAGG*, *VrDA2*, *VrAGB*, *VrGPA1-1*, and *VrGPA1-2*) consisted of five pairs of significant PPIs. Interestingly, four pairs PPIs were found to be in the G-protein-signaling pathway, and one pair of PPIs was found to be in the G-protein-signaling and the ubiquitin proteasome pathways (**Figure 6**; **Supplementary Table 6**). Ubiquitin proteasome pathway is an important pathway for the selective degradation of proteins and seed development (Smalle and Vierstra, 2004), and the G-protein-signaling pathway is a ubiquitous cell transmembrane signal transduction pathway in eukaryotes (Huang et al., 2006). Moreover, mutations in *GPA1* or *AGB1* could cause short flowers (Lease et al., 2001; Ullah et al., 2001). The overexpression of *AGG3* promoted seed and organ growth by increasing cell proliferation, and loss-of-function mutations in *AGG3* caused small seeds and organs (Chakravorty et al., 2011; Li et al., 2012). The ubiquitin receptor DA1 could control seed size by restricting cell proliferation in maternal integuments (Li et al., 2008). DA1 functioned synergistically with DA2 to restrict seed growth, and DA2 physically interacted with DA1 *in vitro* and *in vivo* (Song et al., 2007; Xia et al., 2013). This interaction could mediate the interactions between the G-protein-signaling pathway and the ubiquitin proteasome pathway, which might offer an important clue in the mechanism analysis of seed development.

In addition, 48 genes had higher expressions in the early stage of seed development than in the late maturation stage of seed development, indicating that seed-development-related genes function primarily in the early stages of seed development, which was consistent with the findings of Zuo et al. (2022) in soybean.

Conclusion

This study conducted GWAS for seed-size-related traits in mungbeans. 98 QTNs and 20 QEIs were identified using 3VmrMLM, while 95, >10,000, and 15 QTNs were identified using EMMAX, GEMMA, and CMLM, respectively. A total of 12 key candidate genes were mined, which were homologous to known seed-development genes in rice and *A. thaliana*. *VrEmp24/25* and *VrKIX8* were identified as main candidate genes around two stable QTNs, the two candidate genes were

further confirmed by RT-qPCR and haplotype analysis, and prevalent haplotypes of *VrEmp24/25* and *VrKIX8* may be useful in mungbean breeding.

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: The WGS sequencing data of 196 mungbean accessions was uploaded to NGDC, with subCRA011538, subSAM100395, and PRJCA010704 ID.

Author contributions

JL, XY, and XC conceived of the project and its components. JL, JC, and YL performed the field experiments. JL, QY, CX, and RW performed the bioinformatics analysis and real data analysis. JL, XC, and XY wrote and revised the manuscript. All authors reviewed the manuscript.

Funding

This work was supported by Natural Science Foundation of Jiangsu Province (BK20190257), National Natural Science Foundation of China (31871696), China Agriculture Research System-Food Legumes (CARS-08), Jiangsu Seed Industry Revitalization Project (JBGS[2021]004), and Jiangsu Planned Projects for Postdoctoral Research Funds (2021K393C).

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.997988/full#supplementary-material

References

Alam, M. K., Islam, M. M., Salahin, N., and Hasanuzzaman, M. (2014). Effect of tillage practices on soil properties and crop productivity in wheat-mungbean-rice cropping system under subtropical climatic conditions. *Sci. World J.* 2014:437283. doi: 10.1155/2014/437283

Alfson, K. J., Avena, L. E., Delgado, J., Beadles, M. W., Patterson, J. L., Carrion, R. Jr., et al. (2018). A single amino acid change in the Marburg virus glycoprotein arises during serial cell culture passages and attenuates the virus in a macaque model of disease. *mSphere* 3:e00401-17. doi: 10.1128/mSphere.00401-17

Bai, F., Yu, L., Shi, J., Li-Beisson, Y., and Liu, J. (2022). Long-chain acyl-CoA synthetases activate fatty acids for lipid synthesis, remodeling and energy production in Chlamydomonas. *New Phytol.* 233, 823–837. doi: 10.1111/nph.17813

Barrett, J. C., Fry, B., Maller, J., and Daly, M. J. (2005). Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263–265. doi: 10.1093/bioinformatics/bth457

Bolger, A. M., Marc, L., and Bjoern, U. (2014). Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics* 30, 2114–2120.

Bonaventure, G., Salas, J. J., Pollard, M. R., and Ohlrogge, J. B. (2003). Disruption of the *FATB* gene in *Arabidopsis* demonstrates an essential role of saturated fatty acids in plant growth. *Plant Cell* 15, 1020–1033. doi: 10.1105/tpc.008946

Chakravorty, D., Trusov, Y., Zhang, W., Acharya, B. R., Sheahan, M. B., McCurdy, D. W., et al. (2011). An atypical heterotrimeric G-protein γ-subunit is involved in guard cell K^+-channel regulation and morphological development in *Arabidopsis thaliana*. *Plant J.* 67, 840–851. doi: 10.1111/j.1365-313X.2011.04638.x

Chauffour, F., Bailly, M., Perreau, F., Cueff, G., Suzuki, H., Collet, B., et al. (2019). Multi-omics analysis reveals sequential roles for ABA during seed maturation. *Plant Physiol.* 180, 1198–1218. doi: 10.1104/pp.19.00338

Cheng, P., Li, H., Yuan, L., Li, H., Xi, L., Zhang, J., et al. (2018). The ERA-related GTPase *AtERG2* associated with mitochondria 18S RNA is essential for early embryo development in *Arabidopsis*. *Front. Plant Sci.* 9:182. doi: 10.3389/fpls.2018.00182

Cheng, Z. J., Zhao, X. Y., Shao, X. X., Wang, F., Zhou, C., Liu, Y. G., et al. (2014). Abscisic acid regulates early seed development in *Arabidopsis* by ABI5-mediated transcription of SHORT HYPOCOTYL UNDER BLUE1. *Plant Cell* 26, 1053–1068. doi: 10.1105/tpc.113.121566

Creff, A., Brocard, L., Joubès, J., Taconnat, L., Doll, N. M., Marsollier, A. C., et al. (2019). A stress-response-related inter-compartmental signalling pathway regulates embryonic cuticle integrity in *Arabidopsis*. *PLoS Genet.* 15:e1007847. doi: 10.1371/journal.pgen.1007847

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158.

Duan, Z., Zhang, M., Zhang, Z., Liang, S., Fan, L., Yang, X., et al. (2022). Natural allelic variation of *GmST05* controlling seed size and quality in soybean. *Plant Biotechnol. J.* 20, 1807–1818. doi: 10.1111/pbi.13865

East, A. K., Curnock, S. P., and Dyke, K. G. (1990). Change of a single amino acid in the leader peptide of a staphylococcal beta-lactamase prevents the appearance of the enzyme in the medium. *FEMS Microbiol. Lett.* 57, 249–254. doi: 10.1016/0378-1097(90)90075-2

Evfratov, S. A., Osterman, I. A., Komarova, E. S., Pogorelskaya, A. M., Rubtsova, M. P., Zatsepin, T. S., et al. (2017). Application of sorting and next generation

sequencing to study 5′-UTR influence on translation efficiency in *Escherichia coli*. *Nucleic Acids Res.* 45, 3487–3502.

Fang, N., Xu, R., Huang, L., Zhang, B., Duan, P., Li, N., et al. (2016). SMALL GRAIN 11 controls grain size, grain number and grain yield in rice. *Rice* 9:64.

Fernandez, G., Shanmugasundaram, S., Shanmugasundaram, S., and Mclean, B. T. (1988). *The AVRDC mungbean improvement program: The past, present and future*. Shanhua: AVRDC.

Garcia, D., Saingery, V., Chambrier, P., Mayer, U., Jürgens, G., and Berger, F. (2003). *Arabidopsis* haiku mutants reveal new controls of seed size by endosperm. *Plant Physiol.* 131, 1661–1670. doi: 10.1104/pp.102.018762

Ge, L., Yu, J., Wang, H., Luth, D., Bai, G., Wang, K., et al. (2016). Increasing seed size and quality by manipulating BIG SEEDS1 in legume species. *Proc. Natl. Acad. Sci. U.S.A.* 113, 12414–12419. doi: 10.1073/pnas.1611763113

Gong, C., Zhao, S., Yang, D., Lu, X., Anees, M., He, N., et al. (2022). Genome-wide association analysis provides molecular insights into the natural variation of watermelon seed size. *Hortic. Res.* 9:uhab074. doi: 10.1093/hr/uhab074

Guo, N., Gu, M., Hu, J., Qu, H., and Xu, G. (2020). Rice OsLHT1 functions in leaf-to-panicle nitrogen allocation for grain yield and quality. *Front. Plant Sci.* 11:1150. doi: 10.3389/fpls.2020.01150

Ha, J., Satyawan, D., Jeong, H., Lee, E., Cho, K. H., Kim, M. Y., et al. (2021). A near-complete genome sequence of mungbean (*Vigna radiata* L.) provides key insights into the modern breeding program. *Plant Genome* 14:e20121. doi: 10.1002/tpg2.20121

Hao, J., Wang, D., Wu, Y., Huang, K., Duan, P., Li, N., et al. (2021). The GW2-WG1-OsbZIP47 pathway controls grain size and weight in rice. *Mol. Plant* 14, 1266–1280. doi: 10.1016/j.molp.2021.04.011

Hu, D., Li, X., Yang, Z., Liu, S., Hao, D., Chao, M., et al. (2022). Downregulation of a gibberellin 3β-hydroxylase enhances photosynthesis and increases seed yield in soybean. *New Phytol.* 235, 502–517. doi: 10.1111/nph.18153

Huang, J., Taylor, J. P., Chen, J. G., Uhrig, J. F., Schnell, D. J., Nakagawa, T., et al. (2006). The plastid protein THYLAKOID FORMATION1 and the plasma membrane G-protein GPA1 interact in a novel sugar-signaling mechanism in *Arabidopsis*. *Plant Cell* 18, 1226–1238. doi: 10.1105/tpc.105.037259

Humphry, M. E., Lambrides, C. J., Chapman, S. C., Aitken, E., and Liu, C. J. (2010). Relationships between hard-seededness and seed weight in mungbean (*Vigna radiata*) assessed by QTL analysis. *Plant Breed.* 124, 292–298.

Islam, M. A., Islam, M. R., Haque, M. E., Yeasmin, F., and Hossain, M. A. (2015). Impacts of famers' participation in upscaling technologies on mungbean (*Vigna radiata* l.) production in the south-western region of Bangladesh. *Agriculturists* 12, 39–47. doi: 10.3329/agric.v12i2.21730

Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., et al. (2009). STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* 37, D412–D416.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42, 348–354.

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

Lease, K. A., Wen, J., Li, J., Doke, J. T., Liscum, E., and Walker, J. C. (2001). A mutant *Arabidopsis* heterotrimeric G-protein beta subunit affects leaf, flower, and fruit development. *Plant Cell* 13, 2631–2641. doi: 10.1105/tpc.010315

Lee, S., Lee, J., Ricachenevsky, F. K., Punshon, T., Tappero, R., Salt, D. E., et al. (2021). Redundant roles of four ZIP family members in zinc homeostasis and seed development in *Arabidopsis thaliana*. *Plant J.* 108, 1162–1173. doi: 10.1111/tpj.15506

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760.

Li, M., Zhang, Y. W., Xiang, Y., Liu, M. H., and Zhang, Y. M. (2022a). IIIVmrMLM: The R and C++ tools associated with 3VmrMLM, a comprehensive GWAS method for dissecting quantitative traits. *Mol. Plant* 15, 1251–1253. doi: 10.1016/j.molp.2022.06.002

Li, M., Zhang, Y. W., Zhang, Z. C., Xiang, Y., Liu, M. H., Zhou, Y. H., et al. (2022b). A compressed variance component mixed model for detecting QTNs, and QTN-by-environment and QTN-by-QTN interactions in genome-wide association studies. *Mol. Plant* 15, 630–650. doi: 10.1016/j.molp.2022.02.012

Li, N., Xu, R., and Li, Y. (2019). Molecular networks of seed size control in plants. *Annu. Rev. Plant Biol.* 70, 435–463.

Li, S., Liu, Y., Zheng, L., Chen, L., Li, N., Corke, F., et al. (2012). The plant-specific G protein γ subunit AGG3 influences organ size and shape in *Arabidopsis thaliana*. *New Phytol.* 194, 690–703. doi: 10.1111/j.1469-8137.2012.04083.x

Li, Y., Zheng, L., Corke, F., Smith, C., and Bevan, M. W. (2008). Control of final seed and organ size by the DA1 gene family in *Arabidopsis thaliana*. *Genes Dev.* 22, 1331–1336. doi: 10.1101/gad.463608

Liaw, A., and Wiener, M. (2002). Classification and regression by randomforest. *R News* 2, 18–22.

Ligas, J., Pineau, E., Bock, R., Huynen, M. A., and Meyer, E. H. (2019). The assembly pathway of complex I in *Arabidopsis thaliana*. *Plant J.* 97, 447–459.

Liu, D., Yu, Z., Zhang, G., Yin, W., Li, L., Niu, M., et al. (2021). Diversification of plant agronomic traits by genome editing of brassinosteroid signaling family genes in rice. *Plant Physiol.* 187, 2563–2576. doi: 10.1093/plphys/kiab394

Liu, J. Y., Chen, J. B., Anochar, K., Lin, Y., Xue, C. C., Wu, R. R., et al. (2022a). *High-quality genome assembly and genome-wide association studies provide genetic insights into natural variation in yield-related traits in mungbean*.

Liu, J. Y., Lin, Y., Chen, J. B., Xue, C. C., Wu, R. R., Yan, Q., et al. (2022b). Identification and clarification of *VrCYCA1*: A key genic male sterility-related gene in mungbean by multi-omics analysis. *Agriculture* 12:686.

Liu, J. Y., Xue, C. C., Lin, Y., Yan, Q., Chen, J. B., Wu, R. R., et al. (2022c). Genetic analysis and identification of *VrFRO8*, a salt tolerance-related gene in mungbean. *Gene* 836:146658. doi: 10.1016/j.gene.2022.146658

Liu, Z., Li, N., Zhang, Y., and Li, Y. (2020a). Transcriptional repression of *GIF1* by the KIX-PPD-MYC repressor complex controls seed size in *Arabidopsis*. *Nat. Commun.* 11:1846. doi: 10.1038/s41467-020-15603-3

Liu, J. Y., Zhang, Y. W., Han, X., Zuo, J. F., Zhang, Z., Shang, H., et al. (2020b). An evolutionary population structure model reveals pleiotropic effects of *GmPDAT* for traits related to seed size and oil content in soybean. *J. Exp. Bot.* 71, 6988–7002. doi: 10.1093/jxb/eraa426

Liu, J. Y., Li, P., Zhang, Y. W., Zuo, J. F., Li, G., Han, X., et al. (2020c). Three-dimensional genetic networks among seed oil-related traits, metabolites and genes reveal the genetic foundations of oil synthesis in soybean. *Plant J.* 103, 1103–1124. doi: 10.1111/tpj.14788

Lu, X., Li, Q. T., Xiong, Q., Li, W., Bi, Y. D., Lai, Y. C., et al. (2016). The transcriptomic signature of developing soybean seeds reveals the genetic basis of seed trait adaptation during domestication. *Plant J.* 86, 530–544. doi: 10.1111/tpj.13181

Luo, J., Liu, H., Zhou, T., Gu, B., Huang, X., Shangguan, Y., et al. (2013). An-1 encodes a basic helix-loop-helix protein that regulates awn development, grain size, and grain number in rice. *Plant Cell* 25, 3360–3376. doi: 10.1105/tpc.113.113589

Lynch, T., Née, G., Chu, A., Krüger, T., Finkemeier, I., and Finkelstein, R. R. (2022). ABI5 binding protein2 inhibits ABA responses during germination without ABA-INSENSITIVE5 degradation. *Plant Physiol.* 189, 666–678. doi: 10.1093/plphys/kiac096

Manan, S., Ahmad, M. Z., Zhang, G., Chen, B., Haq, B. U., Yang, J., et al. (2017). Soybean *LEC2* regulates subsets of genes involved in controlling the biosynthesis and catabolism of seed storage substances and seed development. *Front. Plant Sci.* 8:1604. doi: 10.3389/fpls.2017.01604

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Mei, L., Cheng, X. Z., Wang, S. H., Wang, L. X., and Liu, C. J. (2009). Relationship between bruchid resistance and seed mass in mungbean based on QTL analysis. *Genome* 52, 589–596. doi: 10.1139/G09-031

Moreno, J. E., Moreno-Piovano, G., and Chan, R. L. (2018). The antagonistic basic helix-loop-helix partners BEE and IBH1 contribute to control plant tolerance to abiotic stress. *Plant Sci.* 271, 143–150. doi: 10.1016/j.plantsci.2018.03.024

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628.

Nguyen, C. X., Paddock, K. J., Zhang, Z., and Stacey, M. G. (2021). *GmKIX8-1* regulates organ size in soybean and is the causative gene for the major seed weight QTL qSw17-1. *New Phytol.* 229, 920–934. doi: 10.1111/nph.16928

Orozco-Arroyo, G., Paolo, D., Ezquer, I., and Colombo, L. (2015). Networks controlling seed size in *Arabidopsis*. *Plant Reprod.* 28, 17–32. doi: 10.1007/s00497-015-0255-5

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., and Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* 11, 1650–1667. doi: 10.1038/nprot.2016.095

Plackett, A. R., Powers, S. J., Fernandez-Garcia, N., Urbanova, T., Takebayashi, Y., Seo, M., et al. (2012). Analysis of the developmental roles of the *Arabidopsis* gibberellin 20-oxidases demonstrates that *GA20ox1*, *-2,* and *-3* are the dominant paralogs. *Plant Cell* 24, 941–960. doi: 10.1105/tpc.111.095109

Pongpanich, M., Sullivan, P. F., and Tzeng, J. Y. (2010). A quality control algorithm for filtering SNPs in genome-wide association studies. *Bioinformatics* 26, 1731–1737.

Rampey, R. A., Baldridge, M. T., Farrow, D. C., Bay, S. N., and Bartel, B. (2013). Compensatory mutations in predicted metal transporters modulate auxin conjugate responsiveness in *Arabidopsis*. *G3* 3, 131–141. doi: 10.1534/g3.112.004655

Ren, Z., Fan, K., Fang, T., Zhang, J., Yang, L., Wang, J., et al. (2019). Maize empty pericarp602 encodes a P-type PPR protein that is essential for seed development. *Plant Cell Physiol.* 60, 1734–1746. doi: 10.1093/pcp/pcz083

Schnurr, J., Shockey, J., and Browse, J. (2004). The acyl-CoA synthetase encoded by *LACS2* is essential for normal cuticle development in *Arabidopsis*. *Plant Cell* 16, 629–642. doi: 10.1105/tpc.017608

She, K. C., Kusano, H., Koizumi, K., Yamakawa, H., and Shimada, H. (2010). A novel factor floury endosperm2 is involved in regulation of rice grain size and starch quality. *Plant Cell* 22, 3280–3294. doi: 10.1105/tpc.109.070821

Shin, J. H., Blay, S., Mcneney, B., and Graham, J. (2006). LDheatmap: An R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *J. Stat. Softw.* 16, 1–9.

Singh, A. K., Fu, D. Q., El-Habbak, M., Navarre, D., Ghabrial, S., and Kachroo, A. (2011). Silencing genes encoding omega-3 fatty acid desaturase alters seed size and accumulation of bean pod mottle virus in soybean. *Mol. Plant Microbe Interact.* 24, 506–515. doi: 10.1094/MPMI-09-10-0201

Smalle, J., and Vierstra, R. D. (2004). The ubiquitin 26S proteasome proteolytic pathway. *Annu. Rev. Plant Biol.* 55, 555–590.

Smith, D. S., Maxwell, P. W., and De Boer, S. H. (2005). Comparison of several methods for the extraction of DNA from potatoes and potato-derived products. *J. Agric. Food Chem.* 53, 9848–9859. doi: 10.1021/jf051201v

Somta, P., Ammaranan, C., Ooi, P. A.-C., and Srinives, P. (2007). Inheritance of seed resistance to bruchids in cultivated mungbean (*Vigna radiata*, L. Wilczek). *Euphytica* 155, 47–55. doi: 10.1007/s10681-006-9299-9

Song, X. J., Huang, W., Shi, M., Zhu, M. Z., and Lin, H. X. (2007). A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase. *Nat. Genet.* 39, 623–630. doi: 10.1038/ng2014

Sun, J. Y., Hammerlindl, J., Forseille, L., Zhang, H., and Smith, M. A. (2014). Simultaneous over-expressing of an acyl-ACP thioesterase (FatB) and silencing of acyl-acyl carrier protein desaturase by artificial microRNAs increases saturated fatty acid levels in *Brassica napus* seeds. *Plant Biotechnol. J.* 12, 624–637.

Sun, S., Wang, L., Mao, H., Shao, L., Li, X., Xiao, J., et al. (2018). A G-protein pathway determines grain size in rice. *Nat. Commun.* 9:851. doi: 10.1038/s41467-018-03141-y

Ullah, H., Chen, J. G., Young, J. C., Im, K. H., Sussman, M. R., and Jones, A. M. (2001). Modulation of cell proliferation by heterotrimeric G protein in *Arabidopsis*. *Science* 292, 2066–2069.

Wu, Y., Fu, Y., Zhao, S., Gu, P., Zhu, Z., Sun, C., et al. (2016). CLUSTERED PRIMARY BRANCH 1, a new allele of *DWARF11*, controls panicle architecture and seed size in rice. *Plant Biotechnol. J.* 14, 377–386. doi: 10.1111/pbi.12391

Xia, T., Li, N., Dumenil, J., Li, J., Kamenski, A., Bevan, M. W., et al. (2013). The ubiquitin receptor DA1 interacts with the E3 ubiquitin ligase DA2 to regulate seed and organ size in *Arabidopsis*. *Plant Cell* 25, 3347–3359. doi: 10.1105/tpc.113.115063

Xiao, Y. G., Sun, Q. B., Kang, X. J., Chen, C. B., and Ni, M. (2016). SHORT HYPOCOTYL UNDER BLUE1 or HAIKU2 mixepression alters canola and *Arabidopsis* seed development. *New Phytol.* 209, 636–649.

Xiu, Z., Peng, L., Wang, Y., Yang, H., Sun, F., Wang, X., et al. (2020). Empty *Pericarp24* and empty *Pericarp25* are required for the splicing of mitochondrial introns, complex I assembly, and seed development in maize. *Front. Plant Sci.* 11:608550. doi: 10.3389/fpls.2020.608550

Xu, Q., and Liesche, J. (2021). Sugar export from *Arabidopsis* leaves: Actors and regulatory strategies. *J. Exp. Bot.* 72, 5275–5284. doi: 10.1093/jxb/erab241

Xu, Y., Yang, T., Zhou, Y., Yin, S., Li, P., Liu, J., et al. (2018). Genome-wide association mapping of starch pasting properties in maize using single-locus and multi-locus models. *Front. Plant Sci.* 9:1311. doi: 10.3389/fpls.2018.01311

Yan, Q., Wang, Q., Cheng, X., Wang, L., and Chen, X. (2020). *High-quality genome assembly, annotation and evolutionary analysis of the mungbean (Vigna radiata) genome*. Hoboken, NJ: Authorea. doi: 10.22541/au.160587196.63922177/v1

Zhang, B., Li, C., Li, Y., and Yu, H. (2020). Mobile terminal flower1 determines seed size in *Arabidopsis*. *Nat. Plants* 6, 1146–1157. doi: 10.1038/s41477-020-0749-5

Zhang, H., Cheng, F., Xiao, Y., Kang, X., Wang, X., Kuang, R., et al. (2017). Global analysis of canola genes targeted by SHORT HYPOCOTYL UNDER BLUE 1 during endosperm and embryo development. *Plant J.* 91, 158–171. doi: 10.1111/tpj.13542

Zhang, W., Xu, W., Zhang, H., Liu, X., Cui, X., Li, S., et al. (2021). Comparative selective signature analysis and high-resolution GWAS reveal a new candidate gene controlling seed weight in soybean. *Theor. Appl. Genet.* 134, 1329–1341. doi: 10.1007/s00122-021-03774-6

Zhang, Y. F., Suzuki, M., Sun, F., and Tan, B. C. (2017). The mitochondrion-targeted PENTATRICOPEPTIDE REPEAT78 protein is required for nad5 mature mRNA stability and seed development in maize. *Mol. Plant* 10, 1321–1333. doi: 10.1016/j.molp.2017.09.009

Zhang, Y. M., Jia, Z., and Dunwell, J. M. (2019). Editorial: The applications of new multi-locus GWAS methodologies in the genetic dissection of complex traits. *Front. Plant Sci.* 10:100. doi: 10.3389/fpls.2019.00100

Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42, 355–360.

Zhang, Z., Zhao, H., Huang, F., Long, J., Song, G., and Lin, W. (2019). The 14-3-3 protein GF14f negatively affects grain filling of inferior spikelets of rice (*Oryza sativa* L.). *Plant J.* 99, 344–358. doi: 10.1111/tpj.14329

Zhao, X. Y., Wang, J. G., Song, S. J., Wang, Q., Kang, H., Zhang, Y., et al. (2016). Precocious leaf senescence by functional loss of PROTEIN S-ACYL TRANSFERASE14 involves the NPR1-dependent salicylic acid signaling. *Sci. Rep.* 6:20309. doi: 10.1038/srep20309

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44, 821–824.

Zuo, J. F., Ikram, M., Liu, J. Y., Han, C. Y., Niu, Y., Dunwell, J. M., et al. (2022). Domestication and improvement genes reveal the differences of seed size- and oil-related traits in soybean domestication and improvement. *Comput. Struct. Biotechnol. J.* 20, 2951–2964. doi: 10.1016/j.csbj.2022.06.014

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 17 October 2022
DOI 10.3389/fpls.2022.995609

OPEN ACCESS

EDITED BY
Zhenyu Jia,
University of California, Riverside, United States

REVIEWED BY
Jia Wen,
University of North Carolina at Chapel Hill, United States
Jianfang Zuo,
Huazhong Agricultural University, China
HaiYan Lü,
Henan Agricultural University, China

*CORRESPONDENCE
Yangjun Wen
wenyangjun@njau.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

RECEIVED 16 July 2022
ACCEPTED 27 September 2022
PUBLISHED 17 October 2022

CITATION
Zhang J, Wang S, Wu X, Han L, Wang Y and Wen Y (2022) Identification of QTNs, QTN-by-environment interactions and genes for yield-related traits in rice using 3VmrMLM. *Front. Plant Sci.* 13:995609. doi: 10.3389/fpls.2022.995609

Identification of QTNs, QTN-by-environment interactions and genes for yield-related traits in rice using 3VmrMLM

Jin Zhang[1,2†], Shengmeng Wang[1†], Xinyi Wu[1], Le Han[1], Yuan Wang[1] and Yangjun Wen[1,2*]

[1]College of Science, Nanjing Agricultural University, Nanjing, China, [2]Key Laboratory of Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing, China

Rice, which supports more than half the population worldwide, is one of the most important food crops. Thus, potential yield-related quantitative trait nucleotides (QTNs) and QTN-by-environment interactions (QEIs) have been used to develop efficient rice breeding strategies. In this study, a compressed variance component mixed model, 3VmrMLM, in genome-wide association studies was used to detect QTNs for eight yield-related traits of 413 rice accessions with 44,000 single nucleotide polymorphisms. These traits include florets per panicle, panicle fertility, panicle length, panicle number per plant, plant height, primary panicle branch number, seed number per panicle, and flowering time. Meanwhile, QTNs and QEIs were identified for flowering times in three different environments and five subpopulations. In the detections, a total of 7~23 QTNs were detected for each trait, including the three single-environment flowering time traits. In the detection of QEIs for flowering time in the three environments, 21 QTNs and 13 QEIs were identified. In the five subpopulation analyses, 3~9 QTNs and 2~4 QEIs were detected for each subpopulation. Based on previous studies, we identified 87 known genes around the significant/suggested QTNs and QEIs, such as LOC_Os06g06750 (*OsMADS5*) and LOC_Os07g47330 (*FZP*). Further differential expression analysis and functional enrichment analysis identified 30 candidate genes. Of these candidate genes, 27 genes had high expression in specific tissues, and 19 of these 27 genes were homologous to known genes in *Arabidopsis*. Haplotype difference analysis revealed that LOC_Os04g53210 and LOC_Os07g42440 are possibly associated with yield, and LOC_Os04g53210 may be useful around a QEI for flowering time. These results provide insights for future breeding for high quality and yield in rice.

KEYWORDS

3VmrMLM, GWAS, QTN-by-environment interaction, yield traits, rice

Introduction

Rice (*Oryza sativa* L.), one of the most important food crops, supports more than half the population in the world. Therefore, rice is crucial to improving the safety, quality, stability, and sustainability of the global food supply (Muthayya et al., 2014). In China, rice production is second only to maize, accounting for 31.64% of the total grain produced in 2020 (http://www.stats.gov.cn/tjsj/ndsj/, accessed on June 2022). Moreover, from 1994 to 2020, rice accounted for 27.17% of the total grain produced in the world, which is 657.85 million tons per year (http://www.fao.org/faostat/en/#data/QC/visualize, accessed on June 2022). There is an urgent, ongoing global demand for highly productive rice varieties due to growth in the human population in particular in developing nations, in which rice is the primary source of calories (Toriyama, 2005); climate change; and the labor-, land-, and water-intensive nature of rice cultivation (Greenland, 1997). Furthermore, climate has an impact on the most crucial traits of rice, such as production and quality. Weather catastrophes are becoming increasingly severe across the world because of accelerating global climate change, which poses a significant challenge to the production of sustainable food. Developing resilient crops is an efficient strategy for coping with climate change. A wealth of plant breeding and genomic resources have been developed by the scientific community to assist in this endeavor, including high-quality genome sequences (Goff et al., 2002; Yu et al., 2002), dense SNP maps (McNally et al., 2009; Ebana et al., 2010; Huang et al., 2010), extensive germplasm collections (Ebana et al., 2008; McNally et al., 2009; Agrama et al., 2010), and public databases of genomic information (Tanaka et al., 2008; McNally et al., 2009; Huang et al., 2010; Youens-Clark et al., 2011). Yet despite the emergence of these scientific resources, traditional quantitative trait locus linkage mapping is most often used to understand the genetic structures of complex traits in rice.

Genome-wide association study (GWAS) mapping enables the simultaneous screening of huge numbers of accessions for genetic variation in a variety of complex traits. Humongous genetic variants for agronomic and economic traits have been extensively studied using single-locus GWAS methods, such as MLM (Zhang et al., 2005; Yu et al., 2006), EMMA (Kang et al., 2008), and GEMMA (Zhou and Stephens, 2012). Such single-locus GWAS methods have a limited ability in detecting quantitative trait nucleotides (QTNs) with marginal effects that are affected by the polygenic background and stringent Bonferroni correction (Wang et al., 2016). Even if adjusting for polygenic background enhances the statistical power of QTN detection, it is still difficult to identify the majority of small-effect QTNs related to complex traits using single-locus GWAS methods.

To address the issue in single-locus GWAS methods, multi-locus GWAS methods were developed as a multidimensional method of genome analysis, which simultaneously estimate the effects of all markers (Cui et al., 2018). In particular, to address the selection of cofactors in multi-locus GWAS models with millions of markers, researchers have proposed MLMM (Segura et al., 2012), FarmCPU (Liu et al., 2016), mrMLM (Wang et al., 2016), pLARmEB (Zhang et al., 2017), and FASTmrEMMA (Wen et al., 2018). However, the dominance (d) or QTN-by-environment interaction (QEI) were not fully considered in the above models. Moreover, when additive (a) and dominance (d) effects, additive-by-environment (a×e) interaction, dominance-by-environment (d×e) interaction, and their polygenic backgrounds are simultaneously included as random effects in a mixed model of genome-wide analysis, there are 10 variance components, which creates a huge computational burden.

To improve calculation efficiency, a mixed model with three variance components was combined with mrMLM to establish a new methodological framework, namely, 3VmrMLM, that identifies all types of loci and estimates their effects while controlling all possible polygenic backgrounds (Li et al., 2022a). In GWAS, QEI can be used extensively to explore the genetic structures of complex traits to meet the needs of phenotypic plasticity research and global climate change. 3VmrMLM was expanded to cover QEI using the same thinking as in QTN detection models.

The data set of 413 rice accessions with 44,000 SNPs from the Rice Diversity database (www.ricediversity.org, accessed on April 2022) is suitable for GWAS, which has been performed by many researchers. Although this data set contains a wealth of information, including data on yield-related traits closely related to human life, phenotypic data on a given trait in different locations, and data on different subpopulations with the same trait, it has been seldom studied for further both QTN and QEI detection simultaneously. Therefore, in this study, we reanalyzed eight yield-related traits in this natural population of 413 rice accessions using the proposed multi-locus method, 3VmrMLM. Our goals were to detect the significant QTNs and QEIs related to rice yield, mine candidate genes, speed up molecular marker-assisted breeding, and increase rice production.

Material and methods

Phenotypic data and statistical analysis

We used 3VmrMLM (Li et al., 2022a, 2022b) to reanalyze 413 accessions with 36,901 SNPs in rice (*Oryza sativa* L.) in Zhao et al. (2011) to detect significant QTNs and QEIs for eight yield-related traits. Phenotypic data were downloaded from the Rice Diversity database (www.ricediversity.org, accessed on April 2022). The yield-related agronomic traits were florets per panicle (FPP), panicle fertility (PF), panicle length (PL), panicle number per plant (PNPP), plant height (PH), primary panicle

branch number (PPBN), seed number per panicle (SNPP), and flowering time in three environments, Aberdeen (FTAB), Arkansas (FTAR), and Faridpur (FTF). In Zhao et al. (2011), detailed information on the experimental designs is described. Flowering time at the three locations (FTAB, FTAR, and FTF) was used to detect QEIs for multi-environment analysis and also to detect QTNs for single-environment analysis. The other seven traits were phenotyped at the same locations for single-environment analysis to detect QTNs in this study.

To illustrate the variability of gene-environment interactions in subpopulations in rice, we also analyzed rice flowering time in FTAB, FTAR, and FTF for five subpopulations derived from Zhao et al. (2011), including Admixed (ADMIX), Australia (AUS), Indica (IND), Temperate japonica (TEJ), and Tropical japonica (TRJ), with sample sizes of 43, 50, 52, 69, and 78, respectively.

To visualize all eight traits, descriptive statistical analysis for each phenotypic data was performed, including the mean, minimum, maximum, range, standard deviation, and coefficient of variation (CV) for each trait (Table 1). Pearson correlation analysis (Figure 1) for all phenotypic data was performed in R version 4.1.2 (https://www.r-project.org/).

Genotypic data

Genotypic data for the 413 rice accessions were obtained from the Rice Diversity database (www.ricediversity.org, accessed on April 2022). The data set consisted of a well-distributed 36,901 SNP array across the 12 chromosomes of rice with call rate > 70% and minor allele frequency > 0.01 (Zhao et al., 2011). To visualize the genotype in this study, Figures 2A, B illustrate the distribution of the minor allele frequency and the density distribution of loci on each chromosome. These were relatively uniform, which indicates that this data set is suitable for genetics dissection in rice.

GWAS

The IIIVmrMLM software (Li et al., 2022b) of 3VmrMLM method (Li et al., 2022a) was downloaded from github (https://github.com/YuanmingZhang65/IIIVmrMLM). We performed QTN and QEI detection using the *IIIVmrMLM* function, specifying the parameters of "=*Single_env*" for the QTN detection model and "=*Multi_env*" for the QEI detection model. The thresholds of significant and suggested QTN or QEI were set at P-value = 0.05/m and LOD = 3.00, respectively, where m is the number of markers (Li et al., 2022a).

SNP annotation and the identification of known genes

The China Rice Data Center database (https://ricedata.cn/, accessed on June 2022) was used to annotate the genes around significant/suggested QTNs and QEIs identified by 3VmrMLM. For all identified loci, regions within 200 kb were used to search for known genes (which were reported in previous studies and identified by 3VmrMLM simultaneously) according to linkage disequilibrium decay.

Functional enrichment analysis and the identification of candidate genes

We performed differential expression analysis using the online tool GEO2R (https://www.ncbi.nlm.nih.gov/geo/geo2r/, accessed on September 2022) on four data sets (GSE19024, GSE21396, GSE136746, and GSE166053) from the Gene Expression Omnibus database (https://www.ncbi.nlm.nih.gov/geo/, accessed on September 2022). The datasets contain transcriptomic data related to rice development. Differentially expressed genes (DEGs) were screened by adjusted P-values less

TABLE 1 Statistical analysis of eight rice yield-related traits.

Trait	Mean	Max	Min	SD	CV
FPP	5.056	5.836	3.909	0.323	0.064
PF	0.824	0.980	0.372	0.105	0.127
PL	24.375	35.683	15.633	3.537	0.145
PNPP	3.247	4.172	2.234	0.413	0.127
PH	116.583	194.333	67.750	21.092	0.181
PPBN	9.943	17.000	5.556	1.781	0.179
SNPP	4.854	5.635	3.445	0.330	0.068
FTAB[a]	107.050	306.000	45.000	38.957	0.364
FTAR[a]	87.944	150.500	54.500	12.627	0.144
FTF[a]	71.770	110.000	39.000	8.510	0.119

[a]indicates flowering time in three different environments in the single-environment analysis.

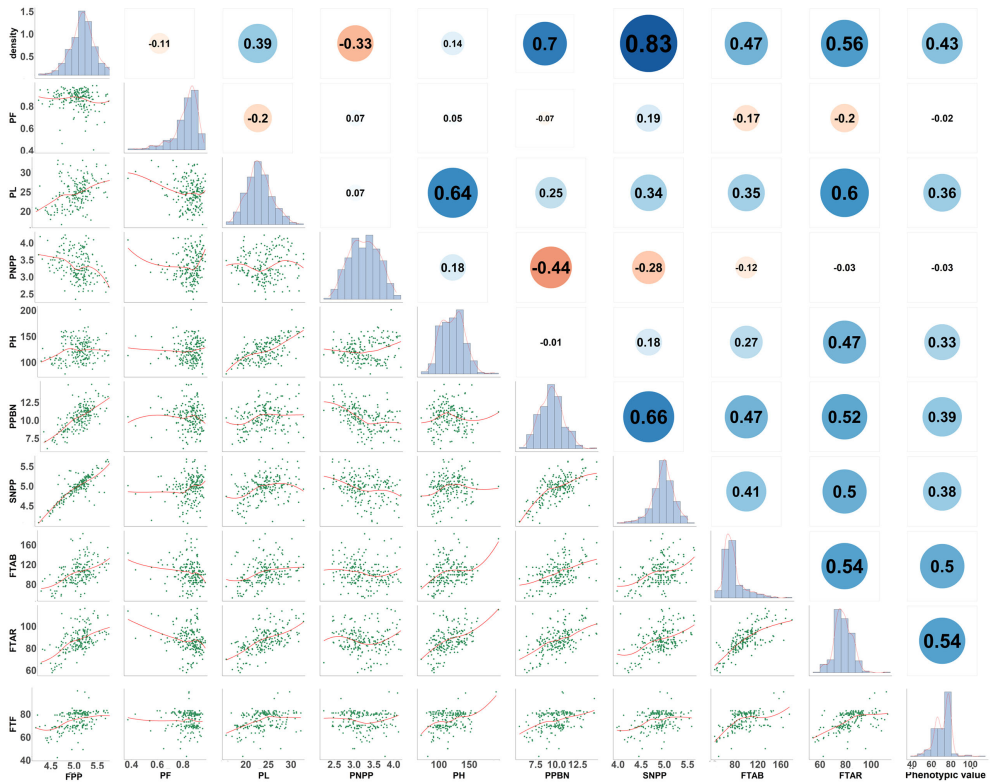


FIGURE 1
Distribution of eight yield-related traits in rice and Pearson coefficients. FTAB, FTAR, and FTF are the flowering time in three different environments in the single-environment analysis. Linear regression statistics between the two traits are below the diagonal, the diagonal histogram represents the distribution of each trait, and correlation coefficients are above the diagonal (positive numbers represent positive correlations, negative numbers represent negative correlations).

than 0.05, and then intersected with genes around significant/suggested QTNs or QEIs to obtain DEGs significantly associated with the target traits. For the functional annotation analysis, information of the above DEGs related to the target traits was submitted to the web-based tool DAVID (https://david.ncifcrf.gov/home.jsp, accessed on September 2022) to perform Kyoto Encyclopedia of Genes and Genomes functional enrichment analysis. Fisher's exact test ($P < 0.05$) was

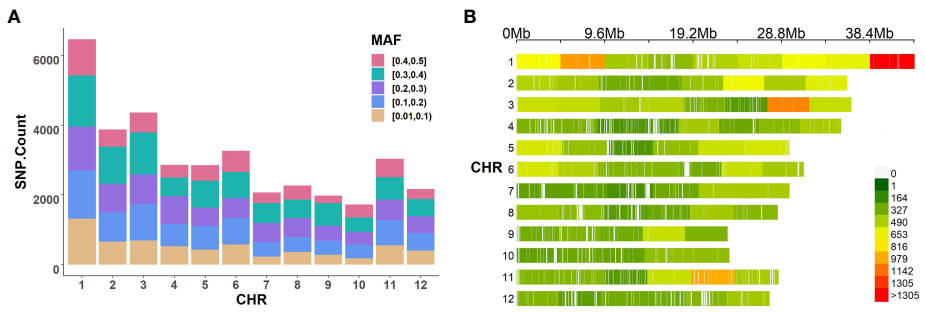


FIGURE 2
The distribution of SNPs in rice. **(A)** The distribution of minor allele frequency. **(B)** The density distribution of SNPs.

used to select enrichment KEGG pathways. Genes that were enriched in these significant pathways were considered as candidate genes.

Tissue specific expression and blast of homologous genes in *Arabidopsis*

The database Rice Genome Annotation Project (http://rice.uga.edu/, accessed on September 2022) was used to investigate the expression of all candidate genes in various tissues to further illustrate the association between genes and phenotypic variations. The R package *pheatmap* was used to create a heatmap of the FPKM expression of the candidate genes. Protein sequence information of the candidate genes was submitted to the Rice Genome Annotation Project (http://rice.uga.edu/analyses_search_blast.shtml, accessed on September 2022) to mine homologous *Arabidopsis* genes.

Analysis of haplotype and phenotypic difference

To validate the associated loci between candidate genes and traits, the *HaploView* software package (http://www.broad.mit.edu/mpg/haploview/; Barrett et al., 2005) was used to perform linkage disequilibrium and haplotype block analyses and to estimate the frequency of haplotype populations in candidate genes. For each gene, significant variations were used for haplotype division, and the phenotypic differences between haplotypes was analyzed via *t* test using the *t.test* function in R.

Results

Phenotypic variation

Eight yield-related traits (including FPP, PF, PL, PNPP, PH, PPBN, SNPP, and flowering time in FTAB, FTAR, and FTF) were reanalyzed to determine whether there exists any significant genetic variation in these traits across 413 rice accessions. Descriptive statistics for all traits are listed in Table 1. Let us consider CV as an example, for flowering time in each single-environment, FTAB had the highest CV at 36.4%, which indicates that flowering time at Aberdeen had the largest variation. Furthermore, the CVs for FTAR and FTF were 14.4% and 11.9%, both relatively large, which indicates large variation and environmentally sensitive for flowering time. In addition, the CVs for the other six traits (PF, PL, PNPP, PH, PPBN, and SNPP) were 12.7%, 14.5%, 12.7%, 18.1%, 17.9%, and 6.8%, and FPP had the lowest CV at 6.4%.

Pearson correlation coefficients (PCCs) were calculated among the eight traits (Figure 1). FPP and PNPP were negatively correlated (PCC = –0.33), and a negative correlation was also observed between FPP and PF (PCC = –0.11). FPP was positively correlated with PPBN (PCC = 0.7) and SNPP (PCC = 0.83). In addition, PL was positively correlated with PH (PCC = 0.64), SNPP (PCC = 0.34), and FPP (PCC = 0.39), which indicates the close genetic relationship between panicle length and panicle number. With regard to flowering time across environments, FTAB was positively correlated with FTAR (PCC = 0.74) and FTF (PCC = 0.50), and FTAR and FTF were positively correlated (PCC = 0.54). These results demonstrate that the eight rice traits play a crucial role in controlling the rice yield and significantly correlate to one another.

Identification of QTNs for yield-related traits using 3VmrMLM

We reanalyzed all eight yield-related traits using the single-environment QTN detection model in 3VmrMLM to identify QTNs, where flowering time was measured in three different environments. A total of 165 significant/suggested QTNs (Supplementary Table S1; Supplementary Figure S1) were detected as associated with at least one of the eight yield-related traits. Of these QTNs, 17, 16, 16, 21, 23, 17, 15, 15, 18, and 7 QTNs (Supplementary Table S1; Supplementary Figure S1) were associated with FPP, PF, PL, PNPP, PH, PPBN, SNPP, FTAB, FTAR, and FTF, respectively. The proportion of total phenotypic variance explained by QTNs for each single trait were 72.61%, 73.29%, 75.48%, 51.99%, 64.17%, 71.64%, 58.55%, 58.04%, 77.07%, and 44.60% calculated by the R package *IIIVmrMLM*. It shows that most QTNs had only additive effects. Note that some QTNs, such as id3005865 for FPP, id5014747 for PF, and id4007762 for PH, had both additive and dominance effects.

A total of 17 QTN hotspots (Supplementary Table S1; Supplementary Figure S1A) were detected as significantly associated with FPP, with P-values of 2.19E-32~7.60E-07 and LOD scores of 5.31~31.66, respectively. A total of 16 QTNs (Supplementary Table S1; Supplementary Figure S1B) associated with PF were detected with P-values of 1.08E-44~1.07E-06 and LOD scores of 4.97~32.90. A total of 16 QTNs (Supplementary Table S1; Supplementary Figure S1C) were associated with PL, with P-values of 2.33E-56~1.02E-05 and LOD scores of 4.23~54.34, and id7004886 located on chromosome 7 had the maximum phenotypic variance explained at 22.04% (Supplementary Table S1). Moreover, 21 QTNs (Supplementary Table S1; Supplementary Figure S1D) associated with PNPP were detected with P-values of 1.78E-37~9.71E-06. For PH, 23 QTNs (Supplementary Table S1; Supplementary Figure S1E) were detected with P-values of 1.15E-38~7.60E-05 and LOD scores of 3.40~37.94. A total of 18 QTNs (Supplementary Table S1; Supplementary Figure S1F) were detected as associated with

PPBN; they were widely located on chromosomes 1, 2, 4, and 9, with P-values of 2.05E-39~1.31E-05 and LOD scores of 4.13~37.47. Among these QTNs, id1009181 located on chromosome 1 explained 16.03% of the phenotypic variance. For SNPP, 15 QTNs (Supplementary Table S1; Supplementary Figure S1G) were detected with P-values of 3.95E-41~2.11E-05 and LOD scores of 3.93~39.18. For the three flowering time environments, 30 QTNs (Supplementary Table S1; Supplementary Figure S1H–J) were detected on all chromosomes except chromosome 12 were detected, with P-values of 1.15E-32~2.17E-05 and LOD scores of 1.40~11.30. id4000121, ud7002024, and id4004217 explained the maximum phenotypic variance, which were 14.47%, 11.30%, and 7.75%, respectively.

Known genes around significant/suggested QTNs

We compared genomic regions of 165 significant/suggested QTNs (200 kb up- and down-stream of each significant/suggested QTN) to the genomic positions of reported genes related to rice yield. A total of 73 known genes were around the significant/suggested QTNs, including 9, 7, 3, 14, 17, 6, 6, 2, 7, and 2 known genes for FPP, PF, PL, PNPP, PH, PPBN, SNPP, FTAB, FTAR, and FTF, respectively (Table 2; Supplementary Figure S1). Marker id1019150 located on chromosome 1 around LOC_Os01g54810 was simultaneously associated with PL and PH (Table 2; Supplementary Figure S1). Moreover, id1002863 and id7004587 around LOC_Os01g07480 and LOC_Os07g41250, respectively, on chromosomes 1 and 7 were associated with FPP and SNPP (Table 2; Supplementary Figure S1). It is interesting that a QTN can overlap with multiple known genes (e.g., three genes, LOC_Os02g45054, LOC_Os02g45070, and LOC_Os02g45110 were simultaneously around id2012042 on chromosome 2, Table 2; Supplementary Figure S1). *sd1* is associated with PH (Zhao et al., 2011). Moreover, *OsRA2*, located on chromosome 1 and simultaneously associated with FPP and SNPP, modifies panicle architecture by regulating pedicel length (Leran et al., 2014; Lu et al., 2017). *OsPTR4* controls FPP and SNPP (Leran et al., 2014).

Detection of QEIs for rice flowering time using 3VmrMLM

In the multi-environment analysis, flowering time at three locations (Aberdeen, Arkansas, and Faridpur) was reanalyzed using the QEI detection model in 3VmrMLM to identify QEIs. A total of 21 significant/suggested QTNs (Table 3; Supplementary Figure S2A) and 13 significant/suggested QEIs (Table 4; Supplementary Figure S2B) were simultaneously detected.

TABLE 2 Known genes identified for rice yield-related traits using the QTN detection model in 3VmrMLM.

Trait	Marker	Chr	Position	add	dom	Variance	r^2(%)	Gene Symbol	ID
FPP	id1002863	1	3481990	-0.049	–	0.002	1.700	*OsRA2*	LOC_Os01g07480
	id1002863	1	3481990	-0.049	–	0.002	1.700	*FIB*	LOC_Os01g07500
	id1003144	1	3801746	-0.057	–	0.003	2.030	*OsRE1*	LOC_Os01g07880
	id3000495	3	871080	0.117	–	0.007	5.714	*Ehd4*[b]	LOC_Os03g02160
	id7004587	7	24790535	0.056	–	0.002	1.338	*OsPTR4*	LOC_Os07g41250
	id7004587	7	24790535	0.056	–	0.002	1.338	*OsMADS18*	LOC_Os07g41370
	id7005660	7	28221129	-0.124	–	0.010	7.409	*OsCOL13*	LOC_Os07g47140
	id7005660	7	28221129	-0.124	–	0.010	7.409	*FZP*	LOC_Os07g47330
	id12009959	12	27218159	0.083	–	0.007	5.143	*OsPAP10c*	LOC_Os12g44020
PF	id1023500	1	37274860	-0.024	–	0.000	2.519	*OsABI5*	LOC_Os01g64000
	id1023500	1	37274860	-0.024	–	0.000	2.519	*REL1*	LOC_Os01g64380
	id3000828	3	1499569	-0.036	–	0.001	7.357	*OsmiR528*	LOC_Os03g03724
	id8007916	8	28208958	0.035	–	0.000	1.994	*OsNTL5*	LOC_Os08g44820
	id9002415	9	7894310	0.027	–	0.001	3.126	*OsEMF2b*	LOC_Os09g13630
	id9002415	9	7894310	0.027	–	0.001	3.126	*SDG724*	LOC_Os09g13740
	id12006848	12	21130413	0.046	–	0.000	2.8543	*OsVIL2*	LOC_Os12g34850
PL	id1017530	1	29565162	0.788	–	0.574	3.375	*OsLFL1*	LOC_Os01g51610
	id1019150	1	31662509	-1.275	–	1.134	6.672	*THIS1*[b]	LOC_Os01g54810
	id10003476	10	13216045	-1.350	–	1.544	9.088	*Brd2*	LOC_Os10g25780
PNPP	id1001128	1	1401052	-0.047	–	0.002	1.257	*MHZ4*	LOC_Os01g03750
	id2000516	2	647801	-0.086	–	0.003	1.662	*DHD4*	LOC_Os02g01990

(Continued)

TABLE 2 Continued

Trait	Marker	Chr	Position	add	dom	Variance	r^2(%)	Gene Symbol	ID
	id2012042	2	27371812	-0.131	–	0.004	2.237	*SID1*	LOC_Os02g45054
	id2012042	2	27371812	-0.131	–	0.004	2.237	*OsAGO1a*	LOC_Os02g45070
	id2012042	2	27371812	-0.131	–	0.004	2.237	*OsMTA2*	LOC_Os02g45110
	id3003977	3	7327105	-0.088	–	0.008	4.398	*OsAPC6*	LOC_Os03g13370
	id3003977	3	7327105	-0.088	–	0.008	4.398	*LPA1*	LOC_Os03g13400
	id3006138	3	12008635	0.046	–	0.002	1.185	*OsPHR1*	LOC_Os03g21240
	id4010447	4	30843940	-0.062	–	0.004	2.125	*OsAP2-39*	LOC_Os04g52090
	id5011783	5	25197731	0.109	–	0.012	6.746	*OsmtSSB1*	LOC_Os05g43440
	id7000258	7	1588172	-0.079	–	0.005	2.844	*OSH15*	LOC_Os07g03770
	id8001120	8	3438707	0.050	–	0.002	1.427	*OsCOMT*	LOC_Os08g06100
	id8001120	8	3438707	0.050	–	0.002	1.427	*OsCCA1*	LOC_Os08g06110
	ud8000279	8	4363409	-0.071	–	0.005	2.710	*DTH8*	LOC_Os08g07740
PH	id1018978	1	31452220	-4.470	–	13.040	2.931	*OsCesA4*	LOC_Os01g54620
	id1024441	1	38537795	7.133	–	17.140	3.853	*sd1*[b]	LOC_Os01g66100
	id1018978	1	31452220	-4.470	–	13.040	2.931	*THIS1*	LOC_Os01g54810
	id1018978	1	31452220	-4.470	–	13.040	2.931	*OsVOZ1*	LOC_Os01g54930
	id1024441	1	38537795	7.133	–	17.140	3.853	*OsCrll3*	LOC_Os01g66590
	id4007762	4	23286717	-7.695	3.480	15.326	3.445	*TDD1*	LOC_Os04g38950
	id4007762	4	23286717	-7.695	3.480	15.326	3.445	*OsALDH10A5*	LOC_Os04g39020
	id4007762	4	23286717	-7.695	3.480	15.326	3.445	*d11*	LOC_Os04g39430
	id4010574	4	31138553	3.210	–	9.981	2.243	*OsAP2-39*	LOC_Os04g52090
	id4010574	4	31138553	3.210	–	9.981	2.243	*OsKS1*	LOC_Os04g52230
	id4010574	4	31138553	3.210	–	9.981	2.243	*FC1*	LOC_Os04g52280
	id6004564	6	7097190	-3.206	–	8.230	1.850	*YPD1*	LOC_Os06g13050
	wd6000736	6	10282460	-3.939	–	10.962	2.464	*OsNF-YB9*	LOC_Os06g17480
	id7005417	7	27547556	-2.096	–	4.341	0.976	*Fd-GOGAT1*	LOC_Os07g46460
	id8006905	8	24940725	5.109	–	17.476	3.928	*RCN11*	LOC_Os08g39380
	id8006905	8	24940725	5.109	–	17.476	3.928	*OsDOG*	LOC_Os08g39450
	id9007929	9	22920706	2.891	–	7.000	1.574	*OsDRP1E*	LOC_Os09g39960
PPBN	id1009181	1	13926463	-0.807	–	0.640	16.026	*IPI1*	LOC_Os01g24880
	id1014302	1	24275703	-0.447	–	0.174	4.361	*OsATG7*	LOC_Os01g42850
	id1022478	1	35621886	0.593	–	0.335	8.400	*LAX1*	LOC_Os01g61480
	id1022478	1	35621886	0.593	–	0.335	8.400	*OsBAG4*	LOC_Os01g61500
	id1024948	1	39308177	-0.460	–	0.122	3.050	*EG1*	LOC_Os01g67430
	id3005659	3	10842947	0.479	–	0.091	2.291	*SSD1*	LOC_Os03g19080
SNPP	id1002863	1	3481990	-0.049	–	0.002	1.894	*OsRA2*	LOC_Os01g07480
	id1013159	1	22950277	0.171	–	0.004	3.481	*LOG*	LOC_Os01g40630
	id3005721	3	10922512	0.087	–	0.003	2.449	*SDG718*	LOC_Os03g19480
	id3005721	3	10922512	0.087	–	0.003	2.449	*SRL2*	LOC_Os03g19520
	id6015132	6	26966327	0.061	–	0.004	3.275	*OsSPL10*	LOC_Os06g44860
	id7004587	7	24790535	0.074	–	0.003	2.665	*OsPTR4*	LOC_Os07g41250
FTAB[a]	id1027324	1	42152363	-10.893	–	27.056	1.783	*OsMLH1*	LOC_Os01g72880
	id6002745	6	3330294	9.162	–	80.673	5.316	*OsMADS5*	LOC_Os06g06750
FTAR[a]	id1021120	1	34082456	-3.468	–	11.006	5.050	*OsGCD1*	LOC_Os01g58750
	id3002064	3	3766414	-4.491	–	19.461	8.930	*DPW*[b]	LOC_Os03g07140
	id3002064	3	3766414	-4.491	–	19.461	8.930	*CYP704B2*[b]	LOC_Os03g07250
	id3002064	3	3766414	-4.491	–	19.461	8.930	*OsSUT1*[b]	LOC_Os03g07480
	ud7001067	7	15702110	4.474	–	17.692	8.119	*ORMDL*	LOC_Os07g26940

(Continued)

TABLE 2 Continued

Trait	Marker	Chr	Position	add	dom	Variance	r^2(%)	Gene Symbol	ID
	id9006822	9	19210667	-2.851	-8.826	3.567	1.637	*OsDFR2A*	LOC_Os09g32025
	id11011548	11	28322308	3.318	–	2.462	1.130	*EDT1*	LOC_Os11g47330
FTF[a]	id4004217	4	14176927	2.677	–	5.804	7.747	*OsACOS12*	LOC_Os04g24530
	id6006288	6	10090472	1.900	–	3.609	4.329	*OsNF-YB9*	LOC_Os06g17480

"-" indicates no dominance effect for this QTN. [a]indicates flowering time in three different environments in the single-environment analysis. [b]indicates known gene which was detected by 3VmrMLM and EMMA simultaneously.

Among them, id6006118 located on chromosome 6 had additive-by-environment interaction and dominance-by-environment interaction in all three environments.

We compared genomic regions of the significant/suggested QTNs or QEIs (200 kb up- and down-stream around the significant/suggested QTNs or QEIs) to the positions of previously reported genes related to rice flowering time. 4 QTNs (Table 3; Supplementary Figure S2A) and 1 QEI (Table 4; Supplementary Figure S2B) overlapped with the known genes. Notably, id6002690, which was adjacent to LOC_Os06g06750 (*OsMADS5*), was demonstrated to have both QTN and QEI effects. Microarray-based expression profiling and genome-wide molecular characterization of the genes that encode the MADS-box transcription factor family was presented by Arora et al. (2007). *OsMADS5* in this gene family is associated with the development of inflorescence. Recently, Zhu et al. (2022) also revealed the function of *OsMADS5* in the development of inflorescence and showed that *OsMADS5* is involved in limiting branching and promoting the transition to spikelet meristem identity, partly by repressing RCN4 expression.

For five different subpopulations (ADMIX, AUS, IND, TEJ, and TRJ), flowering time in FTAB, FTAR, and FTF was also analyzed to illustrate the variability in gene-environment interactions. A total of 25 QTNs and 15 QEIs (Supplementary Table S2; Supplementary Figures S2C–L) were simultaneously detected with the multi-environment detection model in 3VmrMLM, including 3, 3, 6, 9, and 4 QTNs and 4, 2, 4, 3, and 2 QEIs for ADMIX, AUS, IND, TEJ, and TRJ, respectively. Note that there was no overlap in QEI between different subpopulations, which may indicate that these QEIs come from different ecological adaptations.

TABLE 3 Significant/suggested QTNs for rice flowering time in three environments detected using the QTN-by-environment detection model in 3VmrMLM.

Marker	CHR	Positions	LOD	add	dom	Variance	r^2(%)	P-value	Reported Gene	Reference
id1001009	1	1095730	9.492	2.180	–	4.486	1.147	3.810E-11	–	–
id1007272	1	9815262	19.756	-3.224	–	3.810	0.974	1.456E-21	–	–
id1008137	1	11376832	16.619	-2.991	–	8.260	2.112	2.169E-18	-	–
id1012744	1	22493100	11.948	2.765	–	7.583	1.939	1.192E-13	*SaF*	Xie et al., 2017
id1014639	1	24595570	9.958	-2.241	–	2.662	0.681	1.272E-11	-	–
ud3000099	3	1400496	24.223	-3.680	–	12.997	3.323	4.490E-26	-	–
id3004539	3	8656816	34.038	-4.261	–	14.612	3.736	5.814E-36	*OsSTRL2*	Zou et al., 2017
id3008283	3	16551139	4.319	1.506	–	2.092	0.535	8.203E-06	-	–
dd3001061	3	27836287	17.230	3.294	–	8.646	2.211	5.220E-19	-	–
id4001482	4	3628149	8.959	2.138	–	4.134	1.057	1.335E-10	-	–
id4005251	4	17893016	9.178	-2.435	–	5.598	1.431	7.982E-11	-	–
id5000013	5	44370	11.405	2.395	–	4.778	1.222	4.259E-13	-	–
id5008977	5	21268048	21.538	3.386	–	6.812	1.742	2.302E-23	-	–
id5012857	5	26783289	13.681	2.640	–	2.135	0.546	2.068E-15	-	–
id6002690	6	3289852	27.063	3.816	–	13.218	3.380	6.148E-29	*OsMADS5*	Arora et al., 2007; Zhu et al., 2022
id6005322	6	8185001	49.742	-5.502	–	8.540	2.184	9.554E-52	-	–
ud7000660	7	8553942	15.086	-2.944	–	6.661	1.703	7.754E-17	-	–
id7004583	7	24784697	24.039	3.582	–	7.380	1.887	6.889E-26	*OsUAM3*	Konishi et al., 2007
id8000022	8	51045	23.711	-3.509	–	7.457	1.907	1.475E-25	–	–
id10000202	10	1012769	32.396	-4.148	–	12.160	3.109	2.618E-34	–	–
id110107061	11	26711260	13.869	-2.654	–	2.831	0.724	1.332E-15	–	–

"-" indicates no dominance effect or reported gene for this QTN.

TABLE 4 Significant/suggested QEIs for rice flowering time in three environments detected using the QTN-by-environment detection model in 3VmrMLM.

Marker	CHR	Positions	LOD	add1	dom1	add2	dom2	add3	dom3	Variance	r^2 (%)	P-value	Reported Gene	Reference
id1000015	1	149005	21.318	4.569	–	-1.336	–	-3.233	–	11.037	2.822	4.819E-22	–	–
id1000947	1	1042817	10.498	-3.007	–	0.422	–	2.586	–	5.303	1.356	3.180E-11	–	–
id1008137	1	11376832	14.763	-3.871	–	1.145	–	2.726	–	7.910	2.023	1.729E-15	–	–
ud2000978	2	17730153	9.539	-3.236	–	1.034	–	2.202	–	5.465	1.397	2.893E-10	–	–
id4002940	4	8211710	16.440	-4.001	–	1.251	–	2.749	–	8.377	2.142	3.637E-17	–	–
id5000766	5	1128994	10.402	-2.418	–	-0.647	–	3.064	–	5.218	1.334	3.971E-11	–	–
id6002690	6	3289852	6.107	1.782	–	0.622	–	-2.404	–	3.113	0.796	7.813E-07	*OsMADS5*	Arora et al., 2007; Zhu et al., 2022
id6005330	6	8234981	8.562	-2.930	–	1.146	–	1.785	–	4.362	1.115	2.747E-09	–	–
id6006118	6	9651785	33.572	-5.915	-0.010	2.086	1.406	3.829	-1.396	17.945	4.588	2.106E-32	–	–
id6007539	6	12322330	20.565	4.618	–	-1.690	–	-2.928	–	10.917	2.791	2.730E-21	–	–
id7004142	7	23351238	8.336	2.802	–	-0.792	–	-2.010	–	4.174	1.067	4.615E-09	–	–
id10006353	10	20022516	13.901	4.031	–	-1.259	–	-2.772	–	8.506	2.175	1.259E-14	–	–
id11006398	11	17823963	15.862	-3.984	–	1.500	–	2.484	–	8.097	2.070	1.377E-16	–	–

"-" indicates no dominance effect or reported gene for this QEI.

Functional enrichment analysis of candidate genes

In addition to the aforementioned significant/suggested QTNs and QEIs with known genes, we also detected several new QTNs and QEIs that have not been reported in previous studies, such as id2005901, id6007721, id12008098, and id9001769 (Supplementary Table S1; Supplementary Figure S1). To identify the candidate genes, we considered genes in regions 200 kb up- and down-stream around each significant/suggested QTN and QEI, including all studies of population and each subpopulation. There are about 8000 genes within these 200kb regions, of which 755 are DEGs that show different expression between test and control groups of rice accessions.

In the Kyoto Encyclopedia of Genes and Genomes analysis, 30 genes significantly involved in 4 biological processes (terpenoid backbone biosynthesis, butanoate metabolism, carbon metabolism, and alanine, aspartate and glutamate metabolism) were defined as candidate genes. Figure 3A shows results for the candidate genes in the rectangular boxes, the most significant pathways are marked in red.

The results of the functional enrichment analysis (Figure 3A) showed that some candidate genes around the new QTNs and QEIs were involved in many biological and metabolic processes during rice growth, which have not been reported in previous studies, such as flower development, which indicates that these candidate genes have a non-negligible influence on the target traits. For example, LOC_Os01g02020 (Figure 3A), a candidate gene detected in PH and SNPP, was involved in terpenoid backbone biosynthesis, butanoate metabolism, and carbon metabolism. In addition, the candidate gene LOC_Os04g52450 (Figure 3A) was directly involved in butanoate metabolism and in alanine, aspartate, and glutamate metabolism. Moreover, some candidate genes detected in the multi-environment analysis for each subpopulation, including LOC_Os03g16050 for IND, LOC_Os04g53210 for AUS, and LOC_Os07g09060 and LOC_Os07g09190 for TRJ (Figure 3A), were involved in a series of biological and metabolic processes.

Expression profile of candidate genes

The Rice Genome Annotation Project database (http://rice.uga.edu) demonstrates the expression of the candidate genes in various tissues or organs, including shoots, roots, seeds, leaves,

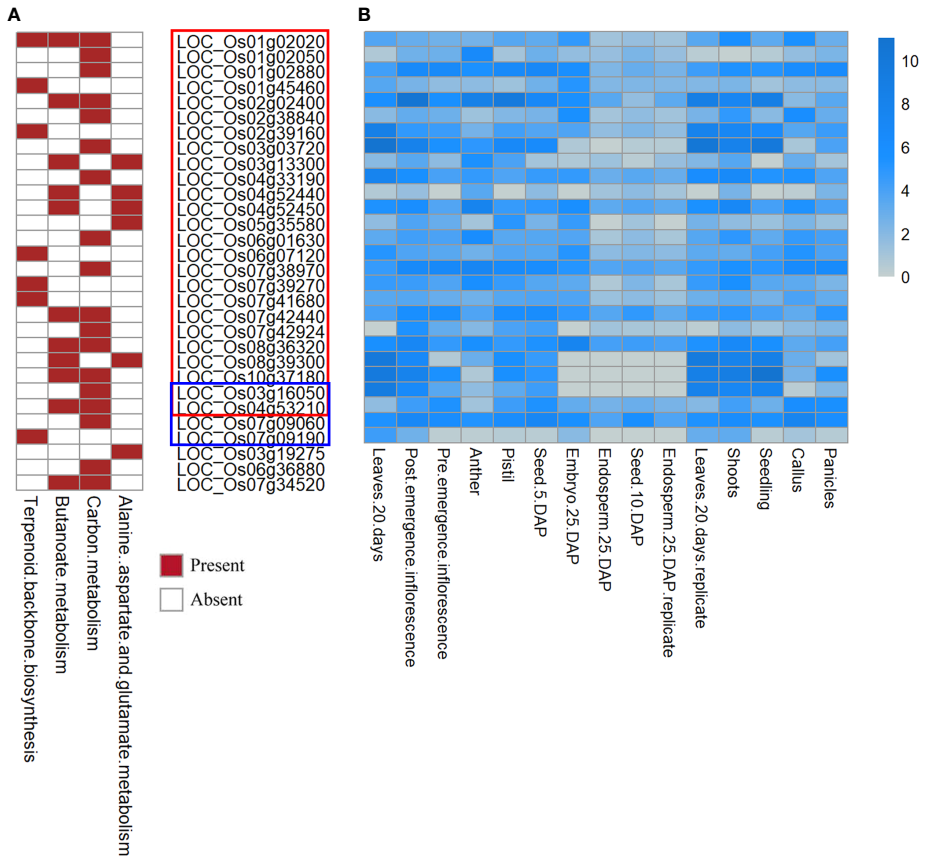


FIGURE 3
Heatmap of the functional enrichment analysis and tissue-specific expression analysis. **(A)** Heatmap of the functional enrichment analysis for the candidate genes. **(B)** Heatmap of FPKM expression for the part of candidate genes. The y-axis is log2(FPKM+1). Candidate genes in the red box correspond to QTNs. Blue box: QEIs for flowering time, remaining: candidate genes not expressed in specific tissue.

panicles, anthers, pistils, post-emergence, pre-emergence, and embryos. The heatmap of the candidate genes presented in Figure 3B shows the FPKM expression of the candidate genes in tissues and organs.

For QTN, LOC_Os04g52450 and LOC_Os08g36320 had high expression in leaves, panicles, shoots, and seedlings in rice (Figure 3B). Furthermore, LOC_Os03g16050 had the highest expression in pre-emergence inflorescence, leaves, shoots, and seedlings. Some earlier studies (Zhao et al., 2011; Weng et al., 2014) suggested that inflorescence, anthers, pistils, and panicles play important roles in regulating yield.

For QEIs of flowering time, LOC_Os03g16050, LOC_Os04g53210, LOC_Os07g09060 had high expression in post-emergence inflorescence and pre-emergence inflorescence, which might indicate a potential association between these candidate genes and flowering time (Figure 3B).

Among the 30 candidate genes, LOC_Os03g19275, LOC_Os06g36880, and LOC_Os07g34520 were not expressed in panicles or inflorescence; thus, these genes were not considered in further analyses. Among the 27 candidate genes identified here after tissue-specific expression analysis, 19 candidate genes are listed in Table 5 for their homologous *Arabidopsis* genes.

Haplotype and phenotypic difference analysis of candidate genes

To further verify the association between the candidate genes and target traits, we performed haplotype analysis of the candidate genes using SNPs within the candidate genes and 2 kb upstream of the candidate genes. LOC_Os04g53210 (CDS coordinates [5′-3′]: 31688717 ~ 31692592) was analyzed to reveal the intragenic variation affecting the rice yield and to identify favorable haplotypes. Figure 4A shows the linkage disequilibrium and haplotype block with two SNPs (id4010894 at 31688182 bp and id4010904 at 31691252 bp). The 413 accessions were classified into 4 haplotypes based on these two

10.3389/fpls.2022.995609

TABLE 5 Orthologous information of candidate genes with higher tissue expression.

Trait	gene	Marker	*Arabidopsis* Orthologous gene	Putative function
FT_Q/ TRJ_Q	LOC_Os01g02880	id1001009/ id1001003	*AT2G01140*	Aldolase superfamily protein
TEJ_Q	LOC_Os01g45460	id1015276	*AT1G26120/ AT3G02410/ AT5G15860*	alpha/beta-Hydrolases superfamily protein/prenylcysteine methylesterase
PPBN	LOC_Os02g38840	id2009400	*AT3G27300/ AT5G40760*	glucose-6-phosphate dehydrogenase 6
PPBN	LOC_Os02g39160	id2009400	*AT5G60600*	4-hydroxy-3-methylbut-2-enyl diphosphate synthase
PNPP	LOC_Os03g13300	id3003977	*AT5G17330*	glutamate decarboxylase
IND_QE/ FT_Q	LOC_Os03g16050	id3004734/ id3004539	*AT3G54050*	high cyclic electron flow 1
PF	LOC_Os04g33190	id4006172	*AT5G36880*	acetyl-CoA synthetase
AUS_QE/ FTAB/FTF	LOC_Os04g53210	id4010914/ id4010930/ id4010984	*AT4G18360*	Aldolase-type TIM barrel family protein
FT_Q	LOC_Os05g35580	id5008977	*AT2G16570/ AT4G34740*	GLN phosphoribosyl pyrophosphate amidotransferase 1/GLN phosphoribosyl pyrophosphate amidotransferase 2
PL/PH	LOC_Os06g01630	id6000302	*AT1G54220/ AT3G13930*	Dihydrolipoamide acetyltransferase, long form protein
FTAB/ FT_Q	LOC_Os06g07120	id6002745/ id6002690	*AT2G17570*	Undecaprenyl pyrophosphate synthetase family protein
TRJ_QE	LOC_Os07g09060	id7000656	*AT2G14170*	aldehyde dehydrogenase 6B2
FT_QE	LOC_Os07g38970	id7004142	*AT5G08300/ AT5G23250*	Succinyl-CoA ligase, alpha subunit
FT_QE	LOC_Os07g39270	id7004142	*AT2G18620/ AT4G36810*	Terpenoid synthases superfamily protein/ geranylgeranyl pyrophosphate synthase 1
FT_Q/FPP/ SNPP	LOC_Os07g41680	id7004583/ id7004587	*AT2G17570*	Undecaprenyl pyrophosphate synthetase family protein
PH	LOC_Os07g42440	id7004779	*AT3G14130/ AT3G14150*	Aldolase-type TIM barrel family protein
PL/FPP	LOC_Os07g42924	id7004886/ id7004865	*AT1G22430/ AT1G22440/ AT4G22110*	GroES-like zinc-binding dehydrogenase family protein/Zinc-binding alcohol dehydrogenase family protein/GroES-like zinc-binding dehydrogenase family protein
PH	LOC_Os08g39300	id8006905	*AT2G13360*	alanine: glyoxylate aminotransferase
FT_QE	LOC_Os10g37180	id10006353	*AT1G32470/ AT2G35370*	Single hybrid motif superfamily protein/ glycine decarboxylase complex H

Q and QE indicate significant/suggested QTNs and QEIs in the multi-environment analysis, respectively. AUS, IND, TEJ, and TRJ indicate subpopulations of the 413 rice accessions. Other abbreviations indicate results of the single-environment analysis.

SNPs (id4010894 and id4010904). Among these haplotypes, haplotypes TT and CT had the highest mean phenotypic values of FTAB (109.54) and FTF (78.25), respectively, whereas haplotype TC presented the lowest FTAB (87.33) and FTF (60.00; Figures 4B, C). A *t* test showed that significant differences in FTAB and FTF existed between haplotypes CT and TT (P-values = 4.93E-02 and 3.84E-04, respectively). There was also a significant difference in FTF between haplotypes CT and CC (P-values = 1.23E-04). Therefore, we infer the candidate gene LOC_Os04g53210 to be associated with flowering in rice.

LOC_Os04g53210 was also detected in the multi-environment analysis for the AUS subpopulation. Supplementary Figure S3A shows the differences in phenotype among the 4 haplotypes. Supplementary Figure S3B shows the results of the haplotype block and phenotype difference in LOC_Os07g42440, which was detected in PH. We infer that the candidate gene LOC_Os04g53210 might be a gene-environment interaction for flowering time and that LOC_Os07g42440 might be associated with yield in rice.

Discussion

Classic single-locus methods, such as MLM and general linear model (GLM), have been used extensively to detect genetic variants in many cereals (Price et al., 2006; Sant'Ana

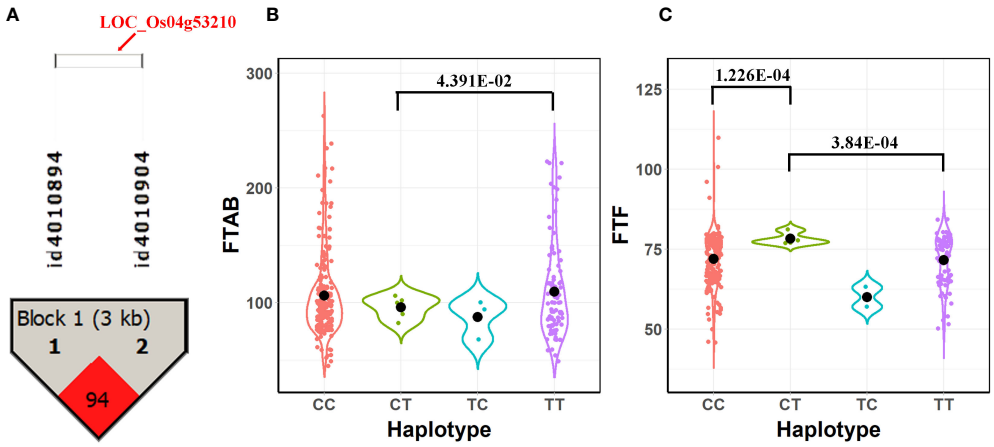


FIGURE 4
Results of haplotype and phenotypic difference analysis for the candidate gene LOC_Os04g53210. **(A)** Linkage disequilibrium and haplotype block with two SNPs inside for LOC_Os04g53210. **(B)** Comparison of FTAB among haplotypes CT, CC, TC, and TT. **(C)** Comparison of FTF among haplotypes CT, CC, TC, and TT.

et al., 2018; He et al., 2019). However, these models suffer from multiple test corrections (e.g., Bonferroni correction) for critical values and neglect the overall effects of multiple loci (Zhong et al., 2021). For example, many robust quantitative trait loci, in particular small-effect quantitative trait loci, are missing because of the stringent threshold (Zhang et al., 2005). Therefore, multi-locus GWAS models, which are relatively closer to the real genetic architecture of animals and plants, have been developed. Geneticists developed these models to reduce the bias associated with estimating effects by controlling the population structure and polygenic background (Zhang et al., 2005; Yu et al., 2006; Zhang et al., 2010). In this study, a multi-locus GWAS method 3VmrMLM was used to detect QTNs for eight yield-related traits in 413 rice varieties with 36,901 SNPs. We detected 17, 16, 16, 21, 23, 17, 15, 15, 18, and 7 significant/suggested SNPs and 9, 7, 3, 14, 17, 6, 6, 2, 7, and 2 known genes for FPP, PF, PL, PNPP, PH, PPBN, SNPP, FTAB, FTAR, and FTF, respectively, using the QTN detection model in 3VmrMLM (Supplementary Table S1). Furthermore, we compared 3VmrMLM to a single-locus method, EMMA (Kang et al., 2008) by Zhao et al. (2011). We detected 4, 3, 3, 6, 5, 2, 1, 14, 6, and 2 QTNs by EMMA; thus, 3VmrMLM detected more significant QTNs than EMMA. Among these significant QTNs, 1, 1, 1, 1, 1, 0, 0, 1, 1, and 0 were detected by the two methods simultaneously, including id3000495, id2004552, id1019150, id12008894, id1101154, id8006573, and id3002064. 1, 0, 2, 1, 1, 0, 0, 1, 3, and 1 known gene were detected by EMMA, which were less than 3VmrMLM. Among these known genes, 6 were detected by EMMA and 3VmrMLM simultaneously, including *End4*, *TH1S1*, *sd1*, *DPW*, *CYP704B2*, and *OsSUT1* (Table 2). In addition to these 6 known genes, we identified 3 candidate genes for EMMA by performing differential expression analysis and functional enrichment analysis, and there was no overlap in candidate genes between the two methods. Moreover, the QTNs detected by 3VmrMLM explained a higher proportion of total phenotypic variance (72.61%, 73.29%, 75.48%, 51.99%, 64.17%, 71.64%, 58.55%, 77.07%, and 44.60%) than those detected by EMMA (17.1%, 8.1%, 10.9%, 7%, 38.6%, 6%, 0.1%, 31.3%, and 8.1%), except for FTAB. Overall, the multi-locus GWAS method are flexible to detect more QTNs and validate more known genes and candidate genes than the single-locus GWAS method.

The contribution of QEI to the genetic analysis of complex traits in plant, animal, and human genetics is growing. As a result of accelerating global climate change, weather disasters in a variety of regions are becoming increasingly severe, posing a substantial obstacle to sustainable food production. An efficient way of adapting to climate change is to develop climate-resilient crops. However, it is first necessary to detect QEIs and mine their genes. In addition, the environment has an impact on important traits, such as quality, yield, adaptability, and resistance, but studies on physiological effects, molecular mechanisms, and functional analyses of QEI genes under a variety of environments are not insightful enough because of the algorithms used. Moreover, joint analysis of multiple environments can enhance statistical power and experimental accuracy in the detection of QTN and QEI. In this study, three flowering time environments were used to identify QEIs for rice using a multi-environment detection model in 3VmrMLM, and 21, 3, 3, 6, 9, and 4 QTNs and 13, 4, 2, 4, 3, and 2 QEIs were detected for all populations and each subpopulation (Tables 3, 4; Supplementary Table S2).

Pleiotropy was verified in this study. Among all the 165 significant/suggested QTNs for the eight traits detected using the QTN detection model in 3VmrMLM, some QTNs were significantly associated with more than one trait. 5 QTNs simultaneously related to FPP and SNPP were detected because of the strong correlation (PCC = 0.83) between these two traits, including id1002863, id3000495, id6009226, id7004587, and id11010822. Around these 5 QTNs, genes the *OsRA2*, *Ehd4*, and *OsPTR4* genes were identified (Gao et al., 2013; Lu et al., 2017; Huang et al., 2019). Id2005901 located on chromosome 2 was associated with both FPP and PPBN (PCC = 0.70). For PH and PL with a positive correlation (PCC = 0.64), id6000302 located on chromosome 6 was simultaneously detected. Moreover, id11011548 located on chromosome 11 was found to affect both PH and FTAR (PCC = 0.47), where the *EDT1* gene was identified (Bai et al., 2019).

Among the total of 117 genes around the significant/suggested QTNs and QEIs in this study, 87 were known genes that have been reported in previous studies. For these known genes with QTN effects (Table 2), *sd1*is associated with PH (Zhao et al., 2011). *OsMADS18* from the MADS-box transcription factor family affects panicle development (Kobayashi et al., 2012). Moreover, *OsRA2*, located on chromosome 1, which simultaneously affects FPP and SNPP, modifies panicle architecture by regulating pedicel length (Lu et al., 2017). Notably, *OsMADS5* was demonstrated to have both QTN effect and QEI effect, which was associated with inflorescence development in several previous studies (Arora et al., 2007; Zhu et al., 2022).

In addition to the above-mentioned 87 known genes, 30 candidate genes around the significant/suggested QTNs and QEIs that have not previously been reported were also detected in this study. These candidate genes were shown to be involved in many biological processes of rice growth, which indicates underlying associations between the identified candidate genes and the target traits (Figure 3A). Among these 30 candidate genes, 27 candidate genes had high expression in specific tissues, such as panicles and inflorescence (Figure 3B). In addition, 19 candidate genes associated with different traits had homologous genes in *Arabidopsis* (Table 5). LOC_Os04g53210 and LOC_Os07g42440 were demonstrated to be potentially associated with flowering and yield, respectively, by haplotype and phenotypic difference analysis (Figure 4; Supplementary Figure S3B). LOC_Os04g53210 especially might be a key gene in gene-environment interaction for flowering time (Supplementary Figure S3A).

3VmrMLM represents a significant advancement in GWAS methodologies and practical applications. First, 3VmrMLM correctly detects both QTNs and QEIs and produces unbiased estimations of their effects, unlike current GWAS methods that only detect QTNs and estimate genetic effects (Li et al., 2022a). Second, despite the fact that Feldmann et al. (2021) discovered that the phenotypic variance explained and the percentage of marker-associated genetic variance of large-effect loci were overestimated in analyses of complex traits, maximum likelihood estimation using ANOVA with the linear invariance property theoretically guarantees accurate loci detection and unbiased estimation of effects. Moreover, 3VmrMLM uses a compressed mixed model with three variance components to overcome the huge computational burden in traditional GWAS models. Therefore, 3VmrMLM is a good choice for detecting QTNs and QEIs associated with rice yield-related traits.

Conclusion

In this study, a compressed mixed model with three variance components in GWAS, 3VmrMLM, was used to detect QTNs and QEIs related to rice yield traits. A total of 165 QTNs were identified. Moreover, 75 known genes were identified adjacent to the QTNs based on genome annotation and previous studies. In terms of QTN-by-environment detection, 21, 3, 3, 6, 9, and 4 QTNs and 13, 4, 2, 4, 3, and 2 QEIs were detected for all populations and each subpopulation. Moreover, 12 known genes were identified adjacent to the QTNs and QEIs. As a result of further differential expression and functional enrichment analysis, 30 candidate genes were detected. LOC_Os04g53210 and LOC_Os07g42440 were confirmed as main candidate genes by tissue-specific expression analysis, comparison of homologous *Arabidopsis* genes, and haplotype and phenotypic difference analysis. LOC_Os04g53210 might be useful in gene-environment interaction for a flowering time trait. These results could be helpful for detecting genes related to rice yield.

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

JZ and SW drafted the manuscript. JZ, SW, XW, LH, and YW analyzed the data. YJW and JZ conceived the study and were in charge of direction and planning. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by grants from the National Natural Science Foundation of China (32070688, 32270694, 31701071), the Postdoctoral Science Foundation of Jiangsu (2020Z330), and the Fundamental Research Funds for the Central Universities (JCQY202108).

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.995609/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Manhattan plots of the single-environment analysis for eight yield-related traits of rice. FTAB, FTAR, and FTF are the flowering time in three different environments. Pink text: known genes for the corresponding significant/suggested SNPs.

SUPPLEMENTARY FIGURE 2
Manhattan plots of the multi-environment analysis for the flowering time of rice. **(A, B)** Manhattan plots of QTNs and QTN-by-environment interactions for all populations. **(C–L)** Manhattan plots of QTNs and QTN-by-environment interactions for each subpopulation. Pink text: known genes for the corresponding significant/suggested SNPs.

SUPPLEMENTARY FIGURE 3
Results of haplotype and phenotypic difference analysis for the candidate genes. **(A)** LOC_Os04g53210. **(B)** LOC_Os07g42440.

References

Agrama, H. A., Yan, W., Jia, M., Fjellstrom, R., and McClung, A. M. J. N. S. (2010). Genetic structure associated with diversity and geographic distribution in the USDA rice world collection. *Natural Sci.* 2 (04), 247. doi: 10.4236/ns.2010.24036

Arora, R., Agarwal, P., Ray, S., Singh, A. K., Singh, V. P., Tyagi, A. K., et al. (2007). MADS-box gene family in rice: genome-wide identification, organization and expression profiling during reproductive development and stress. *BMC Genomics* 8, 242. doi: 10.1186/1471-2164-8-242

Bai, W., Wang, P., Hong, J., Kong, W., Xiao, Y., Yu, X., et al. (2019). Earlier degraded Tapetum1 (EDT1) encodes an ATP-citrate lyase required for tapetum programmed cell death. *Plant Physiol.* 181 (3), 1223–1238. doi: 10.1104/pp.19.00202

Barrett, J. C., Fry, B., Maller, J., and Daly, M. J. (2005). Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263–265. doi: 10.1093/bioinformatics/bth457

Cui, Y. R., Zhang, F., and Zhou, Y. L. (2018). The application of multi-locus GWAS for the detection of salt-tolerance loci in rice. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01464

Ebana, K., Kojima, Y., Fukuoka, S., Nagamine, T., and Kawase, M. (2008). Development of mini core collection of Japanese rice landrace. *Breed. Sci.* 58 (3), 281–291. doi: 10.1270/jsbbs.58.281

Ebana, K., Yonemaru, J., Fukuoka, S., Iwata, H., Kanamori, H., Namiki, N., et al. (2010). Genetic structure revealed by a whole-genome single-nucleotide polymorphism survey of diverse accessions of cultivated Asian rice (*Oryza sativa* l.). *Breed. Sci.* 60 (4), 390–397. doi: 10.1270/jsbbs.60.390

Feldmann, M. J., Piepho, H. P., Bridges, W. C., and Knapp, S. J. (2021). Average semivariance yields accurate estimates of the fraction of marker-associated genetic variance and heritability in complex trait analyses. *PLos Genet.* 17 (8), e1009762. doi: 10.1371/journal.pgen.1009762

Gao, H., Zheng, X. M., Fei, G., Chen, J., Jin, M., Ren, Y., et al. (2013). Ehd4 encodes a novel and oryza-genus-specific regulator of photoperiodic flowering in rice. *PLos Genet.* 9 (2), e1003281. doi: 10.1371/journal.pgen.1003281

Goff, S. A., Ricke, D., Lan, T. H., Presting, G., Wang, R., Dunn, M., et al. (2002). A draft sequence of the rice genome (*Oryza sativa* l. ssp. japonica). *Science* 296 (5565), 92–100. doi: 10.1126/science.1068275

Greenland, D. J. (1997). *The sustainability of rice farming* (Cab International), Wallingford, Oxon, UK.

He, L. Q., Xiao, J., Rashid, K. Y., Yao, Z., Li, P. C., Jia, G. F., et al. (2019). Genome-wide association studies for pasmo resistance in flax (Linum usitatissimum l.). *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01982

Huang, W., Nie, H., Feng, F., Wang, J., Lu, K., and Fang, Z. (2019). Altered expression of OsNPF7.1 and OsNPF7.4 differentially regulates tillering and grain yield in rice. *Plant Sci.* 283, 23–31. doi: 10.1016/j.plantsci.2019.01.019

Huang, X. H., Wei, X. H., Sang, T., Zhao, Q. A., Feng, Q., Zhao, Y., et al. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* 42 (11), 961–U976. doi: 10.1038/ng.695

Kang, H. M., 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 (3), 1709–1723. doi: 10.1534/genetics.107.080101

Kobayashi, K., Yasuno, N., Sato, Y., Yoda, M., Yamazaki, R., Kimizu, M., et al. (2012). Inflorescence meristem identity in rice is specified by overlapping functions of three AP1/FUL-like MADS box genes and PAP2, a SEPALLATA MADS box gene. *Plant Cell* 24 (5), 1848–1859. doi: 10.1105/tpc.112.097105

Konishi, T., Takeda, T., Miyazaki, Y., Ohnishi-Kameyama, M., Hayashi, T., O'Neill, M. A., et al. (2007). A plant mutase that interconverts UDP-arabinofuranose and UDP-arabinopyranose. *Glycobiology* 17 (3), 345–354. doi: 10.1093/glycob/cwl081

Leran, S., Varala, K., Boyer, J. C., Chiurazzi, M., Crawford, N., Daniel-Vedele, F., et al. (2014). A unified nomenclature of NITRATE TRANSPORTER 1/PEPTIDE TRANSPORTER family members in plants. *Trends Plant Sci.* 19 (1), 5–9. doi: 10.1016/j.tplants.2013.08.008

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 (2), e1005767. doi: 10.1371/journal.pgen.1005767

Li, M., Zhang, Y. W., Zhang, Z. C., Xiang, Y., Liu, M. H., Zhou, Y. H., et al. (2022a). A compressed variance component mixed model for detecting QTNs and

QTN-by-environment and QTN-by-QTN interactions in genome-wide association studies. *Mol. Plant* 15 (4), 630–650. doi: 10.1016/j.molp.2022.02.012

Li, M., Zhang, Y. W., Xiang, Y., Liu, M. H., and Zhang, Y. M. (2022b). IIIVmrMLM: The r and c++ tools associated with 3VmrMLM, a comprehensive GWAS method for dissecting quantitative traits. *Mol. Plant* 15 (8), 1251–1253. doi: 10.1016/j.molp.2022.06.002

Lu, H., Dai, Z., Li, L., Wang, J., Miao, X., and Shi, Z. (2017). OsRAMOSA2 shapes panicle architecture through regulating pedicel length. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.01538

McNally, K. L., Childs, K. L., Bohnert, R., Davidson, R. M., Zhao, K., Ulat, V. J., et al. (2009). Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proc. Natl. Acad. Sci. United States America* 106 (30), 12273–12278. doi: 10.1073/pnas.0900992106

Muthayya, S., Sugimoto, J. D., Montgomery, S., and Maberly, G. F. (2014). An overview of global rice production, supply, trade, and consumption. *Ann. N Y Acad. Sci.* 1324, 7–14. doi: 10.1111/nyas.12540

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38 (8), 904–909. doi: 10.1038/ng1847

Sant'Ana, G. C., Pereira, L. F. P., Pot, D., Ivamoto, S. T., Domingues, D. S., Ferreira, R. V., et al. (2018). Genome-wide association study reveals candidate genes influencing lipids and diterpenes contents in coffea arabica l. *Sci. Rep.* 8, 465. doi: 10.1038/s41598-017-18800-1

Segura, V., Vilhjalmsson, B. J., Platt, A., Korte, A., Seren, U., Long, Q., et al. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* 44 (7), 825–U144. doi: 10.1038/ng.2314

Tanaka, T., Antonio, B. A., Kikuchi, S., Matsumoto, T., Nagamura, Y., Numa, H., et al. (2008). The rice annotation project database (RAP-DB): 2008 update. *Nucleic Acids Res.* 36, D1028–D1033. doi: 10.1093/nar/gkm978

Toriyama, K. (2005). Rice is life scientific perspectives for the 21st century. (Tsukuba, Japan:Proceedings of the World Rice Research Conference).

Wang, S. B., Feng, J. Y., Ren, W. L., Huang, B., Zhou, L., Wen, Y. J., et al. (2016). Improving power and accuracy of genome-wide association studies *via* a multi-locus mixed linear model methodology. *Sci. Rep.* 6, 19444. doi: 10.1038/srep19444

Weng, X. Y., Wang, L., Wang, J., Hu, Y., Du, H., Xu, C. G., et al. (2014). Grain number, plant height, and heading Date7 is a central regulator of growth, development, and stress response. *Plant Physiol.* 164 (2), 735–747. doi: 10.1104/pp.113.231308

Wen, Y. J., Zhang, H. W., Ni, Y. L., Huang, B., Zhang, J., Feng, J. Y., et al. (2018). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Briefings Bioinf.* 19 (4), 700–712. doi: 10.1093/bib/bbx028

Xie, Y., Niu, B., Long, Y., Li, G., Tang, J., Zhang, Y., et al. (2017). Suppression or knockout of SaF/SaM overcomes the sa-mediated hybrid male sterility in rice. *J. Integr. Plant Biol.* 59 (9), 669–679. doi: 10.1111/jipb.12564

Youens-Clark, K., Buckler, E., Casstevens, T., Chen, C., DeClerck, G., Derwent, P., et al. (2011). Gramene database in 2010: updates and extensions. *Nucleic Acids Res.* 39, D1085–D1094. doi: 10.1093/nar/gkq1148

Yu, J., Hu, S. N., Wang, J., Wong, G. K. S., Li, S. G., Liu, B., et al. (2002). A draft sequence of the rice genome (*Oryza sativa* l. ssp indica). *Science.* 296 (5565), 79–92, doi: 10.1126/science.1068037

Yu, J. M., 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 (2), 203–208. doi: 10.1038/ng1702

Zhang, Z. W., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42 (4), 355–U118. doi: 10.1038/ng.546

Zhang, J., Feng, J. Y., Ni, Y. L., Wen, Y. J., Niu, Y., Tamba, C., et al. (2017). pLARmEB: integration of least angle regression with empirical bayes for multilocus genome-wide association studies. *Heredity* 118 (6), 517–524. doi: 10.1038/hdy.2017.8

Zhang, Y. M., Mao, Y., Xie, C., Smith, H., Luo, L., and Xu, S. J. G. (2005). Mapping quantitative trait loci using naturally occurring genetic variance among commercial inbred lines of maize (Zea mays l). *Genetics* 169, 2267–2275. doi: 10.1534/genetics.104.033217

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

Zhong, H., Liu, S., Sun, T., Kong, W., Deng, X., Peng, Z., et al. (2021). Multi-locus genome-wide association studies for five yield-related traits in rice. *BMC Plant Biol.* 21 (1), 364. doi: 10.1186/s12870-021-03146-8

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44 (7), 821–U136. doi: 10.1038/ng.2310

Zhu, W. W., Yang, L., Wu, D., Meng, Q. C., Deng, X., Huang, G. Q., et al. (2022). Rice SEPALLATA genes OsMADS5 and OsMADS34 cooperate to limit inflorescence branching by repressing the TERMINAL FLOWER1-like gene RCN4. *New Phytol.* 233 (4), 1682–1700. doi: 10.1111/nph.17855

Zou, T., Li, S., Liu, M., Wang, T., Xiao, Q., Chen, D., et al. (2017). An atypical strictosidine synthase, OsSTRL2, plays key roles in another development and pollen wall formation in rice. *Sci. Rep.* 7 (1), 6863. doi: 10.1038/s41598-017-07064-4

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 27 October 2022
DOI 10.3389/fpls.2022.1026581

OPEN ACCESS

EDITED BY
Yuan-Ming Zhang,
Huazhong Agricultural University, China

REVIEWED BY
Junji Su,
Gansu Agricultural University, China
Hongwei Wang,
Yangtze University, China
Ya-Wen Zhang,
Huazhong Agricultural University, China

*CORRESPONDENCE
Xue Zhao
xuezhao@neau.edu.cn
Yingpeng Han
hyp234286@aliyun.com

SPECIALTY SECTION
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

RECEIVED 24 August 2022
ACCEPTED 04 October 2022
PUBLISHED 27 October 2022

CITATION
Yu K, Miao H, Liu H, Zhou J, Sui M, Zhan Y, Xia N, Zhao X and Han Y (2022) Genome-wide association studies reveal novel QTLs, QTL-by-environment interactions and their candidate genes for tocopherol content in soybean seed. *Front. Plant Sci.* 13:1026581. doi: 10.3389/fpls.2022.1026581

Genome-wide association studies reveal novel QTLs, QTL-by-environment interactions and their candidate genes for tocopherol content in soybean seed

Kuanwei Yu, Huanran Miao, Hongliang Liu, Jinghang Zhou, Meinan Sui, Yuhang Zhan, Ning Xia, Xue Zhao* and Yingpeng Han*

Key Laboratory of Soybean Biology in Chinese Ministry of Education (Key Laboratory of Soybean Biology and Breeding/Genetics of Chinese Agriculture Ministry), Northeast Agricultural University, Harbin, China

Genome-wide association studies (GWAS) is an efficient method to detect quantitative trait locus (QTL), and has dissected many complex traits in soybean [*Glycine max* (L.) Merr.]. Although these results have undoubtedly played a far-reaching role in the study of soybean biology, environmental interactions for complex traits in traditional GWAS models are frequently overlooked. Recently, a new GWAS model, 3VmrMLM, was established to identify QTLs and QTL-by-environment interactions (QEIs) for complex traits. In this study, the GLM, MLM, CMLM, FarmCPU, BLINK, and 3VmrMLM models were used to identify QTLs and QEIs for tocopherol (Toc) content in soybean seed, including δ-Tocotrienol (δ-Toc) content, γ-Tocotrienol (γ-Toc) content, α-Tocopherol (α-Toc) content, and total Tocopherol (T-Toc) content. As a result, 101 QTLs were detected by the above methods in single-environment analysis, and 57 QTLs and 13 QEIs were detected by 3VmrMLM in multi-environment analysis. Among these QTLs, some QTLs (Group I) were repeatedly detected three times or by at least two models, and some QTLs (Group II) were repeatedly detected only by 3VmrMLM. In the two Groups, 3VmrMLM was able to correctly detect all known QTLs in group I, while good results were achieved in Group II, for example, 8 novel QTLs were detected in Group II. In addition, comparative genomic analysis revealed that the proportion of *Glyma_max* specific genes near QEIs was higher, in other words, these QEIs nearby genes are more susceptible to environmental influences. Finally, around the 8 novel QTLs, 11 important candidate genes were identified using haplotype,

and validated by RNA-Seq data and qRT-PCR analysis. In summary, we used phenotypic data of Toc content in soybean, and tested the accuracy and reliability of 3VmrMLM, and then revealed novel QTLs, QEIs and candidate genes for these traits. Hence, the 3VmrMLM model has broad prospects and potential for analyzing the genetic structure of complex quantitative traits in soybean.

KEYWORDS

GWAS, 3VmrMLM, soybean, tocopherol content, QTL, candidate genes

Introduction

Soybean [*Glycine max* (L.) Merr.] is an important crop, and provided a great source of protein, oil, vitamin, and other nutrients for humans around the world. As one of the functional nutrients of soybean, tocopherol (Toc) has strong antioxidative capabilities and benefits to human health. It can scavenge free radicals in the body and increase immune function (Meagher et al., 2001; Kumar et al., 2009). According to the chemical structure, Tocs are composed of four members: α-tocopherol (α-Toc), β-tocopherol (β-Toc), γ-tocopherol (γ-Toc), and δ-tocopherol (δ-Toc) (Wan et al., 2008; Rozanowska et al., 2019; Barouh et al., 2022). Among them, α-Toc has the highest activity (Shaw et al., 2016). Edible oil is one of the main sources of Toc (Packer and Fuchs, 1993). As the most widely produced vegetable oil in the world, soybean oil has the highest total-Toc content, however, γ-Toc in soybean oil accounts for more than 70%. Although γ-Toc has antioxidant and other physiological activities, α-Toc is more excellent (Bramley et al., 2000). Hence, elevating the α-Toc content and total-Toc content in soybean genetics is important for quality improvement.

The Toc content of soybean seed is a typical quantitative trait, and it is difficult to breed this target trait of soybean variety using traditional breeding. This requires a lengthy selection process (Britz et al., 2008; Seguin et al., 2010). As an ancient tetraploid plant (Blanc and Wolfe, 2004), the soybean owing to its large and complex genome background brings great challenges and difficulties in genetic improvement (Young and Bharti, 2012; Tian et al., 2020; Lemay et al., 2022).

Genome-wide association studies (GWAS) is a powerful genomics tool, and it can base on natural populations to detect quantitative trait locus (QTL) underlying complex quantitative traits (Burton et al., 2007; Hamblin et al., 2011). GWAS has the advantage of high-resolution and high-throughput, thus, this method for analysis provides great convenience for the study of genetic variation in soybean (Anderson et al., 2020). Since the first GWAS conducted in soybean until now, almost all the important agronomic traits have been covered and dissected (Zhou et al., 2015; Fang et al., 2017). And yet, different GWAS models yield different GWAS results when we owe high-quality genotype and phenotype data (Chatterjee et al., 2013). Therefore, selecting the most suitable model for GWAS analysis can increase the accuracy to identify QTLs.

The general linear model (GLM) (Price et al., 2006), the mixed linear model (MLM) (Yu et al., 2006), and the compressed mixed linear model (CMLM) (Zhang et al., 2010) are single-marker genome-wide scan models, and these models can comprise a one-dimensional genome scan by testing one marker at a time. Among them, CMLM is frequently used in the genomic dissection of soybean quantitative traits (Jing et al., 2018; Zhao et al., 2019; Sui et al., 2020). However, single-marker genome-wide scan models require Bonferroni correction and multiple tests (Wang et al., 2016). Bonferroni correction is a stringent criterion, although greatly reduced false positive rates, many important loci associated with the target traits were missed (Zhang et al., 2019). With the rapid development of statistical methods, several multi-locus GWAS approaches have been developed to improve the power of QTL detection (Segura et al., 2012; Wen et al., 2018). Such as the Bayesian-information and linkage disequilibrium iteratively nested keyway (BLINK) (Huang et al., 2018), and the fixed and random model circulating probability unification (FarmCPU) (Liu et al., 2016). The obvious advantage of these methods is not a Bonferroni correction, they can reduce the amount of calculation and improve the accuracy.

Recently, a novel model was presented, named 3V multi-locus random-SNP-effect mixed linear model (3VmrMLM) (Li et al., 2022a). It is a multi-marker genome-wide scan model, this model not only provides high QTL detection power and sensitivity, at the same time, but it can also detect the QTL-by-environment interaction (QEI) and the QTL-by-QTL interaction (QQI). In this study, based on 23,149 SNPs and 175 soybean germplasms, we used six models (including 3VmrMLM, BLINK, FarmCPU, GLM, MLM, and CMLM) and conducted GWAS of individual and total-Toc content across three environments. The aim of this study is to reveal novel QTLs and QEIs of soybean Toc content and screen candidate genes.

Materials and methods

Plant materials, field trials, and phenotypic evaluation

The material used in this study included 175 diverse soybean accessions (Table S1), which encompassed most of the northeast regions of China and other countries. These materials were collected from the Chinese National Soybean GeneBank (CNSGB) and can represent the genetic diversity inside and outside of China. In this study, all experimental materials were planted at Harbin (117°17′E, 33°18′N), Liaoning (41°48′N, 123°25′E), and, Jilin (124°82′E, 43°50′N) in 2021. The field trials used a single-row plot (3 m-long rows and spaced 0.65 m) and were arranged in a randomized complete block design with three replicates per test environment. After full maturity, mature kernels of 10 randomly selected plants in each line were collected and used for evaluation of individual and total Toc content. The soybean seed Toc extraction and measurement were performed according to previous reports (Ujiie et al., 2005).

DNA isolation and sequencing

The genomic DNA of each sample from 175 tested accessions was isolated from young leaf was isolated by the method of CTAB (Han et al., 2015), and simplified-sequenced *via* specific locus amplified fragment sequencing (SLAF-seq) (Sun et al., 2013). The digest enzyme group of *Mse*I (EC: 3.1.21.4) and *Hae*III (EC: 3.1.21.4) (Thermo Fisher Scientific Inc, Waltham, MA, USA.) were used to obtain more than 50,000 sequencing tags, each 300-500 bp in length. The obtained markers were evenly distributed in unique genomic regions of the 20 soybean chromosomes. The short oligonucleotide alignment program 2 software (SOAP2) was used to align the raw paired-end reads to the soybean reference genome. Based on over 58,000 high-quality SLAF labels from each test sample, raw reads from the same genomic location were used to define SLAF groups. Genotypes were considered heterozygous if the minor allele depth or total allele depth of the sample was greater than 1/3 (Han et al., 2016).

Population structure evaluation and linkage disequilibrium analysis

The principle component analysis (PCA) was performed using the genome association and prediction integrated tool (GAPIT) R package to analyze the population structure of the natural panel (Lipka et al., 2012). The linkage disequilibrium (LD) parameter (r^2) for estimating the degree of LD between pair-wise SNPs (MAF ≥ 0.05 and missing data ≤ 10%) was calculated by TASSEL 5.0 (Bradbury et al., 2007). Unlike GWAS, missing SNP genotypes were not classified as major alleles prior to LD analysis. Parameters in the program included MAF (≥ 0.05) and completeness (> 80%) for each SNP.

Genome-wide association studies

In total, 23,149 polymorphic SNP markers and 175 tested accessions were used to perform GWAS, it was performed using six models, including three single-locus model: MLM, GLM, CMLM, and three multi-locus models: FarmCPU, BLINK, 3VmrMLM. Among these, the GLM, MLM, CMLM, FarmCPU, and BLINK models were implemented with the R package "GAPIT" and visualization used scripts from the R package "qqman" (https://cran.r-project.org/package=qqman) and "CMplot " (https://github.com/YinLiLin/R-CMplot).

The significant threshold value for the association between SNP and traits were determined by -log10 (P) ≥ 4, which is equivalent to P ≤ 0.0001, for MLM, GLM, CMLM, FarmCPU, and BLINK. The R software IIIVmrMLM (Li et al., 2022b) of the 3VmrMLM method (Li et al., 2022a) was downloaded from GitHub website (https://github.com/YuanmingZhang65/IIIVmrMLM). In this study, we used the single environment and multiple-environment methods to identify QTLs and QEIs. The significant threshold value was determined by LOD score ≥ 4.

Prediction of candidate genes

Candidate genes located in the 200-kb genomic region (100 kb upstream and 100 kb downstream) of each significant or suggested QTL then identified and annotated the candidate genes with the soybean reference genome (Wm82.a2.v1, http://www.soybase.org) (Cheng et al., 2017). The gene ontology (GO) enrichment analysis of candidate genes using the online tool (https://www.soybase.org/goslimgraphic_v2/dashboard.php). In addition, the whole genome and QEIs candidate genes among soybean relatives were compared using OrthoVenn2 (https://orthovenn2.bioinfotoolkits.net/task/create) (Xu et al., 2019).

Association analysis of candidate genes

Genome resequencing data were used to select the SNP variations within candidate genes. These SNP were located in exonic, intronic regions, upstream and downstream regions. Then, we combined the phenotype values of 56 soybean germplasms in three environments, these soybean germplasms were selected from the 175 diverse soybean accessions (Table S1) (including 9 high and low individual and total Toc germplasms), using the general linear model (GLM) in TASSEL 5.0 to identify SNPs of candidate genes that related to individual or total Toc content (Bradbury et al., 2007). Significant SNPs associated with the target trait were claimed when the test statistic was $P < 0.01$.

Haplotype analysis

The haplotypes were classified based on all of the SNPs with an MAF >0.05 in each candidate gene. Best linear unbiased predictors (BLUP) value were calculated using the "Phenotype" (https://cran.r-project.org/package=Phenotype) in R package. For each Toc component, haplotypes containing 18 soybean germplasms accessions were used for comparative analysis. One-way ANOVA and Two-tailed unpaired t -test were used to compare the differences in TC-BLUP value among the haplotypes. Finally, we compared the individual or total Toc content among these different haplotypes.

RNA-Seq data analysis of candidate genes

For candidate genes expression pattern analysis, first, we performed a differential expression pattern analysis at different tissues by downloading the RNA expression data from the plant public RNA seq database (PPRD) (http://ipf.sustech.edu.cn/pub/soybean/), which integrated all publicly available RNA-Seq soybean libraries (4,085) (Yu et al., 2022). Then, we also analyzed the expression of candidate genes in the development stage (R6) at different germplasms using the transcriptome data (unpublished data) from our laboratory. Additionally, we constructed a heat-map plot, and it was performed using the R package pheatmap (Kolde, 2012).

Quantitative real–time PCR (qRT–PCR)

Total RNA was isolated using the RNAprep pure Plant Kit (DP432, Tiangen). First-strand cDNA was synthesized from total RNA using TIANScript RT kits (KR104, Tiangen). And qRT-PCRs were performed using SYBR Green (FP205, Tiangen) reagents on an ABI 7500 fast real-time PCR platform. All qRT-PCRs were performed in three independent repeats, and the relative levels of transcript abundance were calculated using the $2^{-\Delta\Delta CT}$ method (Livak and Schmittgen 2001). The GmActin4 (*Glyma.12G063400*) was used as an internal control for data normalization. Primer sequences for candidate genes were obtained from the qPrimerDB database (Table S2) (Lu et al., 2018).

Statistical analysis

Descriptive statistical analysis of phenotypic data including mean, minimum, maximum, coefficient of variation (CV), heritability, skewness, and kurtosis was performed using IBM SPSS statistics 25.0 (SPSS, Chicago, USA). One-way ANOVA with Dunnett's multiple comparisons test and unpaired two-tailed t-test were performed using GraphPad Prism 9.4.1.

Results

Statistical and variation analysis of Toc content

Statistical analysis showed a wide range of phenotypic variations in the levels of the individual and total Toc content of the 175 soybean accessions from Harbin, Liaoning, and Jilin in 2021 (Table 1). The coefficient of variation (CV%), skewness, and kurtosis of Toc content of the association panel are also presented in Table 1. The CV varied a lot among different Toc content, especially the α-Toc content under three locations were

TABLE 1 Statistical and variation analysis of tocopherol content in the tested soybean population (n = 175).

Traits	Location	Min(μg/g)	Max(μg/g)	Mean(μg/g)	CV	Skewness	Kurtosis	Heritability
α-Toc content	Harbin	6.59	52.43	22.69	35.21%	0.74	0.80	0.51
	Liaoning	5.17	51.12	23.42	44.90%	0.42	-0.64	
	Jilin	5.65	49.62	21.48	41.68%	0.41	-0.33	
γ-Toc content	Harbin	86.97	244.7	164.97	15.65%	0.35	0.17	0.59
	Liaoning	99.01	234.15	161.01	14.63%	0.38	0.55	
	Jilin	88.78	235.8	167.24	15.22%	0.29	0.26	
δ-Toc content	Harbin	53.1	195.1	107.17	27.71%	0.63	-0.12	0.72
	Liaoning	55.6	162.29	93.23	21.04%	0.54	0.21	
	Jilin	43.64	159.24	91.73	25.92%	0.70	0.12	
Total- content	Harbin	179.49	407.31	294.83	13.09%	0.03	0.24	0.64
	Liaoning	190.37	358.14	277.66	12.45%	0.01	-0.21	
	Jilin	188.34	371.91	280.44	11.41%	-0.04	-0.33	

Min, minimum; Max, maximum; CV, coefficient of variation.

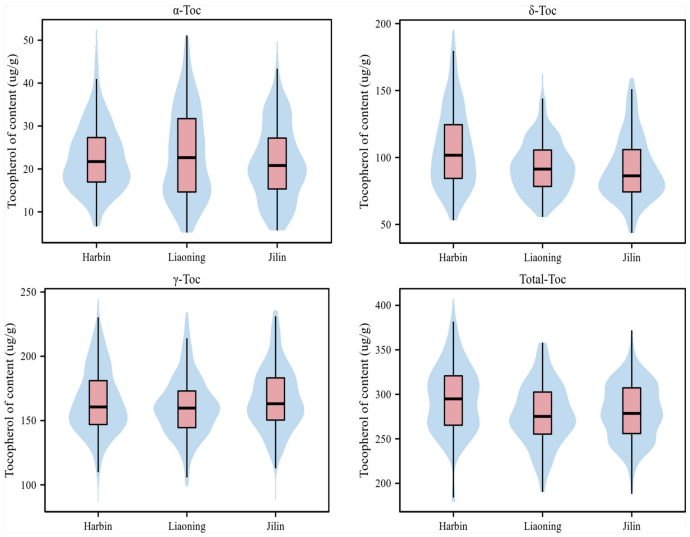


FIGURE 1
Phenotypic variation of Toc content in soybean seeds of the tested accessions at three environments. ('Harbin', 'Liaoning', and 'Jilin'). Variation of Toc content of soybean in the association panel. The black horizontal line represents the median, the black box represents the range from the lower quartile to the upper quartile, and the black vertical line represents the dispersion of phenotypic data.

observed from 35.21% to 44.9%, but all Toc content was no significant skewness or kurtosis (Figure 1). These results showed that Toc content was mainly influenced by genetic factors with less effect by environmental factors. Therefore, the tocopherol content of soybean in this study was appropriate for GWAS.

SNP genotyping, linkage disequilibrium estimating, and population structure for the GWAS panel

The genotyped samples included 175 soybean germplasms (including landraces and elite cultivars). The genomic DNA of these 175 accessions was sequenced using SLAF-seq. A total of 23,149 high-quality markers (MAF ≥ 0.05, missing data ≤ 10%) were identified from 153 million paired-end reads with 45 bp-read lengths and the sequencing depth was about 6.5 fold. The number of SNPs varied across the 20 soybean chromosomes. The highest number of SNPs was observed in Chr.18 (1732) and the lowest was detected in Chr.11 (685) (Figure 2A).

We assessed the mapping power of GWAS by the average distance of LD decay. The mean LD decay of the population was estimated at 97466 bp, when r^2 dropped to 0.2 (Figure 2B). Then, all 23,149 SNPs were used for scanning the population stratification of association panels through the principal component (PC), and evaluation of the variation of the first 10 PCs analysis revealed an inflection point at PC3, which demonstrated that the first 3 PCs dominated the population structure on the association mapping (Figures 2C, D). Additionally, a lower level of genetic relatedness among the 175 tested accessions based on pairwise relative kinship coefficients was observed (Figure 2E).

Quantitative trait locuss associated with Toc content by GWAS

GWAS was conducted using GLM, MLM, CMLM, FarmCPU, BLINK, and 3VmrMLM models. All of which accounted for kinship and population structure. First of all, we used different thresholds of significance (by -log10 (P) or LOD score= 3, 4, 5, 6, 7, 8, and 9) for testing six GWAS models and counted the number of QTLs detected (Figure 3A).Then, when -log10(P) ≥ 4 as significant thresholds, a total of 86 QTLs significantly associated with individual and total Toc concent in soybean seeds were detected *via* GLM, 18 QTLs were detected by MLM, 41 QTLs by CMLM, 41 QTLs by BLINK, and 34 QTLs by FarmCPU (Figure 4A, Figures S1–S5 and Tables S3–S7). Among them, only 4 QTLs were co-detected by all six models (Figure 3B). Furthermore, the largest number of QTLs were detected with the 3VmrMLM model. Among them, the single-environment method detected 101 QTLs (Figure S6, Table S8), the multiple-environments method detected 57 QTLs (Figure S7, Table S9), and 13 QEIs (Figure S8, Table S10). Among them, 11 QTLs were co-detected by single-environment and multiple-environment method (Figure 3C). The results showed that the number of QTLs detected by 3VmrMLM are more abundant and stable under different significance thresholds.

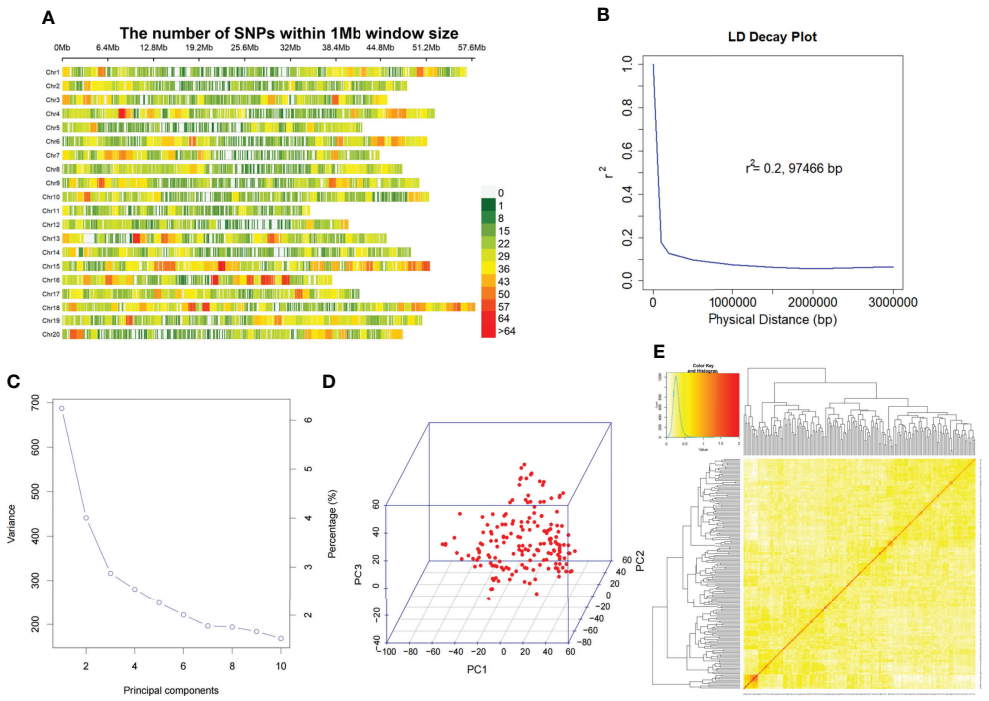


FIGURE 2
SNP density, distribution and mapping genetic data of populations. **(A)**. SNP density and distribution across 20 soybean chromosomes. **(B)**. LD decay of the genome-wide association study (GWAS) population. **(C)**. Population structure of soybean germplasm collection reflected by principal components. **(D)**. The first 3 principal components of the 23,149 SNPs used in GWAS. **(E)**. A heatmap of the kinship matrix of the 175 soybean accessions.

Finally, the QTLs, which were repeatedly detected in multiple GWAS models, were selected as reliable QTLs—group I. As shown in Figure 3B, Table 2, 19 QTLs were co-detected by at least three times or at least two models, which were distributed among 24 genomic regions in 14 chromosomes. Among these, 9 QTLs (rs9337368, rs1834346, rs17125409, rs330000, rs9782629, rs19530677, rs5680781, rs17266245, and rs53062844) were located in genomic regions or QTLs reported by previous studies, confirming the accuracy of QTL detection. We regard the remaining 15 QTLs as the novel QTLs (rs39895210, rs2960931, rs19310064, rs31044180, rs7543892, rs4992837, rs14593163, rs24979561, rs588498, rs19962490, rs6204830, rs8720462, rs37558520, rs34774232, and rs35815938). Moreover, a total of 161 QTLs were identified by 3VmrMLM (Figure 3A), in order to test the reliability of the 3VmrMLM model, we selected the QTLs only detected in 3VmrMLM. 9 QTLs (detected by at least two times) were repeatedly detected as specific QTLs—group II (Table 3), which were distributed among 9 genomic regions in 8 chromosomes. rs41784197 was located in genomic regions or QTLs reported by previous studies. Again, we regard the remaining 8 QTLs as the novel QTLs (rs7167202, rs9140707, rs18105573, rs2669053, rs40595691, rs43000771, rs5779917, and rs46814888).

Prediction of candidate genes for Toc content in soybean seeds

Based on annotations for the soybean reference genome in SoyBase, we further predicted candidate genes within the 200-kb flanking regions of the novel QTLs. In two group novel QTLs, a total of 248 genes were obtained (Table S11). And a total of 134 genes were obtained in QEIs (Table S12). Then, we used GO annotation to perform enrichment analysis for group I and group II genes. The results categorized as molecular function, cellular component, and biological process, were shown in Figure 4. Both group I and group II candidate genes are involved in a variety of functions, such as carbohydrate metabolic process, translation, protein binding, cytoplasm component, DNA binding, and so on.

Comparative genome analysis

In order to predict the authenticity of the QEIs, firstly, we selected four closely related species, *Glyma_max*, *Vigna_radiate*, *Vigna_augularis*, and *Phaseolus_vulgaris*, for comparative genomic analysis. A total of 12847 core gene

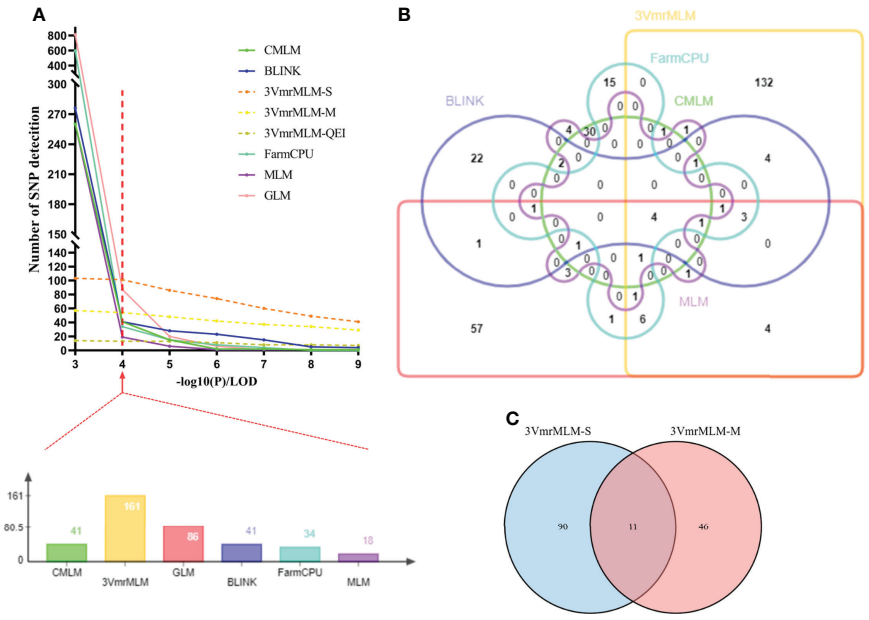


FIGURE 3
Statistics of QTLs in GWAS results under three models. **(A)** Statistics on the number of QTLs detected at different significance thresholds by different models or methods. **(B)**Venn diagram representing the number of unique and shared QTLs with six models. **(C)** Venn diagram representing the number of unique and shared QTLs with 3VmrMLM single-environment method and 3VmrMLM multiple-environment method. Finally determine the red line **(A)** represents the GWAS significance threshold of this study, both **(B, C)** are counted at this significance threshold. 3VmrMLM-S represents 3VmrMLM single-environment method, 3VmrMLM-M represents QTL detection of 3VmrMLM multiple-environment method, 3VmrMLM-QEI represents QEI detection of 3VmrMLM multiple-environment method.

clusters were found in the four species, and 1197 gene clusters were unique to *Glyma_max* (Figure 5A), specific genes clusters account for 5.4% (1197/22159). Then, we used candidate gene of QEIs for comparative genomic analysis, 12 gene clusters were unique to candidate gene of QEIs (Figure 5B), specific genes clusters account 9.23% (12/130), this result shown that these QEIs have more abundant specific genes. As shown in Figure 5C, these specific genes are involved in various biological processes, metabolic processes, response to stimulus, etc. More detailed statistics on the number of shared gene clusters are shown in Figure 5D. Figure 6E is count of proteins by type of cluster.

Gene-based association analysis of candidate genes

Two groups of candidate gene association analysis were performed using the GLM model with the TASSEL, using the genome resequencing of 56 germplasms (including 9 high and low individual and total Toc germplasms). A total of 4537 SNPs with MAF ≥ 0.05 were identified among 248 candidate genes. Among them, a total of 50 SNPs from 11 candidate genes were found to reach the threshold with -log10(P) ≥ 2.0 (Table S13), of these, 4 SNPs are located in upstream regions, 10 SNPs are located in intronic regions, 26 SNPs are located in exonic ;regions, and 10 SNPs are located ;in downstream regions. Those SNPs are considered to be significantly associated with individual and total Toc concentrations in soybean seeds. Among these genes, 4 candidate genes from group I and 7 candidate genes from group II. These genes can be considered potential candidate genes for individual and total Toc-related. For example, as shown in Figure 6A, the significant SNPs correlated to α-Toc and δ-Toc on basis of association analysis for two candidate genes were respectively identified (*Glyma.17G188700* and *Glyma.20G235100* were shown in Figure 6A, others were shown in Figure S9).

Haplotype analysis of candidate genes

For the haplotype analysis, first, all the SNP markers within each gene are used to construct haplotypes. Then, we performed one-way ANOVA with TC-BLUP values of each soybean accession. The results are shown in Table 4, each gene contains haplotypes that are significant differences from TC-BLUP values. In addition, 14 haplotypes of 11 candidate genes respectively conferred an increased individual and total Toc content in soybean seeds (*Glyma.17G188700* and *Glyma.20G235100* were shown in Figure 6B, others were shown in Figure S10). Therefore, these haplotypes are beneficial and can be adjusted for individual and total Toc content in soybean seeds.

TABLE 2 SNPs associated with Toc content of soybean seeds and known QTLs overlapped with peak SNPs of group I.

SNP	Chr.	Position	Allele	Traits	Model/ Method	Significance	Environment	-log10 (P)	Known QTL	References
rs9337368	2	9337368	A/T	δ-Toc content	BLINK		Harbin	6.48	SSR02_0458-SSR02_0520	Sui et al., 2020
				δ-Toc content	FarmCPU		Harbin	6.15		
				δ-Toc content	MLM		Harbin	4.03		
				δ-Toc content	GLM		Harbin	4.49		
				δ-Toc content	3V-M	SIG	–	6.33		
rs39895210	3	39895210	G/A	Total-Toc content	BLINK		Liaoning	4.15		
				Total-Toc content	FarmCPU		Liaoning	4.52		
				Total-Toc content	GLM		Liaoning	4.53		
				Total-Toc content	3V-S	SIG	Liaoning	19.38		
rs2960931	6	2960931	G/A	δ-Toc content	FarmCPU		Liaoning	4.45		
				δ-Toc content	GLM		Liaoning	4.47		
				δ-Toc content	3V-S	SIG	Liaoning	10.47		
				δ-Toc content	3V-M	SIG	–	10.45		
rs1834346	8	1834346	A/T	α-Toc content	MLM		Harbin	4.16	Sat_383-BARC-037229-06749	Li et al., 2016
				α-Toc content	GLM		Harbin	4.06		
				Total-Toc content	3V-M	SIG	–	11.02		
rs19310064	8	19310064	A/C	α-Toc content	CMLM		Harbin	9.02		
				α-Toc content	BLINK		Harbin	11.84		
				α-Toc content	MLM		Harbin	9.02		
				α-Toc content	GLM		Harbin	9.43		
rs31044180	9	31044180	G/T	α-Toc content	FarmCPU		Jilin	4.43		
				γ-Toc content	FarmCPU		Jilin	5.15		
				Total-Toc content	FarmCPU		Jilin	4.22		
				γ-Toc content	MLM		Jilin	4.30		
				α-Toc content	GLM		Jilin	4.43		
				γ-Toc content	GLM		Jilin	5.15		
				Total-Toc content	GLM		Jilin	4.22		

(Continued)

TABLE 2 Continued

SNP	Chr.	Position	Allele	Traits	Model/ Method	Significance	Environment	-log10 (P)	Known QTL	References
				δ-Toc content	3V-S	SIG	Jilin	18.07		
rs7543892	10	7543892	T/G	δ-Toc content	BLINK		Jilin	7.07		
				δ-Toc content	FarmCPU		Jilin	4.55		
				δ-Toc content	GLM		Jilin	4.25		
				δ-Toc content	3V-M	SIG	–	11.13		
rs49928375	10	49928375	G/T	α-Toc content	CMLM		Harbin	5.04		
				α-Toc content	FarmCPU		Harbin	4.67		
				α-Toc content	MLM		Harbin	4.91		
				α-Toc content	GLM		Harbin	5.74		
				α-Toc content	3V-S	SIG	Harbin	17.93		
rs17125409	12	17125409	C/A	α-Toc content	CMLM		Jilin	5.21	–	Zhan et al., 2020
				α-Toc content	BLINK		Harbin	6.09		
				α-Toc content	BLINK		Jilin	10.27		
				α-Toc content	FarmCPU		Harbin	7.63		
				α-Toc content	GLM		Harbin	4.56		
				α-Toc content	3V-M	SIG	–	46.05		
rs330000	13	330000	G/A	δ-Toc content	FarmCPU		Liaoning	4.76	–	Zhan et al., 2020
				δ-Toc content	GLM		Harbin	4.76		
				δ-Toc content	GLM		Liaoning	4.96		
				δ-Toc content	3V-S	SIG	Liaoning	9.65		
				δ-Toc content	3V-M	SUG	–	4.35		
rs9782629	14	9782629	G/T	γ-Toc content	CMLM		Harbin	5.63	BARC-059251-15691-Sct_034	Shaw et al., 2017
				γ-Toc content	BLINK		Harbin	7.27		
				γ-Toc content	FarmCPU		Harbin	4.59		
				γ-Toc content	MLM		Harbin	4.71		
				γ-Toc content	GLM		Harbin	4.89		
				γ-Toc content	3V-QEI	SIG	–	15.98		

(Continued)

TABLE 2 Continued

SNP	Chr.	Position	Allele	Traits	Model/ Method	Significance	Environment	-log10 (P)	Known QTL	References
				Total-Toc content	3V-QEI	SIG	–	18.12		
rs19530677	16	19530677	T/A	γ-Toc content	CMLM		Harbin	7.33	Sat_259-Sat_370	Li et al.,2010/Li et al.,2016
				Total-Toc content	CMLM		Harbin	6.67		
				γ-Toc content	BLINK		Harbin	9.16		
				Total-Toc content	BLINK		Harbin	4.23		
				γ-Toc content	FarmCPU		Harbin	6.10		
				γ-Toc content	MLM		Harbin	5.28		
				γ-Toc content	GLM		Harbin	6.20		
				γ-Toc content	3V-QEI	SIG	–	32.05		
rs14593163	17	14593163	T/G	δ-Toc content	BLINK		Harbin	6.42		
				Total-Toc content	BLINK		Harbin	4.56		
				Total-Toc content	FarmCPU		Harbin	7.49		
				Total-Toc content	MLM		Harbin	5.35		
				δ-Toc content	GLM		Harbin	4.73		
				Total-Toc content	GLM		Harbin	6.12		
rs24979561	17	24979561	G/A	α-Toc content	CMLM		Harbin	5.87		
				α-Toc content	BLINK		Harbin	7.86		
				α-Toc content	FarmCPU		Harbin	7.56		
				α-Toc content	MLM		Harbin	5.87		
				α-Toc content	GLM		Harbin	6.53		
				α-Toc content	3V-S	SIG	Harbin	18.72		
rs588498	18	588498	G/A	α-Toc content	FarmCPU		Liaoning	4.26		
				α-Toc content	GLM		Liaoning	4.26		
				α-Toc content	3V-S	SUG	Liaoning	4.53		
rs5680781	18	5680781	G/T	Total-Toc content	CMLM		Jilin	5.04	–	Zhan et al., 2020
				γ-Toc content	BLINK		Jilin	4.62		
				Total-Toc content	BLINK		Jilin	5.61		

(Continued)

TABLE 2 Continued

SNP	Chr.	Position	Allele	Traits	Model/ Method	Significance	Environment	-log10 (P)	Known QTL	References
				γ-Toc content	3V-S	SIG	Jilin	9.07		
rs17266245	18	17266245	T/G	γ-Toc content	BLINK		Jilin	4.31	Satt038–Sat_164	Sui et al., 2020/ Zhan et al., 2020
				γ-Toc content	3V-S	SIG	Jilin	16.18		
				γ-Toc content	3V-M	SIG	–	11.10		
rs19962490	18	19962490	T/C	δ-Toc content	MLM		Harbin	5.42		
				Total-Toc content	MLM		Harbin	4.39		
				δ-Toc content	GLM		Harbin	4.84		
				Total-Toc content	GLM		Harbin	4.37		
rs53062844	18	53062844	G/T	α-Toc content	CMLM		Liaoning	4.91	Satt472–Satt038	Sui et al., 2020
				α-Toc content	BLINK		Liaoning	12.82		
				α-Toc content	FarmCPU		Liaoning	5.73		
				α-Toc content	MLM		Liaoning	4.87		
				α-Toc content	GLM		Liaoning	5.73		
				α-Toc content	3V-M	SIG	–	23.60		
rs6204830	19	6204830	T/G	α-Toc content	MLM		Liaoning	4.10		
				α-Toc content	3V-S	SIG	Liaoning	15.42		
				α-Toc content	3V-S	SIG	Jilin	7.40		
rs8720462	19	8720462	G/A	δ-Toc content	BLINK		Harbin	7.38		
				δ-Toc content	FarmCPU		Harbin	4.37		
				δ-Toc content	FarmCPU		Liaoning	4.48		
				δ-Toc content	GLM		Harbin	6.82		
				δ-Toc content	GLM		Liaoning	4.48		
				δ-Toc content	3V-S	SUG	Harbin	5.61		
				δ-Toc content	3V-M	SUG	–	4.01		
rs37558520	19	37558520	T/C	Total-Toc content	FarmCPU		Liaoning	4.06		
				Total-Toc content	GLM		Liaoning	4.09		
				α-Toc content	3V-S	SIG	Liaoning	9.34		

(Continued)

TABLE 2 Continued

SNP	Chr.	Position	Allele	Traits	Model/ Method	Significance	Environment	-log10 (P)	Known QTL	References
rs34774232	20	34774232	A/G	δ-Toc content	FarmCPU		Harbin	4.72		
				δ-Toc content	GLM		Harbin	4.82		
				δ-Toc content	3V-M	SIG	–	13.07		
rs35815938	20	35815938	T/C	δ-Toc content	FarmCPU		Liaoning	4.14		
				δ-Toc content	GLM		Liaoning	4.12		
				δ-Toc content	3V-M	SIG	–	10.32		

3V-S represents 3VmrMLM single-environment method, 3V-M represents QTL detection of 3VmrMLM multiple-environment method, 3V-QEI represents QEI detection of 3VmrMLM multiple-environment method, SIG represents significant QTLs, and SUG represents suggested QTLs.

RNA-Seq data analysis of candidate genes for Toc content in soybean

In order to confirm the possible effect of candidate genes in the regulation of Toc content, we firstly used PPRD to analyze the expression patterns of 11 candidate genes in different tissues. The result showed that all candidate genes were expressed in soybean seed (Figure S11), and *Glyma.10G171600* is most abundantly expressed in seed compared with other tissues. Then, for the 11 candidate genes of 56 soybean germplasms at the development stage (R6), RNA-Seq data analysis was done. The result showed that the expression levels of the 11 candidate genes in low and high Toc content germplasms were different. Among them, *Glyma.17G188700* can regulate α-Toc content in soybean seeds. The range of the expression levels of *Glyma.17G188700* in higher α-Toc germplasms was much higher than those of lower. Other

TABLE 3 SNPs associated with Toc content of soybean seeds and known QTLs overlapped with peak SNPs of group Ⅱ.

SNP	Chr.	Position	Allele	Traits	Model/Method	Environment	$-\log_{10}^{(P)}$	Known QTL	References	Significance
rs7167202	1	7167202	G/T	γ-Toc content	3VmrMLM-S	Jilin	5.13			SUG
				Total-Toc content	3VmrMLM-S	Jilin	6.43			SIG
				Total-Toc content	3VmrMLM-M	–	16.28			SIG
rs41784197	1	41784197	T/C	γ-Toc content	3VmrMLM-S	Jilin	11.64	Satt179-Sat_201	Li et al., 2016	SIG
				γ-Toc content	3VmrMLM-M	–	12.32			SIG
rs9140707	7	9140707	G/T	α-Toc content	3VmrMLM-S	Liaoning	17.35			SIG
				α-Toc content	3VmrMLM-M	–	33.30			SIG
rs18105573	8	18105573	A/G	δ-Toc content	3VmrMLM-S	Jilin	6.44			SIG
				δ-Toc content	3VmrMLM-M	–	5.58			SUG
rs2669053	9	2669053	T/C	γ-Toc content	3VmrMLM-S	Harbin	16.30			SIG
				γ-Toc content	3VmrMLM-QEI	–	11.49			SIG
				Total-Toc content	3VmrMLM-S	Harbin	12.13			SIG
				Total-Toc content	3VmrMLM-QEI	–	5.23			SUG
rs40595691	10	40595691	C/T	γ-Toc content	3VmrMLM-M	–	4.35			SUG
				Total-Toc content	3VmrMLM-M	–	7.34			SIG
rs43000771	15	43000771	C/T	γ-Toc content	3VmrMLM-S	Liaoning	4.07			SUG
				Total-Toc content	3VmrMLM-M	–	4.57			SUG
rs5779917	19	5779917	G/T	α-Toc content	3VmrMLM-QEI	–	8.90			SIG
				γ-Toc content	3VmrMLM-S	Harbin	7.76			SIG
rs46814888	20	46814888	T/G	δ-Toc content	3VmrMLM-S	Harbin	8.64			SIG
				δ-Toc content	3VmrMLM-M	–	8.70			SIG

3V-S represents 3VmrMLM single-environment method, 3V-M represents QTL detection of 3VmrMLM multiple-environment method, 3V-QEI represents QEI detection of 3VmrMLM multiple-environment method, SIG represents significant QTLs, and SUG represents suggested QTLs.

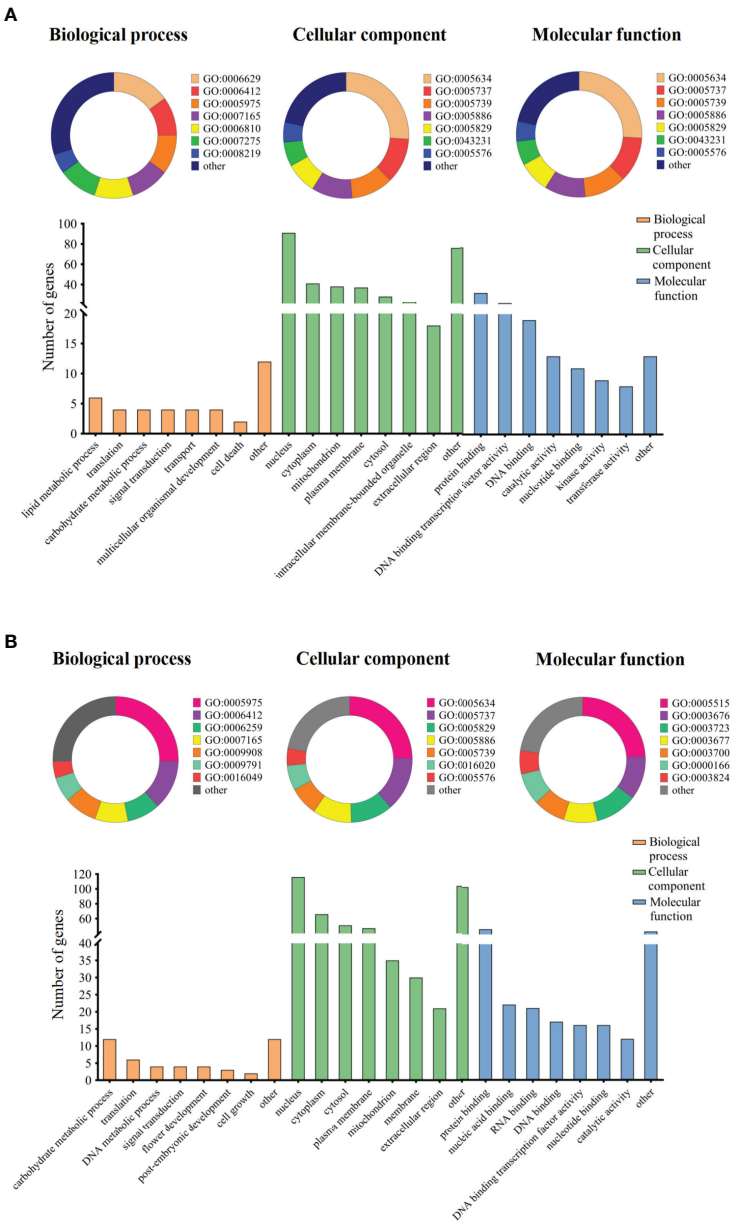


FIGURE 4
Gene ontology term enrichment analysis of candidate genes. Note: The categorized percentage and the quantity statistics of gene ontology term enrichment analysis of candidate genes, **(A)** represents group I candidate genes and **(B)** represents group II candidate genes.

genes regulate Toc content as shown in Figure 7. Interestingly, *Glyma.01G054800*, *Glyma.09G032100*, and *Glyma.10G171600* can regulate both the γ-Toc and Total-Toc content. *Glyma.09G032100* in higher γ-Toc and total-Toc germplasms were much higher than those expression levels of lower. However, *Glyma.01G054800*, and *Glyma.10G171600* in higher γ-Toc germplasms have higher expression levels, but in higher total-Toc germplasms have lower expression levels. Moreover, these candidate genes results of qRT-PCR are consistent with the RNA-seq data (Figure S12).

Discussion

As one of the vitamin E family members, Toc plays a crucial role for humans, plants, and animals (Bramley et al., 2000). For humans, daily Toc supplementation can decrease the risk for cancer and cardiovascular disease (Shaw et al., 2016). For plants, Toc can protection of chloroplasts from photooxidative damage (Munne-Bosch and Alegre, 2002). For animals, Toc must be added to animal feed to improve and maintain growth and

health (Pinelli-Saavedra et al., 2008). Soybean is a major crop used worldwide as a source of food, oil, and animal feed. Soybean oil compared to other oil crops contains a higher total Toc content, but γ-Toc comprises 70% (Park et al., 2019). The physiological activity of γ-Toc was lower than that of α-Toc (Wan et al., 2008). Therefore, increasing the α-Toc and total Toc content in soybean seeds is important to improve the nutritional variety and feed quality of soybean. However, the genetic background of Toc content is complex quantitative inheritance. The reason why quantitative traits are complex is that they are controlled by unequal polygenes and are susceptible to environmental influences. In this study, individual and total Toc content of 175 soybean accessions were evaluated. The results showed that the Toc content of tested germplasms was relatively stable to the environment, and Toc content had a wide range of variation among the different germplasms.

GWAS has been widely used in the mining of QTLs in most crops including soybean. It is a method to identify the genetic variation among the natural populations to establish genetic markers based on linkage disequilibrium (LD) (Yano et al., 2019; Xiao et al., 2022). How improve the power of GWAS has been a major challenge for the last decade. In recent years, a variety of new methods have been proposed, with the rapid development of computing technology and sequencing technology (Wang et al., 2016; Huang et al., 2018; Xiao et al., 2021; Li et al., 2022a). Although this propelled much of the practicability of GWAS, it is particularly important to select the appropriate sequencing method and suitable model for improving the positioning efficiency according to the research needs (Liu et al., 2017; Kim et al., 2021). For this study, we adopted six models (GLM, MLM, CMLM, BLINK, FarmCPU, and 3VmrMLM), to conduct GWAS of Toc content in soybean seeds. And the results were divided into two groups, revealed a total of 23 novel QTLs, other QTLs were located in the regions of QTLs in previous studies or overlapped our previous GWAS studies, and these known QTLs are all covered by 3VmrMLM.

3VmrMLM is a new algorithm, different from other algorithm, the 3VmrMLM use single-marker genome-wide scanning to select potentially associated markers and uses empirical Bayes and the likelihood ratio test in a multi-locus model to identify significant QTLs, this undoubtedly improves its detection capability (Li et al., 2022a). Additionally, it can be simultaneously estimated in a vector manner that QEI and QQI effects. Although the QQI detection in this study did not achieve good results, the 3VmrMLM still showed better detection ability than the GLM, MLM, CMLM, BLINK, and FarmCPU, indicating a more reliable tool for complex trait dissection.

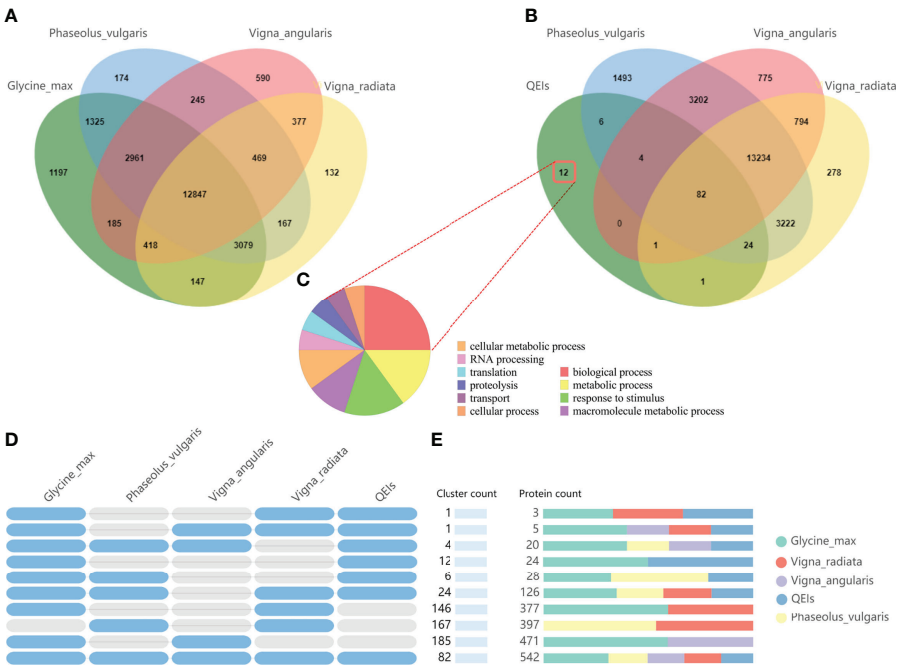


FIGURE 5
Comparative genome analysis candidate genes of QEIs. **(A)**. Venn diagram representing the core orthologs and specific genes cluster for *Glyma_max*, *Vigna_radiate*, *Vigna_augularis*, and *Phaseolus_vulgaris*. **(B)**. Venn diagram representing the core orthologs and specific genes cluster for candidate genes of QEIs, *Vigna_radiate*, *Vigna_augularis*, and *Phaseolus_vulgaris*. **(C)**. Gene ontology term enrichment analysis of unique candidate genes of QEIs. **(D)**. Shared gene clusters of orthologous groups categories. **(E)**. Protein families count shared between *Glyma_max*, *Vigna_radiate*, *Vigna_augularis*, *Phaseolus_vulgaris*, and candidate genes of QEIs.

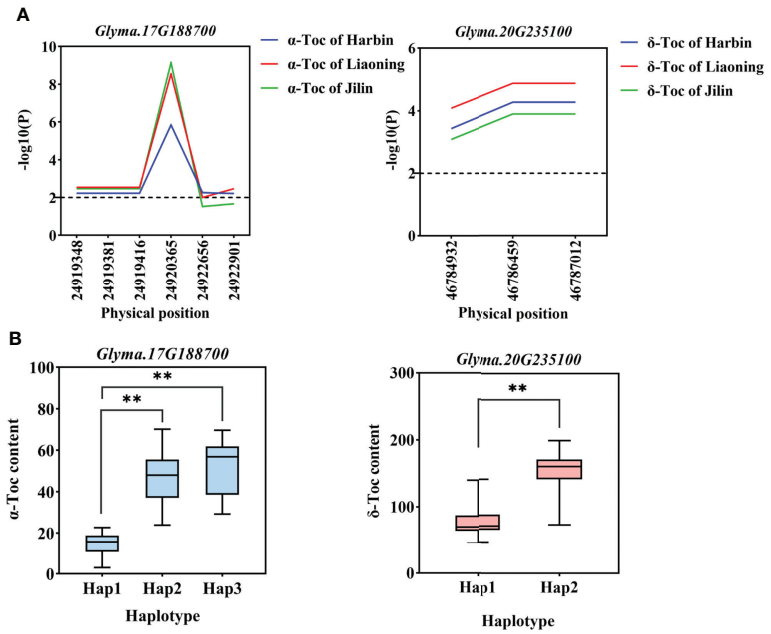


FIGURE 6
Gene-based association analysis and haplotypes analysis. **(A)**. Gene-based association analysis of candidate genes that related to Toc content. **(B)**. Haplotypes analysis of candidate genes that related to Toc content. Horizontal line indicates that the threshold is set to 2.0, the * and ** was significance at $P < 0.05$ and $P < 0.01$, respectively, *Glyma.17G188700* from group I, and *Glyma.20G235100* from group II.

In soybean and other plants, only a few definite genes have been characterized, associated with an individual or total Toc. Among them also includes most of the key enzyme genes (Dwiyanti et al., 2011; Zhang et al., 2013). To accurately screen candidate genes, we selected a total of 248 genes within the 200-kb flanking regions of the 23 novel QTLs and using a gene-based association by the GLM method, a total of 11 genes were finally determined to be significantly related to individual or total Toc in soybean seeds. Moreover, almost all these genes have beneficial haplotypes. *Glyma.06G038000* encoded alpha/beta-Hydrolases

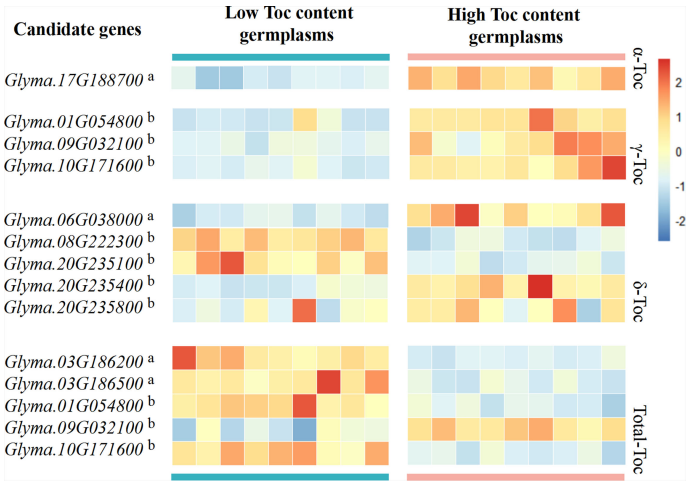


FIGURE 7
Heatmap of candidate gene expression analysis by RNA-Seq data. Candidate gene analysis was performed using different high and low germplasms for each Toc content, the red boxes indicate high transcript levels, and the blue boxes indicate low transcript levels. The letter in the upper right corner a indicates the gene from group I, and the letter in the upper right corner b indicates the gene from group II.

TABLE 4 Haplotype analysis of candidate genes.

Gene ID	Traits	Hap	Total number	Mean TC-BLUP value	P value	Significance	Functional annotation	References
Glyma.03G186200	Total-Toc content	Hap1	9	241.55	–	–	RAB GTPase homolog C2A	
		Hap2	3	307.98	0.0007	***		
		Hap3	6	310.88	<0.0001	****		
Glyma.03G186500	Total-Toc content	Hap1	9	241.55	–	–	Transducin family protein/WD-40 repeat family protein	
		Hap2	7	305.21	<0.0001	****		
		Hap3	2	326.38	0.0002	***		
Glyma.06G038000	δ-Toc content	Hap1	9	79.08	–	–	Alpha/beta-Hydrolases superfamily protein	Albert et al., 2022
		Hap2	9	129.63	<0.0001	****		
Glyma.17G188700	α-Toc content	Hap1	7	12.19	–	–	hAT dimerisation domain-containing protein/ transposase-related	
		Hap2	6	28.82	0.0053	**		
		Hap3	5	33.21	0.0013	**		
Glyma.01G054800	γ-Toc content	Hap1	4	107.76	–	–	Plant protein of unknown function (DUF863)	
		Hap2	5	107.66	>0.9999	ns		
		Hap3	3	206.32	0.0002	***		
		Hap4	6	201.99	<0.0001	****		
	Total-Toc content	Hap1	4	271.77	–	–		
		Hap2	5	298.58	0.6572	ns		
		Hap3	3	262.08	0.9795	ns		
		Hap4	6	266.15	0.9932	ns		
Glyma.08G222300	δ-Toc content	Hap1	4	93.7	–	–	O-fucosyltransferase family protein	
		Hap2	3	120.1	0.5826	ns		
		Hap3	3	156.71	0.0548	ns		
		Hap4	8	199.39	0.0003	***		
Glyma.09G032100	Total-Toc content	Hap1	5	247.64	–	–	MYB domain protein 78	
		Hap2	4	233.93	0.5327	ns		
		Hap3	9	300.91	0.0002	***		
	γ-Toc content	Hap1	5	102.56	–	–		
		Hap2	4	114.14	0.6506	ns		
		Hap3	9	203.43	<0.0001	****		
Glyma.10G171600	Total-Toc content	Hap1	5	306.77	–	–	RAB GTPase homolog A5A	
		Hap2	4	313.84	0.9366	ns		
		Hap3	4	238.98	0.0013	**		
		Hap4	5	243.6	0.0014	**		
	γ-Toc content	Hap1	5	105.48	–	–		
		Hap2	4	110.49	0.9775	ns		
		Hap3	4	200.26	<0.0001	****		
		Hap4	5	205.98	<0.0001	****		
Glyma.20G235100	δ-Toc content	Hap1	8	86.25	–	–	Indeterminate(ID)-domain 2	
		Hap2	10	118.84	0.0089	**		
Glyma.20G235400	δ-Toc content	Hap1	6	84	–	–	P-loop containing nucleoside triphosphate hydrolases superfamily protein	
		Hap2	4	91.82	0.7330	ns		
		Hap3	8	131.88	0.0004	***		
Glyma.20G235800	δ-Toc content	Hap1	6	91.24	–	–	Transducin/WD40 repeat-like superfamily protein	
		Hap2	5	78.23	0.2695	ns		
		Hap3	7	134.25	0.0002	***		

Hap represents Haplotype, TC represents individual and total Toc content. $P < 0.05$ was considered significant, * Significance was $P < 0.05$, ** Significance was $P < 0.01$, *** Significance was $P < 0.001$, **** Significance was $P < 0.0001$ and ns stands for no significance.

superfamily protein. *Glyma.01G054800* encoded plant proteins of unknown function, *Glyma.03G186500* encoded a WD-40 repeat family protein, *Glyma.20G235800* encoded a WD40 repeat-like superfamily protein, Glyma.03G186200 is a RAB GTPase homolog C2A, *Glyma.10G171600* encoded a RAB GTPase homolog A5A, *Glyma.17G188700* encoded transposas, *Glyma.09G032100* encoded a myb domain protein, *Glyma.20G235100* encoded an indeterminate domain protein, *Glyma.20G235400* encoded a P-loop containing nucleoside triphosphate hydrolases superfamily protein. Of these genes, *Glyma.01G054800* and *Glyma.10G171600* are the most special, and these two genes are higher expressed in higher γ-Toc content germplasms, but lower expressed in higher total-Toc content germplasms. The soybean oil contains a higher proportion of γ-Toc, this is very different from the other oil crops (Cahoon et al., 2003). Therefore, we conclude that the *Glyma.01G054800* and *Glyma.10G171600* inhibited the transformation of α-Toc and δ-Toc, resulting in the excessive accumulation of γ-Toc, while the total-Toc content decreased. This requires further experiments to prove. The precise functions and mechanisms of 11 candidate genes will be planned in future studies.

In general, the 3VmrMLM algorithm achieved good results in the GWAS. In this study, Toc content in soybean seed in group I QTLs, 10 known QTLs are all covered by 3VmrMLM. The results of GO enrichment analysis showed that group I; and group II candidate genes had similar GO biological process terms. for the 11 candidate genes finally identified in this study, 7 genes were alone identified by the 3VmrMLM. All candidate genes were able to detected by the 3VmrMLM. In addition, a higher percentage of the *Glyma_max* specific genes have also been found in candidate genes near QEIs by comparative genomic analysis. These results have preliminarily determined the detection efficiency of the 3VmrMLM algorithm. Thus, we hope that using 3VmrMLM could be used to dissect more important complex quantitative traits in the future, and this algorithm is advantageous to promoting the development of soybean breeding.

Data availability statement

The data presented in the study are deposited in the EBI repository, accession number PRJEB55008. Any queries should be directed to the corresponding author.

Author contributions

KWY, and XZ conceived the study and contributed to population development. KWY, HRM, and HLL contributed to phenotypic evaluation. JHZ, and MNS analyzed the data. YHZ, and NX contributed to genotyping. KWY, XZ, and YPH contributed to experimental design and writing the paper. All authors contributed to the article and approved the submitted version.

Funding

This study was financially supported by National Key Research and Development Project of China (2021YFF1001204), the Chinese National Natural Science Foundation (31971967, 31871650), National Key Research and Development Program of China (2021YFD1201604, 2019YFD1002601), the Youth and Middle-aged Scientific and Technological Innovation Leading Talents Program of the Crops (2015RA228), the National Ten Thousand Talent Program (W03020275), Postdoctoral Scientific Research Development Fund of Heilongjiang Province (LBH-Z15017, LBH-Q20004), Program on Industrial Technology System of National Soybean (CARS-04-PS06).

Acknowledgments

This study was conducted in the Key Laboratory of Soybean Biology of the Chinese Education Ministry, Soybean Research & Development Center (CARS) and the Key Laboratory of Northeastern Soybean Biology and Breeding/Genetics of the Chinese Agriculture Ministry.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1026581/full#supplementary-material

References

Albert, E., Kim, S., Magallanes-Lundback, M., Bao, Y., Deason, N., Danilo, B., et al. (2022). Genome-wide association identifies a missing hydrolase for tocopherol synthesis in plants. *PNAS* 119 (23), e2113488119. doi: 10.1073/pnas.2113488119

Anderson, R., Fernandez, C., Yuan, Y., Golicz, A., Edwards, D., and Bayer, P. (2020). Method for genome-wide association study: A soybean example. *Method Microbiol.* 2107, 147–158. doi: 10.1007/978-1-0716-0235-5_7

Barouh, N., Bourlieu-Lacanal, C., Figueroa-Espinoza, M. C., Durand, E., and Villeneuve, P. (2022). Tocopherols as antioxidants in lipid-based systems: The combination of chemical and physicochemical interactions determines their efficiency. *Compr. Rev. Food Sci. F.* 21 (1), 642–688. doi: 10.1111/1541-4337.12867

Blanc, G., and Wolfe, K. H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell.* 16 (7), 1667–1678. doi: 10.1105/tpc.021345

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., Buckler, E.S, and Buckler, E.S (2007). eTASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23 (19), 2633–2635. doi: 10.1093/bioinformatics/btm308

Bramley, P., Elmadfa, I., Kafatos, A., Kelly, F., Manios, Y., Roxborough, H., et al. (2000). Vitamin e. *J. Sci. Food Agric.* 7), 80, 913–938. doi: 10.1002/(SICI)1097-0010

Britz, S. J., Kremer, D. F., and Kenworthy, W. J. (2008). Tocopherols in soybean seeds: genetic variation and environmental effects in field-grown crops. *J. Am. Oil Chem. Soc* 85 (10), 931–936. doi: 10.1007/s11746-008-1286-y

Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., and Duncanson, A. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661–678. doi: 10.1038/nature05911

Cahoon, E. B., Hall, S. E., Ripp, K. G., Ganzke, T. S., Hitz, W. D., and Coughlan, S. J. (2003). Metabolic redesign of vitamin e biosynthesis in plants for tocotrienol production and increased antioxidant content. *J. Nat. Biotechnol.* 21, 1082–1087. doi: 10.1038/nbt853

Chatterjee, N., Wheeler, B., Sampson, J., Hartge, P., Chanock, S. J., and Park, J. H. (2013). Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.* 45 (4), 400–405e4053. doi: 10.1038/ng.2579

Cheng, Y., Ma, Q., Ren, H., Xia, Q., Song, E., Tan, Z., et al. (2017). Fine mapping of a phytophthora-resistance gene RpsWY in soybean (Glycine max l.) by high-throughput genome-wide sequencing. *Theor. Appl. Genet.* 130 (5), 1041–1051. doi: 10.1007/s00122-017-2869-5

Dwiyanti, M.S., Yamada, T., Sato, M., Abe, J., and Kitamura, K. (2011). Genetic variation of γ-tocopherol methyltransferase gene contributes to elevated α-tocopherol content in soybean seeds. *BMC Plant Biol.* 11, 152. doi: 10.1186/1471-2229-11-152

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 (1), 161. doi: 10.1186/s13059-017-1289-9

Hamblin, M., Buckler, E. S., and Jannink, J.-L. (2011). Population genetics of genomics-based crop improvement methods. *Trends Genet.* 27, 98–106. doi: 10.1016/j.tig.2010.12.003

Han, Y., Zhao, X., Cao, G., Wang, Y., Li, Y., Liu, D., et al. (2015). Genetic characteristics of soybean resistance to HG type 0 and HG type 1.2.3.5.7 of the cyst nematode analyzed by genome-wide association mapping. *BMC Genomics* 16, 598. doi: 10.1186/s12864-015-1800-1

Han, Y., Zhao, X., Liu, D., Li, Y., Lightfoot, D. A., Yang, Z., et al. (2016). Domestication footprints anchor genomic regions of agronomic importance in soybeans. *New Phytol.* 209, 871–884. doi: 10.1111/nph.13626

Huang, M., Liu, X., Zhou, Y., Summers, R. M., and Zhiwu, Z. (2018). BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. *Gigascience* 8, 1–12. doi: 10.1093/gigascience/giy154

Jing, Y., Zhao, X., Wang, J., Teng, W., Qiu, L., Han, Y., et al. (2018). Identification of the genomic region underlying seed weight per plant in soybean (*Glycine max* l. merr.) *via* high-throughput single-nucleotide polymorphisms and a genome-wide association study. *Front. Plant Sci.* 91392. doi: 10.3389/fpls.2018.01392

Kim, M.-S., Lozano, R., Kim, J., Bae, D., Kim, S., Park, J.-H., et al. (2021). The patterns of deleterious mutations during the domestication of soybean. *Nat. Commun.* 12 (1), 97. doi: 10.1038/s41467-020-20337-3

Kolde, R. (2012) *Pheatmap: Pretty Heatmaps*. Available at: https://CRAN.Rproject.org/package=pheatmap (Accessed Aug 15, 2022).

Kumar, V., Rani, A., Dixit, A. K., Bhatnagar, D., and Chauhan, G.S (2009). Relative changes in tocopherols, isoflavones, total phenolic content, and antioxidative activity in soybean seeds at different reproductive stages.*J. Agr Food Chem.* 57 (7), 2705–2710. doi: 10.1021/jf803122a

Lemay, M. A., Sibbesen, J. A., Torkamaneh, D., Hamel, J., Levesque, R. C., and Belzile, F. (2022). Combined use of Oxford nanopore and illumina sequencing yields insights into soybean structural variation biology. *BMC Biol.* 20 (1), 53. doi: 10.1186/s12915-022-01255-w

Li, H., Liu, H., Han, Y., Wu, X., Teng, W., Liu, G., et al. (2010). Identification of QTL underlying vitamin e contents in soybean seed among multiple environments. *Theor. Appl. Genet.* 120 (7), 1405–1413.

Li, H., Wang, Y., Han, Y., Teng, W., Zhao, X., Li, Y., et al. (2016). Mapping quantitative trait loci (QTLs) underlying seed vitamin E content in soybean with main, epistatic and QTL x environment effects. *Plant Breed* 135 (2), 208–214. doi: 10.1111/pbr.12346

Lipka, A., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P., et al. (2012). GAPIT: genome association and prediction integrated tool. *Bioinformatics* 28 (18), 2397–2399. doi: 10.1093/bioinformatics/bts444

Liu, H., Cao, G., Wu, D., Jiang, Z., Han, Y., and Li, W. (2017). Quantitative trait loci underlying soybean seed tocopherol content with main additive, epistatic and QTL × environment effects. *Plant Breeding.* 136(6), 924–938. doi: 10.1111/pbr.12534

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 (2), e1005767. doi: 10.1371/journal.pgen.1005767

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *Methods.* 25, 402–408. doi: 10.1006/meth.2001.1262

Li, M., Zhang, Y. W., Xiang, Y., Liu, M. H., and Zhou, Y. H. (2022b). IIIVmrMLM: the r and c++ tools associated with 3VmrMLM, a comprehensive GWAS method for dissecting quantitative traits. *Mol. Plant* 15 (8), 1251–1253. doi: 10.1016/j.molp.2022.06.002

Li, M., Zhang, Y. W., Zhang, Z. C., Xiang, Y., Liu, M. H., Zhou, Y. H., et al. (2022a). A compressed variance component mixed model for detecting QTNs and QTN-by-environment and QTN-by-QTN interactions in genome-wide association studies. *Mol. Plant* 15 (4), 630–650. doi: 10.1016/j.molp.2022.02.012

Lu, K., Li, T., He, J., Chang, W., Zhang, R., Liu, M., et al. (2018). qPrimerDB: a thermodynamics-based gene-specific qPCR primer database for 147 organisms. *Nucleic Acids Res* 46 (D1), 1229–1236. doi: 10.1093/nar/gkx725

Meagher, E. A., Barry, O. P., Lawson, J. A., Rokach, J., and FitzGerald, G. A. (2001). Effects of vitamin e on lipid peroxidation in healthy persons. *JAMA* 285 (9), 1178–1182. doi: 10.1001/jama.285.9.1178

Munne-Bosch, S., and Alegre, L. (2002). The function of tocopherols and tocotrienols in plants. *J. Crit. Rev. Plant Sci.* 21, 31–57. doi: 10.1080/0735-260291044179

Packer, L., and Fuchs, J. (1993). Vitamin e in health and disease. *J. Crc Press.*

Park, C., Dwiyanti, M. S., Nagano, A. J., Liu, B., Yamada, T., and Abe, J. (2019). Identification of quantitative trait loci for increased alpha-tocopherol biosynthesis in wild soybean using a high-density genetic map. *BMC Plant Biol.* 19 (1), 510. doi: 10.1021/jf100455f

Pinelli-Saavedra, A., Calderón de la Barca, A. M., Hernández, J., Valenzuela, R., and Scaife, J. R. (2008). Effect of supplementing sows' feed with alpha-tocopherol acetate and vitamin c on transfer of alpha-tocopherol to piglet tissues, colostrum, and milk: aspects of immune status of piglets. *Res. Vet. Sci.* 85 (1), 92–100. doi: 10.1016/j.rvsc.2007.08.007

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D.. (2006). Principal components analysis corrects for stratification in genomewide association studies. *Nat. Genet.* 38 (8), 904–909. doi: 10.1038/ng1847

Rozanowska, M., Edge, R., Land, E. J., Navaratnam, S., Sarna, T., and Truscott, T. G. (2019). Scavenging of retinoid cation radicals by urate, trolox, and α-, β-, γ-, and δ-tocopherols. *IJMS.* 20 (11), 2799. doi: 10.3390/ijms20112799

Seguin, P., Tremblay, G., and Pageau, D. (2010). Soybean tocopherol concentrations are affected by crop management. *J. Agric. Food Chem.* 58 (9) 5495–5501. doi: 10.1021/jf100455f

Segura, V., Vilhjálmsson, B., Platt, A., Korte, A., Seren, Ü., and Long, Q. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* 44, 825–830. doi: 10.1038/ng.2314

Shaw, E. J., Kakuda, Y., and Rajcan, I. (2016). Effect of genotype, environment, and genotype×environment interaction on tocopherol accumulation in soybean seed. *J. Crop Sci.* 56, 40–50. doi: 10.2135/cropsci2015.02.0069

Shaw, E., and Rajcan, I. (2017). Molecular mapping of soybean seed tocopherols in the cross AC Bayfield X OAC Shire. *Plant Breeding* 136, 83–93. doi: 10.1111/pbr.12437

Sui, M., Jing, Y., Li, H., Zhan, Y., Luo, J., Teng, W., et al. (2020). Identification of loci and candidate genes analyses for tocopherol concentration of soybean seed. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.539460

Sun, X., Liu, D., Zhang, X., Li, W., Liu, H., Hong, W., et al. (2013). SLAF-seq: an efficient method of large-scale *de novo* SNP discovery and genotyping using high-throughput sequencing. *PloS One* 8 (3), e58700. doi: 10.1371/journal.pone.0058700

Tian, D., Wang, P., Tang, B., Teng, X., Li, C., Liu, X., et al. (2020). GWAS atlas: a curated resource of genome-wide variant-trait associations in plants and animals. *Nucleic Acids Res.* 48 (D1), D927–D932. doi: 10.1093/nar/gkz828

Ujiie, A., Yamada, T., Fujimoo, K., Endo, Y., and Kitamura, K. (2005). Identification of soybean varities with high levels of α-tocopherol content. *Breed Sci.* 55 (2), 123–125. doi: 10.1270/jsbbs.55.123

Wang, S.-B., Feng, J.-Y., Ren, W.-L., Huang, B., Zhou, L., Wen, Y.-J., et al. (2016). Improving power and accuracy of genome-wide association studies *via* a multi-locus mixed linear model methodology. *Sci. Rep.* 6, 19444. doi: 10.1038/srep19444

Wan, J., Zhang, W., Jiang, B., Guo, Y., and Hu, C. (2008). Separation of individual tocopherols from soybean distillate by low pressure column chromatography. *J. Am. Oil Chem. Soc* 85 (4), 331–338. doi: 10.1007/s11746-008-1198-x

Wen, Y., Zhang, H., Ni, Y., Huang, B., Zhang, J., and Feng, J. (2018). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief. Bioinf.* 19, 700–712. doi: 10.1093/bib/bbw145

Xiao, Q., Bai, X., Zhang, C., and He, Y. (2022). Advanced high-throughput plant phenotyping techniques for genome-wide association studies: A review. *J. Adv. Res.* 35, 215–230. doi: 10.1016/j.jare.2021.05.002

Xiao, J., Zhou, Y., He, S., and Ren, W.-L. (2021). An efficient score test integrated with empirical bayes for genome-wide association studies. *Front. Genet.* 12. doi: 10.3389/fgene.2021.742752

Xu, L., Dong, Z., Fang, L., Luo, Y., Wei, Z., Guo, H., et al. (2019). OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res.* 47, W52–W58. doi: 10.1093/nar/gkz333

Yano, K., Morinaka, Y., Wang, F., Huang, P., Takehara, S., Hirai, T., et al. (2019). GWAS with principal component analysis identifies a gene comprehensively controlling rice architecture. *Proc. Natl. Acad. Sci.* 116 (42), 21262–21267. doi: 10.1073/pnas.1904964116

Young, N., and Bharti, A. (2012). Genome-enabled insights into legume biology. *Annu. Rev. Plant Biol.* 63, 283–305. doi: 10.1146/annurev-arplant-042110-103754

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

Yu, Y., Zhang, H., Long, Y., Shu, Y., and Zhai, J. (2022). Plant public RNA-seq database: a comprehensive online database for expression analysis of ~45 000 plant public RNA-seq libraries. *Plant Biotechnol. J.* 20 (5), 806–808. doi: 10.1111/pbi.13798

Zhang, Z., Ersoz, E., Lai, C., Todhunter, R., Tiwari, H., and Gore, M. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42, 355–360. doi: 10.1038/ng.546

Zhang, Y., Jia, Z., and Dunwell, J. (2019). Editorial: the applications of new multi-locus GWAS methodologies in the genetic dissection of complex traits. *Front. Plant Sci* 10. doi: 10.3389/fpls.2019.00100

Zhang, L., Luo, Y., Zhu, Y., Zhang, L., Chen, R., Xu, M., et al. (2013). GmTMT2a from soybean elevates the α-tocopherol content in corn and arabidopsis. *Transgenic Res.* 22 (5), 1021-1028. doi: 10.1007/s11248-013-9713-8

Zhan, Y., Li, H., Sui, M., Zhao, X., Jing, Y., Luo, J., et al. (2020). Genome wide association mapping for tocopherol concentration in soybean seeds across multiple environments. *Ind. Crops Products.* 154, 1-15. doi: 10.1016/j.indcrop.2020.112674

Zhao, X., Dong, H., Chang, H., Zhao, J., Teng, W., Qiu, L., et al. (2019). Genome wide association mapping and candidate gene analysis for hundred seed weight in soybean [*Glycine max* (L.) Merrill]. *BMC Genomics* 20 (1), 648. doi: 10.1186/s12864-019-6009-2

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* 33 (4), 408–414. doi: 10.1038/nbt.3096

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 07 November 2022
DOI 10.3389/fpls.2022.1048860

OPEN ACCESS

EDITED BY
Yuan-Ming Zhang,
Huazhong Agricultural University, China

REVIEWED BY
Hailan Liu,
Maize Research Institute of Sichuan Agricultural University, China
Xuehai Zhang,
Henan Agricultural University, China
Yang-Jun Wen,
Nanjing Agricultural University, China

*CORRESPONDENCE
Jie Luo
jie.luo@hainanu.edu.cn
Liqiang He
heliqiang66@126.com

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

RECEIVED 20 September 2022
ACCEPTED 18 October 2022
PUBLISHED 07 November 2022

CITATION
He L, Wang H, Sui Y, Miao Y, Jin C and Luo J (2022) Genome-wide association studies of five free amino acid levels in rice.
Front. Plant Sci. 13:1048860.
doi: 10.3389/fpls.2022.1048860

Genome-wide association studies of five free amino acid levels in rice

Liqiang He[1*†], Huixian Wang[1†], Yao Sui[1†], Yuanyuan Miao[1,2], Cheng Jin[1,2] and Jie Luo[1,2*]

[1]College of Tropical Crops, Hainan University, Haikou, China, [2]Sanya Nanfan Research Institute of Hainan University, Hainan Yazhou Bay Seed Laboratory, Sanya, China

Rice (*Oryza sativa* L.) is one of the important staple foods for human consumption and livestock use. As a complex quality trait, free amino acid (FAA) content in rice is of nutritional importance. To dissect the genetic mechanism of FAA level, five amino acids' (Val, Leu, Ile, Arg, and Trp) content and 4,325,832 high-quality SNPs of 448 rice accessions were used to conduct genome-wide association studies (GWAS) with nine different methods. Of these methods, one single-locus method (GEMMA), seven multi-locus methods (mrMLM, pLARmEB, FASTmrEMMA, pKWmEB, FASTmrMLM, ISIS EM-BLASSO, and FarmCPU), and the recent released 3VmrMLM were adopted for methodological comparison of quantitative trait nucleotide (QTN) detection and identification of stable quantitative trait nucleotide loci (QTLs). As a result, 987 QTNs were identified by eight multi-locus GWAS methods; FASTmrEMMA detected the most QTNs (245), followed by 3VmrMLM (160), and GEMMA detected the least QTNs (0). Among 88 stable QTLs identified by the above methods, 3VmrMLM has some advantages, such as the most common QTNs, the highest LOD score, and the highest proportion of all detected stable QTLs. Around these stable QTLs, candidate genes were found in the GO classification to be involved in the primary metabolic process, biosynthetic process, and catalytic activity, and shown in KEGG analysis to have participated in metabolic pathways, biosynthesis of amino acids, and tryptophan metabolism. Natural variations of candidate genes resulting in the content alteration of five FAAs were identified in this association panel. In addition, 95 QTN-by-environment interactions (QEIs) of five FAA levels were detected by 3VmrMLM only. GO classification showed that the candidate genes got involved in the primary metabolic process, transport, and catalytic activity. Candidate genes of QEIs played important roles in valine, leucine, and isoleucine degradation (QEI_09_03978551 and candidate gene *LOC_Os09g07830* in the Leu dataset), tryptophan metabolism (QEI_01_00617184 and candidate gene *LOC_Os01g02020* in the Trp dataset), and glutathione metabolism (QEI_12_09153839 and candidate gene *LOC_Os12g16200* in the Arg

dataset) pathways through KEGG analysis. As an alternative of the multi-locus GWAS method, these findings suggested that the application of 3VmrMLM may provide new insights into better understanding FAA accumulation and facilitate the molecular breeding of rice with high FAA level.

KEYWORDS

rice, free amino acid level, genome-wide association study, quantitative trait locus, quantitative trait nucleotide-by-environment interactions

Introduction

Rice (*Oryza sativa* L.) is one of the most important crops worldwide and provides energy, amino acid, and dietary fiber for human consumption. In addition to the basic unit in protein biosynthesis, amino acids are involved in several cellular responses to affect physiological processes in plants, such as plant growth and development, intracellular pH control, production of metabolic energy or redox capacity, signal transduction, and response to abiotic and biotic stresses (Moe, 2013; Watanabe et al., 2013; Zeier, 2013; Fagard et al., 2014; Galili et al., 2014; Pratelli and Pilot, 2014: Hausler et al., 2014; Hildebrandt et al., 2015). Free amino acids (FAAs) not only play essential roles in plant growth, development, and responses to stress, but also serve as important nutrients for human health (Pathria and Ronai, 2021; Yang et al., 2022). Of all the amino acids, tryptophan (Trp), isoleucine (Ile), leucine (Leu), and valine (Val) are essential amino acids that are based on plants and cannot synthesize from external sources (Galili et al., 2016). In plants, branched-chain amino acids are important compounds in several aspects. Besides their function as building blocks of proteins, they get involved in the synthesis of a number of secondary products in plants and regulate plant growth by affecting the homeostasis of mineral elements in rice (Diebold et al., 2002; Jin et al., 2019). Arginine (Arg) is a semi-essential amino acid and involved in the regulation of various molecular pathways, which regulates key metabolic, immune, and neural signaling pathways in human cells (Patil et al., 2016). Branched-chain amino acids mainly including leucine, valine, and isoleucine generally participate in regulating protein synthesis, metabolism, food intake, and aging (Le Couteur et al., 2020). Arginine is a precursor of amino acids, polyamines, and nitric oxide (NO) for protein synthesis and is an important metabolite for many cells at the developmental stage (VanEtten et al., 1963; King and Gifford, 1997). Arginine is generally a major nitrogen storage form also in underground storage organs, roots of trees, and other plants (Bausenwein et al., 2001; Rennenberg et al., 2010). Tryptophan (Trp) is an aromatic amino acid that is synthesized through the shikimate/chorismate pathway. Notably, Trp is decarboxylated to tryptamine *in vivo*; subsequently, hydroxylase catalyzes the conversion of tryptamine to 5-hydroxytryptamine (5-HT). 5-HT is an important neurotransmitter associated with a range of human behavior problems such as personality and emotional disorders (Muller et al., 2016). Tryptophan provides the structural backbone for numerous plant secondary metabolites including the indoleamines, auxin [indole-3-acetic acid (IAA)], alkaloids, and benzoxazinoids (Erland and Saxena, 2019). Numerous loci with small effect underlying the natural variation of primary metabolites were found in previous studies (Rowe et al., 2008; Chan et al., 2010; Joseph et al., 2013; Fernie and Tohge, 2017). However, as one of primary metabolites, the genetic mechanism underlying these five FAA levels in rice is largely unknown, which is a limitation to the molecular breeding of rice with high-level FAAs.

Genome-wide association studies (GWAS) provide an insight into unraveling the genetic basis of complex traits in plants, especially for the trait controlled by small-effect genes (Zhu et al., 2008). Since the landmark GWAS of 107 Arabidopsis accessions (Atwell et al., 2010), GWAS of several agronomical traits in plants have been reported, which included starch content in wheat (Hao et al., 2020), flowering time and grain yield in rice (Yang et al., 2014; Liu et al., 2021), and seed protein and oil in soybean (Kim et al., 2021). With the technical progress and cost reduction of metabolomics, metabolite-based genome-wide association study (mGWAS) has been successfully applied in several functional genomics and metabolomics studies in plants (Luo, 2015; Fang et al., 2016; Fang and Luo, 2019).

Previous studies have proven the effectively controlled spurious association of widely adopted single-locus GWAS methods (Yu et al., 2006; Zhou and Stephens, 2012). However, the stringent Bonferroni correction is commonly used as the significant threshold of marker–trait associations (MTAs), which may result in the low power of polygenic loci detection in these methods (Zhang Y.M. et al., 2019). Thus, multi-locus GWAS methods have been proposed and identified quantitative trait nucleotide/locus (QTN/QTL) with small effect in a powerful manner (Segura et al., 2012). For instance, the improved statistical power and short computing time have been shown in the implementation of the FarmCPU method (Liu et al., 2016). The improvement of power and accuracy of the multi-locus GWAS method mrMLM have been reported (Wang et al., 2016). Additionally, a series of multi-locus models were

proposed and released in R package mrMLM, which contained mrMLM (Wang et al., 2016), pLARmEB (Zhang et al., 2017), FASTmrEMMA (Wen et al., 2017), pKWmEB (Ren et al., 2017), FASTmrMLM (https://cran.r-project.org/web/packages/mrMLM/index.html), and ISIS EM-BLASSO (Tamba et al., 2017). However, the additive and dominance effects of trait-associated loci remain unclear. To address this issue, a new multi-locus GWAS method, 3VmrMLM, was proposed to estimate the genetic effects of three marker genotypes (AA, Aa, and aa) by controlling all the possibly polygenic backgrounds. Subsequently, these effects were further divided into additive and dominance effects for QTNs. Moreover, QTN-by-environment interactions (QEIs) were also able to be detected by 3VmrMLM for dissecting the genetic architecture of complex and multi-omics traits in GWAS (Li et al., 2022a).

To identify the QTLs associated with five FAAs levels, GWAS was performed on a genetic panel including 448 accessions with 4,325,832 SNPs from the rice core collection using nine statistical methods. Of these methods, one single-locus method, seven previous released multi-locus methods, and the recent proposed 3VmrMLM method were employed to determine the reliable approaches for main-effect QTLs and QEI detection of five FAA contents.

Materials and methods

Genetic panel for GWAS

A genetic panel of 448 rice accessions from our lab—a previously released core collection by Chen et al. (2014)—was used in Huazhong Agricultural University. It included 293 *indica* and 155 *japonica* accessions, of which 362 varieties are from Asia, 22 varieties are from America, 8 rice accessions are from Africa, 13 accessions are from Europe, 3 varieties are from Oceania and, 40 varieties have unknown geographical information.

Metabolite profiling and sequencing

Two biological replicates of the 448 rice accessions grew in the normal rice growing season at two different blocks of Huazhong Agricultural University, Wuhan, China. For each replicate, randomly designed planting materials were used to harvest leaves at the five-leaf stage in liquid nitrogen of three different plants in each row of the field for metabolite extraction. Then, mix the material for biological replicate of each accession. The broad-sense heritability H^2 was calculated by using the data collected from different biological replicates at two different experimental bases of Huazhong Agricultural University. A scheduled multiple reaction monitoring (MRM) method with an MRM detection window of 80 s and a target scan time of 1.5 s were used to quantify the FAAs (Chen et al., 2013). Log_2-transformed metabolite data were used for further analysis to improve normality.

To identify the genetic variation of 448 rice accessions, approximately 448 Gb high-quality genome sequences of these accessions were obtained from the Illumina HiSeq 2000 platform (Chen et al., 2014). Rice reference genome sequence MSU 6.1 (Nipponbare, version 6.1) and corresponding annotation were downloaded from Rice Genome Annotation Project (http://rice.uga.edu/index.shtml). Clean reads were mapped to the rice reference genome using BWA software (https://sourceforge.net/projects/bio-bwa/) with default settings. The mapping files were processed with SAMtools software (Li et al., 2009). HaplotypeCaller, CombineGVCFs and GenotypeGVCFs functions with default settings in GATK software (https://gatk.broadinstitute.org/hc/en-us) were used for SNP joint-calling and filter of the 448 accessions. Filtered high-quality SNPs (–maf 0.05 and –geno 0.1 in PLINK software, https://zzz.bwh.harvard.edu/plink/) were used for subsequent analysis.

PCA and phylogenetic analysis

To summarize the genetic structure and variation of 448 rice accessions, principal component analysis (PCA) was conducted by PLINK software using the obtained high-quality SNPs. Furthermore, SNP-based phylogenetic analysis of all accessions was performed by MEGA-CC with a pairwise gap deletion method for 1,000 bootstrap replicates (Kumar et al., 2012).

Population structure and linkage disequilibrium

ADMIXTURE software was employed to estimate the population stratification of all accessions (Alexander et al., 2009). To evaluate LD decay across the whole genome, the squared correlation coefficient (r^2) between SNPs was computed and plotted using PopLDdecay software (Zhang C. et al., 2019).

Genome-wide association study

GWAS were performed on the association panel containing 448 rice accessions with 4,325,832 high-quality SNPs. In total, nine models were implemented for GWAS, which included a single-locus model GEMMA (Zhou and Stephens, 2012) and eight multi-locus models, namely, FarmCPU (Liu et al., 2016), mrMLM (Wang et al., 2016), pLARmEB (Zhang et al., 2017), FASTmrEMMA (Wen et al., 2017), pKWmEB (Ren et al., 2017), FASTmrMLM (https://cran.r-project.org/web/packages/mrMLM/index.html), ISIS EM-BLASSO (Tamba et al., 2017), and 3VmrMLM (Li et al., 2022a). The R package mrMLM composed of six multi-locus methods mrMLM, pLARmEB, FASTmrEMMA, pKWmEB, FASTmrMLM,

and ISIS EM-BLASSO was applied to test the marker and trait association. mrMLM parameter for six methods: Likelihood="REML", SearchRadius=20, CriLOD=3, SelectVariable=50, and Bootstrap=FALSE. These six methods in the mrMLM package were developed and released from the same research group that were referred to as "mrMLM series methods". The LOD score ≥ 3 was used to detect the association signals of mrMLM series methods by default. The new released 3VmrMLM method, implemented by the IIIVmrMLM software (Li et al., 2022b), was used to detect main-effect quantitative trait nucleotide (QTN) and QTN by environment interaction (QEI). 3VmrMLM parameter for main-effect QTL: method="Single_env", SearchRadius=20, and svpal=0.01. 3VmrMLM parameter for QEI: method="Multi_env", SearchRadius=20, and svpal=0.01. The threshold of significant association of other methods was determined by a critical *p*-value at the 0.05 significant level subjected to Bonferroni correction (*p*-value = 1.16×10^{-8}). All methods used in this study were implemented with default parameters. Manhattan and QQ plots were drawn using R CMplot, mrMLM, and 3VmrMLM packages with default settings.

Analysis of candidate genes

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway annotation of candidate genes was analyzed by the Plant GeneSet Enrichment Analysis Toolkit (PlantGSEA) (Yi et al., 2013). The annotation of SNP effects on gene body was obtained from the RiceVarMap database (http://ricevarmap.ncpgr.cn/) and further used for haplotype and content analysis of potential candidate genes. Haplotype network was generated according to all information of a candidate gene from RiceVarMap database (http://ricevarmap.ncpgr.cn/). Temporal and spatial expression of potential candidate genes were assayed based on the expression data from electronic fluorescent pictograph Browser (ePlant) (http://bar.utoronto.ca/).

Results

FAA levels of rice genetic panel

The five FAA levels (Val, Leu, Ile, Arg, and Trp) were quantified by LC-MS/MS to evaluate the phenotypic variation in 448 rice accessions. The CV of them were 45.03%, 58.83%, 71.25%, 92.30%, and 58.21%, respectively (Table 1). Furthermore, significant differences on five FAA levels were observed between *indica* and *japonica* accessions in this rice genetic panel (Figure 1). High correlation of five FAA contents was observed among them. For instance, the Val dataset was highly correlated with the Leu (r = 0.83) and Ile (r = 0.90) datasets, and the Leu dataset was highly correlated with the Ile (r = 0.93) dataset (Supplementary Figure 1). The skewness and kurtosis of five FAA levels were less than 1, which showed the nature of quantitative traits (Supplementary Figure 1; Table 1). The broad-sense heritability (H^2) for Val, Leu, Ile, Arg, and Trp ranged from 0.32 to 0.51 (Table 1). These indicated the natural variation of five amino acids present in this genetic panel.

TABLE 1 Descriptive statistics of five FAA content datasets.

Trait	Val	Leu	Ile	Arg	Trp
Number	448	448	448	448	448
Mean	23.68	23.41	21.80	17.91	22.20
Standard deviation	0.65	0.83	0.92	0.97	0.78
Variance	0.42	0.69	0.84	0.95	0.61
Mean squared error	0.03	0.04	0.04	0.05	0.04
Median	23.72	23.40	21.81	17.93	22.20
Trimmed	23.70	23.42	21.80	17.91	22.20
Median absolute deviation	0.61	0.87	0.98	1.02	0.82
Minimum	21.69	20.96	19.42	15.01	20.36
Maximum	25.83	25.51	24.94	21.87	24.47
Range	4.14	4.55	5.52	6.86	4.11
Skewness	−0.26	−0.07	0.05	0.10	0.04
Kurtosis	0.06	−0.31	−0.17	0.49	−0.27
[a] Coefficient of variation (%)	45.03	58.83	71.25	92.30	58.21
Confidence interval of 0.95	0.06	0.08	0.09	0.09	0.07
H^2	0.32	0.51	0.46	0.38	0.43

[a]Calculated from the original dataset.

10.3389/fpls.2022.1048860

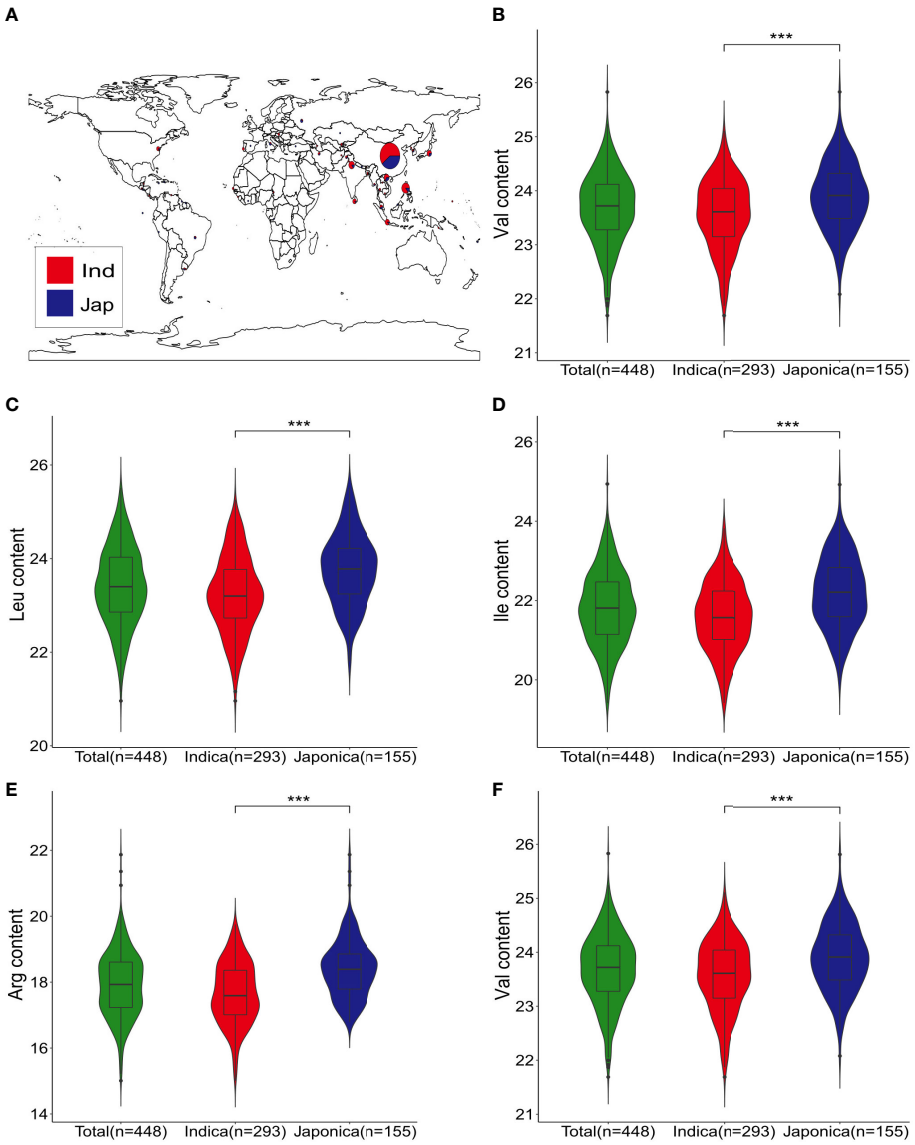


FIGURE 1
Geographic distribution and five FAA levels of genetic panel. **(A)** Geographic distribution of indica and japonica accessions in the genetic panel; indica accessions are indicated in red, and blue represents japionica accessions. **(B–F)** Violin plots of Val, Leu, Ile, Arg, and Trp contents for all, indica, and japonica accessions; *** indicate statistical significance at the 0.1% probability level

Population structure and phylogenetic relationship of rice genetic panel

To dissect the genetic basis underlying the natural variation of FAAs, the relationship assessment of rice genetic panel was based on 4,325,832 SNPs. According to the Neighbor-joining (NJ) phylogenetic tree, 448 rice accessions were mainly divided into two clades which contained 293 *indica* accessions and 155 *japonica* accessions, respectively (Figure 2A). Likewise, the classification of these accessions into two groups were observed in principal component analysis (PCA) (Figure 2B). Moreover, the population structure of rice genetic panel was identical with those obtained in NJ tree and PCA (Figure 2C). Linkage disequilibrium (LD) analysis showed that LD decayed fastest before 122 kb, and subsequently tended to be flat for the rice genetic panel (Figure 2D). Therefore, the 122- kb flanking region of each QTN was used for putative candidate gene prediction hereafter. Additionally, *indica* accessions showed the highest decay rate in Figure 2D.

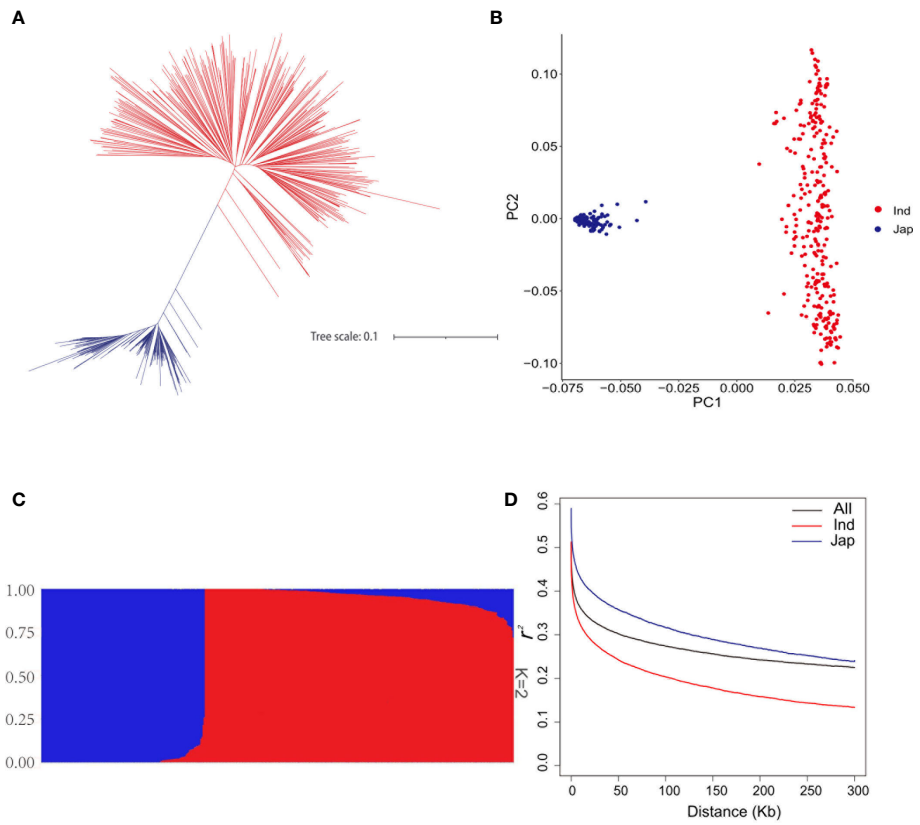


FIGURE 2
Population analyses of the genetic panel. **(A)** Phylogenetic tree of 448 rice accessions. **(B)** Principal component analysis of 448 rice accessions. **(C)** Population structure estimated by ADMIXTURE. **(D)** LD decay analysis of the genetic panel; LD decay of all 448 rice accessions, *indica* accessions, and *japonica* accessions is indicated in black, red, and blue, respectively.

Identification of five FAA-associated QTLs

In this study, a total of 987 QTNs are identified using nine GWAS methods (a single-locus method, seven multi-locus methods, and the recently released 3VmrMLM method) for five FAA content datasets. Detected QTNs varied resulting from statistical methods (Supplementary Table 1). 3VmrMLM detected 160 QTNs and the largest number of common QTNs, while no QTN was detected by GEMMA. In addition, the largest number of QTNs were identified in the Trp dataset (214) by eight multi-locus GWAS methods (3VmrMLM, mrMLM, FASTmrEMMA, pLARmEB, FASTmrMLM, pKWmEB, ISIS EM-BLASSO, and FarmCPU), followed by the Val dataset (207), the Ile dataset (203), the Arg dataset (195), and the smallest number of detected QTNs in the Leu dataset (168) (Figures 3A–E and Supplementary Figures 2A–E; Supplementary Table 1). Six mrMLM series methods were compared together; FASTmrEMMA detected the most QTNs (245), followed by pLARmEB (160), mrMLM (151), FASTmrMLM (145), pKWmEB (77), and ISIS EM-BLASSO, which detected the least QTNs (25) (Supplementary Figures 2A–E; Supplementary Table 1). Different R^2 values of common QTNs across methods were observed, such as the R^2 value (%) of 3VmrMLM-detected QTNs that ranged from 0.78 to 6.95, while the R^2 value (%) of the mrMLM-detected QTN dataset was from 0.43 to 17.61. The average R^2 value (%) of ISIS EM-BLASSO-detected QTNs was the highest (2.93) among nine GWAS methods, whereas the average R^2 value (%) of the QTNs detected by FarmCPU was the lowest (0.24) (Table 2). Tag QTNs were selected and referred to as QTLs hereafter.

In addition, some common QTLs were detected in different FAA datasets. Intriguingly, QTL_01_10944343 (this QTL ID refers to QTL_Chromosome_Position) and QTL_05_19754561 were associated with Val and Ile datasets, respectively; QTL_01_23419417 was co-detected in the Leu and Ile datasets; QTL_02_24189963 was co-localized in the Leu and Trp datasets; QTL_09_16065720 was detected in the Arg and Trp datasets simultaneously; and QTL_10_17905052 was identified in the Ile and Arg datasets (Supplementary Figure 3). Among nine GWAS methods, most *p*-values of the 3VmrMLM-detected common QTLs were the lowest and most of their LOD scores were the highest correspondingly (Table 2; Supplementary Table 1; Supplementary Figure 3). These results

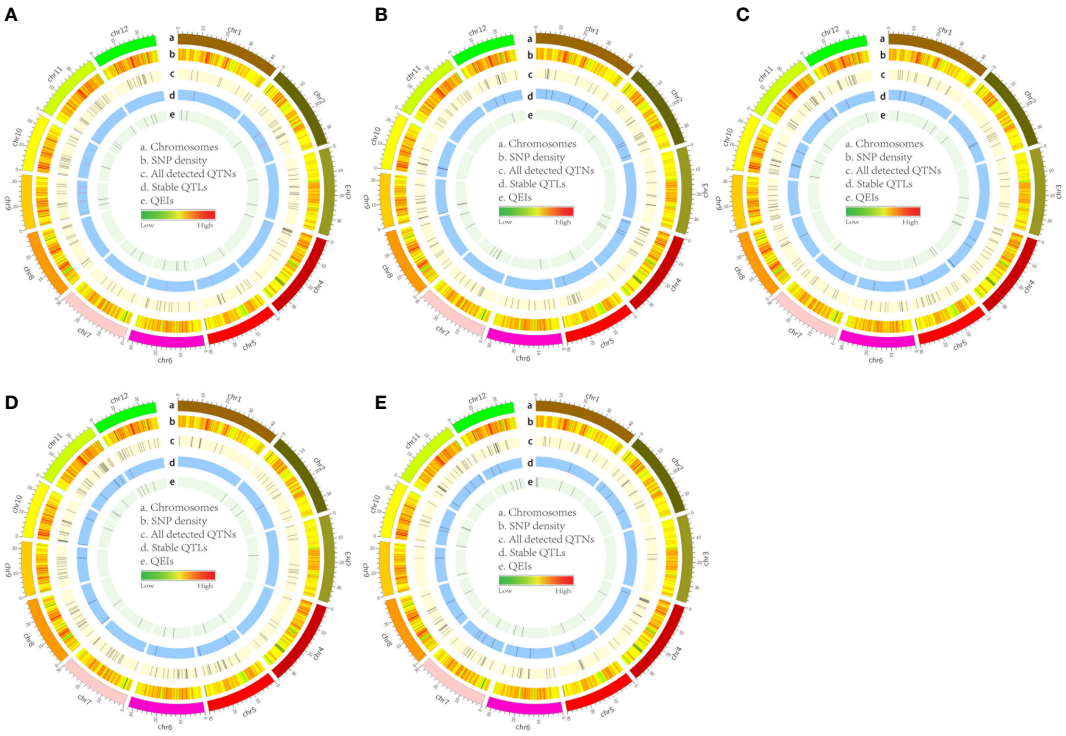


FIGURE 3
Circos map of QTLs and QEIs in rice genome identified from Val **(A)**, Leu **(B)**, Ile **(C)**, Arg **(D)**, and Trp **(E)** datasets. Track A: 12 rice chromosomes; Track B: heatmap of SNP density with bin sizes of 0.1 Mb; Track C: total unique QTNs detcted by all used methods; Track D: stable QTLs co-detected by no more than two methods; Track E: all detected QEIs by the 3VmrMLM method.

indicated that the common QTLs detected by 3VmrMLM across traits were more significant than those detected by other eight GWAS methods.

Stable FAA-associated QTLs and candidate genes

A QTL detected by no less than two methods of 3VmrMLM, mrMLM series methods (mrMLM, pLARmEB, FASTmrEMMA, pKWmEB, FASTmrMLM, and ISIS EM-BLASSO), FarmCPU, and GEMMA was defined as a stable QTL. A total of 88 stable QTLs were identified in five FAA datasets (Supplementary Table 2). Fifteen stable QTLs were detected in the Val dataset (Figures 3A, 4A). In particular, QTL_01_10944343 was identified by seven GWAS methods (3VmrMLM, mrMLM, FASTmrMLM, FASTmrEMMA, pLARmEB, pKWmEB, and FarmCPU), and the QTL was also detected in Ile (Supplementary Figure 3A; Supplementary Table 2). For the Trp dataset, 23 stable QTLs were identified (Figures 3E, 4E). Of these QTLs, QTL_09_16065720 was identified by six GWAS methods (3VmrMLM, FASTmrMLM, FASTmrEMMA, pLARmEB, pKWmEB, and ISIS EM-BLASSO), and it was detected in the Arg dataset simultaneously (Supplementary Figure 3E; Supplementary Table 2). Additionally, 16, 20, and 14 stable QTLs were detected in Leu, Ile, and Arg datasets (Figures 3B–D, 4B–D). Significant correlations between NPQTL (the number of QTL with positive-effect or favorite alleles) and five FAA contents were observed in Figures 5A–E (r = 0.53–0.69). The highest correlation was shown in the Trp dataset (r = 0.69) (Figure 5E).

To understand the molecular basis controlling the five FAA levels, the biological function of candidate genes was investigated. According to functional annotations, these candidate genes were primarily categorized as protein, protein kinase, glycosyltransferase, and transcription factor (Supplementary Table 3). Furthermore, GO analysis showed that these genes were classified into 51 GO terms, such as the primary metabolic process, biosynthetic process, and catalytic activity (Supplementary Figure 4). Meanwhile, KEGG analysis of candidate genes showed that most of them were involved in metabolic pathways; biosynthesis of amino acids; glycine, serine, and threonine metabolism; and tryptophan metabolism (Supplementary Figure 5), for instance, biosynthesis of amino acids in five FAA datasets (Supplementary Figures 5A–E); glycine, serine, and threonine metabolism in the Leu dataset

10.3389/fpls.2022.1048860

TABLE 2 Comparison of QTN/QTL identification for different GWAS methods.

Statistical method	No. of detected QTNs	No. of stable QTLs	Average R^2 (%)	R^2 range (%)	LOD range
3VmrMLM	160	83	1.99	0.78–6.95	3.04–46.29
FASTmrEMMA	245	29	1.01	0.01–8.93	3.01–24.01
FASTmrMLM	145	48	1.14	0.03–5.22	3.03–9.95
ISIS EM-BLASSO	25	9	2.93	0.98–6.89	3.01–10.65
mrMLM	151	19	2.54	0.43–17.61	3.06–21.49
pKWmEB	77	22	2.82	0.79–10.46	3.01–9.20
pLARmEB	160	34	1.46	0.01–14.39	3.02–14.80
FarmCPU	24	9	0.24	0.09–0.50	NA
GMMEA	0	0	NA	NA	NA

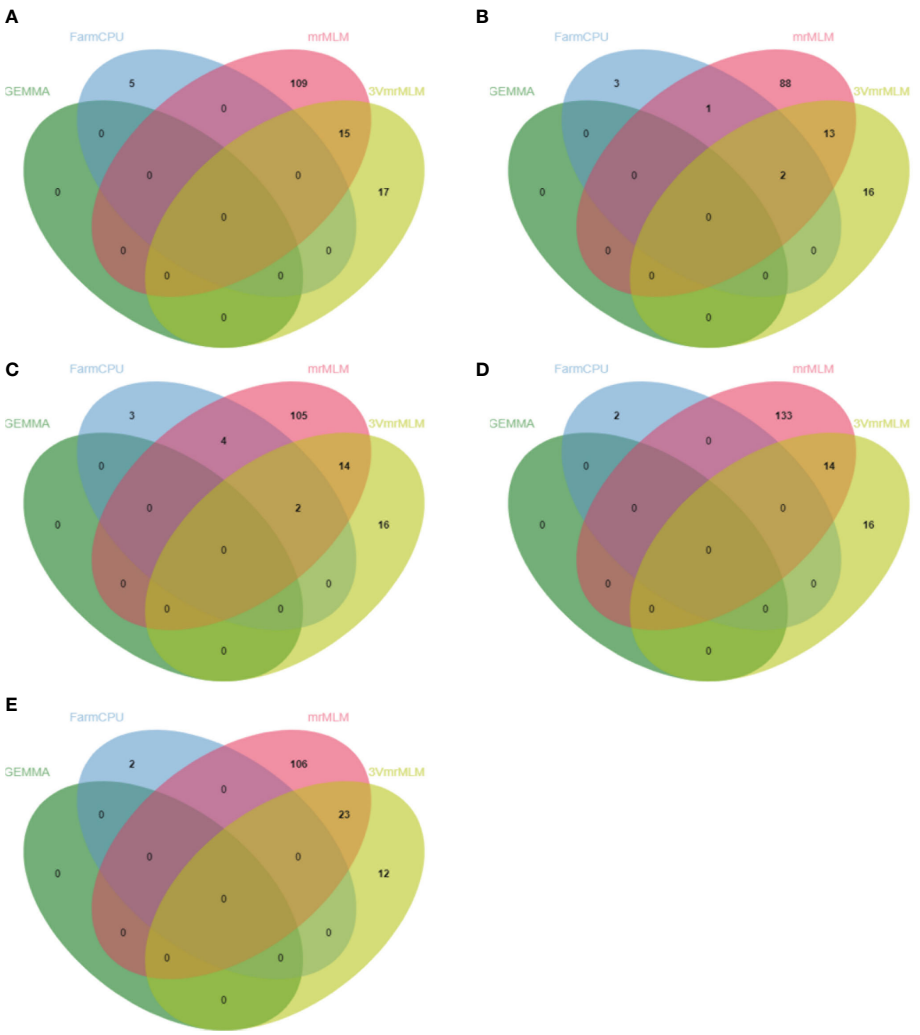


FIGURE 4
Venn diagrams of unique QTNs detected by different GWAS methods from Val **(A)**, Leu **(B)**, Ile **(C)**, Arg **(D)**, and Trp **(E)** datasets. mrMLM represents mrMLM series methods including mrMLM, FASTmrEMMA, pLARmEB, pKWmEB, ISIS EM-BLASSO, and FASTmrMLM.

10.3389/fpls.2022.1048860

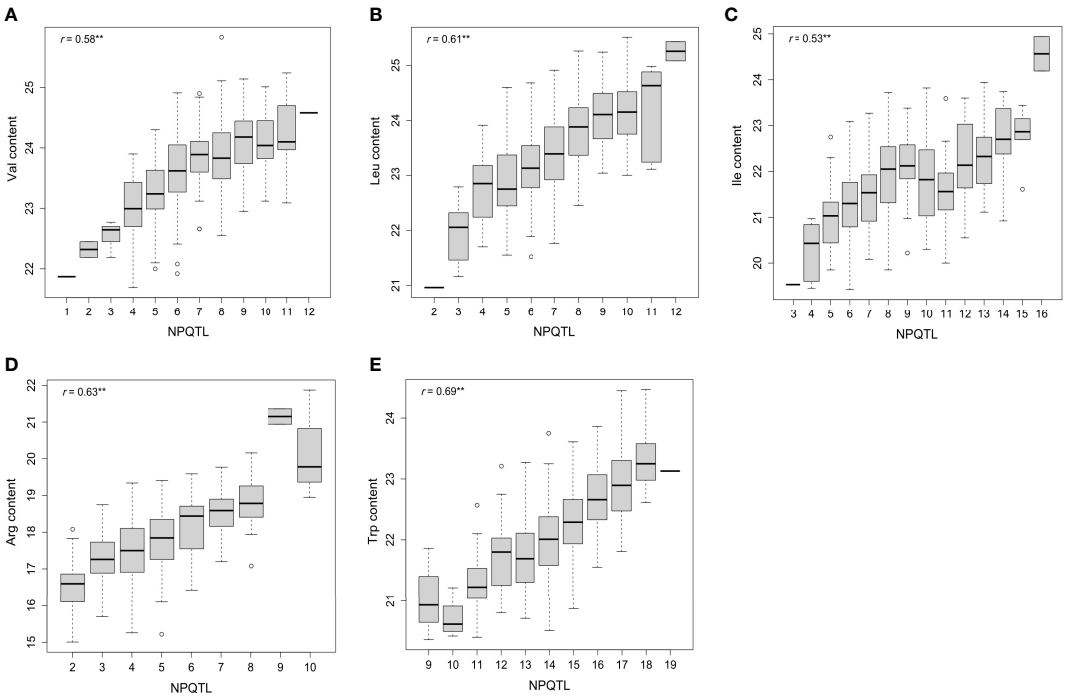
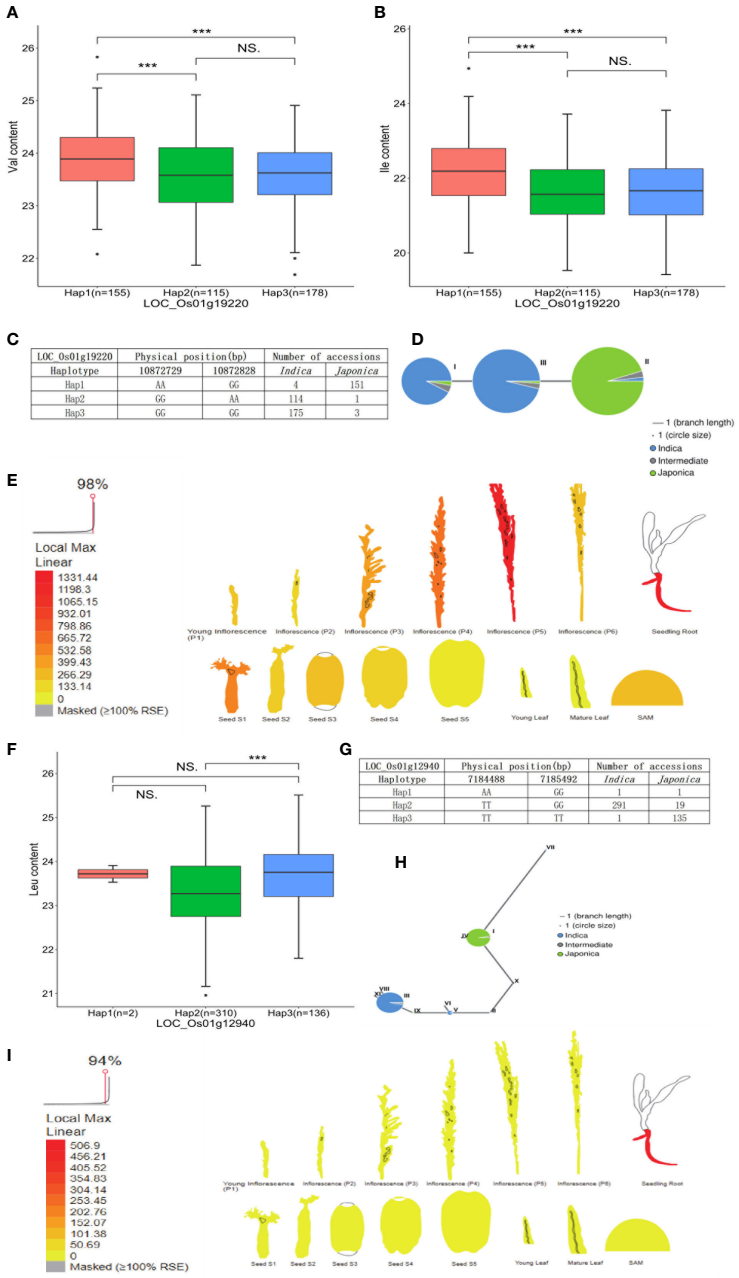


FIGURE 5
Box plots of the number of stable QTL with positive-effect alleles (NPQTL) in relation to Val, Leu, Ile, Arg, and Trp contents **(A–E)**. ** indicates statistical significance at the 1% probability level.

(Supplementary Figure 5B); and tryptophan metabolism in the Trp dataset (Supplementary Figure 5E).

The candidate gene *LOC_Os01g19220* encoding beta-D-xylosidase was identified in the Val and Ile datasets, which presented three types of alleles: Hap1 (AAGG) was concentrated in *japonica* accessions, while Hap2 (GGAA) and Hap3 (GGGG) were mainly concentrated in *indica* accessions, and the Val and Ile content of Hap1 was significantly different with the contents of Hap2 and Hap3. A lower Val and Ile content in Hap2 and Hap3 was observed than that in Hap1, which directly indicated the relatively high Val and Ile content present in *japonica* accessions compared with *indica* accessions (Figures 6A–C; Supplementary Table 4). Based on previous transcriptome and haplotype network analysis, *LOC_Os01g19220* was mainly expressed in seed (S1), inflorescence (P5), and seedling root. In the haplotype network, haplotype II of *LOC_Os01g19220* was mainly presented in *japonica* accessions; however, haplotypes I and III gathered in *indica* accessions (Figures 6D, E). Moreover, the gene *LOC_Os01g12940* encoding the phosphorylase domain containing protein detected in the Leu dataset had three types of allelic variation. Hap2 (TTGG) was concentrated in *indica* accessions, whereas Hap3 (TTTT) was concentrated in *japonica* accessions. A vast majority of *japonica* accessions with Hap3 showed significantly higher Leu level than *indica* accessions *with* Hap2 (Figures 6F, G; Supplementary Table 4). *LOC_Os01g12940* was highly expressed in seedling root. In the haplotype network, haplotype I of *LOC_Os01g12940* was concentrated in *japonica* accessions, while haplotypes III and V were concentrated in *indica* accessions (Figures 6H,I). In addition, the gene *LOC_Os05g49760* encoding the dehydrogenase is identified in the Arg dataset, which was involved in glutathione metabolism and had three types of allelic variation. Hap1 (AAGG) and Hap3 (GGGG) were enriched in *indica* accessions, and Hap2 (GGAA) was enriched in *japonica* accessions. Significant differences of Arg content were observed among accessions with Hap2, Hap1, and Hap3. Correspondingly, the Arg level of *japonica* accessions carrying Hap2 was higher than the *indica* accessions with Hap1 and Hap3 (Figures 7A, B; Supplementary Table 4). Relatively high abundance of *LOC_Os05g49760* was found in SAM (shoot apical meristem), young leaf, and inflorescence (P5). In the haplotype network, haplotype II was concentrated in *japonica* accessions, while haplotypes I and III gathered in *indica* accessions (Figures 7C, D). Moreover, the gene *LOC_Os11g06900* encoding amidase family protein detected in the Trp dataset had two alleles. Hap1 (CC) gathered in *indica* accessions, and Hap2 (TT) was mostly present in *japonica* accessions. Significant differences of Trp content were observed among accessions with Hap2 and Hap1. Subsequently, the Trp level of *japonica* accessions carrying Hap2 was higher than the *indica* accessions with Hap1 (Figures 7E, F; Supplementary Table 4). High expression of *LOC_Os11g06900*

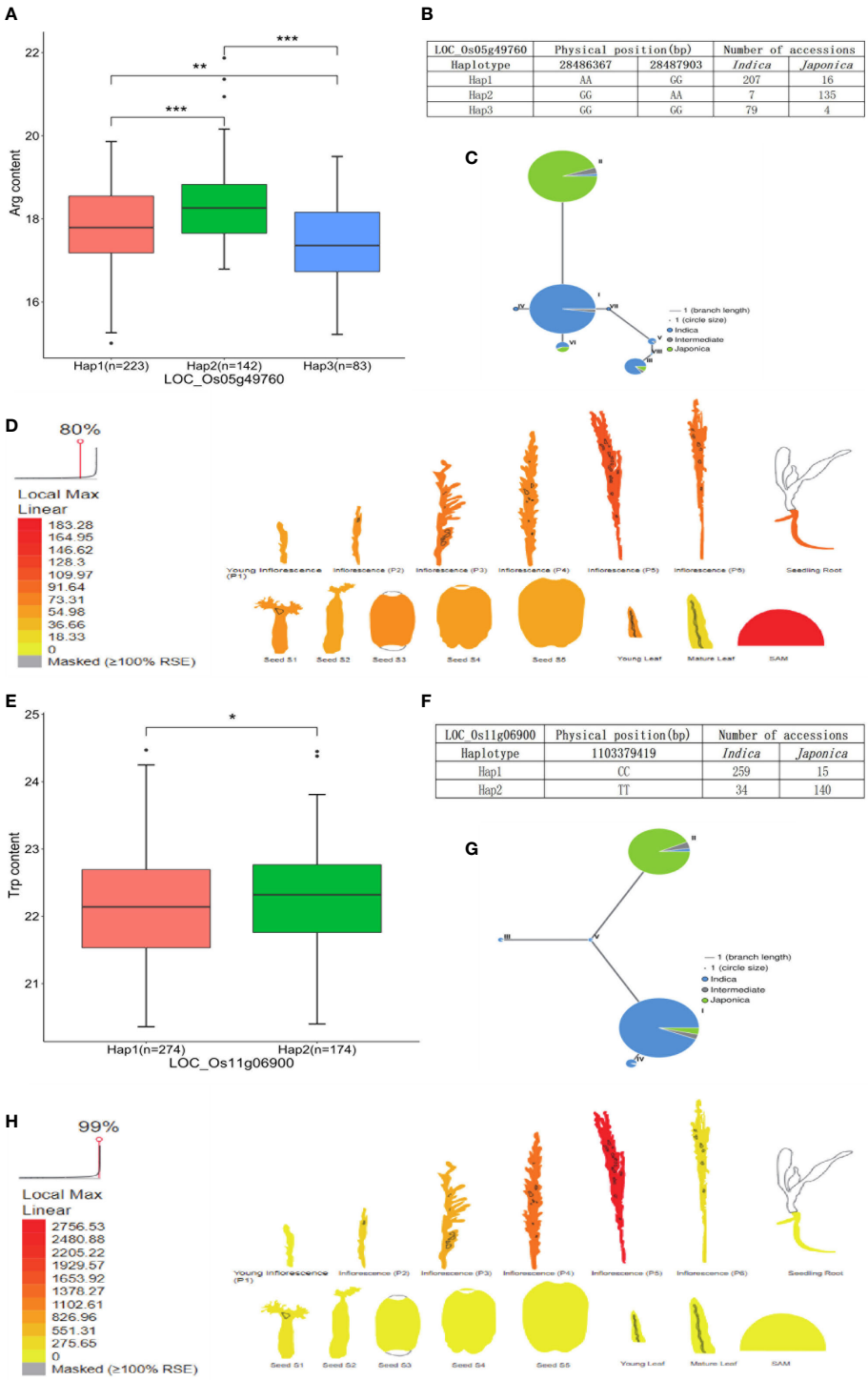
10.3389/fpls.2022.1048860



LOC_Os01g19220	Physical position(bp)		Number of accessions	
Haplotype	10872729	10872828	*Indica*	*Japonica*
Hap1	AA	GG	4	151
Hap2	GG	AA	114	1
Hap3	GG	GG	175	3

LOC_Os01g12940	Physical position(bp)		Number of accessions	
Haplotype	7184488	7185492	*Indica*	*Japonica*
Hap1	AA	GG	1	1
Hap2	TT	GG	291	19
Hap3	TT	TT	1	135

FIGURE 6
Analyses of Val and Ile level associated gene LOC_Os01g19220 and Leu level associated gene LOC_Os01g12940. **(A)** Significant tests between three haplotypes of LOC_Os01g19220 and Val contents. **(B)** Significant tests between three haplotypes of LOC_Os01g19220 and Ile contents. **(C)** Three haplotypes of LOC_Os01g19220 and their distribution in indica and japonica accessions. **(D)** Haplotype network of LOC_Os01g19220. **(E)** Expression profile of LOC_Os01g19220 based on ePlant transcriptome analysis in rice; expression strength coded by color from yellow (low) to red (high). **(F)** Significant tests between three haplotypes of LOC_Os01g12940 and Leu contents. **(G)** Three haplotypes of LOC_Os01g12940 and their distribution in indica and japonica accessions. **(H)** Haplotype network of LOC_Os01g12940. **(I)** Expression profile of LOC_Os01g12940 based on ePlant transcriptome analysis in rice, expression strength coded by color from yellow (low) to red (high). *** and NS indicate statistical significance at the 0.1% probability level and no significant difference, respectively.



LOC_Os05g49760	Physical position(bp)		Number of accessions	
Haplotype	28486367	28487903	*Indica*	*Japonica*
Hap1	AA	GG	207	16
Hap2	GG	AA	7	135
Hap3	GG	GG	79	4

LOC_Os11g06900	Physical position(bp)	Number of accessions	
Haplotype	1103379419	*Indica*	*Japonica*
Hap1	CC	259	15
Hap2	TT	34	140

FIGURE 7
Analyses of Arg level associated gene *LOC_Os05g49760* and Trp level associated gene *LOC_Os11g06900*. **(A)** Significant tests between three haplotypes of *LOC_Os05g49760* and Arg contents. **(B)** Three haplotypes of *LOC_Os05g49760* and their distribution in *indica* and *japonica* accessions. **(C)** Haplotype network of *LOC_Os05g49760*. **(D)** Expression profile of *LOC_Os05g49760* based on ePlant transcriptome analysis in rice, expression strength coded by color from yellow (low) to red (high). **(E)** Significant tests between two haplotypes of *LOC_Os11g06900* and Trp contents. **(F)** Three haplotypes of *LOC_Os11g06900* and their distribution in *indica* and *japonica* accessions. **(G)** Haplotype network of *LOC_Os11g06900*. **(H)** Expression profile of *LOC_Os11g06900* based on ePlant transcriptome analysis in rice, expression strength coded by color from yellow (low) to red (high). *, **, and *** indicate statistical significance at the 5%, 1%, and 0.1% probability level, respectively.

was observed in inflorescence (P5). In the haplotype network, haplotypes I, III, IV, and V of it gathered in *indica* accessions, whereas haplotype II was concentrated in *japonica* accessions (Figures 7G, H).

QEI detection of five FAAs

In total, 95 QEIs of five FAAs were detected by 3VmrMLM (Supplementary Table 5). Of them, 23, 16, 16, 18, and 22 QEIs were identified in the Val, Leu, Ile, Arg, and Trp datasets (Table 3). However, no QEI was detected on some chromosomes in five FAA datasets (Figure 3; Supplementary Figure 6). For instance, no QEI on chromosomes 8 and 3 was found in the Val and Trp datasets, respectively (Figures 3A, E); none of the QEIs on chromosomes 3, 10, and 11 were detected in the Leu dataset (Figure 3B); no QEI located on chromosomes 6, 8, and 9 was identified in the Ile dataset (Figure 3C); and no QEI located on chromosomes 4 and 9 was identified in the Arg dataset (Figure 3D). Based on biological process, molecular function, and cellular component in GO analysis, candidate genes of these detected QEIs were classified into 47 GO terms, such as metabolic process, transferase activity, and transport (Supplementary Figure 7). Furthermore, KEGG pathway analysis showed that candidate genes were mainly involved in glutathione metabolism (QEI_12_09153839 and its candidate gene *LOC_Os12g16200* in the Arg dataset), valine leucine and isoleucine degradation (QEI_09_03978551 and its candidate gene *LOC_Os09g07830* in the Leu dataset), and tryptophan metabolism (QEI_01_00617184 and its candidate gene *LOC_Os01g02020* in the Trp dataset) (Supplementary Figure 8 and Supplementary Table 6). In addition, cysteine and methionine metabolism in the Val dataset (Supplementary Figure 8A); tryptophan metabolism in the Trp dataset (Supplementary Figure 8E); and valine, leucine, and isoleucine degradation in the Leu dataset (Supplementary Figure 8B) are also shown in Supplementary Figure 8. According to ePlant analysis, high expression of *LOC_Os12g16200* encoding glutathione synthetase was observed in seedling root and mature leaf. *LOC_Os09g07830* encoding acetyl-CoA acetyltransferase was highly expressed in seedling root and SAM. Relatively high abundance of *LOC_Os01g02020* encoding acetyl-CoA acetyltransferase was found in young leaf and mature leaf.

Discussion

Methods comparison

Due to the difference of algorithm in different GWAS methods, the varied number of detected QTNs was observed accordingly. The FASTmrEMMA method detected the most QTNs (245), followed by 3VmrMLM (160), pLARmEB (160), mrMLM (151), FASTmrMLM (145), pKWmEB (77), ISIS EM-BLASSO (25), FarmCPU (24), and GEMMA, which detected the least QTNs (0) (Supplementary Table 1). Meanwhile, 3VmrMLM detected the largest number of common QTNs (Figure 4). Similar to the result obtained in this study, no QTN was identified in Xu et al. (2017) and Li et al. (2018) by GEMMA (Xu et al., 2017; Li et al., 2018). These were consistent with previous studies suggesting that multi-locus methods outperform single-locus methods on the statistical power of QTL detection, especially on the accuracy of QTN effect estimation and reduction of false-positive rate (Misra et al., 2017; Chang et al., 2018; Cui et al., 2018; Hou et al., 2018; Ma et al., 2018). The results of 3VmrMLM and mrMLM were compared as 3VmrMLM was a new three-variance component integrated with the mrMLM methodological framework. Most *p*-values of 3VmrMLM-detected QTNs were lower than those in mrMLM, and the LOD value of QTNs measured by 3VmrMLM was larger than the other eight methods (Supplementary Figure 3). These results indicated that the QTNs identified by 3VmrMLM were more significant than those identified by mrMLM. Additionally, the average R^2 value (%) of 3VmrMLM-detected QTNs was lower than that of mrMLM. The average R^2 value of ISIS EM-BLASSO (2.93) was the highest, followed by pKWmEB (2.82), mrMLM (2.54), 3VmrMLM (1.99), pLARmEB (1.46), FASTmrMLM (1.14), FASTmrEMMA (1.01), and FarmCPU (0.24) (Table 2). Notably, in this study, stable QTL_05_19754561 detected by 3VmrMLM/pLARmEB in the Val dataset, QTL_01_07646091 and QTL_07_08680072 detected by 3VmrMLM/mrMLM/pLARmEB/FarmCPU in the Ile dataset, QTL_11_22412156 detected by 3VmrMLM/pLARmEB in the Arg dataset, and

TABLE 3 QTN-by-environment interactions (QEIs) detected from five FAA content datasets.

Trait	No. of detected QEIs	R^2 range (%)	LOD range	add*env1 range	add*env2 range
Val	23	0.33–2.42	5.09–35.28	−0.13–0.15	−0.15–0.13
Leu	16	0.57–2.41	5.07–21.31	−0.15–0.12	−0.12–0.15
Ile	16	0.46–2.94	4.83–29.92	−0.19–0.12	−0.12–0.19
Arg	18	0.34–1.22	6.16–21.15	−0.14–0.14	−0.14–0.14
Trp	22	0.36–2.60	4.63–34.53	−0.16–0.11	−0.11–0.16

QTL_01_23592545 detected by 3VmrMLM/FASTmrEMMA in the Trp dataset were reported in a previous study (Chen et al., 2014). Furthermore, QTN-0315484798 detected by 3VmrMLM only and QTN-0134428638 (~5.55 kb downstream of QTN-vg0134424130 detected by mrMLM in Ile dataset; QTN-0107646091 detected by FarmCPU/mrMLM in the Val/Trp dataset; QTN-0100694213, QTN-0727264573, and QTN-1203473916 detected by mrMLM/ISIS EM-BLASSO/pLARmEB in the Arg dataset; and QTN-0619805830 detected by ISIS EM-BLASSO and QTN-0805618520 detected by mrMLM in the Trp dataset were consistent with previous studies (Chen et al., 2014; Sun et al., 2020). Six QTLs (QTL_01_10944343, QTL_01_23419417, QTL_02_24189963, QTL_05_19754561, QTL_09_16065720, and QTL_10_17905052) were identified in more than one FAA dataset by no less than three methods (Supplementary Figure 3). Thus, the present complementarity of different methods suggested that the combined utilization of various single-locus and multi-locus GWAS methods may facilitate the identification of all potential QTLs with large and small effects in a powerful and robust manner, and the 3VmrMLM method may be used as an alternative for other multi-locus methods.

Candidate genes for five FAA levels

A total of 88 stable QTLs were identified by no less than two methods. Genes co-localized in the 122-kb flanking region of stable QTL were identified for further analysis. Based on GO classification and KEGG pathway analysis, four potential candidate genes were found related to five FAA levels in rice, and the *Beta-glucosidase* gene (*LOC_Os01g19220*) involved in cyano amino acid metabolism (map00460) was a candidate gene of QTL_01_0944343 on chromosome 1, which was identified in both the Val and Ile datasets. According to KEGG pathway information, beta-glucosidase plays an important role in cyano amino acid metabolism, in which L-isoleucine and L-valine are required. The *Adenosylhomocysteine nucleosidase* gene (*LOC_Os01g12940*) associated with Leu content was identified in QTL_01_07089989 on chromosome 1 and involved in biosynthesis of amino acids (map01230) according to KEGG annotation. The *Isocitrate dehydrogenase* gene (*LOC_Os05g49760, IDH*) involved in glutathione metabolism (map00480) was detected in QTL_05_28394307 from the Arg dataset according to KEGG annotation. The *IDH* gene has been reported as a key enzyme in glutathione metabolism (Koh et al., 2004; Reitman et al., 2011; Tang et al., 2020). Glutathione is formed by the binding of γ-glutamate and cysteine *via* peptide bonds *via* the γ-glutamylcysteine synthetase (GSH1) and the binding of glycine catalyzed by glutathione synthetase (GSH2) (Noctor et al., 2012). As the essential precursor of glutathione, glutamate plays an important role in the biosynthetic and catabolism pathway of arginine. For instance, ornithine is synthesized from glutamate either in a cyclic or in a linear pathway and subsequently further converts to arginine; arginine catabolism begins with the degradation of arginine to ornithine, followed by the generation of glutamate through ornithine degradation (Winter et al., 2015; Majumdar et al., 2016). Genetic variation of *LOC_Os05g49760* resulted in the content alteration of Arg in this study (Figure 7A). The *Amidase* gene (*LOC_Os11g06900*) that participated in tryptophan metabolism (map00380) was a candidate gene of QTL_11_03441584 on chromosome 11, which was associated with Trp level in rice. In Arabidopsis, amidase catalyzes the conversion of indole-3-acetamide (IAM) to indole-3-acetic acid (IAA), which is an alternative terminal reaction step of IAA synthesis (Pollmann et al., 2009). IAA is the predominant auxin in plants, which can be synthesized from the Trp-dependent pathway. It has been confirmed that amidase promotes the synthesis of IAA, which is formed from tryptophan (Dharmasiri et al., 2005; Mockaitis and Estelle, 2008; Erland and Saxena, 2019). The natural variation of *LOC_Os11g06900* caused the content alteration of Trp in this study (Figure 7E). Moreover, *bZIP18*, *BCAT2*, and *BCAT4* genes have been validated to control the FAA levels in rice and other plant studies (Schuster et al., 2006; Angelovici et al., 2013; Sun et al., 2020). However, they were not found to be candidate genes of five FAA datasets in this study. Some transcript factors were co-localized with stable QTLs, which may contribute to the natural variation of FAA level in rice. Hence, the molecular mechanism of these candidate genes underlying the variation of FAA levels is warranted for further validation in the laboratory.

Candidate gene prediction based on detected QEI

Compared with the other eight methods, 3VmrMLM is able to detect the QEI of five FAA levels. Based on the 95 detected QEIs, their predicted candidate genes were subjected to further functional analysis (Supplementary Table 6). According to KEGG annotation, the candidate gene *LOC_Os12g16200* of QEI_12_09153839 (this QEI ID refers to QEI_Chromosome_Position) encoding glutathione synthetase was identified in glutathione metabolism (map00480) in the Arg dataset. Glutathione synthetase (GSH) is an important enzyme to catalyze the formation of glutathione *via* the binding of γ-glutamate and cysteine (Noctor et al., 2012). Glutamate not only is an essential precursor for glutathione synthesis, but also participates in the biosynthetic and catabolism pathway of arginine (Noctor et al., 2012; Winter et al., 2015). *LOC_Os09g07830* of QEI_09_03978551 encoding acetyl-CoA acetyltransferase was identified in the Leu dataset, which was involved in valine leucine and isoleucine degradation (map00280) according to KEGG

annotation. In the Trp dataset, *LOC_Os01g02020* gene harbored in QEI_01_00617184 encoding acetyl-CoA acetyltransferase was involved in tryptophan metabolism (map00380). These results suggested that a few QEIs may contribute to a small proportion of total variation on five FAA levels in rice.

Breeding applications of FAA-associated QTLs

Significant correlations between NPQTL and five FAA contents were observed (r = 0.53–0.69), which indicated the additive effect of these QTLs, especially for the Trp dataset (r = 0.69) (Figure 5). It was observed that the highest levels of Arg were present in some rice accessions carrying nine QTLs with positive-effect or favorite alleles (PQTLs), such as C063 and W088. In addition, the Trp levels in accessions with 18 PQTLs (C119, etc.) were higher than those with 19 PQTLs (C197) (Supplementary Table 7). These suggested that the accessions carrying these PQTLs hold the potential in FAA biofortified rice breeding through the pyramiding of loci. This strategy has been successful in the improvement of FHB resistance in wheat (Buerstmayr et al., 2008). In five FAA datasets, FAA content in *japonica* accessions was generally higher than that in *indica* accessions (Figures 1B–F; Supplementary Table 4). This suggested that *japonica* accessions have more breeding potential than *indica* accessions in terms of these five FAA levels. These *japonica* accessions are good parents for genetic improvement of high FAA level by directly hybridizing with elite varieties. The average R^2 value of QTL detected in all five FAA datasets by 3VmrMLM was lower than that by mrMLM (Table 2). QTLs with a small effect have been successfully applied in genomic selection (GS) breeding for the improvement of disease resistance and yield in crops (Crossa et al., 2017; Wang et al., 2018; Xu et al., 2021). Hence, these relatively small-effect QTLs detected by 3VmrMLM might be applicable for genomic selection breeding in rice with high FAA levels; in particular, the 3VmrMLM method is beneficial for the QTL detection of an association mapping population consisting of heterozygous individuals (Li et al., 2022a).

Conclusion

In this study, a total of 987 QTNs were detected in five FAA datasets by nine GWAS methods. The large number of detected QTNs demonstrated five FAA levels in rice were controlled by polygenes. 3VmrMLM has advantages in several aspects compared to other GWAS methods; 3VmrMLM detected the largest number of common QTNs, more significant on QTN detection, and relatively moderate R^2 values of QTLs were detected in multi-locus methods. The combined use of GWAS methods may facilitate the identification of all potential QTLs with large and small effects in a powerful and robust manner. Additionally, 15, 16, 20, 14, and 23 stable QTLs were detected in Val, Leu, Ile, Arg, and Trp datasets. Natural variations of the *LOC_Os01g19220* gene resulting in the content alteration of Val and Ile demonstrated that some potential candidate genes may play an important role in the crosslinking of different pathways. Of these QTLs, KEGG analysis of the candidate genes of five FAA-associated stable QTLs showed that they participated in biosynthesis of amino acids in five FAA datasets; glycine, serine, and threonine metabolism in the Leu dataset; and tryptophan metabolism in the Trp dataset. Moreover, 23, 16, 16, 18, and 22 QEIs were identified in the Val, Leu, Ile, Arg, and Trp datasets. KEGG pathway analysis showed that candidate genes were mainly involved in valine, leucine, and isoleucine degradation (QEI_09_03978551 and its candidate gene *LOC_Os09g07830* in the Leu dataset), tryptophan metabolism (QEI_01_00617184 and its candidate gene *LOC_Os01g02020* in the Trp dataset), and glutathione metabolism (QEI_12_09153839 and its candidate gene *LOC_Os12g16200* in the Arg dataset). To sum up, the combined utilization of 3VmrMLM with other GWAS methods will facilitate the mining of genes controlling complex traits and genomic selection breeding in rice.

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

LH conceived and designed this research project. YS, HW, and YM undertook the analysis of all available data. LH and HW contributed to resources and the writing of the original draft. JL and HL discussed the results, guided the entire study, participated in data analysis, and revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the Natural Science Foundation of Hainan Province (No. 321RC1148), the Key Research and Development Program of Hainan (No. ZDYF2020066), the "111" Project (No. D20024), the Hainan University Startup Fund KYQD (ZR) 1866 to J.L, and the Hainan University Startup Fund (RZ2100003217).

Acknowledgments

We appreciate Wei Chen and other authors in Chen et al. (2014) for their great contribution to rice metabolic research field and public accessed data availability for reusage in this study.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1048860/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Dot plots (lower triangle), histograms (diagonal) and Pearson correlations (upper triangle) between five FAAs datasets. Best curves are fitted in dot plots and histograms. *** indicates statistical significance at the 0.1% probability level probability level, and the size of the coefficient value is proportional to the strength of the correlation.

SUPPLEMENTARY FIGURE 2
Venn diagrams of unique QTNs detected by mrMLM series methods from Val **(A)**, Leu **(B)**, Ile **(C)**, Arg **(D)** and Trp **(E)**.

SUPPLEMENTARY FIGURE 3
Common QTNs detected in different FAA datasets by different methods. **(A)**: QTN-0110944343; **(B)**: QTN-0123419417; **(C)**: QTN-0224189963; **(D)**: QTN-0224189963; **(E)**: QTN-0519754561; **(F)**: QTN-0916065720; **(G)**: QTN-1017905052. The size of the circle is proportional to the significance level.

SUPPLEMENTARY FIGURE 4
GO classification of candidate genes harbored in stable QTLs in Val **(A)**, Leu **(B)**, Ile **(C)**, Arg **(D)**, Trp **(E)** datasets.

SUPPLEMENTARY FIGURE 5
KEGG pathway analysis of candidate genes harbored in stable QTLs in Val **(A)**, Leu **(B)**, Ile **(C)**, Arg **(D)** and Trp **(E)** datasets.

SUPPLEMENTARY FIGURE 6
Manhattan plots for five FAA levels detected QEIs by 3VmrMLM. QEIs in Val **(A)**, QEIs in Leu **(B)**, QEIs in Ile **(C)**, QEIs in Arg **(D)**, QEIs in Trp **(E)**. Black horizontal lines in the Manhattan plots represent the genome-wide significant threshold.

SUPPLEMENTARY FIGURE 7
GO classification of candidate genes harbored in QEIs in Val **(A)**, Leu **(B)**, Ile **(C)**, Arg **(D)** and Trp **(E)** datasets.

SUPPLEMENTARY FIGURE 8
KEGG pathway analysis of candidate genes harbored in QEIs in Val **(A)**, Leu **(B)**, Ile **(C)**, Arg **(D)** and Trp **(E)** datasets.

References

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi: 10.1101/gr.094052.109

Angelovici, R., Lipka, A. E., Deason, N., Gonzalez-Jorge, S., Lin, H., Cepela, J., et al. (2013). Genome-wide analysis of branched-chain amino acid levels in arabidopsis seeds. *Plant Cell.* 25, 4827–4843. doi: 10.1105/tpc.113.119370

Atwell, S., Huang, Y. S., Vilhjalmsson, B. J., Willems, G., Horton, M., Li, Y., et al. (2010). Genome-wide association study of 107 phenotypes in arabidopsis thaliana inbred lines. *Nature* 465, 627–631. doi: 10.1038/nature08800

Bausenwein, U., Millard, P., Thornton, B., and Raven, J. A. (2001). Seasonal nitrogen storage and remobilization in the forb rumex acetosa. *Funct. Ecol.* 15, 370–377. doi: 10.1046/j.1365-2435.2001.00524.x

Buerstmayr, H., Ban, T., and Anderson, J. (2008). QTL mapping and marker assisted selection for fusarium head blight resistance in wheat. *Cereal Res. Commun.* 36, 1–3. doi: 10.1556/CRC.36.2008.Suppl.B.1

Chang, F., Guo, C., Sun, F., Zhang, J., Wang, Z., Kong, J., et al. (2018). Genome-wide association studies for dynamic plant height and number of nodes on the main stem in summer sowing soybeans. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01184

Chan, E. K., Rowe, H. C., Hansen, B. G., and Kliebenstein, D. J. (2010). The complex genetic architecture of the metabolome. *PloS Genet.* 6, e1001198. doi: 10.1371/journal.pgen.1001198

Chen, W., Gao, Y., Xie, W., Gong, L., Lu, K., Wang, W., et al. (2014). Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nat. Genet.* 46, 714–721. doi: 10.1038/ng.3007

Chen, W., Gong, L., Guo, Z., Wang, W., Zhang, H., Liu, X., et al. (2013). A novel integrated method for large-scale detection, identification, and quantification of widely targeted metabolites: application in the study of rice metabolomics. *Mol. Plant* 6, 1769–1780. doi: 10.1093/mp/sst0

Crossa, J., Perez-Rodriguez, P., Cuevas, J., Montesinos-Lopez, O., Jarquin, 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

Cui, Y., Zhang, F., and Zhou, Y. (2018). The application of multi-locus GWAS for the detection of salt-tolerance loci in rice. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01464

Dharmasiri, N., Dharmasiri, S., and Estelle, M. (2005). The f-box protein TIR1 is an auxin receptor. *Nature* 435, 441–445. doi: 10.1038/nature03543

Diebold, R., Schuster, J., Daschner, K., and Binder, S. (2002). The branched-chain amino acid transaminase gene family in arabidopsis encodes plastid and mitochondrial proteins. *Plant Physiol.* 129, 540–550. doi: 10.1104/pp.001602

Erland, L. A. E., and Saxena, P. (2019). Auxin driven indoleamine biosynthesis and the role of tryptophan as an inductive signal in *Hypericum perforatum* (L.). *PloS One* 14. doi: 10.1371/journal.pone.0223878

Fagard, M., Launay, A., Clement, G., Courtial, J., Dellagi, A., Farjad, M., et al. (2014). Nitrogen metabolism meets phytopathology. *J. Exp. Bot.* 65, 5643–5656. doi: 10.1093/jxb/eru323

Fang, C., and Luo, J. (2019). Metabolic GWAS-based dissection of genetic bases underlying the diversity of plant metabolism. *Plant J.* 97, 91–100. doi: 10.1111/tpj.14097

Fang, C., Zhang, H., Wan, J., Wu, Y., Li, K., Jin, C., et al. (2016). Control of leaf senescence by an MeOH-jasmonates cascade that is epigenetically regulated by OsSRT1 in rice. *Mol. Plant* 9, 1366–1378. doi: 10.1016/j.molp.2016.07.007

Fernie, A. R., and Tohge, T. (2017). The genetics of plant metabolism. *Annu. Rev. Genet.* 51, 287–310. doi: 10.1146/annurev-genet-120116-024640

Galili, G., Amir, R., and Fernie, A. R. (2016). The regulation of essential amino acid synthesis and accumulation in plants. *Annu. Rev. Plant Biol.* 67, 153–178. doi: 10.1146/annurev-arplant-043015-112213

Galili, G., Avin-Wittenberg, T., Angelovici, R., and Fernie, A. R. (2014). The role of photosynthesis and amino acid metabolism in the energy status during seed development. *Front. Plant Sci.* 5. doi: 10.3389/fpls.2014.00447

Hao, C., Jiao, C., Hou, J., Li, T., Liu, H., Wang, Y., et al. (2020). Resequencing of 145 landmark cultivars reveals asymmetric Sub-genome selection and strong founder genotype effects on wheat breeding in China. *Mol. Plant* 13, 1733–1751. doi: 10.1016/j.molp.2020.09.001

Hausler, R. E., Ludewig, F., and Krueger, S. (2014). Amino acids–a life between metabolism and signaling. *Plant Sci.* 229, 225–237. doi: 10.1016/j.plantsci.2014.09.011

Hildebrandt, T. M., Nunes Nesi, A., Araujo, W. L., and Braun, H. P. (2015). Amino acid catabolism in plants. *Mol. Plant* 8, 1563–1579. doi: 10.1016/j.molp.2015.09.005

Hou, S., Zhu, G., Li, Y., Li, W., Fu, J., Niu, E., et al. (2018). Genome-wide association studies reveal genetic variation and candidate genes of drought stress related traits in cotton (*Gossypium hirsutum* l.). *Front. Plant Sci.* 9, 1276. doi: 10.3389/fpls.2018.01276

Jin, C., Sun, Y., Shi, Y., Zhang, Y., Chen, K., Li, Y., et al. (2019). Branched-chain amino acids regulate plant growth by affecting the homeostasis of mineral elements in rice. *Sci. China Life Sci.* 62, 1107–1110. doi: 10.1007/s11427-019-9552-8

Joseph, B., Corwin, J. A., Li, B., Atwell, S., and Kliebenstein, D. J. (2013). Cytoplasmic genetic variation and extensive cytonuclear interactions influence natural variation in the metabolome. *Elife* 2, e00776. doi: 10.7554/eLife.00776

Kim, M. S., Lozano, R., Kim, J. H., Bae, D. N., Kim, S. T., Park, J. H., et al. (2021). The patterns of deleterious mutations during the domestication of soybean. *Nat. Commun.* 12, 97. doi: 10.1038/s41467-020-20337-3

King, J. E., and Gifford, D. J. (1997). Amino acid utilization in seeds of loblolly pine during germination and early seedling growth (I. arginine and arginase activity). *Plant Physiol.* 113, 1125–1135. doi: 10.1104/pp.113.4.1125

Koh, H. J., Lee, S. M., Son, B. G., Lee, S. H., Ryoo, Z. Y., Chang, K. T., et al. (2004). Cytosolic $NADP^+$-dependent isocitrate dehydrogenase plays a key role in lipid metabolism. *J. Biol. Chem.* 279, 39968–39974. doi: 10.1074/jbc.M402260200

Kumar, S., Stecher, G., Peterson, D., and Tamura, K. (2012). MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. *Bioinformatics* 28, 2685–2686. doi: 10.1093/bioinformatics/bts507

Le Couteur, D. G., Solon-Biet, S. M., Cogger, V. C., Ribeiro, R., de Cabo, R., Raubenheimer, D., et al. (2020). Branched chain amino acids, aging and age-related health. *Ageing Res. Rev.* 64, 101198. doi: 10.1016/j.arr.2020.101198

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, J., Tang, W., Zhang, Y. W., Chen, K. N., Wang, C., Liu, Y., et al. (2018). Genome-wide association studies for five forage quality-related traits in sorghum (Sorghum bicolor l.). *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01146

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, C., Tu, Y., Liao, S., Fu, X., Lian, X., He, Y., et al. (2021). Genome-wide association study of flowering time reveals complex genetic heterogeneity and epistatic interactions in rice. *Gene* 770, 145353. doi: 10.1016/j.gene.2020.145353

Li, M., Zhang, Y. W., Xiang, Y., Liu, M. H., and Zhang, Y. M. (2022a). IIIVmrMLM: The r and c++ tools associated with 3VmrMLM, a comprehensive GWAS method for dissecting quantitative traits. *Mol. Plant* 15, 1251–1253. doi: 10.1016/j.molp.2022.06.002

Li, M., Zhang, Y. W., Zhang, Z. C., Xiang, Y., Liu, M. H., Zhou, Y. H., et al. (2022b). A compressed variance component mixed model for detecting QTNs and QTN-by-environment and QTN-by-QTN interactions in genome-wide association studies. *Mol. Plant* 15, 630–650. doi: 10.1016/j.molp.2022.02.012

Luo, J. (2015). Metabolite-based genome-wide association studies in plants. *Curr. Opin. Plant Biol.* 24, 31–38. doi: 10.1016/j.pbi.2015.01.006

Majumdar, R., Barchi, B., Turlapati, S. A., Gagne, M., Minocha, R., Long, S., et al. (2016). Glutamate, ornithine, arginine, proline, and polyamine metabolic interactions: The pathway is regulated at the post-transcriptional level. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.00078

Ma, L., Liu, M., Yan, Y., Qing, C., Zhang, X., Zhang, Y., et al. (2018). Genetic dissection of maize embryonic callus regenerative capacity using multi-locus genome-wide association studies. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00561

Misra, G., Badoni, S., Anacleto, R., Graner, A., Alexandrov, N., and Sreenivasulu, N. (2017). Whole genome sequencing-based association study to unravel genetic architecture of cooked grain width and length traits in rice. *Sci. Rep.* 7, 12478. doi: 10.1038/s41598-017-12778-6

Mockaitis, K., and Estelle, M. (2008). Auxin receptors and plant development: a new signaling paradigm. *Annu. Rev. Cell Dev. Biol.* 24, 55–80. doi: 10.1146/annurev.cellbio.23.090506.123214

Moe, L. A. (2013). Amino acids in the rhizosphere: from plants to microbes. *Am. J. Bot.* 100, 1692–1705. doi: 10.3732/ajb.1300033

Muller, C. L., Anacker, A. M. J., and Veenstra-VanderWeele, J. (2016). The serotonin system in autism spectrum disorder: From biomarker to animal models. *Neuroscience* 321, 24–41. doi: 10.1016/j.neuroscience.2015.11.010

Noctor, G., Mhamdi, A., Chaouch, S., Han, Y., Neukermans, J., Marquez-Garcia, B., et al. (2012). Glutathione in plants: an integrated overview. *Plant Cell Environ.* 35, 454–484. doi: 10.1111/j.1365-3040.2011.02400.x

Pathria, G., and Ronai, Z. A. (2021). Harnessing the Co-vulnerabilities of amino acid-restricted cancers. *Cell Metab.* 33, 9–20. doi: 10.1016/j.cmet.2020.12.009

Patil, M. D., Bhaumik, J., Babykutty, S., Banerjee, U. C., and Fukumura, D. (2016). Arginine dependence of tumor cells: targeting a chink in cancer's armor. *Oncogene* 35, 4957–4972. doi: 10.1038/onc.2016.37

Pollmann, S., Duchting, P., and Weiler, E. W. (2009). Tryptophan-dependent indole-3-acetic acid biosynthesis by 'IAA-synthase' proceeds *via* indole-3-acetamide. *Phytochemistry* 70, 523–531. doi: 10.1016/j.phytochem.2009.01.021

Pratelli, R., and Pilot, G. (2014). Regulation of amino acid metabolic enzymes and transporters in plants. *J. Exp. Bot.* 65, 5535–5556. doi: 10.1093/jxb/eru320

Reitman, Z. J., Jin, G., Karoly, E. D., Spasojevic, I., Yang, J., Kinzler, K. W., et al. (2011). Profiling the effects of isocitrate dehydrogenase 1 and 2 mutations on the cellular metabolome. *Proc. Natl. Acad. Sci. U. S. A.* 108, 3270–3275. doi: 10.1073/pnas.1019393108

Rennenberg, H., W ildhagen, H., and Ehlting, B. (2010). Nitrogen nutrition of poplar trees. *Plant Biol. (Stuttg.)* 12, 275–291. doi: 10.1111/j.1438-8677.2009.00309.x

Ren, W. L., Wen, Y. J., Dunwell, J. M., and Zhang, Y. M. (2017). pKWmEB: integration of kruskal-Wallis test with empirical bayes under polygenic background control for multi-locus genome-wide association study. *Heredity. (Edinb).* 120, 208–218. doi: 10.1038/s41437-017-0007-4

Rowe, H. C., Hansen, B. G., Halkier, B. A., and Kliebenstein, D. J. (2008). Biochemical networks and epistasis shape the *Arabidopsis thaliana* metabolome. *Plant Cell.* 20, 1199–1216. doi: 10.1105/tpc.108.058131

Schuster, J., Knill, T., Reichelt, M., Gershenzon, J., and Binder, S. (2006). Branched-chain aminotransferase4 is part of the chain elongation pathway in the biosynthesis of methionine-derived glucosinolates in *Arabidopsis*. *Plant Cell.* 18, 2664–2679. doi: 10.1105/tpc.105.039339

Segura, V., Vilhjalmsson, B. J., Platt, A., Korte, A., Seren, U., Long, Q., et al. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* 44, 825–830. doi: 10.1038/ng.2314

Sun, Y., Shi, Y., Liu, G., Yao, F., Zhang, Y., Yang, C., et al. (2020). Natural variation in the *OsbZIP18* promoter contributes to branched-chain amino acid levels in rice. *New Phytol.* 228, 1548–1558. doi: 10.1111/nph.16800

Tamba, C. L., Ni, Y. L., and Zhang, Y. M. (2017). Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. *PloS Comput. Biol.* 13, e1005357. doi: 10.1371/journal.pcbi.1005357

Tang, X., Fu, X., Liu, Y., Yu, D., Cai, S. J., and Yang, C. (2020). Blockade of glutathione metabolism in IDH1-mutated glioma. *Mol. Cancer Ther.* 19, 221–230. doi: 10.1158/1535-7163.MCT-19-0103

VanEtten, C. H., Wolff, I. A., Jones, Q., and Miller, R. W. (1963). Amino acid composition of seeds from 200 angiospermous plant species. *J. Agric. Food Chem.* 11, 399–410. doi: 10.1021/jf60129a016

Wang, S. B., Feng, J. Y., Ren, W. L., Huang, B., Zhou, L., Wen, Y. J., et al. (2016). Improving power and accuracy of genome-wide association studies *via* a multi-locus mixed linear model methodology. *Sci. Rep.* 6, 19444. doi: 10.1038/srep19444

Wang, X., Xu, Y., Hu, Z., and Xu, C. (2018). Genomic selection methods for crop improvement: Current status and prospects. *Crop J.* 6, 330–340. doi: 10.1016/j.cj.2018.03.001

Watanabe, M., Balazadeh, S., Tohge, T., Erban, A., Giavalisco, P., Kopka, J., et al. (2013). Comprehensive dissection of spatiotemporal metabolic shifts in primary, secondary, and lipid metabolism during developmental senescence in arabidopsis. *Plant Physiol.* 162, 1290–1310. doi: 10.1104/pp.113.217380

Wen, Y. J., Zhang, H., Ni, Y. L., Huang, B., Zhang, J., Feng, J. Y., et al. (2017). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief Bioinform.* 19, 700–712. doi: 10.1093/bib/bbw145

Winter, G., Todd, C. D., Trovato, M., Forlani, G., and Funck, D. (2015). Physiological implications of arginine metabolism in plants. *Front. Plant Sci.* 6. doi: 10.3389/fpls.2015.00534

Xu, Y., Ma, K., Zhao, Y., Wang, X., Zhou, K., Yu, G., et al. (2021). Genomic selection: A breakthrough technology in rice breeding. *Crop J.* 9, 669–677. doi: 10.1016/j.cj.2021.03.008

Xu, Y., Xu, C., and Xu, S. (2017). Prediction and association mapping of agronomic traits in maize using multiple omic data. *Heredity. (Edinb).* 119, 174–184. doi: 10.1038/hdy.2017.27

Yang, W., Guo, Z., Huang, C., Duan, L., Chen, G., Jiang, N., et al. (2014). Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice. *Nat. Commun.* 5, 5087. doi: 10.1038/ncomms6087

Yang, J., Zhou, Y., and Jiang, Y. (2022). Amino acids in rice grains and their regulation by polyamines and phytohormones. *Plants (Basel)* 11, 1581. doi: 10.3390/plants11121581

Yi, X., Du, Z., and Su, Z. (2013). PlantGSEA: a gene set enrichment analysis toolkit for plant community. *Nucleic Acids Res.* 41, W98–103. doi: 10.1093/nar/gkt281

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

Zeier, J. (2013). New insights into the regulation of plant immunity by amino acid metabolic pathways. *Plant Cell Environment.* 36, 2085–2103. doi: 10.1111/pce.12122

Zhang, C., Dong, S.-S., Xu, J.-Y., He, W.-M., Yang, T.-L., and Schwartz, R. (2019). PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* 35, 1786–1788. doi: 10.1093/bioinformatics/bty875

Zhang, J., Feng, J. Y., Ni, Y. L., Wen, Y. J., Niu, Y., Tamba, C. L., et al. (2017). pLARmEB: integration of least angle regression with empirical bayes for multilocus genome-wide association studies. *Heredity. (Edinb).* 118, 517–524. doi: 10.1038/hdy.2017.8

Zhang, Y. M., Jia, Z., and Dunwell, J. M. (2019). Editorial: The applications of new multi-locus GWAS methodologies in the genetic dissection of complex traits. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00100

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

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

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 14 November 2022
DOI 10.3389/fpls.2022.1033120

OPEN ACCESS

EDITED BY
Shang-Qian Xie,
University of Idaho, United States

REVIEWED BY
Liu Jinyang,
Jiangsu Academy of Agricultural Sciences (JAAS), China
Shibo Wang,
University of California, Riverside, United States

*CORRESPONDENCE
Yingpeng Han
hyp234286@aliyun.com
Lijuan Qiu
qiulijuan@caas.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

RECEIVED 31 August 2022
ACCEPTED 04 October 2022
PUBLISHED 14 November 2022

CITATION
Hong H, Li M, Chen Y, Wang H, Wang J, Guo B, Gao H, Ren H, Yuan M, Han Y and Qiu L (2022) Genome-wide association studies for soybean epicotyl length in two environments using 3VmrMLM. *Front. Plant Sci.* 13:1033120. doi: 10.3389/fpls.2022.1033120

Genome-wide association studies for soybean epicotyl length in two environments using 3VmrMLM

Huilong Hong[1,2†], Mei Li[3†], Yijie Chen[4†], Haorang Wang[5], Jun Wang[4], Bingfu Guo[6], Huawei Gao[2], Honglei Ren[7], Ming Yuan[8], Yingpeng Han[1*] and Lijuan Qiu[2*]

[1]Key Laboratory of Soybean Biology in Chinese Ministry of Education (key Laboratory of Soybean Biology and Breeding/Genetics of Chinese Agriculture Ministry), Northeast Agricultural University, Harbin, China, [2]Institute of Crop Science, National Key Facility for Crop Gene Resources and Genetic Improvement (NFCRI) Chinese Academy of Agricultural Sciences, Beijing, China, [3]Crop Information Center, College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, China, [4]College of Agriculture, Yangtze University, Jingzhou, China, [5]Jiangsu Xuhuai Regional Institute of Agricultural Sciences, Xuzhou, China, [6]Nanchang Branch of National Center of Oil crops Improvement, Jiangxi Province Key Laboratory of Oil crops Biology, Crops Research Institute of Jiangxi Academy of Agricultural Sciences, Nanchang, China, [7]Soybean Research Institute, Heilongjiang Academy of Agricultural Sciences, Harbin, China, [8]Qiqihar Branch of Heilongjiang Academy of Agricultural Sciences, Qiqihar, China

Germination of soybean seed is the imminent vital process after sowing. The status of plumular axis and radicle determine whether soybean seed can emerge normally. Epicotyl, an organ between cotyledons and first functional leaves, is essential for soybean seed germination, seedling growth and early morphogenesis. Epicotyl length (EL) is a quantitative trait controlled by multiple genes/QTLs. Here, the present study analyzes the phenotypic diversity and genetic basis of EL using 951 soybean improved cultivars and landraces from Asia, America, Europe and Africa. 3VmrMLM was used to analyze the associations between EL in 2016 and 2020 and 1,639,846 SNPs for the identification of QTNs and QTN-by-environment interactions (QEIs)".A total of 180 QTNs and QEIs associated with EL were detected. Among them, 74 QTNs (ELS_Q) and 16 QEIs (ELS_QE) were identified to be associated with ELS (epicotyl length of single plant emergence), and 60 QTNs (ELT_Q) and 30 QEIs (ELT_QE) were identified to be associated with ELT (epicotyl length of three seedlings). Based on transcript abundance analysis, GO (Gene Ontology) enrichment and haplotype analysis, ten candidate genes were predicted within nine genic SNPs located in introns, upstream or downstream, which were supposed to be directly or indirectly involved in the process of seed germination and seedling development., Of 10 candidate genes, two of them (Glyma.04G122400 and Glyma.18G183600) could possibly affect epicotyl length elongation. These results indicate the genetic basis of EL and provides a valuable basis for specific functional studies of epicotyl traits.

KEYWORDS

genome-wide association analysis, single nucleotide polymorphism, candidate genes, 3VmrMLM, epicotyl length

Introduction

Epicotyl length (EL), an important complicated and agronomically trait, was significantly related to plant density and sowing depth of soybean (Camargos et al., 2019). EL exhibited the higher genetic variability at the early developmental stages of soybean, especially at V_2 and V_3 development stages (Matsuo et al., 2012). EL also affected plant height and yield of soybean (Hanyu et al., 2020). As a typical quantitative trait, EL, with relatively high heritability (more than 95%), was controlled by a few large-effect genes and a series of polygenes (Chaves et al., 2017). EL was significantly affected by environment, genotype their interactions (Chaves et al., 2017; Hanyu et al., 2020). Several studies showed that genetic and environmental variation approximately accounted for half of experimental observation. Although EL has been considered as the important feature of variety during the long-term soybean breeding, development of soybean cultivar with reasonable and stable EL through traditional selection method was still difficult (Chaves et al., 2017). It required evaluation in multiple environments over several years, and traditional selection method was expensive, time-consuming and labor-intensive (Chaves et al., 2017).

Molecular marker could effectively improve traditional selection efficiency by increasing the allele's frequency of desirable quantitative trait loci (QTLs). Presently, linkage analysis and association analysis, were two major strategies utilized to identify QTLs of important traits in crops (Li et al., 2020; Liu et al., 2020; Wang et al., 2021). Segregating population based linkage analysis strategy is a well-known approach to obtain QTLs, followed by fine mapping using larger secondary population or other types of population with sufficient map resolution, then candidate genes could be cloned for functional characterization. (Dinka et al., 2007) mapped four additive QTLs for the length of hypocotyl in soybean. However, none of EL QTLs of soybean has been reported to date. Based on diversegermplasms, Genome-Wide Association Study (GWAS) take advantages of historical recombination events offered another strategy to effectively fine map QTL with rapid decay of linkage disequilibrium (LD) (Flint-Garcia et al., 2003). Due to the advances in next-generation sequencing (NGS) technologies or Chip with high-density SNPs, GWAS has been widely extensively utilized to dissect genetic architecture of important traits in crops including soybean, e.g. biotic stress (Zhao et al., 2015; Zhao et al., 2017), abiotic stress (Zhang et al., 2015; Jia et al., 2017), yield-related trait including seed weight (Yan et al., 2017), maturity time (Contreras-Soto et al., 2017), and seed composition including seed oil content (Cao et al., 2017; Li et al., 2018), seed protein content (Zhang et al., 2019), tocopherol (Sui et al., 2020) and isoflavone concentration (Wu et al., 2020). Liang et al. (2014) identified four additive QTLs for the length of hypocotyl in soybean using linkage analysis. However, no EL QTLs in soybean has been reported to date.

Since the establishment of mixed linear model (MLM) method in genome-wide association studies (GWAS) (Zhang et al., 2005; Yu et al., 2006; Kang et al., 2008), these methods have proven to be useful in controlling for population structure and relatedness of individuals. However, these methods are computationally challenging for large datasets. Thus, a series of fast MLM-based algorithms have been developed and widely-used, such as CMLM (Zhang et al., 2010), EMMAX (Kang et al., 2010), FaST-LMM (Lippert et al., 2011), and GEMMA (Zhou and Stephens, 2012). In these methods, single marker genome scanning was used to identify significant QTNs. This is involved in multiple tests. To control false positive rate, Bonferroni correction is frequently adopted. The stringent significant criterion frequently results in the missing of some important loci, especially in crop GWAS. To overcome this issue, several multi-locus mixed model methods have been proposed and widely used (Segura et al., 2012; Wang et al., 2016; Wen et al., 2017). As we know, there are frequently three genotypes for each marker in GWAS. Two effects should be estimated, while their polygene backgrounds should be controlled. In most GWAS methods, however, only one confound effect is estimated, while its polygene background is controlled. To solve this issue, recently, Li et al. (2022b) established a three-variance-component mixed linear model framework, 3VmrMLM, to identify QTNs, QTN-by-environment interactions (QEIs), and QTN-by-QTN interactions under controlling all the possibly polygene backgrounds.

Cytokinins and light can sometimes elicit similar morphological and biochemical responses. In the absence of light plant seedlings have long epi- or hypocotyls and appressed leaves with the plastid development blocked at the stage of etioplasts or amyloplasts. The l6 ight-i6 ndependent p6 hotomorphogenesis (lip1) mutant of pea shows many of the characteristics normally associated with light-grown seedlings when grown in complete darkness, such as expanded leaves, a short epicotyl and partially developed chloroplast (Frances et al., 1992). Chory et al. the effects of cytokinin treatment on epicotyl growth inhibition of lip1 i n darkness are comparable to a hypocotyl growth inhibition observed in Arabidopsis(Chory et al., 1994), It appears that the effect of cytokinin on the growth of the axis of young hypogeal (e.g., Arabidopsis) and epigeal (e.g., pea) seedlings is similar. The phenotype of wild-type Arabidopsis plants following cytokinin treatment is similar to that of the amp1 mutant of Arabidopsis, suggesting that light and cytokinin act through a common signaling pathway (Chory et al., 1994; Seyedi et al., 2001). genetic analysis of Arabidopsis has provided unequivocal evidence that the brassinosteroids (BRs) are essential phytohormones (He et al., 2003). Brassinolide (BL), an end product of campesterol oxidationis is required for the regulation of cell elongation, stress response, male fertility, pigment biosynthesis, and numerous other developmental and physiological responses in higher plant (Grove et al., 1979), The Arabidopsis CYP90A1 (constitutive

photomorphogenesis and dwarfism, CPD) has been identified to functions as the C-23 hydroxylase in the biosynthetic pathway of brassinosteroids, and cpd mutant exhibited the most pronounced effect in dwarf phenotype than another five cytochrome P450 mutants. The biosynthetic model of BRs has been clearly identified in *Arabidopsis*, we supposed a similar model, It has been proved in 1998 that the transcription of *Arabidopsis* CYP90A1 was negatively controlled by exogenous brassinolide (Mathur et al., 1998).

To address above mentioned issues, 951 landraces and cultivarsselected from Chinese primary core collection in the Chinese National Soybean GeneBank (CNSGB), were phenotyped for EL in 2016 and 2020, and genotyped by 1,639,846 SNPs in order to identify QTNs, QEIs, and their candidate genes for EL in soybean.

Materials and method

Plant materials, filed trials and epicotyl length evaluations

To construct a diversity panel of EL, a total of 951 landraces was selected from more than 20,000 samples, which delegated much of the representatives of diversity of the collection at the Chinese National Soybean GeneBank (CNSGB). These tested materials were planted with the single row plots (3-m long and 0.35-m between rows), which was performed with the completely randomized design and three replications in Sanya, Hainan China in 2016 and 2020.

A total of 3 randomly selected plants from each plot were phenotyped for EL by measuring the distance between the cotilenodary knot and the unifoliate leaves pair knot using vernier caliper.

DNA isolation and genome sequencing

The genomic DNA of each tested samples were isolated from fresh leaves of a single plant, and then resequenced. Sequencing libraries were constructed based on TruseqNano® DNA HT sample preparation Kit (Illumina USA), and index codes were added to attribute sequences to each accession according to the method described by (Li et al., 2020a). The Illumina Hiseq X platform was used to analyze the libraries of these samples. A total of 10.58 Tb raw sequences with 150-bp read length, were obtained. After sequence quality filtering, the clean read of all tested samples, were aligned to soybean reference genome *via* Short Oligonucleotide Alignment Program 2 (SOAP2) software. The SNPs were calling based on MAF ≥ 0.05. The genotype was regarded as heterozygous if the depth of minor allele/the total depth of the sample was more than 1/3.

Population structure evaluation and linkage disequilibrium (LD) analysis

The population structure of GWAS panel were evaluated based on principle component analysis (PCA) programs of Software package GAPIT (Lipka et al., 2012). LD was called with SNP (MAF ≥ 0.04 and missing data ≦ 10%) based on TASSEL version 3.0 (Bradbury et al., 2007).

Association analysis of epicotyl length of soybean

A total of 1,639,846 SNPs from 951 landraces samples were utilized to detect association signals of EL in soybean. Imputed genotype of total sample panel was first transformed in to *.fam, *.bed, and *.bim format, ELS and ELT in two different environments were adopted as phenotype, evolutionary population structure encoded as B (Landrace) and C (Improved cultivar), and kinship were employed as covariates for multi-environment joint analysis with significant level of 0.01 using IIIVmrMLM software of Li et al. (2022b); Li et al. (2022c). Linkage disequilibrium (LD) of 250kb up- and down-stream of significantly associated SNP were calculated by PLINK1.9, and threshold of regional average LD > 0.9 was used to define credible associated region. Functional annotation of candidate genes was performed based on annotation by phytozome (https://phytozome-next.jgi.doe.gov/info/Gmax_Wm82_a2_v1).

Definition and verification of candidate genes

Then SNP variations in the coding region of candidate genes were analyzed to screen candidate genes with mutation type of nonsynonymous, stoploss, stopgain, or alternative splicing. To further screen candidate genes, fixation index (F_{ST}) was calculated by published genome sequences data of 2214 soybeans (Li et al., 2022d) using vcftools (0.1.13) with window size of 100bp, and coding regions with $F_{ST} \geq 0.6$ were regarded as potential domestication gene (Song et al., 2013). Subsequently, spatial and temporal expression of candidates were analyzed using publicly available soybean transcriptome integration dataset (Yu et al., 2022). Functional annotations of all candidate genes were performed based on the SoyBase database (http://www.soybase.org) and the Kyoto Encyclopedia of gene and genomes (KEGG).

Haplotype analysis

Gene region were defined using *.gff, regional genotype of hapmap diploid were extracted from imputed genotype,

then haplotypes were inferred based on regional genotype classified according to its location relative to the gene structure. Significance of traits between different haplotypes were performed by Kruskal-Wallis ($P<0.01$) (Theodorsson, 1986). Haplotype TCS network was inferred using PopART (Bandelt et al., 1999; Clement et al., 2002; French et al., 2014). Geographic mapping of different haplotypes was performed using R scripts.

Results

Distribution of the landraces used in the experiment

Globally, the improved cultivars selected for the experiment mainly comes from America and Asia, with few from Europe and Africa. Landraces were all obtained from Asia (Figure 1). To better understand the genetic architecture of these germplasms, geographical distribution and ecological types were taken into account for classification. Both domestic and foreign varieties can be divided into southern (SR), northern (NR) and central (HR) varieties, namely domestic varieties (SR, HR, NR) and foreign varieties (WDD_SR, WDD_HR, WDD_NR). Domestic NR sources are the maximum, and foreign WDD_HR varieties account for more than half of the total foreign varieties (Figure 2A and Table S1). According to ecological types, domestic cultivars can be divided into northeast spring type (NESp), northern spring type (NSp), Huang-huai spring type (HSp), Huang-huai summer type (HSu), Southern spring type (SSp), Southern summer type (SSu) and Southern autumn type (SAu), with NESp ranking the first place. The selected foreign varieties were mainly divided into spring type (WDD_Sp) and summer type (WDD_Su), and the quantity of WDD_Sp was twice as much as WDD_Su (Figure 2B and Table S2). These results demonstrated that nearly 80% of the varieties used in the experiment came from China, and 60% of the varieties obtained abroad were spring varieties in the central region.

Statistical analysis for inflorescence length of the association panel

The EL of 951 landraces in Sanya, Hainan China in 2016 and 2020, were evaluated, respectively. The skewness and kurtosis of EL the three environments were less than ±1, which exhibited a continuous variation and the near normal distribution (Table S3). Therefore, EL of the association panel in this study, were appropriate.

Distribution of SNPs and analysis of mapping population

Based with the frequency > 0.05 as the minor allele and the missing data less than 0.03, a total of 1,639,846 single nucleotide polymorphisms (SNPs) were unevenly distributedon 20 chromosomes of soybean genome. with a density of578.8 bp per SNP on average, and varied from 337.3bp~1334.4bp per SNP. In detail, there were 168,498 SNPs on Chr1 with the highest density (337.3bp/SNP), 31,650 SNPs on Chr5 with lowest density (1334.4bp/SNP). (Figure 3). Based on these SNPs, principal component analysis and phylogenetic analysis were performed on the association panel. The results showed that the first PCs explained 24.52% of the genetic variation, the 951 varieties were divided into two categories with apparent discrepancy of genetic relatedness (Figure 4). For a preferably clearer study of epicotyl traits, they were also divided into two categories, ELS and ELT. Statistical methods were used to test that ELS and ELT showed normal distribution in different environments among varieties (Figure 5).

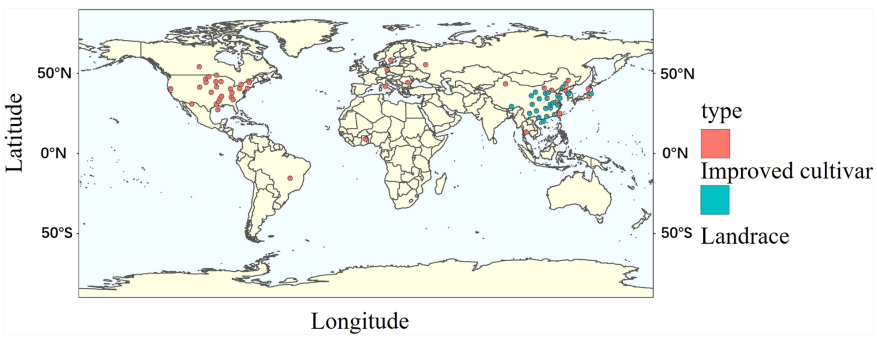


FIGURE 1
The geographical distribution of the tested accessions.

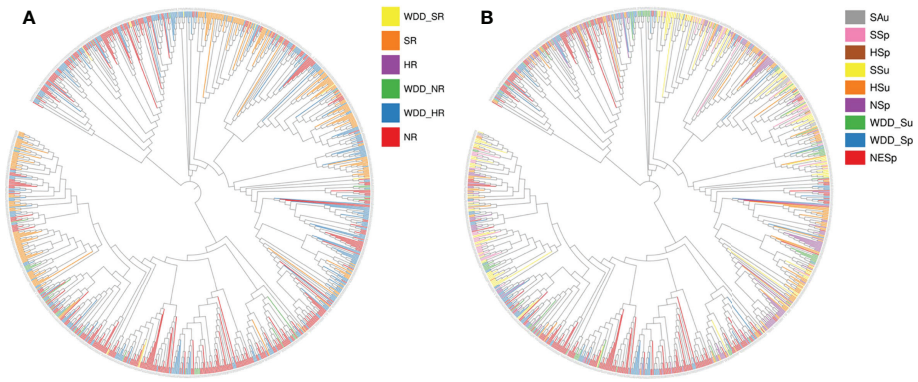


FIGURE 2
951 species construct phylogenetic tree according to geographical distribution and ecological type. **(A)** Variety Geographical Distribution Evolutionary Tree. WDD_: Oversea_; NR: Northern Region; HR: Central Region; SR: Southern Region **(B)** Variety Ecotype Evolutionary Tree. SAu, Southern autumn soybean; SSp, Southern spring soybean; HSp, Huanghuai summer soybean; SSu, Southern summer soybean; HSu, Huanghuai summer soybean; NSp, Northern spring soybean; NESp, Northeast Spring Soybeans; WDD_Su, Oversea summer soybean; WDD_Sp, Oversea spring soybean.

Quantitative trait nucleotide associated with epicotyl length-related traits by GWAS

QTN (Q) and QTN-by-environment interaction (QEI) detection method in the 3VmrMLM was used to analyze SNP-trait associations in two EL two-environment datasets, ELS (2016 and 2020) and ELT (2016 and 2020). A total of 180 QTNs and QEIs associated with epicotyl length were detected. Among them, 74 QTNs (ELS_Q) and 16 QEIs (ELS_QE) were identified to be associated with ELS, and 60 QTNs (ELT_Q) and 30 QEIs (ELT_QE) were identified to be associated with ELT. Figure 6 Of these, three sites (Gm_09_28400545, Gm_11_31100989, Gm_19_557643) could be found in all these four result datasets (Table S4).

Prediction of candidate genes for epicotyl length traits

We performed candidate gene prediction analyses with peak SNP of ±100 kb based on the physical locations of 180 SNPs associated with epicotyl length. A total of 1945 genes were included in these regions (Table S4). Functional annotation of

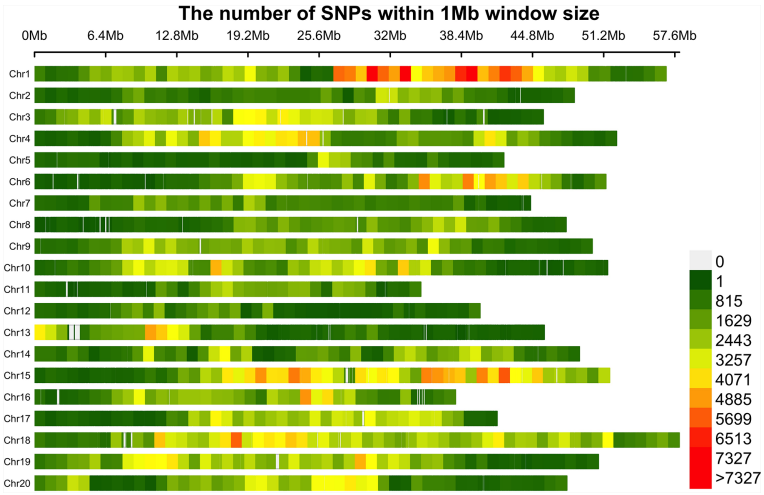


FIGURE 3
Distribution of SNP markers among 20 chromosomes.

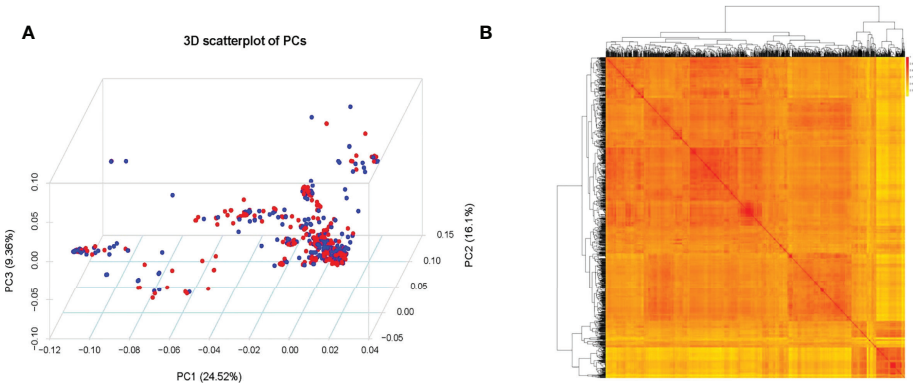


FIGURE 4
(A) Population structure of soybean germplasm. **(B)** Heatmap of the kinship matrix of the 951 soybean accessions.

1945 genes were completed by using *Arabidopsis* annotation information. site contribution rate, Transcription abundance of candidate genes in epicotyl of two representative soybean germplasms including cultivar Williams 82 with a long epicotyl of 3.93 cm and cultivar Jack with a short epicotyl of 2.13 cm were analyzed using publicly available soybean transcriptome integration dataset (Yu et al., 2022). By comparing the epicotyl lengths of Williams 82 and Jack, a very significant difference was found (Figures 7A, B). Based on the transcriptome data of epicotyls from Williams 82 and Jack, 585 out of 1945 genes were not expressed in both epicotyls of Williams 82 and Jack, 94 genes were expressed only in the epicotyl of Jack and 60 genes were expressed only in the epicotyl of Williams 82. A total of 1206 genes were expressed in both epicotyls of Williams 82 and Jack, of them, 157 genes were significantly differentially expressed in Williams 82 and Jack. Combined with Arabidopsis annotation information, 103 genes were identified as potentially candidate genes for epicotyl length (Table S5, Figure 7C). These differentially expressed genes in long and short epicotyl cultivars might be related to the length of epicotyl of soybean.

To further elucidate whether the differentially expressed genes were related to the length of the epicotyl, GO enrichment analysis was performed (http://amigo.geneontology.org/). GO enrichment analysis showed all genes were assigned to one of three GO categories: biological process (BP), molecular function (MF), and Cellular component (CC) (Figure 8).

Further, haplotype analysis was performed for 103 potentially candidate genes screened by the above analysis. epicotyl

In order to determine the role of the selected potential genes in soybean epicotyl growth, 22 potential candidates were screened by combining gene GO annotation and transcriptome differential expression analysis, and referring to Arabidopsis annotation information. Haplotype analysis identified 10 significantly different genesepicotyl. The Hap1 and Hap2 of Glyma.01G005900 in different years of ELS(P=0.0039) and ELT (P=0.039)showed

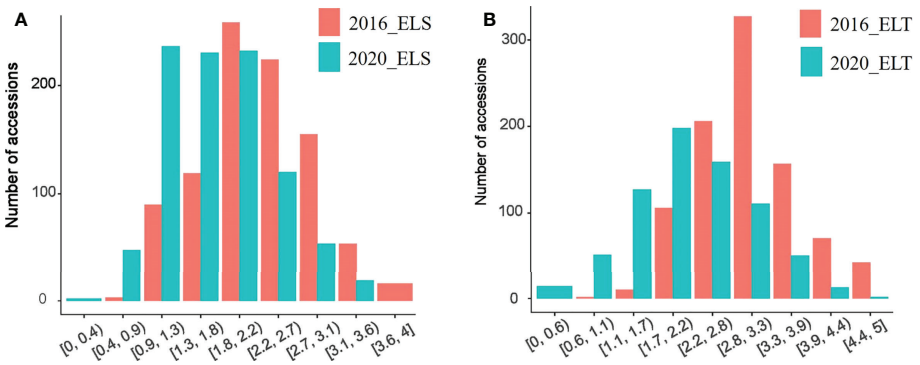


FIGURE 5
ELS and ELT phenotype distriution. **(A)** ELS phenotypes at different ages **(B)** ELT phenotypes at different ages.

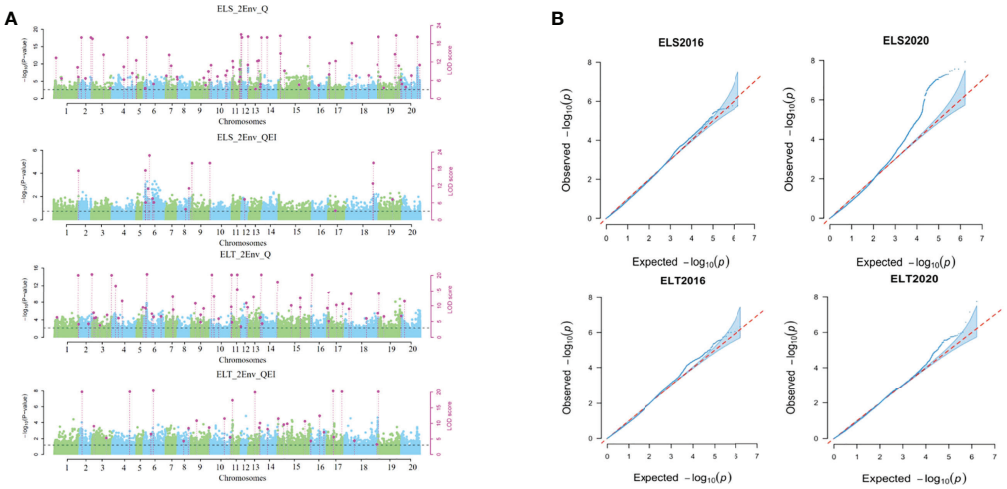


FIGURE 6
Results of association mapping of soybean epicotyl length traits. **(A)** Manhattan plot of locus distribution; **(B)** phenotype fitting results.

extremely significant differences (P<0.01). The Hap1 and Hap3 of Glyma.18G183600(2016_ELS *P*=1.1e-09; 2020_ELS *P*=0.00013; 2016_ELT *P*=3.4e-06; 2020_ELT *P*=0.69), Glyma.18G185300(2016_ELS *P*=0.0083; 2020_ELS *P*=1.2e-08; 2016_ELT *P*=0.02; 2020_ELT *P*=0.0031), exhibited extremely significant differences (*P*<0.01), while the Hap1 and Hap3 of Glyma.01G050100(2016_ELS *P*=4.4e-05; 2020_ELS *P*=0.0021), Glyma.04G122400(2016_ELS *P*=1.6e-08; 2020_ELS *P*=0.0006), Glyma.18G183600(2016_ELS *P*=1.1e-09; 2020_ELS *P*=0.00013) in different years of ELS had a very significant difference in 2016 (*P*<0.01), but there was no significant difference in 2020. The candidate gene *Glyma.18G185300* showed a very significant difference in the two years of EL (*P*<0.01), and the ELT revealed a significant difference in 2016(2016_ELT *P*=0.02) and showed a very significant difference in 2020(2020_ELT *P*=0.0031) (Figure 9). Meanwhile, we counted the variation sites of 10 gene haplotypes (Table S7). The results demonstrated that *Glyma.04G122400*, *Glyma.10G031900* and *Glyma.18G183600* exist in exon variation sites, of which *Glyma.04G122400* and *Glyma.18G183600* exist non-synonymous mutations, hence, we speculate that *Glyma.04G122400* and *Glyma.18G183600* are candidate genes for epicotyl differences. At the same time, we combed the geographical origin of the two gene haplotypes and the distribution of variety characteristics. From the geographical distribution, we could see that Hap1, Hap2, Hap3 and Hap4 haplotypes of the two candidate genes were absolutely dominant in the selected varieties. In terms of ecological characteristics of cultivars, Hap1 and Hap2 haplotypes of the two genes accounted for more than Landrace haplotypes in improved cultivars (Figure 10).

We predicted ten plant growth-related genes, namely *Glyma.03G142200* (Ribosomal protein S10p/S20e family protein), *Glyma.04G122400* (DCD domain protein), *Glyma.04G145000* (nuclear factor Y, subunit B13), *Glyma.10G0319000* (indole-3-acetic acid 7), *Glyma.10G056000* (SAUR-like auxin-responsive protein family), *Glyma.13G270800* (ubiquitin-conjugating enzyme 35), *Glyma.17G005900* (Pollen Ole e 1 allergen and extensin family protein), *Glyma.17G18500* (NAC domain containing protein 83), *Glyma.18G183600* (far-red elongated hypocotyl 1), and *Glyma.18G255300* (thioredoxin H-type 5). These results suggest that soybean epicotyl length may be regulated by multiple signaling pathways (Table 1).Additionally, none of these 10 cadidates were identified to be differentiated among wild soybean, landrace and improved cultivar (Figure S1).

Discussion

As an important feature of soybean variety, many studies indicated that EL affected 43.12% of seeds germination and 57.12% of seedlings emergence for soybean (Hanyu et al., 2020) estimated the genotypic determination coefficient of EL was more than 80% regardless of the evaluation period. (Matsuo et al., 2012) also obtained similar results. The genotypic determination coefficient was significantly related to inheritability, thus, it made the inference about genotypes possible (Vasconcelos et al., 2012; Hanyu et al., 2020). Through screening a large enough and reasonable gene database from more than 20,000 varieties, the SNPs and potential genes related to epicotyl traits were analyzed by GWAS technology. By elucidating the epicotyl related loci, it has a potential role in the study of

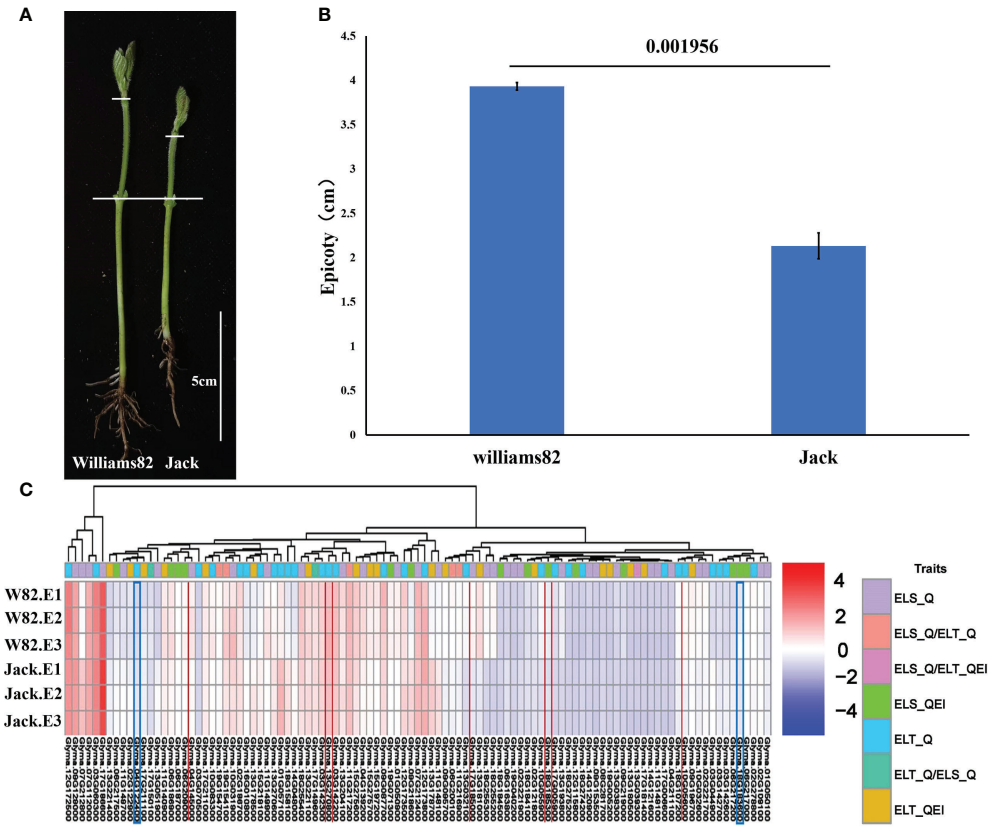


FIGURE 7
Epicotyl length of Williams82 and Jack and expression analysis of 103 candidate genes. **(A)** Epicotyl phenotype of W82 and Jack **(B)** Epicotyl Length Analysis of W82 and Jack **(C)** Transcriptome alignment of 103 candidate genes.

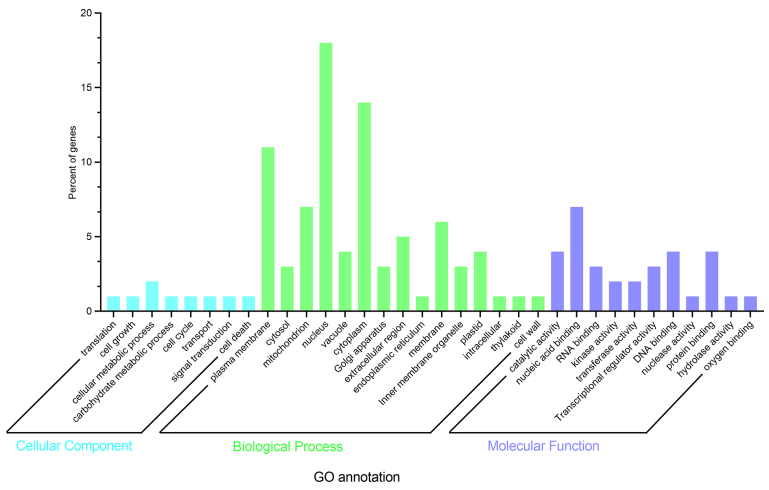


FIGURE 8
Functional categories of the genes in 100kb flanking regions around peak SNPs.

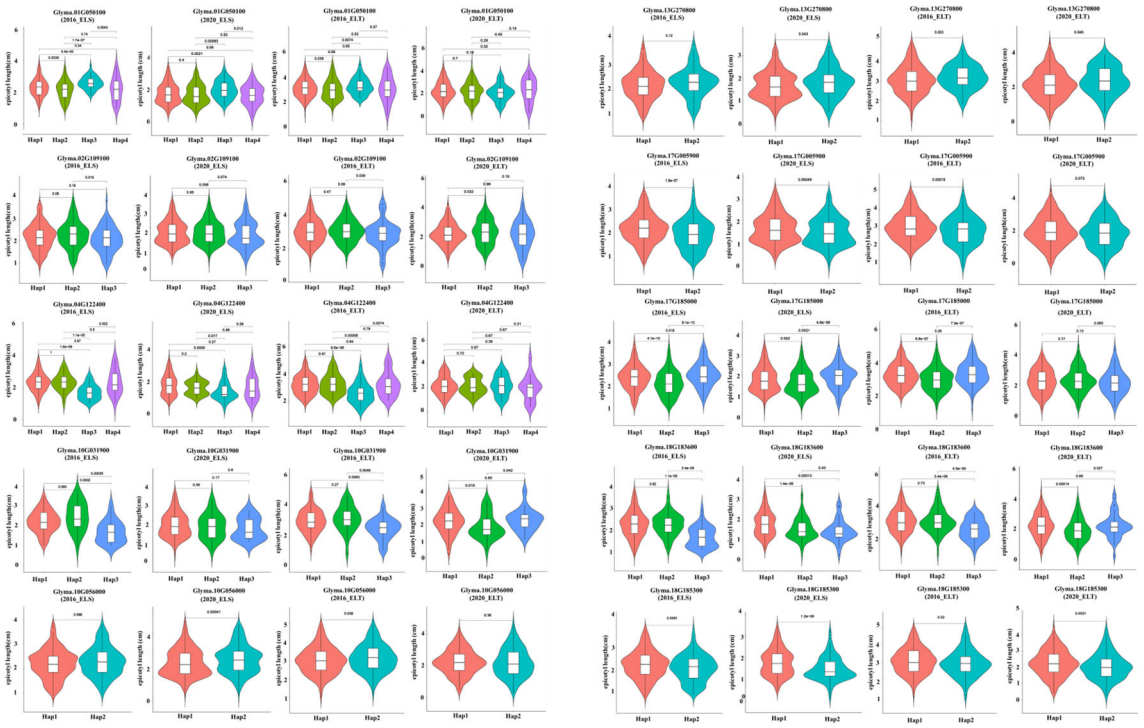


FIGURE 9
Genotyping of potential gene.

early seed germination, seedling germination and stem strength of soybean.

To date, many seedling crop traits have been studied and elucidated, but epicotyl traits have been largely ignored and poorly studied. Four of Chr.2, Chr.4, Chr.7 and Chr.10 were identified in the F2 population of adzuki bean "Tokei1121" (T1121, long epimorph) and cultivar "Erimo167" (ordinary ectomorph) with EL associated SNP) (Mori et al., 2021). There are no reports on EL-related SNP sites in other plants. The genetic mechanism of the hypocotyl length trait (HL) has been extensively studied. SNP mapping of soybean root-related traits at seedling stage revealed that HL is regulated by multiple additive genes. Seven QTLs in HL associated with seedling photomorphology were identified by using recombinant inbred (RIL) populations obtained from biparental crosses between Patagonia (Pat) and Colombia (COL0) (Matsusaka et al., 2021). Compound spacer and epitaxial array localization methods were also used to identify HL loci associated with light-responsive quantitative traits (Wolyn et al., 2004). To pinpoint trait-associated loci, the combination of GWAS and

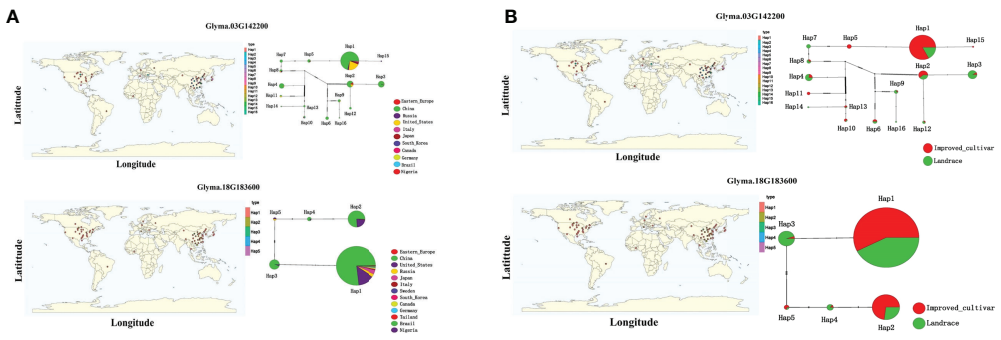


FIGURE 10
Haplotype analysis of candidate genes **(A)** distribution of geographical origin **(B)** distribution of cultivar characteristics.

TABLE 1 Gene based association of candidate genes.

Chr.	Physical position (bp)	Gene model	Trait	R^2 (contribution rate)	Pvalue	Functional annotation
3	35863419	Glyma.03G142200	ELT_Q	0.5768	6.09167E-21	Ribosomal protein S10p/S20e family protein
4	15439303	Glyma.04G122400	ELT_Q	0.4336	8.01377E-07	DCD (Development and Cell Death) domain protein
4	26351924	Glyma.04G145000	ELS_Q	0.2279	4.20387E-22	nuclear factor Y, subunit B13
10	2738580	Glyma.10G031900	ELS_Q	0.5234	1.14306E-11	indole-3-acetic acid 7
10	5143580	Glyma.10G056000	ELT_Q	0.5294	5.84009E-32	SAUR-like auxin-responsive protein family
13	37284883	Glyma.13G270800	ELT_Q	1.5012	7.02497E-35	ubiquitin-conjugating enzyme 35
17	637613	Glyma.17G005900	ELT_Q	0.5942	5.10164E-10	Pollen Ole e 1 allergen and extensin family protein
17	23689587	Glyma.17G185000	ELS_Q	0.7863	4.96905E-13	NAC domain containing protein 83
18	44381201	Glyma.18G183600	ELS_QEI	2.1064	1.12718E-32	far-red elongated hypocotyl 1
18	44381201	Glyma.18G185300	ELS_QEI	2.1064	1.12718E-32	one helix protein

transcriptome can be used to identify major genes affecting HL (Luo et al., 2017). These studies suggest that hypocotyl play a role in root growth and photomorphological responses. (Huang et al., 2006) studied the regulatory effect of brassinolide on epicotyl under low temperature conditions by proteomics. How xylan content in the gravitational bending direction of the epicotyl of adzuki bean affects its internal xylan content (Ikushima et al., 2008). Inhibitory effect of red light of the active form of phytochrome (Pfr) on epicotyl elongation in pea seedlings (Okoloko et al., 1970). These indicate that epicotyl play a non-negligible role in a variety of crops, especially dicotyledonous crops. Faced with this situation, this study used the soybean EL association panel to analyze the natural variation of epicotyl length and the related genetic structure, and analyzed the Hypothetically revealing a set of candidate genes controlling epicotyl development by GWAS analysis is undoubtedly a key step in filling in the relevant loci for epicotyl trait mapping.

Putative genes involved in epicotyl length

Through the Arabidopsis annotation information, candidate gene phenotype contribution rate, and combining with Yu et al. (2022) Williams 82 and Jack transcriptome results of extremely different genes, we screened 22 potential genes from 103 hypothetical genes. These genes are located in SNP peak within 100Kb.10 significantly different candidate genes were identified by haplotype analysis, these genes were genotyped significantly and distinctly of ELS and ELT. *Glyma.03G142200* is a Ribosomal protein S10p/S20e family protein, proteins involved in photosynthesis (Bah et al., 2010). Wycoff found that a lectin protein, analogous to ribosomal proteins, is detected in roots, hypocotyls and leaves and involved in soybean nodule formation (Wycoff et al., 1997).

Glyma.04G122400 DCD (Development and Cell Death) domain protein, thought to be involved in the hypersensitive response and programmed (Ludwig and Tenhaken, 2001, Enhaken et al., 2005), In previous studies, DCD domain proteins was believed to be involved in extracellular matrix or cytoskeleton proteins involved in growth and differentiation processes (Ichinose et al., 1990, Massimiliano et al., 2007).

Glyma.04G145000 nuclear factor Y, subunit B13, Nuclear factor Y is one of the largest transcription factor gene families in plants, The NUCLEAR FACTOR Y (NF-Y) transcription factors are heterotrimeric complexes composed of NF-YA and histone-fold domain (HFD) containing NF-YB/NF-YC (Siriwardana et al., 2016), NF-Y subunits are emerging as transcriptional regulators with essential roles in diverse plant processes (Zanetti et al., 2010). playing key roles in development and in response to adverse environmental conditions (Nelson et al., 2007; Li et al., 2008)AtNF-YB6 (L1L) and AtNF-YB9 (LEC1) are involved in embryo development in seeds (Yamamoto et al., 2009). Overexpression of PdNF-YB7 in Arabidopsis exhibited earlier seedling establishment, longer primary roots, larger leaf areas, and increased photosynthetic rate that conferred drought tolerance and improved WUE in transgenic plants. In Arabidopsis, AtNF-YB3 plays an important role in the promotion of flowering specifically under inductive long-day photoperiodic conditions. Consistent with this, the overexpression of PdNF-YB7 in Arabidopsis caused earlier seedling germination time and enhanced the development of both vegetative and reproductive organs (Xiao et al., 2013), also found that overexpression of AtNF-YB2 enhanced primary root elongation due to a faster cell division and/or elongation(Ballif et al., 2011)

The soybean epicotyl is the basis for the formation of true leaves after seed germination, which ensures the normal development of seedlings, and the synthesis of related hormones is also important. The Glyma.10G056000 and Glyma.17G005900 encoding SAUR-like auxin-responsive protein and allergen and elongation protein, respectively, are annotated through multiple omics networks in the Arabidopsis genome (Depuydt and Vandepoele, 2021). Glyma.10G031900 encodes an indole-3-ACID 7 protein that functions as the

principal component of the ABA-and auxin dependent reactions during post-germination seed growth (Belin et al., 2009). Glyma.13G270800 ubiquitin-conjugating enzyme 35, Previous studies have shown that ubiquitination plays important roles in plant abiotic stress responses, Protein ubiquitinations play crucial roles for numerous cellular processes such as cell growth, development, and response to diverse biotic and abiotic stresses. (Takahashi et al., 2009; Zhou et al., 2010), The ubiquitin-depen-dent protein degradation pathway is involved in photo-morphogenesis, hormone regulation, floral homeosis, senescence, and pathogen defense (Suzuki et al., 2002; Devoto et al., 2003).

Glyma.17G185000 NAC domain containing protein 83, The NAC (for NAM-ATAF1/2-CUC2) transcription factors constitute one of the largest transcription factor families in plant genomes (Ooka et al., 2004; Olsen et al., 2005b). Roles of many NAC transcription factors have been demonstrated in diverse developmental processes and plant responses to biotic and abiotic stresses, such apical meristem formation (Hibara et al., 2003), cell cycle control (Kim et al., 2006), AtNAC2 functioning in root development (He et al., 2005). cell divi-sion (Riechmann et al., 2000; Kim et al., 2006), NTM2 inte- grates auxin and salt signals in regulating Arabidopsis seed germination (Park et al., 2011), In Arabidopsis thaliana, 105 genes are predicted to encode NAC proteins (Ooka et al., 2004). Song et al. study found The highly homologous NAC transcription factors ANAC060, ANAC040 and ANAC089 regulate important transitions in the early phases of plant development. All three genes play a role in the interplay between the environment and the developmental switch that results in germination and/or seedling development (Song et al., 2022). For germination and seedling development to occur, the protein has to be released from the membrane, which for ANAC089 was shown to be directly affected by changes in the cellular redox status (Albertos et al., 2021).

Glyma.18g183600 far-red elongated hypocotyl 1, Phytochrome A (phyA) is the primary photoreceptor for mediating the far-red high irradiance response in Arabidopsis thaliana.FAR-RED ELONGATED HYPOCOTYL1 (FHY1) and its homolog FHY1-LIKE (FHL) define two positive regulators in the phyA signaling pathway (Shen et al., 2009). Most abundant in young seedlings in the dark.encodes FHY1 protein that mediates the transfer of phytochrome A (phyA) to the nucleus. Phytochrome A (phyA) acts as red and far red (FR) sensing photoreceptors to regulate plant growth and development (Helizon et al., 2018). Multiple metabolic pathways are required to regulate the length of soybean epicotyl (Clouse et al., 1992; Hao et al., 2014).

Glyma.18G185300 one helix protein, The cellular functions of two Arabidopsis (Arabidopsis thaliana) one-helix proteins, OHP1 and OHP2 (also named LIGHTHARVESTING-LIKE2 [LIL2] and LIL6, respectively, because they have sequence similarity to light-harvesting chlorophyll a/b-binding proteins), OHP1 and OHP2 play an essential role in chloroplast development as well as in vegetative growth, The photosynthetic capacity of ohp1-1 and ohp1-2 mutants also was decreased significantly (Myouga et al., 2018).The protein is localized to the thylakoid membrane and its transcript is transiently induced by exposure to high light conditions. increased expression of OHP1 is observed under light stress (Jansson et al., 2000). may constitute a novel mechanism of photoprotection in the plant photosynthetic apparatus (Psencik et al., 2020).

We speculate that traits during soybean domestication are gradually selected, and the priority traits are yield-related traits, such as seed size, oil content, and protein content (Wang et al., 2020). The epicotyl length involved in this study is not a major direct yield trait and therefore demonstrated weak signal of domestication selection.

In general, It is certain that most of the above candidate genes are related to the regulation of light and temperature, For example, the candidate gene Glyma.18G183600 is a phytochrome A (phyA) gene, which is the main photoreceptor mediating the far-red high-irradiation response in *Arabidopsis*. Cellular function of Glyma.18G185300 with sequence similarity to light-harvesting chlorophyll a/b binding protein, Glyma.03G142200 is a protein involved in photosynthesis, and the analysis results show that they are all involved in the growth and development of soybean epicotyl. This is consistent with the results that soybean epicotyl length is greatly affected by different environments. These results can be reflected from the haplotype analysis of ten candidate genes, which can be reflected in the significant differences in different environments (Figure 9).epicotyl However, further functional verification is needed to clarify the whole mechanism of action. More importantly, since the epicotyl is located in the country of cotyledons and true leaves, it is not only involved in seed germination and seedling growth, but also affects early morphogenesis of seedlings. Understanding and regulating the molecular regulatory network of epicotyl length has important guiding significance for crop breeding.

Data availability statement

All whole genome sequencing data in this study have been deposited in the NCBI Sequence Read Archive under ac-cession number PRJNA681974.

Author contributions

HH, and ML conceived the study and contributed to population development. YC, HW, JW, and BG contributed to phenotypic evaluation. HG, and HR analyzed the data. MY, and HG contributed to genotyping. LQ, and YH contributed to experimental design and writing the paper. All authors contributed to the article and approved the submitted version.

Funding

This study was financially supported by the National Natural Science Foundation of China (32172005), Agricultural Science and Technology Innovation Program (ASTIP) of Chinese Academy of Agricultural Sciences.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1033120/full#supplementary-material

References

Albertos, P., Tatematsu, K., Mateos, I., Sánchez-Vicente, I., Fernández-Arbaizar, A., Nakabayashi, K., et al. (2021). Redox feedback regulation of *ANAC089* signaling alters seed germination and stress response. *Cell Rep.* 35 (11), 109263. doi: 10.1016/j.celrep.2021.109263

Albertos, P., T atematsu, K., Mateos, I., Sánchez-Vicente, I., and Lorenzo, O. (2021). Redox feedback regulation of *ANAC089* signaling alters seed germination and stress response. *Cell Rep.* 35 (11), 109263. doi: 10.1016/j.celrep.2021.109263

Bah, A. M., Sun, H., Fei, C., Zhou, J., Dai, H., Zhang, G., et al. (2010). Comparative proteomic analysis of typha angustifolia leaf under chromium, cadmium and lead stress. *J. Hazard Mater.* 184 (1-3), 191–203. doi: 10.1016/j.jhazmat.2010.08.023

Ballif, J., Endo, S., Kotani, M., Macadam, J., and Wu, Y. (2011). Over-expression of *HAP3b* enhances primary root elongation in arabidopsis. *Plant Physiol. Bioch.* 49 (6), 579–583. doi: 10.1016/j.plaphy.2011.01.013

Bandelt, H., Forster, P., and Röhl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* 16 (1), 37–48. doi: 10.1093/oxfordjournals.molbev.a026036

Belin, C., Megies, C., Hauserová, E., and Lopez-Molina, L. (2009). Abscisic acid represses growth of the arabidopsis embryonic axis after germination by enhancing auxin signaling. *Plant Cell.* 21 (8), 2253–2268. doi: 10.1105/tpc.109.067702

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 (19), 2633–2635. doi: 10.1093/bioinformatics/btm308

Camargos, T., Campos, N., Alves, G., Ferreira, S., and Matsuo, D. (2019). The effect of soil volume, plant density and sowing depth on soybean seedlings characters. *Agron. Sci. Biotech.* 5 (2), 47. doi: 10.33158/ASB.2019v5i2p47

Chaves, M. V. A., Silva, N. S., Silva, R. H. O., Jorge, G. L., Silveira, I. C., Medeiros, L. A., et al. (2017). Genotype x environment interaction and stability of soybean cultivars for vegetative-stage characters. *Genet. Mol. Res.* 16 (3). doi: 10.4238/gmr16039795

Chory, J., Reinecke, D., Sim, S., Washburn, T., and Brenner, M. (1994). A role for cytokinins in de-etiolation in arabidopsis. *Plant Physiol.* 104 (2), 339–347. doi: 10.1104/pp.104.2.339

Clement, M., Snell, Q., Walker, P., Posada, D., and Crandall, K. (2002). TCS: Estimating gene genealogies. *parallel distributed Process. symposium Int. Proc.* 2, 184.

Clouse, S. D., Zurek, D. M., McMorris, T. C., and Baker, M. E. (1992). Effect of brassinolide on gene expression in elongating soybean epicotyls. *Plant Physiol.* 100 (3), 1377–1383. doi: 10.1104/pp.100.3.1377

Contreras-Soto, R. I., Mora, F., Lazzari, F., de Oliveira, M. A. R., Scapim, C. A., and Schuster, I. (2017). Genome-wide association mapping for flowering and maturity in tropical soybean: implications for breeding strategies. *Breed Sci.* 67 (5). doi: 10.3389/fpls.2017.01222

Depuydt, T., and Vandepoele, K. (2021). Multi-omics network-based functional annotation of unknown arabidopsis genes. *Plant J.* 108 (4), 1193–1212. doi: 10.1111/tpj.15507

Devoto, A., Muskett, P. R., and Shirasu, K. (2003). Role of ubiquitination in the regulation of plant defense against pathogens. *Curr. Opin. Plant Biol.* 6 (4), 307–311. doi: 10.1016/s1369-5266(03)00060-8

Dinka, S. J., Campbell, M. A., Demers, T., and Raizada, M. N. (2007). Predicting the size of the progeny mapping population required to positionally clone a gene[J]. *Genetics* 176 (4), 2035–2054. doi: 10.1534/genetics.107.074377

Enhaken, R., Doerks, T., and Bork, P. (2005). DCD – a novel plant specific domain in proteins involved in development and programmed cell death. *BMC Bioinf.* 6, 169. doi: 10.1186/1471-2105-6-169

Flint-Garcia, S. A., Thornsberry, J. M., and Th, B. E. (2003). Structure of linkage disequilibrium in plants. *Annu. Rev. Plant Biol.* 54 (4), 357–374. doi: 10.1146/annurev.arplant.54.031902

Frances, S., White, M. J., Edgerton, M. D., Jones, A. M., Elliot, R. C., and Thompson, W. F. (1992). Initial characterization of a pea mutant with light-independent photomorphogenesis. *Plant Cell.* 4 (12), 1519–1530. doi: 10.1105/tpc.4.12.1519

French, N., Yu, S., Biggs, P., Holland, B., Fearnhead, P., Binney, B., et al. (2014). "Evolution of campylobacter species in new Zealand," in *Campylobacter ecology and evolution*. Eds. S. K. Sheppard and G. Méric., 221–240. ISBN:978-1-908230-36-2.

Grove, M. D., Spencer, G. F., and Rohwedder, W. K. (1979). Brassinolide, a plant growth-promoting steroid isolated from brassica napus pollen[J]. *Nature* 281. doi: 10.1038/281216a0

Hanyu, J., Costa, S., Cecon, P., and Matsuo, D. (2020). Genetic parameters estimate and characters analysis in phenotypic phase of soybean during two evaluation periods. *Agron. Sci. Biotech.* 6, 1–12. doi: 10.33158/ASB.r104.v6.2020

Hao, H. P., He, Z., Li, H., Shi, L., and Tang, Y. D. (2014). Effect of root length on epicotyl dormancy release in seeds of paeonia ludlowii. *Tibetan peony. Ann. Bot.* 113 (3), 443–452. doi: 10.1093/aob/mct273

He, J. X., Fujioka, S., Li, T. C., Kang, S. G., Seto, H., Takatsuto, S., et al. (2003). Sterols regulate development and gene expression in *Arabidopsis*. *Plant Physiol.* 131 (3), 1258–1269. doi: 10.1104/pp.014605

Helizon, H., Rösler-Dalton, J., Gasch, P., von Horsten, S., Essen, L. O., and Zeidler, M. (2018). Arabidopsis phytochrome a nuclear translocation is mediated by a far-red elongated hypocotyl 1-importin complex. *Plant J.* 96 (6), 1255–1268. doi: 10.1111/tpj.14107

He, X. J., Mu, R. L., Cao, W. H., Zhang, Z. G., Zhang, J. S., and Chen, S. Y. (2005). *AtNAC2*, a transcription factor downstream of ethylene and auxin signaling pathways, is involved in salt stress response and lateral root development. *Plant J.* 44 (6), 903–916.

Hibara, K., Takada, S., and Tasaka, M. (2003). *CUC1* gene activates the expression of SAM-related genes to induce adventitious shoot formation. *Plant J.* 36 (5), 687–696.

Huang, B., Chu, C. H., Chen, S. L., Juan, H. F., and Chen, Y. M. (2006). A proteomics study of the mung bean epicotyl regulated by brassinosteroids under conditions of chilling stress. *Cell Mol. Biol. Lett.* 11 (2), 264–278. doi: 10.2478/s11658-006-0021-7

Ichinose, A., Bottenus, R. E., and Davie, E. W. (1990). Structure of transglutaminases. *J. Biol. Chem.* 265 (23), 13411–13414. doi: 10.1016/0008-6215(90)80036-3

Ikushima, T., Soga, K., Hoson, T., and Shimmen, T. (2008). Role of xyloglucan in gravitropic bending of azuki bean epicotyl. *Physiol. Plant* 132 (4), 552–565. doi: 10.1111/j.1399-3054.2007.01047

Jansson, S., Andersson, J., Kim, S. J., and Jackowski, G. (2000). An *Arabidopsis* thaliana protein homologous to cyanobacterial high-light-inducible proteins. *Plant Mol. Biol.* 42 (2), 345–351. doi: 10.1023/a:1006365213954

Jia, Q., Xiao, Z. X., Wong, F. L., Sun, S., Liang, K. J., and Lam, H. M. (2017). Genome-wide analyses of the soybean f-box gene family in response to salt stress. *Int. J. Mol. Sci.* 18 (4), 818. doi: 10.3390/ijms18040818

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42 (4), 348–354. doi: 10.1038/ng.548

Kang, H. M., 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 (3), 1709–1723. doi: 10.1534/genetics.107.080101

Kan, G., Zhang, W., Yang, W., Ma, D., Zhang, D., Hao, D., et al. (2015). Association mapping of soybean seed germination under salt stress. *Mol. Genet. Genomics* 290 (6), 2147–2162. doi: 10.1007/s00438-015-1066-y

Kim, Y. S., Kim, S. G., Park, J. E., Park, H. Y., Lim, M. H., Chua, N. H., et al. (2006). A membrane-bound NAC transcription factor regulates cell division in *Arabidopsis*. *Plant Cell.* 18 (11), 3132–3144. doi: 10.1105/tpc.106.043018

Liang, H., Yu, Y., Yang, H., Zhang, H., Wei, D., Cui, W., et al. (2014). Epistatic effects and quantitative trait Loci(QTL) x Environment(QE) interaction effects for yield per plot and botanical traits in soybean. *Chin. Bull. Bot*. doi: 10.3724/SP.J.1259.2014.00273

Li, Y. H., Li, D. L., Jiao, Y. Q., Schnable, J. C., Li, Y. F., Li, H. H., et al. (2020a). Identification of loci controlling adaptation in Chinese soya bean landraces *via* a combination of conventional and bioclimatic GWAS. *Plant Biotechnol. J.* 18 (2), 389–401. doi: 10.1111/pbi.13206

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., et al. (2012). GAPIT:genome association and prediction integrated tool. *Bioinf. 15;* 28 (18), 2397–2399. doi: 10.1093/bioinformatics/bts444

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nat. Methods* 8 (10), 833–835. doi: 10.1038/nmeth.1681

Li, Y., Qin, C., Wang, L., Jiao, C. Z., Hong, H. L., Tia, Y., et al. (2022d). Genome-wide signatures of geographic expansion and breeding process in soybean. *Sci. China Life Sci.* 19. doi: 10.1007/s11427-022-2158-7

Li, Y. H., Reif, J. C., Hong, H. L., Li, H. H., Liu, Z. X., Ma, Y. S., et al. (2018). Genome-wide association mapping of QTL underlying seed oil and protein contents of a diverse panel of soybean accessions. *Plant Sci.* 266, 95–101. doi: 10.1016/j.plantsci.2017.04.013

Liu, M., Tan, X., Yang, Y., Liu, P., Zhang, X., Zhang, Y., et al. (2020). Analysis of the genetic architecture of maize kernel size traits by combined linkage and association mapping. *Plant Biotechnol. J.* 18 (1), 207–221. doi: 10.1111/pbi.13188

Li, M., Zhang, Y. W., Xiang, Y., Liu, M. H., and Zhang, Y. M. (2022c). IIIVmrMLM: The r and c++ tools associated with 3VmrMLM, a comprehensive GWAS method for dissecting quantitative traits. *Mol. Plant* 15 (8), 1251–1253. doi: 10.1016/j.molp.2022.06.002

Li, M., Zhang, Y. W., Zhang, Z. C., Xiang, Y., Liu, M. H., Zhou, Y. H., et al. (2022b).). a compressed variance component mixed model for detecting QTNs and QTN-by-environment and QTN-by-QTN interactions in genome-wide association studies. *Mol. Plant* 15 (4), 630–650.

Li, X., Zheng, H., Wu, W., Liu, H., Wang., J., Jia, Y., et al. (2020). QTL mapping and candidate gene analysis for alkali tolerance in japonica rice at the bud stage based on linkage mapping and genome-wide association study. *Rice (N Y).* 1316 (1), 48. doi: 10.1186/s12284-020-00412-5

Ludwig, A. A., and Tenhaken, R. (2001). A new cell wall located n-rich protein is strongly induced during the hypersensitive response in *Glycine max* l. *Eur. J. Plant Pathol.* 107, 323–336. doi: 10.1023/A:1011202225323

Luo, X., Xue, Z., Ma, C., Hu, K., Zeng, Z., Dou, S., et al. (2017). Joint genome-wide association and transcriptome sequencing reveals a complex polygenic network underlying hypocotyl elongation in rapeseed (*Brassica napus* l.). *Sci. Rep.* 7, 41561. doi: 10.1038/srep41561

Mathur, J., Molnár, G., Fujioka, S., Takatsuto, S., Sakurai, A., Yokota, T., et al. (1998). Transcription of the arabidopsis CPD gene, encoding a steroidogenic cytochrome P450, is negatively controlled by brassinosteroids. *Plant J.* 14 (5), 593–602. doi: 10.1046/j.1365-313x

Matsuo, E., Sediyama, T., Cruz, C. D., and Oliveira, R. (2012). Estimates of the genetic parameters, optimum sample size and conversion of quantitative data in multiple categories for soybean genotypes. *Acta Sci. Agron.* 34 (3), 265–273. doi: 10.4025/actasciagron.v34i3.14015

Matsusaka, D., Filiault, D., Sanchez, D. H., and Botto, J. F. (2021). Ultra-High-Density QTL marker mapping for seedling photomorphogenesis mediating arabidopsis establishment in southern patagonia. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.677728

Mori, M., Maki, K., Kawahata, T., Kawahara, D., Kato, Y., Yoshida, T., et al. (2021). Mapping of QTLs controlling epicotyl length in adzuki bean (*Vigna angularis*). *Breed Sci.* 71 (2), 208–216. doi: 10.1270/jsbbs.20093

Myouga, F., Takahashi, K., Tanaka, R., Nagata, N., and Shinozaki, K. (2018). Stable accumulation of photosystem II requires *ONE-HELIX PROTEIN1 (OHP1)* of the light harvesting-like family[J]. *Plant Physiol.* 176 (3), 01782.2017. doi: 10.1104/pp.17.01782

Okoloko, G. E., Lewis, L. N., and Reid, B. R. (1970). Changes in nucleic acids in phytochrome-dependent elongation of the alaska pea epicotyl. *Plant Physio.* 46 (5), 660–665. doi: 10.1104/pp.46.5.660

Olsen, A. N., Ernst, H. A., Leggio, L. L., and Skriver, K. (2005). DNA-Binding specificity and molecular functions of NAC transcription factors. *Plant Sci.* 169 (4), 785–797. doi: 10.1016/j.plantsci.2005.05.035

Ooka, H., Satoh, K., Doi, K., T Nagata, T. S., and Kikuchi, S. (2004). Comprehensive analysis of NAC family genes in *Oryza sativa* and *Arabidopsis thaliana*. *DNA Res.* 10 (6), 239–247. doi: 10.1093/dnares/10.6.239

Park, J., Kim, Y. S., Kim, S. G., Jung, J. H., Woo, J. C., and Park, C. M. (2011). Integration of auxin and salt signals by the *NAC* transcription factor *NTM2* during seed germination in *Arabidopsis*. *Plant Physiol.* 156 (2), 537–549. doi: 10.1104/pp.111.177071

Psencik, J., Hey, D., Grimm, B., and Lokstein, H. (2020). Photoprotection of photosynthetic pigments in plant one-helix protein 1/2 heterodimers. *J. Phys. Chem. Lett.* 11 (21), 9387–9392. doi: 10.1021/acs.jpclett.0c02660

Riechmann, J. L., Heard, J., Martin, G., Reuber, L., Jiang, C. Z., Keddie, J., et al. (2000). *Arabidopsis* transcription factors: Genome-wide comparative analysis among eukaryotes. *Science* 290 (5499), 2105–2110. doi: 10.1126/science.290.5499.2105

Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., et al. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* 44 (7), 825–830. doi: 10.1038/ng.2314

Seyedi, M., Selstam, E., Timko, M. P., and Sundqvist, C. (2001). The cytokinin 2-isopentenyladenine causes partial reversion to skotomorphogenesis and induces formation of prolamellar bodies and protochlorophyllide657 in the lip1 mutant of pea. *Physiol. Plant* 112 (2), 261–272. doi: 10.1034/j.1399-3054

Shen, Y., Zhou, Z., Feng, S., Li, J., Tan-Wilson, A., Qu, L. J., et al. (2009). Phytochrome a mediates rapid red light-induced phosphorylation of arabidopsis FAR-RED *ELONGATED HYPOCOTYL1* in a low fluence response. *Plant Cell.* 21 (2), 494–506. doi: 10.1105/tpc.108.061259

Song, Q., Hyten, D. L., Jia, G., Quigley, C. V., Fickus, E. W., Nelson, R. L., et al. (2013). Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *PloS One* 8 (1), e54985. doi: 10.1371/journal.pone.0054985

Song, S., Willems, L., Jiao, A., Zhao, T., Eric, S. M., and Leónie, B. (2022). The membrane associated NAC transcription factors *ANAC060* and *ANAC040* are functionally redundant in the inhibition of seed dormancy in arabidopsis thaliana. *J. Exp. Bot.* 73 (16), 5514–5528. doi: 10.1093/jxb/erac232

Sui, M., Jing, Y., Li, H., Zhan, Y., Luo, J., Teng, W., et al. (2020). Identification of loci and candidate genes analyses for tocopherol concentration of soybean seed. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.539460

Suzuki, G., Yanagawa, Y., Kwok, S. F., Matsui, M., and Deng, X. W. (2002). *Arabidopsis COP10* is a ubiquitin-conjugating enzyme variant that acts together with *COP1* and the *COP9* signalosome in repressing photomorphogenesis. *Genes Dev.* 16 (5), 554–559. doi: 10.1101/gad.964602

Takahashi, H., Nozawa, A., Seki, M., Shinozaki, K., Endo, Y., and Sawasaki, T. (2009). A simple and high-sensitivity method for analysis of ubiquitination and polyubiquitination based on wheat cell-free protein synthesis. *BMC Plant Biol.* 9 (1), 39. doi: 10.1186/1471-2229-9-39

Theodorsson, N. E. (1986). Kruskal-Wallis test: BASIC computer program to perform nonparametric one-way analysis of variance and multiple comparisons on ranks of several independent samples. *Comput. Meth Prog. Bio.* 23 (1), 57–62. doi: 10.1016/0169-2607(86)90081-7

Wang, S. B., Feng, J. Y., Ren, W. L., Huang, B., Zhou, L., Wen, Y. J., et al. (2016). Improving power and accuracy of genome-wide association studies *via* a multi-locus mixed linear model methodology. *Sci. Rep.* 6, 19444. doi: 10.1038/srep19444

Wang, X., Guan, P., Xin, M., Wang, Y., Chen, X., Zhao, A., et al. (2021). Genome-wide association study identifies QTL for thousand grain weight in winter wheat under normal- and late-sown stressed environments. *Theor. Appl. Genet.* 134 (1), 143–157. doi: 10.1007/s00122-020-03687-w

Wang, S., Liu, S., Wang, J., Yokosho, K., Zhou, B., Yu, Y. C., et al. (2020). Simultaneous changes in seed size, oil content and protein content driven by selection of *SWEET* homologues during soybean domestication. *Natl. Sci. Rev.* 7 (11), 1776–1786. doi: 10.1093/nsr/nwaa110

Wen, Y. J., Zhang, H., Ni, Y. L., Huang, B., Zhang, J., Feng, J. Y., et al. (2017). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief Bioinform.* 19 (4), 700–712. doi: 10.1093/bib/bbw145

Wolyn, D. J., Borevitz, J. O., Loudet, O., Schwartz, C., Maloof, J., Ecker, J. R., et al. (2004). Light-response quantitative trait loci identified with composite interval and eXtreme array mapping in *Arabidopsis thaliana*. *Genetics* 167 (2), 907–917. doi: 10.1534/genetics.103.024810

Wu, D. P., Li, D. M., Zhao, X., Zhan, Y. H., Teng, W. L., Qiu, L. J., et al. (2020). Identification of a candidate gene associated with isoflavone content in soybean seeds using genome-wide association and linkage mapping. *Plant J.* 104 (4), 950–963. doi: 10.1111/tpj.14972

Wycoff, K. L., Rhijn, P. V., and Hirsch, A. M. (1997). The ribosomal protein P0 of soybean (*Glycine max* l. merr.) has antigenic cross-reactivity to soybean seed lectin. *Plant Mol. Biol.* 34 (2), 295–306. doi: 10.1023/a:1005817114562

Xiao, H., Tang, S., An, Y., Zheng, D. C., Xia, X. L., and Yin, W. L. (2013). Overexpression of the poplar NF-YB7 transcription factor confers drought tolerance and improves water-use efficiency in arabidopsis. *J. Exp. Bot.* 64 (14), 4589–4601. doi: 10.1093/jxb/ert262

Yamamoto, A., Kagaya, Y., T oyoshima, R., Kagaya, M., T akeda, S., and Hattori, T. (2009). *Arabidopsis* NF-YB subunits LEC1 and LEC1-LIKE activate transcription by interacting with seed-specific ABRE-binding factors. *Plant J.* 58 (5), 843–856. doi: 10.1111/j.1365-313X.2009.03817.x

Yan, L., Hofmann, N., Li, S., Ferreira, M. E., Song, B., Jiang, G., et al. (2017). Identification of QTL with large effect on seed weight in a selective population of soybean with genome-wide association and fixation index analyses. *BMC Genomics* 18 (1), 529. doi: 10.1186/s12864-017-3922-0

Yu, J., Pressoir, G., Briggs, W. H., Vroh, B. 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 (2), 203–208. doi: 10.1038/ng1702

Yu, Y., Zhang, H., Long, Y. P., Shu, Y., and Zhai, J. X. (2022). PPRD: a comprehensive online database for expression analysis of ~45,000 plant public RNA-seq libraries. *Plant Biotechnol. J.* 20 (5), 806–808. doi: 10.1111/pbi.13798

Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42 (4), 355–360. doi: 10.1038/ng.546

Zhang, Y. M., Mao, Y., Xie, C., Smith, H., Luo, L., and Xu, S. (2005). Mapping quantitative trait loci using naturally occurring genetic variance among commercial inbred lines of maize (*Zea mays* l.). *Genetics* 169 (4), 2267–2275. doi: 10.1534/genetics.104.033217

Zhang, J. P., Singh, A. P., Mueller, D. S., and Singh, A. K. (2015). Genome-wide association and epistasis studies unravel the genetic architecture of sudden death syndrome resistance in soybean. *Plant J.* 84 (6), 1124–1136. doi: 10.1111/tpj.13069

Zhang, T., Wu, T., Wang, L., Jiang, B., Zhen, C., Yuan, S., et al. (2019). A combined linkage and GWAS analysis identifies QTLs linked to soybean seed protein and oil content. *Int. J. Mol. Sci.* 20 (23), 5915. doi: 10.3390/ijms20235915

Zhao, X., Han, Y., Li, Y., Liu, D., Sun, M., Zhao, Y., et al. (2015). Loci and candidate gene identification for resistance to sclerotinia sclerotiorum in soybean (*Glycine max l.* merr.) *via* association and linkage maps. *Plant J.* 82 (2), 245–255. doi: 10.1111/tpj.12810

Zhao, S. P., Lu, D., Yu, T. F., Ji, Y. J., Zheng, W. J., Zhang, S. X., et al. (2017). Genome-wide analysis of the YABBY family in soybean and functional identification of *GmYABBY10* involvement in high salt and drought stresses. *Plant Physiol. Biochem.* 119, 132–146. doi: 10.1016/j.plaphy.2017.08.026

Zhao, X., Teng, W. L., Li, Y. H., Liu, D. Y., Cao, G. L., Li, D. M., et al. (2017). Loci and candidate genes conferring resistance to soybean cyst nematode HG type 2.5.7. *BMC Genomics* 14, 18(1):462. doi: 10.1186/s12864-017-3843-y

Zhou, G. A., Chang, R. Z., and Qiu, L. J. (2010). Overexpression of soybean ubiquitin-conjugating enzyme gene GmUBC2 confers enhanced drought and salt tolerance through modulating abiotic stress-responsive gene expression in arabidopsis. *Plant Mol. Biol.* 72 (s4-5), 357–367. doi: 10.1007/s11103-009-9575-x

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44 (7), 821–824. doi: 10.1038/ng.2310

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 21 November 2022
DOI 10.3389/fpls.2022.1065766

OPEN ACCESS

EDITED BY
Zhenyu Jia,
University of California, Riverside,
United States

REVIEWED BY
Ling Qiao,
Shanxi Agricultural University, China
Yang-Jun Wen,
Nanjing Agricultural University, China

*CORRESPONDENCE
Liang Guo
guoliang@mail.hzau.edu.cn

SPECIALTY SECTION
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

RECEIVED 10 October 2022
ACCEPTED 31 October 2022
PUBLISHED 21 November 2022

CITATION
Han X, Tang Q, Xu L, Guan Z, Tu J, Yi B, Liu K, Yao X, Lu S and Guo L (2022) Genome-wide detection of genotype environment interactions for flowering time in *Brassica napus*. *Front. Plant Sci.* 13:1065766. doi: 10.3389/fpls.2022.1065766

Genome-wide detection of genotype environment interactions for flowering time in *Brassica napus*

Xu Han[1,2], Qingqing Tang[1,2], Liping Xu[1,2], Zhilin Guan[1], Jinxing Tu[1,2], Bin Yi[1,2], Kede Liu[1], Xuan Yao[1,2], Shaoping Lu[1,2] and Liang Guo[1,2]*

[1]National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, China, [2]Hubei Hongshan Laboratory, Wuhan, China

Flowering time is strongly related to the environment, while the genotype-by-environment interaction study for flowering time is lacking in *Brassica napus*. Here, a total of 11,700,689 single nucleotide polymorphisms in 490 *B. napus* accessions were used to associate with the flowering time and related climatic index in eight environments using a compressed variance-component mixed model, 3VmrMLM. As a result, 19 stable main-effect quantitative trait nucleotides (QTNs) and 32 QTN-by-environment interactions (QEIs) for flowering time were detected. Four windows of daily average temperature and precipitation were found to be climatic factors highly correlated with flowering time. Ten main-effect QTNs were found to be associated with these flowering-time-related climatic indexes. Using differentially expressed gene (DEG) analysis in semi-winter and spring oilseed rapes, 5,850 and 5,511 DEGs were found to be significantly expressed before and after vernalization. Twelve and 14 DEGs, including 7 and 9 known homologs in *Arabidopsis*, were found to be candidate genes for stable QTNs and QEIs for flowering time, respectively. Five DEGs were found to be candidate genes for main-effect QTNs for flowering-time-related climatic index. These candidate genes, such as *BnaFLC*s, *BnaFT*s, *BnaA02.VIN3*, and *BnaC09.PRR7*, were further validated by the haplotype, selective sweep, and co-expression networks analysis. The candidate genes identified in this study will be helpful to breed *B. napus* varieties adapted to particular environments with optimized flowering time.

KEYWORDS

***Brassica napus*, flowering time, QTN-by-environment interactions, multiple genome-wide association studies, differentially expressed gene, climatic index**

Introduction

As the world's most important oilseed crop, planting of *Brassica napus* spans a wide range of growth periods and climate zones (Yang et al., 2014). To meet the needs of adaptation, *B. napus* adjusts the correct time to flower. Flowering time determines the transition from the vegetative to the reproductive phase, and therefore, the nutrients are available for remobilization at seed filling (Han et al., 2021). Early flowering facilitates mechanical harvesting and rotation with other crops, whereas late flowering enhances stem development, thus improving lodging resistance (Cui et al., 2021). Although previous studies have revealed the genetic basis of flowering time in *B. napus*, no studies have been reported on the genetic dissection of flowering time plasticity, namely, genotype-by-environment interaction (G by E).

The genetic basis of flowering time has been well-studied in the model plant *Arabidopsis thaliana* (Mouradov et al., 2002; Putterill et al., 2004; Bouché et al., 2016). The genetic networks underlying flowering consist of six major pathways interconnected, namely, photoperiod, vernalization, gibberellin, autonomous, thermal clock, and aging pathways (Putterill et al., 2004). Epigenetic regulation, miRNAs, phytohormones, sugar status, and signaling also play important roles in flowering time control (Bouché et al., 2016). In *B. napus*, the polyploid nature of *B. napus* has resulted in flowering-time-related genes undergoing extensive subfunctionalization (Schiessl, 2020). It has been demonstrated that there is a sophisticated network of interactions among *FLOWERING LOCUS C* homologs with different expression patterns in organs and development stages (Zou et al., 2012). *LOWERING LOCUS T* and *TERMINAL FLOWER 1* were found to have pleiotropic effects on flowering time, despite their redundancy in *B. napus* genome (Guo et al., 2014). Therefore, it demands more genetic basis research on flowering time in *B. napus*.

Flowering time is strongly influenced by the environment. A decrease in day length delays flowering in *B. napus*. A period of cooler temperature will determine vernalization and ensure reproductive development (Matar et al., 2021). Precipitation has been reported to have different effects on flowering phenology in different species (Zhang et al., 2018). Many genes have been reported to influence flowering time in response to the environment. *FLOWERING LOCUS T* (FT) was found to induce flowering through long-distance signaling by activating seasonal changes in day length (Corbesier et al., 2007). The epigenetic silencing of *FLC* accelerates flowering by prolonged cold vernalization (Bastow et al., 2004). H2A.Z incorporates *BraA.FT.a* chromatin at high ambient temperature and delays flowering time in *B. rapa* (Del Olmo et al., 2019). In *B. napus*, *Cycling Dof Factor1* delays the flowering time and was induced in response to low temperature (Xu and Dai, 2016). *BnNAC485* altered flowering time in response to abiotic stress (Ying et al., 2014). However, *B. napus* has developed two eco-types in China, namely, semi-winter oilseed rapes (SWORs) and spring oilseed rapes (SORs), to adapt different geographical environments and climates, leading to more complex molecular mechanisms of flowering time (Song et al., 2020).

In response to climate change, G by E is of fundamental importance in plant breeding and adaptation (Arnold et al., 2019; Zhao et al., 2022). In *B. napus*, the G by E of seed yield and oil content were found to exert specific adaptation to climates (Zhang et al., 2013a). Genotype and temperature interactions of seed oil content were found to be differential at the level of gene expression profiles (Zhu et al., 2012). Moreover, quantitative and population genetics have shown great power to bridge the gap between genomic diversity and phenotypic plasticity (Wu, 1998; Kusmec et al., 2017; Liu et al., 2021). For G by E studies on flowering time, four environmentally sensitive quantitative trait loci for flowering time identified in 473 *Arabidopsis* accessions were found to be related to adaptation (Li et al., 2010). It has been found that interacting flowering-time-related genes differentially respond to the temperature at the early growth stage in rice (Guo et al., 2020). Quantitative trait nucleotide (QTN)-by-environment interaction (QEI) mapping for flowering time has been performed in a doubled haploid *B. napus* population (Shen et al., 2018). Although many genome-wide association studies (GWAS) for flowering time have been reported in *B. napus* (Xu et al., 2016; Song et al., 2020; Helal et al., 2021; Hu et al., 2022), knowledge about QEI for flowering time detected by GWAS is scarce.

Recently, the newly published method 3VmrMLM provides a solution for QEI detection in GWAS (Li et al., 2022a). Here, we investigated the landscape of flowering time plasticity of 490 *B. napus* accessions in eight environments. A total of 11,700,689 single nucleotide polymorphisms (SNPs) were used to detect main-effect QTNs for flowering time and related climatic index and QEIs for flowering time. The transcriptome of SWORs and SORs before and after vernalization was used to identify the candidate genes around QTNs and QEIs. Co-expression, haplotype, and selection sweep analysis were used to further validate the candidate flowering time genes in specific eco-oilseed rapes. Our finding will facilitate the breeding for adaptation to particular environments with optimized flowering time in *B. napus*.

Materials and methods

Germplasm, phenotypic, and genomic data

A diversity panel of 490 *B. napus* accessions collected from Xu et al. (2016) was used in this study. This panel was cultivated

in eight natural environments, i.e., Wuhan 2013 and 2014 (WH2013 and WH2014), Changsha 2013 and 2014 (CS2013 and CS2014), Nanjing 2013 and 2014 (NJ2013 and NJ2014), Ezhou 2013 (EZ2013), and Chongqing 2013 (CQ2013). Additionally, the Gangan and ZS11 cultivars for RNA-seq were planted in Wuhan 2018 at the experimental stations of Huazhong Agricultural University. The design of field trial of the above materials and the acquisition of phenotypic data were the same as those used in the previous study (Xu et al., 2016). The re-sequencing genome data were obtained from Tang et al. (2021). The *B. napus* genome (*B. napus* ZS11 v0) from BnPIR (Song et al., 2020; Song et al., 2021) (http://cbi.hzau.edu.cn/bnapus/index.php) was used as the reference genome.

Statistical analysis for phenotypic data

By using the "lme4" R package (Bates et al., 2015), the best linear unbiased prediction (BLUP) model was fitted to each *B. napus* accession:

$$\text{Phenotype} \sim (1|\text{Accession}) + (1|\text{Environment})$$

Taking into account the variations between eight environments as phenotypic variance derived from environmental factors, broad-sense heritability (h_B^2) was estimated using the following equation by treating populations as a random effect and the environments as an environment effect, where σ_g^2 and σ_e^2 is the variance derived from genetic and environmental effects, respectively (Knapp et al., 1985).

$$h_B^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$$

Identification of flowering-time-related climatic index

Climatic data for daily average temperature (TAVG, °F) and precipitation (PRCP, in) were retrieved from the National Oceanic and Atmospheric Administration (https://www.noaa.gov/weather). Due to the lack of climatic data for Ezhou, there were climatic datasets of seven environments in total, i.e., WH2013 and WH2014 (114.05°E, 30.60°N; Station ID: GHCND: CHM00057494), CS2013 and CS2014 (112.87°E, 28.23°N; GHCND: CHM00057687), NJ2013 and NJ2014 (118.90°E, 31.93°N; GHCND: CHM00058238), and CQ2013 (106.48°E, 29.58°N; GHCND: CHM00057516). Climatic data were obtained from the day after being planted to the 200 days after planting (DAP). For each window from a starting day (3 DAP) to an end day (41 DAP) during *B. napus* growth, the average value of the climatic index and their correlation with the environmental mean vector for flowering time was calculated by CERIS analytical package (Li et al., 2021; https://github.com/jmyu/CERIS_JGRA). The most relevant climatic index for flowering time was chosen according to the highest correlation between environmental means and climatic index with corresponding window. Reaction norms were calculated as described in Guo et al. (2020) and Liu et al. (2020), using environmental mean and environmental climatic index as x-axis and phenotype as y-axis. Each line represented an individual and was shown by fitted linear regression. The intercept and slope were used to perform GWAS further.

Detecting QTNs and QEIs by GWAS

The intersection of the accessions in phenotypic and genotypic datasets, i.e., 490 accessions with 11,700,689 SNPs, were used for GWAS using 3VmrMLM (Li et al., 2022a) *via* software IIIVmrMLM (Li et al., 2022b). Flowering time QTNs were obtained from separate analyses of phenotypic data from eight environments and joint environmental analyses of these datasets. The reaction norms between flowering time and climatic index were also used to conduct GWAS by 3VmrMLM. QEIs for flowering time were obtained by joint environment analyses of the above phenotypic datasets in eight environments. Population structure and kinship matrix were considered in 3VmrMLM analysis, and the "svpal" parameter was set as 0.01. According to Tang et al. (2021), the population structure calculated as K=3 was used in the analysis. The threshold was set at $0.05/m$ for significant QTNs and QEIs and LOD score ≥ 3.0 for suggested QTNs and QEIs, where m is the number of markers (Li et al., 2022a; Li et al., 2022b). According to the LD interval estimated by Tang et al. (2021), stable QTNs were defined as QTNs identified in at least three environments within the 100-kb upstream and downstream regions.

Identification of candidate genes

To identify candidate genes for flowering-time-related QTNs and QEIs, genes within the 100 kb upstream and downstream regions of each QTN or QEI were extracted according to the LD interval estimated by Tang et al. (2021). Then, two strategies were employed. First, the *B. napus* homologs of *Arabidopsis* flowering time genes downloaded from FLOR-ID (http://www.flor-id.org) were selected and considered as known genes. Second, new candidate genes were identified using differentially expressed genes (DEGs) in two SWORs (Gangan and ZS11) before and after vernalization and in two SORs (Westar and No. 2127). The *t*-test was adopted in the hypothesis testing for haplotype analysis; $p< 0.05$, $p< 0.01$, and $p< 0.001$ indicated the significances at 0.05, 0.01, and 0.001 probability levels, respectively.

Differential expression analysis based on RNA-seq

The leaves of Westar, No. 2127, Gangan, ZS11 at 24 and 147 DAP were collected for RNA-seq with two biological replicates. Total RNA was extracted using the TIANGEN RNAprep Pure Plant Kit. Sequencing libraries were generated using the NEBNext® UltraTM RNA Library Prep Kit for Illumina® (NEB, USA) and were sequenced on an Illumina Hiseq 4000 platform. The detailed processes were described in Tan et al. (2022). We used MultiQC (Ewels et al., 2016) to perform quality control and Salmon (Patro et al., 2017) to quantify the RNA-seq reads of annotated genes in the reference ZS11. DESeq2 was used for differential expression analysis (Love et al., 2014). The threshold for DEG is set as the absolute value of $\log_2$FoldChange >1 and adjusted $p< 0.05$ (two-tailed Student's *t*-test; Tan et al., 2022).

Identification of selective sweep signals

To detect the regions under selective sweeps between SWOR and SOR, XP-CLR (v1.1.1), a genome scan using the composite likelihood approach was performed in sub-populations (Chen et al., 2010). Each chromosome was analyzed using the XP-CLR command with the parameters "–ld 0.99 –phased –maxsnps 200 –size 100000 –step 10000." Non-overlapping 20-kp windows within the top 20% XP-CLR scores were merged into one single region, and then, these regions in the top 1% of XP-CLR scores were considered as candidate selective regions (An et al., 2019).

Construction of co-expression network

According to the above RNA-seq datasets, Pearson correlation analysis was calculated between candidate genes and DEGs in SWORs and SORs, respectively. Significant genes were considered to be co-expressed when Pearson correlation coefficient was >0.80 and *p*-value was<0.05. Network visualization was implemented with the Cytoscape package (Shannon et al., 2003).

Results

Flowering time plasticity and related climatic index for *B. napus*

Complex flowering time variation was observed in diversity group of 490 *B. napus* oilseed rapes, including 49 SORs, 20 winter oilseed rapes, 326 SWORs, and 95 mixed type oilseed rapes, grown in eight natural environments (Figure 1A; Supplementary Table S1). The means plus standard deviations of the eight environments WH2013, WH2014, CS2013, CS2014, NJ2013, NJ2014, CQ2013, EZ2013, and BLUP values were 155.49 ± 3.80, 153.56 ± 9.61, 160.27 ± 4.29, 166.55 ± 5.44, 160.50 ± 5.49, 167.55 ± 6.38, 151.31 ± 7.82, 162.57 ± 5.30, and 159.68 ± 4.83 (DAP), respectively (Figure 1B). The correlation of each pair of environments ranged from 0.37 to 0.72 (0.50 ± 0.09). The coefficients of variation, skewness, and kurtosis of the trait in eight environments illustrated that flowering time is a typical quantitative trait (Supplementary Table S2). The broad-sense heritability for flowering time is 0.86. More importantly, joint regression analysis modeled with environmental mean showed the presence of a significant phenotypic plasticity (Figure 1C).

Climate change is altering the environment in which all plants grow. To understand the effect of climatic index on flowering time plasticity, the correlation between environmental means and climatic index (TAVG and PRCP) for different growth windows was predicted by CERIS (Supplementary Table S3). The results of the correlation pattern between TAVG and flowering time showed a positive correlation at early seedling stage and a negative trend after bolting stage, while the pattern of PRCP was exactly opposite (Figure 1D; Supplementary Figure S1A; Supplementary Table S4). The windows with the highest negative ($TAVG_{135–144}$ and $PRCP_{3–41}$) and positive correlations ($TAVG_{10–19}$ and $PRCP_{133–169}$) were chosen as the most related climatic index for further analysis (Figures 1E, F; Supplementary Figures S2A–F; Supplementary Table S4). $TAVG_{135–144}$ (r = −0.986) showed higher correlation with flowering time than $PRCP_{3–41}$ (r = −0.809). $TAVG_{10–19}$ (r = 0.922) showed higher correlation with flowering time than $PRCP_{133–169}$ (r = 0.901). It is noted that these windows are surrounded by other windows with slightly decreasing correlation values (Figure 1D; Supplementary Figure S1).

Detection of QTNs for flowering time

To detect QTNs for flowering time, the phenotypes in each of the eight environments were used to associate with 11,700,689 SNPs using 3VmrMLM under population structure and polygenic background control. As a result, 55, 57, 42, 49, 54, 50, 44, and 43 significant QTNs at the critical *p*-value of 4.27e −09 (=0.05/*m*, where *m* is the number of markers) and 10, 5, 14, 10, 8, 13, 13, and 13 suggested QTNs (with the LOD score ≥ 3.0 but the $p > 0.05/m$) were identified for WH2013, WH2014, CS2013, CS2014, NJ2013, NJ2014, CQ2013, and EZ2013, respectively (Supplementary Table S5; Supplementary Figure S3). In addition, flowering phenotypes from eight environments were used to perform joint analysis by 3VmrMLM. Sixty-eight significant and 11 suggested QTNs were identified. Based on the above QTNs in single and

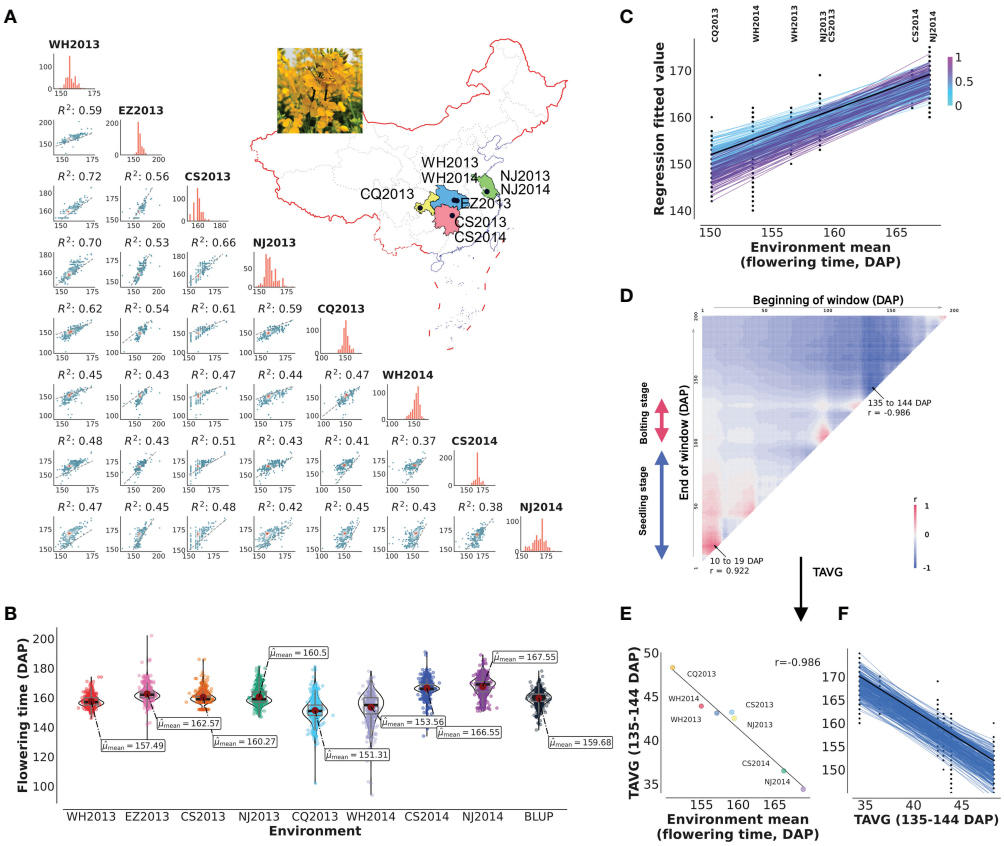


FIGURE 1
Plasticity of flowering and reaction norm of its associated window to daily average temperature (TAVG). **(A, B)** Characteristics and pairwise correlations of flowering time of 490 *B napus* in eight environments. WH2013, Wuhan in 2013; WH2014, Wuhan in 2014; CS2013, Changsha in 2013; CS2014, Changsha in 2014; NJ2013, Nanjing in 2013; NJ2014, Nanjing in 2014; CQ2013, Chongqing in 2013; EZ2013, Ezhou in 2013; BLUP, the best linear unbiased prediction value. **(C)** Reaction norm for flowering time based on a numerical order of environmental mean. Dots are the observed flowering time phenotypic values. The line with black color represents the ZS11 cultivar. The color of the line represents the value of the slope. **(D)** Search for the window to TAVG, which is highly correlated with environmental mean of flowering time (from planting to 200 days after planting, DAP). TAVG within the window of 10–19 and 135–144 DAP was chosen and denoted as $TAVG_{10-19}$ and $TAVG_{135-144}$. **(E, F)** Significant correlation and reaction norm between $TAVG_{135-144}$ and environmental mean of flowering time.

multiple environments analyses, 19 stable QTNs were identified in at least three environments (Figure 2A; Table 1).

Detection of QTN-by-environment interactions for flowering time in multiple environments

All the datasets in eight environments were used to conduct joint analysis for identifying QEIs using 3VmrMLM. As a result, 32 significant QEIs and 4 suggested QEIs were identified, including 10 significant QEIs overlapped with the above stable QTNs (Supplementary Table S6). Among these significant QEIs, 20 were found to have the highest absolute value of additive-by-environment interaction effects in WH2014 than those in other environments (Figures 2B, D), e.g., BnvaC0967693730 has an additive-by-environment interaction effect of −1.85 in WH2014 than those in other environments (Supplementary Table S6; LOD = 67.17; R^2 = 1.07%). The two loci BnvaC0967693730 and BnvaA0406097547 have the highest R^2 (LOD = 67.17; R^2 = 1.07% and LOD = 66.42; R^2 = 1.06%, respectively).

Detection of QTNs for flowering-time-related climatic index

To obtain reaction norms of flowering-time-related climatic index, joint regression analyses were performed on phenotypes and the above flowering-time-related climatic indexes ($TAVG_{135-144}$, $PRCP_{3-41}$, $TAVG_{10-19}$, and $PRCP_{133-169}$; Supplementary Table S4). The intercept and slope of reaction-

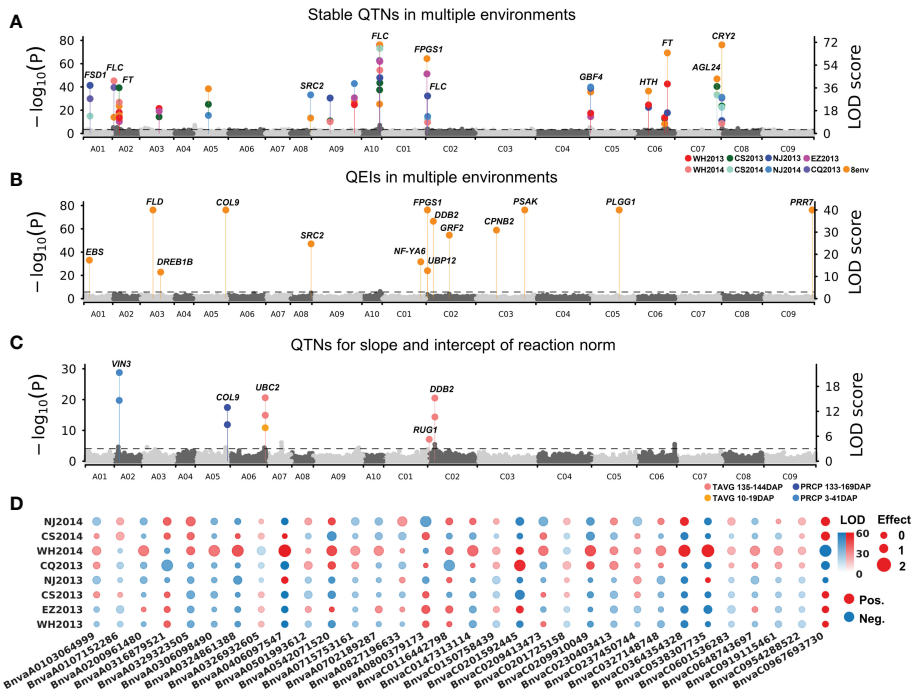


FIGURE 2
Manhattan plots for flowering time of 490 *B napus* accessions. **(A)** Nineteen stable main-effect QTNs and their candidate genes for flowering time in eight single environment analyses and multiple environments joint analysis. **(B)** QTN-by-environment interactions (QEIs) and their candidate genes for flowering time in multiple environments joint analysis. **(C)** Ten main-effect QTNs for slope and intercept of reaction norm for flowering-time-related climatic indexes. **(D)** Additive-by-environment interaction effects of 32 QEIs in eight environments. The size of dot: absolute value of additive-by-environment interaction effect. Red/blue dot: positive/blue value. WH2013, Wuhan in 2013; WH2014, Wuhan in 2014; CS2013, Changsha in 2013; CS2014, Changsha in 2014; NJ2013, Nanjing in 2013; NJ2014, Nanjing in 2014; CQ2013, Chongqing in 2013; EZ2013, Ezhou in 2013.

norm parameters were used to detect QTNs for flowering-time-related climatic indexes using 3VmrMLM. As a result, 10 QTNs for reaction norm parameters of *B. napus* flowering time were commonly identified with the above stable QTNs or QEIs, including 5, 2, 1, and 2 for $TAVG_{135–144}$, $PRCP_{3–41}$, $TAVG_{10–19}$, and $PRCP_{133–169}$, respectively (Figure 2C; Supplementary Table S7).

Prediction of candidate genes for flowering time

To mine candidate genes among the above QTNs and QEIs, DEGs analysis was conducted before and after vernalization. A total of 5,511 DEGs were identified in two SORs before and after vernalization (Figure 3A; Supplementary Table S8), and 5,850 DEGs were identified in two SWORs before and after vernalization (Figure 3A; Supplementary Table S9). Then, according to *Arabidopsis* gene annotation, 12 candidate genes were found to be associated with flowering time in approximately above 19 stable QTNs, including 7 known flowering-time-related homologs in *Arabidopsis* and 5 newly discovered genes (Table 1). Using the same methods, 14 candidate genes were identified to be located in the above 32 QEIs, including 9 homologs of known genes, in which their homologs are related to flowering time and environments in *Arabidopsis* and 5 newly identified genes (Table 2). In addition, five candidate genes were found to be associated with flowering-time-related climatic index, including two genes (*BnaC02.DDB2* and *BnaA05.COL9*) commonly identified in QEIs and three genes (*BnaA02.VIN3*, *BnaC02.RUG1*, and *BnaA06.UBC2*) commonly found to be associated with the flowering time QTNs (Supplementary Table S7).

Among these candidate genes, *BnaFTs*, *BnaA05.COL9*, *BnaA08.SRC2*, and *BnaA03.DREB1B* were significantly upregulated before vernalization in both SWORs and spring SORs, while *BnaFLCs*, *BnaA01.FSD1*, *BnaC02.RUG1*, *BnaC05.PLGG*, and *BnaC03.PSAK* were significantly upregulated after vernalization (Figures 3A, B). Interestingly, *BnaA02.VIN3* and *BnaC02.FPGS* were only significantly upregulated before vernalization in SOR, which may indicate different functions between eco-types.

TABLE 1 Nineteen stable QTNs for *B. napus* flowering time and their candidate genes.

Genome-wide association studies						Comparative genomics analysis			
Chr	Pos (bp)	Marker	LOD	R^2	Environments[a]	Gene ID	Abbr.	Function	Reference
A10	24056113–24056153	*BnvaA1024056153, BnvaA1024056113, BnvaA1024056139*	39.96–117.88	0.49–2.10	E1, E3, E4, E5, E6, E7, E9				
C08	912878	*BnvaC0800912878*	7.94–82.49	0.53–2.42	E1, E3, E4, E5, E6, E7, E8	*BnaC08G0010300ZS*	*CRY2*	Cryptochrome-2	Sharma et al., 2022
C05	1376324	*BnvaC0501376324*	13.31–36.7	0.12–0.81	E1, E2, E6, E7, E9	*BnaC05G0024000ZS*	*GBF4*	G-BOX BINDING FACTOR 4	
A09	56413085–56417605	*BnvaA0956413085, BnvaA0956414961, BnvaA0956417605*	22.79–39.42	0.14–1.43	E1, E2, E9, E7				
A10	23668965–23770033	*BnvaA1023770033, BnvaA1023668965*	23.22–84.79	0.11–2.11	E1, E2, E4, E8	*BnaA10G0244800ZS*	*FLC*	MADS-box protein FLOWERING LOCUS C	Tadege et al., 2001
A02	9020851–9105883	*BnvaA0209020851, BnvaA0209054089, BnvaA0209105883*	9.42–24.67	0.08–1.45	E1, E2, E3, E9	*BnaA02G0156900ZS*	*FT*	Protein FLOWERING LOCUS T	Wang et al., 2009
C02	2400090–2502621	*BnvaC0202402020, BnvaC0202400090, BnvaC0202502621, BnvaC0202402023*	8.82–29.63	0.30–1.47	E3, E5, E6, E7	*BnaC02G0039100ZS*	*FLC*	MADS-box protein FLOWERING LOCUS C	Tadege et al., 2001
C07	55454986–55455005	*BnvaC0755455005, BnvaC0755454986*	30.53–43.11	0.14–0.67	E1, E4, E5	*BnaC07G0458500ZS*	*AGL24*	MADS-box protein AGL24	Yu et al., 2002
C02	1592445	*BnvaC0201592445*	47.10–59.13	0.19–1.06	E1, E9	*BnaC02G0022200ZS*	*FPGS1*	Folylpolyglutamate synthase	
A01	8566494–8643230	*BnvaA0108643230, BnvaA0108602009, BnvaA0108566494*	13.70–37.97	0.55–1.29	E5, E6, E8	*BnaA01G0146300ZS*	*FSD1*	Fe superoxide dismutase. Superoxide dismutase	
A05	19689622	*BnvaA0519689622*	14.27–35.31	0.13–0.48	E1, E4, E7				
A08	27196633–27207043	*BnvaA0827196633, BnvaA0827207043, BnvaA0827196973*	12.21–30.51	0.13–1.71	E1, E7	*BnaA08G0296600ZS*	*SRC2*	soybean gene regulated by cold-2	
A02	8776765–8833814	*BnvaA0208833814, BnvaA0208776765*	16.66–36.12	0.24–0.76	E1, E2, E4				
A03	25426194–25520626	*BnvaA0325426194, BnvaA0325520626*	13.14–19.74	0.37–0.63	E2, E4, E9				
C06	17894906	*BnvaC0617894906*	20.65–33.50	0.15–0.41	E1, E2, E6				
C06	39070745–39079766	*BnvaC0639079766, BnvaC0639070745*	7.60–12.55	0.08–0.55	E1, E2, E4	*BnaC06G0286700ZS*	*HTH*	Omega-Hydroxy Fatty Acyl Dehydrogenase	
C06	42697888	*BnvaC0642697888*	16.35–63.64	0.31–0.74	E1, E2, E6	*BnaC06G0323800ZS*	*FT*	Protein FLOWERING LOCUS T	Wang et al., 2009
A02	1946991–2001373	*BnvaA0201946991, BnvaA0202001373, BnvaA0202001103*	4.53–41.57	0.07–1.03	E1, E3, E8	*BnaA02G0035100ZS*	*FLC*	MADS-box protein FLOWERING LOCUS C	Tadege et al., 2001
A09	24518838–24519761	*BnvaA0924518838, BnvaA0924519761*	9.29–27.99	0.24–0.65	E4, E3, E6				

[a]E1: multi-environments joint GWAS; E2: WH2013; E3: WH2014; E4: CS2013; E5: CS2014; E6: NJ2013; E7: NJ2014; E8: CQ2013; E9: EZ2013.

Validation of candidate genes

To validate the above flowering time candidate genes, we conducted selective sweep, haplotype, and co-expression analysis. First, by performing XP-CLR between SWORs and SORs, 954 selective sweeps were detected (Supplementary Table S10). Eleven candidate genes for flowering time were found in the selective sweep, e.g., *BnaFLCs*, *BnaFTs*, *BnaC02.FPGS1*, *BnaA08.SRC2*, *BnaA01.FSD1*, *BnaA02.VIN3*, and *BnaC09.PRR7*. Second, haplotype analyses were further conducted in these genes. For *BnaA02.FT*, *BnaA10.FLC*, *BnaA02.VIN3*, and *BnaC09.PRR7*, significant difference exists between each haplotype in different environments (Figures 4A–D; Supplementary Figures S4A–D). Interestingly, the haplotype for early flowering tends to exist in SORs, while the haplotype for late flowering prefers to exist in SWORs. Moreover, the co-expression networks of *BnaA02.VIN3*

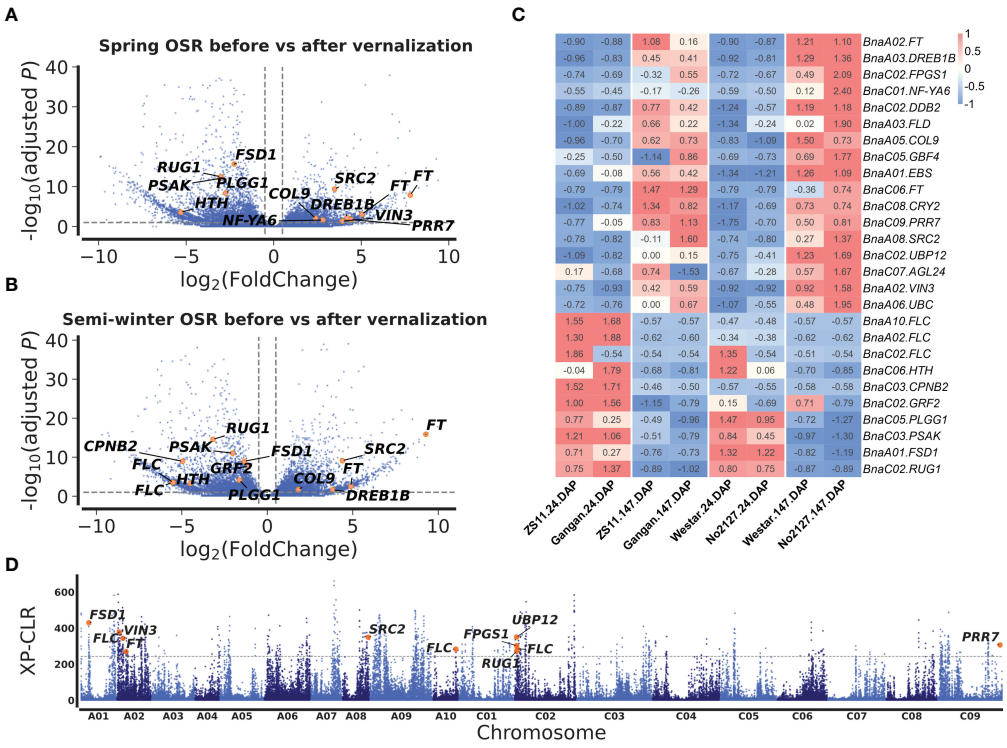


FIGURE 3
Differentially expressed gene (DEG) analysis before and after vernalization and selection sweeps between semi-winter and spring oilseed rapes (SWORs and SORs). Volcano plots of DEGs in SORs **(A)** and SWORs **(B)**. The y-axis is the adjusted *p*-value and the x-axis is $\log_2$ fold-change (FC) before and after vernalization. Gray lines are at the absolute value of $\log_2$FC = 1 or adjusted *p*-value = 0.05. **(C)** The expression profiling of 27 candidate genes around main-effect QTNs and QEIs for flowering time in two SWORs and two SORs in the 2018–2019 growing season in Wuhan. DAP, days after planting. **(D)** Selective sweeps between SWORs and SORs by XP-CLR. The horizontal dashed lines indicate the cutoff in the top 1% of XP-CLR scores. Candidate genes for flowering time are marked above the selective sweep peaks.

and *BnaC09.PRR7* have been constructed using DEGs in SWORs and SORs, respectively (Figures 5A, B). The co-expressed genes of *BnaA02.VIN3* mainly participated in the circadian clock, photoperiodism, light perception, and signaling. Eight genes are specific co-expressed in SORs, including *BnaA07.ZEP* and *BnaC09.ABCG22* in response to water deprivation. Five genes are specific co-expressed in SWORs. On the other hand, the co-expressed genes of *BnaC09.PRR7* mainly participated in the circadian clock and autonomous pathway. Five and one genes are specific co-expressed in SORs and SWORs, respectively. *BnaCKA2*s and *BnaPKDM7*s participated in epigenetic regulation.

Discussion

Although flowering time is strongly related to the environment, G by E studies for flowering time are lacking in *B. napus*. The current study analyzed the G by E for flowering time in the following three aspects. First, four windows of flowering-time-related climatic index were identified ($TAVG_{135–144}$, $PRCP_{3–41}$, $TAVG_{10–19}$, and $PRCP_{133–169}$) by CERIS. Second, 19 stable QTNs and 32 QEIs were found to be significantly associated with flowering time of 490 *B. napus* accessions in eight environments, and 10 QTNs were found to be associated with flowering-time-related climatic index. Finally, based on DEGs and homology with *Arabidopsis*, 12, 14, and 5 candidate genes were found to be associated with stable QTNs, QEIs, and QTNs for flowering-time-related climatic index, respectively. These candidate genes were further validated by the haplotype, selective sweep, and co-expression network analysis.

Flowering-time-related climatic index in *B. napus* whole growth stages

It is well-known that the flowering time regulation of *B. napus* is in response to day length or vernalization (Reeves and Coupland, 2000). This study calculated the correlations between two climatic factors, TAVG and PRCP, and flowering time in seven environments. TAVG correlated positively with flowering time in vernalization and negatively with flowering time after the seedling stage (Figure 1D). In a previous study, a reduction in

TABLE 2 Fourteen candidate genes for *B. napus* flowering time around significant QTN-by-environment interactions.

Genome-wide association studies					Comparative genomics analysis				Evidences for environmental interaction	
Chr	Pos (bp)	Marker	LOD	R^2 (%)	Gene ID	Abbr.	Function	Reference	Environment	Differences of flowering time under various environments
C09	67693730	*BnvaC0967693730*	67.17	1.07	*BnaC09G0614800ZS*	*PRR7*	Two-component response regulator-like APRR7	Nakamichi et al., 2007	Circadian clock	*prr7* single mutant is late flowering under LD conditions
C05	38307735	*BnvaC0538307735*	62.47	1.01	*BnaC05G0345200ZS*	*PLGG1*	Plastidal glycolate/ glycerate translocator			
C03	64354328	*BnvaC0364354328*	60.84	0.97	*BnaC03G0665500ZS*	*PSAK*	Photosystem I reaction center subunit K			
C02	1592445	*BnvaC0201592445*	58.92	0.94	*BnaC02G0022200ZS*	*FPGS1*	Folylpolyglutamate synthase			
A05	42071520	*BnvaA0542071520*	50.18	0.79	*BnaA05G0456200ZS*	*COL9*	Zinc finger protein CONSTANS-LIKE 9	Cheng and Wang, 2005	Circadian clock	*col9* single mutant is early flowering under LD conditions
A03	16879521	*BnvaA0316879521*	42.43	0.70	*BnaA03G0318500ZS*	*FLD*	FOLOWERING LOCUS D	Zhang et al., 2013b	Circadian clock	*fld* single mutant is late flowering under both SD and LD conditions
C02	9413473	*BnvaC0209413473*	34.92	0.55	*BnaC02G0132800ZS*	*DDB2*	Damaged DNA-binding proteins 2 required for UV-B tolerance	Al Khateeb and Schroeder, 2007	Light signaling	ddb2 suppressed the early flowering time of *det1* under long-day conditions
C03	27148748	*BnvaC0327148748*	30.92	0.48	*BnaC03G0400500ZS*	*CPNB2*	Chaperonin 60 subunit beta			
C02	30403413	*BnvaC0230403413*	28.66	0.45	*BnaC02G0311500ZS*	*GRF2*	G-box binding factor GF14 omega encoding a 14-3-3 protein	Liu et al., 2012	Unclear	*BnGRF2a* transgenic lines delays flowering
A08	27196633	*BnvaA0827196633*	24.73	0.38	*BnaA08G0296600ZS*	*SRC2*	Involved in Protein Storage Vacuole targeting.			
A01	7152286	*BnvaA0107152286*	17.37	0.27	*BnaA01G0121900ZS*	*EBS*	PHD finger family protein	López-González et al., 2014	Epigenetic regulation	*ebs* mutants repressed flowering
C01	50758439	*BnvaC0150758439*	16.63	0.27	*BnaC01G0442400ZS*	*NF-YA6*	Nuclear factor Y, subunit A6	Siriwardana et al., 2016	Photoperiod	*NF-YA* can be positive regulators of photoperiod dependent flowering
C02	1725158	*BnvaC0201725158*	12.62	0.19	*BnaC02G0024600ZS*	*UBP12*	Ubiquitin carboxyl-terminal hydrolase 12	Cui et al., 2013	Circadian clock	ubp12 single mutant is slightly early flowering under both SD and LD conditions
A03	26932605	*BnvaA0326932605*	11.96	0.18	*BnaA03G0486700ZS*	*DREB1B*	Dehydration-responsive element-binding protein 1B	Seo et al., 2009	Cold	Response to ABA treatment

autumn or winter chilling delays floral transition in *B. napus* (O'Neill et al., 2019). An elevated growth temperature is equally efficient in inducing the flowering of *Arabidopsis* (Balasubramanian et al., 2006). However, the transition or critical point of these two stages is unclear. For PRCP, this study reported the relationships between PRCP and flowering time in *B. napus* for the first time. Although the correlation coefficients are lower than TAVG, PRCP was found to be correlated negatively with flowering time in early development and positively later (Supplementary Figure S1). This result is consistent with a previous study in *Arabidopsis* that flowering time correlated negatively with fall and winter precipitations and positively with summer precipitation (Vidigal et al., 2016).

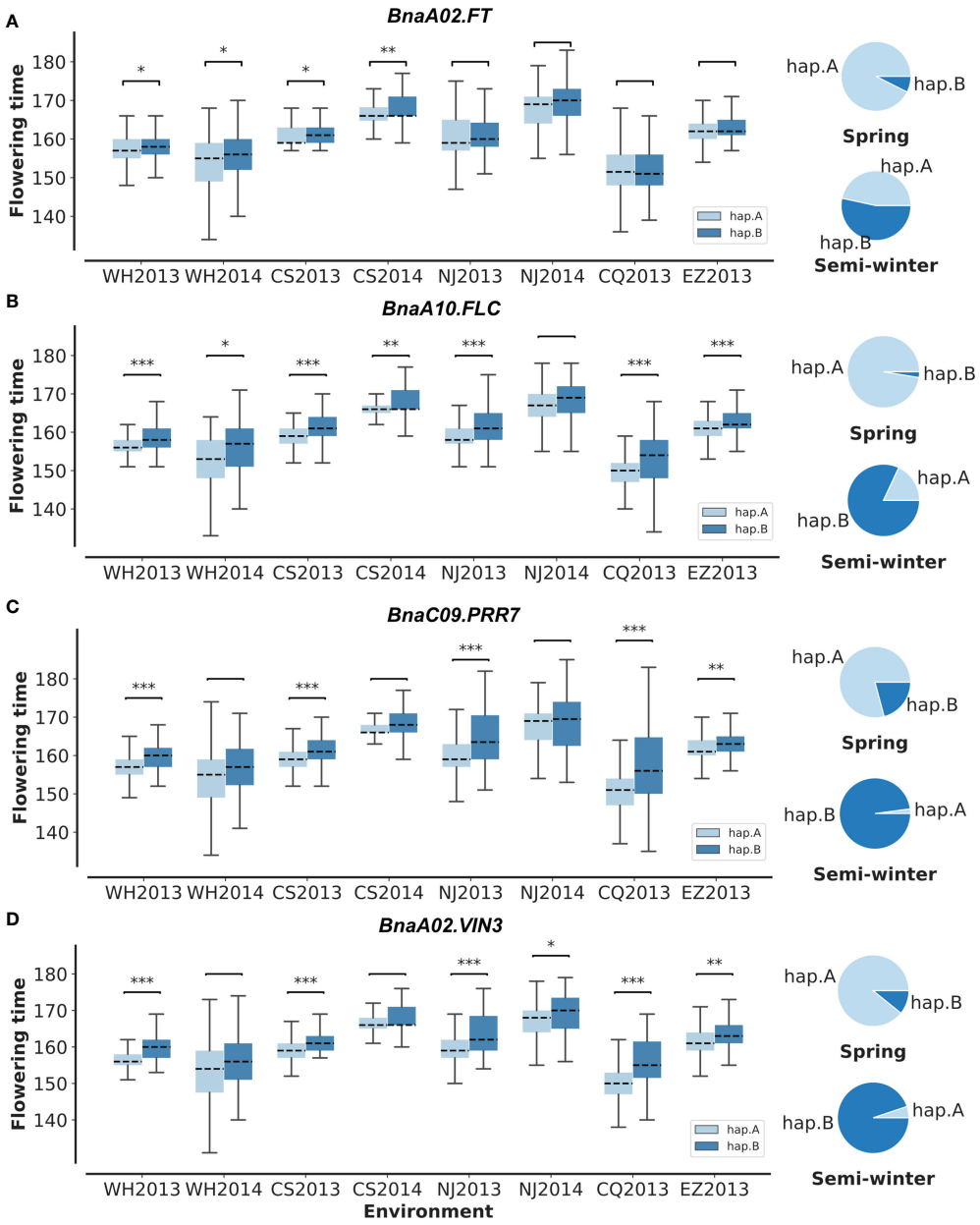


FIGURE 4
Haplotype analysis of *BnaC02.FT*, *BnaA10.FLC*, *BnaC09.PRR7*, and *BnaA02.VIN3* **(A–D)**. In the boxplot, significant differences for flowering time between each haplotype are calculated in eight environments with *t*-test. In pie plots, the haplotype frequencies of each gene in semi-winter and spring oilseed rapes are marked. WH2013, Wuhan in 2013; WH2014, Wuhan in 2014; CS2013, Changsha in 2013; CS2014, Changsha in 2014; NJ2013, Nanjing in 2013; NJ2014, Nanjing in 2014; CQ2013, Chongqing in 2013; EZ2013, Ezhou in 2013. *p = 0.05, **p = 0.01, and ***p = 0.001.

Genetic basis for flowering time in *B. napus*

In this study, multi-environment joint GWAS improved the power on identifying more QTNs than single environment GWAS. We dissected the genetic basis for flowering time in the following three aspects. First, 12 flowering time candidate genes were mined in approximately 19 stable QTNs for flowering time. Seven genes are previously reported, e.g., *BnaFLCs* (Tadege et al., 2001), *BnaFTs* (Wang et al., 2009), *BnaAGL24* (Yu et al., 2002), and *BnaCRY2* (Sharma et al., 2022), whereas five genes are newly identified, which are differentially expressed before and after the vernalization of different ecotypes (Figure 3; Supplementary Tables S8, S9). Second, it is worth

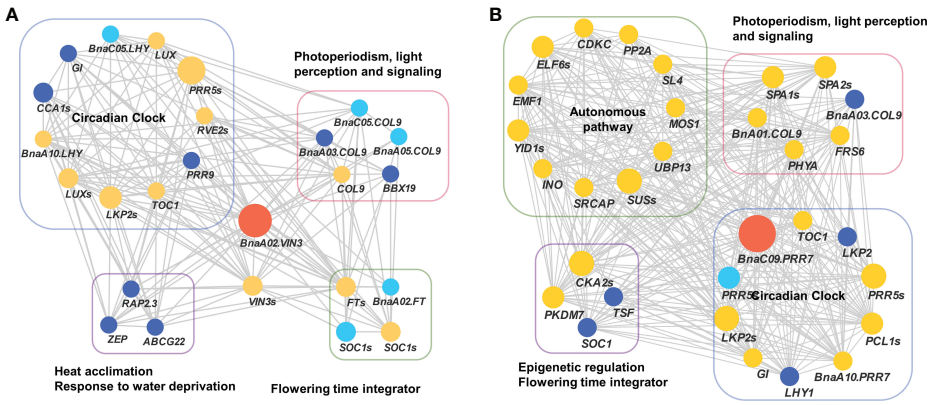


FIGURE 5
Co-expression network of *BnaVIN3* and *BnaPRR7* with co-expressed genes related to flowering time **(A, B)**. Light blue, dark blue, and yellow node indicate co-expressed genes that were detected in semi-winter, spring, and both types of oilseed rapes, respectively. The size of each node represents the number of genes in each gene family.

noting that this study focused on the mining of flowering time genes related to the environments. Fourteen candidate genes were identified around 32 QEIs, including 9 known flowering time genes related to environments. For example, *BnaCOL9* and *BnaUBP12* are regulated by the circadian clock in the photoperiod pathway (Cheng and Wang, 2005; Cui et al., 2013). *BnaFLD* is subjected to the direct regulation by brassinosteroids (Zhang et al., 2013b). It has been reported that the overexpression of *BnaDREB1B* not only delayed flowering but also responded to cold (Seo et al., 2009). *BnaEBS* functions in the chromatin-mediated repression of floral initiation by H3K4me3 (López-González et al., 2014). Finally, five genes were found to be associated with flowering-time-related climatic index. *BnaC02.DDB2* and *BnaA05.COL9* were commonly identified in QEIs, and *BnaA02.VIN3*, *BnaC02.RUG1*, and *BnaA06.UBC2* were commonly found to be associated with the main effect flowering time QTNs.

In this study, the missing heritability exists, in which the total phenotypic variance explained of QEIs and QTNs is much less than the estimated broad-sense heritability. This can be explained in several ways. First, the population is not enough to detect rare variants. Second, allelic heterogeneity may be the reason for this phenomenon. Lastly, epigenetic variation is likely to be a source of missing heritability (Brachi et al., 2011). Moreover, some candidate genes for stable QTNs, e.g., *BnaA02.FT* and *BnaA10.FLC*, were found to be related to environments but were not identified in QEIs (Figure 4). This result is explained by multiple facets, e.g., the difference in phenotypic data among environments, the diversity of population accessions, and the power of QEI detection. In the previous study, *COL9* and *FLD* have been reported to regulate *FT* and *FLC*, respectively (Cheng and Wang, 2005; Jiang et al., 2009). *BnA05.COL9* and *BnaA03.FLD* were found to be candidate genes for QEIs in this study. We hypothesized that QEI may be associated with direct environmental response upstream regulators due to the complexity of transcription and epigenetic regulations of flowering (Bouché et al., 2016).

BnaA02.VIN3 and *BnaC09.PRR7* are potential G by E genes for flowering time

In *Arabidopsis*, *VIN3* acts together with *PRC2* to repress histone marks at *FLC* in response to vernalization (Kim and Sung, 2013). *PRR7* was reported to coordinate with *PRR9* and *PRR5* and regulate flowering time through the canonical *CO*-dependent photoperiodic pathway (Nakamichi et al., 2007). In this study, *BnaA02.VIN3* and *BnaC09.PRR7* have been shown to be crucial G by E genes for flowering time. There are three pieces of evidence. First, *BnaA02.VIN3* is significantly associated with ChrA02-6152101 (LOD = 13.14) for flowering-time-related climatic factors and with ChrA02-6374324 (LOD = 12.17) for flowering time in WH2013. *BnaC09.PRR7* is significantly associated with the QEI, ChrC09-67693730 (LOD = 67.17), by multi-environment GWAS. Second, *BnaA02.VIN3* and *BnaC09.PRR7* are DEGs before and after vernalization and in the selective sweep between SORs and SWORs (Figure 2). Then, in these genes with significant haplotype differences, their haplotypes for early flowering tend to exist more in SORs (Figures 4C, D). Lastly, co-expression networks were constructed for *BnaA02.VIN3* and *BnaC09.PRR7*. Some relationships have been proven, e.g., *PRR7* with *LHY* (Liu et al., 2013), *PRR7* with *PRR5* (Nakamichi et al., 2007), and *VIN3* with *CCA1* and *LHY* (Kyung et al., 2022).

In summary, we dissected the G by E for flowering time for *B. napus* from different eco-types in eight environments. Four

windows of flowering-time-related climatic index were identified. Stable QTNs and QEIs for flowering time and their candidate genes were identified. These findings provide valuable information that can be used to breed *B. napus* varieties with optimized flowering time by pyramiding favorable alleles. The candidate genes will also greatly promote the dissection of flowering time mechanisms in different eco-types.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: NGDC, PRJCA012445 and CRA008501.

Author contributions

LG and XH conceived this study. QT, LX, ZG, JT, and BY performed the field experiments. XH performed the bioinformatics analysis and wrote the manuscript. JT, BY, KL, XY, SL, and LG revised the manuscript. All authors approved the submitted version.

Funding

This work was supported by grants from the National Natural Science Foundation of China (U2102217), Key Research and Development Program of Hubei (2021ABA011) and Higher Education Discipline Innovation Project (B20051).

Acknowledgments

We would like to thank Prof. Yuan-Ming Zhang (College of Plant Science and Technology, Huazhong Agricultural University, Wuhan) for improving the language within the manuscript.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1065766/full#supplementary-material

References

Al Khateeb, W. M., and Schroeder, D. F. (2007). DDB2, DDB1A and DET1 exhibit complex interactions during arabidopsis development. *Genetics* 176, 231–242. doi: 10.1534/genetics.107.070359

An, H., Qi, X., Gaynor, M. L., Hao, Y., Gebken, S. C., Mabry, M. E., et al. (2019). Transcriptome and organellar sequencing highlights the complex origin and diversification of allotetraploid *Brassica napus*. *Nat. Commun.* 10, 2878. doi: 10.1038/s41467-019-10757-1

Arnold, P. A., Kruuk, L. E. B., and Nicotra, A. B. (2019). How to analyse plant phenotypic plasticity in response to a changing climate. *New Phytol.* 222, 1235–1241. doi: 10.1111/nph.15656

Balasubramanian, S., Sureshkumar, S., Lempe, J., and Weigel, D. (2006). Potent induction of *Arabidopsis thaliana* flowering by elevated growth temperature. *PloS Genet.* 2, e106. doi: 10.1371/journal.pgen.0020106

Bastow, R., Mylne, J. S., Lister, C., Lippman, Z., Martienssen, R. A., and Dean, C. (2004). Vernalization requires epigenetic silencing of *FLC* by histone methylation. *Nature* 427, 164–167. doi: 10.1038/nature02269

Bates, D., Mächler, M., Bolker, B. M., and Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Software* 67, 1–48. doi: 10.18637/jss.v067.i01

Bouché, F., Lobet, G., Tocquin, P., and Périlleux, C. (2016). FLOR-ID: an interactive database of flowering-time gene networks in *Arabidopsis thaliana*. *Nucleic Acids Res.* 44, D1167–D1171. doi: 10.1093/nar/gkv1054

Brachi, B., Morris, G. P., and Borevitz, J. O. (2011). Genome-wide association studies in plants: The missing heritability is in the field. *Genome Biol.* 12, 232. doi: 10.1186/gb-2011-12-10-232

Cheng, X. F., and Wang, Z. Y. (2005). Overexpression of *COL9*, a *CONSTANS-LIKE* gene, delays flowering by reducing expression of *CO* and *FT* in *Arabidopsis thaliana*. *Plant J.* 43, 758–768. doi: 10.1111/j.1365-313X.2005.02491.x

Chen, H., Patterson, N., and Reich, D. (2010). Population differentiation as a test for selective sweeps. *Genome Res.* 20, 393–402. doi: 10.1101/gr.100545.109

Corbesier, L., Vincent, C., Jang, S., Fornara, F., Fan, Q., Searle, I., et al. (2007). FT protein movement contributes to long-distance signaling in floral induction of arabidopsis. *Science* 316, 1030–1033. doi: 10.1126/science.1141752

Cui, X., Lu, F., Li, Y., Xue, Y., Kang, Y., Zhang, S., et al. (2013). Ubiquitin-specific proteases UBP12 and UBP13 act in circadian clock and photoperiodic flowering regulation in arabidopsis. *Plant Physiol.* 162, 897–906. doi: 10.1104/pp.112.213009

Cui, Y., Xu, Z., and Xu, Q. (2021). Elucidation of the relationship between yield and heading date using CRISPR/Cas9 system-induced mutation in the flowering pathway across a large latitudinal gradient. *Mol. Breed.* 41, 23. doi: 10.1007/s11032-021-01213-4

Del Olmo, I., Poza-Viejo, L., Piñeiro, M., Jarillo, J. A., and Crevillén, P. (2019). High ambient temperature leads to reduced *FT* expression and delayed flowering in *Brassica rapa via* a mechanism associated with H2A.Z dynamics. *Plant J.* 100, 343–356. doi: 10.1111/tpj.14446

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32, 3047–3048. doi: 10.1093/bioinformatics/btw354

Guo, Y., Hans, H., Christian, J., and Molina, C. (2014). Mutations in single *FT*- and *TFL1*-paralogs of rapeseed (*Brassica napus* l.) and their impact on flowering time and yield components. *Front. Plant Sci.* 5. doi: 10.3389/fpls.2014.00282

Guo, T., Mu, Q., Wang, J., Vanous, A. E., Onogi, A., Iwata, H., et al. (2020). Dynamic effects of interacting genes underlying rice flowering-time phenotypic plasticity and global adaptation. *Genome Res.* 30, 673–683. doi: 10.1101/gr.255703.119

Han, X., Xu, Z. R., Zhou, L., Han, C. Y., and Zhang, Y. M. (2021). Identification of QTNs and their candidate genes for flowering time and plant height in soybean using multi-locus genome-wide association studies. *Mol. Breed.* 41, 39. doi: 10.1007/s11032-021-01230-3

Helal, M. M. U., Gill, R. A., Tang, M., Yang, L., Hu, M., Yang, L., et al. (2021). SNP- and haplotype-based GWAS of flowering-related traits in *Brassica napus*. *Plants* 10, 2475. doi: 10.3390/plants10112475

Hu, J., Chen, B., Zhao, J., Zhang, F., Xie, T., Xu, K., et al. (2022). Genomic selection and genetic architecture of agronomic traits during modern rapeseed breeding. *Nat. Genet.* 54, 694–704. doi: 10.1038/s41588-022-01055-6

Jiang, D., Gu, X., and He, Y. (2009). Establishment of the winter-annual growth habit *via FRIGIDA*-mediated histone methylation at *FLOWERING LOCUS c* in arabidopsis. *Plant Cell* 21, 1733–1746. doi: 10.1105/tpc.109.067967

Kim, D. H., and Sung, S. (2013). Coordination of the vernalization response through a *VIN3* and *FLC* gene family regulatory network in arabidopsis. *Plant Cell* 25, 454–469. doi: 10.1105/tpc.112.104760

Knapp, S. J., Stroup, W. W., and Ross, W. M. (1985). Exact confidence intervals for heritability on a progeny mean basis. *Crop Sci.* 25, 192–194. doi: 10.2135/cropsci1985.0011183X002500010046x

Kusmec, A., Srinivasan, S., Nettleton, D., and Schnable, P. S. (2017). Distinct genetic architectures for phenotype means and plasticities in zea mays. *Nat. Plants* 3, 715–723. doi: 10.1038/s41477-017-0007-7

Kyung, J., Jeon, M., Jeong, G., Shin, Y., Seo, E., Yu, J., et al. (2022). The two clock proteins CCA1 and LHY activate *VIN3* transcription during vernalization through the vernalization-responsive cis-element. *Plant Cell* 34, 1020–1037. doi: 10.1093/plcell/koab304

Li, X., Guo, T., Wang, J., Bekele, W. A., Sukumaran, S., Vanous, A. E., et al. (2021). An integrated framework reinstating the environmental dimension for GWAS and genomic selection in crops. *Mol. Plant* 14, 874–887. doi: 10.1016/j.molp.2021.03.010

Li, Y., Huang, Y., Bergelson, J., Nordborg, M., and Borevitz, J. O. (2010). Association mapping of local climate-sensitive quantitative trait loci in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci.* 107, 21199–21204. doi: 10.1073/pnas.1007431107

Li, M., Zhang, Y. W., Xiang, Y., Liu, M. H., and Zhang, Y. M. (2022a). IIIVmrMLM: The r and c++ tools associated with 3VmrMLM, a comprehensive GWAS method for dissecting quantitative traits. *Mol. Plant* 15, 1251–1253. doi: 10.1016/j.molp.2022.06.002

Li, M., Zhang, Y. W., Zhang, Z. C., Xiang, Y., Liu, M. H., Zhou, Y. H., et al. (2022b). A compressed variance component mixed model for detecting QTNs, and QTN-by-environment and QTN-by-QTN interactions in genome-wide association studies. *Mol. Plant* 0. doi: 10.1016/j.molp.2022.02.012

Liu, T., Carlsson, J., Takeuchi, T., Newton, L., and Farré, 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

Liu, J., Hua, W., Yang, H. L., Zhan, G. M., Li, R. J., Deng, L. B., et al. (2012). The *BnGRF2* gene (*GRF2*-like gene from *Brassica napus*) enhances seed oil production through regulating cell number and plant photosynthesis. *J. Exp. Bot.* 63, 3727–3740. doi: 10.1093/jxb/ers066

Liu, J. Y., Zhang, Y. W., Han, X., Zuo, J. F., Zhang, Z., Shang, H., et al. (2020). An evolutionary population structure model reveals pleiotropic effects of *GmPDAT* for traits related to seed size and oil content in soybean. *J. Exp. Bot.* 71, 6988–7002. doi: 10.1093/jxb/eraa426

Liu, N., Du, Y., Warbuton, 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

López-González, L., Mouriz, A., Narro-Diego, L., Bustos, R., Martínez-Zapater, J. M., Jarillo, J. A., et al. (2014). Chromatin-dependent repression of the arabidopsis floral integrator genes involves plant specific PHD-containing proteins. *Plant Cell* 26, 3922–3938. doi: 10.1105/tpc.114.130781

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550. doi: 10.1186/s13059-014-0550-8

Matar, S., Kumar, A., Holtgräwe, D., Weisshaar, B., and Melzer, S. (2021). The transition to flowering in winter rapeseed during vernalization. *Plant Cell Environ.* 44, 506–518. doi: 10.1111/pce.13946

Mouradov, A., Cremer, F., and Coupland, G. (2002). Control of flowering time: Interacting pathways as a basis for diversity. *Plant Cell* 14, S111–S130. doi: 10.1105/tpc.001362

Nakamichi, N., Kita, M., Niinuma, K., Ito, S., Yamashino, T., Mizoguchi, T., et al. (2007). Arabidopsis clock-associated pseudo-response regulators *PRR9*, *PRR7* and *PRR5* coordinately and positively regulate flowering time through the canonical *CONSTANS*-dependent photoperiodic pathway. *Plant Cell Physiol.* 48, 822–832. doi: 10.1093/pcp/pcm056

O'Neill, C. M., Lu, X., Calderwood, A., Tudor, E. H., Robinson, P., Wells, R., et al. (2019). Vernalization and floral transition in autumn drive winter annual life history in oilseed rape. *Curr. Biol.* 29, 4300–4306.e2. doi: 10.1016/j.cub.2019.10.051

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14, 417–419. doi: 10.1038/nmeth.4197

Putterill, J., Laurie, R., and Macknight, R. (2004). It's time to flower: the genetic control of flowering time. *BioEssays* 26, 363–373. doi: 10.1002/bies.20021

Reeves, P. H., and Coupland, G. (2000). Response of plant development to environment: control of flowering by daylength and temperature. *Curr. Opin. Plant Biol.* 3, 37–42. doi: 10.1016/S1369-5266(99)00041-2

Schiessl, S. (2020). Regulation and subfunctionalization of flowering time genes in the allotetraploid oil crop *Brassica napus*. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.605155

Seo, E., Lee, H., Jeon, J., Park, H., Kim, J., Noh, Y. S., et al. (2009). Crosstalk between cold response and flowering in arabidopsis is mediated through the flowering-time gene *SOC1* and its upstream negative regulator *FLC*. *Plant Cell* 21, 3185–3197. doi: 10.1105/tpc.108.063883

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Sharma, P., Mishra, S., Burman, N., Chatterjee, M., Singh, S., Pradhan, A. K., et al. (2022). Characterization of *Cry2* genes (*CRY2a* and *CRY2b*) of *B. napus* and comparative analysis of *BnCRY1* and *BnCRY2a* in regulating seedling photomorphogenesis. *Plant Mol. Biol.* 110, 161–186. doi: 10.1007/s11103-022-01293-6

Shen, Y., Xiang, Y., Xu, E., Ge, X., and Li, Z. (2018). Major co-localized QTL for plant height, branch initiation height, stem diameter, and flowering time in an alien introgression derived *Brassica napus* DH population. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00390

Siriwardana, C. L., Gnesutta, N., Kumimoto, R. W., Jones, D. S., Myers, Z. A., Mantovani, R., et al. (2016). NUCLEAR FACTOR y, subunit a (NF-YA) proteins positively regulate flowering and act through *FLOWERING LOCUS t*. *PloS Genet.* 12, e1006496. doi: 10.1371/journal.pgen.1006496

Song, J., Guan, Z., Hu, J., Guo, C., Yang, Z., Wang, S., et al. (2020). Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nat. Plants* 6, 34–45. doi: 10.1038/s41477-019-0577-7

Song, J., Liu, D., Xie, W., Yang, Z., Guo, L., Liu, K., et al. (2021). BnPIR: Brassica napus pan-genome information resource for 1689 accessions. *Plant Biotechnol. J.* 19, 412–414. doi: 10.1111/pbi.13491

Tadege, M., Sheldon, C. C., Helliwell, C. A., Stoutjesdijk, P., Dennis, E. S., and Peacock, W. J. (2001). Control of flowering time by *FLC* orthologues in *Brassica napus*. *Plant J.* 28, 545–553. doi: 10.1046/j.1365-313X.2001.01182.x

Tang, S., Zhao, H., Lu, S., Yu, L., Zhang, G., Zhang, Y., et al. (2021). Genome- and transcriptome-wide association studies provide insights into the genetic basis of natural variation of seed oil content in *Brassica napus*. *Mol. Plant* 14, 470–487. doi: 10.1016/j.molp.2020.12.003

Tan, Z., Xie, Z., Dai, L., Zhang, Y., Zhao, H., Tang, S., et al. (2022). Genome- and transcriptome-wide association studies reveal the genetic basis and the breeding history of seed glucosinolate content in *Brassica napus*. *Plant Biotechnol. J.* 20, 211–225. doi: 10.1111/pbi.13707

Vidigal, D. S., Marques, A. C. S. S., Willems, L. A. J., Buijs, G., Méndez-Vigo, B., Hilhorst, H. W. M., et al. (2016). Altitudinal and climatic associations of seed dormancy and flowering traits evidence adaptation of annual life cycle timing in *Arabidopsis thaliana*. *Plant Cell Environ.* 39, 1737–1748. doi: 10.1111/pce.12734

Wang, J., Long, Y., Wu, B., Liu, J., Jiang, C., Shi, L., et al. (2009). The evolution of *Brassica napus FLOWERING LOCUS t* paralogues in the context of inverted chromosomal duplication blocks. *BMC Evol. Biol.* 9, 271. doi: 10.1186/1471-2148-9-271

Wu, R. (1998). The detection of plasticity genes in heterogeneous environments. *Evolution* 52, 967–977. doi: 10.1111/j.1558-5646.1998.tb01826.x

Xu, J., and Dai, H. (2016). *Brassica napus cycling dof Factor1* (*BnCDF1*) is involved in flowering time and freezing tolerance. *Plant Growth Regul.* 80, 315–322. doi: 10.1007/s10725-016-0168-9

Xu, L., Hu, K., Zhang, Z., Guan, C., Chen, S., Hua, W., et al. (2016). Genome-wide association study reveals the genetic architecture of flowering time in rapeseed (*Brassica napus* l.). *DNA Res.* 23, 43–52. doi: 10.1093/dnares/dsv035

Yang, C., Gan, Y., Harker, K. N., Kutcher, H. R., Gulden, R., Irvine, B., et al. (2014). Up to 32 % yield increase with optimized spatial patterns of canola plant establishment in western Canada. *Agron. Sustain. Dev.* 34, 793–801. doi: 10.1007/s13593-014-0218-5

Ying, L., Chen, H., and Cai, W. (2014). *BnNAC485* is involved in abiotic stress responses and flowering time in *Brassica napus*. *Plant Physiol. Biochem. PPB* 79, 77–87. doi: 10.1016/j.plaphy.2014.03.004

Yu, H., Xu, Y., Tan, E. L., and Kumar, P. P. (2002). *AGAMOUS-LIKE 24*, a dosage-dependent mediator of the flowering signals. *Proc. Natl. Acad. Sci. U. S. A.* 99, 16336–16341. doi: 10.1073/pnas.212624599

Zhang, H., Berger, J. D., and Milroy, S. P. (2013a). Genotype×environment interaction studies highlight the role of phenology in specific adaptation of canola (*Brassica napus*) to contrasting Mediterranean climates. *Field Crops Res.* 144, 77–88. doi: 10.1016/j.fcr.2013.01.006

Zhang, Y., Li, B., Xu, Y., Li, H., Li, S., Zhang, D., et al. (2013b). The cyclophilin *CYP20-2* modulates the conformation of *BRASSINAZOLE-RESISTANT1*, which binds the promoter of *FLOWERING LOCUS d* to regulate flowering in arabidopsis. *Plant Cell* 25, 2504–2521. doi: 10.1105/tpc.113.110296

Zhang, J., Yi, Q., Xing, F., Tang, C., Wang, L., Ye, W., et al. (2018). Rapid shifts of peak flowering phenology in 12 species under the effects of extreme climate events in Macao. *Sci. Rep.* 8, 13950. doi: 10.1038/s41598-018-32209-4

Zhao, H., Savin, K. W., Li, Y., Breen, E. J., Maharjan, P., Tibbits, J. F., et al. (2022). Genome-wide association studies dissect the G × E interaction for agronomic traits in a worldwide collection of safflowers (*Carthamus tinctorius* l.). *Mol. Breed.* 42, 24. doi: 10.1007/s11032-022-01295-8

Zhu, Y., Cao, Z., Xu, F., Huang, Y., Chen, M., Guo, W., et al. (2012). Analysis of gene expression profiles of two near-isogenic lines differing at a QTL region affecting oil content at high temperatures during seed maturation in oilseed rape (*Brassica napus* l.). *Theor. Appl. Genet.* 124, 515–531. doi: 10.1007/s00122-011-1725-2

Zou, X., Suppanz, I., Raman, H., Hou, J., Wang, J., Long, Y., et al. (2012). Comparative analysis of *FLC* homologues in brassicaceae provides insight into their role in the evolution of oilseed rape. *PloS One* 7, e45751. doi: 10.1371/journal.pone.0045751

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 02 February 2023
DOI 10.3389/fpls.2023.1119218

OPEN ACCESS

EDITED BY
Jianlong Xu,
Institute of Crop Sciences (CAAS), China

REVIEWED BY
Chenwu Xu,
Yangzhou University, China
Longbiao Guo,
China National Rice Research Institute (CAAS), China
Dali Zeng,
Zhejiang Agriculture and Forestry University, China

*CORRESPONDENCE
Ya-Wen Zhang
yawen@mail.hzau.edu.cn
Jiaming Mi
mjm@mail.hzau.edu.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 09 December 2022
ACCEPTED 13 January 2023
PUBLISHED 02 February 2023

CITATION
Zhao Q, Shi X-S, Wang T, Chen Y, Yang R, Mi J, Zhang Y-W and Zhang Y-M (2023) Identification of QTNs, QTN-by-environment interactions, and their candidate genes for grain size traits in main crop and ratoon rice. *Front. Plant Sci.* 14:1119218. doi: 10.3389/fpls.2023.1119218

Identification of QTNs, QTN-by-environment interactions, and their candidate genes for grain size traits in main crop and ratoon rice

Qiong Zhao[1†], Xiao-Shi Shi[1†], Tian Wang[1,2], Ying Chen[1], Rui Yang[1,2], Jiaming Mi[1,2*], Ya-Wen Zhang[1*] and Yuan-Ming Zhang[1]

[1]College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, China, [2]National Key Laboratory of Crop Genetic Improvement and National Centre of Plant Gene Research (Wuhan), Huazhong Agricultural University, Wuhan, China

Although grain size is an important quantitative trait affecting rice yield and quality, there are few studies on gene-by-environment interactions (GEIs) in genome-wide association studies, especially, in main crop (MC) and ratoon rice (RR). To address these issues, the phenotypes for grain width (GW), grain length (GL), and thousand grain weight (TGW) of 159 accessions of MC and RR in two environments were used to associate with 2,017,495 SNPs for detecting quantitative trait nucleotides (QTNs) and QTN-by-environment interactions (QEIs) using 3VmrMLM. As a result, 64, 71, 67, 72, 63, and 56 QTNs, and 0, 1, 2, 2, 2, and 1 QEIs were found to be significantly associated with GW in MC (GW-MC), GL-MC, TGW-MC, GW-RR, GL-RR, and TGW-RR, respectively. 3, 4, 7, 2, 2, and 4 genes were found to be truly associated with the above traits, respectively, while 2 genes around the above QEIs were found to be truly associated with GL-RR, and one of the two known genes was differentially expressed under two soil moisture conditions. 10, 7, 1, 8, 4, and 3 candidate genes were found by differential expression and GO annotation analysis to be around the QTNs for the above traits, respectively, in which 6, 3, 1, 2, 0, and 2 candidate genes were found to be significant in haplotype analysis. The gene *Os03g0737000* around one QEI for GL-MC was annotated as salt stress related gene and found to be differentially expressed in two cultivars with different grain sizes. Among all the candidate genes around the QTNs in this study, four were key, in which two were reported to be truly associated with seed development, and two (*Os02g0626100* for GL-MC and *Os02g0538000* for GW-MC) were new. Moreover, 1, 2, and 1 known genes, along with 8 additional candidate genes and 2 candidate GEIs, were found to be around QTNs and QEIs for GW, GL, and TGW, respectively in MC and RR joint analysis, in which 3 additional candidate genes were key and new. Our results provided a solid foundation for genetic improvement and molecular breeding in MC and RR.

KEYWORDS

rice, grain size, QTN, QTN-by-environment interaction, ratoon rice, 3VmrMLM

Introduction

Rice (*Oryza sativa* L.) is the principal food for more than half of the population in the world (Rosegrant and Cline, 2003). Effective panicle number per plant, grain number per panicle, and thousand-grain weight (TGW) are three main yield component factors (Xing and Zhang, 2010). Thus, increasing grain weight is an effective way to increase rice yield. TGW is mainly determined by grain size and grouting degree, in which the grain size is determined by grain length (GL), width (GW), and grain thickness (GT). These grain size-related traits are quantitative traits. In addition, grain size not only affects the rice yield but also affects its taste and appearance (Lou et al., 2009; Zhao et al., 2018). Therefore, it is necessary to investigate genetic mechanisms of GL, GW, and TGW.

With the completion of rice genome sequencing, more than 400 quantitative trait nucleotides (QTNs) for rice grain size in different genetic populations have been identified in previous studies (Huang et al., 2013). Among these loci, some of them have been fine-mapped, such as *gw9.1* (Xie et al., 2008), *qGL7* (Bai et al., 2010), *qGL3-2* (Liang et al., 2021), and *qGSN5* (Yuan et al., 2022a). At present, at least 22 QTLs/genes for grain size traits in rice have been cloned and functionally identified (Jiang et al., 2022), for example, *GW2* (Song et al., 2007), *GS2* (Duan et al., 2015), *GS5* (Li et al., 2011), *GS9* (Zhao et al., 2018), *GS3* (Fan et al., 2006), *GL3.1* (Qi et al., 2012), *GL3.3* (Xia et al., 2018), *qGL3* (Zhang et al., 2012), *qTGW2* (Ruan et al., 2020) and *qTGW3* (Hu et al., 2018) were mined by map-based cloning, while *GSE5* (Duan et al., 2017) and *OsSPL13* (Si et al., 2016) were detected by GWAS. Clearly, most were identified by map-based cloning, being a time-consuming work in developing near-isogenic lines. Moreover, it has been shown that the grain size is affected by environmental factors in many previous studies (Arshad et al., 2017; Bahuguna et al., 2017; Wu et al., 2022). However, few QTL-by-environment interactions (QEIs) have been identified in rice grain size. Although many QEIs have been detected in other rice traits in recent years, such as *qGT9* (Rahimsoroush et al., 2021), *qPC6*, *qPC7*, and *qGLU6* (Fiaz et al., 2021), they were identified by linkage analysis rather than genome-wide association studies (GWAS).

Ratoon rice has been considered as an efficient, green, and cost-saving rice cultivation mode, which has been popularized in many countries (Firouzi et al., 2018; Ziska et al., 2018; Wang et al., 2020). Compared with main crop, lower temperature after heading stage affects grain filling to reduce yield and improve quality of ratoon rice (Huang et al., 2020). However, QEIs for grain size between main crop and ratoon rice were rarely reported in previous studies, although main crop is used to identify QTNs and their candidate genes for grain size traits. More importantly, at present, most GWAS report only stable QTNs rather than QEIs, owing to the lack of feasible methodology of QEI detection in multiple environments (Kang et al., 2010; Zhang et al., 2010; Zhou and Stephens, 2012; Jiang et al., 2019b). To address this issue, Li et al. (2022a) and Li et al. (2022b) established a new compressed variance component mixed model method, namely 3VmrMLM, to identify QTNs, QEIs, and QTN-by-QTN interactions under controlling all the possible polygenic backgrounds.

To address the above issues, single environment analysis and two-environment joint analysis *via* 3VmrMLM (Li et al., 2022b) were used to identify QTNs and QEIs for GW, GL, and TGW in main crop (MC) and ratoon rice (RR) of 159 rice accessions with 2,017,495 SNPs. Previously reported genes around QTNs and QEIs for the three traits were mined and their candidate genes were predicted by comparative genomics and confirmed by gene haplotype analysis. In this study, we identified 202 QTNs and 3 QEIs in MC and 191 QTNs and 5 QEIs in RR, 18 previously reported genes around QTNs and two previously reported genes around QEIs were found to be truly associated with grain size in previous studies, and one of two genes around QEIs had the evidence of environmental interaction. Among 25 candidate genes identified by GO annotation and differential expression analysis, 12 were further confirmed by gene haplotype analysis, especially, four candidate genes and one candidate GEI for grain size are more important. In addition, the MC and RR datasets were jointly analyzed as well using 3VmrMLM, as a result, one, two, and one known genes were found to be around QTNs for GW, GL and TGW, respectively, 8 additional candidate genes and 2 candidate GEIs were also mined, in which 3 additional candidate genes are new and key in rice grain size related traits.

Material and methods

Plant materials and phenotyping of grain size related traits

All the 159 *indica* rice accessions were planted, with a randomized complete block design, in Wuhan in 2021. This experiment was replicated two times in different fields, namely environments 1 and 2. Each material was planted in one plot with 10 seedlings, row spacing was 16.7 cm × 20 cm, and one line empty between cells. At yellow ripening stage, GW (mm), GL (mm), and TGW (g) for each accession in MC and RR were measured for three times, and their averages were regarded as their trait phenotypes. GL in MC is abbreviated as GL-MC, and it is true for other traits.

Statistical analysis for the phenotypic data

The minimum, maximum, mean, standard deviation (SD), kurtosis, skewness (S_k), and coefficient of variation (CV), along with broad-sense heritability (H_B^2) , for all the above traits were calculated by R software lme4 v1.1.28. The H_B^2 for each trait was calculated by $H_B^2 = \frac{\sigma_g^2}{\sigma_g^2+\sigma_{ge}^2/l+\sigma_e^2/rl} \times 100\,\%$, where σ_g^2 is genetic variance, σ_e^2 is residual variance, σ_{ge}^2 is the variance of genotype-by-environment interaction, l is the number of environments, and r is the number of replicates. The analysis of variance (ANOVA) for phenotypic data was conducted using the R function *aov*. Normal distribution test for phenotypic data was conducted using the R function *shapiro.test*.

Genotyping data

The genotypic data of the 159 rice accessions used in this study consisted of two parts. The genotypic datasets of 134 accessions were derived from RiceVarMap database (http://ricevarmap.ncpgr.cn/), and the DNAs of leaves were extracted to conduct 1K Genobaits to verify their authenticity. The genotypic datasets of twenty-five modern breeding cultivars were obtained by double-terminal

sequencing with coverage of approximate 10× based on illumina's Hiseq 4000 technology sequencing platform at Novogene Technology Company. Then, extract the common SNPs from the genotype dataset of 134 public database accessions and 25 modern breeding cultivars to obtain new genotypic datasets with 2,019,008 SNPs. The software plink v1.90 was used to filter all the 2,019,008 SNPs based on minimum allele frequencies (MAFs) < 0.05 and all variants with missing call rates > 10%, where sliding window distance, step length, and R^2 were set as 1000 kb, 1, and 0.3, respectively. As a result, a total of 2,017,495 SNPs were used in subsequent GWAS.

Linkage disequilibrium decay and population structure

All the 2,017,495 SNPs were used to conduct linkage disequilibrium (LD) analysis using popLDdecay (https://github.com/BGIShenzhen/PopLDdecay). The LD decay was determined by plotting the r^2 values against the genetic distance of a pair of loci (kb) for each chromosome. G-matrix and cluster analysis for all the 159 accessions were performed using the 2,017,495 SNPs by R package sommer v4.2.0 and amap v0.8.19, respectively. Principal component analysis (PCA) was analyzed using R function *prcomp*, and the first two principal components were plotted using the R package ggplot2 v3.3.6. ADMIXTURE v1.3.0 (http://dalexander.github.io/admixture) was used to determine population structure (Alexander et al., 2009), where the number of subgroups (K) was set from 1 to 10, and the K value corresponding to the minimum CV error is the most likely subgroup number.

Multi-locus genome-wide association studies for grain size related traits

A total of 2,017,495 SNPs of 159 rice accessions were used to associate with GW, GL, and TGW in two environments in MC and RR using the 3VmrMLM method and its IIIVmrMLM software (https://github.com/YuanmingZhang65/IIIVmrMLM; Li et al., 2022a; Li et al., 2022b). All parameters were set as default values. Population structure adopts the first three principal components. The K matrix was calculated using the IIIVmrMLM software. The probability threshold was set at $0.05/m$ = 2.48e-08 for significant QTNs and QEIs, where m was the number of markers. To reduce the loss of important candidate genes, some insignificant QTNs and QEIs with LOD score ≥ 3.0 were regarded as suggested QTNs and QEIs (Li et al., 2022a; Li et al., 2022b).

Identification of candidate genes for grain size related traits in rice

Candidate genes for grain size traits were mined based on the below steps. First, all the genes were found in the 200 kb regions of upstream and downstream around each significant QTN without previously reported gene, because the LD decay distance was 150 kb using popLDdecay. Then, RNA-seq datasets of Zhenshan 97 and Minghui 63 at endosperm 7, 14, and 21 days after pollination (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19024) were used to conduct differential expression analysis using NCBI (https://www.ncbi.nlm.nih.gov) GEO2R online tool, and the thresholds of significant difference were set as p-value < 0.05 and | Log2FC | > 1. Finally, all the differentially expressed genes (DEGs) were further analyzed by GO annotation using AgBase (https://agbase.arizona.edu), and the significant E-value was set as 10e-50. If biological process is related to the reported molecular mechanisms of grain size, the DEGs in the biological process were regarded as candidate genes.

Haplotype analysis of candidate genes

The software plink v1.90 was used to extract all the significant SNP information (P < 0.05) after single marker genome scanning within one candidate gene and its upstream 2 kb, R v4.1.3 was used to calculate its haplotypes of the candidate gene, and the 159 rice accessions were grouped based on these haplotypes. Thus, ANOVA was performed using R function *aov* to test the significance of the QTN-associated trait across these haplotypes at a 5% probability level.

Result

Phenotypic variation

The averages plus standard deviations of GW-MC, GL-MC, TGW-MC, GW-RR, GL-RR, and TGW-RR in 159 rice accessions in two environments were 2.44 ± 0.33 ~ 2.47 ± 0.34 (mm), 8.41 ± 0.85 ~ 8.42 ± 0.84 (mm), 23.92 ± 3.01 ~ 24.03 ± 2.98 (g), 2.38 ± 0.29 ~ 2.43 ± 0.28 (mm), 8.00 ± 0.79 ~ 8.05 ± 0.81 (mm), 21.74 ± 2.84 ~ 22.32 ± 3.09 (g), and their coefficients of variation (CV) were 13.57 ~ 13.70, 9.91 ~ 10.12, 12.40 ~ 12.58, 11.50 ~ 12.17, 9.91 ~ 10.08, and 13.04 ~ 13.83 (%), respectively, having large phenotypic variations (Supplementary Table S1). The analysis of variance was conducted and the results were listed in Supplementary Table S2. As a result, genotypes, environments, and their interactions for all the three traits in MC and RR were significant at the 0.05 probability level (Supplementary Table S2), and the H_B^2 of GW, GL, and TGW ranged from 96.39% to 99.07% in MC and from 90.21% to 98.37% in RR, indicating large genetic variations (Supplementary Table S1). In addition, main crop had higher trait averages than ratoon rice, especially for GL and TGW (Figure 1). The phenotypes of TGW-MC and GW-RR in two environments, GL-RR in environment 2, and TGW-RR in environment 1 were found to obey normal distribution, while GW-MC and GL-MC in two environments, and GL-RR and TGW-RR in environment 1 were found to approximately obey normal distribution (Figure 1; Supplementary Table S1).

Population structure and linkage disequilibrium analysis

To determine the LD decay distance, LD decay analysis was performed using all the 2,017,495 SNP markers. The r^2 gradually decreases with the increase of distance. When it drops to half of the maximum value, the corresponding distance is regarded as the average distance of LD decay. In this study the LD decay distance was 150 kb, when r^2 dropped to half of its maximum value (r^2 = 0.3) (Figure 2C).

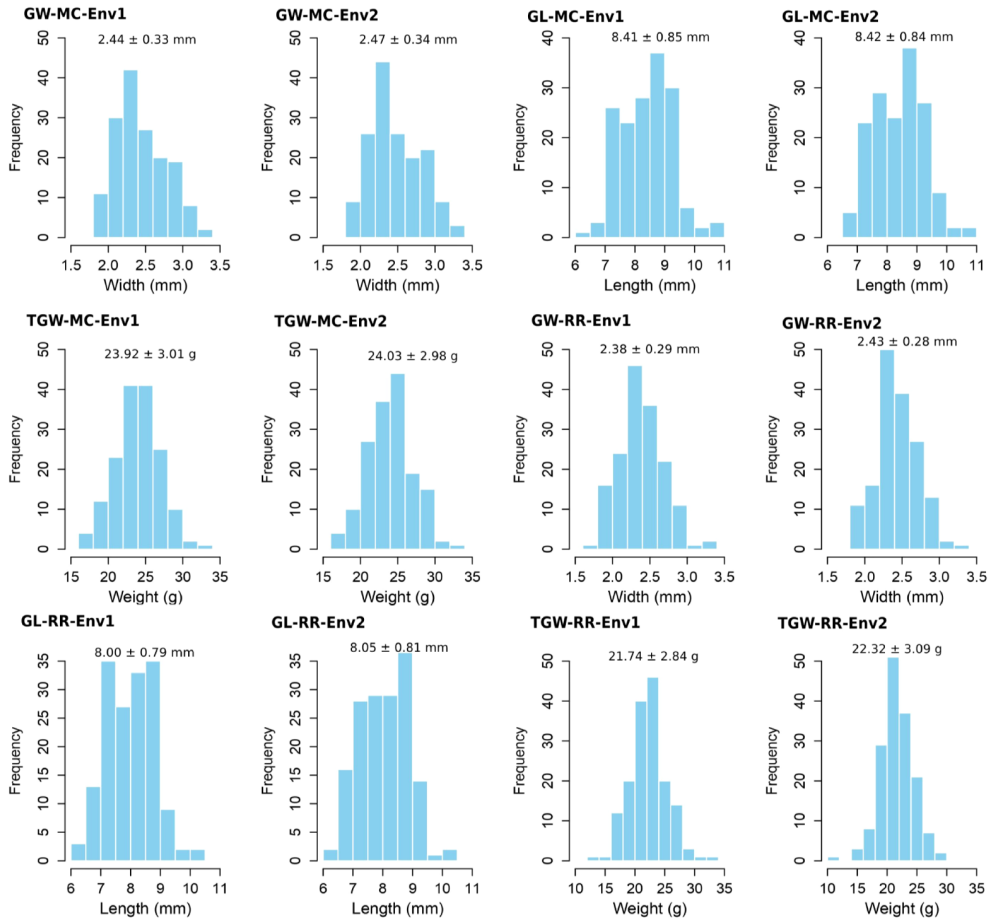


FIGURE 1
Phenotypic distributions for grain length (GL), grain width (GW), and thousand grain weight (TGW) of 159 accessions of main crop and ratoon rice in two environments.

The number of sub-populations was determined by principal component analysis (PCA), population structure analysis, and cluster analysis. The results were showed in Figure 2. In PCA, the first two principal components separated all the 159 accessions into three subgroups: indica I, indica II, and indicia Intermediate (Figure 2A). In population structure analysis *via* the ADMIXTURE software, cross-validation (CV) error is the lowest when the number of subgroups is three (Figures 2D, E), which is consistent with that in cluster analysis (Figure 2B). Thus, the first three principal components were used in genome-wide association studies.

Identification of QTNs and QEIs in main crop and ratoon rice

Identification of QTNs and QEIs when two environments in MC or RR were separately and jointly analyzed *via* 3VmrMLM

The 3VmrMLM method, implemented by its IIIVmrMLM software, was used to identify QTNs and QEIs for the three traits in this study. As a result, we identified 64, 71, and 67 QTNs for GW, GL and TGW in main crop, respectively, and 72, 63, and 56 QTNs for GW, GL, and TGW in ratoon rice, respectively (Supplementary Tables S3–S20). Among these QTNs for the above three traits in MC, there were 18, 17, and 13 significant QTNs and 2, 1, and 3 suggested QTNs in environment 1, there were 10, 18, and 17 significant QTNs and 2, 2, and 1 suggested QTNs in environment 2 (Supplementary Tables S3–S5, S9–S11), and there were 27, 32, and 27 significant QTNs and 5, 1, and 3 suggested QTNs detected in multi-environment joint analysis (Supplementary Tables S15–S17). In ratoon rice, there were 16, 13, and 16 significant QTNs and 4, 3, and 2 suggested QTNs in environment 1, there were 20, 13, and 12 significant QTNs and 0, 3, and 4 suggested QTNs in environment 2 (Supplementary Tables S6–S8, S12–S14), and there were 29, 29, and 20 significant QTNs and 3, 2, and 2 suggested QTNs in multi-environment joint analysis (Supplementary Tables S18–S20). More importantly, one GL and two TGW QEIs were detected in main crop, and two GW, two GL, and one TGW QEIs were detected in ratoon rice (Supplementary Tables S15–S20).

Identification of QTNs and QEIs when the MC and RR datasets were jointly analyzed in each environment *via* 3VmrMLM

The MC and RR datasets in each environment were jointly analyzed using the IIIVmrMLM software. As a result, 34, 35, and 42 QTNs and 7, 1, and 14 QEIs were identified for GW, GL, and TGW, respectively (Supplementary Tables S29–S34). Among these QTNs and QEIs, there were 32, 30, and 37 significant QTNs, 2, 5, and

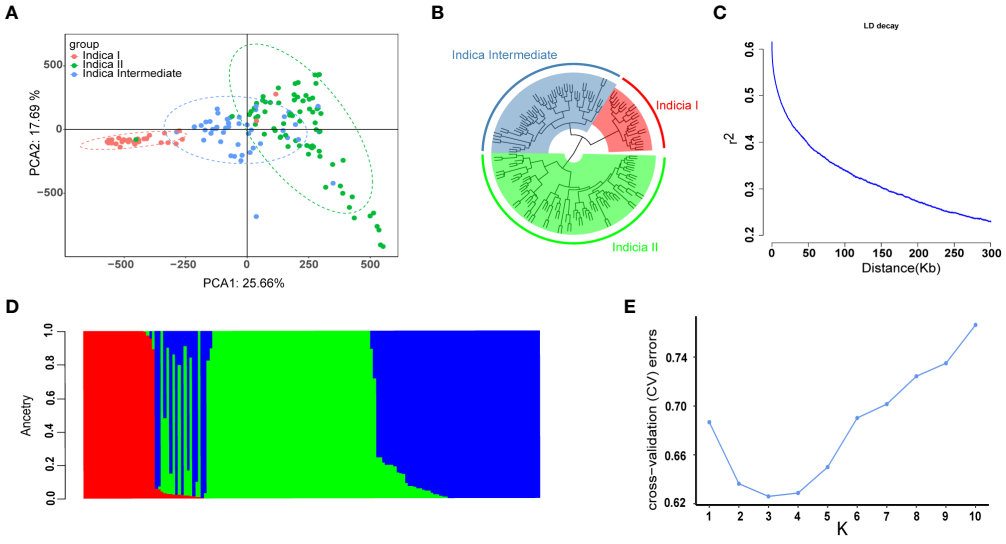


FIGURE 2
Population structure and LD decay of 159 rice accessions. **(A)** Principal component analysis (PCA) of the association panel. **(B)** Cluster analysis results of 159 rice accessions with 2,017,495 SNPs. **(C)** The entire genome LD decay of the population. **(D)** Population structure estimates (K = 3), the areas of the three colors illustrate the proportion of each subgroup. **(E)** cross-validation (CV) error line graph of subgroups (K = 3).

5 suggested QTNs, 4, 0, and 10 significant QEIs, and 3, 1, and 4 suggested QEIs for GW, GL, and TGW, respectively.

Known genes around QTNs and QEIs

Known genes were searched within 200 kb upstream and downstream regions of QTNs and QEIs. Among the QTNs and QEIs, 3, 4, and 7 known genes were found in main crop to be truly associated with the above three traits, respectively, and 2, 4, and 4 known genes were found in ratoon rice to be truly associated with the above three traits, respectively. Among these known genes, 4 were simultaneously found in main crop and ratoon rice, and 3 were found across multiple traits. 3, 5, and 9 known genes were found to be around significant QTNs and QEIs for the above three traits, respectively, and 0, 1, and 2 known genes were found to be around suggested QTNs for the above three traits, respectively (Tables 1, 2).

Around the above QTNs, some known genes were simultaneously mined in single-environment analysis and two-environment joint analysis. *GW5* was identified to be associated with GW-MC and GW-RR in two single-environment analyses and two-environment joint analysis, *VLN2* was identified to be associated with GW-MC in two-environment joint analysis and GW-RR in the first environment analysis (Figures 3A–C; Supplementary Figure 1), *GS3* was found to be associated with GL-MC and GL-RR in two single-environment analysis and two-environment joint analysis, and *GW5* was found to be associated with GL-MC in the second environment analysis and two-environment joint analysis (Supplementary Figures 2A–C, 3). For TGW, all known genes were separately detected in a single-environment analysis or two-environment joint analysis (Supplementary Figures 4A–C, 5).

Around the above QEIs, two known genes, *OsACOT* and *GW6a*, for GL-RR were mined (Table 2). Among the two known genes, *OsACOT* was found to be interacted with environments. In detail, its expression level under moderate soil drying treatment was higher than that under well-watered control (Teng et al., 2022) (Table 2).

In the joint analysis of the MC and RR datasets, 1, 2, and 1 known genes were found to be around significant QTNs and to be truly associated with GW, GL, and TGW, respectively (Supplementary Tables S35; Figures 3D, E; Supplementary Figures 2, 4D, E). Among these known genes, most of them were consistent with the above known genes, such as *GW5, GS3*, and *qTGW3*, but *PGL2* was found only in the MC and RR joint analysis.

Prediction of candidate genes

Around other QTNs without known genes, all the genes within 200 kb upstream and downstream regions were used to conduct differential expression analysis. All the differential expression genes (DEGs) were used to conduct gene annotation analysis. In gene annotation analysis, the significant biological processes were mainly included the below categories: cytokinin, abscisic acid and other plant hormone metabolism (e.g., *Os02g0197600*, *Os02g0621300*, *Os02g0626100*, and *Os02g0178800*), protein ubiquitination (*Os07g0166800*), sucrose starch metabolism (*Os04g0169100*, *Os12g0112500*, and *Os08g0205900*), protein phosphorylation (*Os02g0126400*, and *Os03g0717700*), and endosperm development (*Os02g0538000*, *Os12g0277500*, and *Os01g0280500*) (Supplementary Tables S21, S22), which are highly consistent with the previously reported regulatory pathways in Zuo and Li (2014); Cai et al. (2018), and Li et al. (2019). These genes were regarded as candidate genes. As a result, there were 10, 7 and 1 candidate genes for GW, GL and TGW in main crop, respectively, and 8, 4, and 3 candidate genes for GW, GL and TGW in ratoon rice, respectively (Supplementary Tables S21, S22).

For candidate genes for GW, *Os02g0126400* and *Os03g0717700* were predicted to be related to protein phosphorylation,

TABLE 1 Known genes around QTNs for grain length (GL), grain width (GW), and thousand grain weight (TGW) in main crop (MC) and ratoon rice (RR).

Trait	MC/RR	No.	Chr	Posi (bp)	LOD scores of QTN detection in two environments			r^2 (%)	Significance	Comparative genomics analysis		Reference
					I	II	I + II			Known genes	Distance (kb)	
GW	Both	1	3	13768754~13863861	14.58		34.25	0.92~1.57	Significant	*VLN2*	8.379~75.176	Wu et al., 2015
	Both	2	5	5357438~5361276	28.87~32.99	21.43~44.21	55.83~93.69	3.75~17.29	Significant	*GW5*	3.846~7.684	Liu et al., 2017
	MC	3	7	24771358		18.09		1.96	Significant	*GW7*	102.037	Wang et al., 2015a
GL	Both	1	3	16708508~16845802	32.72~40.25	47.94	22.1~36.57	2.02~13.09	Significant	*GS3*	11.033~110.693	Mao et al., 2010
	MC	2	3	35504491		5.28		0.68	Suggested	*qTGW3*	112.509	Ying et al., 2018
	MC	3	5	5357676~5456085		10.46	33.37	0.73~1.41	Significant	*GW5*	7.446~89.384	Liu et al., 2017
	Both	4	7	24533051~24800887	10.73		12.68	0.15~0.88	Significant	*GW7*	131.277~136.719	Wang et al., 2015a
TGW	MC	1	1	800544		5.97		0.88	Significant	*SPL33*	129.340	Wang et al., 2017
	MC	2	2	8196020		13.19		2.24	Significant	*GW2*	74.369	Song et al., 2007
	RR	3	2	26049877	17.19			2.10	Significant	*OsVPE3*	148.959	Lu et al., 2016
	RR	4	2	28749717	15.39			1.74	Significant	*GS2*	113.557	Hu et al., 2015
	MC	5	3	35437797			11.11	0.43	Significant	*qTGW3*	45.815	Ying et al., 2018
	RR	6	4	4570606	24.24			4.63	Significant	*ETR2*	167.769	Wuriyanghan et al., 2009
	MC	7	5	5356835			31.77	3.19	Significant	*GW5*	8.287	Liu et al., 2017
	MC	8	6	1540336	4.83			1.91	Suggested	*SSG6*	89.442	Matsushima et al., 2016
	MC	9	7	7640833		16.51		2.47	Significant	*SSH1*	90.919	Jiang et al., 2019a
	RR	10	8	6110721	6.11			2.25	Suggested	*UAP1*	126.728	Wang et al., 2015b
	MC	11	8	25154283	13.02			3.19	Significant	*OsSPL14*	120.258	Jiao et al., 2010

I and II: QTN detection in environments I and II, respectively; I + II: joint analysis of datasets in environments I and II. The same is true for Table 3.

10.3389/fpls.2023.1119218

10.3389/fpls.2023.1119218

TABLE 2 Two known genes around QEIs for rice grain size traits in ratoon rice (RR) and the evidence of gene-by-environment interactions.

No.	Trait	QEI					Known gene	Evidence for environmental interaction genes			Reference
		Chr	Posi (bp)	LOD	r^2 (%)	Significance		Environment	Indicator	Difference of indicator under various environments	
1	GL-RR	4	20228091	15.7721	0.7362	Significant	*OsACOT*	Moderate soil drying	Expression level	The expression of *OsACOT* increased after MD treatment	Zhao et al., 2019; Teng et al., 2022
2	GL-RR	6	26752211	21.2258	1.0259	Significant	*GW6a*				Song et al., 2015

GW, grain width; GL, grain length; TGW, thousand grain weight.

Os02g0178800 and *Os03g0592500* were predicted to respond to abscisic acid, *Os02g0197600* was found to be related to cytokinin, *Os07g0166800* was found to be related to the process of protein ubiquitination, *Os02g0538000* and *Os12g0277500* were found to be associated with embryonic development at the end of seed dormancy, and *Os08g0205900* and *Os04g0169100* were predicted to be related to sucrose metabolism and starch synthesis metabolism, respectively.

For candidate genes for GL, *Os02g0197600*, *Os02g0621300*, *Os02g0626100, Os04g0514800*, and *Os03g0108600* were predicted to respond to cytokinin, abscisic acid, gibberellin, auxin, and ethylene, respectively, *Os01g0280500* and *Os02g0538000* were found to affect embryonic development, and *Os04g0169100* and *Os12g0112500* were predicted to be related to starch synthesis. In Wuriyanghan et al. (2009), *Os04g0169100* was reported to affect the ethylene sensitivity of seeds to substantially enhance TGW of mutant. Here there is one issue pending, that is, whether the TGW increase is caused by the GL increase.

For candidate genes for TGW, *Os02g0621300* was predicted to be related to response to abscisic acid, *Os03g0411500* was predicted to be related to photosynthesis, *Os03g0607400* was predicted to be related to positive regulation of unidimensional cell growth, and *Os05g0445900* was predicted to participate in DNA demethylation, which is consistent with that in Zhou et al. (2021), in detail, *Os05g0445900* encodes rice DNA glycosylase, and the mutation of *DNG701* can lead to embryo retardation or abortion of part seeds (Zhou et al., 2021).

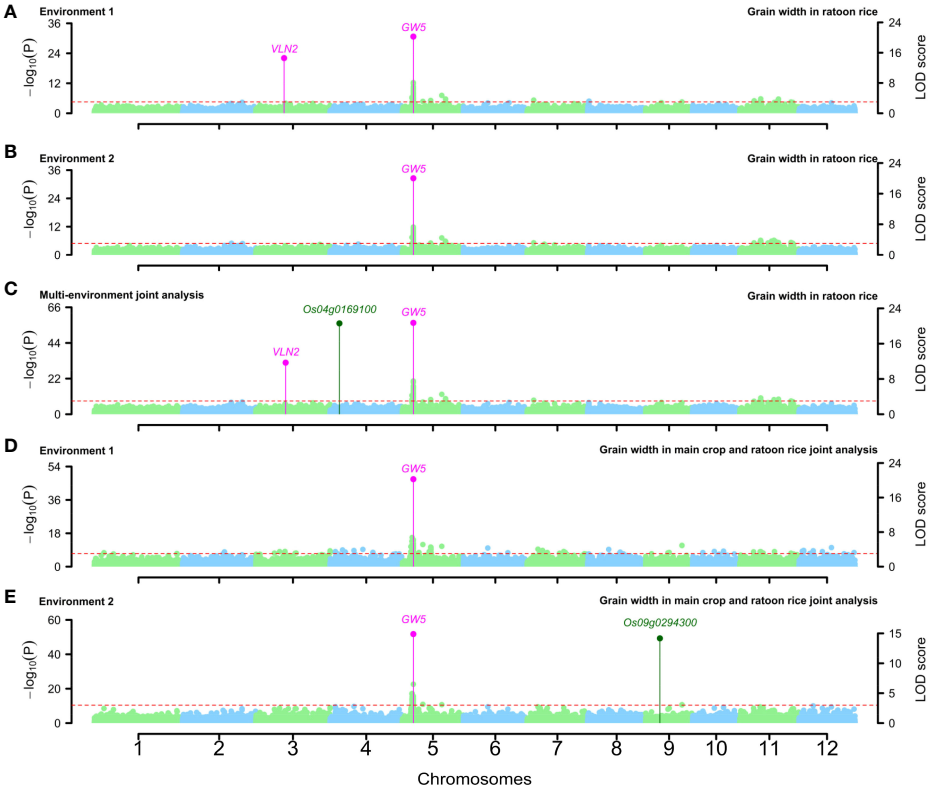


FIGURE 3
Manhattan plots for grain width in ratoon rice **(A–C)** and grain width in the joint analysis of main crop and ratoon rice **(D, E)**. Known genes around QTNs were marked with magenta color, and candidate genes around QTNs were marked with dark green color.

For the DEG *Os03g0737000* (P=7.77E-03, log2FC=-1.29) around the QEI of chr3-30340995 for GL-MC, *Os03g0737000* was predicted to be related to "response to salt stress" (Table 3). In the future, new experiments are necessary to explore these novel gene-trait and GEI-trait associations.

In the joint analysis of the MC and RR datasets, there were 22 candidate genes around QTNs to be responsible for the above three traits, but only two were consistent with the above 25 candidate genes (Supplementary Table S36). The significant biological processes of these candidate genes were mainly included the below categories: plant hormone metabolic pathway (e.g., *Os02g0126400*, *Os05g0563400*, and *Os12g0288000*), protein phosphorylation (*Os03g0838100, Os08g0200500*, and *Os05g0514200*), embryo development (*Os02g0538000* and *Os08g0428100*), and protein ubiquitination (*Os09g0294300* and *Os12g0111500*). In addition, we also mined two additional DEGs for TGW around QEIs, among which *Os06g0154200* was predicted to be related to "positive regulation of response to water deprivation" and *Os11g0600900* was predicted to be related to "response to light intensity" (Table 3).

Haplotype analysis

To further verify the reliability of candidate genes, we conducted haplotype analysis. As a result, 12 of the above 25 candidate genes had significant differences among the phenotypes of the traits corresponding to the haplotypes of each gene (Figure 4). Among the 12 significant candidate genes, 7, 3, and 3 were found to be associated with GW, GL, and TGW, respectively, of which there are 6, 3, and 1 significant candidate genes in main crop and 2, 0, and 2 significant candidate genes in ratoon rice (Figure 4). *Os08g0205900* for GW was mined in both MC and RR, and *Os02g0621300* was found in both GL and TGW (Figure 4). It should be noted that 8 of 22 candidate genes, which were mined in the MC and RR joint analysis, were significant in haplotype analysis, and the eight genes were different from the 12 significant candidate genes in the above haplotype analysis (Figure 4).

Discussion

To address the studies on gene-by-environmental interactions, especially, across main crop and ratoon rice, in this study we conducted genome-wide association studies for GW, GL, and TGW using 3VmrMLM. As a result, a total of 202 QTNs and 3 QEIs in main crop, and 191 QTNs and 5 QEIs in ratoon rice were identified. Around these QTNs and QEIs, 18 and 2 known genes were found to be truly associated with the grain size related traits, in which 4 were common across main crop and ratoon rice, and 12 candidate genes were mined through differential expression analysis, GO annotation, and haplotype analysis, in which one was common across main crop and ratoon rice. More importantly, four key candidate genes around QTNs were predicted, in which two were new and all identified in main crop. In addition, we identified a new candidate GEI *Os03g0737000*, which was predicted to be related "response to salt stress". In the joint analysis of the MC and RR datasets, furthermore, 8 additional candidate genes and two additional GEIs were minded, and 3 of 8 additional candidate genes were new and key for grain size related traits in this study.

Comparison of QTNs, QEIs, known genes, and candidate genes across main crop and ratoon rice

Ratoon rice is a new mode of rice planting, which can effectively save costs and increase benefits (Shen et al., 2021). Although it is generally accepted that RR has higher quality than MC, there are still many controversies in the studies on grain size related traits (Alizadeh and Habibi, 2016; Huang et al., 2020; Yuan et al., 2022b).

In this study, 64, 71, and 64 QTNs and 0, 1 and, 2 QEIs in MC, and 72, 63 and 56 QTNs and 2, 2, and 1 QEIs in RR were identified to be associated with GW, GL and TGW, respectively. Among these QTNs, 4 known genes were commonly detected in main crop and ratoon rice to be truly associated with grain size related traits, including *GW5* and *VLN2* for GW, and *GS3* and *GW7* for GL (Table 1). Some known genes were found only in main crop or ratoon rice, such as *qTGW3*, *SPL33*, and *OsSPL14* were detected only in main crop, and *OsVPE3*, *GS2*, *ETR2*, and *UPA1* were found only in the ratoon rice (Table 1). Among all the candidate genes, one was commonly found in main crop and ratoon rice, and 12 were detected only in main crop or ratoon rice (Figure 4). No common QEIs were detected between main crop and ratoon rice.

Based on the above results, main crop can detect more known genes (14), candidate genes (10) and candidate GEIs (1) than ratoon rice (8, 4, and 0). Although some known and candidate genes can be commonly found in main crop and ratoon rice, there are still some specific candidate genes in main crop or ratoon rice. In the independent and joint analyses of the MC and RR datasets, most candidate genes and candidate GEIs were different across the two analyses. This indicated that more known and candidate genes and GEIs can be identified while the datasets in main crop and ratoon rice are simultaneously or jointly analyzed.

Key candidate genes for GW, GL, and TGW in rice

The candidate genes were mined by expression and GO annotation analysis, and further validated through haplotype analysis. In this study we identified five new and key candidate genes that were predicted to be closely related to the three traits, among which 3 were mined to be around QTNs in the MC and RR joint analysis (Table 3), the evidence was as below.

Os02g0626100 for GL-MC, and *Os02g0538000* for GW-MC were differentially expressed. In GO annotation analysis, the two genes were annotated as "response to gibberellin", and "embryo development ending in seed dormancy", respectively, in which these biological processes are highly consistent with metabolic pathways of important grain size traits in rice (Li et al., 2018; Li et al., 2020; Jiang et al., 2022). In haplotype analysis, significant GL/GW differences were observed across 2 and 5 haplotypes from 1 and 11 significant SNPs within the two genes and their 2 kb upstream. Thus, the two genes may be important candidate genes for GL/GW.

TABLE 3 Key candidate genes and gene-by-environment interactions for grain size related traits in rice.

QTN/ GEI	No.	Trait	Locus		LOD scores			r^2 (%)	Gene differential expression analysis			P-value in haplotype analysis	GO annotation analysis			
			Chr	Posi (bp)	II	I + II	MC+RR		Gene_ID	log2 (Fold Change)	P-value		GO_ID	GO_name	E-value	Reference
QTN	1	GW-MC	2	20073320		11.87		0.47	*Os02g0538000*	1.23	6.47E-03	1.50E-06	GO:0009793	embryo development ending in seed dormancy	0	
QTN	2	GL-MC	2	24992114	13.56			0.75	*Os02g0626100*	-1.27	9.33E-03	7.30E-03	GO:0009739	response to gibberellin	0	
QTN	3	GW-RR	4	4591488		49.91		0.97	*Os04g0169100*	-1.20	2.76E-03	4.06E-15	GO:2000904	regulation of starch metabolic process	0	Wuriyanghan et al., 2009
QTN	4	TGW-MC	5	22017452		10.27		1.82	*Os05g0445900*	1.19	1.48E-02	1.32E-03	GO:0080111	DNA demethylation	0	Zhou et al., 2021
QTN	5	TGW	8	17687290			11.62	1.71	*Os08g0379300*	-1.18	2.72E-02	7.20E-03	GO:0005983	starch catabolic process	0	
QTN	6	GW	9	6986114			14.17	0.79	*Os09g0294300*	-1.6	3.46E-03	1.70E-05	GO:0016567	protein ubiquitination	2.71E-288	
QTN	7	GL	12	22906272			12.45	1.52	*Os12g0557800*	1.79	4.95E-03	1.30E-06	GO:0009737	response to abscisic acid	5.65E-216	
GEI	1	GL-MC	3	30340995		6.39		0.20	*Os03g0737000*	-1.29	7.77E-03		GO:0009651	response to salt stress	0	
GEI	2	TGW	6	2928548			13.19	2.91	*Os06g0154200*	-1.22	1.11E-02		GO:1902584	positive regulation of response to water deprivation	0	
GEI	3	TGW	11	22930659			5.21	1.08	*Os11g0600900*	-1.14	2.01E-02		GO:0009642	response to light intensity	0	

GW, grain width; GL, grain length; TGW, thousand grain weight; MC, main crop; RR, ratoon rice.

10.3389/fpls.2023.1119218

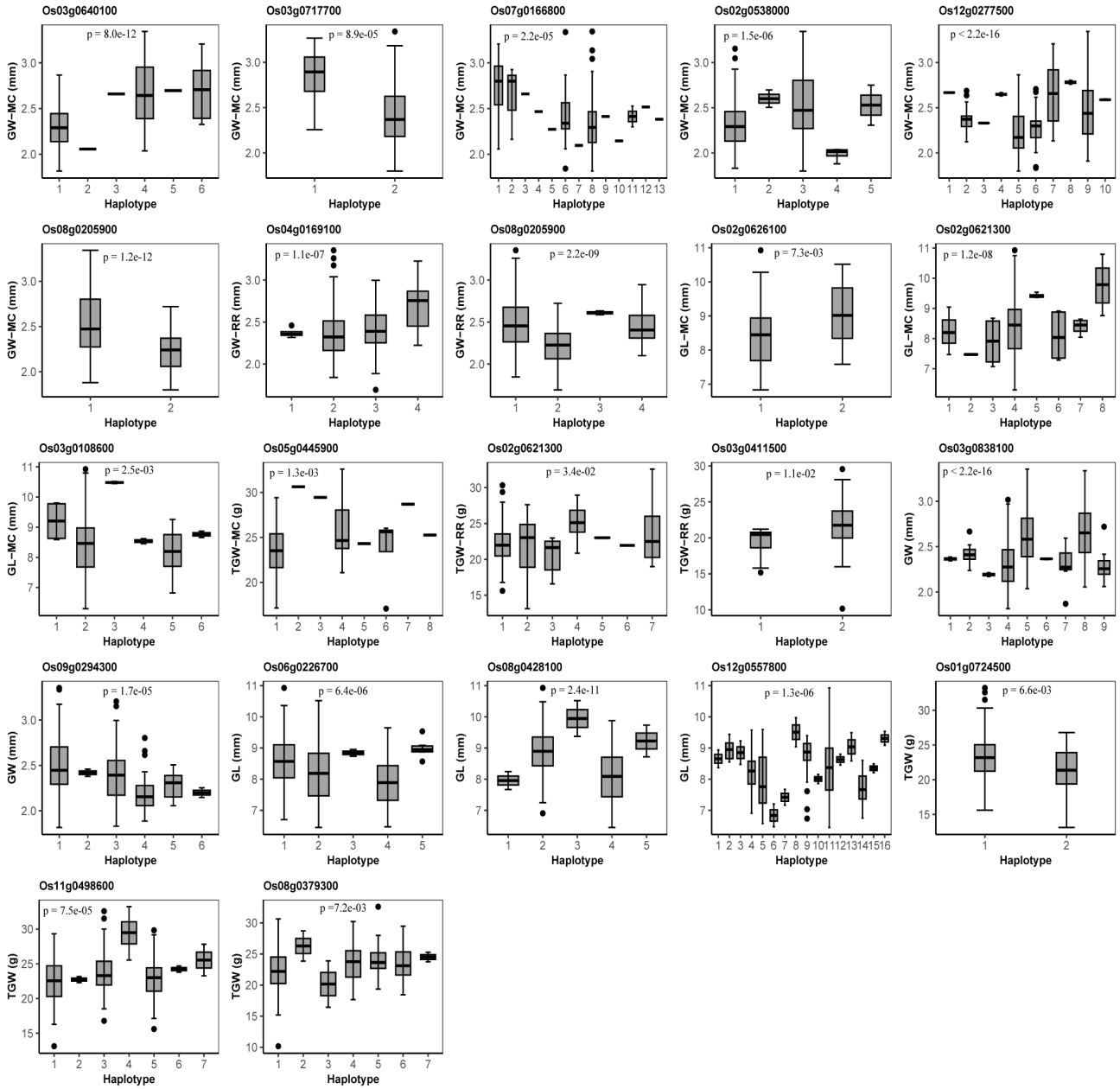


FIGURE 4
Haplotype analysis for candidate genes for grain width (GW), grain length (GL) and thousand grain weight (TGW) in main crop (MC), ratoon rice (RR), and the joint analyses of MC and RR. The P-values indicate the significance of trait averages across gene haplotypes for GW, GL, and TGW in one-way ANOVA.

In the same way, *Os08g0379300* for TGW, *Os09g0294300* for GW, and *Os12g0557800* for GL were found to be DEGs around the QTNs in the MC and RR joint analysis. In GO annotation analysis, the three gene were predicted to be related to "starch catabolic process", "protein ubiquitination", and "response to abscisic acid", respectively, which have been confirmed to be important regulatory pathways of rice grain size (Li et al., 2008; Choi et al., 2018; Gao et al., 2021). In haplotype analysis, significant TGW/GW/GL difference was observed across 7, 6, and 16 haplotypes from 6, 6, and 16 significant SNPs within the genes and their 2 kb upstream. Thus, the three genes may be important candidate genes for TGW/GW/GL.

In addition, two candidate genes have been reported to be related to rice seed development. In Wuriyanghan et al. (2009), *Os04g0169100*, identified for GW-RR in this study, significantly increased TGW of mutants by increasing the sensitivity of seeds to ethylene. In Zhou et al. (2021), the mutant of *Os05g0445900*, identified for TGW-MC in this study, participated in DNA methylation process causing the endosperm of some seeds to be stunted or aborted.

Identification of known and candidate GEIs for grain size traits in rice

Around the QEIs, *OsACOT* for GL-RR has been confirmed to be differentially expressed under two soil moisture treatments (Teng et al., 2022; Table 2).

Around QEIs in the independent analysis of MC or RR, *Os03g0737000* for GL-MC was found to be differentially expressed, and its biological process in GO annotation was predicted to be related to salt stress. We speculate that *Os03g0737000* may be affected by environmental factors, such as different salt treatments. Around QEIs detected in the MC and RR joint analysis, *Os06g0154200* and *Os11g0600900* for TGW were found to be differentially expressed, and the biological processes in their GO annotations were predicted to be related to water deprivation and light intensity, respectively. Thus, we speculate that *Os06g0154200* may be affected by the moisture content of the environment and *Os11g0600900* may be affected by the intensity of external light. The molecular functions of above three candidate GEIs need to be verified by subsequent molecular biology experiments.

Comparison of known genes across two types of interval lengths

To investigate the effect of interval length on mining known genes, two types of interval lengths were compared. One was 200 kb upstream and downstream regions of QTNs and QEIs, which was determined based on LD decay distance, while another was 1000 kb for QTNs and 1500 kb for QEIs. The results are listed in Table 1 and Supplementary Table S37. As a result, 3, 4, 7, 2, 2, and 4 known genes around QTNs and 0, 0, 0, 0, 2, and 0 known genes around QEIs were found to be located on their corresponding 200 kb upstream and downstream regions and to be truly associated with GW-MC, GL-MC, TGW-MC, GW-RR, GL-RR, and TGW-RR, respectively, while 6, 7, 21, 7, 6, and 16 known genes for the above six traits were found to be located on 1000 kb upstream and downstream regions of QTNs, and 0, 0, 1, 2, 2, and 0 known genes for the above six traits were found to be located on 1500 kb upstream and downstream regions of QEIs. This indicates that large intervals can find more known genes. Thus, it is very important to determine a suitable interval length in mining known genes.

Comparison of QTNs, QEIs, and known genes across various population structures

To investigate the effect of population structure on genome-wide association studies, we compared the results from evolutionary population structure (Liu et al., 2020), Q matrix, and PCA in this study. As a result, 323, 283, and 393 QTNs, 9, 6, 8 QEIs, and 11, 12, and 20 known genes were identified from evolutionary population, Q matrix and PCA, respectively (Supplementary Tables S23–S28). Clearly, the PCA result is the best, followed by evolutionary population, and the worst is the Q matrix result in this study. Thus, population structure is an important parameter in genome-wide association study.

Data availability statement

The genotypic datasets of 134 rice accessions can be downloaded from the NCBI under accession numbers PRJNA171289 and PRJEB6180, the phenotype values of grain size traits can be downloaded from figshare (https://doi.org/10.6084/m9.figshare.21957449.v1), RNA-seq data from endosperm tissue of Zhenshan 97 and Minghui 63 are available in the NCBI Gene Expression Omnibus (GEO) database under the accession number GSE19024, and further inquiries can be directed to JMM (mjm@mail.hzau.edu.cn).

Author contributions

Y-WZ and JMM managed the research. JMM was in charge of research experiments. TW and RY measured trait phenotypes and marker genotypes. QZ, X-SS, and YC analyzed datasets and mined candidate genes. QZ wrote the draft. Y-MZ, Y-WZ, and J-MM revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (32200500 and 32070557), the Hubei Key R&D Program (2021BBA225 and 2020BBA031), the Natural Science Foundation of Hubei Province (2022CFB780), and Postdoctoral Innovative Research Position of Hubei Province.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1119218/full#supplementary-material

References

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi: 10.1101/gr.094052.109

Alizadeh, M. R., and Habibi, F. (2016). A comparative study on the quality of the main and ratoon rice crops. *J. Appl. Bot. Food Qual.* 39, 669–674. doi: 10.1111/jfq.12250

Arshad, M. S., Farooq, M., Asch, F., Krishna, J. S. V., Prasad, P. V. V., and Siddique, K. H. M. (2017). Thermal stress impacts reproductive development and grain yield in rice. *Plant Physiol. Biochem.* 115, 57–72. doi: 10.1016/j.plaphy.2017.03.011

Bahuguna, R. N., Solis, C. A., Shi, W., and Jagadish, K. S. (2017). Post-flowering night respiration and altered sink activity account for high night temperature-induced grain

yield and quality loss in rice (*Oryza sativa* L.). *Physiol. Plant* 159, 59–73. doi: 10.1111/ppl.12485

Bai, X., Luo, L., Yan, W., Kovi, M. R., Zhan, W., and Xing, Y. (2010). Genetic dissection of rice grain shape using a recombinant inbred line population derived from two contrasting parents and fine mapping a pleiotropic quantitative trait locus *qGL7*. *BMC Genet.* 11, 16. doi: 10.1186/1471-2156-11-16

Cai, Y., Li, S., Jiao, G., Sheng, Z., Wu, Y., Shao, G., et al. (2018). *OsPK2* encodes a plastidic pyruvate kinase involved in rice endosperm starch synthesis, compound granule formation and grain filling. *Plant Biotechnol. J.* 16, 1878–1891. doi: 10.1111/pbi.12923

Choi, B. S., Kim, Y. J., Markkandan, K., Koo, Y. J., Song, J. T., and Seo, H. S. (2018). *GW2* functions as an E3 ubiquitin ligase for rice expansin-like 1. *Int. J. Mol. Sci.* 19, 1904. doi: 10.3390/ijms19071904

Duan, P., Ni, S., Wang, J., Zhang, B., Xu, R., Wang, Y., et al. (2015). Regulation of *OsGRF4* by *OsmiR396* controls grain size and yield in rice. *Nat. Plants* 2, 15203. doi: 10.1038/nplants.2015.203

Duan, P., Xu, J., Zeng, D., Zhang, B., Geng, M., Zhang, G., et al. (2017). Natural variation in the promoter of *GSE5* contributes to grain size diversity in rice. *Mol. Plant* 10, 685–694. doi: 10.1016/j.molp.2017.03.009

Fan, C., Xing, Y., Mao, H., Lu, T., Han, B., Xu, C., et al. (2006). *GS3*, a major QTL for grain length and weight and minor QTL for grain width and thickness in rice, encodes a putative transmembrane protein. *Theor. Appl. Genet.* 112, 1164–1171. doi: 10.1007/s00122-006-0218-1

Fiaz, S., Sheng, Z. H., Zeb, A., Barman, H. N., Shar, T., Ali, U., et al. (2021). Analysis of genomic regions for crude protein and fractions of protein using a recombinant inbred population in rice (*Oryza sativa* L.). *J. Taibah Univ. Sci.* 15, 579–588. doi: 10.1080/16583655.2021.1991733

Firouzi, S., Nikkhah, A., and Aminpanah, H. (2018). Resource use efficiency of rice production upon single cropping and ratooning agro-systems in terms of bioethanol feedstock production. *Energy* 150, 694–701. doi: 10.1016/j.energy.2018.02.155

Gao, Q., Zhang, N., Wang, W. Q., Shen, S. Y., Bai, C., and Song, X. J. (2021). The ubiquitin-interacting motif-type ubiquitin receptor HDR3 interacts with and stabilizes the histone acetyltransferase *GW6a* to control the grain size in rice. *Plant Cell* 33, 3331–3347. doi: 10.1093/plcell/koab194

Huang, R., Jiang, L., Zheng, J., Wang, T., Wang, H., Huang, Y., et al. (2013). Genetic bases of rice grain shape: So many genes, so little known. *Trends Plant Sci.* 18, 218–226. doi: 10.1016/j.tplants.2012.11.001

Huang, J., Pan, Y., Chen, H., Zhang, Z., Fang, C., Shao, C., et al. (2020). Physiochemical mechanisms involved in the improvement of grain-filling, rice quality mediated by related enzyme activities in the ratoon cultivation system. *Field Crops Res.* 258, 107962. doi: 10.1016/j.fcr.2020.107962

Hu, Z., Lu, S. J., Wang, M. J., He, H., Sun, L., Wang, H., et al. (2018). A novel QTL *qTGW3* encodes the GSK3/SHAGGY-like kinase OsGSK5/OsSK41 that interacts with OsARF4 to negatively regulate grain size and weight in rice. *Mol. Plant* 11, 736–749. doi: 10.1016/j.molp.2018.03.005

Hu, J., Wang, Y., Fang, Y., Zeng, L., Xu, J., Yu, H., et al. (2015). A rare allele of *GS2* enhances grain size and grain yield in rice. *Mol. Plant* 8, 1455–1465. doi: 10.1016/j.molp.2015.07.002

Jiang, L., Ma, X., Zhao, S., Tang, Y., Liu, F., Gu, P., et al. (2019a). The APETALA2-like transcription factor *SUPERNUMERARY BRACT* controls rice seed shattering and seed size. *Plant Cell* 31, 17–36. doi: 10.1105/tpc.18.00304

Jiang, H., Zhang, A., Liu, X., and Chen, J. (2022). Grain size associated genes and the molecular regulatory mechanism in rice. *Int. J. Mol. Sci.* 23, 3169. doi: 10.3390/ijms23063169

Jiang, L., Zheng, Z., Qi, T., Kemper, K. E., Wray, N. R., Visscher, P. M., et al. (2019b). A resource-efficient tool for mixed model association analysis of large-scale data. *Nat. Genet.* 51, 1749–1755. doi: 10.1038/s41588-019-0530-8

Jiao, Y., Wang, Y., Xue, D., Wang, J., Yan, M., Liu, G., et al. (2010). Regulation of *OsSPL14* by OsmiR156 defines ideal plant architecture in rice. *Nat. Genet.* 42, 541–544. doi: 10.1038/ng.591

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42, 348–354. doi: 10.1038/ng.548

Liang, P., Wang, H., Zhang, Q., Zhou, K., Li, M., Li, R., et al. (2021). Identification and pyramiding of QTLs for rice grain size based on short-wide grain CSSL-Z563 and fine-mapping of *qGL3-2*. *Rice* 14, 35. doi: 10.1186/s12284-021-00477-w

Li, Y., Fan, C., Xing, Y., Jiang, Y., Luo, L., Sun, L., et al. (2011). Natural variation in *GS5* plays an important role in regulating grain size and yield in rice. *Nat. Genet.* 43, 1266–1269. doi: 10.1038/ng.977

Li, N., Xu, R., Duan, P., and Li, Y. (2018). Control of grain size in rice. *Plant Reprod.* 31, 237–251. doi: 10.1007/s00497-018-0333-6

Li, N., Xu, R., and Li, Y. (2019). Molecular networks of seed size control in plants. *Annu. Rev. Plant Biol.* 70, 435–463. doi: 10.1146/annurev-arplant-050718-095851

Li, M., Zhang, Y. W., Xiang, Y., Liu, M. H., and Zhang, Y. M. (2022b). IIIVmrMLM: The r and c++ tools associated with 3VmrMLM, a comprehensive GWAS method for dissecting quantitative traits. *Mol. Plant* 15, 1251–1253. doi: 10.1016/j.molp.2022.06.002

Li, M., Zhang, Y. W., Zhang, Z. C., Xiang, Y., Liu, M. H., Zhou, Y. H., et al. (2022a). A compressed variance component mixed model for detecting QTNs and QTN-by-environment and QTN-by-QTN interactions in genome-wide association studies. *Mol. Plant* 15, 630–650. doi: 10.1016/j.molp.2022.02.012

Li, Y., Zheng, L., Corke, F., Smith, C., and Bevan, M. W. (2008). Control of final seed and organ size by the *DA1* gene family in *Arabidopsis thaliana*. *Gene Dev.* 22, 1331–1336. doi: 10.1101/gad.463608

Li, Q. F., Zhou, Y., Xiong, M., Ren, X. Y., Han, L., Wang, J. D., et al. (2020). Gibberellin recovers seed germination in rice with impaired brassinosteroid signalling. *Plant Sci.* 293, 110435. doi: 10.1016/j.plantsci.2020.110435

Liu, J., Chen, J., Zheng, X., Wu, F., Lin, Q., Heng, Y., et al. (2017). *GW5* acts in the brassinosteroid signalling pathway to regulate grain width and weight in rice. *Nat. Plants* 3, 17043. doi: 10.1038/nplants.2017.43

Liu, J. Y., Zhang, Y. W., Han, X., Zuo, J. F., Zhang, Z., Sang, H., et al. (2020). An evolutionary population structure model reveals pleiotropic effects of GmPDAT for traits related to seed size and oil content in soybean. *J. Exp. Bot.* 71, 6988–7002. doi: 10.1093/jxb/eraa426

Lou, J., Chen, L., Yue, G., Lou, Q., Mei, H., Xiong, L., et al. (2009). QTL mapping of grain quality traits in rice. *J. Cereal Sci.* 50, 145–151. doi: 10.1016/j.jcs.2009.04.005J

Lu, W., Deng, M., Guo, F., Wang, M., Zeng, Z., Han, N., et al. (2016). Suppression of *OsVPE3* enhances salt tolerance by attenuating vacuole rupture during programmed cell death and affects stomata development in rice. *Rice* 9, 65. doi: 10.1186/s12284-016-0138-x

Mao, H., Sun, S., Yao, J., Wang, C., Yu, S., Xu, C., et al. (2010). Linking differential domain functions of the *GS3* protein to natural variation of grain size in rice. *Proc. Natl. Acad. Sci. U. S. A.* 107, 19579–19584. doi: 10.1073/pnas.1014419107

Matsushima, R., Maekawa, M., Kusano, M., Tomita, K., Kondo, H., Nishimura, H., et al. (2016). Amyloplast membrane protein *SUBSTANDARD STARCH GRAIN6* controls starch grain size in rice endosperm. *Plant Physiol.* 170, 1445–1459. doi: 10.1104/pp.15.01811

Qi, P., Lin, Y. S., Song, X. J., Shen, J. B., Huang, W., Shan, J. X., et al. (2012). The novel quantitative trait locus *GL3.1* controls rice grain size and yield by regulating *Cyclin-T1;3*. *Cell Res.* 22, 1666–1680. doi: 10.1038/cr.2012.151

Rahimsoroush, H., Nazarian-Firouzabadi, F., and Chaloshtari, M. H. (2021). Identification of main and epistatic QTLs and QTL through environment interactions for eating and cooking quality in Iranian rice. *Euphytica* 217, 25. doi: 10.1007/s10681-020-02759-8

Rosegrant, M. W., and Cline, S. A. (2003). Global food security: Challenges and policies. *Science* 302, 1917–1919. doi: 10.1126/science.1092958

Ruan, B., Shang, L., Zhang, B., Hu, J., Wang, Y., Lin, H., et al. (2020). Natural variation in the promoter of *TGW2* determines grain width and weight in rice. *New Phytol.* 227, 629–640. doi: 10.1111/nph.16540

Shen, X., Zhang, L., and Zhang, J. (2021). Ratoon rice production in central China: Environmental sustainability and food production. *Sci. Total Environ.* 764, 142850. doi: 10.1016/j.scitotenv.2020.142850

Si, L., Chen, J., Huang, X., Gong, H., Luo, J., Hou, Q., et al. (2016). *OsSPL13* controls grain size in cultivated rice. *Nat. Genet.* 48, 447–456. doi: 10.1038/ng.3518

Song, X. J., Huang, W., Shi, M., Zhu, M. Z., and Lin, H. X. (2007). A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase. *Nat. Genet.* 39, 623–630. doi: 10.1038/ng2014

Song, X. J., Kuroha, T., Ayano, M., Furuta, T., Nagai, K., Komeda, N., et al. (2015). Rare allele of a previously unidentified histone H4 acetyltransferase enhances grain weight, yield, and plant biomass in rice. *Proc. Natl. Acad. Sci. U. S. A.* 112, 76–81. doi: 10.1073/pnas.1421127112

Teng, Z., Chen, Y., Yuan, Y., Peng, Y., Yi, Y., Yu, H., et al. (2022). Identification of microRNAs regulating grain filling of rice inferior spikelets in response to moderate soil drying post-anthesis. *Crop J.* 10, 962–971. doi: 10.1016/j.cj.2021.11.004

Wang, W. Q., He, A. B., Jiang, G. L., Sun, H. J., Jiang, M., Man, J. G., et al. (2020). Chapter four - ratoon rice technology: A green and resource-efficient way for rice production. *Adv. Agron.* 159, 135–167. doi: 10.1016/bs.agron.2019.07.006

Wang, S., Lei, C., Wang, J., Ma, J., Tang, S., Wang, C., et al. (2017). *SPL33*, encoding an eEF1A-like protein, negatively regulates cell death and defense responses in rice. *J. Exp. Bot.* 68, 899–913. doi: 10.1093/jxb/erx001

Wang, S., Li, S., Liu, Q., Wu, K., Zhang, J., Wang, S., et al. (2015a). The *OsSPL16-GW7* regulatory module determines grain shape and simultaneously improves rice yield and grain quality. *Nat. Genet.* 47, 949–954. doi: 10.1038/ng.3352

Wang, Z., Wang, Y., Hong, X., Hu, D., Liu, C., Yang, J., et al. (2015b). Functional inactivation of UDP-n-acetylglucosamine pyrophosphorylase 1 (UAP1) induces early leaf senescence and defence responses in rice. *J. Exp. Bot.* 66, 973–987. doi: 10.1093/jxb/eru456

Wu, C., Cui, K., and Fahad, S. (2022). Heat stress decreases rice grain weight: Evidence and physiological mechanisms of heat effects prior to flowering. *Int. J. Mol. Sci.* 23, 10922. doi: 10.3390/ijms231810922

Wuriyanghan, H., Zhang, B., Cao, W. H., Ma, B., Lei, G., Liu, Y. F., et al. (2009). The ethylene receptor ETR2 delays floral transition and affects starch accumulation in rice. *Plant Cell* 21, 1473–1494. doi: 10.1105/tpc.108.065391

Wu, S., Xie, Y., Zhang, J., Ren, Y., Zhang, X., Wang, J., et al. (2015). *VLN2* regulates plant architecture by affecting microfilament dynamics and polar auxin transport in rice. *Plant Cell* 27, 2829–2845. doi: 10.1105/tpc.15.00581

Xia, D., Zhou, H., Liu, R., Dan, W., Li, P., Wu, B., et al. (2018). *GL3.3*, a novel QTL encoding a GSK3/SHAGGY-like kinase, epistatically interacts with *GS3* to produce extra-long grains in rice. *Mol. Plant* 11, 754–756. doi: 10.1016/j.molp.2018.03.006

Xie, X., Jin, F., Song, M. H., Suh, J. P., Hwang, H. G., Kim, Y. G., et al. (2008). Fine mapping of a yield-enhancing QTL cluster associated with transgressive variation in an

Oryza sativa × *O. rufipogon* cross. *Theor. Appl. Genet.* 116, 613–622. doi: 10.1007/s00122-007-0695-x

Xing, Y., and Zhang, Q. (2010). Genetic and molecular bases of rice yield. *Annu. Rev. Plant Biol.* 61, 421–442. doi: 10.1146/annurev-arplant-042809-112209

Ying, J. Z., Ma, M., Bai, C., Huang, X. H., Liu, J. L., Fan, Y. Y., et al. (2018). *TGW3*, a major QTL that negatively modulates grain length and weight in rice. *Mol. Plant* 11, 750–753. doi: 10.1016/j.molp.2018.03.007

Yuan, H., Gao, P., Hu, X., Yuan, M., Xu, Z., Jin, M., et al. (2022a). Fine mapping and candidate gene analysis of *qGSN5*, a novel quantitative trait locus coordinating grain size and grain number in rice. *Theor. Appl. Genet.* 135, 51–64. doi: 10.1007/s00122-021-03951-7

Yuan, S., Yang, C., Yu, X., Zheng, C., Xiao, S., Xu, L., et al. (2022b). On-farm comparison in grain quality between main and ratoon crops of ratoon rice in hubei province, central China. *J. Sci. Food Agric.* 102, 7259–7267. doi: 10.1002/jsfa.12091

Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42, 355–360. doi: 10.1038/ng.546

Zhang, X., Wang, J., Huang, J., Lan, H., Wang, C., Yin, C., et al. (2012). Rare allele of *OsPPKL1* associated with grain length causes extra-large grain and a significant yield increase in rice. *Proc. Natl. Acad. Sci. U. S. A.* 109, 21534–21539. doi: 10.1073/pnas.1219776110

Zhao, D. S., Li, Q. F., Zhang, C. Q., Zhang, C., Yang, Q. Q., Pan, L. X., et al. (2018). *GS9* acts as a transcriptional activator to regulate rice grain shape and appearance quality. *Nat. Commun.* 9, 1240. doi: 10.1038/s41467-018-03616-y

Zhao, Y. F., Peng, T., Sun, H. Z., Teotia, S., Wen, H. L., Du, Y. X., et al. (2019). miR1432-*OsACOT* (Acyl-CoA thioesterase) module determines grain yield *via* enhancing grain filling rate in rice. *Plant Biotechnol. J.* 17, 712–723. doi: 10.1111/pbi.13009

Zhou, S., Li, X., Liu, Q., Zhao, Y., Jiang, W., Wu, A., et al. (2021). DNA Demethylases remodel DNA methylation in rice gametes and zygote and are required for reproduction. *Mol. Plant* 14, 1569–1583. doi: 10.1016/j.molp.2021.06.006

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

Ziska, L. H., Fleisher, D. H., and Linscombe, S. (2018). Ratooning as an adaptive management tool for climatic change in rice systems along a north-south transect in the southern Mississippi valley. *Agr. Forest. Meteorol.* 263, 409–416. doi: 10.1016/j.agrformet.2018.09.010

Zuo, J., and Li, J. (2014). Molecular genetic dissection of quantitative trait loci regulating rice grain size. *Annu. Rev. Genet.* 48, 99–118. doi: 10.1146/annurev-genet-120213-092138

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 02 February 2023
DOI 10.3389/fpls.2023.1118777

OPEN ACCESS

EDITED BY
Zhenyu Jia,
University of California, Riverside,
United States

REVIEWED BY
Jian-Fang Zuo,
Huazhong Agricultural University, China
Ya-Wen Zhang,
Huazhong Agricultural University, China
Haiping Zhang,
Anhui Agricultural University, China

*CORRESPONDENCE
WeiGang Xu
xuwg1958@163.com

SPECIALTY SECTION
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

RECEIVED 08 December 2022
ACCEPTED 02 January 2023
PUBLISHED 02 February 2023

CITATION
Kou C, Peng C, Dong H, Hu L and Xu W (2023) Mapping quantitative trait loci and developing their KASP markers for pre-harvest sprouting resistance of Henan wheat varieties in China.
Front. Plant Sci. 14:1118777.
doi: 10.3389/fpls.2023.1118777

Mapping quantitative trait loci and developing their KASP markers for pre-harvest sprouting resistance of Henan wheat varieties in China

Cheng Kou[1,2], ChaoJun Peng[2,3,4], HaiBin Dong[2,3,4], Lin Hu[2,3,4] and WeiGang Xu[1,2,3,4]*

[1]College of Agronomy, Northwest A&F University, Xianyang, China, [2]Institute of Crop Molecular Breeding, Henan Academy of Agricultural Sciences, Zhengzhou, Henan, China, [3]Henan Key Laboratory of Wheat Germplasm Resources Innovation and Improvement, Zhengzhou, Henan, China, [4]The Shennong laboratory, Zhengzhou, Henan, China

Introduction: Pre-harvest Sprouting (PHS) seriously affects wheat quality and yield. However, to date there have been limited reports. It is of great urgency to breed resistance varieties *via* quantitative trait nucleotides (QTNs) or genes for PHS resistance in white-grained wheat.

Methods: 629 Chinese wheat varieties, including 373 local wheat varieties from 70 years ago and 256 improved wheat varieties were phenotyped for spike sprouting (SS) in two environments and genotyped by wheat 660K microarray. These phenotypes were used to associate with 314,548 SNP markers for identifying QTNs for PHS resistance using several multi-locus genome-wide association study (GWAS) methods. Their candidate genes were verified by RNA-seq, and the validated candidate genes were further exploited in wheat breeding.

Results: As a result, variation coefficients of 50% and 47% for PHS in 629 wheat varieties, respectively, in 2020-2021 and 2021-2022 indicated large phenotypic variation, in particular, 38 white grain varieties appeared at least medium resistance, such as Baipimai, Fengchan 3, and Jimai 20. In GWAS, 22 significant QTNs, with the sizes of 0.06% ~ 38.11%, for PHS resistance were stably identified by multiple multi-locus methods in two environments, e.g., AX-95124645 (chr3D:571.35Mb), with the sizes of 36.390% and 45.850% in 2020-2021 and 2021-2022, respectively, was detected by several multi-locus methods in two environments. As compared with previous studies, the AX-95124645 was used to develop Kompetitive Allele-Specific PCR marker QSS.TAF9-3D (chr3D:569.17Mb~573.55Mb) for the first time, especially, it is available in white-grain wheat varieties. Around this locus, nine genes were significantly differentially expressed, and two of them (TraesCS3D01G466100 and TraesCS3D01G468500) were found by GO annotation to be related to PHS resistance and determined as candidate genes.

Discussion: The QTN and two new candidate genes related to PHS resistance were identified in this study. The QTN can be used to effectively identify the PHS

resistance materials, especially, all the white-grained varieties with QSS.TAF9-3D-TT haplotype are resistant to spike sprouting. Thus, this study provides candidate genes, materials, and methodological basis for breeding wheat PHS resistance in the future.

KEYWORDS

wheat, pre-harvest sprouting, genome-wide association study, RNA-seq, KASP, mrMLM

1 Introduction

Wheat is a major worldwide food crop, and China is the largest wheat producer and consumer in the world. In 2022, Chinese wheat harvest area was 22,911.2 thousand hectares, and the total yield reached 135.76 million tons. In Henan, wheat harvest area and yield accounts for 24.8% and 28.1% in China, respectively, being the largest main wheat producing area in China. Its genetic improvement of wheat varieties has played an important role in its continuous improvement of wheat production capacity.

Pre-harvest Sprouting (PHS) refers to the phenomenon of seeds germinating and sprouting on the spike under rainy or humid conditions before wheat harvest. It is a worldwide natural disaster, and has been reported in China (Zhou et al., 2017), Japan (Kashiwakura et al., 2016), the United States (Nonogaki et al., 2014), Canada (Cabral et al., 2014), Europe (Rakoczy-Trojanowska et al., 2017), South Africa (Sydenham and Barnard, 2018), and Australia (Barrero et al., 2010). In China, the frequent and severe PHS spike hazards happened in the middle and lower reaches of Changjiang River winter wheat zone, southwest winter wheat zone, and northeast spring wheat zone (Jin, 1996; Zhang et al., 2010). In these zones, PHS resistance depends on dormant genes linked to red seed coat. In northern China, such as Henan, however, white-grained wheat varieties are used in production, and with the overall popularization of wheat mechanization harvest, that wheat should be harvested after being fully mature and dehydrated in the field. The varieties with PHS susceptibility have an increased probability of spike sprouting due to rainfall during the mature harvest period. Therefore, it is of great urgency to breed resistance varieties *via* PHS resistance quantitative trait nucleotides (QTNs) or genes of white-grained wheat.

Wheat PHS resistance is a complex quantitative trait controlled by multiple genes (Imtiaz et al., 2008). Thus, it is very important and necessary to identify these resistance loci and develop their molecular markers in crop breeding. In previous linkage analysis, a series of QTLs for PHS resistance has been located on all the 21 wheat chromosomes (Mohan et al., 2009; Cabral et al., 2014; Cao et al., 2016; Fakthongphan et al., 2016), in which repeatedly and stably QTLs were found on chromosome 3 (Kato et al., 2001; Osa et al., 2003; Kulwal et al., 2004; Mori et al., 2005; Liu and Bai, 2010). Currently, red-grained wheat varieties generally exhibit higher PHS resistance, because the PHS resistance genes on chromosomes 3A, 3B and 3D are thought to be closely linked to red seed coat, which is controlled by R dominant allele (Himi et al., 2011). Recently, genome-wide association studies (GWAS) have been used to identify QTLs and their candidate genes for wheat grain weight and plant height (Zanke et al., 2014; Chen et al., 2016; Wang et al., 2017), especially, Zhu et al. (2019) identified some QTLs and developed their molecular markers on wheat chromosomes 1AL, 3BS, and 6BL for PHS resistance, and Lin et al. (2017) identified two candidate genes for PHS resistance in 80 wheat varieties. However, the studies on wheat PHS resistance are relatively limited.

Chinese wheat local varieties showed higher PHS resistance than improved varieties (Wang et al., 2011; Liu et al., 2014), which provided valuable genetic resources for mining the loci of PHS resistance. In this study, 629 wheat varieties were measured for PHS resistance in 2020-2021 and 2021-2022, including 373 wheat local varieties over 70 years ago and 256 wheat improved varieties over the last 70 years in Henan Province, China. To mine some valuable QTNs for PHS resistance, these phenotypes were used to associate with SNP markers in the above 629 wheat varieties using several multi-locus GWAS methods. The results were validated by RNA-seq datasets between PHS resistance and susceptibility varieties, one confirmed QTN was used to develop Kompetitive Allele-Specific PCR (KASP) marker, and the KASP marker was further confirmed to be associated with PHS resistance. Thus, this study provides a valuable locus and white-grained wheat PHS resistance materials, which is available in main producing zones.

2 Materials and methods

2.1 Materials

In association mapping population, there were 629 Chinese wheat varieties, including 373 local wheat varieties from 70 years ago and 256 improved wheat varieties (lines). In autumns of 2020 and 2021, these varieties were planted in the experimental field of Henan Modern Agricultural Research and Development Base (East longitude: 113.707°, North latitude: 35.011°). The winter wheat varieties were provided by Institute of Crops Molecular Breeding, Henan Academy of Agricultural Sciences.

2.2 Measurement of PHS resistance in 629 wheat varieties

Based on the agricultural industry standards of the People's Republic of China, NY/T 1939-2009, namely "standard" hereinafter, we harvested the varieties in the dough stage in turn, and 20 main stem spikes of each variety were stored in the refrigerator at -20°C. After all the varieties were harvested, we measured the PHS

phenotypes of all the varieties on phytotron with temperature of 22°C ± 1°C and relative humidity of 95% ± 5%. Samples were removed from phytotron after 96 hours, and dried at 60°C for counting, spike sprouting (SS) was calculated from the formula x=(n/N)*100%,

where *n* is the number of sprouted grains per spike, and *N* is the total number of grains per spike. The relative SS index "*I*" of each variety to be tested was calculated from I=x1/x2.

where x_1 is the SS of each variety to be tested, and x_2 is the SS of the control variety, being Zhoumai 18 or a local variety with similar PHS phenotype with Zhoumai 18. Based on the criteria of PHS resistance in Supplementary Table S1, pre-harvest sprouting grade of each variety was determined.

2.3 Multi-locus GWAS for wheat PHS resistance in 629 varieties

As described in the reference (Du et al., 2021), all the 629 varieties were genotyped by wheat 660K microarray, and high quality genotypes of 314,548 SNP markers were obtained based on four screening criteria: alleles = 2, minor allele frequency (MAF) ≥ 0.01, missing ≤ 10%, and heterozygosity ≤ 10%. The best linear unbiased prediction (BLUP) values in 2020-2021 and 2021-2022 years was calculated by R language package (R 4.2.1). These marker genotypes were used to associate trait phenotypes or BLUP values in the 629 wheat varieties using the IIIVmrMLM (Li et al., 2022a; Li et al., 2022b) and mrMLM (Zhang et al., 2020) software packages, in which the latter included mrMLM (Wang et al., 2016), FASTmrMLM (Tamba and Zhang, 2018), FASTmrEMMA (Wen et al., 2017), pLARmEB (Zhang et al., 2017), ISIS EM-BLASSO (Tamba et al., 2017), and pKWmEB (Ren et al., 2018) methods. The population structure was determined using admixture_linux-1.3.0 software. The number of subgroups (K) was scanned from 2 to 5 using the admixture software and determined as two. The kinship matrix was calculated using the mrMLM software. The critical LOD score for significant QTLs was set as LOD = 3.0, which is equivalent to P-value = 2e-4. The Manhattan plots were drawn using the mrMLM software. The LD decay distance was calculated using vcftools v0.1.13, plink-v1.07, and PopLDdecay 3.41 softwares. The 2.192 Mb region was regarded as the upstream and downstream of a significant QTL.

2.4 Design and analysis of molecular markers for PHS resistance loci

0.2-0.3g fresh leaves were taken from each of 629 wheat varieties, pre-cooled with liquid nitrogen, crushed, and placed into 1.5mL centrifuge tube, and wheat genomic DNA was extracted based on Gawel and Jarret (1991).

2.4.1 Design of KASP molecular marker for PHS resistance locus

The forward and reverse primers (FT: 5'-ATCAATTATCAG CTCTGGAT-3'; FC: 5'-ATCAATTATCAGCTCTGGAC-3'; R: 5'-AATCTTGACCTGTGTCCCGA - 3') of KASP molecular markers were designed in the upstream 20 bp and downstream 155 bp according to the physical location information of significantly associated locus in reference to the Chinese spring sequence information of wheat Whole Genomics website (http://202.194.139.32/jbrowse-1.12.3-release/?data=Chinese_Spring1.0). HEX (red, 5'-GAAGGTCGGAGTCAACGGATT-3') was added to the 5' end of FT primer sequence and FAM (blue, 5'-GAAGGTGACCAAG TTCATGCT-3') was added to the 5' end of FC primer sequence, respectively. These primers were synthesized by Sangon Biotech (Shanghai) Co., Ltd. PCR reactions were performed in an Hydrocycler-thermal cycler in a total volume of 3μL, including 1.5μL KASP 2× Master Mix (LGC Technology (Shanghai) Co., Ltd.), 80 ng of template DNA, 0.06μL KASP Assay mix(100uM of Forward primer-FT, Forward primer-FC, Reverse primer-R and ddH2O mixed in a 12:12:30:46 volume ratio). PCR amplification were 94°C for 15min, 10 cycles of 94°C for 20s, 61°C-55°C for 60s by 0.6°C decrease per cycle, and with a final extension is 29 cycles of 94°C 20s, 55°C 60s.

2.4.2 Sequence analysis of the KASP marker amplified product

The KASP molecular marker reaction products of Zhoumai 18 and Shengsimai were separated by 1% agarose gel electrophoresis, the target fragments were recovered and purified, which was cloned with pMDTM19-T vector (Takara Biomedical Technology (Beijing) Co., Ltd.), and sequenced by Sangon Biotech (Shanghai) Co., Ltd. At least 10 clones were sequenced for each variety. DNAMAN software was used to analyze the allelic variation of the amplified product sequences of KASP marker primers, and then BLAST (basic local alignment search tool) at EnsemblPlants database (http://plants.ensembl.org/index.html).

2.5 RNA-seq sample selection preparation and differential gene expression analysis

2.5.1 RNA-seq sample selection preparation

According to the identification results of spike sprouting in association population, the highly resistant red-grained variety Shengsimai, the white-grained variety Baipimai, and the highly susceptible white-grained variety Zhoumai 18 were selected as RNA-seq samples. The sample processing method was carried out according to the standard. At the wax-ripening stage, all the three samples were cut from 15 cm below the spike, 9 spikes were taken from each material, which were divided into 3 portions, each spike was a biological replicate. And then, after soaking for 4 hours, one of 3 portions was taken out of liquid

Nitrogen and frozen for the 0-point control of RNA-seq (0h). The remaining two portions were further tested for PHS identification. A total of 96 hours were required for PHS identification. Samples were taken out and frozen in liquid nitrogen at 48 hours (48h) and 96 hours (96h). The subsequent RNA extraction library preparation, sequencing, and analysis results of RNA-seq were provided by Beijing Biomarker Technologies Co., Ltd.

2.5.2 RNA-seq differential expression analysis of genes

Differential gene expression analysis of RNA-seq samples was performed on the website of Beijing Biomarker Technologies Co., Ltd. (http://www.biomarker.com.cn/). FDR < 0.05 was used as the

standard for screening differentially expressed genes, and the difference groups were set according to the PHS resistance and susceptibility, and the PHS resistance of different seed coat colors, as shown in Supplementary Table S2. Using the differential gene expression datasets, the P-values were calculated by GO annotation enrichment tool of the Beijing Biomarker Technologies (https://international.biocloud.net/).

3 Results

3.1 Phenotypic analysis for PHS resistance in 629 wheat varieties

The spike sprouting method was used to identify the phenotypes of 629 wheat varieties (Supplementary Table S3), including 333 red-grained and 296 white-grained varieties. Among them, the numbers of red-grained varieties and white-grained varieties resistant to spike germination in the two years were 293 and 38, respectively. Red-grained varieties were generally more resistance than white-grained varieties. Among the 373 local varieties, 305 were red-grained varieties, 68 were white-grained varieties, and 298 were resistant to spike sprouting in both the two years. Among the 256 improved varieties, 28 were red-grained varieties, 228 were white-grained varieties, and 33 were resistant to spike sprouting in both the two years (Figure 1). This indicates that the grain color of wheat varieties in Henan Province has changed greatly from the local varieties before 1950 to the later improved varieties, and red-grained varieties were gradually changed to white -grained varieties. In the past two decades, all the varieties developed have been white-grained varieties.

In the identification of PHS resistance, the ranges of spike sprouting rates in 2020-2021 and 2021-2022 were 0.51%~99.22% (Mean ± SD: 35.38% ± 0.29%) and 0.00%~97.47% (Mean ± SD: 33.63% ± 0.25%), respectively, indicating abundant phenotypic variation in both environments. Analysis of variance showed that the spike sprouting rates were significant across genotypes, environments, and their interactions, and the heritability was 0.88 (P-value ≤ 0.001; Supplementary Table S4), indicating that wheat PHS resistance was mainly determined by genotypes and modified by environments.

3.2 GWAS for PHS resistance index in 629 wheat varieties

The phenotypes for PHS resistance index in 629 wheat varieties in the two years were used to associate with all the SNP markers using six multi-locus GWAS approaches. As a result, a total of 22 QTNs were stably detected by multiple methods or environments, and their proportions of total phenotypic variation explained by each QTN (R^2) was from 0.00001% to 38.1121% (Table 1). Among these loci, two loci, AX-95124645 on chromosome 3D and AX-109028892 on chromosome 5D, had been identified by Zhou et al. (2017), while other loci were identified for the first time, especially, AX-111020384 on chromosome 3A and AX-95124645 on chromosome 3D were identified by all the seven methods in the two software packages in all the two environments, and their R^2 values were 12.8% and 38.1%, respectively (Figure 2), indicating the major QTN around AX-95124645 for wheat PHS resistance. As compared with the GWAS results for PHS resistance in 272 local varieties genotyped by Wheat660 SNP markers (Zhou et al., 2017), the resistance allele of AX-95124645 was found to be associated with only red-grained varieties in Zhou et al. (2017) and with both red-grained and white-grained varieties in this study. In linkage disequilibrium analysis, the LD decay distance in association mapping population was found to be 2.192Mb. This means that 2.192 Mb upstream and downstream regions of the significant QTL, that is QSS.TAF9-3D (chr3D:569.167Mb ~ 573.551Mb), may be used to mine candidate genes.

3.3 KASP marker of QSS.TAF9-3D

Using the KASP marker around a major QTN AX-95124645, two haplotypes were found in the 629 wheat varieties(Figure 3), namely QSS.TAF9-3D-TT and QSS.TAF9-3D-CC, which is completely consistent with the results of marker AX-95124645 obtained from 629 wheat varieties scanned by 660K chip. We also used T-A cloning and sequencing of the amplified products of Zhoumai18 (QSS.TAF9-3D-CC) and Shengsimai (QSS.TAF9-3D-TT), which there was only a T/C allele mutation at 26 bp in the amplified products of Zhoumai 18 and Shengmai. Using EnsemblPlants database (http://plants.ensembl.org/index.html), it was found that the above T/C alleles in the amplified products are exactly consistent with those at the physical location of marker AX-95124645 (Figure 3). QSS.TAF9-3D-TT and

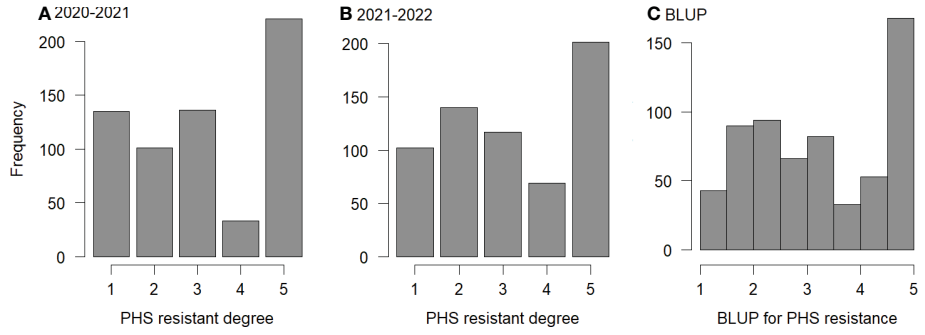


FIGURE 1
Frequency for PHS resistance. **(A–C)**: frequencies of PHS resistance for 629 varieties at 2020-2021, 2021-2022, and their BLUP values, respectively; 1 to 5: highly resistance, resistance, middle resistance, susceptibility, and highly susceptibility, respectively.

TABLE 1 Significant QTNs for PHS resistance index detected by multiple multi-locus GWAS methods in two environments.

QTN	Chr	Marker	Position (Mb)	Env	Methods	−log10(P-value)	R^2 (%)	Reference
QTN1	1A	AX-110511933	12.859	3	2, 4	5.355~5.3664	0.5869~0.6868	
QTN2	1A	AX-109827872	545.852	3	3, 5	4.0032~6.3289	0.4481~1.0833	
QTN3	2A	AX-94559008	21.208	1, 3	1, 4	4.2459~6.2418	1.1445~2.3547	
QTN4	2A	AX-109841146	716.163	1	1, 2, 4	4.3164~6.8085	0.0652~1.6299	
QTN5	3A	AX-111020384	10.159	1, 2, 3	1~7	4.3263~36.0266	3.1246~12.8725	
QTN6	5A	AX-111670342	569.991	1, 3	3	4.4955~5.8588	0.00001~0.7568	
QTN7	6A	AX-110436229	590.075	2	1, 2	4.4267~4.7556	0.7937~1.1922	
QTN8	6A	AX-94617998	608.969	3	1, 2, 5	3.8013~5.7994	0.4255~0.7798	
QTN9	7A	AX-110492207	20.188	1, 3	1~5, 7	3.8284~24.2651	0.6686~3.1246	
QTN10	1B	AX-94741303	3.181	2, 3	1, 2, 4, 6	4.5796~8.3543	0.5791~1.886	
QTN11	2B	AX-111503288	765.578	1, 3	1, 4, 6	5.5209~12.389	1.4445~3.8593	
QTN12	3B	AX-111703196	16.805	2, 3	3, 4, 7	3.9937~9.9601	0.8401~1.7144	
QTN13	5B	AX-94487480	469.826	1, 3	2, 5	3.7531~4.4669	0.49~1.1933	
QTN14	5B	AX-108862465	511.697	1	1, 2, 4	5.009~8.1404	0.0551~1.318	
QTN15	5B	AX-108932221	536.054	3	2, 4	4.4563~5.4732	0.2389~0.3166	
QTN16	1D	AX-94392070	58.087	1, 3	4, 5, 6	4.0595~7.3964	0.915~1.557	
QTN17	1D	AX-110332164	458.942	2, 3	1, 4	4.1832~5.235	0.6609~0.905	
QTN18	3D	AX-95124645	571.359	1, 2, 3	1~7	5.0794~48.9107	4.9300~38.1121a	Zhou et al., 2017
QTN19	4D	AX-108916749	19.09	1, 3	2, 3	4.2911~6.551	0.00001~3.6386	
QTN20	5D	AX-109028892	45.711	1	1, 5, 6	10.6444~27.0823	7.2283~11.2516	Zhou et al., 2017
QTN21	6D	AX-109716798	143.583	3	3, 5	3.9008~5.7168	0.4118~0.5253	
QTN22	6D	AX-109293498	472.945	1, 3	1, 2, 3, 4, 6	4.2369~7.5779	0.5624~1.7955	

Env 1, 2, and 3: the PHS resistance indices in 2020-2021, 2021-2022, and their BLUP values, respectively. Methods 1 to 7: mrMLM, FASTmrMLM, FASTmrEMMA, pLARmEB, ISIS EM-BLASSO, pKWmEB, and IIIVmrMLM, respectively. "a": the R^2 value is greater than 30%.

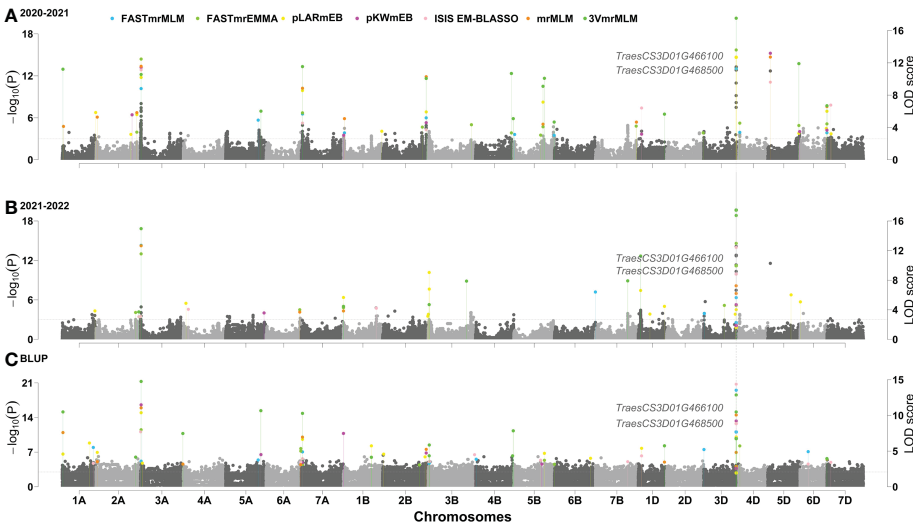


FIGURE 2
(A–C) Manhattan plot for PHS resistance index in 629 wheat varieties. If LOD score ≥ 10, the LOD scores obtained are transformed as LOD' = 10 + (LOD − 10)/5.

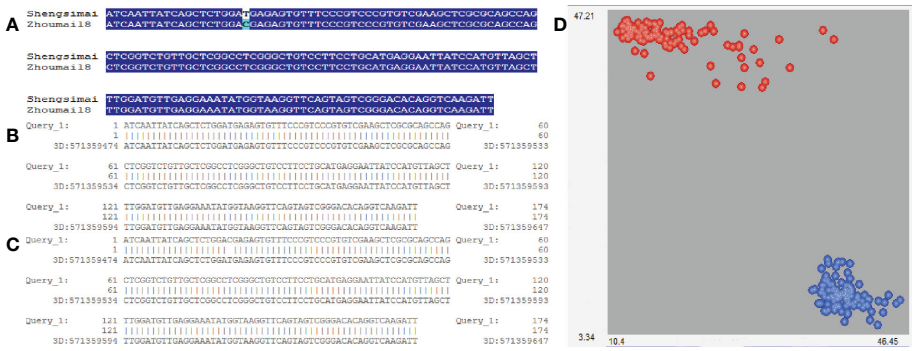


FIGURE 3
Sequence alignment of KASP marker amplification products at significant locus of QSS.TAF9-3D. **(A)** sequence alignment of the amplified products of Shengsimai and Zhoumai18; **(B)** Sequence alignment of Shengsimai amplification products from EnsemblPlants database; **(C)** Sequence alignment of Zhoumai18 amplification products from EnsemblPlants database; **(D)** Red and blue: varieties with QSS.TAF9-3D-TT and QSS.TAF9-3D-CC haplotypes, respectively. Horizontal and vertical coordinates represent the fluorescence signal values of FAM and HEX, respectively.

QSS.TAF9-3D-CC haplotypes could be distinguished by KASP molecular marker, having 261 QSS.TAF9-3D-TT haplotypes and 368 QSS.TAF9-3D-CC haplotypes in 629 wheat varieties.

The KASP marker was used to conduct haplotype analysis in 629 wheat varieties. The results were listed in Supplementary Table S5. The results showed that QSS.TAF9-3D-TT haplotype had significantly higher PHS resistance than QSS.TAF9-3D-CC haplotype. TAF9-3D-TT/CC markers accounted for 36.390% and 45.850% of phenotypic variation in SS_2021 and SS_2022, respectively. The QSS.TAF9-3D-TT haplotype was negatively correlated with the PHS resistance index, indicating that the QSS.TAF9-3D-TT haplotype was mainly distributed in varieties with high PHS resistance. Among 261 varieties with QSS.TAF9-3D-TT haplotype, 253 and 252 were resistant to spike sprouting in 2020-2021 and 2021-2022, respectively. Among the 38 white-grained resistant PHS varieties, 11 white grained varieties with QSS.TAF9-3D-TT showed PHS resistance (Supplementary Table S6). We considered that PHS resistance in the remaining 27 varieties was dependent on other related genes or QTLs.

3.4 RNA-seq analysis

3.4.1 Validation of GWAS results by RNA-seq

Differentially expressed genes (DEGs) in QSS.TAF9-3D region were listed in Table 2. The results showed the existence of differential expressions between the PHS resistance varieties (Baipimai and Shengsimai) and the PHS susceptibility variety (Zhoumai18), indicating the association of QSS.TAF9-3D with PHS resistance. With the increase of treatment time, the number of DEGs between the two resistant varieties and one susceptible variety significantly increased. The number of DEGs in the two resistance varieties of Baipimai and Shengsimai with different seed coat colors increased first and then decreased with the increase of treatment time, indicating the association of seed coat color with PHS resistance.

4.1 Candidate genes around QSS.TAF9-3D

In the region of QTL QSS.TAF9-3D, there were 56 genes. Using the RNA-seq datasets, 9 genes were found to be differentially expressed, as shown in Figure 4. Among them, TraesCS3D01G466100 GO annotation showed that it encodes ubiquitin protein transferase, and the NCBI conserved domain analysis showed that it encodes RING-type E3 ubiquitin ligase. In recent years, a large number of studies have shown that RING E3 is widely involved in abiotic stress processes (Cho et al., 2017). TraesCS3D01G468500 gene encodes initiation transcription factor TAF9. At present, the function of TAF9 has been reported in both human and yeast (Frontini et al, 2005; Knoll et al, 2020). However, TAF9 has limited studied in plants. Thus, the two genes were regarded as new candidate genes in this study.

4 Discussion

Genome-wide association studies for wheat PHS resistance in 629 local and improved varieties (lines) in Henan Province, China provide new insights into the genetic foundation of the important trait and variety breeding. In previous studies, most of them focused on the PHS resistance in southwest and southern wheat zones in China, for example, Zhou et al. (2017) found that the landraces in Chinese wheat zones with high precipitation showed strong PHS resistance in 717 Chinese wheat landraces, but there were few studies on PHS resistance in northern wheat zones with less rain. In this study, 373 local varieties before 1950 and 256 improved varieties after 1950 in Henan Province were included. It was found that genes for wheat PHS resistance were gradually lost in the process of selection and breeding for yield, quality, and other important breeding traits. The main reason is that most of the loci or genes for wheat PHS resistance are found to be linked with red seed coat, while most improved varieties are white grained, resulting in the generally reduced PHS resistance of modern improved varieties. In this study, 38 white grain resistant varieties were observed, and this study provides a material basis for breeders to select white-grained resistant varieties for PHS.

Zhou et al. (2017) found three major PHS resistant QTLs on chromosomes 3A, 3D, and 5D, in which the marker AX-95124645 was located on Chr 3D. In this study, AX-95124645 locus was identified by several multi-locus methods in two environments to be associated with PHS resistance, especially, its R^2 value was 38%. However, this study provides three new results compared with

10.3389/fpls.2023.1118777

TABLE 2 No. of differentially expressed genes in QSS.TAF9-3D region.

Time point	DEG	Comparison		
		Baipimai vs Zhoumai18	Shengsimai vs Zhoumai18	Baipimai vs Shengsimai
0h	Up	1	0	1
	Down	0	2	0
	Total	1	2	1
48h	Up	6	7	5
	Down	1	5	2
	Total	7	12	7
96h	Up	3	3	4
	Down	8	9	0
	Total	11	12	4

Time point indicates sample treatment time; DEG indicates differentially expressed genes, Up indicates that "vs" is less expressed in the former than in the latter, and Down indicates that "vs" is more expressed in the former.

previous studies. First, resistant and susceptible PHS varieties were used to conduct RNA-seq analysis, and 9 DEGs were found in the 2.192 Mb upstream and downstream intervals of AX-95124645, and two candidate genes were predicted. Then, the KASP marker QSS.TAF9-3D-TT/CC was developed based on the AX-95124645 locus. The results showed that QSS.TAF9-3D-TT/CC haplotypes with only one T/C base allele variation could completely distinguish all the PHS resistant and susceptible varieties. Finally, all the QSS.TAF9-3D-TT haplotypes were found in 11 white-grained varieties to be resistant for PHS.

It should be point out that the KASP marker QSS.TAF9-3D developed in this study is valuable. First, the KASP marker was used to select 11 white-grained resistant varieties with excellent haplotype QSS.TAF9-3D-TT, indicating its possibility of marker-assisted selection in white-grained varieties for PHS resistance. But among 629 varieties, the numbers of white-grained varieties resistant to spike germination in the two years was 38. Because Wheat PHS resistance controlled by multiple genes (Imtiaz et al., 2008), so we consider the spike sprouting resistance of the remaining 27 white grain varieties was caused by other genes or QTLs. Second, this marker uses high-throughput KASP genotyping technology. In particular, KASP is based on conventional PCR and fluorescence detection, which can meet the requirements of low, medium, and high throughput genotyping on the basis of ordinary laboratory operation (Semagn et al., 2014), indicating that it is flexible, cheap, high-throughput, automated, and accurate. As we known, KASP, as an alternative to TaqMan, is similar in principle to TaqMan (also based on terminal fluorescence reading), but it differs from TaqMan technology in the following ways. It uses a universal probe, which can be used with a variety of different gene-specific primers, without the need for probe synthesis for each specific site, which greatly reduces the reagent cost of the experiment (Majeed et al., 2018). In conclusion, QSS.TAF9-3D-

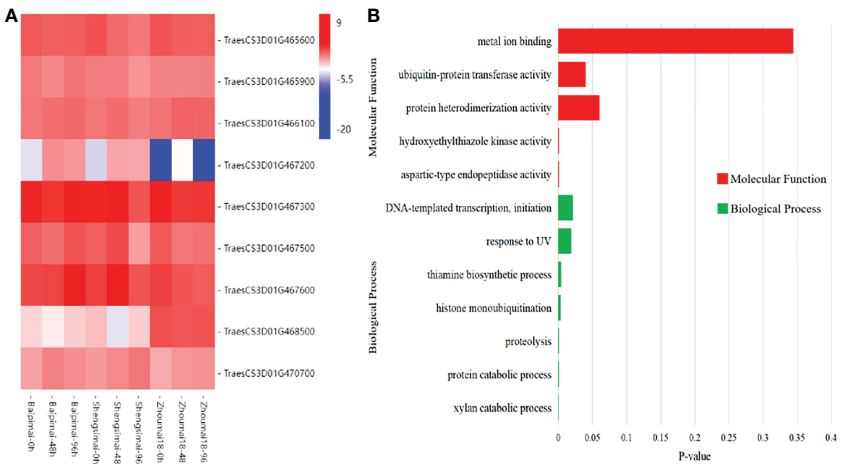


FIGURE 4
Differentially expressed genes in the QSS.TAF9-3D transcriptome. **(A)** heat map of differentially expressed genes (DEGs) between PHS resistant and susceptible. **(B)** GO Molecular Functions and Biological Process of DEGs.

TT/CC markers can be used for higher throughput and more accurate screening of PHS resistance varieties, especially in white-grained varieties, which provides a strong theoretical basis for molecular mark-assisted breeding.

Myb10 is an important regulatory gene in the pathway of pigment synthesis. The earliest MYB-type transcription factor identified was maize Colorless 1 (Paz-Ares et al., 1987). In wheat, *Tamyb10* gene is believed to be related to seed dormancy, because it may affect the sensitivity of wheat embryo to ABA. Lang et al. (2021) found that *myb10-D* gene, as a candidate gene for PHS-3D, not only regulates the synthesis of flavonoid compounds, but also increases the ABA concentration in developing seeds, thus inhibiting the wheat PHS. In this study, the candidate gene TraesCS3D01G468400 was found to be consistent with *Tamyb10-D* in the annotation information of 61 genes in the QSS.TAF9-3D region. Although Himi et al. (2011) designed the *Tamyb10-D* marker to screen PHS resistance materials, *Tamyb10-D* is an important regulatory gene involved in the pigment synthesis of wheat seed coat so that its corresponding molecular marker is mainly used to screen the PHS resistance of red-grained wheat varieties, indicating its difficulty in the application of white-grained varieties. We identified two differentially expressed genes TraesCS3D01G466100 and TraesCS3D01G468500 in the QSS.TAF9-3D region using RNA-seq. TraesCS3D01G466100 GO annotation shows that it encodes C3HC4-RING fifinger E3 ubiquitin ligase. Yang et al. (2016) identified *AtAIRP4* in Arabidopsis, which is induced by ABA and other stress treatments. *AtAIRP4* encodes a cellular protein with a C3HC4-RING finger domain in its C-terminal side, which has *in vitro* E3 ligase activity. A large number of studies have shown that the dormancy period of wheat seeds is negatively correlated with the degree of PHS (Flintham et al., 2000; Biddulph et al., 2008; Shu et al., 2016), and ABA plays a crucial role in promoting seed dormancy and inhibiting seed germination (Martínez-Andújar et al., 2011). Thus, it is possible for *TraesCS3D01G466100* to affect PHS resistance by regulating seed ABA levels. *TraesCS3D01G468500* gene encodes the initiation transcription factor TAF9. Yang (2015) cloned a gene *CpTAF9* in the woody ornamental plant *Chimonanthus melanoides*. Salt stress, high temperature or ABA application promoted the expression of *CpTAF9* gene in leaves. ABA is an important hormone regulating seed dormanness. *TraesCS3D01G468500* gene may affect wheat spike germination by indirectly regulating seed ABA content. We selected these two genes as new PHS resistance candidate genes.

5 Conclusion

We firstly identified 38 white-grained varieties with PHS resistance in 629 wheat varieties (lines) from Henan Province, China, stably identified a major QTN AX-95124645 on chromosome 3D, and developed its KASP marker QSS.TAF9-3D-TT/CC. This marker haplotype can effectively detect the PHS resistance materials, especially, all the white-grained varieties with QSS.TAF9-3D-TT haplotype are resistant to spike sprouting, which can be used for molecular mark-assisted breeding of spike sprouting resistance in white-grained varieties. This study provides material and methodological basis for breeding wheat PHS resistance in the future.

Data availability statement

The sequencing data have been successfully submitted to the GEO database and applied for public disclosure. The GEO accession number is GSE222342, and the BioProject accession number is PRJNA919175.

Author contributions

W-GX conceived and managed the research and revised the manuscript. CK and C-JP analyzed datasets. CK measured the phenotypes of the traits. LH and H-BD provided the research materials and Instruments. CK wrote the draft. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Key Research and Development Program of China (2021YFD1200601-03), the Excellent Youth Science and Technology Foundation of Henan Academy of Agricultural Sciences (2022YQ17), the China Agriculture Research System of MOF and MARA (CARS-03-7), and the Henan Academy of Agricultural Sciences Special Fund for Independent Innovation Foundation (2022ZC73).

Acknowledgments

We are grateful to Prof Yuan-Ming Zhang at Henan Academy of Agricultural Sciences and Huazhong Agricultural University for revising this manuscript.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1118777/full#supplementary-material

References

Barrero, J. M., Millar, A. A., Griffiths, J., Czechowski, T., Scheible, W. R., Udvardi, M., et al. (2010). Gene expression profiling identifies two regulatory genes controlling dormancy and ABA sensitivity in arabidopsis seeds. *Plant J.* 61 (4), 611–622. doi: 10.1111/j.1365-313X.2009.04088.x

Biddulph, T., Plummer, J., Setter, T., and Mares, D. (2008). Seasonal conditions influence dormancy and preharvest sprouting tolerance of wheat (*Triticum aestivum* l.) in the field. *Field Crops Res.* 107 (2), 116–128. doi: 10.1016/j.fcr.2008.01.003

Cabral, A. L., Jordan, M. C., McCartney, C. A., You, F. M., Humphreys, D. G., MacLachlan, R., et al. (2014). Identification of candidate genes, regions and markers for pre-harvest sprouting resistance in wheat (*Triticum aestivum* l.). *BMC Plant Biol.* 14, 340–352. doi: 10.1186/s12870-014-0340-1

Cao, L., Hayashi, K., Tokui, M., Mori, M., Miura, H., and Onishi, K.. (2016). Detection of QTLs for traits associated with pre-harvest sprouting resistance in bread wheat (*Triticum aestivum* l.). *Breed. Sci.* 66 (2), 260–270. doi: 10.1270/jsbbs.66.260

Chen, G. F., Zhang, H., Deng, Z. Y., Wu, R. G., Li, D. M., Wang, M. Y., et al. (2016). Genome-wide association study for kernel weight-related traits using SNPs in a Chinese winter wheat population. *Euphytica* 212 (2), 173–185. doi: 10.1007/s10681-016-1750-y

Cho, S. K., Ryu, M. Y., Kim, J. H., Hong, J. S., and Yang, S. W. (2017). Ring E3 ligases: key regulatory elements are involved in abiotic stress responses in plants. *BMB Rep.* 50 (8), 393–400. doi: 10.5483/BMBRep.2017.50.8.128

Du, X. J., Xu, W. G., Peng, C. J., Li, C. X., Zhang, Y., and Hu, L. (2021). Identification and validation of a novel locus, qpm-3BL, for adult plant resistance to powdery mildew in wheat using multilocus GWAS. *BMC Plant Biol.* 21 (1), 357–370. doi: 10.1186/s12870-021-03093-4

Fakthongphan, J., Graybosch, R. A., and Baenziger, P. S. (2016). Combining Ability for Tolerance to Pre-Harvest Sprouting in Common Wheat (Triticum aestivum L.). *Crop Science* 56 (3), 1025–1035. doi: 10.2135/cropsci2015.08.0490

Flintham, J. E. (2000). Different genetic components control coat-imposed and embryo-imposed dormancy in wheat. *Seed Sci. Res.* 10 (1), 43–50. doi: 10.1017/S0960258500000052

Frontini, M., Soutoglou, E., Argentini, M., Bole-Feysot, C., Jost, B., Scheer, E., et al. (2005). *TAF9b* (Formerly *TAF9L*) is a bona fide TAF that has unique and overlapping roles with *TAF9*. *Mol. And Cell. Biol.* 25 (11), 4638–4649. doi: 10.1128/MCB.25.11.4638

Gawel, N., and Jarret, R. (1991). A modified CTAB DNA extraction procedure for musa and ipomoea. *Plant Mol. Biol. Rep.* 3 (9), 262–266. doi: 10.1007/bf02672076

Himi, E., Maekawa, M., Miura, H., and Noda, K. (2011). Development of PCR markers for *Tamyb10* related to r-1, red grain color gene in wheat. *Theor. Appl. Genet.* 122 (8), 1561–1576. doi: 10.1007/s00122-011-1555-2

Imtiaz, M., Ogbonnaya, F. C., Oman, J., and van Ginkel, M. (2008). Characterization of quantitative trait loci controlling genetic variation for preharvest sprouting in synthetic backcross-derived wheat lines. *Genetics* 178 (3), 1725–1736. doi: 10.1534/genetics.107.084939

Jin, S. B. (1996). *Wheat in china* (Beijing: Chinese agricultural publisher).

Kashiwakura, Y., Kobayashi, D., Jikumaru, Y., Takebayashi, Y., Nambara, E., Seo, M., et al. (2016). Highly sprouting-tolerant wheat grain exhibits extreme dormancy and cold imbibition-resistant accumulation of abscisic acid. *Plant Cell Physiol.* 57 (4), 715–732. doi: 10.1093/pcp/pcw051

Kato, K., Nakamura, W., Tabiki, T., Miura, H., and Sawada, S. (2001). Detection of loci controlling seed dormancy on group 4 chromosomes of wheat and comparative mapping with rice and barley genomes. *Theor. Appl. Genet.* 102 (6-7), 980–985. doi: 10.1007/s001220000494

Knoll, E. R., Zhu, Z. I., Sarkar, D., Landsman, D., and Morse, R. H. (2020). Kin28 depletion increases association of TFIID subunits Taf1 and Taf4 with promoters in saccharomyces cerevisiae. *Nucleic Acids Res.* 48 (8), 4244–4255. doi: 10.1093/nar/gkaa165

Kulwal, P. L., Singh, R., Balyan, H. S., and Gupta, P. K. (2004). Genetic basis of pre-harvest sprouting tolerance using single-locus and two-locus QTL analyses in bread wheat. *Funct. Integr. Genomics* 4 (2), 94–101. doi: 10.1007/s10142-004-0105-2

Lang, J., Fu, Y., Zhou, Y., Cheng, M., Deng, M., Li, M., et al. (2021). *Myb10-d* confers PHS-3D resistance to pre-harvest sprouting by regulating NCED in ABA biosynthesis pathway of wheat. *New Phytol.* 230 (5), 1940–1952. doi: 10.1111/nph.17312

Lin, Y., Liu, S. H., Liu, Y. X., Liu, Y. J., Chen, G. Y., Xu, J., et al. (2017). Genome-wide association study of pre-harvest sprouting resistance in Chinese wheat founder parents. *Genet. Mol. Biol.* 40 (3), 620–629. doi: 10.1590/1678-4685-gmb-2016-0207

Liu, S., and Bai, G. (2010). Dissection and fine mapping of a major QTL for preharvest sprouting resistance in white wheat *Rio blanco*. *Theor. Appl. Genet.* 121 (8), 1395–1404. doi: 10.1007/s00122-010-1396-4

Liu, S., Li, J., Wang, Q., Zhu, X. G., Liu, L., Hu, X. R., et al. (2014). Germplasm screening for resistance to pre-harvest sprouting in southwest china. *Southwest China J. Agric. Sci.* 27 (3), 931–937. doi: 10.16213/j.cnki.scjas.2014.03.060

Li, M., Zhang, Y. W., Xiang, Y., Liu, M. H., and Zhang, Y. M. (2022a). IIIVmrMLM: The r and c++ tools associated with 3VmrMLM, a comprehensive GWAS method for dissecting quantitative traits. *Mol. Plant* 15 (8), 1251–1253. doi: 10.1016/j.molp.2022.06.002

Li, M., Zhang, Y. W., Zhang, Z. C., Xiang, Y., Liu, M. H., Zhou, Y. H., et al. (2022b). A compressed variance component mixed model for detecting QTNs, and QTN-by-environment and QTN-by-QTN interactions in genome-wide association studies. *Mol. Plant* 15 (4), 630–650. doi: 10.1016/j.molp.2022.02.012

Majeed, U., Darwish, E., Rehman, S. U., and Zhang, X. (2018). Kompetitive allele specific PCR (KASP): A singleplex genotyping platform and its application. *Agric. Sci.* 1 (11), 1–11. doi: 10.5539/jas.v11n1p11

Martínez-Andújar, C., Ordiz, M. I., Huang, Z. L., Nonogaki, M., Beachy, R. N. , and Nonogaki, H. (2011). Induction of 9-cis-epoxycarotenoid dioxygenase in *Arabidopsis thaliana* seeds enhances seed dormancy. *Proc. Natl. Acad. Sci. United States America* 108 (41), 17225-17229. doi: 10.1073/pnas.1112151108

Mohan, A., Kulwal, P., Singh, R., Kumar, V., Mir, R. R., Kumar, J., et al. (2009). Genome-wide QTL analysis for pre-harvest sprouting tolerance in bread wheat. *Euphytica* 168 (3), 319–329. doi: 10.1007/s10681-009-9935-2

Mori, M., Uchino, N., Chono, M., Kato, K., and Miura, H. (2005). Mapping QTLs for grain dormancy on wheat chromosome 3A and the group 4 chromosomes, and their combined effect. *Theor. Appl. Genet.* 110 (7), 1315–1323. doi: 10.1007/s00122-005-1972-1

Nonogaki, M., Sall, K., Nambara, E., and Nonogaki, H. (2014). Amplification of ABA biosynthesis and signaling through a positive feedback mechanism in seeds. *Plant J.* 78 (3), 527–539. doi: 10.1111/tpj.12472

Osa, M., Kato, K., Mori, M., Shindo, C., Torada, A., and Miura, H.. (2003). Mapping QTLs for seed dormancy and the Vp1 homologue on chromosome 3A in wheat. *Theor. Appl. Genet.* 106 (8), 1491–1496. doi: 10.1007/s00122-003-1208-1

Paz-Ares, J., Ghosal, D., Wienand, U., Peterson, P. A., and Saedler, H. (1987). The regulatory cl locus of *Zea mays* encodes a protein with homology to myb proto-oncogene products and with structural similarities to transcriptional activators. *EMBO J.* 6 (12), 3553–3558. doi: 10.1002/j.1460-2075.1987.tb02684.x

Rakoczy-Trojanowska, M., Krajewski, P., Bocianowski, J., Schollenberger, M., Wakuliński, W., Milczarski, P., et al. (2017). Identification of single nucleotide polymorphisms associated with brown rust resistance, α-amylase activity and pre-harvest sprouting in rye (*Secale cereale* l.). *Plant Mol. Biol. Rep.* 35 (3), 366–378. doi: 10.1007/s11105-017-1030-6

Ren, W. L., Wen, Y. J., Dunwell, J. M., and Zhang, Y. M. (2018). PKWmEB: Integration of kruskal-Wallis test with empirical bayes under polygenic background control for multi-locus genome-wide association study. *Heredity* 120 (3), 208–218. doi: 10.1038/s41437-017-0007-4

Semagn, K., Babu, R., Hearne, S., and Olsen, M. (2014). Single nucleotide polymorphism genotyping using kompetitive allele specific PCR (KASP): Overview of the technology and its application in crop improvement. *Mol. Breed.* 33 (1), 1–14. doi: 10.1007/s11032-013-9917-x

Shu, K., Liu, X. D., Xie, Q., and He, Z. H. (2016). Two faces of one seed: hormonal regulation of dormancy and germination. *Mol. Plant* 9 (1), 34–45. doi: 10.1016/j.molp.2015.08.010

Sydenham, S. L., and Barnard, A. (2018). Targeted haplotype comparisons between south african wheat cultivars appear predictive of pre-harvest sprouting tolerance. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00063

Tamba, C. L., Ni, Y., and Zhang, Y. (2017). Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. *PloS Comput. Biol.* 13 (1), 1–20. doi: 10.1371/journal.pcbi.1005357

Tamba, C. L., and Zhang, Y. M. (2018). A fast mrMLM algorithm for multi-locus genome-wide association studies. *bioRxiv*, 1–34. doi: 10.1101/341784

Wang, S., Feng, J., Ren, W., Huang, B., Zhou, L., Wen, Y. J., et al. (2016). Improving power and accuracy of genome-wide association studies *via* a multi-locus mixed linear model methodology. *Sci. Rep.* 6 (1), 1–10. doi: 10.1038/srep19444

Wang, J., Liu, Y., Wang, Y., Chen, Z., Dai, S., Cao, W. G., et al. (2011). Genetic variation of *Vp1* in sichuan wheat accessions and its association with pre-harvest sprouting response. *Genes Genomics* 33 (2), 139–146. doi: 10.1007/s13258-010-0125-3

Wang, S., Zhu, Y., Zhang, D., Shao, H., Liu, P., Hu, J. B., et al. (2017). Genome-wide association study for grain yield and related traits in elite wheat varieties and advanced lines using SNP markers. *PloS One* 12 (11), 1–14. doi: 10.1371/journal.pone.0188662

Wen, Y. J., Zhang, H. W., Ni, Y. L., Huang, B., Zhang, J., Feng, J. Y., et al. (2017). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Briefings Bioinf.* 18 (5), 906–906. doi: 10.1093/bib/bbx028

Yang, J. F. (2015). *Preliminary functional analysis of gene CpTAF9 and gene CpTAF10 from chimonanthus praecox (L.) link* (Chongqing, China: Southwest University).

Yang, L., Liu, Q. H., Liu, Z. B., Yang, H., Wang, J. M., Li, X., et al. (2016). *Arabidopsis* C3HC4-RING finger E3 ubiquitin ligase at *AIRP4* positively regulates stress-responsive abscisic acid signaling. *J. Integr. Plant Biol.* 58 (1), 67–80. doi: 10.1111/jipb.12364

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* 9 (11), 1–16. doi: 10.1371/journal.pone.0113287

Zhang, H. P., Chang, C., You, G. X., Zhang, X. Y., Yan, C. S., Xiao, S. H., et al. (2010). Identification of molecular markers associated with seed dormancy in mini core

collections of Chinese wheat and landraces. *Acta Agronomica Sin.* 36 (10), 1649–1656. doi: 10.1016/S1875-2780(09)60077-8

Zhang, J., Feng, J. Y., Ni, Y. L., Wen, Y. J., Niu, Y., Tamba, C. L., et al. (2017). PLARmEB: Integration of least angle regression with empirical bayes for multilocus genome-wide association studies. *Heredity* 118 (6), 517–524. doi: 10.1038/hdy.2017.8

Zhang, Y., Tamba, C. L., Wen, Y. J., Li, P., Ren, W. L., Ni, Y. L., et al. (2020). mrMLM v4.0.2: An r platform for multi-locus genome-wide association studies. *Genomics Proteomics Bioinf.* 18 (4), 481–487. doi: 10.1016/j.gpb.2020.06.006

Zhou, Y., Tang, H., Cheng, M. P., Dankwa, K. O., Chen, Z. X., Li, Z. Y., et al. (2017). Genome-wide association study for pre-harvest sprouting resistance in a large germplasm collection of chinese wheat landraces. *Front. Plant Sci.* 08. doi: 10.3389/fpls.2017.00401

Zhu, Y., Wang, S. X., Wei, W. X., Xie, H. Y., Liu, K., Zhang, C., et al. (2019). Genome-wide association study of pre-harvest sprouting tolerance using a 90K SNP array in common wheat (*Triticum aestivum* l.). *Theor. Appl. Genet.* 132 (11), 2947–2963. doi: 10.1007/s00122-019-03398-x

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 14 February 2023
DOI 10.3389/fpls.2023.1120498

OPEN ACCESS

EDITED BY
Awais Rasheed,
Quaid-i-Azam University, Pakistan

REVIEWED BY
Yingpeng Han,
Northeast Agricultural University, China
Muhammad Qasim Shahid,
South China Agricultural University, China

*CORRESPONDENCE
Bin Li
libin02@caas.cn
Lijuan Qiu
qiulijuan@caas.cn
Junming Sun
sunjunming@caas.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Functional and Applied Plant Genomics, a section of the journal Frontiers in Plant Science

RECEIVED 10 December 2022
ACCEPTED 01 February 2023
PUBLISHED 14 February 2023

CITATION
Azam M, Zhang S, Li J, Ahsan M, Agyenim-Boateng KG, Qi J, Feng Y, Liu Y, Li B, Qiu L and Sun J (2023) Identification of hub genes regulating isoflavone accumulation in soybean seeds *via* GWAS and WGCNA approaches.
Front. Plant Sci. 14:1120498.
doi: 10.3389/fpls.2023.1120498

Identification of hub genes regulating isoflavone accumulation in soybean seeds *via* GWAS and WGCNA approaches

Muhammad Azam[1†], Shengrui Zhang[1†], Jing Li[1†], Muhammad Ahsan[1], Kwadwo Gyapong Agyenim-Boateng[1], Jie Qi[1], Yue Feng[1], Yitian Liu[1], Bin Li[2*], Lijuan Qiu[3*] and Junming Sun[1*]

[1]The National Engineering Research Center of Crop Molecular Breeding, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, China, [2]Ministry of Agriculture and Rural Affairs (MARA) Key Laboratory of Soybean Biology (Beijing), Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, China, [3]The National Key Facility for Crop Gene Resources and Genetic Improvement (NFCRI)/Key Laboratory of Germplasm and Biotechnology Ministry of Agriculture and Rural Affairs (MARA), Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, China

Introduction: Isoflavones are the secondary metabolites synthesized by the phenylpropanoid biosynthesis pathway in soybean that benefits human and plant health.

Methods: In this study, we have profiled seed isoflavone content by HPLC in 1551 soybean accessions grown in Beijing and Hainan for two consecutive years (2017 and 2018) and in Anhui for one year (2017).

Results: A broad range of phenotypic variations was observed for individual and total isoflavone (TIF) content. The TIF content ranged from 677.25 to 5823.29 μg g^{-1} in the soybean natural population. Using a genome-wide association study (GWAS) based on 6,149,599 single nucleotide polymorphisms (SNPs), we identified 11,704 SNPs significantly associated with isoflavone contents; 75% of them were located within previously reported QTL regions for isoflavone. Two significant regions on chromosomes 5 and 11 were associated with TIF and malonylglycitin across more than 3 environments. Furthermore, the WGCNA identified eight key modules: black, blue, brown, green, magenta, pink, purple, and turquoise. Of the eight co-expressed modules, brown (r = 0.68***), magenta (r = 0.64***), and green (r = 0.51**) showed a significant positive association with TIF, as well as with individual isoflavone contents. By combining the gene significance, functional annotation, and enrichment analysis information, four hub genes *Glyma.11G108100*, *Glyma.11G107100*, *Glyma.11G106900*, and *Glyma.11G109100* encoding, basic-leucine zipper (bZIP) transcription factor, MYB4 transcription factor, early responsive to dehydration, and PLATZ transcription factor respectively were identified in brown and green modules. The allelic variation in *Glyma.11G108100* significantly influenced individual and TIF accumulation.

Discussion: The present study demonstrated that the GWAS approach, combined with WGCNA, could efficiently identify isoflavone candidate genes in the natural soybean population.

KEYWORDS

soybean, isoflavone, genome-wide association study (GWAS), WGCNA, RNA-Seq

1 Introduction

Soybean isoflavones are of great importance because of their positive impact on human health, including the treatment and prevention of various types of cancers (prostate cancer, breast cancer etc.) (Nielsen and Williamson, 2007; Phetnoo et al., 2013), cardiovascular disease, osteoporosis, and metabolic syndrome (Cai et al., 2004; Mozaffarian et al., 2011; Bradbury et al., 2014). In plants, isoflavones can resist adverse stress and promote the growth and reproduction of rhizobia, root nodule development, and nitrogen fixation (Sugiyama et al., 2017; Darwish et al., 2022; Wang et al., 2022). Soybean seed isoflavones contain 12 components which are divided into four groups, daidzein, genistein, glycitein (aglycones), daidzin, glycitin, genistin (glycosides), acetyldaidzin, acetylglycitin, and acetylgenistin (acetylglycosides), and malonyldaidzin, malonylglycitin, malonylgenistin (malonylglycosides) (Kim et al., 2014; Azam et al., 2021). The malonyldaidzin, malonylglycitin, and malonylgenistin are the most abundant form of the isoflavones, while aglycones are present in very small amounts but have higher phytoestrogenic activity and more bioavailability in humans (Nielsen and Williamson, 2007; Park et al., 2016; Azam et al., 2020). Improving soybean isoflavone content through conventional breeding and metabolic engineering is a complementary way for the biofortification of food crops to combat isoflavone deficiency (De Steur et al., 2014).

Isoflavone content is controlled by multiple genes, and there are often complex interaction mechanisms among various enzyme genes in its synthesis path, which jointly determine isoflavone biosynthesis. The metabolic pathway controlling the synthesis of soybean isoflavones in plants is very complex (Wang and Murphy, 1994; Bennett et al., 2004). The synthesis of soybean isoflavones starts from the synthesis of phenylpropionic acid. The original substrate of isoflavones is phenylalanine, which is catalyzed by phenylalanine lyase (PAL), cinnamate-4-hydroxylase (C4H), and 4-coumarin coenzyme A ligase (4CL), respectively to produce p-coumaroyl COA, Isoliquiritigenin chalcone and chalcone were formed with malonyl COA of 3 molecules under the co catalysis of chalcone synthase (CHS) and chalcone reductase (CHR). Isoliquiritigenin chalcone is catalyzed by chalcone isomerase (CHI) to produce liquiritigenins (Ralston et al., 2005), which are then catalyzed by isoflavone synthase genes (*IFS1 and IFS2*) to their corresponding isoflavones (Akashi et al., 1999; Jung et al., 2000; Dhaubhadel et al., 2003). Among isoflavone synthase genes, *IFS2* has a higher expression level in the embryo and seed pods, while *IFS1* has higher expression in roots and seed coats. In addition, various kind of MYB transcription factors (CCA1, R2R3, and R1) helps in isoflavone accumulation by regulating the isoflavone synthesis genes related to phenylpropanoid biosynthesis pathways (Bian et al., 2018; Sarkar et al., 2019). The R2R3-MYB transcription factor *GmMYB29*, *GmMYB102*, *GmMYB280*, *MYB502*, *GmMYB100* regulate isoflavone accumulation by activating the *IFS1*, *IFS2* and *CHS8* enzymes (Yan et al., 2015; Sarkar et al., 2019). The CCA1-like R1 MYB transcription factor *GmMYB133* regulates isoflavone biosynthesis by activating the promoters of *CHS8* and *IFS2* (Bian et al., 2018). A dual-function C2H2 zinc-finger transcription factor *GmZFP7* has recently been shown to divert metabolic flow to isoflavone by increasing the expression of *GmC4H*, *Gm4CL*, *GmCHS*, *GmCHR*, and *GmIFS2* while decreasing the expression of *GmF3H1* in soybean seeds. (Feng et al., 2023).

Soybean isoflavones are quantitative traits regulated by multiple genes. The genotyping by sequencing (GBS) approach and SNP genotyping have substantially expanded the application of GWAS to soybeans (Lee et al., 2015; Sonah et al., 2015; Torkamaneh and Belzile, 2015). Natural population based GWAS have more recombination events than biparental populations, resulting in less short LD regions and higher precision and accuracy of marker phenotype association (Duan et al., 2022; Liang et al., 2022). These approaches have been utilized in GWAS to identify genomic regions associated with resistance to biotic and abiotic stress, including soybean cyst nematode, abiotic stress, seed quality traits such as oil and protein content, and yield related traits (Hwang et al., 2014; Cao et al., 2017; Zeng et al., 2017; Zhao et al., 2017). Furthermore, weighted gene co-expression network (WGCNA) analysis is a powerful tool for describing gene expression correlations using microarray or RNA-seq data. The WGCNA is an effective method to narrow down the range of candidate genes (Schaefer et al., 2018). Recently, GWAS combined with WGCNA has been applied to identify the genes responsible for salt tolerance in maize, silique length in *Brassica napus*, and root growth dynamics in rapeseed (Li et al., 2021; Ma et al., 2021; Wang et al., 2021). However, no study has used the GWAS and the WGCNA to explain the gene networks and molecular regulatory mechanisms that govern isoflavone regulation in soybean. Therefore, the present study aimed to identify the genomic regions and candidate genes involved in the isoflavone biosynthesis pathway using GWAS coupled with WGCNA in 1551 soybean accessions.

2 Research materials and methods

2.1 Planting materials

A total of 1551 natural population panel of diverse soybean accessions was used in this study. The accessions were selected from a mini core collection developed by Qiu et al. (2009) based on their availability at the soybean genetic resource research group of the Institute of Crop Sciences, Chinese Academy of Agricultural Sciences (CAAS). The origin and number of soybean accessions from each country are Brazil (8), Canada (6), China (1283), Colombia (1), East Europe (3), Germany (4), Italy (2), Japan (21), Nigeria (1), North Korea (1), Russia (22), South Korea (4), Thailand (1), USA (194). Information on each accession is also presented in Supplementary Table 1. Field trials were conducted at three locations (Changping, Beijing (40° 13′ N and 116° 12′ E), Sanya, Hainan (18° 24′ N and 109° 5′ E) in 2017 and 2018, while, for only 2017, planted in Hefei, Anhui (33°61′ N and 117 °E). A randomized incomplete block design was employed to sow the cultivars, with the various planting sites serving as replications. The cultivars were replicated across different sites due to a large number of cultivars and the scarcity of available land resources. Each cultivar's seeds were sown in 3 m long rows with 0.5 m inter-row and 0.1 m intra-row spacing. Fertilizer containing 30 kg/ha, 40 kg/ha, and 60 kg/ha of nitrogen, phosphorous, and

potassium was applied to the field, respectively. From planting until harvest, the advised agronomic procedures were used. The seeds from each accession were pooled and used for soybean seed isoflavone determination (Azam et al., 2020; Azam et al., 2021).

2.2 Extraction and quantification of isoflavones

The isoflavone contents were determined using a previously reported method (Sun et al., 2011) and as follows. Around 20 g seeds of each accession were grounded by a cyclone mill (IKA, A10 basic, Rheinische, Germany). Approximately 0.1 g of the finely ground powder was placed in a 10 mL tube pre-filled with 5 mL of a solution containing 0.1% (v/v) acetic acid and 70% (v/v) ethanol and shaken for 12 hours on a twist mixer (TM – 300, AS ONE, Osaka, Japan). The mixture was centrifuged for 10 min at 6000 rpm, and the supernatant was filtered using a 0.2 µm YMC Duo filter (YMC Co., Kyoto, Japan). Samples were stored at 4°C prior to use and measured for isoflavones using an Agilent HPLC system (Agilent 1260, Santa Clara, CA, USA) having YMC ODS AM-303 column (250 mm × 4.6 mm I.D., S-5 µm, 120 Å, YMC Co., Kyoto, Japan). The identification and quantification of the isoflavone contents were carried out using the following isoflavone standards: daidzein (DE), glycitein (GLE), genistein (GE), daidzin (D), glycitin (GL), genistin (G), malonyldaidzin (MD), malonylglycitin (MGL), malonylgenistin (MG), acetyldaidzin (AD), acetylglycitin (AGL), and acetylgenistin (AG). The detected isoflavone component concentrations were determined using the formula provided by (Sun et al., 2011).

2.3 Association analysis and candidate gene prediction and annotation

A total number of 6,149,599 SNPs with MAF 0.01 from previously sequenced 2,241 soybean accessions were used for GWAS analysis (Li et al., 2022). GWAS was performed using the compressed mixed linear model (cMLM) in the GAPIT program (Lipka et al., 2012), where the first three principal component analysis (PCA) values were included as fixed effects in the mixed model to correct for stratification. The threshold for significance was estimated to be approximately $P = 1 \times 10^{-6}$ (that is, 1/6,149,599) by the Bonferroni correction method. These 6,149,599 SNPs were distributed equally across the 20 soybean chromosomes (one SNP per 154.3 bp). The extent of model fitting was confirmed using a quantile-quantile (Q-Q) plot for the expected and obtained *p*-values of each SNP to evaluate how much a significant result was produced by the analysis than expected by chance. The Manhattan plots for the isoflavone contents for each of the five environments were generated from GAPIT (Lipka et al., 2012). The Phytozome database (http://www.phytozome.org/) and the SoyBase database (http://www.soybase.org/) were used to predict and annotate the candidate genes.

2.4 RNA seq-analysis

The four soybean varieties Luheidou (LHD), Zhonghuang 13 (ZH13), Zhonghuang 35 (ZH35), and Nanhuizao (NHZ), varying in their isoflavone contents, were used as materials for RNA seq-analysis. About 20 seeds were harvested at different developmental stages (R5 to R8) after 7 days intervals. Each sample was set with three replications for isoflavone contents, and RNA extraction. The total RNAs were extracted using the TRIzol method. The high-quality RNA samples were sent for RNA-seq analysis to BLgene co. LTD (Beijing, China). HISAT2 was used to map the clean RNA-seq data onto the reference genome (Kim et al., 2015). FeatureCounts calculated the transcriptional abundance and gene expression count matrix (Liao et al., 2014). TPM (transcripts per million) was used as the expression level, and log10 (TPM + 1) was used to standardize it (Feng et al., 2023).

2.5 Weighted gene co-expression network analysis

The transcriptome data of (LHD, NHZ, ZH13, and ZH35) at different seed developmental stages was used for the WGCNA. The R WGCNA (v1.47) package was used to create the weighted gene co-expression network (Langfelder and Horvath, 2008). The gene expression values were imported into WGCNA to construct co-expression modules using the automatic network construction with default settings. The phenotype data was imported into the WGCNA package, and correlation-based connections between phenotypes and gene modules were computed using the default settings. Pearson's correlation between all gene pairs was first determined to create a matrix of adjacencies. Using the TOM similarity function, this matrix was transformed into a Topological Overlap Matrix (TOM) (Zhang and Horvath, 2005). Finally, modules on the dendrogram were discovered using the R package dynamicTreeCut method (Langfelder et al., 2008). The hub genes are usually characterized by high gene significance (GS, association between gene expression and traits) and module membership (MM, correlation between gene expression and module eigengene) values.

2.6 Gene ontology analysis

The GO enrichment analysis was performed to identify GO categories based on the SoyBase database (http://soybase.org/) and detect those over/under-represented. The significant enriched GO terms ($P < 0.05$) for biological processes, the cellular process, and molecular processes were further identified using PlantRegMap online tool (http://plantregmap.cbi.pku.edu.cn/go_result.php) and were visualized REVIGO (http://revigo.irb.hr/) (Supek et al., 2011).

3 Results

3.1 Variations among seed isoflavone contents in soybean natural population

The individual and TIF content was profiled in soybean accessions collected from distinct regions of China and other countries that have grown across three locations over two years. The mean TIF content of the 1551 soybean natural population grown across five environments is presented in Supplementary Table 1. The mean TIF content of the soybean accessions ranged from 677.25 to

5823.29 μg g^{-1} (Azam et al., 2020; Azam et al., 2021). The individual and TIF content of the soybean accessions in five environments are presented in Figure 1. The correlations among the five environments for individual and TIF content are presented in Supplementary Figure 1. The higher levels of daidzin (172.7 μg g^{-1}), genistin (290 μg g^{-1}) were observed in Hainan 2018, followed by Hainan 2017 (daidzin (152.4 μg g^{-1}, genistin (218.8 μg g^{-1}). The higher levels of malonyldaidzin (888.3 μg g^{-1}), malonylgenistin (1574.1μg g^{-1}), and TIF (3012.3 μg g^{-1}) were observed in Beijing 2017, followed by Hainan 2017 (malonyldaidzin (789.9 μg g^{-1}), malonylgenistin (1183.1 μg g^{-1}) and TIF (2685.5 μg g^{-1}), while Anhui 2017 showed lower levels of these components (malonyldaidzin (589.2 μg g^{-1}), malonylgenistin (984.1 μg g^{-1}) and TIF (2153.1 μg g^{-1}). While higher levels of malonylglycitin (208.2 μg g^{-1}) were observed in Hainan 2017, followed by Anhui 2017 (168.2 μg g^{-1}) and lowest in Hainan 2018 (100.1 μg g^{-1}) (Figure 1).

Furthermore, Pearson's correlation was performed to reveal the association between individual and TIF content. TIF content was positively associated with individual isoflavone contents (Figure 2). Malonylgenistin, Malonyldaidzin, genistin, and daidzin showed the highest correlation with TIF content (r = 0.93***, r = 0.91***, r = 0.89***, r = 0.82***, respectively), followed by malonylglycitin and glycitin (r = 0.48***, r = 0.47***, respectively). Furthermore, glycosides showed highly significant positive correlations with their respective malonylglycosides, genistin and malonylgenistin (r = 0.90***), daidzin and malonyldaidzin (r = 0.89***), and glycitin and malonylglycitin (r = 0.87***) (Figure 2).

3.2 GWAS reveals candidate loci underlying seed isoflavone contents

The phenotypic and genotypic data for 1551 diverse soybean accessions were used for GWAS analysis to identify putative loci associated with isoflavone contents in the individual environment (Hainan 2017, Hainan 2018, Beijing 2017, Beijing 2018, and Anhui 2017). The principal component analysis (PCA) was used for scanning the population stratification. The landrace group overlapped partially with the improved cultivar group, indicating a broad genetic variation within this set of 1551 soybean accessions. Meanwhile, clear clustering based on planting region was observed; the first two PCs accounted for 40.47% of the genetic variation, demonstrating that the first two PCs uncommonly affect the mapping

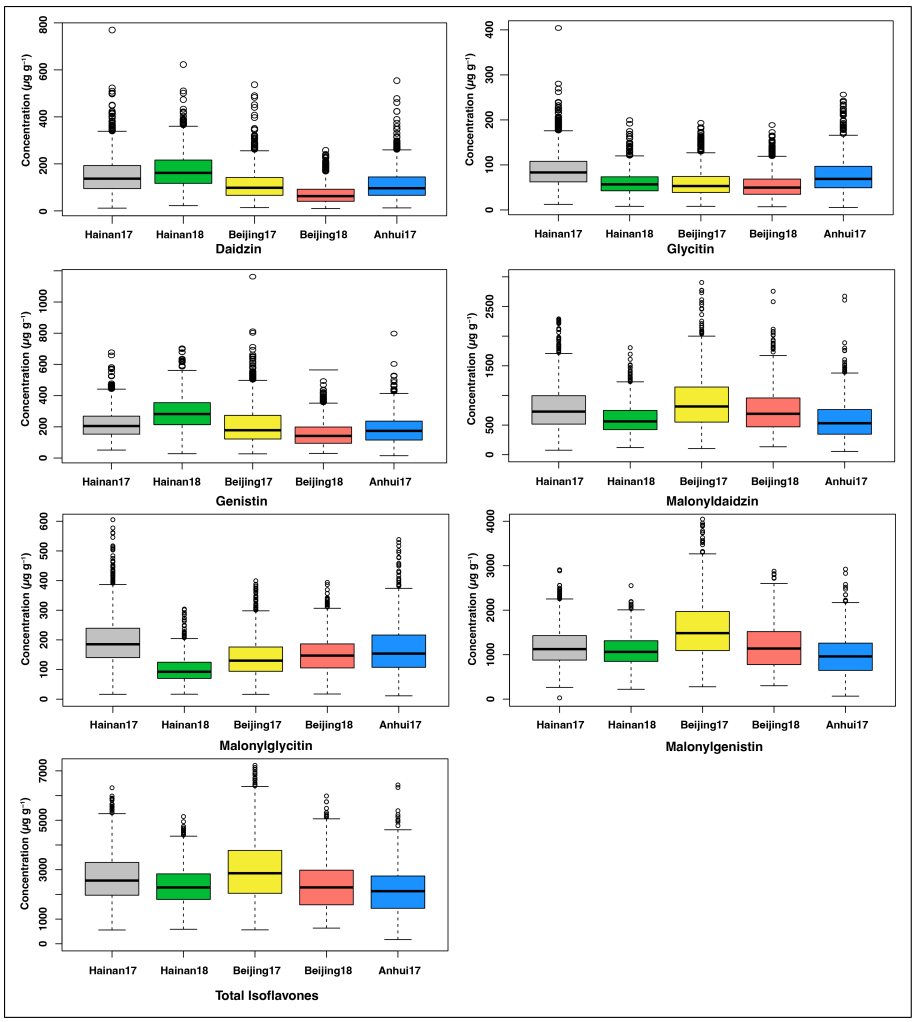


FIGURE 1
Individual and TIF content in five environments (Hainan 2017 Hainan 2018, Beijing 2017, Beijing 2018, and Anhui 2017).

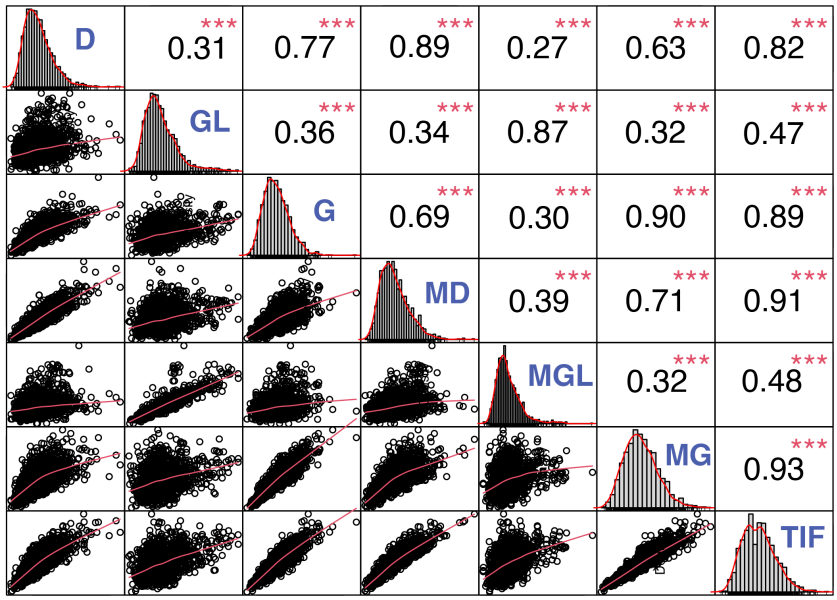


FIGURE 2
Correlation analysis among the individual and TIF content in soybean seeds. *, **, and *** represent significance at $p < 0.05$, 0.01, and 0.001, respectively. D, Daidzin; GL, Glycitin; G, Genistin; MD, Malonyldaidzin; MGL, Malonylglycitin; MG, Malonylgenistin; TIF, Total isoflavone.

population. The average distance over which LD decays to half of its maximum value in soybean was 97kb (Supplementary Figures 2A, B) GWAS identified 11704 genome-wide distributed SNPs that were significantly (-log10P>6) associated with isoflavone levels with P-values ranging from 9.99e-07 to 7.30e-30, the detailed information is listed in Supplementary Table 2. Of the 11704 significant SNPs, 53.8% were annotated in intergenic regions, 19.9% in the upstream and downstream regions, 14% in the intron regions. Herein, 8786 SNPs (75%) identified from the GWAS were located within the regions of previously reported QTLs for isoflavone in soybean. In total, 2,018 known genes were mapped by the significant SNPs, which include 29 isoflavone biosynthesis enzymes and 18 MYB transcription factors; of these, 417, 261, 316, 428, 307, 847, and 230 genes were significantly associated with daidzin, glycitin, genistin, malonyldaidzin, malonylgenistin, malonylglycitin, and TIF content, respectively (Supplementary Tables 2, 3). Interestingly, a significant region (8147595 to 8315102bp) has been identified on chromosome 11 across four environments associated with malonylglycitin and contains 18 genes (Figures 3A, B), including eight enzymes and three transcription factors MYB (1), bZIP (1) and zinc finger (1). Furthermore, a significant region on Chromosome 5 related to TIF content across three environments spanning from 41760764 to 42234431 bp encoded 63 candidate genes (Figures 3C, D), including seven key enzymes, and four transcription factors WD40 (1), bZIP (1) and zinc finger (2) (Supplementary Tables 4, 5).

3.3 Identification of key modules possessing candidate genes *via* WGCNA

The transcriptome data of different seed developmental stages were used for WGCNA, which provided new genomic insights to better understand the molecular mechanisms underlying isoflavone accumulation in soybean seed. The candidate genes identified in the linkage disequilibrium regions obtained through GWAS analysis were blast searched against the transcriptome data of soybean cultivars collected at different seed developmental stages (R5-R8) to identify common genes for WGCNA analysis. The WGCNA identified eight key modules, namely, black, blue, brown, green, magenta, pink, purple, and turquoise, possessing 253, 1251, 316, 426, 82, 113, 83, and 1275 genes, respectively (Figures 4A, B).

To further investigate the modules containing genes involved in isoflavone synthesis, Pearson's correlation analysis was performed. Of the eight co-expressed modules, brown (r = 0.68***), magenta (r = 0.64***), and green (r = 0.51**) showed significant positive correlations with TIF, as well as with individual isoflavone contents. The sample dendrogram and trait heat map also revealed that the isoflavone accumulation is higher at late seed developmental stages (Figures 4C, D). Furthermore, genes in brown, magenta, and green modules showed higher expression patterns at late seed developmental stages. It is already established that higher isoflavone accumulations were observed in the soybean seeds at later developmental stages (Figure 5). To further investigate the relationship of genes in each of the positive modules with isoflavone synthesis, the correlation between gene significance (GS) and module membership (MM) was carried out. Out of 8 modules, the brown module showed a highly positive correlation with TIF (r = 0.71***), followed by magenta (r = 0.7***), while the lowest was observed in the green module (r = 0.44***) (Supplementary Figure 3). Furthermore, the GO enrichment analysis revealed that the brown module possesses genes linked to defense response to bacterium (GO:0042742), defense response to other organism (GO:0098542), defense response, incompatible interaction (GO:0009814), response to reactive oxygen species (GO:0000302). Similarly, genes present in the magenta module are response to stress (GO:0006950), response to water deprivation (GO:0009414), cellular response to red or far-red

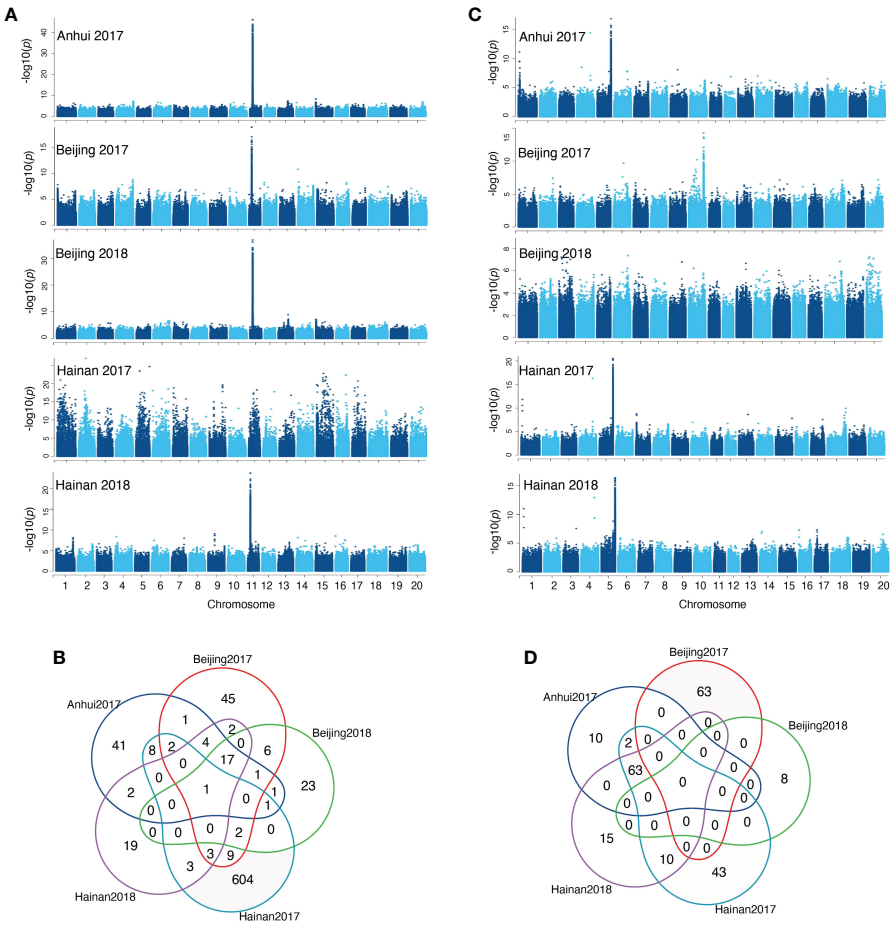


FIGURE 3
(A) Manhattan plots of malonylglycitin for five environments, **(B)** Venn plot for malonylglycitin genes in five environments, **(C)** Manhattan plots of TIF content for five environments, **(D)** Venn plot for TIF content genes in five environments.

light (GO:0071489) enzyme regulator activity (GO:0030234), and genes in the green module are involved in the regulation of circadian rhythm (GO:0042752), response to UV (GO:0009411), response to salt stress (GO:0009651) are engaged in biotic and abiotic stresses (Figures 6A-D). Current results suggest that genes present in the above-mentioned modules, i.e., brown, magenta, and green, might be involved in isoflavone accumulation in soybean seeds; they can play important roles in the isoflavone synthesis pathway.

Further, gene annotation and gene significance information were used to identify hub genes in brown, magenta, and green modules. Based on the gene significance and annotation information, 27 key candidate genes were identified and are presented in Table 1. These candidate genes include 9 transcription factors (4 MYB, 3 WD40, 1 WRKY, and 1 Zinc finger) and 12 key enzymes, including glucosyl transferases, isoflavone 2'-hydroxylase, etc. Two MYB transcription factors in the brown module, *MYB133* (*Glyma.07G066100*) and *MYB121* (*Glyma.15G176000*) were identified as positive regulators of isoflavone biosynthesis from previous studies. While in the magenta module, we identified a cytochrome P450 enzyme, isoflavone 2'-hydroxylase (*Glyma.16G149300*), a positive regulator of isoflavones. Interestingly, four hub genes *Glyma.11G108100*, *Glyma.11G107100*, *Glyma.11G106900*, and *Glyma.11G109100* encoding, basic-leucine zipper (bZIP) transcription factor, MYB4 transcription factor, early responsive to dehydration, and PLATZ transcription factor, respectively were identified in brown and green modules. These four hub (*Glyma.11G108100*, *Glyma.11G107100*, *Glyma.11G106900*, and *Glyma.11G109100*) genes were also present in the candidate region located on Chromosome 11 identified by GWAS and matched with previously identified QTLs. Isoflavones play an important role in biotic and abiotic stress in plants, and MYB transcription factors help in isoflavone accumulation by regulating key isoflavone synthase genes (*IFS1* and *IFS2*). Therefore, the identified transcription factors (bZIP, MYB, PLATZ) might be involved in the isoflavone accumulation as they are also helping plants to adapt to various kinds of biotic and abiotic stresses.

3.4 Natural variation in *Glyma.11G108100* contributes to isoflavone accumulation

Natural variation of *Glyma.11G108100* was identified by using the soybean functional genomics & breeding (SoyFGB v 2.0) database (https://sfgb.rmbreeding.cn/) (Zheng et al., 2022). Based on the phytozome database (https://phytozome-next.jgi.doe.gov), the coding region of *Glyma.11G108100* contains 813 nucleotides, which encodes 270 amino acids with two exons and one intron. The causal SNP was in

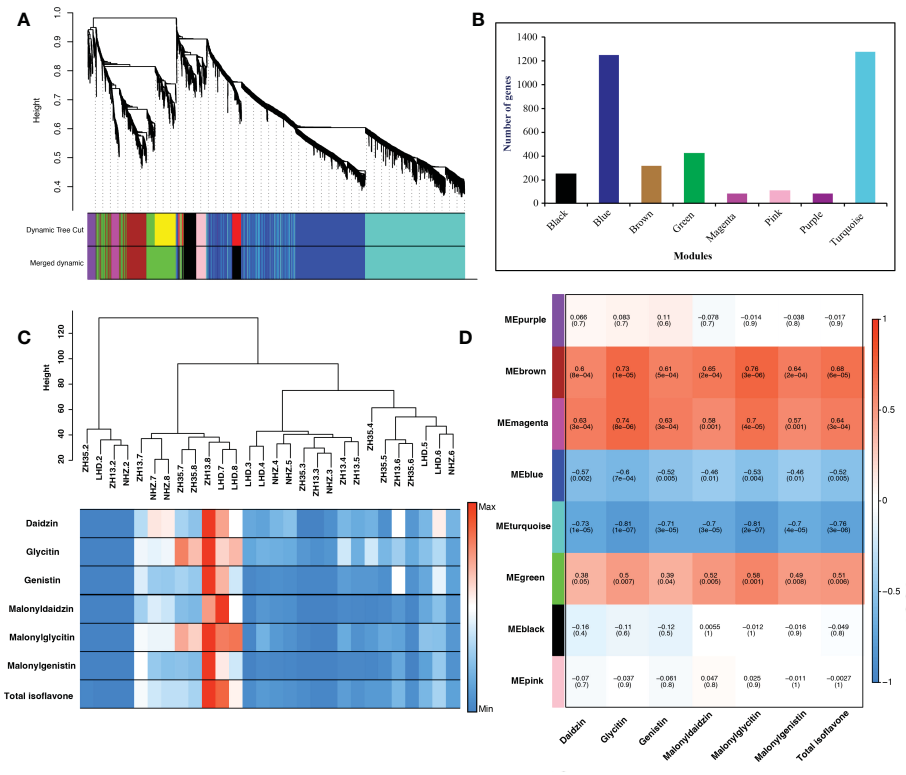


FIGURE 4
(A) Module clustering, different colors represent different modules. **(B)** Number of genes in each module. **(C)** Sample dendrogram and trait heatmap, each row corresponds to the isoflavone content, while each column corresponds to seed samples of four soybean cultivars (LHD, NHZ, ZH13, and ZH35) collected at different seed developmental stages (R5-R8). The right panel represents the minimum (blue color) and maximum (red color) isoflavones accumulation at different seed developmental stages. **(D)** Module trait relationship, each row corresponds to a module, while each column corresponds to the isoflavone content. The left panel shows the modules, while the right panel shows positive (red, 1) and negative (blue, – 1) correlations.

the exonic region (Figure 7A). Williams82 provided the reference allele (C), while the polymorphism that occurred resulted in the alternate allele (G). The geographical distribution of C and G alleles is presented in Figure 7B. The overall variation revealed significant differences in malonylglycitin content for C and G alleles which have 58% and 42% distribution in the soybean germplasm. The C allele had higher malonylglycitin content (183.3 μg g^{-1}) than the G allele (126.8 μg g^{-1}). The regional distribution of these alleles showed significant differences in malonylglycitin content in NR, HR, and SR regions. The distribution of the C allele in NR, HR, and SR regions is 37%, 63%, and 72%, respectively, while the G allele is 63%, 37%, and 28%, respectively. The C allele had higher malonylglycitin content in NR (150.4μg g^{-1}), HR (222.1μg g^{-1}), and SR (159.7μg g^{-1}) compared with the G allele (Figure 7C). Furthermore, the natural variation of *Glyma.11G108100*

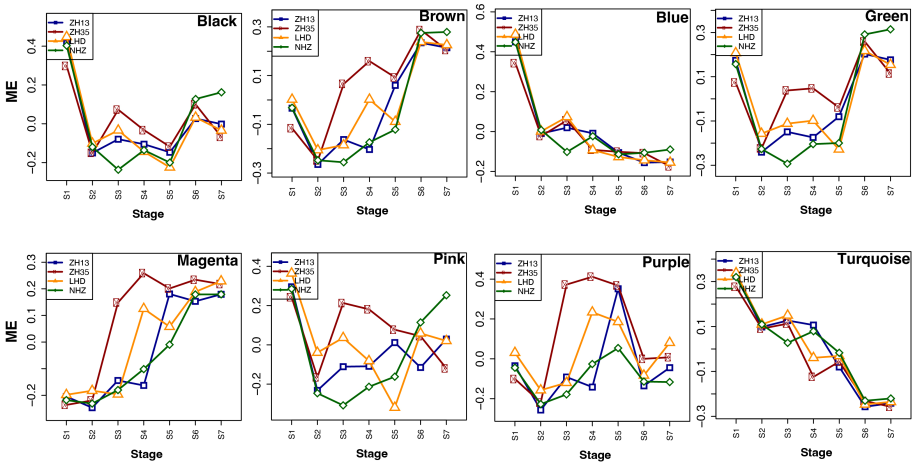


FIGURE 5
Expression profiles of the modules at different seed developmental stages in four soybean cultivars.

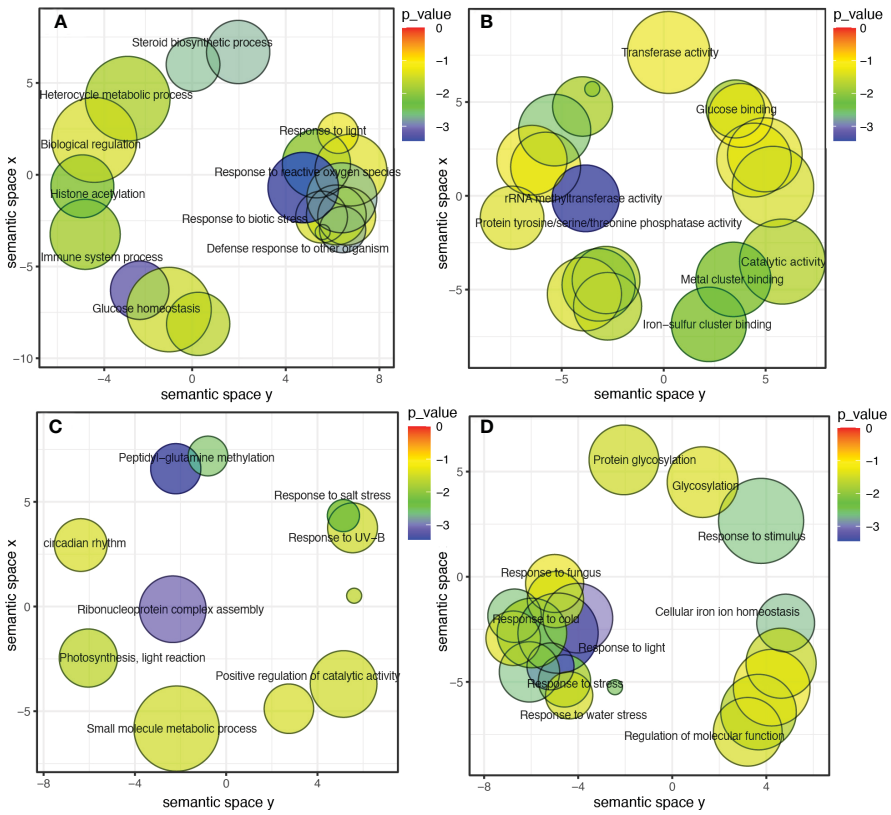


FIGURE 6
(A) GO categories for biological process, brown module. **(B)** GO categories for molecular function, brown module. **(C)** Categories for biological process, green module. **(D)** Categories for biological process, magenta module.

also influenced the TIF content accumulation in soybean seed. The overall variation revealed significant differences in TIF content for C and G alleles, with 58% and 42% distribution in the soybean germplasm. The C allele had higher TIF content (2568.8 μg g^{-1}) compared with the G allele (2387.7 μg g^{-1}). The regional distribution of these alleles showed significant differences for TIF content in the HR region, while non-significant differences for NR and SR regions. The distribution of the C allele in the HR region is 63%, and the G allele is 37%. The TIF content of the C allele (2793.9 μg g^{-1}) was significantly higher than the G allele (2509.5 μg g^{-1}) in the HR region (Figure 7D). The polymorphism in *Glyma.11G108100* showed significant variations for individual and TIF content across soybean germplasm and regions, suggesting that it might be associated with isoflavone accumulation in soybean.

4 Discussion

Soybean isoflavones are of great interest owing to their beneficial impact on plant and human health. Increasing isoflavone concentration in soybean is one of the major goals of soybean breeders; however, the narrow genetic diversity of the soybean germplasm constrains the improvement of the isoflavones (Qiu et al., 2009). In this study, we determined the isoflavone composition from the core germplasm of soybean accessions grown at three locations for two years. Significant differences were observed for individual and TIF content across different environments. The TIF concentration ranged from 677.25 to 5823.29 μg g^{-1} across all the examined environments. Malonylglycosides were identified as major isoflavone contents (Zhang et al., 2014; Azam et al., 2020). Furthermore, glycosides and malonylglycosides showed positive associations as they are synthesized by the action of key isoflavone biosynthesis enzymes (glucosyltransferase and malonyltransferase) *via* common branches in the phenylpropanoid pathway (Yu and Mcgonigle, 2005; Barnes, 2010). The phenotypic variation of individual and TIF content demonstrated significant differences among the soybean accessions, growing environments, and growing years which suggests that genetic as well as environmental factors affect isoflavone accumulation in soybean seeds (Tsai et al., 2007; Rasolohery et al., 2008; Zhang et al., 2014; Pei et al., 2018; Azam et al., 2023).

Isoflavones are typical quantitative traits; many QTLs for individual and TIF content distributed on most soybean chromosomes have been detected in several studies (Akond et al., 2013; Pei et al., 2018; Wu et al., 2020). Alternatively, genome-wide association studies (GWAS) based on the use of natural population, in contrast to linkage analysis using bi-parental populations, have more extensive recombination events and, thus, result in less short LD segments leading to increased resolution and accuracy of marker-phenotype associations (Duan et al., 2022; Liang et al., 2022). In this study, hundreds of SNPs loci were found to be significantly associated with the individual and TIF content, and they were distributed across all 20 chromosomes of soybean. Furthermore, many of these SNPs were simultaneously identified in five environments, as observed in malonylglycitin, malonylgenistin, and four environments like total isoflavones, malonyldaidzin, malonylgenistin, malonylglycitin, etc. Most of the significantly associated SNPs were observed for

10.3389/fpls.2023.1120498

TABLE 1 List of candidate genes for individual and TIF content in brown, magenta, and green modules.

Gene ID	Module	GS.TIF	*p*.GS.TIF	Annotation
Glyma.11G108100	Brown	0.77	1.73E-08	Basic-leucine zipper (bZIP) transcription factor
Glyma.17G085800	Brown	0.76	2.02E-06	S-adenosyl-L-methionine methyltransferase
Glyma.07G100700	Brown	0.76	2.20E-06	MYB transcription factor
Glyma.08G125100	Brown	0.75	3.10E-06	Cytochrome P450
Glyma.06G094900	Brown	0.74	5.79E-06	WD40 repeat family protein
Glyma.11G109100	Brown	0.74	1.73E-08	PLATZ transcription factor
Glyma.11G106900	Brown	0.72	1.26E-05	Early responsive to dehydration
Glyma.13G069200	Brown	0.67	0.000121	Zinc finger family protein
Glyma.14G054400	Brown	0.66	0.000115	UDP-glucosyl transferase
Glyma.07G066100	Brown	0.62	0.000402	MYB transcription factor *MYB133*
Glyma.18G114800	Brown	0.61	0.000646	WD40 repeat family protein
Glyma.15G053400	Brown	0.61	0.000851	Potassium transporter
Glyma.08G240800	Brown	0.59	0.000852	WRKY transcription factor
Glyma.15G176000	Brown	0.56	0.001773	MYB transcription factor *MYB121*
Glyma.03G187700	Green	0.79	3.37E-07	UDP-glucosyl transferase
Glyma.15G048600	Green	0.69	4.01E-05	Mitogen-activated protein kinase
Glyma.01G092100	Green	0.64	0.000216	Zinc finger family protein
Glyma.10G216200	Green	0.55	0.002163	Heat shock protein
Glyma.06G171900	Green	0.53	0.003656	4-coumarate-coa ligase
Glyma.02G267800	Green	0.46	0.013048	WD40 repeat protein
Glyma.05G242800	Green	0.41	0.034201	ATP-dependent RNA helicase A-like protein
Glyma.11G107100	Green	0.44	0.018738	Transcription factor MYB4
Glyma.04G243600	Green	0.36	0.041367	MYB transcription factor
Glyma.17G112400	Magenta	0.66	0.000119	N-acetylglucosaminyltransferase
Glyma.14G198600	Magenta	0.65	0.000161	UDP-Glycosyltransferase
Glyma.02G263500	Magenta	0.61	0.000663	S-adenosyl-L-methionine methyltransferases
Glyma.16G149300	Magenta	0.42	0.023798	Isoflavone 2'-hydroxylase

GS.TIF, gene significance total isoflavone; **p.GS.TIF**, significant level.

individual and total isoflavones, underlying that a high portion of the *G. max* genome has genomic regions harboring many candidate SNPs based on the wide diverse panel of soybean accessions utilized in the current study. These findings are consistent with a previous study (Wu et al., 2020) that found significant loci for both individual and TIF content across several sites in a natural soybean population.

WGCNA analysis is an effective technique for categorizing the transcriptome data into co-expression modules to reduce the number of potential candidate genes (Hollender et al., 2014; Greenham et al., 2017; Schaefer et al., 2018; Azam et al., 2023). In this study, out of eight modules, three modules were positively associated with individual and TIF content. The expression patterns of genes present in these modules revealed a higher expression at the late seed development stage. Previous studies also reported that the accumulation of isoflavones mainly occurs at the late stage of seed development (Jung et al., 2000; Dhaubhadel et al., 2003; Cheng et al., 2008; Azam et al., 2023). In addition, GO analysis of these modules revealed some significant GO terms related to biotic and abiotic stresses. Devi et al. (2020) reported that biotic and abiotic stresses lead to an increase isoflavone accumulation by the upregulation of *IFS1* and *IFS2* genes at the late seed development stage. While Uchida et al. (2020) also found that isoflavone O-methyltransferase (*GmIOMT1*) produced higher levels of glycitein in response to biotic stress. Therefore, identifying genes involved in these modules would provide new genetic resources to better understand the isoflavone biosynthesis pathway.

We have identified 27 key candidate genes from brown, magenta, and green modules. Brown module, which showed the highest correlation and gene significance with TIF, contained a cytochrome P450 (*Glyma.08G125100*). A branch of the phenylpropanoid pathway synthesizes isoflavones. Cytochrome P450 play a crucial role in the biosynthesis of a wide variety of plant metabolites (Chapple, 1998). Isoflavone synthases (*IFS1* and *IFS2*) are the members of cytochrome

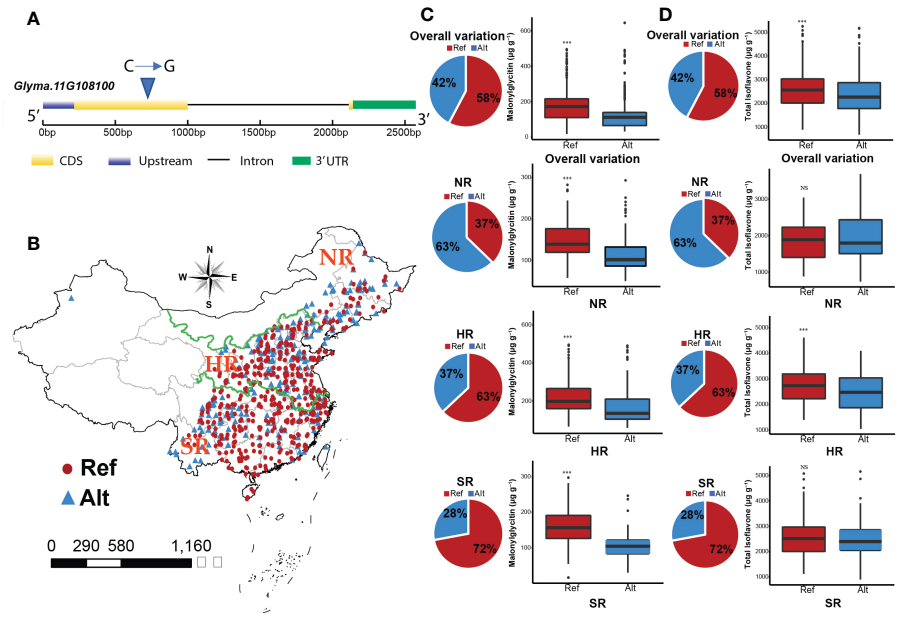


FIGURE 7
(A) Polymorphism that occurred in *Glyma.11G108100*. **(B)** Geographical distribution of *Glyma.11G108100* (NR, Northern region; HR, Huang Huai Hai valley region; SR, Southern region). **(C)** Natural variation of *Glyma.11G108100* for malonylglycitin content. **(D)** Natural variation of *Glyma.11G108100* for TIF content.

P450 super gene family and play a vital role in isoflavone accumulation by producing the 2-hydroxyisoflavone by catalyzing the flavone intermediates (naringenin and liquiritigenin) (Liu et al., 2002). The MYB transcription factors play crucial roles in the regulation of isoflavone biosynthesis by triggering the gene expression of key isoflavonoid biosynthesis enzymes, namely, chalcone isomerases (*CHI*), chalcone synthases (*CHS*), isoflavone synthases (*IFS1* and *IFS2*) (Yi et al., 2010; Chu et al., 2017). We identified *MYB133* as a key candidate gene which was previously identified by (Bian et al., 2018) as a positive regulator of isoflavones through genome-wide analysis, which directly activates *IFS2* and CHS8 and promotes isoflavone accumulation. We identified the natural variation of *MYB133* in the natural population of soybean, which showed a higher TIF level across different regions, landraces, and cultivars (Supplementary Figure 4). Furthermore, the natural variation in the bZIP transcription factor caused synonymous mutation which revealed significant variations for individual and total isoflavones. Previous studies also reported that the synonymous mutations are not just silent but also cause a significant change in the phenotypes (Chu and Wei, 2020; Shen et al., 2022). The bZIP transcription factors are previously reported to control isoflavone accumulation by interacting with MYB transcription factors and play an important role against biotic and abiotic stresses in soybean (He et al., 2020; Yang et al., 2020; Anguraj Vadivel et al., 2021). In addition to MYB and bZIP transcription factors, different zinc-finger transcription factors, such as *GmZFP7*, *GmVOZs*, and *GsVOZs*, regulate isoflavone and stress responses in soybean. (Rehman et al., 2021; Feng et al., 2023)

These findings suggest that most identified key candidate genes include enzymes and transcription factors from important gene families involved in isoflavone biosynthesis. So, the functional validation of these key candidate genes will provide new insights to better understand the molecular mechanism underlying isoflavone biosynthesis.

5 Conclusion

The current study demonstrated that GWAS analysis using natural populations is an effective strategy for identifying candidate genes in soybean. Based on the GWAS and WGCNA, 3 modules were identified that were highly correlated with individual and TIF content. Within these modules, we have identified four key candidate genes and the natural variation present in *Glyma.11G108100* revealed that it influences the isoflavone accumulation in soybean seed. The functional analysis of *Glyma.11G108100* will provide new insight to better understand the isoflavone synthesis pathway.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: https://sfgb.rmbreeding.cn/search/gemplasm, 16NF1005_1006, corresponding accession name Dongnong4hao, ID number ZDD00023.

Author contributions

MAz, Investigation, data curation, visualization, writing-original draft preparation, SZ, LJ, supervision, conceptualization, methodology, investigation, data curation, MAh, KGAB, JQ, resources, formal analysis, software, YF, YL, LQ and BL resources, project administration, conceptualization, writing-review, and editing, JS, funding acquisition, supervision, conceptualization, visualization, writing-review, and editing. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by the National Natural Science Foundation of China (32272178, 32161143033, 31671716, and 32001574) and the Agricultural Science and Technology Innovation Program of CAAS (2060203-2).

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1120498/full#supplementary-material

References

Akashi, T., Aoki, T., and Ayabe, S. (1999). Cloning and functional expression of a cytochrome P450 cDNA encoding *2-hydroxyisoflavanone synthase* involved in biosynthesis of the isoflavonoid skeleton in licorice. *Plant Physiol.* 121, 821–828. doi: 10.1104/pp.121.3.821

Akond, M., Richard, B., Ragin, B., Herrera, H., Kaodi, U., Akbay, C., et al. (2013). Additional quantitative trait loci and candidate genes for seed isoflavone content in soybean. *J. Agric. Sci.* 5, 20. doi: 10.5539/jas.v5n11p20

Anguraj Vadivel, A. K., Mcdowell, T., Renaud, J. B., and Dhaubhadel, S. (2021). A combinatorial action of *GmMYB176* and *GmbZIP5* controls isoflavonoid biosynthesis in soybean (*Glycine max*). *Commun. Biol.* 4, 356. doi: 10.1038/s42003-021-01889-6

Azam, M., Zhang, S., Abdelghany, A. M., Shaibu, A. S., Feng, Y., Li, Y., et al. (2020). Seed isoflavone profiling of 1168 soybean accessions from major growing ecoregions in China. *Food Res. Int.* 130, 108957. doi: 10.1016/j.foodres.2019.108957

Azam, M., Zhang, S., Huai, Y., Abdelghany, A. M., Shaibu, A. S., Qi, J., et al. (2023). Identification of genes for seed isoflavones based on bulk segregant analysis sequencing in soybean natural population. *Theor. Appl. Genet.* 136, 1–12. doi: 10.1007/s00122-023-04258-5

Azam, M., Zhang, S., Qi, J., Abdelghany, A. M., Shaibu, A. S., Ghosh, S., et al. (2021). Profiling and associations of seed nutritional characteristics in Chinese and USA soybean cultivars. *J. Food Compos. Anal.* 98, 103803. doi: 10.1016/j.jfca.2021.103803

Barnes, S. (2010). The biochemistry, chemistry and physiology of the isoflavones in soybeans and their food products. *Lymphat. Res. Biol.* 8, 89–98. doi: 10.1089/lrb.2009.0030

Bennett, J. O., Yu, O., Heatherly, L. G., and Krishnan, H. B. (2004). Accumulation of genistein and daidzein, soybean isoflavones implicated in promoting human health, is significantly elevated by irrigation. *J. Agric. Food Chem.* 52, 7574–7579. doi: 10.1021/jf049133k

Bian, S., Li, R., Xia, S., Liu, Y., Jin, D., Xie, X., et al. (2018). Soybean CCA1-like MYB transcription factor *GmMYB133* modulates isoflavonoid biosynthesis. *Biochem. Biophys. Res. Commun.* 507, 324–329. doi: 10.1016/j.bbrc.2018.11.033

Bradbury, K. E., Appleby, P. N., and Key, T. J. (2014). Fruit, vegetable, and fiber intake in relation to cancer risk: Findings from the European prospective investigation into cancer and nutrition (EPIC). *Am. J. Clin. Nutr.* 100, 394S–398S. doi: 10.3945/ajcn.113.071357

Cai, D. J., Zhao, Y., Glasier, J., Cullen, D., Barnes, S., Turner, C. H., et al. (2004). Comparative effect of soy protein, soy isoflavones, and 17β-estradiol on bone metabolism in adult ovariectomized rats. *J. Bone Miner. Res.* 20, 828–839. doi: 10.1359/JBMR.041236

Cao, Y., Li, S., Wang, Z., Chang, F., Kong, J., Gai, J., et al. (2017). Identification of major quantitative trait loci for seed oil content in soybeans by combining linkage and genome-wide association mapping. *Front. Plant Sci.* 8, 1222. doi: 10.3389/fpls.2017.01222

Chapple, C. (1998). Molecular-genetic analysis of plant cytochrome P450-dependent monooxygenases. *Annu. Rev. Plant Biol.* 49, 311. doi: 10.1146/annurev.arplant.49.1.311

Cheng, H., Yu, O., and Yu, D. (2008). Polymorphisms of *IFS1* and *IFS2* gene are associated with isoflavone concentrations in soybean seeds. *Plant Sci.* 175, 505–512. doi: 10.1016/j.plantsci.2008.05.020

Chu, S., Wang, J., Zhu, Y., Liu, S., Zhou, X., Zhang, H., et al. (2017). An R2R3-type MYB transcription factor, *GmMYB29*, regulates isoflavone biosynthesis in soybean. *PloS Genet.* 13, e1006770. doi: 10.1371/journal.pgen.1006770

Chu, D., and Wei, L. (2020). Genome-wide analysis on the maize genome reveals weak selection on synonymous mutations. *BMC Genom.* 21, 333. doi: 10.1186/s12864-020-6745-3

Darwish, D. B. E., Ali, M., Abdelkawy, A. M., Zayed, M., Alatawy, M., and Nagah, A. (2022). Constitutive overexpression of *GsIMaT2* gene from wild soybean enhances rhizobia interaction and increase nodulation in soybean (*Glycine max*). *BMC Plant Biol.* 22, 431. doi: 10.1186/s12870-022-03811-6

De Steur, H., Mogendi, J. B., Blancquaert, D., Lambert, W., van der Straeten, D., and Gellynck, X. (2014). "Genetically modified rice with health benefits as a means to reduce micronutrient malnutrition: global status, consumer preferences, and potential health impacts of rice biofortification," in *Wheat and rice in disease prevention and health* (San Diego, Academic Press). 283–299.

Devi, M. K. A., Kumar, G., and Giridhar, P. (2020). Effect of biotic and abiotic elicitors on isoflavone biosynthesis during seed development and in suspension cultures of soybean (*Glycine max* l.). *3 Biotech.* 10, 98. doi: 10.1007/s13205-020-2065-1

Dhaubhadel, S., Mcgarvey, B. D., Williams, R., and Gijzen, M. (2003). Isoflavonoid biosynthesis and accumulation in developing soybean seeds. *Plant Mol. Biol.* 53, 733–743. doi: 10.1023/B:PLAN.0000023666.30358.ae

Duan, Z., Zhang, M., Zhang, Z., Liang, S., Fan, L., Yang, X., et al. (2022). Natural allelic variation of *GmST05* controlling seed size and quality in soybean. *Plant Biotechnol. J.* 20, 1807–1818. doi: 10.1111/pbi.13865

Feng, Y., Zhang, S., Li, J., Pei, R., Tian, L., Qi, J., et al. (2023). Dual-function C2H2-type zinc-finger transcription factor *GmZFP7* contributes to isoflavone accumulation in soybean. *New Phytol.* 237, 1794–1809. doi: 10.1111/nph.18610

Greenham, K., Guadagno, C. R., Gehan, M. A., Mockler, T. C., Weinig, C., Ewers, B. E., et al. (2017). Temporal network analysis identifies early physiological and transcriptomic indicators of mild drought in *Brassica rapa*. *Elife* 6, e29655. doi: 10.7554/eLife.29655.026

He, Q., Cai, H., Bai, M., Zhang, M., Chen, F., Huang, Y., et al. (2020). A soybean bZIP transcription factor *GmbZIP19* confers multiple biotic and abiotic stress responses in plant. *Int. J. Mol. Sci.* 21, 4701. doi: 10.3390/ijms21134701

Hollender, C. A., Kang, C., Darwish, O., Geretz, A., Matthews, B. F., Slovin, J., et al. (2014). Floral transcriptomes in woodland strawberry uncover developing receptacle and anther gene networks. *Plant Physiol.* 165, 1062–1075. doi: 10.1104/pp.114.237529

Hwang, E. Y., Song, Q., Jia, G., Specht, J. E., Hyten, D. L., Costa, J., et al. (2014). A genome-wide association study of seed protein and oil content in soybean. *BMC Genom.* 15, 1. doi: 10.1186/1471-2164-15-1

Jung, W., Yu, O., Lau, S. M. C., O'keefe, D. P., Odell, J., Fader, G., et al. (2000). Identification and expression of isoflavone synthase, the key enzyme for biosynthesis of isoflavones in legumes. *Nat. Biotechnol.* 18, 208. doi: 10.1038/72671

Kim, J. K., Kim, E. H., Park, I., Yu, B. R., Lim, J. D., Lee, Y. S., et al. (2014). Isoflavones profiling of soybean [*Glycine max* (L.) Merrill] germplasms and their correlations with metabolic pathways. *Food Chem.* 153, 258–264. doi: 10.1016/j.foodchem.2013.12.066

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi: 10.1038/nmeth.3317

Langfelder, P., and Horvath, S. (2008). WGCNA: An r package for weighted correlation network analysis. *BMC Bioinform.* 9, 559. doi: 10.1186/1471-2105-9-559

Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: The dynamic tree cut package for r. *Bioinform* 24, 719–720. doi: 10.1093/bioinformatics/btm563

Lee, Y. G., Jeong, N., Kim, J. H., Lee, K., Kim, K. H., Pirani, A., et al. (2015). Development, validation and genetic analysis of a large soybean SNP genotyping array. *Plant J.* 81, 625–636. doi: 10.1111/tpj.12755

Li, Y. H., Qin, C., Wang, L., Jiao, C., Hong, H., Tian, Y., et al. (2022). Genome-wide signatures of the geographic expansion and breeding of soybean. *Sci. China Life Sci.* 19, 1–6. doi: 10.1007/s11427-022-2158-7

Li, K., Wang, J., Kuang, L., Tian, Z., Wang, X., Dun, X., et al. (2021). Genome-wide association study and transcriptome analysis reveal key genes affecting root growth dynamics in rapeseed. *Biotechnol. Biofuels.* 14, 178. doi: 10.1186/s13068-021-02032-7

Liang, Q., Chen, L., Yang, X., Yang, H., Liu, S., Kou, K., et al. (2022). Natural variation of *Dt2* determines branching in soybean. *Nat. Commun.* 13, 6429. doi: 10.1038/s41467-022-34153-4

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinform* 30, 923–930. doi: 10.1093/bioinformatics/btt656

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

Liu, C. J., Blount, J. W., Steele, C. L., and Dixon, R. A. (2002). Bottlenecks for metabolic engineering of isoflavone glycoconjugates in arabidopsis. *Proc. Natl. Acad. Sci.* 99, 14578–14583. doi: 10.1073/pnas.212522099

Ma, L., Zhang, M., Chen, J., Qing, C., He, S., Zou, C., et al. (2021). GWAS and WGCNA uncover hub genes controlling salt tolerance in maize (*Zea mays* l.) seedlings. *Theor. Appl. Genet.* 134, 3305–3318. doi: 10.1007/s00122-021-03897-w

Mozaffarian, D., Hao, T., Rimm, E. B., Willett, W. C., and Hu, F. B. (2011). Changes in diet and lifestyle and long-term weight gain in women and men. *N. Engl. J. Med.* 364, 2392–2404. doi: 10.1056/NEJMoa1014296

Nielsen, I. L. F., and Williamson, G. (2007). Review of the factors affecting bioavailability of soy isoflavones in humans. *Nutr. Cancer.* 57, 1–10. doi: 10.1080/01635580701267677

Park, M. R., Seo, M. J., Lee, Y. Y., and Park, C. H. (2016). Selection of useful germplasm based on the variation analysis of growth and seed quality of soybean germplasms grown at two different latitudes. *Plant Breed. Biotechnol.* 4, 462–474. doi: 10.9787/PBB.2016.4.4.462

Pei, R., Zhang, J., Tian, L., Zhang, S., Han, F., Yan, S., et al. (2018). Identification of novel QTL associated with soybean isoflavone content. *Crop J.* 6, 244–252. doi: 10.1016/j.cj.2017.10.004

Phetnoo, N., Werawatganon, D., and Siriviriyakul, P. (2013). Genistein could have a therapeutic potential for gastrointestinal diseases. *Thai J. Gastroenterol.* 2013, 120–125.

Qiu, L., Li, Y., Guan, R., Liu, Z., Wang, L., and Chang, R. (2009). Establishment, representative testing and research progress of soybean core collection and mini core collection. *Acta Agron. Sin.* 35, 571–579. doi: 10.3724/SP.J.1006.2009.00571

Ralston, L., Subramanian, S., Matsuno, M., and Yu, O. (2005). Partial reconstruction of flavonoid and isoflavonoid biosynthesis in yeast using soybean type I and type II chalcone isomerases. *Plant Physiol.* 137, 1375–1388. doi: 10.1104/pp.104.054502

Rasolohery, C. A., Berger, M., Lygin, A. V., Lozovaya, V. V., Nelson, R. L., and Daydé, J. (2008). Effect of temperature and water availability during late maturation of the soybean seed on germ and cotyledon isoflavone content and composition. *J. Sci. Food Agric.* 88, 218–228. doi: 10.1002/jsfa.3075

Rehman, S. U., Qanmber, G., Tahir, M. H. N., Irshad, A., Fiaz, S., Ahmad, F., et al. (2021). Characterization of vascular plant one-zinc finger (*VOZ*) in soybean (*Glycine max* and *Glycine soja*) and their expression analyses under drought condition. *PloS One* 16, e0253836. doi: 10.1371/journal.pone.0253836

Sarkar, M., Watanabe, S., Suzuki, A., Hashimoto, F., and Anai, T. (2019). Identification of novel MYB transcription factors involved in the isoflavone biosynthetic pathway by using the combination screening system with agroinfiltration and hairy root transformation. *Plant Biotechnol.* 36, 241–251. doi: 10.5511/plantbiotechnology.19.1025a

Schaefer, R. J., Michno, J. M., Jeffers, J., Hoekenga, O., Dilkes, B., Baxter, I., et al. (2018). Integrating coexpression networks with GWAS to prioritize causal genes in maize. *Plant Cell.* 30, 2922–2942. doi: 10.1105/tpc.18.00299

Shen, X., Song, S., Li, C., and Zhang, J. (2022). Synonymous mutations in representative yeast genes are mostly strongly non-neutral. *Nature* 606, 725–731. doi: 10.1038/s41586-022-04823-w

Sonah, H., O'donoughue, L., Cober, E., Rajcan, I., and Belzile, F. (2015). Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soyabean. *Plant Biotechnol. J.* 13, 211–221. doi: 10.1111/pbi.12249

Sugiyama, A., Yamazaki, Y., Hamamoto, S., Takase, H., and Yazaki, K. (2017). Synthesis and secretion of isoflavones by field-grown soybean. *Plant Cell Physiol.* 58, 1594–1600. doi: 10.1093/pcp/pcx084

Sun, J., Sun, B. L., Han, F. X., Yan, S. R., Yang, H., and Akio, K. (2011). Rapid HPLC method for determination of 12 isoflavone components in soybean seeds. *Agric. Sci. China* 10, 70–77. doi: 10.1016/S1671-2927(11)60308-8

Supek, F., Bosnjak, M., Skunca, N., and Smuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS One* 6, e21800. doi: 10.1371/journal.pone.0021800

Torkamaneh, D., and Belzile, F. (2015). Scanning and filling: ultra-dense SNP genotyping combining genotyping-by-sequencing, SNP array and whole-genome resequencing data. *PloS One* 10, e0131533. doi: 10.1371/journal.pone.0131533

Tsai, H. S., Huang, L. J., Lai, Y. H., Chang, J. C., Lee, R. S., and Chiou, R. Y. (2007). Solvent effects on extraction and HPLC analysis of soybean isoflavones and variations of isoflavone compositions as affected by crop season. *J. Agric. Food Chem.* 55, 7712–7715. doi: 10.1021/jf071010n

Uchida, K., Sawada, Y., Ochiai, K., Sato, M., Inaba, J., and Hirai, M. Y. (2020). Identification of a unique type of isoflavone O-methyltransferase, GmIOMT1, based on multi-omics analysis of soybean under biotic stress. *Plant Cell Physiol.* 61, 1974–1985. doi: 10.1093/pcp/pcaa112

Wang, J., Fan, Y., Mao, L., Qu, C., Lu, K., Li, J., et al. (2021). Genome-wide association study and transcriptome analysis dissect the genetic control of silique length in *Brassica napus l. Biotechnol. Biofuels.* 14, 214. doi: 10.1186/s13068-021-02064-z

Wang, H. J., and Murphy, P. A. (1994). Isoflavone content in commercial soybean foods. *J. Agric. Food Chem.* 42, 1666–1673. doi: 10.1021/jf00044a016

Wang, X., Song, S., Wang, X., Liu, J., and Dong, S. (2022). Transcriptomic and metabolomic analysis of seedling-stage soybean responses to PEG-simulated drought stress. *Int. J. Mol. Sci.* 23, 6869. doi: 10.3390/ijms23126869

Wu, D., Li, D., Zhao, X., Zhan, Y., Teng, W., Qiu, L., et al. (2020). Identification of a candidate gene associated with isoflavone content in soybean seeds using genome-wide association and linkage mapping. *Plant J.* 104, 950–963. doi: 10.1111/tpj.14972

Yan, J., Wang, B., Zhong, Y., Yao, L., Cheng, L., and Wu, T. (2015). The soybean R2R3 MYB transcription factor GmMYB100 negatively regulates plant flavonoid biosynthesis. *Plant Mol. Biol.* 89, 35–48. doi: 10.1007/s11103-015-0349-3

Yang, Y., Yu, T. F., Ma, J., Chen, J., Zhou, Y. B., Chen, M., et al. (2020). The soybean bZIP transcription factor gene *GmbZIP2* confers drought and salt resistances in transgenic plants. *Int. J. Mol. Sci.* 21, 670. doi: 10.3390/ijms21020670

Yi, J., Derynck, M. R., Li, X., Telmer, P., Marsolais, F., and Dhaubhadel, S. (2010). A single-repeat MYB transcription factor, *GmMYB176*, regulates *CHS8* gene expression and affects isoflavonoid biosynthesis in soybean. *Plant J.* 62, 1019–1034. doi: 10.1111/j.1365-313X.2010.04214.x

Yu, O., and Mcgonigle, B. (2005). Metabolic engineering of isoflavone biosynthesis. *Adv. Agron.* 86, 147–190. doi: 10.1016/S0065-2113(05)86003-1

Zeng, A., Chen, P., Korth, K., Hancock, F., Pereira, A., Brye, K., et al. (2017). Genome-wide association study (GWAS) of salt tolerance in worldwide soybean germplasm lines. *Mol. Breed.* 37, 30. doi: 10.1007/s11032-017-0634-8

Zhang, J., Ge, Y., Han, F., Li, B., Yan, S., Sun, J., et al. (2014). Isoflavone content of soybean cultivars from maturity group 0 to VI grown in northern and southern China. *J. Am. Oil Chem. Soc* 91, 1019–1028. doi: 10.1007/s11746-014-2440-3

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4, 17. doi: 10.2202/1544-6115.1128

Zhao, X., Teng, W., Li, Y., Liu, D., Cao, G., Li, D., et al. (2017). Loci and candidate genes conferring resistance to soybean cyst nematode HG type 2.5. 7. *BMC Genom.* 18, 462. doi: 10.1186/s12864-017-3843-y

Zheng, T., Li, Y., Li, Y., Zhang, S., Ge, T., Wang, C., et al. (2022). A general model for" germplasm-omics" data sharing and mining: A case study of SoyFGB v2. 0. *Sci. Bull.* 67, 1716–1719. doi: 10.1016/j.scib.2022.08.001

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 15 February 2023
DOI 10.3389/fpls.2023.1050313

OPEN ACCESS

EDITED BY
Zhenyu Jia,
University of California, Riverside,
United States

REVIEWED BY
Liu Jinyang,
Jiangsu Academy of Agricultural Sciences (JAAS), China
Suhong Bu,
South China Agricultural University, China
Melaku Gedil,
International Institute of Tropical Agriculture (IITA), Nigeria

*CORRESPONDENCE
Jin Zhang
zhangjin@njau.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

RECEIVED 21 September 2022
ACCEPTED 01 February 2023
PUBLISHED 15 February 2023

CITATION
Wen Y-J, Wu X, Wang S, Han L, Shen B, Wang Y and Zhang J (2023) Identification of QTN-by-environment interactions for yield related traits in maize under multiple abiotic stresses.
Front. Plant Sci. 14:1050313.
doi: 10.3389/fpls.2023.1050313

Identification of QTN-by-environment interactions for yield related traits in maize under multiple abiotic stresses

Yang-Jun Wen[1,2†], Xinyi Wu[1†], Shengmeng Wang[1], Le Han[1], Bolin Shen[1], Yuan Wang[1] and Jin Zhang[1,2*]

[1]College of Science, Nanjing Agricultural University, Nanjing, China, [2]Key Laboratory of Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing, China

Introduction: Quantitative trait nucleotide (QTN)-by-environment interactions (QEIs) play an increasingly essential role in the genetic dissection of complex traits in crops as global climate change accelerates. The abiotic stresses, such as drought and heat, are the major constraints on maize yields. Multi-environment joint analysis can improve statistical power in QTN and QEI detection, and further help us to understand the genetic basis and provide implications for maize improvement.

Methods: In this study, 3VmrMLM was applied to identify QTNs and QEIs for three yield-related traits (grain yield, anthesis date, and anthesis-silking interval) of 300 tropical and subtropical maize inbred lines with 332,641 SNPs under well-watered and drought and heat stresses.

Results: Among the total 321 genes around 76 QTNs and 73 QEIs identified in this study, 34 known genes were reported in previous maize studies to be truly associated with these traits, such as ereb53 (GRMZM2G141638) and thx12 (GRMZM2G016649) associated with drought stress tolerance, and hsftf27 (GRMZM2G025685) and myb60 (GRMZM2G312419) associated with heat stress. In addition, among 127 homologs in Arabidopsis out of 287 unreported genes, 46 and 47 were found to be significantly and differentially expressed under drought vs well-watered treatments, and high vs. normal temperature treatments, respectively. Using functional enrichment analysis, 37 of these differentially expressed genes were involved in various biological processes. Tissue-specific expression and haplotype difference analysis further revealed 24 candidate genes with significantly phenotypic differences across gene haplotypes under different environments, of which the candidate genes GRMZM2G064159, GRMZM2G146192, and GRMZM2G114789 around QEIs may have gene-by-environment interactions for maize yield.

Discussion: All these findings may provide new insights for breeding in maize for yield-related traits adapted to abiotic stresses.

KEYWORDS

multiple abiotic stresses, QTN-by-environment interaction, GWAS, 3VmrMLM, yield-related traits, maize

Introduction

Maize (*Zea mays*) is a vital and strategic cereal crop cultivated in a variety of agroecological zones across the world. Growing on non-irrigated fields exposes them to various environmental stresses, such as drought stress, heat stress, and their combination. Heat waves mixed with acute and persistent drought stress can have disastrous consequences for agriculture, as well as economic and social stability, especially affecting drylands utilized for grain production across the world (Ciais et al., 2005; Mittler, 2006; Zandalinas et al., 2020). The vulnerability of maize to drought and heat stresses can lead to yield losses of 15-20% every year (Khan et al., 2016). Such losses are likely to rise as a result of climate change, especially in emerging nations with rising maize consumption (Campos et al., 2006). To fulfill the future demands of the world's rising population, high yielding and drought tolerant maize cultivars are seen as the most economically feasible answer (Monneveux et al., 2006).

Due to the poor heritability of grain production (Edmeades et al., 1999) and the likelihood of drought occurring at several growth periods, direct selection for grain yield under drought circumstances is frequently challenging (Chen et al., 2012). The use of secondary traits in breeding programs has become one of the finest methods for choosing the genotypes that perform the best under stress situations (Parajuli et al., 2018). Due to the separation of male and female flowers, maize is more vulnerable to drought than any other crop, especially when temperatures are rising above 35°C (Huang et al., 2006). Consequently, the rise in anthesis-silking interval is one of the primary effects of drought stress in maize (Bänziger et al., 2000). The anthesis date keeps a strong genetic correlation with grain yield and remains highly heritable and cost-effective to measure (Cerrudo et al., 2018). These studies demonstrated that the secondary traits comprising anthesis-silking interval and anthesis date have been included in breeding programs to promote indirect selection for grain yield.

As global climate change accelerates, quantitative trait nucleotide (QTN)-by-environment interactions (QEIs) play an increasingly essential role in the genetic dissection of complex traits in plants (Lukens and Doebley, 1999). There are currently accessible methodologies and software tools for identifying QEIs. Crossa et al. (1999) developed a factorial regression model for QEI in tropical maize. In its basic form, an additional covariate needs to be introduced for each putative QTL, thus least squares estimate approaches fail when there are a large number of genotypic or environmental covariables. To detect QEIs, Zhu and Weir (1998) and Wang et al. (1999) developed the mixed-model based composite interval mapping (MCIM) approach, but the results may be susceptible to the specified model of multiple QTL (Piepho, 2000). Li et al. (2015) expanded the inclusive composite interval mapping (ICIM) main-effect genetic model into a QEI model. In real data analysis, it is challenging to uncover small QEIs. However, these approaches are suitable in bi-parental segregation populations. Although Moore et al. (2019) proposed the structured linear mixed model (StructLMM) to detect QEIs, only allelic substitution was detected, and its polygenic background was controlled. To over these issues, recently, Li et al. (2022a, 2022b) proposed a compressed variance component mixed model (3VmrMLM) to detect and estimate all the effects in QTN and QEI detection under controlling all the possibly polygenic backgrounds in genome-wide association studies (GWAS). Based on a full mixed-model framework, the numbers of variance components in QTN and QEI detection were compressed from 5 and 10 to 3, respectively, showing very good performances in computational efficiency. Furthermore, 3VmrMLM can identify QTNs and QEIs accurately and estimate their genetic effects unbiasedly (Zuo et al., 2022; Zhao et al., 2023).

From now, lots of genes response to abiotic stresses were identified in *Arabidopsis*, rice and maize. For example, in *Arabidopsis*, *DREB2A* is one of the transcription factors that activates the expression of heat-stress-responsive genes (Sakuma et al., 2006a). *DREB2A* has a conserved ERF/AP2 DNA-binding domain and recognizes a dehydration-responsive element (DRE). This DRE was reported to function as a heat-stress-responsive element (Sakuma et al., 2006b). Liu et al. (2013a) reported that *di19* functions as a transcriptional regulator and is involved in *Arabidopsis* responses to drought stress through up-regulation of pathogenesis-related *PR1*, *PR2*, and *PR5* gene expressions. In rice, *OsGRAS23* can bind to the promoters of several target genes and modulate the expressions of a series of stress-related genes. Overexpression of *OsGRAS23* conferred transgenic rice plants with improved drought resistance (Xu et al., 2015). The RING finger ubiquitin E3 ligase *OsHTAS* functions in leaf blade to enhance heat tolerance through modulation of hydrogen peroxide-induced stomatal closure. In maize, *ZmHsf11* decreases plant tolerance to heat stress by negatively regulating the expression of oxidative stress-related genes, thus increasing reactive oxygen species levels and decreasing proline content. It is a negative regulator involved in high temperature stress response (Qin et al., 2022). In addition, the overexpression of *ZmPIS* in maize plants under drought stress might lead to the increased synthesis of unsaturated phospholipid and galactolipid species, which are involved in the maintenance of membrane permeability and fluidity that might contribute to plant adaptation to drought stress (Liu et al., 2013b). However, seldom maize gene-by-environment interactions (GEIs) were identified, most of the maize genes were identified by transcriptome analysis and comparative genome analysis (Shi et al., 2017; Zhao et al., 2019). Mining QEIs and related GEIs would provide excellent genes for the genetic improvement of high tolerance to biological stress breeding in maize.

In this study, 3VmrMLM was used to detect QTNs and QEIs for three yield-related traits in an association-mapping panel of 300 tropical and subtropical inbred maize lines each with 955,690 single nucleotide polymorphisms (SNPs) from the DTMA (Drought Tolerant Maize for Africa, https://www.cimmyt.org/projects/drought-tolerant-maize-for-africa-dtma/) in four environments. The transcriptomic data of drought treatment *vs.* well-watered and high *vs.* normal temperature, respectively, were used to identify differentially expressed genes. Functional enrichment, tissue-specific expression, and haplotype and phenotypic difference analysis were used to further validate the candidate maize genes in drought and heat stresses. Multi-environment joint analysis will be helpful for identifying candidate genes related to yield under multiple abiotic stresses in maize.

Materials and methods

Phenotypic data and statistical analysis

The DTMA panel datasets were achieved from International Maize and Wheat Improvement Center (CIMMYT, http://hdl.handle.net/11529/10548156), including 300 inbred lines of tropical and subtropical maize gathered and tested against CML-539 (Wen et al., 2011). Three yield-related traits, grain yield (GY, ton/hectare), anthesis date (AD, day), and anthesis-silking interval (ASI, day), were investigated to detect QTNs and QEIs. The yield trial data were collected from Mexico, Kenya, Thailand, Zimbabwe, and India between 2008 and 2011 under environments of well-watered (WW), drought stress (DS), heat stress (HS), and combined drought and heat stress (DHS). The detailed description and calculated best linear unbiased prediction values for each yield-related trait under the various scenarios were provided by Cairns et al. (2013).

To better understand the patterns of variation of three yield-related traits under various environments, we calculated Pearson correlation coefficients and carried out significance tests for 12 trait-environment combinations using *cor.test* function based on R (Version 4.2.1). The violin plots were adopted to illustrate the variation of three traits under four environments by using the *ggbetweenstats* function in ggstatsplot package of R (Patil, 2021), and the *Kruskal-Wallis* one-way analysis of variance by ranks was conducted with the parameter "type" set to "nonparametric" to test whether the phenotypic mean of each trait differed significantly across four environments.

Genotypic data

We obtained the original genotypic data from http://hdl.handle.net/11529/10548156, with a total of 955,690 SNPs. Then we performed quality control on the SNP dataset by filtering markers with minor allele frequency (MAF) < 0.01 and missing genotype rate > 25% by PLINK (Version 1.9). The imputation of the absent markers was carried out by Beagle (Version 5.4) with the default settings (Browning et al., 2018). Ultimately, we obtained 332,641 SNPs with known physical positions and high quality for further research. To visualize the genotype in this study, PopLDdecay (Version 3.31, https://github.com/BGI-shenzhen/PopLDdecay) was used to calculate linkage disequilibrium (LD) on SNP pairs within a 10-kb window. In addition, the distribution of 332,641 SNPs across 10 chromosomes was plotted by CMplot package in R.

GWAS method

We performed GWAS for the detection of QEIs and QTNs using the IIIVmrMLM package (https://github.com/YuanmingZhang65/IIIVmrMLM; Li et al., 2022b) in R, with high computational efficiency. It mainly used the *IIIVmrMLM* function, where the parameter "method" was set to "Multi_env". The kinship matrix was also calculated *via* the package. In the 3VmrMLM method, the P-value thresholds for significant and suggested QTNs or QEIs were based on Bonferroni correction (P-value < 0.05/m, where m is the number of markers) and logarithm of odds (LOD) score ≥ 3.0, respectively. In the following analysis, as long as one of them was satisfied, we considered it as QTNs or QEIs significantly associated with the target traits. In addition, the package can automatically generate the attractive Manhattan diagrams.

Differential expression and functional enrichment analyses

Genes situated within or contiguous 5 kb (5 kb upstream and downstream, total 10 kb, according to LD decay shown in Figure 1A) of the QTNs and QEIs significantly associated with the target traits were extracted following the B73 AGPV2 (MaizeGDB, https://www.maizegdb.org/) reference genome assembly (Woodhouse et al., 2021). The DNA sequence of all detected genes was used for similarity search on BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) in order to determine the *Arabidopsis* ortholog.

For the above *Arabidopsis* homologous genes, excluding the known genes reported in the literatures, we performed differential expression analysis of the series GSE124340 and GSE154373 from the Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/) database for the unreported genes to identify differentially expressed genes (DEGs) responding to drought stress and heat stress, respectively. The series GSE124340 contains transcript per million (TPM) value of maize under well-watered condition (WW) and drought treatments (DT) at various levels (DT2, DT3, and DT4 represent soil moistures for maize plants were 30-35%, 20-25%, and 10-15% respectively). Each treatment has 2 biological replicates. Meanwhile, the series GSE154373 contains fragments per kilobase of feature per million (FPKM) values for maize plants (inbred line W22) at different temperature treatments (31°C, 33°C, 35°C, and 37°C), with three replicates for each treatment. DEGs between two pairwise samples (DT2 *vs.* WW, DT3 *vs.* WW, DT4 *vs.* WW, 33°C *vs.* 31°C, 35°C *vs.* 31°C, and 37°C *vs.* 31°C) were discovered by limma package in R, with a cutoff of the absolute value of $\log_2$FoldChange greater than 1 and P-value less than 0.05. Simultaneously, these DEGs responding to drought stress and heat stress were intersected with the detected genes, respectively, and thus we obtained the DEGs responding to multiple abiotic stresses for yield-related traits.

For gene ontology-based functional enrichment analysis, information of the above DEGs related to traits were simultaneously submitted to the web-based program AgriGO (Tian et al., 2017). We performed singular enrichment analysis and *Fisher's exact* test with P-value less than 0.05 to select enrichment gene ontology (GO) terms (Xu et al., 2014).

Tissue-specific expression, analysis of haplotype and phenotypic difference, and identification of candidate genes

The database MaizeGDB (https://www.maizegdb.org/) was used to investigate the expression of genes in various tissues to illustrate the association between genes enriched in significant pathways and phenotypic variations. The HaploView software (Version 4.1) was

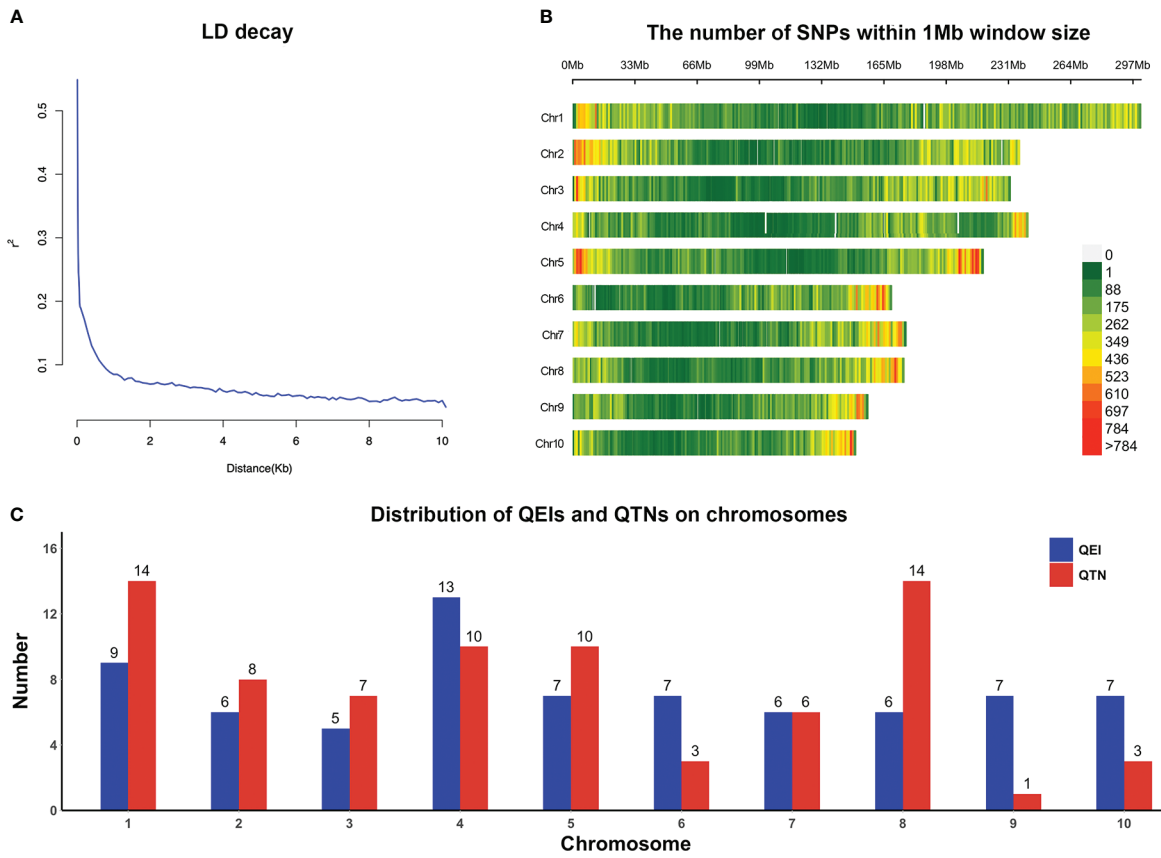


FIGURE 1
(A) LD decay plot for high-quality SNPs. **(B)** Distribution of high-quality SNPs on chromosomes. **(C)** Distribution of QEIs and QTNs across all chromosomes.

used to perform linkage disequilibrium and haplotype block studies, as well as estimate the frequency of haplotype populations in genes widely expressed in various tissues of maize (Barrett et al., 2005), for validating the associated loci between genes and traits. Significant variants were utilized for haplotype division for each gene, and phenotypic differences across haplotypes were examined using the *t.test* function in R. Genes with significant differences in phenotypes across haplotypes under different environments were considered as the candidate genes.

Results

Phenotypic variation and correlation

The phenotypic performance of each trait varied under each environment, suggesting that the DTMA panel seemed to have large variation (Figure 2). All three traits examined under WW condition performed much better than those under stress situations including DS, HS, and DHS. The average performance for trait GY was much higher under WW than under all other situations (Figure 2A). On the other hand, the phenotypic variations for traits AD and ASI measured under WW were smaller than those under stress situations (Figures 2B, C). Except for DHS condition, the average value of AD was larger under WW than that under stress conditions (Figure 2B). The mean ASI value under WW was, however, smaller than that under stress conditions (Figure 2C). The P-values in the *Kruskal-Wallis* test for all three traits under four different environments were 6.98E-209, 1.76E-172, and 1.54E-143, respectively, and the P-values in any *pairwise comparison* test were less than 1.29E-03 (Figure 2), indicating that mean phenotypic values significantly differ across environments.

The phenotypic correlations among all yield-related traits under the same environment varied (Supplementary Figure 1). The correlations for GY under diverse situations were slight, favorable, and significant especially under WW. The correlations were favorable and extremely significant for AD between all situations. Only WW, DS, and HS had significant phenotypic correlations with ASI, while ASI under DHS was strongly linked with DS. On the whole, GY was negatively and strongly correlated with ASI under each condition, with a range of -0.67 to 0.08, confirming the previous findings (Ribaut et al., 2009). Nevertheless, none significant associations were found between GY and AD, or between AD and ASI under the same condition.

The phenotypic correlations between the same traits under various environments also varied (Supplementary Figure 1). For AD, the correlations between any two situations fluctuated from 0.55 to 0.95. The majority of correlations for GY and ASI under diverse situations varied from 0.09 to 0.60. The trait GY under DHS was not strongly correlated with DS or HS circumstance; furthermore, indirect correlations were observed between GY under DHS and that under DS or HS. The trait ASI under WW was positively correlated

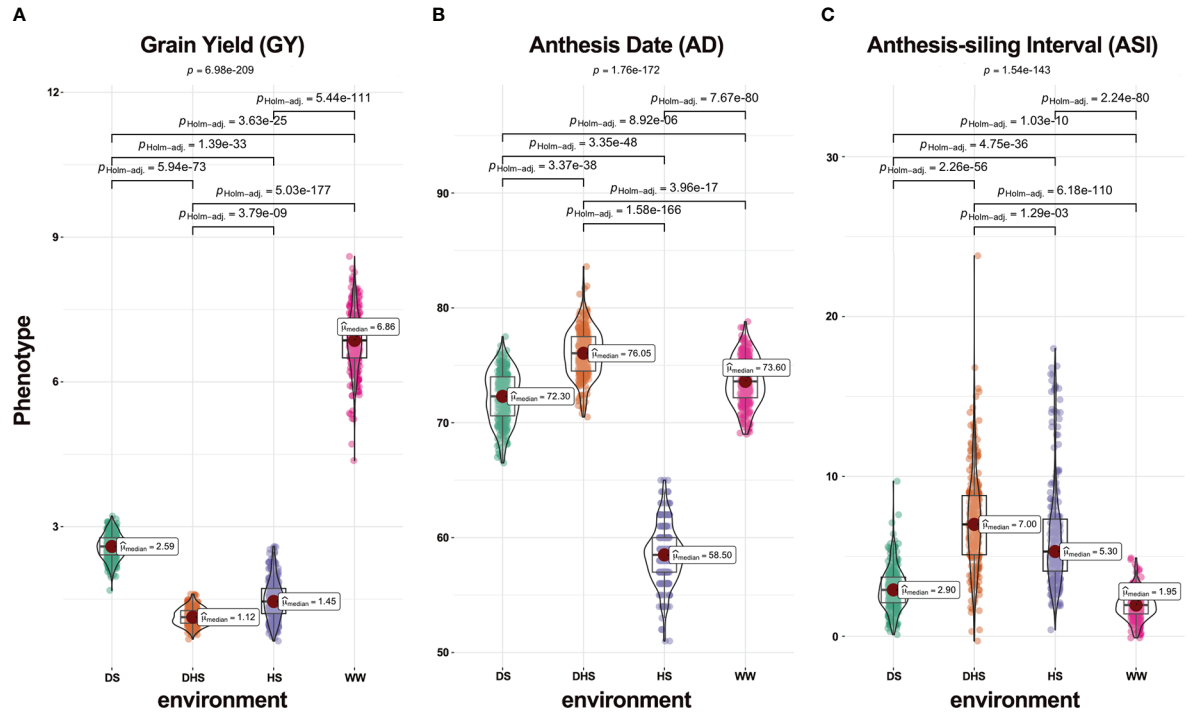


FIGURE 2
Violin plots of phenotypic distribution of three yield-related traits **(A)** grain yield (GY, ton/hectare), **(B)** anthesis date (AD, day), and **(C)** anthesis-silking interval (ASI, day) under the four evaluation conditions, i.e., drought stress (DS), combined drought and heat stress (DHS), heat stress (HS), and well-watered (WW).

with DS or HS situation, but ASI under HS was uncorrelated with DHS situation.

Combined with the above analysis shown in Figure 2 and Supplementary Figure 1, it can be justified that the DTMA panel is suitable for application in multi-environment joint analysis.

Multi-environment joint analysis using 3VmrMLM

In total, 300 inbred lines with 332,641 SNPs were applied to carry out GWAS for each of three traits jointly analyzed in the four environments. LD decay measured the physical distance at which the Pearson's correlation efficient dropped to half of the maximum (Figure 1A). These SNPs were evenly distributed across the 10 chromosomes (Figure 1B). The 3VmrMLM method used in this study identified 73 QEIs (57 significant and 16 suggested QEIs, Supplementary Table 1) and 76 QTNs (64 significant and 12 suggested QTNs, Supplementary Table 2) that were strongly associated with the yield-related traits.

In general, these QEIs and QTNs were distributed on all chromosomes (Figure 1C). For QEIs, the loci were spread out relatively evenly on the chromosomes, it was most distributed on chromosome 4 with 13 and least distributed on chromosome 3 with only 5 (Figure 1C). The highest number of QTNs was found on chromosomes 1 and 8, and the least on chromosome 9 (Figure 1C). On chromosomes 4 and 8, there were relatively more QTNs as well as QEIs, suggesting that these two chromosomes have a greater effect on the genetic variation of yield-related traits; while on chromosome 6, there were twice as many QEIs as QTNs, which may implicate that chromosome 6 may be more susceptible to environmental influences (Figure 1C).

A total of 29 QEIs were detected significantly related to GY, with P-values of 7.176E-129~8.065E-08 and LOD scores of 5.069~132.822, respectively (Figure 3A; Supplementary Table 1). Only 7 QEIs were distinguished for AD, with P-values of 6.123E-62~5.420E-10 and LOD scores of 7.130~65.274 (Figure 3B; Supplementary Table 1). The most QEIs were identified to be significantly associated with ASI in the multi-environment analysis, 37 QEIs were detected with P-values of 5.496E-121~1.978E-08 and LOD scores of 3.063~124.884 (Figure 3C, Table 1, and Supplementary Table 1).

On the other hand, numbers of the significantly associated QTNs of each trait under four environments varied from 20 for ASI to 34 for AD (Supplementary Figure 2, Supplementary Table 2). 22 QTNs related to GY were detected with P-values of 6.021E-30~9.862E-08 and LOD scores of 5.886~29.221(Supplementary Figure 2A, Supplementary Table 2). 34 QTNs were associated with AD, with P-values of 1.414E-41~8.291E-08 and LOD scores of 3.387~40.851 (Supplementary Figure 2B, Supplementary Table 2), and moreover, 20 QTNs associated with ASI were detected with P-values of 3.386E-32~2.295E-08 (Supplementary Figure 2C, Supplementary Table 2). The loci S1_18891169 and S5_205942859 were also identified for AD in the previous study (Yuan et al., 2019).

Meanwhile, the total phenotypic variance explained (PVE) of QEIs for ASI was 71.214% (Table 1 and Supplementary Table 1), higher than the PVE of QTNs 8.966% (Supplementary Table 2). Among these 37 QEIs, S1_29787938 located on chromosome 1 had the maximum PVE of 9.549% (Table 1 and Supplementary Table 1).

10.3389/fpls.2023.1050313

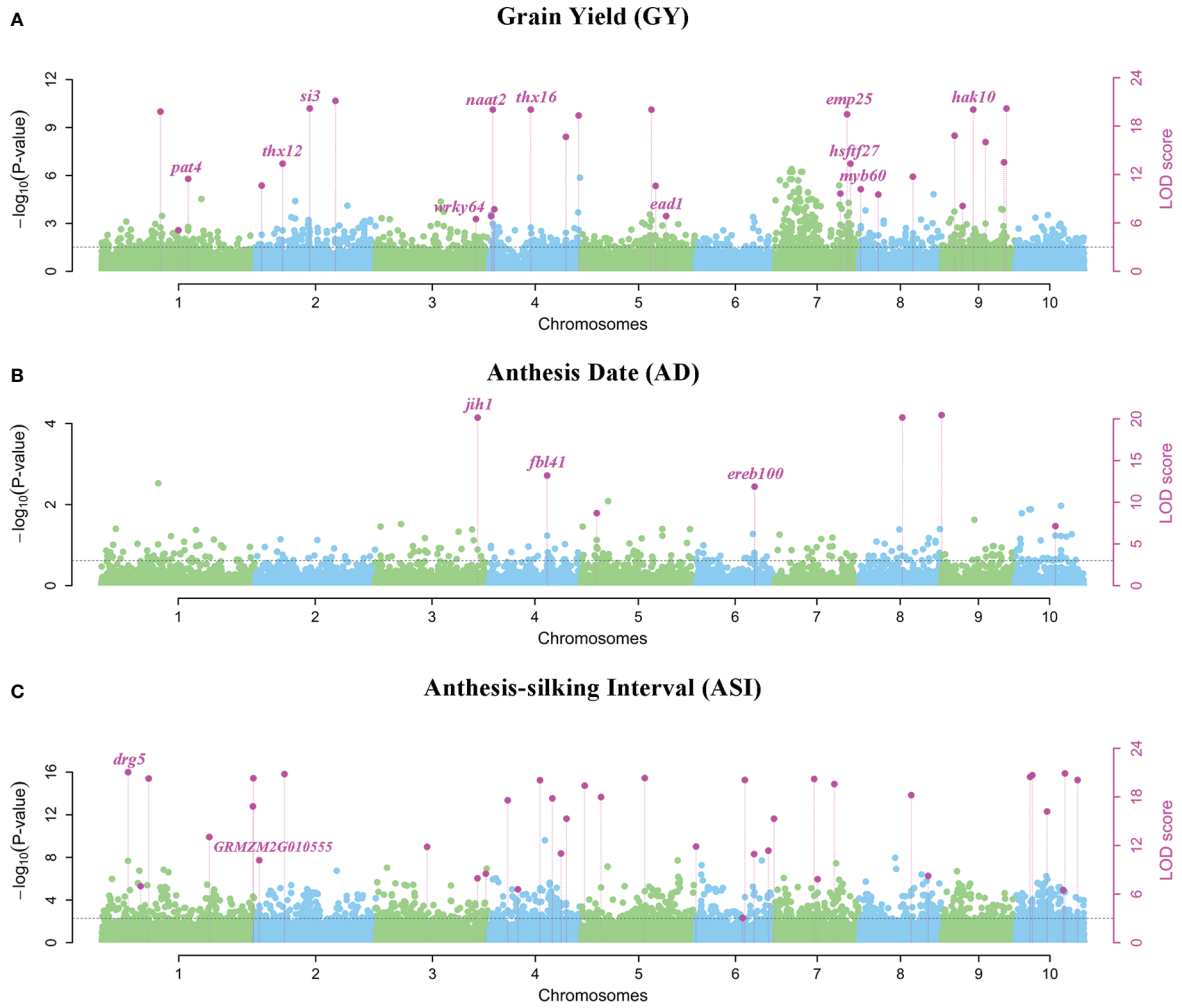


FIGURE 3
Manhattan plots using 3VmrMLM for QEIs on three yield-related traits **(A)** GY, **(B)** AD, and **(C)** ASI under four environments. Y-axis on the left side represents -log10 (P-values) of QEIs, which are obtained from single-marker genome-wide scanning for all markers, while y-axis on the right-side represents LOD scores, which are obtained from likelihood ratio test for QEIs, with the threshold of LOD = 3.0 (dashed line). These LOD scores are shown in points with straight lines. Highlighted text is the corresponding known gene of the loci.

Although the PVE of QTNs for GY was relatively low at 0.515%, the PVE of QEIs was nearly four times higher at 1.974% (Supplementary Tables 1, 2). For AD, the PVE of QTNs was 2.659%, which was higher than the PVE of QEIs (Supplementary Tables 1, 2).

The dominance and additive effects for ASI were relatively significant in all four environments, as listed in Table 1 and Supplementary Table 1. The interaction effect of dominance with the third environment HS for ASI was generally large, with an effect of 8.005 for S1_29787938 located on chromosome 1 and an effect of 4.907 for S6_141276881 located on chromosome 6 (Table 1 and Supplementary Table 1). The interaction effect of additive effect with the first environment DS for AD was positive and moderate, S9_567464 located on chromosome 9, where its effect was 0.488 (Supplementary Table 1). For ASI, the interaction effect of additive with environment DS was also relatively high, the effect of S2_23529006 was 0.647, simultaneously, the effect of S5_160123104 was 0.524 (Table 1 and Supplementary Table 1). In summary, the higher effect of interaction with the environment indicated that the effect of heat and drought stresses on crop yield is not negligible.

Known genes around QEIs and QTNs for yield-related traits under multiple abiotic stresses

In multi-environment joint analysis, a total of 321 genes (5 kb upstream and downstream) were found to be around their significant loci based on MazieGDB against the B73 AGPV2 genome. 161 out of 321 genes were homologous to *Arabidopsis* and their functional annotations were listed in Supplementary Table 3. Number of genes varied among the three traits. In total, 117, 78, and 126 genes were found to be around the significant loci for GY, AD, and ASI, respectively (Supplementary Table 3). For ASI, 74 and 52 genes were found to be around QEIs and QTNs, respectively. At the same time, 63 and 54 genes were found to be around QEIs and QTNs for GY, respectively. However, for AD, 58 genes were found to be around QTNs, but only 20 were found to be around QEIs (Supplementary Table 3). Highlighting in Figure 3 and Supplementary Figure 2, 34 known genes were annotated according to the previous literatures (Augustine et al., 2016; Qi et al., 2017; Li et al., 2019).

TABLE 1 Results of 37 QEIs for trait ASI using multi-environment joint analysis of 3VmrMLM.

Marker	Chr	Pos (bp)	LOD (QEI)	add1	dom1	add2	dom2	add3	dom3	add4	dom4	r^2(%)	P-value	SIG/SUG
S1_29787938	1	29787938	124.884	0.001	-1.342	-1.345	-4.045	1.172	8.005	0.172	-2.618	9.549	5.496E-121	SIG
S1_47457445	1	47457445	6.976	-0.145	0.062	-0.038	-0.140	0.355	0.112	-0.171	-0.034	0.412	1.544E-05	SUG
S1_62226889	1	62226889	46.714	-0.349	-2.227	0.110	-0.494	0.737	5.252	-0.498	-2.531	2.999	1.143E-43	SIG
S1_229206706	1	229206706	13.043	0.020	0.358	0.379	-0.340	-0.403	-0.332	0.004	0.313	0.776	4.369E-11	SIG
S1_297750016	1	297750016	16.830	-0.149	-0.672	-0.040	-0.636	0.363	2.123	-0.175	-0.816	1.029	1.172E-14	SIG
S1_298273269	1	298273269	51.094	0.297	-0.731	0.220	2.949	-0.919	-1.212	0.402	-1.006	3.317	5.696E-48	SIG
S2_2682470	2	2682470	10.180	0.109	-0.002	-0.396	-0.809	0.102	0.371	0.185	0.439	0.599	1.978E-08	SIG
S2_23529006	2	23529006	101.660	0.647	-1.729	0.382	-0.275	-1.446	4.059	0.417	-2.055	7.393	6.103E-98	SIG
S3_147588583	3	147588583	11.834	-0.151	0.284	0.344	0.666	-0.116	-1.342	-0.076	0.392	0.698	5.856E-10	SIG
S3_218123483	3	218123483	7.944	0.010	-0.290	-0.085	-1.586	0.255	2.819	-0.180	-0.943	0.468	2.126E-06	SUG
S3_226979707	3	226979707	8.521	0.067	-0.221	0.301	0.417	-0.321	-0.015	-0.047	-0.181	0.502	6.430E-07	SUG
S4_35625212	4	35625212	17.580	0.197	-1.786	0.083	-0.119	-0.347	4.782	0.067	-2.877	1.055	2.269E-15	SIG
S4_73208150	4	73208150	6.586	-0.056	-0.171	0.301	0.701	-0.131	0.392	-0.115	-0.923	0.385	3.405E-05	SUG
S4_167022069	4	167022069	25.660	-0.044	-0.660	-0.630	-0.003	0.314	-0.027	0.360	0.690	1.566	3.958E-23	SIG
S4_186691903	4	186691903	17.815	-0.031	-0.259	-0.070	-2.715	0.226	2.474	-0.125	0.500	1.069	1.355E-15	SIG
S4_202589250	4	202589250	11.007	-0.211	0.188	-0.047	0.257	0.353	-1.050	-0.095	0.605	0.652	3.426E-09	SIG
S4_223836871	4	223836871	15.310	-0.083	0.063	0.513	0.471	-0.260	-0.125	-0.169	-0.408	0.912	3.224E-13	SIG
S5_2353940	5	2353940	19.387	-0.006	-0.800	0.094	0.025	-0.020	2.037	-0.068	-1.263	1.213	4.279E-17	SIG
S5_14841812	5	14841812	17.978	-0.062	1.120	0.415	0.264	-0.222	-2.051	-0.131	0.667	1.078	9.482E-16	SIG
S5_160123104	5	160123104	52.683	0.524	1.284	-0.732	-2.472	-0.296	-0.251	0.503	1.439	3.378	1.561E-49	SIG
S6_656139	6	656139	11.863	0.154	-0.321	-0.450	0.359	0.226	0.150	0.070	-0.188	0.700	5.511E-10	SIG
S6_137397546	6	137397546	3.063	-0.108	-0.657	0.071	0.423	0.107	-0.297	-0.070	0.531	0.178	2.850E-02	SUG
S6_141276881	6	141276881	29.009	-0.336	-2.635	0.409	0.341	0.041	4.907	-0.114	-2.612	1.776	2.257E-26	SIG
S6_152209037	6	152209037	10.937	0.174	-1.475	-0.212	4.005	-0.117	-1.473	0.155	-1.056	0.576	3.975E-09	SIG
S6_163662312	6	163662312	11.361	0.156	-1.277	-0.004	3.920	-0.096	-0.857	-0.056	-1.785	0.671	1.611E-09	SIG
S6_167325529	6	167325529	15.302	0.010	0.262	-0.459	-0.552	0.383	0.335	0.067	-0.045	0.914	3.280E-13	SIG
S7_126213664	7	126213664	40.770	0.345	-0.367	-0.475	-1.435	-0.499	1.477	0.629	0.326	2.579	7.667E-38	SIG
S7_130495196	7	130495196	7.833	0.015	0.372	-0.281	-0.274	0.232	-0.692	0.033	0.594	0.461	2.672E-06	SUG

(Continued)

10.3389/fpls.2023.1050313

TABLE 1 Continued

Marker	Chr	Pos (bp)	LOD (QEI)	add1	dom1	add2	dom2	add3	dom3	add4	dom4	r^2(%)	P-value	SIG/SUG
S7_155070876	7	155070876	19.580	-0.153	-0.057	0.572	0.977	-0.228	-0.707	-0.190	-0.213	1.176	2.802E-17	SIG
S8_147292704	8	147292704	18.210	0.064	-2.425	-0.194	5.252	0.304	-0.051	-0.173	-2.776	0.975	5.690E-16	SIG
S8_165163196	8	165163196	8.237	-0.086	0.041	-0.173	-0.394	0.373	0.396	-0.114	-0.044	0.486	1.160E-06	SUG
S10_34023703	10	34023703	66.109	-0.039	-0.783	1.082	-0.277	-0.433	2.405	-0.610	-1.345	4.410	9.192E-63	SIG
S10_50775539	10	50775539	87.805	0.109	-0.783	-1.280	-0.277	0.357	2.405	0.814	-1.345	6.153	3.260E-84	SIG
S10_100028483	10	100028483	16.198	-0.224	-2.458	0.196	0.686	0.212	4.198	-0.184	-2.426	0.967	4.663E-14	SIG
S10_133408126	10	133408126	6.483	-0.243	0.140	0.079	0.168	0.265	-0.509	-0.101	0.201	0.379	4.187E-05	SUG
S10_135046780	10	135046780	108.758	0.260	-2.433	0.665	0.778	-1.480	3.735	0.555	-2.080	8.053	5.569E-105	SIG
S10_145183843	10	145183843	27.872	-0.174	0.346	0.583	-0.734	0.137	0.020	-0.546	0.367	1.711	2.858E-25	SIG

Chr, chromosome; Pos, position; LOD, logarithm of odds; addk, additive effect in environment k; domk, dominance effect in environment k; r2 (%): the proportion of total phenotypic variance explained by each QEI. SIG, significant; and SUG, suggested.

For QEIs, 11 known genes related to GY, 3 known genes related to AD, and 2 known genes related to ASI were identified (Figure 3; Supplementary Table 3). The known genes *thx12* (*GRMZM2G016649*, around the locus S2_21790763) and *thx16* (*GRMZM2G063203*, around the locus S4_149899538) related to GY (Figure 3A; Supplementary Table 3) are Trihelix transcription factors (also known as GT transcription factors) that are unique to plants and play important roles in abiotic drought stress (Du et al., 2016). The known gene *hsftf27* (*GRMZM2G025685*) around the locus S7_169176208 (Figure 3A; Supplementary Table 3), which acts as a heat shock transcription factor, helps to resist many environmental stresses and is involved in the regulation of primary metabolism, was also related to GY (Haider et al., 2021). Moreover, the expression of known gene *myb60* (*GRMZM2G312419*) around the locus S8_2763002 (Figure 3A; Supplementary Table 3) in response to jasmonic acid was up-regulated in heat-tolerant maize variety, which is considered to be important signaling substances with respect to plant stress responses (Wang et al., 2020). The known gene *ead1* (*GRMZM2G329229*) around the locus S5_194560419 (Figure 3A; Supplementary Table 3) plays a critical role in malate-mediated female inflorescence development and provides a promising genetic resource for enhancing maize grain yield (Pei et al., 2022). Moreover, *emp25* (*GRMZM2G312954*, around the locus S7_166553957) (Figure 3A; Supplementary Table 3) functions in the splicing of *nad4* introns, and is essential to maize kernel development (Xiu et al., 2020). The known gene *ereb100* (*AC209257.4_FG006*) around the locus S6_153235783 related to AD (Figure 3B; Supplementary Table 3) belongs to the APETALA2/Ethylene-responsive factor (AP2/ERF), which plays an active role in growth, development, and adaptation to abiotic stresses in maize (Zhang et al., 2022). *Drg5* (*GRMZM2G135877*, around the locus S1_29787938) related to ASI (Figure 3C; Supplementary Table 3) is shown to be rhythmically expressed under dark and light-dark cycles (Dong et al., 2020).

For QTNs, 3 known genes were related to GY (Supplementary Figure 2A and Supplementary Table 3), of which *dek2* (*GRMZM2G110851*, around the locus S1_299093763) is a pentatricopeptide repeat protein that affects the splicing of mitochondrial *nad1* intron 1 and is required for mitochondrial function and kernel development (Qi et al., 2017). Meanwhile, 9 known genes were detected for AD (Supplementary Figure 2B and Supplementary Table 3), among which *ereb53* (*GRMZM2G141638*, around the locus S3_166796324) and *ereb60* (*GRMZM2G131266*, around the locus S1_211326173), among the largest transcription factors in plants, were shown to exhibit differential expression patterns at different developmental stages in maize confirmed by the previous study (Zhang et al., 2022), especially in response to three different abiotic stresses, suggesting their important roles in abiotic stress tolerance (Zhang et al., 2022). A total of 7 known genes were found to be related to ASI (Supplementary Figure 2C and Supplementary Table 3), of which *bzip22* (*GRMZM2G043600*, around the locus S7_140710756) is a transcription factor from the basic leucine zipper family, and they are involved in stress responses and hormone signaling (Cao et al., 2019).

There were few overlapped genes detected for the different traits, indicating the genetic divergence between the traits. One common gene homologous to *Arabidopsis* observed for *GRMZM2G064159*

between a QTN of GY and a QEI of AD (Supplementary Table 3). Only one known gene *naat2* (*GRMZM2G006480*) around the locus S4_3890824, which was confirmed to be related to GY, was overlapped between QTN and QEI (Figure 3; Supplementary Figure 2, and Supplementary Table 3). This finding showed the challenge of enhancing maize GY response to numerous abiotic stress tolerances at the same time. The more detailed information about the genes around QTNs and QEIs identified by the 3VmrMLM method can be referred to Supplementary Table 3.

Response to multiple abiotic stresses and GO enrichment pathway

The differential expression analysis was used to determine the response of genes to DS and HS stresses. Among 127 homologs in *Arabidopsis* out of 287 unreported genes, 46 were identified as DEGs under DT *vs*. WW treatments and 47 were identified as DEGs under high temperature *vs*. normal temperature treatments. Among them, 29 DEGs were identified in both DS and HS tolerance (Supplementary Table 4). *GRMZM2G152549* was simultaneously found in six comparison groups (Supplementary Table 4), but it was lowly expressed under different levels of drought treatment relative to WW condition. The absolute value of $\log_2$FoldChange for *GRMZM2G016084* was as high as 205.14, followed by *GRMZM5G896082* and *GRMZM2G048836*, which had absolute values of $\log_2$FoldChange of 200.905 and 198.9, respectively (Supplementary Table 4). The two genes *GRMZM5G896082* and *GRMZM2G048836* were highly expressed after severe drought treatment and heat treatment (Supplementary Table 4).

According to outcomes of the GO functional enrichment analysis, a total of 37 genes among the above 46 and 47 DEGs significantly enriched to 13 GO terms associated with various biological processes (Figure 4A; Supplementary Figure 3, 4). Such as, 17 genes around QEIs and QTNs were enriched to organic substance metabolic process (GO: 0071704), among which 2 genes *GRMZM2G109651* and *GRMZM2G048836* were also participated in the cellular component and molecular function (Supplementary Figures 3 and 4). Pleiotropic gene *GRMZM2G064159* which simultaneously identified around the locus S10_123819112, a QTN for GY and a QEI for AD was also involved in organic substance metabolic process (GO: 0071704, Supplementary Figures 3 and 4). Under adverse environment, plant metabolism is profoundly involved in signaling, physiological regulation, and defense responses (Fraire-Velázquez and Balderas-Hernández, 2013). Cellular components are the complex biomolecules and structures of which cells, and thus living organisms, are composed. In the last layer in Supplementary Figure 3, 6 genes were enriched to intracellular organelle part (GO: 0044446).

Moreover, the expression levels of some genes were significantly different under different treatment conditions. Under drought treatments (Figure 4B), most of the 33 genes were responded to drought stress. *GRMZM2G004377* around the locus S9_149252534, a QEI associated with GY, combined with candidate genes around the QEIs significantly associated with ASI such as *GRMZM2G140609*, *GRMZM2G084767*, and *GRMZM2G070797* had high expression under DT4 treatment and low expression under WW conditions (Figure 4B). In contrast, the gene *GRMZM2G431039* around the locus S7_155070876 associated with ASI had lower expression values under severe drought treatment and higher expression values under sufficient water conditions (Figure 4B). The expression levels of the 25 genes varied under different temperature treatments (Figure 4C). The gene *GRMZM2G146192* around the locus S4_2488289, a QEI associated with GY had a high expression value at 37°C, while *GRMZM2G178829* and *GRMZM2G139600* around QTNs significantly associated with AD had low expression values at high temperature (35°C and 37°C) (Figure 4C). A total of 21 genes responded to drought stress and heat stress, simultaneously (Figures 4B, C). Genes around QEIs significantly associated with ASI, such as *GRMZM2G016084* and *GRMZM2G084806*, were highly expressed under 37°C and DT3 treatment (Figures 4B, C). Gene *GRMZM2G02170* had low expression values under both high temperature at 37°C and extreme drought DT4 treatment (Figure 4B, C). In addition, some genes were expressed at different levels under drought stress and heat stress treatments. For example, the gene *GRMZM2G455476* had high expression value under DT4 treatment but low expression value under high temperature treatment at 37°C (Figures 4B, C). The gene *GRMZM2G070709* had high expression under DT3 treatment, but low expression value under high temperature treatment at 35°C (Figures 4B, C). This information may be useful in providing some biological basis for newly discovered heat and drought tolerant genes in maize.

Haplotype and phenotypic difference analysis of candidate genes and tissue-specific expression profiles

Based on the results of tissue-specific expression, almost all the 37 genes significantly enriched to the pathways, except for *AC202120.3_FG003*, were expressed in various maize tissues. To further confirm the association between the genes and yield-related traits, we performed haplotype analysis of the remaining genes using SNPs within these genes and 2 kb upstream of them. A total of 24 genes differed significantly in phenotypes across haplotypes under different environments, and were considered as the candidate genes (Table 2). Among 24 candidate genes, there were 13 genes around QEIs and 13 genes around QTNs, with two candidate genes, *GRMZM2G006480* and *GRMZM2G064159*, being detected around both QEIs and QTNs. The more detailed results were listed in Table 2 and Supplementary Table 5.

Pleiotropic candidate gene *GRMZM2G064159* (CDS coordinates [5′-3′]: 123811073 ~ 123815007) around the locus S10_123819112, a QEI for AD and a QTN for GY (Table 2; Supplementary Tables 3 and 5), was analyzed to reveal the intragenic variation affecting the yield and to identify favorable haplotypes. Figure 5A exhibited the tissue-specific expression profile of the candidate gene *GRMZM2G064159*, which has a much higher expression value of 747.60 in Anther-2.0mm-W23 and is also commonly expressed in spike, embryo, and root-associated tissues. Figure 5B showed the linkage disequilibrium and haplotype block with 15 SNPs. The 300 inbred lines were classified into 7 haplotypes based on 14 SNPs (S10_123811034, S10_123811055, S10_123811069, S10_123811287, S10_123811289, S10_123814031, S10_123814100, S10_123814124, S10_123814202, S10_123814715, S10_123814731, S10_123814738, S10_123814750, S10_123814751).

For AD, haplotype VI (GCGGCAACAGGACA) had the highest mean phenotypic values in DS (72.63) and DHS (76.17) conditions, whereas haplotype IV (AAGGCAGCGCCGCT) presented the lowest mean phenotypic values in DS (70.45) and DHS (74.48) conditions (Figure 5C). A *t* test showed that significant differences in DS condition existed between haplotypes II and IV (P-value = 4.62E-04, Supplementary Table 5). There was also a significant difference in DHS condition between haplotypes II and IV (P-value = 4.13E-03, Supplementary Table 5). For GY, haplotype VII (GCGGCAGCGCCGCT) had the highest mean phenotypic values in DS (2.63) and DHS (1.21) conditions, while haplotype IV had the lowest mean phenotypic values under DS (2.35) and HS (1.14) conditions (Figure 5D). A *t* test showed that significant differences in HS condition between haplotypes IV and VI (P-value = 1.21E-02, Supplementary Table 5). Therefore, we hypothesized that the candidate gene *GRMZM2G064159* may interact with environments for yield-related traits in maize.

The candidate gene *GRMZM2G146192* (CDS coordinates [5′-3′]: 2481257 ~ 2484641) was detected around the locus S4_2488289, a QEI for GY (Table 2; Supplementary Tables 3 and 5). Supplementary Figure 5A showed the tissue-specific expression profile of *GRMZM2G146192*, with higher expression values in root and leaf-associated tissues. Supplementary Figure 5B, C revealed the results of the haplotype block and phenotype difference. We inferred that the candidate gene *GRMZM2G146192* might also respond to various environment conditions for maize yield.

GRMZM2G114789 (CDS coordinates [5′-3′]: 10541987 ~ 10545884) was also detected around the locus S5_10542293, a QEI for AD (Table 2; Supplementary Tables 3 and 5). Supplementary Figure 6A showed the tissue-specific expression profile of the candidate gene *GRMZM2G114789*, with higher expression values in root and embryo-associated tissues. Supplementary Figures 6B, C revealed the results of the haplotype block and phenotype difference. Haplotype II (CCGGCCCAAGGCT) had the highest mean phenotypic values in DS (75.27), DHS (77.12), HS (60.29), and WW (75.27) conditions, whereas haplotype V (TCGGCCCAAGGCT) presented the lowest mean phenotypic values in DS (69.56), DHS (74.88), HS (56.4), and WW (71.42) conditions. Supplementary Figure 6C showed significant differences in all conditions between haplotypes II and V, haplotypes II and VI (TCGGCCCAAGGTT), and haplotypes II and VII (TCGGCTTCAGGTT). Therefore, we inferred that the candidate gene *GRMZM2G114789* might be also a gene that interacted with environments related to yield in maize.

In summary, we supposed that the three candidate genes around QEIs mentioned above might have potential gene-by-environment interactions, including *GRMZM2G064159*, *GRMZM2G146192*, and *GRMZM2G114789*. In addition, some candidate genes around QTNs differed significantly in phenotypes across haplotypes under different environments (Supplementary Table 5). For example, the candidate gene *GRMZM2G166987* (CDS coordinates [5′-3′]: 213939500 ~ 213945050) identified around the QTN S3_213937689, which was significantly associated with ASI (Table 2; Supplementary Table 3), showed that its haplotype I (GAGGCAG) and haplotype III (GCTACAG) were significantly different to the phenotype under DS, HS, and DHS conditions by *t* test (Supplementary Table 5). However, whether these candidate genes around QTNs have gene-by-environment interactions for yield-related traits in maize needs to be further verified by new experiments.

Discussion

Tolerance to drought and heat stresses

Drought stress and heat stress are the most significant abiotic restrictions in the present and future climate change scenarios. Any additional rise in the frequency and severity of these stressors, either separately or in combination, would have a devastating impact on world agricultural yield and food security. Although they impede agricultural output at all phases of development, the level of damage during the blooming stage, particularly during the seed

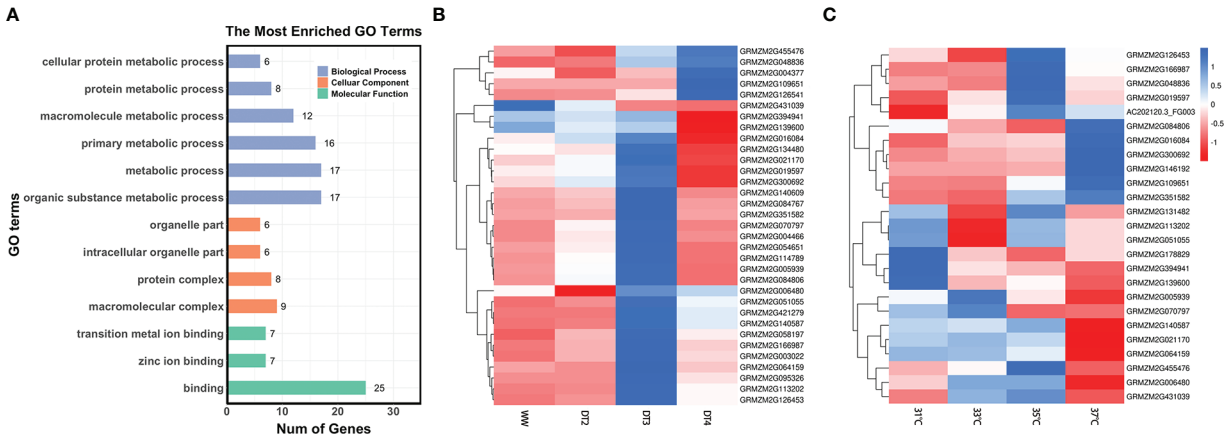


FIGURE 4
(A) Results of gene ontology-based functional enrichment analysis. **(B)** Clustered heatmap of expression values for 33 genes under different drought level treatments. WW stands for well-watered condition, DT2, DT3, and DT4 represent soil moistures for maize plants were 30-35%, 20-25%, and 10-15%, respectively. **(C)** Clustered heatmap of expression values for 25 genes under different temperature treatments (31°C, 33°C, 35°C, and 37°C). The numerical data represent the Z-score of mean TPM of two or three replicates.

filling phase, is essential and causes significant yield losses. Cultivating climate-resilient crops is thus an efficient means of adapting to climate change.

We only obtained the transcriptomic data for drought stress and heat stress, and couldn't obtain ones for combined drought and heat stress. Then, 46 and 47 DEGs were found to be significantly expressed under drought *vs.* well-watered treatments, and high *vs.* normal temperature treatments, respectively. Among them, 29 genes were identified in both DS and HS tolerance (Supplementary Table 4). However, most of the candidate genes did not show significant differences in combined drought and heat stress across haplotypes (Supplementary Table 5). This finding indicated that tolerance to individual stresses in maize is genetically distinct from tolerance to combined drought and heat stress, and tolerance to either stress alone does not confer tolerance to combined drought and heat stress, which was confirmed in the previous study (Cairns et al., 2013). Identification of genes tolerance to combined drought and heat stress will be the further work.

Genetic basis for yield-related traits in maize

3VmrMLM identified 73 QEIs and 76 QTNs significantly associated with three yield-related traits under four environments in this study. The total PVE of all significant QEIs was 73.191%, which is six times that of QTNs (Supplementary Tables 1 and 2). Moreover, this study found a higher contribution by QEIs to total variation (PVE = 71.214%) than QTNs (PVE = 8.967%) for ASI (Table 1; Supplementary Tables 1 and 2). For ASI, 4 out of QEIs had a PVE value greater than 5% (Table 1 and Supplementary Table 1). Among these four QEIs, *drg5* (*GRMZM2G135877*) around the locus S1_29787938 (r^2 = 9.549%, Table 1; Supplementary Tables 1 and 3) is a known gene that has been verified by transcriptome analysis in the previous study (Dong et al., 2020).

The two known genes *thx12* (*GRMZM2G016649*) around the QEI S2_21790763 (P-value = 2.299E-11, LOD = 13.341, Figure 3A; Supplementary Tables 1 and 3) and *thx16* (*GRMZM2G063203*) around the QEI S4_149899538 (P-value = 8.289E-22, LOD =24.292,

TABLE 2 Results of 24 candidate genes and functional annotation of *Arabidopsis* homologous genes.

Trait	QTN/QEI	Marker	Candidate Gene	Phytozome Annotations
GY	QEI	S4_2488289	*GRMZM2G146192*	beta-xylosidase 2
	QTN&QEI	S4_3890825	*GRMZM2G006480*	Tyrosine transaminase family protein
	QEI	S4_238951599	*GRMZM2G019597*	tRNA (guanine-N-7) methyltransferase
	QTN	S6_113109041	*GRMZM2G048836*	FTSH protease 6
	QEI	S7_160600156	*GRMZM2G058197*	C2H2-like zinc finger protein
	QEI	S9_47606538	*GRMZM2G131482*	surp domain-containing protein
	QEI	S9_149252534	*GRMZM2G004466*	seed storage 2S albumin superfamily protein
	QTN	S10_123819112	*GRMZM2G064159*	porphyromonas-type peptidyl-arginine deiminase family protein
AD	QTN	S1_279123888	*GRMZM2G351582*	ZPR1 zinc-finger domain protein
	QTN	S4_6553499	*GRMZM2G054651*	HVA22 homologue A
	QEI	S5_10542294	*GRMZM2G114789*	RNA-binding (RRM/RBD/RNP motifs) family protein
	QTN	S7_161438376	*GRMZM2G178829*	ARM repeat superfamily protein
	QTN	S7_174741307	*GRMZM2G134480*	ubiquitin activating enzyme 2
	QTN	S8_14796428	*GRMZM2G139600*	gamma-glutamyl transpeptidase 4
	QTN	S8_62998618	*GRMZM2G109651*	Cyclin/Brf1-like TBP-binding protein
	QEI	S10_123819112	*GRMZM2G064159*	porphyromonas-type peptidyl-arginine deiminase family protein
ASI	QEI	S1_47457445	*GRMZM2G300692*	galacturonosyltransferase-like 7
	QEI	S1_297750017	*GRMZM2G016084*	Nucleic acid-binding proteins superfamily
	QTN	S3_213937689	*GRMZM2G166987*	GDSL-like Lipase/Acylhydrolase superfamily protein
	QTN	S4_2764858	*GRMZM2G126453*	AAA-type ATPase family protein
	QEI	S6_141276882	*GRMZM2G084806*	Leucine-rich repeat protein kinase family protein
	QEI	S6_152209037	*GRMZM2G140587*	GDA1/CD39 nucleoside phosphatase family protein
	QEI	S6_167325529	*GRMZM2G051055*	casein kinase 1
	QTN	S10_96835918	*GRMZM2G021170*	Nucleic acid-binding OB-fold-like protein
	QTN	S10_127370470	*GRMZM2G005939*	basic helix-loop-helix DNA-binding superfamily protein

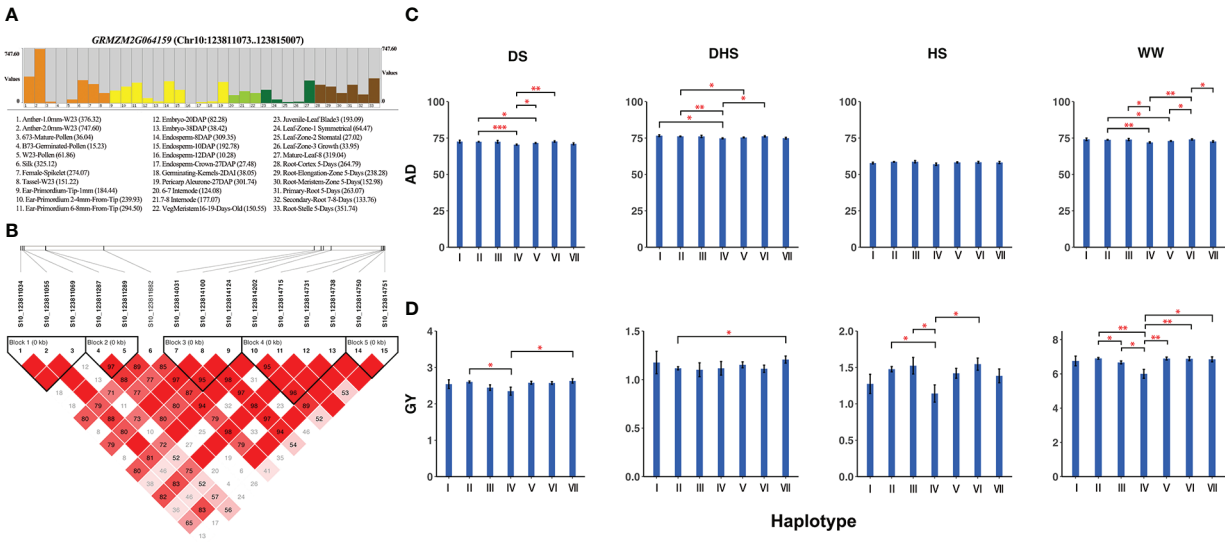


FIGURE 5
(A) Tissue-specific expression profile, **(B)** Linkage disequilibrium, and haplotype block with 14 SNPs inside for the candidate gene GRMZM2G064159. **(C)** Comparison of trait AD among haplotypes I (AACGCAACAGGACA), II (AACGCAGCGCCGCT), III (AACGCAGCGGCATA), IV (AAGGCAGCGCCGCT), V (AAGGCAGCGGCATA), VI (GCGGCAACAGGACA) and VII (GCGGCAGCGCCGCT). **(D)** Comparison of trait GY among haplotypes I, II, III, IV, V, VI, and VII. The number of stars represents the result of *t* test at different significance levels (*:0.05; **:0.01; ***:0.001).

Figure 3A, Supplementary Tables 1 and 43), related to GY and homologous to the *Arabidopsis* gene *AT1G76890*, are the GT factors and play important roles in drought stress (Du et al., 2016). The mRNA expression levels of GT factors were determined for maize under drought stress. Moreover, the known gene *hsftf27* (*GRMZM2G025685*) around the QEI S7_169176208 (P-value = 1.996E-08, LOD = 13.335, Figure 3A; Supplementary Tables 1 and Supplementary Table 1 and 3), which acts as a heat shock transcription factor, helps to resist many environmental stresses and is involved in the regulation of primary metabolism (Haider et al., 2021), was also related to GY. The expression of known gene *myb60* (*GRMZM2G312419*) around the QEI S8_2763002 (P-value = 2.331E-11, LOD = 10.176, Figure 3A; Supplementary Tables 1 and 3) in response to jasmonic acid is up-regulated in heat-tolerant maize variety, which is considered to be important signaling substances with respect to plant stress responses (Wang et al., 2020). *Thx12* and *thx16* exhibited high expression levels in immature leaves and at the base of two leaves stage. *Hsftf27* and *myb60* had higher expression values in root tissue at all stages. Roots and leaves are major tissues in coping with drought and heat stresses (Du et al., 2016).

In addition, the known gene *ereb60* (*GRMZM2G131266*) around the QTN S1_211326173 (P-value = 1.181E-08, LOD = 7.928, Supplementary Figure 2B, Supplementary Tables 2 and 3) significantly associated with AD exhibited obvious spatial and temporal expression profiles, specifically expressed in embryos (Zhang et al., 2022), implying that it was involved in maize growth and development regulation. The known gene *ereb53* (*GRMZM2G141638*) around the QTN S3_166796324 (P-value = 4.437E-11, LOD = 10.353, Supplementary Figure 2B, Supplementary Tables 2 and 3) significantly associated with AD was highly up-regulated after drought stress by transcriptome analysis (Zhang et al., 2022). The known gene *bzip22* (*GRMZM2G043600*) around the QTN S7_140710756 (P-value = 7.000E-13, LOD = 12.155, Supplementary Figure 2C, Supplementary Tables 2 and 3) significantly associated with ASI has been demonstrated to play essential roles in drought stress primarily through the ABA signal transduction pathway in the reported literature (Cao et al., 2019). This finding implied that the main effect of QTNs may also reflect an influence of environmental interactions.

Except for the above known genes, we also detected 24 new candidate genes in this study (Table 2). Among them, *GRMZM2G064159*, *GRMZM2G146192*, and *GRMZM2G114789* around QEIs have been shown the potential gene-by-environment interactions for yield-related traits in maize. First, *GRMZM2G064159* was a pleiotropic candidate gene which was simultaneously identified around the locus S10_123819112, a QEI for AD (P-value = 1.128E-05, LOD = 7.130, Supplementary Table 1) and a QTN for GY (P-value = 3.032E-18, LOD = 17.519, Supplementary Table 2). *GRMZM2G146192* was found to be around the locus S4_2488289, a QEI for GY (P-value = 2.058E-05, LOD = 6.835, Supplementary Table 1). *GRMZM2G114789* was found to be around the locus S5_10542293, a QEI for AD (P-value = 4.598E-07, LOD = 8.6818, Supplementary Table 1). Second, they are homologous to *Arabidopsis* (Table 2; Supplementary Table 3). *GRMZM2G146192* is homologous to *AT1G02640* (*BXL2*, Table 2; Supplementary Table 3), which increased enzymatic saccharification efficiency in *Arabidopsis* (Ohtani et al., 2018). *GRMZM2G064159* is homologous to *AT5G08170* (*EMB1873*, Table 2; Supplementary Table 3), which acted upstream of or within embryo development ending in seed dormancy. EMB genes encoded proteins with an essential function required throughout the life cycle (Muralla et al., 2011). *GRMZM2G114789* is homologous to the RNA-binding family protein *AT4G17720* (*BPL1*, Table 2; Supplementary Table 3) which contains classical RNA recognition motif domains and is implicated in the response to cytokinin (Marondedze et al., 2016). Third, they

were DEGs under DT *vs*. WW treatments or under high *vs*. normal temperature treatments (Figures 4B, C; Supplementary Table 4), and *GRMZM2G064159* and *GRMZM2G146192* both involved in organic substance metabolic process (GO: 0071704, Supplementary Figure 3), *GRMZM2G114789* involved in binding (GO:0005438, Supplementary Figure 3). Moreover, their phenotypic differences across haplotypes were significant under four environments (Figure 5C; Supplementary Figures 5C, 6C, and Supplementary Table 5). Lastly, *GRMZM2G064159* was commonly expressed in spike, embryo, and root-associated tissues (Figure 5A). High expression in embryo implies that it may be involved in maize growth and development regulation (Zhang et al., 2022). The root system is the primary site that perceives drought stress signals (Seo et al., 2009). Besides, *GRMZM2G146192* was highly expressed in root and leaf-associated tissues (Supplementary Figure 5A). *GRMZM2G114789* was expressed at various stages in root, leaf, internode, seed, and embryo-associated tissues, with higher expression values in root and embryo-related tissues (Supplementary Figure 6A). Therefore, we supposed that the candidate genes *GRMZM2G064159*, *GRMZM2G146192*, and *GRMZM2G114789* around QEIs may have gene-by-environment interactions for yield-related traits in maize, although new experiments such as functional validation are necessary to explore these novel GEI-trait associations. Although the results for known genes suggested that genes around QTNs may reflect an influence of environmental interactions (such as *ereb60*, *ereb53*, and *bzip22*, Supplementary Figure 2B, C and Supplementary Table 3), whether the candidate genes identified around QTNs in this study (Table 2) have gene-by-environment interactions needs to be further explored.

In addition, for ASI, the dominance effect in HS situation was positive and significant, ranging from -2.051% to 8.005%. In contrast, the dominance effect in DS situation was relatively negative and moderate, with a range mostly concentrated from -2.635% to 0.284% (Table 1 and Supplementary Table 1). While on the other hand, the overall PVE of QTNs and QEIs significantly associated with GY were relatively low, largely clustered at 0.01% to 0.56% (Supplementary Tables 1 and 2). These findings suggested that trait GY and secondary trait ASI under abiotic stress would be regulated by small effect QTNs or QEIs that are dispersed across the genome in maize. This also suggested that it is relatively difficult to use marker-assisted selection to improve maize yield due to the complexity of traits under multiple environments. And in real data application, introducing secondary yield-related traits to assist maize breeding might be a good choice, which is also consistent with the findings in Bolaños and Edmeades (1996).

Methods comparison

We also performed a single-environment analysis in the DTMA panel using the IIIVmrMLM package. The PVE of QTNs for ASI under each environment ranged from 50.25% to 58.04% (Supplementary Table 6), while the total PVE of QEIs for ASI in the multi-environment joint analysis was as high as 71.214% (Table 1 and Supplementary Table 1). Moreover, 102 QTNs and 221 genes for ASI were detected in the single-environment approach, of which 5 QTNs overlapped with QEIs in the multi-environment joint analysis, and 11 genes overlapped (Supplementary Tables 3 and 6), of which one known gene *drg5* (*GRMZM2G135877*) was confirmed to be dark response gene in the previous literature (Dong et al., 2020). There were few overlapped loci detected in single- and multi-environment analyses, further illustrating that the yield-related traits in maize are complex and relatively susceptible to environmental influences. The more detailed results were listed in Supplementary Table 6. To address this issue, it is necessary to optimize the "SearchRadius" parameter.

Under the framework multiple-locus association studies, a few multi-year and multi-location GWAS methods are applicable for high-dimensional data analysis, and the DTMA panel with 332,641 SNPs has been seldom applied to reveal QEIs. Compared to the above single-environment analysis in 3VmrMLM, the significant loci overlapped fewer. We also compared 3VmrMLM with ICIM method (Li et al., 2015). Firstly, to reduce the computational burden, we used *Levene's* test (Brown and Forsythe, 1974) in R and set the threshold to 0.05 to downscale the DTMA dataset. That is because the ICIM method is very slow in handling high-dimensional dataset and *Levene's* test can be used to detect potential loci for heterogeneity of variances due to potentially interacting SNPs such as QTN-by-environment interactions. 58,000~71,000 significant markers for each trait were identified by *Levene's* test. Then, the linkage map was converted according to the ratio of genetic distance to physical distance of 1.296 cM/Mb (Guo et al., 2015). Finally, we performed a multi-environment joint analysis for the above data using the QTL IciMapping 4.2 software (Meng et al., 2015). A comparison was listed in Supplementary Table 7. The threshold was set to LOD (A) > 3 for additive QTLs and LOD (A by E) > 3 for additive QTLs by environment interactions in ICIM approach. 3vmrMLM detected more QTNs or QEIs than additive QTLs or additive QTLs by environment interactions. In particular, for ASI, 3VmrMLM detected 37 QEIs (PVE = 71.214%), but ICIM detected only 6 additive QTLs by environment interactions (PVE = 9.34%). 3VmrMLM added the polygenic effect and population structure to control the genetic background, thus it might be relatively close to the true genetic models of plants and animals. In addition, the computing time for GY, AD, and ASI ranged from 1~2 days, while 3VmrMLM consumed less than 7 hours for each trait, which took about one fourth of ICIM's. 3VmrMLM reduces the dimensionality of SNPs by single-locus method, and constructs the multi-locus model based on the remaining markers, which decreases computational volume and computational complexity. In summary, 3VmrMLM presents well-performance results with higher statistical power, lower false positive rate and high computational efficiency, and it is a recommended method in multi-environment joint analysis.

Conclusion

In this study, we identified QTN-by-environment interactions for three yield-related traits in maize under four abiotic stresses using the newly proposed 3VmrMLM method. A total of 73 QEIs and 76 QTNs were identified. Moreover, 34 known genes and 24 candidate genes were identified in the vicinity of QEIs and QTNs. Among 34 known genes, *ereb53* (*GRMZM2G141638*) & *thx12* (*GRMZM2G016649*), and *hsftf27* (*GRMZM2G025685*) & *myb60* (*GRMZM2G312419*) were confirmed to play important roles in drought and heat stresses,

respectively, by transcriptome and bioinformatics analysis in previous maize studies. Among 24 candidate genes, 13 genes around QEIs and 13 genes around QTNs were validated functioning in drought and heat stresses by homologous genes miming, differential expression, functional enrichment, tissue-specific expression, and haplotype and phenotypic difference analysis in this study. Importantly, *GRMZM2G064159*, *GRMZM2G146192*, and *GRMZM2G114789* around QEIs may have gene-by-environment interactions for yield. These findings will facilitate the mining of genes involved in maize breeding under the abiotic stresses.

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

JZ conceived the study. JZ, Y-JW, and XW designed the experiment. XW, SW, and LH performed data analyses under the assistance or guidance from JZ and Y-JW. BS and YW contributed resources. Y-JW and XW wrote the manuscript with the participation of all authors. All authors contributed to the article and approved the submitted version.

Funding

The work was supported by the National Natural Science Foundation of China (32270694, 32070688, and 31701071), the Postdoctoral Science Foundation of Jiangsu (2020Z330), and the Fundamental Research Funds for the Central Universities (JCQY202108).

Acknowledgments

We would like to thank the editor and reviewers for their suggestions for improving the framework and language within this manuscript.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1050313/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Pearson correlation coefficients and test for three yield-related traits under four environments in the DTMA panel. (Upper right) Pearson correlation coefficients, when the color is darker, the association is stronger; (Lower left) Pearson correlation test, the number of stars represents the different significance level (*: 0.05; **: 0.01; ***: 0.001). NS indicates non-significant.

SUPPLEMENTARY FIGURE 2
Manhattan plots using 3VmrMLM for QTNs on three yield-related traits **(A)** GY, **(B)** AD and **(C)** ASI under four environments. Y-axis on the left side represents -log10 (P-values) of QTNs, which are obtained from single-marker genome-wide scanning for all markers, while y-axis on the right-side represents LOD scores, which are obtained from likelihood ratio test for QTNs, with the threshold of LOD = 3.0 (dashed line). These LOD scores are shown in points with straight lines. Highlighted text is the corresponding known gene of the loci.

SUPPLEMENTARY FIGURE 3
Hierarchical tree graph of overrepresented GO terms in biological process category generated by singular enrichment analysis. Boxes in the graph represent GO terms labeled by their GO ID, term definition and statistical information. The significant (P-value < 0.05) and non-significant terms are marked with color and white boxes, respectively. The diagram, the degree of color saturation of a box is positively correlated to the enrichment level of the term. Solid, dashed, and dotted lines represent two, one, and zero enriched terms at both ends connected by the line, respectively.

SUPPLEMENTARY FIGURE 4
Expression map of GO for the 37 genes.

SUPPLEMENTARY FIGURE 5
(A) Tissue-specific expression profile, **(C)** Linkage disequilibrium, and haplotype block with 6 SNPs inside for the candidate gene GRMZM2G146192. **(C)** Comparison of trait GY among haplotypes I (GTCTCC), II (CTTGGC), III (CTCTCC), and IV (CACTCT). The number of stars represents the result of *t* test at different significance levels (*: 0.05; **: 0.01; ***: 0.001).

SUPPLEMENTARY FIGURE 6
(A) Tissue-specific expression profile, **(B)** Linkage disequilibrium, and haplotype block with 13 SNPs inside for the candidate gene GRMZM2G114789.. **(C)** Comparison of trait AD among haplotypes I (CCGGCCCAACACT), II (CCGGCCCAAGGCT), III (CCGGCCCAAGGTT), IV (TCGGCCCAACACT), V (TCGGCCCAAGGCT), VI (TCGGCCCAAGGTT), and VII (TCGGCTTCAGGTT). The number of stars represents the result of *t* test at different significance levels (*: 0.05; **: 0.01; ***: 0.001).

References

Augustine, R. C., York, S. L., Rytz, T. C., and Vierstra, R. D. (2016). Defining the SUMO system in maize: SUMOylation is up-regulated during endosperm development and rapidly induced by stress. *Plant Physiol.* 171 (3), 2191–2210. doi: 10.1104/pp.16.00353

Bänziger, M., Edmeades, G., Beck, D., and Bellon, M. (2000). Breeding for drought and nitrogen stress tolerance in maize: From theory to practice. Mexico, CIMMYT.

Barrett, J. C., Fry, B., Maller, J., and Daly, M. J. (2005). Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* 21 (2), 263–265. doi: 10.1093/bioinformatics/bth457

Bolaños, J., and Edmeades, G. O. (1996). The importance of the anthesis-silking interval in breeding for drought tolerance in tropical maize. *Field Crops Res.* 48 (1), 65–80. doi: 10.1016/0378-4290(96)00036-6

Brown, M. B., and Forsythe, A. B. (1974). Robust tests for the equality of variances. *J. Am. Stat. Assoc.* 69 (346), 364–367. doi: 10.1080/01621459.1974.10482955

Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* 103 (3), 338–348. doi: 10.1016/j.ajhg.2018.07.015

Cairns, J. E., Crossa, J., Zaidi, P., 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 (4), 1335–1346. doi: 10.2135/cropsci2012.09.0545

Campos, H., Cooper, M., Edmeades, G., Loffler, C., Schussler, J., and Ibanez, M. (2006). Changes in drought tolerance in maize associated with fifty years of breeding for yield in the US corn belt. *Maydica* 51(2), 369–381.

Cao, L., Lu, X., Zhang, P., Wang, G., Wei, L., and Wang, T. (2019). Systematic analysis of differentially expressed maize *ZmbZIP* genes between drought and rewatering transcriptome reveals bZIP family members involved in abiotic stress responses. *Int. J. Mol. Sci.* 20(17), 4103. doi: 10.3390/ijms20174103

Cerrudo, D., Cao, S., Yuan, Y., Martinez, C., Suarez, E. A., Babu, R., et al. (2018). Genomic selection outperforms marker assisted selection for grain yield and physiological traits in a maize doubled haploid population across water treatments. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00366

Chen, J., Xu, W., Velten, J., Xin, Z., and Stout, J. (2012). Characterization of maize inbred lines for drought and heat tolerance. *J. Soil Water Conserv.* 67 (5), 354–364. doi: 10.2489/jswc.67.5.354

Ciais, P., Reichstein, M., Viovy, N., Granier, A., Ogée, J., Allard, V., et al. (2005). Europe-Wide reduction in primary productivity caused by the heat and drought in 2003. *Nature* 437 (7058), 529–533. doi: 10.1038/nature03972

Crossa, J., Vargas, M., Van Eeuwijk, F., Jiang, C., Edmeades, G., and Hoisington, D. (1999). Interpreting genotype× environment interaction in tropical maize using linked molecular markers and environmental covariables. *Theor. Appl. Genet.* 99 (3-4), 611–625. doi: 10.1007/s001220051276

Dong, C. J., Liu, X. Y., Xie, L. L., Wang, L. L., and Shang, Q. M. (2020). Salicylic acid regulates adventitious root formation *via* competitive inhibition of the auxin conjugation enzyme CsGH3.5 in cucumber hypocotyls. *Planta* 252 (5), 1–15. doi: 10.1007/s00425-020-03467-2

Du, H., Huang, M., and Liu, L. (2016). The genome wide analysis of GT transcription factors that respond to drought and waterlogging stresses in maize. *Euphytica* 208, 113–122. doi: 10.1007/s10681-015-1599-5

Edmeades, G., Bolaños, J., Chapman, S., Lafitte, H., and Bänziger, M. (1999). Selection improves drought tolerance in tropical maize populations: I. gains in biomass, grain yield, and harvest index. *Crop Sci.* 39 (5), 1306–1315. doi: 10.2135/cropsci1999.3951306x

Fraire-Velázquez, S., and Balderas-Hernández, V. E. (2013). Abiotic stress in plants and metabolic responses," in Abiotic Stress-Plant Responses and Applications in Agriculture. 25-48. doi: 10.5772/54859

Guo, J. J., Han, X. T., Zhang, J., and Chen, J. T. (2018). High-density genetic linkage map construction and QTL mapping for kernel test weight and related traits in maize. *J. OF MAIZE Sci.* 26 (6), 27–32. doi: 10.13597/j.cnki.maize.science.20180605

Haider, S., Rehman, S., Ahmad, Y., Raza, A., Tabassum, J., Javed, T., et al. (2021). In silico characterization and expression profiles of heat shock transcription factors (HSFs) in maize (*Zea mays* l.). *Agronomy* 11(11), 2335. doi: 10.3390/agronomy11112335

Huang, R., Birch, C., and George, D. (2006). "Water use efficiency in maize production-the challenge and improvement strategies,"," in *Proceeding of 6th Triennial Conference* (Darlington Point, NSW: Maize Association of Australia).

Khan, N. H., Ahsan, M., Naveed, M., Sadaqat, H. A., and Javed, I. (2016). Genetics of drought tolerance at seedling and maturity stages in *Zea mays* l. *Span J. Agric. Res.* 14 (3), e0705. doi: 10.5424/sjar/2016143-8505

Li, N., Lin, B., Wang, H., Li, X., Yang, F., Ding, X., et al. (2019). Natural variation in *ZmFBL41* confers banded leaf and sheath blight resistance in maize. *Nat. Genet.* 51 (10), 1540–1548. doi: 10.1038/s41588-019-0503-y

Liu, X., Zhai, S., Zhao, Y., Sun, B., Liu, C., Yang, A., et al. (2013b). Overexpression of the phosphatidylinositol synthase gene (*ZmPIS*) conferring drought stress tolerance by altering membrane lipid composition and increasing ABA synthesis in maize. *Plant Cell Environ.* 36 (5), 1037–1055. doi: 10.1111/pce.12040

Liu, W. X., Zhang, F. C., Zhang, W. Z., Song, L. F., Wu, W. H., and Chen, Y. F. (2013a). *Arabidopsis* Di19 functions as a transcription factor and modulates *PR1*, *PR2*, and *PR5* expression in response to drought stress. *Mol. Plant* 6 (5), 1487–1502. doi: 10.1093/mp/sst031

Li, S., Wang, J., and Zhang, L. (2015). Inclusive composite interval mapping of QTL by environment interactions in biparental populations. *PloS One* 10 (7), e0132414. doi: 10.1371/journal.pone.0132414

Li, M., Zhang, Y. W., Zhang, Z. C., Xiang, Y., Liu, M. H., Zhou, Y. H., et al. (2022a). A compressed variance component mixed model for detecting QTNs and QTN-by-environment and QTN-by-QTN interactions in genome-wide association studies. *Mol. Plant* 15 (4), 630–650. doi: 10.1016/j.molp.2022.02.012

Li, M., Zhang, Y. W., Xiang, Y., Liu, M. H., and Zhang, Y. M. (2022b). IIIVmrMLM: The r and c++ tools associated with 3VmrMLM, a comprehensive GWAS method for dissecting quantitative traits. *Mol. Plant* 15 (8), 1251–1253. doi: 10.1016/j.molp.2022.06.002

Lukens, L. N., and Doebley, J. (1999). Epistatic and environmental interactions for quantitative trait loci involved in maize evolution. *Genet. Res.* 74 (3), 291–302. doi: 10.1017/S0016672399004073

Marondedze, C., Thomas, L., Serrano, N. L., Lilley, K. S., and Gehring, C. (2016). The RNA-binding protein repertoire of *Arabidopsis thaliana*. *Sci. Rep.* 6, 29766. doi: 10.1038/srep29766

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 (3), 269–283. doi: 10.1016/j.cj.2015.01.001

Mittler, R. (2006). Abiotic stress, the field environment and stress combination. *Trends Plant Sci.* 11 (1), 15–19. doi: 10.1016/j.tplants.2005.11.002

Monneveux, P., Sanchez, C., Beck, D., and Edmeades, G. (2006). Drought tolerance improvement in tropical maize source populations: evidence of progress. *Crop Sci.* 46(1), 180–191. doi: 10.2135/cropsci2005.04-0034

Moore, R., Casale, F. P., Jan Bonder, M., Horta, D., Franke, L., Barroso, I., et al. (2019). A linear mixed-model approach to study multivariate gene–environment interactions. *Nat. Genet.* 51 (1), 180–186. doi: 10.1038/s41588-018-0271-0

Muralla, R., Lloyd, J., and Meinke, D. (2011). Molecular foundations of reproductive lethality in *Arabidopsis thaliana*. *PloS One* 6 (12), e28398. doi: 10.1371/journal.pone.0028398

Ohtani, M., Ramachandran, V., Tokumoto, T., Takebayashi, A., Ihara, A., Matsumoto, T., et al. (2018). Identification of novel factors that increase enzymatic saccharification efficiency in *Arabidopsis* wood cells. *Plant Biotechnol.* 34 (4), 203–206. doi: 10.5511/plantbiotechnology.17.1107a

Parajuli, S., Ojha, B., and Ferrara, G. (2018). Quantification of secondary traits for drought and low nitrogen stress tolerance in inbreds and hybrids of maize (*Zea mays l.*). *J. Plant Genet. Breed.* 2 (1), 106.

Patil, I. (2021). Visualizations with statistical details: The 'ggstatsplot' approach. *J. Open Source Software* 6(61), 3167. doi: 10.21105/joss.03167

Pei, Y., Deng, Y., Zhang, H., Zhang, Z., Liu, J., Chen, Z., et al. (2022). EAR APICAL DEGENERATION1 regulates maize ear development by maintaining malate supply for apical inflorescence. *Plant Cell* 34 (6), 2222–2241. doi: 10.1093/plcell/koac093

Piepho, H. P. (2000). A mixed-model approach to mapping quantitative trait loci in barley on the basis of multiple environment data. *Genetics* 156 (4), 2043–2050. doi: 10.1093/genetics/156.4.2043

Qin, Q., Zhao, Y., Zhang, J., Chen, L., Si, W., and Jiang, H. (2022). A maize heat shock factor *ZmHsf11* negatively regulates heat stress tolerance in transgenic plants. *BMC Plant Biol.* 22 (1), 1–14. doi: 10.1186/s12870-022-03789-1

Qi, W., Yang, Y., Feng, X., Zhang, M., and Song, R. (2017). Mitochondrial function and maize kernel development requires Dek2, a pentatricopeptide repeat protein involved in nad1 mRNA splicing. *Genetics* 205 (1), 239–249. doi: 10.1534/genetics.116.196105

Ribaut, J. M., Betran, J., Monneveux, P., and Setter, T. (2009). "Drought tolerance in maize," in *Handbook of maize: Its biology*. Eds. J. L. Bennetzen and S. C. Hake (Berlin: Springer), 311–344. doi: 10.1007/978-0-387-79418-1_16

Sakuma, Y., Maruyama, K., Osakabe, Y., Qin, F., Seki, M., Shinozaki, K., et al. (2006a). Functional analysis of an *Arabidopsis* transcription factor, DREB2A, involved in drought-responsive gene expression. *Plant Cell* 18(5), 1292–1309. doi: 10.1105/tpc.105.035881

Sakuma, Y., Maruyama, K., Qin, F., Osakabe, Y., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2006b). Dual function of an *Arabidopsis* transcription factor DREB2A in water-stress-responsive and heat-stress-responsive gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 103 (49), 18822–18827. doi: 10.1073/pnas.0605639103

Seo, P. J., Xiang, F., Qiao, M., Park, J. Y., Lee, Y. N., Kim, S. G., et al. (2009). The MYB96 transcription factor mediates abscisic acid signaling during drought stress response in arabidopsis. *Plant Physiol.* 151 (1), 275–289. doi: 10.1104/pp.109.144220

Shi, J., Yan, B., Lou, X., Ma, H., and Ruan, S. (2017). Comparative transcriptome analysis reveals the transcriptional alterations in heat-resistant and heat-sensitive sweet maize (*Zea mays* l.) varieties under heat stress. *BMC Plant Biol.* 17 (1), 1–10. doi: 10.1186/s12870-017-0973-y

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., et al. (2017). agriGO v2.0: a GO analysis toolkit for the agricultural communit Update. *Nucleic Acids Res.* 45 (W1), W122–W129. doi: 10.1093/nar/gkx382

Wang, H. Q., Liu, P., Zhang, J. W., Zhao, B., and Ren, B. Z. (2020). Endogenous hormones inhibit differentiation of young ears in maize (*Zea mays l.*) under heat stress. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.533046

Wang, D., Zhu, J., Li, Z., and Paterson, A. (1999). Mapping QTLs with epistatic effects and QTL× environment interactions by mixed linear model approaches. *Theor. Appl. Genet.* 99(7), 1255–1264. doi: 10.1007/s001220051331

Wen, W., Araus, J. L., Shah, T., Cairns, J., Mahuku, G., Bänziger, M., et al. (2011). Molecular characterization of a diverse maize inbred line collection and its potential utilization for stress tolerance improvement. *Crop Sci.* 51 (6), 2569–2581. doi: 10.2135/cropsci2010.08.0465

Woodhouse, M. R., Cannon, E. K., Portwood, J. L., Harper, L. C., Gardiner, J. M., Schaeffer, M. L., et al. (2021). A pan-genomic approach to genome databases using maize as a model system. *BMC Plant Biol.* 21 (385), 1–10. doi: 10.1186/s12870-021-03173-5

Xiu, Z., Peng, L., Wang, Y., Yang, H., Sun, F., Wang, X., et al. (2020). *Empty Pericarp24* and *Empty Pericarp25* are required for the splicing of mitochondrial introns, complex I assembly, and seed development in maize. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.608550

Xu, K., Chen, S., Li, T., Ma, X., Liang, X., Ding, X., et al. (2015). *OsGRAS23*, a rice GRAS transcription factor gene, is involved in drought stress response through regulating expression of stress-responsive genes. *BMC Plant Biol.* 15 (1), 1–13. doi: 10.1186/s12870-015-0532-3

Xu, J., Yuan, Y., Xu, Y., Zhang, G., Guo, X., Wu, F., et al. (2014). Identification of candidate genes for drought tolerance by whole-genome resequencing in maize. *BMC Plant Biol.* 14 (83), 1–15. doi: 10.1186/1471-2229-14-83

Yuan, Y., Cairns, J. E., Babu, R., Gowda, M., Makumbi, 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. doi: 10.3389/fpls.2018.01919

Zandalinas, S. I., Fritschi, F. B., and Mittler, R. (2020). Signal transduction networks during stress combination. *J. Exp. Bot.* 71 (5), 1734–1741. doi: 10.1093/jxb/erz486

Zhang, J., Liao, J., Ling, Q., Xi, Y., and Qian, Y. (2022). Genome-wide identification and expression profiling analysis of maize AP2/ERF superfamily genes reveal essential roles in abiotic stress tolerance. *BMC Genomics* 23 (1), 1–22. doi: 10.1186/s12864-022-08345-7

Zhao, Y., Hu, F., Zhang, X., Wei, Q., Dong, J., Bo, C., et al. (2019). Comparative transcriptome analysis reveals important roles of nonadditive genes in maize hybrid an'nong 591 under heat stress. *BMC Plant Biol.* 19 (273), 1–17. doi: 10.1186/s12870-019-1878-8

Zhao, Q., Shi, X. S., Wang, T., Chen, Y., Yang, R., Mi, J., et al. (2023). Identification of QTNs, QTN-by-environment interactions, and their candidate genes for grain size traits in main crop and ratoon rice. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1119218

Zhu, J., and Weir, B. S. (1998). Mixed model approaches for genetic analysis of quantitative traits," in *Advanced Topics in Biomathematics*, 321-330.

Zuo, J. F., Chen, Y., Ge, C., Liu, J. Y., and Zhang, Y. M. (2022). Identification of QTN-by-environment interactions and their candidate genes for soybean seed oil-related traits using 3VmrMLM. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.109645

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 13 August 2023
DOI 10.3389/fpls.2023.1202755

OPEN ACCESS

EDITED BY
Shang-Qian Xie,
University of Idaho, United States

REVIEWED BY
Yang Xu,
Yangzhou University, China
Ya-Wen Zhang,
Huazhong Agricultural University, China

*CORRESPONDENCE
Li Fenghai
524376731@qq.com
Lv Xiangling
lvxiangling521@syau.edu.cn

RECEIVED 09 April 2023
ACCEPTED 18 July 2023
PUBLISHED 13 August 2023

CITATION
Ruidong S, Shijin H, Yuwei Q, Yimeng L, Xiaohang Z, Ying L, Xihang L, Mingyang D, Xiangling L and Fenghai L (2023) Identification of QTLs and their candidate genes for the number of maize tassel branches in F_2 from two higher generation sister lines using QTL mapping and RNA-seq analysis.
Front. Plant Sci. 14:1202755.
doi: 10.3389/fpls.2023.1202755

Identification of QTLs and their candidate genes for the number of maize tassel branches in F_2 from two higher generation sister lines using QTL mapping and RNA-seq analysis

Sun Ruidong, He Shijin, Qi Yuwei, Li Yimeng, Zhou Xiaohang, Liu Ying, Liu Xihang, Ding Mingyang, Lv Xiangling* and Li Fenghai*

Special Corn Institute, Shenyang Agricultural University, Shenyang, China

Tassel branch number is an important agronomic trait that is closely associated with maize kernels and yield. The regulation of genes associated with tassel branch development can provide a theoretical basis for analyzing tassel branch growth and improving maize yield. In this study. we used two high-generation sister maize lines, PCU (unbranched) and PCM (multiple-branched), to construct an F_2 population comprising 190 individuals, which were genotyped and mapped using the Maize6H-60K single-nucleotide polymorphism array. Candidate genes associated with tassel development were subsequently identified by analyzing samples collected at three stages of tassel growth *via* RNA-seq. A total of 13 quantitative trait loci (QTLs) and 22 quantitative trait nucleotides (QTNs) associated with tassel branch number (TBN) were identified, among which, two major QTLs, *qTBN6.06-1* and *qTBN6.06-2*, on chromosome 6 were identified in two progeny populations, accounting for 15.07% to 37.64% of the phenotypic variation. Moreover, we identified 613 genes that were differentially expressed between PCU and PCM, which, according to Kyoto Encyclopedia of Genes and Genomes enrichment analysis, were enriched in amino acid metabolism and plant signal transduction pathways. Additionally, we established that the phytohormone content of Stage I tassels and the levels of indole-3-acetic acid (IAA) and IAA-glucose were higher in PCU than in PCM plants, whereas contrastingly, the levels of 5-deoxymonopolyl alcohol in PCM were higher than those in PCU. On the basis of these findings, we speculate that differences in TBN may be related to hormone content. Collectively, by combining QTL mapping and RNA-seq analysis, we identified five candidate genes associated with TBN. This study provides theoretical insights into the mechanism of tassel branch development in maize.

KEYWORDS

maize, QTL, TBN, SNP array, RNA-seq

1 Introduction

As one of the most important food crops worldwide, maize is widely used in industry, agriculture, and animal husbandry (Huang et al., 2022). Indeed, in recent decades, the demand for maize has steadily increased to meet the needs of a rapidly expanding global population and economy. As such, breeding maize varieties with optimal agronomic traits is a key objective to achieve the desired increases in yield (Wang et al., 2018). In this regard, the tassel of maize, which was domesticated from the wild ancestor teosinte, is considered an important agronomic trait (Doebley et al., 1990; Matsuoka et al., 2002; Wei et al., 2018). During growth, the ear and tassel develop simultaneously and compete for nutrients when the overall nutrient uptake of maize remains unchanged (Lambert and Johnson, 1978; Brown et al., 2011). However, appropriately reducing the tassel volume and branch number can contribute to yield increases (Brewbaker, 2015). Compared with wild-type maize, yield increases of between 5% and 19% can be obtained by using artificially emasculated strains (Hunter et al., 1969; Lambert and Johnson, 1978). Given that reducing the TBN can increase the light transmittance and photosynthetic efficiency of the upper leaves (Duncan et al., 1967; Xu et al., 2017), breeders are more inclined to select for smaller tassels, with the aim of promoting increases in yield (Gao et al., 2007). However, a larger number of tassel branches can ensure sufficient pollen production, which in turn contributes to adequate seed quantity.

TBN is a complex quantitative trait controlled by multiple genes. Previous studies have analyzed the genetics of maize tassels by constructing numerous genetic populations with germplasm materials from different backgrounds. For example, an F_2 population comprising 6,872 individuals was constructed using the LX1 and LX2 lines for QTL mapping, resulting in the identification of *Ub4*, a potential candidate gene located on chromosome 6 (Li et al., 2019). Moreover, SICAU1212 and the maize-inbred lines 3237 and B73 were used to construct BC1S1, the subsequent analysis of which revealed 21 QTLs associated with TBN on chromosomes 2, 3, 5, and 7 (Chen et al., 2017). However, the establishment of high-density genetic maps of single-nucleotide polymorphism (SNP) markers and genome-wide association study (GWAS) analysis of natural populations provide powerful tools for the fine mapping and analysis of quantitative traits. For instance, Qin employed Mo17 as a test inbred line to conduct whole-genome association analysis and identified the tassel branch-related gene Q^{Dtbn1} (Qin et al., 2021). Using a similar strategy, Wu identified 63 QTLs distributed on 10 chromosomes, primarily concentrated on chromosomes 1, 2, and 7, that are associated with tassel branches (Wu et al., 2016). Moreover, several SNPs associated with tassel branching have been obtained based on the GWAS analysis of 513 inbred lines using a nonparametric model (Yang et al., 2014). However, most of the QTLs identified to date have been found to have small effect values or are readily affected by environmental factors, and consequently have not been applied in breeding practices.

With the rapid development of molecular biotechnology and bioinformatics, various key genes associated with tassel branch development have been identified, and their functions have been characterized. For example, ramosal1 (*Ra1*) and *Ra2* are transcription factors, whereas *Ra3* encodes a trehalose 6-phosphate phosphatase (TPP), and it has been established that *Ra2* and *Ra3* promote the expression of *Ra1*. Moreover, it has been observed that *ra1*, *ra2*, and *ra3* are associated with an increased TBN phenotype (Vollbrecht et al., 2005; Bortiri et al., 2006; Satoh-Nagasawa et al., 2006; Claeys et al., 2019). Genes from different transcription factor families are also involved in the regulation of TBN, notable among which is barren stalk 1 (*Ba1*), which encodes a basic helix-loop-helix (bHLH) transcription factor that influences TBN by regulating meristem transformation processes (Gallavotti et al., 2004). The ethylene response factor (ERF) family encoding the APETALA2 (AP2) transcription factor indeterminate spikelet 1 (*Ids1*) and sister of indeterminate spikelet 1 (*Sid1*) has also been demonstrated to regulate tassel development (Chuck et al., 1998; Chuck et al., 2008). Furthermore, three genes, namely, tassel sheath 4 (*Tsh4*), unbranched 2 (*Ub2*), and *Ub3*, belonging to the squamosa promoter binding-box transcription factor family, have been found to contribute to TBN regulation. Notably, these three genes are characterized by functional redundance, with single, double, and triple mutant plants showing marked reductions in TBN and an increase in the number of rows of spikes (Chuck et al., 2014). In addition, mutants of the gene liguleless 2 (*Lg2*), which regulates leaf angle, can also be characterized by lower TBNs (Walsh et al., 1998; Walsh and Freeling, 1999).

TBN development is also regulated by different plant hormones, including auxins, cytokinins (CKs), and strigolactones (SLs) (Isbell and Morgan, 1982; Ongaro and Leyser, 2008; Umehara et al., 2008; McSteen, 2009). Among these, auxins are synthesized in the shoot apical meristem (SAM) and transported downward by polar auxin transport, thereby inhibiting branch formation and inducing apical dominance. Contrastingly, CKs are synthesized in roots and stems and promote the synthesis of auxins and, thus, the development of collateral branches (Mueller and Leyser, 2011). CKs also regulate apical meristem size, whereas a loss of function of the lonely guy (Log) and wuschel (*Wus*) genes influences CK synthesis and transport, leading to early SAM termination, and modification of TBN development (Ongaro and Leyser, 2008; Umehara et al., 2008). As carotenoid-derived plant hormones, SLs are also involved in the regulation of branching. For instance, transgenic corn plants overexpressing maize *Dwarf 53* (*ZmD53*) are characterized by excessive tillering and reduced TBN, whereas *ZmD53* interacts with the SL receptor *ZmD14A/B* in a rac-Gr24-dependent manner (Liu et al., 2021). In this way, SLs influence auxin transport by regulating auxin export carrier proteins, thereby leading to altered TBN (Ongaro and Leyser, 2008; Durbak et al., 2012).

To gain further insights into the genetic regulation of maize TBN, in this study, we employed the Maize6H-60K gene array to produce a high-density genetic linkage map of the F_2 population generated using two sister lines, namely the unbranched inbred line, PCU, and multi-branched inbred line, PCM. Subsequently, the genetic linkage map and two-year phenotypic data were used to map QTLs associated with TBN. By analyzing the RNA-seq data, we compared the changes in gene expression between the two parents at different stages of tassel development. Furthermore, the results of QTL mapping and RNA-seq analysis were combined to screen for

candidate genes regulating TBN. Our findings in this study can be used as a reference for verifying the function of genes associated with TBN and provide a theoretical basis for genetic improvement of the maize tassel branch trait and associated molecular breeding.

2 Materials and methods

2.1 Plant materials and construction of mapping populations

The sister lines PCU and PCM were bred using the parents Xianyu 335 and Zheng 58, in which PCU was the non-branching material (TBN, 0) and PCM was the multi-branched material (TBN, 5–8), both of which were provided by the Special Maize Research Institute of Shenyang Agricultural University (Liaoning, China). A total of 994 pairs of simple sequence repeat (SSR) markers and SNP markers were used to assess PCU and PCM, which were established to have a genetic similarity of 93.17%. Subsequently, a single F_2 population comprising 190 plants was developed by crossing PCU and PCM within the experimental field of Shenyang Agricultural University (Shenyang, Liaoning, 41.48°N, 123.25°E). The $F_{2:3}$ population was planted at the Southern Breeding Base of Shenyang Agricultural University (Sanya, Hainan, 18.15°N, 109.30°E). The width and length of the single-row plot were 65 cm and 4 m, respectively, and the spacing between the plants was 20 cm, according to standard field management methods.

2.2 Determination and analysis of phenotype data

After the maize tassels had matured, we investigated the TBN phenotypes, with branches bearing more than one pair of small flowers being considered effective branches. The average branching number was used as the phenotype data for the $F_{2:3}$ population. The statistical parameters of TBN in the F_2 and $F_{2:3}$ populations were calculated using SPSS software version 24.0. Pearson correlation coefficients and phenotype frequency distribution maps were visualized using the R package ggpubr performance analytics.

2.3 Genetic mapping and QTL and QTN detection

The parent plants and 190 F_2 individuals were genotyped using a Maize6H-60K SNP array (Tian et al., 2021). Linkage analysis was performed using QTL ICIMAPPING 4.2 software (Meng et al., 2015), in which markers with no polymorphism between parents and a deletion rate > 10% were removed. The TBN was assessed using the inclusive composite interval mapping method (ICIM) in the software QTL ICIMAPPING 4.2 (Meng et al., 2015), composite interval mapping method (CIM) in the Windows QTL Cartographer 2.5 (Wang et al., 2012), and genome-wide composite interval mapping (GCIM) (https://cran.r-project.org/web/packages/QTL.gCIMapping/index.html) (Wen et al., 2019) and dQTG-seq2 (https://cran.r-project.org/web/packages/dQTG.seq/index.html) (Li et al., 2022). QTLs were evaluated based on 1,000 permutation tests with a significance level of 0.05 to determine the logarithm of the odds (LOD) threshold and thereby identify QTLs. A slightly more stringent criterion (P-value = 0.00316) was applied to denote significant QTLs, which was converted from an LOD score of 2.50. When adopting the dQTG-seq2 method, we used the 20% plants with the highest TBN as the high pool and the 20% of plants with the lowest TBN as the low pool.

2.4 RNA isolation and RNA-seq

The tissues of PCU and PCM tassels collected at three different stages of development, namely, the growth cone elongation stage (Stage I), the early stage of tassel differentiation (Stage II), and the later stage of tassel differentiation stage (Stage III), were immersed in an RNA storage solution (Li et al., 2019). PCU and PCM had similar tassel-branching stem tips during Stage I. However, it is uncertain as to whether the lateral meristems differentiated into tassel branches during Stage II. During Stage III, tassel branches at the base of PCU and PCM could be clearly distinguished. RNA extraction was performed using the TRIzol method (Rio et al., 2010).

For each line at each stage, we obtained three duplicate biological samples, and used the total 18 samples to construct a cDNA library. Construction and sequencing of the library were performed by Beijing Nohezhiyuan Bioinformation Technology Co., Ltd (Tianjin). Using an Illumina Hiseq™4000 high-throughput sequencing platform to obtain 100-bp double-terminal sequence reads, and FastQC tools (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to control the read quality. Low-quality reads were removed using Trimmomotic 0.36 (Bolger et al., 2014). The reference genome (AGPv4) was obtained from the maize database MaizeGDB (https://maizegdb.org). To calibrate the FastQC output, gene expression levels were normalized based on gene length and the number of reads, and the number of transcription fragments per kilobyte/million mapping reads (FPKM) was calculated. The DESeq software package was used to identify those genes that were differentially expressed (DEGs) between PCU and PCM (Anders and Huber, 2010).

Functional annotation and Gene Ontology (GO) analysis of genes were performed using Blast2go 4.1 (Conesa et al., 2005). whereas Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthology-based Annotation System KOBAS 2.0 software (http://kobas.cbi.pku.edu.cn) was used to perform pathway enrichment analysis. The P-value of each gene was adjusted using the Benjamini and Hochberg method to control the false discovery rate. P-values < 0.05 and $|\log2FC| \geq 1$ were applied as thresholds to identify DEGs.

Venn diagrams are drawn by online sites. (https://bioinfogp.cnb.csic.es/tools/venny/index.html)

2.5 qRT-PCR

RNA derived from tassels collected at the three stages (Stage I, II, and III) was assessed *via* qRT-PCR, for which primers were

designed using Primer BLAST (https://www.ncbi.nlm.nih.gov/tools/primer-blast) (Table S8). All primers were synthesized and supplied by Shenggong Biotech Co., Ltd. The housekeeping gene *Gapdh* was used as the internal reference gene, the relative expression levels of which were calculated using the $2^{-\Delta\Delta ct}$ method.

2.6 Determination of hormone content

Tassels collected at Stage I were exfoliated, flash frozen in liquid nitrogen, and stored at -80°C. A standard plant hormone solution was prepared using a 50% formaldehyde solution, and 10 µL of an internal standard plant hormone solution was added to 50 µL of a concentration gradient of standard plant hormone solutions. Thereafter, 1 mL of methanol/water/formic acid mixture (15:4:1, v/v/v) was added, followed by vortexing for 10 min (until thoroughly mixed), and the resultant mixture was allowed to stand for 12 h. The auxin, CK, ethylene (ETH), abscisic acid (ABA), gibberellin (GA), and SL contents of the tassels were determined by analyzing the resultant supernatant *via* liquid chromatography in conjunction with tandem mass spectrometry (LC-MS/MS).

2.7 Identification of candidate genes

Genes located in the vicinity of large loci with an R^2 value > 10%, and which were stable across 2 years, were used for gene annotation. Gene annotation information was obtained using MaizeGDB (https://maizegdb.org) and Phytozome (http://phytozome.jgi.doe.gov). Gene expression in PCU and PCM was analyzed using RNA-seq data and applied to predict gene function that might be associated with tassel branching in maize.

2.8 Cloning and sequence alignment of Zm00001d038537

The candidate gene *Zm00001d038537* was extracted from the genomic DNA and cDNA of PCU and PCM. The primers used for amplification are listed in Supplementary Table S8. DNA sequence alignment was performed using SnapGene software (https://www.snapgene.com/).

3 Results

3.1 Statistical differences in plant architectural traits and phenotypic analysis in sister lines

Architectural traits of plants of the sister lines PCU and PCM were compared and analyzed. Apart from leaf length, leaf width, leaf angle, and TBN, we detected no significant differences between the two lines with respect to plant architecture (Table 1). Notably, over the 2 years of the study, we detected a significant difference between the parent lines with respect to TBN, with PCM being characterized by a larger number of tassel branches, whereas under certain environmental conditions, PCU had no branches, thereby indicating that these phenotypic traits of the parents are probably stable (Table 2).

The TBN of the F_2 population ranged from 0 to 11, with a coefficient of variation of 99.65%, whereas in the $F_{2:3}$ population, the TBN ranged from 0 to 5.43, with a coefficient of variation of 76.92%. In both offspring populations, the number of tassel branches was maintained at an average of that of the two parents (Table 2). Moreover, we detected a highly significant correlation between F_2 and $F_{2:3}$. The TBN of the two offspring groups was biased toward PCU and exhibited a continuous distribution trend (Figure 1). In addition, the skewness and kurtosis results revealed that both populations conformed to the quantitative trait characteristics of skewed normal distribution and polygene control (Table 2). Accordingly, the two progeny populations were assumed to meet the requirements for QTL mapping.

3.2 QTL and QTN identification and effect calculations

The F_2 population was genotyped using the Maize6H-60K SNP array, which contains 61,214 SNP markers covering the entire maize genome. A genetic linkage map was constructed by screening high-quality genotype-independent SNP markers with deletion rates < 10% between the two parents, from which we obtained 4,136 SNP markers (Table S1). The linkage map covered a distance of 2,095.02 cM, with an average distance of 0.51 cM between markers. The number of SNP markers on each chromosome ranged from 46 to 710, with a linkage distance ranging from 37.22 to 410.78 cM (Table 3). As the two parents are higher generation sister lines with high background similarity, the SNP differences detected on chromosomes 4 and 9 were small (Figure 2).

Combined with phenotype data of the two populations and the F_2 genetic linkage map, QTLs for the TBN of F_2 and $F_{2:3}$ were identified using ICIM, CIM, and GCIM methods. Within the two populations, we detected 13 QTLs associated with TBN on chromosomes 3, 6, and 7, with LOD values ranging from 5.10 to 40.78 and accounting for 6.86% to 37.64% of the phenotypic variation (Table 4; Figure S4). Excluding *qTBN-3-4* and *qTBN-3-5*, which exhibited a positive additive effect attributable to the PCM allele, the other QTL sites showed negative additive effects associated with the PCU allele. In addition, we identified 22 SNPs significantly associated with TBN based on dQTG-seq2 mapping. Compared with other methods, we identified new SNPs on chromosomes 1, 2, 4, and 5 when using dQTG-seq2. The upstream and downstream 50 kb of the significantly associated SNPs were used as the intervals for predicting candidate genes (Table 5; Figure S5) (Li et al., 2013).

On the basis of statistical analysis of QTLs and QTNs, we identified two QTLs on chromosome 6 with R^2 > 10%, namely, *qTBN6.06-1* (157846342–159598073 bp) and *qTBN6.06-2* (159648428–159792909 bp) (Table 4). Moreover, we identified

TABLE 1 Statistical difference of agronomic traits in sister lines.

Traits	PCU		PCM	
	Mean	SD [a]	Mean	SD [a]
Plant height(cm)	223.4	4.4	220.1	3.7
Ear height(cm)	88.4	2.1	85.7	2.4
Leaf angle(°)	31.2	3.0	67.7**	4.0
Leaf length(cm)	76.3	3.0	66.1**	3.0
Leaf width(cm)	10.9	0.4	8.6**	0.5
TBN	0.0	0.0	5.1**	1.2
Stem diameter(mm)	26.5	2.0	25.6	1.9
Ear length(cm)	15.7	1.1	15.4	0.9
Ear diameter(mm)	36.9	0.9	36.4	0.7
Ear rows	14.0	0.0	14.0	0.0
Hundred grain weight(g)	26.7	1.1	25.3	0.9

[a] SD, Standard Deviation. The asterisks (*or **) represent the significant differences at P < 0.05 or P< 0.01, respectively.

candidate genes in the two QTLs based on the physical location of the SNP markers. *qTBN6.06-1* and *qTBN6.06-2* contained 73 and 14 genes, respectively (Table S2). In contrast to the findings of previous studies, we failed to identify any TBN-related genes in *qTBN6.06-1* and *qTBN6.06-2*. Hence, we used the online tool Web Gene Ontology Annotation Plot (WEGO) 2.0 (Ye et al., 2018) to annotate the candidate genes within the two QTL intervals. The results revealed that binding (GO:0005488), metabolic process (GO:0008152), and cellular process (GO:0009987) were the three main GO entries for the 84 genes in the two QTLs (Figures S1; S2), and consequently, we speculate that tassel development is associated with these processes.

3.3 RNA-seq analysis

Despite our GO enrichment analysis of genes within the localized intervals, differences in gene expression during tassel development remained undetermined. Consequently, to identify the genes responsible for tassel branch development, we compared the DEGs (|Log2-fold change| ≥ 1 and *P*-value < 0.05) between PCU and PCM at the three assessed developmental stages. We analyzed DEGs common to Stages I, II, and III, among which, 317 and 292 genes were up- and downregulated, respectively (Figures 3A–D; Table S3). GO enrichment analysis revealed a significant enrichment of 118 biological processes (Table S4), which are primarily associated with the growth and development of tissues or cells, including pollen tube growth, cell tip growth, amino acid kinase activity, developmental cell growth, and the endoplasmic reticulum lumen (Figure 4A). In addition, we identified enrichment of several pathways associated with enzyme activity, including those of endonuclease, endoribonuclease, mitogen-activated protein (MAP) kinase, inositol-3-phosphate synthase, and glyceraldehyde-3-phosphate dehydrogenase (NADP+) (phosphorylating). Therefore, we speculate that the activities of different enzymes also influence TBN.

KEGG enrichment analysis further revealed that DEGs were enriched in glycine, serine, and threonine metabolism; taurine and taurine metabolism; plant hormone signal transduction; ATP-binding cassette (ABC) transporter superfamily (Figure 4B; Table S6). In maize, BARREN INFLORESCENCE2 (*Bif2*) encodes a serine/threonine protein kinase *Bif2* phosphorylates *ZmPIN1a*, *Bif2* regulates auxin transport through direct regulation of *ZmPIN1a* during maize inflorescence development (Skirpan et al., 2009; Forestan et al., 2012). The main functions of ABCB protein in ABC transporter family are auxin transport. In *Arabidopsis thaliana* studies, it was found that *ATABCB1*, *ATABCB6*, ATABCB14, *ATABCB15* and *ATABCB20* all participated in auxin transport in inflorescence axis, which further affected the growth and development of inflorescence axis (Okamoto et al., 2016). Thus, the above pathways may be involved in TBN development. Among these, 12 genes were enriched in plant hormone signaling pathways,

TABLE 2 Mean, extreme, Standard Deviation (SD), Coefficient of Variation (CV), Skewness and Kurtosis of the TBN in parents and F_2, $F_{2:3}$ populations.

Parents			offspring of PCU×PCM						
	PCU [a]	PCM [a]	Min	Max	Mean	SD [b]	CV (%) [c]	Skewness	Kurtosis
F_2	0	5.1	0	11	2.55	2.54	99.65	0.97	0.36
$F_{2:3}$	0	4.9	0	5.43	1.71	1.32	76.92	0.52	-0.57

[a] Mean TBN of PCU and PCM calculated from 10 plants per parent in two rows. [b] SD, Standard Deviation; [c] Coefficient of Variation.

10.3389/fpls.2023.1202755

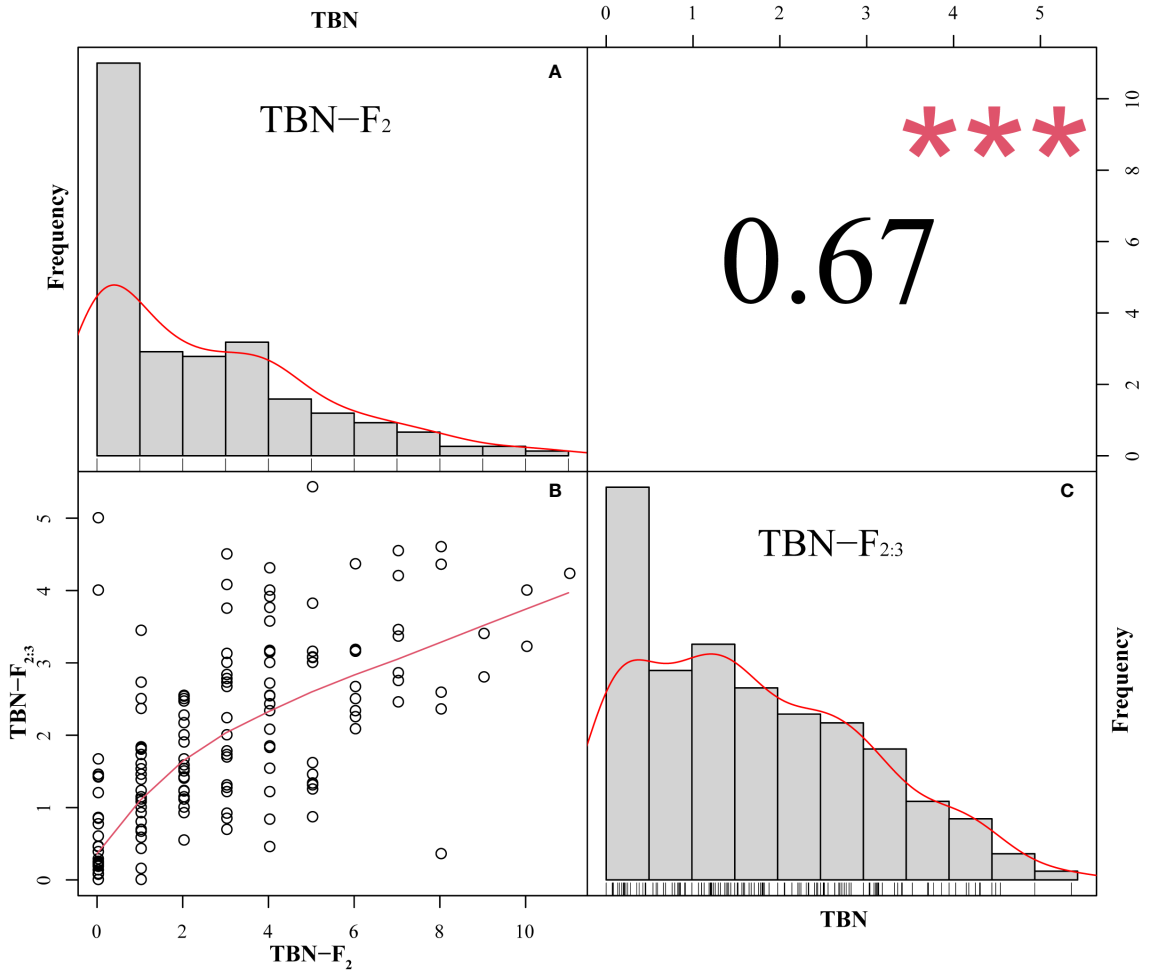


FIGURE 1
Frequency distribution and correlation of TBN of F_2 and $F_{2:3}$, ***P<0.001. **(A)** The horizontal coordinate is TBN and the vertical coordinate is the frequency. **(B)** The TBN distribution of the F_2 population in the horizontal coordinate and the $F_{2:3}$ population in the vertical coordinate. **(C)**: The horizontal coordinate is TBN and the vertical coordinate is the frequency.

TABLE 3 Total SNP numbers and linkage distances of chromosomes in F_2 population.

Chromosome	Number of SNPs	Linkage Distance(cM)	Average Distance between Markers(cM)
1	726	348.72	0.48
2	235	171.88	0.73
3	759	374.11	0.49
4	46	37.22	0.81
5	678	410.78	0.61
6	809	340.48	0.42
7	341	182.79	0.54
8	249	103.97	0.42
9	103	46.73	0.45
10	190	78.34	0.41
Total	4136	2095.02	0.51

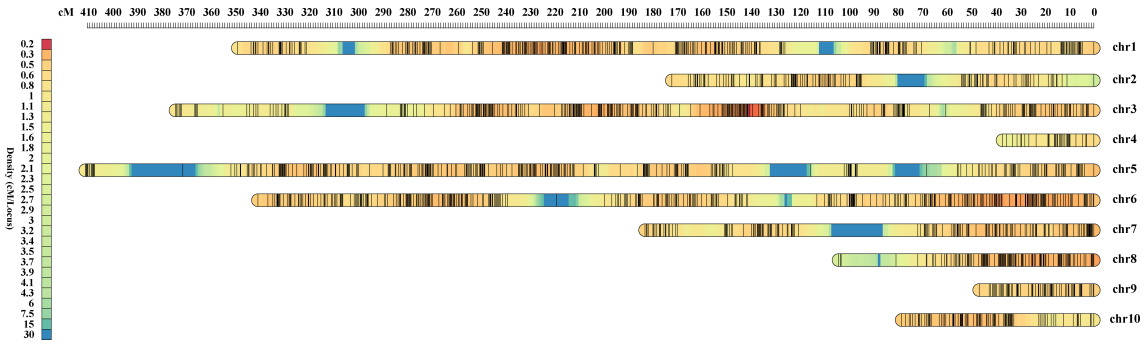


FIGURE 2
Genetic map of F_2 population. The upper ruler shows the distance between SNP markers in centimorgans (cM), the color shades of the left ruler represent the density of SNP markers on each linkage group.

seven of which were associated with indole-3-acetic acid (IAA) signaling. Other pathways were primarily associated with amino acid anabolism. Accordingly, KEGG pathway analysis provided evidence to indicate that tassel branch development might be associated with hormone and energy metabolism (Figure 4C).

Simultaneously, we annotated the DEGs, on the basis of which we retrieved 64 transcription factors, among which myeloblastosis (MYB)-related genes (seven) were the most common, followed by ERF (six), bHLH (five), and C2H2 (Cys2/His2-type; four) genes. In addition, we also identified three auxin response factors (ARFs). Interestingly, the expression of most MYB-related genes in PCU was higher than that in PCM, whereas ERF transcription factor expression was downregulated in PCM (Figure 4D; Table S7).

We also performed GO enrichment analysis for genes differentially expressed in only one of the three assessed stages. Those exclusively identified in Stage I were primarily enriched in the regulation of nitrogen compound metabolic processes, regulation of primary metabolic processes, and regulation of nucleic acid-templated transcription, which are closely associated with plant growth and development (Table S5). Moreover, certain genes known to regulate tassel development in maize were analyzed (Figure 4E), most of which were differentially expressed in Stage I, with the variance fold change being greater than that in the other two stages. On the basis of these findings, we assume that Stage I is critical to the regulation of tassel development.

3.4 Determination of hormone content

Our KEGG results provided evidence to indicate that DEGs were enriched in plant hormone signal transduction, and we speculated that Stage I was the key stage responsible for the observed differences between PCU and PCM with respect to TBN. We thus used samples of Stage I PCU and PCM tassels to quantify hormone content, which revealed that the content of IAA in PCU was slightly higher than that in PCM, whereas the respective contents of tryptamine (TRA) and tryptophan (TRY), two important precursors in the auxin synthesis pathway, were significantly higher in PCU. In addition, the content of IAA-glc, an important form of stored IAA. was found to be 7.9-fold higher in PCU than in PCM, whereas in contrast, the content of 5-deoxymonopolyl alcohol (5DS), the first active product of the SL biosynthetic pathway, was found to be significantly higher in PCM than in PCU. However, we detected no significant differences between the lines with respect to the levels of ABA, trans-zeatin (tZ), or 1-aminocyclopropanecarboxylic acid (ACC). On the basis of these observations, we can speculate that differences in the tassel branching phenotypes of the two parent lines are attributable, at least in part, to differences in the contents of IAA and 5DS (Figure 5).

3.5 Predicting candidate genes

To screen for candidate genes, we selected 614 common DEGs to cross-analyze the mapping interval. The interval *qTBN6.06-1* comprised 73 protein-coding genes, 27 of which were negligibly expressed during the three stages of tassel development, and 38 showed no significant differences. Only two genes, *Zm00001d038519* and *Zm00001d038523*, were differentially expressed at all three stages. Of the 14 protein-encoding genes present within *qTBN6.06-2*, only *Zm00001d038546* and *Zm00001d038552* were identified as being differentially expressed during the three stages.

These four candidate genes were annotated using Phytozome (https://phytozome-next.jgi.doe.gov/), using which, *Zm00001d038519* was predicted to contain a putative S-adenosyl-L-methionine-dependent methyltransferase domain, which regulates plant growth and development *via* methylation. We thus inferred that *Zm00001d038519* might have a similar function. *Zm00001d038546* was found to contain a Myb-like DNA-binding domain and thus could be a member of the MYB family of transcription factors that are primarily involved in inflorescence development and the segregation of lateral organs. However, using this approach, we were unable to predict structures for *Zm00001d038523* or *Zm00001d038552*. The four candidate genes were verified *via* qRT-PCR analysis, and the results were consistent with those obtained based on RNA-seq (Figure 6).

In addition, our annotation of genes in the *qTBN6.06-1* interval revealed a gene encoding the F-box structural domain

10.3389/fpls.2023.1202755

TABLE 4 Analysis of TBN-related QTLs in offspring population from PCU×PCM.

QTL [a]	Chromosome	Mapping interval/ bp [b]	Position	LOD [c]	Additive effect	Dominant effect	R^2(%) [d]	Generation	Method
qTBN-3-1	3	179394655-179625328	96	6.59	-0.96	0.05	9.25	F_2	ICIM
	3	179392238-179900293	96	6.40	-0.95	0.09	6.98	F_2	GCIM
qTBN-3-2	3	134150716-178936874	98	6.37	-1.24	0.10	7.84	F_2	CIM
qTBN-3-3	3	182413848-182508246	48	6.80	-0.40	0.15	6.80	$F_{2:3}$	ICIM
qTBN-3-4	3	2019660-2050620	333	8.35	0.03	0.62	8.34	$F_{2:3}$	ICIM
qTBN-3-5	3	1473821-1548536	353	5.39	0.09	-0.49	5.39	$F_{2:3}$	ICIM
qTBN-6-1	6	157846342-159598073	235.9	38.89	-1.47	-0.23	37.64	$F_{2:3}$	CIM
	6	157846342-159598073	237	40.78	-1.27	-0.08	40.77	$F_{2:3}$	ICIM
	6	159231856-159316218	240	20.83	-1.91	-0.30	34.63	F_2	ICIM
	6	159231856-159316218	240	15.46	-1.88	-0.27	27.63	F_2	GCIM
	6	159116395-159231856	242.5	33.35	-1.27	-0.05	34.81	$F_{2:3}$	CIM
	6	159141240-159355691	243.8	15.62	-1.87	-0.27	15.07	F_2	CIM
	6	159355691-159538438	244.8	4.62	-0.64	-0.68	18.40	$F_{2:3}$	GCIM
qTBN-6-2	6	159648428-159792909	246.5	33.84	-1.36	0.00	37.02	$F_{2:3}$	CIM
qTBN-6-3	6	160665895-160691260	253	5.04	-0.93	-0.82	34.54	$F_{2:3}$	GCIM
qTBN-6-4	6	160691260-160895678	254.7	11.03	-1.80	-0.38	9.39	F_2	CIM
	6	160691260-160895678	254.7	28.63	-1.38	-0.13	28.94	$F_{2:3}$	CIM
qTBN-6-5	6	168094283-168363228	307	4.77	-0.80	-0.43	4.61	F_2	GCIM
	6	168200733-168363228	318	5.10	-0.77	-0.35	6.86	F_2	ICIM
	6	169161160-169372663	319	6.52	-0.37	0.00	6.52	$F_{2:3}$	ICIM
qTBN-7-1	7	127691371-128260837	149	5.39	-0.77	-0.37	7.34	F_2	ICIM
	7	127691371-128260837	149	4.77	-0.72	-0.40	4.61	F_2	GCIM
qTBN-7-2	7	123889115-125102662	145	5.42	-0.32	0.02	5.42	$F_{2:3}$	ICIM
qTBN-7-3	7	128260837-172487130	111.8	5.50	-0.47	-0.15	4.63	$F_{2:3}$	CIM
	7	125921578-127728775	128.6	5.38	-0.37	0.03	4.36	$F_{2:3}$	CIM

[a]QTL detected in different methods and generations at the same, adjacent, or overlapping marker intervals was considered as the same QTL. [b]Physical position of the 95% confidence interval for the detected QTL. [c]LOD (Logarithm of odds) value at the peak likelihood of the QTL. [d]Phenotypic variance (R^2) explained by the detected QTL.

Zm00001d038537. Members of the F-box family of proteins can play roles in forming Skp1-Cullin-F-Box (SCF) structural complexes that ubiquitinate specific proteins and thereby promote their degradation, which is similar to processes that can also occur in the IAA metabolic pathway. The KEGG enrichment results provided evidence to indicate that phytohormone signaling, particularly IAA signaling, might contribute to the observed differences in TBN, as well as differences in the IAA content of parent tassels. Although *Zm00001d038537* was not differentially expressed in the parents, we inferred that *Zm00001d038537* might

TABLE 5 Significant QTNs for TBN in F_2 and $F_{2:3}$ using dQTG-seq2 method.

Generation	Maker	Chromosome	Position	Mapping interval/bp	Gw [a]	Smooth_Gw [b]
F_2	AX-108052314	1	227992408	227942408-228042408	6.91	7.78
	AX-108019986	3	178936874	178886874-178986874	7.52	8.43
	AX-107939474	3	180214656	180164656-180264656	10.77	10.31
	AX-86317565	6	159792909	159742909-159752909	102.67	102.02
	AX-91021926	6	172603449	172553449-172653449	17.68	18.41
$F_{2:3}$	AX-247233306	2	223266472	223176472-223276472	9.03	9.81
	AX-107941057	3	110309750	110259750-110359750	8.41	8.85
	AX-108009558	3	117879603	117829603-117929603	8.84	8.68
	AX-108061753	3	130582492	130532492-130632492	10.67	9.7
	AX-90827906	3	132093889	132043889-132143889	9.84	9.73
	AX-108019986	3	178936874	178886874-178986874	11.11	17.48
	AX-247236770	4	824775	774775-874775	10.9	10.65
	AX-107945551	4	3672068	3622068-3722068	8.29	9.15
	AX-178079230	5	7392849	7342849-7352849	10.66	9.24
	AX-107981631	5	212584879	212534879-212634879	9.47	12.15
	AX-107989634	5	222130069	222080069-222180069	10.13	10.07
	AX-108011870	6	150255513	150205513-150305513	27.67	26.13
	AX-91016539	6	153616434	153566434-153666434	16.88	14.1
	AX-86317565	6	159792909	159742909-159752909	62.31	64.83
	AX-86294633	6	163542081	163492089-163592089	59.83	58.99
	AX-86301494	6	166848539	166798539-166898539	24.81	24.47
	AX-91021926	6	172603449	172553449-172653449	17.57	18.92

[a] Gw: The value of statistic Gw calculated by the dQTGseq2 method. [b] Smooth_Gw: smooth Gw value of one marker via the window size method.

be a candidate gene responsible for TBN differences. Cloning and sequencing of this gene in both parents revealed three SNPs, the first and third of which encoded different amino acids (Figure 7), resulting in different encoded proteins. These differences were found to influence IAA signaling and led to differences in the number of male spike branches. Consequently, *Zm00001d038537* was included as a candidate gene.

4 Discussion

In this study, in which we sought to gain insights into the genetic regulation of tassel development in maize, we used the unbranched parent PCU and multi-branched parent PCM, two high-generation sister lines with high background similarity, to construct a genetic linkage map with a small distribution of markers on a single chromosome. PCU was characterized by an absence of tassel branching under different environmental conditions, thereby indicating that the branching trait in this line is not subjected to environmental control. However, on the basis of our field observations and analysis of natural seed setting rates, we identified no significant differences between the PCU and PCM lines.

Thirteen QTLs were identified on chromosomes 3, 6, and 7. Maize chromosomes 3 and 7 are known hotspots for QTL localization, containing genes associated with tassel development, including *Lg2*, nana plant 1 *(Na1)*, *Ba1*, *Sid1*, *Tsh4*, and *Ra3* (Walsh and Freeling, 1999; Satoh-Nagasawa et al., 2006; Chuck et al., 2007; Chuck et al., 2008; Gallavotti et al., 2008; Hartwig et al., 2011; Phillips et al., 2011). Previously, Chen et al. (2014) constructed an F_2 population comprising 708 individual strains and detected seven TBN-related QTLs, among which the location results obtained for chromosome 3 coincided with *qTBN-3-3*. Moreover, Wang performed similar analyses on the progeny of natural and doubled-haploid populations, and accordingly identified 12 loci (distributed on chromosomes 1, 2, 3, 4, 6, and 7) consistent with multiple environments. Among these, the QTLs located on chromosome 3 overlap with the those observed in the current

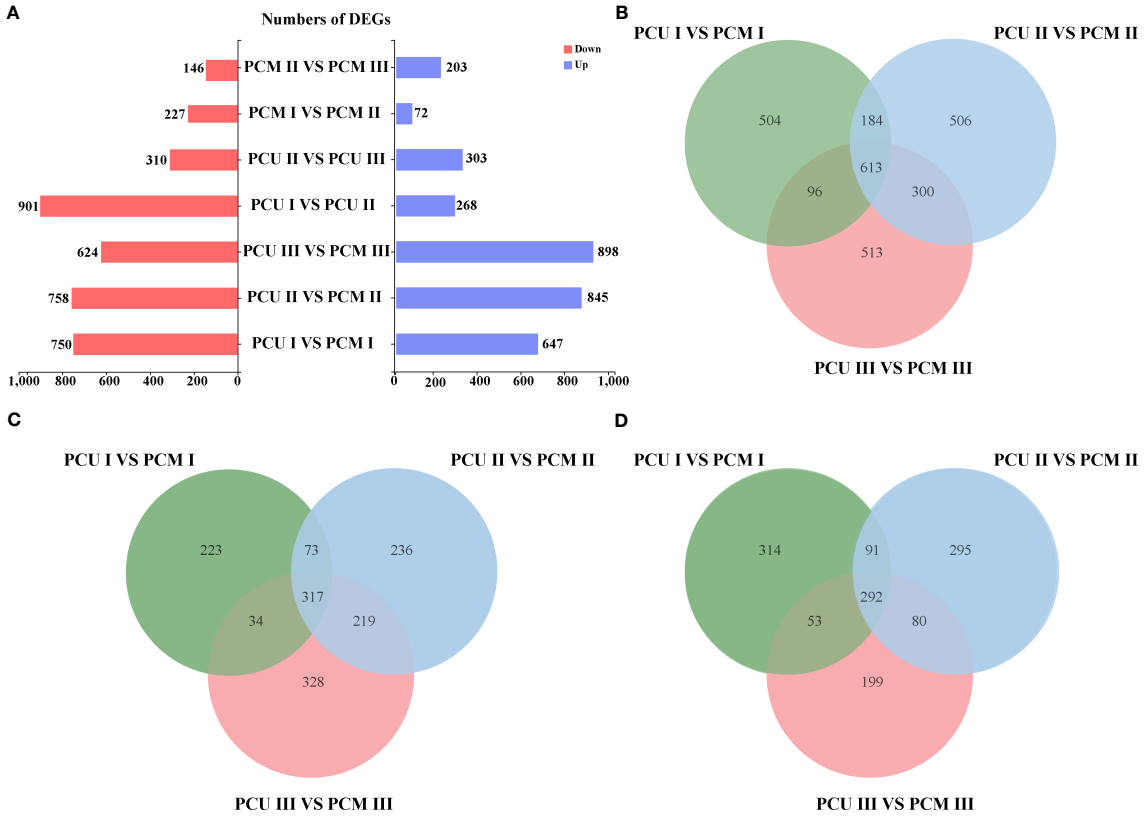


FIGURE 3
Numbers of PCU and PCM differentially expressed genes in Stage I, II and III. **(A)** Comparison of the number of differentially expressed genes in different stages and between different parents. **(B)** Number of co-differentially expressed genes in Stage I, Stage II, and Stage III. **(C)** The number of genes is co-upregulated in three stages. **(D)** The number of genes is co-downregulated in three stages.

study. Moreover, our transcription data also revealed notable differences in the predicted candidate gene *Zm00001d042794* (Wang et al., 2019). Therefore, we identified a new QTL (*qTBN-3-1*) on chromosome 6, which coincides with *Lg2*, a gene that has been established to control leaf angle and TBN in maize. In addition, *qTBN-7-3* was found to harbor *Tsh4* (which is associated with tassel development) and *Ra3* (which is known to regulate the number of tassel branches), which coincide respectively with the *qBTBN7-1* and *qXTBN7-1* loci mapped by Wang et al., 2018.

In this study, the QTL identified on chromosome 6 accounted for 9.39% to 40.77% of the phenotypic variation and was detected in different environments. Similarly, previous studies have identified 14 TBN-related loci on chromosome 6, classified into seven groups on the basis of their physical locations (Li et al., 2019). However, these loci contributed to less than 10% of the observed phenotypic differences and did not coincide with the results of the present study. Furthermore, although the QTLs localized in the present study overlap with those reported by Yi et al. (2018), the distribution range detected by Yi et al. was relatively large, making it difficult to directly compare the respective QTLs.

Auxin is an important hormone involved in plant growth and development and is one of several hormones known to influence tassel branching in plants. *Vt2* (vanishing tassel 2) (Phillips et al., 2011) and *Spi1* (sparse inflorescence 1) (Gallavotti et al., 2008) have been identified as genes involved in auxin synthesis, the mutation of which has been found to coincide with a reduction in maize TBN, thereby providing evidence to indicate that these genes are involved in the initiation and growth of the axillary meristem during maize tassel development. In the present study, we combined our hormone determination results with the findings of KEGG pathway enrichment analysis to elucidate the regulatory pathways from hormones to response genes (Figure 5; Table S3). Auxin synthesis pathways can be divided into two main categories, namely, tryptophan (TRP)-dependent and TRP-independent (Mano and Nemoto, 2012), and the pathways involved in IAA metabolism primarily include IAA oxidation and methylation, resulting in the formation of conjugates with polysaccharides and amino acids (Zhao, 2012). In this study, we assessed the auxin synthesis pathway by synthesizing IAA *via* TAM, which entailed analyses of the contents of TRP, TAM, IAA, IAA-Glu, IAA-glc, IAA-ASP, MeIAA, and oxIAA. By mapping the auxin anabolic and gene response pathways based on KEGG results, we found that the contents of TRP and TAM in the PCU line were significantly higher than those in the PCM line, whereas IAA contents in the two lines was relatively similar, with only slightly higher levels being detected in PCU. Among the assessed IAA metabolites, only the content of IAA-glc was markedly higher in PCU than in PCM. On the basis of these observations, we thus infer that whereas larger amounts of IAA are synthesized in PCU, a large proportion is stored in the form

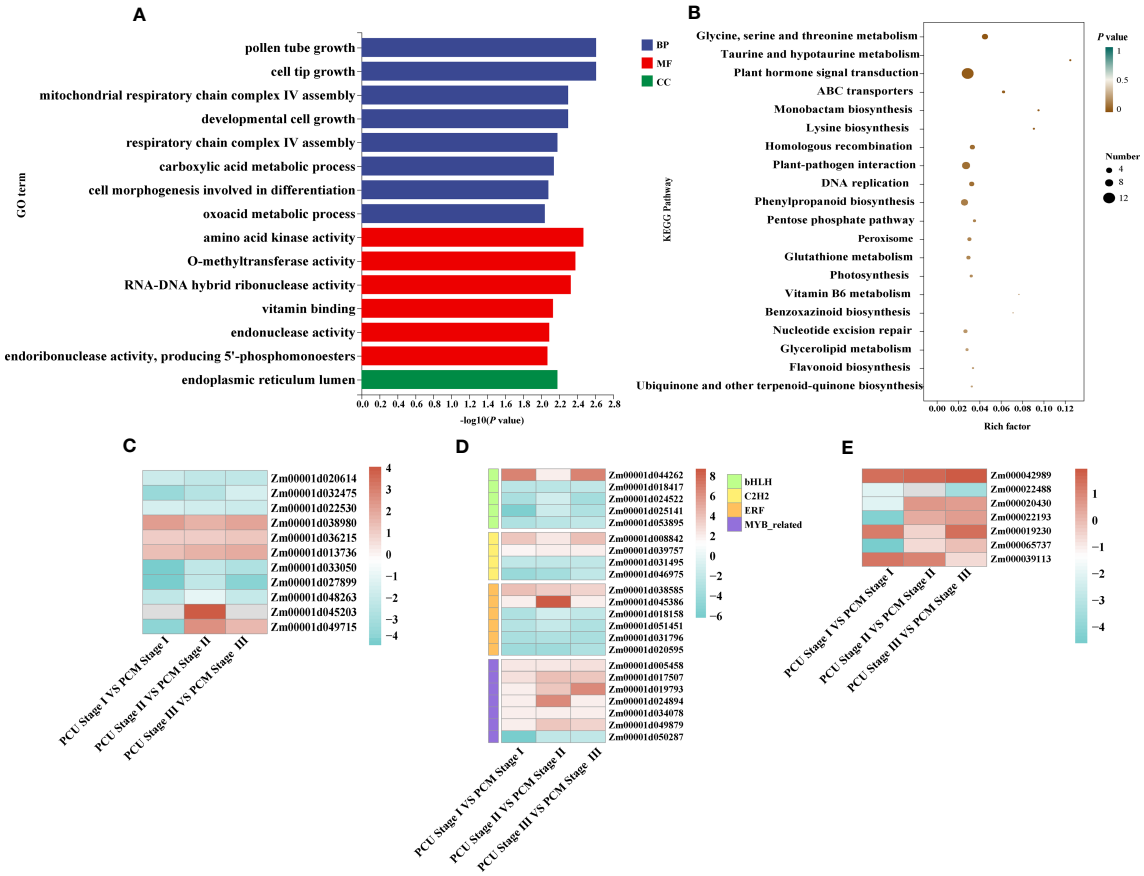


FIGURE 4
GO and KEGG analysis and changes in the expression levels of differentially expressed genes (DEGs). **(A)** GO enrichment analysis was executed with DEGs identified between PCU and PCM. The ordinate and abscissa represent the main biological process GO terms and -Log10(*P*-value), respectively. **(B)** KEGG enrichment analysis was executed with DEGs identified between PCU and PCM. The ordinate and abscissa represent the major KEGG biological pathways and rich factor, the size of the dots represents the number of genes enriched, respectively. **(C)** Expression levels of plant hormone signal transduction pathway-related genes. **(D)** Gene expression levels of different transcription factor families. **(E)** The expression levels of genes related to tassel development are known. The value is the log2 fold-change (log2(FC)) of each gene. The colors of the boxes represent upregulated (red) and downregulated (blue) genes.

of IAA-glc, and thus the levels of IAA detected in the two the parental lines tend to be similar (Figure 8).

Auxin signal transduction is regulated by multiple genes, and IAA enters the cell nucleus through the amino acid permease input carrier protein (auxin resistant-like aux1, *AUX/LAX*) (Swarup and Péret, 2012). In response to low IAA concentrations, auxin/indole-acetic acid genes (*AUX/IAA*) form a heterodimer with *ARFs* (Enders and Strader, 2015), thereby inhibiting the expression of downstream genes. Conversely, when present at high concentrations, IAA combines with transport inhibitor resistant 1/auxin signaling F-box (TIR1/AFB) and AUX/IAA. TIR1/AFB participates in the formation of SCF E3 ubiquitin ligase (Fendrych et al., 2018), resulting in the polyubiquitination of AUX/IAA, subsequent degradation *via* 26S proteasome, and the release of *ARF* inhibition. This also promotes or inhibits the expression of downstream IAA response genes [*AUX/IAA*, *Gh3*, and *SAUR* (small auxin upregulated RNA)]. We speculate that the slightly higher levels of IAA detected in PCU may have resulted in the degradation of AUX/IAA, and a correspondingly enhanced expression of *ARFs*, *AUX/IAA*, and *SAUR*, thus regulating tassel development and branching. Furthermore, given that we detect no significant difference in the expression of the IAA polar transport gene peptidylprolyl *cis*/*trans* isomerase, NIMA-interacting 1 (*PIN1*) between the two parental lines, it is reasonable to assume that the regulation of tassel branching is unrelated to the polar transport of auxin (Figure 8).

Tassel development and branching are assumed to be regulated by multiple hormones. In this regard, CK can alleviate apical dominance and promote lateral branch growth (Bangerth, 1994; Turnbull et al., 1997; Tanaka et al., 2006; Hoyerova and Hosek, 2020). However, CK activity is often regulated by auxin, which in turn promotes the growth of lateral buds by promoting the polar transport of IAA in stems and upregulating IAA synthesis in buds. Furthermore. it has been demonstrated that *ARF19* can inhibit the expression of isopentenyl transferases (IPTs) and control the synthesis of CTK (Li et al., 2006). Although in the present study,

10.3389/fpls.2023.1202755

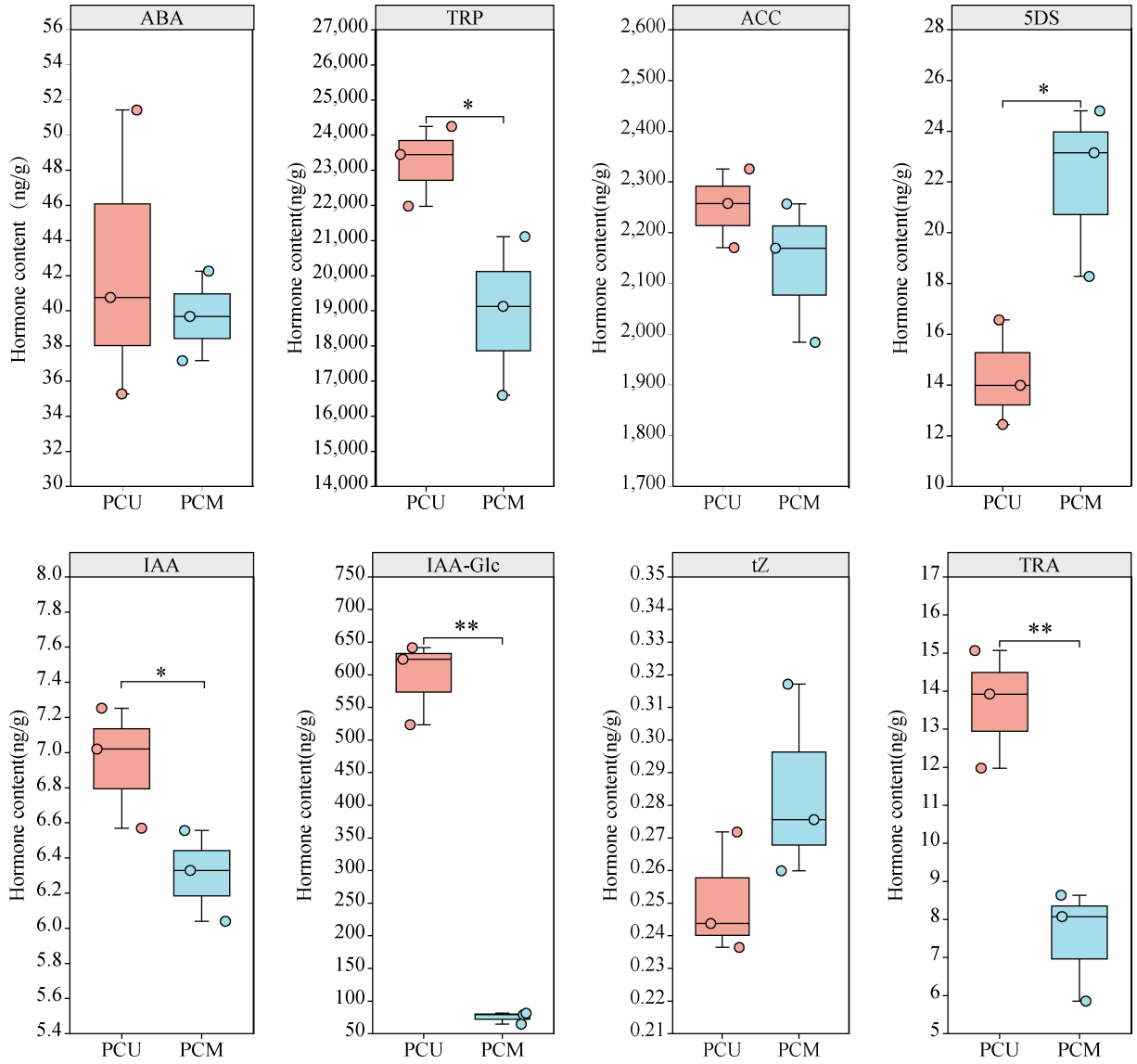


FIGURE 5
Different phytohormone contents of PCU and PCM in Stage I. The asterisks (*or **) represent the significant differences at $P < 0.05$ or $P < 0.01$, respectively.

we detected the upregulated expression of certain ARFs, we observed no significant differences in IPT gene expression or tZ content in the sister lines studied. Moreover, whereas we recorded high levels of N6-isopentenyl-adenine-9-glucoside (iP9G) content in PCM, this compound was not detected in PCU, and we accordingly speculate that iP9G could be involved in the regulation of TBN (Figure S3).

5 Conclusion

In this study, we used the sister maize lines PCU and PCM, characterized by significant differences in tassel branch number, as parents to produce an F_2 population, and applied a genetic microarray to genotype the parents and F_2 population, and to construct an associated genetic linkage map. On the basis of phenotypic and genotypic data, we identified two major QTLs, *qTBN6.06-1* and *qTBN6.06-2*, on chromosome 6. RNA-seq analysis of material collected at three stages of tassel development revealed that DEGs were enriched in amino acid metabolism and phytohormone signaling. Additionally, we established that levels of IAA, IAA-glc, TRP, and TAM were higher in PCU than in PCM, whereas in contrast, PCM was characterized by higher levels of 5DS. By combining our localization results and transcriptome data, we able to identify five candidate genes that putatively contribute to the regulation of tassel branching. Our findings in this study provide a theoretical basis that will potentially contribute to improving tassel traits in maize breeding.

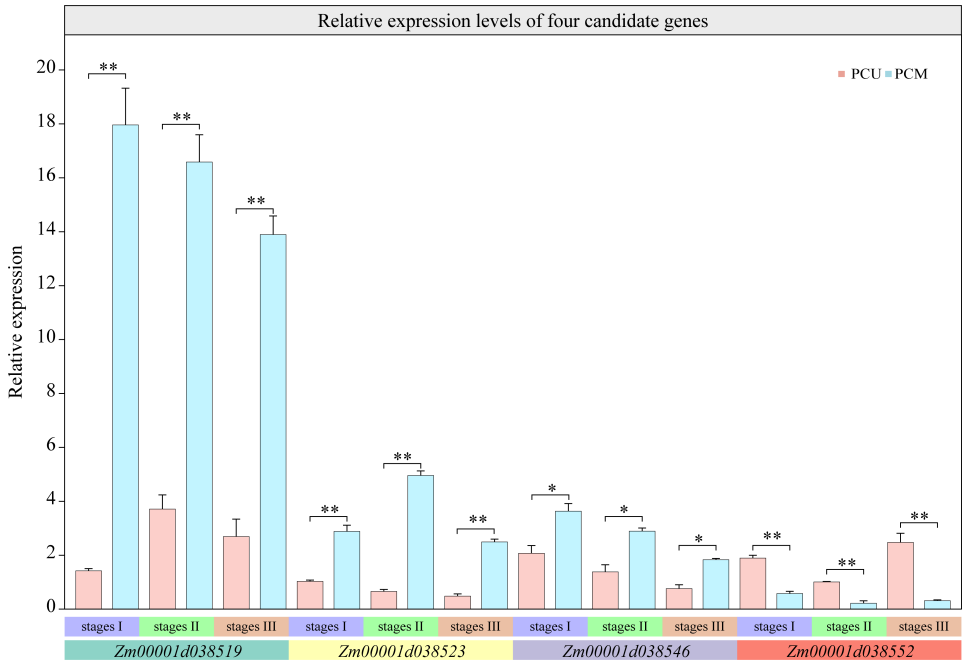


FIGURE 6
Relative expression levels of four candidate genes at three stages analyzed *via* qRT-PCR. The asterisks (*or **) represent the significant differences at $P < 0.05$ or $P < 0.01$, respectively.

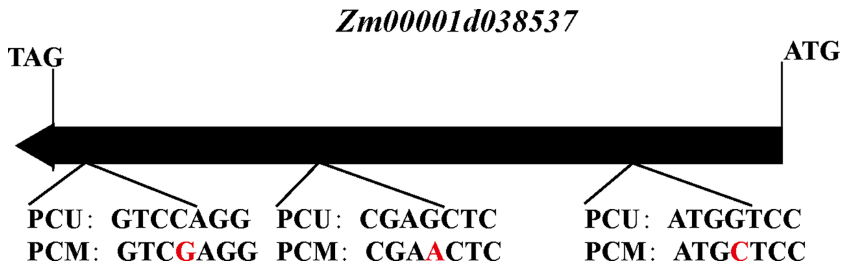


FIGURE 7
The structure of the *Zm00001d038537* between PCU and PCM. Red letters indicate SNP. The direction of the arrow represents the direction of transcription.

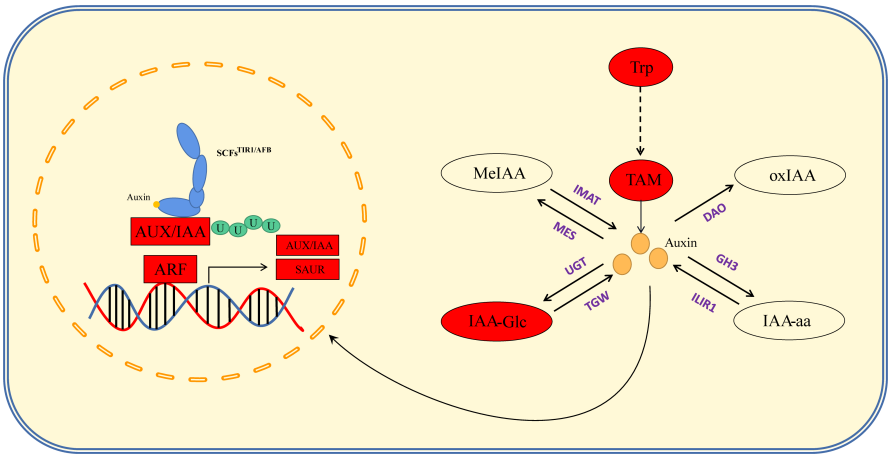


FIGURE 8
Auxin anabolism and signal regulation pathway. The red box indicates that the gene expression level of PCU is higher than that of PCM, and the red oval indicates that the hormone content of PCU is higher than that of PCM. Purple stands for key enzymes in anabolism.

Data availability statement

The data presented in the study are deposited in the SRA repository: https://www.ncbi.nlm.nih.gov/sra/PRJNA998913.

Author contributions

SR: Investigation, data curation, validation, and writing—original draft. HS: Review and editing. QY: Data curation, methodology, formal analysis, software. LYim: Formal analysis and editing. ZX: Investigation and formal analysis. LYin: Investigation. LXih: Formal analysis. LF: Writing—review and editing. DM: Investigation. LXia: Conceptualization, writing—review, editing and funding acquisition. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by Shenyang Science and Technology Plan Seed Industry Innovation Project (21-110-3-16 and 22-318-2-01-02)

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1202755/full#supplementary-material

References

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11. doi: 10.1186/gb-2010-11-10-r106

Bangerth, F. (1994). Response of cytokinin concentration in the xylem exudate of bean (Phaseolus vulgaris L.) plants to decapitation and auxin treatment, and relationship to apical dominance. *Planta* 194 (3), 439–442. doi: 10.1007/BF00197546

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30 (15), 2114–2120. doi: 10.1093/bioinformatics/btu170

Bortiri, E., Chuck, G., Vollbrecht, E., Rocheford, T., Martienssen, R., and Hake, S. (2006). ramosa2 encodes a LATERAL ORGAN BOUNDARY domain protein that determines the fate of stem cells in branch meristems of maize. *Plant Cell* 18 (3), 574–585. doi: 10.1105/tpc.105.039032

Brewbaker, J. L. (2015). Diversity and genetics of tassel branch numbers in maize. *Crop Sci.* 55 (1), 65–78. doi: 10.2135/cropsci2014.03.0248

Brown, P. J., Upadyayula, N., Mahone, G. S., Tian, F., Bradbury, P. J., Myles, S., et al. (2011). Distinct genetic architectures for male and female inflorescence traits of maize. *PloS Genet.* 7 (11). doi: 10.1371/journal.pgen.1002383

Chen, Z., Wang, B., Dong, X., Liu, H., Ren, L., Chen, J., et al. (2014). An ultra-high density bin-map for rapid QTL mapping for tassel and ear architecture in a large F2 maize population. *BMC Genomics* 15. doi: 10.1186/1471-2164-15-433

Chen, Z.-j., Yang, C., Tang, D.-g., Zhang, L., Zhang, L., Qu, J., et al. (2017). Dissection of the genetic architecture for tassel branch number by QTL analysis in two related populations in maize. *J. Integr. Agric.* 16 (7), 1432–1442. doi: 10.1016/s2095-3119(16)61538-1

Chuck, G. S., Brown, P. J., Meeley, R., and Hake, S. (2014). Maize SBP-box transcription factors unbranched2 and unbranched3 affect yield traits by regulating the rate of lateral primordia initiation. *Proc. Natl. Acad. Sci. United States America* 111 (52), 18775–18780. doi: 10.1073/pnas.1407401112

Chuck, G., Meeley, R. B., and Hake, S. (1998). The control of maize spikelet meristem fate by the APETALA2-like gene indeterminate spikelet1. *Genes Dev.* 12 (8), 1145–1154. doi: 10.1101/gad.12.8.1145

Chuck, G., Meeley, R., and Hake, S. (2008). Floral meristem initiation and meristem cell fate are regulated by the maize AP2 genes ids1 and sid1. *Development* 135 (18), 3013–3019. doi: 10.1242/dev.024273

Chuck, G., Meeley, R., Irish, E., Sakai, H., and Hake, S. (2007). The maize tasselseed4 microRNA controls sex determination and meristem cell fate by targeting Tasselseed6/indeterminate spikelet1. *Nat. Genet.* 39 (12), 1517–1521. doi: 10.1038/ng.2007.20

Claeys, H., Vi, S. L., Xu, X., Satoh-Nagasawa, N., Eveland, A. L., Goldshmidt, A., et al. (2019). Control of meristem determinacy by trehalose 6-phosphate phosphatases is uncoupled from enzymatic activity. *Nat. Plants* 5 (4), 352–357. doi: 10.1038/s41477-019-0394-z

Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21 (18), 3674–3676. doi: 10.1093/bioinformatics/bti610

Doebley, J., Stec, A., Wendel, J., and Edwards, M. (1990). Genetic and morphological analysis of a maize-teosinte F_2 population: implications for the origin of maize. *Proc. Natl. Acad. Sci. United States America* 87 (24), 9888–9892. doi: 10.1073/pnas.87.24.9888

Duncan, W. G., Williams, W. A., and Loomis, R. S. (1967). Tassels and the productivity of maize. *Crop Sci.* 7, 37–39. doi: 10.2135/cropsci1967.0011183X000700010013x

Durbak, A., Yao, H., and Mcsteen, P. (2012). Hormone signaling in plant development. *Curr. Opin. Plant Biol.* 15 (1), 92–96. doi: 10.1016/j.pbi.2011.12.004

Enders, T. A., and Strader, L. C. (2015). Auxin activity: Past, present, and future. *Am. J. Bot.* 102 (2), 180–196. doi: 10.3732/ajb.1400285

Fendrych, M., Akhmanova, M., Merrin, J., Glanc, M., Hagihara, S., and Takahashi, K. (2018). Rapid and reversible root growth inhibition by TIR1 auxin signalling. *Nat. Plants* 4 (7), 453–459. doi: 10.1038/s41477-018-0190-1

Forestan, C., Farinati, S., and Varotto, S. (2012). The maize PIN gene family of auxin transporters. *Front. Plant Sci.* 3. doi: 10.3389/fpls.2012.00016

Gallavotti, A., Yang, Y., Schmidt, R. J., and Jackson, D. (2008). The Relationship between auxin transport and maize branching. *Plant Physiol.* 147 (4), 1913–1923. doi: 10.1104/pp.108.121541

Gallavotti, A., Zhao, Q., Kyozuka, J., Meeley, R. B., Ritter, M. K., Doebley, J. F., et al. (2004). The role of barren stalk1 in the architecture of maize. *Nature* 432 (7017), 630–635. doi: 10.1038/nature03148

Gao, S. B., Zhao, M. J., Lan, H., and Zhang, Z. M. (2007). Identification of QTL associated with tassel branch number and total tassel length in maize. *Yi Chuan* 29 (8), 1013–1017. doi: 10.1360/yc-007-1013

Hartwig, T., Chuck, G. S., Fujioka, S., Klempien, A., Weizbauer, R., Potluri, D. P. V., et al. (2011). Brassinosteroid control of sex determination in maize. *Proc. Natl. Acad. Sci. U. S. A.* 108 (49), 19814–19819. doi: 10.1073/pnas.1108359108

Hoyerova, K., and Hosek, P. (2020). New insights into the metabolism and role of cytokininN-glucosides in plants. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.00741

Huang, Y., Wang, H., Zhu, Y., Huang, X., Li, S., Wu, X., et al. (2022). *THP9* enhances seed protein content and nitrogen-use efficiency in maize. *Nature* 612:292–300 doi: 10.1038/s41586-022-05441-2

Hunter, R. B., Daynard, T. B., Hume, D. J., Tanner, J. W., Curtis, J. D., and Kannenberg, L. W. (1969). Effect of tassel removal on grain yield of corn (Zea mays L.). *Crop Sci.* 9, 405–406. doi: 10.2135/cropsci1969.0011183X000900040003x

Isbell, V. R., and Morgan, P. W. (1982). Manipulation of apical dominance in sorghum with growth regulators. *Crop Sci.* 22, 30–35. doi: 10.2135/cropsci1982.0011183X002200010007x

Lambert, R. J., and Johnson, R. R. (1978). Leaf angle, tassel morphology, and the performance of maize hybrids1. *Crop Sci.* 18, 499–502. doi: 10.2135/cropsci1978.0011183X001800030037x

Li, J., Dai, X., and Zhao, Y. (2006). A role for auxin response factor 19 in auxin and ethylene signaling in Arabidopsis. *Plant Physiol.* 140 (3), 899–908. doi: 10.1104/pp.105.070987

Li, P., Li, G., Zhang, Y. W., Zuo, J. F., Liu, J. Y., and Zhang, Y. M. (2022). A combinatorial strategy to identify various types of QTLs for quantitative traits using extreme phenotype individuals in an F_2 population. *Plant Commun.* 3 (3), 100319. doi: 10.1016/j.xplc.2022.100319

Li, J., Meng, D., Yu, H., Zhang, K., Zhu, K., Lv, J., et al. (2019). Fine mapping and identification of ub4 as a candidate gene associated with tassel branch number in maize (Zea mays L.). *Genet. Resour. Crop Evol.* 66 (7), 1557–1571. doi: 10.1007/s10722-019-00805-6

Li, H., Peng, Z., Yang, X., Wang, W., Fu, J., Wang, J., et al. (2013). Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nat. Genet.* 45 (1), 43–50. doi: 10.1038/ng.2484

Liu, Y., Wu, G., Zhao, Y., Wang, H. H., Dai, Z., Xue, W., et al. (2021). DWARF53 interacts with transcription factors UB2/UB3/TSH4 to regulate maize tillering and tassel branching. *Plant Physiol.* 187 (2), 947–962. doi: 10.1093/plphys/kiab259

Mano, Y., and Nemoto, K. (2012). The pathway of auxin biosynthesis in plants. *J. Of Exp. Bot.* 63 (8), 2853–2872. doi: 10.1093/jxb/ers091

Matsuoka, Y., Vigouroux, Y., Goodman, M. M., Sanchez G, J., Buckler, E., and Doebley, J. (2002). A single domestication for maize shown by multilocus microsatellite genotyping. *Proc. Natl. Acad. Sci. United States America* 99 (9), 6080–6084. doi: 10.1073/pnas.052125199

McSteen, P. (2009). Hormonal regulation of branching in grasses. *Plant Physiol.* 149 (1), 46–55. doi: 10.1104/pp.108.129056

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

Mueller, D., and Leyser, O. (2011). Auxin, cytokinin and the control of shoot branching. *Ann. Bot.* 107 (7), 1203–1212. doi: 10.1093/aob/mcr069

Okamoto, K., Ueda, H., Shimada, T., Tamura, K., Koumoto, Y., Tasaka, M., et al. (2016). An ABC transporter B family protein, ABCB19, is required for cytoplasmic streaming and gravitropism of the inflorescence stems. *Plant Signaling Behav.* 11 (3), e1010947. doi: 10.1080/15592324.2015.1010947

Ongaro, V., and Leyser, O. (2008). Hormonal control of shoot branching. *J. Exp. Bot.* 59 (1), 67–74. doi: 10.1093/jxb/erm134

Phillips, K. A., Skirpan, A. L., Liu, X., Christensen, A., Slewinski, T. L., Hudson, C., et al. (2011). vanishing tassel2 encodes a grass-specific tryptophan aminotransferase required for vegetative and reproductive development in maize. *Plant Cell* 23 (2), 550–566. doi: 10.1105/tpc.110.075267

Qin, X., Tian, S., Zhang, W., Dong, X., Ma, C., Wang, Y., et al. (2021). Q(Dtbn1), an F-box gene affecting maize tassel branch number by a dominant model. *Plant Biotechnol. J.* 19 (6), 1183–1194. doi: 10.1111/pbi.13540

Rio, D. C., Ares, M.Jr., Hannon, G. J., and Nilsen, T. W. (2010). Purification of RNA using TRIzol (TRI reagent). *Cold Spring Harbor Protoc.* 2010 (6), pdb.prot5439. doi: 10.1101/pdb.prot5439

Satoh-Nagasawa, N., Nagasawa, N., Malcomber, S., Sakai, H., and Jackson, D. (2006). A trehalose metabolic enzyme controls inflorescence architecture in maize. *Nature* 441 (7090), 227–230. doi: 10.1038/nature04725

Skirpan, A., Culler, A. H., Gallavotti, A., Jackson, D., Cohen, J. D., and McSteen, P. (2009). BARREN INFLORESCENCE2 interaction with ZmPIN1a suggests a role in auxin transport during maize inflorescence development. *Plant Cell Physiol.* 50 (3), 652–657. doi: 10.1093/pcp/pcp006

Swarup, R., and Péret, B. (2012). AUX/LAX family of auxin influx carriers-an overview. *Front. Plant Sci.* 3. doi: 10.3389/fpls.2012.00225

Tanaka, M., Takei, K., Kojima, M., Sakakibara, H., and Mori, H. (2006). Auxin controls local cytokinin biosynthesis in the nodal stem in apical dominance. *Plant J.* 45 (6), 1028–1036. doi: 10.1111/j.1365-313X.2006.02656.x

Tian, H., Yang, Y., Yi, H., Xu, L., He, H., Fan, Y., et al. (2021). New resources for genetic studies in maize (Zea mays L.): a genome-wide Maize6H-60K single nucleotide polymorphism array and its application. *Plant J.* 105 (4), 1113–1122. doi: 10.1111/tpj.15089

Turnbull, C. G. N., Raymond, M. A. A., Dodd, I. C., and Morris, S. E. (1997). Rapid increases in cytokinin concentration in lateral buds of chickpea (Cicer arietinum L.) during release of apical dominance. *Planta* 202 (3), 271–276. doi: 10.1007/s004250050128

Umehara, M., Hanada, A., Yoshida, S., Akiyama, K., Arite, T., Takeda-Kamiya, N., et al. (2008). Inhibition of shoot branching by new terpenoid plant hormones. *Nature* 455 (7270), 195–U129. doi: 10.1038/nature07272

Vollbrecht, E., Springer, P. S., Goh, L., Buckler, E. S. T., and Martienssen, R. (2005). Architecture of floral branch systems in maize and related grasses. *Nature* 436 (7054), 1119–1126. doi: 10.1038/nature03892

Walsh, J., and Freeling, M. (1999). The liguleless2 gene of maize functions during the transition from the vegetative to the reproductive shoot apex. *Plant J. Cell Mol. Biol.* 19 (4), 489–495. doi: 10.1046/j.1365-313X.1999.00541.x

Walsh, J., Waters, C. A., and Freeling, M. (1998). The maize gene liguleless2 encodes a basic leucine zipper protein involved in the establishment of the leaf blade-sheath boundary. *Genes Dev.* 12 (2), 208–218. doi: 10.1101/gad.12.2.208

Wang, S., Basten, C., and Zeng, Z. (2012). *Windows QTL Cartographer 2.5* (Raleigh, NC: Department of Statistics, North Carolina State University). Available at: http://statgen.ncsu.edu/qtlcart/WQTLCart.htm.

Wang, Y., Chen, J., Guan, Z., Zhang, X., Zhang, Y., Ma, L., et al. (2019). Combination of multi-locus genome-wide association study and QTL mapping reveals genetic basis of tassel architecture in maize. *Mol. Genet. And Genomics* 294 (6), 1421–1440. doi: 10.1007/s00438-019-01586-4

Wang, B., Liu, H., Liu, Z., Dong, X., Guo, J., Li, W., et al. (2018). Identification of minor effect QTLs for plant architecture related traits using super high density genotyping and large recombinant inbred population in maize (Zea mays). *BMC Plant Biol.* 18 (1), 17. doi: 10.1186/s12870-018-1233-5

Wei, H., Zhao, Y., Xie, Y., and Wang, H. (2018). Exploiting SPL genes to improve maize plant architecture tailored for high-density planting. *J. Exp. Bot.* 69 (20), 4675–4688. doi: 10.1093/jxb/ery258

Wen, Y. J., Zhang, Y. W., Zhang, J., Feng, J. Y., Dunwell, J. M., and Zhang, Y. M. (2019). An efficient multi-locus mixed model framework for the detection of small and linked QTLs in F_2. *Brief Bioinform.* 20 (5), 1913–1924. doi: 10.1093/bib/bby058

Wu, X., Li, Y., Shi, Y., Song, Y., Zhang, D., Li., C., et al. (2016). Joint-linkage mapping and GWAS reveal extensive genetic loci that regulate male inflorescence size in maize. *Plant Biotechnol. J.* 14 (7), 1551–1562. doi: 10.1111/pbi.12519

Xu, G., Wang, X., Huang, C., Xu, D., Li, D., Tian, J., et al. (2017). Complex genetic architecture underlies maize tassel domestication. *New Phytol.* 214 (2), 852–864. doi: 10.1111/nph.14400

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. doi: 10.1371/journal.pgen.1004573

Ye, J., Zhang, Y., Cui, H., Liu, J., Wu, Y., Cheng, Y., et al. (2018). WEGO 2.0: a web tool for analyzing and plotting GO annotations 2018 update. *Nucleic Acids Res.* 46, W71–W75. doi: 10.1093/nar/gky400

Yi, Q., Liu, Y., Zhang, X., Hou, X., Zhang, J., Liu, H., et al. (2018). Comparative mapping of quantitative trait loci for tassel-related traits of maize in $F_{2:3}$ and RIL populations. *J. Genet.* 97 (1), 253–266. doi: 10.1007/s12041-018-0908-x

Zhao, Y. (2012). Auxin biosynthesis: A simple two-step pathway converts tryptophan to indole-3-acetic acid in plants. *Mol. Plant* 5 (2), 334–338. doi: 10.1093/mp/ssr104

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 30 August 2023
DOI 10.3389/fpls.2023.1250854

OPEN ACCESS

EDITED BY
Yuan-Ming Zhang,
Huazhong Agricultural University, China

REVIEWED BY
Xiangqian Zhao,
Zhejiang Agriculture and Forestry University, China
Yang-Jun Wen,
Nanjing Agricultural University, China
Li Mei,
Huazhong Agricultural University, China

*CORRESPONDENCE
Weiren Wu
wuwr@fafu.edu.cn

†These authors have contributed equally to this work

RECEIVED 30 June 2023
ACCEPTED 14 August 2023
PUBLISHED 30 August 2023

CITATION
Zheng Y, Thi KM, Lin L, Xie X, Khine EE, Nyein EE, Lin MHW, New WW, Aye SS and Wu W (2023) Genome-wide association study of cooking-caused grain expansion in rice (*Oryza sativa* L.).
Front. Plant Sci. 14:1250854.
doi: 10.3389/fpls.2023.1250854

Genome-wide association study of cooking-caused grain expansion in rice (*Oryza sativa* L.)

Yan Zheng[1,2,3†], Khin Mar Thi[2,3†], Lihui Lin[2,3], Xiaofang Xie[1,2,3], Ei Ei Khine[2,3], Ei Ei Nyein[2,3], Min Htay Wai Lin[4], Win Win New[4], San San Aye[4] and Weiren Wu[2,3*]

[1]College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou, Fujian, China, [2]Fujian Provincial Key Laboratory of Crop Breeding by Design, Fujian Agriculture and Forestry University, Fuzhou, Fujian, China, [3]Key Laboratory of Genetics, Breeding and Multiple Utilization of Crops, Ministry of Education, Fujian Agriculture and Forestry University, Fuzhou, Fujian, China, [4]Department of Botany, Mawlamyine University, Mawlamyine, Myanmar

Cooking-caused rice grain expansion (CCRGE) is a critical trait for evaluating the cooking quality of rice. Previous quantitative trait locus (QTL) mapping studies on CCRGE have been limited to bi-parental populations, which restrict the exploration of natural variation and mapping resolution. To comprehensively and precisely dissect the genetic basis of CCRGE, we performed a genome-wide association study (GWAS) on three related indices: grain breadth expansion index (GBEI), grain length expansion index (GLEI), and grain length-breadth ratio expansion index (GREI), using 345 rice accessions grown in two years (environments) and 193,582 SNP markers. By analyzing each environment separately using seven different methods (3VmrMLM, mrMLM, FASTmrMLM, FASTmrEMMA, pLARmEB, pKWmEB, ISIS EM-BLASSO), we identified a total of 32, 19 and 27 reliable quantitative trait nucleotides (QTNs) associated with GBEI, GLEI and GREI, respectively. Furthermore, by jointly analyzing the two environments using 3VmrMLM, we discovered 19, 22 and 25 QTNs, as well as 9, 5 and 7 QTN-by-environment interaction (QEIs) associated with GBEI, GLEI and GREI, respectively. Notably, 12, 9 and 15 QTNs for GBEI, GLEI and GREI were found within the intervals of previously reported QTLs. In the vicinity of these QTNs or QEIs, based on analyses of mutation type, gene ontology classification, haplotype, and expression pattern, we identified five candidate genes that are related to starch synthesis and endosperm development. The five candidate genes, namely, *LOC_Os04g53310* (*OsSSIIIb*, near QTN *qGREI-4.5s*), *LOC_Os05g02070* (*OsMT2b*, near QTN *qGLEI-5.1s*), *LOC_Os06g04200* (*wx*, near QEI *qGBEI-6.1i* and QTNs *qGREI-6.1s* and *qGLEI-6.1t*), *LOC_Os06g12450* (*OsSSIIa*, near QTN *qGLEI-6.2t*), and *LOC_Os08g09230* (*OsSSIIIa*, near QTN *qGBEI-8.1t*), are predicted to be involved in the process of rice grain starch synthesis and to influence grain expansion after cooking. Our findings provide valuable insights and will facilitate genetic research and improvement of CCRGE.

KEYWORDS

rice, grain breadth expansion index (GBEI), grain length expansion index (GLEI), grain length-breadth relative expansion index (GREI), Genome-wide association study (GWAS)

1 Introduction

Rice (*Oryza sativa* L.) is a crucial cereal crop that serves as a staple food for over half of the global population. It is the only cereal crop that is primarily consumed as whole grains, which underscores its significance in the field of rice breeding (Hossain et al., 2009). The quality of rice is assessed based on several factors, including appearance, milling, cooking, sensory properties, and nutrition (Cheng et al., 2005; Feng et al., 2017). Among these factors, cooking quality is a critical determinant for the economic value of rice. The cooking quality of rice refers to the characteristics of cooked rice, including its texture, tenderness, stickiness, and overall palatability. As starch accounts for up to 95% of the dry weight of a polished rice grain (Fitzgerald et al., 2009), the cooking quality of rice is mainly determined by starch. During the cooking process, rice grains absorb water and undergo gelatinization, leading to a noticeable expansion in volume (Golam and Prodhan, 2013). The extent of this cooking-caused rice grain expansion (CCRGE) can affect the texture, tenderness and overall quality of cooked rice, and is significantly influenced by the properties of starch (Pang et al., 2016). In general, rice varieties with a higher amylose content (AC) tend to absorb more water and exhibit greater increase in volume after cooking (Frei et al., 2003). Hence, CCRGE is a complex trait closely related to the cooking quality of rice. As the desired cooking quality can vary depending on the type of rice and the culinary preferences of individuals or cultural cuisines (Suwannaporn and Linnemann, 2008), the corresponding suitable degree of CCRGE is also diverse. To meet the varying demands for the cooking quality of rice, different goals should be established in rice breeding. Dissecting the genetic basis of CCRGE will facilitate the efforts toward the goals.

For this purpose, a number of studies have been conducted to map quantitative trait loci (QTLs) underlying CCRGE. To date, 47 QTLs for grain length expansion (Ahn et al., 1993; Li et al., 2004; Zhang et al., 2004; Ge et al., 2005; Shen et al., 2005; Tian et al., 2005; Wang et al., 2007; Amarawathi et al., 2008; Liu et al., 2008; Govindaraj et al., 2009; Shen et al., 2011; Swamy et al., 2012; Li et al., 2015; Arikit et al., 2019), 10 QTLs for grain breadth expansion (Ge et al., 2005; Govindaraj et al., 2009), and 15 QTLs for grain length-breadth relative expansion (He et al., 2003; Jiang et al., 2008; Liu et al., 2008; Thi et al., 2020; Malik et al., 2022) have been reported, demonstrating that CCRGE is a very complex trait. However, none of these QTLs have been cloned.

All the QTLs reported for CCRGE were identified through conventional linkage analysis methods utilizing various populations derived from bi-parental crosses, including F_2 (Arikit et al., 2019), F_3 (Ahn et al., 1993), $F_{2:3}$ (Jiang et al., 2008; Thi et al., 2020), BC_2F_2 (Swamy et al., 2012), BC_3F_1 (Li et al., 2004), doubled haploid (DH) (Zhang et al., 2004; Tian et al., 2005; Govindaraj et al., 2009), and recombinant inbred lines (RILs) (He et al., 2003; Malik et al., 2022). The linkage-based QTL mapping methods are limited by two main factors. First, it can only investigate the variation between two parents. Second, it has a low mapping resolution due to strong linkage disequilibrium in the mapping population used. Consequently, the mapped QTLs can only account for a small portion of the related genetic variations in the rice germplasm. Therefore, further studies are necessary.

During the domestication process, rice germplasm resources have accumulated a rich array of natural variations in the genome. The advent of high-throughput DNA sequencing technologies has facilitated the use of genome-wide association study (GWAS) as an effective method for identifying natural genomic variations associated with quantitative traits (Huang et al., 2010; Zhao et al., 2011). Unlike the linkage-based QTL mapping method, GWAS utilizes high-density single nucleotide polymorphisms (SNPs) as genetic markers and is performed on diverse natural populations. As linkage disequilibrium is much weaker in natural populations, GWAS achieves higher resolution in QTL mapping (Huang and Han, 2014; Burghardt et al., 2017). GWAS has been successfully employed to map genes or QTLs for numerous important traits in rice, such as flowering time (Huang et al., 2012), grain yield components (Eizenga et al., 2019), grain qualities (Misra et al., 2017; Wang et al., 2020), and so on. However, to date, no GWAS has been conducted to identify QTLs underlying CCRGE.

In this study, we performed GWAS on three traits of CCRGE based on two replicated experiments conducted in two different years (environments) and using seven different methods to analyze the data. We detected 165 related quantitative trait nucleotides (QTNs), including some exhibiting only the effect of QTN-by-environment interaction (QEI). Based on the detected QTNs, we identified five candidate genes through gene ontology (GO), haplotype, and expression pattern analyses. Our findings will facilitate further genetic research and the genetic improvement of CCRGE.

2 Materials and methods

2.1 Plant materials and field experiments

A set of 345 rice accessions among the list of the 3K Rice Genomes Project (2014) were utilized for this research (Table S1). These accessions included 108 japonica, 177 indica, 48 circum-Aus group (cA), 2 circum-Basmati group (cB), and 10 admixed (between major groups) according to Wang et al. (2018). All accessions were grown at the Experimental Farm of Fujian Agriculture and Forestry University in Yangzhong (E118.485841, N26.287161) during the normal growing season (April to October) in 2017 (E1) and 2018 (E2). In both years, 20 seeds of each accession were sown on a seedbed after pregermination, and 14 seedlings were transplanted onto the paddy field 25 days later with a 20-cm spacing between plants and between rows. Field management followed standard agronomic procedures. Mature seeds were harvested from each accession, and subjected to sun, then stored at the room temperature. The newly harvested seeds were utilized for the measurement of CCRGE traits in each year.

2.2 Measure of cooking-caused grain expansion

The procedure for quantifying the characteristics of cooking-caused rice grain expansion was performed according to Thi et al. (2020). The experiment was conducted in three replicates for each

accession. In each replicate, 30 intact white rice grains were soaked (for 30 min) and boiled (for 45 min), and the average length and average breadth of 30 uncooked grains (L_0 and B_0) and 15 unbroken and straight cooked grains (L_1 and B_1) were measured. Subsequently, the grain breadth expansion index (GBEI), grain length expansion index (GLEI) and grain length-breadth relative expansion index (GREI) of each accession were calculated according to the formulae described by Thi et al. (2020), where GLEI = L_1/L_0, GBEI = B_1/B_0, and GREI = $(L_1/B_1)/(L_0/B_0)$ = $(L_1/L_0)/(B_1/B_0)$ = GLEI/GBEI.

2.3 Collection of SNP data

The SNP data of the 345 rice accessions were obtained from the 3K Rice Genomes Project (http://iric.irri.org/resources/3000-genomes-project). The core genome set of 404K SNPs (https://snp-seek.irri.org/download.zul, accessed on 1 September 2021) was employed for the analysis. A stringent quality control process was performed, which involved removal of the SNPs that had more than 20% missing calls and a minor allele frequency (MAF) smaller than 5%. As a result, a total of 193,582 SNPs were retained for subsequent analysis.

2.4 Clustering, population structure and linkage disequilibrium analyses

The genetic distances between 345 accessions were calculated based on SNP data, and a phylogenetic tree was constructed using the MEGA 11 software. Population structure was analyzed using principal component analysis (PCA) plots and the Admixture program as described by Alexander and Lange (2011). The linkage disequilibrium (LD) between pairwise SNPs located within 1 megabase (Mb) on each chromosome or across the entire genome was estimated by computing the determination coefficient (R^2) using the plink software (Purcell et al., 2007).

2.5 Genome-wide association studies

GWAS was performed on GLEI, GBEI and GREI with two strategies: (1) single-environment analysis, namely, analyzing each environment separately; and (2) two-environment analysis, namely, analyzing the two environments jointly. For single-environment analysis, we employed two R packages: 3VmrMLM (Li et al., 2022; https://github.com/YuanmingZhang65/IIIVmrMLM) and mrMLM v4.0.2 (Zhang et al., 2020). The former includes the method 3VmrMLM, while the latter contains six methods, namely, mrMLM (Wang et al., 2016), FASTmrMLM (Tamba and Zhang, 2018), FASTmrEMMA (Wen et al., 2018), pLARmEB (Zhang et al., 2017), pKWmEB (Ren et al., 2018), and ISIS EM-BLASSO (Tamba et al., 2017). The option "method=Single_env" was chosen in 3VmrMLM, while default parameters were used for the other methods. Two-environment analysis was conducted using 3VmrMLM only, with the option set to "method=Multi_env". This method allowed for the estimation of the main effect of a QTN and the effect of QTN-by-environment interaction. For distinction, a QTN showing only the effect of QTN-by-environment interaction was denoted as QEI. Each QTN or QEI was named following the nomenclature "q + trait + chromosome + number + s/t/i", where "s" and "t" indicate that the QTN was detected based on single- or two-environment analysis, respectively, and "i" indicates a QEI. According to Zhang et al. (2019), the QTNs identified by multiple methods were deemed as reliable QTNs, with particular emphasis on those identified in multiple environments, which were considered stable QTNs.

2.6 Prediction of candidate genes

Based on the distinct LD decay in each rice chromosome, the left and right R^2 half-decay regions flanking each QTN or QEI were determined to identify potential candidate genes. The following sequential steps were executed: (1) the SNP effect prediction software snpEff.v1.9 (Cingolani et al., 2012) was employed to evaluate the effects of SNPs on the regional genes, and annotated genes with effective mutation types, such as non-synonymous substitution, splice site, and UTR-5' mutation, were selected; (2) GO classifications related to starch synthesis or endosperm development were searched in the rice database (https://www.ricedata.cn/ontology/), and all genes with these GO classifications were retrieved; and (3) genes that meet both steps 1 and 2 were screened out and then subjected to haplotype analysis, where different haplotypes exhibiting *t*-test significance were considered as candidate genes.

2.7 Tissue specific expression of candidate genes

The expression profiles of the candidate genes in various tissues were obtained from the Rice Genome Annotation Project database (http://rice.uga.edu), including shoots (library name in NCBI: SRR042529), leaves-20 days (OSN_AA and OSN_CA), pre-emergence inflorescence (OSN_AC), post-emergence inflorescence (OSN_AB), anther (OSN_AD), pistil (OSN_AE), seed-5 DAP (days after pollination; OSN_AF), seed-10 DAP (OSN_AK), embryo-25 DAP (OSN_AG) and endosperm-25 DAP (OSN_AH and OSN_BH). A heatmap was generated to visualize the gene expression patterns across the different tissues.

3 Results

3.1 Trait performance

The traits GBEI, GLEI, and GREI exhibited a continuous unimodal distribution in both environments, suggesting that these traits are quantitative and controlled by multiple genes (Figure 1). After performing the Brown-Forsythe Test for assessing homogeneity of variances, the analysis revealed that the error variances of each accession in both environment for the three traits were

10.3389/fpls.2023.1250854

homogeneous, indicating that the collected data is suitable for subsequent analysis of variance (ANOVA). Although the population means of these traits were similar in both environments (GBEI: 1.822 and 1.765; GLEI: 1.752 and 1.740; GREI: 0.990 and 1.016), ANOVA revealed statistically significant variation between the two environments and genotype-by-environment interaction (Table 1). These results indicated that all the three traits exhibited significant variation across macro-environments. However, there were still significant correlations between the two environments in these traits, particularly in GLEI and GREI (Table 2).

GREI exhibited significant positive and negative correlations with GLEI and GBEI, respectively (Table 2). This is understandable, as GREI is a composite trait that is influenced by both GLEI and GBEI. However, the correlation between GLEI and GBEI was found to be low (-0.155 in E1 and -0.101 in E2) (Table 2), implying that grain length expansion and breadth expansion during cooking are two relatively independent processes with potentially distinct genetic bases.

3.2 Population structures and linkage disequilibrium

A set of 193,582 SNPs meeting the requirements of MAF > 5% and missing data < 20% were obtained. The SNPs were not evenly distributed in the genome (Figure 2). SNPs were the densest on chromosome 11 but the sparsest on chromosome 3, respectively (Table 3). On average, there was one SNP every 1928 bp in the genome.

The results of phylogenetic analysis (Figure 3A), PCA (Figure 3B), and admixture analysis (Figures 3C, D) all indicated that the population of the 345 rice accessions could be basically divided into three distinct groups (subpopulations), namely, *indica* group, *japonica* group, and *aus* group (Figures 3C, D).

The average LD (mean R^2) decreased with the increase of physical distance on every chromosome as well as in the whole genome (Figure 4). The average LD half-decay distance (HDD) and the average distance of LD decay to 0.1 (DD0.1) in the whole genome were about 378 kb and 196 kb, respectively (Table 3). However, the HDD and DD0.1 on different chromosomes varied greatly, ranging from 158.4 kb and 62.1 kb on chromosome 2 to 715.7 kb and 712.1 kb on chromosome 7, respectively (Table 3). Therefore, chromosome 2 had the highest LD decay rate, while chromosome 7 had the lowest.

3.3 QTNs detected by single-environment analysis

In total, 386 QTNs were detected by single-environment analysis using seven different methods, with 145, 127 and 128

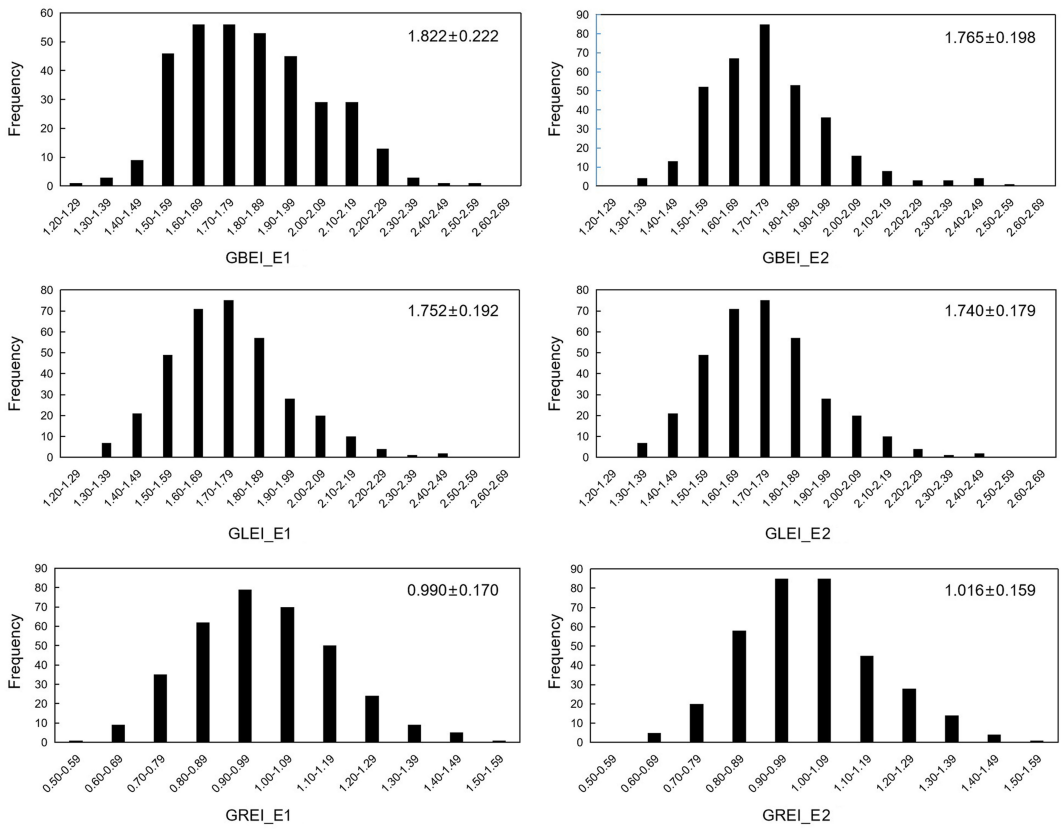


FIGURE 1
Frequency distribution of GBEI, GLEI and GREI in two environments. Values on the top right corner of each diagram are mean ± standard deviation (cm).

TABLE 1 ANOVA of GBEI, GLEI and GREI on genotypes and environments, and their interactions.

	GBEI		GLEI		GREI	
	F value	P value	F value	P value	F value	P value
Genotype (G)	11.279	2.5E-239	15.489	1.01E-306	48.364	0
Environment (E)	74.890	1.36E-17	1.843	0.1748481	130.473	6.15E-29
G×E	6.437	1.4E-138	7.070	1.35E-153	14.398	1.3E-290
Test of HOV	0.780	1.000	0.811	0.999	0.674	1.000

Test of HOV (homogeneity of variance) was performed using the method of Brown-Forsythe Test, in which $F_{0.05}$ = 1.1134 (df_1 = 689, df_2 = 1380).

TABLE 2 Coefficients of correlation between different traits in each environment and between different environments in each trait.

	GBEI	GLEI	GREI
GBEI	0.317**	-0.101	-0.677**
GLEI	−0.155**	0.487**	0.750**
GREI	−0.767**	0.736**	0.542**

The data in the diagonal are correlations between the two years. The data in the lower triangle and the upper triangle are correlations between the three traits in E1 (2017) and in E2 (2018), respectively. ** indicates p-value < 0.01.

QTNs found to be associated with GBEI, GLEI and GREI, respectively (Table 4; Figures S1, S2). However, only 78 (19.5%) QTNs were identified as reliable (Tables 4, S2). The total number of QTNs detected by each method varied greatly, ranging from 32 (FASTmrEMMA) to 131 (3VmrMLM; Table 4). The number and the percentage of reliable QTNs detected by each method also differed significantly (Table 4). Interestingly, there was a positive correlation between the number of reliable QTNs and the total number of QTNs detected by each method (correlation coefficient 80.5%), but a negative correlation between the percentage of reliable QTNs and the total number of QTNs detected by each method (correlation coefficient -88.2%). This indicates that the increase in the number of total QTNs and reliable QTNs detected by a method comes at the cost of a decrease in the percentage of reliable QTNs.

Among the three traits, GBEI had the most reliable QTNs, followed by GREI, and GLEI had the fewest (Table 5). Consistently, GBEI had highest proportion of phenotypic variance explained (PVE) by the reliable QTNs, followed by GREI, and GLEI had the lowest (Table 5). More reliable QTNs were detected and therefore there were higher PVEs in E1 than in E2 for GLEI and GREI, but the results in the two environments were similar for GBEI (Table 5).

Most QTNs identified in this study were found to be reliable because they were detected by multiple methods, while only four QTNs (*qGBEI-5.4s*, *qGLEI-3.3s*, *qGREI-5.2s* and *qGREI-5.6s*) were found to be stable because they were detected in the two environments simultaneously (Table S2). In addition, there were a few SNPs exhibiting pleiotropic effects in one environment, including 3:16774870 (detected as QTNs *qGLEI-3.5s* and *qGREI-3.6s*) and 6:25062099 (*qGLEI-6.4s* and *qGREI-6.4s*), both of which were associated with GLEI and GREI; and 5:5369111 (*qGBEI-5.3s* and *qGREI-5.2s*), which was associated with GBEI and GREI (Table S2).

3.4 QTNs detected by two-environment analysis

The two-environment analysis detected 11, 14 and 19 significant QTNs (P-value ≤ 0.05/m = 2.58E-07, where m = 193,582, the number of markers) and 8, 8 and 6 suggested QTNs (P > 2.58E-07 but LOD ≥ 3.0) associated with GBEI, GLEI and GREI, respectively (Figures 5A-C; Table S3). These QTNs explained 35.41%, 46.37% and 41.49% of the total phenotypic variation in GBEI, GLEI and GREI, respectively. The SNP marker chr5:5369111 was found to be associated with both GBEI and GREI, and was named *qGBEI-5.2t* and *qGREI-5.3t*, respectively. This marker was also detected as QTNs *qGBEI-5.3s* and *qGREI-5.2s* in the single-environment analysis, indicating its reliability. Marker chr6:25000609 was associated with both GLEI and GREI, while chr11:23854971 was associated with both GBEI and GREI. Additionally, SNPs chr2:24264276, chr3:2521638, chr3:35669404 and chr5:14585838 were all detected in both single- and two-environment analyses.

The two-environment analysis also detected 6, 4 and 5 significant QEIs and 3, 1 and 2 suggested QEIs associated with GBEI, GLEI and GREI, respectively. These QEIs accounted for 24.83%, 14.79% and 21.22% of the total phenotypic variation in GBEI, GLEI and GREI, respectively (Figures 5D-F; Table S4). Notably, there was no common site between the QTNs and QEIs detected, indicating that all the SNPs exhibiting significant main (additive and/or dominance) effects in the two-environment analysis did not show significant effects of interaction with the environment, and vice versa (namely, all the SNPs exhibiting significant effects of interaction with the environment did not show significant main effects). Nonetheless, the SNP markers of two QEIs, *qGREI-2.3i* (SNP 2:19642336) and *qGLEI-5.6i* (SNP 5:25726382) were also detected as QTN *qGREI-2.2s* and *qGREI-5.8s* in the single-environment analysis, respectively

10.3389/fpls.2023.1250854

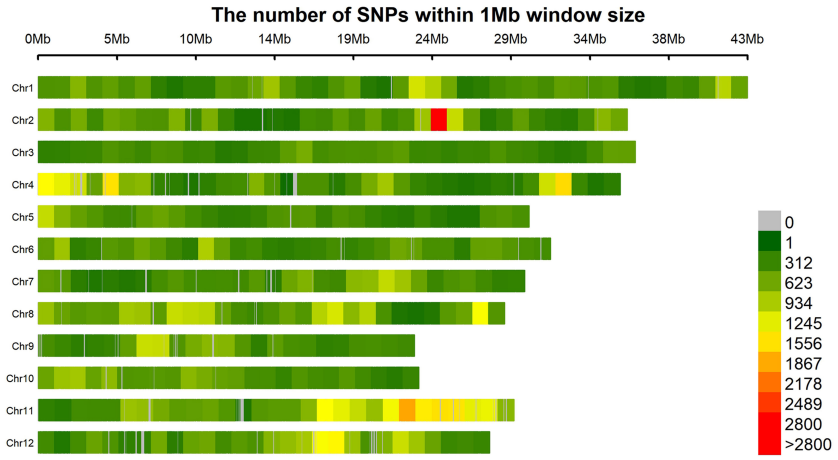


FIGURE 2
Distribution of 193,582 SNPs in the rice genome.

(Tables S2, S4). Interestingly, the targeted traits of *qGLEI-5.6i* and *qGREI-5.8s* were not the same. In addition, the interaction between SNP marker 8:22185608 and environment was found to be associated with both GLEI (as *qGLEI-8.3i*) and GREI (as *qGREI-8.5i*) simultaneously (Table S4).

3.5 Prediction of candidate genes for GBEI, GLEI and GREI

In total, the two-environment analysis detected 66 QTNs and 21 QEIs for the three traits. Plus the 78 reliable QTNs detected in the single-environment analysis, this study detected a total of 165 QTNs/QEIs. These QTNs/QEIs were mainly located on chromosomes 5, 11, 12, 3 and 2, and very rare on chromosomes 1 and 10 (Figure 6).

TABLE 3 Number and density of SNPs and LD decay distances in the rice genome.

Chromosome	Number of SNPs	Average spacing (bp)	HDD (kb)	DD0.1 (kb)
1	20,083	2154.6	651.1	603.9
2	18,756	1916.0	158.4	62.1
3	13,674	2663.0	534.7	507.3
4	19,298	1839.7	223.1	94.6
5	12,058	2484.5	333.3	374.5
6	13,883	2250.9	419.5	420.3
7	13,389	2218.1	715.7	712.1
8	18,850	1508.9	513.7	295.8
9	10,978	2096.3	330.5	249.7
10	11,946	1942.7	688.1	388.2
11	24,068	1205.8	178.9	78.5
12	16,599	1658.6	485.2	83.9
Whole genome	193,582	1928.1	377.9	196.1

HDD, LD half-decay distance; DD0.1, distance of LD decay to 0.1.

Considering that CCRGE may be largely determined by the starch in endosperm, we tried to predict the candidate genes involved in starch metabolism and endosperm development. By searching 20 related Gene Ontology/Term Ontology (GO/TO) classifications on the China Rice Data Center's website (https://www.ricedata.cn/ontology/), 119 genes were found, of which 26 were located within the R^2 half-decay distance around the detected QTNs/QEIs (Table S5). By analyzing the SNP variations in the genes with the software snpEff v1.9, five genes were found to carry effective mutations, including non-synonymous, splice site and UTR-5' mutations (Table 6; Figure S3). So, these genes were considered to be candidate genes.

We then performed haplotype analysis to assess the reliability of the candidate genes. *LOC_Os04g53310* (*OsSSIIIb*), *LOC_Os06g04200* (*wx*) and *LOC_Os08g09230* (*OsSSIIIa*) exhibited significant haplotype differences for GBEI; *LOC_Os04g53310*, *LOC_Os05g02070* (*OsMT2b*) and *LOC_Os06g12450* (*OsSSIIa*) displayed significant haplotype differences for GLEI; and all of the genes except for *LOC_Os06g12450* showed significant haplotype differences for GREI (Figure 7). These findings strongly suggested a close association of these five genes with the CCRGE.

To further verify the potential impact of these candidate genes on the regulation of starch synthesis and endosperm development, we analyzed the expression patterns of the five candidate genes in various tissues based on data from the Rice Genome Annotation Project database (Figure 8). The results showed that *LOC_Os04g53310* (*OsSSIIIb*) was expressed mainly in leaf and pre-emergence inflorescence but not in seed or endosperm; *LOC_Os05g02070* (*OsMT2b*) was expressed mainly in post- and pre-emergence inflorescence and in embryo of 25 DAP (days after pollination), but not in endosperm. This suggests that these two genes maybe not closely or indirectly associated with endosperm

10.3389/fpls.2023.1250854

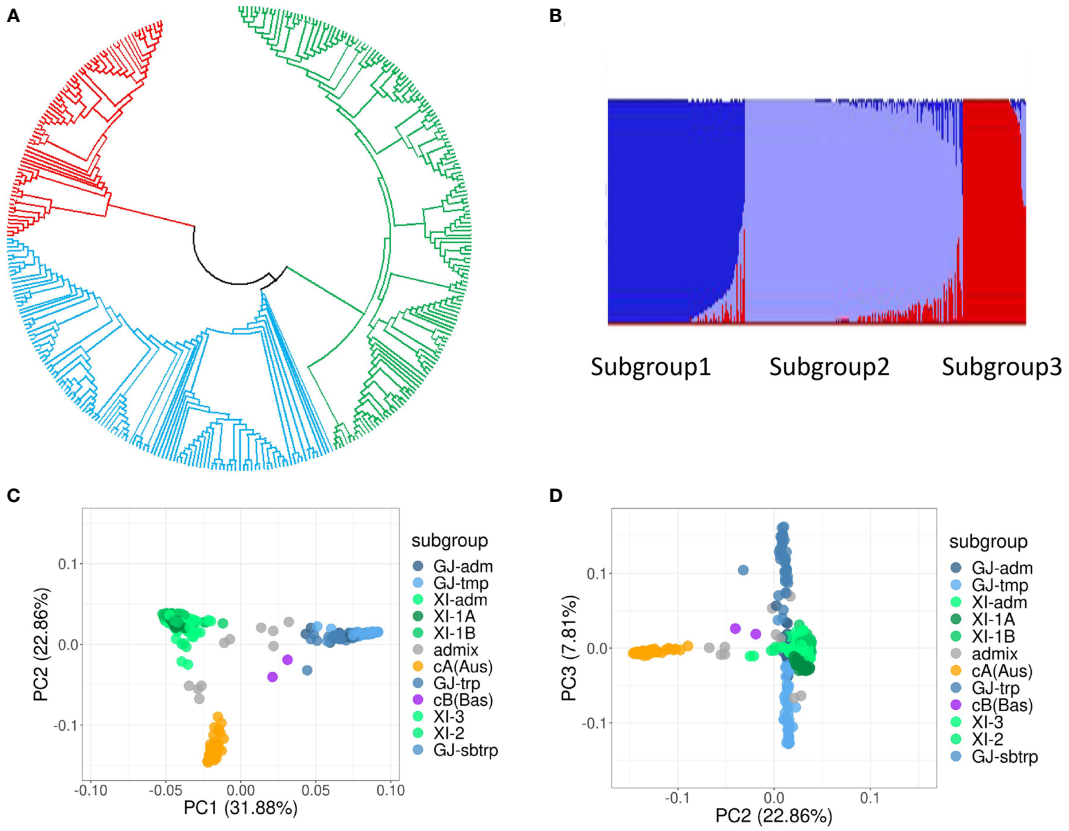


FIGURE 3
Genetic structure analysis of the population of 345 rice accessions. **(A)** Phylogenetic tree. **(B)** Population structure estimated by the software Admixture. **(C, D)** PCA plots of the first three principal components.

development. In contrast, *LOC_Os06g04200* (*wx*), *LOC_Os06g12450* (*OsSSIIa*) and *LOC_Os08g09230* (*OsSSIIIa*) exhibited high expression in 10 DAP seed, and the highest expression in 25 DAP endosperm, but no expression in embryo, indicating their potential involvement in starch synthesis or endosperm development.

4 Discussion

When analyzing single environmental data, only QTNs, *qGREI-5.2s* and *qGREI-5.6s*, were commonly detected in two environments. This may be due to changes in the relative effects of different genes for these traits in different environments,

TABLE 4 Numbers of QTNs for GBEI, GLEI and GREI detected by seven methods in two different environments.

Method	GBEI			GLEI			GREI			Total[1]	Reliable QTNs
	E1	E2	Total[1]	E1	E2	Total[1]	E1	E2	Total[1]		
3VmrMLM	21	19	40	20	24	44	27	21	47	131	31 (23.5%)
mrMLM	10	9	19	7	6	13	16	7	23	55	25 (45.5%)
FASTmrMLM	13	49	62	11	19	29	12	10	22	114	40 (35.1%)
FASTmrEMMA	4	7	11	6	5	11	3	7	10	32	21 (65.6%)
pLARmEB	11	9	20	16	18	34	19	22	41	95	40 (42.1%)
pKWmEB	11	7	18	9	10	19	12	8	20	57	23 (40.4%)
ISIS EM-BLASSO	13	14	27	4	2	6	5	4	9	42	21 (50.0%)
Total[1]	56	90	145	55	73	127	70	60	128	400	78 (19.5%)

1. Redundancy was removed in the totals. 2. The number and proportion of reliable QTNs among the total detected by each method or in the whole experiment.

10.3389/fpls.2023.1250854

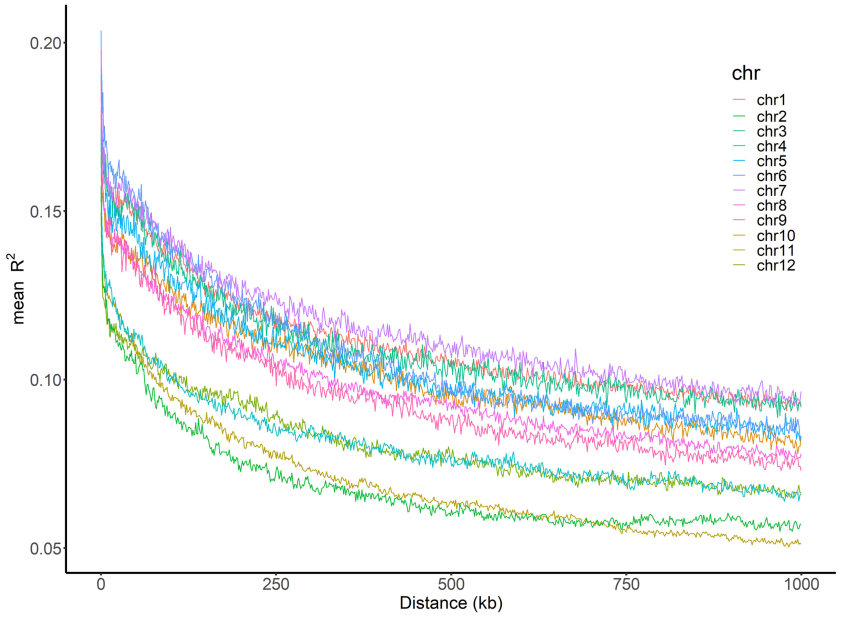


FIGURE 4
Genome-wide LD decay across 12 chromosomes. The x-axis represents the physical distance and the y-axis represents the average pairwise R^2. The color of each chromosome was showed on the top right corner.

TABLE 5 Statistics of reliable QTNs for GBEI, GLEI and GREI detected in each environment.

Trait	No. of reliable QTNs			LOD range		PVE range (%)		Total PVE (%)		
	E1	E2	Total	E1	E2	E1	E2	E1	E2	Average
GBEI	16	17	32	3.9-13.7	3.5-12.2	2.5-7.4	1.7-5.3	63.8	61.9	62.85
GLEI	13	6	19	3.2-14.3	3.9-8.5	0.2-7.2	0.1-4.0	41.5	12.8	27.15
GREI	16	12	27	3.6-14.9	3.9-12.9	1.3-7.7	0.0-9.0	47.8	36.8	42.30
Total	45	35	78							

PVE, proportion of phenotypic variance explained.

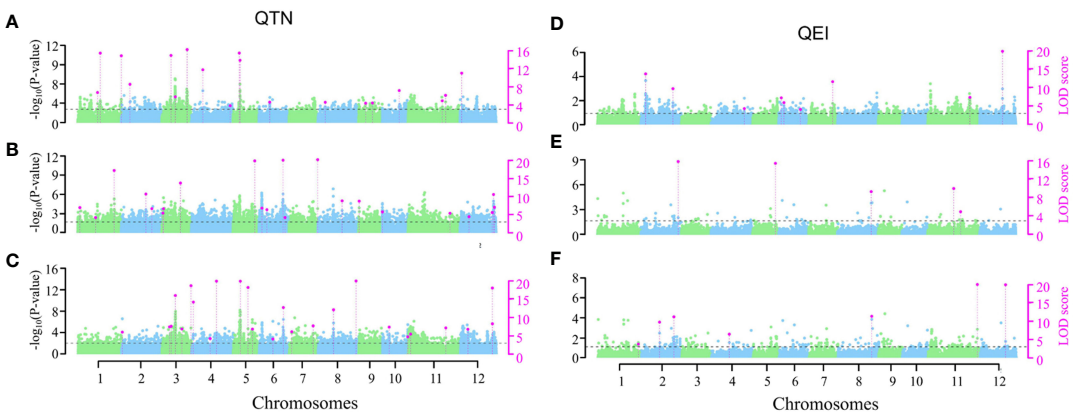


FIGURE 5
Manhattan plots of two-environment analyses on GBEI **(A, D)**, GLEI **(B, E)** and GREI **(C, F)**. The horizontal dashed lines indicate the LOD = 3.0 threshold. The left vertical axis is the $-\log_{10}$ (*P*-value), while the right vertical axis is the LOD score for each SNP marker. Pink dots indicate significant ($-\log_{10}$(*P*-value) ≥ 6.588) or suggested ($-\log_{10}$(*P*-value) < 6.588 but LOD ≥ 3) QTNs **(A–C)** or QEIs **(D–F)**.

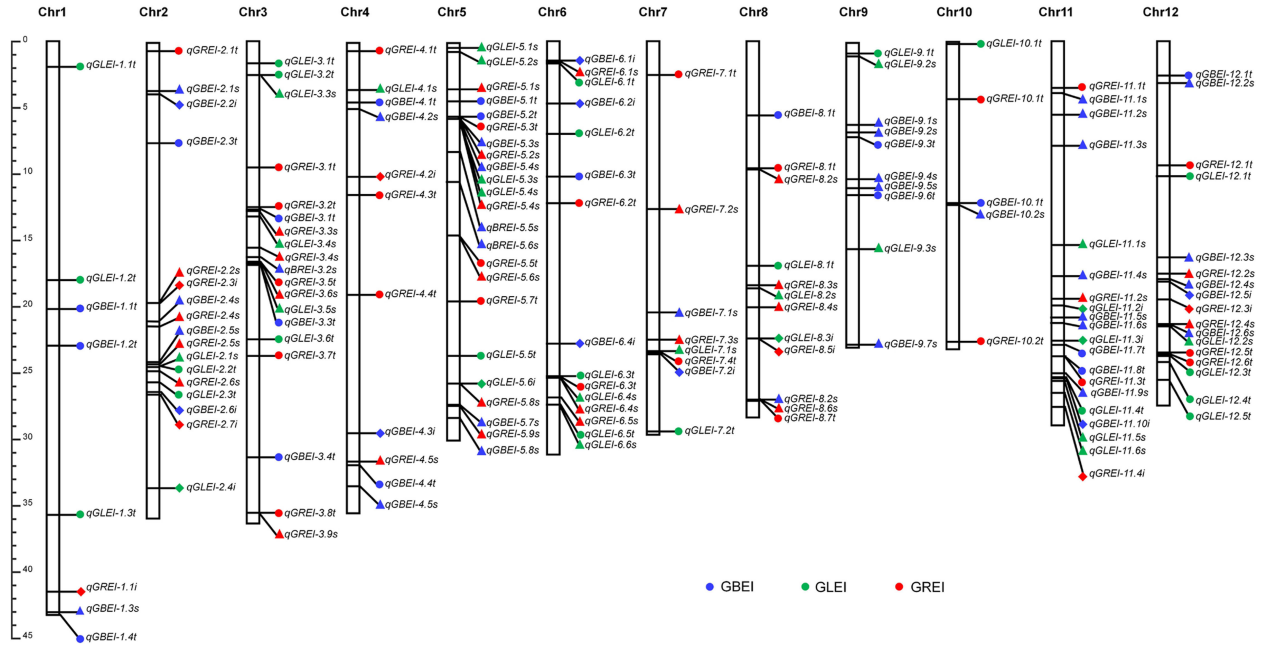


FIGURE 6
Locations of QTNs/QEIs for GBEI, GLEI and GREI in the rice genome. The QTNs detected by single-environment analysis are indicated by solid circles. The QTNs and QEIs detected by two-environment analysis are indicated by filled triangles and solid diamonds, respectively.

TABLE 6 Candidate genes for GBEI, GLEI and GREI .

Gene ID[1]	Gene name	Nearby QTN/QEI	Chr.	No. of Haplotypes	Mutation type	Annotation
Os04g53310	*OsSSIIIb*	*qGREI-4.5s*	4	5	non-synonymous, UTR-5' mutation	soluble starch synthase 3, chloroplast precursor
Os05g02070	*OsMT2b*	*qGLEI-5.1s*	5	2	UTR-5' mutation	metallothionein
Os06g04200	*wx*; *qGC-6*; *Wx-mq*; *Wx-op*	*qGBEI-6.1i*, *qGREI-6.1s*, *qGLEI-6.1t*	6	4	non-synonymous, UTR-5' mutation	granule-bound starch synthase
Os06g12450	*ALK*; *OsSSIIa*	*qGLEI-6.2t*	6	3	non-synonymous, splice site mutation	soluble starch synthase 2-3, chloroplast precursor
Os08g09230	*OsSSIIIa*; *Flo5*	*qGBEI-8.1t*	8	2	non-synonymous mutation	starch synthase III

1. The full gene ID includes a prefix LOC_Os.

indicating that the genes controlling these traits interacted with the environments. Joint analysis of the two environmental datasets using the 3VmrMLM method revealed 21 QEIs for three traits, also indicating the interaction between QTN and environment. Actually, ANOVA results showed significant genotype-by-environment interaction in the three traits (Table 1). However, there were no overlapping sites between QEI and QTNs detected based on two environmental data, indicating that all QEIs had no significant additive or dominant effect, but only the interaction effect between additive or dominant and environment, while all the QTNs in two-environment jointly analyze were opposite. Using the same 3VmrMLM method in previous studies, the overlapping sites between QEI and QTNs were also few, ranging from 1-3 sites (Han et al., 2022; He et al., 2022; Yu et al., 2022; Zhang et al., 2022; Jiang et al., 2023; Zhao et al., 2023), except for the study of Zou et al. (2022), which found 13 overlapping sites. From the perspective of the effect of QEI, since most QEIs do not have a significant additive or dominant effect, their reliability needs to be further confirmed.

In this study, among the 78 QTNs detected by single-environment analysis, only four QTNs were detected in both environments simultaneously (Table 4; Supplemental Table 2), indicating that only a small proportion (~5%) of QTNs exhibited stable significant effects across the environments. Interestingly, these four stable QTNs appear to represent four different types in terms of the way of being detected (Supplemental Table 2). The first type is *qGREI-5.6s*, which was detected by the same method in both environments, and no other methods detected it in either environment. The second type is *qGBEI-5.4s*, which was detected

10.3389/fpls.2023.1250854

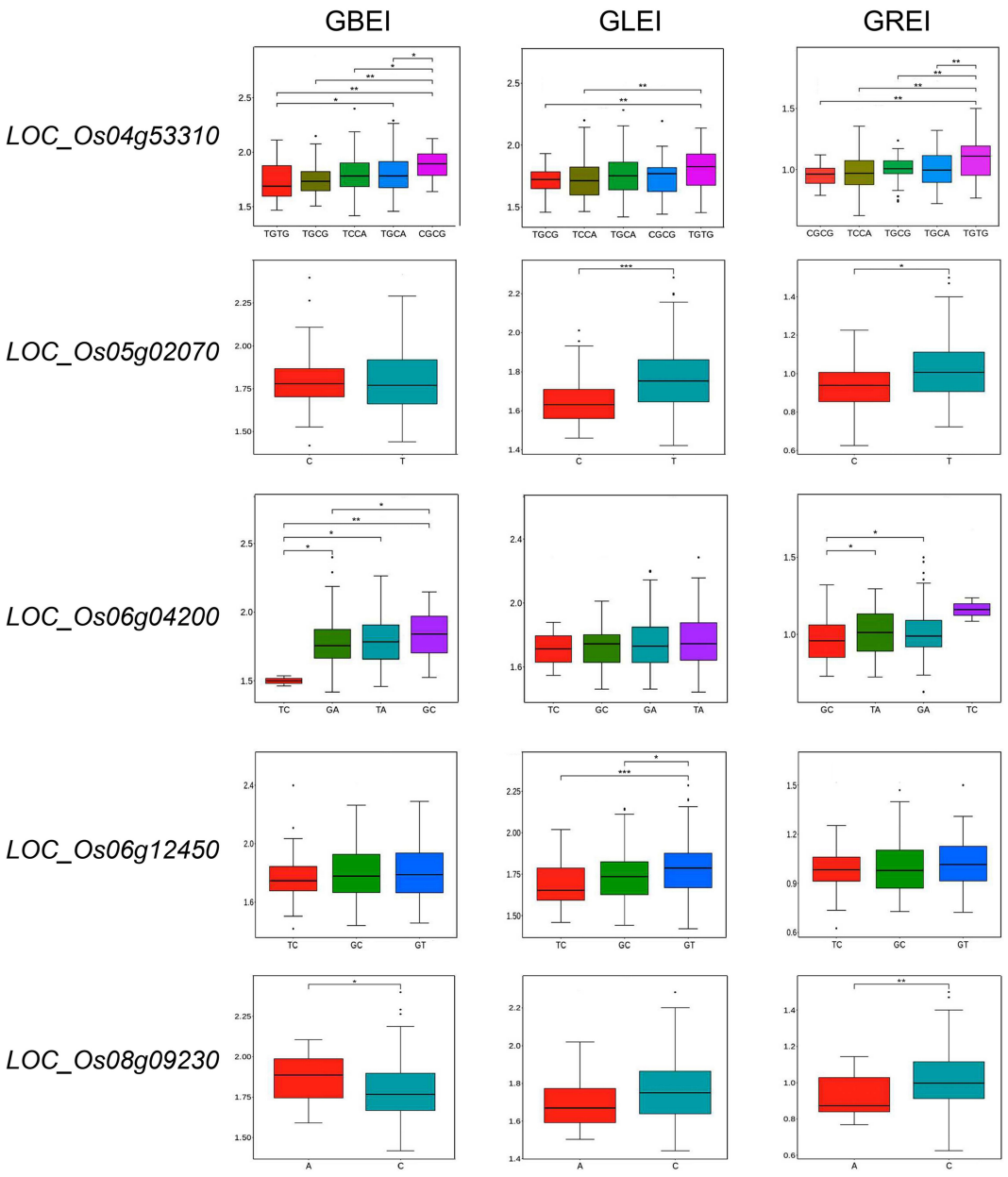


FIGURE 7
Haplotype analysis of candidate genes for GBEI, GLEI and GREI. *, ** and *** indicate significance at P<0.05, P<0.01 and P<0.001, respectively.

by one method in one environment, but by another method in the other environment. The third type is *qGREI-5.2s*, which was detected by one method in one environment, but by multiple other methods in the other environment. The fourth type is *qGLEI-3.3s*, which was also detected by one method in one environment and by multiple methods in the other environment, but with one method being the same in the two environments. It is noticeable that three of the four stable QTNs were detected in two different environments due to the use of multiple methods. These findings highlight the advantages of employing multiple GWAS methods to analyze the data collected from diverse environmental conditions.

According to the definitions, GREI is a composite trait that comprises various levels of component traits, which exhibit correlation with grain length and grain breadth before cooking (L_0 and B_0) or after cooking (L_1 and B_1), and is directly proportional to GLEI while inversely proportional to GBEI. Evidently, genes governing GBEI and GLEI may also impact GREI in principle. In other words, the QTLs for GREI may exhibit pleiotropic effects on its component traits or correlated traits. In this study, we did identify 4 QTNs that simultaneously influence GREI and GLEI, and 2 QTNs that simultaneously affect GREI and GBEI (Table 7). This was consistent with the high correlation between GREI and GLEI and GBEI (Table 2). As expected, there were no QTNs pleiotropic on GLEI and GBEI, which is in line with the conclusion that GLEI and GBEI are independent traits and have different genetic bases. Moreover, 3 QTNs controlling GLEI and

10.3389/fpls.2023.1250854

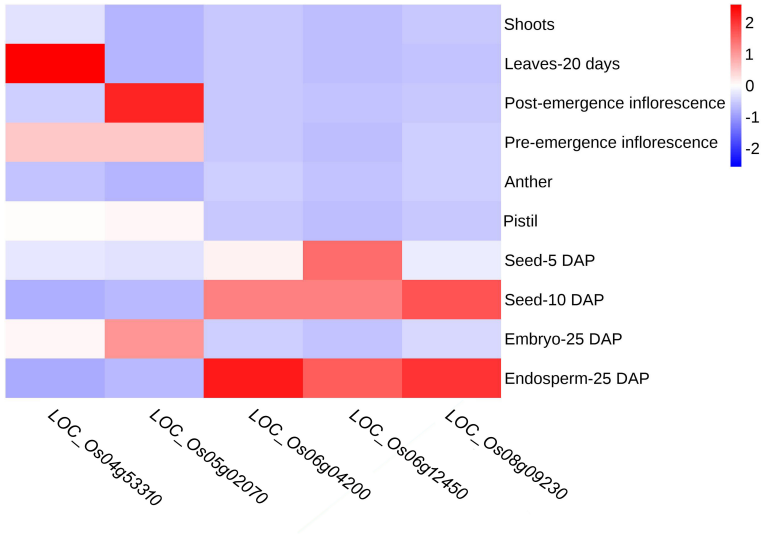


FIGURE 8
Heatmap of candidate gene expression analysis by RNA-seq data from RGAP database. Red boxes indicate high transcript levels and blue boxes indicate low transcript levels. DAP, days after pollination.

GREI respectively were detected simultaneously in single and two environments, demonstrating the stability of these QTNs.

As mentioned above in the introduction, there were 10, 47 and 15 reported QTLs controlling length, width and length-width expansion caused by cooking in rice grain. Upon comparing these QTLs with the QTNs mapped in this study, we observed that 12, 9, and 15 QTNs for GBEI, GLEI, and GREI detected in this study were located within the intervals of one or more previously reported QTLs (Table S6). These comparisons provide evidence for the reliability of the QTLs detected in this study. Notably, the four putative genes (*LOC_Os05g02070*, *LOC_Os06g04200*, *LOC_Os06g12450*, and *LOC_Os08g09230*) identified in this study were found to be in close proximity to four of the aforementioned QTLs).

Due to the swelling of starch granules during cooking, rice grain cooking-caused expansion traits, such as GBEI, GLEI and GREI, is expected to be influenced by starch-related traits which include two typical traits: chalkiness rate and amylose content. Chalkiness rate is a crucial parameter for assessing the visual quality of rice, as high chalkiness rate can lead to easy breakage of grains during processing, low amylose content, and poor eating quality. Thi et al. (2020) utilized a genetic population to map GREI and discovered a positive correlation between amylose content and GREI, with high AC content leading to increased GREI. *OsMT2b* encodes a metallothionein that binds to metal ions and scavenges reactive oxygen species (ROS). Wu et al. (2022) reported that WCR1, a negative regulator of rice chalkiness rate, functions to regulate *OsMT2b* (*LOC_Os05g02070*) transcription level and inhibit 26S proteasome-mediated *OsMT2b* protein degradation, thereby facilitating ROS clearance, delaying programmed cell death (PCD) of endosperm cells, and ultimately increasing the accumulation of

TABLE 7 Common QTLs between GREI, GBEI and GREI, or between different analysis aspects.

Chr.	QTN name	Trait	QTN pos. (bp)
2	*qGREI-2.2s, qGREI-2.3i*	GREI	19,642,336
2	*qGLEI-2.2s, qGLEI-2.2t*	GLEI	24,264,276
3	*qGLEI-3.3s, qGLEI-3.2t*	GLEI	2,521,638
3	*qGLEI-3.5s, qGREI-3.5s*	GLEI	16,774,870
3	*qGREI-3.9s, qGREI-3.8t*	GREI	35,669,404
5	*qGBEI-5.3s, qGREI-5.2s, qGBEI-5.2t, qGREI-5.3t*	GBEI, GREI	5,369,111
5	*qGREI-5.6s, qGREI-5.5t*	GREI	14,585,838
5	*qGLEI-5.6i, qGREI-5.8s*	GLEI, GREI	25,726,382
6	*qGLEI-6.3t, qGREI-6.3t*	GLEI, GREI	25,000,609
6	*qGLEI-6.4s, qGREI-6.4s*	GLEI, GREI	25,062,099
8	*qGLEI-8.3i, qGREI-8.5i*	GLEI, GREI	22,185,608
11	*qGBEI-11.8t, qGREI-11.3t*	GBEI, GREI	23,854,971

storage substances, and reducing chalkiness rate. In this study, a SNP site is present in the 5'-UTR region of *OsMT2b* near *qGLEI-5.1s* (Figure S3), which may disrupt the expression of *OsMT2b*, thereby affecting the change in rice cooking caused expansion in the analyzed population. Furthermore, considering the expression pattern of *OsMT2b*, it is noteworthy that its expression level exhibits a significant reduction in the endosperm. This observation implies its potential indirect influence on starch synthesis or endosperm development.

wx (*LOC_Os06g04200*), *OsSSIIa* (*LOC_Os06g12450*), and *OsSSIIIa* (*LOC_Os08g09230*) are crucial genes involved in the biosynthesis of starch in rice grains. *wx* gene encodes granule-bound starch synthase (GBSS), a major enzyme responsible for amylose synthesis (Kharshiing and Chrungoo, 2021). It exerts a direct influence on the amylose content in the endosperm and pollen of rice, as well as the gel consistency of grains (Su et al., 2011). *OsSSIIa* encodes a soluble starch synthase II, and mutations in this gene may affect the activity of starch synthase, which in turn affects the synthesis of medium-length branched chains of amylopectin, changes the crystal layer structure, and ultimately alters the gelatinization temperature (Gao et al., 2003). *OsSSIIIa* encodes soluble starch synthase III, the second key enzyme involved in rice starch synthesis (Zhou et al., 2016). Mutations in *OsSSIIIa* can affect the structure of amylopectin, amylose content, and physicochemical properties of starch in rice grains. Double mutants of *OsSSIIa* and *OsSSIIIa* exhibited increased chalkiness and amylose content, increased gelatinization temperature, and decreased viscosity (Zhang et al., 2011). In this study, these three genes exhibited the SNP loci with genetic effects. In haplotype analysis, significant differences in GBEI, GLEI, or GREI were observed across different haplotypes caused by SNPs within these genes. In expression pattern analysis, these three genes were highly expressed in the endosperm and seeds 10 days after pollination. All the evidence supported the hypothesis that these three genes were candidate genes controlling CCRGE.

In addition, *OsSSIIIb* (*LOC_Os04g53310*) is a gene that encodes soluble starch synthase in rice. Its expression level and activity directly impact the synthesis and quality of starch in rice endosperm. *OsSSIIIb* can interact coordinately with *OsSSIIIa*, and loss of function of both genes leads to an increase in resistant starch content in cooked rice (Wang et al., 2023). Although its protein function is redundant with OsSSIIIa, its expression pattern differs significantly from *OsSSIIIa* which is expressed in the endosperm. *OsSSIIIb* is mainly expressed in leaves but not endosperm (Figure 8). In this study, the five haplotypes generated by the four SNP loci contained in the *OsSSIIIb* gene exhibit significant differences in three traits. The evidence proves that *OsSSIIIb* may indirectly participate in starch sythesis and subsequently affect CCRGE.

5 Conclusion

In this study, data of GBEI, GLEI and GREI, three traits related to rice grain cooked expansion, were collected from 345 rice accessions in two distinct environments. Utilizing 193,582 SNP markers, seven methods were employed to identify QTNs based on single-environment data, while the 3VmrMLM method was utilized to identify QTNs and QEIs based on two-environment data. A total of 165 reliable QTNs/QEIs were detected, with 60, 46 and 59 of them being associated with GLEI, GBEI and GREI, respectively. Additionally, 26 genes related to starch synthesis or endosperm development were found to be located around these QTNs/QEIs. Further haplotype and expression pattern analyses led to the identification of five candidate genes, namely *LOC_Os04g53310* (*OsSSIIIb*), *LOC_Os05g02070* (*OsMT2b*), *LOC_Os06g04200* (*wx*), *LOC_Os06g12450* (*OsSSIIa*), and *LOC_Os08g09230* (*OsSSIIIa*). These findings can be instrumental in identifying genes and conducting in-depth genetic research on CCRGE.

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

YZ and WW conceived and designed the experiment. KT, LL, EK, EN, ML, WN and SA measured the phenotypes of the traits. YZ, KT and XX analyzed the data. YZ and KT wrote the draft. WW revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by Natural Science Foundation of Fujian Province (CN) (2020I0009, 2022J01596), Cooperation Project on University Industry-Education-Research of Fujian Provincial Science and Technology Plan (CN) (2022N5011), Lancang-Mekong Cooperation Special Fund (2017-2018), International Sci-Tech Cooperation and Communication Program of Fujian Agriculture and Forestry University (KXGH17014).

Acknowledgments

We thank Mr. Jinzhong Li for his help in field experiments. We also thank Dr. Weiqi Tang and Mr. Likun Huang for their help in data analysis.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1250854/full#supplementary-material

SUPPLEMENTARY FIGURE 1
SNP site and its resulting mutation type in five candidate genes. The blue boxes represent exons; the horizontal purple lines represent introns; the white boxes represent 5′ or 3′-UTR. The direction of a white box indicates the direction of the gene in the genome.

SUPPLEMENTARY FIGURE 2
Manhattan plots of single environment anlalyses by six methods in mrMLM R package on GBEI **(A, D)**, GLEI **(B, E)** and GREI **(C, F)**. The horizontal dashed lines indicate the LOD = 3.0 threshold. The left vertical axis is the -$\log_{10}$ (*P*-value), while the right vertical axis is the LOD score for each SNP marker. Pink dots indicate QTNs detected by more than one method. Blue dots indicate QTNs detected by only one method.

SUPPLEMENTARY FIGURE 3
Manhattan plots of two-environment analyses by 3VmrMLM on GBEI **(A, D)**, GLEI **(B, E)** and GREI **(C, F)**. The horizontal dashed lines indicate the LOD = 3.0 threshold. The left vertical axis is the -$\log_{10}$ (*P*-value), while the right vertical axis is the LOD score for each SNP marker. Pink dots indicate significant (-$\log_{10}$(*P*-value) $\geq$ 6.588) or suggested (-$\log_{10}$(*P*-value) < 6.588 but LOD $\geq$ 3).

References

Ahn, S., Bollich, C., Mcclung, A., and Tanksley, S. D. (1993). RFLP analysis of genomic regions associated with cooked- kernel elongation in rice. *Theor. Appl. Genet.* 87 (1-2), 27–32. doi: 10.1007/BF00223739

Alexander, D. H., and Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinf.* 12, 246. doi: 10.1186/1471-2105-12-246

Amarawathi, Y., Singh, R., Singh, A. K., Singh, V. P., Mohapatra, T., Sharma, T. R., et al. (2008). Mapping of quantitative trait loci for basmati quality traits in rice (*Oryza sativa* L.). *Mol. Breed.* 21, 49–65. doi: 10.1007/s11032-007-9108-8

Arikit, S., Wanchana, S., Khanthong, S., Saensuk, C., Thianthavon, T., Vanavichit, A., et al. (2019). QTL-seq identifies cooked grain elongation QTLs near soluble starch synthase and starch branching enzymes in rice (*Oryza sativa* L.). *Sci. Rep.* 9 (1), 1–10. doi: 10.1038/s41598-019-44856-2

Burghardt, L. T., Young, N. D., and Tiffin, P. (2017). A guide to genome-wide association mapping in plants. *Curr. Protoc. Plant Biol.* 2 (1), 22–38. doi: 10.1002/cppb.20041

Cheng, F., Zhong, L., Wang, F., and Zhang, G. P. (2005). Differences in cooking and eating properties between chalky and translucent parts in rice grains. *Food Chem.* 90 (1-2), 39–46. doi: 10.1016/j.foodchem.2004.03.018

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1^{118}; *iso-2*; *iso-3*. *Fly* 6 (2), 80–92. doi: 10.4161/fly.19695

Eizenga, G. C., Jia, M. H., Jackson, A. K., Boykin, D. L., Ali, M. L., and Shakiba, E. (2019). Validation of yield component traits identified by genome-wide association mapping in a *tropical japonica*×*tropical japonica* rice biparental mapping population. *Plant Genome* 12 (1), 1–18. doi: 10.3835/plantgenome2018.04.0021

Feng, F., Li, Y., Qin, X., Liao, Y., and Siddique, K. H. M. (2017). Changes in rice grain quality of Indica and Japonica type varieties released in China from 2000 to 2014. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.01863.eCollection.2017

Fitzgerald, M. A., McCouch, S. R., and Hall, R. D. (2009). Not just a grain of rice: The quest for quality. Trends in. *Plant Sci.* 14, 133–139. doi: 10.1016/j.tplants.2008.12.004

Frei, M., Siddhuraju, P., and Becker, K. (2003). Studies on the in *vitro* starch digestibility and the glycemic index of six different indigenous rice cultivars from the Philippines. *Food Chem.* 83, 395–402. doi: 10.1016/S0308-8146(03)00101-8

Gao, Z., Zeng, D., Cui, X., Zhou, Y., Yan, M., Huang, D., et al. (2003). Map-based cloning of the ALK gene, which controls the gelatinization temperature of rice. *Sci. China C Life Sci.* 46 (6), 661–668. doi: 10.1360/03yc0099

Ge, X., Xing, Y. Z., Xu, C. G., and He, Y. Q. (2005). QTL analysis of cooked rice grain elongation, volume expansion, and water absorption using a recombinant inbred population. *Plant Breed.* 124 (2), 121–126. doi: 10.1111/j.1439-0523.2004.01055.x

Golam, F., and Prodhan, Z. H. (2013). Kernel elongation in rice. *J. Sci. Food Agr.* 93 (3), 449–456. doi: 10.1002/jsfa.5983

Govindaraj, P., Vinod, K., Arumugachamy, S., and Maheswaran, M. (2009). Analysing genetic control of cooked grain traits and gelatinization temperature in a double haploid population of rice by quantitative trait loci mapping. *Euphytica* 166 (2), 165–176. doi: 10.1007/s10681-008-9808-0

Han, X., Tang, Q., Xu, L., Guan, Z., Tu, J., Yi, B., et al. (2022). Genome-wide detection of genotype environment interactions for flowering time in *Brassica napus*. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.1065766

He, L., Wang, H., Sui, Y., Miao, Y., Jin, C., and Luo, J. (2022). Genome-wide association studies of five free amino acid levels in rice. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.1048860

He, Y., Xing, Y., Ge, X., Li, X., and Xu, C. (2003). Gene mapping for elongation index related traits on cooked rice grain quality. *Mol. Plant Breed.* 1 (5/6), 613–619. doi: 10.3969/j.issn.1672-416X.2003.05.004

Hossain, M. S., Singh, A. K., and Zaman, F. U. (2009). Cooking and eating characteristics of some newly identified inter sub-specific (*indica/japonica*) rice hybrids. *ScienceAsia* 35 (4), 320–325. doi: 10.2306/scienceasia1513-1874.2009.35.320

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 (11), 961. doi: 10.1038/ng.695

Huang, X., Zhao, Y., Li, C., Wang, A., Zhao, Q., and Li, W. (2012). Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat. Genet.* 44 (1), 32. doi: 10.1038/ng

Jiang, S., Huang, C., Xu, Z., and Chen, W. (2008). QTL dissection of cooked rice elongation in rice (*Oryza sativa* L. Japonica). *Plant Physiol. Commun.* 44, 1091–1094. doi: 10.13592/j.cnki.ppj.2008.06.022

Jiang, H., Lv, S., Zhou, C., Qu, S., Liu, F., Sun, H., et al. (2023). Identification of QTL, QTL-by-environment interactions, and their candidate genes for resistance HG Type 0 and HG Type 1.2.3.5.7 in soybean using 3VmrMLM. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1177345

Kharshiing, G., and Chrungoo, N. K. (2021). *Wx* alleles in rice: relationship with apparent amylose content of starch and a possible role in rice domestication. *J. Genet.* 100, 65. doi: 10.1007/s12041-021-01311-4

Li, Y., Tao, H., Xu, J., Shi, Z., Ye, W., Wu, L., et al. (2015). QTL analysis for cooking traits of super rice with a high-density SNP genetic map and fine mapping of a novel boiled grain length locus. *Plant Breed.* 134 (5), 535–541. doi: 10.1111/pbr.12294

Li, J., Xiao, J., Grandillo, S., Jiang, L., Wan, Y., Deng, Q., et al. (2004). QTL detection for rice grain quality traits using an interspecific backcross population derived from cultivated Asian (*O. sativa* L.) and African (*O. glaberrima* S.) rice. *Genome* 47 (4), 697–704. doi: 10.1139/g04-029

Li, M., Zhang, Y. W., Xiang, Y., Liu, M. H., and Zhang, Y. M. (2022). IIIVmrMLM: The R and C++ tools associated with 3VmrMLM, a comprehensive GWAS method for dissecting quantitative traits. *Mol. Plant* 15, 1251–1253. doi: 10.1016/j.molp.2022.06.002

Liu, L., Yan, X. Y., Jiang, L., Zhang, W. W., Wang, M. Q., Zhou, S. R., et al. (2008). Identification of stably expressed quantitative trait loci for cooked rice elongation in non-Basmati varieties. *Genome* 51 (2), 104–112. doi: 10.1139/g07-106

Malik, A., Kumar, A., Ellur, R. K., Krishnan, S. G., Dixit, D., Bollinedi, H., et al. (2022). Molecular mapping of QTLs for grain dimension traits in Basmati rice. *Front. Genet.* 13. doi: 10.3389/fgene.2022.932166

Misra, G., Badoni, S., Anacleto, R., Graner, A., Alexandrov, N., and Sreenivasulu, N. (2017). Whole genome sequencing-based association study to unravel genetic architecture of cooked grain width and length traits in rice. *Sci. Rep.* 7 (1), 12478. doi: 10.1038/s41598-017-12778-6

Pang, Y., Ali, J., Wang, X., Franje, N. J., Revilleza, J. E., Xu, J., et al. (2016). Relationship of rice grain amylose, gelatinization temperature and pasting properties for breeding better eating and cooking quality of rice varieties. *PloS One* 11, e0168483. doi: 10.1371/journal.pone.0168483

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Ren, W. L., Wen, Y. J., Dunwell, J. M., and Zhang, Y. M. (2018). pKWmEB: integration of kruskal-Wallis test with empirical bayes under polygenic background control for multi-locus genome-wide association study. *Heredity* 120, 208–218. doi: 10.1038/s41437-017-0007-4

Shen, N., Lai, K., Nian, J., Zeng, D., Hu, J., Gao, Z., et al. (2011). Mapping and genetic analysis of quantitative trait loci for related traits of cooked rice. *Chin. J. Rice Sci.* 25 (5), 475–482. doi: 10.3969/j.issn.1001-7216.2011.05.004

Shen, S., Zhuang, J., Wang, S., Shu, Q., Bao, J., Xia, Y., et al. (2005). Analysis on the QTLs with main, epistasis and genotype- environmental interaction effects for cooked rice elongation. *Chin. J. Rice Sci.* 19, 319–322. doi: 10.3321/j.issn:1001-7216.2005.04.006

Su, Y., Rao, Y., Hu, S., Yang, Y., Gao, Z., Zhang, G., et al. (2011). Map-based cloning proves *qGC-6*, a major QTL for gel consistency of japonica/indica cross, responds by *Waxy* in rice (*Oryza sativa* L.). *Theor. Appl. Genet.* 123 (5), 859–867. doi: 10.1007/s00122-011-1632-6

Suwannaporn, P., and Linnemann, A. (2008). Rice-eating quality among consumers in different rice grain preference contries. *J. Sens. Stud.* 23, 1–13. doi: 10.1111/j.1745-459X.2007.00129.x

Swamy, B. P. M., Kaladhar, K., Rani, S. N., Prasad, G. S. V., Viraktamath, B. C., Reddy, G. A., et al. (2012). QTL analysis for grain quality traits in 2 BC_2F_2 populations derived from crosses between *Oryza sativa* cv. Swarna and 2 accessions of *O. nivara*. *J. Hered.* 103 (3), 442–452. doi: 10.1093/jhered/esr145

Tamba, C. L., Ni, Y., and Zhang, Y. (2017). Iterative sure independence screening EM-bayesian LASSO algorithm for multi-locus genome-wide association studies. *PloS Comput. Biol.* 13, e1005357. doi: 10.1371/journal.pcbi.1005357

Tamba, C., and Zhang, Y. M. (2018). A fast mrMLM algorithm for multi-locus genome-wide association studies. *bioRxiv* 7, 341784. doi: 10.1101/341784

The 3,000 rice genomes project (2014). The 3,000 rice genomes project. *GigaScience* 3, 7. doi: 10.1186/2047-217X-3-7

Thi, K. M., Zheng, Y., Khine, E. E., Nyein, E. E., Lin, M. H. W., Oo, K. T., et al. (2020). Mapping of QTLs conferring high grain length-breadth relative expansion during cooking in rice cultivar Paw San Hmwe. *Breed. Sci.* 70, 551–557. doi: 10.1270/jsbbs.20040

Tian, R., Jiang, G. H., Shen, L. H., Wang, L. Q., and He, Y. Q. (2005). Mapping quantitative trait loci underlying the cooking and eating quality of rice using a DH population. *Mol. Breed.* 15 (2), 117–124. doi: 10.1007/s11032-004-3270-z

Wang, S. B., Feng, J. Y., Ren, W. L., Huang, B., Zhou, L., Wen, Y. J., et al. (2016). Improving power and accuracy of genome-wide association studies via a multilocus mixed linear model methodology. *Sci. Rep.* 6, 19444. doi: 10.1038/srep19444

Wang, A., Jing, Y., Cheng, Q., Zhou, H., Wang, L., Gong, W., et al. (2023). Loss of function of *SSIIIa* and *SSIIIb* coordinately confers high RS content in cooked rice. *Proc. Natl. Acad. Sci. U. S. A.* 120 (19), e2220622120. doi: 10.1073/pnas.2220622120

Wang, L. Q., Liu, W. J., Xu, Y., He, Y. Q., Luo, L. J., Xing, Y. Z., et al. (2007). Genetic basis of 17 traits and viscosity parameters characterizing the eating and cooking quality of rice grain. *Theor. Appl. Genet.* 115 (4), 463–476. doi: 10.1007/s00122-007-0580-7

Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., et al. (2018). Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* 557, 43–49. doi: 10.1038/s41586-018-0063-9

Wang, Q., Tang, J., Han, B., and Huang, X. (2020). Advances in genome-wide association studies of complex traits in rice. *Theor. Appl. Genet.* 133 (5), 1415–1425. doi: 10.1007/s00122-019-03473-3

Wen, Y. J., Zhang, H., Ni, Y. L., Huang, B., Zhang, J., Feng, J. Y., et al. (2018). Methodological implementation of mixed linear models in multi-locus genome wide association studies. *Brief. Bioinform.* 19, 700–712. doi: 10.1093/bib/bbw145

Wu, B., Yun, P., Zhou, H., Xia, D., Gu, Y., Li, P., et al. (2022). Natural variation in *WHITE-CORE RATE 1* regulates redox homeostasis in rice endosperm to affect grain quality. *Plant Cell* 34, 1912–1932. doi: 10.1093/plcell/koac057

Yu, K., Miao, H., Liu, H., Zhou, J., Sui, M., Zhan, Y., et al. (2022). Genome-wide association studies reveal novel QTLs, QTL-byenvironment interactions and their candidate genes for tocopherol content in soybean seed. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.1026581

Zhang, G., Cheng, Z., Zhang, X., Guo, X., Su, N., Jiang, L., et al. (2011). Double repression of soluble starch synthase genes *SSIIa* and *SSIIIa* in rice (*Oryza sativa* L.) uncovers interactive effects on the physicochemical properties of starch. *Genome* 54 (6), 448–459. doi: 10.1139/g11-010

Zhang, J., Feng, J. Y., Ni, Y. L., Wen, Y. J., Niu, Y., Tamba, C. L., et al. (2017). pLARmEB: integration of least angle regression with empirical Bayes formultilocus genome-wide association studies. *Heredity* 118, 517–524. doi: 10.1038/hdy.2017.8

Zhang, Y. M., Jia, Z., and Dunwell, J. M. (2019). Editorial: The applications of new multi-locus GWAS methodologies in the genetic dissection of complex traits. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00100

Zhang, Y. W., Tamba, C. L., Wen, Y. J., Li, P., Ren, W. L., Ni, Y. L., et al. (2020). mrMLM v4.0.2: an R platform for multi-locus genome-wide association studies. *Genomics Proteomics Bioinf.* 18 (4), 481–487. doi: 10.1016/j.gpb.2020.06.006

Zhang, J., Wang, S., Wu, X., Han, L., Wang, Y., and Wen, Y. (2022). Identification of QTNs, QTN-byenvironment interactions and genes for yield-related traits in rice using 3VmrMLM. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.995609

Zhang, G. H., Zeng, D. L., Guo, L. B., Qian, Q., Zhang, G. P., Teng, S., et al. (2004). Genetic dissection of cooked rice elongation in rice (*Oryza sativa* L.). *Yi Chuan* 26 (6), 887–892. doi: 10.3321/j.issn:0253-9772.2004.06.021

Zhao, Q., Shi, X. S., Wang, T., Chen, Y., Yang, R., Mi, J., et al. (2023). Identification of QTNs, QTN-by-environment interactions, and their candidate genes for grain size traits in main crop and ratoon rice. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1119218

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

Zhou, H., Wang, L., Liu, G., Meng, X., Jing, Y., Shu, X., et al. (2016). Critical roles of soluble starch synthase SSIIIa and granule-bound starch synthase Waxy in synthesizing resistant starch in rice. *Proc. Natl. Acad. Sci. U. S. A.* 113 (45), 12844–12849. doi: 10.1073/pnas.1615104113

Zou, J. F., Chen, Y., Ge, C., Liu, J. Y., and Zhang, Y. M. (2022). Identification of QTN-by-environment interactions and their candidate genes for soybean seed oil-related traits using 3VmrMLM. *Front. Plant Sci.* 13, 1096457. doi: 10.3389/fpls.2022.1096457

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 25 October 2023
DOI 10.3389/fpls.2023.1229457

OPEN ACCESS

EDITED BY
Yuan-Ming Zhang,
Huazhong Agricultural University, China

REVIEWED BY
Jian-Fang Zuo,
Huazhong Agricultural University, China
Maria Samsonova,
Peter the Great St. Petersburg Polytechnic University, Russia

*CORRESPONDENCE
Sylvie Cloutier
Sylviej.cloutier@agr.gc.ca
Frank M. You
frank.you@agr.gc.ca
Liqiang He
heliqiang66@126.com

†These authors have contributed equally to this work

RECEIVED 26 May 2023
ACCEPTED 24 July 2023
PUBLISHED 25 October 2023

CITATION
He L, Sui Y, Che Y, Wang H, Rashid KY, Cloutier S and You FM (2023) Genome-wide association studies using multi-models and multi-SNP datasets provide new insights into pasmo resistance in flax. *Front. Plant Sci.* 14:1229457. doi: 10.3389/fpls.2023.1229457

Genome-wide association studies using multi-models and multi-SNP datasets provide new insights into pasmo resistance in flax

Liqiang He[1,2*†], Yao Sui[2†], Yanru Che[2†], Huixian Wang[2], Khalid Y. Rashid[1], Sylvie Cloutier[1*] and Frank M. You[1*]

[1]Ottawa Research and Development Centre, Agriculture and Agri-Food Canada, Ottawa, ON, Canada, [2]School of Tropical Agriculture and Forestry, School of Tropical Crops, Hainan University, Haikou, China

Introduction: Flax (*Linum usitatissimum* L.) is an economically important crop due to its oil and fiber. However, it is prone to various diseases, including pasmo caused by the fungus *Septoria linicola*.

Methods: In this study, we conducted field evaluations of 445 flax accessions over a five-year period (2012–2016) to assess their resistance to pasmo A total of 246,035 single nucleotide polymorphisms (SNPs) were used for genetic analysis. Four statistical models, including the single-locus model GEMMA and the multi-locus models FarmCPU, mrMLM, and 3VmrMLM, were assessed to identify quantitative trait nucleotides (QTNs) associated with pasmo resistance.

Results: We identified 372 significant QTNs or 132 tag QTNs associated with pasmo resistance from five pasmo resistance datasets (PAS2012–PAS2016 and the 5-year average, namely PASmean) and three genotypic datasets (the all SNPs/ALL, the gene-based SNPs/GB and the RGA-based SNPs/RGAB). The tag QTNs had R^2 values of 0.66–16.98% from the ALL SNP dataset, 0.68–20.54%from the GB SNP dataset, and 0.52–22.42% from the RGAB SNP dataset. Of these tag QTNs, 93 were novel. Additionally, 37 resistance gene analogs (RGAs)co-localizing with 39 tag QTNs were considered as potential candidates for controlling pasmo resistance in flax and 50 QTN-by-environment interactions (QEIs) were identified to account for genes by environmental interactions. Nine RGAs were predicted as candidate genes for ten QEIs.

Discussion: Our results suggest that pasmo resistance in flax is polygenic and potentially influenced by environmental factors. The identified QTNs provide potential targets for improving pasmo resistance in flax breeding programs. This study sheds light on the genetic basis of pasmo resistance and highlights the importance of considering both genetic and environmental factors in breeding programs for flax.

KEYWORDS

GWAS, multi-locus model, pasmo, SNP, flax

Introduction

Flax (*Linum usitatissimum* L.) is a valuable economic crop that provides linseed and stem fiber to humans (Singh et al., 2011; You et al., 2017). However, flax production is often constrained by pasmo, a disease caused by the fungus *Septoria linicola*, which reduces seed yield and fiber quality (Halley et al., 2004; He et al., 2018; Islam et al., 2021). The fungus infects flax from the seedling to the ripening stages. At the flowering stage, despite the application of fungicide, susceptible varieties have been reported to experience up to a 75% seed yield loss (Hall et al., 2016; Islam et al., 2021). Therefore, developing resistant varieties is a cost-effective and environmentally-friendly approach to protect flax from pasmo and its effects on yield.

Disease resistance in plants is typically quantitatively inherited and influenced by the environment. It is primarily governed by major resistant genes called *R* genes, which have been the topic of many studies (Marone et al., 2013; Yang et al., 2017). Most cloned *R* genes in plants belong to the nucleotide-binding site-leucine-rich repeat domain (NBS-LRR) class, also known as *NLRs*. For example, a cluster of *NLR* receptor-encoding genes confers durable resistance to *Magnaporthe oryzae* in rice (Deng et al., 2017), and the *rp1* gene in maize and its homolog in barley confer race-specific resistance to rust fungal diseases (Collins et al., 1999; Ayliffe et al., 2000). *Receptor like kinase* (*RLK*) genes also account for a significant proportion of *R* genes. For instance, the RLK-encoding barley *Rpg1* gene confers resistance to stem rust (Brueggeman et al., 2002), and rice *Pi-d2* gene confers resistance against rice blast (Chen et al., 2006). Transmembrane coiled-coil proteins (TM-CC) are another essential type of *R* gene-encoded proteins. The *Rph3* gene, originating from wild barley, is a TM-type *R* gene that encodes a protein that differs from all known plant disease resistance proteins and can significantly enhance barley leaf rust resistance (Dinh et al., 2022). The mutation-induced recessive *mlo* allele of the barley *Mlo* gene also encodes a TM domain protein, and confers broad-spectrum resistance to the fungal pathogen *Erysiphe graminis* (Buschges et al., 1997). Resistance gene analogs (RGAs) are key resistance gene candidates and have been well-characterized in flax (Sekhwal et al., 2015; You et al., 2018b). A total of 1327 RGAs have been categorized into 11 types: RLK (receptor-like protein kinase), TM-CC (transmembrane coiled-coil protein), RLP (receptor-like protein), TNL (TIR-NBS-LRRs), TX (TIR-unknown), NL (NBS-LRR), CNL (CC-NBS-LRR), TN (TIR-NBS), NBS (NBS domain only), CN (CC-NBS), and OTHERS.

Genome-wide association studies (GWAS) have emerged as a powerful and efficient approach for unraveling the genetic basis of complex traits in flax. Compared to traditional linkage mapping, GWAS can achieve higher resolution and more accurate mapping of quantitative trait nucleotides (QTNs) (He et al., 2018; You et al., 2018a; Soto-Cerda et al., 2021; You et al., 2022). However, GWAS has some limitations, including a higher risk of false-positive associations and a lower effectiveness in detecting quantitative trait loci (QTL) associated with rare alleles than biparental populations. Single-locus GWAS models, such as GEMMA and MLM, have proven to be effective in controlling spurious associations using the stringent Bonferroni correction but they are not suited to detecting minor QTL (Yu et al., 2006; Zhou and Stephens, 2012). To enhance the power of polygenic loci detection, multi-locus GWAS models have been developed (Segura et al., 2012; Zhang et al., 2019b). For instance, FarmCPU improves statistical power and reduces confounding associations (Liu et al., 2016), and mrMLM increases power, reduces the false positive rate, and has a shorter running time (Wang et al., 2016). However, these models do not fully assess the effects of QTN-by-environment interactions (QEIs) and QTN-by-QTN interactions (QQIs). To address these, a new multi-locus GWAS model called 3VmrMLM was proposed (Li et al., 2022b). This model estimates the genetic effects of three marker genotypes (AA, Aa and aa) while controlling all possible polygenic backgrounds. It is designed to detect QEIs and QQIs. Our previous study has shown that pasmo resistance in flax is controlled by polygenes (He et al., 2018). However, the small proportion of resistant accessions in the original core collection was limiting and additional research is warranted to detect main-effect QTNs and their corresponding causal genes. Furthermore, the QEIs associated with flax pasmo resistance are still largely unknown. Therefore, the newly released 3VmrMLM model to identify main-effect QTNs and QEIs is expected to improve our understanding of pasmo resistance in flax towards the better design of breeding solutions.

Our previous study has identified a total of 500 QTL associated with pasmo resistance in flax, including 67 stable and large-effect QTL and many additional small effect and environment-specific QTL (He et al., 2018). Here only 8.3% of the flax core collection was found to be resistant or moderately resistant to pasmo, based on the average pasmo severity over five consecutive years (2012–2016). To increase the proportion of resistant lines in the collection while simultaneously improving genetic diversity, 75 sequenced breeding lines were added to the core collection. Pasmo resistance data for these new lines, were collected between 2012 and 2016, alongside data from the existing 370 original accessions of the flax core collection (You et al., 2022; Zheng et al., 2023).

To gain a deeper understanding of pasmo resistance in flax at the genetic level, we conducted a GWAS on a diverse panel of 445 flax accessions, which included 370 accessions of the core collection and 75 selected breeding lines (SBLs). Compared to GWAS that use all SNPs (ALL) as genotypic data, gene-based SNPs (GB) and RGA-based SNPs (RGAB) GWAS have demonstrated higher power and resolution in QTL detection and candidate gene identification (Zhang et al., 2021; You et al., 2022). Thus, three genotypic datasets consisting of 246,035 SNPs (ALL), 65,147 SNPs within genes (GB), and 3,510 SNPs within RGAs (RGAB) were used in the analysis, along with four different GWAS models. These models

included one single-locus model (GEMMA) and three multi-locus models (FarmCPU, mrMLM, and 3VmrMLM), employed to detect quantitative trait nucleotides (QTNs) and QTN-by-environment interactions (QEIs) associated with pasmo resistance across five individual years (2012–2016). Our goal was to identify potential candidate genes conferring pasmo resistance in flax.

Materials and methods

Genetic panel for GWAS

A genetic panel of 445 flax accessions was used for GWAS. The panel included 370 accessions from the flax core collection, which was previously assembled from a worldwide collection of 3,378 flax accessions (Diederichsen et al., 2012; Soto-Cerda et al., 2013; He et al., 2018), and 75 breeding lines that were selected based on their resistance to pasmo, *Fusarium* wilt and powdery mildew diseases (You et al., 2022). The flax core collection included accessions from 11 geographical origins, and were classified based on their morphotype into 80 fibre and 290 linseed accessions. This panel included 17 landraces, 85 breeding lines, 232 cultivars, and 36 accessions of unknown improvement status (Figure 1A) (You et al., 2017). By adding the 75 SBLs to the core collection, the statistical power of the GWAS was increased. This diverse genetic panel allows for a more comprehensive analysis of the genetic variation within flax, and can provide insights into the genetic basis of resistance to pasmo disease and other traits of interest.

Phenotyping of pasmo resistance and statistical analysis

The 445 accessions of the diversity panel were evaluated for field resistance to pasmo over a period of five years (2012–2016) at Agriculture and Agri-Food Canada, Morden Research and Development Center's farm in Morden, Manitoba, Canada. A Type-2 modified augmented design (MAD2) was employed for the field experiments as described by You et al. (2017). The seeds were sown in mid-May each year, and 30-centimeter tall flax plants were inoculated with approximately 200 grams of pasmo-infected chopped straw from the previous growing season. To ensure disease infection and development, a spray system was operated for 5 minutes every half hour for 4 weeks.

Pasmo resistance was evaluated at the early brown boll stage (21–30 days after the flowering) by assessing the leaves and stems of all plants (~300) in a single row plot using a pasmo severity scale of 0–9. Ratings of 0–2 were classified as resistant (R), 3–4 as moderately resistant (MR), 5–6 as moderately susceptible (MS), and 7–9 as susceptible (S). Pasmo severity data were recorded for five individual years (PAS2012, PAS2013, PAS2014, PAS2015, and PAS2016). These five datasets and the five-year average

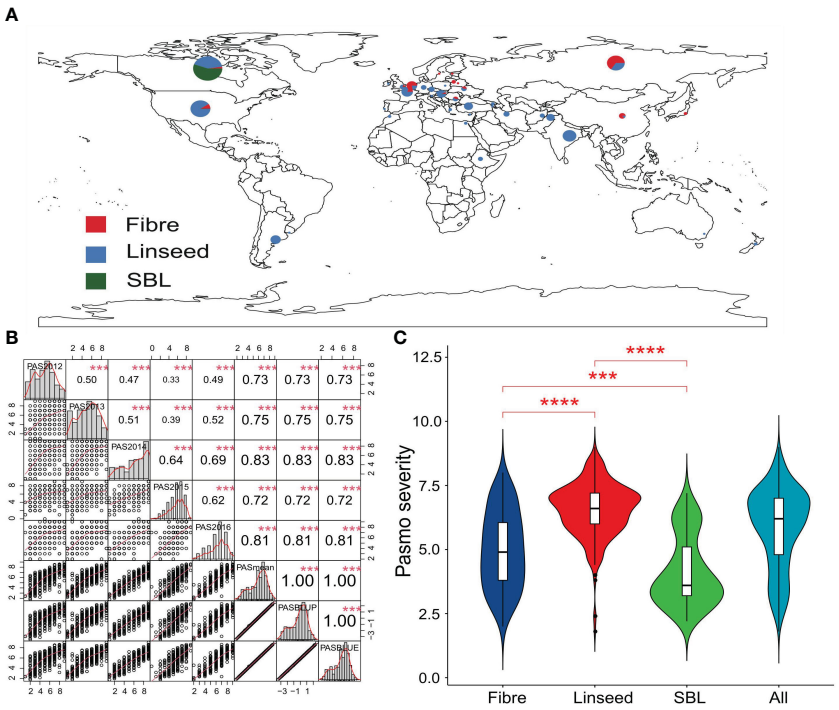


FIGURE 1
Geographic distribution and phenotyping for pasmo resistance in flax accessions. **(A)** Geographic distribution of 445 flax accessions. **(B)** Distribution and correlation matrix of pasmo severity in five consecutive years (2012–2016), mean, BLUP and BLUE pasmo severity over years. *** indicates significant correlation at the 0.1% probability level. **(C)** Violin plot of pasmo severity for the 80 fibre and 290 linseed accessions of the core collection and the 75 selected breeding lines. PAS2012, PAS2013, PAS2014, PAS2015, PAS2016, PASmean, PASBLUP and PASBLUE represent pasmo severity datasets for 2012, 2013, 2014, 2015, 2016, the 5-year average, the best linear unbiased prediction values and the best linear unbiased estimation values of pasmo severity over five years. *** and **** indicate statistical significance at the 0.1% and 0.01% probability level, respectively.

(PASmean) were used as the phenotypic data for all analyses in this study.

To account for environmental variation, the R package lme4 was used to generate the best linear unbiased prediction (BLUP) and best linear unbiased estimate (BLUE) datasets for the pasmo severity of the five years (Bates et al., 2015). A mixed linear model that treated accessions and years as random effects was used to calculate the BLUP values, while another mixed linear model that treated accessions as fixed effects and years as random effects was employed to obtain the BLUE values. The R package PerformanceAnalytics was used to analyze the correlations between the pasmo severity datasets, and to generate histograms and scatter plots (https://cran.r-project.org/web/packages/PerformanceAnalytics/index.html).

Re-sequencing for SNP discovery of the diversity panel

Genome re-sequencing was performed to obtain the genetic variation of 445 flax accessions. As previously described in He et al. (2018), the Illumina HiSeq 2000 platform (Illumina Inc., San Diego, USA) was used to generate 100-bp paired-end reads with an average coverage of ~15.5X of the reference genome. All raw reads were mapped to the flax reference genome using the BWA v0.6.1 mapping tool with a base-quality Q score in Phred scale > 20 and other default parameters (Jo and Koh, 2015). The mapped files were processed using SAMtools and an improved AGSNP pipeline for SNP calling (Li et al., 2009; You et al., 2011; You et al., 2012). The detected SNPs were further filtered with a minor allele frequency (MAF) > 0.05 and a SNP genotyping call rate ≥ 60% using PLINK (https://zzz.bwh.harvard.edu/plink/). After linkage disequilibrium (LD) filtering with pairwise correlation coefficients (r^2) among neighboring SNPs within 200kb > 0.8 and Beagle imputation with default parameters (Browning and Browning, 2007), a total of 246,035 high-quality SNPs were retained for further analysis. The genetic variant annotation and functional effect prediction of each SNP were characterized by snpEff software (Cingolani et al., 2012) based on the reference genome and corresponding annotation (You et al., 2018b).

Population structure analysis

To dissect the genetic structure and variation of the 445 flax accessions, principal component analysis (PCA) was performed using the obtained high-quality SNPs. The analysis was carried out with the PLINK software (Elhaik, 2022). For the SNP-based phylogenetic analysis, MEGA-CC was employed, using a pairwise gap deletion method for 1,000 bootstrap replicates (Kumar et al., 2012). The resulting phylogenetic tree was visualized using the Interactive Tree of Life (iTOL) tool (Letunic and Bork, 2021). The population stratification was estimated using ADMIXTURE (Alexander et al., 2009). The genome-wide LD decay was assessed using PopLDdecay v3.42 software to the squared correlation coefficient (r^2) between SNPs (Zhang et al., 2019a).

Genome-wide association study

The GWAS analysis for pasmo resistance was conducted using the five individual year (PAS2012, PAS2013, PAS2014, PAS2015, and PAS2016) and the five-year average (PASmean) datasets with four GWAS models. The models used included the single-locus model GEMMA and the multi-locus models FarmCPU (Liu et al., 2016), mrMLM (Wang et al., 2016) and 3VmrMLM (Li et al., 2022b). The kinship matrices were estimated using the protocol suggested by each GWAS software package. The genotypic data for the association panel comprised 246,035 high-quality SNPs (ALL) obtained from 445 flax accessions. Of these, the 65,147 SNPs that mapped to the genic regions constituted the gene-based (GB) SNP dataset, and the 3,510 SNPs that mapped to RGAs formed the RGA-based (RGAB) SNP dataset. These datasets were used in sequential analyses. The GEMMA software and R package GAPIT were employed to detect QTNs using default settings (Zhou and Stephens, 2012; Wang and Zhang, 2021). The R package mrMLM was applied to detect QTNs using parameters SearchRadius = 20, CriLOD = 3, and Bootstrap = FALSE (Zhang et al., 2020). The R package IIIVmrMLM implementing the 3VmrMLM model was used to detect main-effect QTNs and the QEIs (Li et al., 2022a). For the detection of the main-effect QTNs, the R package IIIVmrMLM was used with the following parameters: method = "Single_env", SearchRadius = 20, and svpal = 0.01. For QEI detection, the parameters used were method = "Multi_env", SearchRadius = 20, and svpal = 0.01. The association signals of the 3VmrMLM model were detected using a LOD score ≥ 3 (Li et al., 2022a). The threshold of significant association of GEMMA and FarmCPU was determined using a critical *P*-value at the 5% significant level that was subjected to Bonferroni correction (*P*-value = 2.03×10^{-7} for the ALL dataset, *P*-value = 7.67×10^{-7} for the GB dataset, and *P*-value = 1.42×10^{-5} for the RGAB dataset). Manhattan plots were generated using the IIIVmrMLM package with default settings.

QTN identification, candidate gene prediction, allele and haplotype analysis

In order to identify QTNs associated with pasmo resistance in flax, a GWAS was performed using individual year datasets (PAS2012–PAS2016) and a five-year average dataset (PASmean) in combination with the ALL, GB and RGAB genotypic datasets. QTNs detected in different genotypic datasets were analyzed independently and common QTNs were identified based on detection by two or more models or detection in two or more phenotypic datasets. Mann-Whitney U tests were used to validate significant differences between QTN alleles associated with pasmo severity. The significant QTNs were represented by tag QTNs for downstream analyses. R^2 values were calculated to determine the proportion of total variation explained by the pasmo resistance associated QTNs/QEIs. A total of 1,327 RGAs have previously been identified in the flax reference genome (You et al., 2018b). The co-localized RGAs within an estimated 4 kb distance of the averaged whole genome LD decay and local LD block defined flanking

regions of the detected QTNs/QEIs were considered as candidate genes. LDBlockShow v1.40 (Dong et al., 2021) was utilized to estimate the local LD block regions on the chromosomes. For allele analysis, the single SNP with HIGH functional effect prediction on the coding region (CDS) of each candidate gene were selected and tested for significant differences in pasmo severity using the Wilcox non-parametric test at the 5% probability level. Likewise, for haplotype analysis, all the SNPs within each candidate gene that were predicted with HIGH or MODERATE functional effect were considered. Subsequently, these SNPs underwent testing using the Wilcox non-parametric test at the 5% probability level to identify significant differences. A SNP with a HIGH functional effect prediction is assumed to have a disruptive impact on the protein, while a SNP with a MODERATE functional effect prediction is expected to be non-disruptive but could possibly change the protein's effectiveness.

Results

Evaluation of pasmo resistance

Pasmo resistance was evaluated in 445 flax accessions over five consecutive years (PAS2012–PAS2016). The geographic distribution and morphotypes of these accessions are shown in Figure 1A. Correlation coefficients were calculated among PAS2012, PAS2013, PAS2014, PAS2015, PAS2016, PASmean, pasmo best linear unbiased prediction (PASBLUP) and pasmo best linear unbiased estimation (PASBLUE) datasets, and ranged from 0.33 to 1.00, with the highest correlation observed between PASmean and PAS2014 ($r = 0.83$) (Figure 1B). PASmean was further analyzed due to its almost identical correlation coefficients with PASBLUP and PASBLUE ($r = 1.00$). The coefficient of variation (CV) of PAS2012–PAS2016 and PASmean datasets ranged from 24.17% to 39.24% (Supplementary Table S1). Significant differences in pasmo severity were observed between linseed, fibre accessions, and SBLs in this flax genetic panel. High resistance (low severity) to pasmo was observed in the 75 SBLs compared to the 370 accessions from the flax core collection (Figure 1C). The average pasmo severity over five years was 6.56 ± 1.05 for the 290 linseed accessions, 4.98 ± 1.50 for the 80 fibre accessions, and 4.13 ± 1.35 for the 75 breeding lines (Figure 1C). The data distribution and correlation analysis indicated that resistance against pasmo in flax is controlled by polygenes and potentially genetic by environment interactions.

Population structure

To analyze the genetic structure of the 445 flax accessions, a population structure analysis was performed using the ALL SNP dataset of 246,035 SNPs. The results indicated the 445 accessions were divided into five populations (Figure 2A). Population one consisted of 19 linseed accessions and 75 SBLs; population two was composed of 67 fibre accessions and 51 linseed accessions; population three contained 11 fibre accessions and 72 linseed accessions; population four comprised 39 linseed accessions, while population five consisted of only two fibre accessions and 109 linseed accessions. PCA and phylogenetic analysis by neighbor-joining (NJ) (Chen et al., 2014) also showed identical classification of the flax genetic panel into five groups (Figures 2B–D and Supplementary Figure S1). Therefore, a population structure Q matrix with K = 5 was adopted for downstream GWAS analyses. The linkage disequilibrium (LD) analysis showed that the LD decayed rapidly before 4 kb and subsequently became flat for this flax genetic panel (Figure 2E). Therefore, the 4 kb flanking region of each QTN was used for putative candidate gene prediction in subsequent analyses.

Identification of QTNs associated with pasmo resistance

A total of 372 significant QTNs were identified using six pasmo resistance datasets (PAS2012–PAS2016 and PASmean) and three genotypic datasets (ALL, GB and RGAB) using the single-locus model GEMMA and the multi-locus models FarmCPU, mrMLM and 3VmrMLM (Figure 3 and Supplementary Table S2). When the ALL genotypic dataset was used, 3VmrMLM detected the most QTNs (149), followed by mrMLM (89), FarmCPU (25), and GEMMA (4) (Table 1). Forty-seven QTNs were detected by both 3VmrMLM and mrMLM, two by 3VmrMLM, mrMLM, and FarmCPU, and another two by mrMLM, FarmCPU, and GEMMA (Figure 3A). Only one QTN (QTN-Lu4-14738243) was detected in three out of the six phenotypic datasets (PAS2012–PAS2016 and PASmean) (Figure 3B and Supplementary Table S2).

For the GB genotypic dataset, 3VmrMLM detected the most QTNs (105), followed by mrMLM (90), and GEMMA detected a single QTN (Table 1). Among these, 67 were detected by both 3VmrMLM and mrMLM, four by 3VmrMLM, mrMLM, and FarmCPU, and one by mrMLM, FarmCPU, and GEMMA (Figure 3C). Moreover, the same common QTN (QTN-Lu4-14738243) was detected in three out of the six phenotypic datasets (Figure 3D and Supplementary Table S2).

Similarly, 3VmrMLM detected the most QTNs (55) in the RGAB genotypic dataset, followed by mrMLM (28), FarmCPU (10), and GEMMA (2) (Table 1). Interestingly, QTN-Lu10-11656889 was detected by all four models (Figure 3E and Supplementary Table S2). Besides, three common QTNs (QTN-Lu8-23634276, QTN-Lu10-11656889, and QTN-Lu15-14719354) were detected in three out of six phenotypic datasets (Figure 3F and Supplementary Table S2). Notably, QTN-Lu14-2333894 was detected by all three genotypic datasets (Supplementary Figure S2A and Supplementary Table S2).

In summary, 3VmrMLM detected the highest number of total QTNs and common QTNs in the six phenotypic datasets regardless of the genotypic dataset. The largest number of QTNs detected in multiple environments (three out of six phenotypic datasets) was identified using the RGAB genotypic dataset.

All significant QTNs were evaluated for consistency across multiple phenotypic datasets and models, and those detected in ≥ two datasets or ≥ two models were retained for further analysis. A total of 55, 80, and 32 QTNs were thus identified from the ALL, GB,

10.3389/fpls.2023.1229457

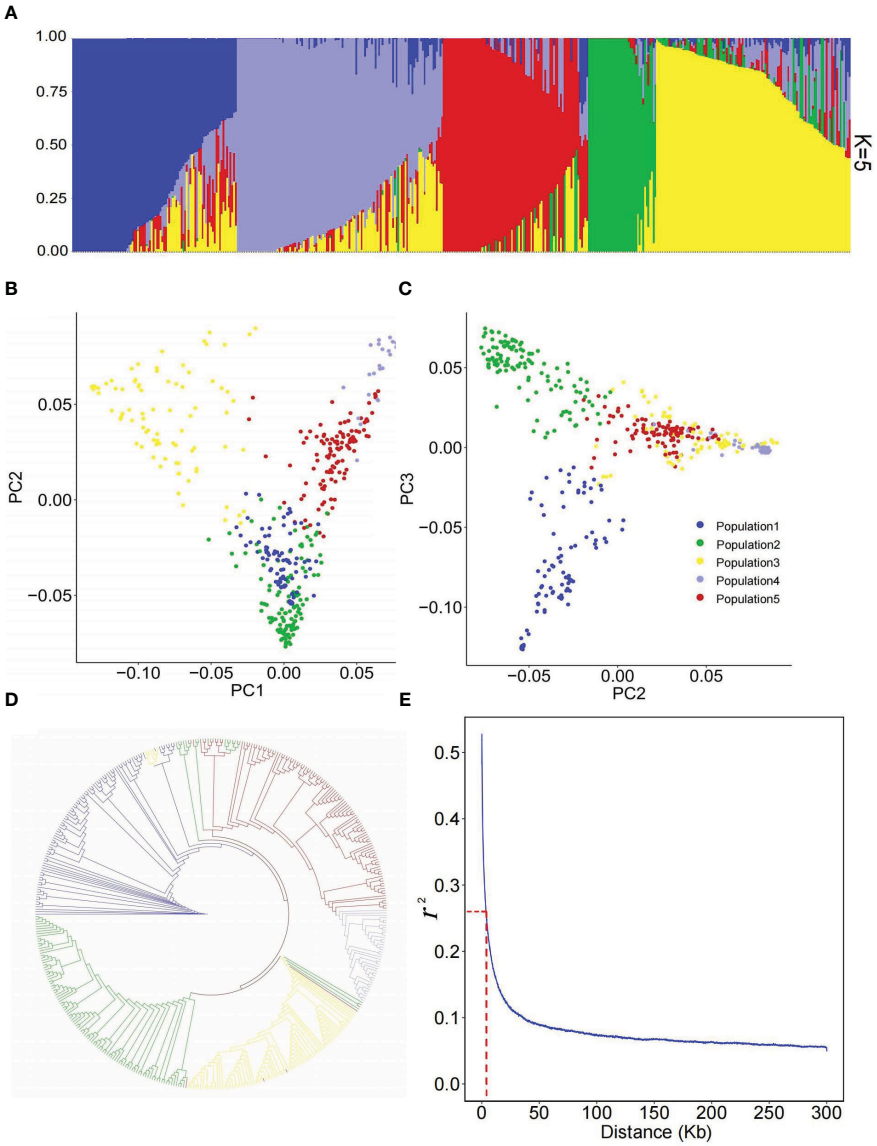


FIGURE 2
Population structure of 445 flax accessions. **(A)** Population structure estimated by ADMIXTURE. **(B, C)** Scatter plots of the first three principal components (PCs) of 445 flax accessions. **(D)** Phylogenetic analysis of 445 flax accessions based on 246,035 single nucleotide polymorphisms (SNPs). Accessions of clades one, two, three, four and five are indicated in blue, green, yellow, mauve and red, respectively. **(E)** Genome-wide LD decay analysis of the genetic panel.

and RGAB genotypic datasets, respectively (Supplementary Table S2). In agreement with the total number of QTNs detected, the majority of the retained QTNs were detected by 3VmrMLM across all three genotypic datasets, with 52 QTNs in ALL, 75 QTNs in GB, and 32 QTNs in RGAB (Table 1 and Supplementary Table S2). Allelic test of significance for these QTNs were performed using the Mann-Whitney U test for the dataset from which the QTNs were detected. A total of 82 non-significant QTNs (U test at the 5% probability level) were removed, leaving 132 significant QTNs used as tag QTNs in subsequent analyses (Figure 4 and Supplementary Tables S2, S3). The majority of the tag QTNs were detected by 3VmrMLM across all three genotypic datasets, with 41 in ALL, 62 in GB, and 30 in RGAB (Table 1). The R^2 values of the 132 tag QTNs ranged from 0.52% to 22.42% (Table 1 and Supplementary Table S3), and varied across the four models due to the differences in statistical models. For example, the R^2 of 3VmrMLM-detected tag QTNs in the ALL genotypic dataset ranged from 0.66% to 16.98%, while the R^2 of GEMMA-detected tag QTNs ranged from 1.11% to 10.00%. Similar results were observed in the GB and RGAB genotypic datasets (Table 1). Of note, eight tag QTNs were identified in both ALL and GB genotypic datasets, and explained 1.06% to 12.72% of the total variation for pasmo severity (Supplementary Table S3 and Supplementary Figure S2B). The position of all tag QTNs for pasmo severity are illustrated on a CIRCOS map (Figure 4). A total of eight tag QTNs were considered large-effect QTNs, i.e., $R^2 \geq 10\%$ (Table 2 and Supplementary Table S4). Based on these QTNs, significant negative correlations were observed between the number of favorable alleles (NFAs) in an

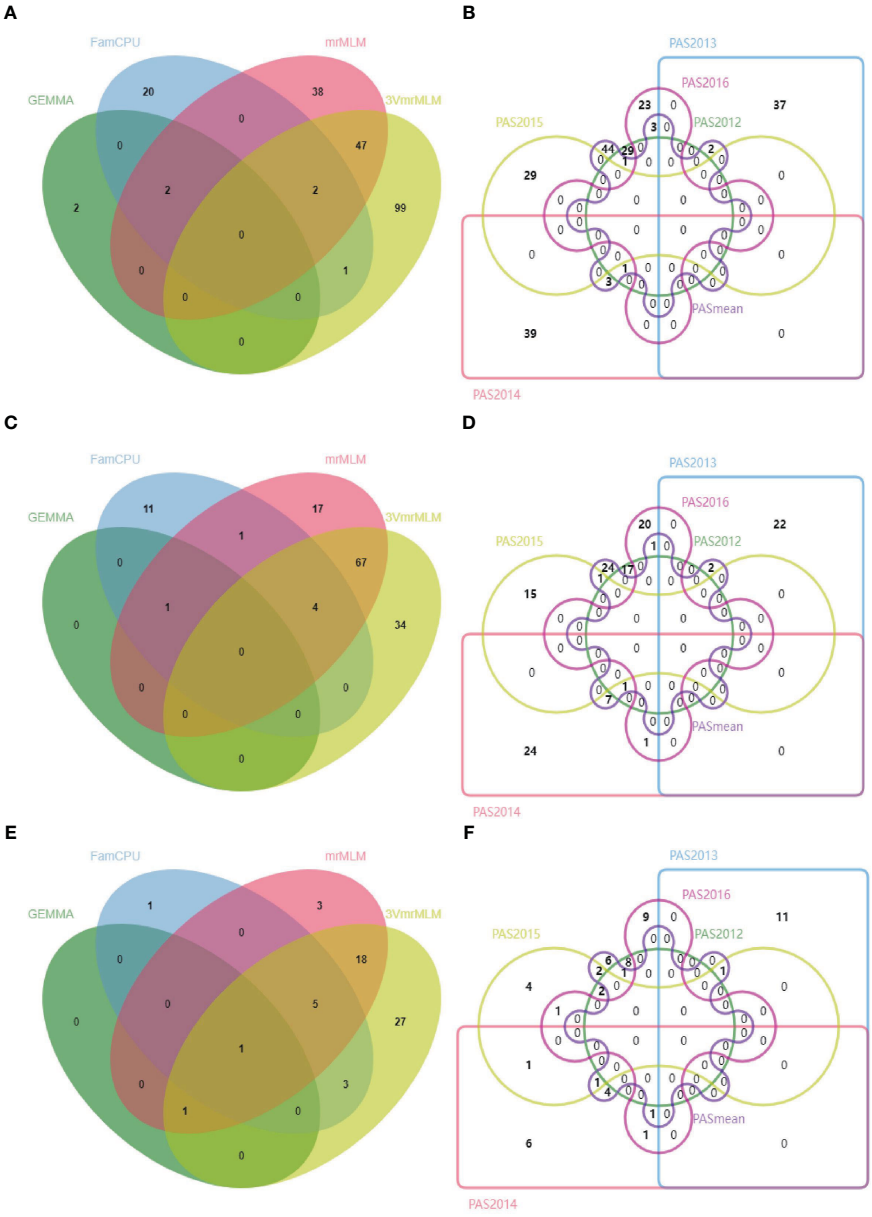


FIGURE 3
Venn diagrams of QTNs detected using four GWAS models (GEMMA, FarmCPU, mrMLM, and 3VmrMLM) for the three single nucleotide polymorphism (SNP) datasets: ALL **(A)**, GB **(C)**, and RGAB **(E)**, and QTNs detected using six different phenotypic datasets (PAS2012–PAS2016 and PASmean) for the three SNP datasets: ALL **(B)**, GB **(D)**, and RGAB **(F)**. ALL, all SNPs; GB, gene-based SNPs; RGAB, resistance gene analog (RGA) -based SNPs.

accession and the six pasmo severity datasets (PAS2012–PAS2016 and PASmean) ($r = -0.39 \sim -0.71$) (Supplementary Figure S3A–F), with the strongest correlation observed in the PASmean dataset ($r = -0.71$) (Supplementary Figure S3F).

Candidate genes for pasmo resistance

To identify the genes putatively involved in pasmo resistance in flax, we scanned resistance gene analogs (RGAs) within the estimated 4 kb flanking region of the QTNs identified from the ALL genotypic dataset, and identified the tag QTNs located within RGAs as candidate genes for the QTNs identified from the GB or RGAB genotypic dataset. The 37 RGAs that co-localized with 39 tag QTNs were considered candidates for pasmo resistance in flax (Supplementary Table S4). These RGAs were mainly classified into eight types, including receptor-like protein (RLP), receptor-like kinase (RLK), TIR-NBS-LRRs (TNL), TIR-unknown (TX), NBS-LRR (NL), TIR-NBS (TN), transmembrane-coiled coil protein (TM-CC), CC-NBS-LRR (CNL), and others. The majority of these RGAs were RLK (19) followed by TM-CC (5) (Figure 5).

Out of the 132 tag QTNs, QTN-Lu10-11656889 was identified by four models from the RGAB genotypic dataset, and explained 22.42% of the total variation. This QTN was located within the *NL* gene

TABLE 1 Comparison of quantitative trait nucleotide (QTN) identification for different GWAS models and genotypic datasets.

Statistical model	Genotypic dataset	NO. of detected QTNs	NO. of common QTNs by models or datasets	NO. of non–significant QTNs	NO. of tag QTNs	R^2 range (%)
GEMMA	ALL	4	2	0	2	1.11–10.00
FarmCPU	ALL	25	6	0	6	1.11–12.11
mrMLM	ALL	89	51	10	41	0.66–12.72
3VmrMLM	ALL	149	52	12	41	0.66–16.98
GEMMA	GB	1	1	0	1	1.11
FarmCPU	GB	17	8	1	7	1.11–13.30
mrMLM	GB	90	74	12	62	0.68–20.54
3VmrMLM	GB	105	75	13	62	0.68–20.54
GEMMA	RGAB	2	2	0	2	9.34–22.42
FarmCPU	RGAB	10	9	2	7	0.54–22.42
mrMLM	RGAB	28	25	4	23	0.52–17.40
3VmrMLM	RGAB	55	32	3	30	0.52–17.40

ALL, all SNPs: GB, gene SNPs: RGAB, resistance gene analog (RGA) based SNPs.

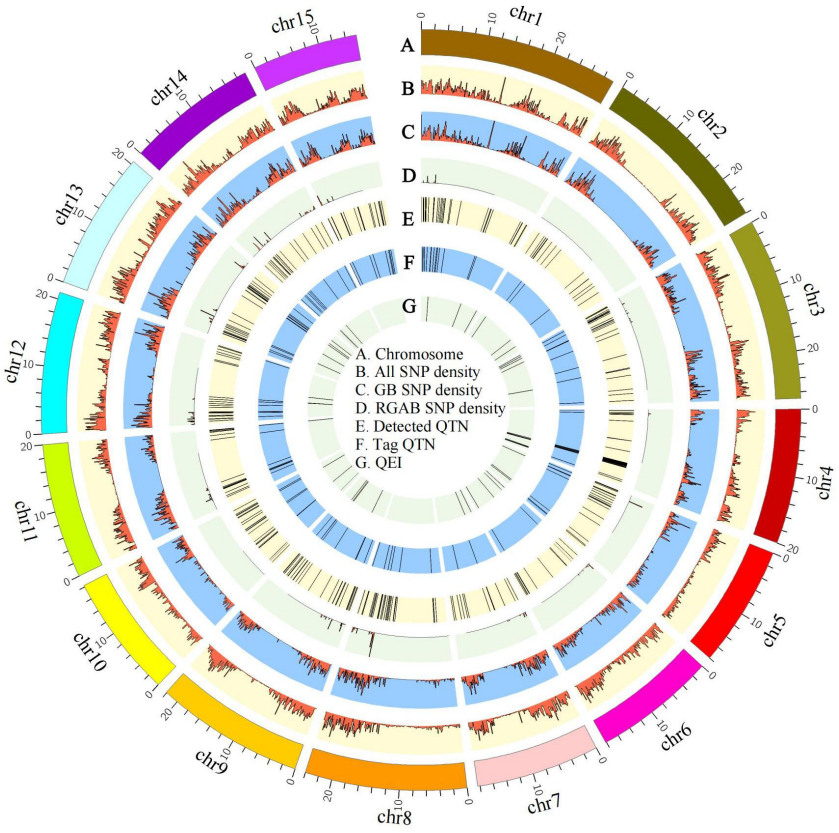


FIGURE 4
Circos map of quantitative trait nucleotides (QTNs) and QTN-by-environment interactions (QEIs) for pasmo severity in flax. Track A: 15 flax chromosomes. Track B: Heatmap of SNP density with bin sizes of 0.1 Mb for the ALL dataset (246,035 SNPs). Track C: Heatmap of SNP density with bin size of 0.1 Mb for the GB dataset (65,147 SNPs). Track D: Heatmap of SNP density with bin size of 0.1 Mb for the RGAB dataset (3,510 SNPs). Track E: QTNs detected using four statistical models: GEMMA, FarmCPU, mrMLM, and 3VmrMLM. Track F: QTNs identified using all four statistical models. Track G: QEIs detected using the 3VmrMLM model. ALL, all SNPs; GB, gene-based SNPs; RGAB, resistance gene analog (RGA)-based SNPs.

10.3389/fpls.2023.1229457

TABLE 2 Large-effect quantitative trait nucleotides (QTNs) and QTN-by-environment interactions (QEIs) detected in two genotypic datasets.

GD	R^2 (%)	QTN/QEI	Chr	Pos	Gene ID	Annotation
RGAB	10.79	QTN-Lu4-14335180	4	14335180	*Lus10041466*	TM-CC
RGAB	27.34	QEI-Lu5-1569144	5	1569144	*Lus10004719*	TNL
RGAB	16.77	QTN-Lu5-1715943	5	1715943	*Lus10008486*	RLK
RGAB	13.34	QTN-Lu5-15543693	5	15543693	*Lus10024053*	TM-CC
RGAB	11.88	QEI-Lu5-15543693	5	15543693	*Lus10024053*	TM-CC
RGAB	10.07	QTN-Lu10-11256857	10	11256857	*Lus10032735*	RLK
RGAB	22.42	QTN-Lu10-11656889	10	11656889	*Lus10032759*	NL
RGAB	17.40	QTN-Lu10-11657307	10	11657307	*Lus10032759*	NL
RGAB	15.77	QTN-Lu12-5214501	12	5214501	*Lus10018309*	TN
GB	13.77	QTN-Lu14-2333894	14	2333894	*Lus10025565*	TM-CC

GD, genotypic dataset: Chr, chromosome: Pos, position: TM-CC, transmembrane coiled-coil protein: TNL, TIR-NBS-LRRs: RLK, receptor-like protein kinase: NL, NBS-LRR. GB, gene-based SNPs: RGAB, resistance gene analog (RGA)-based SNPs.

Lus10032759 (Supplementary Figure S4A and Supplementary Table S4) which had four haplotypes Hap1 (AAAA, n = 336), Hap2 (TTAA, n = 18), Hap3 (TTGG, n = 89), and Hap4 (AAGG, n = 2) (Figure 6A). Significant differences in pasmo severity were observed between accessions with the Hap1 and Hap3 in all six phenotypic datasets, with accessions carrying Hap3 exhibiting lower pasmo severity than those carrying Hap1 (Figure 6A). QTN-Lu5-1715943 also had a relatively large effect (R^2 = 16.77%) in the RGAB genotypic dataset. The candidate gene for this QTN was the RLK-type RGA *Lus10008486* (Supplementary Figure S4B and Supplementary Table S4). The accessions with Hap2 (TTGG, n = 83) showed significantly lower pasmo severity than those with Hap1 (TTAA, n = 333), Hap3 (GGGG, n = 26), and Hap4 (GGAA, n = 3), again in almost all six phenotypic datasets (Figure 6B) In addition, the TM-CC type RGA *Lus10025565*, identified by the QTN-Lu14-2333894, also had a relatively large effect (R^2 = 13.77%), as detected from the GB genotypic dataset (Supplementary Figure S4C and Supplementary Table S4). The pasmo severity of accessions with Hap2 (CCAA, n = 283) was significantly different from those with other two haplotypes, with lower pasmo severity observed in Hap2 accessions than in Hap1 (CCCC, n = 125) and Hap3 (TTAA, n = 37) accessions (Figure 6C).

QEI detection and candidate genes

Using the 3VmrMLM model, a total of 50 QEIs underlying pasmo resistance in flax were identified from the ALL, GB, and RGAB genotypic datasets across the five individual year phenotypic datasets (PAS2012–PAS2016), as shown in Figures 4, 7A–C, and

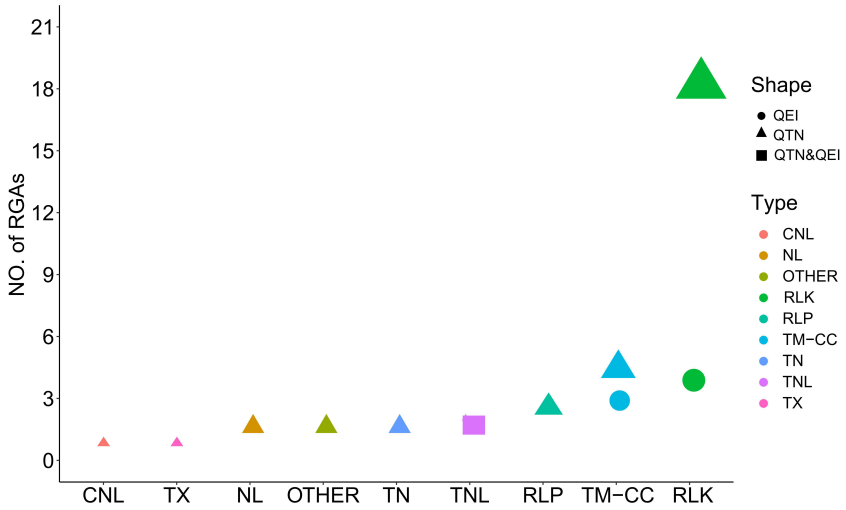


FIGURE 5
Distribution of candidate resistance gene analogs (RGAs) associated with tag quantitative trait nucleotides (QTNs) and QTN-by-environment interactions (QEIs). RLP, receptor like protein: RLK, receptor like kinase: CNL, CC-NBS-LRR: TNL, TIR-NBS-LRRs: TX, TIR-unknown: NL, NBS-LRR: TN, TIR-NBS: TM-CC,transmembrane-coiled coil protein.

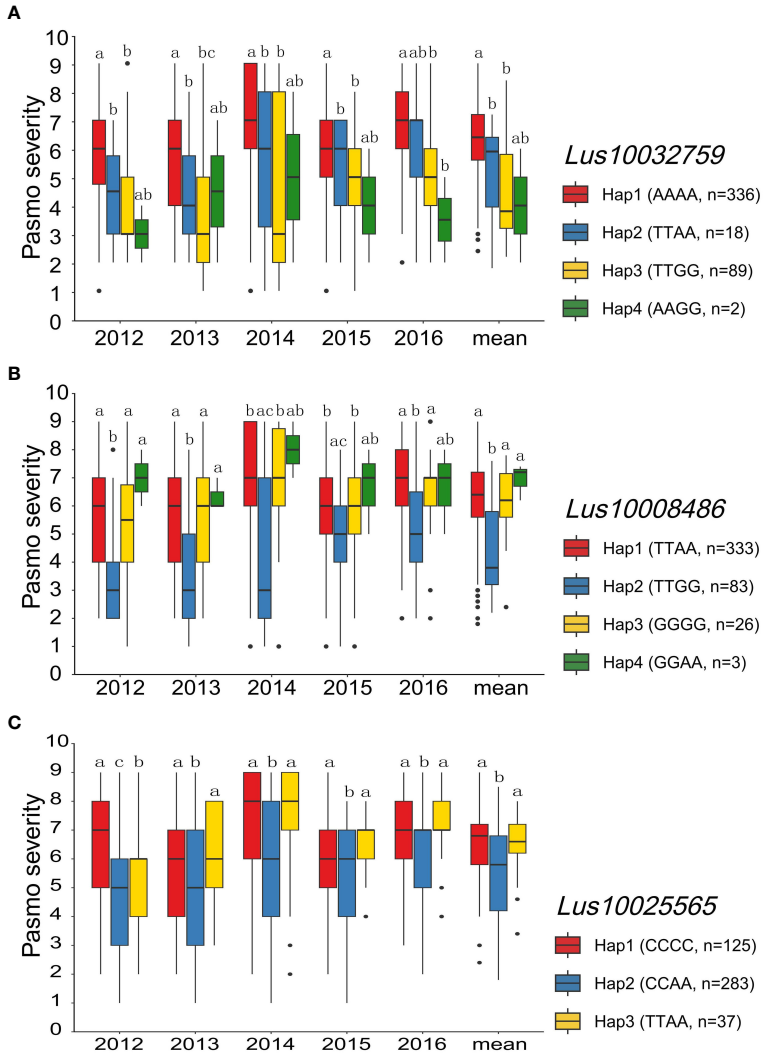


FIGURE 6
Analyses of the candidate genes *Lus10032759*, *Lus10008486* and *Lus10025565* for pasmo resistance for the five individual years and the mean over years. **(A)** Haplotype and pasmo severity analysis of Lus10032759 in 445 flax accessions. **(B)** Haplotype and pasmo severity analysis of *Lus10008486* in 445 flax accessions. **(C)** Haplotype and pasmo severity analysis of *Lus10025565* in 445 flax accessions. Letters indicate significant differences at the 5% probability level.

Supplementary Table S5. Overall, 27, 18, and nine QEIs were identified from the ALL, GB, and RGAB genotypic datasets, respectively. Four of these QEIs were detected in both the ALL and GB genotypic datasets: QEI-Lu1-3346281, QEI-Lu3-4320878, QEI-Lu4-14847340, and QEI-Lu9-17104439. Notably, no QEI loci for pasmo resistance were detected on chromosomes 8 and 15 (Supplementary Table S5).

The following four QEIs located on genes and detected from the GB or RGAB dataset were also identified as tag QTNs: QEI-Lu5-15543693 (R^2 = 11.88%), QEI-Lu11-19819154 (R^2 = 5.10%), QEI-Lu14-2333894 (R^2 = 6.01%), and QEI-Lu14-1935665 (R^2 = 2.85%) (Supplementary Table S2, S5 and Supplementary Figure S5).

The nine RGAs predicted as candidate genes for ten QEIs were further analyzed (Supplementary Table S6 and Figure 5). The TM-CC type RGA *Lus10024053* was the candidate gene for the large-effect QEI-Lu5-15543693, with Hap1 (GGAA, n = 301), Hap2 (GGTT, n = 9), Hap3 (AATT, n = 54), and Hap4 (AAAA, n = 81). The severity of pasmo infection in accessions with Hap4 was significantly lower than that of accessions with the other three haplotypes in the PAS2012, PAS2013, PAS2014, and PAS2016 datasets (Figure 8A; Supplementary Figure S4D; Supplementary Table S6). Additionally, the RLK type RGA *Lus10025492* was identified as the candidate gene of QEI-Lu14-1935665, with Hap1 (AAAA, n = 53), Hap2 (AAGG, n = 269), Hap3 (CCGG, n = 122), and Hap4 (CCAA, n = 1). A significantly lower pasmo severity of Hap2 was observed in PAS2013, PAS2014, and PAS2016 compared to Hap3 (Figure 8B; Supplementary Figure S4E; Supplementary Table S6). Similarly, the RLK RGA *Lus10040160* was identified as the candidate gene of QEI-Lu7-4573781. *Lus10040160* hasHap1 (TTTT, n = 271), Hap2 (GGTT, n = 88), and Hap3 (TTCC, n = 86), and significant differences in pasmo severity were observed between the Hap1 and Hap3 in the PAS2013, PAS2014, and PAS2016 datasets. The pasmo resistance level of accessions with Hap3 was significantly higher than that of accessions with Hap1 in those years (Figure 8C; Supplementary Figure S4F; Supplementary Table S6).

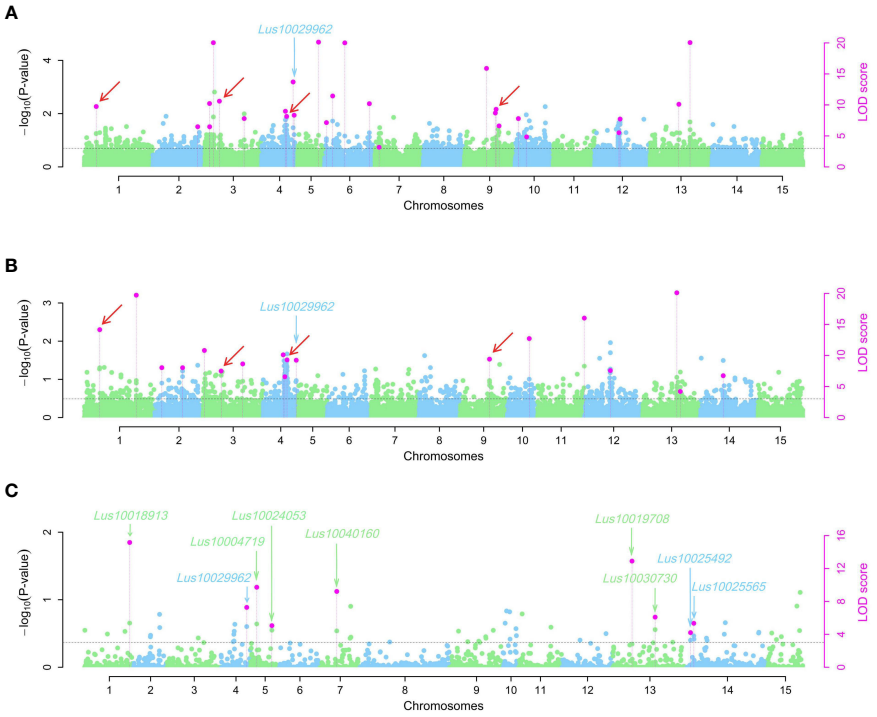


FIGURE 7
Manhattan plots for pasmo resistance associated QTN-by-environment interactions (QEIs) identified using the 3VmrMLM model for three single nucleotide polymorphisms (SNPs) datasets: ALL **(A)**, GB **(B)**, and RGAB **(C)**. Black horizontal lines in the Manhattan plots represent the genome-wide significant threshold. The red arrows indicate the QEIs co-detected in ALL **(A)** and GB **(B)** SNP datasets. The green and blue arrows indicate the candidate genes detected in ALL, GB, and RGAB SNP datasets. ALL, all SNPs; GB, gene SNPs; RGAB, resistance gene analog (RGA)-based SNPs.

Discussion

Comparison across GWAS models

The detection of QTNs in GWAS can vary depending on the statistical algorithms implemented in the models. In this study, three genotypic datasets (ALL, GB, and RGAB) were evaluated across six phenotypic datasets for pasmo resistance. The results showed that the 3VmrMLM model detected the most QTNs, followed by mrMLM and GEMMA. Most of the QTNs detected by at least two models were identified by 3VmrMLM. These findings support previous studies indicating that multi-locus models outperform single-locus models in QTN detection, and suggest that 3VmrMLM high statistical power and low false positive rate are advantageous (Cui et al., 2018; Hou et al., 2018; Zhong et al., 2021; He et al., 2022; Li et al., 2022b; Liu et al., 2022; Yu et al., 2022; Zhang et al., 2022).

After removing non-significant QTNs, the most tag QTNs were also identified by 3VmrMLM, followed by mrMLM and FarmCPU. The largest R^2 ranges were also observed in 3VmrMLM identified tag QTNs in all four models used, indicating its ability to identify tag QTNs with either large or small effects. Taken together, the 3VmrMLM model seems a good alternative to other single-locus and multi-locus models in GWAS. The 3VmrMLM model was developed to effectively detect main-effect QTNs, QEIs, and QQIs while providing unbiased estimates of their effects through an analysis of variance (ANOVA) model. This model builds upon the framework of compressed variance component mixed model (Li et al., 2022a) and presents technical improvements. One key reason for the superior performance of the 3VmrMLM model is its ability to consider all genetic effects in the mixed genetic model while simultaneously controlling for all polygenic backgrounds (Li et al., 2022a; Li et al., 2022b).

Evaluation of QTNs associated with pasmo resistance

Flax pasmo resistance is a quantitative trait, characterized by features of quantitative genetics. The challenge of visually measuring the resistance prompted us to adopt the pasmo severity scale (0–9) as a means to assess the severity of pasmo disease symptoms in our experimental genotypes. This severity scale provides a practical and standardized approach for quantitatively representing pasmo disease symptoms, despite its categorical appearance in scoring pasmo resistance. By utilizing this scale, we were able to capture the gradation in the expression of the trait among different genotypes, enabling a more comprehensive evaluation of the potential genetic factors influencing pasmo severity. Notably, this method has been commonly used for evaluating powdery mildew resistance in flax (You et al., 2022).

Using the multiple years' flax pasmo severity data, a total of the 132 tag QTNs were detected in this study, out of which 29 were previously reported in a study of the flax core collection consisting of

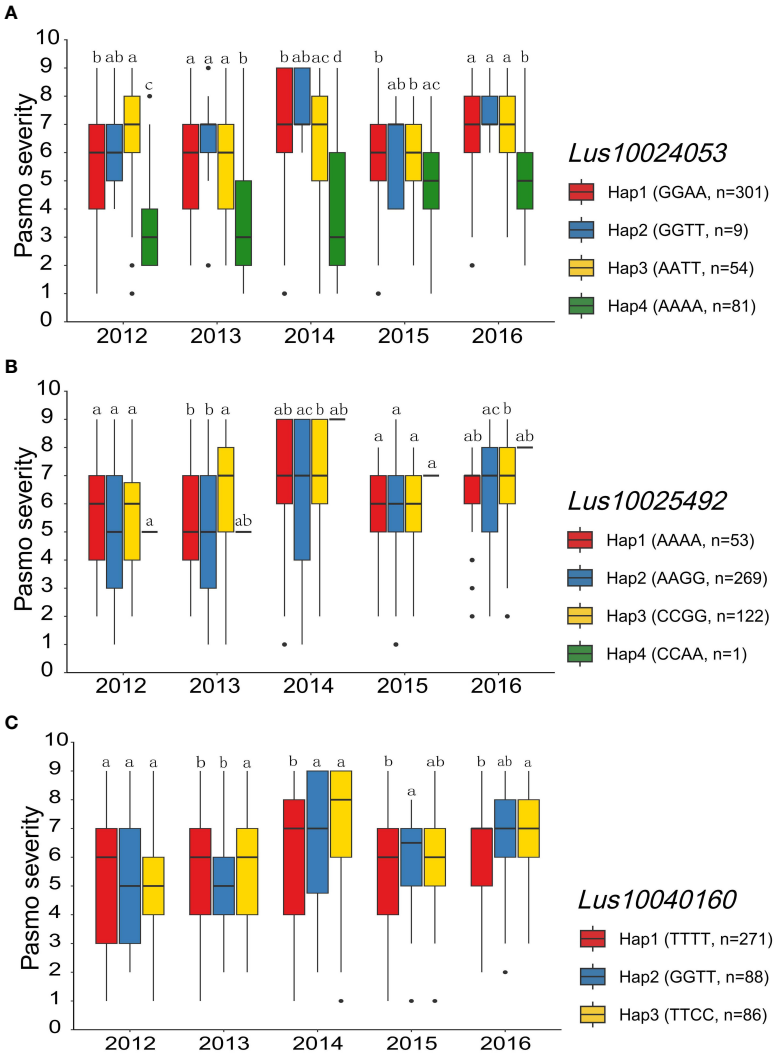


FIGURE 8
Analyses of the candidate gene *Lus10024053*, *Lus10025492* and *Lus10040160* for pasmo resistance associated QTN-by-environment interactions (QEIs) for the five individual years. **(A)** Box plot of pasmo severity of Lus10024053 haplotypes. **(B)** Box plot of pasmo severity of *Lus10025492* haplotypes. **(C)** Box plot of pasmo severity of *Lus10040160* haplotypes. Letters indicate significant differences at the 5% probability level.

370 accessions that utilized the same phenotyping method (He et al., 2018). In the aforementioned study, which focused on the 370 flax accessions, a subset of the current study, a total of 67 QTLs with large effects were identified by GWAS using various models, including GLM, MLM, FarmCPU, GEMMA, mrMLM, FASTmrEMMA, ISIS EM-BLASSO, pLARmEB, pKWmEB and FASTmrMLM models (He et al., 2018). Furthermore, four tag QTNs (QTN-Lu8-17271798, QTN-Lu13-2007925, QTN-Lu15-974597, and QTN-Lu13-14282050) were found to be situated within 1.01–16.97 kb upstream/downstream of QTLs previously reported in He et al. (2018) (Supplementary Table S3). To identify novel QTNs and their corresponding candidate genes associated with pasmo resistance in flax, multi-model and multi-environment GWAS were conducted using the ALL, GB, and RGAB genotypic datasets. A total of 31 (ALL), 49 (GB), and 27 (RGAB) novel tag QTNs were identified using 445 flax accessions (370 core accessions and 75 SBLs), which is an improvement compared to our previous study. Eight tag QTNs (R^2 = 1.11%–12.72%) were identified in both the ALL and GB datasets. Additionally, one and seven out of eight large-effect QTNs ($R^2 \geq 10.00\%$) were identified from the GB and RGAB datasets respectively (Table 2 and Supplementary Table S3). Among the tag QTNs with the top five R^2 (16.98%–22.42%), two, two and one tag QTNs were identified from the GB, RGAB, and ALL datasets, respectively (Supplementary Table S3). These results are consistent with previous studies suggesting that using gene-based or RGA-based SNPs for GWAS is beneficial for detecting QTNs with large effects and predicting key candidate genes (Huang et al., 2011; Zhu et al., 2018; Deng et al., 2020; You et al., 2022; Zhang et al., 2022). Therefore, the use of gene-based or RGA-based SNPs for GWAS is a powerful and efficient approach for identifying QTNs with large and small effects.

Candidate genes associated with pasmo resistance and their effects on main-effect QTNs and QEIs

Main-effect QTNs are QTNs with stable effects across different environments, while QEIs represent loci that may be effective only in some environments. Given the needs of global climate change and phenotypic plasticity research, QEIs have the potential to be exploited to dissect complex traits in future GWAS. In this study, candidate gene prediction of QTNs and QEIs was based on well-characterized RGAs in flax. RGAs have been identified as key candidate genes underlying plant disease resistance in several studies (Kassa et al., 2017; He et al., 2018; Fu et al., 2020; You et al., 2022). A total of 37 RGAs were identified as potential candidate genes of 39 tag QTNs and nine as candidates for ten QEIs. They were summarized into RLK, TM-CC, and NBS-LRR type RGAs. In general, the *RLK*, *TM-CC*, and *NBS-LRR* genes account for a large proportion of *R* genes, playing important roles in plant disease resistance against fungal pathogens. Well-known examples include wheat leaf rust resistance conferred by the *Lr21* (*NBS-LRR*) gene (Huang et al., 2003), resistance to the hemi-biotrophic fungus *Phytophthora infestans* conferred by the potato *R7* (*NBS-LRR*) gene (Leister et al., 1996; Hammond-Kosack and Jones, 1997), broad-spectrum mildew resistance conferred by the Arabidopsis *RPW8* (*TM-CC*) gene (Xiao et al., 2001), and rice blast resistance conferred by the *Pi-d2* (*RLK*) gene (Chen et al., 2006). The RLK, TM-CC, and NBS-LRR type RGAs associated with pasmo resistance in this study may contribute to a better understanding of the genetic mechanisms underlying pasmo resistance in flax. Furthermore, the molecular mechanisms of these candidate genes warrant further validation.

Breeding applications of pasmo resistance associated QTNs

The present study revealed significant differences in pasmo resistance levels between linseed, fibre accessions, and SBLs within a flax genetic panel. Interestingly, 75 SBLs exhibited higher pasmo resistance levels than the flax core collection, which included 370 accessions (Figure 1C). Moreover, the number of favorable alleles (NFA) in fibre accessions was greater than in linseed accessions, and fibre accessions with more favorable alleles were found to be more resistant to pasmo than linseed accessions (Supplementary Figure S6), as demonstrated in a previous study (He et al., 2018). Flax have obtained commercial importance due to the utilization of the stem for high quality fiber (Oomah, 2001; You et al., 2019; Rahman and Hoque, 2023). One of the major objectives in the fiber flax breeding program is to improve fiber yield and quality (Galinousky et al., 2020; Rahman and Hoque, 2023). The productivity of fiber flax is severely affected by devastating fungal disease pasmo, which causes yield loss and fiber quality reduction (Yadav et al., 2022). Therefore, the 75 SBLs represent valuable genetic resources for improving pasmo resistance in elite varieties through direct hybridization.

Negative correlations were observed between the NFA and pasmo resistance of the five-year pasmo severity (PAS2012–PAS2016) and PASmean datasets in Supplementary Figure S3A–F (r = −0.39 ~ −0.71), with the highest correlation found in the PASmean dataset (r = −0.71). This additive effect of identified tag QTNs suggests that accessions carrying more favorable alleles are suitable for high pasmo resistance breeding through the pyramiding of loci. For example, SBL 8031 had 17 favorable alleles (PASmean = 2.2), SBL 8040 had 17 favorable alleles (PASmean = 2.4), and SBL 8032 had 18 favorable alleles (PASmean = 2.4).

Although large-effect tag QTNs, such as QTN-Lu10-11656889 (R^2 = 22.42%) and QTN-Lu12-2992110 (R^2 = 16.68%), may be available for improving pasmo resistance through marker-assisted selection (MAS), several tag QTNs with small effects would be better captured through genomic prediction/selection with the aim to transform flax breeding from a slow and labor-intensive mode into an efficient and accurate one. The breeding values of complex traits, such as pasmo resistance, are predicted by cross-validated models, which are an alternative strategy to MAS (Lipka et al., 2015; Poland and Rutkoski, 2016; He et al., 2019; You et al., 2022). Marker-assisted backcrossing and genomic selection/prediction strategies have already significantly enhanced disease resistance in many crops (Buerstmayr et al., 2008; Buerstmayr et al., 2009; Poland and Rutkoski, 2016; Crossa et al., 2017; He et al., 2019; Xu et al., 2021).

The QEI loci identified in this study constitute an alternative genetic information for improving flax pasmo disease, specifically to cope with environmental changes. These QEI loci can be useful for predicting the performance of flax varieties in specific environments. By identifying specific genetic markers associated with QEI loci, breeders can develop flax varieties that are better adapted to specific environmental conditions. The combined utilization of pasmo resistance-associated QTNs and QEIs holds the promise of driving the molecular breeding of flax with broad-spectrum and durable resistance against *Septoria linicola*.

Conclusion

Our study demonstrates that pasmo resistance in flax is a complex trait, controlled by multiple genes, and influenced by gene-environment interactions. The 3VmrMLM model, which detected more QTNs and QEIs, is a promising alternative to other multi-locus GWAS models. Gene-based and RGA-based SNPs as genotypic datasets in GWAS proved to be efficient for identifying QTNs with both large and small effects and predicting candidate genes. Our research identified 372 significant QTNs and 50 QEIs, providing potential targets for improving pasmo resistance in flax breeding programs. Furthermore, we identified 37 RGAs for 39 tag QTNs and nine RGAs for ten QEIs, suggesting the potential involvement of RLK, TM-CC, and NBS-LRR genes in pasmo resistance. Our findings on gene–environment interactions can guide breeding strategies that account for environmental factors. The 50 QEI loci identified in our study can help improve our understanding of the genetic mechanisms involved in pasmo

resistance and its interactions with environmental factors, ultimately leading to the development of more resilient and better adapted flax varieties. Our study has important implications for the sustainable production of flax and provides valuable information for developing improved flax varieties with enhanced pasmo resistance, which is critical for ensuring the long-term viability of this important oil and fiber crop. The large-effect QTNs and candidate genes identified in this study can be used as molecular markers for marker-assisted selection in future studies to accelerate the breeding process for pasmo-resistant flax varieties.

Data availability statement

The raw sequence data for flax core collection presented in this study are deposited in the NCBI repository (project number: PRJNA707038).

Author contributions

FY, SC and LH conceived and designed this research project. KR produced the breeding lines and provided all phenotypic data. LH, YS, YC and HW undertook the analysis of all available data. LH and YS contributed to the writing of the original draft. FY, SC, and LH discussed the results, guided the entire study, participated in data analysis, and revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by Genome Canada and other industrial stakeholders for the Total Utilization Flax GENomics (TUFGEN) project, Agriculture Development Fund for Diverse Field Crop Cluster project in genomics and molecular markers to identify resistance genes in flax, Hainan Provincial Natural Science Foundation of China (No. 323RC422 and No. 321RC1148), and the Hainan University Startup Fund (KYQD(ZR)-21027).

Acknowledgments

Thanks to Tara Edwards for editing the early version of the manuscript.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1229457/full#supplementary-material

References

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19 (9), 1655–1664. doi: 10.1101/gr.094052.109

Ayliffe, M. A., Collins, N. C., Ellis, J. G., and Pryor, A. (2000). The maize rp1rust resistance gene identifies homologues in barley that have been subjected to diversifying selection. *Theor. Appl. Genet.* 100), 1144–1154. doi: 10.1007/s001220051398

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models usinglme4. *J. Stat. Software* 67 (1), 1–48. doi: 10.18637/jss.v067.i01

Browning, S. R., and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* 81 (5), 1084–1097. doi: 10.1086/521987

Brueggeman, R., Rostoks, N., Kudrna, D., Kilian, A., Han, F., Chen, J., et al. (2002). The barley stem rust-resistance gene Rpg1 is a novel disease-resistance gene with homology to receptor kinases. *Proc. Natl. Acad. Sci. U.S.A.* 99 (14), 9328–9333. doi: 10.1073/pnas.142284999

Buerstmayr, H., Ban, T., and Anderson, J. (2008). QTL mapping and marker assisted selection for Fusarium head blight resistance in wheat. *Cereal Res. Commun.* 36 (Supplement 6), 1–3. doi: 10.1556/CRC.36.2008.Suppl.B.1

Buerstmayr, H., Ban, T., and Anderson, J. (2009). QTL mapping and marker-assisted selection for Fusarium head blight resistance in wheat: a review. *Plant Breed.* 128 (1), 1–26. doi: 10.1111/j.1439-0523.2008.01550.x

Buschges, R., Hollricher, K., Panstruga, R., Simons, G., Wolter, M., Frijters, A., et al. (1997). The barley Mlo gene: a novel control element of plant pathogen resistance. *Cell* 88 (5), 695–705. doi: 10.1016/s0092-8674(00)81912-1

Chen, W., Gao, Y., Xie, W., Gong, L., Lu, K., Wang, W., et al. (2014). Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nat. Genet.* 46 (7), 714–721. doi: 10.1038/ng.3007

Chen, X., Shang, J., Chen, D., Lei, C., Zou, Y., Zhai, W., et al. (2006). A B-lectin receptor kinase gene conferring rice blast resistance. *Plant J.* 46 (5), 794–804. doi: 10.1111/j.1365-313X.2006.02739.x

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 6 (2), 80–92. doi: 10.4161/fly.19695

Collins, N., Drake, J., Ayliffe, M., Sun, Q., Ellis, J., Hulbert, S., et al. (1999). Molecular characterization of the maize Rp1-D rust resistance haplotype and its mutants. *Plant Cell* 11 (7), 1365–1376. doi: 10.1105/tpc.11.7.1365

Crossa, J., Perez-Rodriguez, P., Cuevas, J., Montesinos-Lopez, O., Jarquin, D., de Los Campos, G., et al. (2017). Genomic selection in plant breeding: Methods, models, and perspectives. *Trends Plant Sci.* 22 (11), 961–975. doi: 10.1016/j.tplants.2017.08.011

Cui, Y., Zhang, F., and Zhou, Y. (2018). The application of multi-locus GWAS for the detection of salt-tolerance loci in rice. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01464

Deng, Y., He, T., Fang, R., Li, S., Cao, H., and Cui, Y. (2020). Genome-wide gene-based multi-trait analysis. *Front. Genet.* 11. doi: 10.3389/fgene.2020.00437

Deng, Y., Zhai, K., Xie, Z., Yang, D., Zhu, X., Liu, J., et al. (2017). Epigenetic regulation of antagonistic receptors confers rice blast resistance with yield balance. *Science* 355 (6328), 962–965. doi: 10.1126/science.aai8898

Diederichsen, A., Kusters, P. M., Kessler, D., Bainas, Z., and Gugel, R. K. (2012). Assembling a core collection from the flax world collection maintained by Plant Gene Resources of Canada. *Genet. Resour. Crop Evol.* 60 (4), 1479–1485. doi: 10.1007/s10722-012-9936-1

Dinh, H. X., Singh, D., de la Cruz Gomez, D., Hensel, G., Kumlehn, J., Mascher, M., et al. (2022). The barley leaf rust resistance gene *Rph3* encodes a predicted membrane protein and is induced upon infection by avirulent pathotypes of Puccinia hordei. *Nat. Commun.* 13 (1), 2386. doi: 10.1038/s41467-022-29840-1

Dong, S. S., He, W. M., Ji, J. J., Zhang, C., Guo, Y., and Yang, T. L. (2021). LDBlockShow: a fast and convenient tool for visualizing linkage disequilibrium and haplotype blocks based on variant call format files. *Brief Bioinform.* 22 (4), bbaa227. doi: 10.1093/bib/bbaa227

Elhaik, E. (2022). Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. *Sci. Rep.* 12 (1), 14683. doi: 10.1038/s41598-022-14395-4

Fu, F., Zhang, X., Liu, F., Peng, G., Yu, F., and Fernando, D. (2020). Identification of resistance loci in Chinese and Canadian canola/rapeseed varieties against Leptosphaeria maculans based on genome-wide association studies. *BMC Genomics* 21 (1), 501. doi: 10.1186/s12864-020-06893-4

Galinousky, D., Mokshina, N., Padvitski, T., Ageeva, M., Bogdan, V., Kilchevsky, A., et al. (2020). The toolbox for fiber flax breeding: A pipeline from gene expression to fiber quality. *Front. Genet.* 11. doi: 10.3389/fgene.2020.589881

Hall, L., Booker, H., Siloto, R., Jhala, A., and Weselake, R. (2016). "Flax (Linum usitatissimum L.)," in *Industrial Oil Crops*, ACSO Press, Urbana, IL, 157–194.

Halley, S., Bradley, C. A., Lukach, J. R., McMullen, M., Knodel, J. J., Endres, G. J., et al. (2004). Distribution and severity of pasmo on flax in North Dakota and evaluation of fungicides and cultivars for management. *Plant Dis.* 88 (10), 1123–1126. doi: 10.1094/PDIS.2004.88.10.1123

Hammond-Kosack, K. E., and Jones, J. D. (1997). Plant disease resistance genes. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 48, 575–607. doi: 10.1146/annurev.arplant.48.1.575

He, L., Wang, H., Sui, Y., Miao, Y., Jin, C., and Luo, J. (2022). Genome-wide association studies of five free amino acid levels in rice. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.1048860

He, L., Xiao, J., Rashid, K. Y., Jia, G., Li, P., Yao, Z., et al. (2019). Evaluation of genomic prediction for pasmo resistance in flax. *Int. J. Mol. Sci.* 20 (2), 359. doi: 10.3390/ijms20020359

He, L., Xiao, J., Rashid, K. Y., Yao, Z., Li, P., Jia, G., et al. (2018). Genome-wide association studies for pasmo resistance in flax (*Linum usitatissimum* L.). *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01982

Hou, S., Zhu, G., Li, Y., Li, W., Fu, J., Niu, E., et al. (2018). Genome-wide association studies reveal genetic variation and candidate genes of drought stress related traits in cotton (*Gossypium hirsutum* L.). *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01276

Huang, L., Brooks, S. A., Li, W., Fellers, J. P., Trick, H. N., and Gill, B. S. (2003). Map-based cloning of leaf rust resistance gene *Lr21* from the large and polyploid genome of bread wheat. *Genetics* 164 (2), 655–664. doi: 10.1093/genetics/164.2.655

Huang, H., Chanda, P., Alonso, A., Bader, J. S., and Arking, D. E. (2011). Gene-based tests of association. *PloS Genet.* 7 (7), e1002177. doi: 10.1371/journal.pgen.1002177

Islam, T., Vera, C., Slaski, J., Mohr, R., Rashid, K. Y., Booker, H., et al. (2021). Fungicide management of pasmo disease of flax and sensitivity of septoria linicola to pyraclostrobin and fluxapyroxad. *Plant Dis.* 105 (6), 1677–1684. doi: 10.1094/PDIS-06-20-1175-RE

Jo, H., and Koh, G. (2015). Faster single-end alignment generation utilizing multi-thread for BWA. *BioMed. Mater Eng.* 26 Suppl 1, S1791–S1796. doi: 10.3233/BME-151480

Kassa, M. T., You, F. M., Hiebert, C. W., Pozniak, C. J., Fobert, P. R., Sharpe, A. G., et al. (2017). Highly predictive SNP markers for efficient selection of the wheat leaf rust resistance gene *Lr16*. *BMC Plant Biol.* 17 (1), 45. doi: 10.1186/s12870-017-0993-7

Kumar, S., Stecher, G., Peterson, D., and Tamura, K. (2012). MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. *Bioinformatics* 28 (20), 2685–2686. doi: 10.1093/bioinformatics/bts507

Leister, D., Ballvora, A., Salamini, F., and Gebhardt, C. (1996). A PCR-based approach for isolating pathogen resistance genes from potato with potential for wide application in plants. *Nat. Genet.* 14 (4), 421–429. doi: 10.1038/ng1296-421

Letunic, I., and Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* 49 (W1), W293–W296. doi: 10.1093/nar/gkab301

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25 (16), 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, M., Zhang, Y. W., Xiang, Y., Liu, M. H., and Zhang, Y. M. (2022a). IIIVmrMLM: The R and C++ tools associated with 3VmrMLM, a comprehensive GWAS method for dissecting quantitative traits. *Mol. Plant* 15 (8), 1251–1253. doi: 10.1016/j.molp.2022.06.002

Li, M., Zhang, Y. W., Zhang, Z. C., Xiang, Y., Liu, M. H., Zhou, Y. H., et al. (2022b). A compressed variance component mixed model for detecting QTNs and QTN-by-environment and QTN-by-QTN interactions in genome-wide association studies. *Mol. Plant* 15 (4), 630–650. doi: 10.1016/j.molp.2022.02.012

Lipka, A. E., Kandianis, C. B., Hudson, M. E., Yu, J., Drnevich, J., Bradbury, P. J., et al. (2015). From association to prediction: statistical methods for the dissection and selection of complex traits in plants. *Curr. Opin. Plant Biol.* 24, 110–118. doi: 10.1016/j.pbi.2015.02.010

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 (2), e1005767. doi: 10.1371/journal.pgen.1005767

Liu, J., Lin, Y., Chen, J., Yan, Q., Xue, C., Wu, R., et al. (2022). Genome-wide association studies provide genetic insights into natural variation of seed-size-related traits in mungbean. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.997988

Marone, D., Russo, M. A., Laido, G., De Vita, P., Papa, R., Blanco, A., et al. (2013). Genetic basis of qualitative and quantitative resistance to powdery mildew in wheat: from consensus regions to candidate genes. *BMC Genomics* 14, 562. doi: 10.1186/1471-2164-14-562

Oomah, B. D. (2001). Flaxseed as a functional food source. *J. Sci. Food Agric.* 81 (9), 889–894. doi: 10.1002/jsfa.898

Poland, J., and Rutkoski, J. (2016). Advances and challenges in genomic selection for disease resistance. *Annu. Rev. Phytopathol.* 54, 79–98. doi: 10.1146/annurev-phyto-080615-100056

Rahman, M., and Hoque, A. (2023). "Flax Breeding," in *The Flax Genome*, Springer, Cham, 55–68.

Segura, V., Vilhjalmsson, B. J., Platt, A., Korte, A., Seren, U., Long, Q., et al. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* 44 (7), 825–830. doi: 10.1038/ng.2314

Sekhwal, M. K., Li, P., Lam, I., Wang, X., Cloutier, S., and You, F. M. (2015). Disease resistance gene analogs (RGAs) in plants. *Int. J. Mol. Sci.* 16 (8), 19248–19290. doi: 10.3390/ijms160819248

Singh, K. K., Mridula, D., Rehal, J., and Barnwal, P. (2011). Flaxseed: a potential source of food, feed and fiber. *Crit. Rev. Food Sci. Nutr.* 51 (3), 210–222. doi: 10.1080/10408390903537241

Soto-Cerda, B. J., Aravena, G., and Cloutier, S. (2021). Genetic dissection of flowering time in flax (*Linum usitatissimum* L.) through single- and multi-locus genome-wide association studies. *Mol. Genet. Genomics* 296 (4), 877–891. doi: 10.1007/s00438-021-01785-y

Soto-Cerda, B. J., Diederichsen, A., Ragupathy, R., and Cloutier, S. (2013). Genetic characterization of a core collection of flax (*Linum usitatissimum* L.) suitable for association mapping studies and evidence of divergent selection between fiber and linseed types. *BMC Plant Biol.* 13, 78. doi: 10.1186/1471-2229-13-78

Wang, S. B., Feng, J. Y., Ren, W. L., Huang, B., Zhou, L., Wen, Y. J., et al. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* 6, 19444. doi: 10.1038/srep19444

Wang, J., and Zhang, Z. (2021). GAPIT version 3: Boosting power and accuracy for genomic association and prediction. *Genomics Proteomics Bioinf.* 19 (4), 629–640. doi: 10.1016/j.gpb.2021.08.005

Xiao, S., Ellwood, S., Calis, O., Patrick, E., Li, T., Coleman, M., et al. (2001). Broad-spectrum mildew resistance in *Arabidopsis thaliana* mediated by RPW8. *Science* 291 (5501), 118–120. doi: 10.1126/science.291.5501.118

Xu, Y., Ma, K., Zhao, Y., Wang, X., Zhou, K., Yu, G., et al. (2021). Genomic selection: A breakthrough technology in rice breeding. *Crop J.* 9 (3), 669–677. doi: 10.1016/j.cj.2021.03.008

Yadav, B., Kaur, V., Narayan, O. P., Yadav, S. K., Kumar, A., and Wankhede, D. P. (2022). Integrated omics approaches for flax improvement under abiotic and biotic stress: Current status and future prospects. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.931275

Yang, L., Zhang, X., Zhang, X., Wang, J., Luo, M., Yang, M., et al. (2017). Identification and evaluation of resistance to powdery mildew and yellow rust in a wheat mapping population. *PloS One* 12 (5), e0177905. doi: 10.1371/journal.pone.0177905

You, F. M., Cloutier, S., Rashid, K. Y., and Duguid, S. D. (2019). "Flax (*Linum usitatissimum* L.) Genomics and Breeding," in *Advances in Plant Breeding Strategies: Industrial and Food Crops*, Springer, Cham, 277–317.

You, F. M., Deal, K. R., Wang, J., Britton, M. T., Fass, J. N., Lin, D., et al. (2012). Genome-wide SNP discovery in walnut with an AGSNP pipeline updated for SNP discovery in allogamous organisms. *BMC Genomics* 13, 354. doi: 10.1186/1471-2164-13-354

You, F. M., Huo, N., Deal, K. R., Gu, Y. Q., Luo, M. C., McGuire, P. E., et al. (2011). Annotation-based genome-wide SNP discovery in the large and complex *Aegilops tauschii* genome using next-generation sequencing without a reference genome sequence. *BMC Genomics* 12, 59. doi: 10.1186/1471-2164-12-59

You, F. M., Jia, G., Xiao, J., Duguid, S. D., Rashid, K. Y., Booker, H. M., et al. (2017). Genetic variability of 27 traits in a core collection of flax (*Linum usitatissimum* L.). *Front. Plant Sci.* 8, 1636. doi: 10.3389/fpls.2017.01636

You, F. M., Rashid, K. Y., Zheng, C., Khan, N., Li, P., Xiao, J., et al. (2022). Insights into the genetic architecture and genomic prediction of powdery mildew resistance in flax (*Linum usitatissimum* L.). *Int. J. Mol. Sci.* 23 (9), 4960. doi: 10.3390/ijms23094960

You, F. M., Xiao, J., Li, P., Yao, Z., Jia, G., He, L., et al. (2018a). Genome-wide association study and selection signatures detect genomic regions associated with seed yield and oil quality in flax. *Int. J. Mol. Sci.* 19 (8), 2303. doi: 10.3390/ijms19082303

You, F. M., Xiao, J., Li, P., Yao, Z., Jia, G., He, L., et al. (2018b). Chromosome-scale pseudomolecules refined by optical, physical and genetic maps in flax. *Plant J.* 95 (2), 371–384. doi: 10.1111/tpj.13944

Yu, K., Miao, H., Liu, H., Zhou, J., Sui, M., Zhan, Y., et al. (2022). Genome-wide association studies reveal novel QTLs, QTL-by-environment interactions and their candidate genes for tocopherol content in soybean seed. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.1026581

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 (2), 203–208. doi: 10.1038/ng1702

Zhang, C., Dong, S. S., Xu, J. Y., He, W. M., and Yang, T. L. (2019a). PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* 35 (10), 1786–1788. doi: 10.1093/bioinformatics/bty875

Zhang, Y. M., Jia, Z., and Dunwell, J. M. (2019b). Editorial: The applications of new multi-locus GWAS methodologies in the genetic dissection of complex traits. *Front. Plant Sci.* 10, 100. doi: 10.3389/fpls.2019.00100

Zhang, Y. W., Tamba, C. L., Wen, Y. J., Li, P., Ren, W. L., Ni, Y. L., et al. (2020). mrMLM v4.0.2: An R platform for multi-locus genome-wide association studies. *Genomics Proteomics Bioinf.* 18 (4), 481–487. doi: 10.1016/j.gpb.2020.06.006

Zhang, F., Wang, C., Li, M., Cui, Y., Shi, Y., Wu, Z., et al. (2021). The landscape of gene-CDS-haplotype diversity in rice: Properties, population organization, footprints of domestication and breeding, and implications for genetic improvement. *Mol. Plant* 14 (5), 787–804. doi: 10.1016/j.molp.2021.02.003

Zhang, J., Wang, S., Wu, X., Han, L., Wang, Y., and Wen, Y. (2022). Identification of QTNs, QTN-by-environment interactions and genes for yield-related traits in rice using 3VmrMLM. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.995609

Zheng, C., Rashid, K. Y., Cloutier, S., and You, F. M. (2023). "QTL and candidate genes for flax disease resistance," in *The Flax Genome*, Springer, Cham, 121–148.

Zhong, H., Liu, S., Sun, T., Kong, W., Deng, X., Peng, Z., et al. (2021). Multi-locus genome-wide association studies for five yield-related traits in rice. *BMC Plant Biol.* 21 (1), 364. doi: 10.1186/s12870-021-03146-8

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44 (7), 821–824. doi: 10.1038/ng.2310

Zhu, W., Xu, C., Zhang, J. G., He, H., Wu, K. H., Zhang, L., et al. (2018). Gene-based GWAS analysis for consecutive studies of GEFOS. *Osteoporos Int.* 29 (12), 2645–2658. doi: 10.1007/s00198-018-4654-y

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 02 November 2023
DOI 10.3389/fpls.2023.1247181

OPEN ACCESS

EDITED BY
Shang-Qian Xie,
University of Idaho, United States

REVIEWED BY
Jia Wen,
University of North Carolina at Chapel Hill, United States
Suhong Bu,
South China Agricultural University, China
Shibo Wang,
University of California, Riverside, United States

*CORRESPONDENCE
Jianying Feng
fengjianying@njau.edu.cn

†These authors have contributed equally to this work

RECEIVED 25 June 2023
ACCEPTED 18 October 2023
PUBLISHED 02 November 2023

CITATION
Yang M, Wen Y, Zheng J, Zhang J, Zhao T and Feng J (2023) Improving power of genome-wide association studies via transforming ordinal phenotypes into continuous phenotypes.
Front. Plant Sci. 14:1247181.
doi: 10.3389/fpls.2023.1247181

Improving power of genome-wide association studies via transforming ordinal phenotypes into continuous phenotypes

Ming Yang[1†], Yangjun Wen[2†], Jinchang Zheng[1], Jin Zhang[2], Tuanjie Zhao[1] and Jianying Feng[1*]

[1]Key Laboratory of Biology and Genetics Improvement of Soybean, Ministry of Agriculture/Zhongshan Biological Breeding Laboratory (ZSBBL)/National Innovation Platform for Soybean Breeding and Industry-Education Integration/State Key Laboratory of Crop Genetics & Germplasm Enhancement and Utilization/College of Agriculture, Nanjing Agricultural University, Nanjing, China, [2]College of Science, Nanjing Agricultural University, Nanjing, China

Introduction: Ordinal traits are important complex traits in crops, while genome-wide association study (GWAS) is a widely-used method in their gene mining. Presently, GWAS of continuous quantitative traits (C-GWAS) and single-locus association analysis method of ordinal traits are the main methods used for ordinal traits. However, the detection power of these two methods is low.

Methods: To address this issue, we proposed a new method, named MTOTC, in which hierarchical data of ordinal traits are transformed into continuous phenotypic data (CPData).

Results: Then, FASTmrMLM, one C-GWAS method, was used to conduct GWAS for CPData. The results from the simulation studies showed that, MTOTC +FASTmrMLM for ordinal traits was better than the classical methods when there were four and fewer hierarchical levels. In addition, when MTOTC was combined with FASTmrEMMA, mrMLM, ISIS EM-BLASSO, pLARmEB, and pKWmEB, relatively high power and low false positive rate in QTN detection were observed as well. Subsequently, MTOTC was applied to analyze the hierarchical data of soybean salt-alkali tolerance. It was revealed that more significant QTNs were detected when MTOTC was combined with any of the above six C-GWAs.

Discussion: Accordingly, the new method increases the choices of the GWAS methods for ordinal traits and helps to mine the genes for ordinal traits in resource populations.

KEYWORDS

ordinal trait, genome-wide association study, salt-alkali tolerance, soybean, hierarchical data

1 Introduction

The hierarchical data (HData), phenotypic data for ordinal traits, is commonly used to describe many important traits in crop germplasm resources. This includes count data for quantitative traits and hierarchical data for resistance traits, such as the number of main stem nodes (Chang et al., 2018), the number of branches (Shim et al., 2019), and disease resistance (Megerssa et al., 2020). Ordinal traits are important in crop breeding and have a considerable impact on crop yield and quality. Genome-wide association studies (GWAS) for ordinal traits can further promote the mining of relevant excellent genes, which plays a key role in molecular design breeding and gene cloning. Cuevas et al. (2018) divided the degree of infection of anthracnose-inoculated sorghum leaves into five levels and identified three loci for anthracnose resistance in chromosome 5 using the GWAS methods. Chang et al. (2018) detected three loci significantly associated with "the number of nodes on the main stem" in 368 soybean cultivars with 62,423 SNPs. Meanwhile, Shim et al. (2019) identified five quantitative trait nucleotides (QTNs) for soybean branch number via GWAS and linkage analysis and mined a candidate gene *Glyma.06g210600*.

Ordinal traits are discrete traits that are controlled by multiple genes. However, their phenotypic data is hierarchical and non-continuous and contains relatively limited information; accordingly, GWAS for ordinal traits is more complex than that for continuous quantitative traits. The threshold model represents a reasonable method for the genetic analysis of ordinal traits, and most association mapping methods are developed under this framework (Xu et al., 2005; Osval et al., 2015). Generalized linear model is based on the threshold model and link phenotypic data with latent variables through a link function. They are widely used for genetic analysis of ordinal traits and can deal with non-normal data (Feng et al., 2013; Song et al., 2016; Wang et al., 2018). The logistic regression model is another classical way for dealing with association studies of ordinal traits (Tan et al., 2007; Hoggart et al., 2008; Wu et al., 2009; Jiang et al., 2021). When sample size is limited, the application of a set-valued (SV) system model can improve the statistical power and the accuracy of parameter estimation (Bi et al., 2015). Bayesian and maximum likelihood methods are both widely used for parameter estimation in GWAS (Xu et al., 2005; Hoggart et al., 2008; Wang et al., 2018), while several studies have also employed non-parametric methods for association analysis of ordinal traits (Sun et al., 2016; Wang et al., 2017; He and Kulminski, 2020). However, most of them were either single-locus or were only suitable for the analysis of binary traits, and they had very few applications in crop. GWAS for continuous quantitative traits and single-locus methods are currently the main methods used for association analysis of ordinal traits; however, both have low power in QTN detection.

Accordingly, in this study, we proposed a method for transforming ordinal phenotypes into continuous phenotypes (MTOTC). First, the hierarchical phenotypic data for ordinal traits (HData) was transformed into continuous phenotypic data (CPData). Subsequently, FASTmrMLM (Tamba and Zhang, 2018), one GWAS method suitable for continuous quantitative traits, was used to perform GWAS for CPData. In Monte Carlo simulation studies, we validated the feasibility of the new method through the statistical power, false-positive rate in QTN detection and the accuracies for the estimates of QTN effects and positions, and obtained the number of hierarchical levels suitable for MTOTC +FASTmrMLM. The new method was validated by re-analyzing the salt-alkali resistance traits in soybean germplasm resource population of Zhang et al. (2014) and Zhou et al. (2015). This study provides more choices for association analysis of ordinal traits and helps to identify excellent genes for important complex traits in crops.

2 Theory and methods

Here we proposed a method, named MTOTC, to transform the discrete hierarchical data (HData) of ordinal traits into continuous phenotypic data. Then, GWAS for continuous quantitative traits (C-GWAS) are used to analyze the transformed continuous phenotypic data. The new method was described as below.

2.1 Genetic mapping population

In Monte Carlo simulation studies, 199 *Arabidopsis thaliana* lines harboring 10,000 SNPs with a minimum allele frequency >0.1 (Atwell et al., 2010) were selected as the genetic mapping population. For real data analysis, the population was comprised of 286 soybean cultivars assessed for salt-alkali tolerance, the phenotypic data consisted of the main root length index in 2009 and 2010 (Zhang et al., 2014), and the marker data were 54,296 high-quality SNP markers present in Zhou et al. (2015).

2.2 Method for transforming ordinal phenotypes into continuous phenotypes

To transform ordinal phenotypes into continuous phenotypes, we proposed the MTOTC method. In detail, the Chi-square test and logistic regression were used to initially select the SNPs that were significantly related to the trait. Subsequently, these significant SNPs and ordinal phenotypes were used to construct a multi-locus model, Bayesian method was used to estimate the SNP effects, and the effect estimates were used to predict the continuous phenotypic data (CPData). This is MTOTC. Then, the

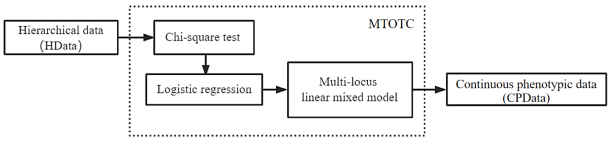


FIGURE 1
Technology framework of the MTOTC method in this study.

predicted CPData is analyzed by C-GWAS methods, such as FASTmrMLM (Figure 1).

2.2.1 The Chi-square test and logistic regression

The Chi-square test in R 4.0.5 (function "chisq.test") was used to scan the SNPs in the whole genome using a single marker method (*P*-value ≤0.05). To further improve the quality of the significant correlated SNPs in the initial screening for reducing interference and improving detection accuracy, logistic regression was used as a secondary SNP screening method. Logistic regression was performed using function "glm" (2 hierarchical levels) and "polr" (the number of hierarchical levels greater than 2) with a *P*-value ≤0.05. The aim of this step was to further eliminate SNPs that were not associated with the traits for simplifying the iterations in the following multi-locus genetic model.

2.2.2 Multi-locus genetic model

Based on the potentially associated markers identified in the above-described initial screening, a multi-locus model was established to transform ordinal phenotypes into continuous phenotypes. The linear model is expressed as:

$$y = W\alpha + \sum_{i=1}^{q} X_i\beta_i + \epsilon \quad (1)$$

where y represents $n \times 1$ ordinal phenotype vector, with n representing sample size; $W = (w_1, w_2, \ldots, w_c)$ represents $n \times c$ matrix of covariates (fixed effects), including a column vector of **1** and population structure, and represents $c \times 1$vector of fixed effects, including intercept; $X_i$and represent respectively $n \times 1$genotype vector and effect of the i-th potential associated SNP; q represents the number of SNPs selected in the initial screening step; $\epsilon \sim \mathrm{MVN}_n(0, \sigma_e^2 \mathrm{I}_n)$ represents $n \times 1$error vector.

The population structure **Q** matrix used in the linear model was calculated using Structure software (Pritchard et al., 2000). Based on the **Q** matrix, the population is divided into corresponding subgroups, and the optimal subgroup number **K** value is determined according to the corresponding standard, yielding the final **Q** matrix. The optimal value of the *Arabidopsis* population structure was calculated as **K**=2, and the optimal value of the salt-alkali tolerant soybean population structure in the actual study was **K**=3.

2.2.3 Parameter estimation

In the second step of the novel method, a multi-locus linear mixed model for transforming ordinal phenotypes into continuous phenotypes was established, based on the empirical Bayesian algorithm (Xu, 2010). And significant loci were screened in threshold value LOD=3.0.

In model (1), set β_i to obey the following prior normal distribution:

$$P(\beta_i|\sigma_i^2) = N(0|\sigma_i^2)$$

$$P(\sigma_i^2|\tau, \omega) \propto (\sigma_i^2)^{-\frac{1}{2}(\tau+2)} \times exp\left(-\frac{\omega}{2\sigma_i^2}\right)$$

The parameters were estimated using empirical Bayes, as follows, and the Newton–Raphson method.

$$\sigma_i^2 = \frac{E(\beta_i^T\beta_i) + \omega}{\tau + 3}$$

$$\alpha = (W^T V^{-1} W)^{-} W^T V^{-1} \mathrm{y}$$

$$\sigma_e^2 = \frac{1}{n}(y - W\alpha)^T \left(y - W\alpha - \sum_{i=1}^{q} X_i E(\beta_i)\right)$$

$$E(\beta_i) = \sigma_i^2 X_i^T V^{-1}(y - W\alpha)$$

Among them,

$$E(\beta_i^T\beta_i) = E(\beta_i^T)E(\beta_i) + tr[Var(\beta_i)]$$

$$Var(\beta_i) = I\sigma_i^2 - \sigma_i^2 X_i^T V^{-1} X_i \sigma_i^2$$

$$(\tau, \omega) = (0, 0)$$

$$V = \sum_{i=1}^{q} X_i X_i^T \sigma_i^2 + \mathrm{I}\sigma_e^2$$

Then, the empirical Bayesian estimates of these SNPs effects were obtained in the multi-locus model (1) based on the selected significant SNP markers and ordinal phenotype, and estimates of these effect were used to predict the phenotype, obtaining the continuous phenotypic data (CPData) of ordinal trait.

2.3 GWAS with MTOTC method for ordinal trait

When continuous phenotypic data was obtained by the above MTOTC method, a C-GWAS method could be used to detect significant loci. In this work, FASTmrMLM, one C-GWAS method, was used. So loci significantly associated with ordinal traits were detected by FASTmrMLM using the obtained continuous phenotypic data and the potential associated markers identified in the above-described initial screening. The GWAS method is henceforth referred to as MTOTC+FASTmrMLM. Moreover, the effects of five other C-GWAS (FASTmrEMMA, mrMLM, ISIS EM-BLASSO, pLARmEB, and pKWmEB) methods are also discussed based on the MTOTC method for ordinal trait, in order to verify the feasibility of MTOTC.

2.4 Monte Carlo simulation datasets for ordinal trait

We conducted six simulation studies to evaluate the feasibility of the new method. For each study, the loci 278, 2143, 2054, 3698, 1716, 6178, and 8501, located on chromosomes 1, 2, 2, 2, 1, 4, and 5, respectively, were selected as the causal loci related to the simulated trait. There were three types of phenotypic data in the simulation experiment—original data (OData), which were continuous and

generated by Monte Carlo simulation; HData, which were generated from the above OData according to specific distribution proportions (i.e., classification proportion of phenotype distribution); and CPData, which were generated from the above HData by MTOTC. Then, FASTmrMLM, one multi-locus C-GWAS algorithm, was used to conduct GWAS for CPData.

3 Results

3.1 Monte Carlo simulation studies

3.1.1 Threshold value in the initial screening

To determine the most suitable threshold value for the Chi-square test and logistic regression in the initial screening, four probability thresholds (0.0001 [i.e., 1/SNP number], 0.01, 0.05, and 0.10) were set for the Chi-square test in the first simulation study, while three probability thresholds (0.0001 [i.e., 1/SNP number], 0.01, and 0.05) were set for logistic regression. The Chi-square test can eliminate a large number of SNPs that are not significantly related to a given phenotype. However, the simulation study showed that some SNPs screened in the above Chi-square test (those with a *P*-value >0.98 and an unusually large absolute value of effect estimate in logistic regression) were not truly related to the phenotype and interfered greatly with subsequent association analysis. Therefore, to further improve the quality of the screened significantly related SNPs and detection accuracy, logistic regression was used as a secondary screening method for SNPs in MTOTC.

In the Chi-square test, the single-locus retention rate decreased with decreasing *P*-values (i.e., threshold values) (Figure 2A). For instance, the single-locus retention rate at loci 278 and 2143 with *P*-values of 0.05 and 0.10 was as high as 96.62%~99.68%, which are very close. When the *P*-value was 0.01, the single-locus retention rate began to decrease, and when the *P*-value was 0.0001, the retention rate dropped to between 59.06% and 68.45%. Moreover, the total retention rate (i.e., the proportion of retained loci among the total loci after chi-square test screening) was the lowest when the *P*-value was 0.0001, followed by 0.01, 0.05, and 0.10 (Figure 3A).

In logistic regression after the Chi-square test, the single-locus retention rate was the highest when the *P*-value was 0.05 (Figure 2B). For instance, the retention rates of loci 278 and 2143 were as high as 97.56%~99.68% when the *P*-value was 0.01 or 0.05; when the *P*-value was 0.0001, the retention rate dropped to between 60.51% and 69.88%. Additionally, the total retention rate was the lowest (only 0.22%) when the *P*-value was 0.0001, followed by 0.01 and 0.05 (Figure 3B).

Owing to too low single-locus retention rate at the *P*-values of 0.01 and 0.0001, the two *P*-values were unsuitable as a threshold for initial screening. Although the total retention rate was high when the *P*-value was 0.10, this *P*-value retains more loci that are not associated with the trait, in which it did not contribute to simplifying the model. Therefore, the probability threshold *P*=0.05, which is commonly used in statistics, was selected as the probability threshold for the Chi-square test and logistic regression of the initial screening in this study. In addition, we also investigated the effect of threshold value on the single-locus retention rate and the total retention rate under different proportions distribution in binary data and the similar results were observed.

3.1.2 MTOTC+FASTmrMLM displayed greater power than other classical mapping methods

In Monte Carlo simulation studies, the GWAS results of hierarchical data using MTOTC+ FASTmrMLM were compared with those using two classical mapping methods (Chi-square test and logistics regression) (Table 1). The results showed that these methods had greater power at the three loci 278, 2143, and 3698, but had less power (<10%) at the other four loci. Compared with the two classical mapping methods, MTOTC+FASTmrMLM had higher power at the three loci 278, 2143, and 3698, and lower false-positive rate, when the number of hierarchical levels of HData was ≤4. The power of the classical methods was higher in a few instances, it was less than 1.5-fold that of MTOTC+FASTmrMLM, but their false-positive rates were 6.8–9.5-fold higher than that of MTOTC+FASTmrMLM. In addition, the results showed that when the number of hierarchical levels was<5, MTOTC+FASTmrMLM was more suitable for HData analysis as compared with FASTmrMLM alone. Moreover, in Table 1, MTOTC +FASTmrMLM had a relatively higher F1 score, especially for binary data (HData with two hierarchical levels). Here the F1

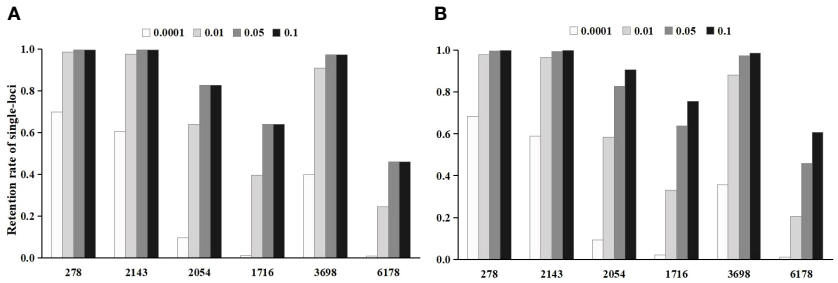


FIGURE 2
The effect of threshold value on the single-locus retention rate after the initial screening. **(A)** is the single-locus retention rate after chi-square test screening; **(B)** is the single-locus retention rate after logistic regression screening.

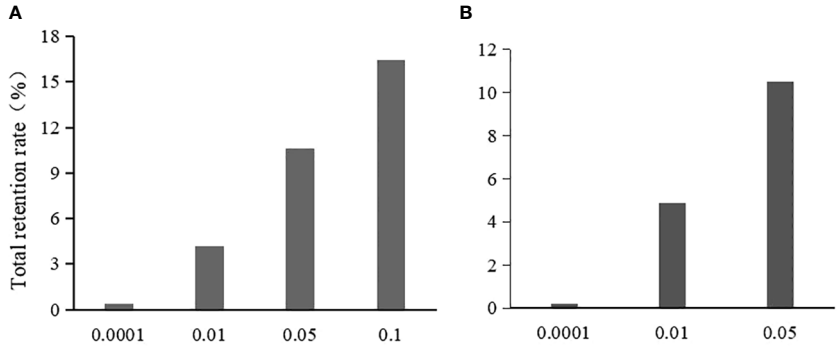


FIGURE 3
The effect of threshold value on the total retention rate after the initial screening. **(A)** is the total retention rate in chi-square test; **(B)** is the total retention rate in logistic regression.

score combines the precision and recall, it is used to effectively measure the accuracy of the statistical methods and balance power and FPR. Therefore, MTOTC is recommended for the analysis of HData under four or fewer hierarchical levels.

3.1.3 The effect of the number of hierarchical levels on the new method

The third simulation study investigated the effect of the number of hierarchical levels on MTOTC. Based on symmetrical distribution, the number of hierarchical levels was set to 2, 3, 4, and 5, respectively, and the number of replicates was 10,000. Meanwhile, we compared the results of OData, HData and CPData using FASTmrMLM.

Compared with CPData from the other hierarchical levels, the distribution of CPData2 (i.e., the CPData converted from the HData of 2 hierarchical levels by MTOTC) was closer to the original data (OData). First, the frequency distribution of the CPData was closer to that of the OData when the hierarchical level was low, especially when it was equal to 2 (Figure 4). As the number of hierarchical levels increased, the peak of CPData began to shift to the right and was far from the peak of the OData, which was expected to affect the GWAS results. The frequency distribution of the OData and the

TABLE 1 Comparison of different genome-wide association study methods.

Hierarchical number	Locus		Chi-square test	logistic regression	FASTmrMLM	MTOTC+ FASTmrMLM
2	Power(%)	278	66.20	22.50	41.85	57.38
		2143	57.70	19.40	28.62	55.98
		3698	18.40	10.10	9.89	22.76
	Mean of Power (%)		20.87	7.60	13.84	19.54
	FPR (‰)		7.27	0.07	0.44	0.77
	F1 score		0.04	0.13	0.16	0.17
3	Power(%)	278	62.00	71.30	53.41	70.87
		2143	56.40	57.20	45.76	66.64
		3698	20.00	26.90	19.66	36.82
	Mean of Power (%)		20.47	23.30	22.06	26.53
	FPR (‰)		6.15	4.77	0.45	0.70
	F1 score		0.04	0.06	0.24	0.24
4	Power(%)	278	68.40	80.30	65.70	75.98
		2143	58.50	65.40	58.71	71.83
		3698	20.80	37.30	27.76	45.69
	Mean of Power (%)		21.77	27.37	28.29	28.53
	FPR (‰)		8.11	5.90	0.45	0.63
	F1 score		0.03	0.06	0.29	0.26

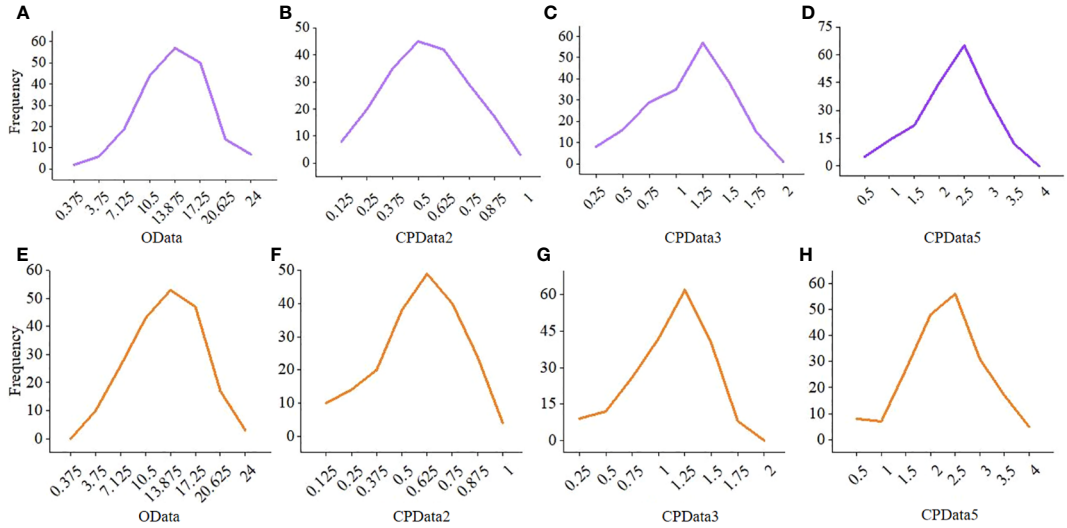


FIGURE 4
The frequency distribution of the OData and the corresponding CPData for different hierarchical levels in the 10th and 613th repetition. **(A–D)** is the 10th repetition, **(E–H)** is the 613th repetition. CPData2 transformed from HData of two hierarchical levels by MTOTC; CPData3 transformed from HData of three hierarchical levels by MTOTC; CPData5 transformed from HData of five hierarchical levels by MTOTC.

corresponding CPData with different hierarchical levels in the 10th and 613th replicates, randomly selected out of the 10,000 replicates using the uniformly distributed random number generator in R, is shown in Figure 4. Second, the range of the coefficient of variation (*CV*) of the OData was between 29.5% and 55.5%. Among the 10,000 replicates, the number of replicates beyond the *CV* range of the OData (4.09%, 18.94%, 21.47%, and 25.37% of CPData2, CPData3, CPData4, and CPData5, respectively) also increased with increasing hierarchical level. Thus, the *CV* range of CPData2 was the closest to that of the OData. Third, among the 10,000 replicates, the skewness range between the CPData and the OData was the closest at 2 hierarchical levels. Among them, the skewness range of the OData was between −1.00 and 0.46 and the range of CPData2 was between −1.28 and 0.35. As the number of hierarchical levels increased, the skewness of the CPData gradually deviated from that of the OData; the kurtosis showed the same tendency as the skewness.

MTOTC performed well for the estimates of QTN position under different numbers of hierarchical levels. The position estimates via MTOTC+FASTmrMLM (i.e., the position estimates of the CPData via FASTmrMLM) were unbiased at loci 278, 2143, and 3698 (Supplementary Table 1). Although the position estimates at loci 2054 and 8501 in CPData2, and at loci 1716 and 6178 in all the CPData were biased, the relative mean absolute deviations of their position estimates were all less than 8.96E-05. The accuracy of the estimates of QTN positions for ordinal traits was significantly improved by MTOTC when the number of hierarchical levels was less than 5, i.e., the estimates of QTN positions for the CPData were better than those for the HData when FASTmrMLM was used (Supplementary Table 1).

The effect of MTOTC on the relative power at loci 278, 2143, and 3698 was the greatest when the number of hierarchical levels is equal to 2 (Supplementary Figure 1). Here, "the effect of MTOTC on the relative power" refers to the increment of the relative power of CPData compared to the relative power of HData. The relative power of the CPData (50%~100%) was significantly higher than that of the HData (22%~88%) and was relatively closer to the power of the OData. When the number of the hierarchical levels of the CPData was less than or equal to 5, the relative power exhibited an increasing trend with increasing the number of hierarchical levels and was significantly superior to that of the HData.

The false-positive rates of CPData2, CPData3, CPData4, and CPData5 via MTOTC+FASTmrMLM were 0.77‰, 0.70‰, 0.63‰, and 0.55‰, respectively.

3.1.4 The effect of the number of replicates on the new method

The fourth simulation study assessed the impact of the number of replicates on the estimates of QTN effects and positions, relative power, and false-positive rate using MTOTC+FASTmrMLM. Based on the results of CPData2 (1:1), CPData3 (1:3:1), and CPData5 (1:2:4:2:1), 10 replicates were set at equal intervals from 1,000 to 10,000. As a result, the results across various numbers of replicates at each locus and for each hierarchical levels (CPData2, CPData3, and CPData5) were insignificant (Figure 5). This indicated that the number of replicates did not affect the power, false-positive rate, and the estimates of QTN effects and positions. Therefore, 1,000 replicates were used in subsequent simulation studies.

3.1.5 The effect of distribution proportion skewness on the new method

In the fifth simulation study, we investigated the effect of distribution proportion skewness on the new method under three hierarchical levels. Here the distribution proportion skewness were set as symmetrical distribution (distribution proportion, 1:2:1), uniform distribution (1:1:1), and skewed distribution (4:2:1). The

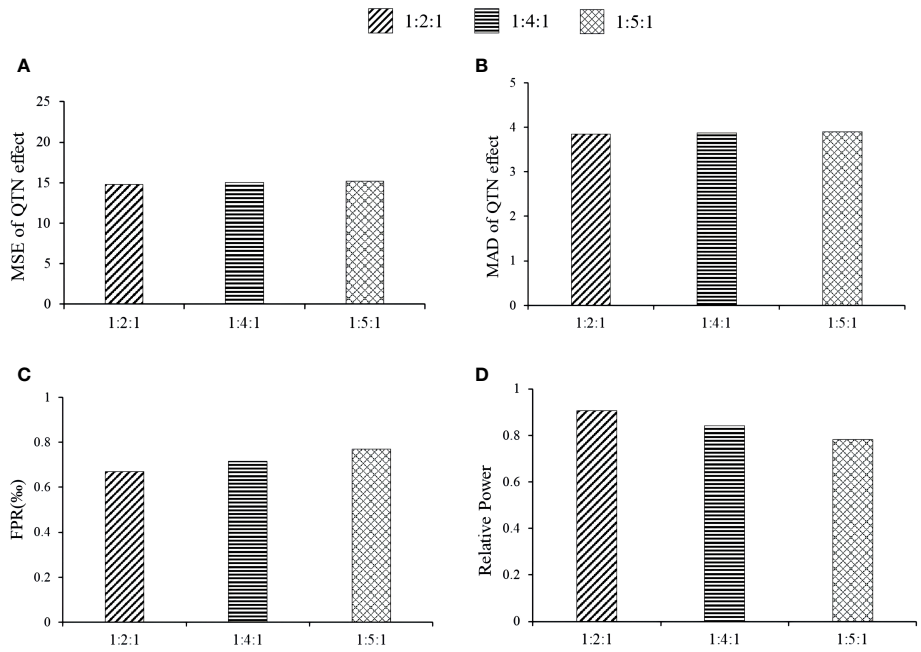


FIGURE 5
The impact of repetition number of simulation experiment on the association analysis results of CPData (2143 Locus). **(A, B)** MSE and MAD of QTN effect at 2143, respectively; **(C)** false-positive rates; **(D)** relative power.

indicators were the relative power, false-positive rate, the estimates of QTN effects and positions. The skewed distribution had the lowest relative power at loci 278, 2143, and 3698, followed by the uniform distribution, and the symmetrical distribution (Supplementary Figure 2). The MAD and mean squared error (MSE) of QTN position estimates showed unbiasedness under the three distribution proportion skewness. The skewed distribution (7.09‰) was slightly higher false-positive rate than symmetrical distribution (6.71‰) and uniform distribution (6.82‰). When the kurtosis values of the three distributions for the CPData and the OData were compared, it was found that the steepness of the CPData under 1:2:1 was closer to that of the OData (the kurtosis values for the OData, 1:2:1 CPData, 1:1:1 CPData, and 4:2:1 CPData ranged from 2.163–5.415, 1.963–5.412, 1.958–5.196, and 1.980–3.830, respectively). The CPData under 1:2:1 and 1:1:1 and the OData were relatively close in terms of skewness (the skewness of OData, 1:2:1 CPData, 1:1:1 CPData, and 4:2:1 CPData were in the range of −1.001~0.462, −1.466~0.319, −1.256~0.282, and −0.812~0.777, respectively). The skewness of the CPData under 4:2:1 and the OData differed markedly. Therefore, the accuracy of symmetric distribution via MTOTC+FASTmrMLM was higher than that of uniform distribution and skewed distribution.

3.1.6 The effect of distribution proportion kurtosis on the new method

Here we studied the effect of distribution proportion kurtosis on the new method. The proportions were set as 1:2:1, 1:4:1, and 1:5:1. The association detection results of the 1:2:1 proportion had the best, e.g., the relative powers of the 1:2:1 proportion at loci 2143, 278, 3698, and 1716 via MTOTC+FASTmrMLM was better than those under others distribution proportion (Figure 6A). The MSE and MAD of effect estimates at locus 278, 2143, and 3698 were lower at 1:2:1 than at 1:4:1 and 1:5:1; however, the differences were insignificant (Figures 6B, C), while the trends at the other loci were unclear. Under the three distribution proportions, the MSE and MAD of QTN position estimates were all unbiased at loci 278, 2143, 2054, and 3698. However, a lower false-positive rate was observed with the 1:2:1 distribution proportion (Figure 6D). Moreover, the steepness of the CPData under distribution proportion 1:2:1 was closer to that of the OData (the kurtosis values of the OData, 1:2:1 CPData, 1:4:1 CPData, and 1:5:1 CPData were 2.163~5.415, 1.963~5.412, 1.967~7.343, and 1.974~7.920, respectively). The skewness showed the same tendency as the kurtosis (the skewness ranges of the OData, 1:2:1 CPData, 1:4:1 CPData, and 1:5:1 CPData were −1.001~0.462, −1.466~0.319, −1.788~0.142, and −1.796~0.150, respectively). In summary, the distribution of the CPData at the 1:2:1 proportion was closer to that of the OData, and MTOTC worked better, compared with the other distribution proportions.

3.1.7 The performance of MTOTC with different GWAS methods

The HData of ordinal trait were transformed by MTOTC, and the obtained CPData were found to be suitable for association analysis via FASTmrMLM when there were five or fewer hierarchical levels, owing to high power. Meanwhile, similar

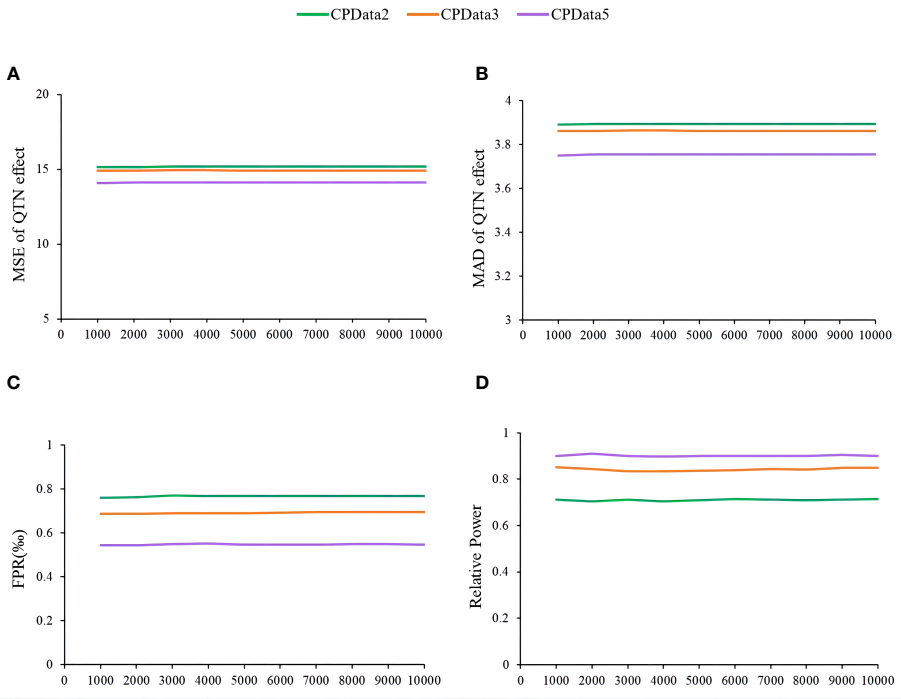


FIGURE 6
The effect of phenotype distribution kurtosis on the association detection results of MTOTC+FASTmrMLM. **(A, B)** MSE and MAD of QTN effect at 2143, respectively; **(C)** false-positive rates; **(D)** relative power.

results were obtained when MTOTC was combined with others methods in the mrMLM software (Zhang et al., 2020) (Supplementary Figure 1; Supplementary Table 1). They were also suitable for GWAS for the CPData of ordinal traits, having the characteristics of high relative power, low false-positive rates, and high accuracy of position and effect estimates. Moreover, similar trends from FASTmrMLM in the simulation experiments with the number of the hierarchical levels and their distribution proportions were observed as well (Supplementary Figure 2). MTOTC + FASTmrMLM had the best performance, followed by mrMLM (Wang et al., 2016), ISIS EM-BLASSO (Tamba et al., 2017), and FASTmrEMMA (Wen et al., 2018); and finally by pLARmEB (Zhang et al., 2017) and pKWmEB (Ren et al., 2018). Therefore, MTOTC can be integrated with different methods to conduct GWAS for ordinal traits. Considering the diversity and complexity of phenotypic data in ordinal traits in practice, multiple methods might be simultaneously used in a complementary manner. Accordingly, MTOTC improves the performance in identifying significant loci for ordinal traits.

3.2 Real data analysis

To validate the new method, the salt-alkali tolerant data in 286 soybean accessions obtained in 2009 and 2010 from Zhang et al. (2014) was re-analyzed in this study. The experiments were conducted in a completely randomized Design, and the number of high-quality SNP markers in this population was 54,296 (Zhou et al., 2015). First, MTOTC was applied to obtain the CPData. Then, the index data, HData5 [hierarchical data generated from the index data by 1:1:1:1:1 (Shao, 1986)], CPData2 (continuous phenotypic data generated from HData2 by MTOTC), and CPData5 (continuous phenotypic data generated from HData5 by MTOTC) for salt-alkali tolerance in soybean were analyzed using the mrMLM, ISIS EM-BLASSO, pLARmEB, FASTmrEMMA, pKWmEB, and FASTmrMLM methods.

3.2.1 QTNs significantly associated with soybean salt-alkali tolerance

For the four types of phenotypic data of salt-alkali tolerance, a greater number of significant QTNs were detected in CPData than in the index data or HData. Six GWAS methods mapped 65 and 99 QTNs in CPData2 and CPData5 of salt tolerance traits, respectively, and 134 and 60 QTNs in CPData2 and CPData5 of alkali tolerance traits, respectively. pLARmEB detected a greater number of QTNs in CPData (116 for salt tolerance traits and 166 for alkali tolerance traits) compared with the other five GWAS methods, which may be related to its relatively higher false-positive rate. Additionally, the numbers of significant QTNs detected by pKWmEB, mrMLM, and FASTmrMLM in CPData (44, 25, and 14 for the salt tolerance trait and 25, 21, and 19 for the alkali-tolerance trait, respectively) were second only to the number of QTNs detected with pLARmEB.

Four QTNs (locus 9682 on chromosome 2 [Chr2-9682], Chr11-54042, Chr13-64738, and Chr13-65248) for salt tolerance were simultaneously detected in the index data and at least one CPData;

however, none of them was detected in HData5. For instance, Chr13-64738 was simultaneously detected in CPData2 by five methods and in the salt tolerance index data by two methods. Chr13-65248 was detected in CPData5 by four methods and in both CPData5 and the index data by FASTmrMLM. Three QTNs (Chr7-34669, Chr13-67342, and Chr20-105040) for alkali tolerance were simultaneously detected in the index data and in at least one CPData, two of them were also detected in HData5.

The results of six GWAS methods for the CPData of salt-alkali tolerance showed that only a few significant QTNs were coincident between 2009 and 2010, which can be explained by the differences in environmental influences between the two years. For salt tolerance, no QTNs were found to overlap between 2009 and 2010 in the six methods. For alkali tolerance, only Chr1-5051 and Chr16-82333 were detected in both years. There was indeed an environmental (year) effect according to variance analysis of the phenotypic results for the two years (Zhang et al., 2014).

3.2.2 Candidate genes for salt-alkali tolerance

Potential candidate genes were mined from 100 kb upstream to 100 kb downstream (Liu et al., 2020) of significant QTNs that were detected in at least two types of data or by two methods (Tables 2 and 3). Functional annotation information in the SoyBase database (**Error! Hyperlink reference not valid.** http://www.Soybase.org/)

TABLE 2 Salt stress-related candidate genes from six genome-wide association study methods.

Candidate genes	QTN positions	Methods	Functional annotation	Arabidopsis homologous
Glyma02g38320	43804331	mrMLM[1]**, pLARmEB[3]**	transmembrane transport	AT5G22900
Glyma02g38350	43804331	mrMLM[1]**, pLARmEB[3]**	Pentatricopeptide repeat (PPR -like) superfamily protein	AT5G37570
Glyma02g38370	43804331	mrMLM[1]**, pLARmEB[3]**	zinc ion binding	AT2G40770
Glyma02g38380	43804331	mrMLM[1]**, pLARmEB[3]**	catalytic activity	AT5G05200
Glyma02g38395	43804331	mrMLM[1]**, pLARmEB[3]**	respiratory burst involved in defense response	AT5G05190
Glyma04g13670	13441084	FASTmrEMMA[3]**, mrMLM[3]**, pLARmEB[3]**	oxidoreductase activity	AT4G25240
Glyma05g25331 #	31519270	FASTmrEMMA[3]*, ISIS EM-BLASSO[3]*, mrMLM[3]*, pKWmEB[3]*, pLARmEB[3]*	WRKY DNA-binding domain	AT2G34830
Glyma05g25420 #	31519270	FASTmrEMMA[3]*, ISIS EM-BLASSO[3]*, mrMLM[3]*, pKWmEB[3]*, pLARmEB[3]*	zinc ion binding	AT5G37930
Glyma05g25450 #	31519270	FASTmrEMMA[3]*, ISIS EM-BLASSO[3]*, mrMLM[3]*, pKWmEB[3]*, pLARmEB[3]*	catalytic activity	AT5G44440
Glyma05g25460 #	31519270	FASTmrEMMA[3]*, ISIS EM-BLASSO[3]*, mrMLM[3]*, pKWmEB[3]*, pLARmEB[3]*	catalytic activity	AT2G34790
Glyma08g13260	9687628	FASTmrEMMA[3]**, FASTmrMLM[3]**, ISIS EM-BLASSO[3]**, mrMLM[3]**, pKWmEB[3]**, pLARmEB[3]**	Serine/threonine protein kinase	AT3G16030
Glyma10g40400 #	47864560	FASTmrMLM[2]*, ISIS EM-BLASSO[2]*, mrMLM[2]*, pKWmEB[2]*, pLARmEB[2]*	zinc ion binding	AT5G67450
Glyma10g40510 #	47864560	FASTmrMLM[2]*, ISIS EM-BLASSO[2]*, mrMLM[2]*, pKWmEB[2]*, pLARmEB[2]*	zinc ion binding	AT4G15090
Glyma10g40520 #	47864560	FASTmrMLM[2]*, ISIS EM-BLASSO[2]*, mrMLM[2]*, pKWmEB[2]*, pLARmEB[2]*	oxidoreductase activity	AT4G33910
Glyma11g14030 #	10094063	mrMLM[1]**, pKWmEB[1]**, pLARmEB[3]**	protein serine/threonine kinase activity	AT3G20830
Glyma11g14040 #	10094063	mrMLM[1]**, pKWmEB[1]**, pLARmEB[3]**	sequence-specific DNA binding transcription factor activity	AT1G51190
Glyma11g14050 #	10094063	mrMLM[1]**, pKWmEB[1]**, pLARmEB[3]**	zinc ion binding	AT1G51200
Glyma11g14081 #	10094063	mrMLM[1]**, pKWmEB[1]**, pLARmEB[3]**	catalytic activity	AT3G18080
Glyma11g14090 #	10094063	mrMLM[1]**, pKWmEB[1]**, pLARmEB[3]**	transmembrane transport	AT3G20870
Glyma11g14100 #	10094063	mrMLM[1]**, pKWmEB[1]**, pLARmEB[3]**	zinc ion binding	AT1G51220
Glyma11g14110 #	10094063	mrMLM[1]**, pKWmEB[1]**, pLARmEB[4]**	Zinc finger, C3HC4 type (RING finger)	AT3G63530

(Continued)

TABLE 2 Continued

Candidate genes	QTN positions	Methods	Functional annotation	Arabidopsis homologous
Glyma12g03490	2356018	FASTmrEMMA[3]*, FASTmrMLM[3]*, ISIS EM-BLASSO[3]*, mrMLM[3]*, pKWmEB[3]*, pLARmEB[3]*	transmembrane transporter	AT2G21050
Glyma12g03570	2356018	FASTmrEMMA[3]*, FASTmrMLM[3]*, ISIS EM-BLASSO[3]*, mrMLM[3]*, pKWmEB[3]*, pLARmEB[3]*	catalytic activity	AT4G34980
Glyma12g03580	2356018	FASTmrEMMA[3]*, FASTmrMLM[3]*, ISIS EM-BLASSO[3]*, mrMLM[3]*, pKWmEB[3]*, pLARmEB[3]*	transmembrane transporter	AT5G09220
Glyma13g25266 #	28469311	FASTmrEMMA[2]**, FASTmrMLM[1,2]**, ISIS EM-BLASSO[2]**, pKWmEB[2]**, pLARmEB[1,2]**	hyperosmotic salinity response	AT1G61120
Glyma13g27630 #	30845044	FASTmrEMMA[3]**, FASTmrMLM[1,3]**, mrMLM[3]**, pKWmEB[3]**	protein serine/threonine kinase activity	AT3G20530
Glyma13g27680 #	30845044	FASTmrEMMA[3]**, FASTmrMLM[1,3]**, mrMLM[3]**, pKWmEB[3]**	transmembrane transport	AT1G61800
Glyma13g27691 #	30845044	FASTmrEMMA[3]**, FASTmrMLM[1,3]**, mrMLM[3]**, pKWmEB[3]**	zinc ion binding	AT4G14220
Glyma13g27701 #	30845044	FASTmrEMMA[3]**, FASTmrMLM[1,3]**, mrMLM[3]**, pKWmEB[3]**	response to oxidative stress	AT3G06050
Glyma13g27710 #	30845044	FASTmrEMMA[3]**, FASTmrMLM[1,3]**, mrMLM[3]**, pKWmEB[3]**	response to oxidative stress	AT3G06050
Glyma13g27740 #	30845044	FASTmrEMMA[3]**, FASTmrMLM[1,3]**, mrMLM[3]**, pKWmEB[3]**	oxidoreductase activity	AT3G06060
Glyma13g27770 #	30845044	FASTmrEMMA[3]**, FASTmrMLM[1,3]**, mrMLM[3]**, pKWmEB[3]**	sequence-specific DNA binding transcription factor activity	AT1G54830
Glyma15g42440	49869431	FASTmrEMMA[2]*, mrMLM[2]*, ISIS EM-BLASSO[2]*, pKWmEB[2]*, pLARmEB[2]*	Myb-like DNA-binding domain	AT2G44430
Glyma15g42460	49869431	FASTmrEMMA[2]*, mrMLM[2]*, ISIS EM-BLASSO[2]*, pKWmEB[2]*, pLARmEB[2]*	Serine/threonine protein kinase	AT2G32850

1: index data; 2: continuous phenotypic data (CPData2) generated from HData2 by MTOTC; 3: continuous phenotypic data (CPData5) generated from HData5 by MTOTC; *: 2009; **: 2010, #: candidate genes were further screened by haplotype block analysis.

was also used to screen candidate genes. A total of 34 potentially candidate genes for salt tolerance and 25 potentially candidate genes for alkali tolerance were mined.

For salt tolerance, 19 candidate genes were detected simultaneously in the index data and CPData5. Among them, *Glyma05g25331*, *Glyma05g25420*, *Glyma05g25450*, and *Glyma05g25460* were all detected by five GWAS methods in CPData5 in 2009. Only one gene, *Glyma13g25266*, was detected in both the index data and CPData2 detected by five GWAS methods in CPData2 and two methods in the index data in 2010. In addition, five candidate genes were detected only in CPData2 by five methods, and nine candidate genes were detected only in CPData5 by three or more methods. No overlapping genes were found between HData5 and the index data or the CPData (Table 2).

For alkali tolerance, 7 candidate genes for alkali stress were concurrently detected in the index data and CPData5. For instance, *Glyma07g20380* was simultaneously detected by 2, 1, and 6 GWAS methods in the index data, HData5, and CPData5 in 2010, respectively (Table 3). Two candidate genes were detected in the index data and CPData2. Ten candidate genes were simultaneously detected in CPData2 and CPData5. *Glyma10g02920* was detected by one GWAS method in CPData2 and five GWAS methods in CPData5 in 2009. *Glyma07g20380* was detected by all six association analysis methods in CPData5 in 2010.

3.2.3 QTN based haplotype and phenotypic difference analysis

Based on the above 34 salt stress-related candidate genes and 25 alkali stress-related candidate genes, Haploview software was used to perform haplotype block analysis. And the phenotypic differences across haplotypes were examined using the t-test in SAS9.4. Four stable QTNs for salt tolerance and six stable QTNs for alkali resistance were screened to form haplotype blocks based on linkage disequilibrium (Supplementary Figures 3 and 4).

In haplotype block with the significant QTNs Chr13-64738 for salt tolerance, t-test showed significant phenotypic differences between haplotypes ACAT and AATT (P=0.0341 in 2009 and P=0.0083 in 2010), between haplotypes TCAT and AATT (P=0.0091) in 2010, and between haplotypes TCAT and TCTT (P=0.0471) in 2010. However, for haplotype blocks of other salt tolerance QTNs, it was showed that the significant phenotypic differences existed between haplotypes only in a single year, and the haplotype pairs with significant differences included haplotype AGTGC and TACCC (P=0.0348), AGTGC and TGTCA (P=0.0345)

TABLE 3 Alkali stress-related candidate genes from six genome-wide association study methods.

Candidate genes	QTN positions	Methods	Functional annotation	Arabidopsis homologous
Glyma01g41510 #	53035914	pLARmEB[2,3]*	Protein serine/threonine kinase activity	AT5G60900
Glyma01g41520 #	53035914	pLARmEB[2,3]*	sequence-specific DNA binding transcription factor activity	AT4G17500
Glyma01g41527 #	53035914	pLARmEB[2,3]*	sequence-specific DNA binding transcription factor activity	AT5G47230
Glyma01g41560 #	53035914	pLARmEB[2,3]*	zinc ion binding	AT5G53110
Glyma01g41581 #	53035914	pLARmEB[2,3]*	sequence-specific DNA binding transcription factor activity	AT5G47370
Glyma01g41610 #	53035914	pLARmEB[2,3]*	sequence-specific DNA binding transcription factor activity	AT3G13540
Glyma03g28210 #	36121029	FASTmrEMMA[2]**, pLARmEB[2]**	F-box family protein	AT2G32560
Glyma03g28222 #	36121029	FASTmrEMMA[2]**, pLARmEB[2]**	F-box family protein	AT2G26850
Glyma03g28234 #	36121029	FASTmrEMMA[2]**, pLARmEB[2]**	F-box family protein	AT2G32560
Glyma03g28247 #	36121029	FASTmrEMMA[2]**, pLARmEB[2]**	F-box family protein	AT2G26850
Glyma07g20380 #	20580766	FASTmrEMMA[3]**, FASTmrMLM[1,3]**, ISIS EM-BLASSO[3]**, mrMLM[3]**, pKWmEB[3]**, pLARmEB[1,3]**	Pentatricopeptide repeat (PPR) superfamily protein	AT3G48810
Glyma13g44560 #	43999096	FASTmrMLM[1]*, pLARmEB[1,3]*, pKWmEB[3]*	transmembrane transport	AT3G19640
Glyma13g44570 #	43999096	FASTmrMLM[1]*, pLARmEB[1,3]*, pKWmEB[3]*	sequence-specific DNA binding transcription factor activity	AT4G37850
Glyma13g44582 #	43999096	FASTmrMLM[1]*, pLARmEB[1,3]*, pKWmEB[3]*	sequence-specific DNA binding transcription factor activity	AT2G22760
Glyma13g44594 #	43999096	FASTmrMLM[1]*, pLARmEB[1,3]*, pKWmEB[3]*	sequence-specific DNA binding transcription factor activity	AT4G37850
Glyma13g44640 #	43999096	FASTmrMLM[1]*, pLARmEB[1,3]*, pKWmEB[3]*	Serine/threonine-protein kinase PBS1	AT1G80640
Glyma13g44660 #	43999096	FASTmrMLM[1]*, pLARmEB[1,3]*, pKWmEB[3]*	sequence-specific DNA binding transcription factor activity	AT5G25190
Glyma16g25280 #	29252235	pLARmEB[2,3]*	sequence-specific DNA binding transcription factor activity	AT2G18350
Glyma16g25310 #	29252235	pLARmEB[2,3]*	transmembrane transport	AT1G75220
Glyma16g25320 #	29252235	pLARmEB[2,3]*	transmembrane transport	AT1G75220
Glyma19g39270	46014852	FASTmrMLM[1]*, pKWmEB[1]*, pLARmEB[1]*	response to oxidative stress	AT4G11290
Glyma19g39320	46014852	FASTmrMLM[1]*, pKWmEB[1]*, pLARmEB[1]*	oxidoreductase activity	AT4G03140
Glyma19g39340	46014852	FASTmrMLM[1]*, pKWmEB[1]*, pLARmEB[1]*	Regulation of transcription	AT5G62000
Glyma20g31790 #	40400845	pLARmEB[1,2]*	zinc ion binding	AT3G52300
Glyma20g31800 #	40400845	pLARmEB[1,2]*	transmembrane transport	AT2G35800

1: index data; 2: continuous phenotypic data (CPData2) generated from HData2 by MTOTC; 3: continuous phenotypic data (CPData5) generated from HData5 by MTOTC; *: 2009; **: 2010, #: candidate genes were further screened by haplotype block analysis.

for Chr5-24153; haplotype GCG and ATA (P=0.0408) for Chr10-52140; haplotypes GTAGA and GTAGT (P=0.0397), GTAGT and AAGTT (P=0.0540) for Chr11-54042.

There were two significant QTNs Chr16-82333 and Chr3-14262 for alkali tolerance with significant phenotypic differences across haplotypes in both years. The Chr16-82333 recorded significant differences between haplotypes CTGACG and CCGGAG (P=0.0158 in 2009, P=0.0614 in 2010), between haplotypes CTGACG and CCGGAG (P=0.0005 in 2009), between haplotypes CTGACG and CCGAAG (P=0.0231 in 2009), between haplotypes TCGAAG and CCGAAG (P=0.0619 in 2009, P=0.0261 in 2010), and between haplotypes CCAAAG and CCGGAG (P=0.0296 in 2010). For Chr3-

14262, the haplotype pairs with significant differences were detected as follows: TTT and TCT (P=0.0217 in 2009, P=0.0085 in 2010), TTT and GCT (P=0.0102 in 2010), GCT and TCT (P=0.0171). The other haplotype blocks of alkali tolerance showed significant phenotypic differences between haplotypes only in a single year and they include: GTGT and TTAT (P<0.0001), TTGT and TTAC (P=0.0038), TTAT and TAGT (P=0.0132) for Chr13-67342; CAG and TGT (P=0.0183) for Chr1-5051; ATCG and GATC (P=0.0009) for Chr7-34669; TAGGCG and AATGCA (P=0.0157), and TAGGCG and TATGCG (P=0.0128) for Chr20-105040.

Genes with significant phenotypic differences across haplotypes were considered as the candidate genes (Tables 2 and 3), including 22 salt stress-related candidate genes and 22 alkali stress-related candidate genes. Among them, six salt stress-related candidate genes (*Glyma05g25420*, *Glyma11g14030*, *Glyma11g14040*, *Glyma11g14050*, *Glyma13g27691*, *Glyma13g27701*) and six alkali stress-related candidate genes (*Glyma03g28222*, *Glyma03g28234*, *Glyma03g28247*, *Glyma16g25320*, *Glyma20g31790*, *Glyma20g31800*) were found in the haplotype block.

4 Discussion

In this study, we established a method for transforming ordinal phenotypes into continuous phenotypes (MTOTC) based on hierarchical data for ordinal trait phenotypes and molecular marker data in resource populations. Therefore, the process of association analysis for ordinal traits is as follows: first, MTOTC is used to transform HData into continuous phenotypic data (CPData), and then a C-GWAS method (i.e. GWAS method for continuous quantitative traits) is selected to analyze the CPData to identify the QTNs that are significantly associated with ordinal traits.

In this study, simulation experiments and soybean saline-alkali tolerance analysis indicated that the new method, MTOTC, is suitable for ordinal traits when they are less than five hierarchical levels. Moreover, the combination of MTOTC with any one of the proposed C-GWAs methods exhibited high power, low false-positive rates, and low bias in estimating the positions and effects of the QTN. The purpose of MTOTC is to provide a different approach for undertaking GWAS for ordinal traits. The feasibility of the MTOTC method was verified in real data analysis of soybean salt-alkaline tolerance using 286 soybean accessions. Compared with HData5 (i.e., the data classified as five hierarchical levels), a greater number of significant QTNs was detected concurrently by at least two GWAS methods or in two years, and more candidate genes for salt and alkali stress were screened in the CPData for salt and alkali tolerance traits. A greater number of QTNs was detected simultaneously by multiple GWAS methods in the CPData than in the index data and HData for salt-alkaline tolerance. For the three types of data, the number of QTNs detected simultaneously was respectively 4, 1, and 1 in salt tolerance and respectively 5, 2, and 3 in alkali resistance.When the phenotype distribution of the CPData generated by the new method were closer to those from the index data of salt-alkali tolerance, the GWAS results were better, and a greater number of candidate genes could be mined. This may be beneficial for selecting the appropriate distribution proportion to obtain hierarchical data of ordinal trait, screening stable QTNs, and promoting the development of molecular breeding. We also applied symmetric distribution (1:2:4:2:1) to generate HData5 for the salt tolerance index data and used MTOTC to generate the corresponding CPData5. The phenotype distribution of CPData5 with symmetric 1:2:4:2:1 exhibited a large deviation from that of the index data, and the phenotype distribution of CPData5 with uniform 1:1:1:1:1 was closer to that of the index data. Under the six methods, there were no overlapping QTNs in CPData5 and the index data for salt tolerance, which was far inferior to the above uniform distribution observed with the distribution proportion 1:1:1:1:1, under which three coincident QTNs were detected in CPData5 and the index data. This result corresponded precisely to the results presented in simulation study 5.

MTOTC performed well in the initial SNP screening. After preliminary screening under a $P \leq 0.05$ threshold, a large number of SNPs that were significantly unrelated to the trait could be eliminated. Meanwhile, the simulation experiment showed that the retention rates of related loci remained high. MTOTC serves to simplify the model and save a substantial amount of computing time for subsequent association studies.

MTOTC helps to improve association analyses of ordinal traits. Regarding coefficient of variation, skewness, kurtosis, and frequency distribution, compared with the HData, the results obtained for the CPData were closer to those of the OData. Meanwhile, the results using six GWAS methods showed that the statistical power, the false-positive rate, and the position estimates in CPData were better than those in HData. Moreover, MTOTC performed better when the frequency distribution of the CPData was close to that of the OData.

The fewer hierarchical levels, the more suitable MTOTC is. Regarding the relative power in CPData under different hierarchical levels, a trend of increasing relative power with increasing number of hierarchical levels was found for all six methods when there were four or less hierarchical levels. When there were five hierarchical levels, the power of MTOTC+FASTmrMLM was close to that of FASTmrMLM in HData, but slightly lower than the power from logistic regression; only three GWAS methods had higher relative power in CPData than in HData. In addition, MTOTC had a tendency to increase variation, especially with increasing numbers of hierarchical levels. This indicates that MTOTC is more suitable for ordinal traits with fewer hierarchical levels, especially those with two or three levels. Among the six GWAS methods, FASTmrEMMA, FASTmrMLM, and mrMLM are significantly better when combined with MTOTC. This is partly attributed to that the distribution and parameter estimation principles set in MTOTC were relatively consistent with those in these three GWAS models.

This study will contribute to further research in association analysis of ordinal traits. This is especially in improving the retention rate of small-effect loci in preliminary screening,

reducing the impact on variability when transforming ordinal phenotypes into continuous phenotypes, and developing novel methods for association analyses of ordinal 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

MY, JF, and YW designed the methodologies. MY, JF, JZ, and JCZ drafted the manuscript, conducted simulation studies, and analyzed the data. TZ and JF revised the paper. All authors contributed to the article and approved the submitted version.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by Major National Agricultural Science and Technology Projects of China (2022ZD0400704), the National Key R & D Program of China (2021YFD1201603), the National Natural Science Foundation of China (32070688).

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1247181/full#supplementary-material

References

Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., et al. (2010). Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature.* 465 (7298), 627–631. doi: 10.1038/nature08800

Bi, W., Kang, G., Zhao, Y., Zhao, Y. L., Cui, Y. H., Yan, S., et al. (2015). SVSI: fast and powerful set-valued system identification approach to identifying rare variants in sequencing studies for ordered categorical traits. *Ann. Hum. Genet.* 79 (4), 294–309. doi: 10.1111/ahg.12117

Chang, F. G., Guo, C. Y., Sun, F. L., Zhang, J. S., Wang, Z. L., Kong, J. J., et al. (2018). Genome-wide association studies for synamic plant height and number of nodes on the main stem in summer sowing soybeans. *Front. Plant science.* 9, 1184. doi: 10.3389/fpls.2018.01184

Cuevas, H. E., Prom, L. K., Cooper, E. A., Knoll, J. E., and Ni, X. Z. (2018). Genome-wide association mapping of anthracnose (Colletotrichum sublineolum) resistance in the U.S. Sorghum association panel. *Plant Genome* 11 (2), 1–13. doi: 10.3835/plantgenome2017.11.0099

Feng, J. Y., Zhang, J., Zhang, W. J., Wang, S. B., Han, S. F., and Zhang, Y. M. (2013). An efficient hierarchical generalized linear mixed model for mapping QTL of ordinal traits in crop cultivars. *PloS One* 8 (4), e59541. doi: 10.1371/journal.pone.0059541

He, L., and Kulminski, A. M. (2020). Fast algorithms for conducting large-scale GWAS of age-at-onset traits using Cox mixed-effects models. *Genetics* 215 (14), 41–58. doi: 10.1534/genetics.119.302940

Hoggart, C. J., Whittaker, J. C., De Iorio, M., and Balding, D. J. (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PloS Genet.* 4 (7), e1000130. doi: 10.1371/journal.pgen.1000130

Jiang, L. D., Zheng, Z. L., Fang, H. L., and Yang, J. (2021). A generalized linear mixed model association tool for biobank-scale data. *Nat. Genet.* 53, 1616–1621. doi: 10.1038/s41588-021-00954-4

Liu, J. Y., Li, P., Zhang, Y. W., Zuo, J. F., Li, G., Han, X., et al. (2020). Three-dimension genetic networks among seed oil-related traits, metabolites and genes reveal the genetic foundations of oil synthesis in soybean. *Plant J.* 103 (3), 1103–1124. doi: 10.1111/tpj.14788

Megerssa, S. H., Ammar, K., Acevedo, M., Brown-Guedira, G., Ward, B., Degete, A. G., et al. (2020). Multiple-race stem rust resistance loci identified in durum wheat using genome-wide association mapping. *Front. Plant Science.* 11, 1934. doi: 10.3389/fpls.2020.598509

Osval, A. M., Abelardo, M., Paulino, P., Gustavo, C., Eskridge, K. M., and Crossa, J. (2015). Threshold models for genome-enabled prediction of ordinal categorical traits in plant breeding. *G3: Genes|Genomes|Genetics.* 5 (2), 291–300. doi: 10.1534/g3.114.016188

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics.* 155 (2), 945–959. doi: 10.1093/genetics/155.2.945

Ren, W. L., Wen, Y. J., Dunwell, J. M., and Zhang, Y. M. (2018). pKWmEB: integration of Kruskal-Wallis test with empirical Bayes under polygenic background control for multi-locus genome-wide association study. *Heredity.* 120 (3), 208–218. doi: 10.1038/s41437-017-0007-4

Shao, G. H. (1986). Field identification method of salt tolerance of soybean germplasm resources. *Crops.* 3, 1001–1986. doi: 10.16035/j.issn.1001-7286.1986.03.031

Shim, S., Ha, J., Kim, M., Choi, M. S., Kang, S., Jeong, S., et al. (2019). GmBRC1 is a candidate gene for branching in soybean [*Glycine max* (L.) Merrill]. *Plant Genet. Mol. Breed.* 20 (1), 135. doi: 10.3390/ijms20010135

Song, X. Y., Iuliana, I. L., Liu, M. L., Reibman, J., and Wei, Y. (2016). A General and robust framework for secondary traits analysis. *Genetics.* 202, 1329–1343. doi: 10.1534/genetics.115.181073

Sun, L. M., Wang, C., and Hu, Y. Q. (2016). Utilizing mutual information for detecting rare and common variants associated with a categorical trait. *PeerJ.* 4, e2139. doi: 10.7717/peerj.2139

Tamba, C. L., Ni, Y. L., and Zhang, Y. M. (2017). Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. *PloS Comput. Biol.* 13 (1), e1005357. doi: 10.1371/journal.pcbi.1005357

Tamba, C. L., and Zhang, Y. M. (2018). A fast mrMLM algorithm for multi-locus genome-wide association studies. *bioRxiv*. doi: 10.1101/341784

Tan, Q. H., Christiansen, L., Charlotte, B. A., Zhao, J. H., Li, S. X., Kruse, T. A., et al. (2007). Retrospective analysis of main and interaction effects in genetic association studies of human complex traits. *BMC Genet.* 8 (1), 70–75. doi: 10.1186/1471-2156-8-70

Wang, S. B., Feng, J. Y., Ren, W. L., Huang, B., Zhou, L., Wen, Y. J., et al. (2016). Improving power and accuracy of genome-wide association studies *via* a multi-locus mixed linear model methodology. *Sci. Rep.* 6 (1), 19444. doi: 10.1038/srep19444

Wang, X., Philip, V. M., Ananda, G., White, C. C., Malhotra, A., Michalski, P. J., et al. (2018). A Bayesian framework for generalized linear mixed modeling identifies new

candidate loci for late-onset Alzheimer's disease. *Genetics.* 209, 51–64. doi: 10.1534/genetics.117.300673

Wang, C., Ruggeri, F., Hsiao, C. K., and Argiento, R. (2017). Bayesian nonparametric clustering and association studies for candidate SNP observations. *Int. J. Approximate Reasoning.* 80, 19–35. doi: 10.1016/j.ijar.2016.07.014

Wen, Y. J., Zhang, H., Ni, Y. L., Huang, B., Zhang, J., Feng, J. Y., et al. (2018). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief Bioinform.* 19 (4), 700–712. doi: 10.1093/bib/bbw145

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange., K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics.* 25 (6), 714–721. doi: 10.1093/bioinformatics/btp041

Xu, S. (2010). An expectation-maximization algorithm for the Lasso estimation of quantitative trait locus effects. *Heredity.* 105 (5), 483–494. doi: 10.1038/hdy.2009.180

Xu, C., Zhang, Y. M., and Xu, S. (2005). An EM algorithm for mapping quantitative resistance loci. *Heredity.* 94, 119–128. doi: 10.1038/sj.hdy.6800583

Zhang, J., Feng, J. Y., Ni, Y. L., Wen, Y. J., Niu, Y., Tamba, C. L., et al. (2017). pLARmEB: integration of least angle regression with empirical Bayes for multilocus genome-wide association studies. *Heredity.* 118 (6), 517–524. doi: 10.1038/hdy.2017.8

Zhang, W. J., Niu, Y., Bu, S. H., Li, M., Feng, J. Y., Zhang, J., et al. (2014). Epistatic association mapping for alkaline and salinity tolerance traits in the soybean germination stage. *PloS One* 9 (1), e84750. doi: 10.1371/journal.pone.0084750

Zhang, Y. W., Tamba, C. L., Wen, Y. J., Li, P., and Zhang, Y. M. (2020). mrMLM v4.0: an R platform for multi-locus genome-wide association studies. *Genomics Proteomies Bioinf.* 18 (4), 481–487. doi: 10.1016/j.gpb.2020.06.006

Zhou, L., Wang, S. B., Jian, J. B., Geng, Q. C., Wen, J., Song, Q. J., et al. (2015). Identification of domestication-related loci associated with flowering time and seed size in soybean with the RAD-seq genotyping method. *Sci. Rep.* 5, 9350. doi: 10.1038/srep09350

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 29 November 2023
DOI 10.3389/fpls.2023.1284403

OPEN ACCESS

EDITED BY
Yuan-Ming Zhang,
Huazhong Agricultural University, China

REVIEWED BY
Jian-Fang Zuo,
Huazhong Agricultural University, China
Gangqiang Cao,
Zhengzhou University, China

*CORRESPONDENCE
Guoping Shu
xugp2011@163.com
Yibo Wang
chigohut@163.com

†These authors have contributed equally to this work

RECEIVED 28 August 2023
ACCEPTED 01 November 2023
PUBLISHED 29 November 2023

CITATION
Shu G, Wang A, Wang X, Chen R, Gao F, Wang A, Li T and Wang Y (2023) Identification of QTNs, QTN-by-environment interactions for plant height and ear height in maize multi-environment GWAS.
Front. Plant Sci. 14:1284403.
doi: 10.3389/fpls.2023.1284403

Identification of QTNs, QTN-by-environment interactions for plant height and ear height in maize multi-environment GWAS

Guoping Shu[1*†], Aifang Wang[1†], Xingchuan Wang[2,3†], Ruijie Chen[2,3], Fei Gao[2,3], Aifen Wang[2,3], Ting Li[1] and Yibo Wang[2,3*]

[1]Center of Biotechnology, Beijing Lantron Seed, LongPing High-tech Corp., Zhengzhou, Henan, China, [2]2Experiment Station, Henan LongPing-Lantron AgriScience and Technology Co., LTD, Zhengzhou, Henan, China, [3]LongPing High-tech Corp., Zhengzhou, Henan, China

Plant height (PH) and ear height (EH) are important traits associated with biomass, lodging resistance, and grain yield in maize. There were strong effects of genotype x environment interaction (GEI) on plant height and ear height of maize. In this study, 203 maize inbred lines were grown at five locations across China's Spring and Summer corn belts, and plant height (PH) and ear height (EH) phenotype data were collected and grouped using GGE biplot. Five locations fell into two distinct groups (or mega environments) that coincide with two corn ecological zones called Summer Corn Belt and Spring Corn Belt. In total, 73,174 SNPs collected using GBS sequencing platform were used as genotype data and a recently released multi-environment GWAS software package IIIVmrMLM was employed to identify QTNs and QTN x environment (corn belt) interaction (QEIs); 12 and 11 statistically significant QEIs for PH and EH were detected respectively and their phenotypic effects were further partitioned into Add*E and Dom*E components. There were 28 and 25 corn-belt-specific QTNs for PH and EH identified, respectively. The result shows that there are a large number of genetic loci underlying the PH and EH GEIs and IIIVmrMLM is a powerful tool in discovering QTNs that have significant QTN-by-Environment interaction. PH and EH candidate genes were annotated based on transcriptomic analysis and haplotype analysis. EH related-QEI *S10_135* (*Zm00001d025947*, *saur76*, small auxin up RNA76) and PH related-QEI *S4_4* (*Zm00001d049692*, *mads32*, encoding MADS-transcription factor 32), and corn-belt specific QTNs including *S10_4* (*Zm00001d023333*, *sdg127*, set domain gene127) and *S7_1* (*Zm00001d018614*, *GLR3.4*, and glutamate receptor 3.4 or *Zm00001d018616*, DDRGK domain-containing protein) were reported, and the relationship among GEIs, QEIs and phenotypic plasticity and their biological and breeding implications were discussed.

KEYWORDS

maize, multi-environment-GWAS, plant height, ear height, QTN, QTN-by-Environment interaction (QEI)

Abbreviations: QEI, QTN that shows QTN-by-environment interaction; GWAS, genome-wide association study.

Introduction

Maize is a cereal plant of the grass family (*Poaceae*) and its domesticated form, the grain corn, is one of the most important crop for food, feed, energy, and industrial materials in the world. China is the second largest grain corn producer after USA and Summer corn belt (33%) and Spring corn belt (47%) are ecological regions that contribute 80% of China's total corn grain output (Shu et al., 2021; Dai et al., 2010). Plant height and ear height are two important maize traits that affect biomass, lodging resistance, and corn grain yield. Enhancing yield and yield stability through genetically controlling plant height and ear height have been important goals in maize genetics and corn breeding. A large number of QTL and QTN loci in maize that associated with plant height and ear height have been identified and reported by quantitative trait loci (QTL) mapping and genome-wide association studies (GWAS) and verified by genetic fine mapping, transcriptomic analyses, and functional genetic analysis (Bai et al., 2010; Zhang et al., 2011; Li et al., 2016; Zheng et al., 2016; Ding et al., 2017; Si et al., 2020; Wang et al., 2023; Jin et al., 2023; Napier et al., 2023; Zhou et al., 2023); among them, *Dwarf 8, Dwarf 9* encodes maize DELLA proteins (Lawit et al., 2010), *Ga3ox2* encodes a GA3 b-hydroxylase (Teng et al., 2013), *ZmTE1*, likely regulates auxin signaling, cell division, and cell elongation (Wang et al., 2022a), *ZmRPH1* that regulate both plant height and ear height, encodes a microtubule-associated protein (Li et al., 2020), *ZmDLE1* is associated with a candidate gene that effectively regulate maize plant height and ear height (Zhou et al., 2023), and a set of growth regulating factors genes (*ZmGRF*) that co-express with a large set of plant height and ear height loci (Si et al., 2020). In the classic Brachytic2 locus (Multani et al., 2003), a number of different alleles or genetic variants have been reported that show various degree of phenotype effect on plant height and ear height and that differentially regulate downstream genes involved in gibberellin and brassinosteroid biosynthesis, auxin transport and cellulose synthesis (Xing et al., 2015; Wei et al., 2018).

Phenotypic plasticity is the property of a given genotype to produce different phenotypes in response to distinct environmental conditions (Pigliucci, 2001) or the ability of a single genotype to produce different phenotypes in response to environmental stimuli (Napier et al., 2023) and it is a joint result of overall environmental effect and genetic effects across environments (Li et al., 2018; Liu et al., 2020b). Genotype x Environment Interaction (GEI) is a special case of environmental plasticity where the two genotypes respond in opposite directions to the changes in the environment (Mather and Caligari, 1974; Laitinen and Nikoloski, 2019). Genotype x Environment Interaction (GEI) on corn yield and agronomic traits has been a major goal of the USA Maize Genomes to Fields Initiative (Alkhalifah et al., 2018; Rogers et al., 2021). Phenotypic plasticity and GEI in maize and other crops have been well-known in plant height and ear height (Wallace et al., 2016; Perrier et al., 2017; Mu et al., 2022). Some environmental factors, such as the difference between day and night temperature (also referred to as DIF) have been shown to influence internode length and plant height (Myster and Moe, 1995). Corn inbred lines with tropical germplasm introgression have been shown to respond to daylength or photoperiod (Coles et al., 2010; Lin et al., 2021; Su et al., 2021; Fei et al., 2022; Osnato et al., 2022). Explaining and predicting phenotypes requires the holistic examination of genomes, environments, and their interaction throughout the spatial and temporal dimensions of an organism's life cycle (Li et al., 2021; Schneider, 2022). In traditional G x E studies, a genotype is treated as a black box of the entire genome, and various statistical models were developed to understand the pattern and mechanism of GEI (Mather and Caligari, 1974; Shu and Fan, 1986; Cooper and DeLacy, 1994; Malosetti et al., 2013). Further partitioning Genome x Environmental interaction or GEI into QTN x E (QEI) or Gene x E (GEI)) is a breakthrough and only becomes feasible in recent years with the availability of whole genome sequencing technology, transcriptomic technology, the availability of abundant DNA polymorphic markers such as SNP and SSR, and improved GWAS methodologies (Xiao et al., 2017; Laitinen and Nikoloski, 2019; Li et al., 2022a; Li et al., 2022b; Jin et al., 2023; Napier et al., 2023).

In this study, we have conducted a multi-environment GWAS using the newly released GWAS software package developed by Li et al. (2022a); Li et al. (2022b) called IIIVmrMLM with the objective of detecting QEIs and QTNs, and estimating their additive-by-environment (add*E) and dominance-by-environment (dom*E) interaction effects of QEIs, and additive effects(add) and dominant effects(dom) of corn-belt specific QTNs. Candidate genes in the surrounding chromosomal regions of these QEIs and QTNs are mined and verified by transcriptomic analysis and haplotype analysis, and their implications to understanding the GEI, and phenotypic plasticity of PH and EH were discussed.

Materials and methods

Germplasm and phenotype evaluation

A diversity panel of 490 inbred lines from Shu et al. (2021) was used for this study, 203 inbred lines (accessions) that grow and seed well in both the Summer Corn Belt and Spring Corn Belt were elected for phenotyping in 2013. Five locations or environments with different latitudes across the Summer and Spring Corn Belt that produce over 80% of China's grain corn were selected for phenotyping, which include a location at the southern end of the Summer Corn Belt, Dancheng (DC, latitude 33.645°N, and longitude 115.177°E) and a location at the northern end of China's Spring Corn Belt, Binxian (BX, latitude 45.759°N, and longitude 127.486°E), and three locations in between: Zhengzhou (ZZ, latitude 34.859°N, and longitude 113.368°E, Summer Corn Belt), Ningjin (NJ, latitude 37.652°N, and longitude 116.800°E, Summer Corn Belt), and Tieling (TL, latitude 42.547°N, and longitude 124.159°E, Spring Corn Belt). At all five locations, the same set of 203 inbreds were planted in the same three-row plots in a complete randomized design (Niu et al., 2013) and five individuals were randomly sampled from each plot to measure plant height and ear height.

Phenotype and environment analysis

The mean values of each inbred for PH and EH in each location (Table S1) were used in the summary statistics, correlation analysis, GGE biplot, and Two-way ANOVA. Summary statistics were obtained by R package 'pastecs', and correlation analysis and plots between different environments for plant height and ear height were completed by R package 'PerformanceAnalytics'. Mega-environments were identified by GGE biplot using the GGEBiplotGUI_1.0-9 package (Frutos et al., 2014) in RStudio software (RStudio, PBC, Boston, MA, USA). Relationships between PH and EH in each location were examined using Pearson correlation coefficients by R. The mean values of plant height and ear height in each mega-environment group were used as phenotype values to identify the significant QTN-by-environment interactions (QEIs). Two-way ANOVA was carried out using the SAS 9.3 (SAS Institute Inc., Cary, NC, USA).

DNA sequencing, genotyping, linkage disequilibrium and population structure

Leaf sample from each inbred line was used for DNA extraction with a CTAB procedure. DNA sequencing follows a protocol of Elshire et al. (2011). Genomic DNA was digested with the restriction enzyme ApeK1. Genotyping-by-Sequencing or GBS libraries were constructed in 96-plex and sequenced on Illumina HiSeq 2000. SNP calling was performed using the TASSEL-GBS pipeline (Glaubitz et al., 2014) and B73 RefGen V2.0 as the reference genome. Initially, 876,297 SNP was filtered with minor allele frequency (MAF) > 5%, missing rate < 20% (Shu et al., 2021; Shu et al., 2023), and data for 73,174 high-quality SNP loci was kept for genome-wide association studies (GWAS). Minor allele frequency (MAF) and proportion heterozygous of filtered SNPs (73,174 SNPs) was calculated by TASSEL 5.2.25. The percentage of SNP with different Minor allele frequency (MAF) and proportion heterozygous was counted and shown in a bar chart (Figure S1).

Linkage disequilibrium (LD) analysis was carried out by TASSEL 5.2.25 (https://www.maizegenetics.net/tassel, Bradbury et al., 2007) with LD window size 50 for all filtered SNP on each chromosome. Structure 2.3.4 (Hubisz et al., 2009) was used to detect the population structure among all 203 maize inbred lines using 7296 Tag-SNP extracted from 73175 SNPs by Haploview 4.2 (Barrett et al., 2005). Burn-in period and Monte Carlo Markov Chain (MCMC) replication number were set as 5,000 and 50,000 respectively for each run. Seven independent runs were performed with subpopulation number k= 3 to 9. The delta K values were estimated and output by Structure 2.3.4.

Genome wide association studies by IIIVmrMLM

IIIVmrMLM, A software package that implements the 3VmrMLM model (Li et al., 2022a; Li et al., 2022b) was employed for genome-wide association studies (GWAS). In the single-locus module, 3VmrMLM includes two steps: 1) genome-scanning was employed, and SNP loci that were significant ($p < 0.01$) in Wald test were kept for the following analysis. A midresult file is output after step 1; 2) all the loci identified in step 1 were incorporated into the Multi-locus Model, all the effects were estimated by empirical Bayes, and the loci with LOD score larger than 3.0 of likelihood ratio test were outputted.

In this study, 73,174 filtered SNPs were used as genotype data, the Q matrix was calculated by the Structure 2.3.4 software under the best K value, the parameter "method" was set to "Multi_env" mode, other parameters were set as default values. The critical P-value and LOD score were set as 0.05/m and 3.0, respectively, for significant and suggested QTNs and QEIs, where m is the number of markers (Li et al., 2022b).

To identify QEIs, the phenotype data from five locations were grouped into the summer corn belt group (E1) containing data from three locations (Dancheng, Zhengzhou, Ningjin) and the spring corn belt group (E2, containing data from Tieling and Binxian), the mean value of all locations within each corn-belt group was calculated for each genotype and used as input data to IIIVmrMLM software under "Multi_env" module. The additive-by-environment (add*E) and dominance-by-environment (dom*E) interaction effects of QEIs were estimated and outputted in the final result.

To identify summer corn belt specific QTNs, the trait phenotype data of a genotype from three locations within the Summer Corn Belt was used, and the phenotype value at each location was used as input data for the IIIVmrMLM software under "Multi_env" module. Similarly, phenotype data from two locations within the spring corn belt was used to identify spring corn belt specific QTNs. The additive effects(add) and dominant effects(dom) of corn-belt specific QTNs were estimated and outputted in the final results of Summer and Spring Corn Belt.

Candidate gene annotations of QEIs and QTNs, and patterns of QTN x E interaction

The fasta sequences containing significant QEIs and QTNs identified by IIIVmrMLM were re-aligned to the B73 v4 reference genome using NCBI BLAST-2.12.0+ (Camacho et al., 2009) to obtain a more accurate physical position for better gene annotations (https://www.maizegdb.org/gbrowse). To identify candidate genes that are associated with a QEI or QTN, we first conducted a primary screening within the chromosomal region 100kb up and down the significant QEI or QTN, then software ANOVAR was used for further screening; ANOVAR only output a candidate that meets the following criteria: the significant QTN or QEI is located within the transcriptional sequence of the candidate (further categorized as in Exon (synonymous or non-synonymous), Intron,3′-UTR, and 5′-UTR or within 1kb upstream or downstream of the candidate. The patterns of key QEIs were visualized by line chart.

Candidate gene identification and tissue-specific expression analysis

The polymorphic SNPs surrounding key significant QEIs and QTNs and their PH and EH phenotype association from the

midresult file and the relationship between SNPs and gene structures was studied using scatter and gene structure diagram. For each candidate gene, transcriptomic databases at MaizeGDB (MaizeGDB, https://www.maizegdb.org/) were searched for its expression profiles in different organs and tissues across different developmental stages. Haplotype analysis was used to verify the phenotype effect of important QTNs.

Results

Phenotypic analyses and mega-environment grouping

The descriptive statistics for PH and EH at five locations or growth environments are presented in Table 1. Variation of PH, measured by CV ranges from 12% to 15% within each location. The range and the degree of variation in PH in the Spring Corn Belts is larger than in the Summer Corn Belt. The absolute values of kurtosis and skewness were all less than 1 (Table 1), indicating that the phenotype data do not significantly depart from a normal distribution and are suitable for GWAS. Variation of EH measured by CV ranges from 19.3% to 29.6% within each location, Which is larger than PH. The range of variation in EH in the Spring Corn Belt is much larger than in the Summer Corn Belts.

The phenotypic correlation between each environment-pair for PH and EH among three Summer Corn Belt locations [Dancheng (DC), Zhengzhou (ZZ), and Ningjin (NJ)] and between two Spring Corn Belt locations Tieling (TL) and Binxian (BX), are shown in Figures 1A and C. As the scatter plot and correlation coefficients in Figure 1A show, the within-corn belt location-pair correlation coefficients (PH*PH) are 0.77, 0.77, and 0.66 for three Summer Corn Belt locations and 0.66 for two Spring Corn Belt locations for PH, which are significant at 0.01 level. Whereas, the six between-corn belt correlation coefficients are from 0.03 to 0.12, which are not significant at the 0.05 level. The same pattern was observed for EH (Figure 1C), suggesting a high location-location correlation within each corn belt and nearly zero location-location correlation between the two corn belts. The lack of phenotypic correlation between the two corn belts was also revealed by biplot for PH (Figure 1B) and EH (Figure 1D), which shows that the location vectors within the same corn belts form tight bundles, and the two vector bundles form a nearly vertical angle. Thus, GGE biplot groups the five locations into two mega environments which fit well with the assignment of five locations into two corn belts widely adopted by maize breeders and grain corn growers. The above analyses revealed the high similarity in a growth environment and in PH and EH phenotype within a corn belt and large divergences in growth environment and PH and EH phenotype between the two corn belts. The correlation coefficients between PH and EH (PH*EH) within each location range from 0.51 to 0.75 (Table 1), which is significant at 0.001 level.

To verify the results of environmental grouping, variance analysis was conducted to reveal the differences between mega environments (Table S2). The results showed that there were significant genotype x mega environment interactions in both PH

TABLE 1 Descriptive statistics for PH, EH among 203 accessions across five environments.

Traits	Corn belt	Environments	Latitude	No. of Inbreds	Max.-Min. (cm)	Mean ± SD	CV (%)	Skewness	Kurtosis	CC(with EH)
PH	E1	DC	33.6° N	202	110-241	178.0 ± 21.4	12	0.13	0.14	0.56***
		ZZ	34.9° N	203	113.8-263.6	174.3 ± 25.1	14.4	0.49	0.42	0.75***
		NJ	37.7° N	201	115-250	184.6 ± 24.3	13.2	0.37	0.04	0.65***
	E2	TL	42.5° N	203	149-290	209.3 ± 29.8	14.2	0.23	-0.46	0.65***
		BX	45.8° N	202	89-245	169.6 ± 25.4	15	-0.16	0.07	0.51***
EH	E1	DC	33.6° N	202	32-109	66.8 ± 13.8	20.7	0.01	-0.29	
		ZZ	34.9° N	203	30.4-115.8	68.7 ± 13.3	19.3	-0.06	0.35	
		NJ	37.7° N	201	40-110	74.3 ± 14.4	19.3	0.21	-0.06	
	E2	TL	42.5° N	203	43-130	85.1 ± 18.7	22	0.06	-0.7	
		BX	45.8° N	201	23.5-114.1	60.2 ± 17.8	29.6	0.21	-0.14	

DC, Dancheng; ZZ, Zhengzhou; NJ, Ningjin; TL, Tieling; BX, Binxian. CC, Correlation coefficient. ***P < 0.001.

10.3389/fpls.2023.1284403

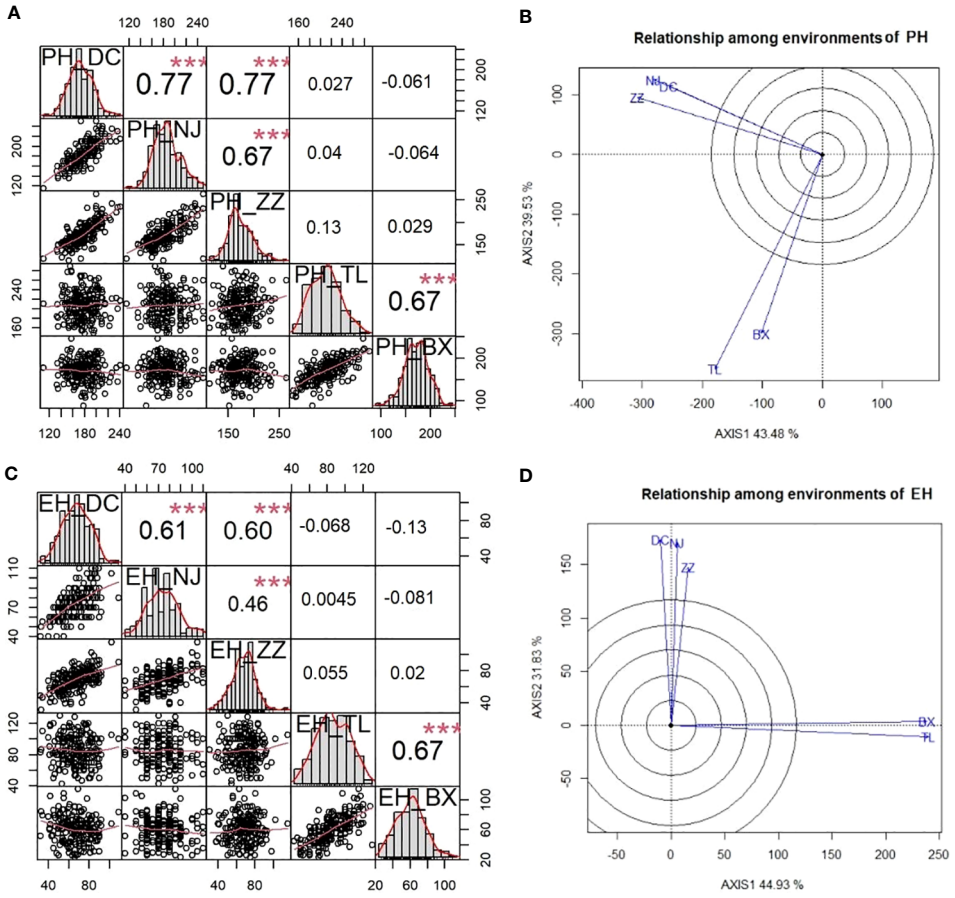


FIGURE 1
Phenotypic correlations between five environments within and between two corn belts viewed by correlation matrix and GGE biplot for PH and EH. **(A)** and **(C)** are correlation matrix among five environments for PH **(A)** and EH **(C)**; **(B)** and **(D)** are GGE biplots for PH **(B)** and EH **(D)**. ***P < 0.001.

and EH. Genotype x mega environments accounted for 30.7% and 31.2% of the total variance for PH and EH respectively. Whereas genotype variance accounted for 32.2% and 29.2% of the total variance for PH and EH, respectively. Therefore, genotype x mega environments interaction is a very important factor in determining the phenotypic plasticity observed in PH and EH.

Characteristics of genotype data, linkage disequilibrium and population structure

Among the 876,297 SNPs collected from 203 inbred lines, 73,174 high-quality SNP loci after a filtering procedure (see Material and Methods) were kept for all analyses in this project. The minor allele frequency (MAF) distribution (see Figure S1A) indicates the existence of abundant allelic polymorphism for genome-wide marker-trait association. About 60% of SNPs with heterozygosity less than 5% are only suitable to additive allelic effect analysis (see Figure S1B), the other 40% of SNPs with heterozygosity higher than 5% are suitable to both additive and dominant allelic effect analysis. The LD decay across all 10 chromosomes reached down to r^2 = 0.1 when the distance between two adjacent SNP increased up to 60 kb (Figure S2A). The population structure analysis showed that the delta K value reached the peak at K=3, indicating that this diversity panel of 203 inbreds can be divided into three subgroups (Figure S2B), namely, M-Reid+P, SS+Iodent+Lan, and LRC+TSPT, respectively (Figure S2C).

Identification of significant QEIs and the patterns of QTN x E interactions

12 significant QEIs for PH and 11 significant QEIs for EH were identified and reported in Table 2 and they are visualized as pink dots on the Manhattan plots (Figure S3A, B), 9 of 12 QEIs for PH and 8 of 11 QEIs for EH are QEIs with additive effect as a key effect, whereas 3 of 12 QEIs for PH and 3 of 11 QEIs for EH are QEIs with dominant effect as a key effect. *S3_224* and *S10_135* are two QEIs for EH with the largest LOD (QE) and variance.

To visualize and verify the QTN x environment interaction in QEIs identified from IIIVmrMLM graphically, the patterns of QTN x environment interaction of five QEIs from Table 2 were shown by line chart (Figure 2). The QTN x environment interaction was

TABLE 2 QEIs between two mega-environmental groups and associated candidate genes for PH and EH.

Trait	Marker (V4, abbr)	Chr#	Position (V4, bp)	Ref/ Alt	LOD (QE)	Add*E1	Dom*E1	Add*E2	Dom* E2	Var	r2 (%)	Het.	|dom|/| add|	Key effect	Gene ID	Gene Symbol	Category
PH	S1_185	1	184855257	G/A	7.0	3.1	4.8	-3.1	-4.8	10.7	2.6	0.09	1.57	add	Zm00001d031277, Zm00001d031278	*ZAT3/ DOF1.6*	Upstream
	S2_85	2	85448512	A/C	10.1	-4.4		4.4		19.3	4.6	0.12	0.00	add	Zm00001d004132	*cl36164_1*	UTR5
	S2_237	2	236504893	G/A	8.1	1.9	11.1	-1.9	-11.1	13.1	3.1	0.08	5.85	dom	Zm00001d007630	*RPS2*	Non-syn.
	S3_156	3	155997977	A/G	9.3	0.2	-7.2	-0.2	7.2	14.5	3.5	0.28	29.39	dom	Zm00001d042199	*PSB28*	Syn.
	S3_159	3	158641942	A/C	6.6	3.4	6.4	-3.4	-6.4	11.9	2.9	0.02	1.90	add	–		Intergenic
	S4_40	4	40463790	T/C	11.5	-4.5	1.4	4.5	-1.4	19.8	4.7	0.02	0.30	add	Zm00001d049691, Zm00001d049692	*mads32*	Syn.
	S6_66	6	66264336	G/A	6.7	-3.4	-3.2	3.4	3.2	11.5	2.7	0.08	0.94	add	Zm00001d036014	*E3/UBPL*	Intronic
	S6_133	6	133125635	A/G	16.6	-5.5	-3.4	5.5	3.4	28.6	6.8	0.11	0.62	add	Zm00001d037655	-	Non-syn.
	S7_48	7	47993521	C/G	10.3	-4.1		4.1		16.8	4.0	0.11	0.00	add	Zm00001d019648	*nbp1*	Syn.
	S8_7	8	7205104	T/G	9.1	0.3	7.7	-0.3	-7.7	14.1	3.4	0.23	24.27	dom	Zm00001d008396	-	UTR5
	S10_149	10	148903473	C/T	13.5	6.4	1.1	-6.4	-1.1	21.6	5.2	0.49	0.17	add	Zm00001d026606	*cdj5*	Non-syn.
EH	*S1_33*	1	32857527	G/T	11.5	-3.1	-1.7	3.1	1.7	7.0	4.4	0.35	0.55	add	*Zm00001d028386*		Downstream
	S1_86	1	86353115	G/A	5.7	-2.3	0.6	2.3	-0.6	2.9	1.9	0.46	0.28	add	*Zm00001d029772*	*prh126*	Non-syn.
	S1_283	1	283402157	A/C	7.7	2.3	-1.9	-2.3	1.9	5.1	3.2	0.01	0.85	add	*Zm00001d034076*	*mmp165*	Non-syn.
	S2_2	2	1669905	T/C	8.7	-1.4	-3.8	1.4	3.8	4.9	3.1	0.24	2.63	dom	*Zm00001d001837*	*myb133*	Non-syn.
	S3_94	3	94315573	C/A	7.0	-2.5	0.7	2.5	-0.7	3.9	2.5	0.41	0.29	add	*Zm00001d041064*	*NUP1*	Non-syn.
	S3_224	3	223519980	C/T	17.3	3.3	1.5	-3.3	-1.5	10.3	6.5	0.04	0.44	add	*Zm00001d044272*	*bhlh94*	UTR5
	S4_38	4	37703788	A/G	5.8	-1.9	4.9	1.9	-4.9	3.7	2.3	0.01	2.57	dom	*Zm00001d049616*	*gpat9*	Syn.
	S4_225	4	224650169	T/C	14.2	3.6	-0.1	-3.6	0.1	8.1	5.1	0.36	0.01	add	-	-	Intergenic
	S5_1	5	1080954	T/C	7.6	-2.6	-0.3	2.6	0.3	4.9	3.1	0.27	0.10	add	*Zm00001d012848*	-	Non-syn.
	S5_215	5	214720899	A/C	10.3	3.0	0.7	-3.0	-0.7	5.6	3.6	0.41	0.23	add	*Zm00001d018122*	*E3/UBPL*	Non-syn.
	S8_174	8	174327122	C/A	5.0	-1.2	2.4	1.2	-2.4	2.7	1.7	0.29	2.07	dom	*Zm00001d012428*	-	Non-syn.
	S10_135	10	134518892	G/C	20.9	3.4	3.4	-3.4	-3.4	11.8	7.4	0.03	0.99	add	*Zm00001d025947*	*saur76*	Intergenic

EH, ear height; PH, plant height; LOD(QE), LOD score for QEIs; Add*E1, additive effect of E1(Summer Corn Belt); Dom*E1, dominant effect of E1(Summer Corn Belt); Add*E2, additive effect of E2(Spring Corn Belt); Dom*E2, dominant effect of E2(Spring Corn Belt); Var, the variance of each QTN; Het., proportion heterozygous; |dom|/|add|, namely |dom*E1|/|add*E1| or |dom*E2|/|add*E2|; Key effect: if |dom|/|add|≤2, or Proportion Heterozygous >0.05, Key effect would be add; if |dom|/|add|>2, and Proportion Heterozygous>0.05, Key effect would be dom. Category: location of SNPs in genes and effect, upsteam, downstream, UTR5, intergenic, intronic represent SNP locate the region of the candidate gene, Non-syn.(non-synonymous) represent the SNP locate in the exonic region of the candidate genes which cause an amino acid change, Whereas syn.(synonymous) represent the SNP locate in the exonic region of the candidate genes which do not cause an amino acid change.

10.3389/fpls.2023.1284403

10.3389/fpls.2023.1284403

further partitioned into add*E and dom*E as shown in Table 2. *S3_156* is a QEI for PH with large negative dom (dominance)*E1 interaction (-7.2) at E1(Summer Corn Belt) locations and large positive dom (dominance)*E2 interaction (7.2) at E2 (Spring Corn Belt) locations, and with an absolute dom/add ratio of 29.39, Figure 2A illustrates the interaction pattern of its three genotypes and shows that heterozygotic AG genotype has significantly shorter PH than both "AA" and "GG" genotype at Summer Corn Belt (E1), but has much taller PH at Spring Corn Belt (E2). Another QEI with a dominant effect as key effect is S8_7 for PH (Figures 2C), with a high absolute dom/add ratio of 24.27. The QEIs *S4_40* for PH and *S3_224* for EH are QEIs with additive effect as key effect and absolute dom/add ratio of 0.3 and 0.44, respectively (Table 2), the genotype CC and TT show opposite phenotype performance in the Summer and Spring Corn Belts (Figures 2B, D). The QEI S10_135 has an absolute dom/add ratio of 0.99 (Table 2), indicating a nearly equal amount of dom*E and add*E interaction (Table 2; Figure 2E). The candidate genes for *S3_224* and *S10_135* are *Zm00001d044272* (*bhlh94*, bHLH-transcription factor 94) and *Zm00001d025947* (*saur76*, small auxin up RNA76), respectively. The candidate genes for *S4_40* are *Zm00001d049691*(*SDH6*, Succinate dehydrogenase subunit 6 mitochondrial) and *Zm00001d049692* (*MADS32*, MADS-transcription factor 32), likely an important QEI for PH.

Identification of significant corn-belt-specific QTNs and annotations

28 and 23 QTNs for PH and EH respectively were identified from Summer Corn Belt data, thus are called summer-corn-belt-specific QTNs (Table S3; Figure 3). 25 and 26 QTNs for PH and EH respectively were identified within the Spring Corn Belt, and thus are called spring corn belt specific QTNs. Among the total 102 corn-belt specific QTNs reported in Table S3, 56 QTNs show an additive effect as key effect (|dom/add|<2.0) and 46 QTNs show a dominant effect as key effect (|dom/add|>2.0).

QTN *S10_4* (*Zm00001d023333*, *sdg127*, set domain gene127) and *S7_1* (*Zm00001d018614*, *GLR3.4*: glutamate receptor 3.4 or *Zm00001d018616*, DDRGK domain-containing protein) are two

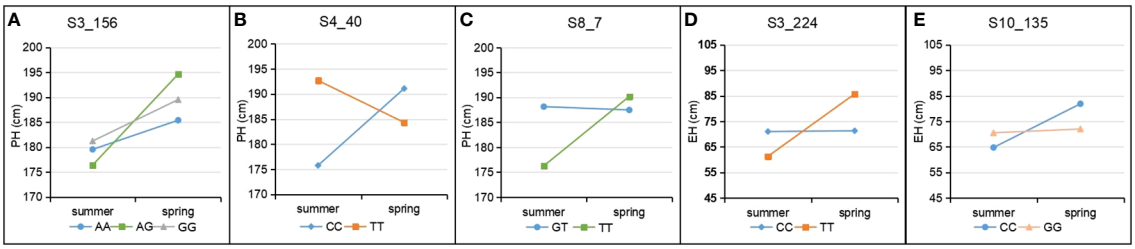


FIGURE 2
Patterns of QTN x E interaction in Summer and Spring Corn Belts for PH and EH. **(A–C)** three QEIs S3_156 **(A)**, S4_40 **(B)** and S8_7 **(C)** for PH; **(D, E)** two QEIs S3_224 **(D)** and S10_135 **(E)** for EH.

TABLE 3 Corn-belt-specific QTNs for PH and EH in Summer and Spring Corn Belt.

Trait	Corn belt	Marker (V4, abbr)	Chr#	Position (V4, bp)	LOD (Q)	Add	Dom	Var	r2 (%)	Het.	|dom|/ |add|	Key effect	Gene ID	Gene Symbol	Category
PH	E1	S1_255	1	255244221	7.8	2.5	0.9	5.6	1.0	0.05	0.37	add	Zm00001d033230	RLK29	Non-syn.
PH	E1	S1_259	1	259066746	55.8	7.3	-0.1	21.6	3.8	0.06	0.01	add	Zm00001d033325	dof39	upstream
PH	E1	S7_1	7	910582	18.5	4.2	-4.8	17.9	3.1	0.06	1.15	add	Zm00001d018614	GLR3.4	Non-syn.
PH	E1	S10_4	10	3618262	9.2	2.9	2.6	8.3	1.4	0.06	0.88	add	Zm00001d023333	sdg127	Non-syn.
PH	E2	S7_151	7	150642747	76.4	17.2	2.1	34.7	3.0	0.12	0.12	add	Zm00001d021386	ZFP2	Non-syn.
PH	E2	S10_15	10	15032123	67.9	7.0	16.4	19.3	1.7	0.73	2.34	dom	Zm00001d023677	sweet13a	Syn.
EH	E1	S1_273	1	273051629	8.8	2.2	-0.3	4.3	2.1	0.07	0.13	add	Zm00001d033765	MAPKK9	upstream
EH	E1	S4_118	4	117960613	29.5	-3.9	-1.0	9.5	4.8	0.01	0.25	add	Zm00001d050715, Zm00001d050716	invan3	upstream
EH	E1	S7_1	7	1024439	6.6	-1.6	1.6	2.5	1.2	0.06	0.97	add	Zm00001d018615	GLR3.4	Non-syn.
EH	E1	S10_4	10	3618262	7.1	1.9	-0.1	3.5	1.7	0.06	0.07	add	Zm00001d023333	sdg127	Non-syn.
EH	E2	S1_7	1	7065140	16.4	-0.9	7.9	15.3	3.0	0.28	8.98	dom	Zm00001d027503, Zm00001d027508	CaBP/ PKs	Non-syn.
EH	E2	S4_41	4	41323782	21.1	5.3		10.2	2.0	0	0	add	Zm00001d049715, Zm00001d049717	iaa16	Syn.

the abbreviation in this table is same as Table 1 and 2. |dom|/|add|: the absolute ratio of dominant effect to additive effect.

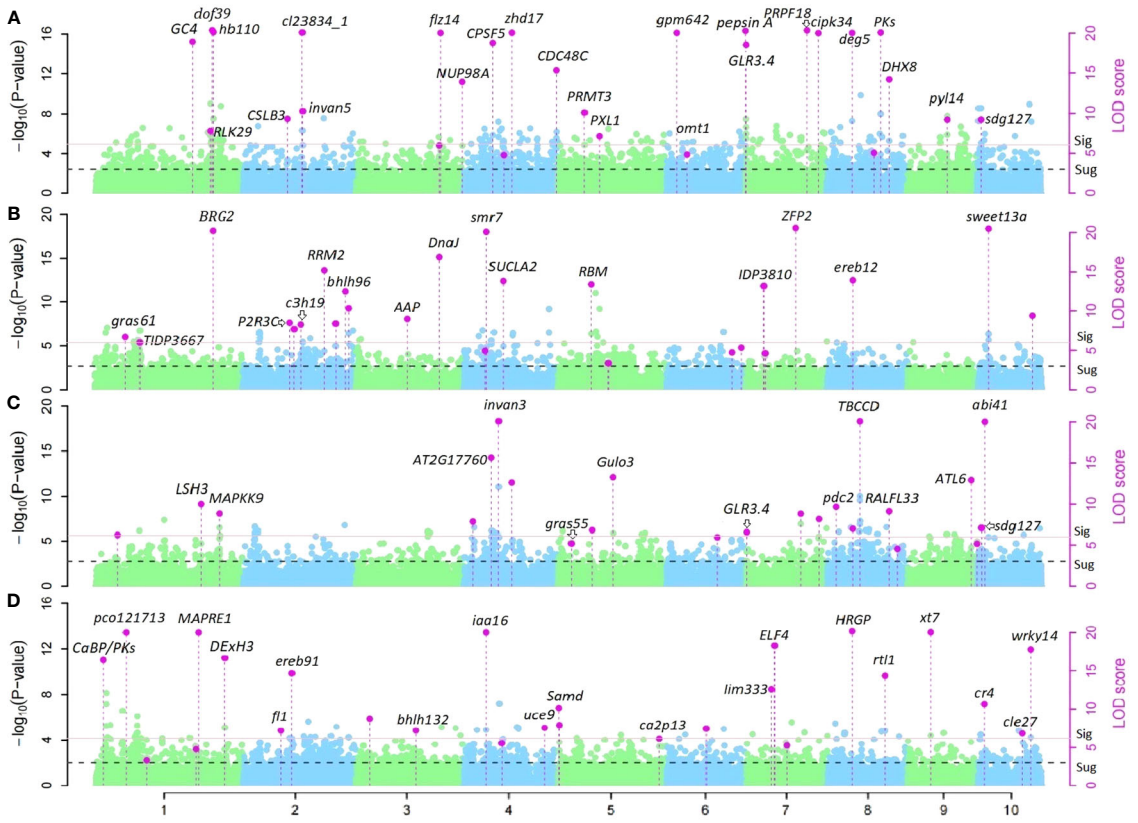


FIGURE 3
Manhattan plots of corn-belt-specific QTNs for PH and EH in Summer and Spring Corn Belt. **(A,B)** corn-belt-specific QTNs and candidate genes for PH in Summer Corn Belt **(A)** and Spring Corn Belt **(B)**; **(C, D)** corn-belt-specific QTNs and candidate genes for EH in Summer Corn Belt **(C)** and Spring Corn Belt **(D)**.

significant summer corn belt specific QTNs for both PH and EH (Tables 3, S3). There are a set of candidate genes located within 7.0 Mb region of chromosome 1, near the three summer corn belt specific QTNs *S1_255*, *S1_259*, and *S1_262*; *Zm00001d033319* (V4: chr1:258878226:258879592, Auxin-responsive protein *IAA4*) is located 200kb from *S1_259* (V4:chr1:259066746) and *Zm00001d033369* (V4:chr1:260633725:260634703, Gibberellin-regulated protein 1) is located between *S1_259* and *S1_262* (Teale et al., 2006; Wang et al., 2017; Luo et al., 2018; Wang and Wang, 2022b; Wu et al., 2023). Another spring corn belt specific QTN, *S1_263* (V4: chr1:262565751) is also located in this region. QTN *S1_255*, *S1_259*, and *S1_262* have additive effects as key effects in the Summer Corn Belt, and the QTN *S1_263* has a dominant effect as key effect in the Spring Corn Belt (Tables 3, S3).

Candidate genes association mapping and tissue-specific expression analysis

Candidate gene search has found that the significant QEI *S3_224* identified by 3VmrMLM is located on the 5'UTR region of *Zm00001d044272* (*bhlh94*), its gene structure is shown in Figure S4. Another QEI, *S4_40* (full ID: S4_40463790, V4: chr4:40463790) is on the exon of two partially overlapping candidate genes *Zm00001d049691*(V4:chr4:40460274 - 40464504) and *Zm00001d049692* (chr4:40462578 - 40464305) (Figure 4. Tables 2, S4). Tissue-specific expression analysis shows *Zm00001d049691* (*SDH6*) expresses in stems, leaves, embryos, roots, spikelets, and silks, *Zm00001d049692* (*MADS32*) expresses in stems, splikelets, and silks, and *Zm00001d049690* (*CY P89A2*) only expresses in roots (Figure S5). *SDH* encodes succinate dehydrogenase, which is activated by salt stress (Fedorin et al., 2023) and is also regulated by light (Eprintsev et al., 2016). Another MADS-transcription factors, *ZmMADS4* and *ZmMADS67* both increase leaf number and delayed flowering, indicating that they promote the floral transition (Sun et al., 2020) and overexpression of *ZmMADS69* causes early flowering (Liang et al., 2019).

Three SNPs surrounding QTN *S10_4* located in *Zm00001d023333* are significant at 0.01 level (-log10-P >2) for PH and EH in the Summer Corn Belt (Figures 5A, B). Two of them: the S10_3620568 and S10_3620675 are located on 5'UTR and the S10_3618266 is located on CDS (Figures 5C, D). *Zm00001d023333* (Chr10:3606398-3621010, *sdg127*, SET domain gene127) encodes a histone-lysine N-methyltransferase ATXR7. Another two SET domain family genes, SET domain group 8 (*SDG 8*) in *Arabidopsis thaliana* (Zhao et al., 2005) and *SDG712* in rice (Zhang et al., 2021) could delay flowering by repressing the expression of FLOWERING LOCUS C (FLC) and florigen genes,

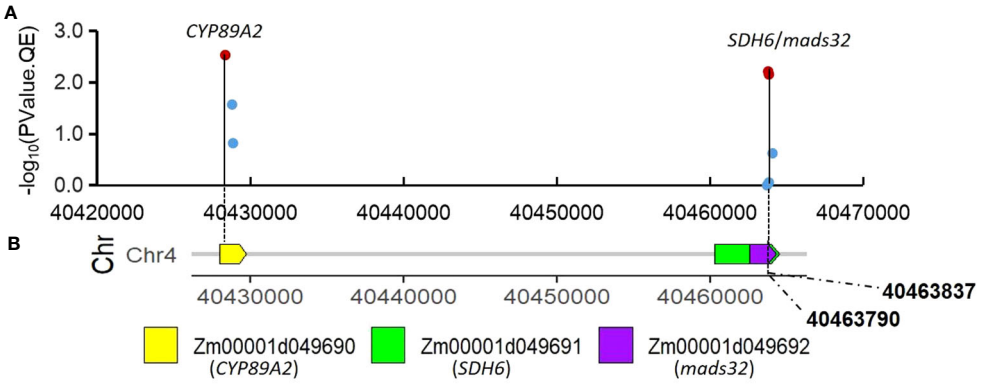


FIGURE 4
Association of SNPs surrounding significant QEI *S4_40* with candidate genes. **(A)** associations of the twelve SNPs using mean value of PH in Summer and Spring Corn Belt; **(B)** gene distribution around *S4_40*(V4:chr4:40463790).

respectively. The above research findings suggest that *Zm00001d023333* we identified in this study might affect PH and EH by delaying flowering time and lengthening vegetative growth. Haplotype analysis has shown that the three SNPs can form six haplotypes (Hap0, Hap1, Hap2, Hap3, Hap4, Hap5) (Figure 5E). Hap 1 (ATA) and Hap 4(GCC) are the major haplotypes, with 36 and 32 inbreds, respectively. Hap 1 (ATA) is higher than Hap 4 (GCC) in terms of both PH and EH (Figures 5F, G).

Several SNPs significantly associated with PH and EH are identified surrounding QTN locus *S7_1*. Some of them are located on the CDS of the two candidate genes *Zm00001d018614* and *Zm00001d018616*. Expression of *Zm00001d018616* (about 30 FPKM) at the mRNA level is ten times higher than *Zm00001d018614* (about 3 FPKM) in the stem (Figure S6). *Zm00001d018614* and *Zm00001d018615* are genes encoding glutamate receptor, which are involved in seed germination inhibition and seedling heat tolerance (Kong et al., 2015; Li et al., 2019). Another candidate gene, *Zm00001d018617* (*ga2ox12*, gibberellin 2-oxidase12, Chr7:1105512-1106576), is a member of gibberellin oxidase gene family which might affect PH (Paciorek et al., 2022), but its expression is not detected in stem tissues of maize (Figure S6).

Three SNPs associated with PH are identified surrounding QTN *S10_15*, a spring-corn belt specific QTN and they are all located in the CDS region of candidate gene *Zm00001d023677* (*sweet13a*, V4:chr10:15030181-15032801) (Figure S7); two SNPs, S10_15032123 and S10_15032153, are synonymous SNV whereas the third SNP, S10_15032160, is nonsynonymous SNV which causes an amino acid change (Table S5). Haplotype analysis has shown that the three SNPs can form four haplotypes (H1, H2, H3, H4). The PH of heterozygous haplotype H2 (CG/CG/TG) is significantly higher than that of the homozygous haplotype H2 (CC/GG/GG) (Figure S7). The candidate gene *Zm00001d023677* (*sweet13a*) encodes a SWEET protein of the MtN3/saliva family (Xuan et al., 2013). Another SWEET protein coding gene *CmSWEET17*, has been reported to be involved in the process of sucrose-induced axillary bud outgrowth in strawberry (*C. morifolium*), possibly via the auxin transport pathway (Liu et al., 2020a).

Discussion

Mega environment, phenotypic plasticity, and mega-environmental GEI and QEI

Partitioning multi-environments into a set of environment clusters or mega environments has been well-studied in which, the multi-environments were grouped using PCA, clustering, and GGE biplot (Shu and Fan, 1986; Yan and Kang, 2003). Yan (2015) defined a mega-environment as a group of geographical environments that share the same (sets of) genotypes consistently across years. Other researchers have defined a mega-environment as a group of growing environments that are similar in terms of genotype response and that show a repeatable relative performance of a set of crop genotypes across years (Yan and Rajcan, 2002). Mega-environments are often identified through the analysis of multiple-environment trial data for a set of genotypes. The purpose of the mega-environment analysis is to understand the nature of environmental variation across experimental locations, whether there is structure or segmentation among the locations. Our result shows that there is significant segmentation among the 5 locations and they can be divided into two mega-environments, there is very little variation among locations within a mega environment and the two segments fall right into the two corn belts that have been widely adopted by breeders and corn growers. Our results also show that the GGE model, with a biplot display, is an effective tool for displaying environment structure and segmentation which explain why it has become popular in analyzing multiple-environment trial data to determine environment cluster (Yan and Kang, 2003; Yan et al., 2011; Yan, 2015; Dai et al., 2010).

Understanding the genetic basis of phenotypic plasticity in general and the genotype x environment interaction (GEI) in particular is of primary importance in traditional crop genetics

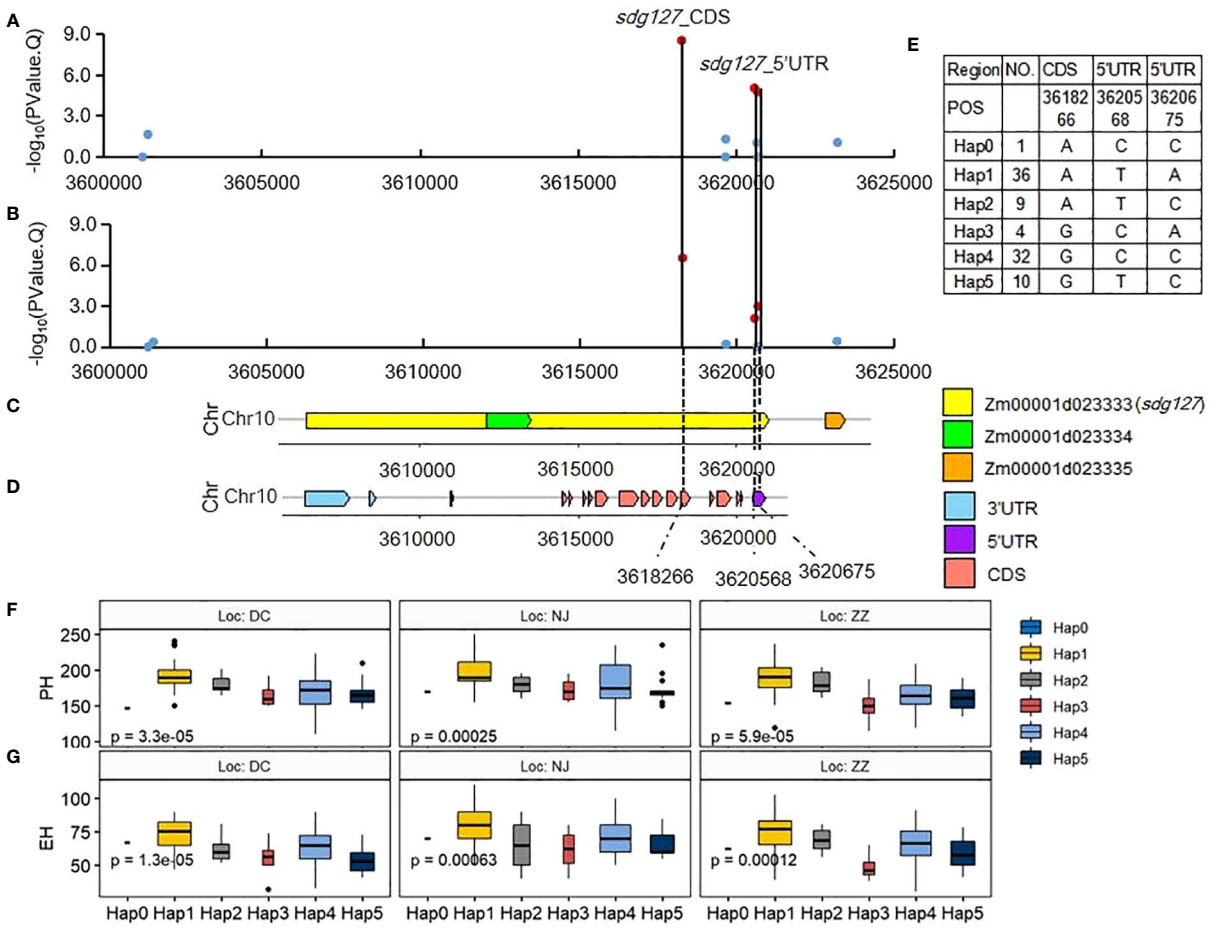


FIGURE 5
Association of SNPs surrounding significant QTN *S10_4* with candidate genes and their haplotype Effects. **(A, B)** associations of the 11 SNPs with PH **(A)** and EH **(B)** in Summer Corn Belt. The dot is red with the threshold of -log10(PValue)>2; **(C)** gene distribution around QTN *S10_4*(V4: chr10:3618266); **(D)** gene structure of *Zm00001d023333*; **(E)** haplotypes of the three significant SNPs; **(F, G)** boxplots of haplotypes for PH **(F)** and EH **(G)** in Summer Corn Belt.

and plant breeding, and a large body of literature on models and strategies is available (Shu and Fan, 1986; Cooper and DeLacy, 1994; Malosetti et al., 2013; Li et al., 2018; Liu et al., 2020b; Schneider, 2022). The genetic bases of genotype x environment interaction (GEI) for PH and EH are difficult to study due to environment structure and segmentation among experiment locations and the multi-locus nature of their genetic control. In this study, we deal with multi-environmental segmentation by grouping multiple locations into mega-environments using GGE biplot and deal with multi-locus nature by dissecting it into QTN x environment interaction or QEIs using multi-environmental GWAS. Our results show that genotype x mega environment interaction (GEI) accounted for about 30% of the total variation for both PH and EH, almost equal to the genotypic variation among 203 inbred lines in proportion (which is also about 30%). Therefore, genotype x mega environments interaction has a significant contribution to the phenotypic plasticity observed in PH and EH.

Understanding the molecular mechanism underlying the detected pattern of phenotypic plasticity in general and G x E, in particular, has been a major effort in the last decade. QTL mapping and genome-wide association studies (GWAS) have been shown effective means in identifying a large number of QTL/QTN and QEIs (Xiao et al., 2017; Jin et al., 2023; Napier et al., 2023) and transcriptomic analysis and functional genomics have been shown as important ways to identify candidate genes and verify their biological functions (Seyfferth et al., 2021; Han et al., 2023; Napier et al., 2023; Wang et al., 2023). Various statistical models and bioinformatic algorithms have been proposed to improve the effectiveness of GWAS but no significant progress has been made on GWAS that can partition GEI and identify QEIs. We have shown that the 3VmrMLM GWAS models and the IIIVmrMLM software package recently released can effectively identify QEIs. The software package has also been applied to data from rice, soybean, and other crops to identify QEIs and hunt candidate genes underlying QEIs (Zhang et al., 2022; Zuo et al., 2022; Zhao et al., 2023). We have shown that by employing 3VmrMLM multi-environment GWAS models, we were able to go beyond the traditional G x E interaction analysis and were able to identify and annotate a set of QEIs for PH and EH.

Among the candidate genes annotated by transcriptomic analysis, *Zm00001d049692* (*MADS32*) surrounding QEI *S4_40*, might affect PH in different ecological zones by both increasing

leaf number, delay flowering time, and lengthen vegetative growth period, similar to *ZmMADS4* and *ZmMADS67* (Sun et al., 2020). *Zm00001d044272* (*bhlh94*) surrounding QEI *S3_224* might be involved in low-temperature responsiveness, MeJA-responsiveness, abscisic acid responsiveness because of its cis-regulatory elements and affect root growth and elongation in response to stressful conditions as the manner of RICE SALT SENSITIVE3 (RSS3) in rice (Toda et al., 2013). These findings will facilitate the understanding of the molecular basis of the G x E observed in PH and EH.

Corn belt-specific QTNs

As has been partly described in the Material and Method section, the summer corn-belt average and spring corn-belt average were used to identify QEI, which is defined as the QTN that shows significant QTN x corn-belt interaction by IIIVmrMLM. When QTN x environment interaction is significant, the significant positive and negative genotype effects were canceled out during averaging, therefore the QTN main effects become less meaningful. We obtain corn belt specific QTNs by feeding the IIIVmrMLM software with multi-location data within a corn belt. A corn belt specific QTN is a QTN that shows a significant genotype effect within either summer or spring corn belt data. QEIs explain the phenotypic plasticity across different corn belts and are frequently the targets to select against by breeders seeking stress tolerance and trait stability whereas corn-belt specific QTNs expain the genetic variation within a corn-belt and are frequently targets to select for by breeders seeking genetic gain and stable phenotypic performance in the corresponding corn belt.

We have identified a set of main effect QTNs or corn belt specific QTNs. In the Summer Corn Belt, four candidate genes *Zm00001d018614*, *Zm00001d018615*, *Zm00001d018616*, and *Zm00001d018617* are identified surrounding QTNs *S7_1* (Figures 6, S6). *Zm00001d018617* is also identified by Zhang et al. (2019) as a candidate gene for PH. *Zm00001d033230* surrounding QTN S1_255 (V4:chr1: 255244221, Tables 3, S3; Figure 3) is associated with PH in the Summer Corn Belt in our study, which is also identified as a candidate gene associated with PH in Zmdle1, a dwarf and low ear maize mutant (Zhou et al., 2023). *Zm00001d049715* (*IAA25*) surrounding QTN *S4_41* is associated with EH in the Spring Corn Belt, which is also identified as a candidate gene for PH by Zheng et al. (2016) through meta-QTL analysis.

3VmrMLM multi-environment GWAS models

The selection of appropriate statistical models to detect and measure association is critical to the success of GWAS. The models should be able to deal with various features of phenotypic and genotype data, such as continuity and normality of phenotypic data, population structure and kinship in genotype data, and various confoundings from other covariables in a model. The R software package provided by Zhang's group, IIIVmrMLM V1.0 (Li et al., 2022a; Li et al., 2022b), is a GWAS model that fits the data of strong G x E. Under the framework of a compressed variance component mixed model, each marker on the maize chromosome was first scanned for statistical significance and a less stringent Banforroni correction was adopted in the statistical test and the significant marker loci identified were then incorporated into a new multi-locus genetic model and their effects were estimated by Empirical Bays and all non-zero effects were further evaluated by the likelihood ratio test. Another feature of the 3VmrMLM model is that it can take advantage of heterozygosity discovered in genomic sequence

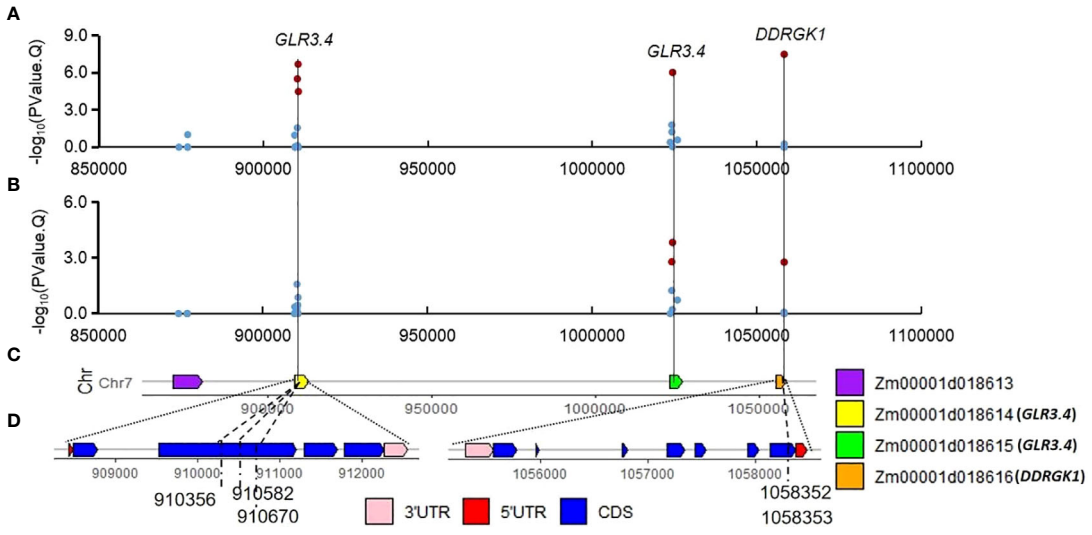


FIGURE 6
Significant QTN S7_1 and associated SNPs on candidate gene *Zm00001d018614* (*GLR3.4*) and *Zm00001d018616* (DDRGK domain-containing protein). **(A, B)** associations of the 28 SNPs for PH **(A)** and EH **(B)** in Summer Corn Belt; **(C)** gene distribution around S7_1 (V4:chr7:910582); **(D, E)** gene structure of *Zm00001d018614* **(D)** and *Zm00001d018616* **(E)**.

data. Heterozygosity has been detected in many DNA sequence projects in corn inbred lines that have been selfed for 6-10 generations, Traditionally, this so-called residual heterozygosity is treated as sequencing errors, or as missing data and is filtered out and ignored. The recent hi-fi sequencing technology has shown this heterozygosity is not a sequencing error and is instead a true variation in inbred lines. The 3VmrMLM model can utilize this important information to reveal QTN x QTN and QTN x environment interaction.

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: European Variation Archive (EVA) at EMBL-EBI, accession number is PRJEB64281 (The European Bioinformatics Institute< EMBL-EBI).

Author contributions

GS: Writing – original draft, Writing – review & editing. AifangW: Writing – review & editing. XW: Data curation, Investigation, Writing – original draft. RC: Validation, Writing – review & editing. FG: Validation, Writing – review & editing. AifenW: Writing – review & editing. TL: Data curation, Writing – review & editing. YW: Funding acquisition, Supervision, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

Authors GS, AifangW, and TL were employed by Beijing Lantron Seed, LongPing High-tech Corp. Authors XW, RC, FG, AifenW and YW were employed by Henan LongPing-Lantron AgriScience and Technology Co., LTD.

The remaining 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.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1284403/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Frequency distribution of minor allele and proportion of heterozygous genotypes in 203 maize inbred lines based on 73175 SNPs dataset. **(A)** minor allele frequency; **(B)** proportion of heterozygous genotypes.

SUPPLEMENTARY FIGURE 2
Linkage disequilibrium decay and 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)** population structure of the 203 lines at K = 3.

SUPPLEMENTARY FIGURE 3
Manhattan Plot of QEIs and associated known candidate genes. **(A)** QEIs and their associated genes for PH identified from mean values of PH in Summer Corn Belt (E1) and Spring Corn Belt (E2). **(B)** QEIs and their associated genes for EH from mean values of EH in Summer Corn Belt (E1) and Spring Corn Belt (E2).

SUPPLEMENTARY FIGURE 4
Association of SNPs surrounding significant QEI *S3_224* with candidate genes. **(A)** associations of the fourteen SNPs using mean values of EH in Summer Corn Belt (E1) and Spring Corn Belt (E2). **(B)** gene structure of *Zm00001d044272*(*bhlh94*).

SUPPLEMENTARY FIGURE 5
Tissue-specific expression profiles of candidate genes around QTN *S10_4* retrieved from maizeGDB.**(A)** *Zm00001d049690* **(B)** *Zm00001d049691* **(C)** *Zm00001d049692*.

SUPPLEMENTARY FIGURE 6
Tissue-specific expression profiles of candidate genes around QTN *S7_1* retrieved from maizeGDB. **(A)** *Zm00001d018614* **(B)** *Zm00001d018615* **(C)** *Zm00001d018616* **(D)** *Zm00001d018617*.

SUPPLEMENTARY FIGURE 7
Association of SNPs surrounding significant QTN S10_15 with candidate genes and their haplotype Effects. **(A)** associations of the SNPs surrounding S10_15 for PH in Spring Corn Belt. **(B)** proportion of heterozygous genotypes of the SNPs surrounding S10_15. **(C)** gene structure of *Zm00001d023677*. **(D)** haplotypes of the three significant SNPs. **(E)** boxplots of haplotypes for PH in five locations.

References

Alkhalifah, N., Campbell, D. A., Falcon, C. M., Gardiner, J. M., Miller, N. D., Romay, M. C., et al. (2018). Maize genomes to fields: 2014 and 2015 field season genotype, phenotype, environment, and inbred ear image datasets. *BMC Res. Notes.* 11 (1), 452. doi: 10.1186/s13104-018-3508-1

Bai, W., Zhang, H., Zhang, Z., Teng, F., Wang, L., Tao, Y., et al. (2010). The evidence for non-additive effect as the main genetic component of plant height and ear height in maize using introgression line populations. *Plant Breed.* 129 (4), 376–384. doi: 10.1111/j.1439-0523.2009.01709.x

Barrett, J. C., Fry, B., Maller, J., and Daly, M. J. (2005). Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263–265. doi: 10.1093/bioinformatics/bth457

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 (19), 2633–2635. doi: 10.1093/bioinformatics/btm308

Camacho, C. G., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinf.* 10, 421. doi: 10.1186/1471-2105-10-421

Coles, N. D., McMullen, M. D., Balint-Kurti, P. J., Pratt, R. C., and Holland, J. B. (2010). Genetic control of photoperiod sensitivity in maize revealed by joint multiple population analysis. *Genetics* 184 (3), 799–812. doi: 10.1534/genetics.109.110304

Cooper, M., and DeLacy, I. H. (1994). Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. *Theor. Appl. Genet.* 88, 561–572. doi: 10.1007/BF01240919

Dai, M., Zhao, J., Yang, G., and Wang, R. (2010). Comparison between different ecological regions on maize yield and agronomic characters. *Chin. Agric. Sci. Bull.* 26 (11), 127–131. doi: 10.11924/j.issn.1000-6850.2009-2795

Ding, X., Wu, X., Chen, L., Li, C., Shi, Y., Song, Y., et al. (2017). Both major and minor qtl associated with plant height can be identified using near-isogenic lines in maize. *Euphytica* 213 (1), 1–9. doi: 10.1007/s10681-016-1825-9

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One* 6 (5), e19379. doi: 10.1371/journal.pone.0019379

Eprintsev, A. T., Fedorin, D. N., Karabutova, L. A., and Pokusina, T. A. (2016). Light regulation of succinate dehydrogenase subunit B gene SDH2-3 expression in maize leaves. *Russ J. Plant Physiol.* 63, 505–510. doi: 10.1134/S102144371604004X

Fedorin, D. N., Eprintsev, A. T., Florez Caro, O. J., and Igamberdiev, A. U. (2023). Effect of salt stress on the activity, expression, and promoter methylation of succinate dehydrogenase and succinic semialdehyde dehydrogenase in maize (*Zea mays* L.) leaves. *Plants* 12 (1), 68. doi: 10.3390/plants12010068

Fei, J., Jiang, Q., Guo, M., Lu, J., Wang, P., Liu, S., et al. (2022). Fine mapping and functional research of key genes for photoperiod sensitivity in maize. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.890780

Frutos, E., Galindo, M. P., and Leiva, V. (2014). An interactive biplot implementation in R for modeling genotype-by-environment interaction. *Stochastic Environ. Res. Risk Assess.* 28, 1629–1641. doi: 10.1007/s00477-013-0821-z

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., et al. (2014). TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PloS One* 9 (2), e90346. doi: 10.1371/journal.pone.0090346

Han, L., Zhong, W., Qian, J., Jin, M., Tian, P., Zhu, W., et al. (2023). A multi-omics integrative network map of maize. *Nat. Genet.* 55 (1), 144–153. doi: 10.1038/s41588-022-01262-1

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

Jin, M., Liu, H., Liu, X., Guo, T., Guo, J., Yin, Y., et al. (2023). Complex genetic architecture underlying the plasticity of maize agronomic traits. *Plant Commun.* 4, 100473. doi: 10.1016/j.xplc.2022.100473

Kong, D., Ju, C., Parihar, A., Kim, S., Cho, D., and Kwak, J. M. (2015). Arabidopsis glutamate receptor homolog3.5 modulates cytosolic Ca2+ level to counteract effect of abscisic acid in seed germination. *Plant Physiol.* 167, 1630–1642. doi: 10.1104/pp.114.251298

Laitinen, R. A. E., and Nikoloski, Z. (2019). Genetic basis of plasticity in plants. *J. Exp. Bot.* 70 (3), 739–745. doi: 10.1093/jxb/ery404

Lawit, S. J., Wych, H. M., Xu, D., Kundu, S., and Tomes, D. T. (2010). Maize DELLA proteins dwarf plant8 and dwarf plant9 as modulators of plant development. *Plant Cell Physiol.* 51 (11), 1854–1868. doi: 10.1093/pcp/pcq153

Li, W., Ge, F., Qiang, Z., Zhu, L., Zhang, S., Chen, L., et al. (2020). Maize ZmRPH1 encodes a microtubule-associated protein that controls plant and ear height. *Plant Biotechnol. J.* 18, 1345–1347. doi: 10.1111/pbi.13292

Li, X., Guo, T., Mu, Q., Li, X., and Yu, J. (2018). Genomic and environmental determinants and their interplay underlying phenotypic plasticity. *Proc. Natl. Acad. Sci. U.S.A.* 115, 6679–6684. doi: 10.1073/pnas.1718326115

Li, X., Guo, T., Wang, J., Bekele, W. A., Sukumaran, S., Vanous, A. E., et al. (2021). An integrated framework reinstating the environmental dimension for GWAS and genomic selection in crops. *Mol. Plant* 14 (6), 874–887. doi: 10.1016/j.molp.2021.03.010

Li, Z. G., Ye, X. Y., and Qiu, X. M. (2019). Glutamate signaling enhances the heat tolerance of maize seedlings by plant glutamate receptor-like channels-mediated calcium signaling. *Protoplasma* 256, 1165–1169. doi: 10.1007/s00709-019-01351-9

Li, M., Zhang, Y. W., Xiang, Y., Liu, M., and Zhang, Y. M. (2022a). IIIVmrMLM: the R and C++ tools associated with 3VmrMLM, a comprehensive GWAS method for dissecting quantitative traits. *Mol. Plant* 15 (8), 1251–1253. doi: 10.1016/j.molp.2022.06.002

Li, M., Zhang, Y. W., Zhang, Z. C., Xiang, Y., Liu, M. H., Zhou, Y. H., et al. (2022b). A compressed variance component mixed model for detecting QTNs, and QTN-by-environment and QTN-by-QTN interactions in genome-wide association studies. *Mol. Plant* 15 (4), 630–650. doi: 10.1016/j.molp.2022.02.012

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. doi: 10.3389/fpls.2016.00833

Liang, Y., Liu, Q., Wang, X., Huang, C., Xu, G., Hey, S., et al. (2019). ZmMADS69 functions as a flowering activator through the ZmRap2.7-ZCN8 regulatory module and contributes to maize flowering time adaptation. *New Phytol.* 221, 2335–2347. doi: 10.1111/nph.15512

Lin, X., Fang, C., Liu, B., and Kong, F. (2021). Natural variation and artificial selection of photoperiodic flowering genes and their applications in crop adaptation. *aBIOTECH* 2 (2), 156–169. doi: 10.1007/s42994-021-00039-0

Liu, N., Du, Y., Warburton, M. L., Xiao, Y., and Yan, J. (2020b). Phenotypic plasticity contributes to maize adaptation and heterosis. *Mol. Biol. Evol.* 38 (4), 1262–1275. doi: 10.1093/molbev/msaa283

Liu, W., Peng, B., Song, A., Jiang, J., and Chen, F. (2020a). Sugar transporter, CmSWEET17, promotes bud outgrowth in Chrysanthemum Morifolium. *Genes (Basel)* 11 (1), 26. doi: 10.3390/genes11010026

Luo, J., Zhou, J., and Zhang, J. (2018). Aux/IAA gene family in plants: molecular structure, regulation, and function. *Int. J. Mol. Sci.* 19 (1), 259. doi: 10.3390/ijms19010259

Malosetti, M., Ribaut, J.-M., and Eeuwijk, F. A. V. (2013). The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. *Front. Physiol.* 4. doi: 10.3389/fphys.2013.00044

Mather, K., and Caligari, P. D. S. (1974). Genotype x environment interactions. *Heredity* 33 (1), 43–59. doi: 10.1038/hdy.1974.63

Mu, Q., Guo, T., Li, X., and Yu, J. (2022). Phenotypic plasticity in plant height shaped by interaction between genetic loci and diurnal temperature range. *New Phytol.* 233, 1768–1779. doi: 10.1111/nph.17904

Multani, D. S., Briggs, S. P., Chamberlin, M. A., Blakeslee, J. J., Murphy, A. S., and Johal, G. S. (2003). Loss of an MDR transporter in compact stalks of maize *br2* and sorghum *dw3* mutants. *Science* 302 (5642), 81–84. doi: 10.1126/science.1086072

Myster, J., and Moe, R. (1995). Effect of diurnal temperature alternations on plant morphology in some greenhouse crops-a mini review. *Scientia Hortic.* 62 (4), 205–215. doi: 10.1016/0304-4238(95)00783-P

Napier, J. D., Heckman, R. W., and Juenger, T. E. (2023). Gene-by-environment interactions in plants: Molecular mechanisms, environmental drivers, and adaptive plasticity. *THE Plant Cell* 35, 109–124. doi: 10.1093/plcell/koac322

Niu, Y., Xu, Y., Liu, X., Yang, S., Wei, S., Xie, F., et al. (2013). Association mapping for seed size and shape traits in soybean cultivars. *Mol. Breed.* 31 (4), 785–794. doi: 10.1007/s11032-012-9833-5

Osnato, M., Cota, I., Nebhnani, P., Cereijo, U., and Pelaz, S. (2022). Photoperiod control of plant growth: flowering time genes beyond flowering. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.805635

Paciorek, T., Chiapelli, B. J., Wang, J. Y., Paciorek, M., Yang, H. P., Sant, A., et al. (2022). Targeted suppression of gibberellin biosynthetic genes *ZmGA20ox3* and *ZmGA20ox5* produces a short staturemaize ideotype. *Plant Biotechnol. J.* 20, 1140–1153. doi: 10.1111/pbi.13797

Perrier, L., Rouan, L., Jaffuel, S., Clement-vidal, A., Roques, S., Soutiras, A., et al. (2017). Plasticity of sorghum stem biomass accumulation in response to water deficit: a multiscale analysis from internode tissue to plant level. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.01516

Pigliucci, M. (2001). *Phenotypic plasticity: beyond nature and nurture* (Baltimore, MD, USA: John Hopkins University Press).

Rogers, A. R., Dunne, J. C., Romay, C., Bohn, M., Buckler, E. S., Ciampitti, I. A., et al. (2021). The importance of dominance and genotype-by-environment interactions on grain yield variation in a large-scale public cooperative maize experiment. *G3* 11 (2), 1–17. doi: 10.1093/g3journal/jkaa050

Schneider, H. M. (2022). Characterization, costs, cues and future perspectives of phenotypic plasticity. *Ann. Bot.* 130, 131–148. doi: 10.1093/aob/mcac087

Seyfferth, C., Renema, J., Wendrich, J., Eekhout, T., Seurinck, R., Vandamme, N., et al. (2021). Advances and opportunities in single-cell transcriptomics for plant research. *Ann. Rev. Plant Biol.* 72, 847–866. doi: 10.1146/annurev-arplant-081720-010120

Shu, G., Cao, G., Li, N., Wang, A., Wei, F., Li, T., et al. (2021). Genetic variation and population structure in China summer maize germplasm. *Sci. Rep.* 11 (1), 8012. doi: 10.1038/s41598-021-84732-6

Shu, G., and Fan, L. (1986). Analysis of the genotype x environment interaction of fertility restoration in Timopheevi CMS hybrid wheat. *Acta Genetica Sin.* 13 (6), 437–446.

Shu, G., Wang, A., Wang, X., Ding, J., Chen, R., Gao, F., et al. (2023). Identification of southern corn rust resistance QTNs in Chinese summer maize germplasm via multi-locus GWAS and post-GWAS analysis. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1221395

Si, W., Wang, H., Dong, J., Li, L., Chen, J., Cheng, B., et al. (2020). Identification of key plant height related ZmGRF genes in maize by weighted gene co-expression network analysis. *J. Maize Sci.* 28 (5), 39–46. doi: 10.13597/j.cnki.maize.science.20200507

Su, H., Liang, J., Abou-Elwafa, S. F., Cheng, H., Dou, D., Ren, Z., et al. (2021). ZmCCT regulates photoperiod-dependent flowering and response to stresses in maize. *BMC Plant Biol.* 21 (1), 1–15. doi: 10.1186/s12870-021-03231-y

Sun, H. Y., Wang, C. L., Chen, X. Y., Liu, H. B., Huang, Y. M., Li, S. X., et al. (2020). dlf1 promotes floral transition bydirectly activating ZmMADS4 and ZmMADS67 in the maize shoot apex. *New Phytol.* 228 (4), 1386–1400. doi: 10.1111/nph.16772

Teale, W. D., Paponov, I. A., and Palme, K. (2006). Auxin in action: signalling, transport and the control of plant growth and development. *Nat. Rev. Mol. Cell Biol.* 7 (11), 847–859. doi: 10.1038/nrm2020

Teng, F., Zhai, L., Liu, R., Bai, W., Wang, L., Huo, D., et al. (2013). *ZmGA3ox2*, a candidate gene for a major QTL, *qPH3.1*, for plant height in maize. *Plant J.* 73, 405–416. doi: 10.1111/tpj.12038

Toda, Y., Tanaka, M., Ogawa, D., Kurata, K., and Takeda, S. (2013). RICE SALT SENSITIVE3 forms a ternary complex with JAZ and class-C bHLH factors and regulates jasmonate-induced gene expression and root cell elongation. *Plant Cell* 25 (5), 1709–1725. doi: 10.1105/tpc.113.112052

Wallace, J. G., Zhang, X., Beyene, Y., Semagn, K., Olsen, M., Prasanna, B. M., et al. (2016). Genome-wide association for plant height and flowering time across 15 tropical maize populations under managed drought stress and well-watered conditions in Sub-Saharan Africa. *Crop Sci.* 56 (5), 1–14. doi: 10.2135/cropsci2015.10.0632

Wang, W., Guo, W., Le, L., Yu, J., Wu, Y., Li, D., et al. (2023). Integration of high-throughput phenotyping, GWAS, and predictive models reveals the genetic architecture of plant height in maize. *Mol. Plant* 16 (2), 354–373. doi: 10.1016/j.molp.2022.11.016

Wang, S., and Wang, Y. (2022b). Harnessing hormone gibberellin knowledge for plant height regulation. *Plant Cell Rep.* 41 (10), 1945–1953. doi: 10.1007/s00299-022-02904-8

Wang, F., Yu, Z., Zhang, M., Wang, M., Lu, X., Liu, X., et al. (2022a). ZmTE1 promotes plant height by regulating intercalary meristem formation and internode cell elongation in maize. *Plant Biotechnol. J.* 20, 526–537. doi: 10.1111/pbi.13734

Wang, Y., Zhao, J., Lu, W., and Deng, D. (2017). Gibberellin in plant height control: old player, new story. *Plant Cell Rep.* 36 (3), 391–398. doi: 10.1007/s00299-017-2104-5

Wei, L., Zhang, X., Zhang, Z., Liu, H., and Lin, Z. (2018). A new allele of the Brachytic2 gene in maize can efficiently modify plant architecture. *Heredity* 121, 75–86. doi: 10.1038/s41437-018-0056-3

Wu, H., Bai, B., Lu, X., and Li, H. (2023). A gibberellin-deficient maize mutant exhibits altered plant height, stem strength and drought tolerance. *Plant Cell Rep.* 42 (10), 1687–1699. doi: 10.1007/s00299-023-03054-1

Xiao, Y., Liu, H., Wu, L., Warburton, M., and Yan, J. (2017). Genome-wide association studies in maize: praise and stargaze. *Mol. Plant* 10, 359–374. doi: 10.1016/j.molp.2016.12.008

Xing, A., Gao, Y., Ye, L., Zhang, W., Cai, L., Ching, A., et al. (2015). A rare SNP mutation in Brachytic2 moderately reduces plant height and increases yield potential in maize. *J. Exp. Bot.* 66 (13), 3791–3802. doi: 10.1093/jxb/erv182

Xuan, Y. H., Hu, Y. B., Chen, L. Q., Sosso, D., Ducat, D. C., Hou, B. H., et al. (2013). Functional role of oligomerization for bacterial and plant SWEET sugar transporter family. *PNAS* 110 (39), E3685–E3694. doi: 10.1073/pnas.1311244110

Yan, W. (2015). Mega-environment analysis and test environment evaluation based on unbalanced multiyear data. *Crop Sci.* 55, 113–122. doi: 10.2135/cropsci2014.03.0203

Yan, W., and Kang, M. S. (2003). *GGE biplot analysis: A graphical tool for breeders, geneticists, and agronomists* (Boca Raton, FL: CRC Press).

Yan, W., Pageau, D., Frégeau-Reid, J., and Durand, J. (2011). Assessing the representativeness and repeatability of test environments for genotype evaluation. *Crop Sci.* 51, 1603–1610. doi: 10.2135/cropsci2011.01.0016

Yan, W., and Rajcan, I. (2002). Biplot analysis of test sites and trait relations of soybean in Ontario. *Crop Sci.* 42, 11–20. doi: 10.2135/cropsci2002.0011

Zhang, S., Hao, H., Liu, X., Li, Y., Ma, X., Liu, W., et al. (2021). *SDG712*, a putative h3k9-specific methyltransferase encoding gene, delays flowering through repressing the expression of florigen genes in rice. *Rice* 14, 73. doi: 10.1186/S12284-021-00513-9

Zhang, Y., Li, Y., Wang, Y., Peng, B., Liu, C., Liu, Z., et al. (2011). Correlations and QTL detection in maize family per se and testcross progenies for plant height and ear height. *Plant Breed.* 130, 617–624. doi: 10.1111/j.1439-0523.2011.01878.x

Zhang, Y., Wan, J., He, L., Lan, H., and Li, L. (2019). Genome-wide association analysis of plant height using the maize f1 population. *Plants* 8, 432. doi: 10.3390/plants8100432

Zhang, J., Wang, S., Wu, X., Han, L., Wang, Y., and Wen, Y. (2022). Identification of QTNs, QTN-by environment interactions and genes for yield-related traits in rice using 3VmrMLM. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.995609

Zhao, Q., Shi, X.-S., Wang, T., Chen, Y., Yang, R., Mi, J., et al. (2023). Identification of QTNs, QTN-byenvironment interactions, and their candidate genes for grain size traits in main crop and ratoon rice. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1119218

Zhao, Z., Yu, Y., Meyer, D., Wu, C., and Shen, W. H. (2005). Prevention of early flowering by expression of FLOWERING LOCUS C requires methylation of histone H3 K36. *Nat. Cell Biol.* 7 (12), 1256–1260. doi: 10.1038/ncb1329

Zheng, L., Zhou, Y., Zeng, X., Di, H., Weng, J., Li, X., et al. (2016). QTL mapping of plant height in maize. *Crops* 2016 (2), 8–13. doi: 10.16035/j.issn.1001-7283.2016.02.002

Zhou, W., Zhang, H., He, H., Gong, D., Yang, Y., Liu, Z., et al. (2023). Candidate gene localization of ZmDLE1 gene regulating plant height and ear height in maize. *Scientia Agricultura Sin.* 56 (5), 821–837. doi: 10.3864/j.issn.0578-1752.2023.05.002

Zuo, J.-F., Chen, Y., Ge, C., Liu, J.-Y., and Zhang, Y.-M. (2022). Identification of QTN-by-environment interactions and their candidate genes for soybean seed oil-related traits using 3VmrMLM. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.1096457

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 08 January 2024
DOI 10.3389/fpls.2023.1283642

OPEN ACCESS

EDITED BY
Yuan-Ming Zhang,
Huazhong Agricultural University, China

REVIEWED BY
Wenlong Ren,
Nantong University, China
Hai Yan Lü,
Henan Agricultural University, China
Liu Jinyang,
Jiangsu Academy of Agricultural Sciences (JAAS), China

*CORRESPONDENCE
Yang-Jun Wen
wenyangjun@njau.edu.cn

†These authors have contributed equally to this work

RECEIVED 26 August 2023
ACCEPTED 15 December 2023
PUBLISHED 08 January 2024

CITATION
Han L, Shen B, Wu X, Zhang J and Wen Y-J (2024) Compressed variance component mixed model reveals epistasis associated with flowering in *Arabidopsis*. *Front. Plant Sci.* 14:1283642. doi: 10.3389/fpls.2023.1283642

Compressed variance component mixed model reveals epistasis associated with flowering in *Arabidopsis*

Le Han[1†], Bolin Shen[1†], Xinyi Wu[1], Jin Zhang[1,2] and Yang-Jun Wen[1,2*]

[1]College of Science, Nanjing Agricultural University, Nanjing, China, [2]State Key Laboratory of Crop Genetics and Germplasm Enhancement and Utilization, Nanjing Agricultural University, Nanjing, China

Introduction: Epistasis is currently a topic of great interest in molecular and quantitative genetics. *Arabidopsis thaliana*, as a model organism, plays a crucial role in studying the fundamental biology of diverse plant species. However, there have been limited reports about identification of epistasis related to flowering in genome-wide association studies (GWAS). Therefore, it is of utmost importance to conduct epistasis in *Arabidopsis*.

Method: In this study, we employed *Levene's* test and compressed variance component mixed model in GWAS to detect quantitative trait nucleotides (QTNs) and QTN-by-QTN interactions (QQIs) for 11 flowering-related traits of 199 *Arabidopsis* accessions with 216,130 markers.

Results: Our analysis detected 89 QTNs and 130 pairs of QQIs. Around these loci, 34 known genes previously reported in *Arabidopsis* were confirmed to be associated with flowering-related traits, such as *SPA4*, which is involved in regulating photoperiodic flowering, and interacts with *PAP1* and *PAP2*, affecting growth of *Arabidopsis* under light conditions. Then, we observed significant and differential expression of 35 genes in response to variations in temperature, photoperiod, and vernalization treatments out of unreported genes. Functional enrichment analysis revealed that 26 of these genes were associated with various biological processes. Finally, the haplotype and phenotypic difference analysis revealed 20 candidate genes exhibiting significant phenotypic variations across gene haplotypes, of which the candidate genes *AT1G12990* and *AT1G09950* around QQIs might have interaction effect to flowering time regulation in *Arabidopsis*.

Discussion: These findings may offer valuable insights for the identification and exploration of genes and gene-by-gene interactions associated with flowering-related traits in *Arabidopsis*, that may even provide valuable reference and guidance for the research of epistasis in other species.

KEYWORDS

epistasis, GWAS, 3VmrMLM, *Arabidopsis thaliana*, flowering-related traits

Introduction

Arabidopsis thaliana, an important flowering plant, has emerged as a model organism for molecular plant genetics research in recent years (Koornneef and Meinke, 2010). Its compact genome, short life cycle, ease of cultivation, and abundant genetic resources make it widely utilized in fundamental biology, crop enhancement, and biotechnology. The flowering phase of *Arabidopsis* plays a crucial role in determining the precise timing of reproduction, seed, and fruit development. Therefore, studying the regulation and molecular mechanisms of flowering time in *Arabidopsis* remains an important area of research. By discovering the genetic factors and regulatory pathways affecting flowering time in *Arabidopsis*, it is possible to identify homologous genes and manipulate their expression in agronomic crops, optimize crop flowering time to adapt to specific environments and agricultural practices, improve crop yields, and produce crops that are more adapted to climate change and stress resistance.

Flowering in *Arabidopsis* has complex regulatory mechanisms and pathways, and the phenotypic material of flowering under different regulatory pathways is particularly important to elucidate the genetic mechanism of flowering (Qi et al., 2018). In the photoperiodic pathway, *Arabidopsis* perceives light signals through photoreceptors and transmits them to its biological clock. The biological clock, responsive to changes in day length, ultimately transforms the light signals into flowering signals via the CONSTANS (CO) gene (Imaizumi and Kay, 2006). Under long-day treatments, the CO gene facilitates flowering, whereas under short-day treatments, it retards the process (Teper-Bamnolker and Samach, 2005; Balasubramanian et al., 2006). In addition, vernalization plays a vital role in regulating flowering. By suppressing the activity of the FLOWERING LOCUS C protein, low-temperature induction during vernalization unlocks *Arabidopsis*'s flowering potential (Helliwell et al., 2015). In additional to the vernalization pathway, it was shown that the flowering time of *Arabidopsis* in 25-27°C short days was similar that in 23°C long days, suggesting that higher temperature promotes flowering in *Arabidopsis* (Balasubramanian et al., 2006). These studies indicate that in the research on flowering-related traits of *Arabidopsis*, factors such as photoperiod, vernalization, and temperature need to be considered.

Epistasis, referred to as loci-locus interactions (He et al., 2019), plays an important role in phenotypic variation and has received much attention over the years. As a major factor in molecular evolution (Breen et al., 2012), epistasis plays a crucial role in quantitative genetic analysis and is now one of the main causes of 'missing heritability' (Mackay and Moore, 2014; Upton et al., 2016). In *Arabidopsis*, flowering time as a complex quantitative trait is regulated by genes such as photoperiod, but also by other physiological processes such as temperature signaling and vernalization, which are both independent and interrelated. Therefore, these physiological processes involve a large number of loci and even genes that often interact with each other, and individual genetic loci or genes may have a small effect on flowering time in *Arabidopsis*, but together with other genes may have a large effect on phenotypic variation (Zhang et al., 2014), making it particularly important to investigate epistatic loci for flowering-related traits in *Arabidopsis*.

Recently, researchers have proposed many epistasis detection algorithms for complex traits based on traditional genome-wide association studies (GWAS) or artificial intelligence (AI). The most basic approach to explore epistasis is regression-based methods such as PLINK (Purcell et al., 2007), which has the advantage of high computational efficiency, rapid analysis of tens of thousands of markers and epistasis, and wide application in case-control datasets, but a high false positive rate. BOOST (Wan et al., 2010), which uses a Boolean representation of genotype data, can save memory space and improve computational speed at the same time, but it can only handle binary phenotype data and not for continuous quantitative traits such as yield and flowering time, which is a very limited application scenario. For continuous traits in plants, mixed linear model (MLM)-based methods perform better due to accounting for environmental factors, controlling for population stratification, and explaining cryptic correlations among individuals. QTXNetwork is a multi-locus mixed model proposed by Zhang et al. (2015). This method first detects each marker to identify potential quantitative trait nucleotides (QTNs), QTN-by-environment interactions (QEIs), and all the pairs of markers to identify potential QTN-by-QTN interactions (QQIs), and then all the potential QTNs, QEIs, and QQIs are placed into a genetic model to identify significant loci. However, the associated polygenic backgrounds in the first step were not taken into account. Ning et al. (2018) proposed a rapid epistatic mixed-model association analysis (REMMA) algorithm, which used the best linear unbiased prediction (BLUP) to predict additive and dominant effects, their epistatic effects and their variances, and then Wald Chi-squared test was used to identify the significance of all the effects. However, their power could be further improved. Multifactor dimensionality reduction (MDR) (Moore, 2004), a classical nonparametric machine learning method, was originally designed for identifying epistasis in case-control studies. Quantitative MDR (QMDR) (Gui et al., 2013; Yu et al., 2015) represents a robust, model-free extension of MDR accommodated for quantitative phenotypes. None of them, however, effectively address the challenges posed by limited interpretability and overfitting in AI and lengthy computation times required for genome-wide markers.

To overcome the above issues, Li et al (2022a; 2022b). established a compressed variance component mixed model method, named 3VmrMLM, to detect QTNs, QEIs, and QQIs while controlling for all the possible polygenic backgrounds. It reveals epistatic effects by reducing the number of variance components, while ensuring high statistical power. Additionally, the method efficiently reduces computation time and effectively addresses potential confounding factors arising from various polygenic backgrounds.

A number of gene-by-gene interactions associated with flowering time have been identified in *Arabidopsis*. For example, Zhao et al. (2022) identified a novel flowering repressor, *UBA2c*, and showed that the expression of a key flowering repressor gene, *FLM*, is promoted by inhibiting the histone modification *H3K27me3*, thereby suppressing premature flowering in plants.

Hanano and Goto (2011) found that the interaction of *FD* with *TFL1* by BiFC assay induces *Arabidopsis* flowering repressor genes to fine-tune flowering time and inflorescence meristem tissue development, which in turn affects flowering time. However, most gene-by-gene interactions related flowering in *Arabidopsis* have been obtained by biological methods such as transcriptome analysis, and few gene-by-gene interactions have been identified by GWAS.

In this study, QQIs and QTNs for eleven flowering-related traits in natural populations of *Arabidopsis* were investigated using 3VmrMLM with data from https://www.Arabidopsis.org. Differentially expressed genes were identified under temperature, photoperiod, and vernalization treatments. Candidate genes and gene-by-gene interactions were identified by functional enrichment, haplotype and phenotypic difference analysis. Epistasis for flowering-related traits of *Arabidopsis* will help identify interacting genes and provide references for studying epistasis in other crops.

Materials and methods

Genotypic and phenotypic data

The dataset of *Arabidopsis* (Atwell et al., 2010) including the phenotypic and genotypic data were obtained from https://www.Arabidopsis.org. The dataset consisted 23 flowering-related traits, 199 individuals, and 216,130 markers.

Among 23 traits, we focused on eleven traits related to flowering under three different environmental conditions, including temperature, photoperiod, and vernalization treatments. They were Days to flowering time under Long Day (LD), Days to flowering time under Long Day with vernalization at 4°C during 5 weeks (LDV), Days to flowering time under Short Day with vernalization at 4°C during 5 weeks (SDV), Days to FT under LD with vernalization for 0 weeks, 2 weeks, 4 weeks, 8 weeks (0W, 2W, 4W, 8W), Flowering time at 10°C, 22°C (FT10, FT22), leaf number at flowering time at 10°C, 22°C (LN10, LN22) (Supplementary Data.zip).

To explore the relationship among the above flowering-related traits, we computed the Pearson correlation coefficients (PCCs) using the *cor.test* function in R (Version 4.2.1) and generated a phenotypic correlation heatmap using the *ggcorrplot* function from the ggcorrplot package. A hierarchical cluster analysis of the phenotypes was also performed using the *hclust* function in R to divide traits into groups that correlated more significantly into the same group (Figure 1A).

GWAS method

To rapidly and accurately analyze epistasis for GWAS, we combined *Levene's* test (Brown and Forsythe, 1974) with 3VmrMLM. Firstly, we conducted *Levene's* test from the OSCA software tool (http://cnsgenomics.com/software/osca; Zhang et al., 2019) for mining the potential epistatic single nucleotide polymorphisms (SNPs) as well as alleviating computational burden. We utilized "*--vqtl -mtd 2*" for *Levene's* test with median and "*--maf 0.01*" for removing data with minor allele frequency (MAF) < 0.01 in OSCA, resulting in the top 5,000 loci for each trait. Subsequently, we used the *IIIVmrMLM* package (https://github.com/YuanmingZhang65/IIIVmrMLM; Li et al., 2022b) in R to detect QQIs and QTNs, with parameter set to "Epistasis". 3VmrMLM determines the significance of QQIs or QTNs using either Bonferroni correction (P-value < $0.05/[m \times (m-1)]/2$, where m is the number of markers) for significant association or a logarithm of odds (LOD) score of 3.0 for suggestive association, either criterion indicates a significant association with the traits. We used $V_p = V_{epi} + V_{add} + V_r$ (Figure 2) for each trait to calculate the proportion of the sum of epistatic variance (V_{epi}) to the phenotypic variance (V_p), where V_{add} is the sum of additive variance of detected QTNs and V_r is the residual variance.

Identification of known genes

We identified genes located within a 20 kb distance around significant loci, specifically focusing on known genes that have been previously reported in relevant articles. Then the *Arabidopsis* Information Resource (TAIR) (https://www.arabidopsis.org/) and National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/) were employed for gene annotation. Known gene mining involved three steps. First, extracting genes within a 20 kb region around significant loci detected by 3VmrMLM from the *Arabidopsis* gene library downloaded from TAIR. Second, screening for genes impacting flowering-related traits and containing relevant keywords. Third, confirming the association between genes and flowering time in *Arabidopsis*, as well as their confirmed epistatic interactions with other genes by retrieving literature from TAIR and NCBI. Finally, known genes will be identified.

Differential expression and functional enrichment analyses

After excluding known genes reported in the literature, we performed differential expression analysis on the remaining unreported genes using the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/). We utilized the GSE197581, GSE190748, and GSE40455 series for targeting differentially expressed genes (DEGs) in response to different temperature, photoperiod, and vernalization treatments. The GSE197581 series included two samples of *Arabidopsis* at 22°C and 10°C, with three biological replicates. The GSE190748 series consisted samples subjected to long-day (16h light/8h dark) and short-day (8h light/16h dark), with two biological replicates. The GSE40455 series included samples to four weeks of vernalization and samples subjected without vernalization treatment, with four biological replicates. For the GSE190748 and GSE40455 series, we used the "*analyze with GEO2R*" tool to identify genes with an absolute $\log_2$FoldChange greater than 1 and a P-value less than 0.05. For the GSE197581 series, we used the provided data from the website and identify genes with an absolute $\log_2$FoldChange greater than 1

10.3389/fpls.2023.1283642

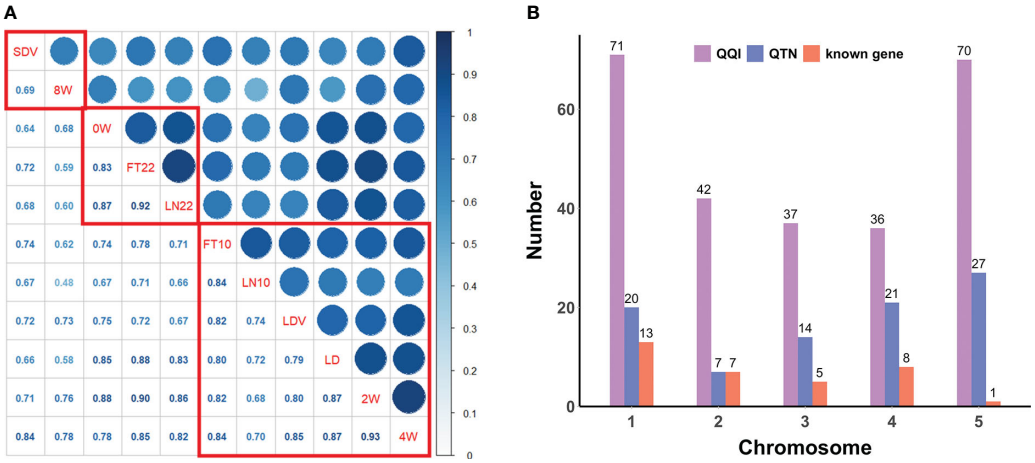


FIGURE 1
(A) Pearson correlation coefficients and correlation clustering of flowering-related traits. The lower diagonal represents the correlation coefficients, and the red boxes indicate the clustering results. **(B)** Distribution of QQIs, QTNs, and known genes across all chromosomes for eleven flowering-related traits.

and the false discovery rate (FDR) less than 0.05. Subsequently, the DEGs obtained above were intersected with the detected unreported genes around QQIs and QTNs, resulting in identification of DEGs associated with flowering-related traits. For gene ontology (GO) based functional enrichment analysis, we submitted the above flowering-related DEGs information to the DAVID platform (https://david.ncifcrf.gov/), and selected the enriched gene ontology terms with a significance threshold of P-value less than 0.05.

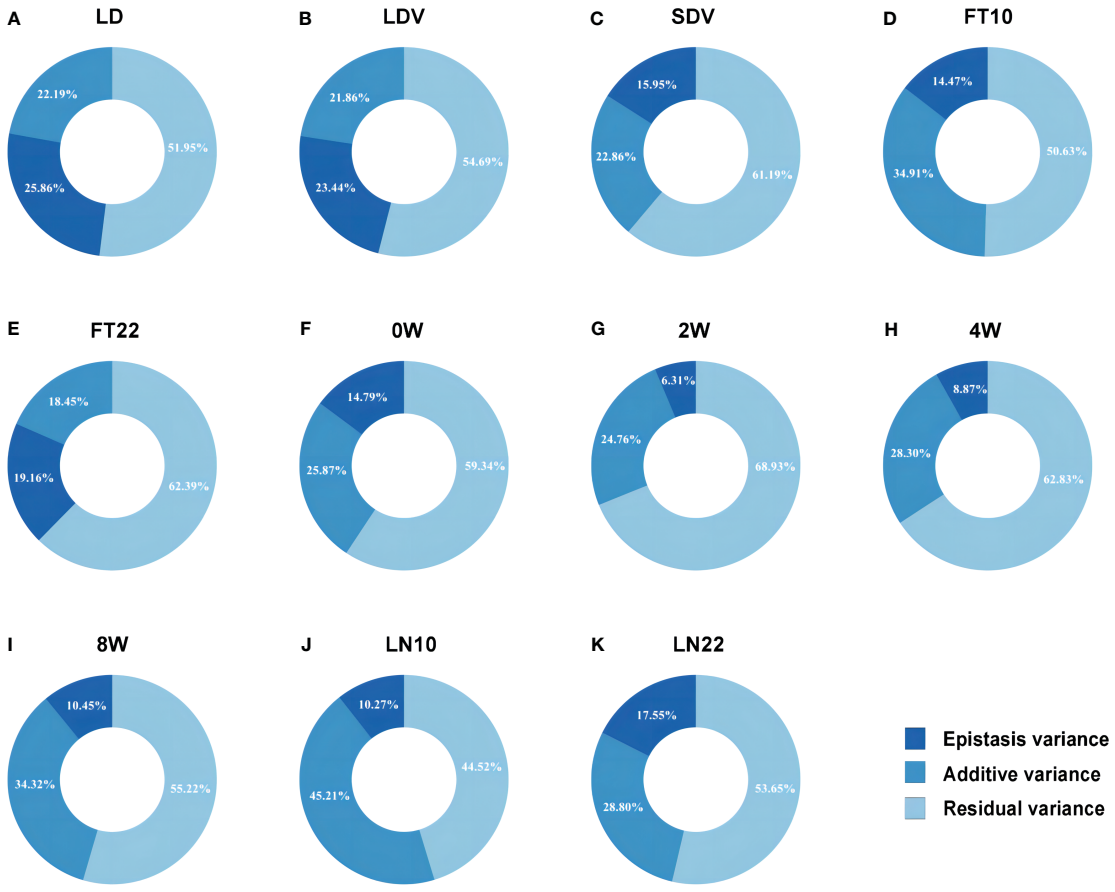


FIGURE 2
Phenotypic variation explained by the epistatic and additive effects for eleven flowering-related traits. **(A–K)** correspond to the traits LD, LDV, SDV, FT10, FT22, 0W, 2W, 4W, 8W, LN10, and LN22, respectively.

Haplotype analysis for identifying candidate genes

We used the HaploView software (Version 4.1) to perform linkage disequilibrium and haplotype block studies (Barrett et al., 2005) based on the SNPs within these genes and 2 kb upstream of them, which are obtained from GO enrichment analysis. Meanwhile, we employed the *t.test* function in R to examine the phenotypic differences among haplotypes. Candidate genes were identified as those exhibiting significant phenotypic differences across various haplotypes. This approach allowed us to identify potential genes associated with the traits of interest.

Results

Phenotypic correlation and clustering

PCCs were obtained from correlation analysis of eleven quantitative traits (Figure 1A). The phenotypic correlations of all flowering-related traits showed positive. There were two pairs of PCCs more than 0.90, 2W and 4W (PCCs = 0.93), FT22 and LN22 (PCCs = 0.92), and only one pair of PCCs less than 0.50, LN10 and 8W, but their PCCs also reached 0.48. The above results indicate that eleven traits play an important role in the regulation of flowering time in *Arabidopsis*, and there is a very significant positive correlation between any two pairs.

Hierarchical cluster analysis of all traits by the *hclust* function in R ranked the phenotypes with more significant correlations and divided them into three groups (Figure 1A). The first group was SDV and 8W with a correlation coefficient of 0.69; the second group was 0W, FT22, and LN22 with PCCs ranging from 0.83 to 0.92; and the third group was FT10, LN10, LDV, LD, 2W, and 4W with PCCs ranging from 0.68 to 0.93. Clustering of these phenotypes revealed a higher overall correlation between these traits and a greater likelihood of interactions between loci, which was further confirmed following by the pleiotropy of known genes (Table 1).

Epistasis mining using 3VmrMLM

After *Levene's* test in the raw dataset, 3VmrMLM used in the top 5,000 markers detected 130 QQIs (107 significant and 23 suggested QQIs; Supplementary Table 1) and 89 QTNs (61 significant and 28 suggested QTNs; Supplementary Table 2) that were strongly associated with the flowering-related traits.

Overall, QQIs and QTNs are distributed on all chromosomes (Figure 1B). For QQIs, 3VmrMLM detected a large number of loci, with the highest distribution on chromosome 1 and 5, with 71 and 70 loci, respectively. Although it has a relatively small distribution on chromosomes 2 and 4, it also has more than 35 loci (Figure 1B). For QTNs, the distribution of loci on chromosome 2 was relatively uniform, with the number ranging from 14 ~27, except for a minimum of 7 loci on chromosome 2 (Figure 1B). On chromosome 1 and chromosome 5, QQIs and QTNs are relatively large, and we can analyze that these two chromosomes have a great influence on the genetic variation of flowering-related traits (Figure 1B). In addition, the number of QQIs far exceeded the number of QTNs, indicating that epistasis is a very important link to explore the genetic mechanism of traits related to flowering time, and the interaction between loci is relatively common.

Six of the 11 traits obtained more than 10 QQIs (Supplementary Table 1). FT22 detected the most QQIs, reaching 19 QQIs, with P values of 2.965E-09~1.386E-04, LOD scores of 3.154~7.645, respectively, and 7 positive effects (Figure 3B; Supplementary Table 1). FT10 detected 11 QQIs with P values of 2.293E-10~9.951E-05 and LOD scores of 3.289~8.730, where SNP72738 on chromosome 2 and SNP167863 on chromosome 5 also were the QQIs for 2W and LN22 traits, respectively (Supplementary Figure 1C; Supplementary Table 1). LN10 detected 16 QQIs, second only to FT22, with P values of 1.327E-10~5.173E-05 and LOD scores of 3.558~8.962, respectively (Figure 3D; Supplementary Table 1). LN22 detected 10 QQIs, with P values of 6.250E-10~1.190E-04 and LOD scores of 3.216~8.304, respectively (Supplementary Figure 1G; Supplementary Table 1). LDV detected 14 QQIs, with P values of 4.326E-15~1.174E-04 and LOD scores of 3.221~13.365, 7 positive effects, respectively (Figure 3A; Supplementary Table 1). SDV detected 14 QQIs, with P values of 4.136E-11~1.379E-04, LOD scores of 3.156~9.457, and 4 positive effects, respectively. Notably, SNP200347 on chromosome 5 was involved in interactions with both SNP179236 and SNP32689. Trait 0W detected 12 QQIs, with P values of 2.605E-14~1.318E-05 and LOD scores of 4.123~12.608, respectively (Supplementary Figure 1D; Supplementary Table 1). Trait 2W detected 14 QQIs, with P values of 3.985E-09~8.515E-05 and LOD scores of 3.353~7.520, respectively, and SNP72738 was found to be involved in intercrossing with SNP2739 and SNP72795 simultaneously in this trait (Figure 3C; Supplementary Table 1).

8 QQIs were detected for both 4W and 8W, with P values of 5.906E-13 ~3.681E-06, LOD scores of 4.652~11.266, respectively, and only 2 positive effects for 4W (Supplementary Figure 1E; Supplementary Table 1). P values of 4.899E-08~1.064E-04 and LOD scores of 3.261~6.462 for 8W (Supplementary Figure 1F; Supplementary Table 1). Although LD obtained the least number of QQIs, only four, with P values of 2.792E-08~8.968E-07 and LOD scores of 5.242~6.699, respectively, the phenotypic contribution of all four pairs of epistatic loci was >4%, with the pair SNP66960 and SNP71678, located on chromosome 2, having the largest percentage of phenotypic variance explained (PVE) of all QQIs at 8.187%. (Supplementary Table 1).

For QTNs, a total of 89 significant/suggestive QTNs were detected to be associated with at least one of the 11 flowering-related traits (Figure 3; Supplementary Figure 1; Supplementary Table 2). Among these QTNs, 3, 4, 8, 10, 6, 6, 13, 11, 11, 13, and 7 QTNs were associated with LD, LDV, SDV, FT10, FT22, 0W, 2W, 4W, 8W, LN10, and LN22, respectively (Supplementary Table 2), and the PVE of all QTNs for each trait were 22.193%, 21.875%, 22.864%, 34.906%, 18.446%, 25.868%, 24.760%, 28.297%, 34.328%, 45.205%, and 28.797%, respectively, with P values ranging from 1.757E-10 to 1.986E-04 and LOD scores of 3.006 to 8.843 (Figure 2; Supplementary Table 2). Notably, SNP31054 and SNP101868 on

10.3389/fpls.2023.1283642

TABLE 1 Pleiotropic genes reported around QQIs/QTNs.

Gene	Bp	Marker	QQI/QTN	Trait	Annotation	Reference
AGL17 (*AT2G22630*)	chr2:9618207..9622163	SNP66970	QQI	LD	MADs domain containing protein involved in promoting flowering	Han et al., 2008
		SNP66990	QQI	LN22		
		SNP67001	QQI	FT22		
LUH (*AT2G32700*)	chr2:13866721..13872246	SNP72705	QTN	2W	WD40 repeat and LUFS domain containing protein that is similar to LUG	Stahle et al., 2009
		SNP72736	QQI	FT22		
		SNP72738	QQI	FT10		
BOP2 (*AT2G41370*)	chr2:17237727..17240609	SNP77354	QQI	2W	cytoplasmic and nuclear-localized NPR1 like protein	Chahtane et al., 2018
		SNP77376	QQI	LN10		
ATH1 (*AT4G32980*)	chr4:15914670..15918153	SNP157833 SNP157883	QQI QQI	LDV	increased levels of ATH1 severely delay flowering	Li et al., 2012
				0W		
CPL3 (*AT4G01060*)	chr4:460395..461246	SNP125917	QTN	2W	Myb-related protein similar to CPC	Zhang and Shen, 2022
		SNP125988	QTN	FT10		

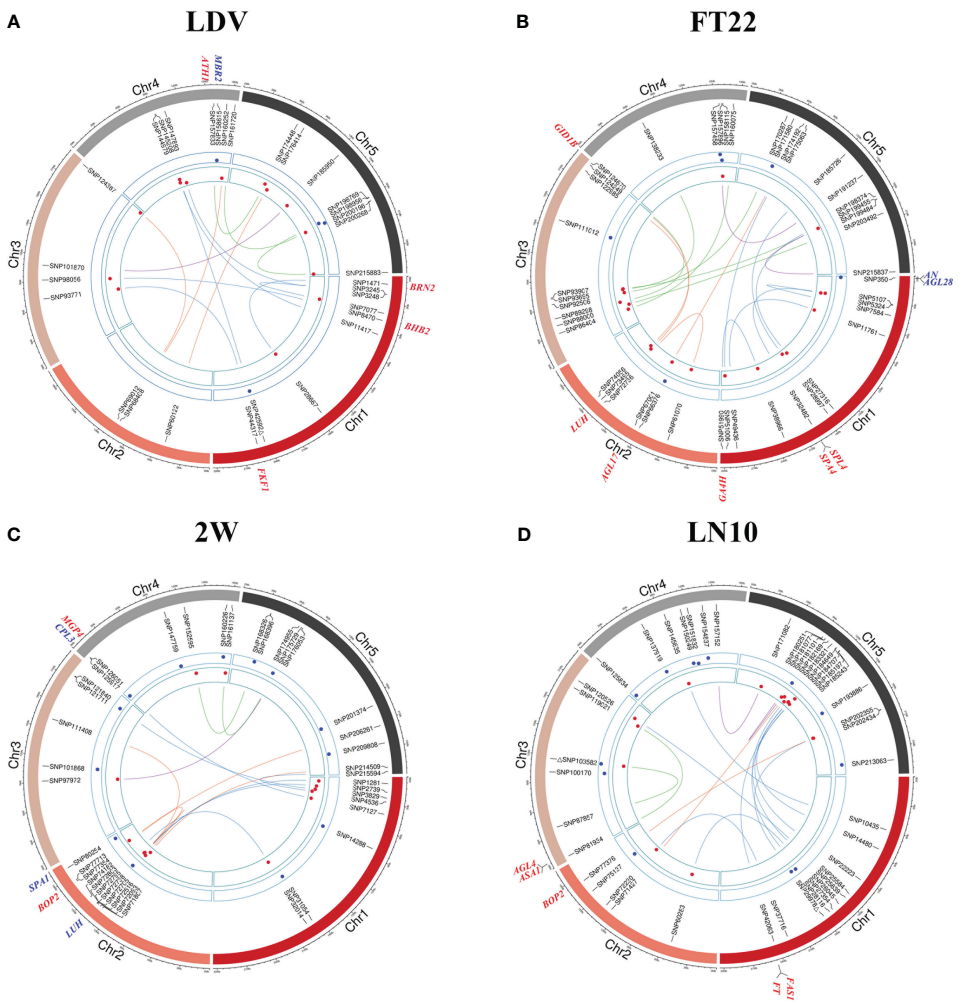


FIGURE 3
Chord diagrams for QQIs and QTNs detected by 3VmrMLM. **(A–D)** correspond to the traits LDV, FT22, 2W, and LN10, respectively. The inner circle displays the detected QQIs or QTNs (△ indicates overlapping loci between QQIs and QTNs), the height of red dots represents the epistatic effects of QQI pairs, and the height of blue dots represents the additive effects of corresponding QTNs. The outer circle indicates the known genes in vicinity of significant loci.

chromosomes 1 and 3 were involved in both 2W and 4W phenotypic variants, and in addition, SNP103582 on chromosome 3 was detected on both LN10 and FT10 (Figure 3D; Supplementary Figure 1C; Supplementary Table 2).

The total PVE for each trait, considering both additive and epistatic effects, was calculated using the IIIVmrMLM package in R, and the results were visualized in Figure 2. The PVE of QQIs for the traits LD, LDV, and FT22 were 25.856%, 23.438%, and 19.163%, respectively, as shown in Figures 2A, B, E. Accordingly, these values were higher than the PVEs of the corresponding QTNs. The analysis of QQIs and QTNs revealed that most locus exhibited either epistatic or additive effects in contributing to phenotypic variation of each trait (Figure 2; Supplementary Tables 1, 2). However, we also identified some specific SNPs, such as SNP42592 for LDV, both SNP103582 and SNP29978 for LN10, SNP200347 for SDV, SNP125854 for 0W, SNP101868 for 4W, both SNP111498 and SNP181717 for 8W, which were involved in both additive and epistatic effects (Figures 3A, D; Supplementary Figures 1B, D, E, F; Supplementary Tables 1, 2).

Known genes around QQIs and QTNs for flowering-related traits in *Arabidopsis*

TAIR (https://www.arabidopsis.org/) was used to mine the known genes around QQIs and QTNs (20 kb upstream and downstream of each locus). A total of 34 known genes were found to be located around the significant/suggested loci, including 29 QQIs and 12 QTNs (Figure 3; Supplementary Figure 1; Supplementary Table 3).

For QQIs, 3, 4, 2, 1, 6, 4, 2, 0, 1, 5, and 1 known genes were explored in LD, LDV, SDV, FT10, FT22, 0W, 2W, 4W, 8W, LN10, and LN22, respectively (Supplementary Table 3). Specifically, the known genes *BRN2* (*AT1G03457*, near SNP1471) and *FKF1* (*AT1G68050*, near SNP44317) associated with LDV (Figure 3A; Supplementary Table 3) interact with the *AtBRN* and *CDF2* protein to promote or repress flowering in *Arabidopsis*, respectively (Kim et al., 2013; Lee et al., 2018). The known gene *SPA4* (*AT1G53090*) associated with FT22 is located near SNP32482 (Figure 3B; Supplementary Table 3). There has been reported that *SPA4* is involved in regulating photoperiodic flowering in *Arabidopsis* and interacts with the flower inducer CO to regulate flowering stability, while it interacts with *PAP1* and *PAP2* and is involved in repressive regulation at the transcriptional level, affecting light conditions growth of *Arabidopsis* under light conditions (Laubinger et al., 2006; Maier et al., 2013). Two known genes, *FT* (*AT1G65480*) and *FAS1* (*AT1G65470*), were detected simultaneously near SNP42063 (Figure 3D; Supplementary Table 3), and two known genes, *ASA1* (*AT3G02260*) and *AGL4* (*AT3G02310*), were detected near SNP81934 under LN10 (Figure 3B; Supplementary Table 3), where *FT* interacts with *FD(AT4G35900)* and 14-3-3 proteins to produce a florigen-activation complex, control flowering time, and correct the expression of floral homologs to promote flowering (Collani et al., 2019); the known gene *AGL4* interacts with DNA and may be involved in forming a tetrameric DNA-binding complex to control flower development and thus affect flowering time (Jetha et al., 2014). The known gene *HOS1* (*AT2G39810*, near SNP76337) associated with trait 0W (Supplementary Figure 1D; Supplementary Table 3) is localized to the nuclear membrane and interacts with *Nup96*, and loss of function of *Nup96* would lead to disruption of *HOS1* protein, resulting in excessive accumulation of CO protein, a key activator of flowering under long-day that suppresses early flowering in *Arabidopsis* under long-day (Lazaro et al., 2015).

For QTNs, 1, 2, 1, 2, 3, 1, and 2 known genes were explored in LDV, SDV, FT10, FT22, 2W, 4W, and LN22, respectively, and only QQI-related genes were obtained for the remaining four traits (Supplementary Table 3). Among the significant loci associated with SDV, *FD* (*AT4G35900*) was found to be located near SNP159681 (Supplementary Figure 1B; Supplementary Table 3), and it was shown that *FD* acts as a transcriptional activator of floral tissue identity genes to regulate flowering time in *Arabidopsis*, while the *FD* transcription factor was shown to interact with *TFL1* by BiFC assay to induce flowering time and inflorescence meristem tissue by *Arabidopsis* repressor genes development is fine-tuned (Hanano and Goto, 2011; Gorham et al., 2018). In the case of FT22, two known genes, *AN* (*AT1G01510*) and *AGL28* (*AT1G01530*), were detected simultaneously near SNP350 (Figure 3B; Supplementary Table 3), and *AN* has been shown to control leaf morphology and thus indirectly affect flowering time in *Arabidopsis*. (Stern et al., 2007); *AGL28* can act as a flower activator by up-regulating the expression of known flower promoters within the autonomous pathway, and its overexpression will up-regulate the expression of *FCA* and *LUMINIDEPENDENS*, leading to early flowering in *Arabidopsis* (Yoo et al., 2006). One known gene associated with LDV, *MBR2* (*AT4G34040*), located near SNP158615 (Figure 3A; Supplementary Table 3), was shown in earlier studies to promote flowering through a *PFT1* dependent and independent mechanism (Iñigo et al., 2012). The gene *SPA1* (*AT2G46340*, near SNP80254) is known to be associated with 2W (Figure 3C; Supplementary Table 3), and is a key repressor of light signaling in the ovary to regulate flowering time by regulating the photoperiod (Ranjan et al., 2011). Near the QTN SNP135761, which is significantly associated with LN22, *CRY1* (*AT4G08920*; Supplementary Figure 1G; Supplementary Table 3) is known to mediate blue light to promote flowering in *Arabidopsis*, which is more sensitive to flowering photoperiod under blue light, suggesting that *CRY1* plays an important role in flowering regulation (Mockler et al., 2003).

Interestingly, out of these 34 known genes, five pleiotropic genes were involved in the performance variation of at least two traits in terms of QQI or QTN (Table 1). In terms of QQI, the known gene *AGL17* (*AT2G22630*), which was detected around SNP67001, SNP66970, and SNP66990 and was associated with FT22, LD, and LN22 (Table 1; Figure 3B; Supplementary Figures 1A, G), has been confirmed to play a role in the photoperiodic pathway of *Arabidopsis* and is positively controlled by the photoperiodic pathway regulator CO. It can promote the flowering of *Arabidopsis thaliana* (Han et al., 2008). At the same time, the known gene *ATH1* (*AT4G32980*, around SNP157833), which is related to LDV and 0W (Table 1; Figure 3A; Supplementary Figure 1D), is necessary for controlling the morphology of

Arabidopsis flower stalk. In addition, there is an interaction between *ATH1* and *KNAT2*, and the protein complex plays a role in regulating flower pedicle development (Li et al., 2012). *BOP2* (*AT2G41370*), detected near SNP77354 and SNP77376, is associated with two traits, 2W and LN10 (Table 1; Figures 3C, D), and studies have shown that the *LFY* and *BOP2* proteins physically interact to inhibit bracteal formation and reduce flowering time in a short period of time under certain conditions (Chahtane et al., 2018). In terms of QTN, a known gene *CPL3* (*AT4G01060*, near SNP125917 and SNP125988) was detected to have additive effects on both 2W and FT10 (Table 1; Figure 3C; Supplementary Figure 1C), and *CPL3* gene has pleiotropic effects on flowering development and epidermal cell size of *Arabidopsis* by regulating internal duplication (Zhang and Shen, 2022).

Notable is, known gene *LUH* (*AT2G32700*), located near SNP72736, SNP72705, and SNP72738, exhibited associations with FT22, 2W, and FT10 (Table 1; Figures 3B, C; Supplementary Figure 1C). Furthermore, it displayed both additive and epistatic effects (Table 1; Figures 3B, C; Supplementary Figure 1C). *LUH* showed epistatic effect at FT10 and FT22, and additive effect at 2W. It was shown that *LUH* interacts with *YAB* to regulate distal axis pattern, lateral organ growth, and inflorescence foliation. At the same time, its leaf-based signaling pathway promotes paraxial cell identity in leaves and initiation and maintenance of embryo bud apical meristem SAM (Stahle et al., 2009). More detailed information about the genes surrounding QTNs and QQIs identified by 3VmrMLM can be found in Supplementary Table 3.

Response to different treatments and GO enrichment pathway

We conducted a comprehensive analysis of gene expression changes under different treatments to gain insights into their responses. Through differential expression analysis on the unreported genes, we successfully identified distinct expression patterns of the 35 genes (Supplementary Table 4). Specifically, we found 18 genes that exhibited significant differential expression between 22°C and 10°C treatments (Figure 4A; Supplementary Table 4), 15 were significantly upregulated at 10°C, while only three genes showed significant downregulation at this temperature. For instance, *AT3G55980*, located near the SNP120225 locus associated with LN22, exhibited a $\log_2$FoldChange of 2.79 and a P-value of 1.05E-07, as illustrated in the upper right corner of the volcano plot. This gene was found to be enriched in the nucleus (Figure 4A; Supplementary Table 4). Similarly, 14 genes showed significant differential expression between long-day and short-day treatments (Figure 4B; Supplementary Table 4), suggesting their involvement in light-dependent processes. Specifically, eight genes exhibited significant upregulation under short-day treatments, while six genes were significantly upregulated under long-day treatments. Additionally, we observed differential expression in 3 genes between 4 weeks and 0 weeks treatments (Figure 4C; Supplementary Table 4), highlighting their role in a time-dependent response. These findings offer valuable insights into the biological underpinnings of the newly identified genes associated with flowering-related traits in *Arabidopsis*.

To gain further functional insights, we performed GO functional enrichment analysis on the identified DEGs. This analysis revealed that out of the 35 DEGs, 26 genes were significantly enriched in 4 distinct GO terms associated with various biological processes (Figure 4D). Furthermore, it was shown that 20 genes located in proximity to QQIs and QTNs were specifically enriched in the nucleus (GO:0005634) (Figure 4D). For example, *AT3G55980*, known as *AtSZF1*, has been reported to be associated with the nucleus and is involved in the *Arabidopsis* salt stress response (Sun et al., 2007). Notably, *AT4G01870* and *AT4G31800* were found to be simultaneously associated with three important biological processes (Figure 4D). Specifically, *AT4G31800*, known as *WRKY18*, enhances developmentally regulated defense responses in transgenic plants without causing significant negative effects on plant growth (Pandey et al., 2010). On the other hand, *AT4G01870* is involved in the chemical reactions and pathways leading to the synthesis of camalexin, an indole phytoalexin (https://www.arabidopsis.org/). In addition, we observed three genes *AT1G52040*, *AT4G03230*, and *AT1G48930* related to carbohydrate binding (Figure 4D), with *AT1G48930* possessing a carbohydrate-binding structural domain (CBM49) that plays a role in *Arabidopsis* root hair and endosperm development, among other functions (del Campillo et al., 2012). Interestingly, we identified a pair of QQIs, *AT1G09950* and *AT1G12990*, in close proximity to the SNP5324 and SNP7584 loci, respectively (Table 2). *AT1G09950* is involved in cellular components. It affects seed germination and early seedling growth by increasing sensitivity to abscisic acid (Ren et al., 2010). Meanwhile, *AT1G12990* is associated with the regulation of the defense response (GO:0031347) and the defense response against bacteria (GO:0042742) for glycosyltransferase activity (https://www.arabidopsis.org/).

Haplotype and phenotypic difference analysis of candidate genes

To further validate the association between genes and flowering-related traits, we performed haplotype analysis on the SNPs within the 2 kb upstream regions of the 26 genes identified from the GO enrichment analysis. In total, 20 candidate genes were identified, which significant phenotypic differences were observed among their haplotypes (Table 2). These genes were associated with six different traits, namely LDV, SDV, FT10, FT22, LN10, and LN22 (Table 2). Among them, 16 genes were located near QQIs, while 4 genes were located near QTNs. It is worth noting that the loci near *AT1G03445* and *AT1G68040*, which correspond to these genes, also contain previously reported known genes. More detailed information was listed in Table 2; Supplementary Table 5.

Figure 5 illustrates the analysis of *AT1G12990* (CDS coordinates [5'-3']: 4433605-4436102), *AT4G01870* (CDS coordinates [5'-3']: 808376-810611), and *AT3G62610* (CDS coordinates [5'-3']: 23154630-23156585) to reveal intragenic variations impacting flowering time and identify favorable haplotypes. Figure 5A presents the linkage disequilibrium and haplotype block with 8 SNPs for the gene *AT1G12990*, located near the SNP7584 locus, a QQI for FT22 (Table 2). After removing 53 missing values from the

10.3389/fpls.2023.1283642

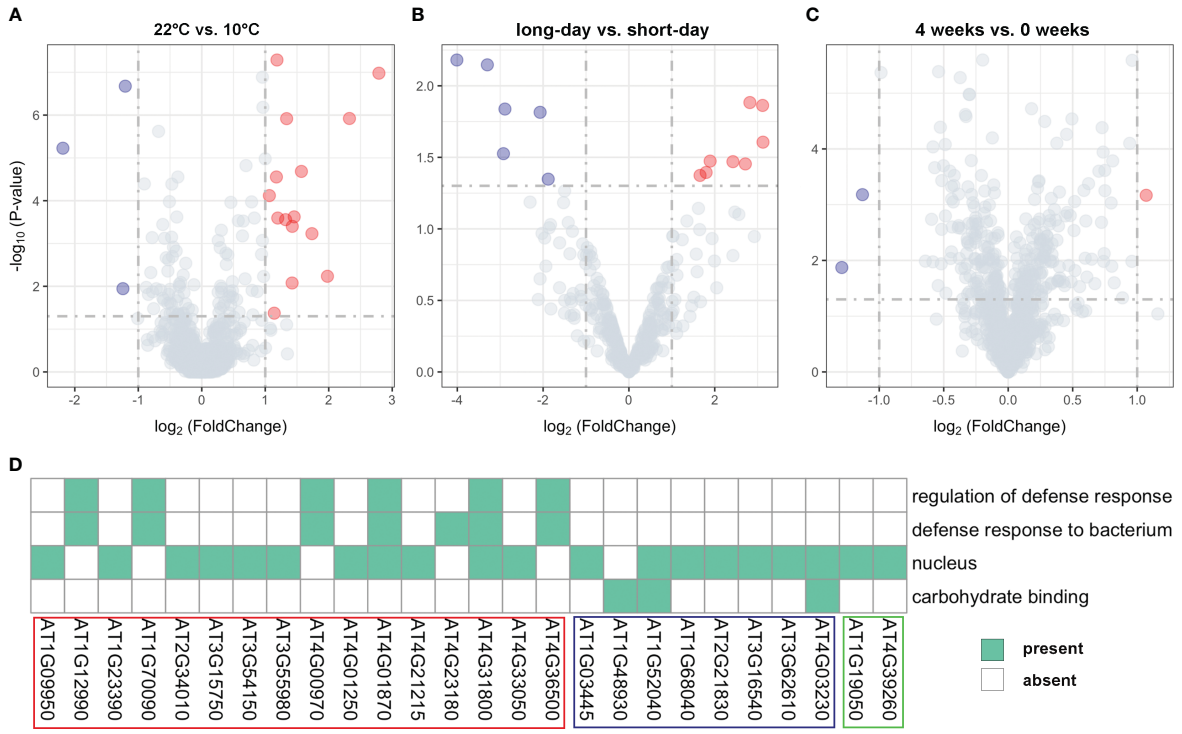


FIGURE 4
Volcano plots for expression values of **(A)** 18 genes under different temperature treatments (22°C vs. 10°C), **(B)** 14 genes under different photoperiod treatments (long-day vs. short-day), and **(C)** 3 genes under different vernalization time treatments (4 weeks vs. 0 weeks). **(D)** Results of functional enrichment analysis based on gene ontology. The genes highlighted within the red, blue, and green boxes belong to the group of significant DEGs between 22°C vs. 10°C treatments, long-day vs. short-day treatments, and 4 weeks vs. 0 weeks treatments, respectively.

phenotypic data, the remaining 146 individuals were classified into four haplotypes based on seven SNPs (SNP7613, SNP7614, SNP7615, SNP7617, SNP7618, SNP7619, and SNP7620). Haplotype IV (TGTGTTT) exhibited significantly higher median phenotypic values for FT22 compared to the other three haplotypes (Figure 5B). Haplotype IV consisted 25 individuals, among which 12 had a maximum phenotypic value of 250 for the FT22 trait, while the other three haplotypes had values of 1, 4, and 1, respectively. Additionally, a *t*-test demonstrated significant differences between haplotype IV and haplotypes I (CGGGGTG, P-value = 5.65E-07), II (CGGGTTG, P-value = 9.16E-06), and III (TGGGTTG, P-value = 7.98E-07; Supplementary Table 5). Similarly, the candidate gene *AT1G09950* (CDS coordinates [5'-3']: 4433605-4436102), located near the SNP5324 locus, showed an interaction effect with the SNP7584 locus for the FT22 trait. Supplementary Figure 2A depicts the linkage disequilibrium and haplotype block analysis using 11 SNPs. After removing 42 missing values from the phenotype data, the remaining 157 individuals were divided into three haplotypes based on seven SNPs (SNP5265, SNP5266, SNP5267, SNP5268, SNP5269, SNP5271, and SNP5272). Supplementary Figure 2B demonstrates significant differences between haplotype I (ATATAGT) and haplotype III (GAGGTCT, P-value = 1.73E-02; Supplementary Table 5). Therefore, we inferred that the candidate genes *AT1G12990* and *AT1G09950* may interact with each other and play a role in flowering time regulation in *Arabidopsis*.

Figures 5C, D present the haplotype block and phenotype differences of the candidate gene *AT4G01870*, detected around the SNP126845 locus, a QQI for FT10 (Table 2; Supplementary Table 5). Haplotype III (TTGTTT) exhibited the highest median phenotypic values and showed significant differences with haplotype I (GTCTGG, P-value = 4.20E-02) and haplotype II (TTGTTG, P-value = 6.87E-03; Supplementary Table 5). Similarly, the candidate gene *AT3G62610* was detected around the SNP124387 locus, a QQI for LDV (Table 2; Supplementary Table 5). Figures 5E, F illustrate the haplotype block and phenotype differences. Hence, we suggest that the candidate genes *AT4G01870* and *AT3G62610* may influence the flowering time in *Arabidopsis*.

Additionally, the candidate gene *AT4G01250* (CDS coordinates [5'-3']: 522530-524249) was detected around the SNP126164 locus, a QTN for FT10, while the candidate gene *AT4G00970* (CDS coordinates [5'-3']: 418327-421885) was detected near the SNP125834 locus, a QTN for LN10 (Table 2; Supplementary Table 5). Supplementary Figures 2C–F display the haplotype block and phenotype differences of these two genes. We hypothesize that the candidate genes *AT4G01250* and *AT4G00970* may also affect the flowering time in *Arabidopsis*.

In summary, we propose that the four candidate genes mentioned above, located near QQIs, may exert potential influence on their corresponding traits, among them *AT1G12990* and *AT1G09950* might have gene-by-gene interaction. Furthermore, several candidate genes near QTNs exhibited significant differences in phenotypes across haplotypes (Supplementary Table 5). However, further experimental

TABLE 2 Results of 20 candidate genes and functional annotation.

Trait	QQI/QTN	Marker	Candidate Gene	Bp	Annotation
LDV	QQI	SNP1471	*AT1G03445*	chr1:854410..859701	erine–threonine protein phosphatase
	QQI	SNP11417	*AT1G19050*	chr1:6577833..6579314	two-component response regulator
	QQI	SNP44317	*AT1G68040*	chr1:25502864..25505263	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein.
	QQI	SNP124387	*AT3G62610*	chr3:23154630..23156585	regulates flavonol biosynthesis.
	QQI	SNP161720	*AT4G39260*	chr4:18273829..18275216	verprolin
SDV	QQI	SNP66659	*AT2G21830*	chr2:9303713..9306025	encodes a putative DegP protease.
	QQI	SNP128333	*AT4G03230*	chr4:1418841..1423337	G-type lectin S-receptor-like Serine/Threonine-kinase.
	QTN	SNP90818	*AT3G16540*	chr3:5626290..5628857	encodes a putative DegP protease.
FT10	QQI	SNP126845	*AT4G01870*	chr4:808376..810611	tolB protein-like protein
	QTN	SNP126164	*AT4G01250*	chr4:522530..524249	involved in regulation of dark induced leaf senescence.
FT22	QQI	SNP5324	*AT1G09950*	chr1:3240531..3241863	response to aba and salt 1
	QQI	SNP7584	*AT1G12990*	chr1:4433605..4436102	beta-1,4-N-acetylglucosaminyltransferase family protein
	QQI	SNP73495	*AT2G34010*	chr2:14368536..14370438	verprolin
LN10	QQI	SNP14480	*AT1G23390*	chr1:8308965..8310916	kelch domain-containing F-box protein
	QQI	SNP119021	*AT3G54150*	chr3:20050564..20052931	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein
	QTN	SNP125834	*AT4G00970*	chr4:418327..421885	encodes a cysteine-rich receptor-like protein kinase.
	QTN	SNP151832	*AT4G23180*	chr4:12137995..12140930	encodes a receptor-like protein kinase.
LN22	QQI	SNP45945	*AT1G70090*	chr1:26400694..26402815	encodes a protein with putative galacturonosyltransferase activity.
	QQI	SNP90174	*AT3G15750*	chr3:5334844..5336485	essential protein Yae1
	QQI	SNP120225	*AT3G55980*	chr3:20776220..20778952	CCCH-type zinc finger protein involved in salt stress and immune responses.

verification is required to determine whether these candidate genes interact with each other in regulating flowering in *Arabidopsis*.

Discussion

Levene's test for potential epistasis

Due to the substantial computational requirements in QQI detection, particularly when considering the population structure and polygenic backgrounds in 3VmrMLM, it is advisable to limit the number of markers to less than 5,000 (Li et al., 2022a; Li et al., 2022b). To obtain the potential epistasis and alleviate the computational burden, we employed *Levene's* test, which can be used to detect potential loci for heterogeneity of variances due to potentially interacting SNPs such as QTN-by-QTN interactions (Zhang et al., 2019). However, the direct application of *Levene's* test to the raw data did not reveal any significant interacting loci due to the large number of markers and the stringent threshold of the Bonferroni correction. Moreover, potential limitations of *Levene*'s test include no covariates are allowed and only equality of variances, but not means, can be tested (Dumitrascu et al., 2019), that is, it could neither consider the population structure nor obtain the effect estimate of markers. Therefore, for each trait, we firstly selected the top 5,000 significantly associated variance-controlling SNPs detected by *Levene's* test, which also exhibited that P values were less than 0.05, and then performed QQI detection of 3VmrMLM using these top 5,000 loci for input. Combining potential epistasis loci selection with 3VmrMLM significantly improves detection accuracy while greatly reducing computation time.

Genetic basis for flowering-related traits in *Arabidopsis*

3VmrMLM detected 130 QQIs and 89 QTNs significantly associated with 11 flowering-related traits in the analysis of epistasis. Among them, the PVE of QQIs for the traits LD, LDV, and FT22 were 25.856%, 23.438%, and 19.163%, respectively (Figures 2A, B, E), which were higher than those of QTNs at 22.193%, 21.863%, and 18.446% (Figures 2A, B, E), indicating that QQIs contribute more to phenotypic variation than QTNs for these three traits and epistasis is a non-negligible factor contributing to phenotypic variation. Notably, A pair of loci SNP66960 and

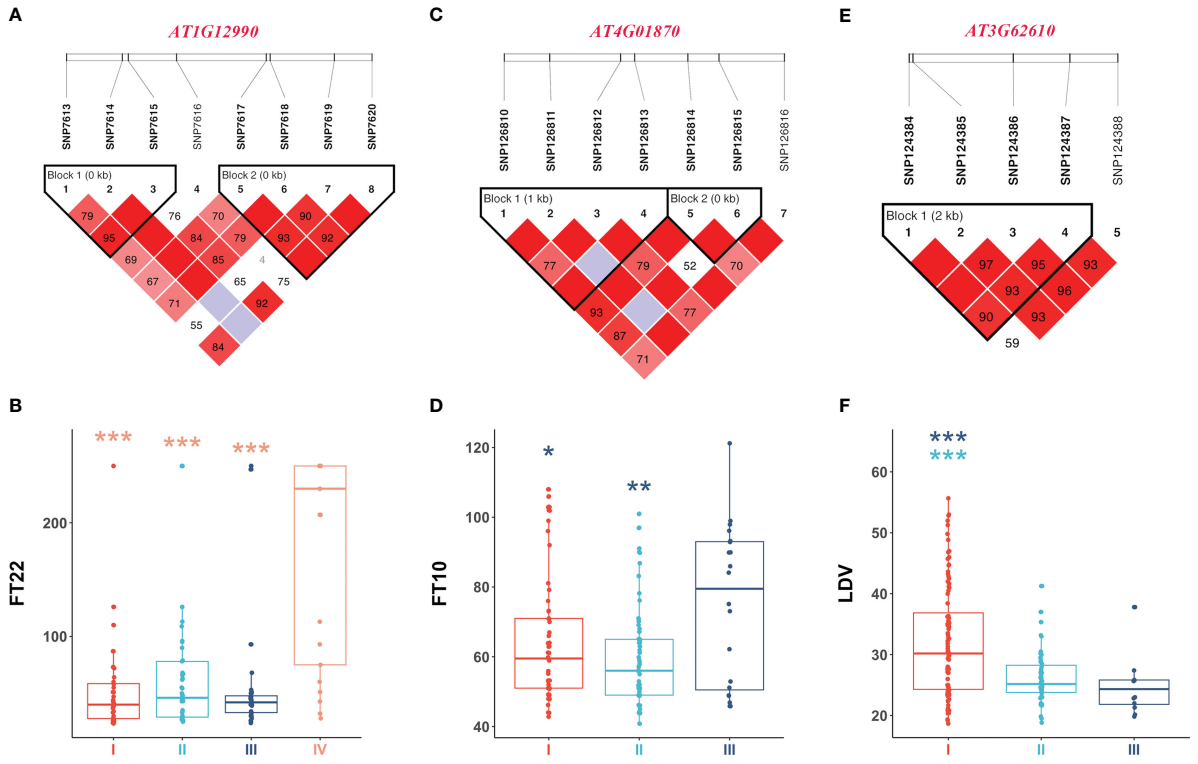


FIGURE 5
Linkage disequilibrium and haplotype block analysis for the candidate genes **(A)** *AT1G12990* associated with FT22, **(C)** *AT4G01870* associated with FT10, and **(E)** *AT3G62610* associated with LDV, respectively. **(B)** Comparison of FT22 across various haplotypes I (CGGGGTG), II (CGGGTTG), III (TGGGTTG), and IV (TGTGTTT). **(D)** Comparison of FT10 across various haplotypes I (GTCTGG), II (TTGTTG), and III (TTGTTT). **(F)** Comparison of LDV across various haplotypes I (AAAG), II (AGTA), and III (CGTA). In the boxplots, the center line represents the median, the box limits indicate the upper and lower quartiles, and the whiskers extend 1.5 times the interquartile range. Data points beyond the whiskers are considered outliers and plotted individually. The number of stars indicates the significance level from *t*-test (*0.05, **0.01, ***0.001).

SNP71678, located on chromosome 2 under LD, had the highest PVE among all traits in terms of QQI, at 8.187% (Supplementary Table 1). In its vicinity, the known gene *SVP* (*AT2G22540*; Supplementary Figure 1A; Supplementary Table 3) has been shown to be an important regulator during the transition to flowering and floral development, while *SVP* interacts with *OsMADS22* and *OsMADS47* to interfere with normal *Arabidopsis* flower development (Fornara et al., 2008).

The known genes *BRN2* (*AT1G03457*) located near QQI SNP1471 (P-value = 4.32628E-15, LOD = 3.2212) and *FKF1* (*AT1G68050*) located near QQI SNP44317 (P-value = 1.37721E-07, LOD = 5.8963; Figure 3A; Supplementary Table 3) are both associated with LDV and interact with *AtBRN*, CDF2 protein to promote or repress flowering in *Arabidopsis*, respectively (Kim et al., 2013). The known gene *SPA4* (*AT1G53090*) associated with FT22 is located near QQI SNP32482 (P-value=1.35181E-08, LOD=7.0044; Figure 3B; Supplementary Table 3). *SPA4* is involved in regulating *Arabidopsis* photoperiodic flowering and was found to interact with both CO, *PAP1* and *PAP2* to jointly regulate flowering stability and growth under light conditions (Laubinger et al., 2006; Maier et al., 2013).Two known genes, *FT* (*AT1G65480*) and *FAS1* (*AT1G65470*), were detected simultaneously near QQI SNP42063 (P-value=9.97104E-07, LOD=5.6226) under the LN10 trait (Figure 3D; Supplementary Table 3), where *FT* interacts with *FD* (*AT4G35900*), and 14-3-3 proteins interact to produce florigen-activation complex to control flowering time and correct expression of floral homologs and promote flowering (Collani et al., 2019). On the other hand, the known genes with QTN effects *FD* (*AT4G35900*, near QTN SNP159681; Hanano and Goto, 2011; Gorham et al., 2018), *AGL28* (*AT1G01530*, near QTN SNP350; Yoo et al., 2006), *MBR2* (*AT4G34040*, near QTN SNP158615; Iñigo et al., 2012) and 8 other genes have been reported to influence flowering through different pathways to exert either facilitative or repressive effects on flowering (Figure 3; Supplementary Figure 1; Supplementary Table 3).

Note that we also uncovered five pleiotropic known genes that act on multiple traits in terms of QQI or QTN. The known gene *AGL17* (*AT2G22630*), detected around QQI SNP67001, SNP66970, and SNP66990, is associated with three traits FT22, LD, and LN22 (Table 1; Figure 3B; Supplementary Figures 1A, G). It has been shown to be positively regulated by the photoperiod pathway regulator CO to promote flowering in *Arabidopsis* (Han et al., 2008). The known genes *ATH1* (*AT4G32980*, around QQI SNP15783; Table 1; Figure 3A; Supplementary Figure 1D) associated with LDV and 0W are required for the control of *Arabidopsis* flower stem morphology and interact with *KNAT2* to help regulate flower tip development (Li et al., 2012). *BOP2* (*AT2G41370*) was detected around QQI SNP77354 and QQI

SNP77376 were detected in the vicinity, associated with 2W and LN10 (Table 1; Figures 3C, D), and BOP2 proteins interaction with *LFY* has been reported to shorten flowering time in a short period of time (Chahtane et al., 2018). The known gene *CPL3* (*AT4G01060*, around QTN SNP125988 and QTN SNP125917) was detected to have additive effects on both FT10 and 2W (Table 1; Figure 3C; Supplementary Figure 1C), confirming a pleiotropic effect on flowering development in *Arabidopsis* (Zhang and Shen, 2022). The known gene *LUH* (*AT2G32700*, around QQI SNP72736, QTN SNP72705, and QQI SNP72738) was uncovered to be involved not only in three traits FT22, 2W, and FT10, but also found to have additive and epistatic effects (Table 1; Figures 3B, C; Supplementary Figure 1C), and studies showed that *LUH* interacts with *YAB* and plays a regulatory role on lateral organ growth and inflorescence leaf management (Stahle et al., 2009). The phenotypic association results of *BOP2* (*AT2G41370*) and *CPL3* (*AT4G01060*) were consistent with the phenotypic clustering results shown in Figure 1A. Additionally, the traits LN22 and FT22 associated with *AGL17* (*AT2G22630*), as well as the traits 2W and FT10 associated with *LUH* (*AT2G32700*), were also grouped together (Figure 1A; Table 1). These findings further support the reliability of our analysis.

Except for known genes, we also identified 20 candidate genes in this study (Table 2). Among them, *AT1G12990*, *AT1G09950*, *AT4G01870*, and *AT3G62610*, located near QQIs, specially, former two genes showed potential gene-by-gene interactions related to flowering traits in *Arabidopsis*. Specifically, *AT1G12990* was found in proximity to the SNP7584 locus, while *AT1G09950* was found near the SNP5324 locus, and remarkably, these loci coincided with a significant pair of QQIs associated with the trait FT22 (P-value = 7.08064E-05, LOD = 3.4287; Supplementary Table 1). *AT4G01870* was detected near the SNP126845 locus, forming a QQI with SNP185421 for FT10 (P-value = 5.12209E-08, LOD = 6.443; Supplementary Table 1). Additionally, *AT3G62610* was found around the SNP124387 locus, forming a QQI with SNP69012 for LDV (P-value = 4.70143E-06, LOD = 4.5505; Supplementary Table 1). These candidate genes also showed differential expression under 22°C vs. 10°C and long-days vs. short-days treatments (Figures 4B, C; Supplementary Table 4). *AT1G12990* and *AT4G01870* were associated with the regulation of defense response (GO:0031347) and defense response to bacterium (GO:0042742), while *AT1G09950*, *AT4G01870*, and *AT3G62610* were involved in nucleus-related functions (GO:0005634). Notably, significant phenotypic differences were observed across different haplotypes. Therefore, we hypothesize that these candidate genes, namely *AT1G12990*, *AT1G09950*, *AT4G01870*, and *AT3G62610*, in proximity of QQIs, may play a role in influencing flowering in *Arabidopsis*. Specially, *AT1G12990* and *AT1G09950* might exist potential gene-by-gene interaction. However, further experimental validation, such as functional validation, is necessary to explore these gene-by-gene interactions for flowering-related traits.

Methods comparison

To better analyze the QQIs results obtained from the 3VmrMLM method, we performed epistasis analysis in the raw dataset using PLINK (Purcell et al., 2007). The command used for detecting pairs of epistatic loci was "*plink –file genotype –pheno phenoq.txt –epistasis –epi1 P-value –allow-no-sex –out result*", with a threshold using Bonferroni correction. The number of significant interacting loci detected for each trait using PLINK ranged from 2,903 to 41,132 (Supplementary Table 6). It is well-known that PLINK uses a simple linear model, which computes quickly even with large sample sizes, but it does not consider the polygenic background, leading to an increased false positive rate (Purcell et al., 2007). In addition, except for trait 0W, the number of significant QQIs detected by PLINK that overlap with those detected by 3VmrMLM ranged from 1 to 34. Among them, for trait FT22, PLINK detected a total of 41,132 QQIs, out of which 34 were simultaneously detected by 3VmrMLM (Supplementary Table 6). This suggests that QQIs detected by 3VmrMLM are likely to be potential interacting loci.

We also employed the REMMA method (Ning et al., 2018), a mixed linear model-based approach, for conducting epistasis analysis in the raw dataset. This method incorporates both additive and dominance relationship matrices, offering theoretical control over Type I errors when examining pairwise epistatic SNPs. Among the eleven traits, three (SDV, FT22, and 8W) showed significant interacting loci, with 429, 72, and 3,541 loci detected, respectively (Supplementary Table 6). The QQIs associated with SDV overlapped with those detected by 3VmrMLM (Supplementary Table 6).

Similarly, we employed the QMDR approach (Yu et al., 2015) based on machine learning to analyze epistasis. Because no results were obtained in the raw dataset due to the large number of markers and strict Bonferroni correction threshold. Thus, the strategy for top 5,000 marker selection and LOD scores greater than 3.0 was identical to that described for 3VmrMLM in order to be comparable. As listed in Supplementary Table 6, only six traits (LD, SDV, FT22, LN22, 4W, and 8W) showed significant interaction loci, while the remaining traits did not. Overall, 3VmrMLM excels in both efficiency and accuracy when analyzing epistasis.

Conclusion

In this study, we performed the novel 3VmrMLM method in GWAS to investigate the epistatic association with eleven flowering-related traits in *Arabidopsis*. A total of 130 pairs of QQIs and 89 QTNs were successfully detected. Furthermore, through genome annotation and previous research, 29 known genes around QQIs and 12 known genes around QTNs were identified. Among the above known genes, five genes, namely *AGL17* (*AT2G22630*), *ATH1* (*AT4G32980*), *BOP2* (*AT2G41370*), *CPL3* (*AT4G01060*), and *LUH* (*AT2G32700*), were demonstrated an epistatic or additive effect for at least two traits. Moreover, 16 candidate genes around QQIs and 4 candidate genes around QTNs were validated using differential expression analysis, functional enrichment analysis, and haplotype and phenotypic difference analysis. Notably, *AT1G12990* and *AT1G09950* around QQIs exhibited potential gene-by-gene interactions influencing flowering. These findings contribute to the identification and exploration of epistasis associated with flowering-related traits in *Arabidopsis*.

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

LH: Data curation, Formal analysis, Investigation, Validation, Visualization, Writing – original draft. BS: Data curation, Formal analysis, Investigation, Validation, Visualization, Writing – original draft. XW: Data curation, Resources, Writing – review & editing. JZ: Writing – review & editing, Funding acquisition. Y-JW: Conceptualization, Funding acquisition, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The work was supported by the National Natural Science Foundation of China (32070688 and 32270694), the Postdoctoral Science Foundation of Jiangsu (2020Z330), and the Fundamental Research Funds for the Central Universities (JCQY202108).

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1283642/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Chord diagrams for QQIs and QTNs detected by 3VmrMLM. **(A~G)** correspond to the traits LD, SDV, FT10, 0W, 4W, 8W, and LN22, respectively. The inner circle displays the detected QQIs or QTNs (△ indicates overlapping loci between QQIs and QTNs), the height of red dots represents the epistatic effects of QQI pairs, and the height of blue dots represents the additive effects of corresponding QTNs. The outer circle indicates the known genes in vicinity of significant loci.

SUPPLEMENTARY FIGURE 2
Linkage disequilibrium and haplotype block for the candidate gene **(A)** *AT1G09950* associated with FT22, **(C)** *AT4G01250* associated with FT10, and **(E)** *AT4G00970* associated with LN10. **(B)** Comparison of FT22 across various haplotypes I (ATATAGT), II (GAGGACT), and III (GAGGTCT). **(D)** Comparison of FT10 across various haplotypes I (TATACTATCT), II (TGGACCATCA), III (TGGACTAAAT), and IV (TGGACTATCT). **(F)** Comparison of LN10 across various haplotypes I (AGCCCACTGA), II (AGCTCGCCGT), III (CAATCGCCGT), and IV (CAATGGCCCT). For boxplots, center line shows median, box limits indicate upper and lower quartiles, and whiskers extend 1.5 times the interquartile range, while data beyond the end of the whiskers are outlying points that are plotted individually. The number of stars represents the result of *t* test at different significance levels (*: 0.05, **: 0.01, ***: 0.001).

SUPPLEMENTARY DATA SHEET
All the phenotypic values of the traits and all the marker genotypes, which are derived from Atwell et al. Nature 2010; 465(7298), 627-631.

References

Atwell, S., Huang, Y. S., Vilhjalmsson, B. J., Willems, G., Horton, M., Li, Y., et al. (2010). Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature* 465 (7298), 627–631. doi: 10.1038/nature08800

Balasubramanian, S., Sureshkumar, S., Lempe, J., and Weigel, D. (2006). Potent induction of Arabidopsis thaliana flowering by elevated growth temperature. *PloS Genet.* 2 (7), e106. doi: 10.1371/journal.pgen.0020106

Barrett, J. C., Fry, B., Maller, J., and Daly, M. J. (2005). Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* 21 (2), 263–265. doi: 10.1093/bioinformatics/bth457

Breen, M. S., Kemena, C., Vlasov, P. K., Notredame, C., and Kondrashov, F. A. (2012). Epistasis as the primary factor in molecular evolution. *Nature* 490 (7421), 535–538. doi: 10.1038/nature11510

Brown, M. B., and Forsythe, A. B. (1974). Robust tests for the equality of variances. *J. Am. Stat. Assoc.* 69 (346), 364–367. doi: 10.1080/01621459.1974.10482955

Chahtane, H., Zhang, B., Norberg, M., LeMasson, M., Thevenon, E., Bako, L., et al. (2018). LEAFY activity is post-transcriptionally regulated by BLADE ON PETIOLE2 and CULLIN3 in Arabidopsis. *New Phytol.* 220 (2), 579–592. doi: 10.1111/nph.15329

Collani, S., Neumann, M., Yant, L., and Schmid, M. (2019). FT modulates genome-wide DNA-Binding of the bZIP transcription factor FD. *Plant Physiol.* 180 (1), 367–380. doi: 10.1104/pp.18.01505

del Campillo, E., Gaddam, S., Mettle-Amuah, D., and Heneks, J. (2012). A tale of two tissues: AtGH9C1 is an endo-β-1,4-glucanase involved in root hair and endosperm development in Arabidopsis. *PloS One* 7 (11), e49363. doi: 10.1371/journal.pone.0049363

Dumitrascu, B., Darnell, G., Ayroles, J., and Engelhardt, B. E. (2019). Statistical tests for detecting variance effects in quantitative trait studies. *Bioinformatics* 35 (2), 200–210. doi: 10.1093/Bioinformatics/bty565

Fornara, F., Gregis, V., Pelucchi, N., Colombo, L., and Kater, M. (2008). The rice StMADS11-like genes OsMADS22 and OsMADS47 cause floral reversions in Arabidopsis without complementing the svp and agl24 mutants. *J. Exp. Bot.* 59 (8), 2181–2190. doi: 10.1093/jxb/ern083

Gorham, S. R., Weiner, A. I., Yamadi, M., and Krogan, N. T. (2018). HISTONE DEACETYLASE 19 and the flowering time gene FD maintain reproductive meristem identity in an age-dependent manner. *J. Exp. Bot.* 69 (20), 4757–4771. doi: 10.1093/jxb/ery239

Gui, J., Moore, J. H., Williams, S. M., Andrews, P., Hillege, H. L., van der Harst, P., et al. (2013). A simple and computationally efficient approach to multifactor dimensionality reduction analysis of gene-gene interactions for quantitative traits. *PloS One* 8 (6), e66545. doi: 10.1371/journal.pone.0066545

Han, P., Garcia-Ponce, B., Fonseca-Salazar, G., Alvarez-Buylla, E. R., and Yu, H. (2008). AGAMOUS-LIKE 17, a novel flowering promoter, acts in a FT-independent photoperiod pathway. *Plant J.* 55 (2), 253–265. doi: 10.1111/j.1365-313X.2008.03499.x

Hanano, S., and Goto, K. (2011). Arabidopsis TERMINAL FLOWER1 is involved in the regulation of flowering time and inflorescence development through transcriptional repression. *Plant Cell* 23 (9), 3172–3184. doi: 10.1105/tpc.111.088641

He, T., Hill, C. B., Angessa, T. T., Zhang, X. Q., Chen, K., Moody, D., et al. (2019). Gene-set association and epistatic analyses reveal complex gene interaction networks affecting flowering time in a worldwide barley collection. *J. Exp. Bot.* 70 (20), 5603–5616. doi: 10.1093/jxb/erz332

Helliwell, C. A., Anderssen, R. S., Robertson, M., and Finnegan, E. J. (2015). How is FLC repression initiated by cold? *Trends Plant Sci.* 20 (2), 76–82. doi: 10.1016/j.tplants.2014.12.004

Imaizumi, T., and Kay, S. A. (2006). Photoperiodic control of flowering: not only by coincidence. *Trends Plant Sci.* 11 (11), 550–558. doi: 10.1016/j.tplants.2006.09.004

Iñigo, S., Giraldez, A. N., Chory, J., and Cerdan, P. D. (2012). Proteasome-mediated turnover of Arabidopsis MED25 is coupled to the activation of FLOWERING LOCUS T transcription. *Plant Physiol.* 160 (3), 1662–1673. doi: 10.1104/pp.112.205500

Jetha, K., Theißen, G., and Melzer, R. (2014). Arabidopsis SEPALLATA proteins differ in cooperative DNA-binding during the formation of floral quartet-like complexes. *Nucleic Acids Res.* 42 (17), 10927–10942. doi: 10.1093/nar/gku755

Kim, H. S., Abbasi, N., and Choi, S. B. (2013). Bruno-like proteins modulate flowering time via 3' UTR-dependent decay of SOC1 mRNA. *New Phytol.* 198 (3), 747–756. doi: 10.1111/nph.12181

Koornneef, M., and Meinke, D. (2010). The development of Arabidopsis as a model plant. *Plant J.* 61 (6), 909–921. doi: 10.1111/j.1365-313X.2009.04086.x

Laubinger, S., Marchal, V., Le Gourrierec, J., Wenkel, S., Adrian, J., Jang, S., et al. (2006). Arabidopsis SPA proteins regulate photoperiodic flowering and interact with the floral inducer CONSTANS to regulate its stability. *Development* 133 (16), 3213–3222. doi: 10.1242/dev.02481

Lazaro, A., Mouriz, A., Pineiro, M., and Jarillo, J. A. (2015). Red light-mediated degradation of CONSTANS by the E3 ubiquitin ligase HOS1 regulates photoperiodic flowering in Arabidopsis. *Plant Cell* 27 (9), 2437–2454. doi: 10.1105/tpc.15.00529

Lee, C. M., Feke, A., Li, M. W., Adamchek, C., Webb, K., Pruneda-Paz, J., et al. (2018). Decoys untangle complicated redundancy and reveal targets of circadian clock F-box proteins. *Plant Physiol.* 177 (3), 1170–1186. doi: 10.1104/pp.18.00331

Li, Y., Pi, L., Huang, H., and Xu, L. (2012). ATH1 and KNAT2 proteins act together in regulation of plant inflorescence architecture. *J. Exp. Bot.* 63 (3), 1423–1433. doi: 10.1093/jxb/err376

Li, M., Zhang, Y. W., Zhang, Z. C., Xiang, Y., Liu, M. H., Zhou, Y. H., et al. (2022a). A compressed variance component mixed model for detecting QTNs and QTN-by-environment and QTN-by-QTN interactions in genome-wide association studies. *Mol. Plant* 15 (4), 630–650. doi: 10.1016/j.molp.2022.02.012

Li, M., Zhang, Y. W., Xiang, Y., Liu, M. H., and Zhang, Y. M. (2022b). IIIVmrMLM: The R and C++ tools associated with 3VmrMLM, a comprehensive GWAS method for dissecting quantitative traits. *Mol. Plant* 15 (8), 1251–1253. doi: 10.1016/j.molp.2022.06.002

Mackay, T. F., and Moore, J. H. (2014). Why epistasis is important for tackling complex human disease genetics. *Genome Med.* 6 (6), 124. doi: 10.1186/gm561

Maier, A., Schrader, A., Kokkelink, L., Falke, C., Welter, B., Iniesto, E., et al. (2013). Light and the E3 ubiquitin ligase COP1/SPA control the protein stability of the MYB transcription factors PAP1 and PAP2 involved in anthocyanin accumulation in Arabidopsis. *Plant J.* 74 (4), 638–651. doi: 10.1111/tpj.12153

Mockler, T., Yang, H., Yu, X., Parikh, D., Cheng, Y. C., Dolan, S., et al. (2003). Regulation of photoperiodic flowering by Arabidopsis photoreceptors. *Proc. Natl. Acad. Sci. U.S.A.* 100 (4), 2140–2145. doi: 10.1073/pnas.0437826100

Moore, J. H. (2004). Computational analysis of gene-gene interactions using multifactor dimensionality reduction. *Expert Rev. Mol. Diagn.* 4 (6), 795–803. doi: 10.1586/14737159.4.6.795

Ning, C., Wang, D., Kang, H., Mrode, R., Zhou, L., Xu, S., et al. (2018). A rapid epistatic mixed-model association analysis by linear retransformations of genomic estimated values. *Bioinformatics* 34 (11), 1817–1825. doi: 10.1093/bioinformatics/bty017

Pandey, S. P., Roccaro, M., Schön, M., Logemann, E., and Somssich, I. E. (2010). Transcriptional reprogramming regulated by WRKY18 and WRKY40 facilitates powdery mildew infection of Arabidopsis. *Plant J.* 64 (6), 912–923. doi: 10.1111/j.1365-313X.2010.04387.x

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81 (3), 559–575. doi: 10.1086/519795

Qi, X. H., Wu, D. T., Li, G. Z., Zhao, J. L., and Li, M. L. (2018). Regulation pathways of flowering in *Arabidopsis thaliana*. *J. Shanxi Agric. Univ. (Nat. Sci.)* 38 (9), 1–8. doi: 10.13842/j.cnki.isn1671-8151.201805004

Ranjan, A., Fiene, G., Fackendahl, P., and Hoecker, U. (2011). The Arabidopsis repressor of light signaling SPA1 acts in the phloem to regulate seedling de-etiolation, leaf expansion and flowering time. *Development* 138 (9), 1851–1862. doi: 10.1242/dev.061036

Ren, Z., Zheng, Z., Chinnusamy, V., Zhu, J., Cui, X., Iida, K., et al. (2010). RAS1, a quantitative trait locus for salt tolerance and ABA sensitivity in Arabidopsis. *Proc. Natl. Acad. Sci. U.S.A.* 107 (12), 5669–5674. doi: 10.1073/pnas.0910798107

Stahle, M. I., Kuehlich, J., Staron, L., von Arnim, A. G., and Golz, J. F. (2009). YABBYs and the transcriptional corepressors LEUNIG and LEUNIG_HOMOLOG maintain leaf polarity and meristem activity in Arabidopsis. *Plant Cell* 21 (10), 3105–3118. doi: 10.1105/tpc.109.070458

Stern, M. D., Aihara, H., Cho, K. H., Kim, G. T., Horiguchi, G., Roccaro, G. A., et al. (2007). Structurally related Arabidopsis ANGUSTIFOLIA is functionally distinct from the transcriptional corepressor CtBP. *Dev. Genes Evol.* 217 (11-12), 759–769. doi: 10.1007/s00427-007-0186-8

Sun, J., Jiang, H., Xu, Y., Li, H., Wu, X., Xie, Q., et al. (2007). The CCCH-type zinc finger proteins AtSZF1 and AtSZF2 regulate salt stress responses in Arabidopsis. *Plant Cell Physiol.* 48 (8), 1148–1158. doi: 10.1093/pcp/pcm088

Teper-Bamnolker, P., and Samach, A. (2005). The flowering integrator FT regulates SEPALLATA3 and FRUITFULL accumulation in Arabidopsis leaves. *Plant Cell* 17 (10), 2661–2675. doi: 10.1105/tpc.105.035766

Upton, A., Trelles, O., Cornejo-Garcia, J. A., and Perkins, J. R. (2016). Review: High-performance computing to detect epistasis in genome scale data sets. *Brief. Bioinform.* 17 (3), 368–379. doi: 10.1093/bib/bbv058

Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N. L., et al. (2010). BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am. J. Hum. Genet.* 87 (3), 325–340. doi: 10.1016/j.ajhg.2010.07.021

Yoo, S. K., Lee, J. S., and Ahn, J. H. (2006). Overexpression of AGAMOUS-LIKE 28 (AGL28) promotes flowering by upregulating expression of floral promoters within the autonomous pathway. *Biochem. Biophys. Res. Commun.* 348 (3), 929–936. doi: 10.1016/j.bbrc.2006.07.121

Yu, W., Kwon, M. S., and Park, T. (2015). Multivariate quantitative multifactor dimensionality reduction for detecting gene-gene interactions. *Hum. Hered.* 79 (3-4), 168–181. doi: 10.1159/000377723

Zhang, F., Boerwinkle, E., and Xiong, M. (2014). Epistasis analysis for quantitative traits by functional regression model. *Genome Res.* 24 (6), 989–998. doi: 10.1101/gr.161760.113

Zhang, F., Chen, W., Zhu, Z., Zhang, Q., Nabais, M. F., Qi, T., et al. (2019). OSCA: a tool for omic-data-based complex trait analysis. *Genome Biol.* 20 (1), 107. doi: 10.1186/s13059-019-1718-z

Zhang, Y., and Shen, L. (2022). CPL2 and CPL3 act redundantly in FLC activation and flowering time regulation in Arabidopsis. *Plant Signal Behav.* 17 (1), 2026614. doi: 10.1080/15592324.2022.2026614

Zhang, F. T., Zhu, Z. H., Tong, X. R., Zhu, Z. X., Qi, T., and Zhu, J. (2015). Mixed linear model approaches of association mapping for complex traits based on omics variants. *Sci. Rep.* 5, 10298. doi: 10.1038/srep10298

Zhao, N., Su, X. M., Liu, Z. W., Zhou, J. X., Su, Y. N., Cai, X. W., et al. (2022). The RNA recognition motif-containing protein UBA2c prevents early flowering by promoting transcription of the flowering repressor FLM in Arabidopsis. *New Phytol.* 233 (2), 751–765. doi: 10.1111/nph.17836

10.3389/fpls.2023.1283642

Glossary

0W	Days to FT under LD with vernalization for 0 weeks
2W	Days to FT under LD with vernalization for 2 weeks
4W	Days to FT under LD with vernalization for 4 weeks
8W	Days to FT under LD with vernalization for 8 weeks
AI	artificial intelligence
BLUP	best linear unbiased prediction
DEGs	differentially expressed genes
FT10	Flowering time at 10°C
FT22	Flowering time at 22°C
FDR	false discovery rate
GWAS	genome-wide association studies
GEO	Gene Expression Omnibus
GO	gene ontology
LD	Days to flowering time under Long Day
LDV	Days to flowering time under Long Day with vernalization at 4°C during 5 weeks
LOD	logarithm of odds
LN10	leaf number at flowering time at 10°C
LN22	leaf number at flowering time at 22°C
MAF	minor allele frequency
MDR	multifactor dimensionality reduction
MLM	mixed linear model
NCBI	National Center for Biotechnology Information
PCCs	Pearson correlation coefficients
QEIs	QTN-by-environment interactions
QMDR	quantitative MDR
QQIs	QTN-by-QTN interactions
QTNs	Quantitative trait nucleotides
REMMA	rapid epistatic mixed-model association analysis
SDV	Days to flowering time under Short Day with vernalization at 4°C during 5 weeks
SNPs	single nucleotide polymorphisms
TAIR	The Arabidopsis Information Resource.

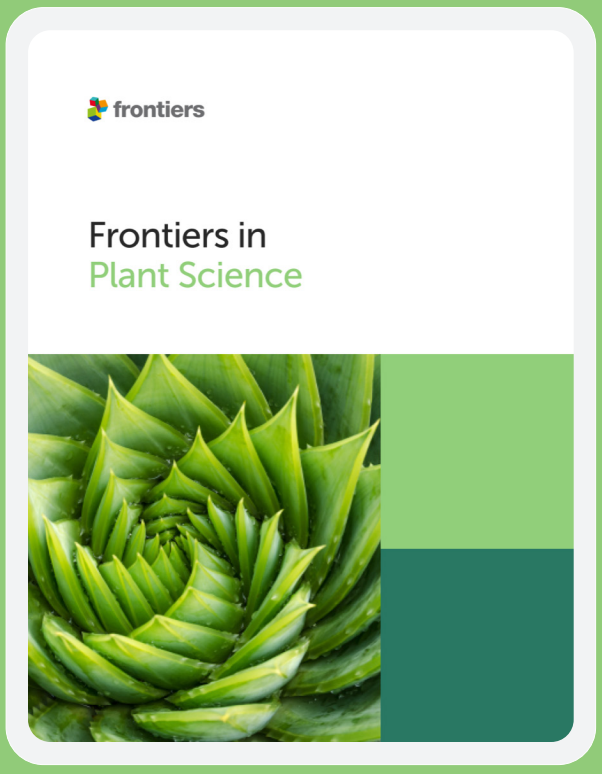
Frontiers in Plant Science

Cultivates the science of plant biology and its applications

The most cited plant science journal, which advances our understanding of plant biology for sustainable food security, functional ecosystems and human health.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact