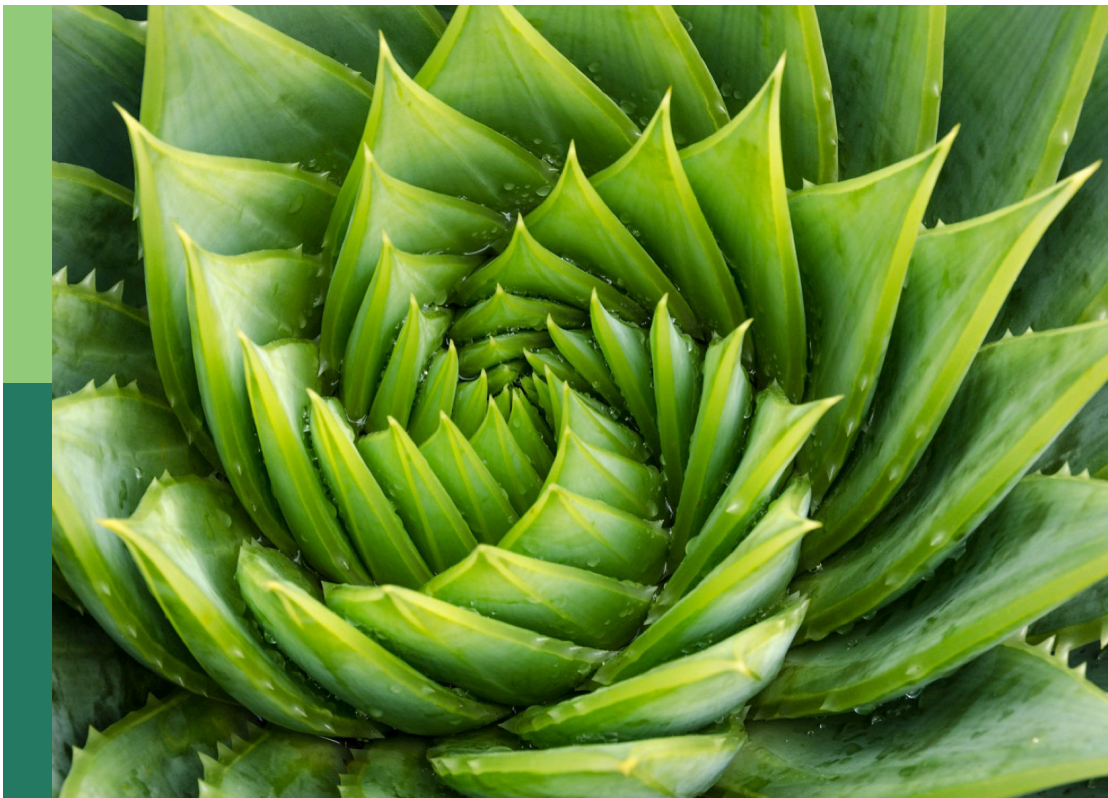
frontiers | Research Topics

Crop improvement by omics and bioinformatics

Edited by
Yan Zhao, Jun Li, Zhichao Wu
and Xueqiang Wang

Published in
Frontiers in Plant Science



ISSN 1664-8714
ISBN 978-2-8325-4774-8
DOI 10.3389/978-2-8325-4774-8

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

Crop improvement by omics and bioinformatics

Topic editors

Yan Zhao — Shandong Agricultural University, China

Jun Li — Zhejiang University, China

Zhichao Wu — National Cancer Institute Bethesda, United States

Xueqiang Wang — Zhejiang University, China

Citation

Zhao, Y., Li, J., Wu, Z., Wang, X., eds. (2024). *Crop improvement by omics and bioinformatics*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4774-8

Table of contents

frontiers | Frontiers in Plant Science

TYPE Editorial
PUBLISHED 03 April 2024
DOI 10.3389/fpls.2024.1391334

OPEN ACCESS

EDITED AND REVIEWED BY
Jihong Hu,
Northwest A&F University, China

*CORRESPONDENCE
Xueqiang Wang
wangxueqiang02@163.com

†These authors have contributed equally to this work

RECEIVED 25 February 2024
ACCEPTED 21 March 2024
PUBLISHED 03 April 2024

CITATION
Li J, Zhao Y, Wu Z and Wang X (2024) Editorial: Crop improvement by omics and bioinformatics.
Front. Plant Sci. 15:1391334.
doi: 10.3389/fpls.2024.1391334

Editorial: Crop improvement by omics and bioinformatics

Jun Li[1,2†], Yan Zhao[3†], Zhichao Wu[4†] and Xueqiang Wang[1,2,5*]

[1]Hainan Institute of Zhejiang University, Sanya, Hainan, China, [2]Zhejiang Provincial Key Laboratory of Crop Genetic Resources, the Advanced Seed Institute, Plant Precision Breeding Academy, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, Zhejiang, China, [3]State Key Laboratory of Crop Biology, Shandong Key Laboratory of Crop Biology, College of Agronomy, Shandong Agricultural University, Tai'an, Shandong, China, [4]National Cancer Institute, National Institutes of Health, Bethesda, MD, United States, [5]Yazhouwan National Laboratory, Sanya, Hainan, China

KEYWORDS

omics, bioinformatics, population genetics, improvement, evolution

Editorial on the Research Topic
Crop improvement by omics and bioinformatics

1 Introduction

Crop improvement in modern era by the genetic and breeding tools is being driven by the requirements of food security and sustainability. The caloric and nutritional needs of the growing population, and respond to environmental changes are the two demands of crop productivity (Zeng et al., 2017; Wang et al., 2018; Chen et al., 2021; Ren et al., 2023). In order to meet these demands, the global food production must increase by one billion tons in the next few decades, but the current growth rate is far from being reached. Moreover, rapid changes in the environment are accelerating land degradation, aggravating pests and diseases, introducing extreme stress and reducing crop productivity (Zeng et al., 2017; Zelm et al., 2020; Liang et al., 2021; Wang et al., 2023).

In the past few decades, remarkable progresses have been achieved in the discovery of yield, quality and resistance genes in crops, and the dissection of plant molecular mechanisms (Zeng et al., 2017; Wang et al., 2018; Zelm et al., 2020; Chen et al., 2021; Liang et al., 2021; Ren et al., 2023; Wang et al., 2023). With the continuous progress in sequencing technology, molecular markers and gene editing, a large number of excellent crop varieties have been cultivated and modern genetic improvement of crops have been realized (Lei et al., 2021; Qin et al., 2021; Tang et al., 2022; Wang and Han, 2022; Shi et al., 2023; Wen et al., 2023). But it is far from enough compared with the rapid changes in the growing population and environment.

Many new omics technologies have been developed in recent years, e.g., genomics, transcriptomics, proteomics, metabolomics, interactomics, and phenomics (Xie et al., 2021; Huang et al., 2022; Shang et al., 2022; Wang and Han, 2022; Wang et al., 2022; Marand et al., 2023; Ren et al., 2023). Integrating multi-omics could clarify the mechanisms of many biological processes and explore the interactions among various substances (Della Coletta et al., 2021; Huang et al., 2022; Luo et al., 2022). This will provide a new perspective for understanding the complex traits of crops and accelerate the breeding programs. The crop improvement is entering a new era of biological information (Shang et al., 2022; Wang and Han, 2022; Shi et al., 2023; Wen et al., 2023).

In this editorial, we set up a Research Topic of Crop Improvement by Omics and Bioinformatics. The goal of this Research Topic is to collect all types of research and review articles describing the latest advances in the discovery of yield, quality and resistance genes in crops, and the dissection of crop molecular mechanisms. In addition, recent discoveries derived from the development or application of new omics technologies in crops as well as new methods for the analysis, mining, and visualization of crop omics datasets are also delightedly accepted. The following themes are included in this Research Topic: (a) Population genetics, haplotype analysis and evolution of important genes in crops; (b) Development of new omics technologies (software or algorithm) for crop improvement; (c) Multi-omics approaches to understand the molecular basis of the genes for important agronomic traits in crops; (d) Integration with multi-omics revealing the origin and evolution of crops; (e) Meta-analysis and comparative analysis of crop omics datasets.

2 Discovery of important genes by multi-omics approaches

Chen et al. systematically evaluated various state-of-the-art object detectors on rice panicle counting and identified YOLOv8-X as the optimal detector. Applying YOLOv8-X to UAV time-series images of 294 rice recombinant inbred lines (RILs) allowed accurate quantification of six heading date-related traits and identified quantitative trait loci (QTL), including verified loci and novel loci, associated with heading date. This research optimized UAV phenotyping and computer vision pipeline that may facilitate scalable molecular identification of heading-date-related genes and guide enhancements in rice yield and adaptation.

Li et al. evaluated the heat tolerance at the seedling stage using 620 diverse rice accessions, and based on the GWAS and transcriptomics integrated results, a hypothetical model modulated by *qHT7* in response to heat stress was proposed. The results provided valuable candidate genes for improving rice heat tolerance through molecular breeding.

Yu et al. identified 5, 6, 6, and 6 QTLs for grain length, grain weight, grain area, and thousand grain weight, respectively, using 55K SNP assay genotyping and large scale phenotyping data of the population and GWAS. A comprehensive analysis of transcriptome data and homologs showed that *TraesCS2D02G414800* could be the real QTL gene for *qGL-2D*. Overall, this study presented several reliable grain size QTLs and candidate genes for grain length for future bread wheat breeding for yield.

Sun et al. screened a total of 15 candidate genes from a genome-wide association study (GWAS) on 8 traits of 150 cotton germplasms under drought conditions and found four genes were highly expressed after drought stress. Three of these genes had the same differential expression pattern. This study provides a theoretical basis for the genetic analysis of cotton yield traits under drought stress, and provides gene resources for improved breeding of cotton yield traits under drought stress.

Wu et al. identified 25 potential earliness related genes from Chinese bayberry (*Myrica rubra*) by analyzing the transcriptome data from early ripening, medium ripening and late ripening varieties, with clustering analysis and comparisons of genes reported related to flowering in *Arabidopsis thaliana*. Finally, through transgenic studies, this study identified an important gene *MrSPL4* in Chinese bayberry, which enhanced growth and flowering, providing important theoretical basis for early-mature breeding of Chinese bayberry.

Gao et al. conducted metabolomic and transcriptomic analyses of 5~8 years old *Cinnamomum cassia*, in order to explore the mechanism of the dynamic accumulation of active ingredients. A total of 72 phenylpropanoids, 146 flavonoids, and 130 terpenoids were found to exhibit marked changes. In addition, transcription factors (TFs) and genes involved in phenylpropanoids and flavonoids synthesis and regulation were identified through co-expression network analyses. The results of this study provide new insights into the synthesis and accumulation of phenylpropanoid, flavonoids and terpenoids in *C. cassia* at four growth stages.

Gao et al. performed full-length transcriptome analysis of *in vitro* bulblet initiation in lily. They compared the expression profiles of crucial genes of carbohydrate metabolism between different stages and different treatments. Significant co-expression was shown between genes involved in carbohydrate metabolism and auxin signaling, together with transcription factors such as bHLHs, MYBs, ERFs and C3Hs. This study indicates the coordinate regulation of bulblet initiation by carbohydrate metabolism and auxin signaling, serving as a basis for further studies on the molecular mechanism of bulblet initiation in lily and other bulbous flowers.

Li et al. presented a co-expression network, involving ABA and other phytohormone signals, based on weighted gene co-expression network analysis of spatiotemporally resolved transcriptome data and phenotypic changes of strawberry receptacles during development and following various treatments. They explored the role of two hub signals, small auxin up-regulated RNA 1 and 2 in receptacle ripening mediated by ABA, which are also predicted to contribute to fruit quality. These results and publicly accessible datasets provide a valuable resource to elucidate ripening and quality formation mediated by ABA and multiple other phytohormone signals in strawberry receptacle and serve as a model for other non-climacteric fruits.

3 Gene family analysis

Liu et al. identified a total of 18 *Whirly* genes from six Triticeae species and found TaWHY1-7A and TaWHY1-7D mainly enhanced the tolerance to oxidative stress in yeast cells. TaWHY2s mainly improved NaCl stress tolerance and were sensitive to oxidative stress in yeast cells. The heterologous expression of *TaWHY1-7D* greatly improved drought and salt tolerance in transgenic *Arabidopsis*. These results provide the foundation for further functional study of *Whirly* genes aiming at improving osmotic stress tolerance in wheat.

Liang et al. identified 37 *StSOS1s* in potato (*Solanum tuberosum*), which were found to be unevenly distributed across 10 chromosomes, with the majority located on the plasma membrane. RT-qPCR results revealed that the expression of *StSOS1s* were significant modulated by various abiotic stresses, in particular salt and abscisic acid stress. This work extends the comprehensive overview of the *StSOS1* gene family and sets the stage for further analysis of the function of genes in SOS and hormone signaling pathways.

Tang et al. identified 57 CCCH genes in the pepper (*Capsicum annuum* L.) genome and explored the evolution and function of the CCCH gene family in *C. annuum*. They found that the expression of CCCH genes was significantly up-regulated during the response to biotic and abiotic stresses, especially cold and heat stresses, indicating that CCCH genes play key roles in stress responses. These results provide new information on CCCH genes in pepper and will facilitate future studies of the evolution, inheritance, and function of CCCH zinc finger genes in pepper.

Zhang et al. identified 59 bZIP genes that were unevenly distributed in the chestnut genome, and found *CmbZIP04*, *CmbZIP13*, *CmbZIP14*, *CmbZIP33*, *CmbZIP35*, *CmbZIP38*, and *CmbZIP56* may be important in regulating starch accumulation in chestnut seeds. This study provided basic information on CmbZIP genes, which can be utilized in future functional analysis and breeding studies.

4 Development of the omics technologies

Shen et al. presented the application of alternative splicing algorithms with or without reference genomes in plants, as well as the integration of advanced deep learning techniques for improved detection accuracy, and discussed alternative splicing studies in the pan-genomic background and the usefulness of integrated strategies for fully profiling alternative splicing.

Zhang et al. induced male sterile mutants by simultaneously editing three cotton *EXCESS MICROSPOROCYTES1* (*GhEMS1*) genes by CRISPR/Cas9. This study would not only facilitates the exploration of the basic research of cotton male sterile lines, but also provides germplasms for accelerating the hybrid breeding in cotton.

Liu et al. developed a new genomic prediction method (RHEPCG) via combining randomized Haseman-Elston (HE) regression (RHE-reg) and preconditioned conjugate gradient (PCG), which avoids the direct inverse of the genomic relationship matrix (GRM). The simulation results demonstrated that RHEPCG not only achieved similar predictive accuracy with GBLUP in most cases, but also significantly reduced computational time, indicating that RHEPCG is a practical alternative to GBLUP with better computational efficiency.

Aparicio et al. developed the Mr.Bean, an accessible and user-friendly tool with a comprehensive graphical visualization interface, to predict the genetic potential of evaluated genotypes. The application integrates descriptive analysis, measures of dispersion and centralization, linear mixed model fitting, multi-environment trial analysis, factor analytic models, and genomic analysis, aiming at helping plant scientists working in agricultural field make informed decisions more quickly.

5 Perspective

It is crucial to identify yield, quality and resistance related genes in crops, and dissect the involved molecular mechanisms. In addition, recent discoveries derived from the development or application of new omics technologies in crops as well as new methods for the analysis, mining, and visualization of crop omics datasets are also urgently needed. These results will provide a new perspective for understanding the complex traits of crops and accelerate the breeding programs.

Author contributions

XW: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Supervision, Validation, Visualization, Writing – review & editing. JL: Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft. YZ: Conceptualization, Data curation, Formal analysis, Investigation, Writing – original draft. ZW: Conceptualization, Investigation, Methodology, Resources, Writing – original draft.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This Research Topic was supported by the Hainan Provincial Natural Science Foundation of China (323QN313) received by XW, and National Natural Science Foundation of China (32200498) received by JL.

Acknowledgments

We extend our heartfelt thanks to all authors and reviewers for their invaluable input and contributions. Additionally, we are grateful to the Journal Committee for providing us the opportunity to develop this Research Topic.

Conflict of interest

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

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

References

Chen, Q., Li, W., Tan, L., and Tian, F. (2021). Harnessing knowledge from maize and rice domestication for new crop breeding. *Mol. Plant* 14, 9–26. doi: 10.1016/j.molp.2020.12.006

Della Coletta, R., Qiu, Y., Ou, S., Hufford, M. B., and Hirsch, C. N. (2021). How the pan-genome is changing crop genomics and improvement. *Genome Biol.* 22, 3. doi: 10.1186/s13059-020-02224-8

Huang, X., Huang, S., Han, B., and Li, J. (2022). The integrated genomics of crop domestication and breeding. *Cell* 185, 2828–2839. doi: 10.1016/j.cell.2022.04.036

Lei, L., Goltsman, E., Goodstein, D., Wu, G. A., Rokhsar, D. S., and Vogel, J. P. (2021). Plant pan-genomics comes of age. *Annu. Rev. Of Plant Biol.* 72, 411–435. doi: 10.1146/annurev-arplant-080720-105454

Liang, Y., Liu, H.-J., Yan, J., and Tian, F. (2021). Natural variation in crops: realized understanding, continuing promise. *Annu. Rev. Of Plant Biol.* 72, 357–385. doi: 10.1146/annurev-arplant-080720-090632

Luo, Z., Xia, H., Bao, Z., Wang, L., Feng, Y., Zhang, T., et al. (2022). Integrated phenotypic, phylogenomic, and evolutionary analyses indicate the earlier domestication of <Em>Geng</Em> upland rice in China. *Mol. Plant* 15, 1506–1509. doi: 10.1016/j.molp.2022.09.011

Marand, A. P., Eveland, A. L., Kaufmann, K., and Springer, N. M. (2023). Cis-regulatory elements in plant development, adaptation, and evolution. *Annu. Rev. Of Plant Biol.* 74, 111–137. doi: 10.1146/annurev-arplant-070122-030236

Qin, P., Lu, H., Du, H., Wang, H., Chen, W., Chen, Z., et al. (2021). Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* 184, 3542–3558.e16. doi: 10.1016/j.cell.2021.04.046

Ren, D., Ding, C., and Qian, Q. (2023). Molecular bases of rice grain size and quality for optimized productivity. *Sci. Bull.* 68, 314–350. doi: 10.1016/j.scib.2023.01.026

Shang, L., Li, X., He, H., Yuan, Q., Song, Y., Wei, Z., et al. (2022). A super pan-genomic landscape of rice. *Cell Res.* 32, 878–896. doi: 10.1038/s41422-022-00685-z

Shi, J., Tian, Z., Lai, J., and Huang, X. (2023). Plant pan-genomics and its applications. *Mol. Plant* 16, 168–186. doi: 10.1016/j.molp.2022.12.009

Tang, D., Jia, Y., Zhang, J., Li, H., Cheng, L., Wang, P., et al. (2022). Genome evolution and diversity of wild and cultivated potatoes. *Nature* 606, 535–541. doi: 10.1038/s41586-022-04822-x

Wang, C., and Han, B. (2022). Twenty years of rice genomics research: from sequencing and functional genomics to quantitative genomics. *Mol. Plant* 15, 593–619. doi: 10.1016/j.molp.2022.03.009

Wang, J., Song, W., and Chai, J. (2023). Structure, biochemical function, and signaling mechanism of plant nlrs. *Mol. Plant* 16, 75–95. doi: 10.1016/j.molp.2022.11.011

Wang, M., Li, W., Fang, C., Xu, F., Liu, Y., Wang, Z., et al. (2018). Parallel selection on A dormancy gene during domestication of crops from multiple families. *Nat. Genet.* 50, 1435–1441. doi: 10.1038/s41588-018-0229-2

Wang, N., Tang, C., Fan, X., He, M., Gan, P., Zhang, S., et al. (2022). Inactivation of A wheat protein kinase gene confers broad-spectrum resistance to rust fungi. *Cell* 185, 2961–2974.e19. doi: 10.1016/j.cell.2022.06.027

Wen, X., Chen, Z., Yang, Z., Wang, M., Jin, S., Wang, G., et al. (2023). A comprehensive overview of cotton genomics, biotechnology and molecular biological studies. *Sci. China Life Sci.* 66, 2214–2256. doi: 10.1007/s11427-022-2278-0

Xie, L., Liu, M., Zhao, L., Cao, K., Wang, P., Xu, W., et al. (2021). Riceencode: A comprehensive epigenomic database as A rice encyclopedia of dna elements. *Mol. Plant* 14, 1604–1606. doi: 10.1016/j.molp.2021.08.018

Zelm, E. V., Zhang, Y., and Testerink, C. (2020). Salt tolerance mechanisms of plants. *Annu. Rev. Of Plant Biol.* 71, 403–433. doi: 10.1146/annurev-arplant-050718-100005

Zeng, D., Tian, Z., Rao, Y., Dong, G., Yang, Y., Huang, L., et al. (2017). Rational design of high-yield and superior-quality rice. *Nat. Plants* 3, 17031. doi: 10.1038/nplants.2017.31

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 30 November 2022
DOI 10.3389/fpls.2022.1072904

OPEN ACCESS

EDITED BY
Jun Li,
Zhejiang University, China

REVIEWED BY
Junzhe Wang,
Northwest A&F University, China
Fa Cui,
Ludong University, China

*CORRESPONDENCE
Shubing Shi
ssb@xjau.edu.cn
Lingrang Kong
lkong@sdau.edu.cn

†These authors have contributed equally to this work

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

RECEIVED 18 October 2022
ACCEPTED 08 November 2022
PUBLISHED 30 November 2022

CITATION
Yu H, Hao Y, Li M, Dong L, Che N, Wang L, Song S, Liu Y, Kong L and Shi S (2022) Genetic architecture and candidate gene identification for grain size in bread wheat by GWAS. *Front. Plant Sci.* 13:1072904. doi: 10.3389/fpls.2022.1072904

Genetic architecture and candidate gene identification for grain size in bread wheat by GWAS

Haitao Yu[1,2†], Yongchao Hao[3†], Mengyao Li[3†], Luhao Dong[3], Naixiu Che[3], Lijie Wang[2], Shun Song[2], Yanan Liu[2], Lingrang Kong[3*] and Shubing Shi[1*]

[1]College of Agriculture, Xinjiang Agricultural University, Urumqi, Xinjiang, China, [2]Wheat Research Institute, Weifang Academy of Agricultural Sciences, Weifang, Shandong, China, [3]State Key Laboratory of Crop Biology, Shandong Key Laboratory of Crop Biology, College of Agronomy, Shandong Agricultural University, Taian, China

Grain size is a key trait associated with bread wheat yield. It is also the most frequently selected trait during domestication. After the phenotypic characterization of 768 bread wheat accessions in three plots for at least two years, the present study shows that the improved variety showed significantly higher grain size but lower grain protein content than the landrace. Using 55K SNP assay genotyping and large-scale phenotyping population and GWAS data, we identified 5, 6, 6, and 6 QTLs associated with grain length, grain weight, grain area, and thousand grain weight, respectively. Seven of the 23 QTLs showed common association within different locations or years. Most significantly, the key locus associated with grain length, *qGL-2D*, showed the highest association after years of multi-plot testing. Haplotype and evolution analysis indicated that the superior allele of *qGL-2D* was mainly hidden in the improved variety rather than in landrace, which may contribute to the significant difference in grain length. A comprehensive analysis of transcriptome and homolog showed that *TraesCS2D02G414800* could be the most likely candidate gene for *qGL-2D*. Overall, this study presents several reliable grain size QTLs and candidate gene for grain length associated with bread wheat yield.

KEYWORDS

wheat, grain size, mapping, GWAS, yield

Introduction

Bread wheat is one of the major crops, accounting for nearly 20% of calories in our diet (http://faostat.fao.org). Improvement of yield remains a challenge under heavy population pressure and projected global change (Ray et al., 2013). Grain size is a major determinant of grain weight, besides the number of panicles per plant and the number of grains per panicle (Fan et al., 2006). In wheat breeding, grain size is usually evaluated by grain weight, which is positively correlated with grain length, grain width and grain thickness (Evans, 1972; Fan et al., 2006). Thus, it is vital to identify and introduce favorable genes or alleles controlling grain traits to improve the grain yield in bread wheat breeding.

Using linkage mapping, hundreds of grain size quantitative trait loci (QTLs) have been identified in the past few years (Zhang et al., 2018; Mora-Ramirez et al., 2021; Guo et al., 2022). Recently, multiple signals associated with grain size were detected in different populations *via* genome-wide association study (GWAS) (Breseghello and Sorrells, 2006a; Breseghello and Sorrells, 2006b; Pang et al., 2020). These QTLs are distributed on all the 21 chromosomes of bread wheat. However, the real genes underlying these QTLs have yet to be identified due to the complexity of parental mapping, QTL effect, QTL × genotype and QTL × QTL interactions. Using homology cloning, several orthologous genes associated with grain traits have been isolated and characterized in bread wheat. For instance, TaGW2 and TaGS5 were isolated in wheat based on *OsGW2* and *OsGS5* orthologs in rice (Wang et al., 2016; Zhai et al., 2018). *TaGW2* is involved in regulation of grain weight and grain number in bread wheat (Zhai et al., 2018). *TaGS5* is associated with thousand grain weight (Wang et al., 2016), *TaGW8* is related to grain size in bread wheat (Yan et al., 2019). It is still hard to determine the variation in natural elite alleles of these known genes that can be used in marker assisted selection (MAS) of bread wheat. Therefore, it is still very important to explore and identify new QTLs and their natural allelic variation in wheat breeding.

In this study, we constructed a GWAS panel with 768 bread wheat accessions. After phenotypic evaluation in multiple plots for several years, we performed GWAS to the identify grain size of QTLs. A total of 23 grain size QTLs were identified. For a major grain length QTL *qGL-2D*, we investigated the signatures of natural variation *via* comprehensive analysis of haplotype and evolutionary features. Finally, one candidate gene associated with *qGL-2D* was identified. The results suggest that grain size QTLs and grain length candidate genes as well as information may facilitate MAS of these loci/genes in breeding high-yield wheat in the future.

Materials and methods

Materials

A total of 768 bread wheat accessions were used to identify QTLs of grain size, including 683 Chinese resources (560 improved varieties and 123 landraces) and 85 introduced accessions. Field experiments were performed at three locations, i: the Shandong Agricultural University Agronomy Experimental Station in Tai'an from 2016 to 2019, ii: Weifang Academy of Agricultural Sciences in Weifang in 2019, and iii: Jining Academy of Agricultural Sciences in Jining in 2019. Each accession was planted in five-row plots with 5 cm distance between plants and 25 cm distance between rows. The interval between adjacent plots was 50 cm. At the mature stage, we harvest 10 spikes without any mechanical damage, disease or insect infestation. After threshing, we measured thousand grain weight (TGW), grain length (GL), grain width (GW), grain area (GA), grain perimeter (GP), grain roundness (GR), grain diameter (GD), length-to-width ratio (LWR), grain protein content (GPC) and grain starch content (GSC) for each accession using a Crop Grain Appearance Quality Scanning Machine (SC-E, Wanshen Technology Company, Hangzhou, China).

Genotyping

Genomic DNA was extracted from the seedling leaves of all 768 wheat accessions, followed by further genotyping *via* an Illumina 55K assay. Finally, a total of 47,743 of 53,063 SNPs were identified in the wheat panel. We estimated the whole-genome distribution and minor allele frequency (MAF) of these SNPs using an in-house Python script. Additionally, we performed quality control of SNPs to exclude those with high missing rate (> 50%) and low MAF (< 5%) for further analysis.

Population structure

We first extracted 45,298 SNPs with miss rate ≤ 0.5 and MAF ≥ 0.05 from 53,063 SNPs using an in-house Python script. Using PLINK (window size 50, step size 50, $r^2 \geq 0.3$), a total of 4,360 independent SNPs were further screened out based on r^2 of LD ≤ 0.3 (Purcell et al., 2007). The software STRUCTURE was used to calculate varying levels of K (K = 1-20) (Pritchard et al., 2000). We also performed principal component analysis (PCA) and kinship analysis using these independent SNPs and GAPIT software (Lipka et al., 2012; Tang et al., 2016). The phylogenetic analysis of *qGL-2D* was performed by generating a neighbor-joining tree using Mega 7 (Kumar et al., 2016).

Association mapping

Only 45,298 un-imputed SNPs with miss rate ≤ 0.5 and MAF ≥ 0.05 were used to conduct GWAS for GL, GW, GA and TGW, respectively. The first three PCs were used to construct the PC matrix. We performed GWAS with a Compressed Mixed Linear

Model (CMLM) *via* PCA and kinship analysis using default settings of GAPIT (Lipka et al., 2012; Tang et al., 2016). Additionally, the threshold to determine significant association was set at 1.0×10^{-5} after Bonferroni-adjusted correction (Pang et al., 2020).

Expression analysis and epidermal cell observation

Gene expression data from different wheat cultivars were used to analyze the gene expression profiles of the candidate region. Expression data were download from wheat-URGI website (https://wheat-urgi.versailles.inra.fr/Seq-Repository/Expression). Then the transcriptomic information of candidate genes were exacted by a custom python script. Epidermal tissues were peeled off using tweezers under a stereomicroscope. Then, the cell layers were stained with safranin and mounted on glass slides (Matsunami Glass Ind., Japan). The tissue specimens were subjected to observation with a light microscope (BX50F Olympus Optical Co., Ltd, Japan).

Screening of candidate genes for *qGL-2D*

In order to identify candidate genes for *qGL-2D*, LD heatmaps surrounding peaks were constructed using the R package "LD heatmap" (Shin et al., 2006). Using pairwise LD correlation ($r^2 > 0.6$), we mined the candidate regions of *qGL-2D* (Yano et al., 2016). We further investigated the expression of these candidate genes in bread wheat grain using typical materials belonging to different haplotypes.

Results

Population structure and grain characterization of 768 bread wheat accessions

To identify genetic loci associated with grain weight, a panel of 768 bread wheat accessions were constructed, including 560 improved varieties, 123 landraces and 85 introduced accessions. Using a 55K SNP assay, we obtained 47,743 SNPs of the panel. Subsets of these data were further filtered and used in additional analyses (Figure S1). A reasonable assessment of population structure facilitates the identification of real marker-trait associations (Crowell et al., 2016; Juliana et al., 2019). Therefore, we calculated varying levels of K means using un-imputed SNPs and STRUCTURE software (Golbeck, 1987). Landrace, improved and introduced varieties appeared clearly at K = 3 (Figure 1A). Further PCA indicated that top three PCs accounted for 17.09%, 6.15% and 3.38% of genetic variation within the bread wheat panel (Figure 1B). The results suggested obvious genetic differentiation between landrace and improved varieties of bread wheat.

A total of 10 traits were identified in three different plots for two years, including eight grain shape components (TGW, GL, GW, GA, GP, GR, GD, and LWR) and two grain quality components (GPC and GSC). All traits showed high heritability from 89.30% (GSC) to 95.27% (TGW) (Table S1). After obtaining the best linear unbiased prediction (BLUP) of each accession with respect to each trait across all traits, the coefficient of variation (CV) of all traits ranged from 1.44% GSC to 15.48% TGW (Table S1). GPC was proved to be negatively correlated with the eight grain size components, suggesting that larger, heavier and longer bread wheat grains usually had lower GPC (Figure S2). During the domestication of landrace to improved variety, bread wheat grains increased in size, weight, and length, but their GPC decreased (Figures 1C, D, 2C).

Identification of grain shape QTLs by GWAS

Focusing on four key grain shape traits (GL, GW, GA and TGW), GWAS was performed to identify QTLs based on their respective multi-year and multi-location data and BLUP. A total of 23 QTLs were detected on 12 chromosomes, including 5, 6, 6 and 6 QTLs for GL, GW, GA and TGW, respectively (Table 1 and Figures S3, S4). Seven of 23 QTLs showed common association within different locations or years, including *qGW-2B*, *qGL-2D*, *qGW-2D.1*, *qTGW-4A*, *qTGW-5A.1*, *qGA-6D*, *qTGW-6D* and *qTGW-7D*. Consistent with the positive correlations between GL, GW, GA and TGW (Figure S2), close linkage, and overlapping or one-factor-to-many-effects (pleiotropy) were detected on chromosome 2D (for *qGA-2D* and *qGL-2D*), chromosome 5A (for *qGA-5A*, *qTGW-5A.1* and *qGL-5A.1)*, chromosome 6D (for *qGA-6D* and *qTGW-6D)*, and chromosome 7D (for *qGA-7D*, *qGW-7D* and *qTGW-7D* (Table 1).

To validate the results of GWAS, we compared the localization of the QTLs identified in this study with previously detected QTLs associated with bi-parental mapping population. Twelve of 23 QTLs in this study were co-localized with previously reported QTLs, including 1, 6, 3 and 6 QTLs for GL, GW, GA, and TGW, respectively (Table 1). The *qGA-6D* and *qTGW-6D* were detected most frequently (five times), followed by *qGA-5A*, *qTGW-5A.1*, *qGL-5A.1*, *qGL-5A.2*, *qGA-7D*, *qGW-7D* and *qTGW-7D* (twice), whereas *qGW-2A*, *qGW-*

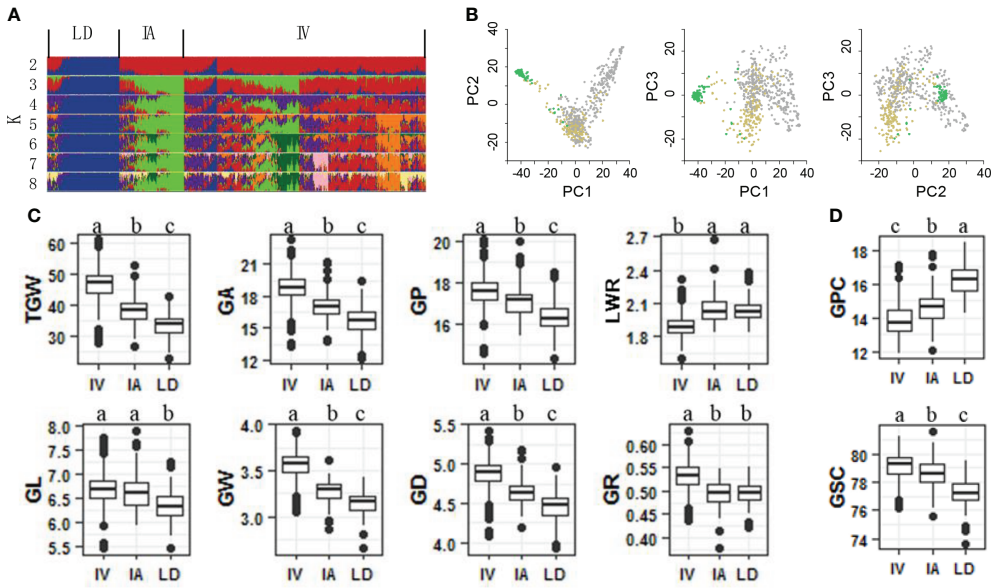


FIGURE 1
Genetic architecture and characteristic of grain size and grain quality of 768 bread wheat accessions. **(A)** Genetic structure of the panel analyzed using the program STRUCTURE. Landrace (LD), improved variety (IV) and introduced variety (IA) groups appeared at K = 3. **(B)** Principle components analysis reveals that the first 3 principle components explain 17.09%, 6.15% and 3.38% of the genetic variance within the panel. Comparison of grain size traits **(C)** and grain quality traits **(D)** among LD, IA, IV. Different letters above the boxes indicate significant differents ($p < 0.05$) when analyzed by Duncan's test.

2B, *qGW-2D.1*, *qGW-2D.2*, *qGW-3D*, *qTGW-4A*, *qGL-5A.2*, *qTGW-5A.2* and *qTGW-5B* were detected rarely (once). Additionally, we also identified six new grain size QTLs, including *qGA-1D.1*, *qGA-1D.2*, *qGA-2D*, *qGL-2D* and *qGL-4B*.

Haplotype analysis of *qGL-2D*

The *qGL-2D* was a key locus for GL, as it was detected using the data for each location every year and BLUP (Figures 3A, B and Figure S3). Using BLUP of GL yielded five significant SNPs (-log(p) > 5) representing *qGL-2D*. Thus, the five SNPs were identified *via qGL-2D* haplotype analysis. A total of seven haplotypes were detected, including two high-frequency haplotypes (HAP1 and HAP4, 36.6% and 56.4%), two low-frequency haplotypes (HAP2 and HAP3, 3.6% and 2.8%) and three rare haplotypes (HAP5-7, < 1%) (Figure 3C). Among them, GL was the shortest in HAP1 (6.56 mm), followed by HAP2 (6.57 mm) and HAP3 (6.70 mm), whereas HAP4 had the longest GL (Figure 3C). For other five traits were related to grain shape (GA, GW, GD and HGW) and grain quality (GPC). The HAP4 exhibited the greatest GA, GW, GD, and HGW, and the lowest GPC (Figure 3D). The results suggested that *qGL-2D* was widely involved in grain shape and grain quality.

To determine the evolutionary features of *qGL-2D*, we conducted a phylogenetic analysis of the seven haplotypes. Two major clades were formed (Figure 3E). One clade contained a widely divergent group, including HAP4, HAP3, HAP2 and HAP7, the most prevalent haplotypes associated with improved varieties of bread wheat. Another major haplotype in bread wheat landrace, HAP1, was clustered in the other clade (Figure 3E). In summary, the *qGL-2D* allele associated with improved varieties of bread wheat showed substantial genetic differences compared with bread wheat landrace, which could be attributed to selective effects on large grain during the process of modern bread wheat improvement.

Determination of candidate genes within *qGL-2D*

To analyze the candidate gene within *qGL-2D*, we defined the QTL region based on local LD. As indicated in the LD heatmap, an interval from 522,544,495 to 533,987,666 bp on chromosome 2D was an LD block with $r^2 > 0.6$ (Figure 2A). The *qGL-2D* contains 125 annotated genes. To further reduce the candidate number, we performed transcriptome analysis using one short-grain accession (Chinese Spring (HAP1)), two long-

TABLE 1 QTL identified for grain weight or shape by combined analysis of six environments and BLUP.

Chr.	QTL	Trait	Environments	Peak SNP	Position	-log10 (p)	QTL reported
1D	*qGA-1D.1*	GA	18T	AX-109817000	79999712	5.13	–
	qGA-1D.2		19T	AX-86164003	95559483	5.79	–
2A	*qGW-2A*	GW	19T	AX-109994744	721725535	5.56	Wang et al. (2012)
			BLUP	AX-109994744	721725535	5.65	
2B	*qGW-2B*		18T	AX-108936154	720581605	5.45	Zanke et al. (2015)
			19T	AX-108936154	720581605	5.53	
			BLUP	AX-108936154	720581605	5.89	
2D	*qGA-2D*	GA	20W	AX-108767381	528101770	5.29	–
			BLUP	AX-108767381	528101770	5.27	–
	qGL-2D	GL	17T	AX-110982403	525904353	6.41	–
			18T	AX-108767381	528101770	6.60	–
			19T	AX-108767381	528101770	8.63	–
			20J	AX-108767381	528101770	6.99	–
			20T	AX-108767381	528101770	7.00	–
			20W	AX-108767381	528101770	6.09	–
			BLUP	AX-108767381	528101770	8.31	–
	qGW-2D.1	GW	17T	AX-109910122	587284788	6.01	Ramya et al. (2010)
			18T	AX-94632592	593270570	6.13	
			19T	AX-109464110	585470933	5.87	
			20J	AX-109449735	590677250	6.79	
			20T	AX-94632592	593270570	6.70	
			20W	AX-111098468	593217154	5.99	
			BLUP	AX-111098468	593217154	7.06	
	qGW-2D.2		18T	AX-111956072	34428803	6.10	Wang et al. (2019a)
3D	*qGW-3D*		20T	AX-111624595	572830156	5.18	Ma et al. (2019)
4A	*qTGW-4A*	TGW	18T	AX-108908317	681180867	5.13	Zanke et al. (2015)
			19T	AX-108908317	681180867	5.22	
			BLUP	AX-108908317	681180867	5.30	
4B	*qGL-4B*	GL	20T	AX-110919438	643312159	5.69	–
5A	*qGA-5A*	GA	18T	AX-111136203	430037627	5.34	Cheng et al., (2017); Wu et al. (2015).
	qTGW-5A.1	TGW	19T	AX-110508884	428416559	5.02	
			20W	AX-110508884	428416559	5.15	
	qGL-5A.1	GL	18T	AX-111136203	430037627	5.14	
	qGL-5A.2		20J	AX-108762108	595372901	5.32	Wang et al. (2019b)
	qTGW-5A.2	TGW	20J	AX-109504344	704583912	5.25	Zanke et al. (2015)
5B	*qTGW-5B*		20T	AX-110427093	34285686	5.09	Yang et al. (2020)
5D	*qGL-5D*	GL	20W	AX-110985437	404832095	5.13	–
6D	*qGA-6D*	GA	17T	AX-110007215	93614544	5.12	Lopes et al. (2013), McCartney et al. (2005), Shi et al. (2017)
			18T	AX-110007215	93614544	6.82	
			20J	AX-110007215	93614544	5.60	
			20W	AX-110007215	93614544	5.87	
			BLUP	AX-110007215	93614544	5.79	
	qTGW-6D	TGW	17T	AX-110007215	93614544	5.52	
			18T	AX-110007215	93614544	8.46	
			19T	AX-110007215	93614544	6.49	
			20W	AX-110007215	93614544	6.03	
			BLUP	AX-110007215	93614544	6.26	

(Continued)

10.3389/fpls.2022.1072904

TABLE 1 Continued

Chr.	QTL	Trait	Environments	Peak SNP	Position	-log10 (p)	QTL reported
7D	*qGA-7D*	GA	20T	AX-110826147	65503524	5.60	Liu et al. (2014), Tang et al. (2017)
	qGW-7D	GW	20T	AX-110826147	65503524	5.55	
	qTGW-7D	TGW	20T	AX-110826147	65503524	5.87	
			18T	AX-111843581	67448018	5.25	

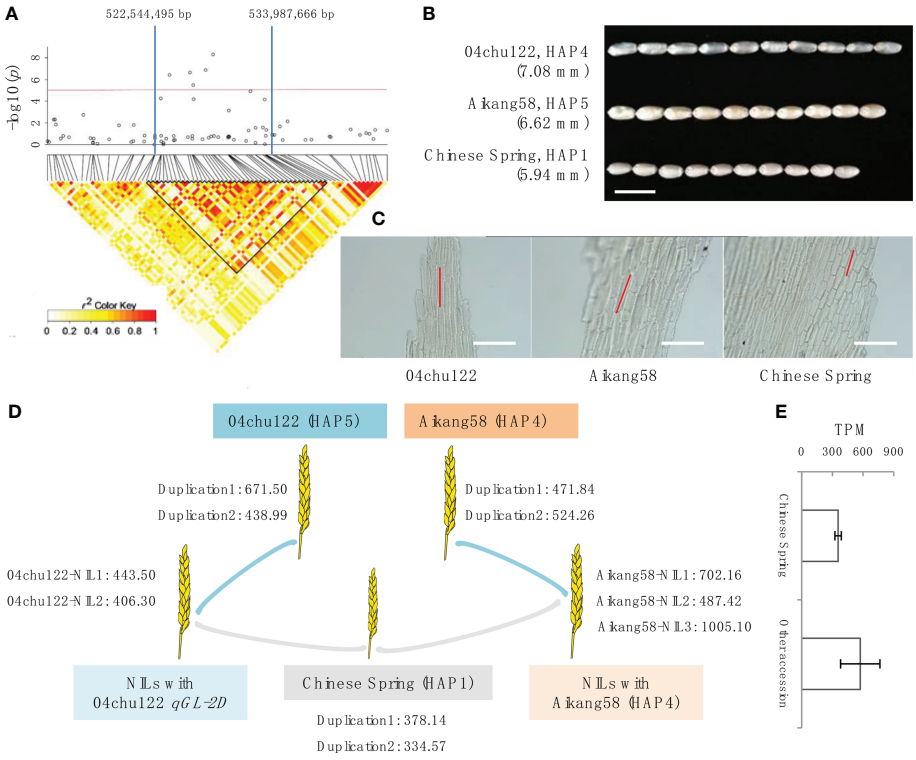


FIGURE 2
Identification and haplotype analysis of grain length QTL *qGL-2D*. Quantile-quantile (Q-Q) plot **(A)** and manhattan plot **(B)** using 7 groups of GL data of multi-year and multi-plots. *qGL-2D* is an association signal detected in all tests. **(C)** *qGL-2D* haplotype analysis and comparisons of grain length (GL) among four *qGL-2D* haplotypes. **(D)** Comparison of grain area (GA), grain perimeter (GP), grain width (GW), length-width ratio (LWR), grain diameter (GD), grain roundness (GR), hundred grain weight (HGW) and grain protein content (GPC) among four *qGL-2D* haplotypes. For better chart presentation, TGW is replaced by HGW. **(E)** Phylogenetic tree of the four *qGL-2D* haplotypes. The number of landrace (LD), improved variety (IV) and introduced variety (IV) are marked for four haplotypes.

grain bread wheat accessions (Aikang 58 (HAP4), 04chu122 (HAP5)) and 5 BC_2 near isogenic lines (NILs) carrying *qGL-2D* 04chu122 or aikang58 segment (Figure 2B). A total of 29 expressed genes were identified in eight accessions mentioned above (Table S2), and only *TraesCS2D02G414800* showed higher expression within two long-grain and eight NILs than in one short-grain accession (Figures 2D, E). Homology analysis showed that *TraesCS2D02G414800* encodes oleosin, which is involved in seed maturation and germination. Taken together, the results provide possible key candidates for further investigation of the molecular mechanism underlying GL within bread wheat.

Discussion

Grain size is one of the most frequently selected traits during domestication (Meyer and Purugganan, 2013; Zuo and Li, 2014).

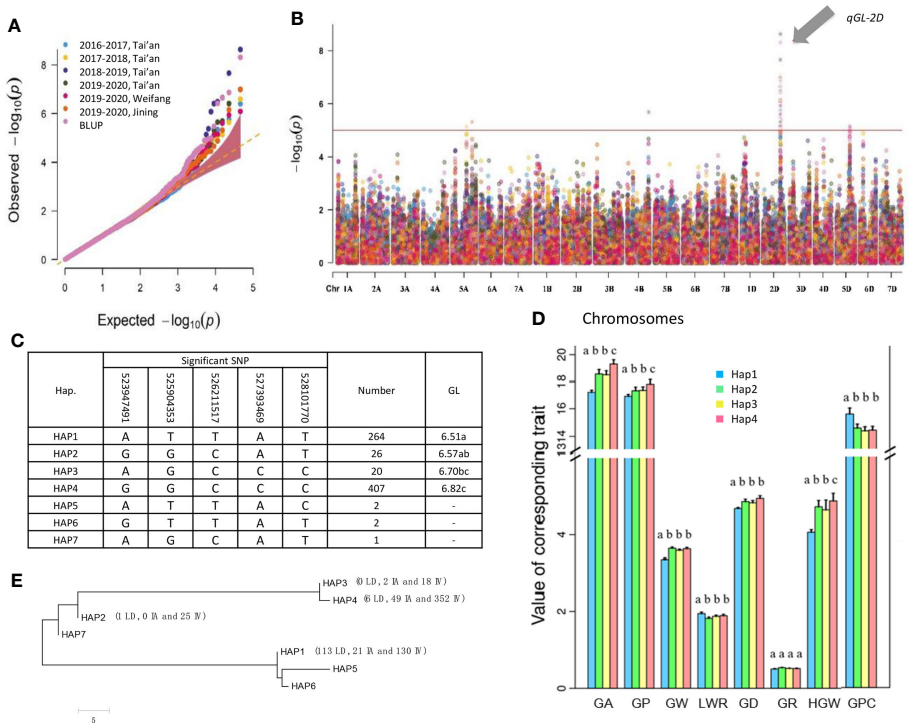


FIGURE 3
Determination of candidate genes within *qGL-2D*. **(A)** Association signals (top) and LD heatmap (bottom) of *qGL-2D*. Triangular block shows region with strong local LD (r^2 > 0.6). **(B)** Grain length of 04chu122, Aikang58 and Chinese Spring. Scale bar, 10 mm. **(C)** Epidermal cell length of 04chu122, Aikang58 and Chinese Spring. Scale bars, 200 um. **(D)** Expression level of TraesCS2D02G414800 in 04chu122, Aikang58, Chinese Spring and NILs carried 04chu122 *qGL-2D* or Aikang58 *qGL-2D*. **(E)** Comparison of expression level of Chinese Spring (HAP1), and the other accessions including 04chu122 (HAP5), Aikang58 (HAP4) and NILs carried 04chu122 *qGL-2D* and Aikang58 *qGL-2D*.

Among the many yield-related traits, increased grain size is the main factor associated with increased grain yield at a certain stage of domestication (Zheng et al., 2011). The grains of wild relatives are usually small and round in shape, and domestication has greatly increased the diversity of grain shape and size together with other changes (Fan et al., 2006). Grain size is predominantly determined by genetic factors, whereas grain filling is controlled by both genetic and environmental factors (Sakamoto and Matsuoka, 2008). Our study validated the significant changes in grain size of landrace to improved variety of bread wheat, and also suggested further accumulation of large-size alleles within improved variety rather than landrace. The most significant finding of the present study was the key locus for GL, *qGL-2D*, which showed the highest association after years of multi-plot testing. Haplotype and evolution analysis indicated that the superior allele of *qGL-2D* was mainly hidden in the improved variety rather than in landrace, which may result in significant difference in GL. Identification of the differential expression yielded a single candidate gene of *qGL-2D*. The results provide the opportunity for the delineation of the regulatory mechanism and related processes during grain development.

The coordination of grain size (weight) and grain quality is a major goal in breeding, as the increased grain size often reduces grain quality (Sakamoto and Matsuoka, 2008; Wang et al., 2012). Correlations between traits are a common biological phenomenon, especially those associated with determination of spike, growth duration, yield, and root and shoot (Crowell et al., 2016; Li et al., 2018; Zhao et al., 2019; Zhao et al., 2021). The present study indicated that the grain size increased while the GPC of bread wheat decreased from landrace to improved variety. The long-grain allele of *qGL-2D* showed a lower GPC, while the short-grain allele of *qGL-2D* showed a higher GPC. Pleiotropy and LD in natural population are usually considered as the main factors underlying this phenomenon, which is a major challenge in future breeding programs (Chen and Lübberstedt, 2010; Crowell et al., 2016). The role of two complementary genes associated with grain yield and grain quality requires further analysis (Zuo and Li, 2014).

Data availability statement

The data presented in the study are deposited in the OMIX repository (https://ngdc.cncb.ac.cn/omix/), accession number OMIX002373.

Author contributions

S.B.S. and L.R.K. designed and supervised the work; H.T.Y., M.Y.L., L.H.D., N.X.C., L.J.W., S.S. and Y.N.L. performed the research; H.T.Y. and Y.C.H. analyzed the data; H.T.Y. and S.B.S. wrote the paper. All authors read and approved the final manuscript.

Funding

This work was supported by the Natural Science Foundation of Shandong Province (ZR2020MC096 and ZR2021ZD31).

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.2022.1072904/full#supplementary-material

References

Breseghello, F., and Sorrells, M. E. (2006a). Association analysis as a strategy for improvement of quantitative traits in plants. *Crop Sci.* 46, 1323–1330. doi: 10.2135/cropsci2005.09-0305

Breseghello, F., and Sorrells, M. E. (2006b). Association mapping of kernel size and milling quality in wheat (Triticum aestivum l.) cultivars. *Genetics* 172, 1165–1177. doi: 10.1534/genetics.105.044586

Chen, Y., and Lübberstedt, T. (2010). Molecular basis of trait correlations. *Trends Plant Sci.* 15, 454–461. doi: 10.1016/j.tplants.2010.05.004

Cheng, R., Kong, Z., Zhang, L., Xie, Q., Jia, H., Yu, D., et al. (2017). Mapping QTLs controlling kernel dimensions in a wheat inter-varietal RIL mapping population. *Theor. Appl. Genet.* 130, 1405–1414. doi: 10.1007/s00122-017-2896-2

Crowell, S., Korniliev, P., Falcão, A., Ismail, A., Gregorio, G., Mezey, J., et al. (2016). Genome-wide association and high-resolution phenotyping link oryza sativa panicle traits to numerous trait-specific QTL clusters. *Nat. Commun.* 7, 10527. doi: 10.1038/ncomms10527

Evans, L. T. (1972). Storage capacity as a limitation on grain yield. *Rice Breeding.*, 499–511.

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

Golbeck, J. H. (1987). Structure, function and organization of the photosystem I reaction center complex. *Biochim. Biophys. Acta* 895, 167–204. doi: 10.1016/S0304-4173(87)80002-2

Guo, X., Fu, Y., Lee, Y. J., Chern, M., Li, M., Cheng, M., et al. (2022). The PGS1 basic helix-loop-helix protein regulates Fl3 to impact seed growth and grain yield in cereals. *Plant Biotechnol. J.* 20, 1311–1326. doi: 10.1111/pbi.13809

Juliana, P., Poland, J., Huerta-Espino, J., Shrestha, S., Crossa, J., Crespo-Herrera, L., et al. (2019). Improving grain yield, stress resilience and quality of bread wheat using large-scale genomics. *Nat. Genet.* 51, 1530–1539. doi: 10.1038/s41588-019-0496-6

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

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. doi: 10.1093/bioinformatics/bts444

Li, F., Xie, J., Zhu, X., Wang, X., Zhao, Y., Ma, X., et al. (2018). Genetic basis underlying correlations among growth duration and yield traits revealed by GWAS in rice (Oryza sativa l.). *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00650

Liu, G., Jia, L., Lu, L., Qin, D., Zhang, J., Guan, P., et al. (2014). Mapping QTLs of yield-related traits using RIL population derived from common wheat and Tibetan semi-wild wheat. *Theor. Appl. Genet.* 127, 2415–2432. doi: 10.1007/s00122-014-2387-7

Lopes, M. S., Reynolds, M. P., Mcintyre, C. L., Mathews, K. L., Jalal Kamali, M. R., Mossad, M., et al. (2013). QTL for yield and associated traits in the Seri/Babax population grown across several environments in Mexico, in the West Asia, north Africa, and south Asia regions. *Theor. Appl. Genet.* 126, 971–984. doi: 10.1007/s00122-012-2030-4

Ma, J., Zhang, H., Li, S., Zou, Y., Li, T., Liu, J., et al. (2019). Identification of quantitative trait loci for kernel traits in a wheat cultivar Chuannong16. *BMC Genet.* 20, 77. doi: 10.1186/s12863-019-0782-4

Mccartney, C. A., Somers, D. J., Humphreys, D. G., Lukow, O., Ames, N., Noll, J., et al. (2005). Mapping quantitative trait loci controlling agronomic traits in the spring wheat cross RL4452x'AC domain'. *Genome* 48, 870–883. doi: 10.1139/g05-055

Meyer, R. S., and Purugganan, M. D. (2013). Evolution of crop species: genetics of domestication and diversification. *Nat. Rev. Genet.* 14, 840–852. doi: 10.1038/nrg3605

Mora-Ramirez, I., Weichert, H., Wirén, N., Frohberg, C., Bodt, S., Schmidt, R. C., et al. (2021). Theda1 mutation in wheat increases grain size under ambient and elevated CO2 but not grain yield due to trade-off between grain size and grain number. *Plant-Environment Interact.* 2, 61–73. doi: 10.1002/pei3.10041

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

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

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

Ramya, P., Chaubal, A., Kulkarni, K., Gupta, L., Kadoo, N., Dhaliwal, H. S., et al. (2010). QTL mapping of 1000-kernel weight, kernel length, and kernel width in

bread wheat (*Triticum aestivum* l.). *J. Appl. Genet.* 51, 421–429. doi: 10.1007/BF03208872

Ray, D. K., Mueller, N. D., West, P. C., and Foley, J. A. (2013). Yield trends are insufficient to double global crop production by 2050. *PloS One* 8, e66428. doi: 10.1371/journal.pone.0066428

Sakamoto, T., and Matsuoka, M. (2008). Identifying and exploiting grain yield genes in rice. *Curr. Opin. Plant Biol.* 11, 209–214. doi: 10.1016/j.pbi.2008.01.009

Shi, S., Azam, F. I., Li, H., Chang, X., and Jing, R. (2017). Mapping QTL for stay-green and agronomic traits in wheat under diverse water regimes. *Euphytica* 213, 246. doi: 10.1007/s10681-017-2002-5

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. Software* 16, 1-9. doi: 10.18637/jss.v016.c03

Tang, Y., Liu, X., Wang, J., Li, M., Wang, Q., Tian, F., et al. (2016). GAPIT version 2: An enhanced integrated tool for genomic association and prediction. *Plant Genome* 9, 1–9. doi: 10.3835/plantgenome2015.11.0120

Tang, H., Wang, H., He, M., Zhang, M., Hu, Y., Li, Z., et al. (2017). QTL analysis for wheat falling number in a recombinant inbred line population segregated with 1BL/1RS translocation in a rainfed agricultural area of China. *Euphytica* 213, 235. doi: 10.1007/s10681-017-2028-8

Wang, S., Kun, W., Yuan, Q., Liu, X., Liu, Z., Lin, X., et al. (2012). Control of grain size, shape and quality by OsSPL16 in rice. *Nat. Genet.* 44, 950–954. doi: 10.1038/ng.2327

Wang, S., Yan, X., Wang, Y., Liu, H., Cui, D., and Chen, F. (2016). Haplotypes of the TaGS5-A1 gene are associated with thousand-kernel weight in chinese bread wheat. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.00783

Wang, X., Dong, L., Hu, J., Pang, Y., Hu, L., Xiao, G., et al. (2019b). Dissecting genetic loci affecting grain morphological traits to improve grain weight *via* nested association mapping. *Theor. Appl. Genet.* 132, 3115–3128. doi: 10.1007/s00122-019-03410-4

Wang, J., Li, X., Do Kim, K., Scanlon, M. J., Jackson, S. A., Springer, N. M., et al. (2019a). Genome-wide nucleotide patterns and potential mechanisms of genome divergence following domestication in maize and soybean. *Genome Biol.* 20, 74. doi: 10.1186/s13059-019-1683-6

Wu, Q. H., Chen, Y. X., Zhou, S. H., Fu, L., Chen, J. J., Xiao, Y., et al. (2015). High-density genetic linkage map construction and QTL mapping of grain shape and size in the wheat population Yanda1817 × Beinong6. *PloS One* 10, e0118144. doi: 10.1371/journal.pone.0118144

Yang, L., Zhao, D., Meng, Z., Xu, K., Yan, J., Xia, X., et al. (2020). QTL mapping for grain yield-related traits in bread wheat *via* SNP-based selective genotyping. *Theor. Appl. Genet.* 133, 857–872. doi: 10.1007/s00122-019-03511-0

Yano, K., Yamamoto, E., Aya, K., Takeuchi, H., Lo, P.-C., Hu, L., et al. (2016). Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nat. Genet.* 48, 927–934. doi: 10.1038/ng.3596

Yan, X., Zhao, L., Ren, Y., Dong, Z., Cui, D., and Chen, F. (2019). Genome-wide association study revealed that the TaGW8 gene was associated with kernel size in Chinese bread wheat. *Sci. Rep.* 9, 2702. doi: 10.1038/s41598-019-38570-2

Zanke, C., Ling, J., Plieske, J., Kollers, S., Ebmeyer, E., Korzun, V., et al (2015). Analysis of main effect QTL for thousand grain weight in European winter wheat (Triticum aestivum L.) by genome-wide association mapping. *Frontiers in Plant Science* 6. doi: 10.3389/fpls.2015.00644

Zhai, H., Feng, Z., Du, X., Song, Y., Liu, X., Qi, Z., et al. (2018). A novel allele of TaGW2-A1 is located in a finely mapped QTL that increases grain weight but decreases grain number in wheat (Triticum aestivum l.). *Theor. Appl. Genet.* 131, 539–553. doi: 10.1007/s00122-017-3017-y

Zhang, Y., Li, D., Zhang, D., Zhao, X., Cao, X., Dong, L., et al. (2018). Analysis of the functions of TaGW2 homoeologs in wheat grain weight and protein content traits. *Plant J.* 94, 857–866. doi: 10.1111/tpj.13903

Zhao, Y., Jiang, C., Rehman, R. M. A., Zhang, H., Li, J., and Li, Z. (2019). Genetic analysis of roots and shoots in rice seedling by association mapping. *Genes Genomics* 41, 95–105. doi: 10.1007/s13258-018-0741-x

Zhao, Y., Yin, Z., Wang, X., Jiang, C., Aslam, M. M., Gao, F., et al. (2021). Genetic basis and network underlying synergistic roots and shoots biomass accumulation revealed by genome-wide association studies in rice. *Sci. Rep.* 11, 13769. doi: 10.1038/s41598-021-93170-3

Zheng, T. C., Zhang, X. K., Yin, G. H., Wang, L. N., Han, Y. L., Chen, L., et al. (2011). Genetic gains in grain yield, net photosynthesis and stomatal conductance achieved in henan province of China between 1981 and 2008. *Field Crops Res.* 122, 225–233. doi: 10.1016/j.fcr.2011.03.015

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 21 December 2022
DOI 10.3389/fpls.2022.1089937

OPEN ACCESS

EDITED BY
Xueqiang Wang,
Zhejiang University, China

REVIEWED BY
Zitong Li,
Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia
Li Li,
University of New England, Australia

*CORRESPONDENCE
Hailan Liu
lhlzju@hotmail.com

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

RECEIVED 04 November 2022
ACCEPTED 06 December 2022
PUBLISHED 21 December 2022

CITATION
Liu H, Xia C and Lan H (2022) An efficient genomic prediction method without the direct inverse of the genomic relationship matrix. *Front. Plant Sci.* 13:1089937. doi: 10.3389/fpls.2022.1089937

An efficient genomic prediction method without the direct inverse of the genomic relationship matrix

Hailan Liu* , Chao Xia and Hai Lan

Maize Research Institute, Sichuan Agricultural University, Chengdu, Sichuan, China

GBLUP, the most widely used genomic prediction (GP) method, consumes large and increasing amounts of computational resources as the training population size increases due to the inverse of the genomic relationship matrix (GRM). Therefore, in this study, we developed a new genomic prediction method (RHEPCG) that avoids the direct inverse of the GRM by combining randomized Haseman–Elston (HE) regression (RHE-reg) and a preconditioned conjugate gradient (PCG). The simulation results demonstrate that RHEPCG, in most cases, not only achieves similar predictive accuracy with GBLUP but also significantly reduces computational time. As for the real data, RHEPCG shows similar or better predictive accuracy for seven traits of the *Arabidopsis thaliana* F2 population and four traits of the *Sorghum bicolor* RIL population compared with GBLUP. This indicates that RHEPCG is a practical alternative to GBLUP and has better computational efficiency.

KEYWORDS

genomic prediction, GBLUP, genomic relationship matrix, randomized Haseman–Elston regression, preconditioned conjugate gradient

Introduction

Currently, genomic prediction (GP) has been widely applied to many species, such as dairy cattle, dairy sheep, maize, and wheat (Pszczola et al., 2011; Duchemin et al., 2012; Crossa et al., 2014). For example, significant achievements have been made in the genetic improvement of dairy cattle *via* GP in many countries, such as the United States, Australia, Canada, New Zealand, and France (Hayes et al., 2009; Winkelman et al., 2015; García-Ruiz et al., 2016; Weller et al., 2017). Moreover, GP helps to optimize the breeding procedure when used with many other breeding technologies. For example, it can accelerate the selection of superior pure lines from the large numbers of those generated by doubled haploid (DH) technology, which is otherwise a significant problem in terms of the consumption of time and money (Wang et al., 2020). Additionally, GP can rapidly

increase the frequencies of favorable alleles when combined with genome editing (GE) (Jenko et al., 2015; Bastiaansen et al., 2018).

Many computational methods of GP have been proposed and GBLUP is the most widely used (Meuwissen et al., 2001; Daetwyler et al., 2013; Mouresan et al., 2019). For conventional GBLUP, restricted maximum likelihood (REML) is often used to estimate heritability, and its computational complexity is cubic of the training population size (Xu et al., 2014). The fact that the inverse of the genomic relationship matrix (GRM) is essential when estimating the heritability in REML and calculating the best linear unbiased prediction (BLUP) contributes to the decrease of the computational efficiency of GBLUP when the size of the training population increases. To improve computational efficiency, methods such as IBS-based HE regression, algorithm for proven and young (APY), updating the inverse, recursive algorithm, spectral decomposition, and the preconditioned conjugate gradient (PCG) algorithm are employed (Kang et al., 2008; Legarra and Misztal, 2008; Misztal et al., 2009; Endelman, 2011; Faux et al., 2012; Meyer et al., 2013; Chen, 2014; Misztal, 2016; Liu and Chen, 2017; Masuda et al., 2017; Vandenplas et al., 2018; Vandenplas et al., 2020). In particular, PCG solves mixed model equations (MMEs) *via* iteration instead of by directly inversing the GRM. Recently, we proposed a fast GP method (SHEAPY) combining randomized Haseman–Elston regression (RHE-reg) and a modified APY (Liu and Chen, 2022). In the SHEAPY, RHE-reg is used to estimate heritability because of its high computational speed. In this study, we continue to combine it with PCG, calculating marker values to develop a new GP method (RHEPCG), which can significantly improve computational efficiency without the direct inverse of GRM.

Materials and methods

Genetic model and the linear system of MMEs

Herein, we only focus on additive effects, and the basic model is described as:

$$\mathbf{y} = \mathbf{X\theta} + \mathbf{Zu} + \mathbf{e} \quad , \quad (1)$$

in which y is the $n\times1$ vector of the standardized phenotypic values; $\boldsymbol{\theta}$ is a fixed effect; $\mathbf{X}$ is the $n\times1$ vector of the incidence; $\boldsymbol{Z}$ is the $n\times m$ matrix of the standardized genotypic values; $\mathbf{u}$ is the $m\times1$ vector of SNP marker effects; and $\mathbf{e}$ is the $n \times 1$ vector of the residual error.

On the basis of the above genetic model of additive effect, the linear system of MMEs was as follows:

$$\begin{bmatrix} \mathbf{X^TX} & \mathbf{X^TZ} \\ \mathbf{Z^TX} & \mathbf{Z^TZ} + \mathbf{I}\frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\theta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T y \\ \mathbf{Z}^T y \end{bmatrix} \quad , \quad (2)$$

in which I is the identity matrix, σ_g^2 is the variance of SNP marker effects, and σ_e^2 is the residual variance.

To solve MMEs, randomized HE-reg based on IBS was used to estimate heritability, which was then introduced into Equation 2 with residual variance. Then, PCG was used to solve Equation 2 to obtain marker values.

Estimating heritability *via* randomized HE-reg based on IBS

IBS-based RHE regression is a method of moment that can reduce computational time and memory to $\mathcal{O}(\frac{nmk}{\max(log_3^{(n)}log_3^{(m)})} + nm)$ and $O(nm)$, respectively (n , m , and k represent the number of samples, the number of markers, and the length of random vector, respectively) (Wu and Sankararaman, 2018; Liu and Chen, 2022). Here, it was used to estimate the heritability:

$$y_i y_i = b_0 + b_1 \omega_{ij} + e \quad (3)$$

in which y_i and y_j represent the phenotypic values of individuals i and. from the training population; b_0 is the intercept; b_1 is the regression coefficient; is the genetic relatedness ($\omega_{ij} = \frac{\mathrm{ZiZj}\}}{\mathrm{m}}$) between a pair of individuals i and j; z_i and z_j are the genotype vector of individuals i and j; and is residual error. For a trait, its phenotypic variance is $\hat{\sigma}_y^2 = \frac{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}{n-1}$, its additive genetic variance is $\hat{\sigma}_g^2 = \hat{b}_1$, and its error variance is $\hat{\sigma}_e^2 = \hat{\sigma}_y^2 - \hat{\sigma}_g^2$. The computational equation of $\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$ is described as:

$$\begin{bmatrix} tr[\boldsymbol{\omega}^2] & tr[\boldsymbol{\omega}] \\ tr[\boldsymbol{\omega}] & n \end{bmatrix} \begin{bmatrix} \hat{\sigma}_g^2 \\ \hat{\sigma}_e^2 \end{bmatrix} = \begin{bmatrix} y^T\boldsymbol{\omega}y \\ y^T y \end{bmatrix}, \quad (4)$$

in which $\omega = \frac{ZZ'}{m}$ corresponds to the genomic related matrix between individuals and $\widehat{tr[\omega]} = n$. To accelerate computational efficiency, $\widehat{tr[\omega^2]}$ was calculated *via* a randomized estimation. The equation is as below:

$$\widehat{tr[\boldsymbol{\omega}^2]} = \frac{1}{s}\frac{1}{m^2}\sum_{s=1}^{S}(\mathbf{w}_s'\mathbf{ZZ}^T)(\mathbf{ZZ}^T\mathbf{w}_s). \quad (5)$$

In the equation, S represents the rounds of randomization implemented, and was set as 5 throughout the study; each entry of $\boldsymbol{w}_s$ comes from a standard normal distribution N (0,1).

Preconditioned conjugate gradient (PCG) algorithm

When the MMEs are described as $\mathbf{Ax} = \mathbf{b}$, in which $\mathbf{A}$ is the coefficient matrix, $\mathbf{x}$ is the vector of solutions, and $\mathbf{b}$ is the right-hand side, the PCG is used to solve the linear system of MMEs and compute marker effects. As it does not need to invert GRM like conventional methods, a much higher efficiency can be achieved (Vandenplas et al., 2019). Its code is as follows (Tsuruta et al., 2001; Vandenplas et al., 2018):

When $\boldsymbol{n} = \mathbf{0}$,

x_0=1 ; e_0=0 ; α_0=1 (1 is a vector of containing 1.);

$$r_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0 \quad ;$$

$$p_0 = \mathbf{M}^{-1}\mathbf{r}_0 \quad ;$$

When n=1, 2….,

$$w_n = M^{-1}\mathbf{p}_{\mathbf{n}-1} \quad ;$$

$$\alpha_n = p'_{n-1}w_n \quad ;$$

$$\beta_n = {}^{\alpha_n}\!/_{\alpha_{n-1}} \quad ;$$

$$\alpha_{n-1} = \alpha_n \quad ;$$

$$e_n = w_n + e_{n-1}\beta_n \quad ;$$

$$q_n = \mathbf{A}e_n \quad ;$$

$$\epsilon_n = {}^{p'_{n-1}w_n}\!/_{e'_n q_n} \quad ;$$

$$x_n = x_{n-1} + e_n\epsilon_n \quad ;$$

$$r_n = r_{n-1} - \epsilon_n q_n$$

Until convergence.

End.

Here, $\mathbf{A} = \begin{bmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} + \mathbf{I}\frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix}$, M is the preconditioner matrix, and $\mathbf{M}$ = diag ($\mathbf{A}$); r , p , and w are vectors, $\mathbf{x} = \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\mu}} \end{bmatrix}$, and $\mathbf{b} = \begin{bmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \end{bmatrix}$. To solve MMEs, $\hat{\sigma}_g^2$ *via* RHE regression based on IBS and $\hat{\sigma}_e^2$ are introduced into $\mathbf{A}$ matrix.

Simulated data

The F2 population was simulated to evaluate the performance and cost time of GBLUP and RHEPCG. We simulated a chromosome with a length of 2,000 cM (the recombination rate was c between the i^{th} and $(i+1)^{th}$ markers), and all markers in this chromosome were defined as QTL, the effects of which followed a standard normal distribution. A series of different training population sizes (1,000, 1,200, 2,000, 6,000, 10,000, 15,000, and 20,000), candidate population sizes (100, 200, 300, and 400), and heritability (0.2, 0.4, 0.6, 0.65, and 0.8) were simulated. Each simulation scenario included 10 replications.

Real data

Two sets of data (*Arabidopsis thaliana* and *Sorghum bicolor*) were used to evaluate the predictive accuracy of GBLUP and RHEPCG. (1) An *A. thaliana* F2 population (P15) with 434 individuals derived from a cross between Br-0 and C24 was obtained from the study by Salomé et al. It consisted of a total of 233 SNP markers and seven traits, including DTF1 (days until visible flower buds in the center of the rosette), DTF2 (days until inflorescence stem reached 1 cm in height), DTF3 (days until first open flower), RLN (rosette leaf number), CLN (cauline leaf number), TLN (total leaf number: sum of RLN and CLN), and LIR1 (leaf initiation rate [RLN/DTF1]) (Salomé et al., 2011). (2) A *S. bicolor* RIL population with 399 individuals derived from a cross between *S. bicolor* BTx623 and *S. bicolor* IS3620C was obtained from the study by Kong et al. It consisted of a total of 381 bins and five traits, including PH (plant height), BTF (base to flag length), FTR (flag to rachis length), ND (number of nodes), and FL (days to flowering). The phenotype data were obtained from the University of Georgia Plant Science Farm, Watkinsville, GA, USA on May 10, 2011 (Kong et al., 2018).

Implementation and computations

The GBLUP and RHEPCG were written in R language (R Core Team, 2017) and run on a server of the CentOS Linux operating system (Intel (R) Xeon (R) CPU E7-4870 @2.40GHz) with 80 CPUs and 755G memory. The RHEPCG program is available from the authors. The squared correlation coefficient (r^2) between the phenotypes and the predicted genotypic values were defined as the prediction accuracy.

Results and discussion

Comparison of GBLUP and RHEPCG in simulated F2 population studies

A series of simulations of the F2 population at different levels of parameters, including training population size, candidate population size, and heritability were used to assess the estimated heritability, predictive accuracy, and consumption time of GBLUP and RHEPCG.

Table 1 shows the predictive accuracy and computational time of the GBLUP and RHEPCG at different training population sizes (1,000, 2,000, 6,000, 10,000, 15,000, and 20,000). As the training population size increased, both methods demonstrated an obvious uptrend of predictive accuracy. When the training population size was 1,000, RHEPCG was slightly better than GBLUP in predictive accuracy ($r^2_{RHEPCG} = 0.597 \pm 0.007$ vs $r^2_{GBLUP} = 0.541 \pm 0.027$), but when the training population size was 10,000, an opposite result was achieved ($r^2_{GBLUP} = 0.640 \pm 0.014$ vs $r^2_{RHEPCG} = 0.607 \pm 0.023$). In other conditions, both methods performed similarly (for example, when a training population size was 20,000 $r^2_{GBLUP} = 0.653 \pm 0.017$ vs. $r^2_{RHEPCG} = 0.648 \pm 0.021$). With the enlargement of the training population, the predictive accuracy approximated the true heritability, which as some studies have demonstrated, is

TABLE 1 Comparison of the estimated heritability, predictive accuracy, and computational time of GBLUP and RHEPCG at the different training population sizes based on 10 simulations in the *Arabidopsis thaliana* F2 population.

Training size	GBLUP			RHEPCG		
	$\hat{h}^2_{GBLUP}$	r^2_{GBLUP}	Average time of each simulation (s)	$\hat{h}^2_{RHEPCG}$	r^2_{RHEPCG}	Average time of each simulation (s)
1,000	0.718 ±0.055	0.541 ±0.027	135	0.637 ±0.042	0.597 ±0.007	31
2,000	0.648 ±0.029	0.576 ±0.011	465	0.625 ±0.051	0.570 ±0.010	58
6,000	0.668 ±0.052	0.615 ±0.011	1,580	0.621 ±0.055	0.594 ±0.013	179
10,000	0.680 ±0.042	0.640 ±0.014	7,932	0.600 ±0.043	0.607 ±0.023	526
15,000	0.650 ±0.038	0.651 ±0.013	22,576	0.724 ±0.059	0.643 ±0.024	820
20,000	0.698 ±0.042	0.653 ±0.017	53,666	0.728 ±0.043	0.648 ±0.021	1,237

The training population sizes were 1,000, 2,000, 6,000, 10,000, 15,000, and 20,000; the candidate population size was 100; the length of the chromosome was 2,000; heritability (h^2) was set as 0.65; the recombination rate © was set as 0.01. $\hat{h}^2_{GBLUP}$ and $\hat{h}^2_{RHEPCG}$ represent the estimated heritability via GBLUP and RHEPCG, respectively; r^2_{GBLUP} and r^2_{RHEPCG} represent the squared correlation coefficient between the phenotypes and the predicted genotypic values and are defined as the prediction accuracy; the values after ± represent the corresponding standard error.

the upper bound of predictive accuracy (de los Campos et al., 2013; Liu and Chen, 2018). Meanwhile, RHEPCG was significantly faster than GBLUP (for example, when a training population size was 20,000 T_{GBLUP}=53666s vs T_{RHEPCG}=1237s). When the training population size was 1,000, 2,000, 6,000, 10,000, 15,000, and 20,000, the computational time of GBLUP was 4, 8, 9, 15, 28, and 43 times that of RHEPCG, respectively. In other words, the larger the training population size, the more obvious the advantage of the computational efficiency of RHEPCG becomes.

In Table 2, the predictive accuracy of both methods was similar at different candidate population sizes (100, 200, 300, and 400), which means the latter has no significant impact on the former. Table 3 shows that the predictive accuracy of GBLUP and RHEPCG increased when heritability varied from 0.2 to 0.4, 0.6, and 0.8. According to further analysis, the correlation between the estimated heritability and the predictive accuracy was 0.999 for both GBLUP and RHEPCG ($P_{Two-tailed}$=0.001), and our results are consistent with Daetwyler et al. (2008) in that heritability can significantly influence predictive accuracy.

Currently, IBS-based RHE regression is used to estimate gene-environmental heritability and multi-trait genetic correlation (Kerin and Marchini, 2020; Wu et al., 2022). Therefore, RHEPCG can also be applied to such data *via* the incorporation of these effects into the model in the future.

Comparison of GBLUP and RHEPCG in studies of the *A. thaliana* F2 and *S. bicolor* RIL populations

A comparison of GBLUP and RHEPCG based on seven traits of the *A. thaliana* F2 population was performed in this study. Table 4 shows a significant difference between the estimated heritability *via* GBLUP and that *via* RHEPCG in

TABLE 2 Comparison of the predictive accuracy of GBLUP and RHEPCG at the different candidate population sizes based on 10 simulations in the *Arabidopsis thaliana* F2 population.

Candidate population	r^2_{GBLUP}	r^2_{RHEPCG}
100	0.581 ± 0.022	0.560 ± 0.016
200	0.561 ± 0.015	0.568 ± 0.016
300	0.561 ± 0.014	0.578 ± 0.008
400	0.580 ± 0.008	0.571 ± 0.011

The training population sizes was 1,200; the candidate population sizes were 100, 200, 300, and 400; the length of the chromosome was 2,000; heritability (h^2) was set as 0.65; the recombination rate (c) was set as 0.01. r^2_{GBLUP} and r^2_{RHEPCG} represent the squared correlation coefficient between the phenotypes and the predicted genotypic values and are defined as the prediction accuracy; the values after ± represent the corresponding standard error.

TABLE 3 Comparison of the estimated heritability and predictive accuracy of GBLUP and RHEPCG at different levels of heritability based on 10 simulations in the *Arabidopsis thaliana* F2 population.

h^2	GBLUP		RHEPCG	
	$\hat{h}^2_{GBLUP}$	r^2_{GBLUP}	$\hat{h}^2_{RHEPCG}$	$\hat{h}^2_{GBLUP}$
0.2	0.220 ± 0.014	0.165 ± 0.021	0.203 ± 0.009	0.127 ± 0.017
0.4	0.460 ± 0.021	0.346 ± 0.027	0.406 ± 0.014	0.304 ± 0.018
0.6	0.708 ± 0.032	0.535 ± 0.025	0.611 ± 0.019	0.506 ± 0.016
0.8	0.910 ± 0.034	0.667 ± 0.032	0.817 ± 0.024	0.726 ± 0.012

The training population sizes was 1,200; the candidate population size was 100; the length of the chromosome was 2,000; heritability (h^2) was set as 0.2, 0.4, 0.6, and 0.8; the recombination rate (c) was set as 0.01. $\hat{h}^2_{GBLUP}$ and $\hat{h}^2_{RHEPCG}$ represent the estimated heritability via GBLUP and RHEPCG, respectively; r^2_{GBLUP} and r^2_{RHEPCG} represent the squared correlation coefficient between the phenotypes and the predicted genotypic values and are defined as the prediction accuracy; the values after ± represent the corresponding standard error.

TABLE 4 Comparison of the predictive accuracy between GBLUP and RHEPCG in seven traits from the *Arabidopsis thaliana* F2 (P15) population based on 10 simulations.

Trait	Training	Candidate	GBLUP		RHEPCG	
			$\hat{h}^2_{GBLUP}$	r^2_{GBLUP}	$\hat{h}^2_{RHEPCG}$	r^2_{RHEPCG}
DTF1	300	133	0.731 ± 0.025	0.383 ± 0.017	0.323 ± 0.034	0.368 ± 0.024
DTF2	300	134	0.604 ± 0.035	0.406 ± 0.024	0.389 ± 0.029	0.401 ± 0.023
DTF3	300	134	0.645 ± 0.033	0.351 ± 0.024	0.311 ± 0.030	0.321 ± 0.025
RLN	300	131	0.923 ± 0.025	0.555 ± 0.046	0.524 ± 0.030	0.640 ± 0.015
CLN	300	130	0.570 ± 0.017	0.350 ± 0.017	0.282 ± 0.019	0.342 ± 0.018
TLN	300	130	0.910 ± 0.022	0.572 ± 0.029	0.421 ± 0.028	0.619 ± 0.011
LIR1	300	131	0.449 ± 0.033	0.208 ± 0.012	0.162 ± 0.011	0.197 ± 0.017

$\hat{h}^2_{GBLUP}$ and $\hat{h}^2_{RHEPCG}$ represent the estimated heritability via GBLUP and RHEPCG, respectively; r^2_{GBLUP} and r^2_{RHEPCG} represent the squared correlation coefficient between the phenotypes and the predicted genotypic values and are defined as the prediction accuracy; the values after ± represent the corresponding standard error.

TABLE 5 Comparison of the predictive accuracy between GBLUP and RHEPCG in five traits from the *Sorghum bicolor* RIL population based on 10 simulations.

Trait	Training	Candidate	GBLUP		RHEPCG	
			$\hat{h}^2_{GBLUP}$	$\hat{r}^2_{GBLUP}$	$\hat{h}^2_{RHEPCG}$	$\hat{r}^2_{RHEPCG}$
PH	300	88	0.653 ± 0.013	0.289 ± 0.011	0.331 ± 0.038	0.257 ± 0.016
BTF	300	88	0.619 ± 0.011	0.335 ± 0.019	0.436 ± 0.047	0.302 ± 0.017
FTR	300	88	0.337 ± 0.008	0.182 ± 0.018	0.493 ± 0.056	0.155 ± 0.026
ND	300	88	0.490 ± 0.011	0.306 ± 0.015	0.747 ± 0.065	0.197 ± 0.039
FL	300	93	0.629 ± 0.014	0.371 ± 0.014	0.622 ± 0.043	0.393 ± 0.031

$\hat{h}^2_{GBLUP}$ and $\hat{h}^2_{RHEPCG}$ represent the estimated heritability via GBLUP and RHEPCG, respectively; $\hat{r}^2_{GBLUP}$ and$\hat{r}^2_{RHEPCG}$ represent the squared correlation coefficient between the phenotypes and the predicted genotypic values and are defined as the prediction accuracy; the values after ± represent the corresponding standard error.

seven traits of *A. thaliana* F2 (P15). Meanwhile, the seven traits were used to evaluate the predictive accuracy of GBLUP and RHEPCG (Table 4). The two methods showed similar predictive accuracy in six traits: DTF1, DTF2, DTF3, CLN, TLN and LIR1 (for example, the predictive accuracies of DTF1 were r^2_{GBLUP} = 0.383 ± 0.017 and r^2_{RHEPCG} = 0.368 ± 0.024). RHEPCG was significantly better than GBLUP for RLN (the predictive accuracies of RLN were r^2_{GBLUP} = 0.555 ± 0.046 and r^2_{RHEPCG} = 0.640 ± 0.015).

In addition, the predictive accuracy of GBLUP and RHEPCG was evaluated based on five traits of the *S. bicolor* RIL population (Table 5). The estimated heritability of PH, BTF, FTR, and ND

via GBLUP differed significantly from that *via* RHEPCG, and the two methods had similar predictive accuracy for PH, BTF, FTR, and FL (for example, the predictive accuracies of PH were $r^2_{GBLUP} = 0.289 \pm 0.011$ and $r^2_{RHEPCG} = 0.257 \pm 0.016$), and the predictive accuracy of GBLUP was significantly superior to that of RHEPCG for ND (the predictive accuracies of ND were $r^2_{GBLUP} = 0.306 \pm 0.015$ and $r^2_{RHEPCG} = 0.197 \pm 0.039$).

These results show that GBLUP and RHEPCG have different estimated heritability in some traits of *A. thaliana* and *S. bicolor*. According to Chen (2016), strong selection can lead to differences in the estimated heritability *via* LMM and HE, and therefore, these traits are very likely to have undergone strong selection. In the future, we will investigate the influence of strong selection on predictive accuracy.

Conclusion

We present a new computing method of genomic prediction (RHEPCG) that does not require direct inversion of the GRM. Compared with GBLUP, it can significantly reduce computational time while maintaining predictive accuracy.

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

HLL conceived and performed the study, interpreted the results, and wrote the manuscript. HL and CX interpreted the results and wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by the National Natural Science Foundation of China (32271984).

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.

References

Bastiaansen, J. W. M., Bovenhuis, H., Groenen, M. A. M., Megens, H. J., and Mulder, H. A. (2018). The impact of genome editing on the introduction of monogenic traits in livestock. *Genet. Selection Evol.* 50, 18. doi: 10.1186/s12711-018-0389-7

Chen, G. B. (2014). Estimating heritability of complex traits from genome-wide association studies using IBS-based haseman-elston regression. *Front. Genet.* 5. doi: 10.3389/fgene.2014.00107

Chen, G. B. (2016). On the reconciliation of missing heritability for genome-wide association studies. *Eur. J. Hum. Genet.* 24, 1810–1816. doi: 10.1038/ejhg.2016.89

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

Daetwyler, H. D., Calus, M. P. L., Pong-Wong, R., de los Campos, G., and Hickey, J. M. (2013). Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics* 193, 347–365. doi: 10.1534/genetics.112.147983

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

de los Campos, G., Vazquez, A. I., Fernando, R., Klimentidis, Y. C., and Sorensen, D. (2013). Prediction of complex human traits using the genomic best linear unbiased predictor. *PLos Genet.* 9, e1003608. doi: 10.1371/journal.pgen.1003608

Duchemin, S. I., Colombani, C., Legarra, A., Baloche, G., Larroque, H., Astruc, J. M., et al. (2012). Genomic selection in the French lacaune dairy sheep breed. *J. Dairy Sci.* 95, 2723–2733. doi: 10.3168/jds.2011-4980

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

Faux, P., Gengler, N., and Misztal, I. (2012). A recursive algorithm for decomposition and creation of the inverse of the genomic relationship matrix. *J. Dairy Sci.* 95, 6093–6102. doi: 10.3168/jds.2011-5249

García-Ruiz, A., Cole, J. B., VanRaden, P. M., Wiggans, G. R., Ruiz-López, F. J., and van Tassell, C. P. (2016). Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. *Proc. Natl. Acad. Sci. U.S.A.* 113, E3995–E4004. doi: 10.1073/pnas.1519061113

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

Jenko, J., Gorjanc, G., Cleveland, M. A., Varshney, R. K., Whitelaw, C. B. A., Woolliams, J. A., et al. (2015). Potential of promotion of alleles by genome editing to improve quantitative traits in livestock breeding programs. *Genet. Selection Evol.* 47, 55. doi: 10.1186/s12711-015-0135-3

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, 1709–1723. doi: 10.1534/genetics.107.080101

Kerin, M., and Marchini, J. (2020). A non-linear regression method for estimation of gene–environment heritability. *Bioinformatics* 36, 5632–5639. doi: 10.1093/bioinformatics/btaa1079

Kong, W., Kim, C., Zhang, D., Guo, H., Tan, X., Jin, H., et al. (2018). Genotyping by sequencing of 393 *Sorghum bicolor* BTx6233×IS3620C recombinant inbred lines

improves sensitivity and resolution of QTL detection. *G3 Genes|Genomes|Genetics* 8, 2563–2572. doi: 10.1534/g3.118.200173

Legarra, A., and Misztal, I. (2008). Technical note: Computing strategies in genome-wide selection. *J. Dairy Sci.* 91, 360–366. doi: 10.3168/jds.2007-0403

Liu, H., and Chen, G. B. (2017). A fast genomic selection approach for large genomic data. *Theor. Appl. Genet.* 130, 1277–1284. doi: 10.1007/s00122-017-2887-3

Liu, H., and Chen, G. B. (2018). A new genomic prediction method with additive-dominance effects in the least-squares framework. *Heredity* 121, 196–204. doi: 10.1038/s41437-018-0099-5

Liu, H., and Chen, G. B. (2022). A novel genomic prediction method combining randomized haseman-elston regression with a modified algorithm for proven and young for large genomic data. *Crop J.* 10, 550–554. doi: 10.1016/j.cj.2021.09.001

Masuda, Y., Misztal, I., Legarra, A., Tsuruta, S., Lourenco, D. A. L., Fragomeni, B. O., et al. (2017). Technical note: Avoiding the direct inversion of the numerator relationship matrix for genotyped animals in single-step genomic best linear unbiased prediction solved with the preconditioned conjugate gradient. *J. Anim. Sci.* 95, 49–52. doi: 10.2527/jas.2016.0699

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

Meyer, K., Tier, B., and Graser, H. U. (2013). Technical note: updating the inverse of the genomic relationship matrix. *J. Anim. Sci.* 91, 2583–2586. doi: 10.2527/jas.2012-6056

Misztal, I. (2016). Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. *Genetics* 202, 401–409. doi: 10.1534/genetics.115.182089

Misztal, I., Legarra, A., and Aguilar, I. (2009). Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *J. Dairy Sci.* 92, 4648–4655. doi: 10.3168/jds.2009-2064

Mouresan, E. F., Selle, M., and Rönnegård, L. (2019). Genomic prediction including SNP-specific variance predictors. *G3 Genes|Genomes|Genetics* 9, 3333–3343. doi: 10.1534/g3.119.400381

Pszczola, M., Mulder, H. A., and Calus, M. P. L. (2011). Effect of enlarging the reference population with (un)genotyped animals on the accuracy of genomic selection in dairy cattle. *J. Dairy Sci.* 94, 431–441. doi: 10.3168/jds.2009-2840

R Core Team (2017). *R: A language and environment for statistical computing* (Vienna, Austria: R Foundation for Statistical Computing).

Salomé, P. A., Bomblies, K., Laitinen, R. A. E., Yant, L., Mott, R., and Weigel, D. (2011). Genetic architecture of flowering-time variation in arabidopsis thaliana. *Genetics* 188, 421–433. doi: 10.1534/genetics.111.126607

Tsuruta, S., Misztal, I., and Strandén, I. (2001). Use of preconditioned conjugate gradient algorithm as a generic solver for mixed-model equations in animal breeding applications. *J. Anim. Sci.* 79, 1166–1172. doi: 10.2527/2001.7951166x

Vandenplas, J., Calus, M. P. L., Eding, H., and Vuik, C. (2019). A second-level diagonal preconditioner for single-step SNPBLUP. *Genet. Selection Evol.* 51, 30. doi: 10.1186/s12711-019-0472-8

Vandenplas, J., Eding, H., Bosmans, M., and Calus, M. P. L. (2020). Computational strategies for the preconditioned conjugate gradient method applied to ssSNPBLUP, with an application to a multivariate maternal model. *Genet. Selection Evol.* 52, 24. doi: 10.1186/s12711-020-00543-9

Vandenplas, J., Eding, H., Calus, M. P. L., and Vuik, C. (2018). Deflated preconditioned conjugate gradient method for solving single-step BLUP models efficiently. *Genet. Selection Evol.* 50, 51. doi: 10.1186/s12711-018-0429-3

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

Weller, J. I., Ezra, E., and Ron, M. (2017). Invited review: A perspective on the future of genomic selection in dairy cattle. *J. Dairy Sci.* 100, 8633–8644. doi: 10.3168/jds.2017-12879

Winkelman, A. M., Johnson, D. L., and Harris, B. L. (2015). Application of genomic evaluation to dairy cattle in new Zealand. *J. Dairy Sci.* 98, 659–675. doi: 10.3168/jds.2014-8560

Wu, Y., Burch, K. S., Ganna, A., Pajukanta, P., Pasaniuc, B., and Sankararaman, S. (2022). Fast estimation of genetic correlation for biobank-scale data. *Am. J. Hum. Genet.* 109, 24–32. doi: 10.1016/j.ajhg.2021.11.015

Wu, Y., and Sankararaman, S. (2018). A scalable estimator of SNP heritability for biobank-scale data. *Bioinformatics* 34, i187–i194. doi: 10.1093/bioinformatics/bty253

Xu, S., Zhu, D., and Zhang, Q. (2014). Predicting hybrid performance in rice using genomic best linear unbiased prediction. *Proc. Natl. Acad. Sci. U.S.A.* 111, 12456–12461. doi: 10.1073/pnas.1413750111

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 09 January 2023
DOI 10.3389/fpls.2022.1102938

OPEN ACCESS

EDITED BY
Yan Zhao,
Shandong Agricultural University, China

REVIEWED BY
Hailan Liu,
Maize Research Institute of Sichuan Agricultural University, China
Chengjun Zhang,
Kunming Institute of Botany (CAS), China

*CORRESPONDENCE
Xianjin Qiu
xjqiu216@yangtzeu.edu.cn
Fan Zhang
zhangfan03@caas.cn
Jianlong Xu
xujianlong@caas.cn

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

RECEIVED 19 November 2022
ACCEPTED 19 December 2022
PUBLISHED 09 January 2023

CITATION
Li P, Jiang J, Zhang G, Miao S, Lu J, Qian Y, Zhao X, Wang W, Qiu X, Zhang F and Xu J (2023) Integrating GWAS and transcriptomics to identify candidate genes conferring heat tolerance in rice.
Front. Plant Sci. 13:1102938.
doi: 10.3389/fpls.2022.1102938

Integrating GWAS and transcriptomics to identify candidate genes conferring heat tolerance in rice

Pingping Li[1], Jing Jiang[2], Guogen Zhang[3], Siyu Miao[2], Jingbing Lu[2], Yukang Qian[2], Xiuqin Zhao[2], Wensheng Wang[2,3], Xianjin Qiu[1]*, Fan Zhang[2,3]* and Jianlong Xu[2,4]*

[1]Ministry of Agriculture and Rural Affairs (MARA) Key Laboratory of Sustainable Crop Production in the Middle Reaches of the Yangtze River (Co-construction by Ministry and Province), College of Agriculture, Yangtze University, Jingzhou, China, [2]Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, China, [3]College of Agronomy, Anhui Agricultural University, Hefei, China, [4]Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China

Introduction: Rice (*Oryza sativa L.*) production is being challenged by global warming. Identifying new loci and favorable alleles associated with heat tolerance is crucial to developing rice heat-tolerant varieties.

Methods: We evaluated the heat tolerance at the seedling stage using 620 diverse rice accessions. A total of six loci associated with heat tolerance were identified by a genome-wide association study (GWAS) with ~2.8 million single nucleotide polymorphisms (SNPs).

Results: Among the six detected loci, *qHT7* harbored the strongest association signal and the most associated SNPs. By comparing the transcriptomes of two representative accessions with contrasting heat tolerance, *LOC_Os07g48710* (*OsVQ30*) was selected as a promising candidate gene in *qHT7* due to the significant difference in its expression level between the two accessions. Haplotype 4 (Hap4) of *LOC_Os07g48710* was determined as the favorable haplotype for heat tolerance via the gene-based haplotype analysis. The heat-tolerant haplotype *LOC_Os07g48710*Hap4 is highly enriched in the tropical *Geng/Japonica* accessions, and its frequency has decreased significantly during the improvement process of rice varieties.

Discussion: Based on the GWAS and transcriptomics integrated results, a hypothetical model modulated by *qHT7* in response to heat stress was proposed. Our results provide valuable candidate genes for improving rice heat tolerance through molecular breeding.

KEYWORDS

GWAS, transcriptome analysis, heat stress, candidate genes, germplasm resource, rice

Introduction

Rice (*Oryza sativa* L.) is one of the major food crops in the world. By 2050, global rice production will need to increase by 1.0%–1.2% annually to meet the growing food demand brought by population growth and economic development (Normile, 2008; Cramer et al., 2011). Unfortunately, heat stress has become a major limiting factor for rice growth and yield in recent years due to the rising global warming trend (Ahuja et al., 2010). Rice is sensitive to high temperature at almost all stages of growth and development. High-temperature stress can hasten the physiological maturity of rice, diminish assimilate accumulation, and cause permanent yield losses (Korres et al., 2017). Therefore, improving the heat tolerance of rice cultivars has become one of the major objectives of rice breeding worldwide.

High-temperature stress in rice induces an increase in reactive oxygen species (ROS), membrane damage, protein degradation, and a cascade of other heat stress reactions (Wahid et al., 2007; Bita and Gerats, 2013; Chen et al., 2021a). ROS can act as crucial signaling messengers in the early stages of heat stress. However, the ROS generated in the late stages of heat stress may cause damage to the cellular components of rice (Wahid et al., 2007; Zhang et al., 2019). For example, the fluidity of the plasma membrane increases during the early stages of heat stress, and cyclic nucleotide-gated channel proteins are responsible for signal transduction (Finka and Goloubinoff, 2014; Li et al., 2018). ROS, nitric oxide (NO), and Ca^{2+}, which are second messengers that can trigger the expression of downstream genes and ROS-scavenging genes, contribute greatly to heat tolerance by controlling ROS concentrations in rice (Liu et al., 2006; Liu et al., 2008; Mittler et al., 2012; Zhu, 2016; Liu et al., 2020b; Chen et al., 2021a). The NAC transcription factors have been identified as vital regulators of stress responses. Under heat stress, the membrane-associated NAC gene *OsNTL3*, which directly binds to the *OsbZIP74* promoter and regulates its expression, may influence the level of H_2O_2 and malondialdehyde (MDA) and electrolyte leakage (Liu et al., 2020a). The heat stress-sensitive rice mutant *hts1* showed increased H_2O_2 accumulation, Ca^{2+} influx, as well as membrane and chloroplast damage in response to heat stress. In *hst1* mutants, the transcriptional activity of *HsfA2s* and its downstream target genes are repressed due to the disruption of heat signal transduction (Chen et al., 2021a). In *Arabidopsis thaliana*, the VQ (FxxxVQxLTG) motif-containing proteins interacting with WRKY transcription factors (TFs) may improve heat tolerance by regulating ROS production (Cheng et al., 2012; Cheng et al., 2021). Similarly, the functional module of *WRKY10-VQ8* plays a role in regulating thermotolerance by modulating the ROS balance in rice (Chen et al., 2022b).

As a complex trait in rice, heat tolerance is controlled by multiple genes and genetic networks. To date, at least 58 quantitative trait loci (QTLs) responsible for heat tolerance at different developmental stages have been identified in rice (Xu et al., 2021). Moreover, more than 23 genes involved in heat tolerance have also been cloned and functionally verified (Huang et al., 2022), leading to a better understanding of the genetic mechanisms underlying heat tolerance. Several studies have demonstrated that plant cells rapidly accumulate misfolded toxins when subjected to severe heat stress (Liu et al., 2020a). The proteasome degrades these toxic proteins more efficiently than they are reactivated by heat shock proteins (Zhang et al., 2019). *TT1*, which encodes the α2 subunit of the 26S proteasome, protects rice against heat stress by eliminating cytotoxic denatured proteins and balancing the heat response process (Li et al., 2015). *TT2*, encoding a Gγ subunit, confers heat tolerance in rice and is associated with wax retention at high temperatures (Kan et al., 2022). A major QTL *TT3*, consisting of two genes named *TT3.1* and *TT3.2*, enhances rice thermotolerance by transducing heat signals from the plasma membrane to the chloroplasts (Zhang et al., 2022). The tRNA 2-thiolation process is a highly conserved form of tRNA modification among organisms. Compared with *Geng* (*japonica*) rice, *Xian* (*indica*) rice exhibits higher heat tolerance, possibly due to a higher level of tRNA thiolation controlled by *SLG1*, which encodes the cytoplasmic tRNA2-thiolated protein 2 (Xu et al., 2020). As a $tRNA^{HIS}$ guanylate transferase, *AET1* contributes to the modification of pre-$tRNA^{His}$ and possibly regulates auxin signaling in rice to enable normal growth under high-temperature conditions (Chen et al., 2019).

Genome-wide association studies (GWAS), a powerful approach for identifying genotype-phenotype associations in natural populations, have been applied to dissect the genetic architecture of many complex traits in rice. Over the past decade, the loci underlying tens of rice traits were identified by GWAS, and several important genes were successfully verified by further transgenic experiments (Wang et al., 2020; Chen et al., 2022a). For heat tolerance, Wei et al. identified 77 loci associated with survival rate after heat treatment at the seedling stage by GWAS based on a panel of 255 rice accessions and identified *LOC_Os02g12890* as an important candidate gene that may respond to high-temperature stress based on integrated transcriptome analysis (Wei et al., 2021). In addition, Yang et al. detected ten heat-associated QTL by GWAS with 221 rice accessions and selected 11 promising candidate genes by combining GWAS and transcriptome data (Yang et al., 2022). However, the genetic basis of heat tolerance in rice remains unclear due to the small size and limited diversity of the previous panels used for GWAS.

In this study, we conducted a GWAS on heat tolerance at the rice seedling stage using 620 diverse accessions and compared the transcriptomes between heat-tolerant and heat-sensitive representative accessions. One potential candidate gene was identified at the major locus *qHT7* on chromosome 7, and the

possible genetic pathways in response to heat stress were approached. This candidate gene could be employed for improving heat tolerance in future rice breeding. Our findings may also provide insight into the genetic mechanisms of heat stress response in rice.

Materials and methods

Plant materials and heat-stress treatment conditions

A panel of 620 rice accessions from the 3K Rice Genome Project (3K RG) (Wang et al., 2018) was used to evaluate heat tolerance at the seedling stage. The accessions contained 173 *Geng*, 411 *Xian*, 19 *admix*, 7 *Aus*, 9 *Basmati* and 1 unknown accessions (Supplementary Table S1). Twenty-four uniformly germinated seeds per replicate of each accession were sown in 96-well plates with holes at the bottom of each well. Then, the seeds were soaked in a container with tap water by placing the plates on scaffolds and were cultured in a phytotron at 28°C/25°C, 70% relative humidity and a 13-h light/11-h dark photoperiod. After 7 d, the seeds were transferred to Yoshida solution (pH 5.8-6.0), which was replaced every 3 d (Li et al., 2015). 13-day-old seedlings were exposed to 45°C for 3 d in a phytotron and then returned to normal conditions (28°C) for 7 d of recovery. The phytotron was set at 60% relative humidity and low light intensity (50–80 μM m^{-2} s^{-1}) to minimize the influence of high light and hydrophobic stress (Hasanuzzaman et al., 2013). Then, the survival rate (SR) was calculated as the proportion of surviving seedlings. Based on the evaluation system (Table 1 and Supplementary Figure S1), the leaf score of heat tolerance (SHT) was determined by visual inspection. At least three biological replicates were performed.

GWAS for heat tolerance

A total of 2,802,578 SNPs with a missing rate < 0.1 and minor allele frequency (MAF) ≥ 0.05 in the GWAS panel were filtered from the 3K-RG 4.8M SNP dataset (Alexandrov et al., 2014) by PLINK (Purcell et al., 2007). The GWAS based on a mixed linear model was performed with EMMAX (Kang et al., 2010) to identify the associations between SNPs and heat tolerance. The kinship matrix was calculated with an identical-by-state matrix using the pruned SNP subset (with the parameter "indep-pairwise 50 10 0.1" in PLINK) as a measure of relatedness between accessions. The eigenvectors of the kinship matrix were calculated using GCTA (with the parameter "-make-grm") (Yang et al., 2011) and the first three principal components were used as covariates to control population structure. The effective number of SNPs (N) was calculated by the GEC software (Li et al., 2012), and a suggestive significance threshold of association (P = 2.29E-06) was determined by the Bonferroni correction method (1/N) for claiming significant SNPs. Manhattan plots of the GWAS results were plotted by the R package "qqman" (Turner, 2014). The significant SNPs within the 300-kb region were considered as a locus based on the previously reported linkage disequilibrium (LD) decay in 3K RG (Wang et al., 2018). The leading SNP within a locus was defined as the SNP with the lowest P value. Local LD block analysis was performed within 150 kb upstream and downstream of the leading SNP using the LDBlockShow (Dong et al., 2021).

TABLE 1 The scale for leaf score of heat tolerance.

Score	Observation
1	The tip of the leaf is less than 1 cm
3	Tip drying extended up to ⅓ length in most leaves
5	⅓-⅔ of all leaves dried
7	More than ⅔ of all leaves fully dried
9	All seedlings apparently dead

Haplotype analysis for candidate genes

The haplotype analysis was performed on each annotated gene in *qHT7* to identify candidate genes and unearth favorable haplotypes. The gene haplotypes were constructed with all SNPs in the coding sequence (CDS) and 1-kb promoter regions, respectively. The synonymous SNPs were merged into one haplotype following the method by Zhang et al. (Zhang et al., 2021). Duncan's multiple range *post-hoc* tests were used to compare phenotypic differences between haplotypes ($n \geq 40$ rice accessions). The module of Custom Genotyping and Comment (Rice) in MBKbase database (http://www.mbkbase.org/rice/customGT) (Peng et al., 2019) was used to construct the candidate gene's variety groups based on the SNP genotype of the published wild rice accessions and 3K RG with parameters "Sample Num: ≥ 40, ALT ≥ 5%, Missing ≤ 20%". The haplotype network of a candidate gene was drawn by the minimum-spanning tree in Popart (Leigh and Bryant, 2015).

Transcriptome analysis

One representative heat-tolerant *Xian* accession, FACAGRO 64 (F64), and a representative heat-sensitive *Xian* accession, PUILLIPINA KATARI (PK), were selected for transcriptome analysis. Shoot samples before and after 24 h of heat-stress treatment were collected and stored in liquid nitrogen, each with three biological replicates. Total RNA was

extracted from shoot samples using the TRIzol reagent (Invitrogen) and then treated with RNase-free DNase I (Takara) to remove genomic DNA. Sequencing libraries were constructed according to the standard protocols provided by Illumina. The libraries were sequenced using Illumina NovaSeq 6000platform (150-bp paired ends) in Novogene (China). The raw sequence data reported have been deposited in the Genome Sequence Archive (Chen et al., 2021b) in National Genomics Data Center (CNCB-NGDC Members and Partners, 2022) with accession number CRA008760 that are publicly accessible at https://ngdc.cncb.ac.cn/gsa.

After removing adaptor and low-quality reads, clean reads were aligned to the Nipponbare reference genome (MSU v7.0) using HISAT2 (Kim et al., 2015). The gene expression levels based on fragments per kilobase of exon per million mapped fragments (FPKM) were calculated by StringTie (Pertea et al., 2015). Differentially expressed genes (DEGs) between two samples were identified with the DESeq2 package (Love et al., 2014) in R. The threshold for claiming DEGs was set as adjusted *P*-value (FDR) ≤ 0.05 and log fold change (FC) absolute value ≥ 1. Functional enrichment analysis of Gene Ontology (GO) and KEGG pathway was performed by clusterProfiler software (Yu et al., 2012). The threshold of adjusted *P*-value (FDR) < 0.05 was used to identify significantly enriched GO terms and KEGG pathways.

Quantitative real-time PCR

Total RNA (1 ug) was reverse-transcribed into cDNA using FastKing gDNA Dispelling RT SuperMix kit (Tiangen; KR118-02). qRT-PCR analyses were performed with SuperReal PreMix Plus (SYBR Green) kit (Tiangen; FP205-2), including three biological replicates. *UBQ* was used as the internal control, and the relative expression levels of the target genes were quantified using the comparative cycle threshold ($2^{-\Delta\Delta CT}$) method (Livak and Schmittgen, 2001). Primers used for qRT-PCR are listed in Supplementary Table S2.

Measurement of malondialdehyde content, ROS levels and enzyme activity

The shoots of F64 and PK under heat stress (45°C) for 72 h and control conditions were collected, respectively. The levels of superoxide dismutase (SOD), malondialdehyde (MDA), peroxidase (POD) and hydrogen peroxide (H_2O_2) in shoot tissue were determined using SOD, MDA, POD, H_2O_2 commercial kits following the manufacturer's instructions (Suzhou Grace Biotechnology Co., Ltd.). Three biological replicates were included.

Results

Phenotypic variation in heat tolerance

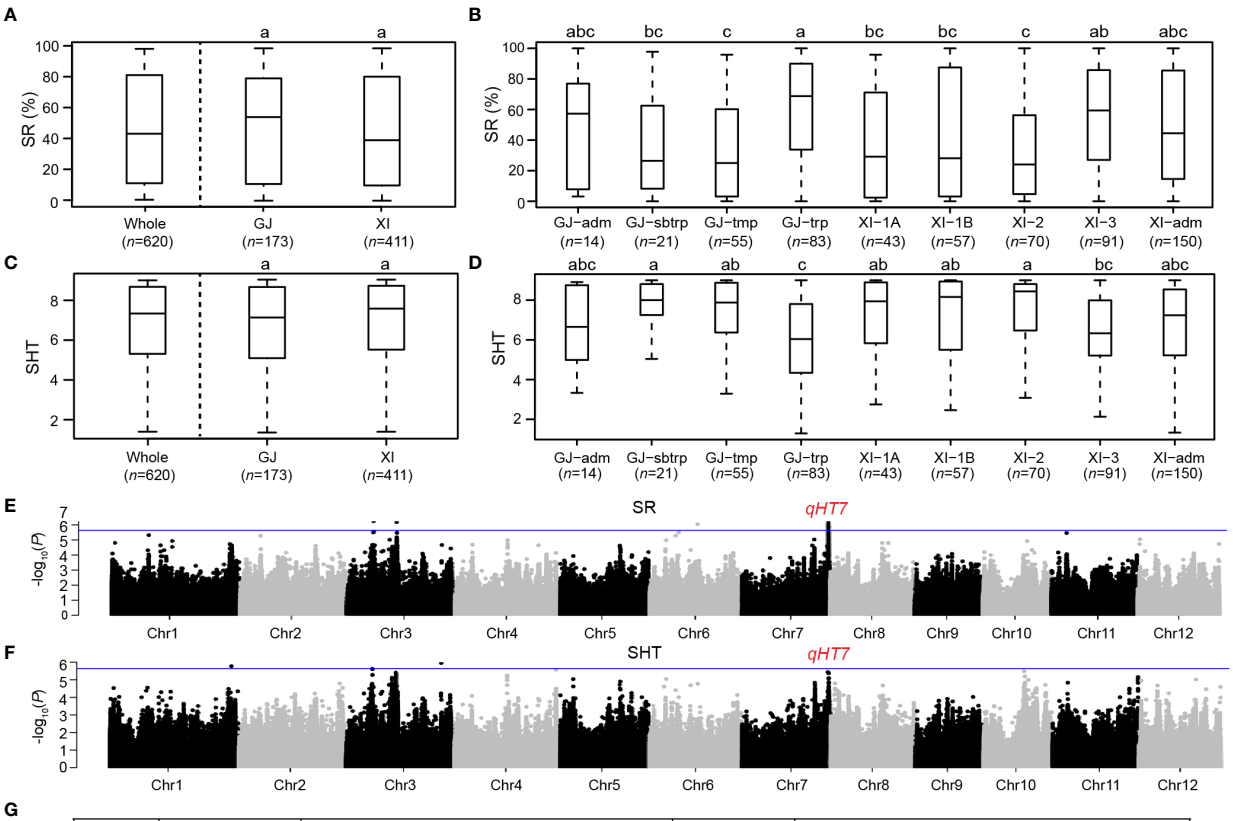
The phenotypic measurements of SR and SHT showed a considerable variation in heat tolerance at the seedling stage among the 620 rice accessions (Figure 1 and Supplementary Table S1). The mean SR in the whole population was 46.9%, ranging from 0 to 100.0%. Similarly, the mean SHT was 6.82, with a range of 1.29 to 9.00. Heat tolerance at the seedling stage did not vary significantly between *Xian* and *Geng* subpopulations (Figures 1A, C). Among the four *Geng* subgroups, most accessions with high heat tolerance belonged to *GJ-trp* (Figures 1B, D). Within *Xian* subpopulation, the average heat tolerance of *XI-3* accessions was higher than those of other *Xian* subgroups. The heat tolerance of the *GJ-trp* accessions was similar to that of the *XI-3* subgroups. Moreover, most *GJ-trp* and *XI-3* accessions are both from Southeast Asia islands.

GWAS for heat tolerance

Thirty-one SNPs significantly associated with heat tolerance were identified in the 620 accessions, including 29 and 2 SNPs associated with SR and SHT, respectively (Figures 1E, F and Supplementary Table S3). Among the significant SNPs, 6, 3 and 22 were located in the promoter, CDS and intergenic regions of 17 annotated genes, respectively. Based on the local LD block analysis, we combined these significant SNPs into six loci distributed on rice chromosomes 1, 3, 6, and 7 (Figure 1G). By comparing the 58 previously reported QTLs for heat tolerance (Xu et al., 2021) and 23 known genes involved in heat tolerance (Huang et al., 2022), three genes/QTLs, *qHTB1* (Zhu et al., 2017), *qHTB3-3* (Jagadish et al., 2010; Zhu et al., 2017; Kilasi et al., 2018), and *TT2* (Kan et al., 2022) were also found in the region of *qHT1*, *qHT3.2*, and *qHT3.3*, respectively. Out of the six loci, *qHT7* (Chr7: 29067638-29223510 bp) was determined as the major locus since it contained the most and strongest association signals (Figure 1G).

Physiological comparison of two rice accessions with different levels of heat tolerance in response to high temperature

F64 was highly tolerant to heat stress, with an average 95.8% ± 7.2% SR, which only slightly dried at the tips after 7 d of recovery from heat stress (Figures 2A, B). PK was extremely sensitive to heat stress, with an average SR of 0% ± 0%, and all of the seedlings were apparently dead. To examine the cell



G

Locus	Trait	Chr.	Position range (bp)	NO. of significant SNPs	Lead SNP position	*P* value	Known related genes
qHT1	SHT	1	40765723-40831675	1	40816016	1.69E-06	*qHTB1* (Zhu *et al.*, 2017)
qHT3.1	SR	3	8838773-9031506	1	8941759	5.70E-07	
qHT3.2	SHT	3	31992761-32174819	1	32025558	1.11E-06	*qHTB3-3* (Jagadish *et al.*, 2010; Kilasi *et al.*, 2018; Zhu *et al.*, 2017)
qHT3.3	SR	3	16713954-16735605	1	16733441	6.44E-07	*TT2* (Kan *et al.*, 2022)
qHT6	SR	6	16222807-16422807	1	16283219	9.03E-07	
qHT7	SR	7	29067638-29223510	26	29140282	1.16E-07	

FIGURE 1
Phenotypic variations of heat tolerance and identification of loci associated with heat tolerance by GWAS in 620 rice accessions. **(A)** Box-plots of survival rate (SR) for the whole population, *Xian/Indica* (*XI*), and *Geng/Japonica* (*GJ*) subpopulations. **(B)** Box-plots of SR among *GJ-adm*, *GJ-subtropical* (*GJ-sbtrp*), *GJ-temperate* (*GJ-tmp*), *GJ-tropical* (*GJ-trp*), *XI-1A*, *XI-1B*, *XI-2*, *XI-3*, and *XI-adm* accessions. **(C)** Box-plots of leaf score of heat tolerance (SHT) for the whole population, *XI* and *GJ* subpopulations. **(D)** Box-plots of SHT among *GJ-adm*, *GJ-sbtrp*, *GJ-tmp*, *GJ-trp*, *XI-1A*, *XI-1B*, *XI-2*, *XI-3*, and *XI-adm* accessions. **(E)** Manhattan plots of GWAS results for SR. **(F)** Manhattan plots of GWAS results for SHT. **(G)** Genome-wide significant loci for SR and SHT. In A-D, different letters indicate significant differences ($P < 0.05$, Duncan's multiple range *post-hoc* test). In E, F, the horizontal blue lines represent the suggestive significant threshold ($P = 2.29E-6$).

membrane damage and redox homeostasis caused by heat stress in the two rice accessions, we compared the physiological traits between F64 and PK under heat stress for 72 h (Figures 2C–F). Although there was no statistically significant difference in the relative MDA content between the two accessions (Figure 2F), the relative H_2O_2 content of F64 after 72 h of heat stress was significantly lower than that of PK (Figure 2E). Moreover, the relative activity of the antioxidant enzymes SOD and POD were significantly higher in F64 than in PK (Figures 2C, D). These results suggested that F64 suffered less damage to cell membranes under heat stress than PK, possibly due to more effective active detoxification by ROS scavenging regulation in F64.

Comparative transcriptome profiling between two rice accessions differing in their heat tolerance

In order to reveal the differences in transcriptome response to heat stress at the seedling stage between rice accessions with different levels of heat tolerance, we compared F64 (a

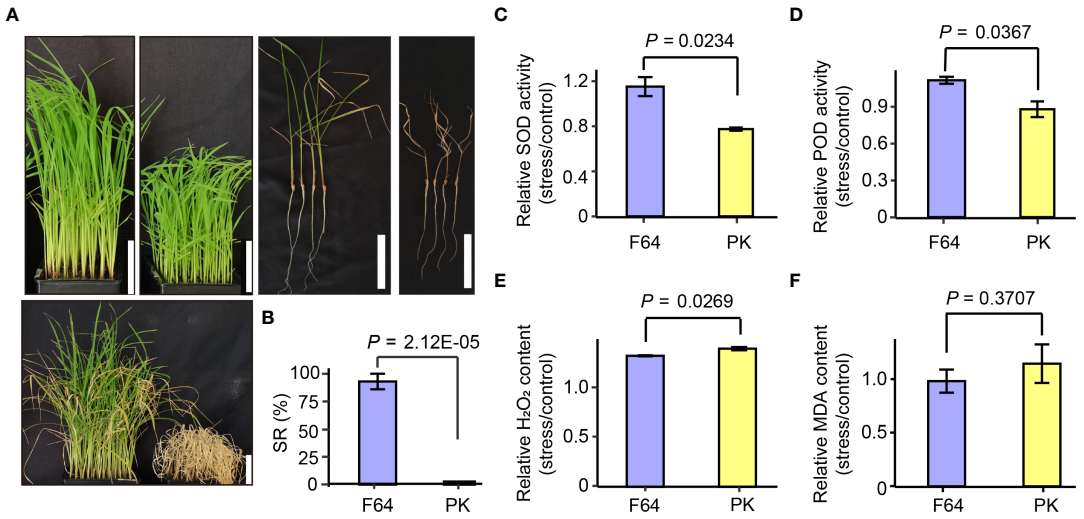


FIGURE 2
The differences in morphological and physiological performances between two rice accessions differing in their tolerance to heat stress. **(A)** Growth images of the heat-tolerant accession FACAGRO 64 (F64) and the heat-sensitive accession PUILLIPINA KATARI (PK) before and after heat stress treatment. Scale bars = 5 cm. **(B)** SR of the two rice accessions in 7 d after heat stress. Data are extracted from Supplementary Table S1. **(C–F)** The relative SOD, POD, H_2O_2 and MDA content between heat stress for 72 h and control conditions. Data shown in the form mean ± standard deviation of three biological replicates. The significant difference between the two groups was calculated using two-tailed Student's *t*-test.

representative heat-tolerant accession) with PK (a representative heat-sensitive accession) using RNA-seq analysis (Supplementary Table S4). A total of 2056, 8303, 4070 and 8717 DEGs were identified for G1 (F64 vs PK under control conditions), G2 (heat stress vs control in F64), G3 (F64 vs PK under heat stress) and G4 (heat stress vs control in PK), respectively. Among them, 1202, 4287, 2311 and 4365 DEGs were upregulated, and 854, 4016, 1759, and 4352 DEGs were down-regulated in G1, G2, G3 and G4, respectively (Figures 3A, B). A series of biological processes and pathways involved in response to heat stress were commonly identified in both heat-tolerant and heat-sensitive accessions. Specifically, gene ontology (GO) analysis for the G2 and G4 DEGs, which were significantly regulated by high temperature in heat-tolerant and heat-sensitive accessions, respectively, showed that the common biological processes were mainly upregulated in protein folding (GO: 0006457) and RNA processing (GO: 0006396) (Figure 3C), and were primarily downregulated in carbohydrate metabolic process (GO: 0005975), biosynthetic process (GO: 0009058), metal ion transport (GO: 0030001), and glycolytic process (GO: 0006096) (Figure 3E). Similarly, six KEGG pathways, including spliceosome (map03040), protein processing in endoplasmic reticulum (map04141), RNA degradation (map03018), RNA transport (map03013), ribosome biogenesis in eukaryotes (map03008), and valine, leucine and isoleucine degradation (map00280), were significantly enriched both in the G2 and G4 upregulated DEGs (Figure 3D). For the G2 and G4 downregulated DEGs, 12 common KEGG pathways were significantly enriched, such as carbon metabolism (map01200), biosynthesis of amino acids (map01230), etc (Figure 3F). The results suggest that the aforementioned biological processes and pathways mentioned above should be the components of regulatory mechanisms underlying heat tolerance in rice.

Moreover, several unique GO terms and KEGG pathways were identified in G2 DEGs compared to the G4 DEGs. Four specific GO terms and one KEGG pathway were significantly enriched in G2 upregulated DEGs compared to G4 upregulated DEGs, including cell redox homeostasis (GO: 0045454), ribosome biogenesis (GO: 0042254), carbohydrate metabolic process (GO: 0005975), protein metabolic process (GO: 0019538), and RNA polymerase (map03020). In contrast, the divergence between G2 and G4 downregulated DEGs was much greater according to the number of unique GO terms and KEGG pathways (Figures 3E, F). Interestingly, cell redox homeostasis (GO: 0045454) was specifically enriched in G4 downregulated DEGs compared to G2 downregulated DEGs.

Genes involved in cell redox homeostasis are usually triggered in plants tolerant to abiotic stresses (Awasthi et al., 2015). Given that the genes related to cell redox homeostasis exhibited contrasting responses to heat stress in the two rice accessions with different heat tolerance, we further compared the expression profiles of 62 DEGs related to cell redox homeostasis between F64 and PK under control and heat stress conditions (Supplementary Figure S2). There were 34 (55%) common, 12 (19%) F64-specific and 10 (16%) PK-specific DEGs regulated by heat stress in the two accessions. The results suggest that cell redox homeostasis should play an important role in rice heat tolerance.

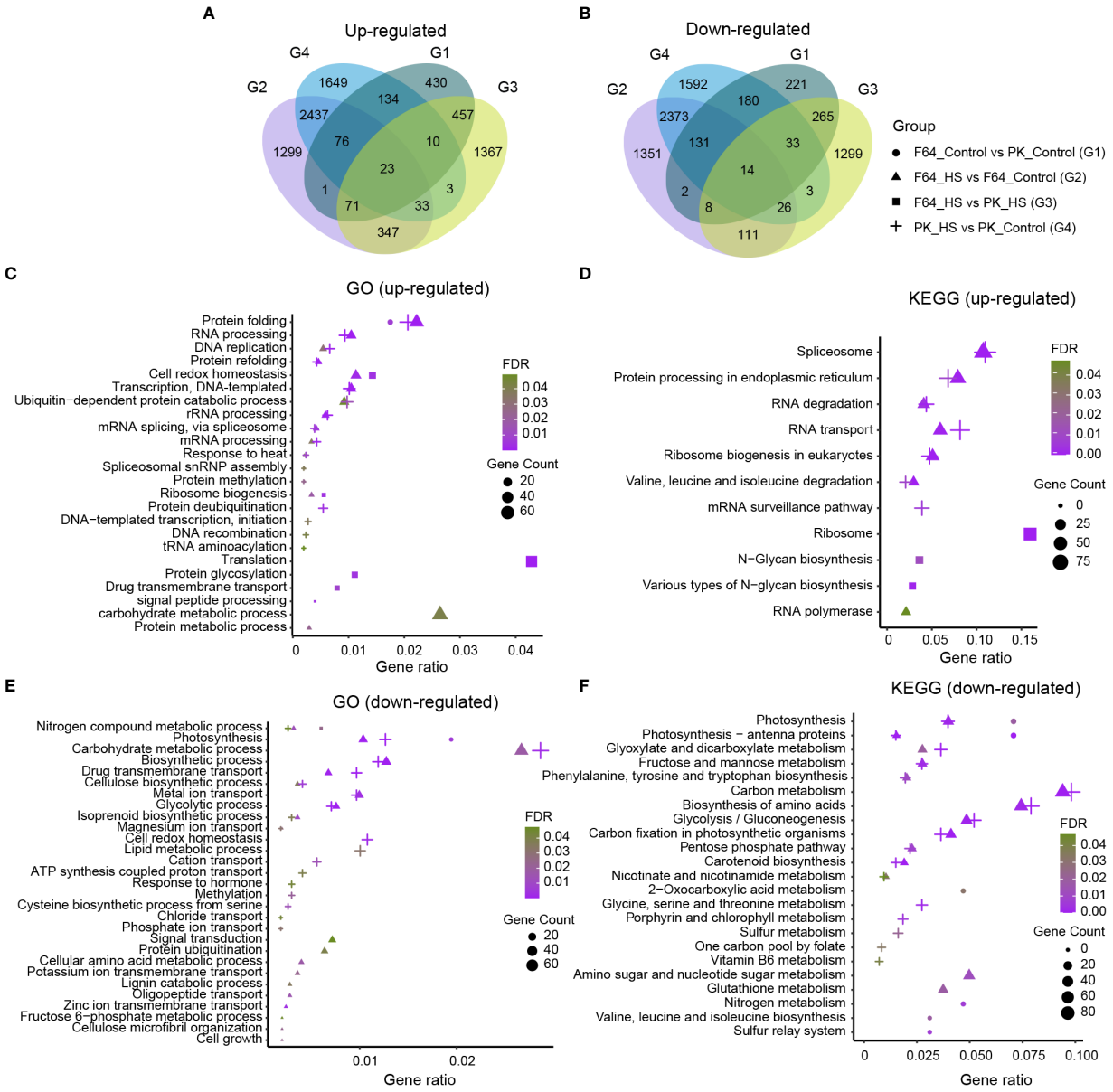


FIGURE 3
The transcriptome analysis of two rice accessions differing in their tolerance to heat stress (HS). **(A)** Venn diagrams showed the up-regulated differentially expressed genes (DEGs). **(B)** Venn diagrams showed the down-regulated DEGs. **(C, D)** GO and KEGG enrichment analysis of up-regulated DEGs. **(E, F)** GO and KEGG enrichment analysis of down-regulated DEGs. In c, e, only biological process GO terms were shown. G1: F64_Control vs PK_Control; G2: F64_HS vs F64_Control; G3: F64_HS vs PK_HS; G4: PK_HS vs PK_Control.

Integrating GWAS and RNA-seq to identify candidate genes for *qHT7*

Based on the Nipponbare reference genome IRGSP 1.0, 28 genes were annotated in *qHT7* (Figure 4A and Supplementary Table S5). Candidate genes for heat tolerance were selected based on the following criteria: (1) functionally related to abiotic stresses based on the annotation of Nipponbare reference genome, GO annotation, and literature search; (2) significant differences in heat tolerance among gene haplotypes. Consequently, 14 candidate genes were identified (Figure 4B and Supplementary Table S5). To further screen the promising candidate genes, we examined the expression profiles of the 14 candidate genes using the transcriptomic datasets of F64 and PK. As a result, five DEGs (*LOC_Os07g48830*, *LOC_Os07g48630*, *LOC_Os07g48710*, *LOC_Os07g48570*, and *LOC_Os07g48760*) were selected (Figure 4B). We also verified the expression of the five genes by qRT-PCR (Figure 4E and Supplementary

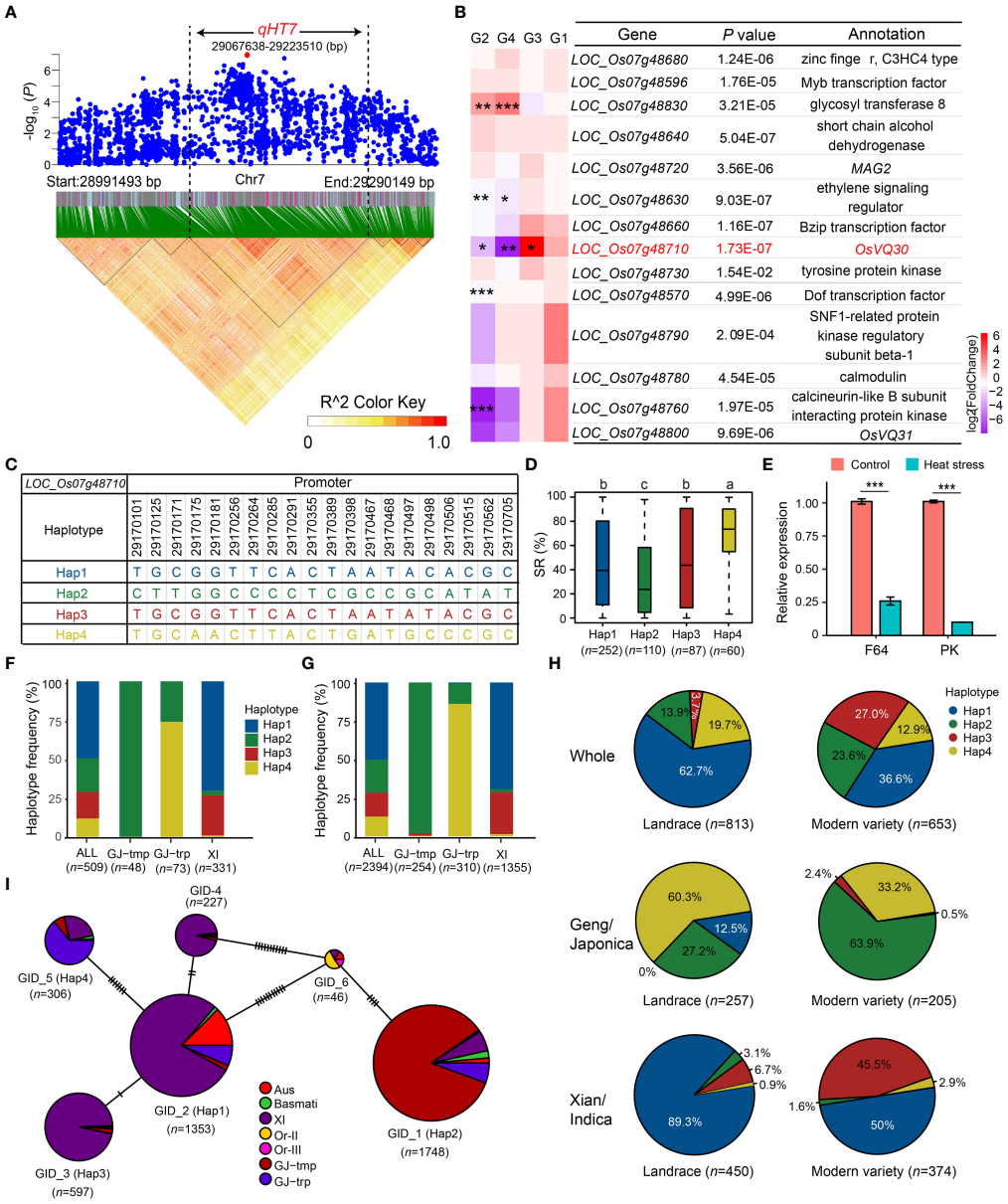


FIGURE 4
Candidate gene analysis of *qHT7*. **(A)** Local Manhattan plot (top) and LD analysis (bottom) of 150-kb upstream and downstream around the lead SNP rs7_29140282. The red dot is the lead SNP, and its LD block region is marked by the black dotted lines. **(B)** Relative expression of 14 annotated genes in *qHT7*. *** FDR < 0.001, ** FDR < 0.01, * FDR < 0.05. G1: F64_Control vs PK_Control; G2: F64_HS vs F64_Control; G3: F64_HS vs PK_HS; G4: PK_HS vs PK_Control. **(C)** Haplotype of *LOC_Os07g48710*, which is the promising candidate gene of *qHT7*. **(D)** The distribution of SR in the whole population for the four major haplotypes (n > 40 accessions) of *LOC_Os07g48710*. Different letters above each boxplot indicate significant differences among haplotypes (P < 0.05, Duncan's multiple range *post-hoc* test). **(E)** Verification of the relative expression of *LOC_Os07g48710* in F64 and PK under heat stress 24 h by qRT-PCR. *UBQ* was used as an internal control. The figure presents the relative expression levels of *LOC_Os07g48710* relative to that under control conditions in each accession. Bars represent standard deviation of three biological replicates. ***P < 0.001 (two-tailed Student's t-test). **(F, G)** Frequency of the four major haplotypes of *LOC_Os07g48710* in the GWAS panel **(F)** and in 3K RG **(G)**. **(H)** Haplotype frequency distribution of *LOC_Os07g48710* in landrace and modern variety of 3K RG. The type of each accession was from the metadata of 3K RG (Wang et al., 2018). **(I)** Haplotype network of *LOC_Os07g48710* retrieved from MBKbase (Peng et al., 2019) (http://www.mbkbase.org/rice/, query date: October 25th, 2022). Circle size of a given haplotype is proportioned to its number of accessions. Letter n indicates the number of rice accessions belonging to the corresponding haplotype in **D** and **I**, subpopulation in **F** and **G**, or variety type in **H**, respectively.

Figure S3). Among the five genes, only the expression level of *LOC_Os07g48710* was significantly higher in the heat-tolerant accession F64 than that in the heat-sensitive accession PK under heat stress (Figure 4E), which was consistent between the RNA-seq and qRT-PCR results. Thus, *LOC_Os07g48710*, encoding a VQ domain-containing protein, was determined as a promising candidate gene for the further analysis.

Mining heat-tolerant allele is helpful in improving the heat tolerance of rice through molecular breeding. To examine the favorable haplotype of the promising candidate gene of *qHT7*, *LOC_Os07g48710*, we performed the haplotype analysis using CDS and 1-kb promoter SNPs in 3K RG. Due to no SNPs detected in the CDS of *LOC_Os07g48710*, we identified four major haplotypes ($n \geq 40$ accessions) using 20 SNPs (MAF ≥ 0.05 and heterozygous rate < 0.05) in its 1-kb promoter region in the GWAS panel (Figure 4C). Among the four haplotypes, Hap4 with significantly higher SR was determined as the favorable haplotype (Figure 4D), which was significantly enriched (P = 2.63E-54) in *GJ-trp* accessions of the GWAS panel (Figure 4F). For the 3K RG, the heat-tolerant haplotype *LOC_Os07g48710*^{Hap4} was also highly enriched in the *GJ-trp* accessions (P = 1.60E-287) (Figure 4G). In contrast, *LOC_Os07g48710*^{Hap4} was virtually absent in *GJ-tmp* subpopulation and *Xian* subpopulation (Figures 4F, G). To explore the origin and spread of Hap4, the haplotype network of *LOC_Os07g48710* was analyzed, showing that Hap4 possibly evolved from Hap1 (Figure 4I). Furthermore, the proportion of *Geng* accessions with *LOC_Os07g48710*^{Hap4} dropped dramatically from 60.3% in landrace to 33.2% in modern variety (Figure 4H).

Discussion

Understanding the genetic mechanisms underlying heat tolerance is vital to developing heat-tolerant rice varieties to adapt to global warming. In this study, different rice subpopulations exhibited different responses to heat stress at the seedling stage. Most *GJ-trp* and *XI-3* accessions, mainly from Southeast Asia islands, showed more tolerant to heat stress than other accessions, suggesting that the high temperature of the tropical environment may be the driving force in the evolution and breeding selection of heat tolerance in rice.

Heat tolerance is a quantitative trait controlled by a complex genetic network in rice. Fortunately, integrating GWAS and transcriptome analysis is now available as a powerful method for identifying candidate genes associated with complex traits. In this study, six loci associated with heat tolerance at the seedling stage were identified by GWAS. By comparing the previously reported cloned genes for heat tolerance with the GWAS results, *TT2*, a well-known heat-tolerant QTL (Kan et al., 2022), was co-localized with *qHT3.3*. Loss-of-function *TT2* allele has been found to exhibit increased thermotolerance and wax retention at high temperatures. In addition, we identified two loci, *qHT1* and *qHT3.2*, which were co-localized with previously reported QTL for heat tolerance, *qHTB1* and *qHTB3-3*, respectively (Jagadish et al., 2010; Zhu et al., 2017; Kilasi et al., 2018).

Notably, a novel major locus *qHT7* (Chr7: 29067638-29223510 bp) associated with heat tolerance at rice seedling stage was identified, and a promising candidate gene (*LOC_Os07g48710*) was predicted. The coding sequence of *LOC_Os07g48710* is highly conserved in the 3K RG with only one major gene-CDS-haplotype (Zhang et al., 2021). In contrast, at least four major haplotypes based on the natural variations in the promoter region exist in rice germplasm (Figure 4C). Moreover, although the expression levels of *LOC_Os07g48710* were both inhibited in heat-tolerant accession F64 (with the favorable haplotype *LOC_Os07g48710*^{Hap4}) and heat-sensitive accession PK (with the non-favorable haplotype *LOC_Os07g48710*^{Hap2}) under heat stress, the expression level of *LOC_Os07g48710* was significantly higher in F64 than in PK, implying natural variations in its promoter region are likely to be causal SNPs responsible for heat tolerance. The heat-tolerant haplotype *LOC_Os07g48710*^{Hap4} is subpopulation-specific, which is preferentially carried by *GJ-trp* accessions rather than *GJ-tmp* and *Xian* accessions (Figures 4F, G), suggesting that *qHT7* may partially explain the phenotypic variation of heat tolerance in rice germplasm. Thus, the favorable haplotype, *LOC_Os07g48710*^{Hap4}, may serve as a potential alternative for improving the heat tolerance of rice varieties by gene editing or marker-assisted selection. Further experiments should be conducted to validate the function of *LOC_Os07g48710* on heat tolerance and evaluate the breeding value of its favorable haplotype in developing new rice varieties with enhanced tolerance to heat stress.

The VQ proteins are plant-specific transcriptional regulatory factors that can fine-tune the regulatory pathway in response to abiotic stresses *via* interacting with TFs (Kim et al., 2013; Jing and Lin, 2015). The rice genome contains at least 39 VQ genes (numbered *OsVQ1* to *OsVQ39*), in which *LOC_Os07g48710* (*OsVQ30*) can be induced by drought stress rather than ABA treatment (Kim et al., 2013). Different VQ proteins can bind to the WRKY DNA-binding domain to modulate the expression of downstream genes and phytohormone signaling pathways in response to high-temperature stress (Cheng et al., 2012; Wang et al., 2015; Zhou et al., 2016; Jiang et al., 2017; Jiang et al., 2018; Cheng et al., 2021; Chen et al., 2022b). Cheng et al. (Cheng et al., 2021) have reviewed the WRKY-VQ protein interaction regulatory mechanism that regulates plant growth under high-temperature stress. For example, *WRKY39* activates SA- and JA-activated signaling pathways that promote the response to heat stress (Li et al., 2010). The functional module of WRKY10-VQ8 regulates heat tolerance by modulating the ROS balance in rice (Chen et al., 2022b). In this study, we identified 22 G3-DEGs with a similar expression pattern as *LOC_Os07g48710*, including a *WRKY* gene (*OsWRKY36*), 12 genes related to hormone biosynthesis/signaling, and nine genes related to cell redox homeostasis (Supplementary Table S6). Based on these genes

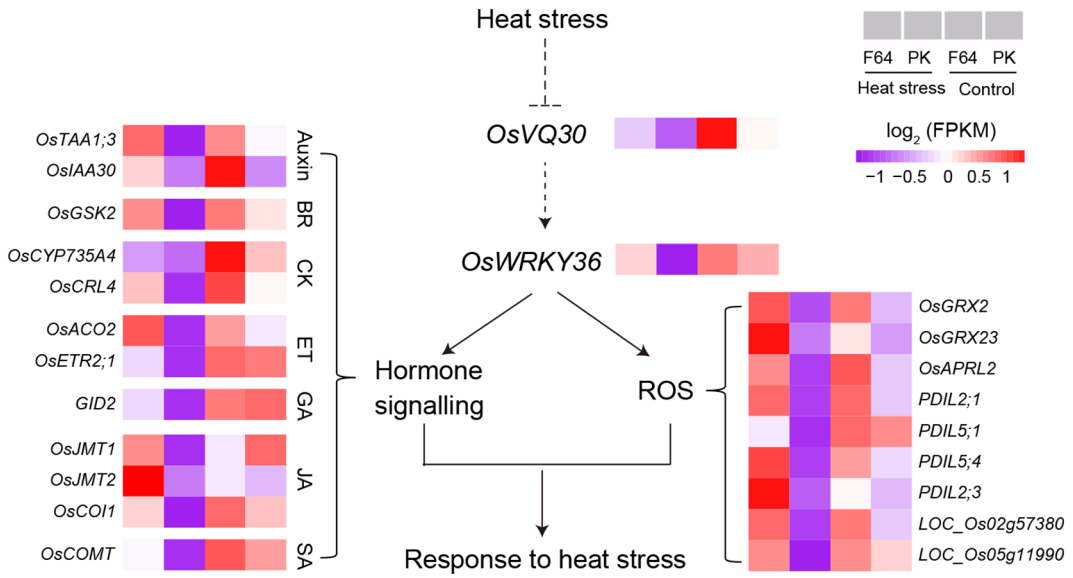


FIGURE 5
Hypothetical model of *qHT7* responses to high-temperature stress in rice. The model is proposed based on the genome and transcriptome analysis in this study. A complete list of genes is shown in Supplementary Table S6.

that are likely connected to *LOC_Os07g48710* (*OsVQ30*), we hypothesize the putative regulatory model mediated by *qHT7* in response to heat stress in rice (Figure 5). In this model, heat stress strongly inhibits the expression of *OsVQ30* and *OsWRKY36*, and the WRKY-VQ module may regulate their target gene expression to respond to high-temperature stress in rice. Further studies are required to verify the hypothesis.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://ngdc.cncb.ac.cn/gsa , CRA008760.

Author contributions

JX, FZ and XQ designed the experiment; PL, JJ, GZ, SM, JL and YQ performed all the phenotypic evaluation; PL, FZ, XZ and WW performed analysis and interpretation of the data; PL and FZ drafted the manuscript; FZ and JX revised the MS. All authors contributed to the article and approved the submitted version.

Funding

This work was funded by the National Natural Science Foundation of China (U21A20214), the Key Research and Development Project of Hainan Province (ZDYF2021XDNY128), and the Agricultural Science and Technology Innovation Program and the Cooperation and Innovation Mission (CAAS-ZDXT202001).

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.2022.1102938/full#supplementary-material

References

Ahuja, I., de Vos, R. C., Bones, A. M., and Hall, R. D. (2010). Plant molecular stress responses face climate change. *Trends Plant Sci.* 15 (12), 664–674. doi: 10.1016/j.tplants.2010.08.002

Alexandrov, N., Tai, S., Wang, W., Mansueto, L., Palis, K., Fuentes, R. R., et al. (2014). SNP-seek database of SNPs derived from 3000 rice genomes. *Nucleic Acids Res.* 43 (D1), D1023–D1027. doi: 10.1093/nar/gku1039%JNucleicAcidsResearch

Awasthi, R., Bhandari, K., and Nayyar, H. (2015). Temperature stress and redox homeostasis in agricultural crops. *Front. Environ. Sci.* 3. doi: 10.3389/fenvs.2015.00011

Bita, C. E., and Gerats, T. (2013). Plant tolerance to high temperature in a changing environment: scientific fundamentals and production of heat stress-tolerant crops. *Front. Environ. Sci.* 4. doi: 10.3389/fpls.2013.00273

Chen, S., Cao, H., Huang, B., Zheng, X., Liang, K., Wang, G.-L., et al. (2022b). The WRKY10-VQ8 module safely and effectively regulates rice thermotolerance. *Plant Cell Environ.* 45 (7), 2126–2144. doi: 10.1111/pce.14329

Chen, T., Chen, X., Zhang, S., Zhu, J., Tang, B., Wang, A., et al. (2021b). The genome sequence archive family: Toward explosive data growth and diverse data types. *Genomics Proteomics Bioinf.* 19, 578–583. doi: 10.1016/j.gpb.2021.08.001

Chen, R., Deng, Y., Ding, Y., Guo, J., Qiu, J., Wang, B., et al. (2022a). Rice functional genomics: decades' efforts and roads ahead. *Sci. China Life Sci.* 65 (1), 33–92. doi: 10.1007/s11427-021-2024-0

Chen, F., Dong, G., Wang, F., Shi, Y., Zhu, J., Zhang, Y., et al. (2021a). A β-ketoacyl carrier protein reductase confers heat tolerance *via* the regulation of fatty acid biosynthesis and stress signaling in rice. *New Phytol.* 232 (2), 655–672. doi: 10.1111/nph.17619

Cheng, Z., Luan, Y., Meng, J., Sun, J., Tao, J., and Zhao, D. (2021). WRKY transcription factor response to high-temperature stress. *Plants (Basel)* 10 (10), 2211. doi: 10.3390/plants10102211

Chen, K., Guo, T., Li, X. M., Zhang, Y. M., Yang, Y. B., Ye, W. W., et al. (2019). Translational regulation of plant response to high temperature by a dual-function tRNA(His) guanylyltransferase in rice. *Mol. Plant* 12 (8), 1123–1142. doi: 10.1016/j.molp.2019.04.012

Cheng, Y., Zhou, Y., Yang, Y., Chi, Y.-J., Zhou, J., Chen, J.-Y., et al. (2012). Structural and functional analysis of VQ motif-containing proteins in arabidopsis as interacting proteins of WRKY transcription factors. *Plant Physiol.* 159 (2), 810–825. doi: 10.1104/pp.112.196816%JPlantPhysiology

CNCB-NGDC Members and Partners. (2022). Database resources of the national genomics data center, China national center for bioinformation in 2022. *Nucleic Acids Res.* 50, D27–D38. doi: 10.1093/nar/gkab951

Cramer, G. R., Urano, K., Delrot, S., Pezzotti, M., and Shinozaki, K. (2011). Effects of abiotic stress on plants: a systems biology perspective. *BMC Plant Biol.* 11 (1), 163. doi: 10.1186/1471-2229-11-163

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), 1–6. doi: 10.1093/bib/bbaa227

Finka, A., and Goloubinoff, P. (2014). The CNGCb and CNGCd genes from physcomitrella patens moss encode for thermosensory calcium channels responding to fluidity changes in the plasma membrane. *Cell Stress Chaperones* 19 (1), 83–90. doi: 10.1007/s12192-013-0436-9

Hasanuzzaman, M., Nahar, K., Alam, M. M., Roychowdhury, R., and Fujita, M. (2013). Physiological, biochemical, and molecular mechanisms of heat stress tolerance in plants. *Int. J. Mol. Sci.* 14 (5), 9643–9684. doi: 10.3390/ijms14059643

Huang, F., Jiang, Y., Chen, T., Li, H., Fu, M., Wang, Y., et al. (2022). New data and new features of the FunRiceGenes (Functionally characterized rice genes) database: 2021 update. *Rice* 15 (1), 23. doi: 10.1186/s12284-022-00569-1

Jagadish, S. V. K., Cairns, J., Lafitte, R., Wheeler, T. R., Price, A. H., and Craufurd, P. Q. (2010). Genetic analysis of heat tolerance at anthesis in rice. *Crop Breed. Genet.* 50 (5), 1633–1641. doi: 10.2135/cropsci2009.09.0516

Jiang, J., Ma, S., Ye, N., Jiang, M., Cao, J., and Zhang, J. (2017). WRKY transcription factors in plant responses to stresses. *J. Integr. Plant Biol.* 59 (2), 86–101. doi: 10.1111/jipb.12513

Jiang, S. Y., Sevugan, M., and Ramachandran, S. (2018). Valine-glutamine (VQ) motif coding genes are ancient and non-plant-specific with comprehensive expression regulation by various biotic and abiotic stresses. *BMC Genomics* 19 (1), 342. doi: 10.1186/s12864-018-4733-7

Jing, Y., and Lin, R. (2015). The VQ motif-containing protein family of plant-specific transcriptional regulators. *Plant Physiol.* 169 (1), 371–378. doi: 10.1104/pp.15.00788%JPlantPhysiology

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

Kan, Y., Mu, X. R., Zhang, H., Gao, J., Shan, J. X., Ye, W. W., et al. (2022). TT2 controls rice thermotolerance through SCT1-dependent alteration of wax biosynthesis. *Nat. Plants* 8 (1), 53–67. doi: 10.1038/s41477-021-01039-0

Kilasi, N. L., Singh, J., Vallejos, C. E., Ye, C., Jagadish, S. V. K., Kusolwa, P., et al. (2018). Heat stress tolerance in rice (Oryza sativa l.): Identification of quantitative trait loci and candidate genes for seedling growth under heat stress. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01578

Kim, D. Y., Kwon, S. I., Choi, C., Lee, H., Ahn, I., Park, S. R., et al. (2013). Expression analysis of rice VQ genes in response to biotic and abiotic stresses. *Gene* 529 (2), 208–214. doi: 10.1016/j.gene.2013.08.023

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12 (4), 357–360. doi: 10.1038/nmeth.3317

Korres, N. E., Norsworthy, J. K., Burgos, N. R., and Oosterhuis, D. M. (2017). Temperature and drought impacts on rice production: An agronomic perspective regarding short- and long-term adaptation measures. *Water Resour. Rural Dev.* 9, 12–27. doi: 10.1016/j.wrr.2016.10.001

Leigh, J. W., and Bryant, D. (2015). Popart: full-feature software for haplotype network construction. *Br. Ecol. Soc.* 6 (9), 1110–1116. doi: 10.1111/2041-210X.12410

Li, X. M., Chao, D. Y., Wu, Y., Huang, X., Chen, K., Cui, L. G., et al. (2015). Natural alleles of a proteasome alpha2 subunit gene contribute to thermotolerance and adaptation of African rice. *Nat. Genet.* 47 (7), 827–833. doi: 10.1038/ng.3305

Li, B., Gao, K., Ren, H., and Tang, W. (2018). Molecular mechanisms governing plant responses to high temperatures. *J. Integr. Plant Biol.* 60 (9), 757–779. doi: 10.1111/jipb.12701

Liu, H. T., Gao, F., Cui, S. J., Han, J. L., Sun, D. Y., and Zhou, R. G. (2006). Primary evidence for involvement of IP3 in heat-shock signal transduction in arabidopsis. *Cell Res.* 16 (4), 394–400. doi: 10.1038/sj.cr.7310051

Liu, H. T., Gao, F., Li, G. L., Han, J. L., Liu, D. L., Sun, D. Y., et al. (2008). The calmodulin-binding protein kinase 3 is part of heat-shock signal transduction in arabidopsis thaliana. *Plant J.* 55 (5), 760–773. doi: 10.1111/j.1365-313X.2008.03544.x

Liu, Y., Liu, X., Wang, X., Gao, K., Qi, W., Ren, H., et al. (2020b). Heterologous expression of heat stress-responsive AtPLC9 confers heat tolerance in transgenic rice. *BMC Plant Biol.* 20 (1), 514. doi: 10.1186/s12870-020-02709-5

Liu, X. H., Lyu, Y. S., Yang, W., Yang, Z. T., Lu, S. J., and Liu, J. X. (2020a). A membrane-associated NAC transcription factor OsNTL3 is involved in thermotolerance in rice. *Plant Biotechnol. J.* 18 (5), 1317–1329. doi: 10.1111/pbi.13297

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 (4), 402–408. doi: 10.1006/meth.2001.1262

Li, M. X., Yeung, J. M., Cherny, S. S., and Sham, P. C. (2012). Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum. Genet.* 131 (5), 747–756. doi: 10.1007/s00439-011-1118-2

Li, S., Zhou, X., Chen, L., Huang, W., and Yu, D. (2010). Functional characterization of arabidopsis thaliana WRKY39 in heat stress. *Molecules Cells* 29 (5), 475–483. doi: 10.1007/s10059-010-0059-2

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 (12), 550. doi: 10.1186/s13059-014-0550-8

Mittler, R., Finka, A., and Goloubinoff, P. (2012). How do plants feel the heat? *Trends Biochem. Sci.* 37 (3), 118–125. doi: 10.1016/j.tibs.2011.11.007

Normile, D. (2008). Reinventing rice to feed the world. *Science* 321 (5887), 330–333. doi: 10.1126/science.321.5887.330

Peng, H., Wang, K., Chen, Z., Cao, Y., Gao, Q., Li, Y., et al. (2019). MBKbase for rice: an integrated omics knowledgebase for molecular breeding in rice. *Nucleic Acids Res.* 48 (D1), D1085–D1092. doi: 10.1093/nar/gkz921%JNucleicAcidsResearch

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 (3), 290–295. doi: 10.1038/nbt.3122

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., 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

Turner, S. D. (2014). Qqman: an r package for visualizing GWAS results using q-q and manhattan plots. *J. Open Source Software* 3 (25), 1731. doi: 10.1101/005165%JbioRxiv

Wahid, A., Gelani, S., Ashraf, M., and Foolad, M. R. (2007). Heat tolerance in plants: An overview. *Environ. Exp. Bot.* 61 (3), 199–223. doi: 10.1016/j.envexpbot.2007.05.011

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 (7703), 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

Wang, M., Vannozzi, A., Wang, G., Zhong, Y., Corso, M., Cavallini, E., et al. (2015). A comprehensive survey of the grapevine VQ gene family and its transcriptional correlation with WRKY proteins. *Front. Plant Sci.* 6. doi: 10.3389/fpls.2015.00417

Wei, Z., Yuan, Q., Lin, H., Li, X., Zhang, C., Gao, H., et al. (2021). Linkage analysis, GWAS, transcriptome analysis to identify candidate genes for rice seedlings in response to high temperature stress. *BMC Plant Biol.* 21 (1), 85. doi: 10.1186/s12870-021-02857-2

Xu, Y., Chu, C., and Yao, S. (2021). The impact of high-temperature stress on rice: Challenges and solutions. *Crop J.* 9 (5), 963–976. doi: 10.1016/j.cj.2021.02.011

Xu, Y., Zhang, L., Ou, S., Wang, R., Wang, Y., Chu, C., et al. (2020). Natural variations of SLG1 confer high-temperature tolerance in indica rice. *Nat. Commun.* 11 (1), 5441. doi: 10.1038/s41467-020-19320-9

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88 (1), 76–82. doi: 10.1016/j.ajhg.2010.11.011

Yang, Y., Zhang, C., Zhu, D., He, H., Wei, Z., Yuan, Q., et al. (2022). Identifying candidate genes and patterns of heat-stress response in rice using a genome-wide association study and transcriptome analyses. *Crop J* 10 (6), 1633–1643. doi: 10.1016/j.cj.2022.1002.1011

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an r package for comparing biological themes among gene clusters. *Omics* 16 (5), 284–287. doi: 10.1089/omi.2011.0118

Zhang, J., Li, X. M., Lin, H. X., and Chong, K. (2019). Crop improvement through temperature resilience. *Annu. Rev. Plant Biol.* 70, 753–780. doi: 10.1146/annurev-arplant-050718-100016

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, H., Zhou, J.-F., Kan, Y., Shan, J.-X., Ye, W.-W., Dong, N.-Q., et al. (2022). A genetic module at one locus in rice protects chloroplasts to enhance thermotolerance. *Science* 376 (6599), 1293–1300. doi: 10.1126/science.abo5721

Zhou, Y., Yang, Y., Zhou, X., Chi, Y., Fan, B., and Chen, Z. (2016). Structural and functional characterization of the VQ protein family and VQ protein variants from soybean. *Sci. Rep.* 6, 34663. doi: 10.1038/srep34663

Zhu, J. K. (2016). Abiotic stress signaling and responses in plants. *Cell* 167 (2), 313–324. doi: 10.1016/j.cell.2016.08.029

Zhu, S., Huang, R., Wai, H. P., Xiong, H., Shen, X., He, H., et al. (2017). Mapping quantitative trait loci for heat tolerance at the booting stage using chromosomal segment substitution lines in rice. *Physiol. Mol. Biol. Plants* 23 (4), 817–825. doi: 10.1007/s12298-017-0465-4

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 09 January 2023
DOI 10.3389/fpls.2022.1102196

OPEN ACCESS

EDITED BY
Zhichao Wu,
National Institutes of Health (NIH),
United States

REVIEWED BY
Rabarijaona Romer,
University of Antananarivo,
Madagascar
Xiaoping Lian,
Yunnan University, China

*CORRESPONDENCE
Xiaoping Guo
xpguo@mail.hzau.edu.cn
Ling Min
lingmin@mail.hzau.edu.cn

†These authors have contributed equally to this work

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

RECEIVED 18 November 2022
ACCEPTED 19 December 2022
PUBLISHED 09 January 2023

CITATION
Zhang J, Wu P, Li N, Xu X, Wang S, Chang S, Zhang Y, Wang X, Liu W, Ma Y, Manghwar H, Zhang X, Min L and Guo X (2023) A male-sterile mutant with necrosis-like dark spots on anthers was generated in cotton. *Front. Plant Sci.* 13:1102196. doi: 10.3389/fpls.2022.1102196

A male-sterile mutant with necrosis-like dark spots on anthers was generated in cotton

Jun Zhang[1,2†], Peng Wu[1†], Ning Li[1†], Xiaolan Xu[1], Songxin Wang[1], Siyuan Chang[1], Yuping Zhang[1], Xingxing Wang[1], Wangshu Liu[2], Yizan Ma[1], Hakim Manghwar[1,3], Xianlong Zhang[3], Ling Min[1*] and Xiaoping Guo[1*]

[1]National Key Laboratory of Crop Genetic Improvement & Hubei Hongshan Laboratory, Huazhong Agricultural University, Wuhan, China, [2]Zhejiang Provincial Key Laboratory of Crop Genetic Resources, Institute of Crop Science, Plant Precision Breeding Academy, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, China, [3]State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources, South China Agricultural University, Guangzhou, China

Although conventional hybrid breeding has paved the way for improving cotton production and other properties, it is undoubtedly time and labor consuming, while the cultivation of male sterile line can fix the problem. Here, we induced male sterile mutants by simultaneously editing three cotton *EXCESS MICROSPOROCYTES1* (*GhEMS1*) genes by CRISPR/Cas9. Notably, the *GhEMS1* genes are homologous to *AtEMS1* genes, which inhibit the production of middle layer and tapetum cells as well, leading to male sterility in cotton. Interestingly, there are necrosis-like dark spots on the surface of the anthers of *GhEMS1s* mutants, which is different from *AtEMS1* mutant whose anther surface is clean and smooth, suggesting that the function of *EMS1* gene has not been uncovered yet. Moreover, we have detected mutations in *GhEMS1* genes from T_0 to T_3 mutant plants, which had necrosis-like dark spots as well, indicating that the mutation of the three *GhEMS1* genes could be stably inherited. Dynamic transcriptomes showed plant hormone pathway and anther development genetic network were differential expression in mutant and wild-type anthers. And the lower level of IAA content in the mutant anthers than that in the wild type at four anther developmental stages may be the reason for the male sterility. This study not only facilitates the exploration of the basic research of cotton male sterile lines, but also provides germplasms for accelerating the hybrid breeding in cotton.

KEYWORDS

cotton, GhEMS1s, CRISPR/Cas9, male-sterile line, necrosis-like dark spots

Introduction

Male sterility is an important tool for the utilization of heterosis such as increasing cotton production and quality with less labor and time. For now, artificial emasculation is still the dominant method used for the production of cotton hybrids in China (Yang et al., 2018). However, the cost of hybrid breeding has been increasing year by year due to the shortage of rural labor, resulting in dramatic decreases of production in the planting area of hybrid cotton (Yang et al., 2018; Zheng et al., 2021). In this way, the creation of male sterile lines is a new breakthrough for the acquisition of hybrid seedlings.

Recently, the CRISPR/Cas9 technology has been widely used in gene editing and the plants acquired can be used for hybrid seed production. Due to its precision, simple operation and high efficiency, the CRISPR/Cas9 technology has been applied for a variety of species such as maize, wheat, soybean and rice (Chen et al., 2018; Ma et al., 2019; Okada et al., 2019; Chen et al., 2021). Novel "transgene clean" thermo-sensitive genic male sterility (TGMS) lines have been created on the basis of the induced specific mutations in *TMS5* with the CRISPR/Cas9 technology. To test the combinatorial capacity of the obtained new male sterile mutants, the rice *TMS5* mutants have crossed with other lines and found that the offspring have better phenotypes and provide higher yields (Zhou et al., 2016). In addition, using the CRISPR/Cas9 technology to target *ABORTED MICROSPORES* (*AMS*) congeners in soybeans to produce stable male sterility lines. Furthermore, they have eventually figured out that the editing of *GmAMS1* is related to not only the formation of the pollen wall but also the degradation of the tapetum (Chen et al., 2021). Ramadan et al. have successfully generated a wide scale of genotypically and phenotypically mutagenesis using CRISPR/Cas9 mediated pooled sgRNAs assembly, paving the way for creation of cotton male sterile lines (Ramadan et al., 2021).

As shown in the anther and pollen-related gene regulatory network diagram (Wilson and Zhang, 2009), the early anther cell differentiation gene *EXCESS MICROSPOROCYTES1/EXTRA SPOROGENOUS CELLS* (*EMS1/EXS*) encodes leucine-rich repeat (LRR) receptor kinase which is located on the cell membrane, and the protein is expressed in primary cell wall and tapetum cells. Hence, mutations in *EMS1/EXS* gene is related to the absence of tapetum in *Arabidopsis thaliana*, eventually resulting in pollen abortion (Canales et al., 2002). Moreover, the *MULTIPLE SPOROCYTE* (*OsMSP1*) gene in rice is homologous to the *AtEMS1* genes, which also encodes LRR receptor kinases as well, and the *OsMSP1* mutant exhibits a highly similar phenotypes to the *Arabidopsis EMS1/EXS* mutant (Ken-Ichi Nonomura et al., 2003). Therefore, the *EMS1* is an important candidate gene for obtaining male sterile lines. At the same time, we also found that the expression of *GhEMS1* was affected in the high temperature sensitive line by high temperature. Here, we have created a complete male sterile line by knocking out the *GhEMS1s* using CRISPR/Cas9 technology. Through the comparison of pollen fertility and tapetum development of edited male sterile lines, the most suitable mutant was identified. This study not only facilitates the exploration of the basic research of cotton sterile lines, but also provides germplasms for accelerating the hybrid breeding using male sterile lines in cotton.

Materials and methods

Plant materials and growth conditions

Jin668, an upland cotton (*Gossypium hirsutum* L.) line developed by the National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University. We have described the transformation system of Jin668 previously (Jin et al., 2006; Li et al., 2019). The wild-type (negative control) and transgenic lines were planted in Wuhan, Hubei under normal farming practices or grown in the greenhouse during the winter in 2018 - 2020. The greenhouses were kept at a temperature of 28–35/20–28°C day/night.

Vector construction and transformation of cotton

We conducted a genome-wide assessment and chose sgRNA through the CRISPR-P 2.0 (http://crispr.hzau.edu.cn/cgi-bin/CRISPR2/CRISPR) (Liu et al., 2017). The process of vector construction refers to our previous report (Wang et al., 2018). The different vectors were transformed into Agrobacterium *GV3101* which then was transformed into cotton Jin668. Refer to published articles for cotton transgene (Jin et al., 2006).

Hi-TOM and gene editing efficiency

In order to detect the editing efficiency of transgenic lines, the targeted genomic DNA was amplified by PCR with a pair of site-specific primers at the 5' end with common bridging sequences. Specific steps were similar to previous reports (Liu et al., 2019; Ramadan et al., 2021), and the primers used ware shown in Supplementary Table 1. PCR products were sequenced on Illumina HiSeq platform (Illumina, USA) after recovery. Hi-TOM website (http://www.hi-tom.net/hi-tom/) was used to analyze the sequencing results. In order to detect the off-target situation in the transgene lines editing process, "sgRNAcas9_3.0.5" (Xie et al., 2014) was used to predict off-target sites, following the software default settings. The "extract_targetSeq.pl" script was used in the software package to extract the flanking sequence of the off-target site on the genome. We designed off-target site primers in batches through the "batchprimer3" website (http://batchprimer3.bioinformatics.

ucdavis.edu), PCR products were sequenced on Illumina HiSeq platform (Illumina, USA). "CRISPResso2" (Clement et al., 2019) was employed for sequencing results to analyze the off-target sites. The off-target sites and sequences are shown in Supplementary Table 2, 3. The primers for amplification of off-target sites are shown in Supplementary Table 1.

Observation of anther phenotype

To detect pollen viability, the anther at 0 days post-anthesis (DPA) of WT and mutants was immersed in 2,3,5-Triphenyl tetrazolium chloride (TTC) solution (8 g TTC dissolved in 1 L phosphate buffer) according to a previous report (Min et al., 2013). After being cultured in a 37°C incubator for 30 min, the staining reaction was terminated with 2% (v/v) sulfuric acid solution. Pollen grains were placed on a microscope slide and the Zeiss (Oberkochen, Germany) Axio Scope A1 microscope was used to collect images.

Polyacrylamide gel electrophoresis

In polyacrylamide gel electrophoresis (PAGE) separations, the 8% non-denaturing polyacrylamide gel (acrylamide: methylene bisacrylamide = 29:1) containing the PCR amplification products were placed in the electrophoresis chamber, and the driving force was set to 60 W. After 1 hour, the sample products were immersed in 0.2% silver nitrate solution for 10 minutes. After that, the products were washed twice with ddH_2O and then put them in the chromogenic solution (1.5% sodium hydroxide, 0.4% formaldehyde) for 5 minutes. Finally, protein band patterns could be visualized and subjected to adequate analysis.

Tissue dissection and PCD assays

Anthers from transgenic lines and wild-type at different developmental stages were immersed in 50% FAA (50% ethanol, 5% propionic acid, and 3.7% formaldehyde) and vacuum infiltrated for 2 h at 4°C, and placed at 4°C for 24 h to fix the tissue. For dehydration, a graded ethanol series (50, 70, 80, 95, and 100%) was used and samples were embedded in paraffin. The embedded tissues were sectioned into 10 μm sections. Anther sections were stained with toluidine blue solution (1%) and the Zeiss (Oberkochen, Germany) Axio Scope A1 microscope was used to collect images. TUNEL detection of apoptosis was performed similar to previous report (Min et al., 2013). Paraffin sections of the anthers were used for TUNEL analysis of the fragmented DNA of apoptotic cells using the DeadEnd™ Fluorometric TUNEL System (G3250, Promega). The analytical wavelengths of fluorescein and propidium iodide were 520 ± 10 nm and 640 ± 10 nm by a confocal microscope (TCS SP2; Leica), respectively.

RNA extraction and RNA-seq

Anthers from transgenic lines and wild-type were sampled and total RNA was extracted. The library preparations were sequenced on an Illumina Novaseq platform and 150 bp paired-end reads were generated. Raw data of fastq format were firstly processed through FastQC (Andrianov et al., 2010). Paired-end clean reads were aligned to the *G. hirsutum* genome using Hisat2 v2.0.5 (Kim et al., 2015). FeatureCounts v1.5.0-p3 was used to count the reads numbers mapped to each gene (Liao et al., 2014). And then fragments per kilobase of exon model per million mapped fragments of each gene was calculated based on the length of the gene and reads count mapped to this gene. Differential expression analysis of two groups was performed using the DESeq2 R package (1.16.1) (Love et al., 2014). Genes with Padj <0.05 and $|\log_2\text{FoldChange}| > 1$ were assigned as differentially expressed. The GO enrichment was performed by the R package 'clusterProfiler' (Yu et al., 2012).

Hormone determination

Extraction and measurement of the endogenous IAA were as described by Miao et al. (Miao et al., 2019). Three replicates, each of 100 mg of anthers from transgenic lines and wild-type, were sampled at anther developmental stage 6, 7, 9 and 10, mixed with 750 μL of ice-cold 80% methanol containing $^2H_5$-IAA (OlChemlm Ltd, CAS: 76934-78-5, 10 ng ml^{-1}) as internal standard, and shake for 16 hours in the dark at 4°C. After centrifugation at 13,000 rpm for 5 minutes, the supernatant was dried with nitrogen, and the residue was reconstituted in 300 μL of 80% methanol. Finally, the IAA content was measured using an Agilent 4000Q-TRAP HPLC-MS system.

Results

Identification of EMS1 genes in *G. hirsutum*

The *Arabidopsis EXCESS MICROSPOROCYTES1* (*AT5G07280*, *AtEMS1*) controls somatic and reproductive cell development in anthers (Zhao et al., 2002). To determine whether the *EMS1* gene participated in the reproductive cell development in cotton, we used *AtEMS1* as a query to perform BLASTP searches and identified 11 *EMS1* members in *G.*

hirsutum. To get a better understanding of the phylogenetic relationships between *EMS1*, a phylogenetic tree was constructed based on these 11 *G. hirsutum EMS1* and *AtEMS1* protein sequences. Clearly, the EMS1s were classified into four branches, the *Ghir_A08G010860* (*GhEMS1_A08*), *Ghir_D08G010810* (*GhEMS1_D08*), and *Ghir_A09G018830* (*GhEMS1_A09*) in the same branch with *AtEMS1* (Figure 1A). *GhEMS1_A08*, *GhEMS1_D08* and *GhEMS1_A09* have leucine rich repeat N-terminal domain and leucine-rich repeat sequences (Supplementary Figure 1). This result indicated that the three *GhEMS1* genes may have similar functions with *AtEMS1*.

To further identify the biological function of the *EMS1* genes involved in cotton-specific developmental processes, we have summarized the expression of *EMS1* genes in different organs/tissues (including roots, leaves, anthers in different length buds) of *G. hirsutum*. As shown in Figure 1B, most of *GhEMS1* genes have expressed in anthers, and *Ghir_A08G010860* (*GhEMS1_A08*), *Ghir_D08G010810* (*GhEMS1_D08*), and *Ghir_A09G018830* (*GhEMS1_A09*) were all predominantly expressed in early-stage anthers (stage 4/5, bud length: 3~5 mm) and their expression gradually decreased with anthers development, implying that these genes may play crucial roles in identification of reproductive cell development in early stage anthers.

Knockout of *Gh*EMS1 genes using CRISPR/Cas9 caused male sterility with necrosis-like dark spots on the anther surface

Due to combination of two sgRNAs with tRNA can improve the transcription and knockout efficiency (Xie et al., 2015; Wang et al., 2018). Thus, two sgRNAs targeting the same *GhEMS1* genes were combined by overlap extension PCR and then ligated to the expression vector pRGEB32-GhU6.7, to produce polycistronic tRNA-gRNA genes *PTG1* and *PTG2* vectors (Figure 1C). The *PTG1* contained sgRNA1 and sgRNA2, which was designed to knock out *GhEMS1_A08*, *GhEMS1_D08*, and *GhEMS1_A09* (Figure 1D). The *PTG2* contained sgRNA3 and sgRNA4, which was designed to knock out *GhEMS1_A08* and *GhEMS1_D08* at the same time (Figure 1D). The prepared vectors were transformed the cotton by Agrobacterium (*GV3101*) (Supplementary Figure 2). We obtained 3 sterile plants (KO1~3) for *PTG1*, and 2 sterile

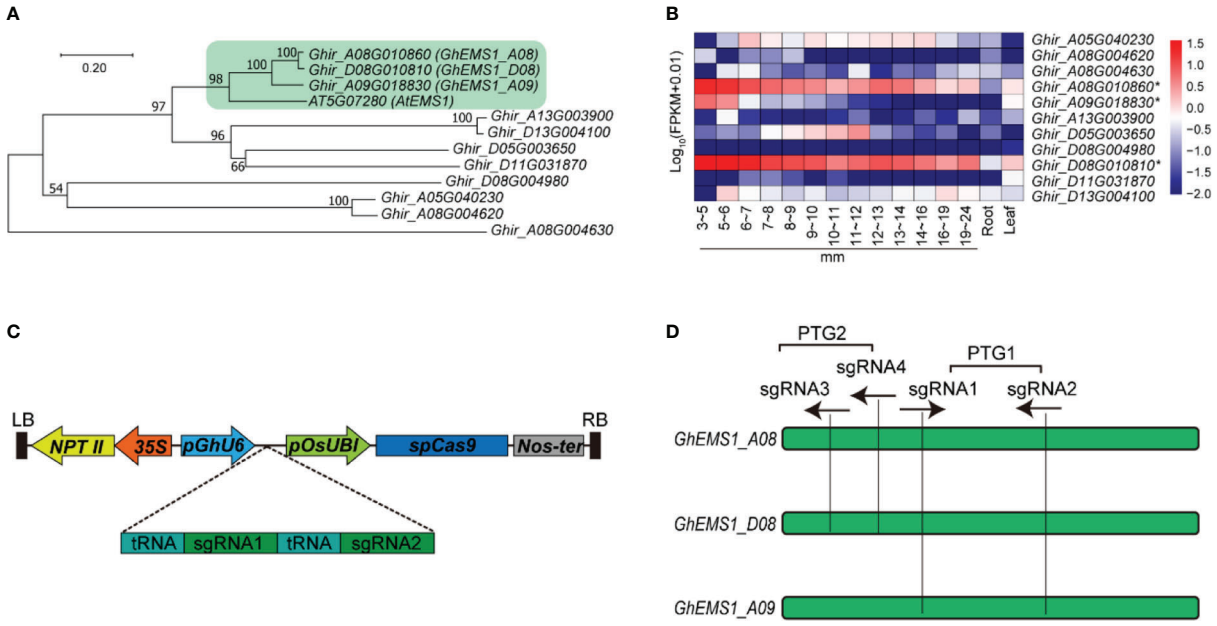
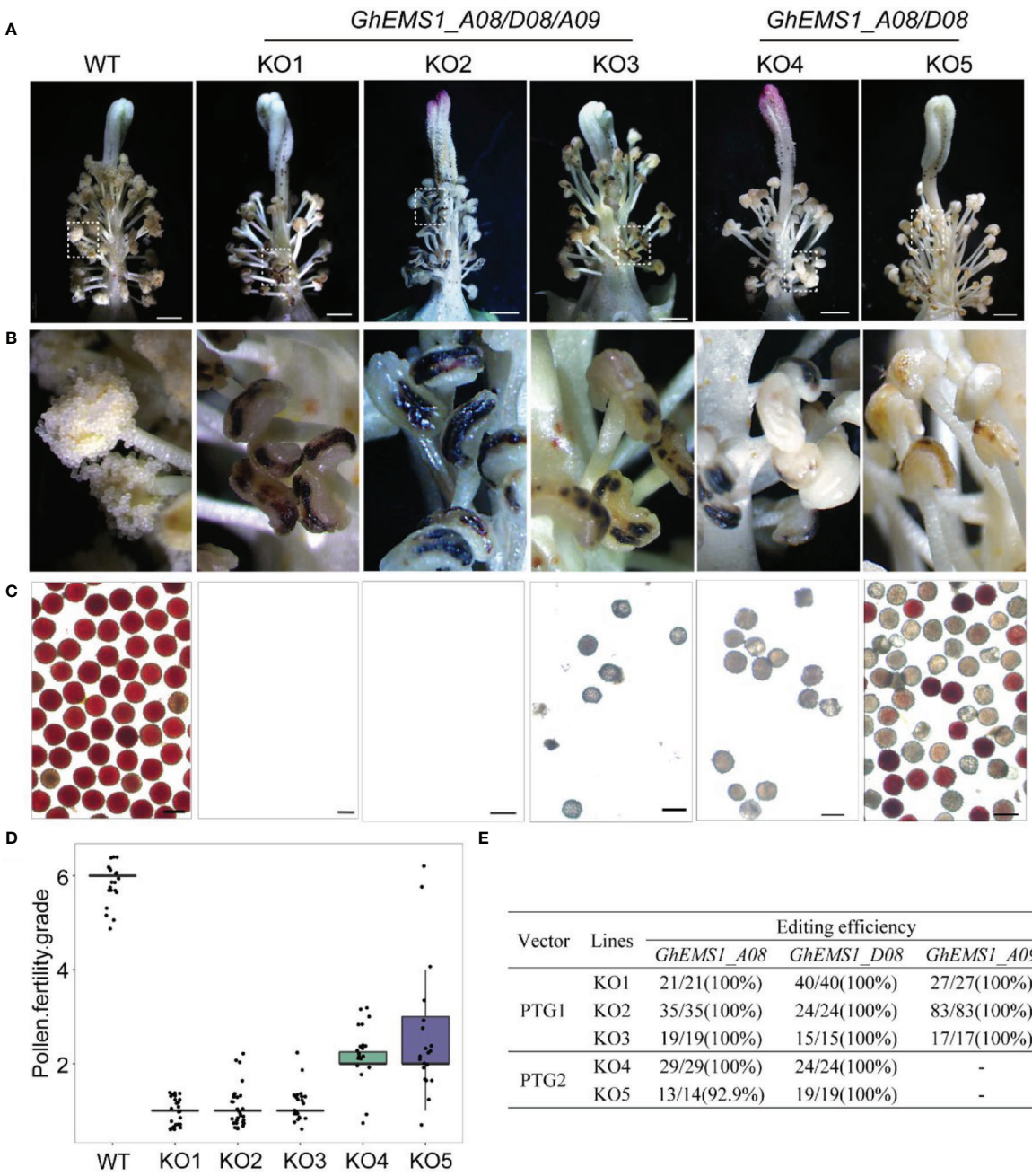


FIGURE 1
Creation of *Gh*EMS1 genes mutants by using CRISPR/Cas9. **(A)** A phylogenetic tree for EMS1 genes in *Gossypium hirsutum* and AtEMS1 was constructed using the neighbor-joining method in MEGA 10.1.8 followed by bootstrapping with 1,000 replicates; **(B)** Expression profiles of *GhEMS1* genes in flower buds of different lengths (3~5 mm, 5~6 mm, 6~7 mm, 7~8 mm, 8~9 mm, 9~10 mm, 10~11 mm, 11~12 mm, 12~13 mm, 13~14 mm, 14~16 mm, 16~19 mm, 19~24 mm)in *G. hirsutum* H05 determined by transcriptome sequencing (Zhang et al., 2022); **(C)** Creation of a male-sterile line pool by CRISPR/Cas9. Two sgRNAs were serially connected to each vector to increase the knockout efficiency; **(D)** Two pairs of sgRNAs on exons were designed and vectors containing polycistronic tRNA–gRNA genes (*PTG*). *Ghir_A08G010860 (GhEMS_A08)*, *Ghir_D08G010810 (GhEMS_D08)*, and *Ghir_A09G018830 (GhEMS_A09)* were knocked out by sgRNA1 and sgRNA2 (*PTG1*). *GhEMS_A08* and *GhEMS_D08 (PTG2)* were knocked out by sgRNA3 and sgRNA4 (*PTG2*).

10.3389/fpls.2022.1102196



Vector	Lines	Editing efficiency		
		GhEMS1_A08	*GhEMS1_D08*	*GhEMS1_A09*
PTG1	KO1	21/21(100%)	40/40(100%)	27/27(100%)
	KO2	35/35(100%)	24/24(100%)	83/83(100%)
	KO3	19/19(100%)	15/15(100%)	17/17(100%)
PTG2	KO4	29/29(100%)	24/24(100%)	-
	KO5	13/14(92.9%)	19/19(100%)	-

FIGURE 2
Phenotypes of *GhEMS1s* mutants. **(A)** The phenotypes of the WT and KO1 - KO5 transgenic lines, three genes *Ghir_A08G010860* (*GhEMS_A08*), *Ghir_D08G010810* (*GhEMS_D08*), and *Ghir_A09G018830* (*GhEMS_A09*) of *PTG1* in KO1- KO3 sterile plants. *GhEMS_A08* and *GhEMS_D08* (*PTG2*) were knocked out in KO4 and KO5 plants. Scale bars: 2 mm; **(B)** Partial enlarged view of anther in picture (a); **(C)** TTC (2,3,5-triphenyl tetrazolium chloride) was used to detect the pollen viability. No pollen was observed in the anthers of KO1 and KO2 plants. Scale bars: 100 μm. **(D)** Statistics for pollen fertility in the *GhEMS1* mutants. During the phenotypic investigation, we divided plant fertility into six levels: Grade 1 indicates that anthers with necrosis-like dark spots and all anthers without pollen; Grade 2 indicates that < 25% of the anthers have a few inactive pollen grains without dehiscence; Grade 3, 4, and 5 indicate that 25%, 50%, and 75% of the anthers spread pollen, respectively; Grade 6 indicates that all anthers dehiscence and release active pollen. The phenotype of every flower was recorded, and the fertility of different plants was counted. **(E)** Analysis of gene editing efficiency in different *GhEMS* mutant lines by Sanger sequencing.

plants (KO4~5) for *PTG2* (Figure 2A). Five independent transformation plants showed different degrees of necrosis-like dark spots on the surface of anther and have different pollen viability (Figures 2A-C). KO1-KO3, showed obvious necrosis-like dark spots on the surface of the anthers (Figures 2A, B). There was no pollen in the anthers of KO1 and KO2, and only a few shriveled pollen grains in the KO3 anthers (Figure 2C). KO4 and KO5 showed a few anthers with dark spots, and the pollen quantity and viability were lower than the wild type (WT) but higher than those of KO1- KO3 (Figures 2A-C).

Furthermore, a three-month fertility assay was performed on WT and the five transgenic plants. We divided plant fertility into six levels: Grade 1 indicates that anthers with necrosis-like dark spots and all anthers without pollen; Grade 2 indicates that < 25% of the anthers have a few inactive pollen grains without dehiscence with a few anthers have dark spots; Grade 3, 4, and 5 indicate that 25%, 50%, and 75% of the anthers spread pollen, respectively; Grade 6 indicates that all anthers dehiscence and release active pollen. The phenotype of every flower was recorded, and the fertility of different plants was counted. The results showed that the fertility of the WT plants was relatively stable at grade 6, with pollen viability higher than 99.5% and anthers normal dehiscence and no dark spots; KO1- KO3 were below grade 2, with no pollen or very few inactive pollen grains, and dark spots on the anther surface (Figures 2C, D). KO1 was the most stable, with 100% anthers of all flowers having necrosis-like dark spots on the anther surface, and no pollen (Figure 2D). The fertility of KO4 and KO5 was above grade 2 and below grade 6, with a few anthers have dark spots (Figures 2B-D).

Identification of target gene editing in male sterile plants

To check the mutation at the selected target site in KO1-KO5 lines, the sanger sequencing was performed. We found that three genes, *GhEMS1_A08*, *GhEMS1_D08* and *GhEMS1_A09*, all were 100% edited in KO1-KO3 plants (Figure 2E), and the deletion length was in the range of 2 to 29 bp (Supplementary Figure S3). In the KO4 and KO5 plants, *GhEMS1_A08* and *GhEMS1_D08* were successfully edited with 100% editing efficiency, and no editing in the *GhEMS1_A09* (Figure 2E and Supplementary Figure S3B). Moreover, the expression level of *GhEMS1_A09* was lower in the early stage anthers, compared with the expression of *GhEMS1_A08* and *GhEMS1_D08* in the same stage anthers, and the expression level of *GhEMS1_A09* decreased earlier (Figure 1B). In all, the male fertility of KO4 and KO5 was better than that of KO1-KO3 (Figure 2C), indicating that the three *GhEMS1* genes were essential for male fertility and played crucial roles in male fertility. In addition, 38 potential off-target genes of the two sgRNAs in KO1 (complete male sterility plant) were analyzed, and no off-target effects were found in KO1 (Supplementary Tables 2, 3), this result suggested the formation of male sterility of KO1 only caused by the mutations of *GhEMS1_A08*, *GhEMS1_D08* and *GhEMS1_A09* genes.

GhEMS1 mutants displayed the genetic stability of the necrosis-like dark spots as a marker of male sterility

Whether the sterile phenotype can be stably inherited to the offspring is related to the successful application of sterile mutant to cross breeding. To test the genetic stability of the necrosis-like dark spots as a marker of sterility, we applied WT pollen to the stigma of sterile KO1 plants. Then, the target site fragments in *PTG1* of *GhEMS1_A08*, *GhEMS1_D08* and *GhEMS1_A09* in WT, KO1, T1 generation (KO1×WT) were amplified by PCR, and polyacrylamide gel electrophoresis (PAGE) was used to identify the editing. The results showed that the T1 generation had the same fragment with T0 generation, but due to WT pollination, some new editing types were generated, such as in T1-3, for which two sgRNAs generated new editing types (Supplementary Figure 3C). The individual sgRNAs showed different efficiencies in the detection of PAGE, indicating that it is necessary to connect two sgRNAs in tandem with one vector (Supplementary Figure 3C). In addition, the necrosis-like dark spots on the anthers could also be observed from the T0 to T3 (Figures 2A, B and 3A), T3 plants (n=64) had 35.9% (23/64) complete necrosis anthers (Figure 3A). Gene editing types were identified by Hi-TOM (Liu et al., 2019), which revealed that the T3 generation had the same editing type as the T0 generation, such as the mutations (-1 bp, -5 bp, -6 bp) at the sgRNA1 target site (Figures 4A, C, E, G), and mutations (-2 bp, -20 bp, -1 bp) at the sgRNA2 target site (Figures 4B, D, F, H). However, due to WT pollination and the retention of Cas9, some new editing types have also been generated, such as the mutation (+1 bp) at the sgRNA1 target site (Figure 4A). Moreover, the pollen quantity and viability of T3 plants were analyzed, the results were consistent with the gene editing efficiency and the number of necrosis-like dark spots on the anthers in these plants, suggested the *GhEMS1* mutants displayed the genetic stability of the necrosis-like dark spots as a marker of male sterility.

GhEMS1s mutants lack of middle and tapetum layers might cause delayed microsporocytes development

The necrosis-like dark spots on the mature anthers could be inherited, but when and how the necrosis-like dark spots appeared on the anther surface? To explore the formation period of necrosis-like dark spots, we obtained the stage 6 to stage 14 anthers of *KO1*. At stage 7, yellow spots appeared on the anthers of sterile phenotype plants, which gradually deepened at

10.3389/fpls.2022.1102196

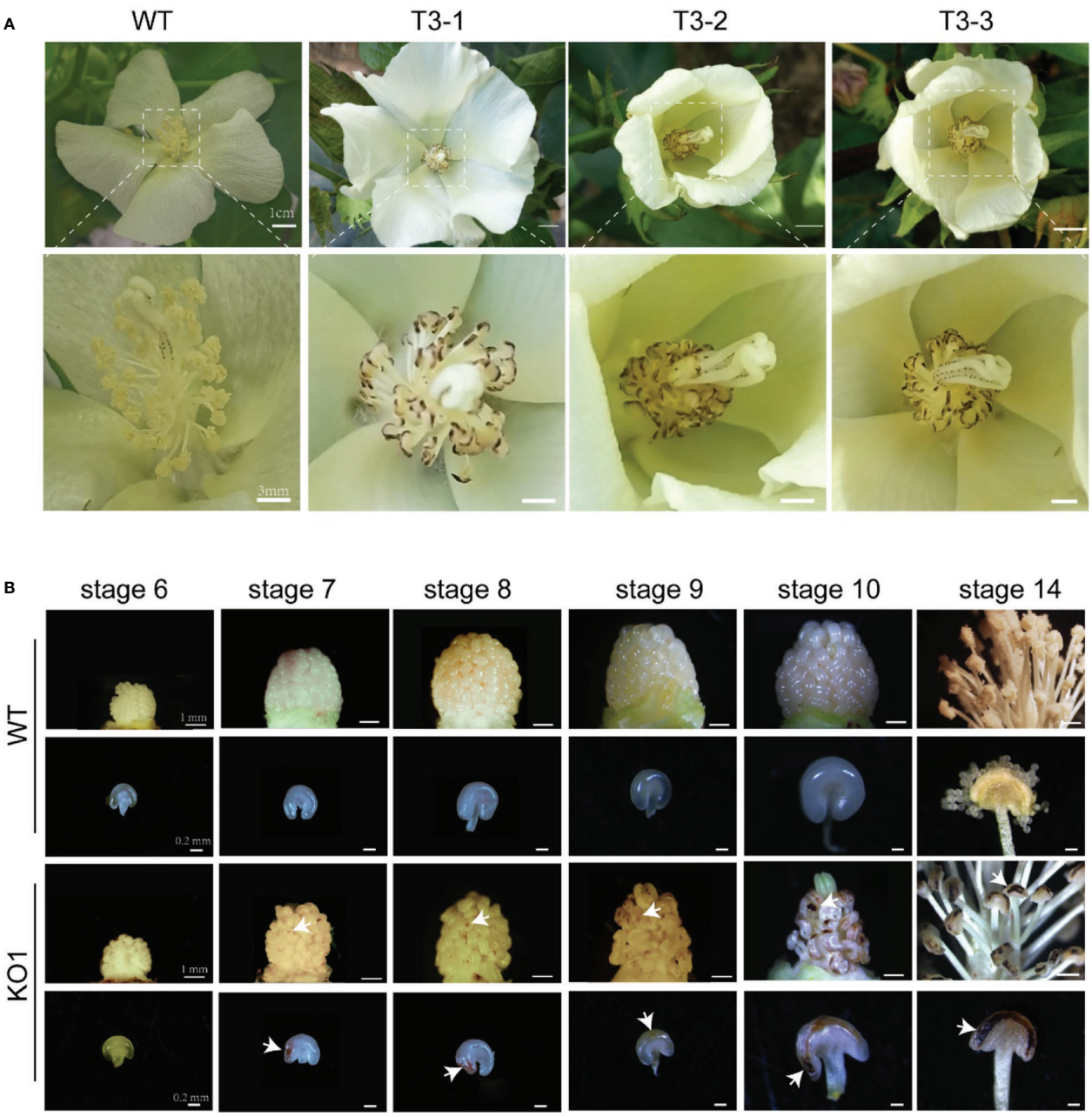


FIGURE 3
The sterile phenotype of surface necrosis of T3. **(A)** wild type, KO1 T3 plants, and enlarged images. **(B)** Changes in the anthers with different degrees of necrosis-like spots. Compared with those of the WT, yellow spots appeared on the anthers of sterile KO1 plants at stage 7, these spots gradually deepened at stage 10 and stage 14 to form necrosis-like dark spots eventually. KO1, *GhEMS* 3-genes simultaneous mutant. The white arrows indicate necrotic spots.

stage 10 and stage 14 to form necrosis-like dark spots eventually (Figure 3B). Therefore, the dark spots on the surface of the anthers started earlier, which can help us to screen sterile plants at the early stage.

To observe the microspore development of *KO1*, anther tissue cross-sections were made. At stage 6, the WT exhibited four complete anther cell layers (from outside to inside: epidermis, endothecium, middle layer, and tapetum layer) and microsporocytes. However, the sterile mutant anthers lacked middle layer and tapetum cells (Figures 5A, C). Furthermore, in the WT anthers, the microsporocytes completed nuclear division at stage 6 (Figure 5A), tetrads formed at stage 7 (Figure 5A),

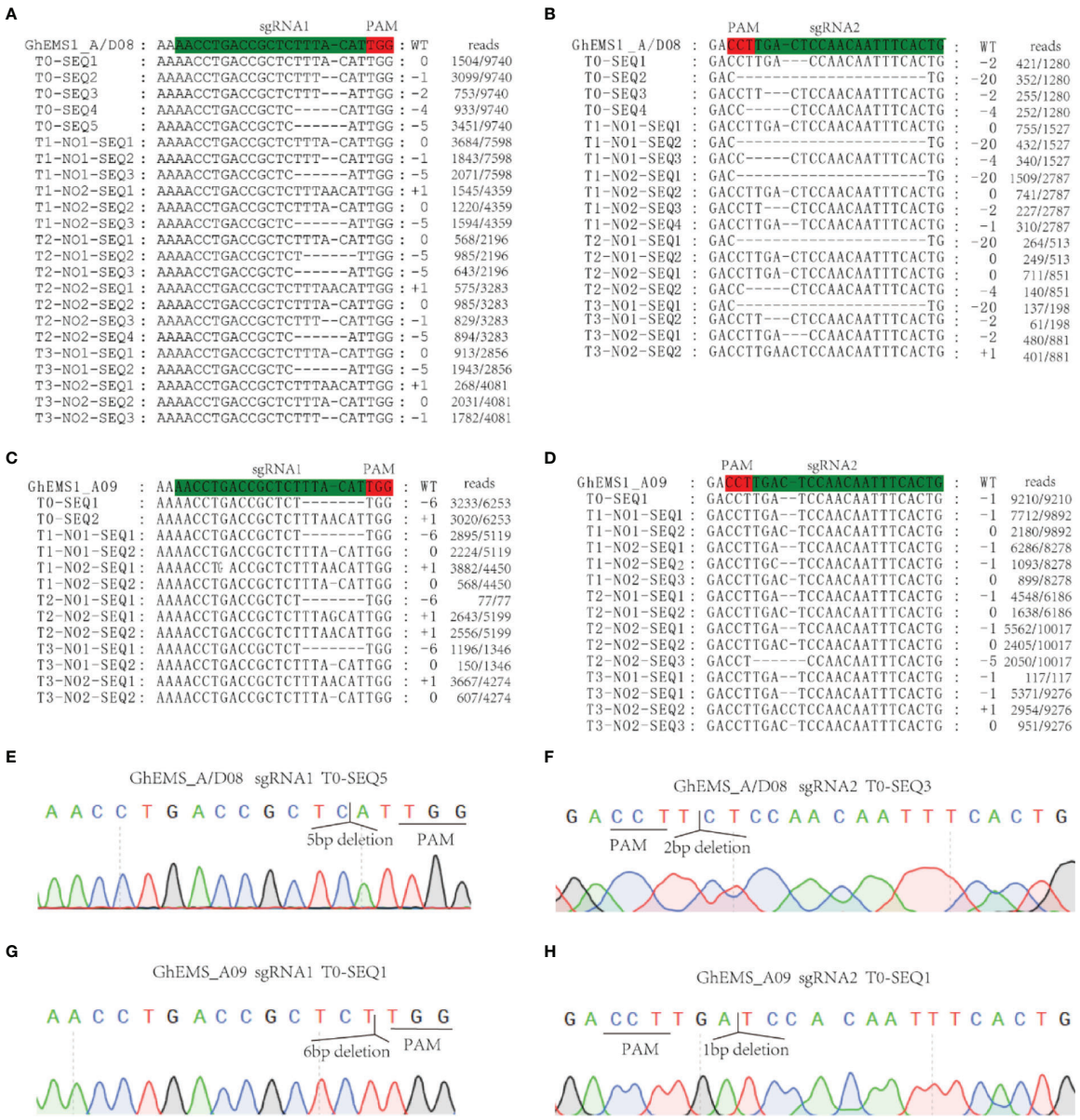


FIGURE 4
KO1 editing types were inherited in T0 to T3. **(A, B)** *GhEMS_A/D08* were edited in sgRNA1 and sgRNA2, respectively. **(C, D)** *GhEMS_A09* were edited in sgRNA1 and sgRNA2, respectively. **(E)** Sequencing peak photos of *GhEMS_A/D08* with 5 bp deletions at the sgRNA1 site. **(F)** Sequencing peak photos of *GhEMS_A/D08* with 2 bp deletions at the sgRNA2 site. **(G)** Sequencing peak photos of *GhEMS_A09* with 6 bp deletions at the sgRNA1 site. **(H)** Sequencing peak photos of *GhEMS_A09* with 1 bp deletions at the sgRNA2 site.

microspores were released from tetrads at stage 8 (Figure 5A), and then microspores developed into mature pollen during stage 9 to stage 11 with the tapetum development and degeneration (Figure 5A). However, the mutants could complete nuclear division but not cytoplasmic division, causing the enlarged microsporocytes and undetached microsporocytes at stages 7 to 9 (Figure 5A). At stage 10, the microsporocytes of mutants started to degrade, and they were completely degraded at stage 11 (Figure 5A). What's more, the degree of DNA fragmentation in the anthers was detected by TUNEL (terminal deoxynucleotidyl transferase dUTP nick end labeling). In the WT, there were a few yellow fluorescence signals in the tapetum layer, a few microspores at stage 9, and the yellow fluorescence signals were observed enhanced at stage 10. However, no fluorescence signals were observed in the stage 9 anthers of the mutant, and only faint fluorescence signals were observed appearing in the degrading microsporocytes in the stage 10 anther locules in the mutant (Figures 5B, C). This result

10.3389/fpls.2022.1102196

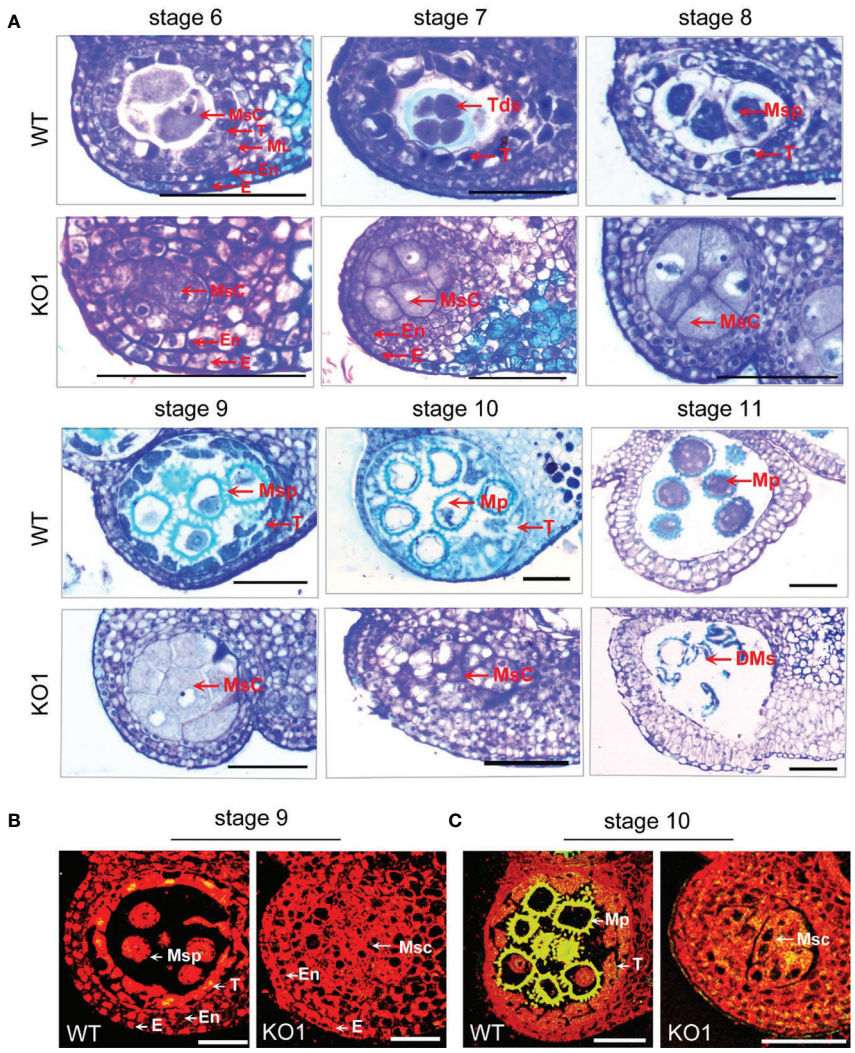


FIGURE 5
Comparison of the histological characteristics of the WT and KO1 sterile line T3 anthers. **(A)** Stage 6-11 histological characteristics of the WT and KO1, Scale bars: 100 μm **(B, C)** Analysis of DNA damage in anthers of WT and KO1 male-sterile plants. The degree of DNA fragmentation of anthers was detected by TUNEL. Scale bars: 100 μm. DMs: degenerated microspores; E, epidermis; En, endothecium; Msc, microsporocyte; Mp, mature pollen; ML, middle layer; Msp, microspore; T, tapetum; Tds, tetrads; WT, wild type.

indicated that there was no normally developed pollen in the chamber of the mutant and pollen deformity was caused by the absence of tapetum formation and the PCD process of microspore mother cells.

Dynamic transcriptomes analysis between *Ghems1s* male sterile mutants and wild-type anthers

To explore the molecular mechanisms of anther abortion in *GhEMS1* mutants, we compared the transcriptomes of WT and *Ghems1* anthers at four developmental stages (stages 6, 7, 9, 10). A total of 7,172 genes were found to be differentially expressed between *Ghems1* and WT at four anther developmental stages (Supplementary Table 4). Of these differentially expressed genes (DEGs), 2,070 (28.86%) were up-regulated and 5,102 (71.14%) were down-regulated ($|\log_2(\text{fold change})| \geq 1$ and padj < 0.05) in *Ghems1* (Figure 6A). Compared with WT, *Ghems1* had more up-regulated genes than down-regulated genes at stages 6 and 7 (Figure 6A). However, at stages 9 and 10, *Ghems1* had more down-regulated genes than up-regulated genes (Figure 6A). Of these genes, 239, 1,250, 551 and 631 genes were unique at stages 6, 7, 9 and 10, respectively, and 23 genes showed differential expression in all four developmental stages (Supplementary Table 5 and Figure 6B). These 23 genes included three

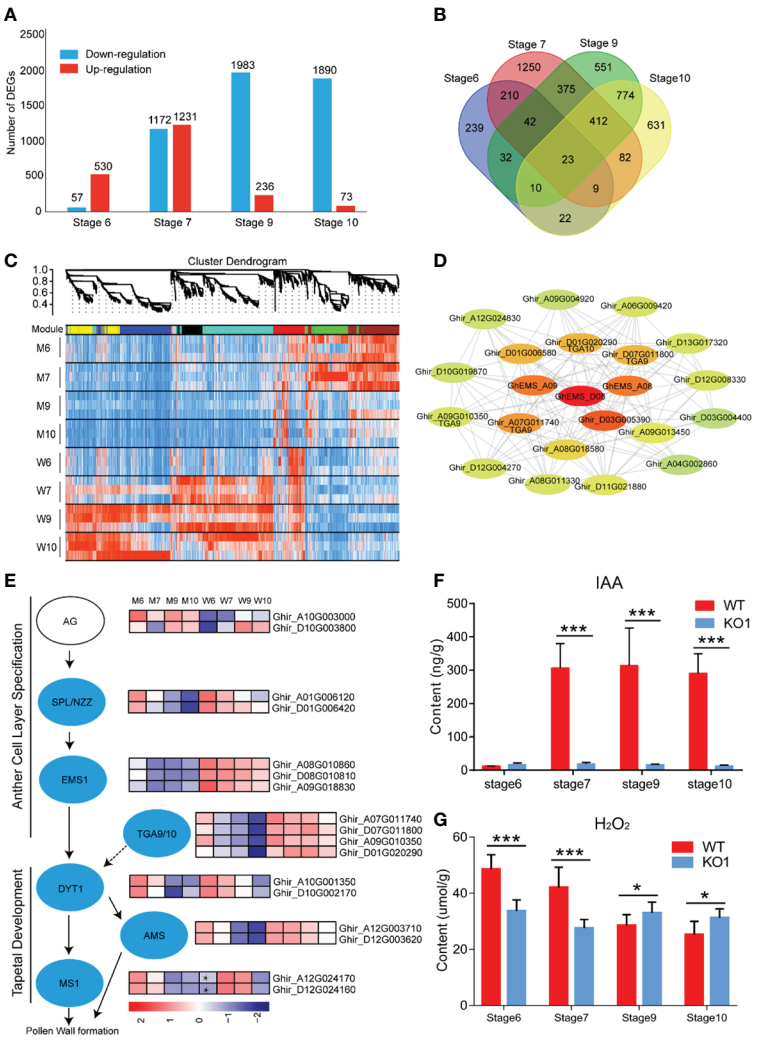


FIGURE 6
Transcriptome analysis of four developmental stages. **(A)** Difference gene statistics of wild type and mutant in four periods; **(B)** Intersection of differential genes at different stages; **(C)** WGCNA co-expression module, a total of 8 different modules; **(D)** Analysis of the co-expression network of the red module. The shades of color represent correlations; **(E)** Gene expression of the regulatory network of pollen development; **(F)** IAA content at four developmental stages; **(G)** H_2O_2 content at four developmental stages. W, wild type; M, mutant; *AG, AGAMOUS; AMS, ABORTED MICROSPORE; DYT1, DYSFUNCTIONAL TAPETUM 1; EMS1, EXCESS MICROSPOROCYTES 1; MS1, MALE STERILITY 1; NZZ/SPL, SPOROCYTELESS/NOZZLE*. Data are presented as means ± SE from five biologically independent experiments. Asterisks indicate statistically significant differences (***, P <0.001; *, P <0.05); by Student's t-test.

GhEMS1 genes that have been edited, and the expression of three *GhEMS1* genes was downregulated in all four stage mutant anthers (Figure 6E). The expression level of *GhEMS_D08* was higher than that of *GhEMS_A08/A09*, and gradually decreased following the anther development. GO enrichment analysis, "peroxidase activity" was highly enriched in stage 6 and 7 anthers of mutants, and "monooxygenase activity" is highly enriched in stage 6, 7, 9, which is related to peroxide in GO enrichment analysis (Supplementary Figure 4 and Table 6). So, we detected the content of peroxide at four developmental stages of *Ghems1* and WT anthers (Figure 6G). The results showed that compared with the WT, the mutants had lower peroxide content in stage 6 and 7, but higher peroxide content in stage 9 and 10. Black spots began to appear in stage 7 and appeared in large numbers in stage 9, might be responsible for the appearance of black spots on the surface of anthers. In mutant stage 6 down-regulated genes, "activation of MAPKK activity" and "MAP kinase kinase kinase activity" were enriched. In stage 6, 9 down-regulated genes, "pollen exine formation" was enriched, while in stage 9 and 10 DEGs was enriched in "plant- type cell wall modification", possibly related to pollen formation (Supplementary Figure 4 and Table 6).

To generate co-expression networks for all DEGs and biological samples, the weighted gene co-expression network analysis was performed. A total of 8 gene modules were identified (Figure 6C). *EMS1* genes were distributed in the red module. We extracted the red module genes that may be related to *GhEMS1* genes to do a further co-expression network analysis and showed that there are four leucine-zipper transcription factors *TGACG9/10* (*TGA9/10*, *Ghir_D01G020290*, *Ghir_A07G011740*, *Ghir_D07G011800* and *Ghir_A09G010350*) genes, one tetrapeptide alpha-pyrone reductase 1 (*TKPR1*, *Ghir_D03G005390*), and one indole-3-acetic acid-amido synthetase (*GH3.6*, *Ghir_D01G006580*), were associated with *GhEMS1*. *TGA9* and *TGA10* are expressed throughout early anther primordia, and mutations in *TGA9* and *TGA10* lead to male sterility and differential defects in abaxial (Murmu et al., 2010). In the genetic framework for control of anthers development (Wilson and Zhang, 2009), the expression trends of *SPL/NZZ*, *EMS1*, *TGA9/10*, *DYT1* and *AMS* were the same, but *MS1* in the stage 6 anthers of mutant was higher than WT (Figure 6E). Previously reported that *TGA9/10* was located downstream of *SPL/NZZ* and upstream or in parallel with *DYT1* in the genetic hierarchy that controls anther development (Murmu et al., 2010), similar to *EMS1*. Interestingly, we also found that there was co-expression of *TGA9/10* and *EMS1* (Figure 6D), it was suggested that *TGA9/10* and *EMS1* may interact with each other to regulate anther development.

In the co-expression network, we found *TKPR1* which co-expressed with *EMS1* (Figure 6D), *TKPR1* involved in the biosynthesis of hydroxylated tetraketide compounds that serve as sporopollenin precursors (the main constituents of exine) was essential for pollen wall development (Tang et al., 2009; Grienenberger et al., 2010). And GO enrichment analysis shows that "pollen exine formation" was highly enriched in stage 6 and 9 anthers of mutants. To confirm whether the synthesis of sporopollenin precursors in the mutants had been affected, we measured the autofluorescence intensity of sporopollenin of *Ghems1* and WT anthers at stages 6, 7, 9, 10 by microscopy with UV light illumination (Supplementary Figure 5) (Ma et al., 2022). In the mutant, the autofluorescence of sporopollenin was not detected in the four stages, but in WT, the autofluorescence of sporopollenin was detected in the 9 and 10 stages. This showed that the synthesis of sporopollenin was affected in the mutant.

Phytohormones play an important role in the regulation of anther development. In the co-expression network, we found *GH3.6* which co-expressed with *EMS1* (Figure 6D), *GH3.6* catalyzes the synthesis of indole-3-acetic acid (IAA)-amino acid conjugates, providing a mechanism for the plant to cope with the presence of excess auxin (Staswick et al., 2005). Compared with WT, the expression of *GH3.6* in the mutant gradually decreased (Supplementary Figure 6). So, we detected the dynamic changes of auxin (IAA) in anthers of *Ghems1s* male sterile line (Figure 6F). During the four anther developmental stages, the free IAA content of male sterile plants *KO1* changed little with the development of anther and was at a lower level. However, the free IAA content of WT increased greatly between the stage 6 and stage 7 of anthers development, and maintained a high level after that. The lower IAA content in the stage 7, 9, 10 anthers of *Ghems1s* mutants may be closely related to the occurrence of male sterility.

From the above results, we found that after the editing of both *GhEMS1_A08* and *GhEMS1_D08*, there were yellow spots on the surface of the anthers and a few fertile pollen grains in the chamber. Moreover, the results of simultaneously editing three *GhEMS1* genes showed that there were necrosis-like dark spots on the surface of the anthers, which contained completely aborted pollen grains, thus the necrosis-like dark spots can serve as a marker of completely male-sterile cotton lines with *GhEMS1s* mutants, and can help breeders to screen sterile plants at the early anther stage.

Discussion

In recent years, CRISPR/Cas9 technology has played an important role in the creation of sterile materials. The sterile genes cloned in model plants *Arabidopsis* and rice have potential applications in cotton. *AtEMS1*, *OsMSP1* encode LRR receptor kinases, and the mutants have the same phenotype, including the production of a large number of microspore mother cells, no tapetum layer and middle layer. Compared with *Arabidopsis thaliana*, cotton *GhEMS1* mutant anther surface has obvious necrotic phenotype, and the molecular mechanism needs to be further studied.

Previous reports have shown that the young microspore stage and flowering stage in rice are very sensitive to high temperature (HT) stress, and HT stress destroys the function of tapetum during microspore formation and leads to poor anther dehiscence (Endo et al., 2009). It was worth noting that the expression of *GhEMS1* genes at the tetrad stage were differently respond to HT in the HT-tolerant and -sensitive lines, could provide a theoretical basis for the study of male sterility of cotton caused by high temperature (Supplementary Figure 7).

Morphological trait markers have broad application prospects in cotton production. Among them, pigment glands, okra leaf shape, and virescently traits were more commonly studied (Zhu et al., 2008; Ma et al., 2013; Ma et al., 2016). The virescently marker is linked to sterility gene and can be used to identify sterile lines at the early seedling stage (Ma et al., 2013). For now, there has been no marker found on the anther associated with fertility. In our study, the necrosis-like dark spots appeared on the anthers in the early stage, which could be

screened during early bud periods, reducing waste of resources and allowing for hybrid breeding.

Upland cotton is a polyploid species with a larger genome (2.5 Gb), so most genes have multiple copies and high sequence similarity due to the polyploidization of *At* and *Dt* sub-genomes, which makes cotton gene engineering very difficult (Wang et al., 2019). In this study, many *EMS1*-like genes were aligned in the upland cotton genome through the amino acid sequence of the *Arabidopsis AtEMS1* gene. The three genes of *GhEMS1_A08, GhEMS1_D08* and *GhEMS1_A09* may have functional redundancy because they are in one branch and the mutant is completely male sterile by knocking out *GhEMS1_A08, GhEMS1_D08* and *GhEMS1_A09* at the same time, but the two gene mutants show partial infertility by knocking out *GhEMS1_A08, GhEMS1_D08*. Multiple genes control male sterility, which makes it difficult to find *EMS* genes using map-based cloning. Completely sterile plants cannot produce tapetum and intermediate layers, confirming that cotton *GhEMS1* is a key gene regulating cotton anther and microspore development. Thus, to establish a cotton hybrid system, the *GhEMS1* mutant can be used as the male sterile line, and the negative sterile plants (Cas9-free) can be selected and crossed with other cultivars to create excellent hybrids. The positive sterile plants (with Cas9) can be crossed with transgenic acceptors or cultivars to breed sterile lines (Supplementary Figure 8).

The effect of IAA on plant male sterility was often reported. The reduction of the express of IAA related gene in wheat was associated with the occurrence of male-sterility (Su et al., 2019). At the third stage, the contents of IAA, GA_3 and ZR in CMS were remarkably lower than its maintainer of Chinese Cabbage (Liu et al., 2014). When IAA was depleted, the vascular bundles develop abnormally, and the passage of water and nutrients into the drug compartment was blocked, resulting in abnormal microspore development and pollen abortion (Liu et al., 2014). ABA and IAA are involved in PCD of microsporocytes during meiosis in *Petunia hybrida* L. (Kovaleva et al., 2018). Sugar and IAA may be the key regulators of cotton anther response to high temperature stress (Min et al., 2014). In this study, we found that among the genes co-expressed by *EMS1* through the co-expression network, there was an auxin synthesis related gene *GH3.6*. At the same time, there was a significant difference in IAA content between the male sterile mutant and the wild type during the 6-10 stage of anthers development, which is very likely to be the cause of male sterility. We also tested other endogenous hormones, but found no regular changes. As to whether the male sterility can be restored by spraying IAA, further study is needed.

In summary, we created a new male-sterile cotton line, which may further promote the utilization of the genic male-sterile lines and the development of cotton hybrid breeding. In addition, the cotton line with three mutated *EMS1* homolog genes provides the basis for the study of cotton anther tapetum and microspore development.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: NCBI, PRJNA827503.

Author contributions

XG, XZ, LM and JZ designed the studies. JZ, XX, NL, PW, SW, YZ, XW, SC, YM performed the experiments and data analysis. WL, and HM improved the grammar of the manuscript. JZ, PW, XG and LM wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by funding from the National Key Research and Development Program of China (2022YFD1200800), Fundamental Research Funds for the Central Universities (2662019PY073), the National Natural Science Foundation of China (32072024), and the Fundamental Research Funds for the Central Universities (2021ZKPY019).

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.2022.1102196/full#supplementary-material

References

Andrianov, V., Borisjuk, N., Pogrebnyak, N., Brinker, A., Dixon, J., Spitsin, S., et al. (2010). Tobacco as a production platform for biofuel: Overexpression of arabidopsis DGAT and LEC2 genes increases accumulation and shifts the composition of lipids in green biomass. *Plant Biotechnol. J.* 8 (3), 277–287. doi: 10.1111/j.1467-7652.2009.00458.x

Canales, C., Bhatt, A. M., Scott, R., and Dickinson, H. (2002). EXS, a putative LRR receptor kinase, regulates male germline cell number and tapetal identity and promotes seed development in arabidopsis. *Curr. Biol.* 12 (20), 1718–1727. doi: 10.1016/S0960-9822(02)01151-X

Chen, R., Xu, Q., Liu, Y., Zhang, J., Ren, D., Wang, G., et al. (2018). Generation of transgene-free maize Male sterile lines using the CRISPR/Cas9 system. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01180

Chen, X., Yang, S., Zhang, Y., Zhu, X., Yang, X., Zhang, C., et al. (2021). Generation of male-sterile soybean lines with the CRISPR/Cas9 system. *Crop J.* 9 (6), 1270–1277. doi: 10.1016/j.cj.2021.05.003

Clement, K., Rees, H., Canver, M. C., Gehrke, J. M., Farouni, R., Hsu, J. Y., et al. (2019). CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat. Biotechnol.* 37 (3), 224–226. doi: 10.1038/s41587-019-0032-3

Endo, M., Tsuchiya, T., Hamada, K., Kawamura, S., Yano, K., Ohshima, M., et al. (2009). High temperatures cause Male sterility in rice plants with transcriptional alterations during pollen development. *Plant Cell Physiol.* 50 (11), 1911–1922. doi: 10.1093/pcp/pcp135

Grienenberger, E., Kim, S. S., Lallemand, B., Geoffroy, P., Heintz, D., Souza, C. D. A., et al. (2010). Analysis of TETRAKETIDE α-PYRONE REDUCTASE function in arabidopsis thaliana reveals a previously unknown, but conserved, biochemical pathway in sporopollenin monomer biosynthesis. *Plant Cell* 22 (12), 4067–4083. doi: 10.1105/tpc.110.080036

Jin, S., Liang, S., Zhang, X., Nie, Y., and Guo, X. (2006). An efficient grafting system for transgenic plant recovery in cotton (Gossypium hirsutum l.). *Plant Cell Tissue Organ Cult.* 85 (2), 181–185. doi: 10.1007/s11240-005-9068-9

Ken-Ichi Nonomura, K. M., Eiguchi, M., Suzuki, T., Miyao, A., Hirochika, H., and Kurata, N. (2003). The MSP1 gene is necessary to restrict the number of cells entering into Male and female sporogenesis and to initiate anther wall formation in rice. *Plant Cell* 15 (8), 1728–1739. doi: 10.1105/tpc.012401

Kim, D., Landmead, B., and Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* 12 (4), 357–360. doi: 10.1038/Nmeth.3317

Kovaleva, L. V., Voronkov, A. S., Zakharova, E. V., and Andreev, I. M. (2018). ABA and IAA control microsporogenesis in petunia hybrida l. *Protoplasma* 255 (3), 751–759. doi: 10.1007/s00709-017-1185-x

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30 (7), 923–930. doi: 10.1093/bioinformatics/btt656

Liu, H., Ding, Y., Zhou, Y., Jin, W., Xie, K., and Chen, L. L. (2017). CRISPR-p 2.0: An improved CRISPR-Cas9 tool for genome editing in plants. *Mol. Plant* 10 (3), 530–532. doi: 10.1016/j.molp.2017.01.003

Liu, Q., Wang, C., Jiao, X., Zhang, H., Song, L., Li, Y., et al. (2019). Hi-TOM: A platform for high-throughput tracking of mutations induced by CRISPR/Cas systems. *Sci. China Life Sci.* 62 (1), 1–7. doi: 10.1007/s11427-018-9402-9

Liu, H., Wu, K., Yang, M., Zhou, X., and Zhao, Y. (2014). Variation of soluble sugar, starch and plant hormones contents in sesame dominant genic male sterile line during bud development. *Oil Crop Sci.* 36 (2), 175–180. doi: 10.7505/j.issn.1007-9084.2014.02.006

Li, J., Wang, M., Li, Y., Zhang, Q., Lindsey, K., Daniell, H., et al. (2019). Multi-omics analyses reveal epigenomics basis for cotton somatic embryogenesis through successive regeneration acclimation process. *Plant Biotechnol. J.* 17 (2), 435–450. doi: 10.1111/pbi.12988

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 (12), 1–21. doi: 10.1186/s13059-014-0550-8

Ma, K., Han, J., Hao, Y., Yang, Z., Chen, J., Liu, Y. G., et al. (2019). An effective strategy to establish a male sterility mutant mini-library by CRISPR/Cas9-mediated knockout of anther-specific genes in rice. *J. Genet. Genomics* 46 (5), 273–275. doi: 10.1016/j.jgg.2019.03.005

Ma, D., Hu, Y., Yang, C., Liu, B., Fang, L., Wan, Q., et al. (2016). Genetic basis for glandular trichome formation in cotton. *Nat. Commun.* 7, 10456. doi: 10.1038/ncomms10456

Ma, J., Wei, H., Liu, J., Song, M., Pang, C., Wang, L., et al. (2013). Selection and characterization of a novel photoperiod-sensitive male sterile line in upland cotton. *J. Integr. Plant Biol.* 55 (7), 608–618. doi: 10.1111/jipb.12067

Ma, H., Wu, Y., Lv, R., Chi, H., Zhao, Y., Li, Y., et al. (2022). Cytochrome P450 mono-oxygenase CYP703A2 plays a central role in sporopollenin formation and ms5ms6 fertility in cotton. *J. Integr. Plant Biol.* 64, 2009–2025. doi: 10.1111/jipb.13340

Miao, Y. H., Xu, L., He, X., Zhang, L., Shaban, M., Zhang, X. L., et al. (2019). Suppression of tryptophan synthase activates cotton immunity by triggering cell death *via* promoting SA synthesis. *Plant J.* 98 (2), 329–345. doi: 10.1111/tpj.14222

Min, L., Li, Y., Hu, Q., Zhu, L., Gao, W., Wu, Y., et al. (2014). Sugar and auxin signaling pathways respond to high-temperature stress during anther development as revealed by transcript profiling analysis in cotton. *Plant Physiol.* 164 (3), 1293–1308. doi: 10.1104/pp.113.232314

Min, L., Zhu, L. F., Tu, L. L., Deng, F. L., Yuan, D. J., and Zhang, X. L. (2013). Cotton GhCKI disrupts normal male reproduction by delaying tapetum programmed cell death via inactivating starch synthase. *Plant J.* 75 (5), 823–835. doi: 10.1111/tpj.12245

Murmu, J., Bush, M. J., DeLong, C., Li, S., Xu, M., Khan, M., et al. (2010). Arabidopsis basic leucine-zipper transcription factors TGA9 and TGA10 interact with floral glutaredoxins ROXY1 and ROXY2 and are redundantly required for anther development. *Plant Physiol.* 154 (3), 1492–1504. doi: 10.1104/pp.110.159111

Okada, A., Arndell, T., Borisjuk, N., Sharma, N., Watson-Haigh, N. S., Tucker, E. J., et al. (2019). CRISPR/Cas9-mediated knockout of Ms1 enables the rapid generation of male-sterile hexaploid wheat lines for use in hybrid seed production. *Plant Biotechnol. J.* 17 (10), 1905–1913. doi: 10.1111/pbi.13106

Ramadan, M., Alariqi, M., Ma, Y., Li, Y., Liu, Z., Zhang, R., et al. (2021). Efficient CRISPR/Cas9 mediated pooled-sgRNAs assembly accelerates targeting multiple genes related to male sterility in cotton. *Plant Methods* 17 (1), 16. doi: 10.1186/s13007-021-00712-x

Staswick, P. E., Serban, B., Rowe, M., Tiryaki, I., Maldonado, M., Maldonado, M. C., et al. (2005). Characterization of an arabidopsis enzyme family that conjugates amino acids to indole-3-Acetic acid. *Plant Cell* 17 (2), 616–627. doi: 10.1105/tpc.104.026690

Su, Q., Yang, J., Fu, Q. Y., Jia, F. Y., Li, S. P., Li, Y., et al. (2019). Profiling of indole metabolic pathway in thermo-sensitive bainong male sterile line in wheat (Triticum aestivum l.). *Physiol. Mol. Biol. Plants* 25 (1), 263–275. doi: 10.1007/s12298-018-0626-0

Tang, L. K., Chu, H., Yip, W. K., Yeung, E. C., and Lo, C. (2009). An anther-specific dihydroflavonol 4-reductase-like gene (DRL1) is essential for male fertility in arabidopsis. *New Phytol.* 181 (3), 576–587. doi: 10.1111/j.1469-8137.2008.02692.x

Wang, M., Tu, L., Yuan, D., Shen, C., Li, J., Liu, F., et al. (2019). Reference genome sequences of two cultivated allotetraploid cottons, gossypium hirsutum and gossypium barbadense. *Nat. Genet.* 51 (2), 224–229. doi: 10.1038/s41588-018-0282-x

Wang, P., Zhang, J., Sun, L., Ma, Y., and Zhang, X. (2018). High efficient multisites genome editing in allotetraploid cotton (Gossypium hirsutum) using CRISPR/Cas9 system. *Plant Biotechnol. J.* 16 (1), 137–150. doi: 10.1111/pbi.12755

Wilson, Z. A., and Zhang, D. B. (2009). From arabidopsis to rice: pathways in pollen development. *J. Exp. Bot.* 60 (5), 1479–1492. doi: 10.1093/jxb/erp095

Xie, K., Minkenberg, B., and Yang, Y. (2015). Boosting CRISPR/Cas9 multiplex editing capability with the endogenous tRNA-processing system. *Proc. Natl. Acad. Sci. U.S.A.* 112 (11), 3570–3575. doi: 10.1073/pnas.1420294112

Xie, S., Shen, B., Zhang, C., Huang, X., and Zhang, Y. (2014). sgRNAcas9: A software package for designing CRISPR sgRNA and evaluating potential off-target cleavage sites. *PloS One* 9 (6), e100448. doi: 10.1371/journal.pone.0100448

Yang, L., Wu, Y., Zhang, M., Zhang, J., Stewart, J. M., Xing, C., et al. (2018). Transcriptome, cytological and biochemical analysis of cytoplasmic male sterility and maintainer line in CMS-D8 cotton. *Plant Mol. Biol.* 97 (6), 537–551. doi: 10.1007/s11103-018-0757-2

Yu, G. C., Wang, L. G., Han, Y. Y., and He, Q. Y. (2012). clusterProfiler: an r package for comparing biological themes among gene clusters. *OMICS* 16 (5), 284–287. doi: 10.1089/omi.2011.0118

Zhang, R., Zhou, L. L., Li, Y. L., Ma, H. H., Li, Y. W., Ma, Y. Z., et al. (2022). Rapid identification of pollen- and anther-specific genes in response to high-temperature stress based on transcriptome profiling analysis in cotton. *Int. J. Mol. Sci.* 23 (6), 3378. doi: 10.3390/ijms23063378

Zhao, D. Z., Wang, G. F., Speal, B., and Ma, H. (2002). The excess microsporocytes1 gene encodes a putative leucine-rich repeat receptor protein kinase that controls somatic and reproductive cell fates in the arabidopsis anther. *Genes Dev.* 16 (15), 2021–2031. doi: 10.1101/gad.997902

Zheng, H., Wang, R., Jiang, Q., Zhang, D., Mu, R., Xu, Y., et al. (2021). Identification and functional analysis of a pollen fertility-associated gene GhGLP4 of gossypium hirsutum l. *Theor. Appl. Genet.* 134 (10), 3237–3247. doi: 10.1007/s00122-021-03888-x

Zhou, H., He, M., Li, J., Chen, L., Huang, Z., Zheng, S., et al. (2016). Development of commercial thermo-sensitive genic Male sterile rice accelerates hybrid rice breeding using the CRISPR/Cas9-mediated TMS5 editing system. *Sci. Rep.* 6, 37395. doi: 10.1038/srep37395

Zhu, W., Liu, K., and Wang, X.-D. (2008). Heterosis in yield, fiber quality, and photosynthesis of okra leaf oriented hybrid cotton (Gossypium hirsutum l.). *Euphytica* 164 (1), 283–291. doi: 10.1007/s10681-008-9732-3

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 30 January 2023
DOI 10.3389/fpls.2023.1117156

OPEN ACCESS

EDITED BY
Zhichao Wu,
National Institutes of Health (NIH),
United States

REVIEWED BY
Mengyao Li,
Sichuan Agricultural University, China
Chunying Kang,
Huazhong Agricultural University, China

*CORRESPONDENCE
Gui-Hua Jiang
jgh2004267@sina.com
Jocelyn K. C. Rose
jr286@cornell.edu
Kun-Song Chen
akun@zju.edu.cn

†These authors have contributed equally to this work

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

RECEIVED 06 December 2022
ACCEPTED 06 January 2023
PUBLISHED 30 January 2023

CITATION
Li B-J, Shi Y-N, Jia H-R, Yang X-F, Sun Y-F, Lu J, Giovannoni JJ, Jiang G-H, Rose JKC and Chen K-S (2023) Abscisic acid mediated strawberry receptacle ripening involves the interplay of multiple phytohormone signaling networks. *Front. Plant Sci.* 14:1117156. doi: 10.3389/fpls.2023.1117156

Abscisic acid mediated strawberry receptacle ripening involves the interplay of multiple phytohormone signaling networks

Bai-Jun Li[1,2,3†], Yan-Na Shi[1,2,3†], Hao-Ran Jia[1], Xiao-Fang Yang[4], Yun-Fan Sun[1], Jiao Lu[1], James J. Giovannoni[5,6], Gui-Hua Jiang[4*], Jocelyn K. C. Rose[5*] and Kun-Song Chen[1,2,3*]

[1]College of Agriculture and Biotechnology, Zhejiang University, Zijingang Campus, Hangzhou, China, [2]Zhejiang Provincial Key Laboratory of Horticultural Plant Integrative Biology, Zhejiang University, Zijingang Campus, Hangzhou, China, [3]State Agriculture Ministry Laboratory of Horticultural Plant Growth, Development and Quality Improvement, Zhejiang University, Hangzhou, China, [4]Institute of Horticulture, Zhejiang Academy of Agricultural Sciences, Hangzhou, China, [5]Plant Biology Section, School of Integrative Plant Science, Cornell University, Ithaca, NY, United States, [6]United States Department of Agriculture – Agricultural Research Service and Boyce Thompson Institute for Plant Research, Cornell University, Ithaca, NY, United States

As a canonical non-climacteric fruit, strawberry (*Fragaria* spp.) ripening is mainly mediated by abscisic acid (ABA), which involves multiple other phytohormone signalings. Many details of these complex associations are not well understood. We present an coexpression network, involving ABA and other phytohormone signalings, based on weighted gene coexpression network analysis of spatiotemporally resolved transcriptome data and phenotypic changes of strawberry receptacles during development and following various treatments. This coexpression network consists of 18,998 transcripts and includes transcripts related to phytohormone signaling pathways, MADS and NAC family transcription factors and biosynthetic pathways associated with fruit quality. Members of eight phytohormone signaling pathways are predicted to participate in ripening and fruit quality attributes mediated by ABA, of which 43 transcripts were screened to consist of the hub phytohormone signalings. In addition to using several genes reported from previous studies to verify the reliability and accuracy of this network, we explored the role of two hub signalings, small auxin up-regulated RNA 1 and 2 in receptacle ripening mediated by ABA, which are also predicted to contribute to fruit quality. These results and publicly accessible datasets provide a valuable resource to elucidate ripening and quality formation mediated by ABA and involves multiple other phytohormone signalings in strawberry receptacle and serve as a model for other non-climacteric fruits.

KEYWORDS

strawberry, phytohormone signalings, abscisic acid, coexpression network, fruit qualities, ripening

Introduction

Fleshy fruits can be classified into those that exhibit either climacteric or non-climacteric ripening: the former type involves a peak of respiration and emission of the gaseous hormone ethylene, which acts as the main regulator of this process, while the latter does not. The phytohormone abscisic acid (ABA) can play either a dominant or supportive role in modulating non-climacteric and climacteric fruit ripening, respectively (Kou et al., 2021a). Although many studies have reported that the ABA controls non-climacteric fruit ripening and influence fruit quality traits (Kou et al., 2021b; Li et al., 2022a), there is limited understanding of this process compared with that of ethylene in climacteric fruit (Wang et al., 2022a). Better elucidation of mechanisms of ABA-mediated fruit ripening has considerable potential for enhancing our understanding of both climacteric and non-climacteric ripening and for developing novel traits and varieties, especially concerning non-climacteric fruit.

Strawberry (*Fragaria* spp.) fruit is a pseudocarp that consists of a receptacle with many achenes (true fruit) embedded in the epidermis. It has typical characteristics of non-climacteric fruit and modern cultivated strawberry (*Fragaria* × *ananasssa*) represents a particularly important fruit crop (FAO, 2020). Strawberry has also been adopted as an experimental model for non-climacteric fruit, which is reflected in the development of effective transgenic systems and ever-growing genomic resources (Edger et al., 2019; Zhou et al., 2020; Kou et al., 2021b; Liu et al., 2021). Through the development of such resources, genes that affect strawberry fruit quality, including coloration (i.e. anthocyanin biosynthesis) (Fischer et al., 2014; Castillejo et al., 2020; Gao et al., 2020), sugar accumulation (Jia et al., 2013a; Jia et al., 2016), aroma (Raab et al., 2006; Medina-Puche et al., 2015; Molina-Hidalgo et al., 2017) production and softening (Molina-Hidalgo et al., 2013; Paniagua et al., 2016), have been identified. Moreover, members of transcription factor (TF) families have been revealed as inducers or suppressors of strawberry fruit ripening (Li et al., 2022a), including the MADS genes *SHATTERPROOF*-like (*FaSHP*; Daminato et al., 2013), *FaMADS1a* (Lu et al., 2018), *FaMADS9* (Vallarino et al., 2020), and *FveSEP3* (Pi et al., 2021), and the *NAC*, *Ripening Inducing Factor* (*FaRIF*; Martín-Pizarro et al., 2021). The expression levels of most of these genes are affected by ABA, and the ABA biosynthetic pathway in strawberry fruit has been also well described (Li et al., 2022a).

Notably, most studies investigating the roles of phytohormones in fruit development have used exogenous hormone treatments. In strawberry, auxin production, which supports the development of the receptacle, and which is antagonistic to ABA (Li et al., 2022a), occurs in the achenes (Thompson, 1969). Accordingly, removing achenes from the receptacle causes reduced auxin levels and, consequently, an elevation in ABA levels and a promotion of receptacle ripening in the late developmental stage (Li et al., 2022b). This experimental manipulation therefore provides a means to study ABA-associated receptacle ripening, in addition to the use of exogenous ABA treatments. In summary, strawberry provides an excellent model system in which to characterize the ABA-mediated fruit ripening of non-climacteric fruit, analogous to the adoption of tomato (*Solanum lycopersicum*.), as principal model for ethylene-regulated climacteric ripening (Liu et al., 2020; Fenn & Giovannoni, 2021; Kou et al., 2021b).

Multiple phytohormone signaling genes participating in strawberry ripening regulated by ABA have been documented (Gu et al., 2019; Fenn & Giovannoni, 2021; Wang et al., 2022a), and ABA can act synergistically or antagonistically with auxin, gibberellins (GAs), ethylene, and jasmonic acids (JAs) in strawberry (Li et al., 2022a). In addition, the roles of ABA signaling genes in ripening, including *FaPYR1* (*Pyrabactin resistance 1*; Chai et al., 2011), *FaABI1* (*ABSCISIC ACID-INSENSITIVE 1* encoding a PP2C protein; Jia et al., 2013b), *FaSnRK2.6* (*SNF1-related protein kinase 2.6*; Han et al., 2015), and *FaABAR* (*Magnesium-protoporphyrin IX chelatase H subunit*; Jia et al., 2011), have been well characterized. These results are consistent with the existence of complicated strawberry ripening mechanisms, involving multiple phytohormones signalings, and disproportionately influenced by a predominant ABA-signaling pathway. However, there are much remains to be learnt about the hub phytohormone signalings in strawberry and additional non-climacteric fruit mediated by ABA.

In this study, we investigated phytohormone signaling pathways in the strawberry receptacle ripening mediated by ABA, using transcriptome profiling of the receptacle at three developmental stages from unripe to ripe and following changes of ABA levels *via* exogenous and removing achenes treatments. Following weighted gene coexpression network analysis (WGCNA), we described a coexpression network and predicted the hub phytohormone signaling genes in ABA-mediated receptacle ripening. Additionally, we identified two hub signaling genes, *small auxin-up RNA*s (*FaSAUR1* and *FaSAUR2*), shown by transient RNA interference (RNAi), to promote receptacle quality formation. Finally, the full-length transcript sequences and their spatiotemporally resolved expressional profiles provide new insights into ABA mediating associated phytohormone signalings in ripening strawberry fruit.

Materials and methods

Plant materials and sampling

Fragaria × *ananassa* 'Yuexin' fruit were sampled at the green (G, green receptacle embedded green achenes), turning (T, pale green receptacle embedded with some brown or green achenes) and half red (HR, half a receptacle with some red and brown achenes) stages (Supplementary Figure S1), achenes were removed as described below from a subset of fruit, and then the samples were immediately frozen in liquid nitrogen and stored at –80°C. All 'Yuexin' fruit used in this study were grown in a Zhejiang Academy of Agricultural Sciences plastic greenhouse (Zhejiang, China) under natural light with daytime and night-time temperatures of 10-24°C.

Removal of achenes and exogenous hormone treatments

The fruit were carefully removed half achenes using a tweezer along the centra axis from the receptacles at G stage. Water (sterile ultrapure water as control for exogenous hormone treatments) or NAA (500 μM; Sigma-Aldrich, USA), and ABA (500 μM; Sigma-

Aldrich, USA) was then injected into the whole receptacles, using a 1 mm injection syringe. NAA and ABA were dissolved in sterile ultrapure water to a final concentration of 500 μM. Each of exogenous hormone or water treatment had three biological replicates.

Determination of ABA content

Each freeze-dried sample (0.1-0.2 mg) was placed in a 2 ml centrifuge tube, 1 ml of acetonitrile (Sigma-Aldrich, USA) containing 0.1% formic acid was added and the sample was incubated for12 h at 4°C. The centrifuge tube was then ultrasonicated in an ice water bath for 10 min and centrifuged at 13,500 g and 4°C for 10 min. The supernatant was then collected and filtered through a 0.22 μm membrane filter (organic phase) into a 1.5 ml centrifuge tube. The supernatant was dried in a stream of nitrogen gas. The dried sample was then redissolved in 50 μl acetonitrile, ultrasonicated in an ice water for 5 min and centrifuged at 13,500 g and 4°C for 5 min, then 40 μl of the supernatant was transferred to a brown chromatography vial sample bottle and subjected to high performance liquid chromatography (HPLC, Waters e2695 Separations Module, Waters, USA). A SunFire C18 5μm, 4.6 × 250 mm, column (Waters, USA) was selected and the following program use: sample introduction for 10 μl, sample temperature at 8°C, and maximum and minimum psi of 4,000 and 0, respectively. The mobile phase consisted of 0.1% formic acid (Phase A) and acetonitrile containing 0.1% formic acid (Phase B) and the program was A: B (95: 5) at 0 min; A: B (80: 20) at 20 min; A: B (35: 65) at 40 min; A: B (0: 100) at 40.5; A: B (0: 100) at 45.5 min; A: B (5: 95) at 46.5; A: B (5: 95) at 52 min. An ABA standard (Sigma-Aldrich, USA) was used for identification and quantification.

Measurement of qualities related to receptacle ripening

The furanone of the receptacle was extracted and measured using a gas chromatography mass spectrometry (GC/MS, Agilent 7890A GC System, Agilent Technologies Inc., MA, USA) as previously described (Zhang et al., 2018). Sugars were extracted and estimated as in our previously study (Li et al., 2022c) and anthocyanins were extracted and measured using an ultraviolet spectrophotometer (UV-2600, SHIMADZU, Japan) as previously described (Wang et al., 2022b).

Total RNA extraction and quality assessment

All freeze-dried samples were powdered in liquid nitrogen and 50 mg samples used to extract total RNA using a CTAB method (Shan et al., 2008). The purity and concentration of RNA samples were measured using a NanoDrop™ One/OneC system (Thermo Fisher Scientific, MA, USA), and the integrity of each RNA samples were estimated using an Agilent 2100 Bioanalyzer (Agilent Technologies Inc., CA, USA) and agarose gel electrophoresis.

PacBio *Iso-Seq* library preparation, sequencing and data analysis

To obtain full-length transcriptome sequences expressed during receptacle development, the total RNA of the basal and apical of the receptacle from G, T, HR stages were fully mixed in equal quantity to construct PacBio *Iso-Seq* libraries, and sequenced using a PacBio Sequel2 platform. The library preparation and sequencing were performed as previously described (Li et al., 2020).

The SMRT Link v8.0.0 pipeline (Gordon et al., 2015) was used to generate unique full-length transcriptome sequences (isoforms) from the raw sequence data. Briefly, the circular consensus sequence (CCS) reads were first extracted from the BAM file and the full-length (FL) reads (i.e., the CCS reads containing 5' primer, 3'primer and poly A structures) were then obtained from CCS reads. Second, the primers, barcodes, poly A tail trimming and concatemer of full passes were removed from the FL reads to obtain full-length non-chimeric (FLNC) reads. Subsequently, the FLNC reads were clustered hierarchically using Minimap2 (Li, 2021) to generate the consensus FLNC reads. Third, the Quiver algorithm (Gordon et al., 2015) was used for further correcting the consensus FLNC reads to obtain the high-quality consensus FLNC reads. Finally, CD-HIT-v4.6.7 (Li & Godzik, 2006) with a threshold of 0.99 identity was used to eliminate redundancy from high-quality consensus FLNC reads to obtain isoforms.

The isoforms were annotated by BLAST searches of the nonredundant protein (NR) database (http://www.ncbi.nlm.nih.gov), the Swiss-Prot protein database (http://www.expasy.ch/sprot), the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (http://www.genome.jp/kegg), COG/KOG database (http://www.ncbi.nlm.nih.gov/COG) with an E-value threshold of 1e−5. We used ANGEL (Shimizu et al., 2006) to predict the coding sequences (CDSs), protein sequences, and UTR sequences of the isoforms.

Illumina transcriptome (*RNA-Seq*) library preparation, sequencing and expression level estimation

The total RNA of the basal and apical of the receptacle from G, T, HR stages, and the achened and de-achened sides of the receptacles with the treatments after 9 and 12 days were used to generate RNA-Seq data using an Illumina HiSeq™ 4000, The RNA-Seq library construction, sequencing, and the clean data acquisition from the raw data were performed as previously described (Li et al., 2020). The clean data from each sample were mapped into the isoform set to estimate expression levels *via* FPKM of isoforms using RSEM (Li & Dewey, 2011).

Weighted gene coexpression network analysis

The WGCNA was constructed using WGCNA (v1.47) package in R (Langfelder & Horvath, 2008). The expression values of isoforms (FPKM ≥ 7) were used to establish weighted gene coexpression

modules under the automatic network construction function blockwiseModules with default settings, and the power was 5; the TOMType was unsigned; the mergeCutHeigh was 0.9; the isoform number of minModuleSize was 50. Finally, the isoforms were clustered into 21 modules. The correlation between modules and traits were estimated using the Pearson's correlation analysis (http://omicshare.com/tools/) between values of module eigengene and phenotypic data, which was displayed using a heatmap analysis. Additionally, the Pearson's correlation analysis between GS (Gene significance value, a Pearson's correlation between expression level of each gene in a module and phenotypic data) and MM (Module membership, a Pearson's correlation between expression level of each gene in a module and the values of module eigengene) was used to further identify the most relevant modules associated with receptacle ripening. The coexpression network of each module was visualized using Cytoscape v3.3.0 (Shannon et al., 2003).

Phylogenetic and heatmap analyses

To identify the phylogenetic relationships between FaSAUR and SAUR proteins from other plants, a Neighbor-joining tree was constructed with bootstrap values evaluated from 1,000 replicate runs using MEGA7 (Kumar et al., 2016). The alignment of the amino acid sequences was performed using Clustal W (Larkin et al., 2007). All heatmap analyses in this study were conducted using Omiscshare tools (http://omicshare.com/tools/).

Transient silencing of *FaSAUR*s by *Agrobacterium* infiltration

To verify the function of *FaSAUR1* and *FaSAUR2*, we used RNA interference (RNAi) to silence these two genes in 'Yuexin' receptacles, using the RNAi methodology as previously described (Shi et al., 2021) with some modifications. Briefly, the partial fragments of *FaSAUR1* and *FaSAUR2* (Supplementary Figure S5) were ligated into pHELLSGATE 2 vector (Shi et al., 2021) using BP Clonase (Invitrogen, MA, USA) to construct *35Spro : FaSAUR1*-RNAi and *35Spro : FaSAUR2*-RNAi, respectively. The recombinant plasmids were transformed into *Agrobacterium tumefaciens* GV3101 by electroporation[46]. After incubation, the GV3101 suspension containing the RNAi vector solution were centrifuged and the cell pellets were resuspended in liquid infection medium (sterile ultrapure water containing 10 mM 2-morpholinoethanesulphonic acid, 10 mM $MgCl_2$, and 150 μM acetosyringone) to an OD_{600} = 1 (Zhang et al., 2020). The final suspensions were injected into the whole receptacle at the T stage using a syringe. The infected fruits were then cultivated in the greenhouse, photographed and sampled. The primers of RNAi fragments of *FaSAUR1* and *FaSAUR2* were designed based on transcriptome data (Supplementary Table S1).

Statistical analysis

Statistical significance was assessed with Student's paired *t*-test using Omiscshare tools (http://omicshare.com/tools/).

Results

Evaluation of ABA levels and quality metrics in the receptacle during development and following removing achenes with exogenous ABA treatments

Using *F.* × *ananasssa* 'Yuexin' as a model allo-octoploid strawberry cultivar, receptacle ripening progressed from the apical to the basal region based on progression of color change (Supplementary Figure S1). Levels of ABA and additional ripening-related compounds increased in the whole receptacle during development, and were notably higher in the apical as compared to the basal region of the fruit (Supplementary Figure S1). This was especially clear for ABA at the T and HR stages, while there was a similar difference in the other three ripening-related compounds, including anthocyanin, HDMF, and sugars, only at HR, consistent with ABA acting as a ripening promoter of receptacle tissue (Supplementary Figure S1).

To confirm and deeply explore ABA as a dominant role in strawberry receptacle ripening, we removed half of the achenes from the receptacle at the G stage and injected water, ABA, or the synthetic auxin naphthylacetic acid (NAA) due to auxin as a repressor for ABA biosynthesis[24], into the receptacle. After injecting water or ABA, the pigmentation on the side of receptacle from which the achenes had been removed ('de-achened side') developed more rapidly than the side with achenes ('achened side'), while there was no visible difference between the two sides following the NAA treatment (Figure 1A), consistent with achene-derived auxin inhibiting ripening. On day 9 after treatment, the de-achened side of receptacle was more pigmented after the ABA treatment than after the water treatment, and at day 12 the achened side was fully red after ABA treated fruit, but only half-red after the water treatment (Figure 1A). These results support ABA as a ripening promoter of receptacle tissue.

The receptacles of the treated fruit were divided into four parts, corresponding to the basal and apical parts of de-achened (DA_BA and DA_AP) and achened sides (Achened_BA and Achened_AP), and ABA levels were measured. ABA contents in the de-achened side, including basal and apical parts, were higher than those in the achened side after treatments, which suggested that removing the achenes promoted ABA biosynthesis (Figure 1B). Moreover, ABA levels in the basal parts of the de-achened and achened sides were lower than those in the apical parts. The receptacle with NAA treatment had the lowest ABA contents among the various treatments and fruit regions, again suggesting that auxin produced by the achenes acts as a repressor of ABA, and consistent with the non-coloration phenotype (Figure 1A).

The levels of ripening-related compounds in the de-achened and achened sides were quantified to investigate the degree of ripeness. Anthocyanin accumulated in the de-achened side of the receptacle following water and ABA treatments at day 9, and ABA treatment resulted in the highest levels (Figure 1C). At day 12, anthocyanin levels were similar in the de-achened tissues under ABA and water treatments, while in the achened side they were higher under ABA treatment than with water. The receptacle following NAA treatment

10.3389/fpls.2023.1117156

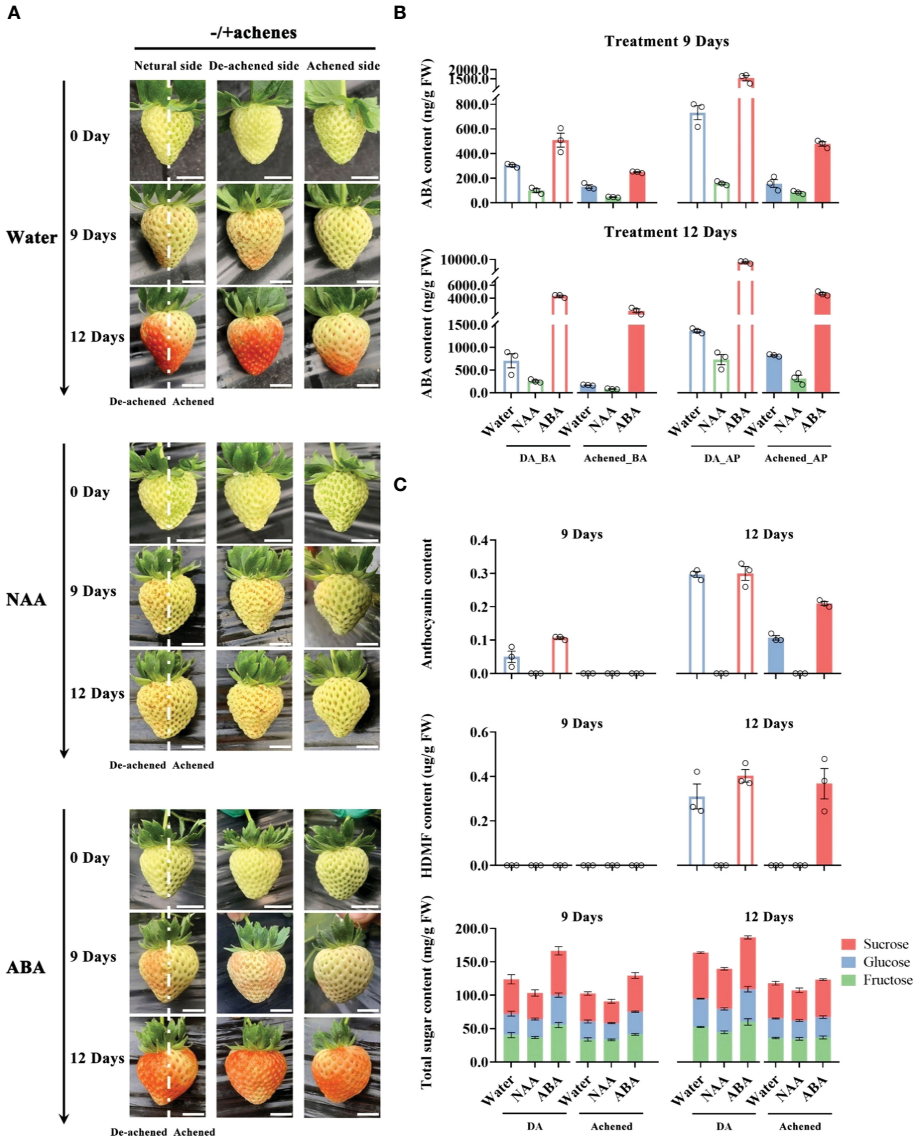


FIGURE 1
Levels of ABA, HDMF, sugars, and anthocyanins the receptacle after removing and exogenous hormone treatments. **(A)** Whole receptacle from which half the achenes were removed along the central axis at the G stage were injected with water, ABA (500 μM), or naphthylacetic acid (NAA, 500 μM), and photographed after 0, 9 and 12 days. Photographs are shown of individual fruit from three perspectives: 'neutral side', with the achenes removed from the left side of the dotted lines ('De-achened side'), or from the perspectives of the left or right sides of the dotted lines ('Achened side'), as indicated. Scale bars = 1 cm. **(B)** ABA levels in the different parts of the receptacles at 6 and 9 days after the treatments. The DA, Achened, BA, and AP in labels represent the de-achened and achened sides, basal and apical parts of receptacles, respectively. **(C)** Levels of ripening-related compounds in the DA and achened sides of the receptacles at 6 and 9 days after the exogenous hormone treatments. DA and Achened indicate the de-achened and achened sides of the receptacles, respectively. The data values are the mean ± SD of three biological replicates.

showed no evidence of anthocyanin accumulation, consistent with the lack of coloration (Figure 1A). The important strawberry aroma compound, HDMF, was detected in both de-achened and achened tissues of the receptacles under ABA treatment and in the de-achened side under water treatment at 12 days (Figure 1C). In addition, the total sugar content of the de-achened sides were higher than in the achened sides, and samples from the ABA and NAA treatments had the highest and lowest contents, respectively, among the different treatments (Figure 1C). Together, these results demonstrate a positive correlation between the measured ripening-related compounds (Figure 1C) and ABA accumulation (Figure 1B).

Transcriptome sequencing and analyses of the receptacles during development and following hormone treatments

The genome of strawberry (*F.* × *ananassa* 'Camarosa') comprises four parental subgenomes (Edger et al., 2019), which complicates calculating expression levels and profiling gene expression using RNA-Seq. To obtain the full-length transcript sequences (Isoforms) expressed during receptacle development, the total mRNAs extracted from basal and apical parts of receptacles at the G, T, HR stages were pooled and sequenced

using the PacBio platform (Supplementary Figure S2). This resulted in the identification of 52,455 transcript isoforms. The different parts of the receptacles during development and under the treatments (a total of 54 samples), were subjected to RNA-Seq using an Illumina platform, and corresponding gene expression profiles were generated (Supplementary Figure S2). Approximately 375.50 gigabytes (GBs) of cleaned sequence data were produced and mapped to the isoform set, with a high mapping ratio (97.2–97.5%). Thus, almost all genes expressing during receptacle development were identified. The gene expressional profiles consisted of three categories: spatial expression, temporal expression during development, or related to changes in ABA levels manipulated by removing achene and exogenous treatments, and these variable and complicated expression patterns can deeply explore potential relationship between phenotypes and gene expressions, and between gene expressions. And then, the isoforms were clustered into 21 modules according to their fragments per kilobase per million (FPKM) using WGCNA (Supplementary Figure S2). Among these modules, the turquoise module had the highest number of isoforms (7,246), while the lowest number was in the grey module (4) (Supplementary Figure S2).

The above data sets were used to identify the modules that had a high correlation with receptacle ripening and these were used to describe the hub phytohormone signaling network regulated by ABA. The expression patterns of the brown, tan, and red modules had a positive relationship with the changes in ABA accumulation, ripening-related compounds qualities, and phenotypes, while the blue, turquoise, and yellow modules were opposite to them (Figures 1, 2A, B; Supplementary Figure S1). Moreover, a Pearson correlation analysis further verified that the changes in the physiological indices displayed a significantly positive relationship with the expressional profiles of brown, red, and tan modules and negative with blue, turquoise, and yellow (Figure 2C). Subsequently, the Pearson correlations between the expression profiles of genes in each module and physiological indices indicated that these six modules ranked in the top seven of all of the modules (Figure 2D). In addition, an analysis of the correlation between gene significance (i.e. the correlation between each isoform and ABA level), and module membership (i.e. the correlation between expressional profile of each isoform and module) showed that the brown, tan, blue, turquoise, and yellow modules had correlation coefficient values > 0.6, with a significance that was higher than others. This suggested that the expression of genes in these modules had a strong relationship with the changes of ABA levels (Figure 2E; Supplementary Figure S3). Together, these results indicated that the brown and tan modules were positively related to ABA-mediated receptacle ripening, while the blue turquoise, and yellow modules had a highly negative relationship, which suggested that the genes in these modules might participate in receptacle ripening.

Construction of coexpression networks of phytohormone signaling during receptacle ripening mediated by ABA

Based on above results, the red and tan modules had the most positive relationship with receptacle ripening mediated by ABA, while the blue, turquoise, and yellow modules had a negative relationship. The isoforms involved in phytohormone signaling, ripening, and data related to levels of fruit quality related compounds were used to construct coexpression networks related to phytohormones that collectively regulate receptacle ripening. The brown module contained genes associated with six phytohormone: ABA, ethylene, GA, JA, auxin, and brassinosteroids (BR). These genes showed a positive relationship with pigmentation, cell wall metabolism, and sugar accumulation, suggesting that these phytohormone signaling networks may underly the expression of genes that affect commercially important fruit quality traits (Figure 3A). We identified sets of ABA (6), auxin (13), ethylene (3), and JA (6) related genes that putatively promote ripening and quality, and equivalent sets plus GA and BR (13, 16, 13, 3, 16, 1 respectively) that putatively suppress ripening and associated traits (Figure 3A; Table 1). Genes in the anthocyanin biosynthesis pathway and the associated TF regulator, MYB10, have been well studied in strawberry fruit, and their expression is upregulated by ABA (Kadomura-Ishikawa et al., 2015). We detected the expression of several anthocyanin biosynthetic genes and multiple MYB10 genes in the brown module and their expression were up-regulated by ABA, indicating a relationship between the module and ripening and providing validation of the reliability and accuracy of the coexpression network (Figure 3A; Supplementary Figure S4; Supplementary Dataset S1).

We also determined that the expression of genes, including *sucrose-phosphate synthase* (*SPS*) and *beta-fructofuranosidase/invertase* (*INV*), involved in sugar accumulation, was positively related to the brown module and changes in sugar levels in the receptacle (Figures 1C, 3A; Supplementary Figure S1). Additionally, we identified TF genes in the *MADS* and *NAC* families that have been widely associated with ripening (Kou et al., 2021a; Kou et al., 2021b). Specifically, we observed that the expression of two *NAC* genes and one *MADS* gene had a negative relationship with the brown module, while 9 *NAC* genes and 8 *MADS* genes showed a positive relationship (Figure 3A). Among these MADS genes, Isoform0048909 has a positive relationship with the brown module and are down-regulated by ABA (Supplementary Figure S4). Notably, its predicted amino acid sequence corresponds to FaSHP, which participates in receptacle ripening mediated by ABA (Daminato et al., 2013) (Supplementary Datasets S1, Dataset S2). In the tan module, which corresponded to a positive correlation with ripening regulated by ABA, one BR signaling gene and three ABA and two auxin signaling genes were negatively and positively correlated with ripening, respectively. These genes were coexpressed with several *NAC* genes and a *MADS* gene possibly participating in the metabolism of sugars and cell walls (Figure 3B).

Members of phytohormone signaling pathways were also coexpressed in other modules, including blue, turquoise, and yellow, negatively related to receptacle ripening mediated by ABA (Figure 4). The coexpression network showed that 40 and 185 members from eight phytohormone signaling pathways were positively and negatively correlated with ripening, respectively (Table 1). Among the phytohormones, only GA signaling did not show a positive relationship with receptacle ripening (Figure 4; Table 1). We also identified 22 *NAC* genes that were positively correlated with ripening, in addition to 3 *NAC* genes and 16 *MADS*

10.3389/fpls.2023.1117156

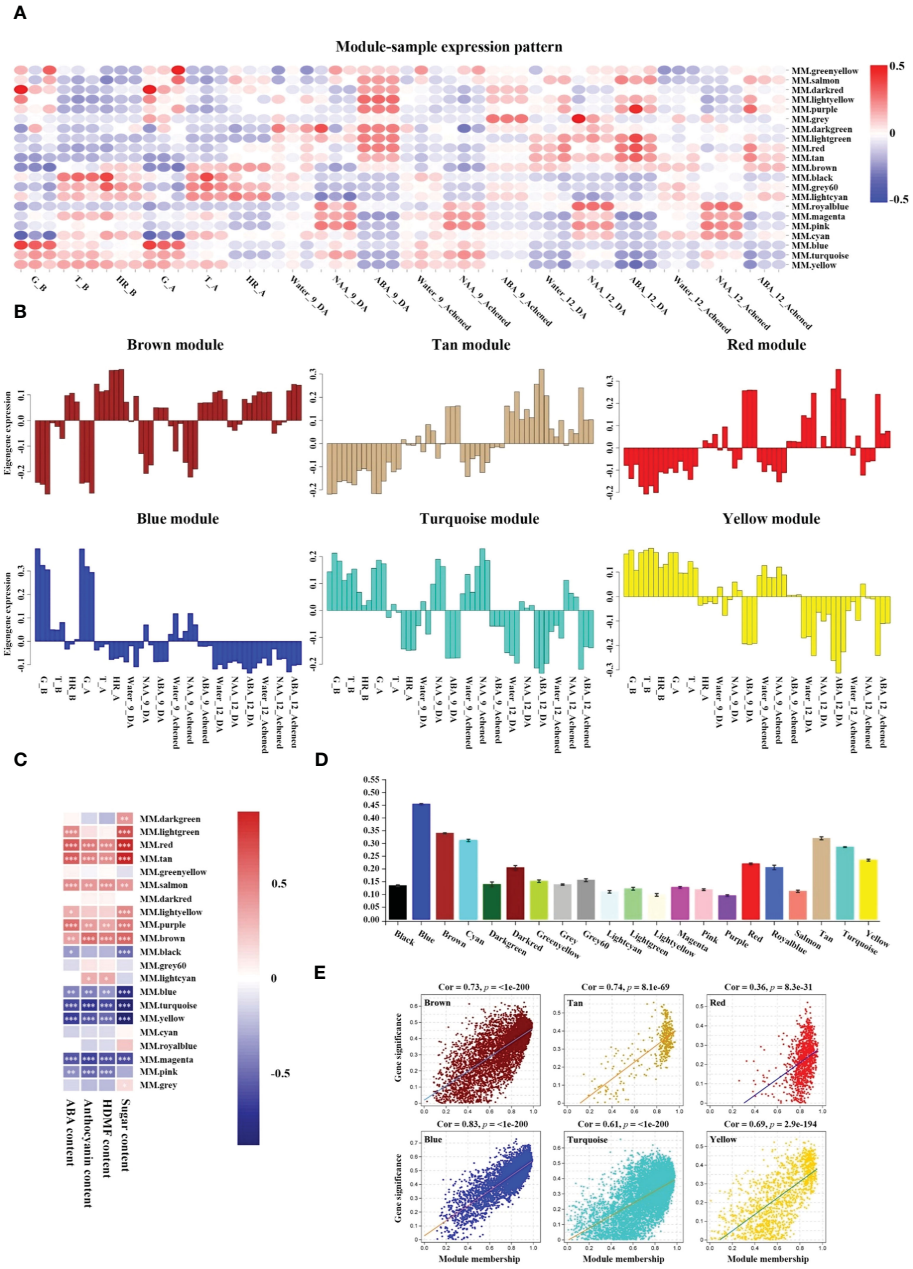


FIGURE 2
Screening positive and negative modules relevant to receptacle ripening. **(A)** Heatmap analysis of module expression pattern in samples based on gene expression profiles. The red and blue colors denote up- and down-regulation of gene expressions in the samples, respectively. **(B)** Histograms of expression patterns of six modules in each sample. **(C)** The Pearson correlation between expression profiles of genes in modules and ABA, anthocyanin, HDMF, and sugar contents, respectively. According to Student's paired t-test, white '*', '**', and '***' in the heatmap represent P < 0.05, P < 0.01, and P < 0.001, respectively. **(D)** The mean of the Pearson correlation between the expression profiles of genes in each module and physiological indices. **(E)** Analysis of correlation between gene significance, the correlation between each isoform and ABA level, and module membership, the correlation between expressional profile of each isoform and module. *P* values were analyzed using a Student's paired *t*-test.

genes with a negative relationship. Among these NACs, the predicted amino acid sequences of both Isoform0048861 and 0046736, which belong to the blue module, had ~98% identity to FaRIF, which has been shown to promote for strawberry ripening and is positively regulated by ABA (Martín-Pizarro et al., 2021). This is consistent with the expressional profiles of these two isoforms and the positive relationship between the them and receptacle ripening (Supplemetary Figure S4; Supplementary Datasets S1, Dataset S2). Additionally, the homolog of FaMADS1a, (Isoform0046787), which represses receptacle ripening and is negatively regulated by ABA at the transcriptional level (Lu et al., 2018), was present in the turquoise module and its expression was repressed by ABA (Supplementary Figure S4; Supplementary Dataset S2).

We also identified genes involved in coloration, sweetness, cell wall metabolism, and aroma that were coexpressed with these phytohormone signaling pathways (Figure 4). These included genes involved in sugar biosynthesis, such as *SPS*, *sucrose-6-phosphatase* (*SPP*) and *INV*, anthocyanin accumulation, HDMF formation,

10.3389/fpls.2023.1117156

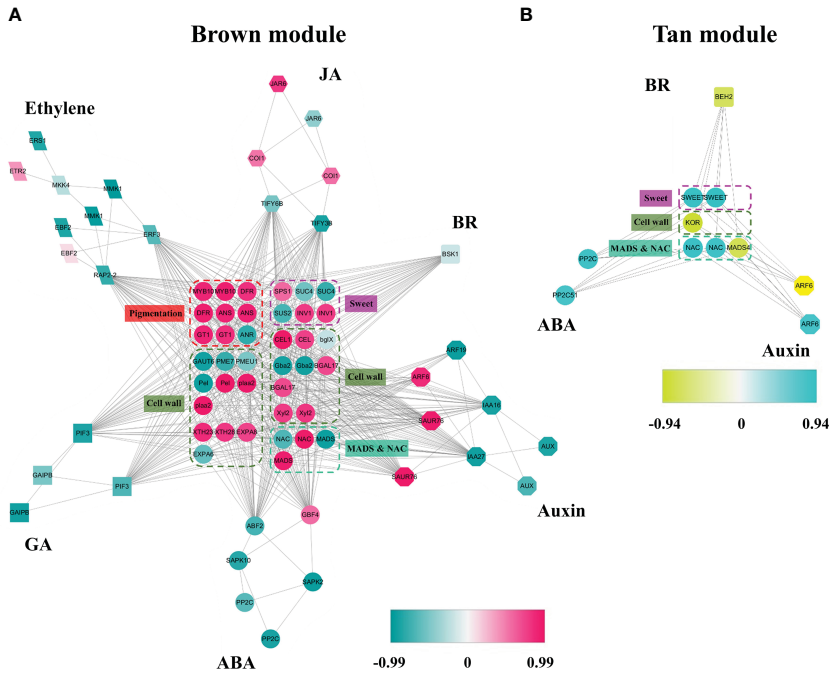


FIGURE 3
Coexpression networks of ABA and other phytohormone signaling pathways during receptacle ripening, in brown and red modules. **(A)** The coexpression network in the red module is positively related to ripening based on WGCNA. Each of shaped block present an isoform, and the colors denote a positively or negatively relationship with the module as indicated by the scale bar. If more than two isoforms of a gene were detected in the module, only the isoforms with highest and lowest levels of correlation with the module are shown. In the figure, the isoforms associated with phytohormone signaling revolve around the isoforms involved in quality traits, such as pigmentation, softening, and sweetness, as well as TFs including *MADS* and *NAC* genes related to ripening. **(B)** The coexpression network in the tan module positively related to ripening. JA, jasmonic acid; BR, brassinosteroid; GA, gibberellins. The full names of the abbreviations of the isoforms are shown in Supplementary Dataset S1.

including *quinone oxidoreductase* (*QR*), which had a negative correlation with the modules indicating a positive relationship with receptacle ripening and were up-regulated by ABA, consistent with observed phenotypic changes (Figures 1C, 4; Supplementary Figures S1, S4). In addition, Isoform0037314 and 0014205, which are positively related to turquoise module (Supplementary Datasets S1, S2) are homologs of FaSnRK2.6 and FaABI1, respectively, which are involved in ABA signaling and suppress receptacle ripening (Jia et al., 2013b; Han et al., 2015), and they were down-regulated by ABA (Supplementary Figure S4).

In summary, in this coexpression network, which consisted of five modules, were eight phytohormone signaling pathways. All GA signaling genes (25 isoforms), showed a negative association with receptacle ripening (Figures 3, 4). Moreover, ABA, auxin, ethylene, JA, BR, and GA signaling pathways had at least 25 isoforms in the coexpression networks, which was considerably more than the corresponding numbers for cytokinin (CTK) and salicylic acid (SA) signaling (Table 1). Among these phytohormone signalings, 43 isoforms from seven signaling pathways, including ABA, auxin, ethylene, GA, BR, SA, and CTK, showed high correlation with their modules (correlation coefficient values > |± 0.9|), which suggested that they acted as hub phytohormone signalings to the most potentially participating in the receptacle ripening mediated by ABA (Table 2). In these hub signalings, the numbers of isoforms of Auxin and BR signaling pathways were at least ten while only several numbers were found in other pathways, including ABA. Notably, the most members of gene expression of hub phytohormone signalings were down-regulated by ABA, while only *small* auxin *up-regulated RNA* (*SAUR*) genes, belonged to auxin signaling pathway, and *a regulatory protein NPR* (*NPR*), belonged to SA signaling pathway, were up-regulated (Figure 5; Table 2). Among these hub phytohormone signalings, only FaSnRK2.6 (Isoform0037314) has been verified to be a negative regulator in receptacle ripening mediated by ABA, and the roles of others are still unclear.

Roles of FaSAURs, the hub phytohormone signalings, in receptacle ripening mediated by ABA

Although Small auxin up-regulated RNA (SAUR) genes are important components of auxin signaling and participate in many aspects of plant growth and development (Ren & Gray, 2015), there is limited understanding of roles in non-climacteric fruit ripening. Among hub phytohormone signalings, five SAUR isoforms had the highest positively relationship with ripening-related quality traits, and were upregulated by ABA (Figure 5; Table 2), which suggested that they might act as a positive role in receptacle quality formation. To verify the prediction of the hub phytohormone singnaling network, we therefore firstly explored the function of these SAUR in receptacle ripening mediated by ABA. Based on an alignment of amino acid sequences, these isoforms were divided into FaSAUR1 (Isoform0051199 and 0051690) and FaSAUR2 (Isoform0051401, 0051699, and

10.3389/fpls.2023.1117156

TABLE 1 The list of isoforms associated with phytohormone signaling in five modules.

Module	Phytohormones	Positive	Negative	Total members
Brown	ABA	6	13	19
	Auxin	13	16	29
	Ethylene	3	13	16
	JA	6	3	9
	GA	NA	16	16
	BR	NA	1	1
Tan	ABA	3	NA	3
	Auxin	2	NA	2
	BR	NA	1	1
Blue	ABA	NA	11	11
	Auxin	1	23	24
	Ethylene	3	8	11
	JA	1	3	4
	GA	NA	4	4
	BR	NA	13	13
	CTK	NA	1	1
	SA	2	1	3
Turquoise	ABA	4	20	24
	Auxin	8	19	27
	Ethylene	2	32	34
	JA	8	5	13
	GA	NA	5	5
	BR	NA	11	11
	CTK	3	2	5
	SA	NA	1	1
Yellow	ABA	2	1	3
	Auxin	1	2	3
	Ethylene	2	12	14
	JA	1	4	5
	BR	NA	7	7
	SA	2	NA	2

Positive and Negative represent expression profiles of isoforms positively and negatively related, respectively, to receptacle ripening mediated by ABA. NA, not available.

0052033) sequences (Supplementary Figure S5). Difference among gene isoforms with the same amino acid sequence were observed in their untranslated regions (UTRs) (Supplementary Figure S6). Among these isoforms, the full-length mRNA sequences of Isoform0051199 and 0051401 were the longest in the *FaSAUR1* and *FaSAUR2* types, respectively (Supplementary Figure S6). To verify the function of *FaSAUR1* and *FaSAUR2*, RNAi targets, including the partial domains of CDS and 3'UTR, specific to each of the two genes were designed and used to silence each gene individually in strawberry fruit (Supplementary Figure S5). Transient RNAi assays showed that silencing *FaSAUR2* (RNAi-*FaSAUR2* fruit) generated a red area that was smaller than the areas caused by silencing *FaSARU1* (RNAi-*FaSAUR1* fruit) and both were less than in empty vector RNAi-Control fruit (Figure 6A). Notably, both *FaSAUR1* and *FaSAUR2* were silenced in RNAi-*FaSAUR1* and -*FaSAUR2* fruit, possibly due to sequence similarity between the two RNAi fragments (Figure 6B; Supplementary Figures S5, S6). Based on the transcriptome analysis, the expression levels of *FaSAUR1* and *FaSAUR2* in RNAi-*FaSAUR2* and RNAi-Control fruits were lowest and

10.3389/fpls.2023.1117156

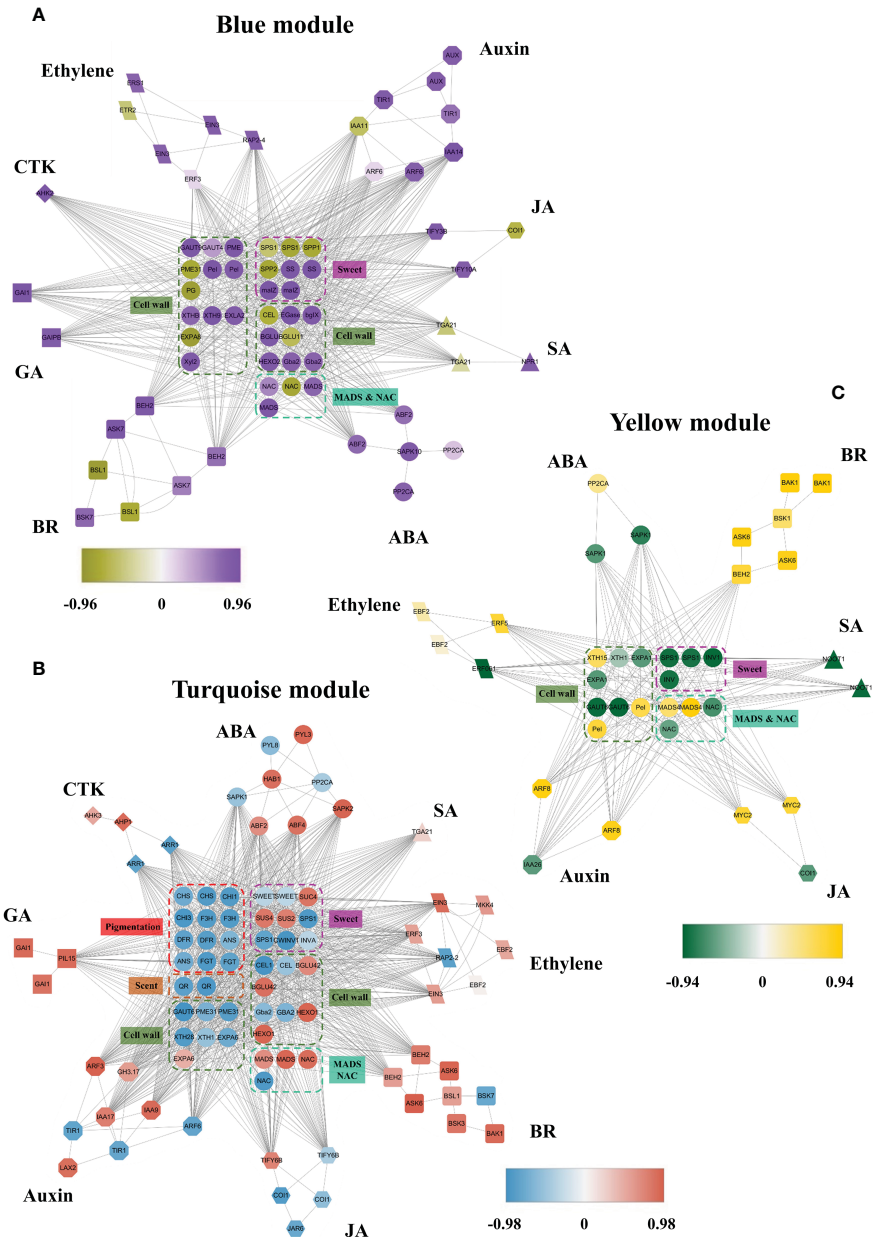


FIGURE 4
The coexpression networks of ABA and other phytohormone signaling pathways during receptacle ripening in blue, turquoise and tan modules. **(A)** The coexpression network in the blue module was negatively related to ripening based on WGCNA. **(B)** The coexpression network in the turquoise module was negatively related to ripening. CTK, cytokinin; SA, salicylic acid. **(C)** The coexpression network in the yellow module was negatively related to ripening. The full names of the isoforms are shown in Supplementary Dataset S1.

highest, respectively (Figure 6B). Moreover, the expressional levels of *FaMYB10* and the anthocyanin biosynthetic genes and the difference of anthocyanin content in the fruit corresponded with the fruit phenotypes and the expressional profiles of *FaSAUR1* and *FaSAUR2* (Figure 6). FaQR, a key gene in the biosynthesis of HDMF (Raab et al., 2006), was highly expressed in RNAi-Control fruit and was expressed at higher levels than in RNAi-*FaSAUR1* and *FaSAUR2* fruits, while HDMF was only was detected in RNAi-Control fruit (Figure 6C, D). In addition, most *FaSPS* genes, corresponding to the rate-limiting gene in sucrose biosynthesis, were down-regulated in the RNAi-*FaSAUR1* and -*FaSAUR2* fruits compared to RNAi-Control and their expressional levels in RNAi-*FaSAUR2* were higher than in RNAi-*FaSAUR1* (Figure 6C). Sucrose contents were approximately 23% and 31% lower in RNAi-*FaSAUR1* and *FaSAUR2* fruit, respectively, compared to RNAi-Control, which was likely the primary reason for the total sugar content in the control fruit being higher than that in RNAi-*FaSAUR1* and -*FaSAUR2* fruits (Figure 6C).

10.3389/fpls.2023.1117156

TABLE 2 The list of isoforms of hub phytohormone signalings in receptacle ripening mediated by ABA.

Phytohormones	Gene name	ID	Modules	Correlation coefficient
ABA	PLY	Isoform0048338	turquoise	0.928147
	SnRK2	Isoform0045396	brown	-0.9327
		Isoform0049457	brown	-0.90388
		Isoform0037314	turquoise	0.913194
		Isoform0046532	turquoise	0.943011
Auxin	IAA	Isoform0044117	brown	-0.91165
		Isoform0045361	blue	0.901957
		Isoform0048541	blue	0.969533
		Isoform0044379	turquoise	0.909236
	ARF	Isoform0000933	brown	-0.93677
		Isoform0004095	brown	-0.92903
		Isoform0005225	turquoise	0.923927
		Isoform0007925	turquoise	0.914834
		Isoform0014472	turquoise	0.964573
		Isoform0014826	turquoise	0.944029
		Isoform0018714	turquoise	0.916653
	SAUR	Isoform0051199	brown	0.975982
		Isoform0051401	brown	0.954076
		Isoform0051690	brown	0.975727
		Isoform0051699	brown	0.951596
		Isoform0052033	brown	0.955492
Ethylene	MMK1	Isoform0039281	brown	-0.95045
		Isoform0042149	brown	-0.96244
		Isoform0043269	brown	-0.9588
GA	DELLA	Isoform0018223	brown	-0.91531
		Isoform0020374	brown	-0.91339
		Isoform0026278	blue	0.905991
		Isoform0027058	blue	0.964394
		Isoform0035101	blue	0.962586
		Isoform0031485	turquoise	0.908676
	PIF	Isoform0016169	turquoise	0.906372
SA	NPR	Isoform0032962	yellow	-0.90934
CTK	AHK	Isoform0000635	blue	0.93727
BR	BAK1	Isoform0016249	blue	0.908892
		Isoform0024480	yellow	0.900491
	BIN2	Isoform0027563	blue	0.908759
		Isoform0032204	blue	0.950279
		Isoform0033990	blue	0.916183
		Isoform0038746	turquoise	0.950838
		Isoform0040492	turquoise	0.94507

(Continued)

TABLE 2 Continued

Phytohormones	Gene name	ID	Modules	Correlation coefficient
		Isoform0041750	turquoise	0.954566
		Isoform0043949	turquoise	0.975739
	BZR1/2	Isoform0044842	blue	0.933731

The abbreviations of gene names in the table as follow: PYL, abscisic acid receptor PYR/PYL family; SnRK2, subfamily 2 SNF1-related kinases; IAA, auxin-responsive protein IAA; ARF, auxin response factor; SAUR, small auxin up-regulated RNA; MMK1, MAPK KINASE1; DELLA, DELLA protein; PIF, phytochrome-interacting factor; NPR, regulatory protein NPR; AHK, Arabidopsis histidine kinase; BAK1, BRASSINOSTEROID INSENSITIVE 1-associated receptor kinase; BIN2, BRASSINOSTEROID INSENSITIVE 2; BZR1/2, BRASSINAZOLE RESISTANT 1/2. Red and blue in the correlation coefficient section represent expression profiles of isoforms positively and negatively related, respectively, to receptacle ripening mediated by ABA.

Discussion

ABA is widely described as a dominant regulator of non-climacteric fruit ripening (Fenn & Giovannoni, 2021). In strawberry fruit, ABA has also been found to interact with other phytohormones, including auxin, GAs, JAs, and ethylene, in receptacle ripening (Li et al., 2022a), suggesting multiple interlinked phytohormone signaling networks participate in ripening mediated by ABA. However, this idea has yet to be investigated in detail. Here we provide evidence of a coexpression network of phytohormone signals and hub signalings in the strawberry receptacle mediated by ABA, based on transcriptome profiling of the receptacle associated with changes in ABA levels, following removing achene and exogenous treatments, as well as data sets quantifying ripening compounds and associated phenotypes (Supplementary Figure S2).

The Previous study shows that the expression profiles of multiple genes of phytohormone signaling pathway, such as ABA, auxin, GA, and ethylene, are influenced in the receptacle with the changes of ABA levels (Gu et al., 2019), which suggests that ABA manipulates receptacle ripening not only *via* itself signaling pathway but also *via* controlling other phytohormone signalings. A total of 328 full-length mRNAs, including those associated with ABA, auxin, GA, JA, ethylene, BR, SA, and CTK signaling, TF genes, and genes related to coloration, sugar accumulation, softening, and aroma, were detected in this coexpression network (Figures 3, 4). The clear coexpression relationship during receptacle ripening among these genes related to different phytohormone signaling components suggests them as regulators of this process. The expression profiles and sequences of these genes and others, totally 18,998, included in this coexpression network provide valuable datasets (Supplementary Datasets S2, S3) for studies of strawberry and potentially of other non-climacteric fruit. The coexpression network included homologs of *FaABI* (Jia et al., 2013) and *FaSnRK2.6* (Han et al., 2015), which negatively regulate receptacle ripening, and displayed decreasing expression levels with receptacle development and were down-regulated by ABA (Figure 4A; Supplementary Figure S4). Moreover, previous studies show that FaSHP (Daminato et al., 2013) and FaRIF (Martín-Pizarro et al., 2021) promote receptacle ripening and their expression is induced by ABA, while that of FaMADS1a (Lu et al., 2018) shows the opposite pattern. We identified homologs of these genes in our coexpression network and their predicted functions in receptacle ripening, based on the expressional profiles, were consistent with the previous studies (Figures 3A, 4B; Supplementary Figure S4). Additionally, homologs of genes that have been shown to participate in strawberry fruit ripening and quality traits, such as *FaMYB10* (Kadomura-Ishikawa et al., 2015) and *FaQR* (Raab et al., 2006), were found in the coexpression network and their expressional profiles were consistent with changes in ripening phenotypes (Figures 3, 4; Supplementary Figure S4; Supplementary Dataset S1). All of these results are consistent with high predictive power of the coexpression network, which provide many clues for studying mechanisms of strawberry receptacle ripening.

In addition, we identified 43 isoforms respectively belonged to seven phytohormone signaling pathways, including auxin, ABA, ethylene, GA, CTK, BR, and SA from the coexpression network. The expression of ABA (*PYL*, *SnRK2*s), auxin (*auxin response factor*s, *auxin-responsive protein IAA*s), ethylene (*MAPK KINASE1*s), GA (*DELLA*s, *phytochrome-interacting factor*), CTK (*arabidopsis histidine kinase*), and BR (*BRASSINOSTEROID INSENSITIVE 1-associated receptor kinase*s, BRASSINOSTEROID INSENSITIVE 2s, *BRASSINAZOLE RESISTANT 1/2*) signaling genes were negatively

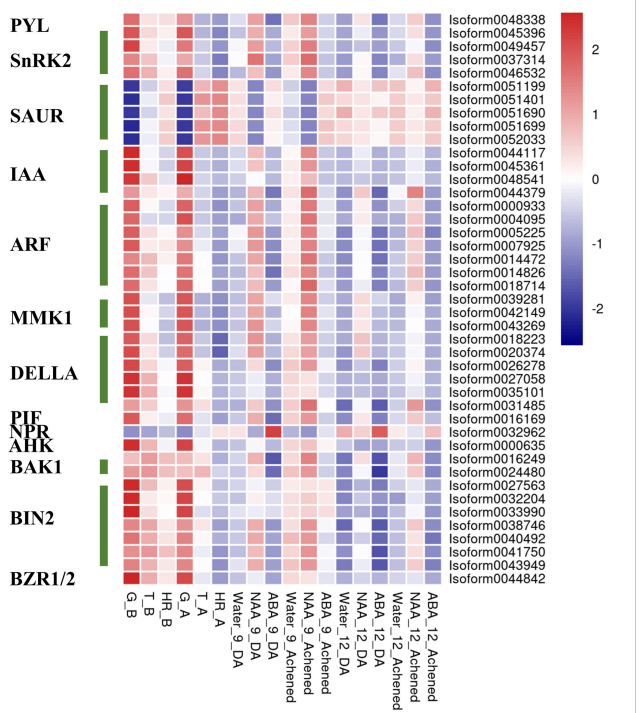


FIGURE 5
The expression profiles of hub phytohormone signaling genes in the receptacle during development and under treatments. The expression profiles of hub phytohormone signaling genes in the receptacle using heatmap analysis.

10.3389/fpls.2023.1117156

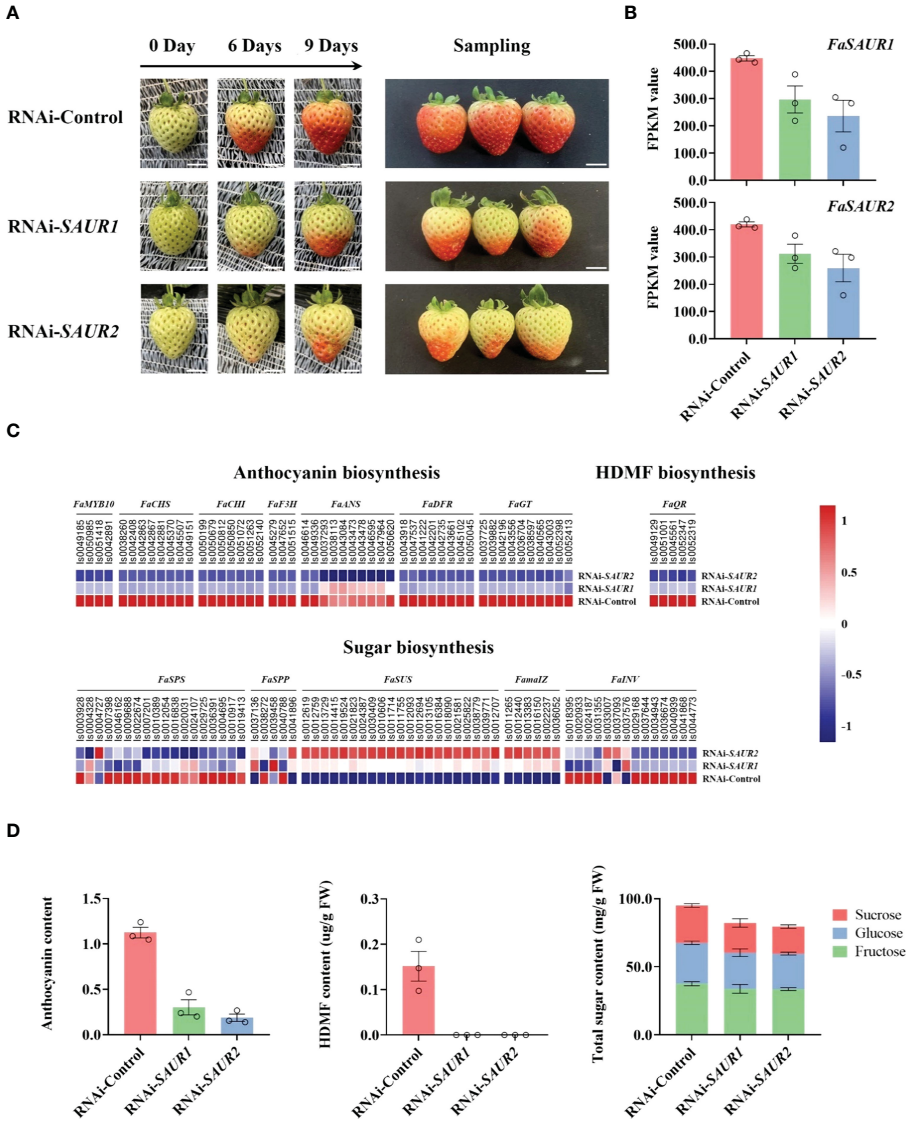


FIGURE 6
Transiently RNAi assays of *FaSAUR1* and *FaSAUR2* in strawberry fruit. **(A)** Phenotypes of fruit after transiently suppressing *FaSAUR1* or *FaSAUR2* using RNAi. The left picture shows the control (injecting empty vector, named as RNAi-Control) and RNAi (named RNAi-*FaSAUR1* and RNAi-*FaSARU2*) fruit at 0, 6, and 9 days, The right picture shows the fruit at sampling stage (9 days). Scale bar = 1 cm. **(B)** Expression levels of *FaSAUR1* or *FaSAUR2* in the control and RNAi fruit based on the FPKM of Isoform0051199 and 0051401. **(C)** Heatmap of expression profiles of genes involved in anthocyanin, HDMF, and sugar biosynthesis in the samples of transient assays, based on FPKM values. **(D)** Levels of anthocyanin, HDMF, and total sugars in the RNAi-Control and RNAi-*FaSAUR1* and RNAi-*FaSAUR2* fruit. The data represent the mean ± SD of three biological replicates.

regulated by ABA (Table 2), which suggested that they might act as repressors for strawberry receptacle ripening. The most of them are not explored the function in the strawberry receptacle ripening but the homolog (Isoform0037314) of FaSnRK2.6. Additionally, FvMAPK3 (MITOGEN-ACTIVATED PROTEIN KINASE3) has been found to repress anthocyanin biosynthesis *via* phosphorylating CHALCONE SYNTHASE1 in wild strawberry *(F. vesca)* fruit (Mao et al., 2022). Although the homolog of FvMAPK3 was not found in the hub phytohormone signalings, three MMK1 belonging to MAPK cascades were identified and they also were predicted to negatively regulate ripening and quality formation. On the other hand, only auxin and SA pathways had the members, *SAUR*s and *NPR*, positively controlled by ABA, which suggested that they might promote quality formation in the receptacle. Based on these results, ABA promote receptacle ripening primarily through down- and up-regulating these hub phytohormone genes, which the specific roles in this process as an important point needs to be investigated in the future.

Using genes that emerged from the expression network analysis, we also extended knowledge of phytohormone signaling that modulates receptacle ripening mediated by ABA. Among the prediction of hub phytohormone signaling genes, only *SAUR*s and *NPR* were up-regulated by ABA, and the former showed the highest correlation coefficient with the ripening and ABA level (Table 2). Therefore, we firstly explored the function of SAURs in the receptacle ripening mediated by ABA, which also were used to verify the reliability of our prediction of hub signaling genes. SAURs are the

largest family of genes that respond to auxin and their expression is also influenced by other phytohormones (Ren & Gray, 2015; Gu et al., 2019). However, the function of SAURs in non-climacteric ripening has yet been previously characterized. Recently, SlSAUR69 was found to promote tomato (*S. lycopersicum* cv. MicroTom) fruit ripening by altering the balance of auxin and ethylene (Shin et al., 2019). Based on predictions of the hub phytohormone signalings combining the analysis of amino acid sequence, expressional profiles and changes in receptacle quality traits, two SAUR homologs, FaSAUR1 and FaSAUR2, were identified as candidates of hub phytohormone signalings for participating in receptacle ripening mediated by ABA (Figure 5; Supplementary Figures S5, S6). We determined through transient RNAi assays and RNA-Seq that the anthocyanin, HDMF, and total sugar levels in RNAi-*FaSAUR1* and -*FaSAUR2* receptacles were lower than in RNAi-Control, in accordance with the changes in expression of the related genes (Figure 6). This is consistent with supposedly roles of *FaSAUR1* and *FaSAUR2* in positively regulating receptacle quality formation. Interestingly, a phylogenetic analysis indicated that FaSAUR1 and FaSAUR2 are closely related to AtSAUR76/77/78 (Supplementary Figure S7), which is a negative regulator of leaf growth (Markakis et al., 2013). Thus, these results further verified the predictive capacity of the coexpression network of multiple phytohormone signaling networks in receptacle ripening mediated by ABA. However, the specific mechanism of receptacle ripening mediated by FaSAUR1 and FaSAUR2 needs further study.

In summary, we describe a coexpression network of phytohormone signaling in the ripening receptacle mediated by ABA and present high-resolution expressional profiles and full-length RNA sequences of suites of genes included in this network (Supplementary Datasets S2, S3), and predict the hub phytohormone signaling genes involving in receptacle ripening mediated by ABA. In addition, we present a strategy for using these data to identify additional ripening factors from multiple phytohormone signaling systems and tested two auxin signaling pathway factors, *FaSAUR1* and *FaSAUR2*, which are up-regulated by ABA and that promote anthocyanin, HDMF, and sugar biosynthesis. These results have great potential for elucidating ripening and quality formation in strawberry receptacle with implications that can additionally be tested in other fruits.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://ngdc.cncb.ac.cn/search/?dbId=gsa&q=CRA006989&page=1, CRA006989; https://ngdc.cncb.ac.cn/search/?dbId=gsa&q=CRA006997&page=1, CRA006997.

Author contributions

K-SC and G-HJ managed the project. K-SC, JR, Y-NS and B-JL designed experiment and coordinated the project. K-SC, JR, Y-NS and B-JL wrote the paper. K-SC, JR, G-HJ, Y-NS, B-JL and JG discussed about the results of experiments and reviewed the paper. G-HJ and X-FY grew the plant material. B-JL, H-RJ, X-FY, Y-FS and JL collected and prepared samples. JL assisted B-JL in analyzing transcriptome data. H-RJ and Y-FS assisted B-JL in participating in the experiments. 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 (2022YFD2100100), National Natural Science Foundation of China (nos. 32102345) and the 111 project (B17039).

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.2023.1117156/full#supplementary-material

SUPPLEMENTARY DATA SHEET 1
The information of genes involved in phytohormone signalings, related to qualities, NAC and MADS families in five core modules.

SUPPLEMENTARY DATA SHEET 2
All of full-length transcript sequences in five core modules.

SUPPLEMENTARY DATA SHEET 3
The information of all of genes in five core modules.

References

Castillejo, C., Waurich, V., Wagner, H., Ramos, R., Oiza, N., Muñoz, P., et al. (2020). Allelic variation of *MYB10* is the major force controlling natural variation in skin and flesh color in strawberry (*Fragaria* spp.) fruit. *Plant Cell* 32, 3723–3749. doi: 10.1105/tpc.20.00474

Chai, Y. M., Jia, H. F., Li, C. L., Dong, Q. H., and Shen, Y. Y. (2011). FaPYR1 is involved in strawberry fruit ripening. *J. Exp. Bot.* 62, 5079–5089. doi: 10.1093/jxb/err207

Daminato, M., Guzzo, F., and Casadoro, G. (2013). A *SHATTERPROOF*-like gene controls ripening in non-climacteric strawberries, and auxin and abscisic acid antagonistically affect its expression. *J. Exp. Bot.* 64, 3775–3786. doi: 10.1093/jxb/ert214

Edger, P. P., Poorten, T. J., VanBuren, R., Hardigan, M. A., Colle, M., McKain, M. R., et al. (2019). Origin and evolution of the octoploid strawberry genome. *Nat. Genet.* 51, 541–547. doi: 10.1038/s41588-019-0356-4

FAO (2020). *Food and agriculture organization of the united nations* (Italy: FAOSTAT). Available at: http://www.fao.org/faostat/en/#data/QC.

Fenn, M. A., and Giovannoni, J. J. (2021). Phytohormones in fruit development and maturation. *Plant J.* 105, 446–458. doi: 10.1111/tpj.15112

Fischer, T. C., Mirbeth, B., Rentsch, J., Sutter, C., Ring, L., Flachowsky, H., et al. (2014). Premature and ectopic anthocyanin formation by silencing of anthocyanidin reductase in strawberry (*Fragaria* × *ananassa*). *New Phytol.* 201, 440–451. doi: 10.1111/nph.12528

Gao, Q., Luo, H., Li, Y., Liu, Z., and Kang, C. (2020). Genetic modulation of *RAP* alters fruit coloration in both wild and cultivated strawberry. *Plant Biotechnol. J.* 18, 1550–1561. doi: 10.1111/pbi.13317

Gordon, S. P., Tseng, E., Salamov, A., Zhang, J. W., Meng, X. D., Zhao, Z. Y., et al. (2015). Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. *PLoS One* 10, e0132628. doi: 10.1371/journal.pone.0132628

Gu, T., Jia, S., Huang, X., Wang, L., Fu, W., Huo, G., et al. (2019). Transcriptome and hormone analyses provide insights into hormonal regulation in strawberry ripening. *Planta* 250, 145–162. doi: 10.1007/s00425-019-03155-w

Han, Y., Dang, R., Li, J., Jiang, J., Zhang, N., Jia, M., et al. (2015). SUCROSE NONFERMENTING1-RELATED PROTEIN KINASE2.6, an ortholog of OPEN STOMATA1, is a negative regulator of strawberry fruit development and ripening. *Plant Physiol.* 167, 915–930. doi: 10.1104/pp.114.251314

Jia, H. F., Chai, Y. M., Li, C. L., Lu, D., Luo, J. J., Qin, L., et al. (2011). Abscisic acid plays an important role in the regulation of strawberry fruit ripening. *Plant Physiol.* 157, 188–199. doi: 10.1104/pp.111.177311

Jia, H., Sun, M., Li, B., Han, Y., Zhao, Y., Li, X., et al. (2013a). Sucrose functions as a signal involved in the regulation of strawberry fruit development and ripening. *New Phytol.* 198, 453–465. doi: 10.1111/nph.12176

Jia, H. F., Lu, D., Sun, J. H., Li, C. L., Xing, Y., Qin, L., et al. (2013b). Type 2C protein phosphatase ABI1 is a negative regulator of strawberry fruit ripening. *J. Exp. Bot.* 64, 1677–1687. doi: 10.1093/jxb/ert028

Jia, H., Jiu, S., Zhang, C., Wang, C., Tariq, P., Liu, Z., et al. (2016). Abscisic acid and sucrose regulate tomato and strawberry fruit ripening through the abscisic acid-stress-ripening transcription factor. *Plant Biotechnol. J.* 14, 2045–2065. doi: 10.1111/pbi.12563

Kadomura-Ishikawa, Y., Miyawaki, K., Takahashi, A., Masuda, T., and Noji, S. (2015). Light and abscisic acid independently regulated *FaMYB10* in *Fragaria* × *ananassa* fruit. *Planta* 241, 953–965. doi: 10.1007/s00425-014-2228-6

Kou, X., Yang, S., Chai, L., Wu, C., Zhou, J., Liu, Y., et al. (2021a). Abscisic acid and fruit ripening: Multifaceted analysis of the effect of abscisic acid on fleshy fruit ripening. *Sci. Hortic.* 281, 109999. doi: 10.1016/j.scienta.2021.109999

Kou, X., Zhou, J., Wu, C. E., Liu, Y., Chai, L., and Xue, Z. (2021b). The interplay between ABA/ethylene and NAC TFs in tomato fruit ripening: a review. *Plant Mol. Biol.* 106, 223–238. doi: 10.1007/s11103-021-01128-w

Kumar, S., Stecher, G., and Tamura, K. (2016). Mega7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Langfelder, P., and Horvath, S. (2008). WGCNA: an r package for weighted correlation network analysis. *BMC Bioinf.* 9, 559. doi: 10.1186/1471-2105-9-559

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., et al. (2007). Clustal W and clustal X version 2.0. *Bioinformatics* 23, 2947–2948. doi: 10.1093/bioinformatics/btm404

Li, H. (2021). New strategies to improve minimap2 alignment accuracy. *Bioinformatics* 37, 4572–4574. doi: 10.1093/bioinformatics/btab705

Li, T., Dai, Z., Zeng, B., Li, J., Ouyang, J., Kang, L., et al. (2022b). Autocatalytic biosynthesis of abscisic acid and its synergistic action with auxin to regulate strawberry fruit ripening. *Hortic. Res.* 9, uhab076. doi: 10.1093/hr/uhab076

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinf.* 12, 323. doi: 10.1186/1471-2105-12-323

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Li, B. J., Grierson, D., Shi, Y., and Chen, K. S. (2022a). Roles of abscisic acid in regulating ripening and quality of strawberry, a model non-climacteric fruit. *Hortic. Res.* 9, uhac08. doi: 10.1093/hr/uhac089

Li, S. J., Liu, S. C., Lin, X. H., Grierson, D., Yin, X. R., and Chen, K. S. (2022c). Citrus heat shock transcription factor CitHsfA7-mediated citricacid degradation in response to heat stress. *Plant Cell Environ.* 45, 95–104. doi: 10.1111/pce.14207

Liu, T., Li, M., Liu, Z., Ai, X., and Li, Y. (2021). Reannotation of the cultivated strawberry genome and establishment of a strawberry genome database. *Hortic. Res.* 8, 41. doi: 10.1038/s41438-021-00476-4

Liu, Y., Tang, M., Liu, M., Su, D., Chen, J., Gao, Y., et al. (2020). The molecular regulation of ethylene in fruit ripening. *Small Methods* 4, 1900485. doi: 10.1002/smtd.201900485

Li, B. J., Zheng, B. Q., Wang, J. Y., Tsai, W. C., Lu, H. C., Zou, L. H., et al. (2020). New insight into the molecular mechanism of colour differentiation among floral segments in orchids. *Commun. Biol.* 3, 89. doi: 10.1038/s42003-020-0821-8

Lu, W., Chen, J., Ren, X., Yuan, J., Han, X., Mao, L., et al. (2018). One novel strawberry MADS-box transcription factor *FaMADS1a* acts as a negative regulator in fruit ripening. *Sci. Hortic.* 227, 124–131. doi: 10.1016/j.scienta.2017.09.042

Mao, W., Han, Y., Chen, Y., Sun, M., Feng, Q., Li, L., et al. (2022). Low temperature inhibits anthocyanin accumulation in strawberry fruit by activating FvMAPK3-induced phosphorylation of FvMYB10 and degradation of chalcone synthase 1. *Plant Cell* 34, 1226–1249. doi: 10.1093/plcell/koac006

Markakis, M. N., Boron, A. K., Loock, B. V., Saini, K., Cirera, S., Verbelen, J. P., et al. (2013). Characterization of a small auxin-up RNA (SAUR)-like gene involved in *Arabidopsis thaliana* development. *PLoS One* 8, e82596. doi: 10.1371/journal.pone.0082596

Martín-Pizarro, C., Vallarino, J. G., Osorio, S., Meco, V., Urrutia, M., Pillet, J., et al. (2021). The NAC transcription factor FaRIF controls fruit ripening in strawberry. *Plant Cell* 33, 1574–1593. doi: 10.1093/plcell/koab070

Medina-Puche, L., Molina-Hidalgo, F. J., Boersma, M., Schuurink, R. C., López-Vidriero, I., Solano, R., et al. (2015). An R2R3-MYB transcription factor regulates eugenol production in ripe strawberry fruit receptacles. *Plant Physiol.* 168, 598–561. doi: 10.1104/pp.114.252908

Molina-Hidalgo, F. J., Franco, A. R., Villatoro, C., Medina-Puche, L., Mercado, J. A., Hidalgo, M. A., et al. (2013). The strawberry (*Fragaria×ananassa*) fruit-specific *rhamnogalacturonate lyase 1* (*FaRGLyase1*) gene encodes an enzyme involved in the degradation of cell-wall middle lamellae. *J. Exp. Bot.* 64, 1471–1483. doi: 10.1093/jxb/ers386

Molina-Hidalgo, F. J., Franco, A. R., Villatoro, C., Medina-Puche, L., Mercado, J. A., Hidalgo, M. A., et al. (2017). The fruit-specific transcription factor FaDOF2 regulates the production of eugenol in ripe fruit receptacles. *J. Exp. Bot.* 68, 4529–4543. doi: 10.1093/jxb/erx257

Paniagua, C., Blanco-Portales, R., Barceló-Muñoz, M., García-Gago, J. A., Waldron, K. W., Quesada, M. A., et al. (2016). Antisense down-regulation of the strawberry β-galactosidase gene *FaβGal4* increases cell wall galactose levels and reduces fruit softening. *J. Exp. Bot.* 67, 619–631. doi: 10.1093/jxb/erv462

Pi, M., Hu, S., Cheng, L., Zhong, R., Cai, Z., Liu, Z., et al. (2021). The MADS-box gene *FveSEP3* plays essential roles in flower organogenesis and fruit development in woodland strawberry. *Hortic. Res.* 8, 247. doi: 10.1038/s41438-021-00673-1

Raab, T., López-Ráez, J. A., Klein, D., Caballero, J. L., Moyano, E., Schwab, W., et al. (2006). *FaQR*, required for the biosynthesis of the strawberry flavor compound 4-hydroxy-2,5-dimethyl-3(2H)-furanone, encodes an enone oxidoreductase. *Plant Cell* 18, 1023–1037. doi: 10.1105/tpc.105.039784

Ren, H., and Gray, W. M. (2015). SAUR proteins as effectors of hormonal and environmental signals in plant growth. *Mol. Plant* 8, 1153–1164. doi: 10.1016/j.molp.2015.05.003

Shan, L. L., Li, X., Wang, P., Cai, C., Zhang, B., Sun, C. D., et al. (2008). Characterization of cDNAs associated with lignification and their expression profiles in loquat fruit with different lignin accumulation. *Planta* 227, 1243–1254. doi: 10.1007/s00425-008-0696-2

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

Shimizu, K., Adachi, J., and Muraoka, Y. (2006). Angle: a sequencing errors resistant program for predicting protein coding regions in unfinished cDNA. *J. Bioinf Comput. Biol.* 4, 649–664. doi: 10.1142/S0219720006002260

Shin, J. H., Mila, I., Liu, M., Rodrigues, M. A., Vernoux, T., Pirrello, J., et al. (2019). The RIN-regulated small auxin-up RNA SAUR69 is involved in the unripe-to-ripe phase transition of tomato fruit *via* enhancement of the sensitivity to ethylene. *New Phytol.* 222, 820–836. doi: 10.1111/nph.15618

Shi, Y., Vrebalov, J., Zheng, H., Xu, Y., Yin, X., Liu, W., et al. (2021). A tomato LATERAL ORGAN BOUNDARIES transcription factor, *SlLOB1*, predominantly regulates cell wall and softening components of ripening. *P Natl. Acad. Sci. U.S.A.* 118, e2102486118. doi: 10.1073/pnas.2102486118

Thompson, P. A. (1969). The effect of applied growth substances on development of the strawberry fruit. II. interactions of auxins and gibberellins. *J. Exp. Bot.* 20, 629–647. doi: 10.1093/jxb/20.3.629

Vallarino, J. G., Merchante, C., Sánchez-Sevilla, J. F., Balaguer, M. A., Pott, D. M., Ariza, M. T., et al. (2020). Characterizing the involvement of *FaMADS9* in the regulation of strawberry fruit receptacle development. *Plant Biotechnol. J.* 18, 929–943. doi: 10.1111/pbi.13257

Wang, W., Fan, D., Hao, Q., and Jia, W. (2022a). Signal transduction in non-climacteric fruit ripening. *Hortic. Res.* 9, uhac190. doi: 10.1093/hr/uhac190

Wang, W. Q., Moss, S. M. A., Zeng, L., Espley, R. V., Wang, T., Lin-Wang, K., et al. (2022b). The red flesh of kiwifruit is differentially controlled by specific activation–repression systems. *New Phytol.* 235, 630–645. doi: 10.1111/nph.18122

Zhang, Z., Shi, Y., Ma, Y., Yang, X., Yin, X., Zhang, Y., et al. (2020). The strawberry transcription factor FaRAV1 positively regulates anthocyanin accumulation by activation of *FaMYB10* and anthocyanin pathway genes. *Plant Biotechnol. J.* 18, 2267–2279. doi: 10.1111/pbi.13382

Zhang, Y., Yin, X., Xiao, Y., Zhang, Z., Li, S., Liu, X., et al. (2018). An ETHYLENE RESPONSE FACTOR-MYB transcription complex regulates furaneol biosynthesis by activating *QUINONE OXIDOREDUCTASE* expression in strawberry. *Plant Physiol.* 178, 189–201. doi: 10.1104/pp.18.00598

Zhou, J., Li, D. D., Wang, G., Wang, F., Kunjal, M., Joldersma, D., et al. (2020). Application and future perspective of CRISPR/Cas9 genome editing in fruit crops. *J. Integr. Plant Biol.* 62, 269–286. doi: 10.1111/jipb.12793

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 01 February 2023
DOI 10.3389/fpls.2023.1127228

OPEN ACCESS

EDITED BY
Jun Li,
Zhejiang University, China

REVIEWED BY
Zhiyuan Zhang,
Zhejiang University, China
Wei Ma,
Hebei Agricultural University, China

*CORRESPONDENCE
Xingjiang Qi
qxj@zaas.ac.cn

†These authors share first authorship

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

RECEIVED 19 December 2022
ACCEPTED 18 January 2023
PUBLISHED 01 February 2023

CITATION
Wu X, Zhang S, Yu Z, Sun L, Liang S, Zheng X, Qi X and Ren H (2023) Molecular cloning and functional analysis of Chinese bayberry *MrSPL4* that enhances growth and flowering in transgenic tobacco. *Front. Plant Sci.* 14:1127228. doi: 10.3389/fpls.2023.1127228

Molecular cloning and functional analysis of Chinese bayberry *MrSPL4* that enhances growth and flowering in transgenic tobacco

Xiangqi Wu[1†], Shuwen Zhang[2†], Zheping Yu[2], Li Sun[2], Senmiao Liang[2], Xiliang Zheng[2], Xingjiang Qi[1,2,3*] and Haiying Ren[2]

[1]College of Chemistry and Life Sciences, Zhejiang Normal University, Jinhua, China, [2]State Key Laboratory for Managing Biotic and Chemical Threats to Quality and Safety of Agro-products, Institute of Horticulture, Zhejiang Academy of Agricultural Sciences, Hangzhou, China, [3]Biotechnology Research Institute, Xianghu Laboratory, Hangzhou, China

Chinese bayberry (*Myrica rubra*) is an important tree in South China, with its fruit being of nutritional and high economic value. In this study, early ripening (ZJ), medium ripening (BQ) and late ripening (DK) varieties were used as test materials. Young leaves of ZJ, BQ and DK in the floral bud morphological differentiation periods were selected for transcriptome sequencing to excavate earliness related genes. A total of 4,538 differentially expressed genes were detected. Based on clustering analysis and comparisons with genes reportedly related to flowering in *Arabidopsis thaliana*, 25 homologous genes were identified. Of these, one gene named *MrSPL4* was determined, with its expression down-regulated in DK but up-regulated in ZJ and BQ. *MrSPL4* contained SBP domain and the target site of miR156, and its total and CDS length were 1,664 bp and 555 bp respectively. The overexpression vector of *MrSPL4* (35S::35S::*MrSPL4*-pCambia2301-KY) was further constructed and successfully transfected into tobacco to obtain *MrSPL4*-positive plants. Based on the results of qRT-PCR, the relative expression of *MrSPL4* was up regulated by 3,862.0-5,938.4 times. Additionally, the height of *MrSPL4*-positive plants was also significantly higher than that of wild-type (WT), with the bud stage occurring 12 days earlier. Altogether, this study identified an important gene -*MrSPL4* in Chinese bayberry, which enhanced growth and flowering, which provided important theoretical basis for early-mature breeding of Chinese bayberry.

KEYWORDS

Chinese bayberry, growth and flowering, MrSPL4, gene function, qRT-PCR

Introduction

Myrica rubra (Lour.) *Sieb. et Zucc.*, of the family *Myricaceae*, is a native, economically important tree in South China where it is particularly concentrated in the south of the Yangtze River Basin (Zhang et al., 2022). Its fruit, Chinese bayberry, is not only soft, juicy and rich in flavor, but in addition to ecological benefits, it also has medical uses. Chinese bayberry is widely favored by consumers, especially in the Zhejiang Province where its fame has, to a large extent, promoted the healthy development of the Chinese bayberry industry to drive the economic development of planting areas (Jia et al., 2019; Ren et al., 2019).

The maturation period of the existing main varieties of *Myrica rubra* is about 15 days and occurs in the middle to late June. In addition, the subsequent ripening period is very short and coincides with the plum rain season in the South, which causes great harm and serious economic losses. In this context, early maturing varieties can effectively avoid the influence of plum rain, lengthen the maturity period, reduce the market pressure caused by concentrated maturation and significantly improve the economic benefits of cultivation. Therefore, the development of early maturing germplasm as well as the cultivation of new varieties with characteristics of early maturation have become important for the sustainable development of this industry. After years of observation and studies on the development of different flower buds, it was found that early flowering is an important phenotype related to early ripening. Hence, identifying flowering genes and elucidating their mechanism of action can be useful to regulate the ripening stage and create new germplasm with the improved characteristics. In this context, the SQUAMOSA promoter-binding protein-like (*SPL*) gene family is known to be important for the regulation of plant flowering, but its role in the flowering process of Chinese bayberry is yet to be reported.

The *SPL* gene family, also known as SBP protein, is a unique type of transcription factor in green plants whose members have a highly conserved SBP domain (Yang et al., 2021). The latter, which is about 80 amino acid residues in length, contains two zinc finger structures (Cys-Cys-His-Cys;Cys-Cys-Cys-His or Cys-Cys-Cys-Cys) as well as a nuclear localization signal (NLS) located at the C-terminal. Most *SPL* genes also contain highly conserved microRNA156/157 (miR156/miR157) targeting sites that regulate more complex physiological processes (Birkenbihl et al., 2005; Li et al., 2020). *SPL*s were first identified from *Antirrhinum majus* but with the rise of plant genomics, they have been isolated, identified and analyzed from a number of other plants, including *Oryza sativa* (Yang et al., 2008), *Arabidopsis thaliana* (Wu et al., 2009), *Solanum lycopersicum* (Salinas et al., 2012), *Malus* × *domestica Borkh.* (Li et al., 2013), *Vitis Vinifera* (Hou et al., 2013), peony (*Paeonia Suffruticosa*) (Zhu et al., 2018), *Fragaria Vesca* (Xiong et al., 2018), *Citrus sinensis* (Liu et al., 2017). Currently, *SPLs* are considered to be key genes that regulate biological processes in plants, especially since they show variations in their functions. For instance, these genes can regulate the flowering process of plants (Lei et al., 2018; Guo et al., 2019), coordinate plant root, stem and leaf development (Yu et al., 2015; Wang et al., 2018; Wang et al., 2019; Li et al., 2021), influence response to biotic and abiotic stress (Ning et al., 2017; Feyissa et al., 2019) as well as participate in secondary metabolic processes (Yang et al., 2021).

In recent years, other functions of *SPL* genes in plants have also been widely studied. For example, in *Arabidopsis thaliana*, *SPL10* was found to be highly expressed in plant leaf and root tissues, resulting in earlier flowering, narrower leaf shapes, smaller and fewer rosette leaves as well as reduced root length and root number by binding to the *AGL79* promoter (Gao et al., 2018). In addition, the expression of *AtSPL9* and *AtSPL10* in leaf primordia was also reported to affect the differentiation of apical meristem into leaves (Wang et al., 2008), while in leaf tissues, *SPL2* could control floral organs, long silique development and plant fertility by activating *AS2* (Wang et al., 2016). In rice, the GO function analysis of differentially expressed genes in blade leaves of *SPL4* mutant rice showed that *OsSPL4* gene mutations affected protein phosphorylation as well as the binding of iron ions in rice leaves, maintaining the normal plant type of rice (Hu et al., 2021). In pea (*Pisum sativum L.*), *PsSPL3a/3c* was found to be mainly expressed at the transcriptional level in leaves, hence indicating its possible involvement in leaf phase transition in the pea aging pathway (Vander Schoor et al., 2022). In maize (*Zea Mays*), *SPL4* plays an important role in bract development and meristem establishment (Chuck et al., 2010), while *SPL10/14/26* not only regulates the expression of *ZmWOX3A* and auxin related genes but is also involved in the development of epidermal hair on maize leaves (Kong et al., 2021). Finally, an analysis of the expression of *MdSBP* genes in apple leaves after different hormone treatments showed that many of the genes responded to different plant hormones, thereby suggesting that *MdSBP* genes could be involved in response to hormone signals during stress or apple development (Li et al., 2013).

Therefore, based on the previous genome sequencing of *Myrica rubra* (Ren et al., 2019) the latter's *SPL* gene family was identified and analyzed based on bioinformatics methods. This was followed by the cloning of an *SPL* gene and its subsequent heterologous expression in *Nicotiana benthamiana* L. by constructing an overexpression vector to validate the functions of the gene. Altogether, this study is expected to provide a theoretical basis for revealing the regulatory pathway of flowering in Chinese bayberry.

Materials and methods

Material information

By referring to the expression of genes related to male and female flowering as described by Jia et al. (2019), three experimental materials with different flowering stages were selected from the Lanxi International Chinese bayberry Research Center(Latitude 29.30°N, longitude 119.60°E) in November 2019 (period during which floral buds can be morphologically differentiated). These included the early maturing variety "Zaojia" (ZJ), the medium maturing variety "Biqizhong" (BQ) and the late maturing variety "Dongkui" (DK) (Table 1). The ages of all selected trees were around 15 years, and they were in consistent cultivation conditions. Each variety was sampled in triplicates, and for transcriptome sequencing, young leaves (not unfolded) were taken from annual branches facing south and at 1 m above the ground.

TABLE 1 Phenological periods of different test materials.

Test material	No.	Flower bud formation time	First flowering period (month-day)	Maturity period (month-day)
Zaojia	ZJ	Mid-November	April 1st	June 6th
Biqizhong	BQ	Mid-December	April 4th	June 15th
Dongkui	DK	Mid-January	April 10th	June 25th

Transcriptome sequencing and screening of differentially expressed genes

A polysaccharide polyphenol RNA extraction kit was used for extracting total RNA from the samples (TIANGEN, Beijing), and after RNA detection, Biomarker Biotechnology Co., Ltd. was commissioned to carry out the transcriptome sequencing. For this purpose, magnetic beads with Oligo (dT) were used to enrich the total RNA of the samples before fragmenting the mRNA with the fragmentation buffer. The first cDNA strand was then synthesized with random hexamers using mRNA as template, and this was followed by the synthesis of the second cDNA strand by adding buffer, dNTPs, RNase H and DNA polymerase I. After purification with the QiaQuick PCR kit and elution with EB buffer, end repair was performed, poly (A) tails were added and sequencing adaptors were connected. Appropriate fragments were then selected by gel electrophoresis prior to PCR-based amplification. The resulting libraries were eventually sequenced on an Illumina HiSeq4000.

After gene splicing, protein sequences were aligned with those from eight public databases (COG, GO, KEGG, KOG, Pfam, Swissport, eggNOG and Nr) using a threshold of $e \leq e^{-10}$. The BLAST algorithm was then used for sequence similarity comparison, with the resulting sequence similarities subsequently used for functional annotations. Relative gene expression was assessed based on RPKM (Reads Per Kilobase of exon model per Million mapped reads) where larger RPKM values were indicative of higher expression levels (Trapnell et al., 2010).

Differentially expressed genes were screened by the false discovery rate (FDR) (Zhao et al., 2020), with a $|\log_2$ fold change$| \geq 2$ and an FDR<0.5 selected as thresholds for a gene to be considered as being differentially expressed.

SPL gene family analysis

The *SPLs* of Chinese bayberry were isolated and identified by tBLASTn analysis of AtSPL amino acid sequences obtained from the genomic data of Chinese bayberry (Ren et al., 2019). The Chinese bayberry *SPLs* and target sites of miRNA156 were then predicted and confirmed using Genscan Web (http://genes.mit.edu/GENSCAN.html) as well as the BLASTx algorithm (http://www.ncbi.nlm.nih.gov/BLAST). After obtaining the nucleotide and amino acid sequences of 16 Arabidopsis and 46 apple *SPL* family genes from the Plant Transcription Factor Database (PlnTFDB3.0) (http://plntfdb.bio.uni-potsdam.de/v3.0/), phylogenetic trees were also constructed using the NJ method in MEGA 7.0, along with full-length protein sequences and the test parameter (bootstrap) set to 1000. The exon and intron structures of Chinese bayberry *SPL* genes were obtained by Gene Structure Display Sever (http://gsds.cbi.pku.edu.cn/index.php).

Strains and vectors

Escherichia coli competent cell DH-5α (Shanghai Jinchao Technology Development Co., LTD.), *Agrobacterium tumefaciens* strain GV3101 and pCambia2301-KY vectors (Shanghai Kaiyi Biotechnology Co., LTD.) were the main requirements of the study.

Primer design and gene cloning

Using the genome sequence of Chinese bayberry (Ren et al., 2019), specific primers for both sides of the open reading framework (ORF) of the target gene were designed with Primer Premier 5.0 software for gene cloning. Total RNA extraction was also performed on healthy Chinese bayberry leaves using the modified cetyl trimethyl ammonium bromide (CTAB) method, with the extracted RNA acting as template to synthesize cDNA according to the instructions of the HiScript 1st Strand cDNA Synthesis Kit (Vazyme). This was followed by PCR amplification with the Phanta Max ultra-fidelity DNA polymerase (Vazyme), using the cDNA as template. In this case, each reaction consisted of the following component: 1 μL of Phanta Max super-Fidelity DNA Polymease, 2 μL of cDNA, 2 μL each of both forward and reverse primers, 25 μL of 2×Phanta Max Buffer, 1 μL of dNTP Mix and ddH_2O (for making up the volume to 50 μL), while the PCR procedure involved an annealing temperature of 49 °C and an extension rate of 1 kb/min, carried out for 39 cycles. Other operations shall follow the product instructions of Vazyme Company. The amplified products were finally detected on 1.5% agarose gel, before being sent to the company for sequencing to verify the accuracy of cloning results.

Construction of an overexpression vector and Agrobacterium transformation

The restriction enzyme *BamHI* (Takara Company) was first used to linearize the vector before extracting the pCambia2301-KY plasmid for digestion with the same enzyme. The overexpression vector was then constructed with Vazyme recombinant enzyme at 37 °C by using the following components: 2 μL of 5×CE II Buffer, 1 μL of Exnase II, 4 μL of linearized carrier, 1 μL of insert fragment and ddH_2O (to make up the volume to 10 μL). After 30 min of reaction, the vector was placed on ice for cooling. The cells were then transfected into competent *E. coli* DH-5α cells and cultured in LB medium containing 50 mg/L Kan. This enabled the selection and subsequent culture of resistant colonies for the positive detection of the gene by PCR. The amplified products were finally sent for sequencing. The positive transformer colony plasmid was extracted and transfected into *Agrobacterium tumefaciens* GV3101 and sterile glycerol was added to preserve the bacteria at -80 °C until required for the next transfection.

Agrobacterium tumefaciens-mediated transfection of tobacco

Tobacco Benn was selected for this set of transformation experiment. WT tobacco was infected with Agrobacterium carrying recombinant vector plasmids of target genes using the leaf disk method. After four times of continuous screening/subculture, resistant buds were eventually recovered and transferred to a rooting medium to induce roots. Once the root system was vigorous, healthy and completely regenerated plants were transplanted to the soil (nutrient soil-vermiculite ratio was 1:1 or 2:1) where they were maintained until the T_0 generation for seed collection.

Collected seeds were sterilized with 70% ethanol, 30% sodium hypochlorite or 40% of 84 disinfectant and rinsed with sterile water 5-6 times. Seeds were then added to 1/2 MS solid selective medium containing 80 mg/L of Kan, and vernalized at 4°C for 2 days to break dormancy. They were subsequently cultured in a light incubator of the laboratory of Zhejiang Academy of Agricultural Sciences (light 28°C, 16 h, Darkness 25°C, 8 h, humidity 50%-70%). After about a week, the seeds were transferred to the soil to maintain grow. The leaves of the transgenic resistant plants and the wild-type ones of the T_1 generation were randomly sampled and stored at -80°C after being frozen in liquid nitrogen.

Determination of relative gene expression

Transgenic positive plants of T_1 generation were obtained through screening with 80 mg/L Kan, 1/2 MS solid selective medium and PCR. The leaves of grown plants were collected, and total RNA was extracted with the RNA simple Total RNA Kit (TIANGEN) after quick-freezing in liquid nitrogen. In addition, synthesis reactions were also performed in 10-μL reaction volumes with the first Strand cDNA synthesis kit. For this purpose, the following components were used as required by the FastFire qPCR PreMix (SYBR Green) Kit (TIANGEN): 5 μL of 2×FastFire qPCR PreMix, 1 μL of forward primer and reverse primers (10 μm) and 1 μL of cDNA template. The reaction was performed on a Light Cycler 96 real-time PCR instrument under the following conditions: 95°C for 60 s, followed by 45 cycles, each at 95°C for 5 s, 63°C for 10 s and 72°C 15 s. Three technical replicates were set for each sample. Quantitative primers were designed according to gene sequences, with *Ntactin*-F/R selected as the reference gene (Zhao et al., 2020), and WT tobacco acting as the control to determine the relative expression of target genes. The $2^{-\triangle\triangle Ct}$ method was used to process the data (Livak and Schmittgen, 2001), while the IBM SPSS Statistics 22 and Origin 2022/Microsoft Excel 2010 were used for statistical analysis and plotting respectively.

Cloning, structural analysis and construction of overexpression vector of *MrSPL4*

Primers were therefore designed based on the reference genome sequence (Table S1), with ORF sequences of the *MrSPL4* gene in ZJ, BQ and DK subsequently obtained by PCR amplification. Therefore, it was inferred that the expression of this gene was different between the different test materials, probably due to the promoter element, but this remained to be experimentally verified. The full length and CDS of *MrSPL4* were 1,664 bp and 555 bp respectively. The amplified product was first recovered, and the vector was digested with *BamHI*. The resulting enzymatically digested product was then recombined with the amplified product to construct a plant overexpression vector. The latter was transformed into *E. coli* competent cells DH-5α before identifying the transformed bacterial solution by PCR.

Verification of *MrSPL4* positive tobacco plants

Agrobacterium-mediated transformation of *Nicotiana benthamiana* with the recombinant plasmid 35S::*MrSPL4*-pCambia2301-KY was performed. The tobacco leaves infected by *Agrobacterium tumefaciens* were directly transferred to a selective medium containing kanamycin (Kan) to induce differentiation and budding. After the buds had grown to 2-3cm, they were inserted into a rooting medium to induce the formation of roots. Once the root system was vigorous, the seedlings were then tempered and transplanted to soil for culture to obtain completely regenerated tobacco with Kan resistance. Leaf DNA from the resistant regenerated tobacco plants to be tested was used as a template for PCR-based validation.

Results

Evaluation of transcriptome data of young leaves from different flowering materials

The transcriptome sequencing of nine samples of young leaves (three biological replicates for each variety) in the floral bud morphological differentiation period was completed, and a total of 59.78 Gb of clean data, with an average GC content of 47.28% and a Q30 base ratio of 93.55%, were obtained. After comparison with the reference genome (Ren et al., 2019), the percentage of clean reads aligned to the reference genome was found to be 95.39% (Table S2, the transcriptome data of BQ and DK were uploaded to https://bigd.big.ac.cn/gsa/browse/CRA008253, and the datasets of ZJ generated and analyzed during the current study are available in the NCBI repository https://www.ncbi.nlm.nih.gov/sra/PRJNA733585). The number of differentially expressed genes between the three samples was then compared. In this case, 623 genes were differentially expressed between ZJ and DK, and of these, 476 were up-regulated and 147 were down-regulated. Similarly, 2,343 genes were differentially expressed between ZJ and BQ, with 1,385 and 958 genes being up-regulated and down-regulated respectively. Finally, the number of differentially expressed genes between DK and BQ was 1,572, with 734 and 838 being up-regulated and down-regulated respectively (Table 2).

KEGG analysis of differentially expressed genes

Through KEGG enrichment analysis, it was found that the differentially expressed genes mainly involved functions such as cellular processes, environmental information processing, genetic

TABLE 2 Differentially expressed genes between pairs of samples.

DEG Set	DEG Number	Up-regulated	Down-regulated
BQ Vs DK	1,572	734	838
ZJ Vs BQ	2,343	1,385	958
DK Vs ZJ	623	476	147

information processing, metabolism and organic systems (Figures 1A–C). The pathways that were significantly enriched in all the three groups included phytohormone signal transduction, sulfur and carbon metabolism, fatty acid, phenylpropionic acid, pyruvic acid, α-linolenic acid metabolism, glycine, serine, threonine, arginine, proline, cyano amino acids, cysteine and methionine metabolism, terpenoid skeleton biology, carotenoid biosynthesis, glutathione and glycerophosphatide metabolism, glycolysis/gluconeogenesis, starch and sucrose, amino sugar and nucleotide sugar metabolism, protein processing in the endoplasmic reticulum. The results indicated that these pathways may participate in the regulation network of Chinese bayberry flowering or other important pathways.

Identification of *MrSPL4* based on flowering-related differentially expressed genes

The differentially expressed genes mentioned above were compared with 306 flowering genes reported in *Arabidopsis* (Bouché et al., 2016) and 25 genes were found to be homologous in Chinese bayberry (Figure 2A). In particular, one of the differentially expressed genes, MRNA_003335_1, was down-regulated in DK but up-regulated in ZJ and BQ. This gene also contained the *SBP* domain and belonged to the *SPL* gene family, named *MrSPL4*. The relative expression of *MrSPL4* was therefore verified by qRT-PCR (Figure 2B), and the results showed that ZJ had the highest expression, followed by BQ, with DK showing the lowest expression level. These results were, in fact, consistent with the expression determined by the transcriptome.

Gene analysis of *SPL*s gene family in Chinese bayberry

Through the screening of all genes in the reported genome provided in Ren et al. (2019), 17 *SPL* family genes with *SBP* domains were found in the Chinese bayberry genome (Table 3). Of these, 12 genes including *MrSPL4*, contained the target site of miR156 in the CD region. The software MEGA7.0 was then used to analyze the evolution of 17 *SPL* genes in *Myrica rubra* (*MrSPL*), 16 *SPL* genes

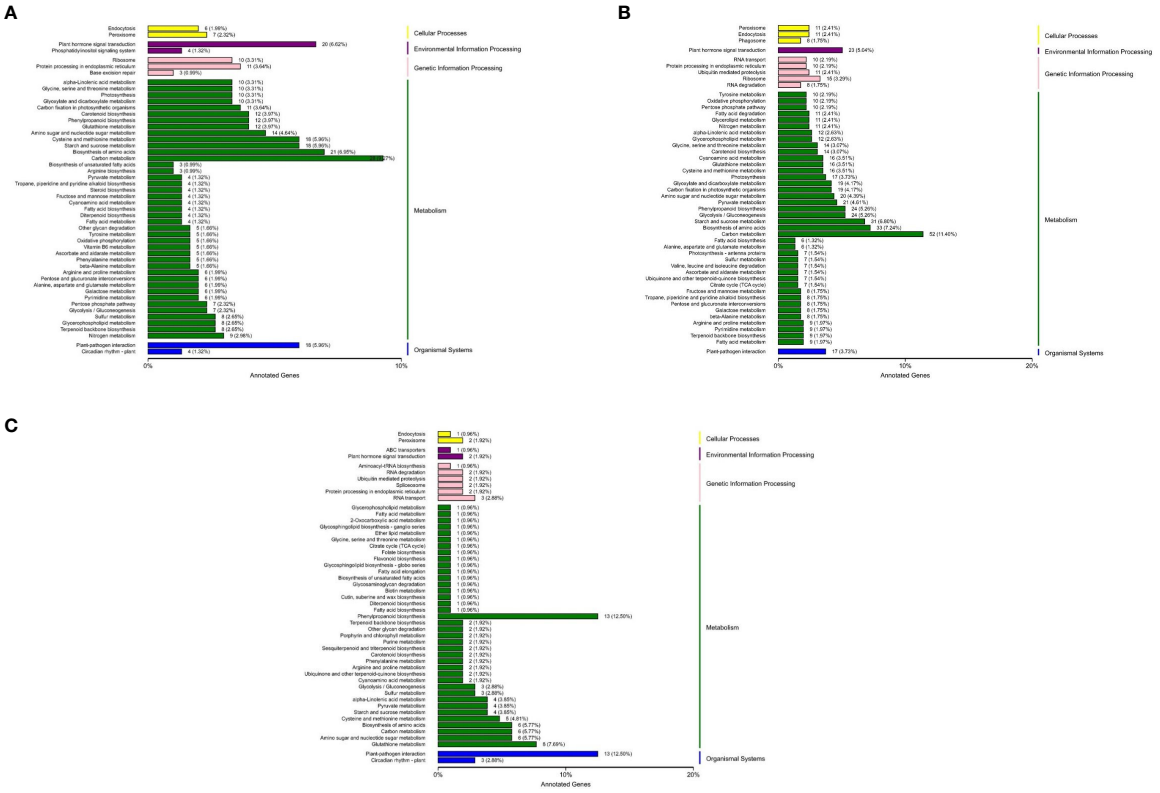


FIGURE 1
KEGG classification of differentially expressed genes in the different test materials. **(A)** KEGG classification of differentially expressed genes between BQ and DK. **(B)** KEGG classification of differentially expressed genes between BQ and ZJ. **(C)** KEGG classification of differentially expressed genes between DK and ZJ.

10.3389/fpls.2023.1127228

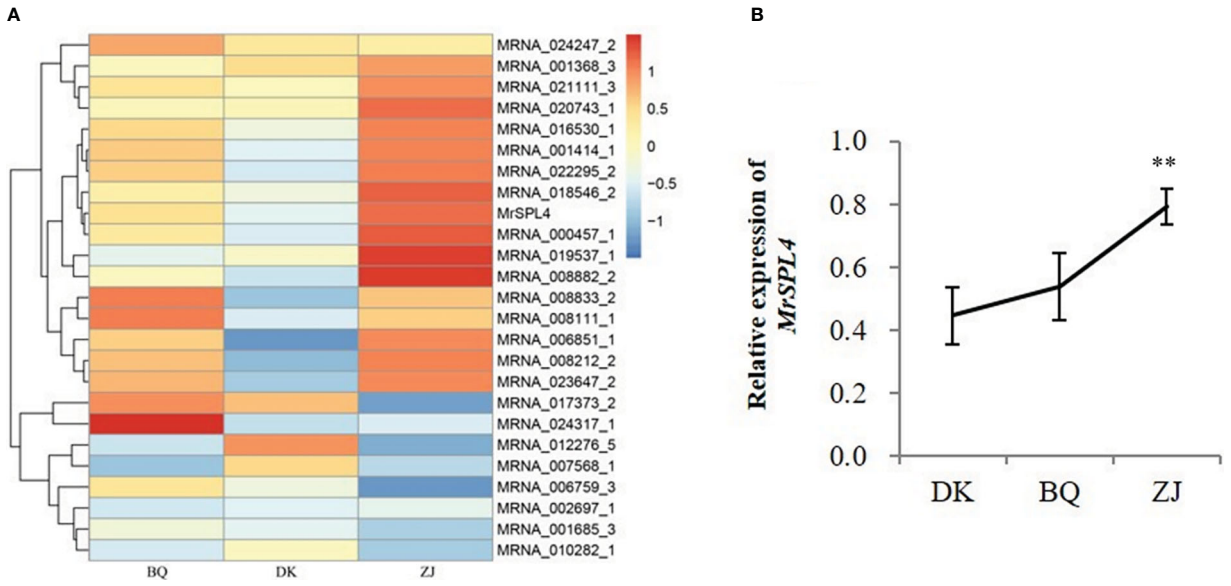


FIGURE 2
Expression of flowering-related differential genes. **(A)** The expression heat map of flowering related genes. **(B)** The relative expression of *MrSPL4* in different samples. Values are the mean (± SD) of three replicates, ** indicates 0.001< *p* <0.05 by the Student's t-test.

in *Arabidopsis thaliana* and 46 *SPL* genes in apple (Figure 3). In this case, it was observed that the *SPL* proteins could be divided into four different groups (I, II, III and IV), with each containing at least one *MrSPL* gene. More specifically, groups I and II contained one *MrSPL* each, group IV contained twelve *MrSPLs* and group III contained three Chinese bayberry *SPL* genes, including *MrSPL4*. *MrSPL4* had the highest homology with the AT1G20980 (*AtSPL14*) gene in *Arabidopsis thaliana*, with previous studies showing that this gene (*AtSPL14*) not only promoted the normal growth and development of *Arabidopsis thaliana*, but also played a crucial role in the development

TABLE 3 Gene status of the SPL family in *Myrica rubra* genome.

Gene ID	Conserved domain	Target site of miR156
MRNA_003237_1	*SBP* domain	No
MRNA_013273_1	*SBP* domain	No
MRNA_013732_2	*SBP* domain	Yes
MrSPL4	*SBP* domain	Yes
MRNA_019617_1	*SBP* domain	No
MRNA_009540_1	*SBP* domain	Yes
MRNA_018907_1	*SBP* domain	Yes
MRNA_022003_2	*SBP* domain	Yes
MRNA_023074_1	*SBP* domain	Yes
MRNA_023065_2	*SBP* domain	No
MRNA_000299_1	*SBP* domain	No
MRNA_007729_2	*SBP* domain	Yes
MRNA_013669_4	*SBP* domain	Yes
MRNA_016389_2	*SBP* domain	Yes
MRNA_001354_1	*SBP* domain	Yes
MRNA_008983_1	*SBP* domain	Yes
MRNA_000723_2	*SBP* domain	Yes

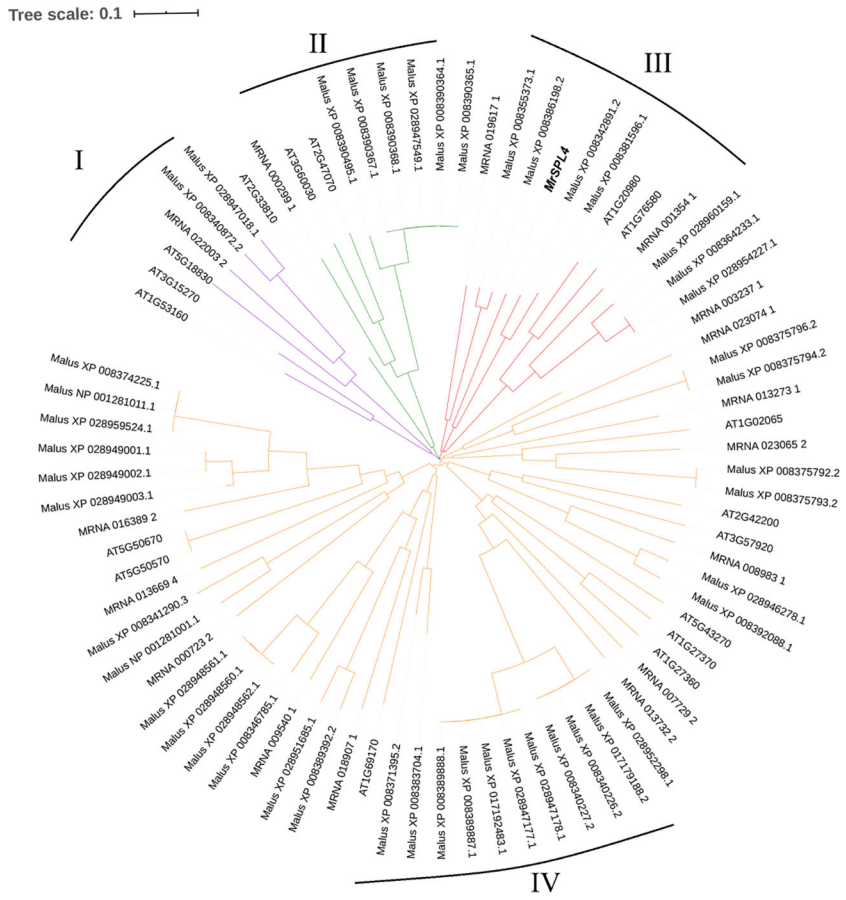


FIGURE 3
Evolutionary analysis of the SPL family gene in *Myrica rubra* and other species.

of flowering as well as the transformation from a vegetative to reproductive growth. Thus, it was speculated that *MrSPL4* could be playing an important role in the flowering process of Chinese bayberry.

Sequence alignment and overexpression vector construction of *MrSPL4*

The electrophoresis results (Figure 4A) showed that the bands were consistent with the expected amplification product size, hence indicating that the sequence of the coding region of *MrSPL4* was successfully obtained. No differences in the gene sequence were noted between the three samples (Figure 4B). *MrSPL4* also contained two exons and one intron (Figure 4C), along with a binding site of *miRNA156* in the CDS1 region (Figure 4D).

The overexpression vector of *MrSPL4* was constructed and transformed into *E. coli* competent cells DH-5α before identifying the transformed bacterial solution by PCR (Figure 4E). The results showed that bands 1, 3, 5, 6, 7 and 8 were consistent with the expected target fragments. In fact, preliminary results further showed that the *MrSPL4* gene was successfully inserted into the vector to yield six positive transformant colonies. Two of these positive colonies were randomly selected for sequencing, with the results being still consistent. This experiment therefore showed that the overexpression vector of *MrSPL4* gene was successfully constructed and named as 35S::*MrSPL4*-pCambia2301-KY.

Regeneration and identification of *MrSPL4*-positive tobacco plants

The tobacco leaves infected by *Agrobacterium tumefaciens* were induced differentiation and budding (Figure 5A), and small seedlings were further grown and induced roots (Figures 5B, C). The regenerated tobacco with Kan resistance was finally transplanted to soil (Figure 5D). Additionally, *MrSPL4*-F and *MrSPL4*-R were used as primers, water was set as a blank control, the 35S::*MrSPL4*-pCambia2301-KY expression vector plasmid was used as a positive control and leaf DNA of WT plants was used as a negative control. The results (Figure 5E) showed that for all resistant regenerated tobacco, the target band was amplified, with its size being consistent with that of the positive control. In addition, no specific bands were observed in WT tobacco and the blank control. Hence, the results showed that the *MrSPL4* gene was successfully transferred into tobacco.

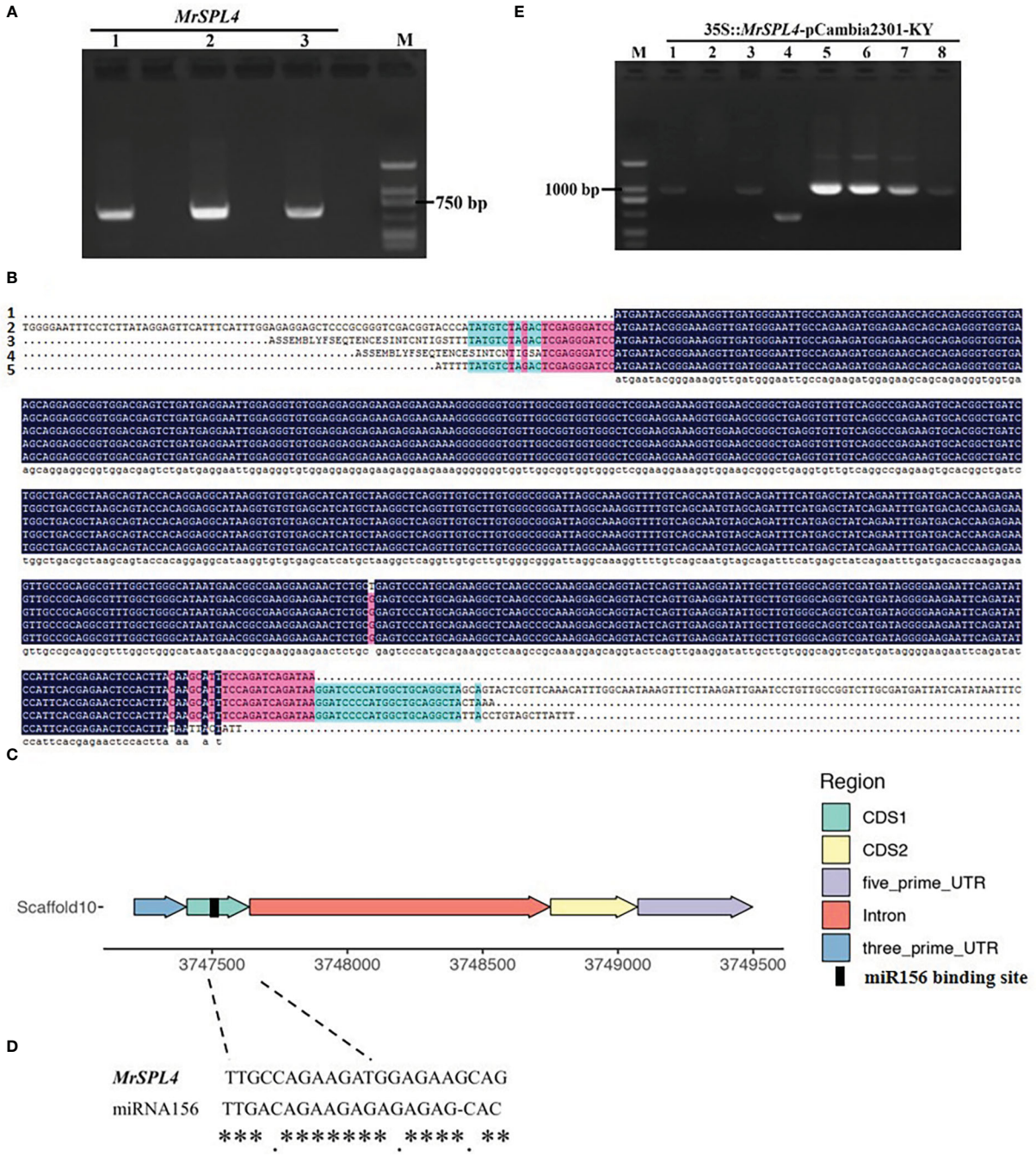


FIGURE 4
Cloning, sequence alignment, structure, miRNA156 target and construction of overexpression vector of *MrSPL4* gene in different test materials. **(A)** The result of the coding region of *MrSPL4* electrophoresis, M: DL2000 DNA marker, 1-3: *MrSPL4* gene cloning in different samples. **(B)** The CDS sequence of *MrSPL4*, 1 refers to genome sequence, and 2, 3, 4 and 5 refer to gene sequence in ZJ, ZJ, BQ and DK, respectively. **(C)** The gene structure of *MrSPL4*. **(D)** The binding site of *miRNA156*. ** represent the consistency of the binding site. **(E)** The PCR identifying result of the transformed bacterial solution, M: DL2000 DNA marker, 1-8: 35S::*MrSPL4*-pCambia2301-KY.

Relative expression and phenotype of *MrSPL4* in positive tobacco plants

Three positive tobacco lines (35S::*MrSPL4*) from the T_1 generation were randomly selected to detect the expression level of the *MrSPL4* gene. Results showed that the relative expression level was significantly higher than that of WT tobacco, with the up-regulation multiple being between 3,862.0-5,938.4 (Figure 6A). This was a clear indication that the gene was overexpressed in transformed tobacco.

The plant heights for 35S::*MrSPL4*-transformed tobacco and WT ones were measured on 34 d, 41 d and 78 d after transplanting and found to be significantly higher for the transformed plants compared with the WT (Figure 6B). Furthermore, the growth rate was also significantly faster than for the WT. In terms of the budding stages, those of 35S::*MrSPL4*-transformed tobacco and WT plants were 34 days and 46 days after transplanting, respectively. Based on the above, it could be concluded that 35S::*MrSPL4*-transformed tobacco showed characteristics of rapid plant growth and early flowering (Figure 6C).

10.3389/fpls.2023.1127228

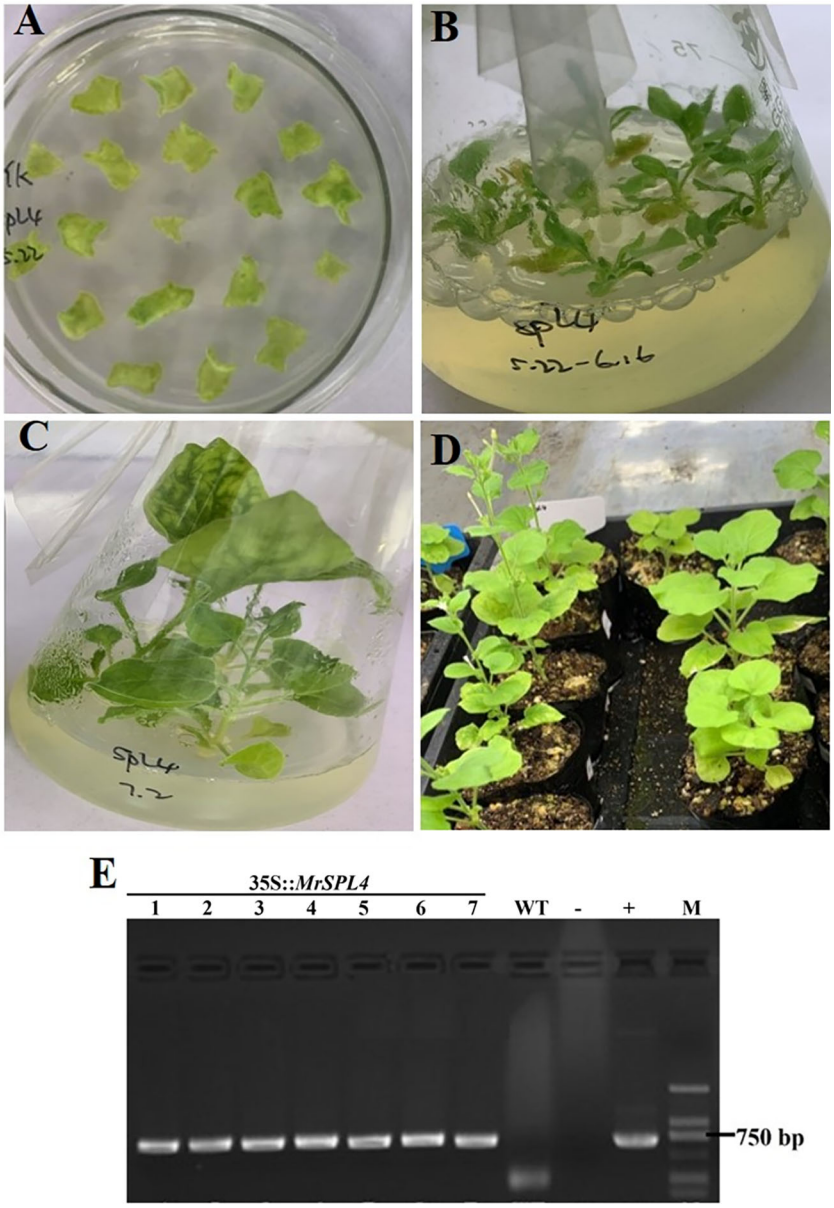


FIGURE 5
Regeneration of positive tobacco plants and PCR identification of Kan resistance. **(A)** Kanamycin medium for bud differentiation. **(B)** Resistant plant culture. **(C)** Rooting of resistant plants. **(D)** Transplanted seedling culture. **(E)** Transgenic tobacco identification using genomic PCR, 1~7: resistant regenerated tobacco, W: negative control, -: blank control, +: positive control, M: DL2000 DNA marker.

Therefore, it was speculated that *MrSPL4* gene affected the phenotype of transgenic tobacco to promote plant growth and flowering.

Discussion

As a specific and important transcription factor in plants, the *SPL* gene family has a highly conserved SBP domain which plays an important regulatory role in plant growth and development. Although the *SPL* gene family has been widely isolated and identified in many plants such as Arabidopsis and rice, research on its role in *Myrica rubra*, an economically important fruit in South China, has not been reported. Previous studies (Yang et al., 2008) have shown that the number of *SPL* gene family members varies in different species, thereby leading to the diversification of gene functions and this was confirmed in the current study. In addition, 17 *MrSPL* gene family members with SBP domains were identified in the genome of *Myrica rubra*, with this number being close to that of *SPL* gene family members in *Arabidopsis* (Wu et al., 2009) and tomato (Cui et al., 2020), but greatly different from that of *Gossypium hirsutum* L (Cai et al., 2018). and apple (Li et al., 2013). In general, members in the same subgroup are likely to have the same or quite similar functions. For example, *AtSPL2*, *AtSPL10* and *AtSPL11*

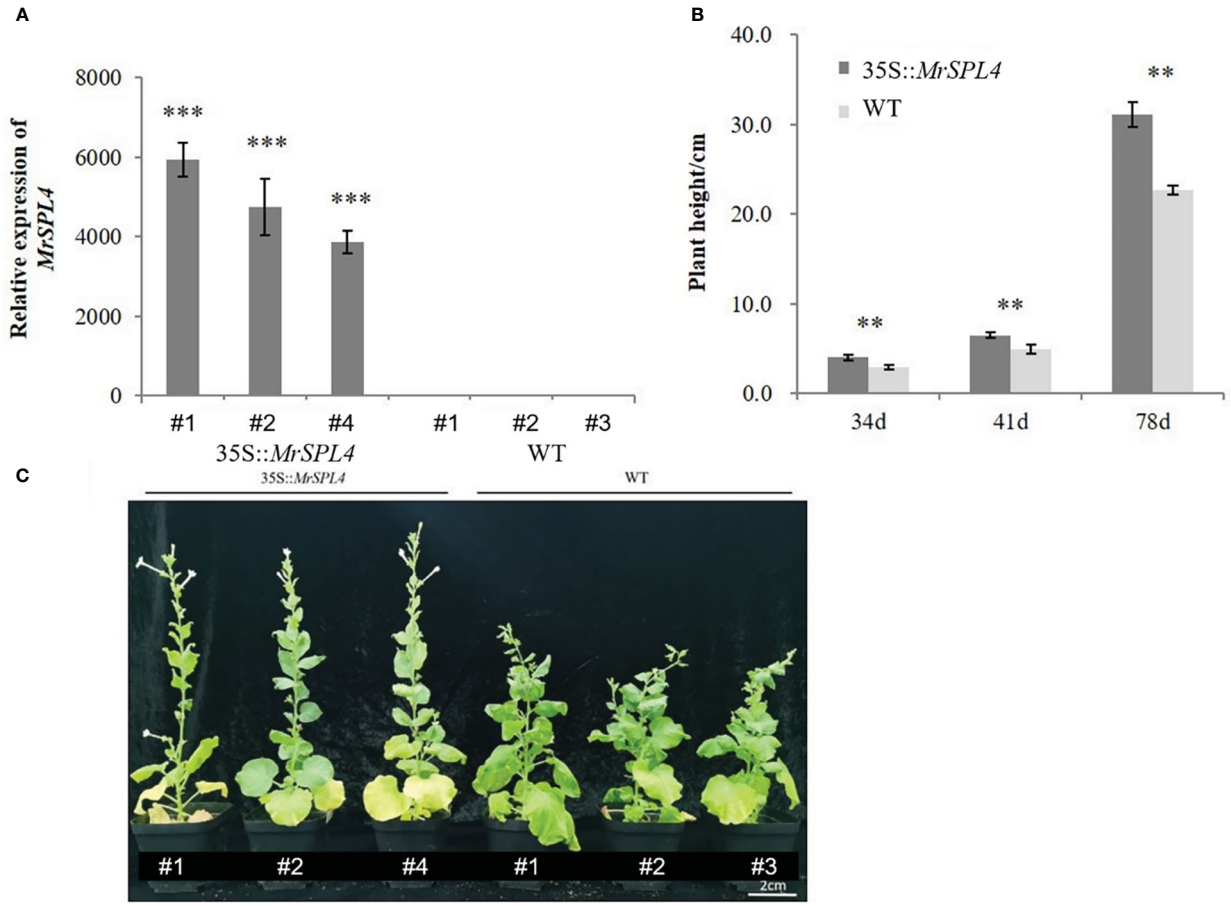


FIGURE 6
Analysis in the T_1 generation of positive tobacco lines. **(A)** *MrSPL4* expression in positive and WT tobacco. **(B)** The plant height of positive and WT tobacco. 34d, 41d and 78d indicate the number of days after transplanting. **(C)** The phenotype of positive and WT tobacco. Values are the mean (± SD) of three replicates, ** indicates 0.001< p <0.05, *** indicates p <0.001 by the Student's t-test.

inhibit root growth, while other members of this group, *CsSPL2* and *CsSPL10*, also participate in the regulation of root development (Yang et al., 2021). In the present study, phylogenetic-based analyses showed that *SPLs* could be divided into four groups, with each containing at least one *MrSPL* gene. Since *MrSPL4* was found to be homologous with the *AtSPL14* gene, it was therefore speculated that *MrSPL4* could be playing a similarly important role in plant growth and development, although it is likely that the gene could also have different functions.

Previous studies have found that flowering is an important sign of plant growth and development, and consequently, research on the role of the *SPL* gene family in the regulation of plant flowering has attracted significant interest. For example, *AtSPL3/4/5* participates in the photoperiod and the age pathway, and as such, it can promote the early flowering of *Arabidopsis* by upregulating the expression of downstream genes (Hyun et al., 2016). Similarly, overexpression of the *EjSPL3/4/5/9* genes in loquat causes transgenic *Arabidopsis thaliana* to exhibit characteristics of early flowering (Jiang et al., 2019), while strawberry *FvSPL10-OE* plants were shown to bloom 3-5 days earlier (Xiong et al., 2019). Despite the above observations, the functions of *MrSPL4* in the flowering process of Chinese bayberry remains unknown. In order to verify its role, the gene was cloned from Chinese bayberry to yield transgenic tobacco overexpressing *MrSPL4*. In this study, the relative expression of *MrSPL4* positive tobacco plants was significantly increased by 3,862.0-5,938.4 times compared with WT under long sunshine conditions. Moreover, the plant heights of transformed tobacco plants were significantly higher than WT tobacco, with the budding period also occurring 12 days earlier. This indicated that the *MrSPL4* gene responded to the flowering process of transgenic tobacco, showing early flowering and increased plant height. In addition, the current study found that the sequence of the *MrSPL4* gene in different Chinese bayberry varieties had no differences, although its expression level did differ in different Chinese bayberry varieties. It was speculated that these differences could be linked to promoter elements but this would need follow-up experiments for validation.

To sum up, 17 members of the *SPL* gene family with SBP domains were identified in *Myrica rubra*. Of these, the *MrSPL4* gene was isolated, cloned and verified in tobacco. The results showed that *MrSPL4* could regulate the flowering process of plants, accelerate their growth and endow the plants with early flowering phenotypes, thus supporting the view that this gene exerted multiple regulatory functions on plant growth and development. These results also provide a basis for further elucidating *MrSPL4*'s regulatory mechanism for flowering in *Myrica*

rubra in order to achieve genetic improvement and gene breeding of this plant in the future. Therefore, the *MrSPL4* gene needs to be further studied, especially with regards to its promoter region.

Data availability statement

The datasets presented in this study can be found in online repositories. A link to the data can be found below: https://bigd.big.ac.cn/gsa/browse/CRA008253.

Author contributions

XW, ZY, and SZ performed the experiments. XQ assisted with design of the project. LS, SL, XZ, and HR assisted with the primary data analysis. XW and SZ wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was funded from the Special breeding program for new varieties in Zhejiang (2021C02066-2) and the Key R & D in Zhejiang (2021C02009) projects.

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.2023.1127228/full#supplementary-material

References

Birkenbihl, R. P., Jach, G., Saedler, H., and Huijser, P. (2005). Functional dissection of the plant-specific SBP-domain: overlap of the DNA binding and nuclear localization domains. *J. Mol. Biol.* 352 (3), 585–596. doi: 10.1016/j.jmb.2005.07.013

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 (D1), D1167–D1171. doi: 10.1093/nar/gkv1054

Cai, C. P., Guo, W. Z., and Zhang, B. H. (2018). Genome-wide identification and characterization of SPL transcription factor family and their evolution and expression profiling analysis in cotton. *Sci. Rep.* 8 (1), 762. doi: 10.1038/s41598-017-18673-4

Chuck, G., Whipple, C., Jackson, D., and Hake, S. (2010). The maize SBP-box transcription factor encoded by tasselsheath4 regulates bract development and the establishment of meristem boundaries. *Development.* 137 (8), 1243–1250. doi: 10.1242/dev.048348

Cui, L., Zheng, J. F., Wang, J. F., Zhang, F. M., Xiao, F. M., Ye, J., et al. (2020). MiR156a-targeted SBP-box transcription factor SlSPL13 regulates inflorescence morphogenesis by directly activating *SFT* in tomato. *Plant Biotechnol. J.* 18, 1670–1682. doi: 10.1111/pbi.13331

Feyissa, B. A., Arshad, M., Gruber, M. Y., Kohalmi, S. E., and Hannoufa, A. (2019). The interplay between *miR156/SPL13* and *DFR/WD40-1* regulate drought tolerance in alfalfa. *BMC Plant Biol.* 19 (1), 434. doi: 10.1186/s12870-019-2059-5

Gao, R. M., Wang, Y., Gruber, M. Y., and Hannoufa, A. (2018). MiR156/SPL10 modulates lateral root development, branching and leaf morphology in arabidopsis by silencing *AGAMOUS-LIKE 79*. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.02226

Guo, J. Q., Tang, C. R., Chen, N. C., Wang, H., Debnat, S., Sun, L., et al. (2019). *SPL7* and *SPL8* represent a novel flowering regulation mechanism in switchgrass. *New Phytol.* 222 (3), 1610–1623. doi: 10.1111/nph.15712

Hou, H. M., Li, J., Gao, M., Stacy, D. S., Wang, H., Mao, L. Y., et al. (2013). Genomic organization, phylogenetic comparison and differential expression of the SBP-box family genes in grape. *PloS One* 8 (3), e59358. doi: 10.1371/journal.pone.0059358

Hu, J. H., Huang, L. Y., Chen, G. L., Liu, H., Zhang, Y. S., Zhang, S. L., et al. (2021). The elite alleles of *OsSPL4* regulate grain size and increase grain yield in rice. *Rice.* 14 (1), 90. doi: 10.1186/s12284-021-00531-7

Hyun, Y., Richter, R., Vincent, C., Martinez-Gallegos, R., Porri, A., and Coupland, G. (2016). Multi-layered regulation of SPL15 and cooperation with SOC1 integrate endogenous flowering pathways at the *Arabidopsis* shoot meristem. *Dev. Cell.* 37 (3), 254–266. doi: 10.1016/j.devcel.2016.04001

Jia, H. M., Jia, H. J., Cai, Q. L., Wang, Y., Zhao, H. B., Yang, W. F., et al. (2019). The red bayberry genome and genetic basis of sex determination. *Plant Biotechnol. J.* 17 (2), 397–409. doi: 10.1111/pbi.12985

Jiang, Y. Y., Peng, J. R., Wang, M., Su, W. B., Gan, X. Q., Jing, Y., et al. (2019). The role of *EjSPL3*, *EjSPL4*, *EjSPL5*, and *EjSPL9* in regulating flowering in loquat (*Eriobotrya japonica* lindl.). *Int. J. Mol. Sci.* 21 (1), 248. doi: 10.3390/ijms21010248

Kong, D. X., Pan, X., Jing, Y. F., Zhao, Y. P., Duan, Y. P., Yang, J., et al. (2021). ZmSPL10/14/26 are required for epidermal hair cell fate specification on maize leaf. *New Phytol.* 230 (4), 1533–1549. doi: 10.1111/nph.17293

Lei, M., Li, Z. Y., Wang, J. B., Fu, Y. L., Ao, M. F., and Xu, L. (2018). Constitutive expression of *Aechmea fasciata SPL14 (AfSPL14)* accelerates flowering and changes the plant architecture in *Arabidopsis*. *Int. J. Mol. Sci.* 19 (7), 2085. doi: 10.3390/ijms19072085

Li, B. B., Zhao, Y. J., Wang, S., Zhang, X. H., Wang, Y. W., Shen, Y., et al. (2021). Genome-wide identification, gene cloning, subcellular location and expression analysis of SPL gene family in *P. granatum l. BMC Plant Biol.* 21 (1), 400. doi: 10.1186/S12870-021-03171-7

Li, J., Gao, X. Y., Zhang, X., and Liu, C. (2020). Dynamic expansion and functional evolutionary profiles of plant conservative gene family SBP-box in twenty two flowering plants and the origin of miR156. *Biomolecules.* 10 (5), 757. doi: 10.3390/blom10050757

Li, J., Hou, H. M., Li, X. Q., Xiang, J., Yin, X. J., Gao, H., et al. (2013). Genome-wide identification and analysis of the SBP-box family genes in apple(*Malus* × *domestica* borkh.). *Plant Physiol. Biochem.* 70, 100–114. doi: 10.1016/j.plaphy.2013.05.021

Liu, M. Y., Wu, X. M., Long, J. M., and Guo, W. W. (2017). Genomic characterization of miR156 and SQUAMOSA *promoter binding protein-like* genes in sweet orange (*Citrus sinensis*). *Plant Cell Tiss Organ Cult.* 130 (1), 103–116. doi: 10.1007/s11240-017-1207-6

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the $2^{\wedge}$(-delta delta CT) method. *Methods.* 25 (4), 402–408. doi: 10.1006/meth.2001.1262

Ning, K., Chen, S., Huang, H. J., Jiang, J., Yuan, H. M., and Li, H. Y. (2017). Molecular characterization and expression analysis of the *SPL* gene family with *BpSPL9* transgenic lines found to confer tolerance to abiotic stress in betula platyphylla suk. *Plant Cell Tiss Organ Cult.* 130 (3), 469–481. doi: 10.1007/s11240-017-1226-3

Ren, H. Y., Yu, Z. P., Zhang, S. W., Liang, S. M., Zheng, X. L., Zhang, S. J., et al. (2019). Genome sequencing provides insights into the evolution and antioxidant activity of Chinese bayberry. *BMC Genomics* 20 (1), 458. doi: 10.1186/s12864-019-5818-7

Salinas, M., Xing, S. P., Höhmann, S., Berndtgen, R., and Huijser, P. (2012). Genomic organization, phylogenetic comparison and differential expression of the SBP-box family of transcription factors in tomato. *Planta.* 235, 1171–1184. doi: 10.1007/s00425-011-1565-y

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28 (5), 511–515. doi: 10.1038/nbt.1621

Vander Schoor, J. K., Hecht, V., Aubert, G., Burstin, J., and Wellerd, J. L. (2022). Defining the components of the miRNA156-SPL-miR172 aging pathway in pea and their expression relative to changes in leaf morphology. *Plant Gene* 30, 100354. doi: 10.1016/j.plgene.2022.100354

Wang, J. W., Park, M. Y., Wang, L. J., Koo, Y. J., Chen, X. Y., Weigel, D., et al. (2018). MiRNA control of vegetative phase change in trees. *PloS Genet.* 13 (7), e0200762. doi: 10.1371/journal.pgen.1002012

Wang, J. W., Schwab, R., Czech, B., Mica, E., and Weigei, D. (2008). Dual effects of miR156-targeted *SPL* genes and *CYP78A5/KLUH* on plastochron length and organ size in *Arabidopsis thaliana*. *Plant Cell.* 20 (5), 1231–1243. doi: 10.1105/tpc.108.058180

Wang, L., Zhou, C. M., Mai, Y. X., Li, L. Z., Gao, J., Shang, G. D., et al. (2019). A spatiotemporally regulated transcriptional complex underlies heteroblastic development of leaf hairs in *Arabidopsis thaliana*. *EMBO J.* 38 (8), e100063. doi: 10.15252/embj.2018100063

Wang, Z. S., Wang, Y., Kohalmi, S. E., Amyot, L., and Hannoufa, A. (2016). SQUAMOSA PROMOTER BINDING PROTEIN-LIKE 2 controls floral organ development and plant fertility by activating *ASYMMETRIC LEAVES 2* in *Arabidopsis thaliana*. *Plant Mol. Biol.* 92, 661–674. doi: 10.1007/s11103-016-0536-x

Wu, G., Mee, Y. P., Susan, R. C., Wang, J. W., Detlef, W., and R. Scott, P. (2009). The sequential action of miR156 and miR172 regulates developmental timing in *Arabidopsis*. *Cell.* 138 (4), 750–759. doi: 10.1016/j.cell.2009.06.031

Xiong, J. S., Bai, Y., Ma, C. J., Zhu, H. Y., Zheng, D., and Cheng, Z. M. (2019). Molecular cloning and characterization of *SQUAMOSA*-promoter binding protein-like gene *FvSPL10* from woodland strawberry *(Fragaria vesca*). *Plants.* 8 (9), 342. doi: 10.3390/plants8090342

Xiong, J. S., Zheng, D., Zhu, H. Y., Chen, J. Q., Na, R., and Cheng, Z. M. (2018). Genome-wide identification and expression analysis of the *SPL* gene family in woodland strawberry *Fragaria vesca*. *Genome.* 61 (9), 1–9. doi: 10.1139/gen-2018-0014

Yang, J., Guo, Z. L., Wang, W. T., Cao, X. Y., and Yang, X. Z. (2021). Genome-wide characterization of SPL gene family in *Codonopsis pilosula* reveals the functions of *CpSPL2* and *CpSPL10* in promoting the accumulation of secondary metabolites and growth of c. *pilosula* hairy root. *Genes.* 12 (10), 1588. doi: 10.3390/genes12101588

Yang, Z. F., Wang, X. F., Gu, S. L., Hu, Z. Q., Xu, H., Xu, C. W., et al. (2008). Comparative study of SBP-box gene family in arabidopsis and rice. *Gene* 407 (1-2), 1–11. doi: 10.1016/j.gene.2007.02.034

Yu, N., Niu, Q. W., Ng, K. H., and Chua, N. H. (2015). The role of miR156/SPLs modules in arabidopsis lateral root development. *Plant J.* 83 (4), 673–685. doi: 10.1111/tpj.12919

Zhang, S. W., Yu, Z. P., Sun, L., Ren, H. Y., Zheng, X. L., Liang, S. M., et al. (2022). An overview of the nutritional value, health properties, and future challenges of Chinese bayberry. *PeerJ.* 10, e13070. doi: 10.7717/peerj.13070

Zhao, Q., Fan, Z. H., Qiu, L., Che, Q. Q., Wang, T., Li, Y. Y., et al. (2020). *MdbHLH130*, an apple bHLH transcription factor, confers water stress resistance by regulating stomatal closure and ROS homeostasis in transgenic tobacco. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.543696

Zhu, F. Y., Wang, S. L., Xue, J. Q., Li, D. D., Ren, X. X., Xue, Y. Q., et al. (2018). Morphological and physiological changes, and the functional analysis of *PdSPL9* in the juvenile-to-adult phase transition of *paeonia delavayi*. *Plant Cell Tiss Organ Cult.* 133 (3), 325–337. doi: 10.1007/s11240-018-1384-y

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 03 April 2023
DOI 10.3389/fpls.2023.1166717

OPEN ACCESS

EDITED BY
Zhichao Wu,
National Institutes of Health (NIH),
United States

REVIEWED BY
Yang Yu,
Institute of Crop Sciences (CAAS), China
Chang Yuansheng,
Shandong Academy of Agricultural Sciences Shandong Institute of Pomology, China
Changwei Bi,
Nanjing Forestry University, China

*CORRESPONDENCE
Xuan Wang
newwxuan@163.com

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

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

RECEIVED 15 February 2023
ACCEPTED 21 March 2023
PUBLISHED 03 April 2023

CITATION
Zhang P, Liu J, Jia N, Wang M, Lu Y, Wang D, Zhang J, Zhang H and Wang X (2023) Genome-wide identification and characterization of the *bZIP* gene family and their function in starch accumulation in Chinese chestnut (*Castanea mollissima* Blume). *Front. Plant Sci.* 14:1166717. doi: 10.3389/fpls.2023.1166717

Genome-wide identification and characterization of the *bZIP* gene family and their function in starch accumulation in Chinese chestnut (*Castanea mollissima* Blume)

Penglong Zhang[1,2†], Jing Liu[1,2†], Nan Jia[3], Meng Wang[1], Yi Lu[1], Dongsheng Wang[2], Jingzheng Zhang[2], Haie Zhang[1,2] and Xuan Wang[1,2*]

[1]Engineering Research Center of Chestnut Industry Technology, Ministry of Education, Qinhuangdao, Hebei, China, [2]Hebei Key Laboratory of Horticultural Germplasm Excavation and Innovative Utilization, College of Horticulture Science and Technology, Hebei Normal University of Science and Technology, Changli, Hebei, China, [3]Changli Institute of Pomology, Hebei Academy of Agriculture and Forestry Science, Changli, Hebei, China

The transcription factors of basic leucine zipper (bZIP) family genes play significant roles in stress response as well as growth and development in plants. However, little is known about the *bZIP* gene family in Chinese chestnut (*Castanea mollissima* Blume). To better understand the characteristics of *bZIP*s in chestnut and their function in starch accumulation, a series of analyses were performed including phylogenetic, synteny, co-expression and yeast one-hybrid analyses. Totally, we identified 59 *bZIP* genes that were unevenly distributed in the chestnut genome and named them *CmbZIP01* to *CmbZIP59*. These *CmbZIP*s were clustered into 13 clades with clade-specific motifs and structures. A synteny analysis revealed that segmental duplication was the major driving force of expansion of the *CmbZIP* gene family. A total of 41 *CmbZIP* genes had syntenic relationships with four other species. The results from the co-expression analyses indicated that seven *CmbZIP*s in three key modules may be important in regulating starch accumulation in chestnut seeds. Yeast one-hybrid assays showed that transcription factors CmbZIP13 and CmbZIP35 might participate in starch accumulation in the chestnut seed by binding to the promoters of *CmISA2* and *CmSBE1_2*, respectively. Our study provided basic information on *CmbZIP* genes, which can be utilized in future functional analysis and breeding studies

KEYWORDS

***Castanea mollissima*, bZIP, gene family, starch accumulation, yeast one-hybrid, CmbZIP13, CmbZIP35**

1 Introduction

Starch is an important form of carbon storage for the majority of plant species. Throughout the lifecycle of a plant, starch plays roles in development and response to the environment (MacNeill et al., 2017). Short-term storage of starch occurs in the leaves of plants, whereas long-term storage takes place in seeds and tubers, providing material for energy, development, and reproduction (MacNeill et al., 2017). A previous study has reported that starch can be degraded by a kind of α-amylase in response to osmotic stress (Thalmann et al., 2016). In many plants, the starch biosynthesis pathway is catalyzed by enzymes, such as ADP-glucose pyrophosphorylase, starch synthase, starch branching enzyme, and starch de-branching enzyme (Qu et al., 2018), and regulated by transcription factors (TFs) (MacNeill et al., 2017). For example, as an AP2/EREBP TF family member, rice starch regulator 1 negatively regulates the expression of starch synthesis-related genes in rice seeds and is involved in the amylose content of the seed (Fu and Xue, 2010). The endosperm-specific TF TaNAC019 can bind to the promoters of starch metabolism genes, regulate starch accumulation, and improve the quality of wheat grains (Gao et al., 2021).

TFs play an indispensable role in plant growth, development, and resistance to biotic and abiotic stresses (Wang et al., 2022). As one of the largest and most diverse TF families, the basic leucine zipper (bZIP) family has been studied extensively (Lee et al., 2006; Schlögl et al., 2008; Wang et al., 2013; An et al., 2018; Song et al., 2020; Duan et al., 2022). The members of the bZIP TF family harbor a highly conserved, 60–80 amino acid long domain, which is composed of a basic region and a leucine zipper region. The sequence of the basic region consists of approximately 20 amino acid residues with a fixed nuclear localization structure N-x7-R/K, which can specifically bind to DNA cis-elements (Lee et al., 2006; Nijhawan et al., 2008). The leucine zipper region, containing the core sequences of L-x6-L-x6-L, consists of various repetitions of leucine or other hydrophobic amino acids, which facilitates hetero- or homo-dimerization of bZIP proteins (Jakoby et al., 2002). The core sequence recognized by bZIP TFs is ACGT, which includes an A-box (TACGTA), C-box (GACGTC), and G-box (CACGTG) (Izawa et al., 1993). Most abscisic acid induced genes have these ACGT cis-elements in their promoter regions. In addition, bZIP TFs can also recognize non-palindrome sequences, such as H-box (CCTACC), GCN4-like motif (GTGAGTCAT), and prolamin box-like (TGAAAA) elements (Kim et al., 2014).

The bZIP TF genes play an important role in many plant biological processes, such as regulating plant morphology and growth. For example, over-expression of the pepper *CabZIP1* gene in *Arabidopsis* slowed plant growth and reduced the number of petals (Lee et al., 2006). The bZIP TFs of tobacco regulate its transition from vegetative growth to reproductive growth (Heinekamp et al., 2002). Plant bZIP transcription factors are induced by plant hormones such as salicylic acid, methyl jasmonate, ethylene, or abscisic acid (Meng et al., 2005; Lee et al., 2006; Schlögl et al., 2008). In addition, many bZIP transcription factors can regulate abscisic acid synthesis, which regulates gene expression (Finkelstein and Lynch, 2000; Uno et al., 2000). bZIP family genes participate in the regulation of plant resistance to biotic and/or abiotic stresses. Overexpression of bZIP-like proteins in plants under stress conditions can improve the photosynthetic capacity of plants, improving their resistance to salt, cold, herbicides, drought, heat, and other stresses (Kim et al., 2004; Lee et al., 2006; Liao et al., 2008; Zhang et al., 2008). For example, the silencing of the rice endogenous *rT-GA2.1* gene (a member of the bZIP family) mediated by dsRNA can improve the resistance of rice to bacterial pathogens, such as *Xanthomonas oryzae* pv. oryzae, indicating that *rTGA2.1* plays a negative role in response to bacterial pathogens (Fitzgerald et al., 2005). Some members of the bZIP family also act as regulators in the starch synthesis pathway. In rice, OsbZIP20, OsbZIP33, and OsbZIP58 TFs interact with *granule-bound starch synthase* (*GBSS*) and *starch branching enzyme 1* (*SBE1*) genes by binding to their promoters, and are capable of regulating starch synthesis (Cai et al., 2002; Wang et al., 2013). In wheat, *TubZIP28* (from *Triticum urartu*) and *TabZIP28* (from *Triticum aestivum*) also participate in the regulation of starch synthesis by interacting with starch synthesis-related genes (Song et al., 2020)

Many bZIP TF families have been identified in important plant species. For example, 75 *bZIP* genes have been predicted in *Arabidopsis thaliana* (Jakoby et al., 2002). One hundred twenty-five members of the *bZIP* genes family were identified in maize (*Zea mays*) (Wei et al., 2012), 114 in apple (*Malus domestica*) (Li et al., 2016), 92 in pear (*Pyrus breschneideri*) (Ma et al., 2021), 77 in tobacco (*Nicotiana tabacum*) (Duan et al., 2022), 65 in pomegranate (*Punica granatum*) (Wang et al., 2022), and 227 in wheat (*Triticum aestivum*) (Liang et al., 2022). However, there are no reports on the identification and functionality of *bZIP* genes in Chinese chestnut (*Castanea mollissima*), despite Chinese chestnut being an economically important dry fruit tree species that is favorited for its sweet, fragrant, and waxy characteristics. The waxiness is a critical parameter for chestnut quality and is determined by the starch content in the kernel (Lin et al., 2012; Shi et al., 2021). Over the past four years, several versions of the *C. mollissima* genome have been assembled and published (Xing et al., 2019; Sun et al., 2020; Wang et al., 2020; Hu et al., 2022). These assemblies are resources for the identification of gene families and genetic improvement of chestnut.

In this study, we aimed to identify *bZIP* genes from Chinese chestnut using the whole genome of N11-1, a seedling Chinese chestnut cultivar line (Wang et al., 2020). We performed phylogenetic, conserved motif, and gene structure analyses to study the relationships between the identified *CmbZIP* family members. Whole genome duplication (WGD) analysis revealed that segmental duplication might be the main factor that led to the expansion of the *CmbZIP* family. Transcriptomic data from different development stages of chestnut seeds showed that *CmbZIP* genes had different expression patterns, and some of them may be related to starch accumulation. These results from this study provide a theoretical reference for *CmbZIP* genes and insight useful for future studies on Chinese chestnut.

2 Results

2.1 Identification and characterization of *CmbZIP* genes

In this study, we identified 59 *bZIP* gene family members from the whole genome of the seedling Chinese chestnut cultivar N11-1, which version provided relatively complete annotative information at the chromosome level (Wang et al., 2020). For subsequent analysis, we named these genes *CmbZIP01* to *CmbZIP59* based on the chromosome and/or contig location (Table S1). The molecular weight of chestnut bZIP family proteins ranged from 16,067.34 to 122,096.53 Da, the theoretical isoelectric point ranged from 4.62 to 9.76, and the protein length ranged from 140 (CmbZIP52) to 1,075 aa (CmbZIP28). Fifty-seven CmbZIP proteins were located in the nuclear region, whereas CmbZIP35 and CmbZIP41 were located in the chloroplast and vacuole, respectively (Table S1). These results provide a theoretical basis for further purification, activity, and functional studies of CmbZIP proteins.

2.2 Phylogenetic analysis and classification of the chestnut bZIP TF family

To explore the homologous evolutionary relationships and classification of the bZIP family, we constructed an unrooted neighbor-joining phylogenetic tree using bZIP protein sequences from chestnut, *Arabidopsis*, and five other bZIP proteins reported in previous studies: OsbZIP20 (Izawa et al., 1994), OsbZIP33 (Nakase et al., 1997), OsbZIP58 (Wang et al., 2013), TubZIP28, and TabZIP28 (Song et al., 2020). The 59 CmbZIP proteins were divided into 13 clades (A, B, C, D, E, F, G, H, I, J, K, L, and Un) according to their homology in *Arabidopsis* (Jakoby et al., 2002) (Figure 1). The number of CmbZIP proteins in the 13 clades differed greatly in size. The largest clade (H) had 13 members. CmbZIP24 was clustered into the smallest, unique clade (Un) in the phylogenetic tree and might have an evolutionary trajectory unrelated to other clades. Two of the clades had only one CmbZIP TF member: clade I and clade Un (Figure 1; Table S1). The bZIP proteins of the four species included in this analysis were separately distributed throughout the 13 clades in the phylogenetic tree, indicating that the bZIP proteins showed similar divergences in gene function in chestnut, *Arabidopsis*, rice, and wheat. Some bZIP proteins clustered together in a small clade, suggesting that a co-speciation event and species-specific duplication events occurred during the divergence of the bZIP TF family. Our analysis revealed that three homologous proteins, CmbZIP16, CmbZIP17, and CmbZIP58, in clade D and CmbZIP41 in clade I, were able to influence starch accumulation, which is similar to the evolutionary relationships in *Arabidopsis*.

2.3 Conserved motif and structure analyses of chestnut *bZIP* genes

In order to study the characteristics of the 59 CmbZIP proteins, we identified 20 conserved motifs varying from eight (motif 8 and 14) to 100 (motif 2, 9, 10, 12 and 13) aa residues long (Figure 2; Table S2). Motifs 1 and 3 were widely distributed in tandem in almost all (55 of 59) CmbZIP proteins; further sequence analysis indicated that these two motifs constitute the DNA binding basic region and the leucine zipper region of the bZIP domain, respectively (Figure S1). In addition, the distribution of 16 motifs showed clade specificity in the phylogenetic tree presented in Figure 2 (Table S3). Motifs 2, 4, 7, and 17 formed two dimers (motifs 2–7 and motifs 17–4), which were present in seven members of clade L. Motif 17 was identified in clade A. Similarly, motif 11 was present in both clades B and I. Motif 5 was only distributed in clade H, except for CmbZIP08, which contained this motif one to five times. Furthermore, motifs 6 and 15 that had a similar pattern were predicted in all members of clades E and F; these two clades were sister to each other in the phylogenetic tree (Figure 2). Eight other motifs were specifically distributed in clade B (motifs 18, 19, and 20), C (motif 12), and D (motifs 9, 10, 13, and 16) (Figure 2).

For further insight into the evolution of *bZIP* genes in chestnut, we compared the DNA sequences and examined the organization of exons and introns in open reading frames of *CmbZIP* genes. In total, the number of introns in the *CmbZIP*s ranged from 0 to 13. Fifteen *CmbZIP*s contained three introns, accounting for the largest proportion of identified *bZIP* genes (25.4%). *CmbZIP15* (with eight introns) and *CmbZIP53* (with 13 introns), respectively, accounted for the smallest proportion of identified *bZIP* genes (1.6%) (Table S1). As expected, members of the same clade had relatively conservative numbers of introns. Seven *CmbZIP*s, all members of clade A, did not have any introns. All *bZIP* genes in clade F contained three introns. Furthermore, the number of introns in clade B varied from five to 13, two to seven in H, and seven to 11 in clade L (Figure 2C; Table S1). Overall, similar exon–

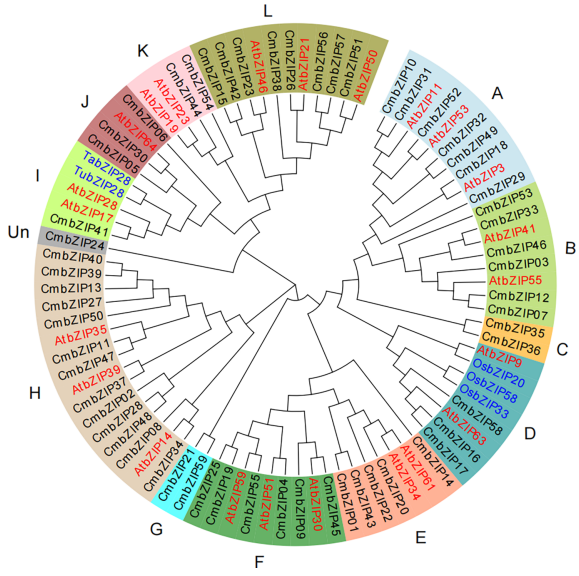


FIGURE 1
The phylogenetic tree of bZIP proteins in chestnut, *Arabidopsis*, rice, and wheat. The proteins are clustered in 13 clades (A–L, Un) by pre-grouping with AtbZIPs (Jakoby et al., 2002). Different background colors indicate different bZIP protein clades. The black, red, and blue font indicate the bZIP proteins of chestnut, *Arabidopsis*, and rice and wheat, respectively.

10.3389/fpls.2023.1166717

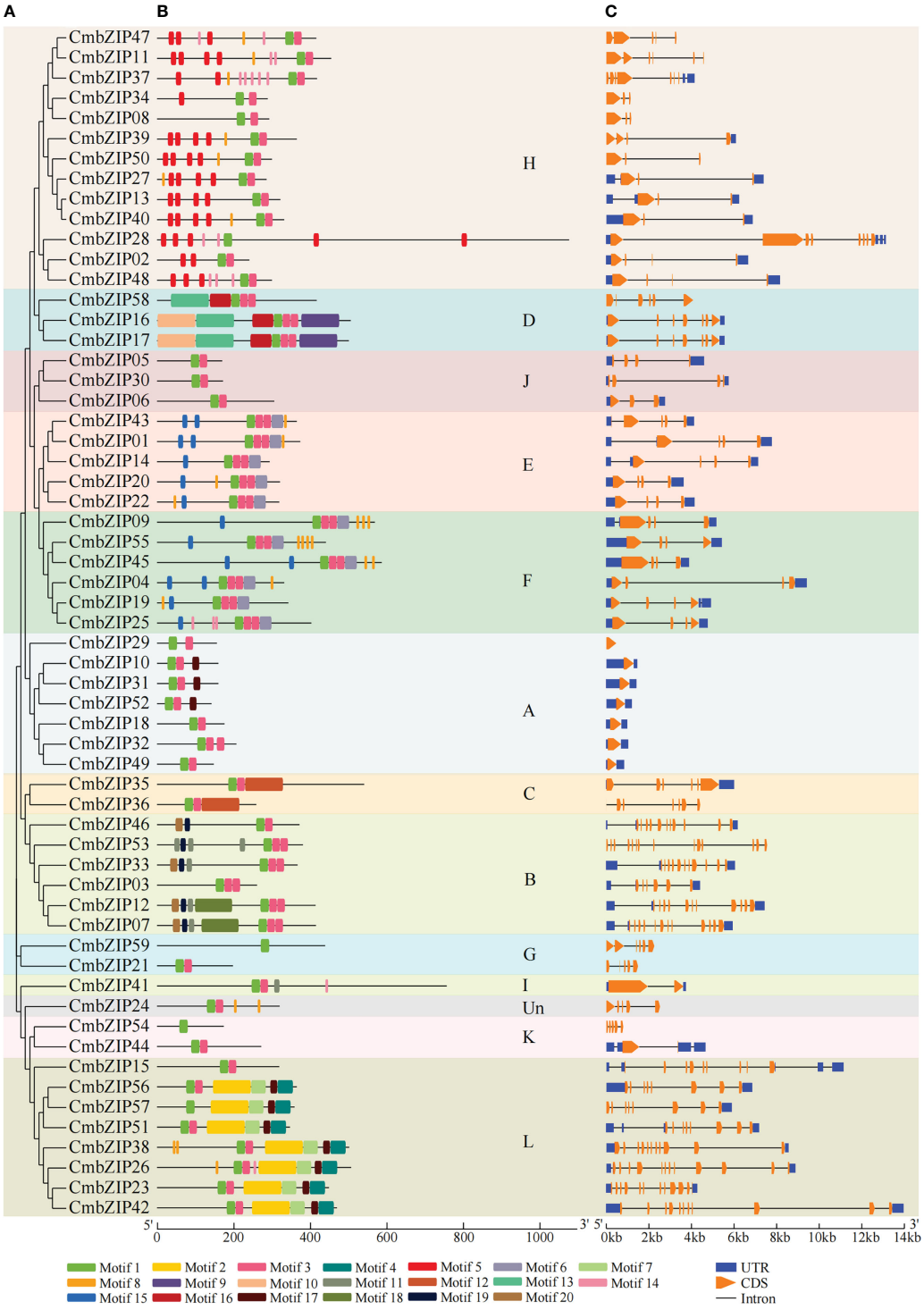


FIGURE 2
Conserved motif patterns and structure schematics of CmbZIPs. **(A)** The phylogenetic tree was derived from 59 CmbZIP proteins. **(B)** Conserved motif analysis of CmbZIP proteins with 20 separate patterns depicted with different colors. **(C)** Analysis of *CmbZIP* gene structure. The clade names A–L and Un are the same as in Figure 1.

intron structure and motif composition were observed for *bZIP* genes of chestnut in the same clade; the structure and motif composition differ between clades, illustrating that the evolution and divergence of *CmbZIP*s might have occurred at an early stage of the evolution of *C. mollisima*.

2.4 Chromosome location and duplication of *bZIP*s in chestnut

Fifty-nine *CmbZIP* genes were distributed unevenly on 11 of the 12 chromosomes (Chr) of chestnut as well as five contigs (Ctg)

(Figure 3; Table S1). Chr02 did not contain any *CmbZIP* genes. There was only one *bZIP* gene on the following chromosomes: Chr07 (*CmbZIP28*), Chr10 (*CmbZIP43*), Ctg1 (*CmbZIP54*), Ctg2 (*CmbZIP55*), Ctg3 (*CmbZIP56*), and Ctg4 (*CmbZIP57*); there were two *bZIP* genes on Ctg5 (*CmbZIP58* and *CmbZIP59*). Eleven *CmbZIP* genes (18.64%) were located on Chr01, which contained the greatest number of *bZIP* genes, with 1 and 10 *CmbZIP*s, respectively, located on the proximate and distal ends of this chromosome. Five *CmbZIP*s were relatively evenly dispersed throughout Chr06. Three to eight *bZIP* genes were located on the remaining eight chromosomes.

We detected 10 pairs of segmental duplications in *CmbZIP* genes, and no tandem duplications were observed, indicating that segmental duplication events were the major cause of expansion of the *CmbZIP* family (Figure 4; Table 1). Chromosomal distribution analysis revealed that the 20 analogous *CmbZIP*s were unevenly located on the Chinese chestnut genome. We also found that every pair of duplicated *CmbZIP*s were in clades A, B, D, E, F, and H (Table 1). Furthermore, we calculated the synonymous substitution rate (Ks) to estimate the segmental duplication events for *CmbZIP*s (Table 1). The divergence time of all duplicated *CmbZIP*s varied greatly from 9.50–116.09 million years ago (Mya). To predict the selection pressure driving the divergence of *CmbZIP*s, we also calculated the nonsynonymous substitution rate (Ka) and the Ka/Ks ratio. Seven pairs of duplicated *CmbZIP*s might have undergone purifying selection from 27.74–116.09 Mya. The selection pressure on *CmbZIP20/22* was the strongest (Ka/Ks=0.06). Conversely, *CmbZIP09/45* and *CmbZIP19/25* might have recently undergone positive selection from 9.50–11.65 Mya.

2.5 Synteny analysis of *bZIP*s between genomes

To gain deeper insight into the evolutionary relationships of the *bZIP* genes family among different species, we constructed four comparative syntenic maps of chestnut associated with *Arabidopsis*, rice, wheat, and apple (Figure 5; Table S4). In total, there were 41 orthologous *bZIP* gene pairs between chestnut and the other four species. Further, we found 31 *CmbZIP*s associated with at least two syntenic gene pairs, suggesting that these genes might play an important role in the evolutionary process of the *bZIP* family. Thirty-four *CmbZIP*s showed syntenic relationships with apple *bZIP* genes, 27 with *Arabidopsis*, five with rice, and one with wheat. There was a far greater number of syntenic *bZIP* pairs between chestnut and the two dicots (i.e., apple and *Arabidopsis*) than between chestnut and the two monocots (i.e., rice and wheat), which might indicate that most of these orthologous pairs occurred after the divergence of dicotyledons and monocotyledons.

2.6 Analysis of expression patterns and identification of *CmbZIP*s related to starch accumulation

To confirm the expression patterns of *CmbZIP* genes related to starch synthesis, we used published transcriptome data of all genes from the N11-1 version of the reference genome to determine the fragments per kilobase transcript per million mapped reads (FPKM) (Table S5). All *CmbZIP* genes were clustered to four

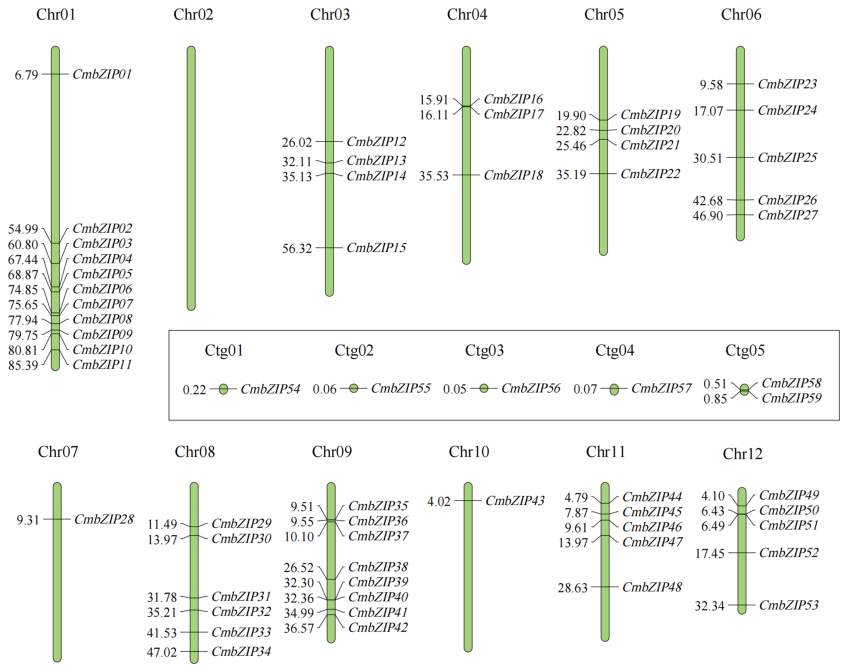


FIGURE 3
Chromosome locations of *CmbZIP* genes. Vertical green bars represent the chromosomes of chestnut. The chromosome number is stated at the top of each chromosome. The numbers on the left of vertical bars indicate the corresponding physical positions of *CmbZIP*s. The locations of *CmbZIP*s on the contigs are shown in the rectangle.

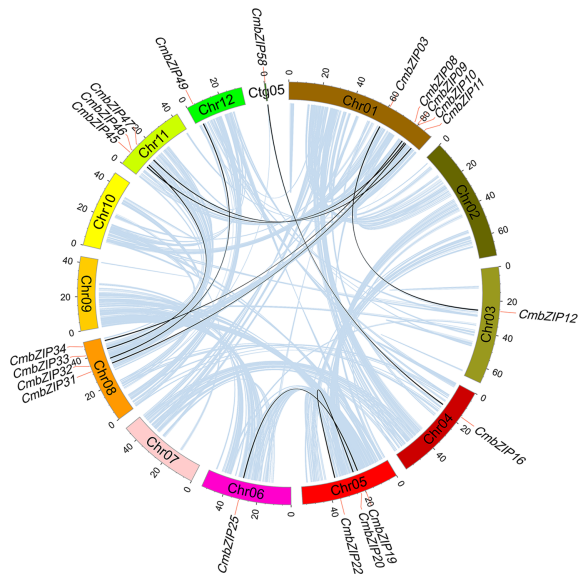


FIGURE 4
The syntenic pairs of *CmbZIP* genes from different modes of duplication events. Colored rectangles represent chromosome, and the chromosome number is inside these rectangles. Light blue lines indicate all synteny blocks in the chestnut genome; black lines indicate duplicated *bZIP* gene pairs. Names of *bZIP* genes are marked at the position corresponding to the red lines.

subclades by their expression profiles (Figure 6A). The greatest number of *CmbZIP*s were in subclade II, but almost all members (10/19) had very low levels of expression (FPKM < 0.5) in every sample. The 15 members of subclade IV had relatively high expression levels, with an average FPKM ranging from 29.1 to 130.3 (Table S5).

Furthermore, we performed a correlation analysis between expression patterns of 59 *CmbZIP*s and four physiological characteristics: total starch content, amylopectin content, amylose content, and starch synthase activity (Figures 6B, S2; Table S6). Fourteen highly expressed *CmbZIP*s, namely *CmbZIP33*, *CmbZIP47*, *CmbZIP14*, *CmbZIP38*, *CmbZIP48*, *CmbZIP56*, *CmbZIP35*, *CmbZIP04*, *CmbZIP39*, *CmbZIP40*, *CmbZIP45*, *CmbZIP06*, *CmbZIP13*, and *CmbZIP07*, were identified to be significantly associated ($|r| \geq 0.67$, $p < 0.05$) with starch accumulation (content of total starch or amylopectin) or activity of starch synthase during chestnut seed development (Figure 6B; Table S6).

2.7 Identification of co-expression networks related to starch accumulation

To investigate the co-expression networks related to starch accumulation, eight modules with gene numbers ranging from 921 to 4,890 were identified by weighted gene co-expression network analysis (WGCNA) (Figure S3; Table S5). Analysis of module–trait relationships revealed that three key modules (MEred, MEgreen, and MEbrown) were significantly associated with starch accumulation, with $|r| \geq 0.6$ and $p < 0.05$ (Figure 7A). In detail, MEred module was negatively related to contents of total starch ($r = -0.79$, $p = 0.01$) and amylopectin ($r = -0.8$, $p = 0.01$), and starch synthase activity ($r = -0.82$, $p = 0.006$). In contrast, MEbrown module was positively related to contents of total starch ($r = 0.73$, $p = 0.03$) and amylopectin ($r = 0.76$, $p = 0.02$), and starch synthase activity ($r = 0.76$, $p = 0.02$). MEgreen module was also related to amylopectin content ($r = 0.69$, $p = 0.04$) and starch synthase activity ($r = 0.86$, $p = 0.003$) positively, but not with total starch content. We further identified seven starch accumulation related genes co-expressing with five *CmbZIP*s (*CmbZIP04*, *CmbZIP14*, *CmbZIP33*, *CmbZIP38*, and *CmbZIP56*) in the MEbrown module. Six starch accumulation related genes were found to be co-expressing with one *CmbZIP* (*CmbZIP35*) in the MEgreen module. Two starch accumulation related genes were found to be co-expressing with

TABLE 1 Estimation of the date of segmental duplication events for *CmbZIP*s.

Duplicated gene pairs		Clade		Ka	Ks	Ka/Ks	Divergence time (Mya)
gene 1	gene 2	gene 1	gene 2				
CmbZIP20	*CmbZIP22*	E	E	0.22	3.48	0.06	116.09
CmbZIP03	*CmbZIP12*	B	B	0.55	2.42	0.23	80.65
CmbZIP08	*CmbZIP34*	H	H	0.39	2.29	0.17	76.43
CmbZIP49	*CmbZIP32*	A	A	0.55	1.60	0.35	53.17
CmbZIP46	*CmbZIP33*	B	B	0.48	1.43	0.34	47.79
CmbZIP10	*CmbZIP31*	A	A	0.21	1.35	0.15	45.01
CmbZIP11	*CmbZIP47*	H	H	0.47	0.83	0.56	27.74
CmbZIP58	*CmbZIP16*	D	D	1.01	0.96	1.05	32.08
CmbZIP09	*CmbZIP45*	F	F	0.57	0.35	1.63	11.65
CmbZIP19	*CmbZIP25*	F	F	0.47	0.29	1.64	9.50

*Ka, the nonsynonymous substitution rate; Ks, the synonymous substitution rate; Mya, million years ago.

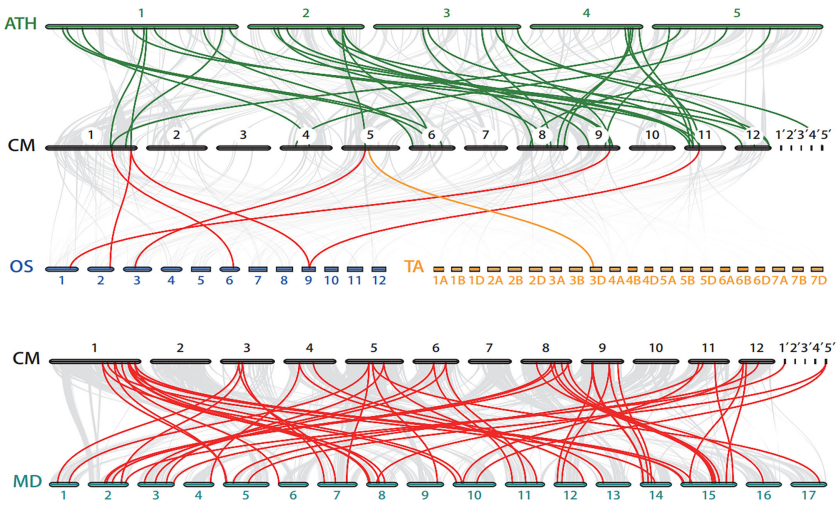


FIGURE 5
Synteny analysis of the *bZIP* genes between chestnut and four other plant species. The grey lines in the background indicate the collinear blocks between the chestnut genome and other genomes. Bold, colored lines highlight the syntenic *bZIP* gene pairs. The colored bars represent chromosomes of different species. 1'-5': Ctg1-Ctg5, contigs in chestnut genome. CM, *Castanea mollissima*; ATH, *Arabidopsis thaliana*; OS, *Oryza sativa*; TA, *Triticum aestivum*; MD, *Malus domestica*.

one *CmbZIP* (*CmbZIP13*) in the MEred module (Figure 7B; Table S7). The expression profiles of these genes were evaluated by FPKM (Figure 7C).

bZIP transcription factors can recognize ACGT sequences in gene promoter regions, particularly A-box (TACGTA), C-box (GACGTC), and G-box (CACGTG) sequences (Izawa et al., 1993). Interestingly, all starch accumulation related genes from key modules contained at least one ACGT sequence in the promoter region, except for *CmPUL1* (Figure S5; Table S7). A-box occurred at -733 bp and -1,377 bp in promoters of *CmSBE1_2* and *CmAA3*, respectively. C-box was only predicted in the promoter region of *CmBA1_1*, with the positions of -726 bp and -294 bp. G-box was the most frequently identified ACGT sequence in the promoter regions of *CmBA1_1* (-240 bp, -98 bp, and -75 bp), *CmBA4_1* (-289 bp), *CmISA2* (-771 bp), and *CmSSS3* (-648 bp and -595 bp). Based on these findings, we can infer that *CmbZIP*s may regulate the expression of starch accumulation related genes through ACGT cis-elements, and then participate in the accumulation of starch in chestnut seeds.

2.8 Identification of interactions of *CmbZIP*s with promoters of starch accumulation related genes

We selected the only two *bZIP* genes in MEgreen and MEred by our interest, *CmbZIP35* and *CmbZIP13*, to identify their potential functions of binding to promoters of starch accumulation related genes. In yeast one-hybrid (Y1H) assays, the Y1H Gold yeast strains containing pCmSBE1_2-AbAi×pGADT7-CmbZIP35 were able to grow on the screening synthetic dropout medium lacking uracil and leucine (SD-UL) containing 100 ng/mL Aureobasidin A (AbA) (Figure 8A). Similarly, the Y1H Gold yeast strains containing pCmISA2-AbAi×pGADT7-CmbZIP13 were able to grow under the same conditions (Figure 8B). These observations suggested that CmbZIP35 can directly bind to the promoter of *CmSBE1_2*, and CmbZIP13 can bind to the promoter of *CmISA2*.

3 Discussion

Previous studies have shown that the bZIP TF family plays an important role in the regulation of plant growth and development as well as resistance to biotic and abiotic stresses (Lee et al., 2006; Schlögl et al., 2008; Song et al., 2020; Wang et al., 2022). Numerous studies have been conducted on bZIP TFs, such as in *Arabidopsis* (Gibalova et al., 2017), wheat (Song et al., 2020), rice (Wang et al., 2013), tobacco (Duan et al., 2022), and apple (An et al., 2018). Although several versions of the Chinese chestnut genome have been published (et al., 2019; Sun et al., 2020; Wang et al., 2020; Hu et al., 2022), to our knowledge, there have been no published studies on the function of bZIP TFs in chestnut. In this study, we identified 59 *bZIP* genes in the chestnut genome, which has a complete genome size of 689.98 Mb (Wang et al., 2020), and further analyzed the characteristics of the *bZIP* genes.

The bZIP domain consists of a basic region and a leucine zipper region (Jakoby et al., 2002). In the present study, we identified the two structural features *via* analyzing conserved motifs, and they were named motif 1 and motif 3 (Figure S1); this is similar to the bZIPs in tobacco (Duan et al., 2022). We also identified another 18 motifs in the 59 CmbZIP proteins (Figure 2B). Strong clade-specificity and conservation were detected in the motif distribution and phylogenetic analyses of CmbZIPs, and the findings were similar to those reported for pomegranate (Wang et al., 2022), pear (Ma et al., 2021), tobacco (Duan et al., 2022), wheat (Liang et al., 2022), and apple (Li et al., 2016). The bZIPs containing the same motifs might have similar functions (Wang et al., 2022). For example, all members in clade D, which contained

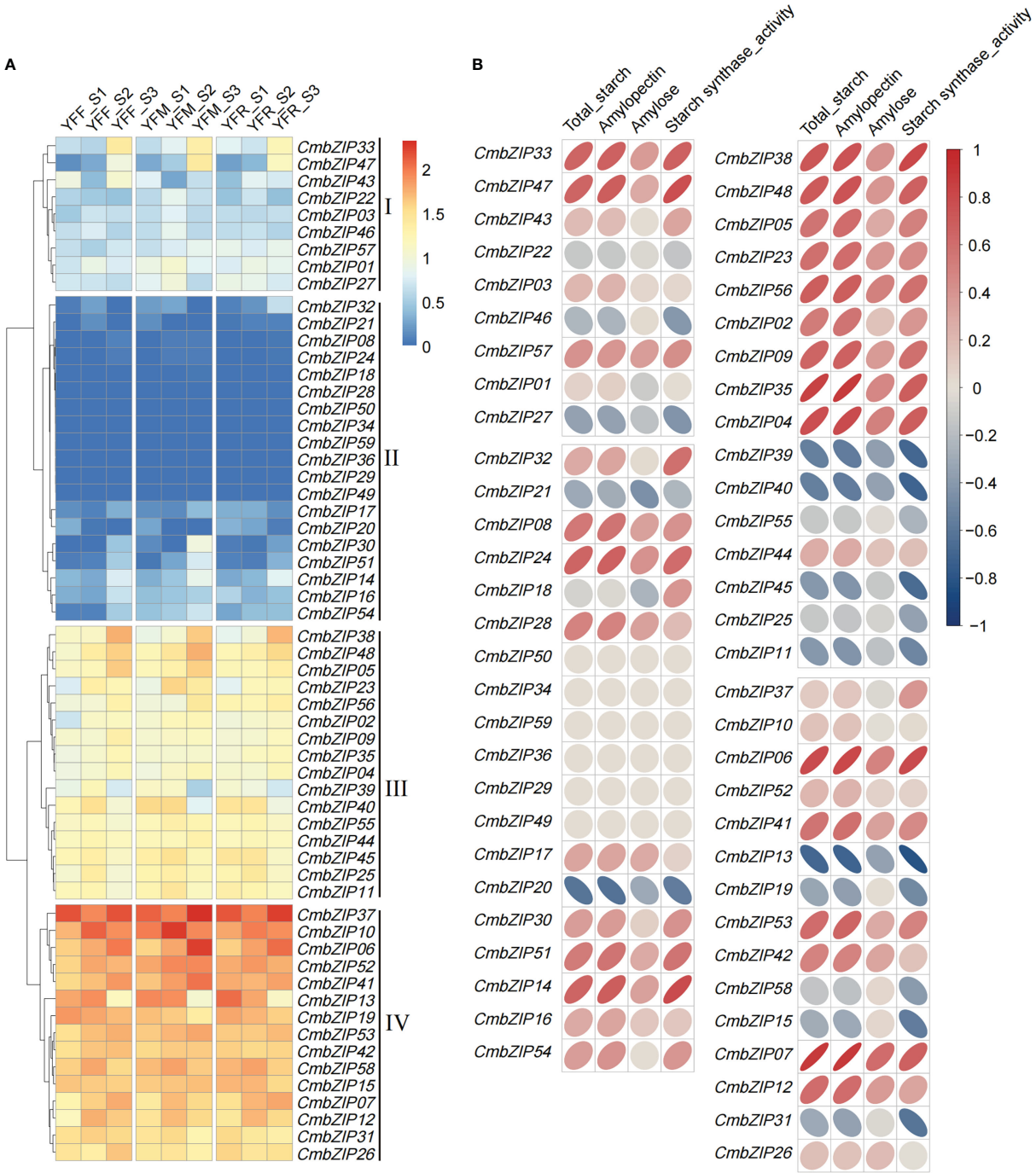


FIGURE 6
Expression pattern of 59 *CmbZIP* genes and correlation analysis of *CmbZIP* expression and physiological characteristics of three chestnut crosses (Li et al., 2021). **(A)** The heatmap shows the expression pattern of *CmbZIP* genes. YFF, YFM, and YFR indicate seeds from crosses of 'Yongfeng 1'×'Yongfeng 1,' 'Yongfeng 1'×'Yimen 1,' and 'Yongfeng 1'×'Yongren Zao,' respectively. S1, S2, and S3 indicate 70, 82, and 94 days after pollination, respectively. The Roman numerals along the right-hand side of the figure indicate $\log_{10}$ FPKM. **(B)** The correlation coefficients between the expression of *CmbZIP*s and four traits are shown by elliptical bubbles. The flatter the bubble, the higher the correlation. Dark red (from top right to bottom left) and dark blue (from top left to bottom right) ellipses represent positive and negative correlations, respectively. The numbers in the legend indicate the correlation coefficient.

motifs 13 and 16, encode the light-inducible protein CPRF2; and most members of clade H, with motifs 5 and 8, may participate in the pathways responding to abiotic stress by encoding ABSCISIC ACID-INSENSITIVE 5-like proteins (Figure 2B; Table S1). In addition, clade-specificity was observed in the number of introns (Figure 2C; Table S1), similar to previous studies (Duan et al., 2022; Liang et al., 2022; Wang et al., 2022). These analyses suggested that conserved motifs and gene structure were critical for members in same clade during evolution and functional differentiation (Wang et al., 2022).

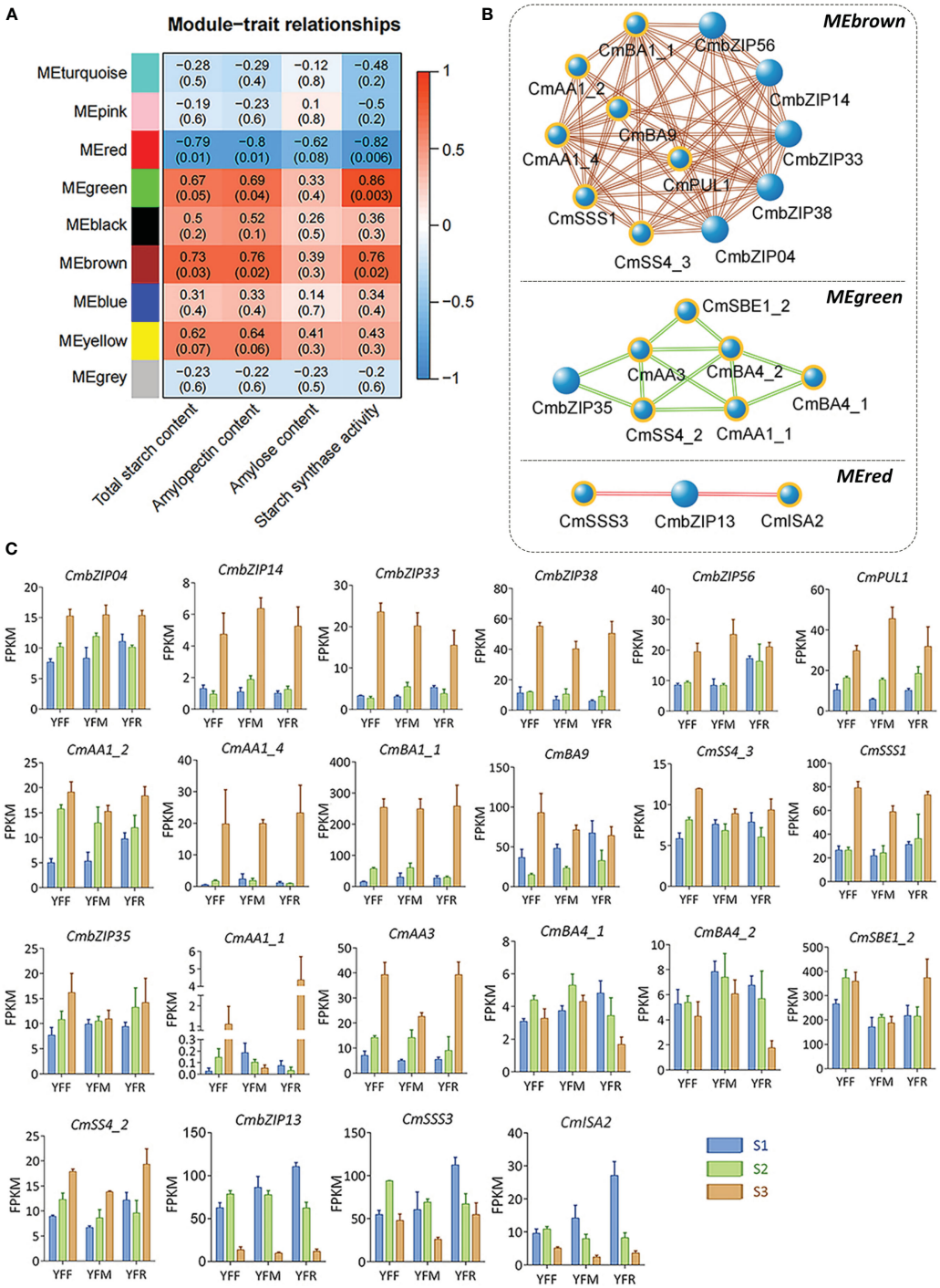


FIGURE 7
Module–trait relationships, co-expression networks, and module-specific gene expression profiles based on results from the WGCNA. **(A)** The heatmap represents relationships between WNCGA modules and traits. The top and bottom numbers in the heat grid represent the correlation coefficients and p-values (shown in parentheses), respectively. **(B)** The networks represent co-expression relationships of *CmbZIP*s and starch accumulation related genes in three key modules. The blue balls highlighted in indicate starch accumulation related genes; the larger balls indicate *CmbZIP*s. Bold and italic characters indicate the names of key modules. **(C)** The column diagrams describe the expression profiles of genes in panel **(B)** The abbreviations YFF, YFM, YFR, S1, S2, and S3 are the same as those used in Figure 6.

WGD drives the evolution and differentiation of plant genome structure (Paterson et al., 2012). WGD, especially segmental and tandem duplications, drive the expansion of gene families (Li et al., 2019; Sun et al., 2019; Wu et al., 2019; Duan et al., 2022). In this study, ten pairs of *CmbZIP*s were identified to have derived from segmental duplication events, and these duplicated genes were clustered into the same clade (Table 1). Based on the Ka/Ks ratio, we found that *CmbZIP* genes have undergone purifying and/or

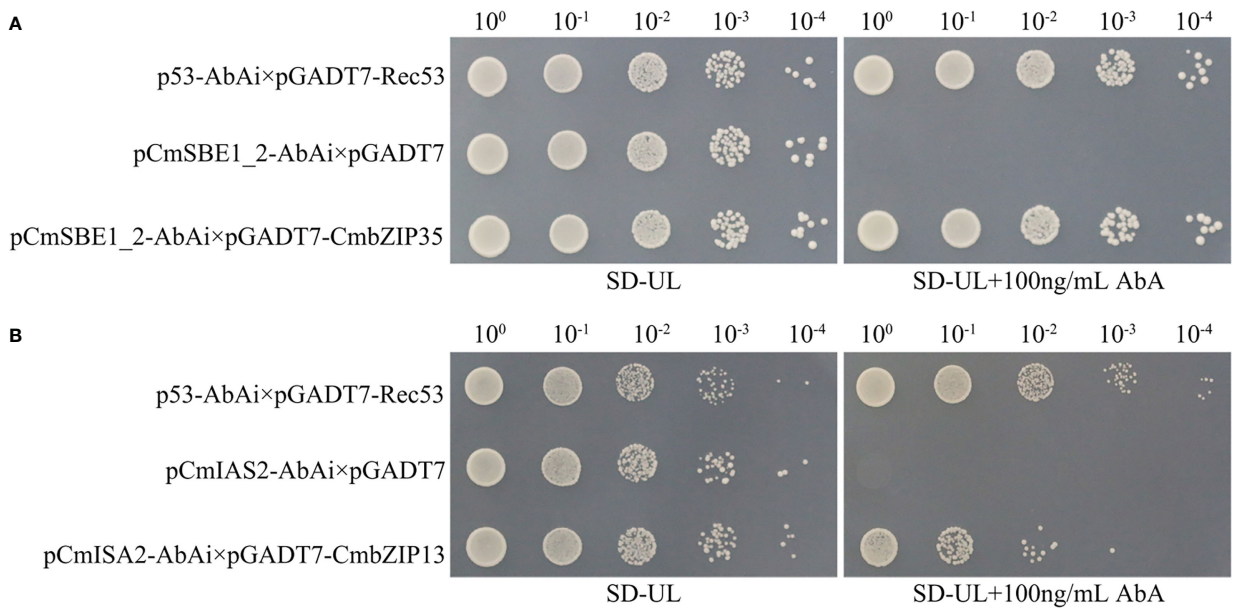


FIGURE 8
Detection of interactions between *CmbZIP*s and starch accumulation related genes. Yeast one-hybrid assays identified that CmbZIP35 interacted with the promoter of *CmSBE1_2* **(A)**, and CmbZIP13 interacted with the promoter of *CmISA2* **(B)**. The yeast strain containing pCmSBE1_2-AbAi×pGADT7-CmbZIP35 **(A)** or pCmISA2-AbAi×pGADT7-CmbZIP13 **(B)** is the experimental group. The yeast strain containing p53-AbAi×pGADT7-Rec53 is the positive control. The yeast strain containing pCmSBE1_2-AbAi×pGADT7 or pCmISA2-AbAi×pGADT7 is the negative control. All yeast strains were selected on SD-UL+100 ng/mL AbA medium. The numbers (10^0, 10^{-1}, 10^{-2}, 10^{-3}, and 10^{-4}) shown along the top of **(A)** and **(B)** represent the dilution ratios of the yeast solution.

positive selection events. It is likely that seven pairs of *CmbZIP*s have undergone purifying selection pressures before the divergence of *C. mollisima* and *Quercus robur* (18.3 Mya) (Wang et al., 2020). These genes may have maintained similar functions. *CmbZIP09/45* and *CmbZIP19/25* might have undergone positive selection after the formation of the genome of *C. mollisima*, which might play an important role in promoting the evolution of *C. mollisima*.

In the phylogenetic analysis, CmbZIP proteins were grouped according to their homology with AtbZIPs (Jakoby et al., 2002). The bZIPs with highly similar protein sequences were clustered into the same clade, in which the members maintained similar functions. Previous studies have shown that two G-box binding factor (GBF) encoding genes *AtbZIP41* (*GBF1*) and *AtbZIP55* (*GBF3*) might play roles during seed maturation (Chern et al., 1996; Jakoby et al., 2002). Therefore, it is possible that CmbZIPs in clade B, especially CmbZIP33, CmbZIP12, and CmbZIP07 (Figure 1), may be involved in the maturation of the chestnut kernel. Furthermore, AtbZIPs in clades F, H, and L might be involved in abiotic and biotic stress response (Jakoby et al., 2002; Liu et al., 2010; Skubacz et al., 2016; Lapham et al., 2018). CmbZIP16/17/58 and CmbZIP41 were, respectively, closely related to OsbZIP33/58 and TabZIP28/TubZIP28, which have been shown to be involved in starch synthesis (Cai et al., 2002; Song et al., 2020). Therefore, we hypothesize that CmbZIP16/17/58 and CmbZIP41 may participate in the regulation of starch synthesis.

In starchy seeds, such as rice and wheat, starch acts as a sink of carbon allocation. In developing seeds, starch is synthesized from sucrose, which is catalyzed by enzymes and regulated by TFs (MacNeill et al., 2017). Previous studies found that OsbZIP58 participates in the regulation of starch synthesis in rice by binding to the promoters of *OsAGPL3*, *OsGBSS*, *OsSSIIa*, *OsSBE1*, *OsBEIIb*, and *OsISA2*, which encode enzymes critical to the starch synthesis process (Wang et al., 2013). Similarly, a previous study reported that TubZIP28 and TabZIP28 are both capable of binding to the promoter of cytosolic AGPase encoding gene to enhance the total starch content (Song et al., 2020). In our study, the expression pattern of *CmbZIP* genes in developing chestnut seeds was further analyzed (Figure 6). The results showed that most *bZIP* genes were highly expressed 70–94 DAP, except for *CmbZIP08*, *CmbZIP18*, *CmbZIP24*, *CmbZIP28*, *CmbZIP29*, *CmbZIP34*, *CmbZIP36*, *CmbZIP49*, *CmbZIP50*, and *CmbZIP59*, which showed overall low levels of expression. This suggested that *CmbZIP* genes might participate in the regulation of chestnut seed maturation. In the correlation and co-expression analysis, we identified that the expression of seven *bZIP*s from the modules containing *CmbZIP04*, *CmbZIP13*, *CmbZIP14*, *CmbZIP33*, *CmbZIP35*, *CmbZIP38*, and *CmbZIP56* was closely related to starch (especially amylopectin) accumulation in chestnut seeds (Figure 7; Table S6) (Li et al., 2021). In the analysis of cis-elements, several 'ACGT' elements, including A- and G-boxes, were found in the promoter regions of *CmISA2* (encoding an isoamylase-type starch de-branching enzyme) and *CmSBE1_2* (encoding a starch branching enzyme) (Figure S5). *CmISA2* and *CmSBE1_2* can modify glucan. In the Y1H assays, CmbZIP13 and CmbZIP35 TFs were found to directly bind to the promoters of *CmISA2* and *CmSBE1_2*, respectively (Figure 8). Therefore, we hypothesize that *CmbZIP13* and *CmbZIP35* genes might participate in starch accumulation

in the chestnut seed by interacting with *CmISA2* and *CmSBE1_2*, respectively.

4 Materials and methods

4.1 Identification of *CmbZIP* genes

The protein sequence data from the previously published genome of the N11-1 Chinese chestnut, a seedling chestnut cultivar, were obtained from the Genome Warehouse in BIG Data Center under accession number GWHANWH00000000 (https://bigd.big.ac.cn/gwh) (Wang et al., 2020). A hidden Markov model (HMM) was used to identify chestnut bZIP candidates, and the HMM profile of bZIP (PF00170) was downloaded from the Pfam protein database (http://pfam.xfam.org/) (Finn et al., 2016). To identify *CmbZIP* genes, the hmmsearch tool of HMMER 3.0 software (Potter et al., 2018) was used to retrieve a domain similar to the bZIP domain in chestnut. The hmmbuild tool was used to rebuild the new HMM profile to re-identify CmbZIP protein sequences. Finally, these protein sequences were confirmed as bZIPs *via* the conserved domain using Pfam (http://pfam.xfam.org/) and Batch CD-Search web tool (https://www.ncbi.nlm.nih.gov/cdd/) (Lu et al., 2020).

4.2 Sequence and phylogenetic analyses

All CDS sequences of *CmbZIP* genes were submitted to ExPASy (https://www.expasy.org) to determine gene length, amino acid length, relative molecular weight, isoelectric point, hydrophilicity, stability, and other physicochemical properties analysis. The protein sequences of 59 *CmbZIP* genes, 23 *AtbZIP* genes (Jakoby et al., 2002), *OsbZIP20* (Izawa et al., 1994), *OsbZIP33* (Cai et al., 2002), *OsbZIP58* (Wang et al., 2013), *TubZIP28*, and *TabZIP28* (Song et al., 2020) were imported into MEGA X, and ClustalW was used for multiple sequence alignments (Kumar et al., 2018). A Neighbor-Joining (NJ) phylogenetic tree was constructed using MEGA X software with bootstrapping set to 1,000. The optional parameters substitution model and gaps data treatment were set to p-distance and partial deletion, respectively. EvolView (https://www.evolgenius.info/evolview/) was used to annotate and visualize the phylogenic trees (Subramanian et al., 2019). CmbZIP proteins were grouped according to their homology with AtbZIPs (Jakoby et al., 2002).

4.3 Gene structure and conserved motif analyses

A gene structure analysis of 59 *CmbZIP* genes was performed with general feature format (GFF) file, and visualized using the online software Gene Structure Display Server (http://gsds.cbi.pku.edu.cn/) (Hu et al., 2015). Conserved motifs were identified using Multiple Expectation Maximization for Motif Elicitation (MEME version 5.1.0) (Bailey et al., 2015) with motif width set to 8–100 and the parameter of maximum motif number set to 20.

4.4 Chromosomal localization and synteny analyses

All *CmbZIP* genes were mapped to Chinese chestnut chromosomes based on physical location information from the GFF file using Mapchart 2.32 software (Voorrips, 2002). Multiple Collinearity Scan toolkit (MCScanX) was used to identify the syntenic gene pairs within the genome, using the results from all-*vs*-all BLASTP analysis (Wang et al., 2012). Results were displayed with Circos (version 0.69-8) software (Krzywinski et al., 2009). The values of Ka and Ks were calculated using the KaKs-calculator 2.0 (Wang et al., 2010). The divergence-times of duplicated gene pairs were estimated using the Ks value with the formula $T = Ks/2r$, where T is the divergence-time and r is the divergence rate of nuclear genes from plants ($r = 1.5 \times 10^{-8}$) (Koch et al., 2000; Huang et al., 2016). We used the Python version of MCscan to analyze the synteny between the genomes of *C. mollissima*, *A. thaliana*, *O. sativa*, *T. aestivum*, and *M. domestica* (Tang et al., 2014).

4.5 Expression profile and co-expression analyses based on RNA-seq

The published RNA-seq data were obtained from the sequence read archive (SRA) in NCBI (accession number PRJNA540079) (Li et al., 2021). All seed samples were collected at 70, 82, and 94 days after pollination from three crosses: 'Yongfeng 1'×'Yongfeng 1,' 'Yongfeng 1'×'Yimen 1,' and 'Yongfeng 1'×'Yongren Zao' (Li et al., 2021). RNA-seq read-files were converted from SRA to fastq format using sratoolkit3.0 (Goldberg et al., 2009). The mapping genome used in the RNA-seq analysis was changed from the previous version of the genome assembly to N11-1. The expression profiles of *CmbZIP*s were determined using Tophat2 software (Kim et al., 2013; Wang et al., 2020; Li et al., 2021). The FPKM value was used to evaluate the expression level of each gene. The correlation coefficients and p values between the $\log_{10}$FPKM of *CmbZIP*s and four physiological characteristics (i.e., total starch content, amylopectin content, amylose content, and starch synthase activity) published in a previous study were estimated using GraphPad Prism version 6.02 for Windows (Li et al., 2021; https://www.graphpad.com/). The co-expression modules were identified using WGCNA with the R package (Langfelder and Horvath, 2008). The co-expression networks were generated using Cytoscape software (Otasek et al., 2019).

4.6 Identification of cis-elements in gene promoter regions

We retrieved sequences 1,500 bp upstream of the transcription start site of starch accumulation related genes. These sequences

were submitted to PlantCARE to identify cis-elements, which might affect gene expression and function (Rombauts et al., 1999).

4.7 Sequence cloning and yeast one-hybrid assays

The open reading frames of *CmbZIP35* and *CmbZIP13* were cloned by reverse transcription PCR using RNA from developing seeds of *C. mollissima* cultivar 'Yanbao.' We confirmed these sequences using DNAMAN 6.0 software. Subsequently, we fused the two sequences into the pGADT7 vector to construct the pGADT7-CmbZIP35 and pGADT7-CmbZIP13 recombinant plasmids, respectively. The *CmSBE1_2* and *CmISA2* promoter fragments were cloned by PCR using DNA from 'Yanbao' seeds. The fragments were confirmed and inserted into the pAbAi vector to construct the pCmSBE1_2-AbAi and pCmISA2-AbAi recombinant plasmids. All primer sequences are listed in Table S8.

To determine the optimal AbA concentration, the Y1H Gold yeast strain containing the recombinant pAbAi plasmids were grown on SD-UL screening medium supplemented with different AbA concentrations (Yang et al., 2019). Then, Y1H Gold yeast cells were co-transformed with pGADT7-CmbZIP35 and pCmSBE1_2-AbAi plasmids, as well as pGADT7-CmbZIP13 and pCmISA2-AbAi plasmids. Interactions were detected on SD-UL selection medium that was supplemented with 100 ng/mL AbA.

5 Conclusions

A total of 59 *CmbZIP* genes were identified in Chinese chestnut. These *CmbZIP*s were clustered into 13 clades with clade-specific motifs and structures. Segmental duplication was determined as the major driving force of the expansion of the *CmbZIP* gene family. *CmbZIP04*, *CmbZIP13*, *CmbZIP14*, *CmbZIP33*, *CmbZIP35*, *CmbZIP38*, and *CmbZIP56* were confirmed to be highly correlated with starch accumulation in chestnut seeds. We demonstrated that *CmbZIP13* and *CmbZIP35* may regulate starch accumulation in the chestnut seed by binding to the promoters of *CmISA2* and *CmSBE1_2*, respectively. These results indicated that *CmbZIP* genes contained information related to starch accumulation in chestnut seeds, which can be used in future functional analysis and breeding studies.

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

XW and HZ designed the experiments. MW and YL collected data of Chinese chestnut genome and RNA-seq. NJ and JL drew all figures of this study. PZ and JL carried out the yeast one-hybrid experiments. PZ and XW performed the bioinformatic analysis and wrote the paper. DW and JZ directed and revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was funded by a subproject of Hebei Collaborative Innovation Center of Chestnut Industry (202202) and Scientific Research Foundation of Hebei Normal University of Science and Technology (2022YB002).

Acknowledgments

The authors thank the Hebei Normal University of Science and Technology, and Engineering Research Center of Chestnut Industry Technology, Ministry of Education for the experimental materials and technical assistance provided.

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.2023.1166717/full#supplementary-material

References

An, J., Yao, J., Xu, R., You, C., Wang, X., and Hao, Y. (2018). Apple bZIP transcription factor MdbZIP44 regulates abscisic acid-promoted anthocyanin accumulation. *Plant Cell Environ.* 41 (11), 2678–2692. doi: 10.1111/pce.13393

Bailey, T. L., Johnson, J., Grant, C. E., and Noble, W. S. (2015). The MEME suite. *Nucleic Acids Res.* 43 (W1), W39–W49. doi: 10.1093/nar/gkv416

Cai, Y., Xie, D., Wang, Z., and Hong, M. (2002). Interaction of rice bZIP protein REB with the 5′-upstream region of both rice sbe1 gene and waxy gene. *Chin. Sci. Bulletin.* 47, 310–314. doi: 10.1360/02tb9074

Chern, M. S., Bobb, A. J., and Bustos, M. M. (1996). The regulator of MAT2 (ROM2) protein binds to early maturation promoters and represses PvALF-activated transcription. *Plant Cell.* 8 (2), 305–321. doi: 10.1105/tpc.8.2.305

Duan, L., Mo, Z., Fan, Y., Li, K., Yang, M., Li, D., et al. (2022). Genome-wide identification and expression analysis of the bZIP transcription factor family genes in response to abiotic stress in nicotiana tabacum l. *BMC Genomics* 23 (1), 318. doi: 10.1186/s12864-022-08547-z

Finkelstein, R. R., and Lynch, T. J. (2000). The arabidopsis abscisic acid response gene ABI5 encodes a basic leucine zipper transcription factor. *Plant Cell.* 12 (4), 599–609. doi: 10.1105/tpc.12.4.599

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., et al. (2016). The pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res.* 44 (D1), D279–DD85. doi: 10.1093/nar/gkv1344

Fitzgerald, H. A., Canlas, P. E., Chern, M. S., and Ronald, P. C. (2005). Alteration of TGA factor activity in rice results in enhanced tolerance to xanthomonas oryzae pv. oryzae. *Plant J.* 43 (3), 335–347. doi: 10.1111/j.1365-313X.2005.02457.x

Fu, F., and Xue, H. (2010). Coexpression analysis identifies rice starch Regulator1, a rice AP2/EREBP family transcription factor, as a novel rice starch biosynthesis regulator. *Plant Physiol.* 154 (2), 927–938. doi: 10.1104/pp.110.159517

Gao, Y., An, K., Guo, W., Chen, Y., Zhang, R., Zhang, X., et al. (2021). The endosperm-specific transcription factor TaNAC019 regulates glutenin and starch accumulation and its elite allele improves wheat grain quality. *Plant Cell.* 33 (3), 603–622. doi: 10.1093/plcell/koaa040

Gibalova, A., Steinbachova, L., Hafidh, S., Blahova, V., Gadiou, Z., Michailidis, C., et al. (2017). Characterization of pollen-expressed bZIP protein interactions and the role of ATbZIP18 in the male gametophyte. *Plant Reprod.* 30 (1), 1–17. doi: 10.1007/s00497-016-0295-5

Goldberg, D. H., Victor, J. D., Gardner, E. P., and Gardner, D. (2009). Spike train analysis toolkit: Enabling wider application of information-theoretic techniques to neurophysiology. *Neuroinformatics* 7, 165–178. doi: 10.1007/s12021-009-9049-y

Heinekamp, T., Kuhlmann, M., Lenk, A., Strathmann, A., and Droge-Laser, W. (2002). The tobacco bZIP transcription factor BZI-1 binds to G-box elements in the promoters of phenylpropanoid pathway genes *in vitro*, but it is not involved in their regulation *in vivo*. *Mol. Genet. Genomics* 267, 16–26. doi: 10.1007/s00438-001-0636-3

Hu, B., Jin, J., Guo, A.-Y., Zhang, H., Luo, J., and Gao, G. (2015). GSDS 2.0: an upgraded gene feature visualization server. *Bioinformatics* 31 (8), 1296–1297. doi: 10.1093/bioinformatics/btu817

Hu, G., Cheng, L., Cheng, Y., Mao, W., Qiao, Y., and Lan, Y. (2022). Pan-genome analysis of three main Chinese chestnut varieties. *Front. Plant Sci.* 13::916550. doi: 10.3389/fpls.2022.916550

Huang, Z., Duan, W., Song, X., Tang, J., Wu, P., Zhang, B., et al. (2016). Retention, molecular evolution, and expression divergence of the auxin/indole acetic acid and auxin response factor gene families in brassica rapa shed light on their evolution patterns in plants. *Genome Biol. Evol.* 8 (2), 302–316. doi: 10.1093/gbe/evv259

Izawa, T., Foster, R., and Chua, N.-H. (1993). Plant bZIP protein DNA binding specificity. *J. Mol. Biol.* 230 (4), 1131–1144. doi: 10.1006/jmbi.1993.1230

Izawa, T., Foster, R., Nakajima, M., Shimamoto, K., and Chua, N.-H. (1994). The rice bZIP transcriptional activator RITA-1 is highly expressed during seed development. *Plant Cell.* 6 (9), 1277–1287. doi: 10.1006/jmbi.1993.1230

Jakoby, M., Weisshaar, B., Dröge-Laser, W., Vicente-Carbajosa, J., Tiedemann, J., Kroj, T., et al. (2002). bZIP transcription factors in arabidopsis. *Trends Plant Sci.* 7 (3), 106–111. doi: 10.1016/S1360-1385(01)02223-3

Kim, S., Kang, Jy, Cho, D. I., Park, J. H., and Kim, S. Y. (2004). ABF2, an ABRE-binding bZIP factor, is an essential component of glucose signaling and its overexpression affects multiple stress tolerance. *Plant J.* 40 (1), 75–87. doi: 10.1111/j.1365-313X.2004.02192.x

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14 (4), 1–13. doi: 10.1186/gb-2013-14-4-r36

Koch, M. A., Haubold, B., and Mitchell-Olds, T. (2000). Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in arabidopsis, arabis, and related genera (Brassicaceae). *Mol. Biol. Evol.* 17 (10), 1483–1498. doi: 10.1093/oxfordjournals.molbev.a026248

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: An information aesthetic for comparative genomics. *Genome Res.* 19 (9), 1639–1645. doi: 10.1101/gr.092759.109

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35 (6), 1547. doi: 10.1093/molbev/msy096

Langfelder, P., and Horvath, S. (2008). WGCNA: An r package for weighted correlation network analysis. *BMC Bioinf.* 9 (1), 1–13. doi: 10.1186/1471-2105-9-559

Lapham, R., Lee, L.-Y., Tsugama, D., Lee, S., Mengiste, T., and Gelvin, S. B. (2018). VIP1 and its homologs are not required for agrobacterium-mediated transformation, but play a role in botrytis and salt stress responses. *Front. Plant Sci.* 9, 749. doi: 10.3389/fpls.2018.00749

Lee, S. C., Choi, H. W., Hwang, I. S., Choi, D. S., and Hwang, B. K. (2006). Functional roles of the pepper pathogen-induced bZIP transcription factor, CAbZIP1, in enhanced resistance to pathogen infection and environmental stresses. *Planta* 224, 1209–1225. doi: 10.1007/s00425-006-0302-4

Li, Y., Meng, D., Li, M., and Cheng, L. (2016). Genome-wide identification and expression analysis of the bZIP gene family in apple (Malus domestica). *Tree Genet. Genomes.* 12, 1–17. doi: 10.1007/s11295-016-1043-6

Li, S., Shi, Z., Zhu, Q., Tao, L., Liang, W., and Zhao, Z. (2021). Transcriptome sequencing and differential expression analysis of seed starch accumulation in Chinese chestnut metaxenia. *BMC Genomics* 22, 1–14. doi: 10.1186/s12864-021-07923-5

Li, M., Wang, R., Liu, Z., Wu, X., and Wang, J. (2019). Genome-wide identification and analysis of the WUSCHEL-related homeobox (WOX) gene family in allotetraploid brassica napus reveals changes in WOX genes during polyploidization. *BMC Genomics* 20 (1), 1–19. doi: 10.1186/s12864-019-5684-3

Liang, Y., Xia, J., Jiang, Y., Bao, Y., Chen, H., Wang, D., et al. (2022). Genome-wide identification and analysis of bZIP gene family and resistance of TaABI5 (TabZIP96) under freezing stress in wheat (Triticum aestivum). *Int. J. Mol. Sci.* 23 (4), 2351. doi: 10.3390/ijms23042351

Liao, Y., Zou, H.-F., Wei, W., Hao, Y.-J., Tian, A.-G., Huang, J., et al. (2008). Soybean GmbZIP44, GmbZIP62 and GmbZIP78 genes function as negative regulator of ABA signaling and confer salt and freezing tolerance in transgenic arabidopsis. *Planta* 228, 225–240. doi: 10.1007/s00425-008-0731-3

Lin, S., Pang, L., and Zhu, M. (2012). Correlation between chestnut starch complex and glutinous characteristics. *Food Sci.* 33, 308–311.

Liu, Y., Kong, X., Pan, J., and Li, D. (2010). VIP1: Linking agrobacterium-mediated transformation to plant immunity? *Plant Cell Rep.* 29, 805–812. doi: 10.1007/s00299-010-0870-4

Lu, S., Wang, J., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., et al. (2020). CDD/SPARCLE: The conserved domain database in 2020. *Nucleic Acids Res.* 48 (D1), D265–D2D8. doi: 10.1093/nar/gkz991

Ma, M., Chen, Q., Dong, H., Zhang, S., and Huang, X. (2021). Genome-wide identification and expression analysis of the bZIP transcription factors, and functional analysis in response to drought and cold stresses in pear (Pyrus breschneideri). *BMC Plant Biol.* 21 (1), 583. doi: 10.1186/s12870-021-03356-0

MacNeill, G. J., Mehrpouyan, S., Minow, M. A., Patterson, J. A., Tetlow, I. J., and Emes, M. J. (2017). Starch as a source, starch as a sink: The bifunctional role of starch in carbon allocation. *J. Exp. Bot.* 68 (16), 4433–4453. doi: 10.1093/jxb/erx291

Meng, X., Zhao, W., Lin, R., Wang, M., and Peng, Y. (2005). Identification of a novel rice bZIP-type transcription factor gene, OsbZIP1, involved in response to infection of magnaporthe grisea. *Plant Mol. Biol. Reporter.* 23, 301–302. doi: 10.1007/BF02772762

Nakase, M., Aoki, N., Matsuda, T., and Adachi, T. (1997). Characterization of a novel rice bZIP protein which binds to the α-globulin promoter. *Plant Mol. Biol.* 33, 513–522. doi: 10.1023/A:1005784717782

Nijhawan, A., Jain, M., Tyagi, A. K., and Khurana, J. P. (2008). Genomic survey and gene expression analysis of the basic leucine zipper transcription factor family in rice. *Plant Physiol.* 146 (2), 333. doi: 10.1104/pp.107.112821

Otasek, D., Morris, J. H., Bouças, J., Pico, A. R., and Demchak, B. (2019). Cytoscape automation: Empowering workflow-based network analysis. *Genome Biol.* 20, 1–15. doi: 10.1186/s13059-019-1758-4

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 (7429), 423–427. doi: 10.1038/nature11798

Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., and Finn, R. D. (2018). HMMER web server: 2018 update. *Nucleic Acids Res.* 46 (W1), W200–W2W4. doi: 10.1093/nar/gky448

Qu, J., Xu, S., Zhang, Z., Chen, G., Zhong, Y., Liu, L., et al. (2018). Evolutionary, structural and expression analysis of core genes involved in starch synthesis. *Sci. Rep.* 8 (1), 12736. doi: 10.1038/s41598-018-30411-y

Rombauts, S., Déhais, P., Van Montagu, M., and Rouzé, P. (1999). PlantCARE, a plant cis-acting regulatory element database. *Nucleic Acids Res.* 27 (1), 295–296. doi: 10.1093/nar/27.1.295

Schlögl, P. S., Nogueira, F. T. S., Drummond, R., Felix, J. M., De Rosa, V. E., Vicentini, R., et al. (2008). Identification of new ABA-and MEJA-activated sugarcane bZIP genes by data mining in the SUCEST database. *Plant Cell Rep.* 27, 335–345. doi: 10.1007/s00299-007-0468-7

Shi, L., Wang, J., Liu, Y., Ma, C., Guo, S., Lin, S., et al. (2021). Transcriptome analysis of genes involved in starch biosynthesis in developing Chinese chestnut (Castanea mollissima blume) seed kernels. *Sci. Rep.* 11 (1), 1–13. doi: 10.1038/s41598-021-82130-6

Skubacz, A., Daszkowska-Golec, A., and Szarejko, I. (2016). The role and regulation of ABI5 (ABA-insensitive 5) in plant development, abiotic stress responses and phytohormone crosstalk. *Front. Plant Science.* 7, 1884. doi: 10.3389/fpls.2016.01884

Song, Y., Luo, G., Shen, L., Yu, K., Yang, W., Li, X., et al. (2020). TubZIP28, a novel bZIP family transcription factor from triticum urartu, and TabZIP28, its homologue from triticum aestivum, enhance starch synthesis in wheat. *New Phytol.* 226 (5), 1384–1398. doi: 10.1111/nph.16435

Subramanian, B., Gao, S., Lercher, M. J., Hu, S., and Chen, W.-H. (2019). Evolview v3: a webserver for visualization, annotation, and management of phylogenetic trees. *Nucleic Acids Res.* 47 (W1), W270–W2W5. doi: 10.1093/nar/gkz357

Sun, W., Ma, Z., Chen, H., and Liu, M. (2019). MYB gene family in potato (Solanum tuberosum l.): Genome-wide identification of hormone-responsive reveals their potential functions in growth and development. *Int. J. Mol. Sci.* 20 (19), 4847. doi: 10.3390/ijms20194847

Sun, Y., Lu, Z., Zhu, X., and Ma, H. (2020). Genomic basis of homoploid hybrid speciation within chestnut trees. *Nat. Commun.* 11 (1), 3375. doi: 10.1038/s41467-020-17111-w

Tang, J., Wang, F., Hou, X., Wang, Z., and Huang, Z. (2014). Genome-wide fractionation and identification of WRKY transcription factors in Chinese cabbage (Brassica rapa ssp. pekinensis) reveals collinearity and their expression patterns under abiotic and biotic stresses. *Plant Mol. Biol. Reporter.* 32, 781–795. doi: 10.1007/s11105-013-0672-2

Thalmann, M., Pazmino, D., Seung, D., Horrer, D., Nigro, A., Meier, T., et al. (2016). Regulation of leaf starch degradation by abscisic acid is important for osmotic stress tolerance in plants. *Plant Cell.* 28 (8), 1860–1878. doi: 10.1105/tpc.16.00143

Uno, Y., Furihata, T., Abe, H., Yoshida, R., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2000). Arabidopsis basic leucine zipper transcription factors involved in an abscisic acid-dependent signal transduction pathway under drought and high-salinity conditions. *Proc. Natl. Acad. Sci.* 97 (21), 11632–11637. doi: 10.1073/pnas.190309197

Voorrips, R. (2002). MapChart: Software for the graphical presentation of linkage maps and QTLs. *J. Heredity.* 93 (1), 77–78. doi: 10.1093/jhered/93.1.77

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40 (7), e49–e4e. doi: 10.1093/nar/gkr1293

Wang, J., Tian, S., Sun, X., Cheng, X., Duan, N., Tao, J., et al. (2020). Construction of pseudomolecules for the Chinese chestnut (Castanea mollissima) genome. *G3: Genes Genomes Genet.* 10 (10), 3565–3574. doi: 10.1534/g3.120.401532

Wang, J., Xu, H., Zhu, Y., Liu, Q., and Cai, X. (2013). OsbZIP58, a basic leucine zipper transcription factor, regulates starch biosynthesis in rice endosperm. *J. Exp. Bot.* 64 (11), 3453–3466. doi: 10.1093/jxb/ert187

Wang, S., Zhang, X., Li, B., Zhao, X., Shen, Y., and Yuan, Z. (2022). Genome-wide identification and characterization of bZIP gene family and cloning of candidate genes for anthocyanin biosynthesis in pomegranate (Punica granatum). *BMC Plant Biol.* 22 (1), 1–18. doi: 10.1186/s12870-022-03560-6

Wang, D., Zhang, Y., Zhang, Z., Zhu, J., and Yu, J. (2010). KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinf.* 8 (1), 77–80. doi: 10.1016/S1672-0229(10)60008-3

Wei, K., Chen, J., Wang, Y., Chen, Y., Chen, S., Lin, Y., et al. (2012). Genome-wide analysis of bZIP-encoding genes in maize. *DNA Res.* 19 (6), 463–476. doi: 10.1093/dnares/dss026

Wu, A., Hao, P., Wei, H., Sun, H., Cheng, S., Chen, P., et al. (2019). Genome-wide identification and characterization of glycosyltransferase family 47 in cotton. *Front. Genet.* 10, 824. doi: 10.3389/fgene.2019.00824

Xing, Y., Liu, Y., Zhang, Q., Nie, X., Sun, Y., Zhang, Z., et al. (2019). Hybrid *de novo* genome assembly of Chinese chestnut (*Castanea mollissima*). *Gigascience* 8 (9), giz112. doi: 10.1093/gigascience/giz112

Yang, G., Chao, D., Ming, Z., and Xia, J. (2019). A simple method to detect the inhibition of transcription factor-DNA binding due to protein–protein interactions *In vivo*. *Genes* 10 (9), 684. doi: 10.3390/genes10090684

Zhang, X., Wollenweber, B., Jiang, D., Liu, F., and Zhao, J. (2008). Water deficits and heat shock effects on photosynthesis of a transgenic arabidopsis thaliana constitutively expressing ABP9, a bZIP transcription factor. *J. Exp. Bot.* 59 (4), 839–848. doi: 10.1093/jxb/erm364

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 12 April 2023
DOI 10.3389/fpls.2023.1135302

OPEN ACCESS

EDITED BY
Xueqiang Wang,
Zhejiang University, China

REVIEWED BY
Hejun Lu,
Zhejiang University, China
Guanghui Hu
Northeast Agricultural University, China

*CORRESPONDENCE
Yanlong Yang
18152930615@163.com

†These authors have contributed equally to this work

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

RECEIVED 31 December 2022
ACCEPTED 10 March 2023
PUBLISHED 12 April 2023

CITATION
Sun F, Ma J, Shi W and Yang Y (2023) Genome-wide association analysis revealed genetic variation and candidate genes associated with the yield traits of upland cotton under drought conditions. *Front. Plant Sci.* 14:1135302. doi: 10.3389/fpls.2023.1135302

Genome-wide association analysis revealed genetic variation and candidate genes associated with the yield traits of upland cotton under drought conditions

Fenglei Sun[1,2†], Jun Ma[3†], Weijun Shi[3] and Yanlong Yang[3*]

[1]State Key Laboratory of Cotton Biology, Institute of Cotton Research of the Chinese Academy of Agricultural Sciences, Anyang, China, [2]Hainan Yazhou Bay Seed Laboratory, Sanya, Hainan, China, [3]Research Institute of Economic Crops, Xinjiang Academy of Agricultural Sciences, Urumqi, China

Drought is one of the major abiotic stresses seriously affecting cotton yield. At present, the main cotton-producing areas in China are primarily arid and semiarid regions. Therefore, the identification of molecular markers and genes associated with cotton yield traits under drought conditions is of great importance for stabilize cotton yield under such conditions. In this study, resequencing data were used to conduct a genome-wide association study (GWAS) on 8 traits of 150 cotton germplasms. Under drought stress, 18 SNPs were significantly correlated with yield traits (single-boll weight (SBW) and seed (SC)), and 8 SNPs were identified as significantly correlated with effective fruit shoot number (EFBN) traits (a trait that is positively correlated with yield). Finally, a total of 15 candidate genes were screened. The combined results of the GWAS and transcriptome data analysis showed that four genes were highly expressed after drought stress, and these genes had significantly increased expression at 10, 15 and 25 DPA of fiber development. qRT-PCR was performed on two samples with drought tolerance extremes (drought-resistant Xinluzao 45 and drought-sensitive Xinluzao 26), revealing that three of the genes had the same differential expression pattern. This study provides a theoretical basis for the genetic analysis of cotton yield traits under drought stress, and provides gene resources for improved breeding of cotton yield traits under drought stress.

KEYWORDS

upland cotton, drought tolerance, genome-wide associated study, SNPs, candidate genes

1 Introduction

As an important cash crop in China, cotton plays a major role in domestic economic development. However, with the increase in extreme climatic events in recent years and the decrease in water resources, abiotic stress (drought) has become increasingly serious (Sun et al., 2017; Sun et al., 2021). Drought, in response to global warming, is becoming more prominent and serious, and it has become a focus in climate change research (Cattivelli et al., 2008). Some arid or semiarid areas are experiencing annual decreases in precipitation and fresh water resources. Even in areas with relatively sufficient water resources, the effect of extreme climate on local precipitation is noticeable, and it is particularly obvious in China (Wang et al., 2017). Cotton cultivation in Chian is located mainly in Xinjiang, the Yangtze River Basin and the Huang-Huai-Hai region (Huang et al., 2017). According to data from the National Bureau of Statistics, the sown area of cotton in Xinjiang in 2021 was 37.5926 million Mus, accounting for 82.76% of the national planting area (Ding et al., 2021). Xinjiang is located in an arid and semiarid region with little precipitation, and agricultural water consumption accounts for approximately 94% of total water consumption (Xiao, 2020). The cotton planting area in Xinjiang accounts for approximately 45% of the total sown area of crops in Xinjiang. With the frequent occurrence of extreme weather resulting from such as climate warming, lack of fresh water resources and high temperature, drought has become an important factor restricting cotton production in Xinjiang; furthermore, drought stress can affect the yield and quality of cotton by changing its metabolic activities and biological functions.

To date, many genes associated with drought tolerance have been identified. Li et al. (2019) studied the drought resistance of 316 upland cotton germplasms at the seedling stage by GWAS and identified *WRKY70*, *GhCIPK6*, *SnRK2.6* and *NET1A* as genes induced by drought stress. In rice, Sun et al. found that *DROT1* can improve drought tolerance, mainly by regulating the cell wall fiber content and crystal structure in microtubule tissues to enhance drought resistance (Sun et al., 2022). Du et al. found that *TaERF87* (ethylene response factor (ERF)) could interact with *TaAKS1* to enhance the expression of *TaP5CS1* and *TaP5CR1*, thus improving proline synthesis and drought resistance in wheat (Du et al., 2022). Through a GWAS in maize, it was found that three SNP mutations in *ZmSRO1d* significantly increased the reactive oxygen species (ROS) content of guard cells, promoted stomatal closure, enhanced drought resistance, and increased the yield of overexpressed *ZMSRO1D-R* by 60% compared with that in the control (Gao et al., 2022). Although these identified drought tolerance genes are associated with different traits, they can all improve the drought tolerance of related crops.

Previously, most drought-resistant quantitative trait loci (QTLs) were identified in genetic populations through simple sequence repeat (SSR) markers (Sang et al., 2017), but with the advancement of sequencing technology and the completion of cotton genome sequencing, GWAS has become an important analytical tool (Li et al., 2019). Some QTL sites associated with drought tolerance traits have also been identified by the GWAS approach. Saleem et al. (2015) used 524 SSR markers to perform linkage analysis of F_2 populations of drought-tolerant (B-557) and drought-resistant (FH-1000) varieties and detected 22 drought-related QTLs. These included two QTLs related to water content and QTLs on chromosome 23 that were associated with leaf water loss. Shukla et al. (2021) performed drought tolerance genetic mapping and QTL analysis of drought-tolerant (AS2) and drought-susceptible (MCU13) terrestrial cotton recombinant inbred line (RIL) populations based on genotyping by sequencing (GBS) and SSRs and identified 19 QTLs associated wtih field drought-tolerance traits, with 3 QTLs on chromosome 8 related to relative water content. Hou et al. (2018) genotyped 319 land cotton accessions through a high-density CottonSNP80K array, found that 20 quantitative trait nucleotides (QTNs) distributed on 16 chromosomes were associated with 6 drought resistance traits, and finally identified two candidate genes related to soluble sugar content and one gene related to root dry matter and hypocotyl length. Abdelraheem et al. (2020a); Abdelraheem et al. (2021) constructed a multiparent advanced generation intercross (MAGIC) population with 11 upland cotton accessions as parents and performed a GWAS of drought resistance traits in 550 strains. A total of 23 and 20 QTLs were detected under normal and drought-resistant treatment conditions, respectively. A GWAS performed on 376 upland cotton seedlings in the United States to investigate drought tolerance revealed 13 QTL clusters at 11 sites. Based on 372 strains derived from MAGIC populations with 8 upland cotton accessions as parents, Huang et al. (2021) used specific locus amplified fragment sequencing (SLAF-seq) to map genome-wide associations, and found that 177 SNPs were significantly associated with 9 stable agronomic traits in multiple environments, and 8 candidate genes with known functions were identified. Ul-Allah et al. (2021) reported that the effects of drought stress on cotton fiber development can lead to a yield loss of approximately 45%. Abdelraheem et al. (2020b) showed that water deficit during flowering can lead to a decrease in cotton fiber strength, an increase in staple fiber content, and a decrease in quality. Most association analyses of drought resistance in cotton populations were based on SSR markers, GBS and gene chips, and there are few reports that use resequencing to locate drought resistance sites. Moreover, studies on the localization of key trait QTLs in cotton have focused mainly on fiber quality, while there have been relatively few studies on the localization of QTLs associated with yield traits under drought conditions.

Although some genes or QTLs associated with yield traits have been identified in genetic populations and natural populations, effective analysis of the genetic basis of yield traits is still incomplete. Therefore, in this study, we collected phenotypic data from 150 upland cotton cultivars with large yield differences in the Shihezi and Korla areas of Xinjiang. Furthermore, we analyzed and explored the genetic loci and key candidate genes related to yield under drought conditions through the GWAS approach, which laid a foundation for studying the molecular mechanism underlying cotton drought resistance and the genetic improvement of cotton.

2 Materials and methods

2.1 Plant material and drought stress treatment

A total of 152 land cotton germplasms (Supplementary Table S1) were collected, all of which are cultivars grown in Northwest

China, and were collected and preserved by the Xinjiang Academy of Agricultural Sciences. In 2019 and 2020, 152 land cotton germplasms were planted in 2 natural environments, namely Shihezi (85.94°E, 44.27°N) in 2019 and Korla (86.06°E, 35.05°N) in 2020. All accessions were planted following a random complete block design (RCBD) with two replicates per environment and two rows per replicate. In both Korla and Shihezi, the row length was 2 meters, the row spacing was 66 + 10 cm (width/narrow), and the plant distance was 10 cm. The conditions for drought stress treatment were achieved by artificial water control. The treatment was mainly applied at the flowering and boll stages, with irrigation stopped in the drought stress treatment group and continued in the control group. During the boll opening period, irrigation was reinitiated in the drought stress treatment followed by normal irrigation.

2.2 Phenotypic data collection and analysis

After cotton maturation, three yield-related traits and five agronomic traits, namely seed cotton (SC), single boll weight (SBW), lint cotton (LC), plant height (PH), fruit branch number (FBN), effective fruit branch number (EFBN), boll number (BN) and effective boll number (EBN), were measured under each environmental condition to analyze phenotypic changes in cotton. For each variety, 10 plants were randomly selected from the middle of each row. The five agronomic traits were measured, with ten biological replicates for each germplasm. Twenty mature bolls were randomly harvested from the middle part of the cotton plant and weighed (BW), with 2 bolls per plant. After ginning, the LC and SC were weighed and counted separately. The survey method followed the Specification for the Description of Cotton Germplasm Resources and Data Standard guidelines (Du and Zhou, 2005). In this study, data for eight traits (including five agronomic traits and three yield traits) in two environments were statistically analyzed. SPSS 25.0 was used for descriptive statistical analysis of all traits as well as analysis of variance. Correlation analysis of all traits in the cotton panel across different environments was performed in R software. Since the seedlings of 2 accessions were not sufficient for phenotypic studies, we used data from eight phenotypic traits of 150 accessions for the subsequent GWAS.

2.3 Genotypic data analysis

Young leaves were collected from plants of the 150 accessions, and genomic DNA was extracted to construct paired end-sequencing libraries for resequencing with 10× genome coverage using the HiSeq 2000 platform (Illumina, Inc., San Diego, California, USA) (He et al., 2021). Clean reads from 150 germplasms were matched with the *Gossypium hirsutum* reference genome TM-1 [CRI v1 (Yang et al., 2019)] using BWA version 0.7.10. After alignment, SNP calling was performed at the population scale with a unified genotype approach using Genomic Analysis Toolkit (GATK, v3.1) (McKenna et al., 2010). Subsequently, high-quality SNPs with a reserved minor allele frequency (MAF) greater than 0.05 were used for further analysis.

2.4 LD analysis, population structure, haplotype analysis and clustering

Population linkage disequilibrium (LD) was analyzed by PopLDdecay software (Zhang et al., 2019), and r^2 was calculated for SNPs within a 1 Mb window. Population structure was analyzed by the Admixture 1.3 program, which was run 1000 times for K values of 2-10 to generate admixture ratios. Then, the optimal value of K was determined by cross-validation (CV) scores and log-likelihood estimates. Haplotypes were detected and analyzed by software such as IGV (Helga et al., 2012), Tassel (Bradbury et al., 2007), Figtree (Rambaut, 2009), and R. First, strong and continuous SNP regions were identified by IGV software, and these regions were named target SNP intervals. Then, Tassel software was used to perform LD segment analysis and to identify numerical genotypes of target SNP intervals. The digital genotypes of "Minor", "Major" and "Hereozygous" were filled with the color scale function in Excel, and haplotype classification was performed according to the color change in the target SNP interval. To construct a phylogenetic tree, the neighbor-joining (NJ) method in Tree Best (v1.9.2) software was used, and the tree was visually edited by Figtree software (Vilella et al., 2009).

2.5 Genome-wide association studies

To ensure the accuracy of the results, SNPs with a missing genotype frequency greater than 0.05 or a MAF less than 0.05 were filtered without imputation. A total of 2,499,987 SNPs were identified in the association panel for the final 150 samples, and the SNPs for the entire genome were viewed using the sliding window method (defaults of 50 bp window size and 10 bp steps). A GWAS between SNPs and traits was performed using Efficient Mixed Model Association Acceleration (EMMAX) software (Kang et al., 2010) and Fixed and random model Circulating Probability Unification (FarmCPU) models, where the threshold for association detection was set to -log(1/N) (where N is a valid value for the SNP label) (Li et al., 2012; Li et al., 2019).

2.6 Prediction of candidate genes and qRT-PCR

In this study, the upstream and downstream 200-600 kb windows of the genomes were scanned to screen for genes near each significant marker-trait association. The screened genes were identified and analyzed, and information on the annotated genes in upland cotton was downloaded from CottonFGD (https://cottonfgd.org) to search for additional potential annotated genes. Transcriptome data of ovule (3, 0, 1, 3, 5, 10, 15, 20, and 25DPA) and fiber tissues (10, 15, 20, and 25DPA) were also downloaded from the NCBI Sequence Read Archive collection PRJNA490626

(Hu et al., 2019). The role of potential candidate genes in responding to abiotic stresses, especially drought, was further analyzed by consulting the relevant literature.

qRT-PCR was used to analyze and screen the relative expression levels of candidate genes related to drought tolerance traits in cotton after drought stress. Leaf samples were collected from drought-tolerant Xinluzao 45 and drought-sensitive Xinluzao 26 (Sun et al., 2021), and total RNA was extracted by a TRIzol kit (Thermo Fisher, Beijing, China). cDNA was synthesized by a one-step RT-PCR kit (Novoprotein Scientific, China). The *GhUBQ7* gene was used as an internal control for data normalization. Gene expression was calculated by the $2^{-\Delta\Delta Ct}$ method (Tanino et al., 2017). The primers selected for this experiment are shown in Supplementary Table S2.

3 Results

3.1 Analysis of variations in yield and agronomic traits

The phenotypic variation in drought tolerance in 150 upland cotton materials was analyzed by measuring 8 drought tolerance related traits, including PH, FBN, EFBN, BN, EBN, SC, LC and SBW. There were differences in all traits between control and drought treatment (Supplementary Table S3 and Supplementary Figure S1). In the control and the two-year average, the PH of the different materials ranged from 52.4-157.2 cm, the FBN ranged from 5.82-15, the EFBN ranged from 5.2-13, the BN ranged from 3.5-15.1, and the EBN ranged from 3.8-14.1 (Supplementary Table S3). SC ranged from 103.15-181.45 g, LC ranged from 35.27-65.03 g, and SBW ranged from 5.16-9.08 (Supplementary Table S3). After exposure to drought stress, averaged over 2 years, the PH of the different materials ranged from 33.02-78.8 cm, the FBN ranged from 3.69-11.1, the EFBN ranged from 3.8-9.05, the BN ranged from 1.34-10, and the EBN ranged from 1.09-4.45. SC ranged from 79.67-138.5 g, LC ranged from 27.9-57.95g, and SBW ranged from 3.98-6.93 g (Supplementary Table S3). Thus, the eight phenotypic traits were affected by drought stress in all samples, with PH, FBN, EFBN, BN, EBN, SC, LC and SBW decreasing by 24.05, 24.89, 21.54, 44.95, 68.05, 20.20, 21.29 and 20.20%, respectively, under drought stress (Supplementary Table S3). Based on the data collected over two years, the coefficient of variation of BN was higher (25.39 and 47.46% for the control and drought stress treatments, respectively), while that of SC was lower (9.17 and 9.59% for the control and drought stress treatments, respectively) (Supplementary Table S3). Except for the control of individual traits, the phenotypic differences of all traits were extremely significant or significant ($p<0.05$) when considering the single year and two-year averages, and the frequency distribution of all traits was consistent with a normal distribution (Supplementary Figure S2). The correlation analysis results for the normal and drought stress treatments in 2019 and 2020 are shown in Supplementary Table S4. In the control, there were extremely significant positive correlations between the five agronomic traits ($p<0.01$). Among the yield component traits, LC and SBW had extremely significant positive correlations with FBN, BN, EBN, EFBN, EBN and SC. However, after drought stress treatment, there were extremely significant positive correlations between all the traits except EFBN and SC; in particular, the yield trait SBW was positively correlated with the other traits ($p<0.01$).

3.2 Group characteristic analysis and LD analysis

To identify drought tolerance genes, we resequenced all the resource samples using upland cotton TM-1 (Yang et al., 2019) as the reference genome, and finally obtained 2,499,987 SNPs (the screening conditions were missing data <20% and MAF<1%). The highest density of SNPs was detected on chromosome A01, while the lowest density of SNPs was detected on chromosome A02, with an average marker density of 1.28 SNPs per kb (Supplementary Table S5). To explore the population structure characteristics and genotype structure of the tested upland cotton germplasm resources, we used ADMIX software. This analysis was based on the maximum likelihood estimation model and cross-validated for the number of subpopulations (k), thus the optimal number of ancestral components was determined (k=1-10). The results of the structural simulation analysis showed that when k=4, the CV error was minimized (Figure 1A). Therefore, a k value of 4 was selected to assess the genetic structure of the 150 cotton genotypes. In a principal component analysis (PCA) of these 150 cotton materials, 36.4% of the genetic variation was explained by the first two PCs (Figure 2). There was abundant genetic variation among the cotton varieties examined in this study. To further analyze the genetic differentiation of genotypes, NJ-based clustering was performed for the samples. Consistent with the ADMIX results, the stratified cluster tree showed significant differences among the variety complexes (Figure 1C). Four main clusters were defined in the tree; these groups corresponded to each of the major subgroups of the ADMIX analysis, which supports the division of the population into four major subgroups (Figure 1B). The corresponding Q matrix at k=4 was used for further marker-trait association mapping.

All identified high-quality SNP markers were used to estimate the degree of LD in the associated population. At the $r^2 = 0.428$ threshold for all chromosomes, the average LD decay distance was approximately 500 kb (Supplementary Figure S4).

3.3 Genome-wide association analysis

To analyze and screen important genetic loci and candidate genes related to yield traits under drought stress conditions, different models were used for GWASs of 8 traits in each single environment and in multiple environments. The FarmCPU software program was used to analyze the associations between the screened SNP markers and the 8 traits in the 150 genotypes to detect marker-trait associations. SNP loci that were significantly associated with yield traits were identified under drought stress, and the loci were stable across environments (Figure 3).

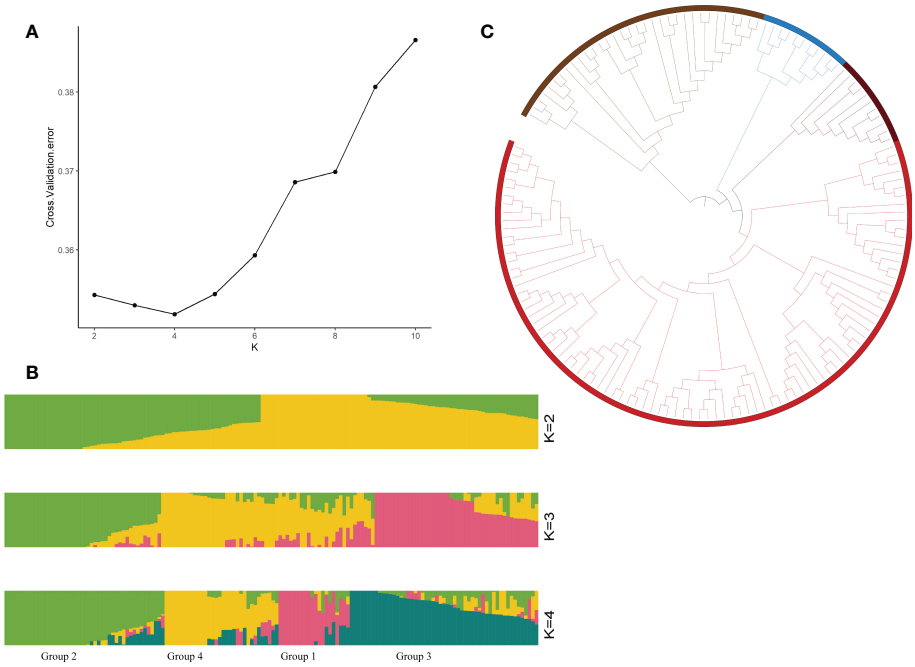


FIGURE 1
Genotyping analyses of 150 cotton germplasms. **(A)** Cross-validation diagram of the SNP dataset, **(B)** population structure analyzed by STRUCTURE at K=2, 3, and 4 (Group 1: High drought resistance, Group 2: Medium drought resistance, Group 3: Drought tolerance, Group 4: Drought sensitivity), and **(C)** phylogenetic tree of the population.

3.4 Yield traits

3.4.1 Control dataset

Under normal control conditions, 718 SNP markers were found to be significantly associated with the three yield traits, scattered across 26 chromosomes (Supplementary Figure S5). The Manhattan plot (Supplementary Figure S5) showed that of the 470 SNPs significantly associated with SBW, 9 were located on chromosome A03, 251 on chromosome A11, 191 on chromosome A12, and 19 on chromosome D06.

3.4.2 Drought treatment dataset

Under drought stress, 126 significant SNPs associated with yield traits were identified. The Manhattan plot (Figure 3) showed that a total of 22 SNP markers above the threshold (Supplementary Table S6) that were associated with SBW were distributed on chromosome D08. The most significantly correlated SNP marker was ChrD08_48059786 (-log (*P*) =6.47). SNP4 (SNP D06_47161952) also had a high -log (*P*) value (5.52) (Supplementary Table S6).

Five important sites associated with SBW were identified. Importantly, SNP 7 (SNP D08_48059786) was located upstream of Gh_D08G143300, and a continuous signal was observed near this site in the Manhattan plot (Figures 2, 3 and Supplemental Table S6). We analyzed LD blocks in the 300 kb region upstream and downstream of this site and found that the SNP was closely related to block LD_SBW (D08: 47.56-48.10) (Figure 2A and Supplementary Table S6). Further haplotype analysis of this region revealed that all materials could be classified into four haplotypes and that HapD08_2, which was located in LD_SBW in this region, was associated with a higher SBW than the other three haplotypes (Figures 2C).

3.5 EFBN

A GWAS was performed for the EFBN phenotype in each environment. A total of nine important EFBN SNPs were identified under drought conditions in both environments, all of which were located on chromosome D04. Two SNPs significantly associated with EFBN were found on this chromosome, located at ChrD04_42670538 (SNP39, -log (*P*) =5.63) and ChrD04_42942695 (SNP45, -log (*P*) =5.55) (Figure 4 and Supplementary Table S6). LD blocks in the 300 kb region upstream and downstream of this site were also analyzed, and LD block analysis showed that the peak SNP was mainly located in 42.90-43.01 Mb of chromosome D04 (Figures 4A, B). The haplotype analysis of this region showed that all materials could be divided into two haplotypes, and the HapD04_1 haplotype samples showed higher EFBN values than the HapD04_2 haplotype samples (Figure 4C).

3.6 EBN

A total of 77 SNPs were significantly associated with EBN. Eighteen of them were consecutive and located on chromosome D08. LD block analysis revealed that the peak SNP D08_48059786 was in the closely linked LD_EBN region (D08: 46.63-48.27 Mb)

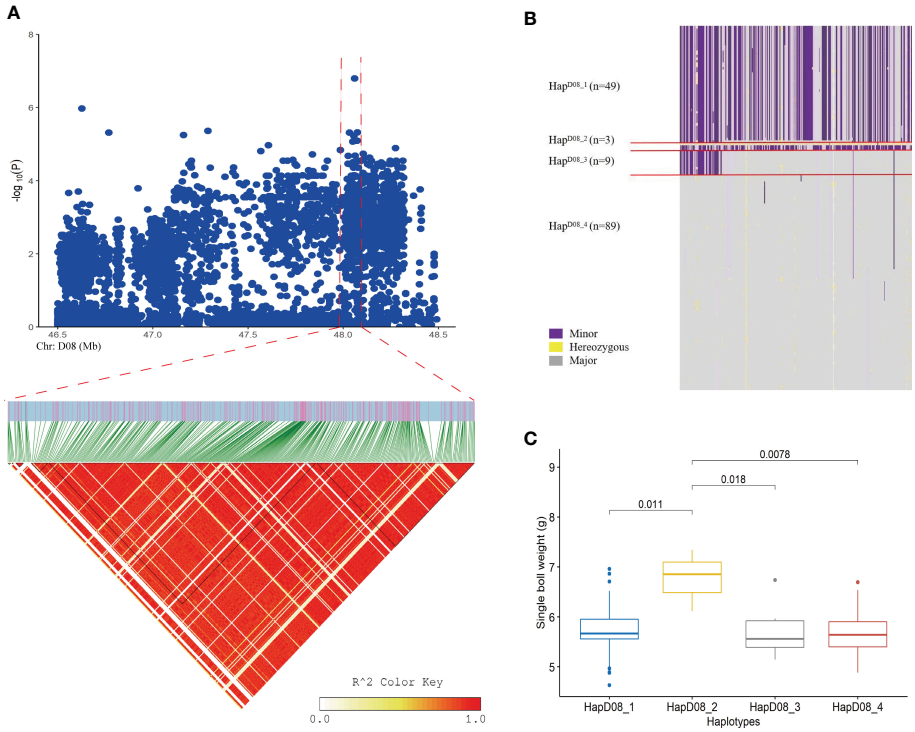


FIGURE 2
Loci related to the SBW trait found on chromosome D08. **(A)** Manhattan plot and LD block analysis of SBW from the GWAS; **(B)** Chr: D08:47.56-48.10 (Mb) interval haplotype analysis; **(C)** Difference analysis of EBN in different haplotypes.

(Supplementary Figure S6A and Supplementary Table S6). Further haplotype analysis of LD_EBN led to the classification of four haplotypes. The difference analysis of four haplotypes in the LD_EBN region showed that the EBN of HapD08_2 was higher than that of the other three haplotypes (Supplementary Figures S6B, C).

3.7 PH

Eleven SNPs were found to be significantly associated with PH under drought stress (Supplementary Table S6). Among the significant SNPs, there was one consecutive SNP signal on chromosomes A03 and A05 in one environment in 2019. In

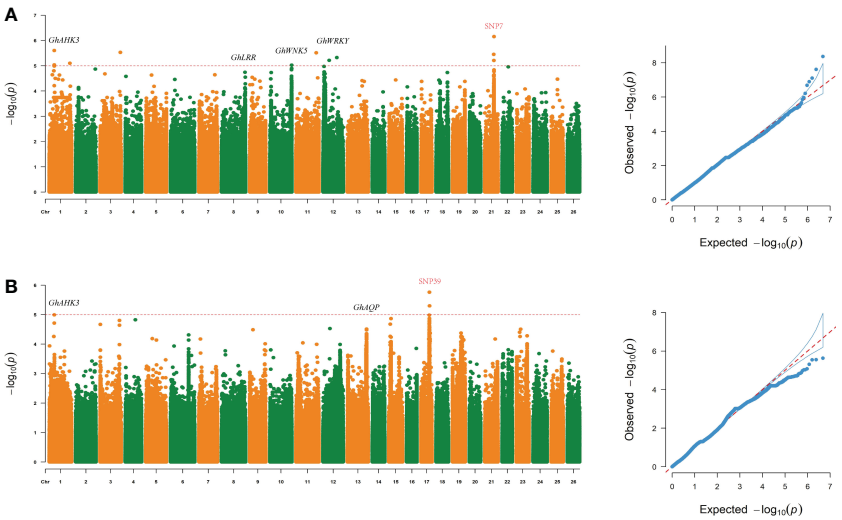


FIGURE 3
Loci related to the SBW and EFBN traits were found on chromosome 17 (chromosome D04) and chromosome 21 (chromosome D08) under drought stress. **(A, B)** Manhattan and QQ plots of GWAS results for SBW and EFBN.

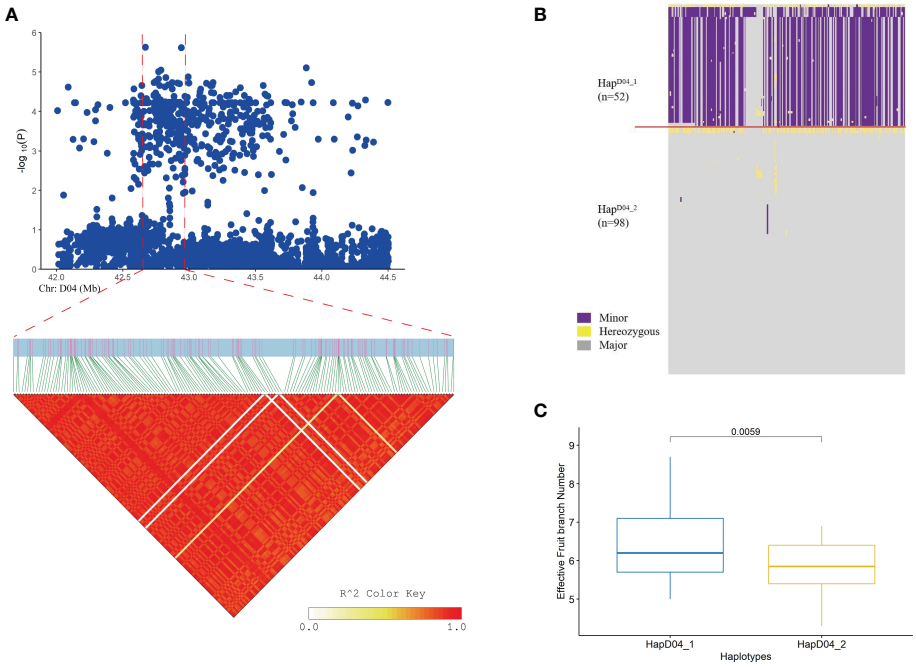


FIGURE 4
Loci related to the EFBN trait found on chromosome D04. **(A)** Manhattan plot and LD block analysis of EFBN from the GWAS; **(B)** Chr: D04.42.90-43.01 (Mb) interval haplotype analysis; **(C)** Difference analysis of EBN in different haplotypes.

another environment, there was a continuous SNP signal on chromosome A08 (Supplementary Table S6).

3.8 FBN

A GWAS of FBN in each environment revealed 91 SNPs significantly related to this trait. There were 9 significant SNPs on chromosome A13 in a single environment, and 7 significant SNPs on chromosome A13 under drought stress. Another continuous SNP signal was detected on chromosome D08 under drought stress, with a total of 19 significant SNP loci, 17 of which were on chromosome D08 (Supplementary Table S6).

3.9 BN

Fifty-nine SNPs were found to be significantly associated with BN (Supplementary Table S6). Under the control conditions in the two environments, continuous SNP signals were found on chromosome A07, with a total of 40 significant SNP sites, and the peak value of the SNPs was mainly distributed between 2.13-2.20 Mb on chromosome A07 (Supplementary Table S6).

3.10 Candidate gene screening and qRT-PCR expression analysis

The LD decay distance can be used as the confidence interval of candidate genes, but due to the characteristics of the cotton genome, the LD decay distance is long. Therefore, in our analysis of the location of significant SNPs in the upland cotton genome, we searched within 500 kb-1 Mb on each side of significant SNP markers to analyze and identify candidate genes associated with drought tolerance traits. Gene functions associated with identified SNPs were assigned using the Universal Protein Database (UniProt) and the Cotton Genome Database (Table 1). In this study, the strongest signals identified on chromosomes D04 and D08 were novel, which resulted in the identification of 15 candidate genes within the candidate intervals on the chromosomes, including 5 on D04 and 10 on D08. Under drought stress, SBW, SC and EBN were mapped to the same region on chromosome D08, while EFBN, another trait significantly associated with yield, was mapped to a region located on chromosome D04 (Figure 3). Among the yield traits, 10 common genes were identified in the common interval of chromosome D08 (Table 1). Gene Ontology (GO) enrichment analysis showed that candidate genes in all ranges were significantly enriched in two functional categories (inorganic diphosphatase activity and phosphate-containing compound metabolic process) (Supplementary Table S7). Kyoto Encyclopedia of Genes and Genomes (KEGG) annotation revealed that the metabolic pathways of these candidate genes are closely related to the biosynthesis of secondary metabolites and plant hormone signal transduction (Supplementary Table S8 and Figure 5B). These candidate genes are important because they are the candidate genes most likely to enhance cotton drought tolerance and mitigate yield loss under drought stress. To further reduce the number of candidate genes, according to published cotton RNA-seq data, we found that there were significant differences in the expression levels of 4 of the 12 functionally annotated genes

TABLE 1 Candidate genes and their annotation of loci related to yield traits under drought stress in GWAS analysis.

Gene ID	Strand	Start	End	Annotation
Gh_D04G138000	–	42952898	42956108	Probable starch synthase 4, chloroplastic/amyloplastic
Gh_D04G138100	+	42971241	42973724	ADP-ribosylation factor GTPase-activating protein AGD7
Gh_D04G138200	–	42980192	42980500	Ribonuclease HI
Gh_D04G138300	–	42997896	42998282	Serine/threonine-protein phosphatase 7 long form homolog
Gh_D04G138400	–	42999850	43003930	Cation/H(+) antiporter 20
Gh_D08G142500	+	47579389	47583876	NAC domain-containing protein 69
Gh_D08G142600	+	47585542	47587607	Unkown
Gh_D08G142700	–	47587523	47589628	Soluble inorganic pyrophosphatase 1
Gh_D08G142800	–	47592151	47594025	3-ketoacyl-CoA synthase 1
Gh_D08G142900	–	47735907	47737840	Unkown
Gh_D08G143000	–	47766366	47767796	Transcription factor bHLH14
Gh_D08G143100	–	47954594	47957599	Rho GTPase-activating protein 3
Gh_D08G143200	–	47963917	47965509	Serine/threonine-protein phosphatase 7 long form homolog
Gh_D08G143300	–	47975232	47976089	Auxin-responsive protein SAUR32
Gh_D08G143400	+	48071810	48072772	Unkown

The positive and negative directions of the chain are chain are indicated by the plus sign + and the minus sign – respectively.

between the control and drought stress treatments. All four of these genes showed upregulated expression after drought stress treatment; moreover, the expression trend of the remaining 8 genes after drought stress treatment was not obvious (Supplementary Figure S7A). In addition, these four genes were highly expressed in ovules and fibers, and three of the genes were significantly highly expressed at 10, 15 and 25 day psot anthesis (DPA) (Figure 5A).

These three highly expressed candidate genes were selected to verify the RNA-seq data. The three candidate genes included one encoding auxin reactive protein (SAUR) and two activating protein

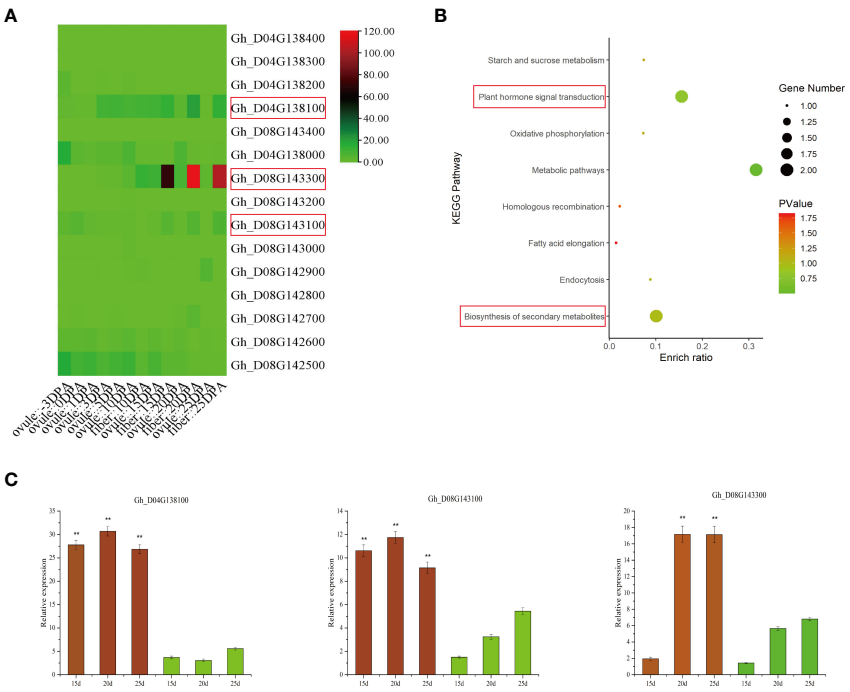


FIGURE 5
Transcriptome data and qRT-PCR analysis of 15 candidate genes. **(A)** Analysis of the transcriptome data of 15 candidate genes at 10, 15 and 20 DPA of ovule and fiber development. **(B)** KEGG analysis of 15 candidate genes. **(C)** qRT-PCR was used to analyze the expression of three genes at 10, 15 and 20 d of fiber development, and significant differences are indicated by ** ($p < 0.01$).

GTPases (Table 1). To determine whether the expression of the three genes was induced by drought stress, one drought-tolerant material (Xinluzao 45) and one drought-sensitive material (Xinluzao 26) were selected to analyze the expression levels of the three genes after drought stress. qRT-PCR was used to detect the relative expression levels of these genes, and the value of the *GhUBQ7* gene was used as the threshold (internal control) for normalization. The results were then compared with the transcriptome results. These results showed that the genes were expressed differently in resistant lines after drought stress, and their expression patterns were basically consistent with the RNA-seq data. The qRT-PCR results for these genes showed significantly higher expression in the drought-tolerant materials than in the drought-sensitive materials (Supplementary Figure S7B). In addition, at 10, 15 and 20 d of fiber development, the expression levels of these three genes in the two materials with different drought tolerance levels were analyzed by qRT-PCR. The results showed that the expression of these three genes in the materials with strong drought tolerance was higher than that in the drought sensitive material (Figure 5C). Therefore, these three genes are significantly related to the drought tolerance of cotton, suggesting their role as candidate drought tolerance genes related to yield traits of cotton under drought.

4 Discussion

Eight yield-related traits of 150 upland cotton germplasms were analyzed by the GWAS approach. In addition, the leaf tissue and fibrous tissue of two materials with different levels of drought resistance were collected and used for qRT-PCR analysis. The results of this study add to the understanding of the variation in yield traits under drought stress. The results can provide a reference for the improvement of cotton molecular breeding under drought conditions.

In all the tested materials, yield traits and other traits were significantly different between the control and drought conditions, indicating a large amount of genetic variation in drought tolerance among the materials. Phenotypic analysis showed that drought treatment significantly affected the traits of the different materials (Supplementary Table S3). SBW, EFBN and EBN were all significantly positively correlated under drought stress, indicating that improving these traits at the same time would lead to an increase in SC yield (Supplementary Table S4) (Sun et al., 2018). The population structure in all the tested materials was analyzed according to the K value, and the population was divided into four categories (Figure 1B), indicating some variation within the population. Phylogenetic analyses showed similar results (Figure 1C), indicating that these analyses can play a role in preventing false positives in GWASs (Soto-Cerda and Cloutier, 2012; Eltaher et al., 2018). The genome-wide LD decayed to half the r^2 (0.428) at 500 kb, and there were a large number of significant SNP markers in LD, suggesting that significant marker-trait association can be found using a GWAS (Park et al., 2008; Schwarz et al., 2015). The population structure shown in the analysis results of the Q-Q diagram is well explained because most of the points are on the diagonal for all traits (Figure 3) (Burghardt et al., 2017; Paterne et al., 2021). Cotton yield is a complex quantitative trait that is greatly affected by the environment. Although cotton resources are abundant, due to the large genome of cotton, the yield traits of cotton have not been fully explored, especially under drought conditions (Sun et al., 2018; Said et al., 2015). Yield traits can indirectly reflect the drought tolerance of cotton (Sun et al., 2021), among which SBW is an important trait related to yield, and EBN is another trait with a significant contribution to yield per plant (Sun et al., 2018; Sun et al., 2021). The results of this study showed that under drought stress, SBW and EBN were stably associated with SNPs on chromosome D08 in both environments, and there were 29 significant SNPs related to SBW and EBN (Figures 3, 2; Supplementary Figure S6A) (Wang et al., 2019), which were different from those located on A07, D03, D06, D09 and D12 (Chen et al., 2008; Ma et al., 2008; Wu et al., 2009; Vollmer et al., 2011; Ning et al., 2013; Yu et al., 2013; Sun et al., 2018). However, Fang et al. identified a significant SNP related to BN that was located adjacent to D08, and there were only three genes in the LD block of this site. One of the genes was differentially expressed in the ovule and fiber of the two different materials, and the haplotype analysis verified this result (Fang et al., 2017). However, the SNP site on chromosome D08 identified in this study is a novel locus associated with yield under drought conditions.

In this study, we identified three candidate genes by GWAS that were supported by published RNA-seq data, one of which is *Gh_D08G143300*, a homolog of Arabidopsis *SAUR32* (Ren and Gray, 2015; Stortenbeker and Bemer, 2019; Zhou et al., 2022). The other two genes are *Gh_D08G143100*, which is homologous to Arabidopsis *ROPGAP3*, and *Gh_D04G138100*, which is homologous to Arabidopsis *AGD7* (Myung et al., 2007; Yoshihisa and Hiroo, 2012). Therefore, *Gh_D08G143300* may also affect auxin synthesis and transport in the fiber under drought stress, leading to the redistribution of auxin and thus promoting the growth of cotton fiber. However, *Gh_D08G143100* may alternatively initiate a unique pattern in the secondary cell wall of fibers under drought stress. *Gh_D04G138100* is activated under drought stress and may be involved in protein transport. The RNA-seq and qRT-PCR results showed that the three candidate genes were differentially expressed in the materials with large differences in drought tolerance (Figure 5 and Supplementary Figure S7). These results suggest that these three candidate genes may be one of the important genes in determining cotton yield formation under drought stress. More studies are needed to further analyze and verify how these three genes affect cotton yield under drought stress and to verify their functions in yield formation under such conditions.

5 Conclusion

To explore the regulatory mechanism of cotton yield variation under drought stress, 150 upland cotton germplasms were selected, and GWAS was conducted on three yield traits (SBW, SC and LC) and five agronomic traits (related to plant height and fruit branch number) that are closely related to yield. The GWAS results revealed a total of 46 significant SNPs under drought stress, and 15 candidate genes were screened. Three differentially expressed genes (*Gh_D04G138100*, *Gh_D08G143100* and *Gh_D08G143300*)

were screened by combining published RNA-seq data. Two materials with extreme drought resistance differences, Xinluzao 45 and Xinluzao 26, were selected, and these two materials also showed significant differences in drought resistance in the field experiment. qRT-PCR was used to verify the expression patterns of the three candidate genes after drought stress in the two materials with drought resistance extremes. The results showed that high expression of these genes was induced by drought stress. At the same time, there were significant differences in the expression of these three genes in the developed fibers. In this paper, we further analyzed the molecular markers and candidate genes related to upland cotton yield under drought stress, and the findings will be helpful for studying the molecular mechanism of cotton yield traits under drought stress.

Data availability statement

The data presented in the study are deposited in the NCBI repository, accession number PRJNA 605345.

Author contributions

FS analyzed the data and drafted the manuscript. JM and YY provided ideas, designed and supervised the experiment. JM and WS provided cotton seeds, and all authors reviewed the manuscript. All authors contributed to the article and approved the submitted version.

Funding

The work was supported by the Major Science and Technology Project of Xinjiang Uygur Autonomous Region, "Study on the key techniques of cotton germplasm resource collection and excellent gene mining" (2022A03004-2).

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.2023.1135302/full#supplementary-material

References

Abdelraheem, A., Adams, N., and Zhang, J. (2020b). Effects of drought on agronomic and fiber quality in an introgressed backcross inbred line population of upland cotton under field conditions. *Field Crops Res.* 254 (1), 107850. doi: 10.1016/j.fcr.2020.107850

Abdelraheem, A., Kuraparthy, V., Hinze, L., Stelly, D., Wedegaertner, T., and Zhang, J. F. (2021). Genome-wide association study for tolerance to drought and salt tolerance and resistance to thrips at the seedling growth stage in US upland cotton. *Ind. Crops Products* 169, 113645. doi: 10.1016/j.indcrop.2021.113645

Abdelraheem, A., Thyssen, G. N., Fang, D. D., Jenkins, J. N., McCarty, J. C., Wedegaertner, T., et al. (2020a). GWAS reveals consistent QTL for drought and salt tolerance in a MAGIC population of 550 lines derived from intermating of 11 upland cotton (*Gossypium hirsutum*) parents. *Mol. Genet. Genomics* 296 (1), 119–129. doi: 10.1007/s00438-020-01733-2

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

Burghardt, L. T., Young, N. D., and Tiffin, P. (2017). A guide to genome-wide association mapping in plants. *Curr. Protoc. Plant Biol.* 2, 22–38. doi: 10.1002/cppb.20041

Cattivelli, L., Rizza, F., Badeck, F. W., Mazzucotelli, E., Mastrangelo, A. M., Francia, E., et al. (2008). Drought tolerance improvement in crop plants: An integrated view from breeding to genomics. *Field Crops Res.* 105, 1–14. doi: 10.1016/j.fcr.2007.07.004

Chen, L., Zhang, Z. S., Hu, M. C., Wang, W., Zhang, J., Liu, D. J., et al. (2008). Genetic linkage map construction and QTL mapping for yield and fiber quality in upland cotton (*Gossypium hirsutum l.*). *Acta Agronomica Sinica.* 34 (7), 1199–1205. doi: 10.3724/SP.J.1006.2008.01199

Ding, F., Lv, J., Liu, Q., Guo, Y., He, W. Q., Wang, L., et al. (2021). Migraton of cotton planting regions and residual pollution of mulch film in China. *J. Huazhong Agric. Univ.* 40 (06), 60–67.

Du, L. Y., Huang, X. L., Ding, L., Wang, Z. X., Tang, D. L., Chen, B., et al. (2022). TaERF87 and TaAKS1 synergistically regulate TaP5CS1/TaP5CR1-mediated proline biosynthesis to enhance drought tolerance in wheat. *New Phytologist.* 20, 18549. doi: 10.1111/nph.18549

Du, X. M., and Zhou, Z. L. (2005). *Description specifications and data standards for cotton germplasm resources* (Beijing: China Agriculture Press).

Eltaher, S., Sallam, A., Belamkar, V., Emara, H. A., Nower, A. A., Salem, K. F., et al. (2018). Genetic diversity and population structure of $F_{3:6}$ Nebraska winter wheat genotypes using genotyping-by-sequencing. *Front. Genet.* 9, 76. doi: 10.3389/fgene.2018.00076

Fang, L., Wang, Q., Hu, Y., Jia, Y. H., Chen, J. D., Liu, B. L., 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

Gao, H. J., Cui, J. J., Liu, S. X., Wang, S. H., Lian, Y. Y., Bai, Y. T., et al. (2022). Natural variations of ZmSRO1d modulate the trade-off between drought resistance and yield by affecting ZmRBOHC-mediated stomatal ROS production in maize. *Mol. Plant* 15 (10), 1558–1574. doi: 10.1016/j.molp.2022.08.009

He, S. P., Sun, G. F., Geng, X. L., Gong, W. F., Dai, P. H., Jia, Y. H., et al. (2021). The genomic basis of geographic differentiation and fiber improvement in cultivated cotton. *Nat. Genet.* 53, 916–924. doi: 10.1038/s41588-021-00844-9

Helga, T., Robinson, J. T., and Mesirov, J. P. (2012). Integrative genomics viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.* 14, 178–192. doi: 10.1093/bib/bbs017

Hou, S., Zhu, G., Li, Y., Li, W. E., 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, 1–15. doi: 10.3389/fpls.2018.01276

Hu, Y., Chen, J. D., Fang, L., Zhang, Z. Y., Ma, W., Niu, Y. C., et al. (2019). Gossypium barbadense and gossypium hirsutum genomes provide into the origin and evolution of allotetraploid cotton. *Nat. Genet.* 51 (4), 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, C., Shen, C., Wen, T., Gao, B., Zhu, D., Li, D. G., et al. (2021). Genome-wide association mapping for agronomic traits in an 8-way upland cotton MAGIC population by SLAF-seq. *Theor. Appl. Genet.* 134 (8), 2459–2468. doi: 10.1007/s00122-021-03835-w

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 genomewide association studies. *Nat. Genet.* 42, 348–354. doi: 10.1038/ng.548

Li, Y., Cao, K., Zhu, G., Fang, W., Chen, C., Wang, X., et al. (2019). Genomic analyses of an extensive collection of wild and cultivated accessions provide new insights into peach breeding history. *Genome Biol.* 20, 36. doi: 10.1186/s13059-019-1648-9

Li, H. M., Liu, S. D., Ge, C. W., Zhang, X. M., Zhang, S. P., Chen, J., et al. (2019). Analysis of drought tolerance and associated traits in upland cotton at the seedling stage. *Int. J. Mol. Sci.* 20, 3888. doi: 10.3390/ijms20163888

Li, M., Yeung, J. M., Cherny, S. S., and Sham, P. C. (2012). Evaluating the efective numbers of independent tests and signifcant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum. Genet.* 131, 747–756. doi: 10.1007/s00439-011-1118-2

Ma, X. X., Ding, Y. Z., Zhou, B. L., Guo, W. Z., and Zhang, T. Z. (2008). QTL mapping in a-genome diploid Asiatic cotton and their congruence analysis with AD-genome tetraploid cotton in genus gossypium. *J. Genet. Genomics* 35 (12), 751–762. doi: 10.1016/S1673-8527(08)60231-3

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

Myung, K. M., Soo, J. K., Yansong, M., Juyoun, S., Jiang, L. W., and Inhwan, H. (2007). Overexpression of arabidopsis AGD7 causes relocation of golgi-localized proteins to the endoplasmic reticulum and inhibits protein trafficking in plant cells. *Plant Physiol.* 143 (4), 1601–1614. doi: 10.1104/pp.106.095091

Ning, Z. Y., Zhao, R., Chen, H., Ai, N. J., Zhan, X., Zhao, J., et al. (2013). Molecular tagging of a major quantitative trait locus for broad-spectrum resistance to verticillium wilt in upland cotton cultivar prema. *Crop Sci.* 53, 2304–2312. doi: 10.2135/cropsci2012.12.0694

Park, H. Y., Seok, H. Y., Park, B. K., Kim, S. H., Goh, C. H., Lee, B. H., et al. (2008). Overexpression of arabidopsis ZEP enhances tolerance to osmotic stress. *Biochem. Biophys. Res. Commun.* 375, 80–85. doi: 10.1016/j.bbrc.2008.07.128.

Paterne, A. A., Norman, P. E., Asiedu, R., and Asfaw, A. (2021). Identification of quantitative trait nucleotides and candidate genes for tuber yield and mosaic virus tolerance in an elite population of white guinea yam (*Dioscorea rotundata*) using genome-wide association scan. *BMC Plant Biol.* 21, 552. doi: 10.1186/s12870-021-03314-w.

Rambaut, A. (2009). FigTree v1.3.1. 2006–2009.

Ren, H., and Gray, W. M. (2015). SAUR proteins as effectors of hormonal and environmental signals in plant growth. *Mol. Plant* 8, 1153–1164. doi: 10.1016/j.molp.2015.05.003.

Said, J. I., Song, M., Wang, H., Lin, Z., Zhang, X., Fang, D. D., et al. (2015). A comparative meta-analysis of QTL between intraspecifc *Gossypium hirsutum* and interspecifc *G. hirsutum* × *G. barbadense* populations. *Mol. Genet. Genomics* 290, 1003–1025. doi: 10.1007/s00438-014-0963-9.

Saleem, M. A., Malik, T. A., Shakeel, A., and Ashraf, M. (2015). QTL mapping for some important drought tolerant traits in upland cotton. *J. Anim. Plant Sci.* 25, 502–509..

Sang, X. H., Zhao, Y. L., Wang, H. M., Chen, W., Gong, H. Y., Zhao, P., et al. (2017). Association analysis of drought tolerance and SSR markers in upland cotton. *Cotton Sci.* 29 (03), 241–252.

Schwarz, N., Armbruster, U., Iven, T., Brückle, L., Melzer, M., Feussner, I., et al. (2015). Tissue-specific accumulation and regulation of zeaxanthin epoxidase in arabidopsis reflect the multiple functions of the enzyme in plastids. *Plant Cell Physiol.* 56, 346–357. doi: 10.1093/pcp/pcu167

Shukla, R. P., Tiwari, G. J., Joshi., B., Song-Beng, K., Tamta, S., Boopathi, N. M., et al. (2021). GBS-SNP and SSR based genetic mapping and QTL analysis for drought tolerance in upland cotton. *Physiol. Mol. Biol. Plants* 27 (8), 1731–1745. doi: 10.1007/s12298-021-01041-y

Soto-Cerda, B. J., and Cloutier, S. (2012). Association mapping in plant genomes. *Genet. Divers. Plants*, 29–54. doi: 10.5772/3305

Stortenbeker, N., and Bemer, M. (2019). The SAUR gene family: The plant's toolbox for adaptation of growth and development. *J. Exp. Bot.* 70 (1), 17–27. doi: 10.1093/jxb/ery332

Sun, F., Chen, Q., Chen, Q., Jiang, M., Gao, W., and Qu, Y. (2021). Screening of key drought tolerance indices for cotton at the flowering and boll setting stage using the dimension reduction method. *Front. Plant Sci.* 12, 619926. doi: 10.3389/fpls.2021.619926

Sun, Z., Wang, X., Liu, Z., Gu, Q., Zhang, Y., Li, Z., et al. (2017). Genome-wide association study discovered genetic variation and candidate genes of fibre quality traits in gossypium hirsutum l. *Plant Biotechnol. J.* 15 (8), 982–996. doi: 10.1111/pbi.12693

Sun, Z. W., Wang, X. F., Liu, Z. W., Gu, Q. S., Zhang, Y., Li, Z. K., et al. (2018). A genome-wide association study uncovers novel genomic regions and candidate genes of yield-related traits in upland cotton. *Theor. Appl. Genet.* 131, 2413–2425. doi: 10.1007/s00122-018-3162-y

Sun, X. M., Xiong, H. Y., Jiang, C. H., Zhang, D. M., Yang, Z. L., Huang, Y. P., et al. (2022). Natural variation of DROT1 confers drought adaptation in upland rice. *Nat. Commun.* 13, 1–17. doi: 10.1038/s41467-022-31844-w

Tanino, Y., Kodama, M., Daicho, H., Yoshito, M., Towa, Y., Yukiji, Y., et al. (2017). Selection of laboratory procedures to detect toxigenic by the 2-step method. *Rinsho Biseibutshu Jinsoku Shindan Kenkyukai shi.* 27, 9–14.

Ul-Allah, S., Rehman, A., Hussain, M., and Faroop, M. (2021). Fiber yield and quality in cotton under drought: Effects and management. *Agric. Water Manage.* 255, 106994. doi: 10.1016/j.agwat.2021.106994

Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). Ensem-blCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* 19, 327–335. doi: 10.1101/gr.073585.107

Vollmer, A. H., Youssef, N. N., and Dewald, D. B. (2011). Unique cell wall abnormalities in the putative phosphoinositide phosphatase mutant AtSAC9. *Planta* 234, 993–1005. doi: 10.1007/s00425-011-1454-4

Wang, X. L., Feng, W., Wang, H. R., Wang, Q. K., Wei, Z., Zhang, G. H., et al. (2019). Multi-environments and multi-models association mapping identified candidate genes of lint percentage and seed index in *Gossypium hirsutum l. Mol. Breed.* 39, 149. doi: 10.1007/s11032-019-1063-7

Wang, H. R., Hng, S. Y., and Qin, D. Q. (2017). Discussion on related issues of drought and water shortage. *Water Resour. Prot.* 33 (05), 1–4. doi: 10.3880/j.issn.1004-6933.2017.05.001

Wu, J. X., Gutierrez, O. A., Jenkins, J. N., Mccarty, J. C., and Zhu, J. (2009). Quantitative analysis and QTL mapping for agronomic and fiber traits in an RI population of upland cotton. *Euphytica* 165, 231–245. doi: 10.1007/s10681-008-9748-8

Xiao, J. (2020). Exploration of agricultural water management system in xinjiang. *Henan Water Resour. South-to-North* 49 (01), 31–32.

Yang, Z. E., Ge, X. Y., Yang, Z. R., Qin, W. Q., Sun, G. F., Wang, Z., et al. (2019). Extensive intraspecific gene order and gene structural variations in upland cotton cultivars. *Nat. Commun.* 10, 2989. doi: 10.1038/s41467-019-10820-x

Yoshihisa, O., and Hiroo, F. (2012). Initiation of cell wall pattern by a rho- and microtubule-driven symmetry breaking. *Science* 337 (6100), 1333–1336. doi: 10.1126/science.1222597

Yu, J. W., Zhang, K., Li, S. Y., Yu, S. X., Zhai, H. H., Wu, M., et al. (2013). Mapping quantitative trait loci for lint yield and fiber quality across environments in a *Gossypium hirsutum* × *Gossypium barbadense* backcross inbred line population. *Theor. Appl. Genet.* 126, 275–287. doi: 10.1007/s00122-012-1980-x

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

Zhou, Y., Lan, Q., Yu, W., Zhou, Y., Ma, S., Bao, Z., et al. (2022). Analysis of the small auxin-up RNA (SAUR) genes regulating root growth angle (RGA) in apple. *Genes.* 13 (11), 2121. doi: 10.3390/genes13112121

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 30 May 2023
DOI 10.3389/fpls.2023.1189038

OPEN ACCESS

EDITED BY
Yan Zhao,
Shandong Agricultural University, China

REVIEWED BY
Fang Wang,
Shandong Agricultural University, China
Zhi Zou,
Chinese Academy of Tropical Agricultural Sciences, China
Jun Zhao,
Jiangsu Academy of Agricultural Sciences, China

*CORRESPONDENCE
Jinhong Chen
jinhongchen@zju.edu.cn
Ziji Liu
Liuziji1982@163.com
Fenglin Gu
Xiaogu4117@163.com

RECEIVED 18 March 2023
ACCEPTED 26 April 2023
PUBLISHED 30 May 2023

CITATION
Tang W, Hao Y, Ma X, Shi Y, Dang Y, Dong Z, Zhao Y, Zhao T, Zhu S, Zhang Z, Gu F, Liu Z and Chen J (2023) Genome-wide analysis and identification of stress-responsive genes of the CCCH zinc finger family in *Capsicum annuum* L. *Front. Plant Sci.* 14:1189038. doi: 10.3389/fpls.2023.1189038

Genome-wide analysis and identification of stress-responsive genes of the CCCH zinc finger family in *Capsicum annuum* L.

Wenchen Tang[1,4], Yupeng Hao[1,4], Xinyu Ma[1,4], Yiqi Shi[1,4], Yongmeng Dang[1,4], Zeyu Dong[1,4], Yongyan Zhao[1,4], Tianlun Zhao[1,4], Shuijin Zhu[1,4], Zhiyuan Zhang[1,4], Fenglin Gu[2]*, Ziji Liu[3]* and Jinhong Chen[1,4]*

[1]Hainan Institute, Zhejiang University, Sanya, China, [2]Spice and Beverage Research Institute, Sanya Research Institute, Chinese Academy of Tropical Agricultural Sciences/Hainan Key Laboratory for Biosafety Monitoring and Molecular Breeding in Off-Season Reproduction Regions, Sanya, China, [3]Tropical Crops Genetic Resources Institute, Chinese Academy of Tropical Agricultural Sciences/Key Laboratory of Crop Gene Resources and Germplasm Enhancement in Southern China, Ministry of Agriculture, Haikou, China, [4]College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, China

The CCCH zinc finger gene family encodes a class of proteins that can bind to both DNA and RNA, and an increasing number of studies have demonstrated that the CCCH gene family plays a key role in growth and development and responses to environmental stress. Here, we identified 57 CCCH genes in the pepper (*Capsicum annuum* L.) genome and explored the evolution and function of the CCCH gene family in *C. annuum*. Substantial variation was observed in the structure of these CCCH genes, and the number of exons ranged from one to fourteen. Analysis of gene duplication events revealed that segmental duplication was the main driver of gene expansion in the CCCH gene family in pepper. We found that the expression of CCCH genes was significantly up-regulated during the response to biotic and abiotic stress, especially cold and heat stress, indicating that CCCH genes play key roles in stress responses. Our results provide new information on CCCH genes in pepper and will aid future studies of the evolution, inheritance, and function of CCCH zinc finger genes in pepper.

KEYWORDS

gene family, CCCH, phylogenetic analysis, *Capsicum annuum*, stress

Introduction

Zinc finger proteins (ZFPs), which are named for their ability to bind zinc to form a stable finger-like structure, are sequence-specific transcription factors that usually contain varying numbers of cysteine (Cys) and histidine (His) residues. Cys and His are used to chelate zinc ions to form a zinc finger structure, which can recognize and bind to DNA

(Hall, 2005). Zinc finger proteins are also associated with the metabolism of different types of RNAs in organisms (Hall, 2005) and can specifically bind to DNA, RNA, and DNA–RNA complexes to regulate gene expression. Several gene families have been identified in plants based on their function and structure, including the RING finger (Freemont, 1993; Kosarev et al., 2002), CCCH (Li et al., 2001), DOF (Lijavetzky et al., 2003), WRKY (Zhang and Wang, 2005), ERF (Nakano et al., 2006), and LIM (Arnaud et al., 2007) families. Zinc finger protein motifs can be divided into different types according to the number of conserved Cys and His residues and the spacing between these residues, such as C2H2, C8, C6, C3HC4, C2HC5, C4, C4HC3, and CCCH (Berg and Shi, 1996; Takatsuji, 1998; Moore and Ullman, 2003; Schumann et al., 2007). CCCH zinc finger proteins generally contain at least one zinc finger motif. Three Cys and one His residue are the most important components of this motif. The common sequence of the CCCH motif can be defined as C-X_{4-15}-C-X_{4-6}C-X_{3-4}-H (where X stands for any amino acid, numbers indicate the number of amino acids, C is Cys, and H is His), and C-X_{7-8}-C-X_5-C-X_3-H is the largest sequence among CCCH proteins (Wang et al., 2008).

CCCH zinc finger proteins are involved in plant development, adaptation, hormonal regulation, and the regulation of processes related to physiological adversity, especially responses to biotic and abiotic stress. In *Arabidopsis*, *AtTZF1*, which consists of two zinc finger motifs separated by 18 amino acids, is a CCCH-type zinc finger protein (Iuchi and Kuldell, 2005). Overexpression of *AtTZF1* enhances the tolerance of *Arabidopsis thaliana* to cold and drought stress and affects the growth and stress responses mediated by abscisic acid (ABA) and gibberellic acid (GA) (Lin et al., 2011). The expression patterns of *AtTZF1*, *AtTZF2*, and *AtTZF3* are similar (Lee et al., 2012). *AtC3H49*/*AtTZF3* and *AtC3H20*/*AtTZF2* can regulate growth rate, plant size, leaf and flower morphology, as well as aging and lifespan. Overexpression of these two genes can attenuate transpiration, enhance drought tolerance, alter growth patterns, and delay senescence (Lee et al., 2012). In addition, the CCCH zinc finger proteins HUA1 and HUA2 play a role in AGAMOUS pre-mRNA processing and in floral reproductive organ identity (Li et al., 2001; Cheng et al., 2003). In rice, OsTZF1 improves stress tolerance by regulating the RNA metabolism of stress-responsive genes (Jan et al., 2013). GhZFP1 in cotton contains two typical zinc finger motifs (C-X_8-C-X_5-C-X_3-H and C-X_5-C-X_4-C-X_3-H) that improve drought and disease resistance in transgenic tobacco (Guo et al., 2009). The overexpression of *GmZF351* in transgenic soybeans activates lipid biosynthesis genes, accelerates the accumulation of seed oil, and thus increases the seed oil content (Li et al., 2017). In cucumber, *CsSEF1* encodes protein containing three conserved zinc finger motifs, two of which are CCCH motifs. The expression of *CsSEF1* is up-regulated in leaves and flowers; it plays a role in later developmental stages after embryogenesis and the signal transduction pathway of fruits from photoassimilate limitation to the sink organs (Grabowska et al., 2009; Tazuke and Asayama, 2013). In pepper, the CCCH zinc finger protein CaC3H14 regulates antagonistic interactions between salicylic acid (SA) and jasmonic acid (JA)/ethylene (ET) signaling, which enhances the resistance of plants to *Ralstonia solanacearum* infection (Qiu et al., 2018).

A total of 68, 67, 68, 91, 34, 62, 80, 89, 103, 116, 31, and 86 CCCH zinc finger family genes have been identified in *Arabidopsis* (Wang et al., 2008), rice (Wang et al., 2008), maize (Peng et al., 2012), poplar (Chai et al., 2012), *Medicago truncatula* (Zhang et al., 2013), citrus (Liu et al., 2014), tomato (Xu, 2014), banana (Mazumdar et al., 2017), cabbage (*Brassica rapa*) (Pi et al., 2018), soybean (Hu and Zuo, 2021), rose (Li et al., 2021), and tobacco (Tang C. et al., 2022), respectively. Although CCCH zinc finger proteins play an important role in many aspects of plant growth and development, no systematic studies have been conducted to analyze and identify members of the CCCH gene family in pepper to date.

Pepper has the highest vitamin C content among all vegetables, which can promote appetite and improve digestion. Whole-genome sequencing and bioinformatics analysis can be used to identify and analyze CCCH zinc finger genes involved in the growth and development, metabolism, and adaptation to stress in pepper plants (Kim et al., 2014; Qin et al., 2014). Here, we identified 57 CCCH zinc finger genes in the pepper genome. We also systematically analyzed the phylogenetic structure, domains, conserved motifs, chromosome localization, duplication events, collinearity, and tissue-specific expression patterns of these CCCH zinc finger genes, and this provided insights into the roles of CCCH gene family members in the growth and development of pepper plants. Finally, the published RNA sequencing (RNA-seq) data were used to investigate the expression of CCCH genes in different tissues, such as the roots, stems, and leaves, and the expression patterns of the genes were validated using quantitative real-time polymerase chain reaction (qRT-PCR). These results provide new insights that will aid future studies of the functions of candidate genes involved in the growth, development, adaptation, hormone regulation, and stress physiology of pepper plants.

Materials and methods

Identification and characterization of CCCH zinc finger family members in pepper

In this study, we used genomic data from *Capsicum annuum* cv. CM334. First, we downloaded amino acid sequences for all *Capsicum* proteins from the Phytozome database[1] (Tuskan et al., 2006; Goodstein et al., 2012) and amino acid sequences for CCCH (PF00642, Zinc finger C-X_8-C-X_5-C-X_3-H type, and similar sequences) from the Pfam database[2] (El-Gebali et al., 2019). The CCCH motif was used to retrieve the amino acid sequence of peppers in hmmsearch[3] with a threshold of E-value < 1×10^{-5}. All the obtained protein sequences were submitted to the Pfam

1 https://phytozome-next.jgi.doe.gov

2 http://pfam.xfam.org/

3 http://www.hmmer.org/

database and SMART domain search database[3] to confirm the structural integrity of the zf_CCCH domain. Furthermore, we made use of the Pfam[2] and SMART[4] databases to clarify the structural integrity of the ZF_CCCH domain (Schultz et al., 2000). We extracted sequences of the conserved domains from the identified pepper CCCH proteins. We used the ExPASy tool[5] (Gasteiger et al., 2005) to calculate the number of amino acids, isoelectric point (pI), molecular weight (Mw), and other physical and chemical properties of the zinc finger CCCH protein sequences.

Classification and sequence analysis of the CCCH genes

We downloaded amino acid sequences for pepper, tomato, and rice from the Phytozome database[1]. *Arabidopsis* CCCH zinc finger genes were identified from the *Arabidopsis* information resource website[6]. Sequences were aligned using the neighbor-joining method, and the evolutionary tree was constructed in MEGA 11 software (Kumar et al., 2018). Branch support was tested by performing 1,000 bootstrap replications. The phylogenetic tree was uploaded in Newick format to the EvolView web server[7] to visualize the tree. The subfamily classification of the *Capsicum* CCCH gene family was based on a previously published classification for *Arabidopsis thaliana* (Wang et al., 2008). MCScanX[8] was used to characterize syntenic relationships among *CCCH* genes in *Arabidopsis*, tomato, and pepper.

Gene structure and conserved motif analysis

We downloaded genome sequences and coding sequences from the Phytozome database[1] to analyze the structure of CCCH gene family members. The structure of the CCCH genes was plotted using TBtools (Chen et al., 2020). MEME Suite Version 5.4.1[9] was used to identify the conserved motifs of CCCH gene family members in pepper, with the maximum motif search number set to 10, and other parameters set to their default values. Any repetitions were considered a motif position that was distributed throughout the sequence (Bailey et al., 2009).

Chromosome location and collinearity analysis

Detailed chromosomal mapping was obtained from GFF genomic files downloaded from the Phytozome database[1] to visualize the chromosomal distribution of the CCCH genes in pepper in TBtools (Chen et al., 2020). We also identified tandem duplication events in CCCH family genes using MCScanX in TBtools. MCScanX in TBtools and BLASTP searches were used to identify the segmental duplication events of CCCH genes in pepper and clarify collinearity relationships between genes in different species (Wang et al., 2012; Chen et al., 2020). The non-synonymous (Ka) and synonymous (Ks) substitutions between gene pairs was calculated by using TBtools.

Analysis of CCCH gene expression by RNA-seq under different conditions

We analyzed the expression profiles of pepper CCCH zinc finger genes in different tissues, under different types of biotic stress and abiotic stress, and in the presence of different phytohormones by downloading the following RNA-seq datasets from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus[10]: flower, root, stem, placenta, and pericarp (stage 1, 2, and 3) of pepper plants during the mature green (MG) stage, breaker (B) stage, and 5 and 10 days after the breaker stage (BioProject ID: PRJNA223222); 30 min, 4 h, 1 day, 2 days, and 3 days after infection with PepMoV and TMV (BioProject ID: PRJNA223222); 1, 3, 6, 12, and 24 h under cold, heat, drought, and salt stress (BioProject ID: PRJNA525913); and 1, 3, 6, 12, and 24 h after MeJA, SA, ET, and ABA treatment (BioProject ID. PRJNA634831) (Kim et al., 2014; Kang et al., 2020; Lee et al., 2020). The fragments per kilobase of exon model per million mapped reads (FPKM) values were calculated using Hisat2 (v2.0.5) and Sringtie (v2.1.7) software with the following formula: log(FPKM+1). These data were then visualized using the 'pheatmap' package in R software.

Stress treatments and collection of materials

In this experiment, gene expression levels of CCCH genes were detected using the pepper cultivar CM334. All pepper plants were sown and grown under greenhouse conditions (16 h light/8 h dark, 25-28°C). When peppers had six true leaves, the experimental groups were subjected to cold treatment (16 h light/8 h dark, 10°C) and heat treatment (16 h light/8 h dark, 40°C) in the incubator, and the leaves were collected at 0, 3, 6, 12, 24, and 72 h after the treatment. Three replicates were collected from three different plants, immediately frozen in liquid nitrogen, and then stored in a -80° C refrigerator.

4 http://smart.embl.de/smart/batch.pl

5 http://web.expasy.org/

6 https://www.arabidopsis.org/

7 http://www.evolgenius.info/evolview/

8 https://github.com/wyp1125/MCScanX

9 https://meme-suite.org/meme/index.html

10 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi

qRT-PCR verification

The RNA sample was extracted using an RNAprep Pure Plant Plus Kit (Tiangen) according to the manufacturer's instructions. The DNAase-treated RNA was reverse-transcribed with M-MLV (RNase H-) reverse transcriptase. qRT-PCR was performed using a CFX96TM Real-Time system (Applied Biosystems). Primers (20-24 bp) were designed using the Primer-BLAST tool in NCBI, and the amplicon lengths were 80-220 bp (Supplementary Table 1). All settings were set to their default values. Three technical replicates were performed for each gene, and *UBI3* was used as the internal reference gene. The total volume of each reaction was 20 μL, which consisted of 2 μL of cDNA, 1 μL of gene-specific primers, 7 μL of ddH_2O, and 10 μL of 2× ChamQ Universal SYBR qPCR Master Mix reagent. The thermal cycling conditions were as follows: 95°C for 10 min, followed by 40 cycles at 95°C for 15 s and 60°C for 1 min. At the end of the cycle, a solubility-free curve was generated to analyze the expression of each gene tested.

Results

Identification and characterization of CCCH transcription factor family members in pepper

In this study, 57 CCCH genes were identified from the *C. annuum* cv. CM334 genome using the Hidden Markov Model of LEA against the genome database of *C. annuum*. These CCCH genes were renamed from *PEPTY1* to *PEPTY57* according to their order on chromosome 1-12 (Supplementary Table 2). All identified CCCH genes encoded proteins ranging from 295 to 1015 amino acids, and their predicted isoelectric points (pI) ranged from 4.7 to 9.39. To investigate the sequence characteristics of the most common CCCH motifs in the pepper CCCH zinc finger proteins, we extracted amino acid sequences from CCCH conserved regions (Thompson et al., 1997). The CCCH domain mainly consisted of a triple cysteine and a histidine, and the following motif was commonly observed ($C\text{-}X_{7\text{-}8}\text{-}C\text{-}X_5\text{-}C\text{-}X_3\text{-}H$) (Supplementary Figure 1).

Phylogenetic tree and sequence structure analysis

We constructed a phylogenetic tree using the entire amino acid sequence of each member of pepper, *Arabidopsis*, tomato, and rice to explore the evolutionary relationships among CCCH zinc finger genes. As shown in Figure 1, the pepper CCCH zinc finger genes were divided into 12 groups based on previous studies of *Arabidopsis*. The number of CCCH zinc finger genes in each group was uneven. Group XII was the largest (13 CCCH zinc finger genes), followed by Group I (8 CCCH zinc finger genes) and Group II, VII, and VIII (each with 2 CCCH zinc finger genes). Group III, IV, V, VI, IX, X, and XI have 3, 4, 5, 3, 3, 6, and 6 CCCH zinc finger genes, respectively.

We performed a structure analysis of the 57 CCCH zinc finger genes in pepper. All the CCCH genes had introns and exons, but they varied greatly in size and number. The number of exons ranged from 1 to 14 (Supplementary Table 3). Most of the genes had less than 10 exons. The average number of exons per gene was 5.4. Genes in Group VI and VII both had two exons, and genes in Group XII contained only one exon. However, genes in Group VIII had 10 exons. Subsequently, the conserved motifs of the CCCH genes in pepper were identified using the online MEME suite program. Ten conserved motifs were detected, ranging from 6 to 50 amino acids in length (Figure 2; Supplementary Table 3). Unsurprisingly, the structure of the genes in the same subclade was similar. The five conserved motifs 1, 5, 6, 7, and 8, were all found in Group I. Motif 5, 7, and 8 had the $C\text{-}X_8\text{-}C\text{-}X_5\text{-}C\text{-}X_3\text{-}H$ structure. Motif 4 was only present in Group X, motif 5 was widely present in Group V and VI, motif 9 was only present in Group XII, and motif 10 ($C\text{-}X_7\text{-}C\text{-}X_5\text{-}C\text{-}X_3\text{-}H$) was only present in Group XI. Most genes in the same branch had similar conserved motif compositions and structures, which suggests that they were functionally similar.

Chromosomal locations and duplications of CCCH zinc finger genes in pepper

Using the pepper genome annotation information and TBtools (Tuskan et al., 2006; Chen et al., 2020), we characterized the chromosomal distribution of CCCH zinc finger genes. A total of 55 of the 57 CCCH zinc finger genes identified could be mapped on chromosomes; *PEPTY56* and *PEPTY57* were the two genes that could not be mapped. As shown in Figure 3, these 55 CCCH genes were unevenly distributed across the 12 chromosomes, and the number of genes on each chromosome was not related to chromosome size. For example, the largest chromosome (Chr 01) contained seven CCCH genes; however, the chromosome containing the most genes was Chr 11, which had eight CCCH genes. Chr 05 and 12 had only two CCCH genes, which was the same number of CCCH genes contained on the shortest chromosome (Chr 08).

Next, we identified tandem duplication events using the Multiple Collinearity Scan toolkit (MCScanX) in TBtools. No tandem duplication events were identified. Thus, we identified segmental duplication events using MCScanX in TBtools and BLASTP searches (Wang et al., 2012; Chen et al., 2020). A total of 5 segmentally duplicated gene pairs were detected, and these were detected across nine chromosomes (Figure 4). On chromosomes 10, 2 pairs of genes (*PEPTY42/PEPTY45* and *PEPTY43/PEPTY44*) on the same chromosomes appear to be products of segmental duplication events. Segmental duplication events were not detected on Chr 01, 04, 07, 09, and 12. These findings indicate that segmental duplication events appear to have played a key role in shaping the diversity of CCCH genes in pepper.

We also investigated collinearity relationships between pepper CCCH genes and associated genes from *Arabidopsis* and *Solanum lycopersicum* to identify homologous genes. Collinearity relationships were observed between 14 pepper genes and 20

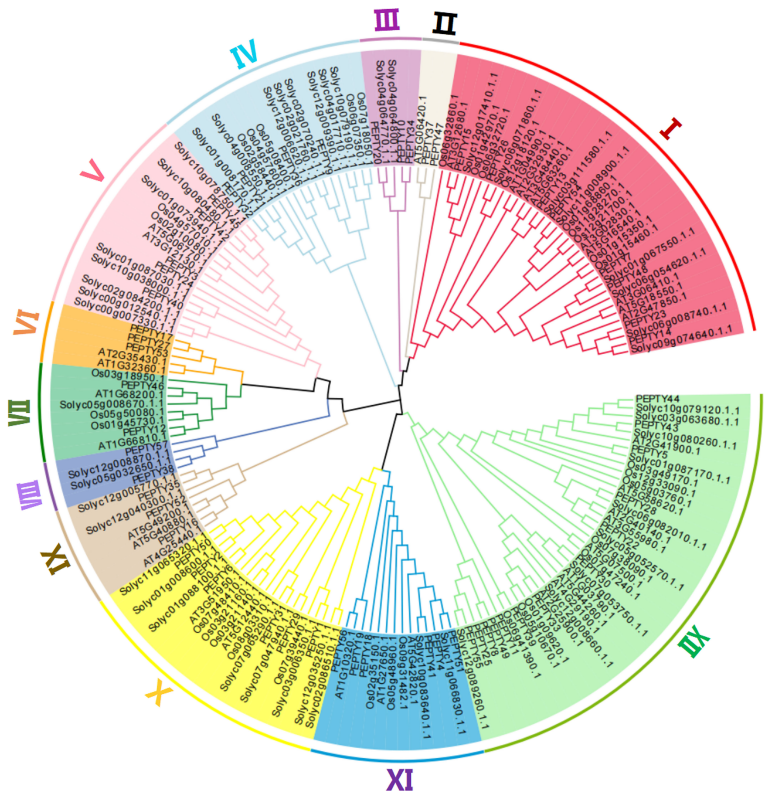


FIGURE 1
Evolutionary tree of CCCH genes in *Arabidopsis thaliana*, *Oryza sativa*, *Solanum lycopersicum*, and *Capsicum annuum*. The different shades of color correspond to different subgroups.

Arabidopsis genes and between 40 pepper genes and 42 tomato genes. A total of 21 pairs of homologous genes were identified between pepper and *Arabidopsis*, and 47 pairs of homologous genes were identified between pepper and tomato (Supplementary Figure 3). The logarithm of homologous genes with tomato was twice that of homologous genes with *Arabidopsis*; and this is likely because the closer phylogenetic relationship between pepper and tomato (both in the family Solanaceae) than between pepper and *Arabidopsis*. To assess the selective constraint pressure of gene pairs, Ka/Ks calculations were performed in TBtools (Supplementary Table 4). Most gene pairs have Ka/Ks ratios below 1, indicating that purification selection may have been undertaken during evolution.

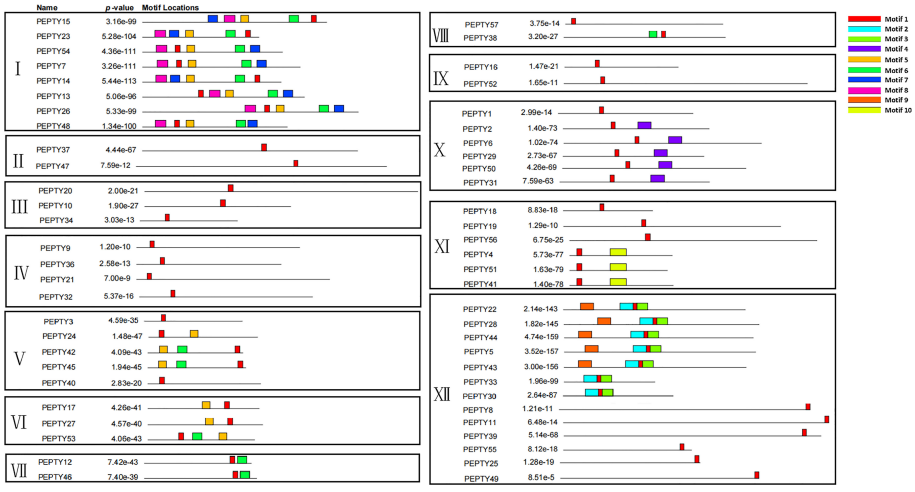


FIGURE 2
Protein motifs of the CCCH gene family in pepper. The colorful boxes delineate different motifs. The clustering was performed according to the results of the phylogenetic analysis.

10.3389/fpls.2023.1189038

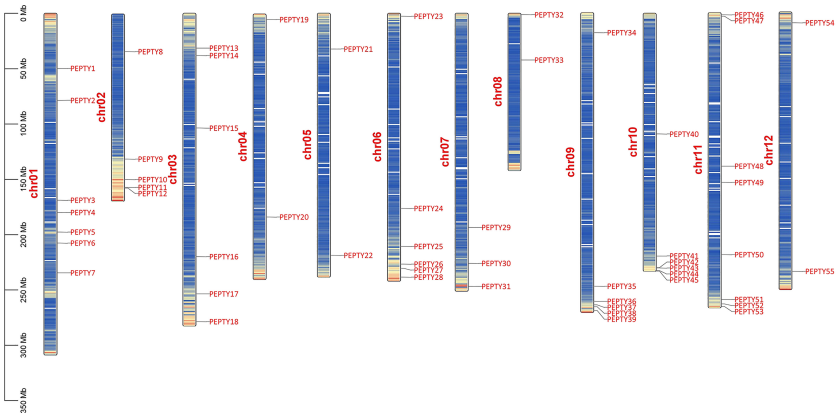


FIGURE 3
Chromosomal distribution of CCCH genes in pepper. Chr01–12 indicate chromosomes 01–12. Bands on the chromosomes indicate gene density.

Expression analysis of *PEPTY* genes in different pepper tissues

We characterized the expression of pepper CCCH genes in five tissues: flower, root, stem, placenta, and pericarp tissue (Figure 5; Supplementary Table 5). *PEPTY24* was expressed at high levels in flowers and at low levels in the roots and stems; *PEPTY12* and *PEPTY46* were expressed at high levels in stems, but their expression gradually decreased in the roots and flowers as development advanced. *PEPTY29* was most highly expressed in the flowers, followed by the roots and stems. In placenta period, the expression of *PEPTY10* gradually increased with developmental stage. The expression of *PEPTY30* was the highest in the initial breaker stage. The expression of *PEPTY35* was up-regulated at the early developmental stage in the placenta and was down-regulated at the breaker stage. In pericarp period, the expression of *PEPTY10* was significantly up-regulated at day 10 of the breaker stage. The expression of *PEPTY2* was high at stage 1 in both the placenta and pericarp period (PL1 and PR1) and decreased thereafter. The expression of CCCH might vary among organs and at different

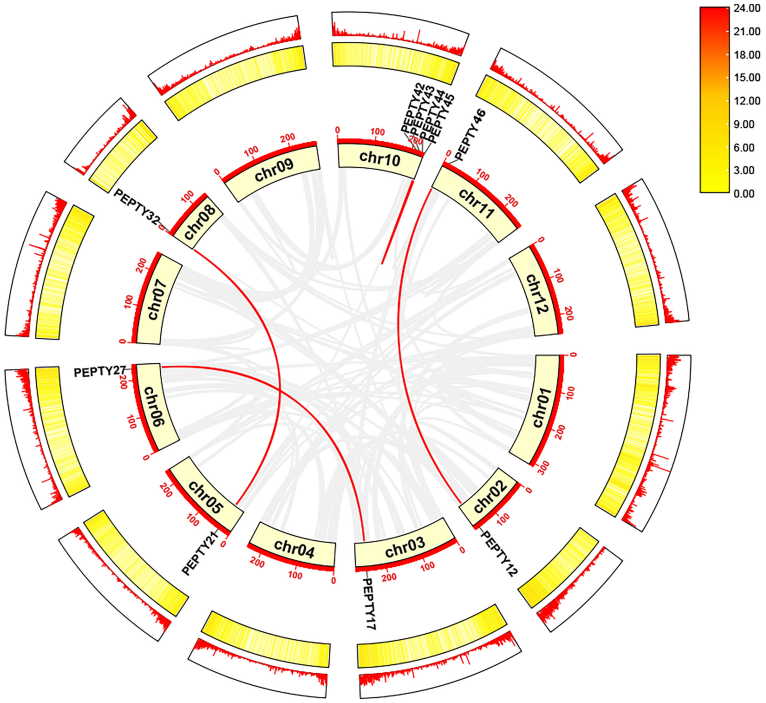


FIGURE 4
Collinearity analysis of the CCCH gene family in pepper. Chromosomes 01–12 are represented by yellow rectangles. The gray lines indicate syntenic blocks in the pepper genome, and the red lines between chromosomes delineate segmentally duplicated gene pairs.The outermost heatmap and lines represent gene density on the chromosomes.

10.3389/fpls.2023.1189038

FIGURE 5
Hierarchical clustering of expression profiles of pepper CCCH genes in different organs. The heatmap was constructed using the 'pheatmap' package in R software, and the fragments per kilobase of exon model per million mapped reads (FPKM) values of the CCCH genes were converted to log(FPKM+1) values. The different tissues included flower, root, stem, placenta (PL), and pericarp (PR). MG denotes mature green, and B denotes breaker. 1, 2, and 3 indicate stage. 5 and 10 indicate days. Red indicates a high relative abundance of transcripts. Green indicates a low relative abundance of transcripts.

growth and developmental stages. Some of these genes such as *PEPTY24* and *PEPTY30* are likely involved in the growth and development of pepper.

Expression analysis of *PEPTY* genes under different stress conditions and phytohormone treatments

Analysis of the relative transcript abundance of *PEPTY* genes under different types of abiotic stress revealed that the expression of many of these genes was significantly up-regulated under cold, heat, drought (D-mannitol) and salt (sodium chloride, NaCl) stress (Figure 6; Supplementary Table 5). The expression of *PEPTY2, PEPTY5, PEPTY7, PEPTY8, PEPTY11, PEPTY16, PEPTY36, PEPTY45*, and *PEPTY57* was up-regulated under cold stress. The expression of *PEPTY4, PEPTY9, PEPTY26, PEPTY32, PEPTY34, PEPTY42, PEPTY51*, and *PEPTY52* was significantly up-regulated at all time points under heat stress. The expression of *PEPTY6, PEPTY31, PEPTY32*, and *PEPTY48* was highest at 12, 6, 24, and 12 h, respectively. By contrast, the expression of *PEPTY14, PEPTY30, PEPTY40*, and *PEPTY46* was up-regulated at 24, 72, 24, and 72 h, respectively, under salt stress. Under drought stress, the expression of *PEPTY5, PEPTY10, PEPTY14, PEPTY23, PEPTY39*, and *PEPTY40* was up-regulated.

The expression of CCCH genes after treatment with two viruses was performed to clarify their responses to biotic stress (Figure 7;

10.3389/fpls.2023.1189038

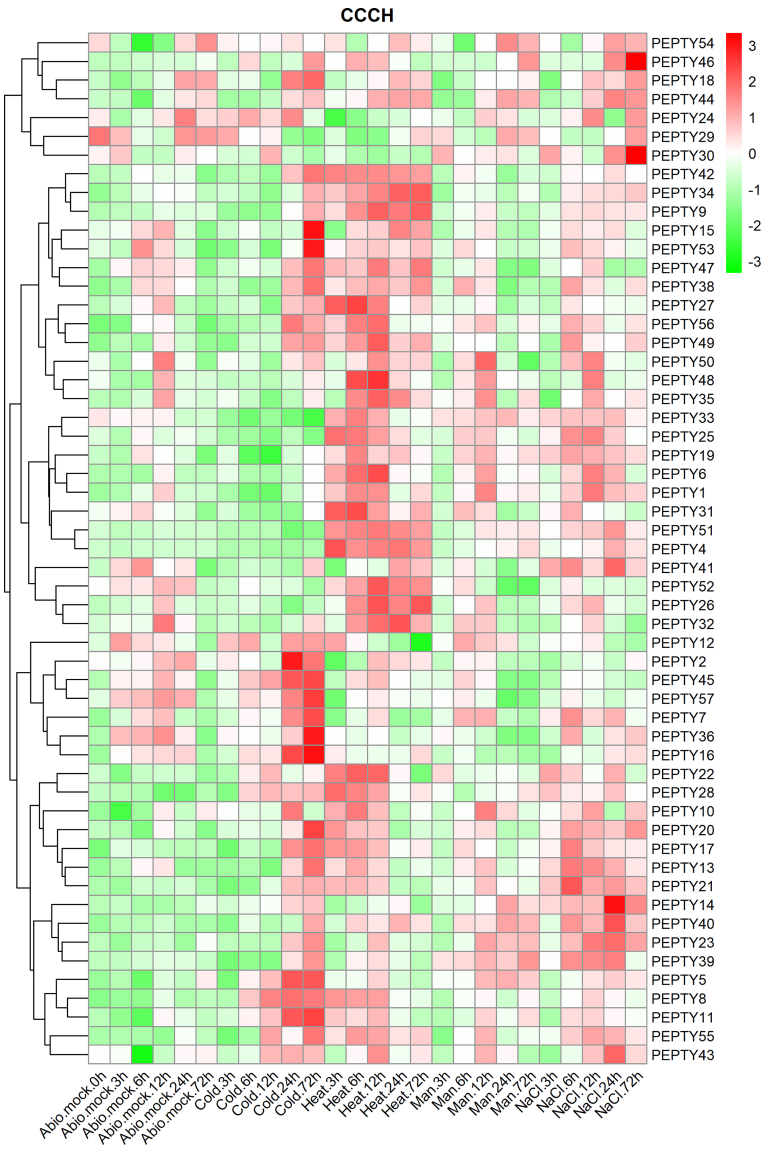


FIGURE 6
Expression profiles of pepper CCCH genes under different types of abiotic stress. Abiotic stresses included cold, heat, drought (D-mannitol), and salt (NaCl). Time points include 1, 3, 6, 12, and 24 h. The control group is indicated by Abio.mock labels. Red indicates a high relative abundance of transcripts. Green indicates a low relative abundance of transcripts.

Supplementary Table 5). The expression of *PEPTY22* following pepper mottle virus (PepMoV) treatment was highest 30 min post-treatment and decreased thereafter. The expression of most genes, such as *PEPTY8*, *PEPTY11*, and *PEPTY54*, was up-regulated 4 h post-treatment. By contrast, the expression of *PEPTY22* was significantly up-regulated 30 min after treatment with tobacco mosaic virus (TMV), which was consistent with its response to PepMoV treatment. The expression of *PEPTY4* and *PEPTY46* was high 4 h after TMV treatment. In addition, the expression of *PEPTY20, PEPTY28, PEPTY30, PEPTY40*, and *PEPTY53* was high 2 days after TMV treatment. The expression of *PEPTY25* and *PEPTY33* was high 3 days after TMV treatment. The responses of most CCCH genes were more pronounced to TMV treatment than to PepMoV treatment.

Ultimately, the expression profiles of CCCH genes were further analyzed under treatment with four phytohormones. The results are shown in Figure 8. The expression of *PEPTY8, PEPTY14, PEPTY22, PEPTY35, PEPTY44, PEPTY55*, and *PEPTY56* was increased after methyl jasmonate (MeJA) treatment. The expression of 13 genes (*PEPTY4, PEPTY13, PEPTY15, PEPTY26, PEPTY27, PEPTY28, PEPTY34, PEPTY35, PEPTY41, PEPTY42, PEPTY43, PEPTY53*, and *PEPTY56*) increased after SA treatment. The expression of *PEPTY35, PEPTY41, PEPTY42, PEPTY43, PEPTY53*, and *PEPTY56* was up-regulated after SA treatment. The expression of *PEPTY37* significantly increased 3 h after ET treatment. This gene was not expressed in the other treatments or the control. However, the expression of *PEPTY9, PEPTY11, PEPTY20, PEPTY21*, and *PEPTY49* was down-regulated. The expression of *PEPTY21* and *PEPTY43* was up-regulated after ABA treatment, especially at 12 h, and the expression of *PEPTY46* was more significantly up-regulated at 24 h. These results suggest that CCCH genes play a role in the response to phytohormones.

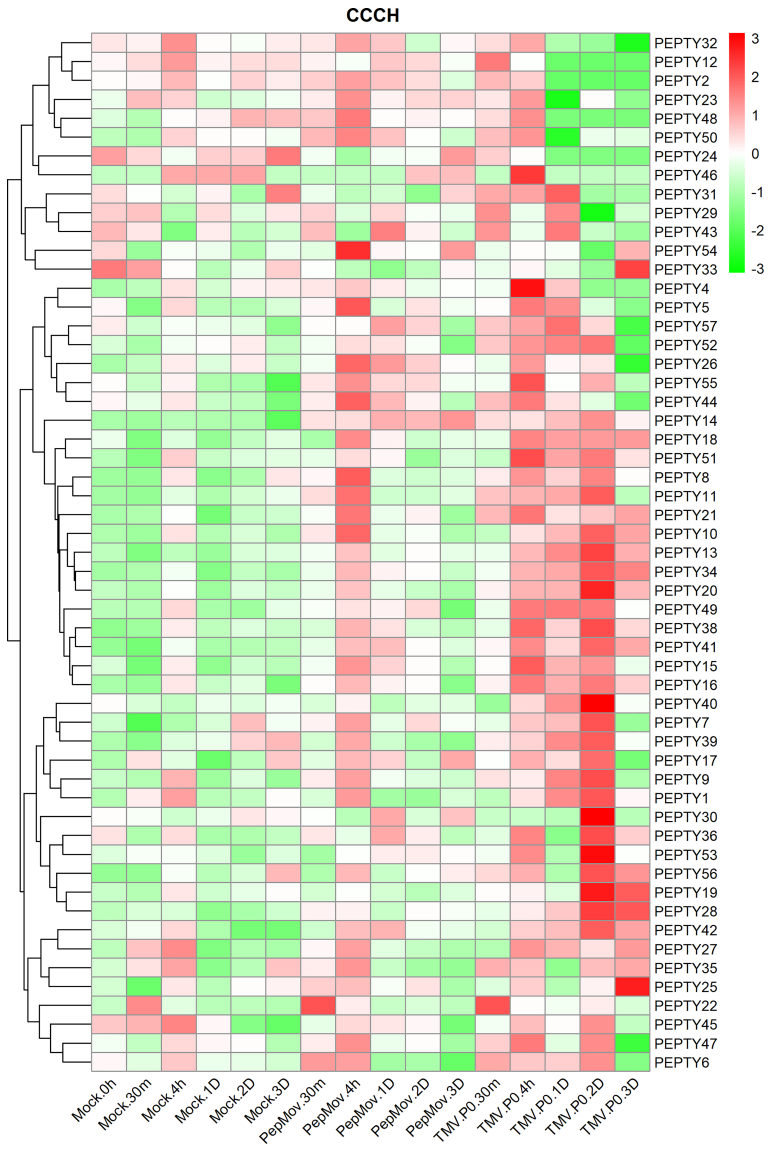


FIGURE 7
Expression profiles of pepper CCCH genes under different types of biotic stress. Biotic stresses included pepper mottle virus (PepMoV) and tobacco mosaic virus (TMV). Time points include 30 min, 4 h, 1 d, 2 d, and 3 d. The control group is indicated by mock labels. Red indicates high relative abundance of transcripts. Green indicates low relative abundance of transcripts.

qRT-PCR validation of the CCCH genes under cold and heat stress

We conducted qRT-PCR analysis on 5 genes that were significantly up-regulated under cold treatment and 7 genes with expression patterns that varied under heat treatment in the heat map (Figure 9). Under cold stress, the expression of four genes (*PEPTY12, PEPTY16, PEPTY36*, and *PEPTY57*) peaked at 72 h, whereas the expression of *PEPTY45* peaked at 24 h. The expression of all these genes did not significantly differ from that of the control under cold treatment in the early stage; however, at 72 h, the expression of genes under cold treatment was at least two-fold higher than that of genes in the control group. A similar pattern was observed for *PEPTY4, PEPTY9, PEPTY26, PEPTY27, PEPTY34, PEPTY51*, and *PEPTY52* under heat treatment, and the significance of differences was even more pronounced. The expression of *EPTY4, PEPTY9, PEPTY27, PEPTY34, PEPTY51*, and *PEPTY52* peaked at 72 h, whereas the expression of *PEPTY26* peaked at 24 h. Differences in the expression of *PEPTY4, PEPTY9*, and *PEPTY51* between the heat treatment and control group gradually increased over time.

Discussion

C. annuum is one of the most widely grown solanaceous vegetables worldwide and capsaicin produced from seed of *C. annuum* is an economically important spice, medicine, vegetable, and biopesticide. However, previous studies have shown that pepper plants are highly sensitive to biotic and abiotic stresses, such as pathogens, drought, cold, and heat (Kim et al., 2014; Kang

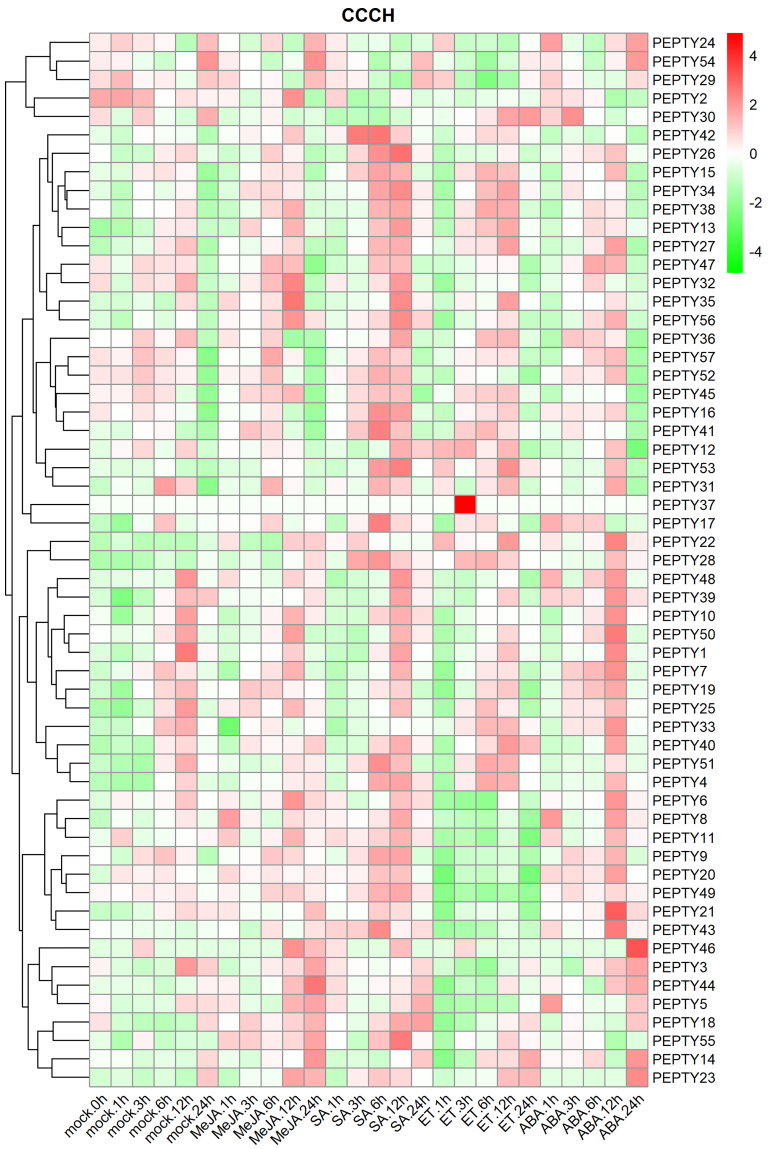


FIGURE 8
Expression profiles of pepper CCCH genes under phytohormone treatments. The phytohormone treatments included methyl jasmonate (MeJA), salicylic acid (SA), ethylene (ET), and abscisic acid (ABA). Time points include 1, 3, 6, 12, and 24 h. The control group is indicated by a mock label. Red indicates a high relative abundance of transcripts. Green indicates a low relative abundance of transcripts.

et al., 2020; Lee et al., 2020). CCCH proteins have been identified in plants. These proteins are rather unusual in that they can regulate the expression of genes by binding to mRNA in addition to DNA (Kim et al., 2014; Qin et al., 2014). Functional analyses of CCCH genes in *Arabidopsis*, rice, maize, poplar, alfalfa (*Medicago truncatula*), citrus, tomato, banana, cabbage, soybean, rose, tobacco, and other plants have been conducted (Wang et al., 2008; Chai et al., 2012; Peng et al., 2012; Zhang et al., 2013; Liu et al., 2014; Xu, 2014; Mazumdar et al., 2017; Pi et al., 2018; Hu and Zuo, 2021; Li et al., 2021; Tang C. et al., 2022).

We identified 57 CCCH zinc finger genes in the genome of *C. annuum* cv. CM334. A total of 80 CCCH genes have been identified in tomato belonging to (Xu, 2014), which is also a member of the family Solanaceae. We searched for CCCH genes in the *C. annuum* L. Zunla-1 genome. However, this species only had 69 CCCH genes (Supplementary Table 6), which was lower than in tomato. The CCCH genes in CM334 could be divided into 12 subfamilies, and Group III and VIII genes were only present in pepper and tomato, but not in *Arabidopsis thaliana* and rice (Figure 1). These subfamilies are likely unique to the Solanaceae family.

Structural analysis of the CCCH genes revealed that the CCCH motifs are highly conserved, motif type and motif position were highly similar within each subfamily, but motif type and motif position varied among most subfamilies. The similarity and specificity within and between subfamilies, respectively, indicated that genes in the same subfamily may have similar functions, and genes in different subclades may perform different functions. No motifs in *PEPTY*35 were in Group IX, and 56.1% of pepper CCCH genes had at least two motifs. The main structures present were C-X_5-C-X_4-C-X_3-H and C-X_{7-8}-C-X_5-C-X_3-H.

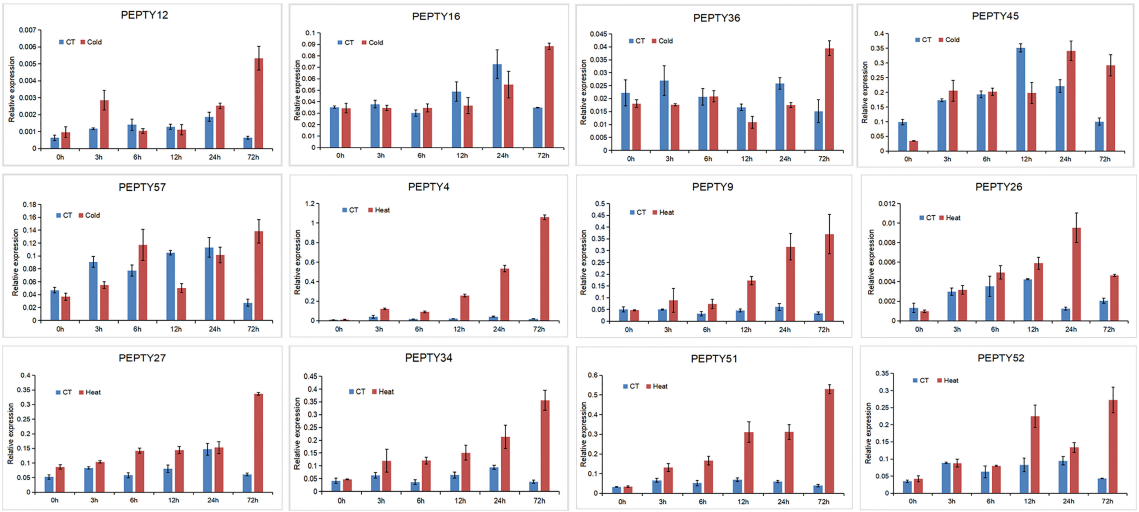


FIGURE 9
qRT-PCR analysis of 12 pepper CCCH genes under different stress treatments. The x-axis shows the time points after stress treatments. The y-axis shows the relative expression levels normalized to the reference gene *UBI3*. Data are mean ± SD of three technical replicates.

Gene duplication is one of the primary drivers of the evolution of genomic and genetic systems. Duplicated genes have the potential to develop new functions. Gene family expansion in the genome generally stems from tandem and segmental duplication events (Moore and Purugganan, 2003; Cannon et al., 2004; Levasseur and Pontarotti, 2011). In Group V, there are five pepper CCCH genes (*PEPTY3, PEPTY24, PEPTY40, PEPTY42*, and *PEPTY45*), but only two *Arabidopsis* CCCH genes (*AtC3H36* and *AtC3H52*) and two rice CCCH genes (*OsC3H14* and *OsC3H31*). Two homologs of *Arabidopsis* or rice were likely generated by segmental duplication, and the pepper CCCH genes likely underwent one round of whole-genome duplication and one tandem duplication.

The expression levels of CCCH genes in pepper varied significantly among tissues and developmental stages (Chai et al., 2012; Li et al., 2021). Only the expression of *PEPTY24, PEPTY29*, and *PEPTY54* was up-regulated in flowers. The expression of *PEPTY24* was specific to flowers, which may be involved in the regulation of flowering in pepper. *PEPTY29* was expressed in flower, root, and stem, but not in placenta and pericarp; this gene might thus be involved in regulating flower, root, and stem development. Twenty-five genes were expressed in the roots, and 27 genes were expressed in the stems. The expression patterns of CCCH genes in pepper differ from those of CCCH genes in *Arabidopsis* and rice, where most CCCH genes are expressed in the roots, inflorescences, leaves, and seeds (Wang et al., 2008).

The expression profiles of CCCH genes under biotic stress, abiotic stress, and phytohormone treatments showed that most *PEPTY* genes were highly expressed under these conditions. Comparison with other studies confirmed that the activity of most CCCH zinc finger proteins can be induced by hormones such as ABA and GA; they may play a role in hormone-mediated signaling pathways (Verma et al., 2016; Han et al., 2021). This pattern of activity is similar to that observed under biotic and abiotic stress; it is even likely that a particular gene could respond to multiple different treatments. For example, in rice, the *OsTZF1* gene responds to GA, MeJA, and salicylate (Jan et al., 2013). In *Arabidopsis*, the expression of *AtOZF1* was highly induced by ABA and salinity treatment (Huang et al., 2011). High expression of *AtTZF2* and *AtTZF3* enhances tolerance to high salt stress, and the silencing of these two genes reduces the tolerance of plants to salt and drought stress (Huang et al., 2011; Huang et al., 2012; Lee et al., 2012). In addition, *AtTZF4, 5*, and *6* are positive regulators of ABA (Bogamuwa and Jang, 2013). These results enhance our understanding of the growth of pepper plants, as well as the response of pepper to various types of stress and hormone treatments.

After identifying CCCH genes in pepper that play significant roles in responses to cold and heat stress, the expression patterns of five candidate genes that were highly induced by cold stress and seven candidate genes that were highly induced by heat stress were validated by qRT-PCR. *PEPTY4* and *PEPTY51*, which were both in Group XI, were not expressed under cold stress and in the control environment, but they were highly expressed under heat stress. However, both *PEPTY16* and *PEPTY52* belonged to Group XI; the former was highly expressed under cold stress, and the latter was highly expressed under heat stress. *PEPTY36* in Group IV was highly expressed under cold treatment at 72 h. *PEPTY9*, which also belongs to the same subfamily as *PEPTY36*, was not significantly expressed under cold stress, but its expression was gradually up-regulated under heat stress. Thus, the expression patterns were not always the same among each subfamily member of each CCCH gene in pepper. One plausible explanation for this observation is that pepper is more sensitive to low-temperature and high-temperature stress. In addition, the responses of different genes to cold and heat might vary (Wang et al., 2019; Wang et al., 2021; Yang et al., 2021; Gao et al., 2022; Tang B. et al., 2022; Zhang et al., 2022). Therefore, further functional studies of these CCCH genes are needed to clarify the pathways underlying their responses to cold stress and heat stress.

Conclusion

In this study, the phylogenetic relationships, structure, conserved motifs, chromosomal localization, duplication events, and expression profiles of CCCH genes were analyzed and 57 CCCH zinc finger genes were identified in pepper. A phylogenetic tree was constructed using CCCH sequences from *Arabidopsis*, tomato, and rice. Based on studies of *Arabidopsis*, we divided the pepper CCCH genes into 12 subfamilies. The exon/intron structure and motif composition were conserved in most subfamily. These genes were unevenly distributed on 12 chromosomes, and segmental duplication events appear to have been the major driver of gene expansion in the CCCH family. We characterized the expression profiles of CCCH genes in different tissues of pepper and under various types of stress and validated these expression patterns using qRT-PCR analysis. We found that CCCH zinc finger genes play important roles in biological processes such as growth and development and adaptation to stress. Overall, our findings will aid future studies aimed at examining the evolution, inheritance, and function of CCCH zinc finger genes in pepper and other plants.

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

JC, ZZ, SZ, TZ, ZL, FG and WT designed the research. WT, YZ, XM, YS and YD performed the research. WT, YH, and ZD analyzed the data. WT, ZZ, ZL, FG, and JC wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

The research was supported by the Project of Central Public-interest Scientific Institution Basal Research Fund, Grant No: 1630032022009, Project of Sanya Yazhou Bay Science and Technology City, Grant No: SCKJ-JYRC-2022-05 and SCKJ-JYRC-2022-25, and Hainan Provincial Natural Science Foundation of China (No. 322MS132).

Conflict of interest

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

The reviewer ZZ declared a shared affiliation with the authors FG, ZL at the time of review.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1189038/full#supplementary-material

References

Arnaud, D., Déjardin, A., Leplé, J. C., Lesage-Descauses, M. C., and Pilate, G. (2007). Genome-wide analysis of LIM gene family in *Populus trichocarpa*, *Arabidopsis thaliana*, and *Oryza sativa*. *DNA Res.* 14, 103–116. doi: 10.1093/dnares/dsm013

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–W208. doi: 10.1093/nar/gkp335

Berg, J. M., and Shi, Y. (1996). The galvanization of biology: a growing appreciation for the roles of zinc. *Science* 271, 1081–1085. doi: 10.1126/science.271.5252.1081

Bogamuwa, S., and Jang, J. C. (2013). The *Arabidopsis* tandem CCCH zinc finger proteins *AtTZF4*, *5* and *6* are involved in light-, abscisic acid- and gibberellic acid-mediated regulation of seed germination. *Plant Cell Environ.* 36, 1507–1519. doi: 10.1111/pce.12084

Cannon, S. B., Mitra, A., Baumgarten, A., Young, N. D., and May, G. (2004). The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biol.* 4, 10. doi: 10.1186/1471-2229-4-10

Chai, G., Hu, R., Zhang, D., Qi, G., Zuo, R., Cao, Y., et al. (2012). Comprehensive analysis of CCCH zinc finger family in poplar (*Populus trichocarpa*). *BMC Genomics* 13, 253. doi: 10.1186/1471-2164-13-253

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: was plotted using TBtools analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp.2020.06.009

Cheng, Y., Kato, N., Wang, W., Li, J., and Chen, X. (2003). Two RNA binding proteins, *HEN4* and *HUA1*, act in the processing of *AGAMOUS* pre-mRNA in *Arabidopsis thaliana*. *Dev. Cell* 4, 53–66. doi: 10.1016/s1534-5807(02)00399-4

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., et al. (2019). The pfam protein families database in 2019. *Nucleic Acids Res.* 47, D427–D432. doi: 10.1093/nar/gky995

Freemont, P. S. (1993). The RING finger. a novel protein sequence motif related to the zinc finger. *Ann. N Y Acad. Sci.* 684, 174–192. doi: 10.1111/j.1749-6632.1993.tb32280.x

Gao, C., Mumtaz, M. A., Zhou, Y., Yang, Z., Shu, H., Zhu, J., et al. (2022). Integrated transcriptomic and metabolomic analyses of cold-tolerant and cold-sensitive pepper species reveal key genes and essential metabolic pathways involved in response to cold stress. *Int. J. Mol. Sci.* 23 (12), 6683. doi: 10.3390/ijms23126683

Gasteiger, E., Hoogland, C., Alexandre, G., Duvaud, S., Wilkins, M. R., Appel, R. D., et al. (2005). "Protein identification and analysis tools on the ExPASy server," In: Walker, J.M. (ed) *The proteomics protocols handbook*. Humana press: Springer protocols Handbooks. 571–607. doi: 10.1385/1-59259-890-0:571

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40, D1178–D1186. doi: 10.1093/nar/gkr944

Grabowska, A., Wisniewska, A., Tagashira, N., Malepszy, S., and Filipecki, M. (2009). Characterization of *CsSEF1* gene encoding putative CCCH-type zinc finger protein

expressed during cucumber somatic embryogenesis. *J. Plant Physiol.* 166, 310–323. doi: 10.1016/j.jplph.2008.06.005

Guo, Y. H., Yu, Y. P., Wang, D., Wu, C. A., Yang, G. D., Huang, J. G., et al. (2009). GhZFP1, a novel CCCH-type zinc finger protein from cotton, enhances salt stress tolerance and fungal disease resistance in transgenic tobacco by interacting with GZIRD21A and GZIPR5. *New Phytol.* 183, 62–75. doi: 10.1111/j.1469-8137.2009.02838.x

Hall, T. M. (2005). Multiple modes of RNA recognition by zinc finger proteins. *Curr. Opin. Struct. Biol.* 15, 367–373. doi: 10.1016/j.sbi.2005.04.004

Han, G., Qiao, Z., Li, Y., Wang, C., and Wang, B. (2021). The roles of CCCH zinc-finger proteins in plant abiotic stress tolerance. *Int. J. Mol. Sci.* 22 (15), 8327. doi: 10.3390/ijms22158327

Hu, X., and Zuo, J. (2021). The CCCH zinc finger family of soybean (Glycine max l.): genome-wide identification, expression, domestication, GWAS and haplotype analysis. *BMC Genomics* 22, 511. doi: 10.1186/s12864-021-07787-9

Huang, P., Chung, M. S., Ju, H. W., Na, H. S., Lee, D. J., Cheong, H. S., et al. (2011). Physiological characterization of the *Arabidopsis thaliana* oxidation-related zinc finger 1, a plasma membrane protein involved in oxidative stress. *J. Plant Res.* 124, 699–705. doi: 10.1007/s10265-010-0397-3

Huang, P., Ju, H. W., Min, J. H., Zhang, X., Chung, J. S., Cheong, H. S., et al. (2012). Molecular and physiological characterization of the *Arabidopsis thaliana* oxidation-related zinc finger 2, a plasma membrane protein involved in ABA and salt stress response through the ABI2-mediated signaling pathway. *Plant Cell Physiol.* 53, 193–203. doi: 10.1093/pcp/pcr162

Iuchi, S., and Kuldell, N. (2005). "Zinc finger proteins: from atomic contact to cellular function,", vol. 276. (New York: Kluwer Academic/Plenum Publishers). doi: 10.1007/b139055

Jan, A., Maruyama, K., Todaka, D., Kidokoro, S., Abo, M., Yoshimura, E., et al. (2013). *OsTZF1*, a CCCH-tandem zinc finger protein, confers delayed senescence and stress tolerance in rice by regulating stress-related genes. *Plant Physiol.* 161, 1202–1216. doi: 10.1104/pp.112.205385

Kang, W. H., Sim, Y. M., Koo, N., Nam, J. Y., Lee, J., Kim, N., et al. (2020). Transcriptome profiling of abiotic responses to heat, cold, salt, and osmotic stress of *Capsicum annuum* l. *Sci. Data* 7, 17. doi: 10.1038/s41597-020-0352-7

Kim, S., Park, M., Yeom, S. I., Kim, Y. M., Lee, J. M., Lee, H. A., et al. (2014). Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. *Nat. Genet.* 46, 270–278. doi: 10.1038/ng.2877

Kosarev, P., Mayer, K. F., and Hardtke, C. S. (2002). Evaluation and classification of RING-finger domains encoded by the *Arabidopsis* genome. *Genome Biol.* 3, RESEARCH0016. doi: 10.1186/gb-2002-3-4-research0016

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096

Lee, S. J., Jung, H. J., Kang, H., and Kim, S. Y. (2012). *Arabidopsis* zinc finger proteins *AtC3H49*/*AtTZF3* and *AtC3H20*/*AtTZF2* are involved in ABA and JA responses. *Plant Cell Physiol.* 53, 673–686. doi: 10.1093/pcp/pcs023

Lee, J., Nam, J. Y., Jang, H., Kim, N., Kim, Y. M., Kang, W. H., et al. (2020). Comprehensive transcriptome resource for response to phytohormone-induced signaling in capsicum annuum l. *BMC Res. Notes* 13, 440. doi: 10.1186/s13104-020-05281-1

Levasseur, A., and Pontarotti, P. (2011). The role of duplications in the evolution of genomes highlights the need for evolutionary-based approaches in comparative genomics. *Biol. Direct* 6, 11. doi: 10.1186/1745-6150-6-11

Li, C.-H., Fang, Q.-X., Zhang, W.-J., Li, Y.-H., Zhang, J.-Z., Chen, S., et al. (2021). Genome-wide identification of the CCCH gene family in rose (*Rosa chinensis* jacq.) reveals its potential functions. *Biotechnol. Biotechnol. Equip.* 35, 517–526. doi: 10.1080/13102818.2021.1901609

Li, J., Jia, D., and Chen, X. (2001). *HUA1*, a regulator of stamen and carpel identities in *Arabidopsis*, codes for a nuclear RNA binding protein. *Plant Cell* 13, 2269–2281. doi: 10.1105/tpc.010201

Li, Q. T., Lu, X., Song, Q. X., Chen, H. W., Wei, W., Tao, J. J., et al. (2017). Selection for a zinc-finger protein contributes to seed oil increase during soybean domestication. *Plant Physiol.* 173, 2208–2224. doi: 10.1104/pp.16.01610

Lijavetzky, D., Carbonero, P., and Vicente-Carbajosa, J. (2003). Genome-wide comparative phylogenetic analysis of the rice and *Arabidopsis* dof gene families. *BMC Evol. Biol.* 3, 17. doi: 10.1186/1471-2148-3-17

Lin, P. C., Pomeranz, M. C., Jikumaru, Y., Kang, S. G., Hah, C., Fujioka, S., et al. (2011). The *Arabidopsis* tandem zinc finger protein *AtTZF1* affects ABA- and GA-mediated growth, stress and gene expression responses. *Plant J.* 65, 253–268. doi: 10.1111/j.1365-313X.2010.04419.x

Liu, S., Khan, M. R., Li, Y., Zhang, J., and Hu, C. (2014). Comprehensive analysis of CCCH-type zinc finger gene family in citrus (Clementine mandarin) by genome-wide characterization. *Mol. Genet. Genomics* 289, 855–872. doi: 10.1007/s00438-014-0858-9

Mazumdar, P., Lau, S. E., Wee, W. Y., Singh, P., and Harikrishna, J. A. (2017). Genome-wide analysis of the CCCH zinc-finger gene family in banana (*Musa acuminata*): an insight into motif and gene structure arrangement, evolution and salt stress responses. *Trop. Plant Biol.* 10, 177–193. doi: 10.1007/s12042-017-9196-5

Moore, R. C., and Purugganan, M. D. (2003). The early stages of duplicate gene evolution. *Proc. Natl. Acad. Sci. U.S.A.* 100, 15682–15687. doi: 10.1073/pnas.2535513100

Moore, M., and Ullman, C. (2003). Recent developments in the engineering of zinc finger proteins. *Brief Funct. Genomic Proteomic* 1, 342–355. doi: 10.1093/bfgp/1.4.342

Nakano, T., Suzuki, K., Fujimura, T., and Shinshi, H. (2006). Genome-wide analysis of the ERF gene family in *Arabidopsis* and rice. *Plant Physiol.* 140, 411–432. doi: 10.1104/pp.105.073783

Peng, X., Zhao, Y., Cao, J., Zhang, W., Jiang, H., Li, X., et al. (2012). CCCH-type zinc finger family in maize: genome-wide identification, classification and expression profiling under abscisic acid and drought treatments. *PloS One* 7, e40120. doi: 10.1371/journal.pone.0040120

Pi, B., He, X., Ruan, Y., Jang, J. C., and Huang, Y. (2018). Genome-wide analysis and stress-responsive expression of CCCH zinc finger family genes in *Brassica rapa*. *BMC Plant Biol.* 18, 373. doi: 10.1186/s12870-018-1608-7

Qin, C., Yu, C., Shen, Y., Fang, X., Chen, L., Min, J., et al. (2014). Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. *Proc. Natl. Acad. Sci. U.S.A.* 111, 5135–5140. doi: 10.1073/pnas.1400975111

Qiu, A., Lei, Y., Yang, S., Wu, J., Li, J., Bao, B., et al. (2018). *CaC3H14* encoding a tandem CCCH zinc finger protein is directly targeted by *CaWRKY40* and positively regulates the response of pepper to inoculation by *Ralstonia solanacearum*. *Mol. Plant Pathol.* 19, 2221–2235. doi: 10.1111/mpp.12694

Schultz, J., Copley, R. R., Doerks, T., Ponting, C. P., and Bork, P. (2000). SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* 28, 231–234. doi: 10.1093/nar/28.1.231

Schumann, U., Prestele, J., O'geen, H., Brueggeman, R., Wanner, G., and Gietl, C. (2007). Requirement of the C3HC4 zinc RING finger of the *Arabidopsis PEX10* for photorespiration and leaf peroxisome contact with chloroplasts. *Proc. Natl. Acad. Sci. U.S.A.* 104, 1069–1074. doi: 10.1073/pnas.0610402104

Takatsuji, H. (1998). Zinc-finger transcription factors in plants. *Cell Mol. Life Sci.* 54, 582–596. doi: 10.1007/s000180050186

Tang, C., Deng, Z., Liu, Q., Jin, W., Xiang, S., Xie, P., et al. (2022). Genome-wide identification and expression analysis of CCCH type zinc-finger gene family in *Nicotiana tabacum*. *J. Henan Agric. Sci.* 51, 48–58. doi: 10.15933/j.cnki.1004-3268.2022.04.006

Tang, B., Li, X., Zhang, X., Yin, Q., Xie, L., Zou, X., et al. (2022). Transcriptome data reveal gene clusters and key genes in pepper response to heat shock. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.946475

Tazuke, A., and Asayama, M. (2013). Expression of *CsSEF1* gene encoding putative CCCH zinc finger protein is induced by defoliation and prolonged darkness in cucumber fruit. *Planta* 237, 681–691. doi: 10.1007/s00425-012-1787-7

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25, 4876–4882. doi: 10.1093/nar/25.24.4876

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. (2006). The genome of black cottonwood, populus trichocarpa (Torr. & Gray). *Science* 313, 1596–1604. doi: 10.1126/science.1128691

Verma, V., Ravindran, P., and Kumar, P. P. (2016). Plant hormone-mediated regulation of stress responses. *BMC Plant Biol.* 16, 86. doi: 10.1186/s12870-016-0771-y

Wang, D., Guo, Y., Wu, C., Yang, G., Li, Y., and Zheng, C. (2008). Genome-wide analysis of CCCH zinc finger family in *Arabidopsis* and rice. *BMC Genomics* 9, 44. doi: 10.1186/1471-2164-9-44

Wang, J., Liang, C., Yang, S., Song, J., Li, X., Dai, X., et al. (2021). iTRAQ-based quantitative proteomic analysis of heat stress-induced mechanisms in pepper seedlings. *PeerJ* 9, e11509. doi: 10.7717/peerj.11509

Wang, J., Lv, J., Liu, Z., Liu, Y., Song, J., Ma, Y., et al. (2019). Integration of transcriptomics and metabolomics for pepper (*Capsicum annuum* l.) in response to heat stress. *Int. J. Mol. Sci.* 20 (20), 5042. doi: 10.3390/ijms20205042

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40, e49. doi: 10.1093/nar/gkr1293

Xu, R. (2014). Genome-wide analysis and identification of stress-responsive genes of the CCCH zinc finger family in *Solanum lycopersicum*. *Mol. Genet. Genomics* 289, 965–979. doi: 10.1007/s00438-014-0861-1

Yang, Y., Guang, Y., Wang, F., Chen, Y., Yang, W., Xiao, X., et al. (2021). Characterization of phytochrome-interacting factor genes in pepper and functional analysis of *CaPIF8* in cold and salt stress. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.746517

Zhang, J., Liang, L., Xie, Y., Zhao, Z., Su, L., Tang, Y., et al. (2022). Transcriptome and metabolome analyses reveal molecular responses of two pepper (*Capsicum annuum* l.) cultivars to cold stress. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.819630

Zhang, Y., and Wang, L. (2005). The WRKY transcription factor superfamily: its origin in eukaryotes and expansion in plants. *BMC Evol. Biol.* 5, 1. doi: 10.1186/1471-2148-5-1

Zhang, C., Zhang, H., Zhao, Y., Jiang, H., Zhu, S., Cheng, B., et al. (2013). Genome-wide analysis of the CCCH zinc finger gene family in *Medicago truncatula*. *Plant Cell Rep.* 32, 1543–1555. doi: 10.1007/s00299-013-1466-6

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 30 June 2023
DOI 10.3389/fpls.2023.1201730

OPEN ACCESS

EDITED BY
Yan Zhao,
Shandong Agricultural University, China

REVIEWED BY
Wenjiao Zhu,
Nanjing Agricultural University, China
Gennady L. Burygin,
Institute of Biochemistry and Physiology of Plants and Microorganisms (RAS), Russia

*CORRESPONDENCE
Gang Gao
ggsxnu@126.com
Weizhong Liu
liuwzh@sxnu.edu.cn

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

RECEIVED 07 April 2023
ACCEPTED 14 June 2023
PUBLISHED 30 June 2023

CITATION
Liang L, Guo L, Zhai Y, Hou Z, Wu W, Zhang X, Wu Y, Liu X, Guo S, Gao G and Liu W (2023) Genome-wide characterization of *SOS1* gene family in potato (*Solanum tuberosum*) and expression analyses under salt and hormone stress.
Front. Plant Sci. 14:1201730.
doi: 10.3389/fpls.2023.1201730

Genome-wide characterization of *SOS1* gene family in potato (*Solanum tuberosum*) and expression analyses under salt and hormone stress

Liqin Liang†, Liuyan Guo†, Yifan Zhai, Zhiling Hou, Wenjing Wu, Xinyue Zhang, Yue Wu, Xiaona Liu, Shan Guo, Gang Gao* and Weizhong Liu*

College of Life Science, Shanxi Normal University, Taiyuan, China

Salt Overly Sensitive 1 (SOS1) is one of the members of the Salt Overly Sensitive (SOS) signaling pathway and plays critical salt tolerance determinant in plants, while the characterization of the *SOS1* family in potato (*Solanum tuberosum*) is lacking. In this study, 37 *StSOS1s* were identified and found to be unevenly distributed across 10 chromosomes, with most of them located on the plasma membrane. Promoter analysis revealed that the majority of these *StSOS1* genes contain abundant *cis*-elements involved in various abiotic stress responses. Tissue specific expression showed that 21 of the 37 *StSOS1s* were widely expressed in various tissues or organs of the potato. Molecular interaction network analysis suggests that 25 StSOS1s may interact with other proteins involved in potassium ion transmembrane transport, response to salt stress, and cellular processes. In addition, collinearity analysis showed that 17, 8, 1 and 5 of orthologous *StSOS1* genes were paired with those in tomato, pepper, tobacco, and Arabidopsis, respectively. Furthermore, RT-qPCR results revealed that the expression of *StSOS1s* were significant modulated by various abiotic stresses, in particular salt and abscisic acid stress. Furthermore, subcellular localization in *Nicotiana benthamiana* suggested that StSOS1-13 was located on the plasma membrane. These results extend the comprehensive overview of the *StSOS1* gene family and set the stage for further analysis of the function of genes in SOS and hormone signaling pathways.

KEYWORDS

***Solanum tuberosum* L., *SOS1*, expression profiles, abiotic stress, genome-wide**

1 Introduction

High soil salinity is a major abiotic stress that significantly affects plant growth and ultimately reduces plant productivity by preventing the absorption of water and nutrients (Brindha et al., 2021; You et al., 2022). The Salt Overly Sensitive (SOS) signaling pathway plays an essential role in the response of plants to salt stress. It consists of three

components: *SOS1*, *SOS2*, and *SOS3* (Cheng et al., 2019). *SOS1* is a Na^+/H^+ antitransporter that governs the efflux of Na^+ into the root and loading into the xylem vessel for long-distance transport out of the root (Świeżawska et al., 2018). *SOS2* exists as a form of protein kinase in the SOS signaling pathway, which in turn activates *SOS1* to bring about sodium ion homeostasis and salt tolerance (Ali et al., 2021). *SOS3*, which encodes an EF-handed Ca^{2+} binding protein, can sense calcium signals elicited by salt stress, interact with SOS2, and activate SOS2 (Zhu et al., 2021).

SOS1 genes were firstly identified in Arabidopsis (Keisham et al., 2018) and designated as *AtNHX1-AtNHX8*. *AtNHX7* (or *AtSOS1*) is a critical player in the SOS signaling pathway (Zhao C. et al., 2021). *AtSOS1* locates in the plasma membrane (Shi et al., 2000). *AtSOS1* is primarily expressed in epidermal cells at the root tip and in the parenchyma at the xylem-symplast boundary of root, stem, and leaf, hinting at the role of this transporter in the extrusion of Na^+ into the growing medium and in controlling long-distance Na^+ transport in plants (Gao et al., 2016). *SOS1* behaves as a homodimer, with each monomer having 12 transmembrane domains at its N-terminal region and a long C-terminal region containing a cytosolic domain, a cyclic nucleotide binding domain, and an auto-inhibitory domain (Wu et al., 1996; Núñez-Ramírez et al., 2012). SOS proteins were involved in the regulation of plant tolerance to salinity (Zhu et al., 1998). Overexpression of *SOS1* led to reduction of Na^+ accumulation in the xylem and shoot (Shi et al., 2003).

In addition to Arabidopsis, the physiological roles of the associated *SOS1* genes have been investigated in cash crop plants, such as soybean, maize, tomato, cotton (Chen et al., 2017; Wang Z. et al., 2021; Zhang M. et al., 2022; Zhou et al., 2022), and so on. In soybeans, significant accumulation of Na^+ in the roots of *GmSOS1* mutants resulted in an imbalance of Na^+ and K^+, suggesting that *GmSOS1* played a critical role in soybean salt tolerance by maintaining Na^+ homeostasis (Zhang et al., 2022). In maize, SOS pathway has a conserved salt tolerant effect, and its components (*ZmSOS1* and *ZmCBL8*) have Na^+ regulation and natural variations of salt tolerance, providing an important gene target for breeding salt-tolerant maize (Zhou et al., 2022). However, its role has not yet been investigated in potato (*Solanum tuberosum*).

Potato is an important crop in human food systems around the world (Dahal et al., 2019; Ceci et al., 2022) and their cultivation and production are often severely threatened by the various environmental stresses such as salinity and pathogens (Li et al., 2021; Yang et al., 2022). Identification and characterization of resistance genes to salt stress would therefore be helpful in improving potato production. Since the role of *SOS1* in controlling ion homeostasis has been shown in several plants, this gene family is thought to also be valuable in the salt tolerance mechanism and quality improvement of potato. However, limited efforts have been made to identify gene families in the potato, and their expression patterns and regulatory mechanisms remain unclear.

In this study, we identified and analyzed the *SOS1* gene family in potato. Extensive analysis including chromosomal localization, gene structure, and upstream promoter *cis*-acting elements of these gene family were conducted. The physicochemical properties, motifs, gene ontologies, and phylogenetic relationships between the encoded proteins were predicted using bioinformatics tools. Furthermore, the expression profiles of specific *StSOS1s* at salt stress were examined using RT-qPCR. In addition, their expression profiles in response to the exogenous phytohormone abscisic acid (ABA), methyl jasmonate (MeJA), gibberellin (GA) and salicylic acid (SA) were also investigated. The results indicate a diverse pattern of responses to abiotic stress *via* SOS and hormone signaling pathways. It may be beneficial to elucidate the resistance of the potato to abiotic stress, providing some theoretical basis for molecular breeding.

2 Materials and methods

2.1 Plant material and treatments

The potato (diploid cultivar *Solanum phureja*, DM1-3 516 R44) plants used in this study were obtained from Institute of Vegetable and Flowers, Chinese Academy of Agricultural Sciences (CAAS). The potato was grown in a growth chamber at 26 °C/18 °C (day/night) with a 16:8 light: dark cycle and 60-70% relative humidity according to (Ali et al., 2014). The roots of 7-8-leaves-old plantlets were watered with 200 mM NaCl solution (Ma et al., 2021). And the leaves were sprayed with 100 μM ABA, 50 μM MeJA, 350 μM GA and 50 μM SA, respectively. When spraying, moisten the positive and negative sides of all leaves with condensed water droplets without dropping. After the spraying, the plants were immediately wrapped in black plastic bags and treated only once (Yu et al., 2021). Then, the 1, 2, 3, 4 and 5 d (0 d as control) treated plant leaves were respectively quickly frozen in liquid nitrogen at -80 °C for later use (Li et al., 2021). And each treatment was repeated three times.

2.2 *SOS1* genes identification in the potato

All protein sequences were obtained from potato genome data (SolTub_3.0)[1]. First, the HMM profile for the SOS1s domain (PF00999) was downloaded from the Pfam server[2]. Then, the HMMER program[3] was used to identify the SOS1 proteins in the potato genome (Liang et al., 2017). Finally, the SOS1 (Na^+/H^+ exchanger, NHX) domain of all putative SOS1 proteins were determined through CDD[4] and SMART databases[5]. A total of 37 putative *SOS1* genes were identified.

1 http://plants.ensembl.org/index.html

2 http://pfam.xfam.org

3 http://hmmer.janelia.org/

4 http://www.ncbi.nlm.nih.gov/cdd

5 http://smart.emblheidelberg.de/

2.3 Biophysical properties and chromosomal location analysis

Biophysical characteristics of SOS1 proteins were analyzed through ExPASy webserver[6] (Wang T. et al., 2021) and NetPhos 3.1[7] (Naureen et al., 2023). The online prediction tool UniProt[8] (Ilzhöfer et al., 2022) was applied to predict the tertiary structures of potato SOS1s. Subcellular location of protein was predicted using the Cell-PLoc 2.0 prediction tool[9]. The physical positions of the *StSOS1s* along each chromosome were identified from the potato genome database and the distribution of *StSOS1s* was plotted (Xiang et al., 2016).

2.4 *StSOS1s cis*-acting element analysis

The 2000 bp upstream region of the ATG start codon was submitted to PlantCARE[10] (Koul et al., 2019) to identify the *cis*-acting elements and calculate the number of each element. These promoter sequences were represented as word clouds with the help of the WordArt tool[11] (Sharma et al., 2021).

2.5 Conserved motifs and gene structure analysis

The conserved motifs in StSOS1s were identified to use the MEME website[12] (Multiple Em for Motif Elicitation) (Zhang et al., 2021) with the maximum number of motifs was set to 10. Figures of phylogenetic tree along with gene conserved motifs and CDS/UTR structure of StSOS1s were drawn with TBtools (v1.098) (Chen et al., 2020) software. Gene Structure Display Server (GSDS)[13] (Sun et al., 2022) and MEME webserver were employed for gene structure analysis.

2.6 *StSOS1s* tissue-specific expressions and GO enrichment

RNA-Seq data (fragments per kilobase of exon per million mapped, FPKM) (NCBI accession number ERP000527) in potato DM genotype (Wang J. et al., 2021) was used to analyze the expression level of *StSOS1* genes. PlantRegMap[14] (Li H. et al., 2020) was used to functionally re-annotate the proteome of up or down-regulated genes and to plot gene ontology (GO) annotations. Protein-protein interaction (PPI) enrichment was computed by STRING[15] (Fayez et al., 2022) tool, in which Cytoscape software was used for reconstructing the PPI network, modules and to detect the relationship between overall targeted genes.

2.7 Evolutionary tree construction and collinearity analysis

The SOS1 protein sequences of Arabidopsis, tomato, pepper and tobacco were downloaded from the EnsemblPlants (Contreras-Moreira et al., 2022). Homologous sequences were fed into the MEGA7 software and the Clustalw program was used to perform multi-sequence alignment. The results of the output multi-sequence alignment were used to construct an evolutionary tree using the proximity method (He et al., 2022). The collinearity of the sequences of potato with other four species was extracted using TBtools (Zhang C. et al., 2022).

2.8 RNA isolation and RT-qPCR analysis

The leaves samples were ground into powder in liquid nitrogen, total RNA was extracted using *TransZol* Up Plus RNA kit (Trans, Beijing, China), following the manufacturer's instruction. Then the extracted RNA was employed as a template with *TransScript*® One-Step gDNA Removal and cDNA Synthesis SuperMix for qPCR (Trans, Beijing, China) for the first strand cDNA synthesis. All primer sequences used in this study were designed by Primer Blast website[16] of NCBI (Table S1). The RT-qPCR was performed on a QuantStudio-3 system (Thermo Fisher Scientific, Shanghai, China). The reaction system was 20 μL (cDNA 1 μL, *SOS1*-F 0.4μL, *SOS1*-R 0.4μL, SuperMix 10 μL, DyeII 0.4μL, Water 7.8 μL). The reaction system was 94 °C 30 s, (94 °C 5 s, 60 °C 30 s) ×40 Cycles. Three replications were performed and the expression values were calculated by using the $2^{-\Delta\Delta CT}$ method (Mo et al., 2022).

2.9 Subcellular localization of StSOS1-13

For the localization and expression of StSOS1-13 in potato, the CDS without the stop codon was cloned into pCAMBIA1300. Firstly, the complete coding region of *StSOS1-13* (1 734bp) was amplified from the cDNA by PCR using a pair of primers with a homologous arm and inserted into the pCAMBIA1300 vector linearized by the restriction enzyme *Nco*I. Then, the obtained pStSOS1-13-GFP fusion plasmid was converted into *Escherichia*

6 http://www.expasy.org/

7 https://services.healthtech.dtu.dk/services/NetPhos-3.1/

8 https://www.uniprot.org/

9 http://www.csbio.sjtu.edu.cn/bioinf/Cell-PLoc-2/

10 http://bioinformatics.psb.ugent.be/webtools/plantcare/html/

11 https://wordart.com

12 http://memesuite.org/

13 http://gsds.cbi.pku.edu.cn/

14 http://plantregmap.gao-lab.org/go.php

15 https://cn.string-db.org/

16 https://www.ncbi.nlm.nih.gov/tools/primer-blast/

coli DH5α for verified by bacterial liquid PCR and company sequencing (Sangon, Shanghai, China), further inserted into individual *Agrobacterium tumefaciens* strain GV3101 cells and a single colony was selected for PCR positive identification. Finally, the expression vectors were injected into tobacco leaves for the transient expression experiments (Luo et al., 2022). GFP expression was analyzed using scanning confocal laser microscopy.

3 Results

3.1 Identification of *SOS1* genes in the potato

To identify the *SOS1s* family members in potato, the similar protein sequences were searched in the HMMER program with the query sequence SOS1s motif (PF00999). The SMART tool was then used to confirm whether the candidates contained the Na^+/H^+ exchanger (NHX) domain. In total, 37 *SOS1* genes were retrieved from the potato genome and renamed *StSOS1-1* to *StSOS1-37* based on their relative linear order on each chromosome, following the widely used nomenclature (Figure 1). Meantime, we found four pairs of tandem duplicated genes existed in 37 *StSOS1* genes. The analysis showed that there was one pair of tandem duplicated genes (*StSOS1-2* and *StSOS1-3*) on Chr1, one pair (*StSOS1-7* and *StSOS1-8*) on Chr2, one pair (*StSOS1-26* and *StSOS1-27*) on Chr6, and one pair (*StSOS1-30* and *StSOS1-31*) on Chr9.

We further determined the biophysical properties of the potato *SOS1* genes including the locus ID, protein length (aa), predicted protein molecular weight (MW), isoelectric points (pI), and NHX domain. The statistical results showed that the protein length ranged from 209 (*StSOS1-15*) to 1153 (*StSOS1-1*) amino acids, the average amino acids length and molecular weights ranged from 22.51 KDa (*StSOS1-15*) to 127.86 KDa (*StSOS1-1*). PI varying from 4.96 (*StSOS1-20*) to 10.12 (*StSOS1-3*). The subcellular localization of these *StSOS1s* predicted through Cell-PLoc 2.0 tool revealed that most of the StSOS1 proteins were localized in the plasma membrane (Table 1). The results of the NetPhos 3.1 server revealed that StSOS1 proteins were phosphorylated, and phosphorylated residues were Serine (Ser), threonine (Thr) and tyrosine (Tyr) (Table S2), among which serine prediction sites ranged from 11 (StSOS1-24) to 72 (StSOS1-1). The threonine prediction sites ranged from 5 (StSOS1-15) to 32 (StSOS1-1), and the tyrosine prediction sites ranged from 0 (StSOS1-15 and StSOS1-20) to 9 (StSOS1-21). Three-dimensional protein models were constructed by sequence similarity search using UniProt PDB database and the homology modeling was predicted by DS Visualizer (Figure S1). The structures of *StSOS1-7*, *StSOS1-8*, *StSOS1-9*, *StSOS1-16*, *StSOS1-17*, *StSOS1-21*, *StSOS1-22*, *StSOS1-25*, *StSOS1-28*, *StSOS1-29*, *StSOS1-31*, *StSOS1-32*, *StSOS1-36*, and *StSOS1-37* are similar and suggest shared functionality, as do *StSOS1-18* and *StSOS1-19*. These provide an initial basis for understanding the molecular function of the StSOS1 proteins.

3.2 Prediction of *cis*-elements in the promoter sequences of *StSOS1* genes

To clarify which hormonal, environmental stress, or developmental-related signal elements are involved in these *StSOS1s*, we performed a promoter analysis using the PlantCARE server. A large number of basic components were discovered in the upstream sequence (2000 bp) regions, including WRE3, GATA-motif, CAT-box and G-Box, but also P-box, TCA-element, AuxRR-core, TGACG-motif, ABRE and ERE hormonal response-related elements; as-1, LTR, ARE, GC-motif, MBS environmental stress-related components and A-box development-related elements (Figures 2A, B). Hormonal response elements were detected in the promoters of 37 potato *StSOS1* genes, including 15 SA, 19 MeJA, 26 ABA and 30 auxin response. The *cis*-elements involved in the GA response are present in all promoters of *StSOS1s*. The promoters of 10, 16, and 20 *StSOS1* genes contained MYB binding sites involved in low-temperature response, defense and stress response *cis*-elements and drought-inducibility, respectively (Figure 2C). These results suggest that the *StSOS1* genes may play a critical role not only in phytohormones, but also in biological and abiotic responses in the potato.

3.3 Gene structure and conserved motifs of StSOS1s

In order to better understand the relationship between the structure and function of these StSOS1 proteins, gene structure

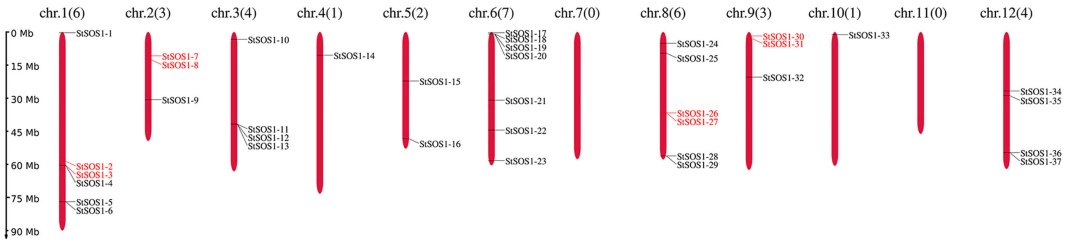


FIGURE 1
Distribution of the *StSOS1* genes in the potato on 12 chromosomes. The nomenclature for *StSOS1* members was based on the physical position from top to bottom on the chromosome, the names were displayed on the right-hand side of each chromosome, the number of chromosomes and the *StSOS1* genes were indicated at the top of each chromosome, and the scale of the genome size was given on the left-hand side. All protein sequences were obtained from potato genome data (SolTub_3.0).

10.3389/fpls.2023.1201730

TABLE 1 Detailed information regarding StSOS1 proteins in the potato.

Gene Name	Gene ID	Transcript ID	AA Number	MW (KDa)	pI	Na^+/H^+ Exchanger Domain (start-end)	Localization
StSOS1-1	PGSC0003DMG400022786	PGSC0003DMT400058653	1153	127.86	5.87	29-459	Plasma membrane
StSOS1-2	PGSC0003DMG400010663	PGSC0003DMT400027658	537	59.45	8.55	21-445	Extracellular
StSOS1-3	PGSC0003DMG400010663	PGSC0003DMT400027657	252	28.25	10.12	1-160	Extracellular
StSOS1-4	PGSC0003DMG400010663	PGSC0003DMT400027656	478	53.16	7.70	21-430	Extracellular
StSOS1-5	PGSC0003DMG400022490	PGSC0003DMT400057914	411	45.49	7.80	8-320	Extracellular
StSOS1-6	PGSC0003DMG400022490	PGSC0003DMT400057913	536	58.81	7.70	26-445	Extracellular
StSOS1-7	PGSC0003DMG400021928	PGSC0003DMT400056443	694	76.70	5.53	1-685	Plasma membrane
StSOS1-8	PGSC0003DMG400021928	PGSC0003DMT400056445	813	89.81	5.69	12-804	Extracellular
StSOS1-9	PGSC0003DMG400009710	PGSC0003DMT400025130	823	89.10	8.82	8-783	Plasma membrane
StSOS1-10	PGSC0003DMG400018689	PGSC0003DMT400048101	790	87.29	5.97	28-777	Plasma membrane
StSOS1-11	PGSC0003DMG400031029	PGSC0003DMT400079669	294	32.09	8.45	1-262	Membrane
StSOS1-12	PGSC0003DMG400031029	PGSC0003DMT400079670	500	54.21	8.82	92-464	Plasma membrane
StSOS1-13	PGSC0003DMG400031029	PGSC0003DMT400079671	577	62.96	7.14	169-541	Plasma membrane
StSOS1-14	PGSC0003DMG400027255	PGSC0003DMT400070102	791	87.48	7.89	43-775	Plasma membrane
StSOS1-15	PGSC0003DMG400009808	PGSC0003DMT400025403	209	22.51	4.50	162-204	Plasma membrane
StSOS1-16	PGSC0003DMG400011649	PGSC0003DMT400030419	793	87.85	6.74	23-790	Plasma membrane
StSOS1-17	PGSC0003DMG400007292	PGSC0003DMT400018809	807	89.34	8.19	24-806	Plasma membrane
StSOS1-18	PGSC0003DMG402021988	PGSC0003DMT400056557	269	30.01	8.81	2-179	Extracellular
StSOS1-19	PGSC0003DMG402021988	PGSC0003DMT400056556	306	34.27	9.11	3-216	Extracellular
StSOS1-20	PGSC0003DMG402021988	PGSC0003DMT400056555	252	27.51	4.96	25-231	Extracellular
StSOS1-21	PGSC0003DMG402021988	PGSC0003DMT400061554	832	91.61	7.08	13-773	Plasma membrane
StSOS1-22	PGSC0003DMG400013814	PGSC0003DMT400035881	841	91.99	6.61	11-827	Plasma membrane
StSOS1-23	PGSC0003DMG400030375	PGSC0003DMT400078102	738	80.33	7.10	1-692	Plasma membrane
StSOS1-24	PGSC0003DMG400035252	PGSC0003DMT400085681	424	45.08	9.03	3-408	Plasma membrane
StSOS1-25	PGSC0003DMG400030154	PGSC0003DMT400077544	832	91.88	5.37	17-778	Plasma membrane
StSOS1-26	PGSC0003DMG400029945	PGSC0003DMT400076994	599	64.77	7.6	179-551	Plasma membrane
StSOS1-27	PGSC0003DMG400029945	PGSC0003DMT400076993	389	41.77	5.65	175-367	Plasma membrane

(Continued)

TABLE 1 Continued

Gene Name	Gene ID	Transcript ID	AA Number	MW (KDa)	pI	Na^+/H^+ Exchanger Domain (start-end)	Localization
StSOS1-28	PGSC0003DMG400012169	PGSC0003DMT400031718	802	87.00	8.64	3-798	Plasma membrane
StSOS1-29	PGSC0003DMG400012168	PGSC0003DMT400031717	802	86.63	8.57	3-778	Plasma membrane
StSOS1-30	PGSC0003DMG400008849	PGSC0003DMT400022808	679	74.74	9.02	12-672	Plasma membrane
StSOS1-31	PGSC0003DMG400008849	PGSC0003DMT400022809	796	87.69	8.71	12-774	Plasma membrane
StSOS1-32	PGSC0003DMG400004171	PGSC0003DMT400010686	789	87.81	6.80	14-781	Plasma membrane
StSOS1-33	PGSC0003DMG400034953	PGSC0003DMT400085382	548	61.77	7.73	40-490	Extracellular
StSOS1-34	PGSC0003DMG400014998	PGSC0003DMT400038811	628	69.31	5.98	26-623	Plasma membrane
StSOS1-35	PGSC0003DMG400014998	PGSC0003DMT400038812	777	85.91	8.40	31-772	Plasma membrane
StSOS1-36	PGSC0003DMG400005009	PGSC0003DMT400012866	793	86.34	8.16	5-773	Plasma membrane
StSOS1-37	PGSC0003DMG400005009	PGSC0003DMT400012865	791	86.13	8.16	5-771	Plasma membrane

* http://plants.ensembl.org/index.html.

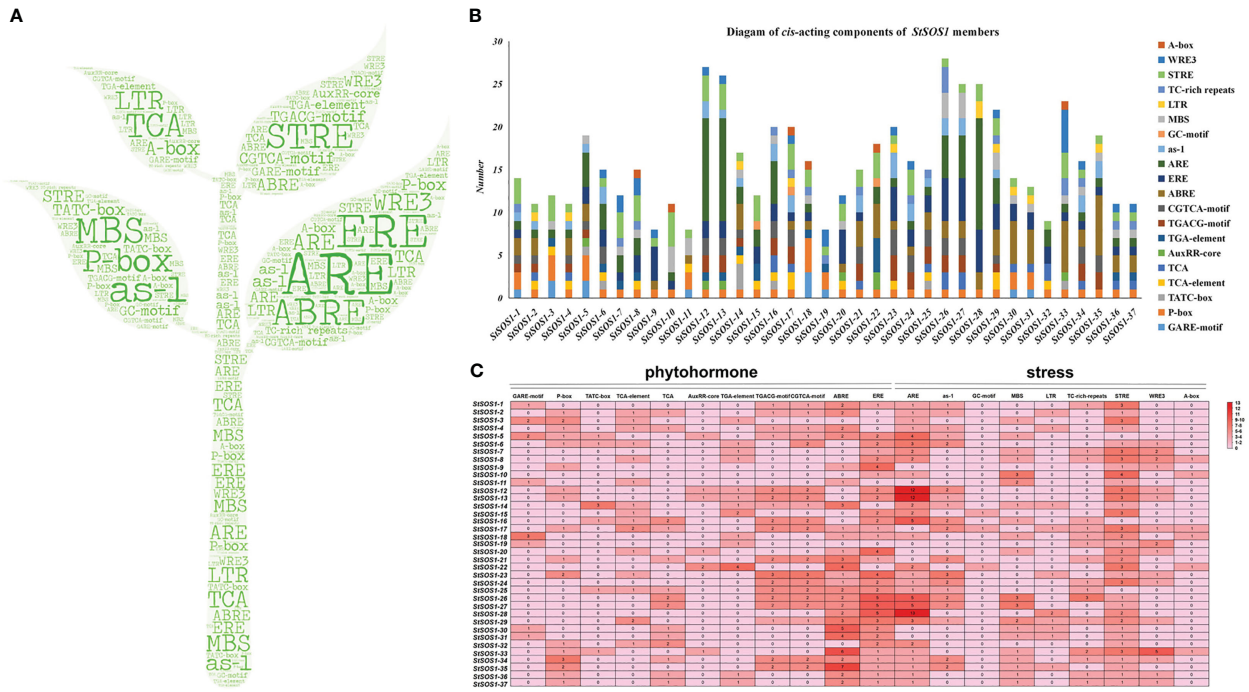


FIGURE 2
Analysis of the *cis*-acting elements. **(A)** Word clouds representing different *cis*-regulatory elements present at 2000 bp upstream of *StSOS1* genes sequences. **(B)** Graphical representation of 37 *StSOS1* genes with various roles in hormonal response, abiotic stress, plant development, defense, and stress response. **(C)** *Cis*-elements were denoted by different colors according to their number. The darker the color, the higher the occurrence frequency, and the number indicates the number of *cis*-elements.

and conserved motifs were analyzed to construct individual phylogenies. Depending on the different branches of the evolutionary tree, it has been found that the motif architectures remain consistent within the same evolutionary branch, and thus they may have a similar function (Figures 3A, B). The results showed that the number of intron in *StSOS1* genes ranged from two (*StSOS1-22*, *StSOS1-23*, *StSOS1-34*) to 20 (*StSOS1-13*, *StSOS1-26*, *StSOS1-27*). Furthermore, closely related genes share a similar structural architectures with different introns lengths (Figure 3C). The shortest StSOS1 protein was just 209 aa in length (StSOS1-15), while the longest was StSOS1-1, with a length of 1153 aa (Table 1). The functional sites in the conserved motifs were analyzed using the Eukaryotic Linear Motif resource server (ELM) and the results showed that there was a great functional divergency among these sites and most of the functional sites are related to phosphorylation, kinase phosphorylation, binding and sorting signal responsible for the interaction (Table S3).

3.4 Expression characterization of *StSOS1s*

To investigate the biological function of *StSOS1s* in different tissues, expression profiles of all identified *StSOS1* genes were analyzed in six different tissues, including roots, tubers, stolons, leaves, whole mature flowers, and mature whole fruit (Figure 4A). Of all the 21 *StSOS1* genes, *StSOS1-2* exhibits the highest levels of expression in almost all the tissues except the tubers. Some members of *StSOS1* exhibit highly tissue-specific expression, such as the expression of *StSOS1-10*, *StSOS1-16*, *StSOS1-17*, *StSOS1-19*, *StSOS1-22*, *StSOS1-23*, *StSOS1-32*, and *StSOS1-35* throughout the mature flower, suggesting that the *StSOS1* genes exhibit differential tissue-specific expression patterns. Then we analyzed spatio-temporal expression patterns in stolon, tuber pith, tuber peel, tuber cortex, young tuber, mature tuber and tuber sprout using RNA-seq data (Figure 4B). It showed that two genes (*StSOS1-14* and *StSOS1-32*) had a very low abundance in these tissues or organs. *StSOS1-16*, *StSOS1-19*, and *StSOS1-31* were predominantly expressed in stolon; *StSOS1-6* and *StSOS1-13* were predominantly expressed in tuber sprout. *StSOS1-1* was highly expressed in tuber pith, tuber peel, tuber cortex, young tuber and tuber sprouts. To have a better understand the function of *StSOS1s* under biotic stress, the expression pattern was observed responding to *Phytophthora infestans*, β-aminobutyric acid (BABA) and benzothiadiazole (BTH) treatment (Figure 4C). *StSOS1-2* was the only member to exhibit down-regulation under all three biotic stress conditions. Some genes show up-regulation, in particular one type of stress treatment; *StSOS1-6*, *StSOS1-13*, *StSOS1-28* and *StSOS1-29* showed up-regulation only in response to BABA treatment. For the abiotic stresses and phytohormones responsiveness of *StSOS1s*, we analyzed their transcript profiling in response to three abiotic stress and four phytohormone conditions mannitol, water-stress, heat, IAA, GA_3, BAP and ABA (Figure 4D). *StSOS1-6* was found to

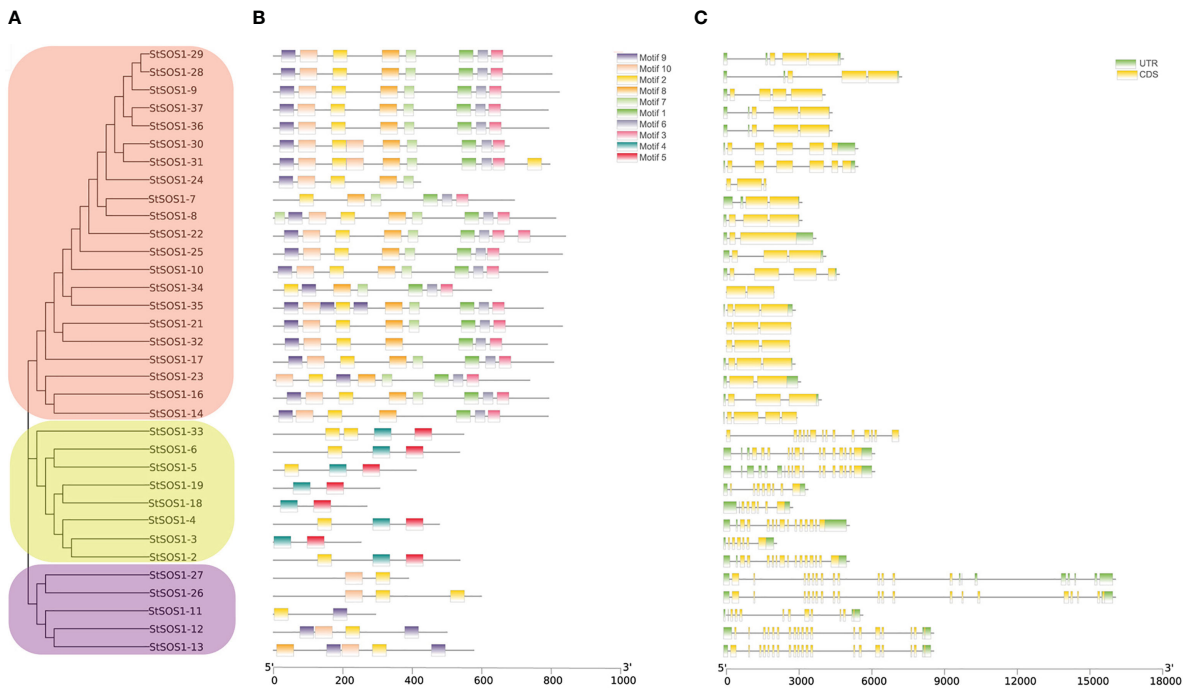


FIGURE 3
Phylogenetic relationships, structures, and motifs of members of the *StSOS1s* family [StSOS1-1 (1153 aa), StSOS1-15 (209 aa), and StSOS1-20 (252 aa) excepted]. **(A)** The phylogenetic tree of the StSOS1 proteins was constructed using the Maximum Likelihood method, which was based on conserved motifs and CDS/UTR structure. Different subgroups were represented by different background colors. **(B)** The conserved motifs of the StSOS1 proteins. Different patterns were represented by boxes of various colors, 5 'and 3' represent the N and C ends. **(C)** Gene structures, exons and untranslated regions (UTR) are shown in green and yellow boxes, while black lines indicated introns. Phylogenetic trees, conserved motifs, and gene structures were predicted using TBtools, and their lengths were estimated using bottom ruler.

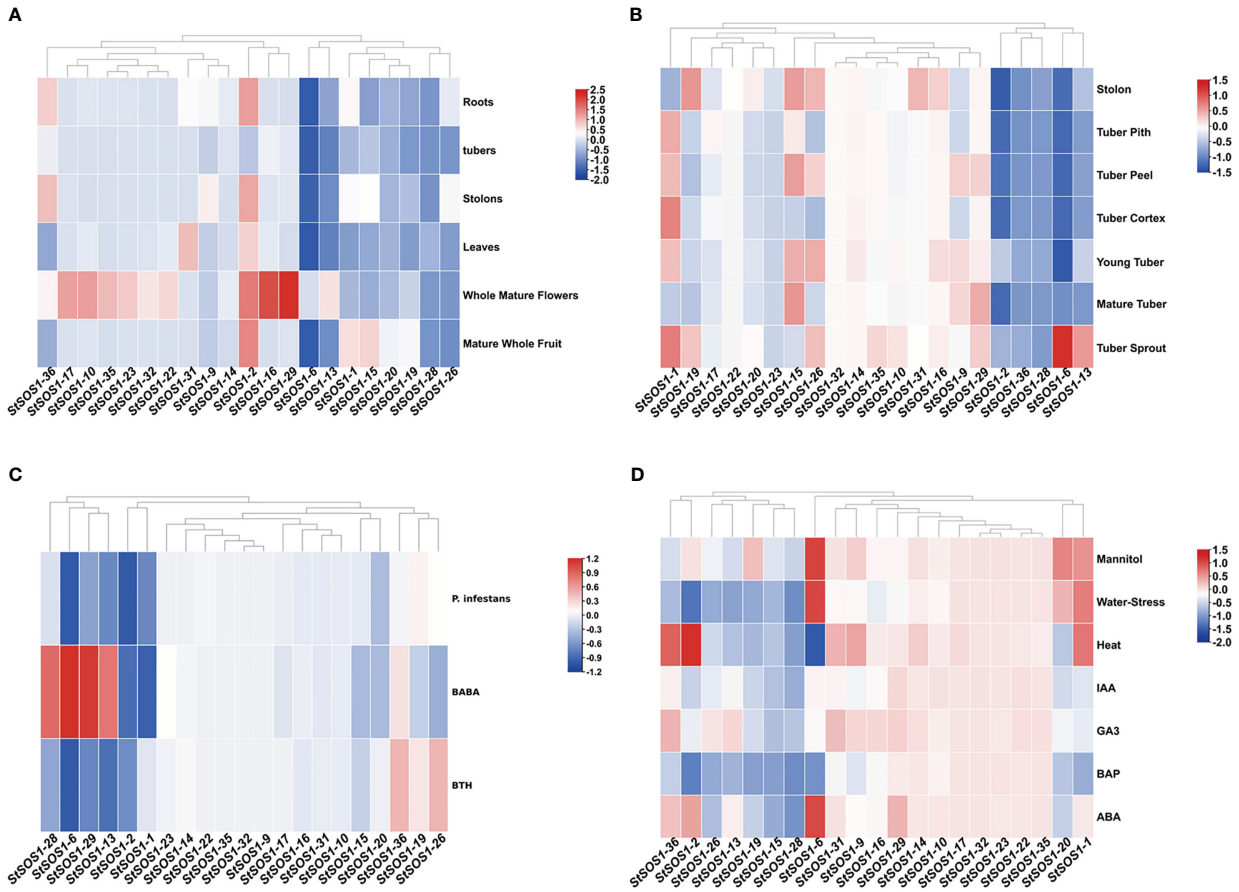


FIGURE 4
Expression levels of *StSOS1* genes in biotic, abiotic stress, and in different tissues and developmental stages. **(A)** Expression profiles of *StSOS1s* in different tissues and developmental stages. **(B)** Expression profiles of *StSOS1s* in different tissues and developmental stages of the potato tuber. **(C)** Expression of *StSOS1s* transcripts was altered in response to biotic stress. **(D)** Expression profiles of *StSOS1s* at abiotic stress and phytohormones. In the heat map, red, blue and white represent up-regulated, down-regulated, and unchanged (log_{10} ratio), respectively. Heat map and hierarchical clustering were performed by average linage (default) method.

be highly up-regulated in the three stress conditions of mannitol, water-stress and ABA. *StSOS1-10*, *StSOS1-17*, *StSOS1-22*, *StSOS1-23*, *StSOS1-32*, and *StSOS1-35* showed low or no expression in the eight tissues.

3.5 Gene ontology analyses of *StSOS1s*

To identify functions of up and down-regulated genes, GO analysis was performed and genes belonging to different categories of Biological Processes (BP), Molecular Functions (MF) and Cellular Compartments (CC) were identified (Figure 5). The BP categorized results showed that the up-regulated genes were significantly enriched in transport and cellular process. For MF, these up-regulated states enriched in transport activity. Moreover, up-regulated genes in the CC category are significantly enriched in both membrane and membrane-like components. In addition, the most significantly enriched GO terms for down-regulated genes were detection of hydrogen transport (BP), and transporter activity and antiporter activity (MF). It is important to note that the membrane integral, the membrane intrinsic and membrane fraction are all present in both up- and down-regulated genes in CC. The difference is that the up-regulated genes have a late endosome while the down-regulated genes do not. In summary, most GO terms are involved in membrane transport and composition, suggesting that they are likely to play an important role in maintaining proper ion homeostasis in the cytoplasm.

The verification of PPI is a defining aspect of molecular biology. PPI analysis was conducted to analyze the interactions among the SOS1s (Figure S2). The biological pathways and cellular compartments (retrieved from the GO) associated with these proteins were similar. Here, the interaction network between 96 SOS1-related genes was also mapped using the STRING database and Cytoscape software for function analyse, seven clusters were identified, including the pathways of biological regulation, membrane, ion transmembrane transporter activity, calcium ion binding, potassium ion transmembrane transport, response to salt stress and cellular process (Figure S3). Only 25 *StSOS1s* interact with other genes, and the most PPI was observed between proteins involved in potassium ion transmembrane transport, response to salt stress and cellular processes. These studies inform the biochemical mechanism of *StSOS1* and provide a new reference

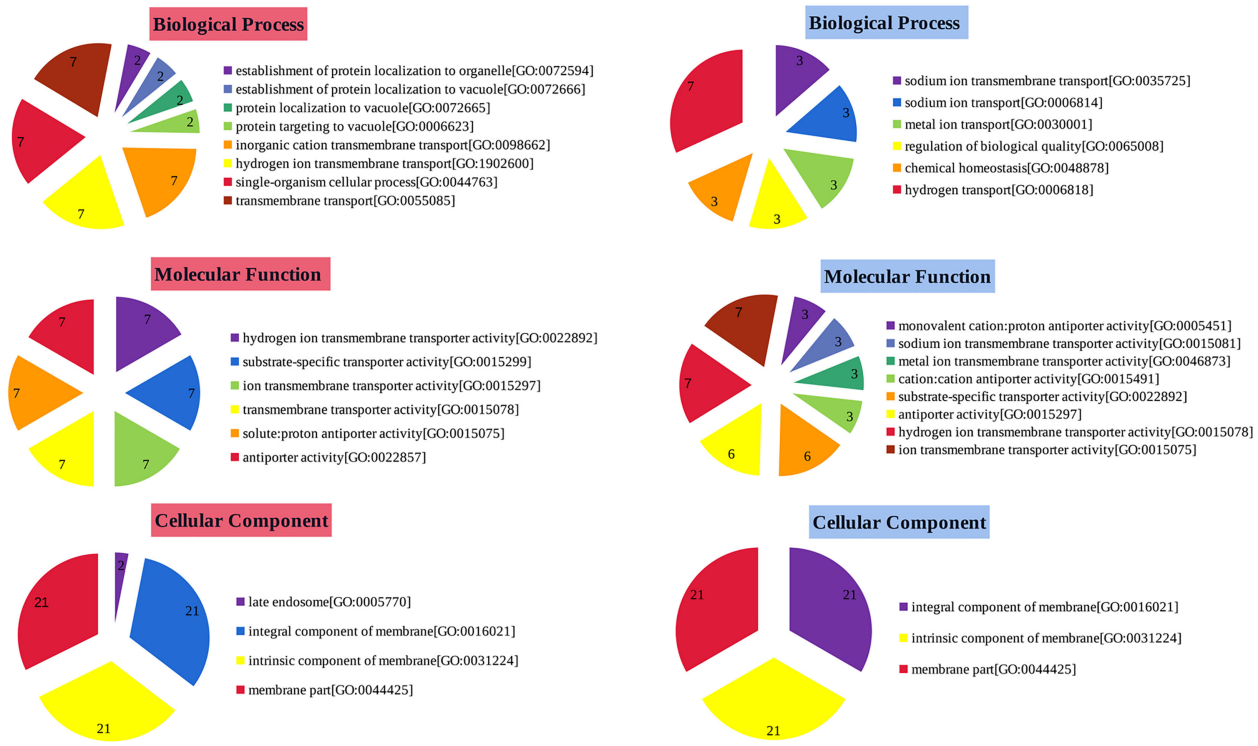


FIGURE 5
Gene ontology analyse of *StSOS1s*. Functional classification of up- and down-regulated genes by GO analysis into categories of Biological Process, Molecular Function and Cellular Component. The number in each pie represents the number of times the function (BP, CC, MF) was expressed in the data (abundance). GO-based classifications of up-regulated genes were shown in red, while those of down-regulated genes were shown in blue.

for the interplay between ion homeostasis and transmembrane transport during plant salt tolerance.

3.6 Phylogenetic and collinearity analyses of *StSOS1s*

For the evolutionary relationship of SOS1s among Arabidopsis, tomato, pepper, potato and tobacco, we extracted and compared the protein sequences of SOS1s in these species, and constructed the phylogenetic tree of neighbor junction (NJ) (Figure 6A). Potato SOS1s are named based on their position relative to orthologs from four other species on the tree. 134 SOS1 candidates of five species were grouped into four distinct classes (I-IV) based on sequence conservation. Among them, the subgroup I had 13 members (11.19%), subgroup II 27 (20.14%) and subgroup III 37 (27.61%), respectively. The subgroup IV contained 57 genes and had the most members (42.54%). The phylogenetic relationships indicate that the SOS1 proteins in the potato are more strongly homologous to pepper and tomato than to Arabidopsis and tobacco. Gene duplication has always played a key role in the expansion of genes and the occurrence of novel functions of genes. To explore the evolution of *SOS1* genes, we studied the replication patterns of the five species and performed genetic correlation analysis (Figure 6B). The results showed that there were 17, 8, 5, and 1 *SOS1* members participating in the potato-tomato, potato-pepper, potato-Arabidopsis and potato-tobacco synteny relations, respectively. Among the above collinear gene pairs, *StSOS1-11* with *SlSOS1-23*, *CaSOS1-10*, *AtSOS1-58*, respectively; *StSOS1-28* with *SlSOS1-26*, *CaSOS1-1*, *AtSOS1-47*, respectively; and *StSOS1-37* with *SlSOS1-26*, *CaSOS1-37*, *AtSOS1-55*, respectively, had simultaneously collinear relations.

3.7 Expression analysis of *StSOS1* genes under different abiotic stresses

The SOS pathway plays an important role in maintaining proper ion homeostasis in the cytoplasm and in regulating plant tolerance to salinity. However, there is limited information on *SOS1*'s response to potato salt stress. In order to investigate the potato response to salt stress, the *StSOS1* genes were analyzed using the transcriptomic data of potato exposed to NaCl treatment. Only 21 *StSOS1* genes showed differential gene expression pattern and were identified and visualized in a heat map (Figure 7A). Furthermore, six *StSOS1* genes in potato leaves of different grow stages under salt stress were randomly selected and quantitative analyzed by RT-qPCR (Figures 7B–G). These results suggested that these six genes were significantly differentially up-regulated under salt stress, which may positively regulate salt tolerance in the potato, this is not consistent with the heat map, which may be related with different levels of expression under different levels of salt stress treatment. *StSOS1-2*, *StSOS1-6* and *StSOS1-28* occurred two up-regulated expressions phenomenon under salt stress, this could be

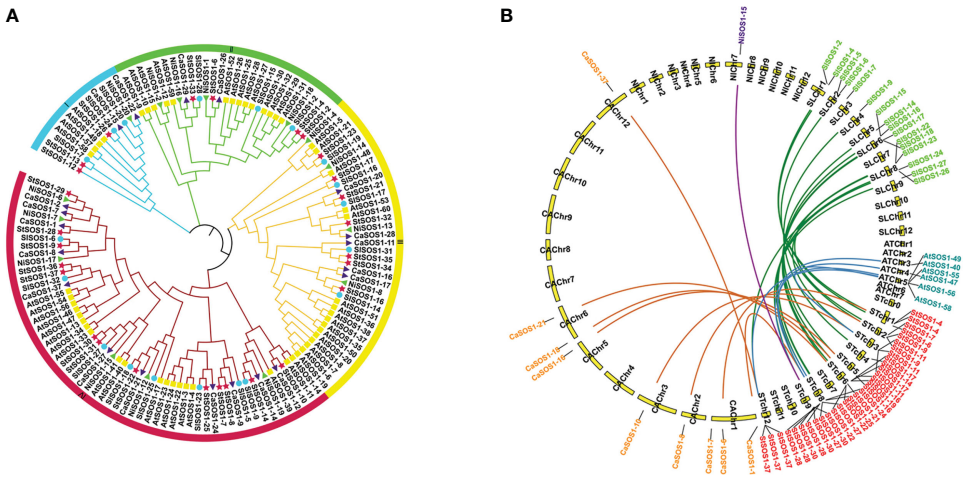


FIGURE 6
A Phylogenetic analysis of SOS1 proteins. **(A)** phylogenetic tree of SOS1 proteins was constructed with neighbor-junction (NJ) phylogenetic tree. The four subgroups were shown in different colors. The red stars represent potato SOS1s (StSOS1s), the green triangles represent tobacco SOS1s (NiSOSs), the purple triangles represent pepper SOS1s (CaSOSs), the yellow boxes represent Arabidopsis SOS1s (AtSOSs) and the blue circles represent tomato SOS1s (SlSOSs). **(B)** Collinearity analysis of SOS1s in potato and other plants. The green, orange, blue and purple lines in the background correspond to collinear gene pairs in potato and tomato, potato and pepper, and potato and Arabidopsis, potato and tobacco, respectively.

related to the response period of the SOS1 signaling pathway. Notably, the expression of *StSOS1-13* was 14-fold higher at 3 d after salt treatment compared to expression levels before salt stress, and then reached 32-fold higher at 4 d, suggesting that *StSOS1-13* may be an important candidate gene involved in the salt stress response.

To further understand potential function changes in *StSOS1-13* gene in response to abiotic stress, RT-qPCR was used to analyze the expression patterns of the selected *StSOS1-13* gene in phytohormone treatment (Figure 8). It was observed that the *StSOS1-13* was up-regulated on exposure to ABA, GA, and SA treatment, and the magnitude of up-regulation was higher in ABA treatment as compared to GA, SA treatment. Conversely, for the MeJA treatment, expression in the leaves decreased after 0-2 d and then increased continuously, with the highest levels of expression in the leaves at 5 d. Overall, these results indicated that *StSOS1-13* may play a critical regulatory role in response to abiotic stress.

3.8 Subcellular localization of StSOS1-13

Detecting the subcellular localization of StSOS1-13 is essential to elucidate their function. The subcellular localization of StSOS1-13 predicted by the Cell-PLoc 2.0 tool revealed that the StSOS1-13 protein was localized in the plasma membrane. To further verify the location of StSOS1-13 protein, the full-length coding sequence of StSOS1-13 deleted stop codon was fused with green fluorescence protein (GFP) and the transient expression was performed under the control of 35S promoter in tobacco. The results showed that the StSOS1-13 protein is localized in the plasma membrane (Figure 9), this is consistent with the result of bioinformatics analysis.

4 Discussion

Soil salinity is one of the most significant abiotic stresses faced by crop plants in agricultural fields worldwide (Świeżawska et al., 2018), reducing crop yield and production (Rolly et al., 2020). Plants have evolved the SOS pathway to achieve salt tolerance (Cha et al., 2022), the SOS pathway comprising *SOS1*, *SOS2* and *SOS3* has been proposed to regulate cellular signaling during salt stress to mediate ion homeostasis (Luo et al., 2022). *SOS1* is a critical salt tolerance determinant in plants (Świeżawska et al., 2018). *SOS1* genes have been reported to improve the tolerance to salt stresses in plants such as Arabidopsis (Wu et al., 1996), soybean (Zhang et al., 2022), and maize (Zhou et al., 2022). Potato is one of the most crucial crops in the world due to its nutritional quality (Takeuchi et al., 2022). The crop can also be used as a commercial health food because it is high in antioxidants, minerals, and dietary fibers (Kumar et al., 2021). In addition, potato plants are often subjected to various types of abiotic stress during growth and development (Yang et al., 2020; Kumar et al., 2021). It was reported that soil salinization negatively affected the growth and yield of potato crops, especially in arid and semi-arid climates (Li et al., 2022), which caused osmotic and oxidative stress, ion imbalance, mineral deficiency, and ion toxicity problems (Hamooh et al., 2021). Therefore, the selection and breeding of salt-tolerant genes has become a promising approach for improving the yield and adaptability of potato (Zhu et al., 2022). Previous studies have shown that a gene encoding SOS2 (PGSC0003DMG400006384) is up-regulated, indicating that this gene plays an active regulatory role in salt stress response. However, the complete SOS pathway for salt stress response in potato has not been established, and only a few genes of this pathway have been

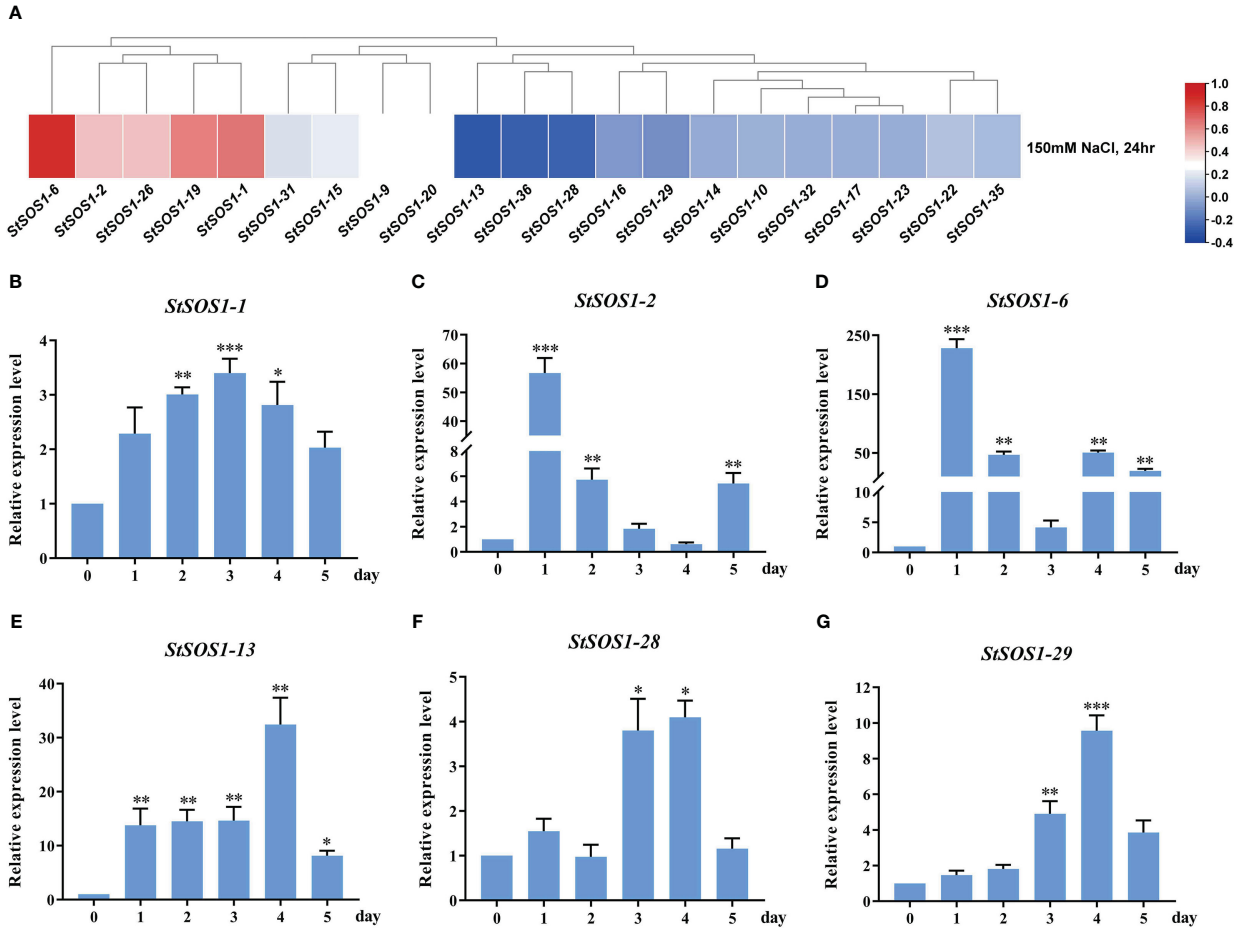


FIGURE 7
The expression pattern of *StSOS1s* under salt stress. **(A)** Expression profiles of *StSOS1s* at NaCl stress based on RNA seq-date. **(B-G)** RT-qPCR profiles of *StSOS1* genes under salt stress. The expression level of *StSOS1s* on control was normalized as "1". The vertical bars indicate the standard error of the mean. Asterisks indicate significant differences based T test (*, $p < 0.05$, **, $p < 0.01$, ***, $p < 0.001$, ****, $p < 0.0001$).

reported (Li Q. et al., 2020). The aim of this study is to screen for key *StSOS1* genes that are more sensitive to abiotic stress and to lay the groundwork for further unraveling the regulatory mechanisms of *SOS1* genes in potato.

In this study, a total of 37 *SOS1* family members were identified in potato (Table 1) and they locate in 10 of 12 chromosomes (Figure 1) which were significantly lower than Arabidopsis (60 *SOS1s* in Figure 6A). Gene duplication may explain the difference in the number of *SOS1* family members between the potato and Arabidopsis. A possible explanation for this is that *SOS1* genes in the potato may have a higher rate of gene loss than in Arabidopsis, and frequent gene loss has been reported in various plant species during genome duplication events (Li et al., 2020), indicating a key role of gene duplication over the course of evolution in various species (Zhang et al., 2021). Some of the duplicated genes may be retained in its descendants, which could provide the original genetic resource for the adaptive evolution of plants (Flagel and Wendel, 2009). The number of *SOS1* genes in the potato was similar to that in the pepper. Phylogenetic analysis demonstrated that the *Solanaceae SOS1* genes were generally classified into four clades (Figure 6A). Interestingly, four subfamilies were present in all five plant species, suggesting that genetic expansion occurred prior to the divergence of these plant species. By comparing the syntenic analysis of *SOS1* genes in potato and four other plants (tomato, pepper, Arabidopsis and tobacco), we found that the sequence similarity between the *SOS1* gene pairs within potato was much higher than that between the tomato and the pepper (Figure 6B), which is consistent with the phenomenon in chrysanthemum (Gao et al., 2016), indicating the similarity of evolutionary relationship among different species in the same group. The conserved motif analysis of *SOS1s* family revealed the occurrence of 10 conserved motifs (Figure 3) might be related to specific functions shared among *SOS1* family members. In addition, *StSOS1s* within the same subfamily share a high degree of similarity in exon-intron structures and conserved motifs. The loss and gain of introns may reflect evolutionary trends in genes with similar functions (Rogozin et al., 2003), which had been demonstrated in *Brassica juncea* (Cheng et al., 2019).

In salt-acclimated tobacco, the compartmentalization of Na^+ in vacuoles may be mediated by vesicle transport (Garcia de la Garma et al., 2015), which represents an over-sensitive mechanism of the Na^+/H^+ antitransporter SOS1 to accommodate salt stress (Hamaji

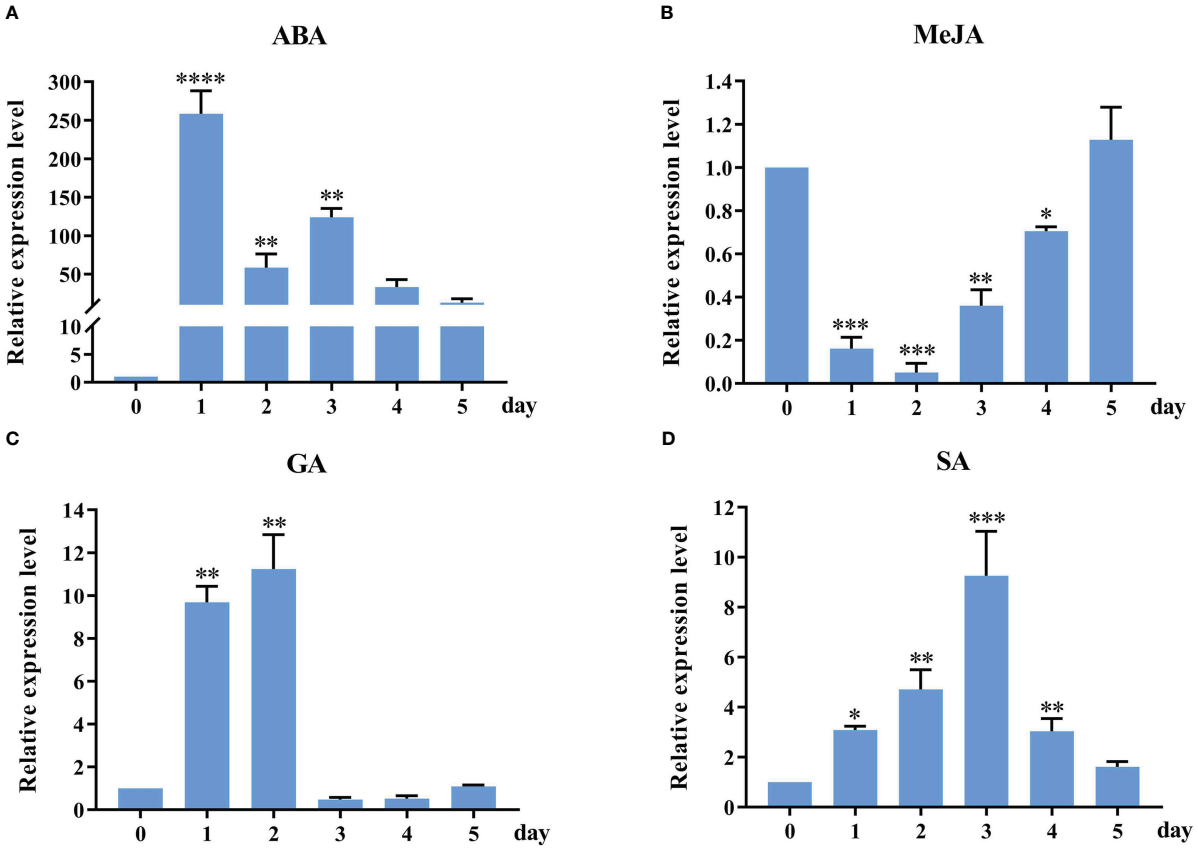


FIGURE 8
RT-qPCR profiles of *StSOS1-13* gene under phytohormone treatment. The *StSOS1-13* expression level of control was normalized as "1". The vertical bars indicate the standard error of the mean. Asterisks indicate significant differences based on T test (*, $p < 0.05$, **, $p < 0.01$, ***, $p < 0.001$, ****, $p < 0.0001$).

et al., 2009; Zhao S. et al., 2021). When the SOS signaling pathway is activated, the Na^+/H^+ antiport activity of *SOS1* is enhanced and the accumulated Na^+ is transported out of the cell (Xie et al., 2022). For further functional analysis, we use GO annotation enrichment analysis to functionally annotate different *StSOS1s*. Gene ontology is a fundamental analysis that predicts the contribution of putative functions across living organisms. In the present study, GO analysis revealed the significant role of *StSOS1s* with cellular process, transport and component of membrane (Figure 5). To support this argument, we have constructed an additional PPI network with StSOS1 proteins as the core (Figure S2). Among the numerous functional modulated by the SOS1 network, there are the regulation pathways of biological regulation, membrane, ion transmembrane transporter activity, calcium ion binding, potassium ion transmembrane transport, response to salt stress and cellular process (Figure S3). Most of the *StSOS1* genes are involved in cellular transport process (Figure S2), suggesting that they probably play a vital role in maintaining appropriate ion homeostasis in the cytoplasm.

The *cis*-elements and functional characteristics of *SOS* genes promoters have been identified in many species, such as *Brassica juncea* var. *Tumida* (Cheng et al., 2019), *B. juncea* (Kaur et al., 2015), and Arabidopsis (Feki et al., 2015). To further explore the possible function of *SOS1s* in potato, we performed an analysis of *cis*-acting regulatory elements in the promoter region in this study. *Cis*-regulatory elements were found to include phytohormone (SA, MeJA, ABA, auxin, GA) and abiotic stresses (cold, defense and stress response, drought) (Figure 2), which is consistent with the report about the previous studies in other species. More importantly, the *cis*-elements involved in the GA response are present in the promoters of all *StSOS1s*, and more than half of the promoters of *StSOS1s* have MYB elements involved in drought-inducibility. Interestingly, in the heat map (Figure 4D), *StSOS1s* could be induced by both auxin and GA, two important plant hormones in regulation. Most of the *StSOS1s* notably up-regulate under both mannitol and NaCl stress conditions. Overall, the results presented above revealed that *StSOS1s* may play a significant role in the response to phytohormone and abiotic stresses.

In wheat, most *TaSOS1* genes expressed in different tissues, including shoots, leaves, spikes, and grains (Jiang et al., 2021). In Arabidopsis, *AtSOS1* promoter-driven GUS expressed primarily in the roots, inflorescences and leaves (Yang et al., 2009). Our results revealed *StSOS1-2* and *StSOS1-31* were specifically expressed in leaves whereas *StSOS1-10*, *StSOS1-16*, *StSOS1-17*, *StSOS1-19*, *StSOS1-22*, *StSOS1-23*, *StSOS1-32* and *StSOS1-35* had clear expression preference in whole mature flowers (Figure 4A). These suggested that the *StSOS1s* played a significant role in the growth and development of different potato organs. In addition, the

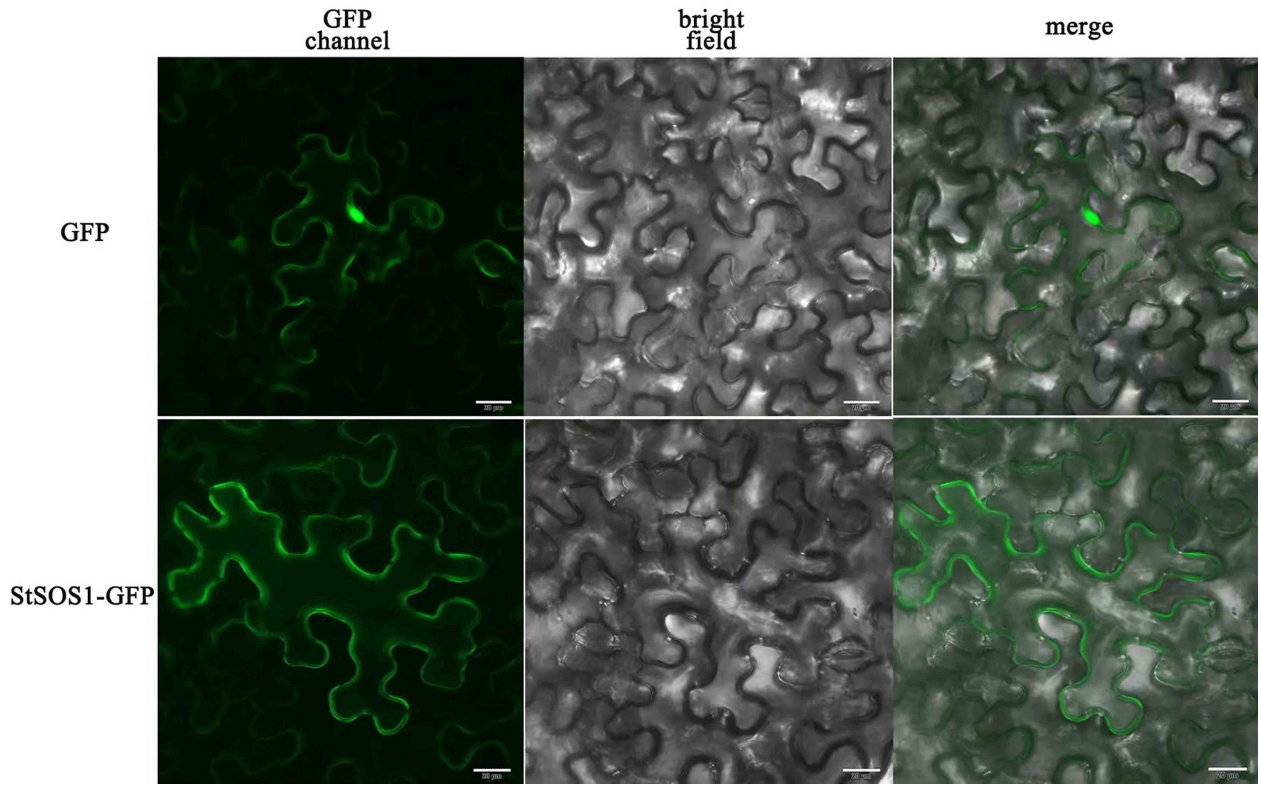


FIGURE 9
Subcellular localization of StSOS1-13 in *Nicotiana benthamiana*. The leaves were injected with a strain of *Agrobacterium tumefaciens* containing 35S::StSOS1-13-GFP, and the empty vector 35S::GFP were used as a control. After 48 h of injection, pStSOS1-13-GFP fusion protein and GFP alone transiently expressed separately in leaves, the dark field was green fluorescence and the white field was cell morphology, with Confocal combined detection. GFP, GFP fluorescence (green). Bright, bright fields. Merge, superimpose GFP and bright-field images. The experiment repeated three times with similar results. Scale bar, 20 μm.

StSOS1s had the similar tissue-specific expression patterns with the *AtSOS1s*, this suggested that the *SOS1* gene family played a conserved function in both Arabidopsis and potato.

Under salt stress, *SOS1* gene expression levels of *Populus euphratica* and *Chrysanthemum crassum* were up-regulated (Wu et al., 2007; Song et al., 2012). There were differences in *SOS1* gene expression in cotton at different time intervals (Akram et al., 2020). In this study, compared with the control, the expression level of *StSOS1s* in leaves was immediately up-regulated under salt stress, and the results of RT-qPCR of *StSOS1-1*, *StSOS1-2* and *StSOS1-6* were highly consistent with the results of heat map (Figures 7B–D). In addition, in wheat, the expression of *SOS1* in leaves under salt stress was consistent with mRNA abundance (Xu et al., 2008). However, the RT-qPCR results of *StSOS1-13*, *StSOS1-28* and *StSOS1-29* were contrary to the down-regulated results of heat map within 24 h (Figures 7E–G). In purslane (*Sesuvium portulacastrum*), the RT-qPCR results also differ from the heat map results. That is, the quantitative expression level of *SpSOS1* in roots increased sharply within 3-6 h and then decreased to the basic level, while the transcription abundance of *SOS1* in leaves did not change significantly within 48 h of NaCl treatment (Zhou et al., 2015). In addition, the expressions of *StSOS1-2*, *StSOS1-6* and *StSOS1-28* in leaves were up-regulated twice (Figures 7C–D, F). Similarly, the expression level of *GhSOS1* under salt stress also showed this phenomenon (Chen et al., 2017). In conclusion, the mechanism of *SOS1* in potato salt stress resistance is relatively complex and more studies are needed to determine the function of SOS1s in potato in the future.

Exogenous ABA, MeJA, SA treatment can improve the yield of potato (Pérez-Alonso et al., 2021). Under ABA stress, the expression of *BjSOS* genes increased with increasing stress duration in both contrasting genotypes (Nutan et al., 2018). Several reports have suggested co-expression of many stress-responsive genes at both salinity and ABA (Takahashi et al., 2004). Our results of RT-qPCR analysis indicated that the *StSOS1-13* was expressed under four phytohormone treatment (Figure 8). *StSOS1-13* was significantly up-regulated about 3, 9, and 250 times at 1 d in leaves under SA, GA, and ABA treatment, respectively, while *StSOS1-13*, was down-regulated under MeJA treatment. The promoter biological function is further corroborated by the expression analysis of *StSOS1-13* in response to hormonal stress. *StSOS1s* may regulate the expression of genes involved in the transduction of hormone signals, and thus participate in plant growth and development.

Studies have reported that the excessive Na^+ ions in soil can cause imbalance *in vivo*, moisture deficiency and ion toxicity (Tester and Davenport, 2003), so some plants formed a Na^+ efflux and Na^+ segment processing. As a result, some plants have developed Na^+

efflux and Na^+ segment treatments to maintain low intracellular Na^+ concentrations to accommodate the effects of salt stress on plant growth and development. The SOS pathway studied previously is a more classical salt signaling pathway (Chinnusamy et al., 2004). Arabidopsis salt-tolerant site SOS1 encodes Na^+/H^+ antiporter. Confocal imaging of a green fluorescent protein fusion protein of SOS1 in a transgenic Arabidopsis plant revealed that SOS1 is localized in the plasma membrane (Shi et al., 2002). SOS3 and SOS2, which are located in the cytoplasm, regulate SOS1 on the cytoplasmic membrane, which will therefore achieve an intracellular balance of Na^+ (Hill et al., 2013). Protein subcellular location is key in determining the function and accumulation patterns of plant proteins (Hooper et al., 2020). In *Chrysanthemum crassum*, CcSOS1 was expressed close to the plasma membrane in transiently transformed onion epidermal cells (Song et al., 2012). Like the *A. thaliana* homologue *AtSOS1* (Shi et al., 2002), CcSOS1 is regulated by salinity, especially in the roots after stress, and could play an important role in salt tolerance in *C. crassum*. In rice (Gupta et al., 2021) and cotton (Guo et al., 2020), SOS1 genes were also predicted to express in plasma membrane. To investigate the subcellular localization of *StSOS1-13*, the cassette encoding *StSOS1-13*-Green Fluorescent protein (GFP) fusion protein driven by the CaMV 35S promoter (35S::*StSOS1-13*-GFP) was transformed into *Nicotiana benthamiana* leaves, and the fluorescence was observed using the confocal microscope. Fluorescence localization verified that the selected StSOS1-13 was expressed in the plasma membrane (Figure 9), demonstrating the reliability and accuracy of the predicted results.

5 Conclusions

This study provides a genome-wide analysis of the *StSOS1* genes, with 37 *StSOS1*s in the potato identified and divided into three subfamilies. We found that segmental and tandem duplication contribute to the expansion of *StSOS1* gene family. These *StSOS1s* phylogenetically cluster with *SlSOS1s* and *CaSOS1s*. The exon-intron structures and motifs of *StSOS1s* further suggest that the potato SOS1 proteins were highly conserved within the subfamilies. In addition, subcellular localization in *Nicotiana benthamiana* suggested that StSOS1-13 was located on the plasma membrane. The RT-qPCR results suggested the crucial role of the *StSOS1s* in response to salt and homologous stress, and suggested that some specific up-regulated genes such as *StSOS1-1*, *StSOS1-13*, and *StSOS1-29* would be potential candidates for potato salt-tolerant seeding. The results presented in this study will provide essential clues in elucidating the role of the *StSOS1s* in abiotic stress and the mechanisms underlying the tolerance to salt stress in potato mediated by the StSOS1 proteins.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the corresponding author GG (ggsxnu@126.com), without undue reservation. The potato RNA-Seq data in this article can be download in NCBI with accession number ERP000527.

Author contributions

LL conceived and designed the study. LG analyzed and mapped the bioinformatics content, designed and performed the experimental work, interpreted and analyzed the data, and wrote the manuscript. YZ and ZH carried out the experimental work. WW, XZ, YW, XL, and SG helped to supplement the bioinformatics content and beautify the images. GG and WL supervised the project and critically revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the National Natural Science Foundation of China (31771858), Natural Science Foundation of Shanxi Province (202203021211249, 202203021211259), Postgraduate Innovation Project of Shanxi Normal University (2021XSYO27, 2022XSY004), Innovation and Entrepreneurship Training Program for College Students of Shanxi Province (20220303).

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.2023.1201730/full#supplementary-material

References

Świeżawska, B., Duszyn, M., Jaworski, K., and Szmidt-Jaworska, A. (2018). Downstream targets of cyclic nucleotides in plants. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01428

Akram, U., Song, Y., Liang, C., Abid, M. A., Askari, M., Myat, A. A., et al. (2020). Genome-wide characterization and expression analysis of NHX gene family under salinity stress in *Gossypium barbadense* and its comparison with *Gossypium hirsutum*. *Genes (Basel)* 11, 803. doi: 10.3390/genes11070803

Ali, A., Alexandersson, E., Sandin, M., Resjö, S., Lenman, M., Hedley, P., et al. (2014). Quantitative proteomics and transcriptomics of potato in response to *Phytophthora infestans* in compatible and incompatible interactions. *BMC Genomics* 15, 497. doi: 10.1186/1471-2164-15-497

Ali, A., Raddatz, N., Pardo, J. M., and Yun, D. J. (2021). HKT sodium and potassium transporters in *Arabidopsis thaliana* and related halophyte species. *Physiol. Plant* 171, 546–558. doi: 10.1111/ppl.13166

Brindha, C., Vasantha, S., Raja, A. K., and Tayade, A. S. (2021). Characterization of the salt overly sensitive pathway genes in sugarcane under salinity stress. *Physiol. Plant* 171, 677–687. doi: 10.1111/ppl.13245

Ceci, A. T., Franceschi, P., Serni, E., Perenzoni, D., Oberhuber, M., Robatscher, P., et al. (2022). Metabolomic characterization of pigmented and non-pigmented potato cultivars using a joint and individual variation explained (JIVE). *Foods* 11, 1708. doi: 10.3390/foods11121708

Cha, J. Y., Kim, J., Jeong, S. Y., Shin, G. I., Ji, M. G., Hwang, J. W., et al. (2022). The Na(+)/H(+) antiporter SALT OVERLY SENSITIVE 1 regulates salt compensation of circadian rhythms by stabilizing GIGANTEA in *Arabidopsis*. *Proc. Natl. Acad. Sci. U. S. A.* 119, e2207275119. doi: 10.1073/pnas.2207275119

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp

Chen, X., Lu, X., Shu, N., Wang, D., Wang, S., Wang, J., et al. (2017). *GhSOS1*, a plasma membrane Na^+/H^+ antiporter gene from upland cotton, enhances salt tolerance in transgenic *Arabidopsis thaliana*. *PloS One* 12, e0181450. doi: 10.1371/journal.pone.0181450

Cheng, C., Zhong, Y., Wang, Q., Cai, Z., Wang, D., and Li, C. (2019). Genome-wide identification and gene expression analysis of SOS family genes in tuber mustard (*Brassica juncea* var. *tumida*). *PloS One* 14, e0224672. doi: 10.1371/journal.pone

Chinnusamy, V., Schumaker, K., and Zhu, J. K. (2004). Molecular genetic perspectives on cross-talk and specificity in abiotic stress signalling in plants. *J. Exp. Bot.* 55, 225–236. doi: 10.1093/jxb/erh005

Contreras-Moreira, B., Naamati, G., Rosello, M., Allen, J. E., Hunt, S. E., Muffato, M., et al. (2022). Scripting analyses of genomes in ensembl plants. *Methods Mol. Biol.* 2443, 27–55. doi: 10.1007/978-1-0716-2067-0_2

Dahal, K., Li, X. Q., Tai, H., Creelman, A., and Bizimungu, B. (2019). Improving potato stress tolerance and tuber yield under a climate change scenario - a current overview. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00563

Fayez, A. G., Esmaiel, N. N., Salem, S. M., Ashaat, E. A., El-Saiedi, S. A., and El Ruby, M. O. (2022). miR-454-3p and miR-194-5p targeting cardiac sarcolemma ion exchange transcripts are potential noninvasive diagnostic biomarkers for childhood dilated cardiomyopathy in Egyptian patients. *Egypt Heart J.* 74, 65. doi: 10.1186/s43044-022-00300-x

Feki, K., Brini, F., Ben Amar, S., Saibi, W., and Masmoudi, K. (2015). Comparative functional analysis of two wheat Na(+)/H (+) antiporter SOS1 promoters in *Arabidopsis thaliana* under various stress conditions. *J. Appl. Genet.* 56, 15–26. doi: 10.1007/s13353-014-0228-7

Flagel, L. E., and Wendel, J. F. (2009). Gene duplication and evolutionary novelty in plants. *New Phytol.* 183, 557–564. doi: 10.1111/j.1469-8137

Gao, J., Sun, J., Cao, P., Ren, L., Liu, C., Chen, S., et al. (2016). Variation in tissue na (+) content and the activity of *SOS1* genes among two species and two related genera of chrysanthemum. *BMC Plant Biol.* 16, 98. doi: 10.1186/s12870-016-0781-9

Garcia de la Garma, J., Fernandez-Garcia, N., Bardisi, E., Pallol, B., Asensio-Rubio, J. S., Bru, R., et al. (2015). New insights into plant salt acclimation: the roles of vesicle trafficking and reactive oxygen species signalling in mitochondria and the endomembrane system. *New Phytol.* 205, 216–239. doi: 10.1111/nph.12997

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40 (Database issue), D1178–D1186. doi: 10.1093/nar/gkr944

Guo, W., Li, G., Wang, N., Yang, C., Zhao, Y., Peng, H., et al. (2020). A Na(+)/H(+) antiporter, K2-NhaD, improves salt and drought tolerance in cotton (*Gossypium hirsutum* l.). *Plant Mol. Biol.* 102, 553–567. doi: 10.1007/s11103-020-00969-1

Gupta, B. K., Sahoo, K. K., Anwar, K., Nongpiur, R. C., Deshmukh, R., Pareek, A., et al. (2021). Silicon nutrition stimulates salt-overly sensitive (SOS) pathway to enhance salinity stress tolerance and yield in rice. *Plant Physiol. Biochem.* 166, 593–604. doi: 10.1016/j.plaphy.2021.06.010

Hamaji, K., Nagira, M., Yoshida, K., Ohnishi, M., Oda, Y., Uemura, T., et al. (2009). Dynamic aspects of ion accumulation by vesicle traffic under salt stress in arabidopsis. *Plant Cell Physiol.* 50, 2023–2033. doi: 10.1093/pcp/pcp143

Hamooh, B. T., Sattar, F. A., Wellman, G., and Mousa, M. A. A. (2021). Metabolomic and biochemical analysis of two potato (*Solanum tuberosum* l.) cultivars exposed to *in vitro* osmotic and salt stresses. *Plants (Basel)* 10, 98. doi: 10.3390/plants10010098

He, F., Shi, Y. J., Chen, Q., Li, J. L., Niu, M. X., Feng, C. H., et al. (2022). Genome-wide investigation of the *PtrCHLP* family reveals that *PtrCHLP3* actively mediates poplar growth and development by regulating photosynthesis. *Front. Plant Sci.* 13. doi: 10.3389/fpls

Hill, C. B., Jha, D., Bacic, A., Tester, M., and Roessner, U. (2013). Characterization of ion contents and metabolic responses to salt stress of different arabidopsis *AtHKT1;1* genotypes and their parental strains. *Mol. Plant* 6, 350–368. doi: 10.1093/mp/sss125

Hooper, C. M., Castleden, I. R., Aryamanesh, N., Black, K., Grasso, S. V., and Millar, A. H. (2020). CropPAL for discovering divergence in protein subcellular location in crops to support strategies for molecular crop breeding. *Plant J.* 104, 812–827. doi: 10.1111/tpj.14961

Ilzhöfer, D., Heinzinger, M., and Rost, B. (2022). SETH predicts nuances of residue disorder from protein embeddings. *Front. Bioinform.* 2. doi: 10.3389/fbinf

Jiang, W., Pan, R., Buitrago, S., Wu, C., Abou-Elwafa, S. F., Xu, Y., et al. (2021). Conservation and divergence of the *TaSOS1* gene family in salt stress response in wheat (*Triticum aestivum* l.). *Physiol. Mol. Biol. Plants.* 27, 1245–1260. doi: 10.1007/s12298-021-01009-y

Kaur, C., Kumar, G., Kaur, S., Ansari, M. W., Pareek, A., Sopory, S. K., et al. (2015). Molecular cloning and characterization of salt overly sensitive gene promoter from *Brassica juncea* (*BjSOS2*). *Mol. Biol. Rep.* 42, 1139–1148. doi: 10.1007/s11033-015-3851-4

Keisham, M., Mukherjee, S., and Bhatla, S. C. (2018). Mechanisms of sodium transport in plants-progresses and challenges. *Int. J. Mol. Sci.* 19, 647. doi: 10.3390/ijms19030647

Koul, A., Sharma, D., Kaul, S., and Dhar, M. K. (2019). Identification and in silico characterization of *cis*-acting elements of genes involved in carotenoid biosynthesis in tomato. *3 Biotech.* 9, 287. doi: 10.1007/s13205-019-1798-1

Kumar, P., Kumar, P., Sharma, D., Verma, S. K., Halterman, D., and Kumar, A. (2021). Genome-wide identification and expression profiling of basic leucine zipper transcription factors following abiotic stresses in potato (*Solanum tuberosum* l.). *PloS One* 16, e0247864. doi: 10.1371/journal.pone.0247864

Li, H., Guan, H., Zhuo, Q., Wang, Z., Li, S., Si, J., et al. (2020). Genome-wide characterization of the abscisic *acid-*, *stress-* and *ripening-induced* (*ASR*) gene family in wheat (*Triticum aestivum* l.). *Biol. Res.* 53, 23. doi: 10.1186/s40659-020-00291-6

Li, J., Li, X., Han, P., Liu, H., Gong, J., Zhou, W., et al. (2021). Genome-wide investigation of *bHLH* genes and expression analysis under different biotic and abiotic stresses in *Helianthus annuus* l. *Int. J. Biol. Macromol.* 189, 72–83. doi: 10.1016/j.ijbiomac

Li, Q., Qin, Y., Hu, X., Jin, L., Li, G., Gong, Z., et al. (2022). Physiology and gene expression analysis of potato (*Solanum tuberosum* l.) in salt stress. *Plants (Basel)* 11, 1565. doi: 10.3390/plants11121565

Li, Q., Qin, Y., Hu, X., Li, G., Ding, H., Xiong, X., et al. (2020). Transcriptome analysis uncovers the gene expression profile of salt-stressed potato (*Solanum tuberosum* l.). *Sci. Rep.* 10, 5411. doi: 10.1038/s41598-020-62057-0

Liang, Y., Wan, N., Cheng, Z., Mo, Y., Liu, B., Liu, H., et al. (2017). Whole-genome identification and expression pattern of the vicinal oxygen chelate family in rapeseed (*Brassica napus* l.). *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.00745

Luo, B., Guang, M., Yun, W., Ding, S., Ren, S., and Gao, H. (2022). Camellia sinensis chloroplast fluoride efflux gene *CsABCB9* is involved in the fluoride tolerance mechanism. *Int. J. Mol. Sci.* 23, 7756. doi: 10.3390/ijms23147756

Ma, R., Liu, W., Li, S., Zhu, X., Yang, J., Zhang, N., et al. (2021). Genome-wide identification, characterization and expression analysis of the *CIPK* gene family in potato (*Solanum tuberosum* l.) and the role of *StCIPK10* in response to drought and osmotic stress. *Int. J. Mol. Sci.* 22, 13535. doi: 10.3390/ijms222413535

Mo, F., Li, L., Zhang, C., Yang, C., Chen, G., Niu, Y., et al. (2022). Genome-wide analysis and expression profiling of the *Phenylalanine ammonia-lyase* gene family in *Solanum tuberosum*. *Int. J. Mol. Sci.* 23, 6833. doi: 10.3390/ijms23126833

Naureen, U., Khosa, A. N., Mukhtar, M. A., Nabi, F., Ahmed, N., and Saleem, M. (2023). Genetic biodiversity and posttranslational modifications of protease serine endopeptidase in different strains of *Sordaria fimicola*. *BioMed. Res. Int.* 2023, 2088988. doi: 10.1155/2023/2088988

Núñez-Ramírez, R., Sánchez-Barrena, M. J., Villalta, I., Vega, J. F., Pardo, J. M., Quintero, F. J., et al. (2012). Structural insights on the plant salt-overly-sensitive 1 (SOS1) Na(+)/H(+) antiporter. *J. Mol. Biol.* 424, 283–294. doi: 10.1016/j.jmb.2012.09.015

Nutan, K. K., Kumar, G., Singla-Pareek, S. L., and Pareek, A. (2018). A salt overly sensitive pathway member from *Brassica juncea* BjSOS3 can functionally complement *ΔAtsos3* in arabidopsis. *Curr. Genomics* 19, 60–69. doi: 10.2174/1389202918666170228133621

Pérez-Alonso, M. M., Ortiz-García, P., Moya-Cuevas, J., and Pollmann, S. (2021). Mass spectrometric monitoring of plant hormone cross talk during biotic stress responses in potato (*Solanum tuberosum* l.). *Methods Mol. Biol.* 2354, 143–154. doi: 10.1007/978-1-0716-1609-3_7

Rogozin, I. B., Wolf, Y. I., Sorokin, A. V., Mirkin, B. G., and Koonin, E. V. (2003). Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr. Biol.* 13, 1512–1517. doi: 10.1016/s0960-9822(03)00558-x

Rolly, N. K., Imran, Q. M., Lee, I. J., and Yun, B. W. (2020). Salinity stress-mediated suppression of expression of salt overly sensitive signaling pathway genes suggests negative regulation by *AtbZIP62* transcription factor in *Arabidopsis thaliana*. *Int. J. Mol. Sci.* 21, 1726. doi: 10.3390/ijms21051726

Sharma, B., Saxena, H., and Negi, H. (2021). Genome-wide analysis of HECT E3 ubiquitin ligase gene family in *Solanum lycopersicum*. *Sci. Rep.* 11, 15891. doi: 10.1038/s41598-021-95436-2

Shi, H., Ishitani, M., Kim, C., and Zhu, J. K. (2000). The *Arabidopsis thaliana* salt tolerance gene *SOS1* encodes a putative Na^+/H^+ antiporter. *Proc. Natl. Acad. Sci. U. S. A.* 97, 6896–6901. doi: 10.1073/pnas.120170197

Shi, H., Lee, B. H., Wu, S. J., and Zhu, J. K. (2003). Overexpression of a plasma membrane Na^+/H^+ antiporter gene improves salt tolerance in *Arabidopsis thaliana*. *Nat. Biotechnol.* 21, 81–85. doi: 10.1038/nbt766

Shi, H., Quintero, F. J., Pardo, J. M., and Zhu, J. K. (2002). The putative plasma membrane Na(+)/H(+) antiporter SOS1 controls long-distance na(+) transport in plants. *Plant Cell.* 14, 465–477. doi: 10.1105/tpc.010371

Song, A., Lu, J., Jiang, J., Chen, S., Guan, Z., Fang, W., et al. (2012). Isolation and characterisation of *Chrysanthemum crassum SOS1*, encoding a putative plasma membrane Na(+)/H(+) antiporter. *Plant Biol. (Stuttg).* 14, 706–713. doi: 10.1111/j.1438-8677

Sun, H., Ren, M., and Zhang, J. (2022). Genome-wide identification and expression analysis of fibrillin (*FBN*) gene family in tomato (*Solanum lycopersicum* l.). *PeerJ* 10, e13414. doi: 10.7717/peerj.13414

Takahashi, S., Seki, M., Ishida, J., Satou, M., Sakurai, T., Narusaka, M., et al. (2004). Monitoring the expression profiles of genes induced by hyperosmotic, high salinity, and oxidative stress and abscisic acid treatment in arabidopsis cell culture using a full-length cDNA microarray. *Plant Mol. Biol.* 56, 29–55. doi: 10.1007/s11103-004-2200-0

Takeuchi, A., Akatsu, Y., Asahi, T., Okubo, Y., Ohnuma, M., Teramura, H., et al. (2022). Procedure for the efficient acquisition of progeny seeds from crossed potato plants grafted onto tomato. *Plant Biotechnol. (Tokyo).* 39, 195–197. doi: 10.5511/plantbiotechnology.21.1119a

Tester, M., and Davenport, R. (2003). Na^+ tolerance and na^+ transport in higher plants. *Ann. Bot.* 91, 503–527. doi: 10.1093/aob/mcg058

Wang, T., Gao, X., Chen, S., Li, D., Chen, S., Xie, M., et al. (2021). Genome-wide identification and expression analysis of ethylene responsive factor family transcription factors in *Juglans regia*. *PeerJ* 9, e12429. doi: 10.7717/peerj.12429

Wang, Z., Hong, Y., Li, Y., Shi, H., Yao, J., Liu, X., et al. (2021). Natural variations in *SlSOS1* contribute to the loss of salt tolerance during tomato domestication. *Plant Biotechnol. J.* 19, 20–22. doi: 10.1111/pbi.13443

Wang, J., Zhang, Y., Xu, N., Zhang, H., Fan, Y., Rui, C., et al. (2021). Genome-wide identification of *CK* gene family suggests functional expression pattern against Cd(2+) stress in *Gossypium hirsutum* l. *Int. J. Biol. Macromol.* 188, 272–282. doi: 10.1016/j.ijbiomac

Wu, Y., Ding, N., Zhao, X., Zhao, M., Chang, Z., Liu, J., et al. (2007). Molecular characterization of *PeSOS1*: the putative Na(+)/H (+) antiporter of *Populus euphratica*. *Plant Mol. Biol.* 65, 1–11. doi: 10.1007/s11103-007-9170-y

Wu, S. J., Ding, L., and Zhu, J. K. (1996). SOS1, a genetic locus essential for salt tolerance and potassium acquisition. *Plant Cell.* 8, 617–627. doi: 10.1105/tpc.8.4.617

Xiang, X. H., Wu, X. R., Chao, J. T., Yang, M. L., Yang, F., Chen, G., et al. (2016). Genome-wide identification and expression analysis of the *WRKY* gene family in common tobacco (*Nicotiana tabacum* l.). *Yi Chuan.* 38, 840–856. doi: 10.16288/j.yczz

Xie, Q., Zhou, Y., and Jiang, X. (2022). Structure, function, and regulation of the plasma membrane Na(+)/H(+) antiporter salt overly sensitive 1 in plants. *Front. Plant Sci.* 13. doi: 10.3389/fpls

Xu, H., Jiang, X., Zhan, K., Cheng, X., Chen, X., Pardo, J. M., et al. (2008). Functional characterization of a wheat plasma membrane Na^+/H^+ antiporter in yeast. *Arch. Biochem. Biophys.* 473, 8–15. doi: 10.1016/j.abb

Yang, Q., Chen, Z. Z., Zhou, X. F., Yin, H. B., Li, X., Xin, X. F., et al. (2009). Overexpression of *SOS* (*Salt overly sensitive*) genes increases salt tolerance in transgenic *Arabidopsis*. *Mol. Plant* 2, 22–31. doi: 10.1093/mp/ssn058

Yang, G., Xu, C., Varjani, S., Zhou, Y., Wc Wong, J., and Duan, G. (2022). Metagenomic insights into improving mechanisms of Fe(0) nanoparticles on volatile fatty acids production from potato peel waste anaerobic fermentation. *Bioresour. Technol.* 361, 127703. doi: 10.1016/j.biortech.2022.127703

Yang, X., Yuan, J., Luo, W., Qin, M., Yang, J., Wu, W., et al. (2020). Genome-wide identification and expression analysis of the class III peroxidase gene family in potato (*Solanum tuberosum* l.). *Front. Genet.* 11. doi: 10.3389/fgene.2020.593577

You, X., Nasrullah, Wang, D., Mei, Y., Bi, J., Liu, S., et al. (2022). N(7) -SSPP fusion gene improves salt stress tolerance in transgenic arabidopsis and soybean through ROS scavenging. *Plant Cell Environ.* 45, 2794–2809. doi: 10.1111/pce.14392

Yu, R. M., Suo, Y. Y., Yang, R., Chang, Y. N., Tian, T., Song, Y. J., et al. (2021). *StMBF1c* positively regulates disease resistance to *Ralstonia solanacearum via* it's primary and secondary upregulation combining expression of *StTPS5* and resistance marker genes in potato. *Plant Sci.* 307, 110877. doi: 10.1016/j.plantsci

Zhang, M., Cao, J., Zhang, T., Xu, T., Yang, L., Li, X., et al. (2022). A putative plasma membrane Na(+)/H(+) antiporter *GmSOS1* is critical for salt stress tolerance in *Glycine max*. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.870695

Zhang, Q., Hou, S., Sun, Z., Chen, J., Meng, J., Liang, D., et al. (2021). Genome-wide identification and analysis of the *MADS-box* gene family in *Theobroma cacao*. *Genes (Basel)* 12. doi: 10.3389/fpls.2022.870695

Zhang, C., Liu, S., Liu, D., Guo, F., Yang, Y., Dong, T., et al. (2022). Genome-wide survey and expression analysis of GRAS transcription factor family in sweetpotato provides insights into their potential roles in stress response. *BMC Plant Biol.* 22, 232. doi: 10.1186/s12870-022-03618-5

Zhao, C., William, D., and Sandhu, D. (2021). Isolation and characterization of salt overly sensitive family genes in spinach. *Physiol. Plant* 171, 520–532. doi: 10.1111/ppl.13125

Zhao, S., Zhang, Q., Liu, M., Zhou, H., Ma, C., and Wang, X. X. X. P. (2021). Regulation of plant responses to salt stress. *Int. J. Mol. Sci.* 22, 4609. doi: 10.3390/ijms22094609

Zhou, X., Li, J., Wang, Y., Liang, X., Zhang, M., Lu, M., et al. (2022). The classical SOS pathway confers natural variation of salt tolerance in maize. *New Phytol.* 236, 479–494. doi: 10.1111/nph.18278

Zhou, Y., Yin, X., Duan, R., Hao, G., Guo, J., and Jiang, X. (2015). *SpAHA1* and *SpSOS1* coordinate in transgenic yeast to improve salt tolerance. *PloS One* 10, e0137447. doi: 10.1371/journal.pone.0137447

Zhu, L., Li, M., Huo, J., Lian, Z., Liu, Y., Lu, L., et al. (2021). Overexpression of *NtSOS2* from halophyte plant *N. tangutorum* enhances tolerance to salt stress in arabidopsis. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.716855

Zhu, J. K., Liu, J., and Xiong, L. (1998). Genetic analysis of salt tolerance in arabidopsis. evidence for a critical role of potassium nutrition. *Plant Cell.* 10, 1181–1191. doi: 10.1105/tpc.10.7.1181

Zhu, X., Wang, F., Li, S., Feng, Y., Yang, J., Zhang, N., et al. (2022). Calcium-dependent protein kinase 28 maintains potato photosynthesis and its tolerance under water deficiency and osmotic stress. *Int. J. Mol. Sci.* 23, 8795. doi: 10.3390/ijms23158795

frontiers | Frontiers in Plant Science

TYPE Perspective
PUBLISHED 18 September 2023
DOI 10.3389/fpls.2023.1232466

OPEN ACCESS

EDITED BY
Xueqiang Wang,
Zhejiang University, China

REVIEWED BY
Sen Yang,
Henan Agricultural University, China
Weiping Mo,
Chinese Academy of Sciences (CAS), China
Xiaodong Zheng,
Qingdao Agricultural University, China

*CORRESPONDENCE
Jirong Zhao
zjr520999@126.com
Xiaozeng Yang
yangxz@sRNAworld.com

†These authors have contributed equally to this work

RECEIVED 31 May 2023
ACCEPTED 28 August 2023
PUBLISHED 18 September 2023

CITATION
Shen F, Hu C, Huang X, He H, Yang D, Zhao J and Yang X (2023) Advances in alternative splicing identification: deep learning and pantranscriptome. *Front. Plant Sci.* 14:1232466. doi: 10.3389/fpls.2023.1232466

Advances in alternative splicing identification: deep learning and pantranscriptome

Fei Shen[1†], Chenyang Hu[1,2†], Xin Huang[1], Hao He[1], Deng Yang[1], Jirong Zhao[2*] and Xiaozeng Yang[1*]

[1]Institute of Biotechnology, Beijing Academy of Agriculture and Forestry Sciences, Beijing, China, [2]Shanxi Key Lab of Chinese Jujube, College of Life Science, Yan'an University, Yan'an, Shanxi, China

In plants, alternative splicing is a crucial mechanism for regulating gene expression at the post-transcriptional level, which leads to diverse proteins by generating multiple mature mRNA isoforms and diversify the gene regulation. Due to the complexity and variability of this process, accurate identification of splicing events is a vital step in studying alternative splicing. This article presents the application of alternative splicing algorithms with or without reference genomes in plants, as well as the integration of advanced deep learning techniques for improved detection accuracy. In addition, we also discuss alternative splicing studies in the pan-genomic background and the usefulness of integrated strategies for fully profiling alternative splicing.

KEYWORDS

alternative splicing, RNA-seq, Iso-seq, detection algorithm, deep learning, pantranscriptome

1 The alternative splicing event in plants

1.1 Definition and classification of alternative splicing

Alternative splicing (AS) is a crucial mechanism for gene expression regulation, which entails the selection of different splice sites, removal of introns, and subsequent combine various exons to generate multiple mature mRNA isoforms in plants (Barbazuk et al., 2008). Plants generate extensive AS to increase the diversity of their transcriptomes, especially faced with complex environmental changes (Nilsen and Graveley, 2010; Szakonyi and Duque, 2018; Jia et al., 2022; Lam et al., 2022). There are several types of AS events in plants, including exon skipping (ES), intron retention (IR), alternative 5′ splice site (AE5′), alternative 3′ splice site (AE3′), mutually exclusive alternate exon splicing (MEE), alternative first exon (AFE), and alternative last exon (ALE) (Filichkin et al., 2010; E et al., 2013; Chen et al., 2020b). Among them, IR is the predominant type (Syed et al., 2012; Zhu et al., 2017).

1.2 Generation of alternative splicing

The spliceosome is a large ribonucleoprotein complex that interacts with various trans-acting factors and is involved in controlling AS in plants (Will and Luhrmann, 2010; Ule and Blencowe, 2019; Liu et al., 2021; Jia et al., 2022). The U2 and U12 spliceosomal RNA are the focus RNAof most studies on the spliceosome (Hartmann, 2007; Reddy et al., 2012; Zhang et al., 2020). The spliceosome splices intron-exon junction sites, which are characterized by the conserved 5′-GT sequence and AG-3′ sequence. Non-snRNA (small nuclear RNA) splicing factors, such as serine/arginine-rich proteins and heterogeneous ribonucleoproteins, are known to facilitate the localization of splicing enhancers and inhibitors, thereby regulating the selection of splice sites (Geuens et al., 2016; Jeong, 2017; Chen et al., 2020a). Pre-mRNA undergoes two consecutive reactions to complete the splicing process: (i) introns form a unique chain-like structure; (ii) intron are rapidly degraded as a chain-like structure, and exons at the left and right ends are joined by phosphodiester bonds, achieving intron excision and exon joining (Black, 2003; Wan et al., 2019).

1.3 Functionality of alternative splicing

AS plays a crucial role in regulating plant growth, development and responses to abiotic stresses. AS generally occurs during seed germination, plant growth, and flowering stages. For example, AS of the *NAC transcription factor 109* (*NACTF109*) during maize embryo development regulates seed dormancy by controlling ABA content in seeds (Thatcher et al., 2016). *FLOWERING LOCUS C (FLC)* is an important repressor of flowering in Arabidopsis (Andersson et al., 2008; Sharma et al., 2020), and *AtU2AF65b* is a splicing factor involved in ABA-mediated regulation of flowering time in *Arabidopsis* by splicing *FLC* pre-mRNA (Xiong et al., 2019; Lee et al., 2023). JASMONATE ZIM-DOMAIN (JAZ) is a key regulators of jasmonate (JA) signaling in plants (Yan et al., 2009). In *Arabidopsis*, the JAZ protein binds to the transcription factor MYC2 and inhibits JA signaling during quiescence. Binding to the hormone receptor *CORONATINE INSENSITIVE 1 (COI1)* upon hormone induction leads to degradation of JAZ. This degradation allows *AtMED25* to activate MYC2 and promote JA signaling. *AtMED25* regulates JAZ gene replacement splicing by recruiting splicing factors PRP39a and PRP40a, preventing excessive desensitization of JA signaling mediated by JAZ splice variants (Pauwels and Goossens, 2011; Wu et al., 2020). In rice (*Oryza Sativa*), *OsDREB2* activates the expression of downstream genes involved in heat shock stress response and tolerance. The direct homolog of *OsDREB2B* enhances the ability of plants to cope with drought stress through AS by directly producing *OsDREB2B2* by splicing I1, E2, and I2 at once under drought stress (Matsukura et al., 2010).

Different gene variants affecting alternative splicing (AS) have been observed in numerous functional gene studies. These variants play a crucial role in phenotypic changes. For instance, in poplar (*Populus tomentosa*), age-dependent AS triggers an aberrant splicing event in the pre-mRNA encoding *PtRD26*. This event leads to the production of a truncated protein, PtRD26IR, which acts as a dominant negative regulator of senescence by interacting with multiple senescence-associated NAC family transcription factors, inhibiting their DNA-binding activity (Wang et al., 2021). In *Arabidopsis*, the RNA-binding splicing factor SUPPRESSOR-OF-WHITE-APRICOT/SURP RNA-BINDING DOMAIN-CONTAINING PROTEIN1 (SWAP1) interacts with the splicing factor complexes SPLICING FACTOR FOR PHYTOCHROME SIGNALING (SFPS) and REDUCED RED LIGHT RESPONSES IN CRY1CRY2 BACKGROUND 1 (RRC1). These complexes regulate pre-mRNA splicing and induce alterations in photo morphology (Kathare et al., 2022). In bread wheat (*Triticum aestivum*), two variable splicers, Pm4b_V1 and Pm4b_V2, of the powdery mildew resistance gene Pm4b interact. In brief, Pm4b_V2 enhances wheat disease resistance by recruiting Pm4b_V1 from the cytoplasm to the endoplasmic reticulum (ER) by forming an ER-related complex (Sanchez-Martin et al., 2021).

2 Detection of alternative splicing using transcriptome sequencing

The continuous advancement of RNA sequencing (next generation sequencing) and long-read isoform sequencing (Iso-seq) has significantly enhanced our ability to study alternative splicing comprehensively. Two primary computational approaches have been employed to investigate splicing diversity using RNA-seq data.

Transcript reconstruction methods: These approaches focus on inferring isoform usage frequency by utilizing probabilistic models to reconstruct each isoform based on the read distribution mapped to a specific gene. Typical software packages include Cufflinks (Trapnell et al., 2010), StringTie (Pertea et al., 2015), MISO (Yarden et al., 2010), SpliceGrapher (Mark et al., 2012). Indeed, transcriptome reconstruction is an exceptionally challenging problem in the field of bioinformatics and computational biology (Estefania et al., 2021). Single-molecule long-read sequencing technology has emerged as a valuable tool in transcriptome sequencing due to its ability to generate long reads with high throughput. The utilization of Iso-seq has become a preferred approach for sequencing more comprehensive and full-length transcriptomes, enabling the prediction and validation of gene models with greater accuracy and completeness. By producing long reads that can span entire transcript isoforms, Iso-seq overcomes some of the challenges associated with transcriptome reconstruction, such as accurately detecting complex splicing events and resolving alternative isoforms that may be missed by short-read sequencing. However, they are not suitable to pinpoint splicing events but whole sequences of transcripts. For instance, degraded and immature RNA as well as DNA fragments in the RNA samples can be erroneously identified as novel genes and transcripts in the Iso-seq data. In practice, tools such as TAMA software (Sim et al., 2020) could determine splice junctions and transcription start and

end sites accurately. Unfortunately, the current cost of third-generation sequencing is high, and the detection of all transcripts may be limited by the depth of sequencing and the number of samples. Therefore, the development of tools combining RNA-seq and Iso-seq could effectively solve these problems. Regrettably, no mature tools have been released so far.

The second computational approach involves utilizing junction and/or exon information to infer, annotate, and identify novel splicing events (Table 1). Several methods, such as rMATS (Shen et al., 2014), MAJIQ (Vaquero-Garcia et al., 2016), and LeafCutter (Li et al., 2018), utilize junction information to identify these splicing events. On the other hand, DEXSeq (Anders and Huber, 2010) specifically focuses on analyzing the differential usage of exons between different experimental conditions. Two main methodologies are commonly used to quantify alternative splicing (AS) events: the percent spliced-in (PSI) and the splicing index (SI). PSI provides an estimate of the relative usage of each alternative pathway of an AS event. In contrast, the splicing index (SI) measures the relative signal or coverage of an exon or a junction compared to the entire gene.

In addition to detecting different AS events, it is important to directly compare direct AS differences across samples. The Cuffdiff (Cufflinks) (Trapnell et al., 2010) package can test for differential splicing between isoforms in different samples. In addition, CASH (Wu et al., 2018), DEXseq (Anders and Huber, 2010), DiffSplice (Hu et al., 2013), Gess (Ye et al., 2014), rMATS (Shen et al., 2014), SplAdder (Kahles et al., 2016) and other software can use different algorithms to detect different AS events between different samples. But unfortunately, none of these AS analysis software takes into account the existence of variants. Direct analysis at the allele-aware level cannot be achieved. Allele-aware AS analysis software is of great significance in analyzing the causes of variable AS, such as comparing the differences in AS between different genomic haplotypes.

3 Deep learning based alternative splicing study

Several models have been developed for predicting and identifying alternative splicing events combining deep learning approaches (Table 2). For example, DeepASmRNA is a convolutional neural network (CNN) model capable of identifying alternative splicing events with over 90% accuracy (Cao et al., 2022). The Deep Splicing Code model uses raw RNA sequences to classify exons based on their alternative splicing behavior and performs well in identifying splice sites and motifs (Louadi et al., 2019). The deep-learning model AbSplice predicts anomalous splicing, increasing the accuracy of traditional DNA-based anomalous splicing prediction to 48% at a 20% call rate. Furthermore, integrating RNA-Seq raises the accuracy to 60% (Wagner et al., 2023). Additionally, the deep learning based computational framework called DARTS (deep-learning augmented RNA-seq analysis of transcript splicing) utilizes deep neural networks and Bayesian hypothesis testing for identifying exons based on their sequence characteristics, attaining a more than 95% accuracy rate in recognizing alternative splicing (Zhang et al., 2019). Finally the hybrid model combining CNN, recurrent neural network, and Long Short-Term Memory (LSTM) network has a splice locus identification accuracy of 96% (Nazari et al., 2019). In summary, deep learning models for alternative splicing detection have high detection accuracy, event classification, and splice site identification.

4 Pan-genomics-based alternative splicing study

During the lengthy process of evolution, each plant develops unique genetic influenced by geographical and environmental factors. Consequently, the genome of a single plant can no longer fully represent all the genetic information of a species, and pan-genome of a species encompasses all the genetic information of a species and captures most of its genetic diversity and can help to explore plant genome evolution (Alonge et al., 2020; Liu et al., 2020; Long et al., 2021; Qin et al., 2021), crop molecular breeding (Tao et al., 2019; Yu et al., 2021b), and construction of genotype databases (Gui et al., 2020; Peng et al., 2020; Song et al., 2021). Similarly, the pan-transcriptome is a recalling concept of the pan-genome, which reflects the set of all transcripts of a species or an organism. The aggregation group integrating AS events from different genomes in a species can better represent the whole transcriptomes of the species and can better promote the study of AS biological processes. A tool RPVG (Sibbesen et al., 2023) was released to construct spliced pangenome graphs, to map RNA sequencing data to these graphs, and to perform haplotype-aware expression quantification of transcripts in a pantranscriptome.

5 Conclusions and prospects

The recent the developments of third-generation sequencing technologies and detection algorithms have led to significant advances in the study of alternative splicing. While much has been identified regarding the mechanism of alternative splicing generation and some of its functions, challenges remain in the detection of alternative splicing events without reference genomes. Using the third-generation reconstruction technology can reconstruct the AS version very well, but cannot directly determine the coordinates of the AS sites. Therefore, the algorithm combined with the second generation and the third generation sequencing technologies can solve most of such problems well. Compared with state-of-the-art methods, deep learning-based models have been used to improve the detection accuracy and the number of splicing events. Allele-aware AS analysis software is of great significance in analyzing the causes of variable AS, such as comparing the differences in AS between different genomic haplotypes. In the pan-genome context, it is of great significance to integrate different transcript information from

TABLE 1 Algorithms for the identification of Alternative Splicing events.

Algorithm	E.	S.	V.	PSI	D.	Information used for quantification	References
Aspli	✓	✓	×	✓	×	Only junctions	(Mancini et al., 2021)
Leafcutter	×	✓	✓	✓	×	Only junctions	(Li et al., 2018)
CASH	✓	✓	×	~	✓	Exons and junctions	(Wu et al., 2018)
SplAdder	✓	✓	×	✓	✓	Exons and junctions	(Kahles et al., 2016)
SGSeq	✓	×	✓	✓	×	Exons and junctions	(Xing et al., 2016)
MAJIQ+VOILA	✓	✓	✓	✓	×	Only junctions	(Vaquero-Garcia et al., 2016)
EventPointer	✓	✓	✓	✓	×	Exons and junctions	(Romero et al., 2016)
SUPPA	✓	✓	✓	✓	×	Expression of isoforms involved in event	(Alamancos et al., 2015)
SplicingTypesAnno	✓	✓	✓	✓	×	Exons and junctions	(Sun et al., 2015)
SplicingExpress	✓	✓	✓	×	×	Expression of isoforms involved in event	(Kroll et al., 2015)
SplicePie	~	✓	×	✓	×	Exons and junctions	(Pulyakhina et al., 2015)
Vast-Tools	✓	✓	✓	✓	×	Exons and junctions	(Irimia et al., 2014)
SpliceR	✓	×	×	×	✓	Expression of isoforms involved in event	(Vitting-Seerup et al., 2014)
rMATS	✓	✓	×	✓	✓	Exons and junctions	(Shen et al., 2014)
Gess	✓	✓	×	✓	✓	Only exons	(Ye et al., 2014)
SplicingCompass	×	✓	✓	×	×	Exons and junctions	(Aschoff et al., 2013)
ASprofile	✓	×	×	×	×	Expression of isoforms involved in event	(Florea et al., 2013)
DSGseq	×	✓	×	×	✓	Only exons	(Wang et al., 2013)
DiffSplice	✓	✓	✓	✓	✓	Exons and junctions	(Hu et al., 2013)
SpliceSeq	✓	✓	×	✓	×	Exons and junctions	(Ryan et al., 2012)
SpliceTrap	✓	✓	×	✓	×	Only exons	(Wu et al., 2011)
JuncBASE	✓	✓	×	×	×	Only junctions	(Brooks et al., 2011)
DEXseq	×	✓	×	×	✓	Only exons	(Anders and Huber, 2010)
AltAnalyze	✓	✓	✓	×	×	Exons and junctions	(Emig et al., 2010)

*There is not a peer-reviewed reference for this algorithm. E, event classification; S, this method provides statistics; V, visualization; PSI, whether the PSI is returned; D, Whether to make discrepancy detection.
✓, this algorithm provides this result; ×, this algorithm does not provide this result; ~, this algorithm does not provide this result, but it is easily computed.

TABLE 2 Deep learning algorithms for predicting and recognizing Alternative Splicing events.

Algorithm	Neural Network	References
AbSplice	Deep Neural Network	(Wagner et al., 2023)
CI-SpliceAI	Deep Neural Network	(Strauch et al., 2022)
Deep Splicer	Convolutional Neural Network	(Fernandez-Castillo et al., 2022)
DeepASmRNA	Convolutional Neural Network	(Cao et al., 2022)
DeepIsoFun	Deep Neural Network	(Yu et al., 2021a)
DMIL-IsoFun	Convolutional Neural Network	(Yu et al., 2021a)
LSTM_Splice	Long Short-Term Memory Network	(Regan et al., 2021)
SQUIRLS	Deep Neural Network	(Danis et al., 2021)
Deep SHAP	Deep Neural Network	(Jha et al., 2020)

(Continued)

TABLE 2 Continued

Algorithm	Neural Network	References
ESPRNN	Recurrent Neural Network	(Lee et al., 2020)
Splice2Deep	Convolutional Neural Network	(Albaradei et al., 2020)
DARTS	Deep Neural Networks	(Zhang et al., 2019)
Deep Splicing Code	Convolutional Neural Network	(Louadi et al., 2019)
DIFFUSE	Deep Neural Network	(Chen et al., 2019)
MMSplice	Deep Neural Network	(Cheng et al., 2019)
SpliceAI	Deep Neural Network	(Jaganathan et al., 2019)
COSSMO	Long Short-Term Memory Network	(Bretschneider et al., 2018)
DeepSplice	Convolutional Neural Network	(Zhang et al., 2018)
SpliceRover	Convolutional Neural Network	(Zuallaert et al., 2018)
DeepCode	Deep Neural Network	(Xu et al., 2017)

different samples. Exploring the relationship between different alternative splicing events and mutations detected by different algorithms is of great significance for mining the influence of mutations on AS events.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding authors.

Author contributions

XY, JZ designed the study and methodology. FS, CH, XH, HH and JZ wrote the manuscript draft. XY performed writing-review, editing and supervision. All authors contributed to the article and approved the submitted version.

Funding

The authors declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Natural Science Foundation of China (32102339), Beijing Academy of Agriculture and Forestry Sciences (YXQN202203, QNJJ202106).

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

Alamancos, G. P., Pages, A., Trincado, J. L., Bellora, N., and Eyras, E. (2015). Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA* 21 (9), 1521–1531. doi: 10.1261/rna.051557.115

Albaradei, S., Magana-Mora, A., Thafar, M., Uludag, M., Bajic, V. B., Gojobori, T., et al. (2020). Splice2Deep: An ensemble of deep convolutional neural networks for improved splice site prediction in genomic DNA. *Gene* 763, 100035. doi: 10.1016/j.gene.2020.100035

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., et al. (2020). Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* 182 (1), 145–161. doi: 10.1016/j.cell.2020.05.021

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11 (10), R106. doi: 10.1186/gb-2010-11-10-r106

Andersson, C. R., Helliwell, C. A., Bagnall, D. J., Hughes, T. P., Finnegan, E. J., Peacock, W. J., et al. (2008). The FLX gene of Arabidopsis is required for FRI-dependent activation of FLC expression. *Plant Cell Physiol.* 49 (2), 191–200. doi: 10.1093/pcp/pcm176

Aschoff, M., Hotz-Wagenblatt, A., Glatting, K. H., Fischer, M., Eils, R., and Konig, R. (2013). SplicingCompass: differential splicing detection using RNA-seq data. *Bioinformatics* 29 (9), 1141–1148. doi: 10.1093/bioinformatics/btt101

Barbazuk, W. B., Fu, Y., and McGinnis, K. M. (2008). Genome-wide analyses of alternative splicing in plants: opportunities and challenges. *Genome Res.* 18 (9), 1381–1392. doi: 10.1101/gr.053678.106

Black, D. L. (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* 72, 291–336. doi: 10.1146/annurev.biochem.72.121801.161720

Bretschneider, H., Gandhi, S., Deshwar, A. G., Zuberi, K., and Frey, B. J. (2018). COSSMO: predicting competitive alternative splice site selection using deep learning. *Bioinformatics* 34 (13), i429–i437. doi: 10.1093/bioinformatics/bty244

Brooks, A. N., Yang, L., Duff, M. O., Hansen, K. D., Park, J. W., Dudoit, S., et al. (2011). Conservation of an RNA regulatory map between Drosophila and mammals. *Genome Res.* 21 (2), 193–202. doi: 10.1101/gr.108662.110

Cao, L., Zhang, Q., Song, H., Lin, K., and Pang, E. (2022). DeepASmRNA: Reference-free prediction of alternative splicing events with a scalable and interpretable deep learning model. *iScience* 25 (11), 105345. doi: 10.1016/j.isci.2022.105345

Chen, H., Shaw, D., Zeng, J., Bu, D., and Jiang, T. (2019). DIFFUSE: predicting isoform functions from sequences and expression profiles via deep learning. *Bioinformatics* 35 (14), i284–i294. doi: 10.1093/bioinformatics/btz367

Chen, M. X., Zhang, K. L., Gao, B., Yang, J. F., Tian, Y., Das, D., et al. (2020a). Phylogenetic comparison of 5' splice site determination in central spliceosomal proteins of the U1-70K gene family, in response to developmental cues and stress conditions. *Plant J.* 103 (1), 357–378. doi: 10.1111/tpj.14735

Chen, M. X., Zhang, K. L., Zhang, M., Das, D., Fang, Y. M., Dai, L., et al. (2020b). Alternative splicing and its regulatory role in woody plants. *Tree Physiol.* 40 (11), 1475–1486. doi: 10.1093/treephys/tpaa076

Cheng, J., Nguyen, T. Y. D., Cygan, K. J., Celik, M. H., Fairbrother, W. G., Avsec, Z., et al. (2019). MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol.* 20 (1), 48. doi: 10.1186/s13059-019-1653-z

Danis, D., Jacobsen, J. O. B., Carmody, L. C., Gargano, M. A., McMurry, J. A., Hegde, A., et al. (2021). Interpretable prioritization of splice variants in diagnostic next-generation sequencing. *Am. J. Hum. Genet.* 108 (9), 1564–1577. doi: 10.1016/j.ajhg.2021.06.014

E, Z., Wang, L., and Zhou, J. (2013). Splicing and alternative splicing in rice and humans. *BMB Rep.* 46 (9), 439–447. doi: 10.5483/BMBRep.2013.46.9.161

Emig, D., Salomonis, N., Baumbach, J., Lengauer, T., Conklin, B. R., and Albrecht, M. (2010). AltAnalyze and DomainGraph: analyzing and visualizing exon expression data. *Nucleic Acids Res.* 38 (Web Server issue), W755–W762. doi: 10.1093/nar/gkq405

Estefania, M., Andres, R., Javier, I., Marcelo, Y., and Ariel, C. (2021). ASpli: Integrative analysis of splicing landscapes through RNA-Seq assays. *Bioinformatics* 37 (17), 2609–2616. doi: 10.1093/bioinformatics/btab141

Fernandez-Castillo, E., Barbosa-Santillan, L. I., Falcon-Morales, L., and Sanchez-Escobar, J. J. (2022). Deep splicer: A CNN model for splice site prediction in genetic sequences. *Genes (Basel)* 13 (5), 907. doi: 10.3390/genes13050907

Filichkin, S. A., Priest, H. D., Givan, S. A., Shen, R., Bryant, D. W., Fox, S. E., et al. (2010). Genome-wide mapping of alternative splicing in Arabidopsis thaliana. *Genome Res.* 20 (1), 45–58. doi: 10.1101/gr.093302.109

Florea, L., Song, L., and Salzberg, S. L. (2013). Thousands of exon skipping events differentiate among splicing patterns in sixteen human tissues. *F1000Res* 2, 188. doi: 10.12688/f1000research.2-188.v2

Geuens, T., Bouhy, D., and Timmerman, V. (2016). The hnRNP family: insights into their role in health and disease. *Hum. Genet.* 135 (8), 851–867. doi: 10.1007/s00439-016-1683-5

Gui, S., Yang, L., Li, J., Luo, J., Xu, X., Yuan, J., et al. (2020). ZEAMAP, a comprehensive database adapted to the maize multi-omics era. *iScience* 23 (6), 101241. doi: 10.1016/j.isci.2020.101241

Hartmann, T. (2007). From waste products to ecochemicals: fifty years research of plant secondary metabolism. *Phytochemistry* 68 (22-24), 2831–2846. doi: 10.1016/j.phytochem.2007.09.017

Hu, Y., Huang, Y., Du, Y., Orellana, C. F., Singh, D., Johnson, A. R., et al. (2013). DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. *Nucleic Acids Res.* 41 (2), e39. doi: 10.1093/nar/gks1026

Irimia, M., Weatheritt, R. J., Ellis, J. D., Parikshak, N. N., Gonatopoulos-Pournatzis, T., Babor, M., et al. (2014). A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* 159 (7), 1511–1523. doi: 10.1016/j.cell.2014.11.035

Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., et al. (2019). Predicting splicing from primary sequence with deep learning. *Cell* 176 (3), 535–548 e524. doi: 10.1016/j.cell.2018.12.015

Jeong, S. (2017). SR proteins: binders, regulators, and connectors of RNA. *Mol. Cells* 40 (1), 1–9. doi: 10.14348/molcells.2017.2319

Jha, A., Aicher, J. ,. K., Gazzara, M. ,. R., Singh, D., and Barash, Y. (2020). Enhanced Integrated Gradients: improving interpretability of deep learning models using splicing codes as a case study. *Genome Biol.* 21 (1), 149. doi: 10.1186/s13059-020-02055-7

Jia, Z. C., Yang, X., Hou, X. X., Nie, Y. X., and Wu, J. (2022). The importance of a genome-wide association analysis in the study of alternative splicing mutations in plants with a special focus on maize. *Int. J. Mol. Sci.* 23 (8), 4201. doi: 10.3390/ijms23084201

Kahles, A., Ong, C. S., Zhong, Y., and Rätsch, G. (2016). SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data. *Bioinformatics* 32 (12), 1840–1847. doi: 10.1093/bioinformatics/btw076

Kathare, P. K., Xin, R., Ganesan, A. S., June, V. M., Reddy, A. S. N., and Huq, E. (2022). SWAP1-SFPS-RRC1 splicing factor complex modulates pre-mRNA splicing to promote photomorphogenesis in Arabidopsis. *Proc. Natl. Acad. Sci. U.S.A.* 119 (44), e2214565119. doi: 10.1073/pnas.2214565119

Kroll, J. E., Kim, J., Ohno-MaChado, L., and de Souza, S. J. (2015). Splicing Express: a software suite for alternative splicing analysis using next-generation sequencing data. *PeerJ* 3, e1419. doi: 10.7717/peerj.1419

Lam, P. Y., Wang, L., Lo, C., and Zhu, F. Y. (2022). Alternative splicing and its roles in plant metabolism. *Int. J. Mol. Sci.* 23 (13), 7355. doi: 10.3390/ijms23137355

Lee, H. T., Park, H. Y., Lee, K. C., Lee, J. H., and Kim, J. K. (2023). Two arabidopsis splicing factors, U2AF65a and U2AF65b, differentially control flowering time by modulating the expression or alternative splicing of a subset of FLC upstream regulators. *Plants (Basel)* 12 (8), 1655. doi: 10.3390/plants12081655

Lee, D., Zhang, J., Liu, J., and Gerstein, M. (2020). Epigenome-based splicing prediction using a recurrent neural network. *PloS Comput. Biol.* 16 (6), e1008006. doi: 10.1371/journal.pcbi.1008006

Li, Y. I., Knowles, D. A., Humphrey, J., Barbeira, A. N., Dickinson, S. P., Im, H. K., et al. (2018). Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* 50 (1), 151–158. doi: 10.1038/s41588-017-0004-9

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., et al. (2020). Pan-genome of wild and cultivated soybeans. *Cell* 182 (1), 162–176. doi: 10.1016/j.cell.2020.05.023

Liu, L., Tang, Z., Liu, F., Mao, F., Yujuan, G., Wang, Z., et al. (2021). Normal, novel or none: versatile regulation from alternative splicing. *Plant Signaling Behav.* 16 (7), e1917170. doi: 10.1080/15592324.2021.1917170

Long, Y., Liu, Z., Wang, P., Yang, H., Wang, Y., Zhang, S., et al. (2021). Disruption of topologically associating domains by structural variations in tetraploid cottons. *Genomics* 113 (5), 3405–3414. doi: 10.1016/j.ygeno.2021.07.023

Louadi, Z., Oubounyt, M., Tayara, H., and Chong, K. T. (2019). Deep splicing code: classifying alternative splicing events using deep learning. *Genes (Basel)* 10 (8), 587. doi: 10.3390/genes10080587

Mancini, E., Rabinovich, A., Iserte, J., Yanovsky, M., and Chernomoretz, A. (2021). ASpli: integrative analysis of splicing landscapes through RNA-Seq assays. *Bioinformatics* 37 (17), 2609–2616. doi: 10.1093/bioinformatics/btab141

Mark, F. R., Julie, T., Anireddy, S. R., and Asa, B. (2012). SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data. *Genome Biol.* 13 (1), R4. doi: 10.1186/gb-2012-13-1-r4

Matsukura, S., Mizoi, J., Yoshida, T., Todaka, D., Ito, Y., Maruyama, K., et al. (2010). Comprehensive analysis of rice DREB2-type genes that encode transcription factors involved in the expression of abiotic stress-responsive genes. *Mol. Genet. Genomics* 283 (2), 185–196. doi: 10.1007/s00438-009-0506-y

Nazari, I., Tayara, H., and Chong, K. T. (2019). Branch point selection in RNA splicing using deep learning. *IEEE Access* 7, 1800–1807. doi: 10.1109/access.2018.2886569

Nilsen, T. W., and Graveley, B. R. (2010). Expansion of the eukaryotic proteome by alternative splicing. *Nature* 463 (7280), 457–463. doi: 10.1038/nature08909

Pauwels, L., and Goossens, A. (2011). The JAZ proteins: a crucial interface in the jasmonate signaling cascade. *Plant Cell* 23 (9), 3089–3100. doi: 10.1105/tpc.111.089300

Peng, H., Wang, K., Chen, Z., Cao, Y., Gao, Q., Li, Y., et al. (2020). MBKbase for rice: an integrated omics knowledgebase for molecular breeding in rice. *Nucleic Acids Res.* 48 (D1), D1085–D1092. doi: 10.1093/nar/gkz921

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 (3), 290–295. doi: 10.1038/nbt.3122

Pulyakhina, I., Gazzoli, I., 't Hoen, P.-B., Verwey, N., den Dunnen, J., Aartsma-Rus, A., et al. (2015). SplicePie: a novel analytical approach for the detection of alternative, non-sequential and recursive splicing. *Nucleic Acids Res.* 43 (12), e80–e80. doi: 10.1093/nar/gkv242

Qin, P., Lu, H., Du, H., Wang, H., Chen, W., Chen, Z., et al. (2021). Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* 184 (13), 3542–3558 e3516. doi: 10.1016/j.cell.2021.04.046

Reddy, A. S., Rogers, M. F., Richardson, D. N., Hamilton, M., and Ben-Hur, A. (2012). Deciphering the plant splicing code: experimental and computational approaches for predicting alternative splicing and splicing regulatory elements. *Front. Plant Sci.* 3. doi: 10.3389/fpls.2012.00018

Regan, K., Saghafi, A., and Li, Z. (2021). Splice junction identification using long short-term memory neural networks. *Curr. Genomics* 22 (5), 384–390. doi: 10.2174/1389202922666211011143008

Romero, J. P., Muniategui, A., De Miguel, F. J., Aramburu, A., Montuenga, L., Pio, R., et al. (2016). EventPointer: an effective identification of alternative splicing events using junction arrays. *BMC Genomics* 17, 467. doi: 10.1186/s12864-016-2816-x

Ryan, M. C., Cleland, J., Kim, R., Wong, W. C., and Weinstein, J. N. (2012). SpliceSeq: a resource for analysis and visualization of RNA-Seq data on alternative splicing and its functional impacts. *Bioinformatics* 28 (18), 2385–2387. doi: 10.1093/bioinformatics/bts452

Sanchez-Martin, J., Widrig, V., Herren, G., Wicker, T., Zbinden, H., Gronnier, J., et al. (2021). Wheat Pm4 resistance to powdery mildew is controlled by alternative splice variants encoding chimeric proteins. *Nat. Plants* 7 (3), 327–341. doi: 10.1038/s41477-021-00869-2

Sharma, N., Geuten, K., Giri, B. S., and Varma, A. (2020). The molecular mechanism of vernalization in Arabidopsis and cereals: role of Flowering Locus C and its homologs. *Physiol. Plant* 170 (3), 373–383. doi: 10.1111/ppl.13163

Shen, S., Park, J. W., Lu, Z. X., Lin, L., Henry, M. D., Wu, Y. N., et al. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U.S.A.* 111 (51), E5593–E5601. doi: 10.1073/pnas.1419161111

Sibbesen, J. A., Eizenga, J. M., Novak, A. M., Siren, J., Chang, X., Garrison, E., et al. (2023). Haplotype-aware pantranscriptome analyses using spliced pangenome graphs. *Nat. Methods* 20 (2), 239–247. doi: 10.1038/s41592-022-01731-9

Sim, M., Lee, J., Lee, D., Kwon, D., and Kim, J. (2020). TAMA: improved metagenomic sequence classification through meta-analysis. *BMC Bioinf.* 21 (1), 185. doi: 10.1186/s12859-020-3533-7

Song, J. M., Liu, D. X., Xie, W. Z., Yang, Z., Guo, L., Liu, K., et al. (2021). BnPIR: Brassica napus pan-genome information resource for 1689 accessions. *Plant Biotechnol. J.* 19 (3), 412–414. doi: 10.1111/pbi.13491

Strauch, Y., Lord, J., NIranjan, M., and Baralle, D. (2022). CI-SpliceAI-Improving machine learning predictions of disease causing splicing variants using curated alternative splice sites. *PloS One* 17 (6), e0269159. doi: 10.1371/journal.pone.0269159

Sun, X., Zuo, F., Ru, Y., Guo, J., Yan, X., and Sablok, G. (2015). SplicingTypesAnno: annotating and quantifying alternative splicing events for RNA-Seq data. *Comput. Methods Programs BioMed.* 119 (1), 53–62. doi: 10.1016/j.cmpb.2015.02.004

Syed, N. H., Kalyna, M., Marquez, Y., Barta, A., and Brown, J. W. (2012). Alternative splicing in plants–coming of age. *Trends Plant Sci.* 17 (10), 616–623. doi: 10.1016/j.tplants.2012.06.001

Szakonyi, D., and Duque, P. (2018). Alternative splicing as a regulator of early plant development. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01174

Tao, Y., Zhao, X., Mace, E., Henry, R., and Jordan, D. (2019). Exploring and exploiting pan-genomics for crop improvement. *Mol. Plant* 12 (2), 156–169. doi: 10.1016/j.molp.2018.12.016

Thatcher, S. R., Danilevskaya, O. N., Meng, X., Beatty, M., Zastrow-Hayes, G., Harris, C., et al. (2016). Genome-wide analysis of alternative splicing during development and drought stress in maize. *Plant Physiol.* 170 (1), 586–599. doi: 10.1104/pp.15.01267

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28 (5), 511–515. doi: 10.1038/nbt.1621

Ule, J., and Blencowe, B. J. (2019). Alternative splicing regulatory networks: functions, mechanisms, and evolution. *Mol. Cell* 76 (2), 329–345. doi: 10.1016/j.molcel.2019.09.017

Vaquero-Garcia, J., Barrera, A., Gazzara, M. R., Gonzalez-Vallinas, J., Lahens, N. F., Hogenesch, J. B., et al. (2016). A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife* 5, e11752. doi: 10.7554/eLife.11752

Vitting-Seerup, K., Porse, B. T., Sandelin, A., and Waage, a.J. (2014). spliceR: an R package for classification of alternative splicing and prediction of coding potential from RNA-seq data. *BMC Bioinf.* 15, 81. doi: 10.1186/1471-2105-15-81

Wagner, N., Celik, M. H., Holzlwimmer, F. R., Mertes, C., Prokisch, H., Yepez, V. A., et al. (2023). Aberrant splicing prediction across human tissues. *Nat. Genet.* 55 (5), 861–870. doi: 10.1038/s41588-023-01373-3

Wan, R., Bai, R., and Shi, Y. (2019). Molecular choreography of pre-mRNA splicing by the spliceosome. *Curr. Opin. Struct. Biol.* 59, 124–133. doi: 10.1016/j.sbi.2019.07.010

Wang, W., Qin, Z., Feng, Z., Wang, X., and Zhang, X. (2013). Identifying differentially spliced genes from two groups of RNA-seq samples. *Gene* 518 (1), 164–170. doi: 10.1016/j.gene.2012.11.045

Wang, H. L., Zhang, Y., Wang, T., Yang, Q., Yang, Y., Li, Z., et al. (2021). An alternative splicing variant of PtRD26 delays leaf senescence by regulating multiple NAC transcription factors in Populus. *Plant Cell* 33 (5), 1594–1614. doi: 10.1093/plcell/koab046

Will, C. L., and Luhrmann, R. (2010). Spliceosome structure and function. *Cold Spring Harbor Perspect. Biol.* 3 (7), a003707–a003707. doi: 10.1101/cshperspect.a003707

Wu, J., Akerman, M., Sun, S., McCombie, W. R., Krainer, A. R., and Zhang, M. Q. (2011). SpliceTrap: a method to quantify alternative splicing under single cellular conditions. *Bioinformatics* 27 (21), 3010–3016. doi: 10.1093/bioinformatics/btr508

Wu, F., Deng, L., Zhai, Q., Zhao, J., Chen, Q., and Li, C. (2020). Mediator subunit MED25 couples alternative splicing of JAZ genes with fine-tuning of jasmonate signaling. *Plant Cell* 32 (2), 429–448. doi: 10.1105/tpc.19.00583

Wu, W., Zong, J., Wei, N., Cheng, J., Zhou, X., Cheng, Y., et al. (2018). CASH: a constructing comprehensive splice site method for detecting alternative splicing events. *Brief Bioinform.* 19 (5), 905–917. doi: 10.1093/bib/bbx034

Xing, Y., Goldstein, L. D., Cao, Y., Pau, G., Lawrence, M., Wu, T. D., et al. (2016). Prediction and quantification of splice events from RNA-seq data. *PloS One* 11 (5), e0156132. doi: 10.1371/journal.pone.0156132

Xiong, F., Ren, J. J., Yu, Q., Wang, Y. Y., Lu, C. C., Kong, L. J., et al. (2019). AtU2AF65b functions in abscisic acid mediated flowering via regulating the precursor messenger RNA splicing of ABI5 and FLC in Arabidopsis. *New Phytol.* 223 (1), 277–292. doi: 10.1111/nph.15756

Xu, Y., Wang, Y., Luo, J., Zhao, W., and Zhou, X. (2017). Deep learning of the splicing (epi)genetic code reveals a novel candidate mechanism linking histone modifications to ESC fate decision. *Nucleic Acids Res.* 45 (21), 12100–12112. doi: 10.1093/nar/gkx870

Yan, J., Zhang, C., Gu, M., Bai, Z., Zhang, W., Qi, T., et al. (2009). The Arabidopsis CORONATINE INSENSITIVE1 protein is a jasmonate receptor. *Plant Cell* 21 (8), 2220–2236. doi: 10.1105/tpc.109.065730

Yarden, K., Eric, T. W., Edoardo, M. A., and Christopher, B. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* 7 (12), 1009–1015. doi: 10.1038/nmeth.1528

Ye, Z., Chen, Z., Lan, X., Hara, S., Sunkel, B., Huang, T. H., et al. (2014). Computational analysis reveals a correlation of exon-skipping events with splicing, transcription and epigenetic factors. *Nucleic Acids Res.* 42 (5), 2856–2869. doi: 10.1093/nar/gkt1338

Yu, H., Lin, T., Meng, X., Du, H., Zhang, J., Liu, G., et al. (2021b). A route to *de novo* domestication of wild allotetraploid rice. *Cell* 184 (5), 1156–1170. doi: 10.1016/j.cell.2021.01.013

Yu, G., Zhou, G., Zhang, X., Domeniconi, C., and Guo, M. (2021a). DMIL-IsoFun: predicting isoform function using deep multi-instance learning. *Bioinformatics* 37 (24), 4818–4825. doi: 10.1093/bioinformatics/btab532

Zhang, D., Chen, M.-X., Zhu, F.-Y., Zhang, J., and Liu, Y.-G. (2020). Emerging functions of plant serine/arginine-rich (SR) proteins: lessons from animals. *Crit. Rev. Plant Sci.* 39 (2), 173–194. doi: 10.1080/07352689.2020.1770942

Zhang, Y., Liu, X., MacLeod, J., and Liu, J. (2018). Discerning novel splice junctions derived from RNA-seq alignment: a deep learning approach. *BMC Genomics* 19 (1), 971. doi: 10.1186/s12864-018-5350-1

Zhang, Z., Pan, Z., Ying, Y., Xie, Z., Adhikari, S., Phillips, J., et al. (2019). Deep-learning augmented RNA-seq analysis of transcript splicing. *Nat. Methods* 16 (4), 307–310. doi: 10.1038/s41592-019-0351-9

Zhu, F.-Y., Chen, M.-X., Ye, N.-H., Shi, L., Ma, K.-L., Yang, J.-F., et al. (2017). Proteogenomic analysis reveals alternative splicing and translation as part of the abscisic acid response in Arabidopsis seedlings. *Plant J.* 91 (3), 518–533. doi: 10.1111/tpj.13571

Zuallaert, J., Godin, F., Kim, M., Soete, A., Saeys, Y., and De Neve, W. (2018). SpliceRover: interpretable convolutional neural networks for improved splice site prediction. *Bioinformatics* 34 (24), 4180–4188. doi: 10.1093/bioinformatics/bty497

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 20 September 2023
DOI 10.3389/fpls.2023.1236315

OPEN ACCESS

EDITED BY
Yan Zhao,
Shandong Agricultural University, China

REVIEWED BY
XingWen Zhou,
Fujian University of Technology, China
Yun-peng Du,
Beijing Academy of Agricultural and Forestry Sciences, China

*CORRESPONDENCE
Ziming Ren
zimingren@zju.edu.cn
Yun Wu
yunwu@zju.edu.cn

†PRESENT ADDRESS
Lin Zhang,
Department of Ophthalmology and Neurobiology, University of California, Los Angeles, Los Angeles, CA, United States

‡These authors have contributed equally to this work

RECEIVED 07 June 2023
ACCEPTED 01 September 2023
PUBLISHED 20 September 2023

CITATION
Gao C, Zhang L, Xu Y, Liu Y, Xiao X, Cui L, Xia Y, Wu Y and Ren Z (2023) Full-length transcriptome analysis revealed that 2,4-dichlorophenoxyacetic acid promoted *in vitro* bulblet initiation in lily by affecting carbohydrate metabolism and auxin signaling.
Front. Plant Sci. 14:1236315.
doi: 10.3389/fpls.2023.1236315

Full-length transcriptome analysis revealed that 2,4-dichlorophenoxyacetic acid promoted *in vitro* bulblet initiation in lily by affecting carbohydrate metabolism and auxin signaling

Cong Gao[1‡], Lin Zhang[1†‡], Yunchen Xu[1], Yue Liu[1], Xiao Xiao[1], Liu Cui[2], Yiping Xia[1], Yun Wu[2*] and Ziming Ren[2*]

[1]Genomics and Genetic Engineering Laboratory of Ornamental Plants, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, Zhejiang, China, [2]Laboratory of Flower Bulbs, Department of Landscape Architecture, Zhejiang Sci-Tech University, Hangzhou, Zhejiang, China

Bulblet initiation, including adventitious bud initiation and bulblet formation, is a crucial process for lily and other bulbous flowers that are commercially propagated by vegetative means. Here, by a hybrid strategy combining Pacific Biosciences (PacBio) full-length sequencing and Illumina RNA sequencing (RNA-seq), high-quality transcripts of *L. brownii* (*Lb*) and its variety, *L. brownii* var. giganteum (*Lbg*), during *in vitro* bulblet initiation were obtained. A total of 53,576 and 65,050 high-quality non-redundant full-length transcripts of *Lbg* and *Lb* were generated, respectively. Morphological observation showed that *Lbg* possessed a stronger capacity to generate bulblets *in vitro* than *Lb*, and 1 mg L^{-1} 2,4-dichlorophenoxyacetic acid (2,4-D) significantly increased bulblet regeneration rate in two lilies. Screening of differentially expressed transcripts (DETs) between different stages and Mfuzz analysis showed 0 DAT to 1 DAT was the crucial stage with the most complex transcriptional change, with carbohydrate metabolism pathway was significantly enriched. In addition, 6,218 and 8,965 DETs were screened between the 2,4-D-treated group and the control group in *Lbg* and *Lb*, respectively. 2,4-D application had evident effects on the expression of genes involved in auxin signaling pathway, such as TIRs, ARFs, Aux/IAAs, GH3s and SAURs. Then, we compared the expression profiles of crucial genes of carbohydrate metabolism between different stages and different treatments. SUSs, SUTs, TPSs, AGPLs, GBSSs and SSs showed significant responses during bulblet initiation. The expression of CWINs, SUTs and SWEETs were significantly upregulated by 2,4-D in two lilies. In addition, 2,4-D increased the expression of starch degradation genes (AMYs and BAMs) and inhibited starch synthesis genes (AGPLs, GBSSs and SSs). SBEs were significantly upregulated in *Lbg* but not in *Lb*. Significant co-expression was showed between genes involved in carbohydrate metabolism and auxin signaling, together with transcription factors such as bHLHs, MYBs, ERFs and C3Hs. This study indicates the coordinate regulation of bulblet initiation by carbohydrate metabolism and

auxin signaling, serving as a basis for further studies on the molecular mechanism of bulblet initiation in lily and other bulbous flowers.

KEYWORDS

starch synthesis and degradation, sucrose unloading, auxin signaling, full-length sequencing, *Lilium*

1 Introduction

Bulbous flowers, which are highly popular in the world floriculture market, are usually commercially propagated by vegetative means, especially by bulbs, to maintain phenotypic uniformity and genetic purity. Bulblet initiation, including the process of adventitious/axillary bud initiation and bulblet formation, has been reported in multiple bulbous plants from plant tissues *in vitro* (Van Aartrijk and Blom-Barnhoorn, 1984; Li et al., 2014; Ren et al., 2017; Lv et al., 2020). Lily (*Lilium* spp.), a perennial monocotyledon of the family Liliaceae, is one of the major bulbous crops in the floriculture industry with high ornamental, medical and edible value (Lee et al., 2013; Xu et al., 2017; Wu et al., 2019). Due to its strong bulblet formation capacity *in vitro*, lily is considered as an appropriate experimental material for the study of bulblet initiation and its underlying mechanisms. Studies on bulblet initiation in lily have focused mainly on the influence of wounds, temperature treatment and exogenous phytohormones (Van Aartrijk and Blom-Barnhoorn, 1984), changes in endogenous carbohydrate and hormone contents, and the expression level of genes involved in carbohydrate and phytohormone metabolism (Li et al., 2014; Wu et al., 2020).

Carbohydrate metabolism plays a vital role in bulblet initiation of multiple bulbous flowers. Therein, starch metabolism has been reported to be ubiquitously involved. Storage starch in mother scales could act as a carbon source for bulblet initiation. At the bulblet appearance and enlargement stage, the enzymes involved in the starch synthetic direction, such as ADP-glucose Pyrophosphorylase (AGPase, EC 2.7.7.27), Starch Synthase (SS, EC 2.4.1.21), Starch Branching Enzyme (SBE, EC 2.4.1.18) and Granule-bound Starch Synthase (GBSS, EC 2.4.1.242), showed a decreasing trend in mother scales but higher gene expression levels in newly formed bulblets, while the enzyme in the starch cleavage direction, Starch Debranching Enzyme (DBE, EC 3.2.1.10), showed higher expression levels in scales than in bulblets in lily (Li et al., 2014). Similarly, in *Lycoris*, soluble sugars derived from starch degradation in the outer scales were transported into the inner scales and promote bulblet initiation and development through starch synthesis, especially through AGPases (Xu et al., 2020a). Sucrose, the main form of transported sugar in higher plants, were also considered to regulate bulblet initiation. Sucrose Synthase (SUS, EC 2.4.1.13) and Invertase (INV, EC 3.2.1.26, including Cell Wall Invertase (CWIN), Vacuolar Invertase (VIN) and Cytoplasmic Invertase (CIN)), mainly hydrolyzing sucrose, presented higher expression levels in mother scales and bulblets at stages of bulblet appearance and enlargement in lily (Li et al., 2014). A clear shift was observed from CWIN-catalyzed to SUS-catalyzed sucrose cleavage patterns, meanwhile, sucrose unloading pathway changed from apoplasmically to symplasimically at the key shoot-to-bulblet transition stage in *Lilium* Oriental Hybrids 'Sorbonne' (Wu et al., 2021). Similarly, *CWIN* and *SUS* exhibited exactly opposite expression patterns during the competence stage of bulblet regeneration in *Lycoris* (Ren et al., 2021). The above results indicated that the transition from bud initiation to bulblet enlargement was usually accompanied by the change of dominant sucrose unloading pathway.

Auxin, a key phytohormone, regulates diverse aspects of plant growth and developmental processes through its dynamic differential distribution (Vanneste and Friml, 2009). The biosynthesis of indole-3-acetic acid (IAA), the main naturally occurring auxin, in higher plants requires two steps: first, tryptophan is converted to indole-3-pyruvate (IPA) by Tryptophan Aminotransferase (TAA) or Tryptophan Aminotransferase Related (TAR); second, IAA is produced from IPA by YUC family (YUC) (Zhao, 2012). A previous study revealed that adventitious bulblets of lily were formed at the basal edge of the explant under tissue culture conditions, which caused by basipetal auxin transport (Van Aartrijk and Blom-Barnhoorn, 1984). Different auxin concentrations showed different effects on the process of bulblet formation. Auxin likely promoted the initiation of bulbils and then inhibited further bulbil formation in lily (Yang et al., 2017). In *Lycoris*, endogenous IAA content showed an increase and then a decrease during bulblet initiation and development, which were consistent with the expression patterns of genes involved in IAA synthesis and signal transduction (Xu et al., 2020a).

To obtain more genetic information, Illumina next-generation sequencing (NGS) technology has been widely used in the study of the process of bulblet formation in multiple bulbous plants, including *Lilium* (Li et al., 2014; Yang et al., 2017), *Lycoris* (Ren et al., 2022) and sweet potato (*Ipomoea batatas*) (Firon et al., 2013). In *Lilium*, transcriptome analysis was used to elucidate the molecular mechanism of bulblet/bulbil formation and development. (Li et al., 2014; Yang et al., 2017). However, the early stage of bulblet formation in *Lilium* and even in bulbous flowers remains unclear. Recently, the single-molecule real-time (SMRT) sequencing technology of the PacBio system has offered a new third-generation sequencing platform, which possesses advantages such as long read lengths (length > 10 kb), high consensus accuracy and a low degree of bias (Hu et al., 2022), which is an available and reliable strategy to generate more accurate

and comprehensive genetic information. A recent study obtained *Lilium* Oriental Hybrids 'Sorbonne' transcriptome during induction of aerial bulbil using the combination of SMRT and NGS technology (Li et al., 2022).

Here, we explore the *in vitro* bulblet initiation process of *Lilium brownii* (*Lb*) and *Lilium brownii* var. *giganteum* (*Lbg*), a variant of *Lb* (Li et al., 2007), through careful morphological observation, and then divided the process into four stages. Then, a hybrid strategy combining Pacific Biosciences (PacBio) full-length sequencing and Illumina sequencing was conducted. Through differentially expression transcript (DET) screening and Mfuzz analysis, we identified key metabolic pathways and candidate genes during bulblet initiation. In addition, DET screening was also performed between the 2,4-dichlorophenoxyacetic acid (2,4-D)-treated group and the control group, to explore how 2,4-D, a synthetic auxin analog, affect the process of bulblet initiation. Furthermore, we hypothesized that carbohydrate metabolism and auxin signaling coordinately regulate bulblet initiation, with several transcriptional factors (TFs) involved in the regulation of this process. Our findings provide a comprehensive understanding of the molecular mechanism underlying the process of bulblet initiation in lily.

2 Materials and methods

2.1 Plant materials and growth conditions

Bulblet induction experiments were conducted at the Physiology & Molecular Biology Laboratory of Ornamental Plants and Tissue Culture Laboratory of Ornamental Plants at Zhejiang University, Hangzhou (118°21'-120°30'E, 29°11′-30°33'N), China. *In vitro* seedlings of *Lb* and *Lbg* were cultured at 25 ± 2°C under a 12:12 h light:dark photoperiod with 60 μmol photons m^{-2} s^{-1}. Healthy outer scales without damage were removed carefully from fresh *in vitro* bulbs (4~6 cm in circumference) and then cultured for 14 days on basal Murashige and Skoog (MS) medium (Murashige and Skoog, 1962) containing 6% sucrose and 0.3% Phytagel (P8169, Sigma-Aldrich, St. Louis, MO, USA) (pH 5.8), to which was added 0 mg L^{-1} or 1 mg L^{-1} 2,4-D, with the adaxial side facing upward. Each treatment contained three biological replicates, and each replicate included 120 scales.

2.2 Morphological and histological observation

Morphological changes of bulblet initiation in *Lb* and *Lbg* were observed under a stereomicroscope (SZM745T, OPLENIC, China). The regeneration rate and propagation efficiency of bulblets were calculated as follows: Regeneration rate = number of scales that produced adventitious buds/total number of scales; Propagation efficiency = total number of produced buds/total number of scales. Representative data were supported by three biological replicates, each containing 120 repeats. Transverse sections of scales at the proximal end where adventitious buds initiated were stained with periodic acid-Schiff (PAS) and Naphthol Yellow S as previously described (Ren et al., 2017), and then observed using an upright light microscope (Eclipse E100, Nikon, Japan).

2.3 Sample collection

The bulblet initiation process of *Lb* and *Lbg* was divided into four crucial stages according to the results of morphological and histological observations: stage of scale detachment (0 DAT; DAT, days after the treatment of detaching the scale from the basal plate), stage of wounding response and early regeneration competence (1 DAT), stage of adventitious bud initiation (8 DAT) and stage of adventitious bud swelling and bulblet formation (14 DAT). The entire scales used for bulblet induction at 0 DAT, 1 DAT, 8 DAT and 14 DAT were sampled, frozen in liquid nitrogen and stored at -80°C for total RNA extraction. Sampling was performed with three biological replicates for each stage.

2.4 Generation of the full-length reference transcripts for *Lb* and *Lbg*

Total RNA of scale samples at four stages of *Lb* and *Lbg* cultured on 2,4-D-free (0 mg L^{-1} 2,4-D) and 2,4-D-containing (1 mg L^{-1} 2,4-D) medium was extracted using an EASYspin Plus Complex RNA Kit (RN53, Aidlab Bio, China) according to the manufacturer's instructions. Total RNA from each sample was equally mixed to generate a pool and then synthesized to first-strand cDNA using Clontech SMARTer PCR cDNA Synthesis Kit (Clontech, Mountain View, CA, USA). Large-scale PCR was performed using the BluePippin™ Size Selection System (Sage Science, Beverly, MA, USA). The SMRTbell template libraries were constructed and then sequenced on the PacBio Sequel platform. The following methods for generating full-length reference transcripts referenced Li et al. (2022) with some modifications. The high-quality full-length transcripts were removed rebundancy using CD-HIT v4.6.142 (Li and Godzik, 2006). BUSCO 5.2.0 was used to assess the quality and completeness of the reference transcripts using the official BUSCO datasets (liliopsida_odb10) (Manni et al., 2021).

Total RNA of scale samples was sequenced with the Illumina Xten 4000 platform (Illumina, San Diego, CA, USA). Quality control (QC) for each Illumina transcriptome was performed by fastp (v0.19.7) with default parameters (Chen et al., 2018). Clean reads were separately mapped to their corresponding reference transcripts by Bowtie2 (v2.3.4), and the expression levels of each transcript, including reads count and fragments per kilobase million (FPKM), were calculated by RSEM (v1.3.1) (Li and Dewey, 2011; Langmead and Salzberg, 2012).

2.5 Transcript annotation

Each transcript was annotated by the eggNOG-mapper webserver (http://eggnog-mapper.embl.de/) with an e-value of ≤

$1e^{-5}$ and identity of ≥ 60% with Viridiplantae (green plants) selected as the taxonomic scope (Huerta-Cepas et al., 2019; Cantalapiedra et al., 2021). As a result, transcripts were functionally annotated based on the following databases: Pfam, NCBI nonabundant (NR), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO). The PlantTFDB v5.0 database (Jin et al., 2017) was used to predict TFs and their homologous genes in *Arabidopsis thaliana* of the full-length reference transcripts in *Lb* and *Lbg*. Venn diagrams were generated using Evenn (http://www.ehbio.com/test/venn/#/, Chen et al., 2021).

2.6 DET screening and enrichment analysis

Differential expression analysis was performed using R package DESeq2 (v1.32.0) (Love et al., 2014). Transcripts with a false discovery rate (FDR) ≤ 0.05 and a threshold of |log2-fold changes| ≥ 2 were recognized as DETs. Each group of DETs was calculated for KEGG pathways by clusterProfiler (v4.0.5) using an p-value of 0.05 to find significant enrichments (Wu et al., 2021). Time-series cluster analysis was performed by the Mfuzz package (v2.50.0) in R software, with the low expression level transcripts with FPKM ≤ 5 removed.

2.7 Quantitative real-time PCR validation

To confirm the results of the transcript expression levels from RNA-Seq, six DETs of *Lbg* and six DETs of *Lb* were selected for expression analysis using quantitative real-time PCR (qRT-PCR). Total RNA (1 μg) of each sample was reverse transcribed by the PrimeScript™ RT reagent Kit with gDNA Eraser (RR047A, TaKaRa, Dalian, China). The diluted (1:30) cDNA was used as the template for qRT-PCR analysis. Gene-specific primers were designed by the NCBI Primer-BLAST tool (Ye et al., 2012) and are listed in Supplementary Table S2. Then, qRT-PCR was performed with TB Green™ Premix Ex Taq™ Kit (RR420A, TaKaRa, Dalian, China) in a Bio-Rad Connect™ Optics Module (Bio-Rad, CA, USA). All reactions were conducted in triplicate, and the $2^{-\Delta\Delta CT}$ method was applied to calculate the relative expression level using *GAPDH* as the reference gene.

2.8 Nonstructural carbohydrate content assay

Total starch contents of scales at different stages were measured using a Starch Content Kit (A148-1-1, Nanjing Jiancheng Bioengineering Institute, Nanjing, China) following the manufacturer's protocol. Three replicates were included in each assay. Sucrose, fructose, and glucose contents were measured by high-performance liquid chromatography (HPLC) (e2695, Waters, MA, USA) equipped with Refractive Index (RI) Detector 2414 (Waters) according to the previous method in Liu et al. (2020).

2.9 Determination of IAA concentration

IAA extraction and quantification were performed using a previously described method with slight modifications (Guo et al., 2016). Briefly, frozen scale sample (100 mg) of each treatment at each stage was weighed in a 10-mL centrifuge tube, and homogenized in 1 mL of ethyl acetate that had been spiked with D5-IAA (C/D/N Isotopes) as an internal standard at a final concentration of 100 ng mL^{-1}. The tubes were centrifuged at 12 000 rpm for 10 min at 4°C. The resulting supernatant was dried by blowing under N2. The residue was resuspended in 0.5 mL of 70% (v/v) methanol and centrifuged, and the supernatants were then analyzed in a triple quadrupole mass spectrometer (6470, Agilent Technologies, CA, USA).

2.10 Statistical analysis

One-way analysis of variance (ANOVA) was used to compare differences among different indices or treatments via SPSS 26.0 (IBM Corp., Armonk, NY, USA). Correlation analyses between gene expression data in *Lb* and *Lbg* were performed using Pearson's two-tailed tests and visualized by Cytoscape v3.7.1 (Shannon et al., 2003).

3 Results

3.1 Stage division according to morphological and histological observation

Throughout the bulblet initiation and development process (0 DAT to 49 DAT), the regeneration rate and propagation efficiency were calculated. It is showed that visible adventitious buds occurred at 8 DAT, and the number of newly formed bulblets was significantly increased until 14 DAT and then tended to plateau until 25 DAT or later in both *Lbg* and *Lb* (Figure 1A). In addition, 1 DAT was considered as the crutial wounding response stage according to our previous studies. Based on morphological observations, browning substances began to accumulate in the transverse section at 1 DAT (Figures 1B3, B8). At 8 DAT, adventitious buds formed at the adaxial side rather than the abaxial side of the scale, mainly around lateral vascular bundles (Figures 1B4, B9; Supplementary Figures S1B, E). Then, adventitious buds swelled, and visible bulblets occurred at 14 DAT (Figures 1B5, B10; Supplementary Figures S1C, F). Taken together, we divided the early bulblet initiation process into four stages: 0 DAT (scale detachment), 1 DAT (wounding response and early regeneration competence), 8 DAT (adventitious bud initiation) and 14 DAT (bud swelling and bulblet formation). Moreover, the bulblet regeneration rate and propagation efficiency of *Lbg* were both significantly higher than those of *Lb* from 8 DAT to 49 DAT (Figure 1A). At 14 DAT, the regeneration rate and propagation efficiency of *Lbg* were 0.619 and 1.677, respectively, significantly higher ($p < 0.001$ and $p < 0.01$, respectively) than those in *Lb* (0.236

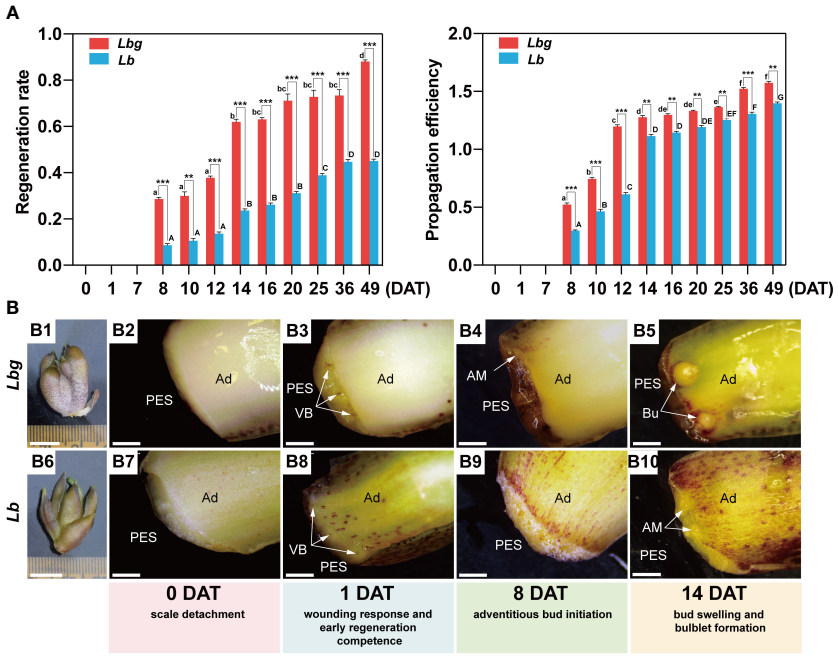


FIGURE 1
Morphological observation during *in vitro* bulblet initiation in *Lbg* and *Lb*. **(A)** Regeneration rate and propagation efficiency during *in vitro* bulblet initiation. Regeneration rate, number of scales that produced adventitious buds/total number of scales. Propagation efficiency, total number of produced buds/total number of scales. Lowercase and uppercase letters represent significant differences ($p < 0.001$) for relevant parameters within *Lbg* and *Lb*, respectively. Asterisks indicate significant differences for relevant parameters between *Lbg* and *Lb* (**Differences significant at $p < 0.01$; ***Differences significant at $p < 0.001$). Representative data were supported by three biological replicates containing 120 repeats each. **(B)** *In vitro* bulb of which scales were used for bulblet induction **(B1, B6)** and key stages during bulblet initiation in *Lbg* **(B2-B5)** and *Lb* **(B7-B10)**. PES, proximal end of scale; Ad, adaxial side of scale; VB, vascular bundle; AM, adventitious meristem; Bu, bulblet. The white arrows represent vascular bundles (B3 and B4), adventitious meristems (B4 and B10) or bulblets (B9). Bars, 1 cm **(B1, B6)** and 1 mm **(B2-B5, B7-B10)**.

and 1.114, respectively) (Figure 1A). The above results indicated that *Lbg* possesses a stronger capacity to generate bulblets *in vitro* than *Lb*.

3.2 Quality assessment of obtained reference transcripts

Full-length transcriptome sequencing was conducted to generate complete and accurate gene information during bulblet initiation (Figure 2A). The reference transcripts for *Lbg* were first obtained with 53,576 nonredundant transcripts of an average length of 3,108 bp, and 80.79% of the *Lbg* clean reads obtained by Illumina RNA-Seq mapped to the *Lbg* reference transcripts (Supplementary Table S1). Similarly, the *Lb* full-length reference transcripts were obtained, consisting of 65,050 nonredundant transcripts of an average length of 2,965 bp, with 81.85% *Lb* clean reads mapped to them (Supplementary Table S1). Moreover, the *Lbg* and *Lb* reference transcripts had 73.1% and 75.9% of the conserved plant genes by BUSCO 5.2.0, respectively (Supplementary Table S1). These results confirmed the reliability of these two transcriptomes for downstream analysis. The length of transcripts ranged from 294 bp to 13,738 bp in *Lbg* and from 292 bp to 14,486 bp in *Lb*, with a median length of 2,139 and 2,021 bp and a mean length of 3,108 and 2,965 bp in *Lbg* and *Lb*, respectively (Figure 2B). Moreover, a total of 38,725 (72.3%) and 45,967 (70.7%) transcripts were annotated to the NR, GO, KEGG and Pfam databases in *Lbg* and *Lb*, respectively (Figure 2C). Among them, 13,174 and 15,472 transcripts were annotated in all the four databases.

Principal component analysis (PCA) showed that there were obvious differences of expression patterns among different stages under the same treatment conditions, except for 8 DAT and 14 DAT in *Lb* in the medium supplemented with 1 mg L^{-1} 2,4-D (Figure 2D). In addition, a large deviation was also observed between the 2,4-D treatment group and the control group at the same stage. The results of sample clustering were consistent with the PCA results (Supplementary Figure S2). The above results further validated that obvious differences existed among stages during bulblet initiation, and 2,4-D influenced this process in both *Lbg* and *Lb*.

Six transcripts of *Lbg* (Isoform 22016, 21863, 28761, 46833, 29525 and 20315) and six transcripts of *Lb* (Isoform 31916, 25438, 39080, 36307, 34089 and 27602) were randomly selected for qRT-PCR to validate the differential expression by RNA-Seq. The results showed that the differential expression levels of these selected transcripts by qRT-PCR were highly consistent with those obtained by RNA-Seq (Supplementary Figure S3), confirming the reliability and accuracy of the RNA-Seq data.

10.3389/fpls.2023.1236315

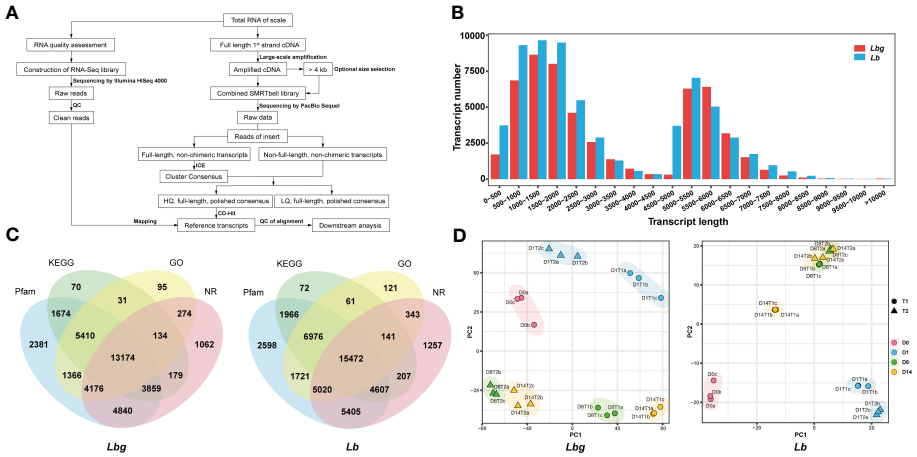


FIGURE 2
Reference transcripts generation of lily samples. **(A)** Workflow of lily sample sequencing. **(B)** Distribution of the length of reference transcripts. **(C)** Functional annotation of reference transcripts by Pfam, KEGG, GO and Nr databases. **(D)** Principal component analysis (PCA) plot of the samples. T1 and T2 represent samples of the control group and 2,4-D-treated group, respectively. D0, D1, D8 and D14 represent samples of 0 DAT, 1 DAT, 8 DAT and 14 DAT, respectively. The letters (a-c) represent three biological replicates.

3.3 Stage-specific DET screening and Mfuzz analysis revealed possible events of different stages

To identify transcriptional changes between distinct bulblet initiation stages, we compared the transcript expression profiles of adjacent stages in each lily, including 1 DAT versus (vs) 0 DAT (Group 1), 8 DAT vs 1 DAT (Group 2) and 14 DAT vs 8 DAT (Group 3). In total, 11,964 and 18,834 stage-specific DETs were screened in *Lbg* and *Lb*, respectively (Figures 3A, B). Clearly, Group 1 had the largest number of DETs among the three groups in both *Lbg* and *Lb* (Figures 3A, B), indicating that there were more complex transcriptional changes from 0 DAT to 1 DAT than in the following stages.

To define the temporal characteristics of the transcript dataset, we performed clustering analysis of 41,680 and 48,314 transcripts by Mfuzz in *Lbg* and *Lb*, respectively. The transcripts were divided into eight and ten clusters in *Lbg* and *Lb*, respectively (Figures 3C,

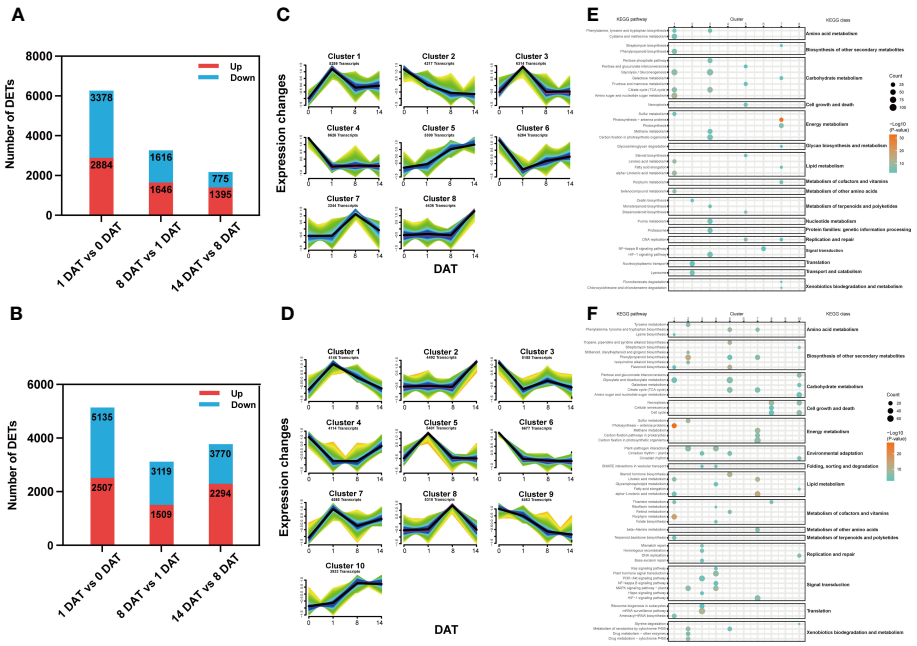


FIGURE 3
Stage-specific DETs screening and Mfuzz analysis of reference transcripts. **(A, B)** Number of DETs (|log2-fold changes| ≥ 2) identified between different stages in *Lbg* **(A)** and *Lb* **(B)**. **(C, D)** Stage-specific DETs were classified into eight clusters of *Lbg* **(C)** and ten clusters of *Lb* **(D)** through Mfuzz analysis. **(E, F)** Bubble charts of KEGG pathway enrichment analysis of each cluster in *Lbg* **(E)** and *Lb* **(F)**.

D), and KEGG pathway enrichment analysis of these clusters was conducted (Figures 3E, F). The transcripts of clusters 1 and 3 in *Lbg* and clusters 5 and 7 in *Lb* could be candidates for the early response to scale detachment, since the highest expression of these clusters was exhibited at 1 DAT. The pathway "citrate cycle (TCA cycle)" (map00020), which is involved in carbohydrate metabolism, was upregulated in all the four candidate clusters (Figures 3E, F). The transcripts of cluster 7 in *Lbg* and cluster 8 in *Lb* were highly expressed at 8 DAT, with "DNA replication" (map03030) and "glycosaminoglycan degradation" (map00531) enriched in cluster 7 (*Lbg*), and "cell cycle" (map04110), "cellular senescence" (map04218) and "necroptosis" (map04217) enriched in cluster 8 (*Lb*), indicating that this stage could be associated with cell growth and death (Figures 3E, F). Moreover, the expression levels of transcripts in cluster 8 (*Lbg*) and cluster 2 (*Lb*) were increased at 14 DAT, and transcripts in cluster 2 (*Lb*) were mainly enriched in "biosynthesis of other secondary metabolites" and "xenobiotics biodegradation and metabolism" classes, which might play important functional roles in the bulblet swelling stage (Figures 3E, F).

3.4 2,4-D treatment promoted the process of *in vitro* bulblet initiation

To explore the effect of the exogenous application of 2,4-D on *in vitro* bulblet initiation, 1 mg L^{-1} 2,4-D was added to the medium for bulblet induction. Results showed that the regeneration rate was significantly higher ($p < 0.001$) in 2,4-D-treated group than in the control group in both *Lbg* and in *Lb* (Figure 4A). More specifically, the 2,4-D treatment increased the regeneration rate by 1.84-fold in *Lbg* and 3.55-fold in *Lb* at 8 DAT, and 1.36-fold in *Lbg* and 1.51-fold in *Lb* at 14 DAT, indicating that the promotion effect of 2,4-D was stronger in *Lb*. Next, we screened 6,218 (3,175 upregulated, 3,043 downregulated) and 8,965 (5,382 upregulated, 3,583 downregulated) DETs between 2,4-D-treated group and the control group at 1 DAT, 8 DAT and 14 DAT in *Lbg* and *Lb*, respectively (Figure 4B). KEGG pathway enrichment analysis showed that, with 2,4-D treatment, "plant hormone signal transduction" (map04075) was upregulated in all the three stages in *Lb* and was also upregulated at 1 DAT and 8 DAT in *Lbg* (Figure 4C), indicating that 2,4-D could promote phytohormone responsiveness throughout the bulblet initiation process. Thus, 2,4-D had a promoting effect on *in vitro* bulblet initiation, and this effect might be closely related to phytohormone signal transduction.

3.5 Changes in auxin-related genes during bulblet initiation

Considering the effect of auxin on bulblet initiation in various bulbous flowers, and the "plant hormone signal transduction" pathway was in response to exogenous 2,4-D application, we focused on the expression patterns of all screened DETs involved in auxin biosynthesis and signaling (Figure 5A). It is obvious that all screed DETs encoding Transporter Inhibitor Response 1 (TIR1) (six in *Lbg* and four in *Lb*) were significantly downregulated ($p <$ 0.001) at 1 DAT in both *Lbg* and *Lb* (Figures 5B, C). Particularly, in

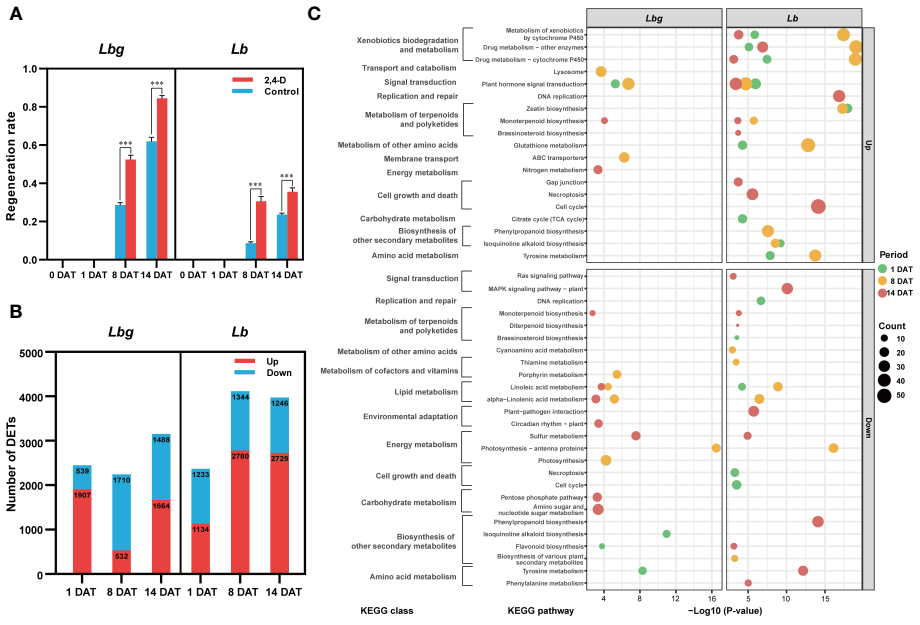


FIGURE 4
2,4-D-related DETs screening between the control group and 1 mg L^{-1} 2,4-D-treated group of reference transcripts. **(A)** Regeneration rate of the control group and 2,4-D-treated group during *in vitro* bulblet initiation. Regeneration rate = number of scales that produced adventitious buds/total number of scales. ***Differences significant at $p < 0.001$. **(B)** Number of DETs (|log2-fold changes| ≥ 2) identified between the control group and 2,4-D-treated group. **(C)** Bubble charts of KEGG pathway enrichment analysis of 2,4-D-related DETs.

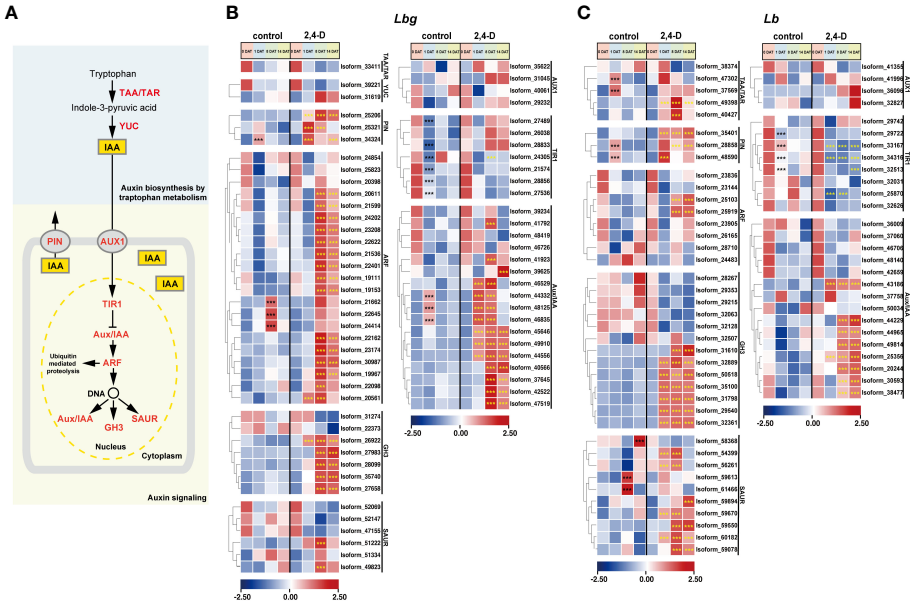


FIGURE 5
Expression patterns of stage-specific and 2,4-D related DETs involved in pathways of auxin biosynthesis and signaling. **(A)** Pathway of indole-3-acetic acid (IAA) biosynthesis and signaling. TAA, tryptophan aminotransferase; TAR, tryptophan aminotransferase related; YUC, YUC family; PIN, PIN-formed protein family; AUX1, AUX1/LAX symporters; TIR1, transporter inhibitor response 1, Aux/IAA, indole-3-acetic acid inducible; ARF, auxin response factor; GH3, GH3 family; SAUR, small auxin upregulated RNA. **(B, C)** Expression patterns of DETs involved in auxin biosynthesis and signaling in *Lbg* **(B)** and *Lb* **(C)**. ***Differences significant at $p < 0.001$. The black asterisk represents a significant difference compared to 0 DAT in the control group. The yellow asterisk represents a significant difference in the 2,4-D-treated group compared to the control group at the same stage.

Lb, the downregulation of four TIR1s (Isoform 33167, 34316, 32513 and 25870) were significantly enhanced ($p < 0.001$) by 2,4-D treatment (Figure 5C). Besides, 2,4-D significantly promoted ($p < 0.001$) the expression of two Tryptophan Aminotransferase (TAA) and Tryptophan Aminotransferase Related (TAR) (Isoform 49388 and 40427) in *Lb*, which could promote the endogenous synthesis of auxin (Figure 5C). Correspondingly, the content of endogenous IAA in the scales was significantly higher ($p < 0.05$) under 2,4-D treatment than that of the control group at 8 DAT and 14 DAT in *Lb* (Supplementary Figure S4). These changes above may lead to a stronger increase of regeneration rate in *Lb* than in *Lbg*.

Three Auxin Response Factors (ARFs) of *Lbg* (Isoform 21662, 22645 and 24414) were significantly ($p < 0.001$) upregulated at 8 DAT in the control group, and 15 ARFs in *Lbg* and two ARFs in *Lb* were significantly ($p < 0.001$) upregulated at 8 DAT with 2,4-D treatment, following the downregulation of TIR1s (Figures 5B, C), indicating that 2,4-D could promote auxin signaling pathway during bulblet initiation. The significantly higher expression levels of DETs ($p < 0.001$) encoding some GH3 (five in *Lbg* and seven in *Lb*), Small Auxin Upregulated RNA (SAUR) (two in *Lbg* and seven in *Lb*) and Indole-3-acetic Acid Inducible (Aux/IAA) (14 in *Lbg* and eight in *Lb*) in the 2,4-D-treated group than in the control group at 1 DAT or 8 DAT further support the above idea (Figures 5B, C). Moreover, 2,4-D also promoted all screed DETs encoding PIN-formed protein family (PIN) (three in *Lbg* and three in *Lb*) after 1 DAT (Figures 5B, C). Overall, the promotion of auxin signaling pathway by 2,4-D application might contribute to the enhancement of bulblet regeneration ability.

3.6 Changes in key genes involved in sucrose and starch metabolism during bulblet initiation

Sucrose and starch metabolism was repeatedly reported to play an essential role in bulblet initiation in various bulbous flowers. Here, we focused on the expression patterns of key enzymes, transporters, and regulators involved in sucrose and starch metabolism pathway during bulblet initiation (Figure 6A). Without 2,4-D treatment, all screened DETs encoding Sucrose Synthase (SUS) (16 in *Lbg* and 17 in *Lb*) were significantly upregulated ($p < 0.001$) at 1 DAT or 8 DAT (Figures 6B, C). One Cell Wall Invertase (CWIN) in *Lbg* (Isoform 29531) were upregulated at 1 DAT and then downregulated, and the expression level of four CWINs (Isoform 29531, 28761 in *Lbg* and Isoform 39080, 34537 in *Lb*) were significantly increased ($p < 0.001$) in the 2,4-D-treated group (Figures 6B, C). Correspondingly, the content of glucose, *one of the* hexose hydrolytic *products of sucrose*, in the scales was significantly higher ($p < 0.05$) in the 2,4-D treated group than that in the control group in both *Lbg* and *Lb* (Supplementary Figure S4). Three Trehalose 6-Phosphate Synthases (TPSs) in *Lbg* (Isoform 36005, 19880 and 20796) and five TPSs in *Lb* (Isoform 24997, 32954, 34599, 24963 and 43346) were significantly upregulated ($p < 0.001$) at 1 DAT, 8 DAT or 14 DAT (Figures 6B, C). Notably, the expression patterns of some Sucrose Transporters (SUTs) and SWEET Sucrose-Efflux Transporters (SWEETs) were strongly affected by 2,4-D application. All screened differentially expressed SUTs (two in

10.3389/fpls.2023.1236315

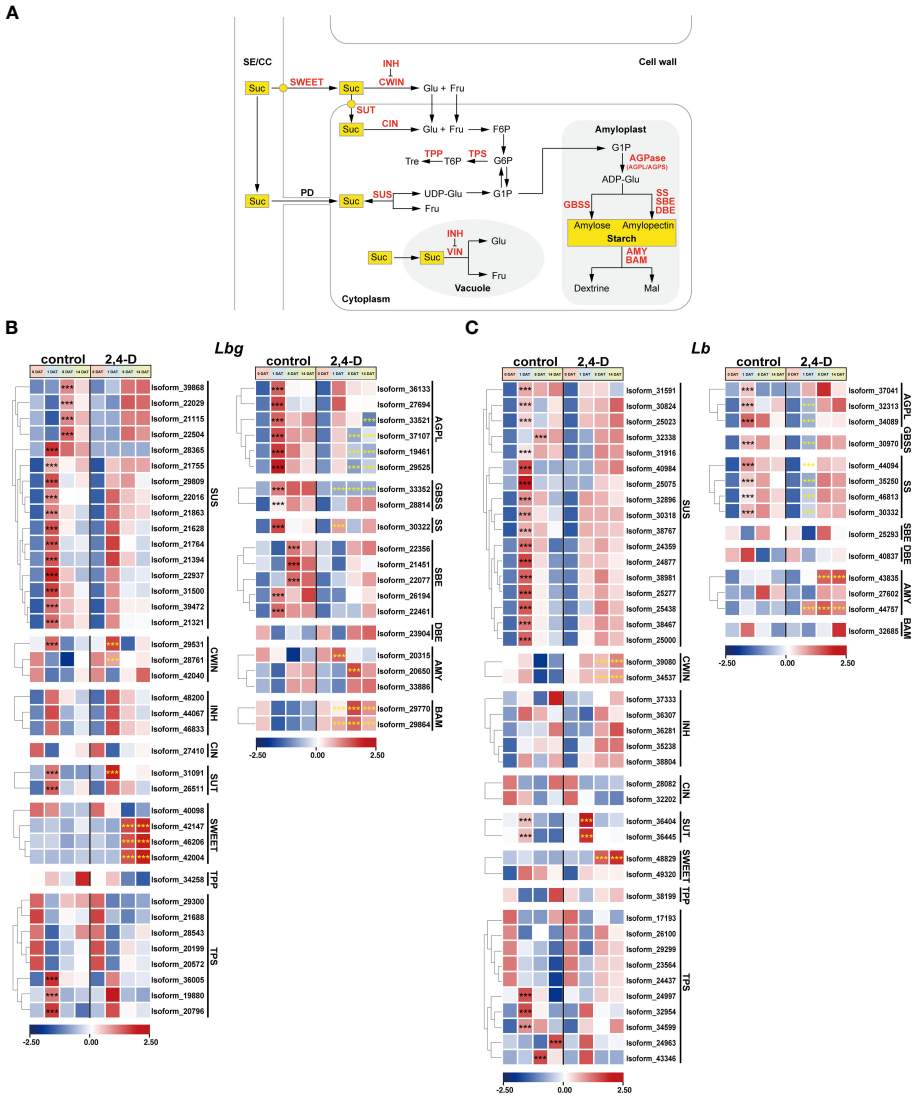


FIGURE 6
Expression patterns of stage-specific and 2,4-D related DETs involved in sucrose and starch metabolism pathway. **(A)** Sucrose and starch metabolism pathway. SE/CC, sieve element/companion cell complex; PD, plasmodesma; Suc, sucrose; Glu, glucose; Fru, fructose; Tre, trehalose; Mal, maltose; UDP-Glu, UDP-glucose; F6P, fructose-6-phosphate; G6P, glucose-6-phosphate; G1P, glucose-1-phosphate; ADP-Glu, ADP-Glucose; SWEET, SWEET sucrose-efflux transporter family; SUT, sucrose transporter; SUS, sucrose synthase; CWIN, cell wall invertase; VIN, vacuolar invertase; INH, invertase inhibitor; CIN, cytoplasmic invertase; TPP, trehalose 6-phosphate phosphatase; TPS, trehalose 6-phosphate synthase; AGPL/AGPS, large/small subunit of ADP-glucose pyrophosphorylase (AGPase); GBSS, granule-bound starch synthase; SS, starch synthase; SBE; starch branching enzyme; DBE, starch debranching enzyme; AMY, amylase, BAM, β-amylase. **(B, C)** Expression patterns of DETs involved in sucrose and starch metabolism in *Lbg* **(B)** and *Lb* **(C)**. ***Differences significant at $p < 0.001$. The black asterisk represents a significant difference compared to 0 DAT in the control group. The yellow asterisk represents a significant difference in the 2,4-D-treated group compared to the control group at the same stage.

Lbg and two in *Lb*) were significantly upregulated ($p < 0.001$) at 1 DAT, among them, the upregulation of one SUT in *Lbg* (Isoform 31091) and two SUTs (Isoform 36404 and 36445) in *Lb* was significantly enhanced ($p < 0.001$) by 2,4-D (Figures 6B, C). In addition, there was no significant change in the expression level of three SWEETs in *Lbg* (Isoform 42147, 46206 and 42002) and one in *Lb* (Isoform 48829) without 2,4-D during bulblet initiation, but these SWEETs significantly upregulated ($p < 0.001$) at 8 DAT with 2,4-D application (Figures 6B, C).

Several genes encoding key enzymes involved in starch synthesis were upregulated during bulblet initiation. All screened DETs encoding large subunit of AGPase (AGPL), Granule-Bound Starch Synthase (GBSS), and Starch Synthase (SS) were significantly upregulated ($p < 0.001$) at 1 DAT without 2,4-D, but some of these upregulations were inhibited (Isoform 33352 in *Lbg*), attenuated (Isoform 33521, 37101, 19461, 29525, 30322 in *Lbg* and Isoform 44094 in *Lb*) or delayed (Isoform 28814 in *Lbg* and Isoform 32313, 32089, 30970, 44094, 35250, 46813, 30332 in *Lb*) in the 2,4-D-treated group (Figures 6B, C), which indicated that 2,4-D suppressed starch synthesis during bulblet initiation. On the contrary, the expression levels of some Amylases (AMYs) (Isoform 20315, 20650 in *Lbg* and Isoform 43835, 44457 in *Lb*) and β-Amylases (BAMs) (Isoform 29770, 29864 in *Lbg*), which were involved in starch degradation, were significantly upregulated

($p < 0.001$) by 2,4-D (Figures 6B, C), indicating that 2,4-D promoted starch degradation. In *Lbg*, the starch content of scales was significantly decreased ($p < 0.05$) by 2,4-D at 1 DAT, 8 DAT and 14 DAT (Supplementary Figure S4). Similar decreases were also observed in *Lb*, but there were no significant differences ($p < 0.05$) (Supplementary Figure S4). In addition, we identified five significantly upregulated Starch Branching Enzymes (SBEs) ($p < 0.001$) in *Lbg* (Isoform 22356, 21451, 22077, 26194 and 22461) but not in *Lb* (Figures 6B, C).

3.7 Identification of potential regulators of *in vitro* bulblet initiation

In total, we identified 1209 and 1363 TFs from stage-specific DETs and 2,4-D-related DETs in *Lbg* and *Lb*, respectively. In stage-specific TFs, the number of MYB, bHLH and ERF ranked in the top three in *Lbg*, and MYB, bHLH and WRKY ranked in the top three in *Lb* (Figure 7A). In 2,4-D-related TFs, the three most were bHLH, GRAS and C3H in *Lbg*, and bHLH, C3H and ERF in *Lb* (Figure 7B). Given these, the expression patterns of TFs belonging to these TF families were analyzed, including 25 bHLHs, 24 GRASs, 22 C3Hs, 21 MYBs and 25 ERFs in *Lbg*, and 28 bHLHs, 25 C3Hs, 23 ERFs, 31 MYBs and 23 WRKYs in *Lb* (Figures 7C, D).

Further, the correlation analysis ($|r| \geq 0.8$. r, Pearson correlation coefficient) between these candidate TFs and DETs involved in auxin signaling and carbohydrate metabolism was conducted. In *Lbg*, we found six bHLHs, seven ERFs, eight MYBs and one GRAS were co-expressed with ten SUSs, two INHs, one CIN, three AGPLs and two PINs (Figure 7E). Eight bHLHs and five C3Hs were co-expressed with three SWEETs, two AMYs, one BAM, one SUS and one AUX1, six Aux/IAAs, five ARFs and five GH3s (Figure 7E). In *Lb*, the expression level of six MYBs and six bHLHs were co-expressed with three SUSs and one SWEET (Figure 7F). Eight C3Hs three ERFs and three bHLHs were co-expressed with an AMY and a SWEET (Figure 7F). Ten WRKYs, six MYBs, three bHLHs and one ARF were co-expressed (Figure 7F). Above all, the above-mentioned TFs are possibly involved in the regulation of auxin signaling and carbohydrate metabolism, thus regulate the process of *in vitro* bulblet initiation.

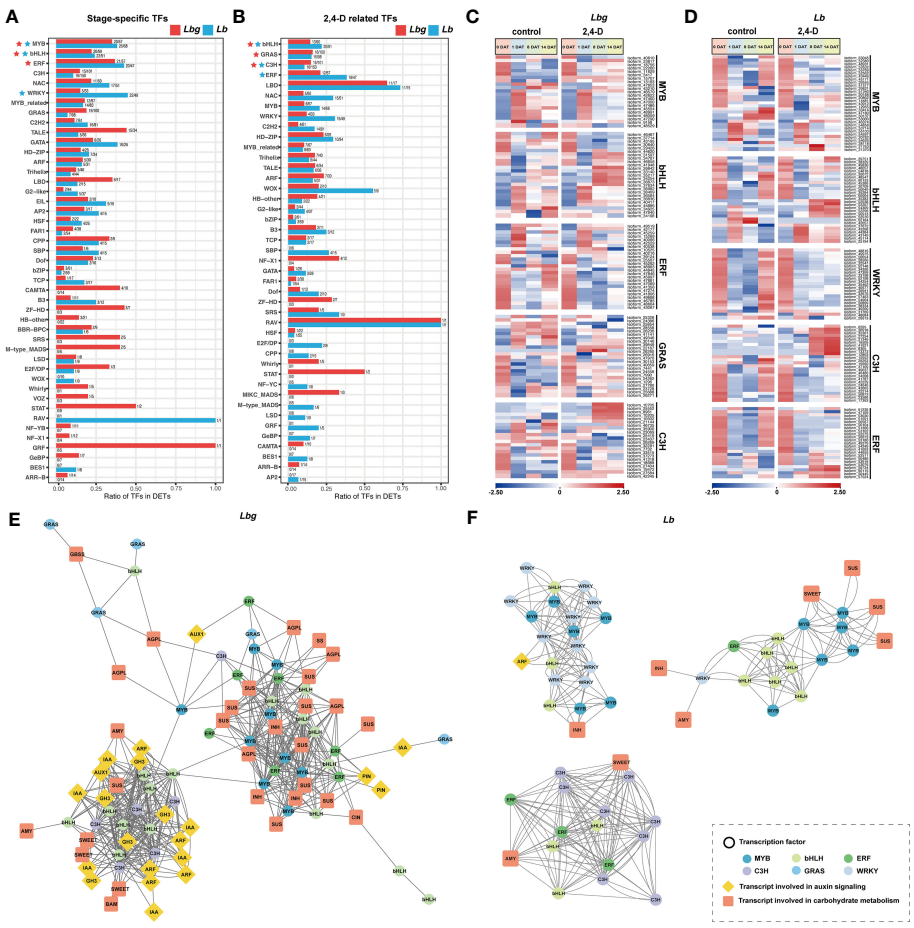


FIGURE 7
(A) TF families identified from the stage-specific DETs. **(B)** TF families identified from the 2,4-D-related DETs. The TF families in **(A)** and **(B)** rank according to the number of differentially expressed TFs they contained. **(C)** Express patterns of differentially expressed TFs belonging to MYB, bHLH, ERF, GRAS and C3H families in *Lbg*. **(D)** Express patterns of differentially expressed TFs belonging to MYB, bHLH, WRKY, C3H and ERF families in *Lb*. **(E, F)** Correlation analysis of DETs involved in auxin signaling and carbohydrate metabolism, and differentially expressed TFs in *Lbg* **(E)** and *Lb* **(F)**. The correlation analysis was conducted with Pearson's two-tailed test. DETs with significant correlations ($|r| \geq 0.8$. r, Pearson correlation coefficient) were linked. stering analysis of reference transcripts of *Lbg* **(A)** and *Lb* **(B)**.

4 Discussion

4.1 High-quality full-length transcripts of lily during *in vitro* bulblet initiation were constructed by a hybrid sequencing strategy

Tissue culture is a main asexual reproduction method for many bulbous crops and has a significant advantage in promoting regeneration efficiency and shortening the breeding and propagation cycle (Bakhshaie et al., 2016). Adventitious bud initiation and bulblet formation are critical steps during micropropagation of bulbous flowers, especially for direct organogenesis via shoot induction, which depends heavily on efficient nutritional allocation and hormone regulation (Xu et al., 2020a; Ren et al., 2022). Although many reports have been published on the process of bulblet formation and development in the lily (De Klerk, 2012; Li et al., 2014; Wu et al., 2021), there are relatively few reports on the detailed mechanism of early bulblet initiation.

In the present study, comprehensive full-length transcriptomes of *Lbg* and *Lb* were obtained through PacBio Iso-seq together with Illumina short-read sequencing during *in vitro* bulblet initiation. To date, Illumina sequencing technology has been widely applied to explore the process of bulblet formation and development in lily (Li et al., 2014; Du et al., 2017; Yang et al., 2017; Lazare et al., 2019). Recently, a full-length transcriptome of *Lilium* Oriental Hybrids 'Sorbonne' was generated (Li et al., 2022). Here, we conducted construct two high-quality full-length reference transcriptomes (*Lbg* and *Lb*). The N50 of transcripts of *Lbg* and *Lb* was 5,422 bp and 5,199 bp, respectively, and the mean length of transcripts of *Lbg* and *Lb* were 3,108 bp and 2,965 bp, respectively (Supplementary Table S1). These data contribute to further studies on molecular mechanisms of bulblet formation and other biological process of lily.

4.2 Exogenous 2,4-D application promote *in vitro* bulblet initiation by enhancing auxin signaling

Recent studies have indicated that auxin contributes to bulblet initiation in several bulbous flowers (Yang et al., 2017; Xu et al., 2020a). The content of IAA, the main naturally occurring auxin, increased consistently during the process of bulblet initiation and development in *Lycoris*, and exogenous IAA improved bulblet growth (Zhao, 2012; Xu et al., 2020a). In the present study, 1 mg L^{-1} 2,4-D significantly increase the regeneration rate of *in vitro* bulblet in *Lbg* and *Lb* (Figure 5A). Generally, auxin binds to TIR1 nuclear receptors, and then the auxin signal is modulated by the quantitative and qualitative responses of the Aux/IAAs and ARFs (Chapman and Estelle, 2009; Hayashi, 2012). ARFs are crucial regulators involved in the auxin signaling pathway, and then induce three major families: SAUR, GH3 and Aux/IAA genes (Guilfoyle and Hagen, 2007; Hayashi, 2012). In a previous study, differentially expressed TIRs, ARFs, and SAURs were identified during bulblet in *Lycoris* through transcriptome analysis (Xu et al., 2020a). Here, we found that 2,4-D affected the expression of genes involved in auxin signaling, thus promoted *in vitro* bulblet initiation. The obvious downregulations were found in many differentially expressed TIR1s at 1 DAT in both *Lbg* and *Lb*, which could be enhanced by 2,4-D application (Figures 5B, C). In addition, many ARFs, SAUR, GH3 and Aux/IAAs, which were not apparently responsive in the control group during bulblet initiation, were significantly upregulated in the 2,4-D-treated group (Figures 5B, C). Particularly, three ARFs in *Lbg* were significantly upregulated at 8 DAT in the control group (Figure 5B). Taken together, we proposed ARFs as key response factors during *in vitro* bulblet initiation in lily.

The distribution of auxin in cells depends largely on auxin transport, especially auxin efflux, which is directed by the polar subcellular localization of the PIN1 auxin efflux transporter in the plasma membrane (Vanneste and Friml, 2009; Hayashi, 2012). The expression and gradual polarization of PINs induced by auxin promote the formation of new vascular strands originating from the position of auxin application (Sauer et al., 2006). Three *PIN* genes were significantly upregulated during bulblet initiation in *Lycoris* (Xu et al., 2020a). In our study, all differentially expressed PINs were upregulated by 2,4-D in both *Lbg* and *Lb* (Figures 5B, C), indicating that 2,4-D might promote auxin transportation to promote bulblet initiation.

4.3 Starch and sucrose metabolism are crucial processes during *in vitro* bulblet initiation

The process of bulblet initiation involves carbohydrate transport from the source to the sink. The starch storage in the mother scales and exogenous carbon supply can be considered as the carbon source, and the basal of the mother scale, where the bulblets initiate, act as the sink tissue. Sucrose is unloaded from the phloem into sink cells either apoplasmically or symplasmically, then utilized to produce energy for cellular process (Ruan, 2014; Figure 6A).

Starch accounts for approximately 70% of the dry weight of lily bulbs (Wu et al., 2021). Starch synthesis is considered to be a crucial pathway for bulblet initiation. Several studies have indicated that whether a meristem can produce scale primordia depends on its capacity to accumulate starch (Bourque et al., 1987; Wu et al., 2021). In *Lycoris*, abscisic acid (ABA) upregulated the expression level of *LrSS1*, *LrSS2*, and *LrGBSS1* genes, which could enhance carbohydrate accumulation in the bulblets, thus promoted their development (Xu et al., 2020b). Similarly, starch synthesis was positively correlated with bulbil formation in *Lilium lancifolium* with upregulation of *AGPL*, *SS*, *GBSS* and *SBE* (Yang et al., 2017). In the present study, all screened DETs encoding AGPL, GBSS, and SS were upregulated at 1 DAT or 8 DAT without 2,4-D in both *Lbg* and *Lb* (Figures 6B, C), indicating the enhancement of starch synthesis process. Especially, five significantly upregulated SBEs were identified in *Lbg* (Figures 6B, C), indicating that *Lbg* might have a stronger ability of starch accumulation for bulblet initiation than *Lb*. During bulblet formation, starch is degraded in the mother scales and synthesized at the bulblet regeneration site and in the newly formed bulblets. In *Lilium*, the enzymes involved in starch

synthetic direction, such as AGPase, GBSS, SS, and SBE, showed a decreasing trend in mother scales but an increasing trend in bulblets during bulblet formation (Li et al., 2014; Wu et al., 2021). Moreover, starch content in basal scales and basal plates of *Lycoris* (the major sites of bulblet regeneration) showed a rapid decline during bulblet initiation in the efficient bulblet regeneration system (Ren et al., 2021). Here, we found in the 2,4-D-treated group, the more efficient group for *in vitro* bulblet initiation, key enzymes involved in starch synthesis (AGPL, SS and GBSS) were downregulated, while key enzymes involved in starch degradation (AMY and BAM) were upregulated compared to the control group (Figures 6B, C). In addition, 2,4-D reduced the starch content in the scales during bulblet initiation (Supplementary Figure S4). Taken together, we suggested that 2,4-D accelerate the starch degradation process to increase carbon supply for newly bulblet initiation.

Soluble sugars in mother scales were transported into the region where bulblets were initiated to supply the follow-up bulblet development (Xu et al., 2020a). Sucrose is the dominant transport form of sugars in higher plants (Lalonde et al., 1999; Rolland et al., 2006). SUS and CWIN are considered the most important sucrose hydrolases involved in the sucrose unloading pathway (Ruan, 2014; Figure 6A). SUS contribute to starch synthesis and accumulation, functioning during the later bulblet initiation and development (Yang et al., 2017; Wu et al., 2021). In *Lilium*, *SUS*s and *INV*s were both highly expressed in the mother scales and bulblets during bulblet emergence and swelling (Li et al., 2014). Similarly, *SUS*s and *INV*s were greatly upregulated accompanied by a decrease in sucrose content in mother scales during bulblet initiation in *Lycoris* (Xu et al., 2020a). In this study, 16 SUSs in *Lbg* and 17 SUSs in *Lb* were upregulated at 1 DAT or 8 DAT, and one CWIN in *Lbg* and two CWINs in *Lb* were upregulated at 1 DAT and then downregulated (Figures 6B, C). For example, *SUS* and *CWIN* often presented an opposite expression pattern during bulblet initiation, and this change was considered to be a possible sign of the transition from bulblet initiation to development (Ren et al., 2021; Wu et al., 2021). In particular, *CWINs* were highly expressed during the early bulblet initiation stage and produced glucose, which might act as sugar signaling rather than carbon resources (Wu et al., 2021). In *Lycoris*, the more highly *LsCWIN2* was expressed, the more bulblets were produced (Ren et al., 2021). Here, the expression levels of CWINs in *Lbg* and *Lb* were significantly increased by 2,4-D, showing a possible role of CWINs in increasing bulblet regeneration rate. Recent study showed that *LbgCWIN1* significantly upregulated endogenous starch was degraded during *in vitro* bulblet initiation in lily (Gao et al., 2023), indicating that *CWIN* can be selected as a candidate gene subsequent function verification.

Interestingly, the application of 2,4-D in the medium for bulblet induction had significant effects on the expression of genes involved in carbohydrate metabolism, especially SUTs and SWEETs (Figure 6A). SWEETs probably mediate sucrose efflux from SE/CC to apoplasm, and then sucrose can be taken up by SUTs, which are key steps proceeding phloem unloading (Ruan, 2014). The upregulation of one SUT in *Lbg* and two SUTs in *Lb* was significantly enhanced by 2,4-D at 1 DAT (Figures 6B, C). Three SWEETs in *Lbg* and one SWEET had no significant expression change in *Lb* without 2,4-D during bulblet initiation, but significantly upregulated at 8 DAT and 14 DAT with 2,4-D application (Figures 6B, C). Thus, 2,4-D might facilitate *in vitro* bulblet initiation mainly through promoting sucrose unloading from the SE/CC to the sink cells, and *SWEETs* and *SUTs* can be considered as good candidates for future functional studies.

4.4 Candidate TFs might be involved in the regulation of *in vitro* bulblet initiation

Although carbohydrate metabolism and the auxin signaling pathway have been respectively demonstrated to participate in bulblet initiation, their cooperative function during this process has not yet been reported. Here, we found that in *Lbg*, three SWEETs, two AMYs, one BAM and one SUS were co-expressed with one AUX1, six Aux/IAAs, five ARFs and five GH3s, and ten SUSs, two INHs, one CIN and three AGPLs were co-expressed with two PINs (Figure 7E), indicating the possible coregulation by carbohydrate metabolism and auxin signaling of *in vitro* bulblet initiation in lily. A recent study showed that bHLH, bZIP, WRKY, TCP, MYB, YABBY, NAC, C2H2 were identified to be involved in bulbil induction of 'Sorbonne' lily by coexpression analysis (Li et al., 2022). In *Lycoris*, coexpression analysis revealed that transcripts encoding WOX14, MYB117, and ULT1 coexpressed with *LsCWIN2*, and transcripts encoding ERFs were coexpressed with *LsSUS4* during bulb vegetative production (Ren et al., 2022). In the present study, MYB, bHLH, ERF, C3H, GRAS, WRKY families were identified to be candidate regulators of *in vitro* bulblet initiation. Among them, several TFs might be involved in the regulation the expression of key enzymes in carbohydrate metabolism during *in vitro* bulblet initiation in both *Lbg* and *Lb*, for example, C3Hs and bHLHs were co-expressed with AMYs and SWEETs, and some MYBs and bHLHs were co-expressed with SUSs in both *Lbg* and in *Lb* (Figures 7E, F). The above results may assist in the understanding of molecular mechanism of lily bulblet initiation.

Data availability statement

The data presented in the study are deposited in the GenBank repository, accession number PRJNA933000.

Author contributions

LZ, YW and ZMR designed the experiment. LZ, CG, YCX, YL and XX carried out the experiment. CG and LC analyzed the data.

CG drafted the manuscript. ZR, YW and YPX revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was financially supported by the Zhejiang Science and Technology Major Program on Agricultural New Variety Breeding (2021C02071-6) and the National Natural Science Foundation of China (Grant Nos. 32002071, 32101571).

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.2023.1236315/full#supplementary-material

SUPPLEMENTARY TABLE 1
Statistical summary of the reference transcriptome.

SUPPLEMENTARY TABLE 2
Primers of selected transcripts for quantitative real-time PCR validation.

References

Bakhshaie, M., Khosravi, S., Azadi, P., Bagheri, H., and van Tuyl, J. M. (2016). Biotechnological advances in *lilium*. *Plant Cell Rep.* 35 (9), 1799–1826. doi: 10.1007/s00299-016-2017-8

Bourque, J. E., Miller, J. C., and Park, W. D. (1987). Use of an *in vitro* tuberization system to study tuber protein gene expression. *In Vitro Cell. Dev. Biol.* 23 (5), 381–386. doi: 10.1007/BF02620996

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). EggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* 38 (12), 5825–5829. doi: 10.1093/molbev/msab293

Chapman, E. J., and Estelle, M. (2009). Mechanism of auxin-regulated gene expression in plants. *Annu. Rev. Genet.* 43 (1), 265–285. doi: 10.1146/annurev-genet-102108-134148

Chen, T., Zhang, H., Liu, Y., and Liu, Y. X. (2021). EVenn: Easy to create repeatable and editable Venn diagrams and Venn networks online. *J. Genet. Genomics* 48 (9), 863–866. doi: 10.1016/j.jgg.2021.07.007

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34 (17), i884–i890. doi: 10.1093/bioinformatics/bty560

De Klerk, G.-J. (2012). Micropropagation of bulbous crops: technology and present state. *Floricult. Ornamental Biotechnol.* 6, 1–8.

Du, F., Fan, J., Wang, T., Wu, Y., Grierson, D., Gao, Z., et al. (2017). Identification of differentially expressed genes in flower, leaf and bulb scale of *lilium* oriental hybrid 'Sorbonne' and putative control network for scent genes. *BMC Genomics* 18 (1), 899. doi: 10.1186/s12864-017-4303-4

Firon, N., LaBonte, D., Villordon, A., Kfir, Y., Solis, J., Lapis, E., et al. (2013). Transcriptional Profiling of Sweetpotato (*Ipomoea batatas*) Roots Indicates Down-regulation of Lignin Biosynthesis and Up-regulation of Starch Biosynthesis at an Early Stage of Storage Root Formation. *BMC Genomics* 14, 460. doi: 10.1186/1471-2164-14-460

Gao, C., Li, S., Xu, Y., Liu, Y., Xia, Y., Ren, Z., et al. (2023). Molecular cloning, characterization and promoter analysis of *LbgCWIN1* and its expression profiles in response to exogenous sucrose during in *vitro* bulblet initiation in lily. *Hortic. Plant J.* doi: 10.1016/j.hpj.2022.09.009

Guilfoyle, T. J., and Hagen, G. (2007). Auxin response factors. *Curr. Opin. Plant Biol.* 10 (5), 453–460. doi: 10.1016/j.pbi.2007.08.014

Guo, Z., Wang, F., Xiang, X., Ahammed, G. J., Wang, M., Onac, E., et al. (2016). Systemic induction of photosynthesis via illumination of the shoot apex is mediated sequentially by phytochrome B, auxin and hydrogen peroxide in tomato. *Plant Physiol.* 172 (2), 1259–1272. doi: 10.1104/pp.16.01202

Hayashi, K. (2012). The interaction and integration of auxin signaling components. *Plant Cell Physiol.* 53 (6), 965–975. doi: 10.1093/pcp/pcs035

Hu, X. G., Zhuang, H., Lin, E., Borah, P., Du, M., Gao, S., et al. (2022). Full-Length Transcriptome Sequencing and Comparative Transcriptomic Analyses Provide Comprehensive Insight into Molecular Mechanisms of Cellulose and Lignin Biosynthesis in *Cunninghamia lanceolata*. Front. *Plant Sci.* 13. doi: 10.3389/fpls.2022.883720

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 (D1), D309–D314. doi: 10.1093/nar/gky1085

Jin, J., Tian, F., Yang, D. C., Meng, Y. Q., Kong, L., Luo, J., et al. (2017). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.* 45 (D1), D1040–D1045. doi: 10.1093/nar/gkw982

Lalonde, S., Boles, E., Hellmann, H., Barker, L., Patrick, J. W., Frommer, W. B., et al. (1999). The dual function of sugar carriers: transport and sugar sensing. *Plant Cell* 11 (4), 707–726. doi: 10.1105/tpc.11.4.707

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nat. Methods* 9 (4), 357–359. doi: 10.1038/nmeth.1923

Lazare, S., Bechar, D., Fernie, A. R., Brotman, Y., and Zaccai, M. (2019). The proof is in the bulb: glycerol influences key stages of lily development. *Plant J.* 97 (2), 321–340. doi: 10.1111/tpj.14122

Lee, E., Yun, N., Jang, Y. P., and Kim, J. (2013). *Lilium lancifolium* thunb. Extract attenuates pulmonary inflammation and air space enlargement in a cigarette smoke-exposed mouse model. *J. Ethnopharmacol.* 149 (1), 148–156. doi: 10.1016/j.jep.2013.06.014

Li, G., Chen, Z., and Yan, F. (2007). *Lilium brownii* var. *giganteum*: A New Variety of *Lilium* from Wenling, Zhejiang. *J. Zhejiang Forest. Coll.* 06, 767–768.

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinf.* 12, 323. doi: 10.1186/1471-2105-12-323

Li, W., and Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22 (13), 1658–1659. doi: 10.1093/bioinformatics/btl158

Li, J., Sun, M., Li, H., Ling, Z., Wang, D., Zhang, J., et al. (2022). Full-length transcriptome-referenced analysis reveals crucial roles of hormone and wounding during induction of aerial bulbils in lily. *BMC Plant Biol.* 22 (1), 415. doi: 10.1186/s12870-022-03801-8

Li, X., Wang, C., Cheng, J., Zhang, J., da Silva, J. A. T., Liu, X., et al. (2014). Transcriptome Analysis of Sarbohydrate Metabolism during Bulblet Formation and Development in *Lilium davidii* var. *unicolor*. *BMC Plant Biol.* 14 (1), 358. doi: 10.1186/s12870-014-0358-4

Liu, B., Wang, X., Cao, Y., Arora, R., Zhou, H., and Xia, Y. (2020). Factors affecting freezing tolerance: a comparative transcriptomics study between field and artificial cold acclimations in overwintering evergreens. *Plant J.* 103, 2279–2300. doi: 10.1111/tpj.14899

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 (12), 550. doi: 10.1186/s13059-014-0550-8

Lv, X., Zhang, D., Min, R., Li, S., Li, Z., Ren, Z., et al. (2020). Effects of Exogenous Sucrose on Bulblet Formation of *Lycoris sprengeri in vitro*. *Acta Hortic.* 47 (8), 1475–1489. doi: 10.16420/j.issn.0513-353x.2019-0415

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., and Zdobnov, E. M. (2021). BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* 38 (10), 4647–4654. doi: 10.1093/molbev/msab199

Murashige, T., and Skoog, F. (1962). A revised medium for rapid growth and bio assays with tobacco tissue cultures. *Physiol. Plant* 15 (3), 473–497. doi: 10.1111/j.1399-3054.1962.tb08052.x

Ren, Z., Xia, Y., Zhang, D., Li, Y., and Wu, Y. (2017). Cytological analysis of the bulblet initiation and development in *lycoris* species. *Sci. Hortic.* 218, 72–79. doi: 10.1016/j.scienta.2017.02.027

Ren, Z., Xu, Y., Lvy, X., Zhang, D., Gao, C., Lin, Y., et al. (2021). Early sucrose degradation and the dominant sucrose cleavage pattern influence *lycoris sprengeri* bulblet regeneration *in vitro*. *Int. J. Mol. Sci.* 22 (21), 11890. doi: 10.3390/ijms222111890

Ren, Z., Zhang, D., Jiao, C., Li, D., Wu, Y., Wang, X., et al. (2022). Comparative transcriptome and metabolome analyses identified the mode of sucrose degradation as a metabolic marker for early vegetative propagation in bulbs of *lycoris*. *Plant J.* 112 (1), 115–134. doi: 10.1111/tpj.15935

Rolland, F., Baena-Gonzalez, E., and Sheen, J. (2006). Sugar sensing and signaling in plants: conserved and novel mechanisms. *Annu. Rev. Plant Biol.* 57, 675–709. doi: 10.1146/annurev.arplant.57.032905.105441

Ruan, Y. L. (2014). Sucrose metabolism: gateway to diverse carbon use and sugar signaling. *Annu. Rev. Plant Biol.* 65 (1), 33–67. doi: 10.1146/annurev-arplant-050213-040251

Sauer, M., Balla, J., Luschnig, C., Wisniewska, J., Reinöhl, V., Friml, J., et al. (2006). Canalization of auxin flow by aux/IAA-ARF-dependent feedback regulation of PIN polarity. *Genes Dev.* 20 (20), 2902–2911. doi: 10.1101/gad.390806

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 (11), 2498–2504. doi: 10.1101/gr.1239303

Van Aartrijk, J., and Blom-Barnhoorn, G. J. (1984). Adventitious bud formation from bulb-scale explants of *lilium speciosum* thunb. *In vitro* interacting effects of NAA, TIBA, wounding, and temperature. *J. Plant Physiol.* 116 (5), 409–416. doi: 10.1016/S0176-1617(84)80132-7

Vanneste, S., and Friml, J. (2009). Auxin: A trigger for change in plant development. *Cell* 136 (6), 1005–1016. doi: 10.1016/j.cell.2009.03.001

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., et al. (2021). ClusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation* 2 (3), 100141. doi: 10.1016/j.xinn.2021.100141

Wu, Y., Ma, Y., Li, Y., Zhang, L., and Xia, Y. (2019). Plantlet Regeneration from Primary Callus Cultures of *Lilium brownii* F.E.Br. Ex Miellez var. *giganteum* G. Y. Li and Z. H. Chen, a Rare Bulbous Germplasm. *In Vitro Cell. Dev. Biol.: Plant* 55 (1), 44–59. doi: 10.1007/s11627-018-09955-1

Wu, Y., Ren, Z., Gao, C., Sun, M., Li, S., Min, R., et al. (2020). Change in sucrose cleavage pattern and rapid starch accumulation govern lily shoot-to-bulblet transition in *vitro*. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.564713

Xu, J., Li, Q., Li, Y., Yang, L., Zhang, Y., and Cai, Y. (2020b). Effect of exogenous gibberellin, paclobutrazol, abscisic acid, and ethrel application on bulblet development in *lycoris radiata*. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.615547

Xu, J., Li, Q., Yang, L., Li, X., Wang, Z., and Zhang, Y. (2020a). Changes in Carbohydrate Metabolism and Endogenous Hormone Regulation during Bulblet Initiation and Development in *Lycoris radiata*. *BMC Plant Biol.* 20, 180. doi: 10.1186/s12870-020-02394-4

Xu, L., Yang, P., Feng, Y., Xu, H., Cao, Y., Tang, Y., et al. (2017). Spatiotemporal transcriptome analysis provides insights into bicolor tepal development in *lilium* "Tiny padhye". *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.00398

Yang, P., Xu, L., Xu, H., Tang, Y., He, G., Cao, Y., et al. (2017). Histological and Transcriptomic Analysis during Bulbil Formation in *Lilium lancifolium*. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.01508

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T. L. (2012). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinf.* 13, 134. doi: 10.1186/1471-2105-13-134

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 05 December 2023
DOI 10.3389/fpls.2023.1297228

OPEN ACCESS

EDITED BY
Xueqiang Wang,
Zhejiang University, China

REVIEWED BY
Mehdi Mansouri,
Shahid Bahonar University of Kerman, Iran
Muhammad Waseem,
Hainan University, China

*CORRESPONDENCE
Yanfang Wang
wyf863@126.com
Qin Wang
wangqin@bzmc.edu.cn
Yanhong Zhao
zyhbob@163.com

RECEIVED 19 September 2023
ACCEPTED 20 November 2023
PUBLISHED 05 December 2023

CITATION
Liu H, Wang X, Yang W, Liu W, Wang Y, Wang Q and Zhao Y (2023) Identification of Whirly transcription factors in Triticeae species and functional analysis of *TaWHY1-7D* in response to osmotic stress. *Front. Plant Sci.* 14:1297228. doi: 10.3389/fpls.2023.1297228

Identification of Whirly transcription factors in Triticeae species and functional analysis of *TaWHY1-7D* in response to osmotic stress

Hao Liu[1], Xiaoyu Wang[2], Wenbo Yang[3], Wenyan Liu[1], Yanfang Wang[4]*, Qin Wang[5]* and Yanhong Zhao[1]*

[1]College of Agriculture, Ludong University, Yantai, China, [2]College of Chemical and Biological Engineering, Shandong University of Science and Technology, Qingdao, China, [3]Institute of Cereal Crops, Henan Academy of Agricultural Sciences, Zhengzhou, China, [4]College of Life Science, Ludong University, Yantai, China, [5]Department of Biochemistry and Molecular Biology, Binzhou Medical University, Yantai, China

Osmotic stress poses a threat to the production and quality of crops. *Whirly* transcription factors have been investigated to enhance stress tolerance. In this study, a total of 18 *Whirly* genes were identified from six Triticeae species, which were classified into *Whirly1* and *Whirly2*. The exon–intron structure, conserved motif, chromosomal location, collinearity, and regulatory network of *Whirly* genes were also analyzed. Real-time PCR results indicated that *TaWHY1* genes exhibited higher expression levels in leaf sheaths and leaves during the seedling stage, while *TaWHY2* genes were predominantly expressed in roots. Under PEG stress, the expression levels of *TaWHY1-7A*, *TaWHY2-6A*, *TaWHY2-6B*, and *TaWHY2-6D* were increased, *TaWHY1-7D* was reduced, and *TaWHY1-4A* had no significant change. All *TaWHY* genes were significantly up-regulated in response to NaCl stress treatment. In addition, TaWHY1-7A and TaWHY1-7D mainly enhanced the tolerance to oxidative stress in yeast cells. TaWHY2s mainly improved NaCl stress tolerance and were sensitive to oxidative stress in yeast cells. All TaWHYs slightly improved the yeast tolerance to D-sorbitol stress. The heterologous expression of *TaWHY1-7D* greatly improved drought and salt tolerance in transgenic *Arabidopsis*. In conclusion, these results provide the foundation for further functional study of *Whirly* genes aimed at improving osmotic stress tolerance in wheat.

KEYWORDS

Triticeae species, wheat, *Whirly* gene, gene expression, osmotic stress

Introduction

Wheat (*Triticum aestivum* L.) is one of the most important staple crops worldwide and a major source of calories for the expanding world population. As a sessile organism, wheat has to suffer from a variety of adverse conditions during the growth and development stages, such as drought and salinization, which contribute to a great reduction in the overall wheat yield and quality (Gupta et al., 2020). Therefore, mining stress-resistant genes and developing improved varieties are the most important strategies to improve wheat yield and quality.

Whirly (WHY) proteins are plant-specific transcription factors binding to single-stranded DNA (ssDNA) to modulate growth and defense responses and located in the chloroplasts, mitochondria, and nucleus (Desveaux et al., 2005; Krupinska et al., 2022; Taylor et al., 2022). Whirly domain consists of four structural topologies, which are characterized by two antiparallel four-stranded β-sheets stabilized by a C-terminal helix-loop-helix motif (Desveaux et al., 2005; Cappadocia et al., 2013; Taylor et al., 2022). Due to the structural similarity with "whirligig," Whirly transcription factors are named Whirly (Desveaux et al., 2005). The conserved "KGKAAL" motif in the Whirly domains exists extensively in higher plants, which participate in binding to ssDNA and hexamerization of the tetramers forming hollow sphere structures of 12 nm in diameter (Desveaux et al., 2002; Cappadocia et al., 2012). Additionally, Whirly proteins contain a conserved cysteine residue, which might play a vital role in the formation of disulfide bridges between two Whirly proteins (Foyer et al., 2014).

Whirly was initially identified as p24/PBF-2 protein that binds to the elicitor response element (ERE) on the promoter of the pathogen response gene *PR-10a* in potato (Desveaux et al., 2000). In *Arabidopsis*, AtWHY1 is targeted to chloroplasts and nucleus (Krause et al., 2005; Ren et al., 2017), which plays a crucial role in regulating telomere length homeostasis (Yoo et al., 2007), maintaining the stability of plastid genome (Marechal et al., 2009), modulating reactive oxygen species (ROS) homeostasis, controlling leaf senescence (Lin et al., 2019), and responding to salicylic acid (SA)-dependent disease resistance (Desveaux et al., 2004). AtWHY1 protein represses the expression of *WRKY53* and delays leaf senescence in *Arabidopsis* (Miao et al., 2013). AtWHY2 is located in the mitochondria and nucleus (Krause et al., 2005; Golin et al., 2020). Overexpression of *AtWHY2* leads to mitochondrial dysfunction, early accumulation of senescence-related transcripts (Marechal et al., 2008; Golin et al., 2020), slower growth of pollen tubes, elevation of mtDNA content, and ROS levels in pollen (Cai et al., 2015). AtWHY3 is targeted to chloroplasts, mitochondria, and nucleus in compensation for the lack or mutation of AtWHY1 and AtWHY2 (Krause et al., 2005; Golin et al., 2020). In tomato (*Solanum lycopersicum*), *SlWHY1* and *SlWHY2* can be induced by drought and salt stresses (Akbudak and Filiz, 2019). Overexpression of *SlWHY1* enhances heat and cold stress tolerance and reduces ROS levels in tomato (Zhuang et al., 2020a; Zhuang et al., 2020b), and SlWHY2 can maintain mitochondrial function under drought stress through interacting with SlRECA2 in tomato (Meng et al., 2020). MeWHY1/2/3 can interact with MeCIPK23 to activate abscisic acid (ABA) biosynthesis and regulate drought resistance in cassava (*Manihot esculenta*) (Yan et al., 2020). In barley (*Hordeum vulgare* L.), overexpression of HvWHY1 delays drought-induced leaf senescence (Manh et al., 2023).

Whirly genes have been identified in various plant species, such as *Arabidopsis*, strawberry, tomato, cassava, and barley (Desveaux et al., 2005; Janack et al., 2016; Yan et al., 2020; Hu and Shu, 2021), however, a comprehensive genome-wide analysis of *Whirly* genes in Triticeae species has not been investigated. In this study, a genome-wide analysis of *Whirly* genes was performed in Triticeae species including *Triticum aestivum*, *Triticum urartu*, *Triticum dicoccoides*, *Aegilops tauschii*, *Hordeum vulgare*, and *Secale cereale* to characterize their sequences, gene structures, evolutionary relationships, expression patterns, and stress tolerance under osmotic stress. These results will provide a valuable foundation for further functional investigations of *Whirly* genes in response to osmotic stress.

Materials and methods

Plant material and growth conditions

Bread wheat cv. Chinese Spring preserved in our laboratory was used in this study, and the sterilized bread wheat seeds were soaked with ddH_2O in dark and 4°C condition overnight, then cultured on filter paper wetted with ddH_2O in a culture room at 25/18°C with 16-h light/8-h dark for 1 week. Next, 7-day-old bread wheat seedlings with uniform leaf size and root length were selected for subsequent experiments. For drought and salt stress treatments, 7-day-old bread wheat seedlings were cultured under 20% PEG6000 (w/v) and 300 mM NaCl treatments, respectively. In each treatment, the root, leaf sheath, and leaf tissues were collected at 0 h, 1 h, and 6 h, then frozen in liquid nitrogen and stored at −80°C for further investigation.

Genome-wide identification of *Whirly* gene family

The protein sequences of *Triticum aestivum* (Chinese Spring, IWGSC.52), *Triticum urartu* (Tu 2.0), *Triticum dicoccoides* (WEWSeq_v1.0), *Aegilops tauschii* (Aet_v4.0), *Hordeum vulgare* (IBSC_v2), *Brachypodium distachyon* (IBI_v3.0), *Oryza sativa* (Japonica, IRGSP 1.0), *Zea mays* (B73 RefGen_v4), *Solanum lycopersicum* (SL3.0), and *Arabidopsis thaliana* (TAIR10) were downloaded from EnsemblPlants database (http://plants.ensembl.org/index.html). The protein sequence data of *Secale cereale* (Weining v1) was acquired from the China National Center for Bioinformation (CNCB-NGDC Members and Partners, 2022). To identify candidate Whirly protein sequences, the Hidden Markov Model (HMM) profile of the typical Whirly transcription factor domain (PF08536) (Mistry et al., 2021) was used as a query to search against the protein sequences of these 11 plant species with TBtools software (Chen et al., 2020a). Next, the Pfam (https://www.ebi.ac.uk/interpro/) (Mistry et al., 2021) and

SMART (Simple Modular Architecture Research Tool, http://smart.embl.de/) (Letunic et al., 2021) online services were used to further confirm the putative Whirly proteins. The protein length, molecular weight, isoelectric point (*p*I), and grand average of hydropathy (GRAVY) of the Whirly proteins were analyzed by WheatOmics 1.0 (Ma et al., 2021).

Multiple sequence alignment and phylogenetic tree construction

Multiple sequence alignment of Whirly amino acid sequences was performed with ClustalW using the default options in MEGA 11 (Tamura et al., 2021) and visualized by ESPript 3.0 (Gouet et al., 2003). The phylogenetic tree was constructed by using the neighbor-joining (NJ) method with 1,000 bootstrap replicates in MEGA 11 software (Tamura et al., 2021) and visualized by Evolview service (Subramanian et al., 2019).

Gene structure, conserved motif, domain, and 3D structure analyses

The exon–intron structures of *Whirly* genes were analyzed based on TGT (Triticeae-Gene Tribe) (Chen et al., 2020b). The conserved motifs and domains of Whirly family proteins were annotated using the MEME program (Bailey et al., 2009) and SMART website (Letunic et al., 2021) and visualized by TBtools (Chen et al., 2020a). The Swiss-Model program was used to predict the three-dimensional (3D) structure of Whirly proteins (Waterhouse et al., 2018).

Chromosome localization, gene duplication, and micro-collinearity analysis

The chromosome localization, micro-collinearity, and paralogous/orthologous gene pairs of *Whirly* genes were identified by using Triticeae-Gene Tribe (TGT) (Chen et al., 2020b). The gene duplication events were determined by Multiple Collinear Scanning Toolkits (MCScanX) (Wang et al., 2012). TBtools was used to calculate the nonsynonymous rate (K_a), synonymous rate (K_s), and the nonsynonymous and synonymous substitution ratio (K_a/K_s) values of the paralogous gene pair with the Nei–Gojobori (NG) method (Chen et al., 2020a).

Regulatory network analysis

The upstream transcription factors and downstream target genes of *TaWHYs* were predicted by using the wheat integrative gene regulatory network (wGRN) (Chen et al., 2023). Protein–protein interactions (PPIs) were analyzed using the STRING database (Von Mering et al., 2003) and WheatOmics 1.0 (Ma et al., 2021).

Gene expression analysis by RNA-seq data

To investigate the gene expression patterns in bread wheat under drought stress, bread wheat cv. Chinese Spring was planted in a growth chamber under a photoperiod of 16 h/8 h (light/dark). For drought stress, the seedlings were subjected to water-deficit condition during the seedling stage. The leaf tissues were harvested after 0 days, 2 days, 6 days, and 10 days of treatment, and the total RNA of all the collected samples was extracted. A Nanodrop2000 spectrophotometer was used to determine the quantity and quality of the RNA. A total of 12 bread wheat samples (three biological replicates were conducted for each treatment) were sequenced at Majorbio Bio-Pharm Technology Co. Ltd. (Shanghai, China), and paired-end sequencing was performed with an Illumina Novaseq 6000. The transcriptome data have been submitted to NCBI (BioProject ID: PRJNA1003680).

The transcriptome data of different bread wheat tissues (root and shoot) were obtained from NCBI SRA (DRX002485, DRX002486, DRX002487, DRX002491, DRX002492, and DRX002493). The transcriptome data SRX9781249, SRX9781250, SRX9781251, SRX9781252, SRX9781253, SRX9781254, SRX9781255, SRX9781256, SRX9781257, SRX9781258, SRX9781259, and SRX9781260 were used to analyze the gene expression profiles under NaCl stress in leaves during bread wheat seedling stage.

RNA extraction and real-time PCR

Real-time PCR was performed to detect the expression pattern of *Whirly* genes according to a previous study (Liu et al., 2022). The total RNA was isolated using RNApure Plant Kit (CWBIO), and the first-strand cDNA was synthesized from 1 μg of total RNA using Prescript III RT ProMix (CISTRO). The real-time PCR was performed using gene-specific primers (Supplementary Table S1) with 2× Ultra SYBR Green qPCR Mix (CISTRO), and the *TaActin* gene was selected as a reference control. The real-time PCR cycling parameters were 95°C for 30 s, followed by 45 cycles at 95°C for 5 s and 60°C for 30 s, with a melting curve analysis. All reactions were performed on three technical and biological replicates. The relative expression levels of target genes were calculated using the $2^{-\triangle\triangle CT}$ method (Livak and Schmittgen, 2001).

Stress tolerance assay in yeast cells

The coding sequences (CDS) of *Whirly* genes were cloned into a pGADT7 vector using the ClonExpress II One Step Cloning Kit (Vazyme, Nanjing), then transformed into *Saccharomyces cerevisiae* (*S. cerevisiae*) BY4741 or its stress-sensitive mutant BY4741 (Δ*hog1*). The primers are shown in Supplementary Table S1. For osmotic and oxidative stress, the yeast cells Δ*hog1* carrying the recombinant vector pGADT7-*TaWHY2-6A*/*TaWHY2-6B*/*TaWHY2-6D*/*TaWHY1-7A*/*TaWHY1-7D* were cultured in YPD liquid medium (1% yeast extract, 2% peptone, and 2% glucose) at 30°C until density reached an OD_{600} of 1.0, then serially diluted (10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}) with ddH_2O. The

cells were spotted onto YPD medium plates (1% yeast extract, 2% peptone, 2% glucose, and 2% agar) containing 1.2 M D-sorbitol, 0.4 M NaCl, or 4.0 mM H_2O_2 and cultured at 30°C for 3–5 days. The wild-type yeast cells BY4741 and stress-sensitive mutant Δ*hog1* carrying the empty vector pGADT7 were used as positive and negative controls, respectively.

Drought and salt tolerance assay in *Arabidopsis*

The coding sequences of *TaWHY1-7D* were cloned into the pCAMBIA3301-GFP vector, then transformed into *Agrobacterium tumefaciens* EHA105, and generated *35S:TaWHY1-7D* transgenic *Arabidopsis* lines via the floral dip method. The primers are shown in Supplementary Table S1. The transgenic *Arabidopsis* lines were selected via spraying 0.5‰ Basta solution. For drought and salt tolerance assays, WT and *35S:TaWHY1-7D* transgenic *Arabidopsis* were treated with drought (water-deficit) and 500 Mm NaCl conditions.

Results

Genome-wide identification and phylogenetic relationship analysis of *Whirly* genes

A total of 29 *Whirly* genes were identified from the protein sequences of 11 plant species via a hidden Markov model (HMM) search. In total, 24 *Whirly* genes were identified from nine monocotyledon species, comprising six Triticeae species (*T. aestivum* (6), *T. urartu* (2), *T. dicoccoides* (4), *Ae. tauschii* (2), *H. vulgare* (2), and *S. cereal* (2)) and three other monocotyledon species (*B. distachyon* (2), *O. sativa* (2), and *Z. mays* (2)), while five *Whirly* genes were identified in two dicotyledon species, including *S. lycopersicum* (2) and *A. thaliana* (3) (Figure 1A; Supplementary Table S2). To further confirm the reliability of the identified *Whirly* genes, the expression of *Whirly* genes was analyzed in *T. urartu*, *T. dicoccoides*, *S. cereale*, *H. vulgare*, and *T. aestivum* based on previous published transcriptomic data (Supplementary Table S3). The length of the identified 29 Whirly proteins varied from 223 (*HvWHY2-6H*) to 286 (*ZmWHY1*) amino acid residues, with the molecular weights ranging from 24.24 to 31.71 kDa. The *p*I values ranged from 8.84 (*TdWHY1-4A*) to 10.81 (*SlWHY2*), with the calculated grand average of hydrophilic index (GRAVY) varying from −0.207 (*AtWHY1*) to −0.459 (*TaWHY1-4A*), suggesting that these 29 *Whirly* genes encoded highly hydrophilic proteins (Supplementary Table S2).

To elucidate the evolutionary relationship of *Whirly* genes, a phylogenetic tree was constructed using these 29 Whirly proteins (Figure 1A). According to the results, *Whirly* genes were classified into two categories, named group 1 (Whirly1) and group 2 (Whirly2). Bread wheat *T. aestivum* (AABBDD, hexaploid) has undergone two rounds of natural hybridization events (Levy and Feldman, 2022). Thus, the number of gene family members in *T. aestivum* (AABBDD) is approximately 1.5- and 3-fold than that in *T. dicoccoides* (AABB, tetraploid) and other diploid Triticeae species, respectively. Consistently, three *Whirly1* or *Whirly2* genes

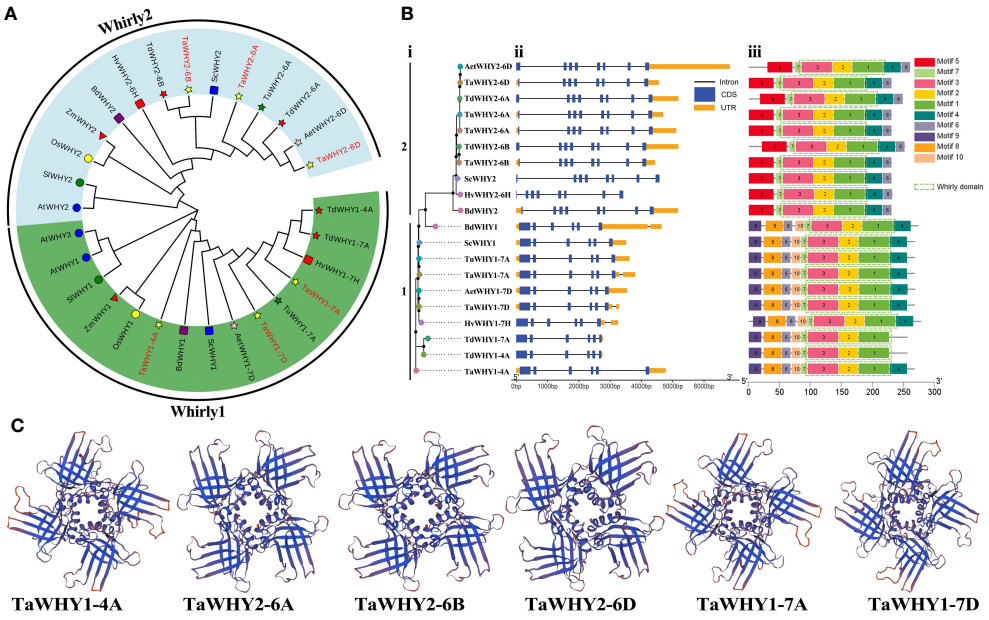


FIGURE 1
The neighbor-joining (NJ) phylogenetic tree **(A)**, gene structures **(B)**, and 3D structures **(C)** of Whirly proteins. **(A)** The tree was constructed using Whirly protein sequences from *T. aestivum* (Ta), *T. urartu* (Tu), *T. dicoccoides* (Td), *Ae. tauschii* (Aet), *H. vulgare* (Hv) and *S. cereal* (Sc), *B. distachyon* (Bd), *O. sativa* (Os) and *Z. mays* (Zm), *S. lycopersicum* (Sl), and *A. thaliana* (At) with bootstrap values of 1,000 replicates. Different groups of Whirly proteins are marked by different colors. **(B)** Phylogenetic classification (i), exon–intron structure (ii), and conserved domain (iii) analyses of *Whirly* genes in Triticeae species. **(C)** The Swiss Model program was used to predict the three-dimensional (3D) structure of the Whirly proteins.

were found in *T. aestivum*, while *T. dicoccoides* and other diploid Triticeae included two and one *Whirly1* or *Whirly2* gene, respectively (Figure 1A; Supplementary Table S2).

Gene structure and conserved motif analysis

To investigate the functional divergence of *Whirly* genes, the exon–intron structures, conserved motifs, and 3D structures of *Whirly* genes were analyzed in Triticeae species (Figure 1; Supplementary Figure S1). The results revealed that *Whirly1* and *Whirly2* genes contained six and eight exons in the Triticeae species, respectively. The conserved motif analysis showed that all Whirly proteins contained the Whirly transcription factor domain (PF08536), which consisted of motifs 1, 2, 3, and 7. These also confirmed the reliability of the identified *Whirly* gene family members. Motif 3 contained the conserved "KGKAAL" sequence, which participated in binding to ssDNA (Supplementary Figure S1) (Desveaux et al., 2002; Cappadocia et al., 2012). Almost all Whirly proteins contained motif 4, except for TdWHY1-4A and TdWHY1-7A, which lacked a portion of the amino acid sequences of motif 4 (Supplementary Figure S1). Motifs 8, 9, and 10 were present in group 1 members, while they were absent in group 2 members. Motif 5 was unique to group 2 members. In addition, all TaWHY proteins contained two anti-parallel four-stranded β-sheets that extend like blades from an α-helical core (Figure 1C), which were consistent with its "whirligig" structure (Desveaux et al., 2005).

Chromosomal location, collinearity, and K_a/K_s analysis of *Whirly* genes

The distribution of *Whirly* genes on the chromosome in six Triticeae species (*T. aestivum*, *T. urartu*, *T. dicoccoides*, *Ae. tauschii*, *H. vulgare*, and *S. cereal*), three other monocotyledon species (*B. distachyon*, *O. sativa*, and *Z. mays*), and two dicotyledon species (*S. lycopersicum* and *A. thaliana*) are shown in Supplementary Table S2. In *T. aestivum* (AABBDD, hexaploid), *Whirly1* genes were distributed on chromosomes 4A (*TaWHY1-4A*), 7A (*TaWHY1-7A*), and 7D (*TaWHY1-7D*). *Whirly2* genes had three copies in its subgenomes A, B, and D, i.e., *TaWHY2-6A*, *TaWHY2-6B*, and *TaWHY2-6D* were distributed on chromosomes 6A, 6B, and 6D, respectively (Figure 2). Similarly, *TdWHY1-4A*, *TdWHY2-6A*, *TdWHY2-6B*, and *TdWHY1-7A* were located on chromosomes 4A, 6A, 6B, and 7A in *T. dicoccoides* (AABB, tetraploid), respectively. *AetWHY2-6D* and *AetWHY1-7D* were distributed on chromosomes 6D and 7D in *Ae. tauschii* (DD, diploid), respectively. *TuWHY2-6A* and *TuWHY1-7A* were located on chromosomes 6A and 7A in *T. urartu* (AA, diploid), respectively. *HvWHY2-6H* and *HvWHY1-7H* were located on chromosomes 6H and 7H in *H. vulgare* (HH, diploid), respectively. In *S. cereale* (RR, diploid), *ScWHY1* and *ScWHY2* were distributed on chromosomes 1R and 6R, respectively. Interestingly, the orthologous genes of *TaWHY1-4A* were not distributed on chromosomes 4A in *T. urartu* and 4H in *H. vulgare*, whereas *TdWHY1-4A* existed on chromosome 4A of *T. dicoccoides* (Supplementary Table S2). This result suggested that the expansion events of *Whirly* genes occurred through hybridization and polyploidization.

To further investigate the evolutionary process of *TaWHYs*, gene duplication, and micro-collinearity analyses of the *Whirly* genes were performed (Figure 3; Supplementary Table S4). A total of six paralogous gene pairs of *TaWHYs* (*TaWHY1-4A*/*TaWHY1-7A*/*TaWHY1-7D*, and *TaWHY2-6A*/*TaWHY2-6B*/*TaWHY2-6D*) were identified in bread wheat genome and expanded by whole-genome duplication (WGD) or segmental duplication events (Figure 2B; Supplementary Table S4). The K_a/K_s values of paralogous gene pairs were all less than 1, indicating that *TaWHY* genes underwent purifying selection to avoid functional divergence (Supplementary Table S4). Micro-collinearity analysis contributes to the investigation of the inheritance and variation of specific genes in local regions and detecting the origin of specific genes during the hybridization and polyploidization process (Chen et al., 2020b). To explore the origin of *Whirly* genes in Triticeae species, *TaWHY1-4A*, *TaWHY2-6A*, and *TaWHY1-7A* were used as query genes to analyze the micro-collinearity by TGT (Figure 3). The homologous genes of *TaWHY2-6A* were detected in the collinearity regions of *T. urartu*, *Ae. tauschii*, subgenomes A and

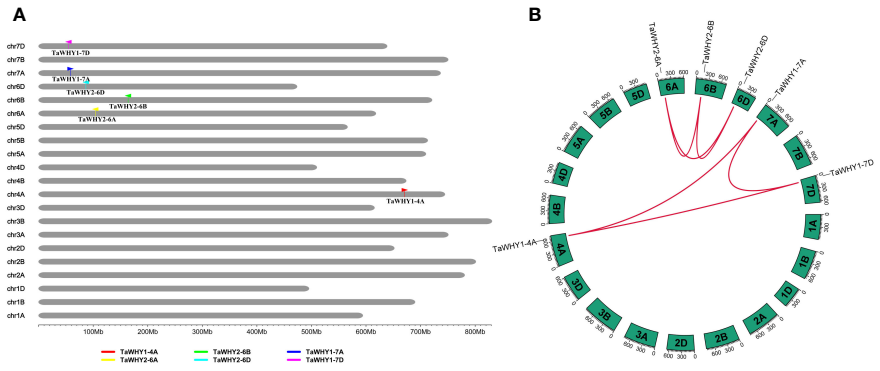


FIGURE 2
Chromosomal localizations **(A)** and syntenic relationships **(B)** among *TaWHY* genes in *T. aestivum*. **(B)** Red lines in the highlight indicate the syntenic *TaWHY* gene pairs.

10.3389/fpls.2023.1297228

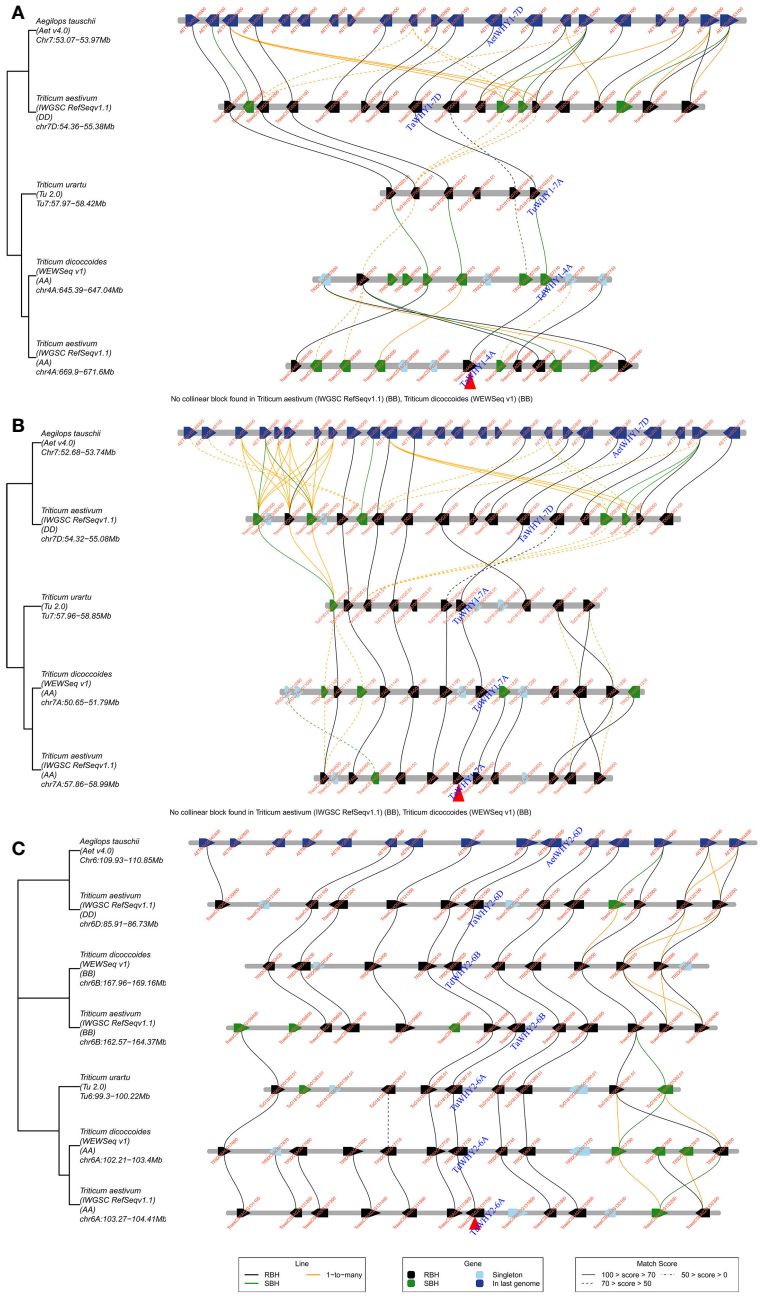


FIGURE 3
Micro-collinearity analysis of the *Whirly* gene in Triticeae species. *TaWHY1-4A* **(A)**, *TaWHY1-7A* **(B)**, and *TaWHY2-6A* **(C)** were used as the query gene, respectively.

B of *T. dicoccoides*, and subgenomes B and D of *T. aestivum*, suggesting that the *Whirly2* genes and their adjacent genes in the collinearity regions were relatively conserved during evolutionary processes in Triticeae species. However, no homologous genes of *TaWHY1-4A* and *TaWHY1-7A* were found in the collinearity regions of subgenome B of *T. dicoccoides*, and subgenome B of *T. aestivum*. In addition, the homologous genes of *TaWHY1-4A* were present in the collinearity regions on chromosome 7A of *T. urartu*, and 7D of *Ae. tauschii*, and 7D of *T. aestivum*, but absent on chromosome 4 of *T. urartu*, suggesting that *TaWHY1-4A* and *TdWHY1-4A* might originate from *TuWHY1-7A* or *AetWHY1-7D*.

Expression patterns analysis of *TaWHYs*

To insight into the biological function of *TaWHY* genes, the transcriptome data and real-time PCR were used to determine the expression patterns of six *TaWHY* genes in different tissues (leaves, leaf sheaths, and roots during bread wheat seedling stage) and in response to osmotic (drought and salt) stress (Figures 4, 5). The analysis of the transcriptome data revealed that the *TaWHY1* genes (*TaWHY1-4A*, *TaWHY1-7A*, and *TaWHY1-7D*) exhibited the highest expression levels in leaves, and the *TaWHY2* genes (*TaWHY2-6A*, *TaWHY2-6B*, and *TaWHY2-6D*) showed the

highest expression levels in roots (Figure 4A). Consistently, the real-time PCR results showed that *TaWHY1* genes (*TaWHY1-4A*, *TaWHY1-7A*, and *TaWHY1-7D*) were highest expressed in leaf sheaths, followed by leaves, and roots during the bread wheat seedling stage. *TaWHY2* genes (*TaWHY2-6A*, *TaWHY2-6B*, and *TaWHY2-6D*) exhibited the highest expression level in roots, followed by leaf sheaths, and finally in leaves (Figure 4B).

After drought stress treatment, RNA-seq analysis revealed that the *TaWHY1* genes exhibited the highest expression levels after 6 days of drought treatment, and the expression of *TaWHY2* genes increased with the progression of drought stress duration (Figure 5A). The real-time PCR results demonstrated the expression of *TaWHY1-7A* was up-regulated under PEG stress, peaking at 1 h with 1.6-fold compared with the control, *TaWHY1-*

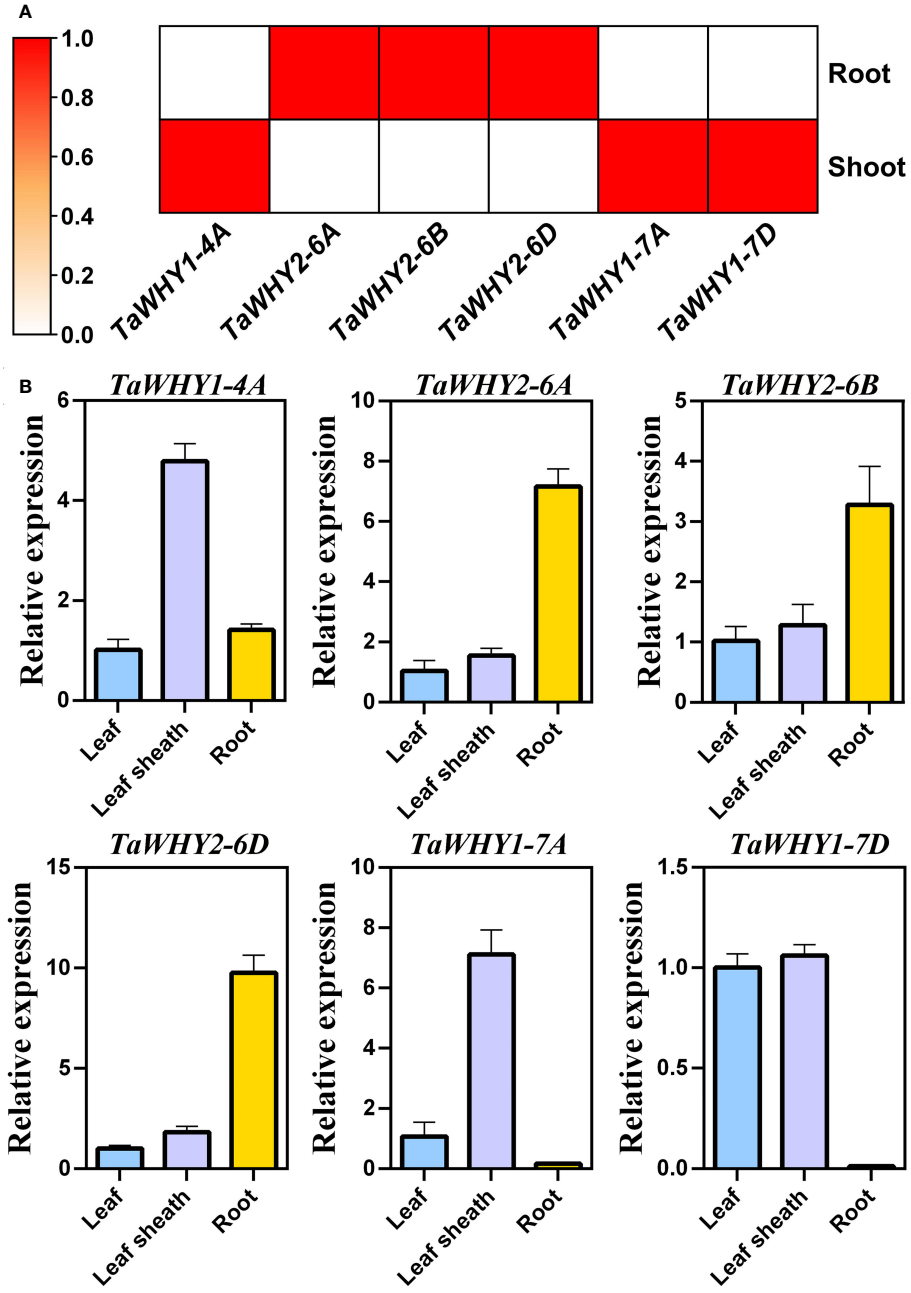


FIGURE 4
Expression pattern analysis of *TaWHYs* in different tissues. **(A)** The expression levels of *TaWHY* genes in root and shoot were determined through RNA-seq analysis. Fragments per kilobase of exon per million mapped fragments (FPKM) values were used to measure the expression levels of genes. **(B)** The expression levels of *TaWHY* genes in the root, leaf sheath, and leaf during the bread wheat seedling stage were determined by real-time PCR. The expression level of the bread wheat *actin* gene was used as the reference control to standardize the RNA samples for each reaction. Data represent the mean ± SD of three replicates.

10.3389/fpls.2023.1297228

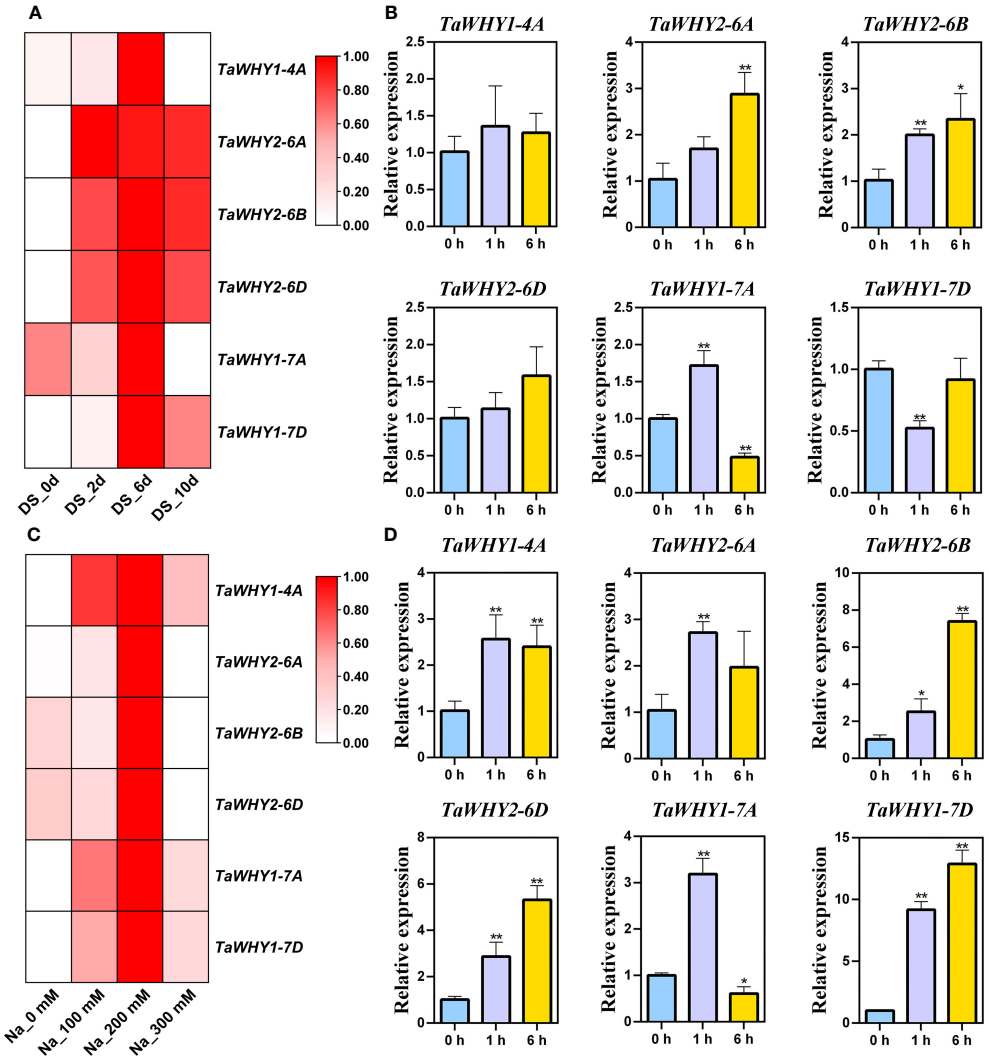


FIGURE 5
Expression patterns of *TaWHY* genes in response to osmotic stress. **(A)** RNA-seq analysis of the expression profiles of *TaWHY* genes in response to drought stress for 0 days, 2 days, 6 days, and 10 days, respectively. Fragments per kilobase of exon per million mapped fragments (FPKM) values were used to measure the expression levels of genes. **(B)** The expression profiles of *TaWHY* genes in bread wheat seedling leaves at 0 h, 1 h, and 6 h after PEG stress treatment. **(C)** RNA-seq analysis of the expression profiles of *TaWHY* genes in response to 0 mM, 100 mM, 200 mM, and 300 mM NaCl treatment. FPKM values were used to measure the expression levels of genes. **(D)** The expression profiles of *TaWHY* genes in bread wheat seedling leaves at 0 h, 1 h, and 6 h after NaCl stress treatment. The expression level of the bread wheat *actin* gene was used as the reference control to standardize the RNA samples for each reaction. Data represent the mean ± SD of three replicates. The asterisk indicates significant differences compared with 0 h (control, as 1) based on Student's *t*-test (*$p < 0.05$; **$p < 0.01$).

7D was down-regulated, and *TaWHY1-4A* was not significantly changed. The expression of *TaWHY2-6A*, *TaWHY2-6B*, and *TaWHY2-6D* (group 2) was gradually up-regulated and reached the highest expression level at 6 h under PEG stress with approximately 2.9-, 2.3-, and 1.6-fold compared with the control, respectively (Figure 5B).

After NaCl treatment, the expression levels of *TaWHY* genes were significantly up-regulated (Figures 5C, D). The real-time PCR results revealed that the expression levels of *TaWHY1-4A*, *TaWHY2-6A*, *TaWHY2-6B*, *TaWHY2-6D*, *TaWHY1-7A*, and *TaWHY1-7D* were all increased, peaking at 1 h, 1 h, 6 h, 6 h, 1 h, and 6 h with approximately 2.6-, 2.7-, 7.4-, 2.9-, 3.2-, and 12.9-fold compared with the control, respectively (Figure 5D). Therefore, we speculated that *TaWHYs* might play an important role under osmotic stress.

Upstream transcription factors, downstream target genes, and interacting proteins analysis of TaWHYs

Transcription factors can interact with different *cis*-elements in the promoter region of target genes, exerting diverse functions in plant growth, development, and stress response (Strader et al., 2022). To determine the functions of *TaWHY* genes, upstream transcription

factors and downstream target genes of *TaWHYs* were predicted by using the wheat integrative gene regulatory network (wGRN) (Figure 6; Supplementary Table S5) (Chen et al., 2023). Then, 22, 28, 33, 44, 195, and 187 transcription factors were predicted to regulate the expression of *TaWHY1-4A*, *TaWHY1-7A*, *TaWHY1-7D*, *TaWHY2-6A*, *TaWHY2-6B*, and *TaWHY2-6D*, respectively (Supplementary Table S5). We also conducted an analysis of the expression patterns for the top 30 potential upstream transcription factors and downstream target genes associated with *TaWHYs*. Under drought stress, the expression patterns of the cytokinin-responsiveGATA transcription factor 1-like gene (*TraesCS6B03G0575900*) were most similar to *TaWHY1-4A*. Additionally, the most similar expression patterns were observed in the transcription factor GLK2 (*TraesCS3D03G0362600*) and TCP family transcription factor TCP5 (*TraesCS3A03G0743200*, *TraesCS3B03G0849100*) with *TaWHY1-7A*. The MYB transcription factor (*TraesCS6B03G0466300*) and the cytokinin-responsive GATA transcription factor 1-like gene (*TraesCS6B03G0575900*) exhibited the most similar expression patterns to *TaWHY1-7D*. Furthermore, the nuclear transcription factor Y subunit C-4-like (*TraesCS6A03G0382200*) showed the most similar expression patterns to *TaWHY2-6A*. The transcription factor bHLH49-like gene (*TraesCS4D03G0108100*) demonstrated the most similar expression patterns to *TaWHY2-6B* and *TaWHY2-6D* (Figure 6A; Supplementary Figure S2A). Under salt stress, transcription factors LSD1 (*TraesCS1A03G0706000* and *TraesCS1B03G0806900*) and GLK2 (*TraesCS3A03G0376200*) exhibited the most similar expression patterns to *TaWHY1-4A*. The transcription factors GLK2 (*TraesCS3A03G0376200*), LSD1 (*TraesCS1A03G0706000*), GATA transcription factor 17-like (*TraesCS6A03G0279700*), RAP2-9-like (*TraesCS7B03G0076700*), and Zinc finger CCCH domain-containing protein 44-like (*TraesCS7A03G0973900*) displayed the most similar expression patterns to *TaWHY1-7A*, *TaWHY1-7D*, *TaWHY2-6A*, *TaWHY2-6B*, and *TaWHY2-6D*, respectively (Figure 6A; Supplementary Figure S2B). These transcription factors are highly likely to regulate the expression of *TaWHY* genes under drought and salt stress.

TaWHYs, as transcription factors, also regulate downstream target genes in response to osmotic stress. The result suggested that *TaWHY1-4A*, *TaWHY1-7A*, *TaWHY1-7D*, *TaWHY2-6A*, *TaWHY2-6B*, and *TaWHY2-6D* might bind to the promoter of 1,345, 1,181, 1,404, 999, 3,413, and 3,662 downstream target genes, respectively (Supplementary Table S6). Under drought stress, the similar expression patterns were observed in fructokinase-like 1 (*TraesCS3A03G0869600*), protein fluorescent in blue light (*TraesCS5D03G0431900*), 2-carboxy-1,4-naphthoquinone phytyltransferase (*TraesCS4A03G1008500*), 50S ribosomal protein (*TraesCS4A03G0332200* and *TraesCS6B03G1250700*), and starch synthase (*TraesCS4D03G0513300*) with *TaWHY1-7A*. The gene of glutamyl-tRNA (Gln) amidotransferase (*TraesCS2A03G0645400*), CDK5RAP1-like protein (*TraesCS4D03G0338300*), chaperone protein dnaJ 6-like (*TraesCS6A03G0385500*), OSB (*TraesCS3B03G1336700*), and flap endonuclease (*TraesCS1B03G1029400*) exhibited the most similar expression patterns to *TaWHY1-4A*, *TaWHY1-7D*, *TaWHY2-6A*, *TaWHY2-6B*, and *TaWHY2-6D*, respectively (Figure 6B; Supplementary Figure S3A). Similarly, the downstream target genes of the most similar expression patterns with *TaWHYs* under salt stress were also detected, i.e., *TaWHY1-4A* with transcription termination/antitermination protein NusG-like (*TraesCS5B03G1215400*), *TaWHY1-7A* with protein fluorescent in blue light (*TraesCS5D03G0431900*), *TaWHY1-7D* with superoxide dismutase (*TraesCS4A03G1080200*) and transcription termination/antitermination protein NusG-like (*TraesCS5B03G1215400*), *TaWHY2-6A* with eukaryotic translation initiation factor 3 subunit F-like (T*raesCS6A03G0205500*), *TaWHY2-6B* with HSP20-like chaperones superfamily protein (*TraesCS7D03G0654000*) and eukaryotic translation initiation factor 3 subunit K (*TraesCS4B03G0785500*), and *TaWHY2-6D* with HSP20-like chaperones superfamily protein (*TraesCS7D03G0654000*), DNA polymerase delta small subunit-like (*TraesCS4B03G0833700*) and flap endonuclease 1-A-like (*TraesCS1B03G1029400* and *TraesCS1A03G0881400*) (Figure 6B; Supplementary Figure S3B). The GO enrichment analysis result showed the downstream target genes of TaWHY1s mainly participated in translation, glutaminyl-tRNAGln biosynthesis, protoporphyrinogen IX biosynthetic process, and heme biosynthetic process (Supplementary Figure S4). TaWHY2s might take part in mRNA splicing, RNA binding, and DNA replication (Supplementary Figure S4). It was worth noting that TaWHY1-7D and TaWHY2-6D were predicted to respond to hydrogen peroxide (H_2O_2) and oxidative stress (Supplementary Figure S4), suggesting *TaWHY1-7D* and *TaWHY2-6D* might respond to osmotic stress via regulating ROS homeostasis.

The protein–protein interactions (PPIs) analysis suggested that TaWHY1-4A, TaWHY1-7A, and TaWHY1-7D could interact with 16, 37, and 36 proteins, respectively. TaWHY2-6A, TaWHY2-6B, and TaWHY2-6D interact with 102 proteins (Supplementary Table S7). We identified the interacting proteins with similar expression patterns to TaWHYs under drought stress (Supplementary Figure S5), i.e., TaWHY1-4A was found to interact with fructokinase-like 2 (TraesCS2A02G013600). TaWHY1-7A showed interactions with glutamate-rich WD repeat-containing protein (TraesCS4B02G157000), fructokinase-like 2 (TraesCS2A02G013600), and serine/arginine-rich splicing factor SR34A (TraesCS4D02G168700). TaWHY1-7D demonstrated an interaction with fructokinase-like 2 (TraesCS2A02G013600). Additionally, TaWHY2-6A interacted with DnaJ protein homolog (TraesCS5B02G374900), while TaWHY2-6B and TaWHY2-6D showed interactions with methionine aminopeptidase 1B (TraesCS2B02G448000) and protein OSB2 (TraesCS3B02G536700) (Figure 7; Supplementary Figure S5A). After NaCl treatment, TaWHY1 (TaWHY1-4A, TaWHY1-7A, and TaWHY1-7D) showed the most similar expression patterns with interacting protein single-stranded DNA-binding protein (TraesCS3A02G231400). TaWHY2 (TaWHY2-6A, TaWHY2-6B, and TaWHY2-6D) demonstrated the most similar expression patterns with glutamate-rich WD repeat-containing protein (TraesCS5B02G137200), actin-related protein (TraesCS5B02G422700), chaperone protein dnaJ A6 (TraesCS6B02G274600), and methionine aminopeptidase 1B (TraesCS2D02G231000) (Figure 7; Supplementary Figure S5B).

10.3389/fpls.2023.1297228

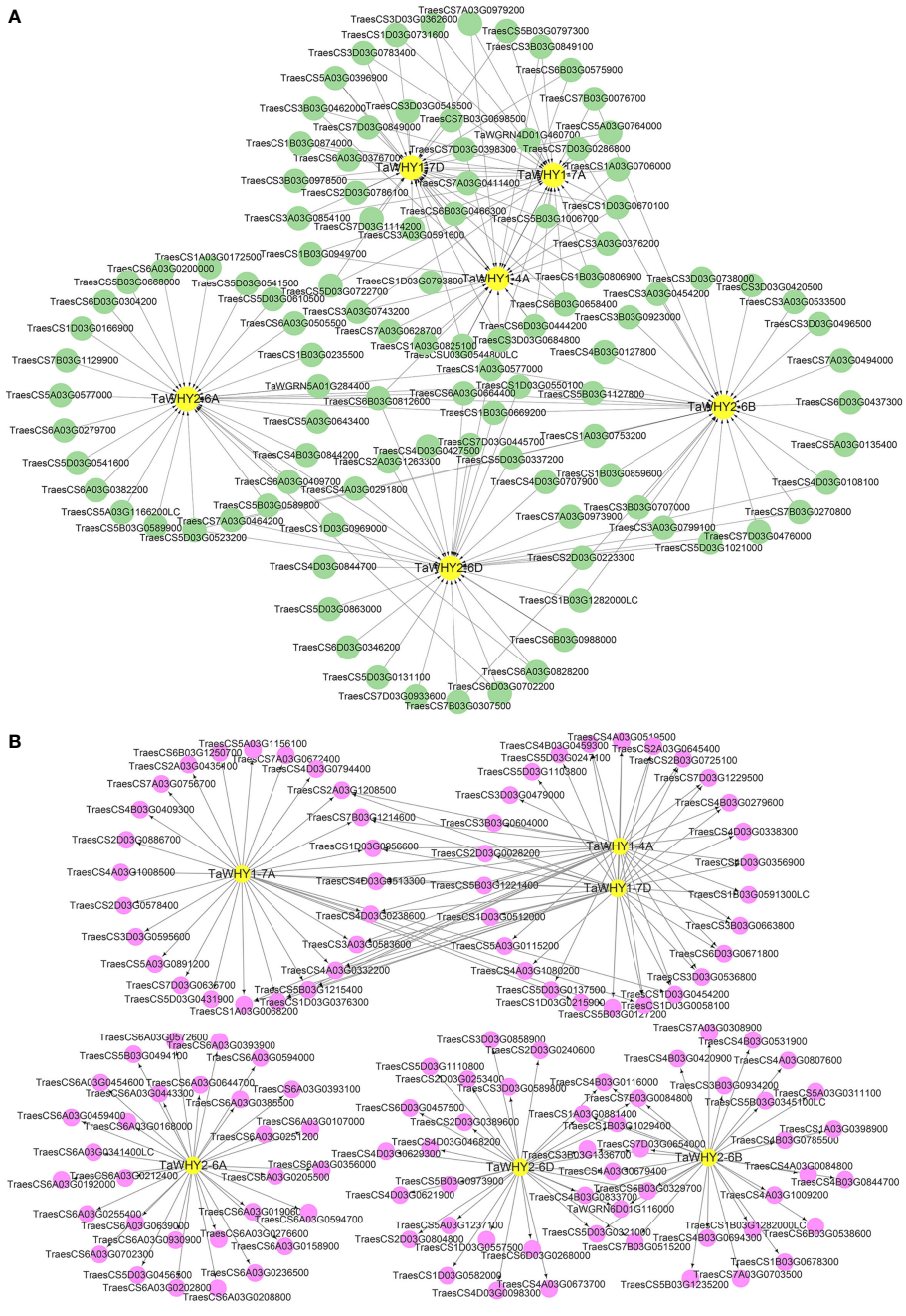


FIGURE 6
The upstream transcription factor **(A)** and downstream target gene **(B)** analyses of *TaWHY* genes.

These results suggested the regulatory mechanism of *TaWHY* genes to avoid or defend against osmotic stress.

TaWHYs improve the tolerance to osmotic and oxidative stresses in yeast cells

To further investigate the function of *TaWHY* genes under osmotic (D-sorbitol and NaCl) and oxidative (H_2O_2) stresses, *TaWHY2-6A*, *TaWHY2-6B*, *TaWHY2-6D*, *TaWHY1-7A*, and *TaWHY1-7D* were cloned into the pGADT7 vector, and then transformed into the yeast cells BY4741 or stress-sensitive yeast mutant BY4741 (Δ*hog1*) to confirm the ability to improve stress resistance in yeast cells (Figure 8). The results suggested that the growth of the BY4741 or Δ*hog1* yeast cells carrying these *TaWHY* genes was not obviously different compared with the control (pGADT7 empty vector) under normal growth conditions. After D-sorbitol treatment, Δ*hog1* yeast cells overexpressing *TaWHYs* slightly enhanced their tolerance to D-sorbitol stress in comparison to the negative control. The Δ*hog1* yeast overexpressing *TaWHY2-*

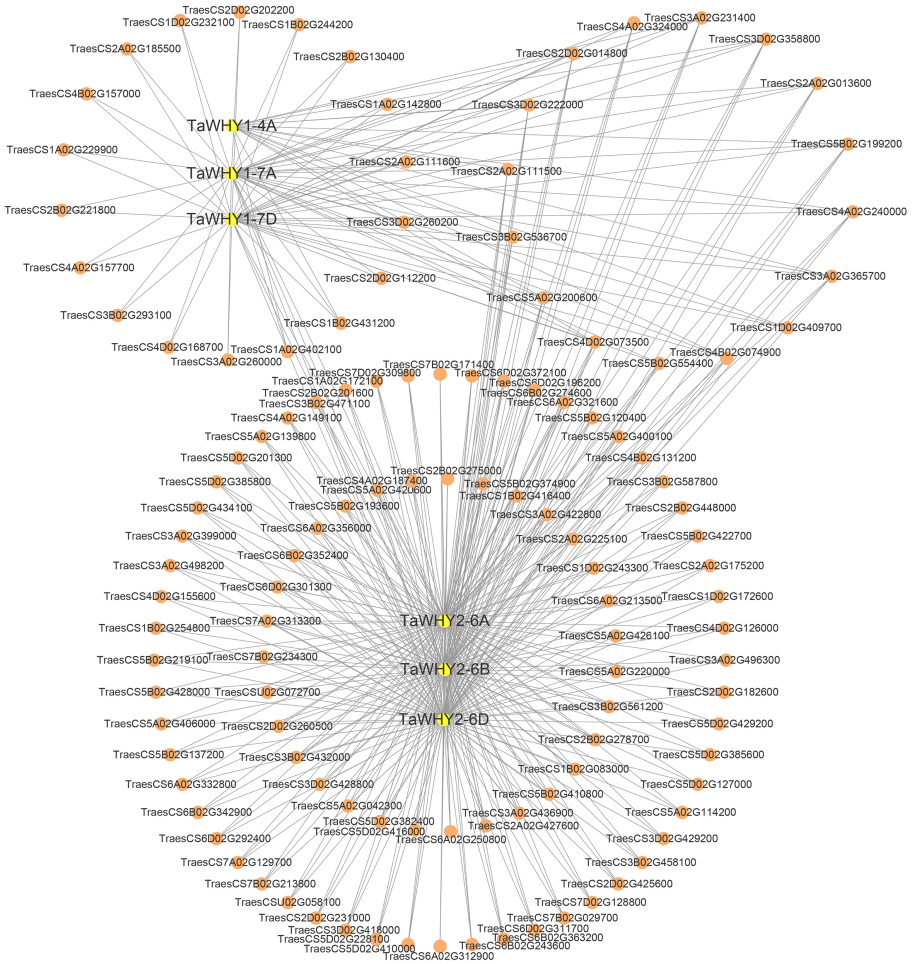


FIGURE 7
Protein–protein interaction (PPI) network analysis of TaWHY proteins.

6A, *TaWHY2-6B*, and *TaWHY2-6D* obviously improved the resistance to NaCl stress, but the colonies of Δ*hog1* with *TaWHY1-7A* and *TaWHY1-7D* were slightly increased compared with the negative control under NaCl stress.

Adverse environmental conditions induce ROS production; ROS accumulation can cause oxidative damage to membranes, proteins, and RNA and DNA molecules and even lead to the oxidative destruction of the cell in a process termed oxidative stress; thereby, ROS scavenging is essential for plants to avoid or defend against adverse stress (Choudhury et al., 2017). To determine whether *TaWHYs* enhanced stress tolerance by scavenging ROS in yeast cells, Δ*hog1* yeast cells carrying pGADT7-*TaWHYs* or pGADT7 were grown on YPD medium containing 4.0 mM H_2O_2, suggesting *TaWHY1-7A* and *TaWHY1-7D* strongly enhanced the oxidative stress tolerance in yeast, but the colonies of Δ*hog1* overexpressing *TaWHY2-6A*, *TaWHY2-6B*, and *TaWHY2-6D* were reduced compared with control. These results indicated that the *TaWHY1* and *TaWHY2* genes performed diverse functions. *TaWHY1* mainly enhanced the tolerance to oxidative stresses; *TaWHY2* mainly improved NaCl stress tolerance and was sensitive to oxygen stress; and *TaWHY1* and *TaWHY2* genes slightly improved the tolerance to D-sorbitol stress.

TaWHY1-7D confers drought and salt tolerance in *Arabidopsis*

In order to further confirm the potential role of *TaWHY1-7D* in response to drought and salt stresses, we generated *35S:TaWHY1-7D* transgenic *Arabidopsis* lines. Three independent transgenic lines (OE4, OE8, and OE10) and wild-type (WT) were chosen for the functional analysis of *TaWHY1-7D* in response to drought and salt stresses (Figure 9; Supplementary Figure S6). The results showed that there were no obvious phenotypic differences between transgenic and WT plants under normal conditions. After an 8-day drought treatment, the wild-type (WT) plants exhibited wilting and subsequent yellowing. In contrast, the transgenic *Arabidopsis* overexpressing *TaWHY1-7D*

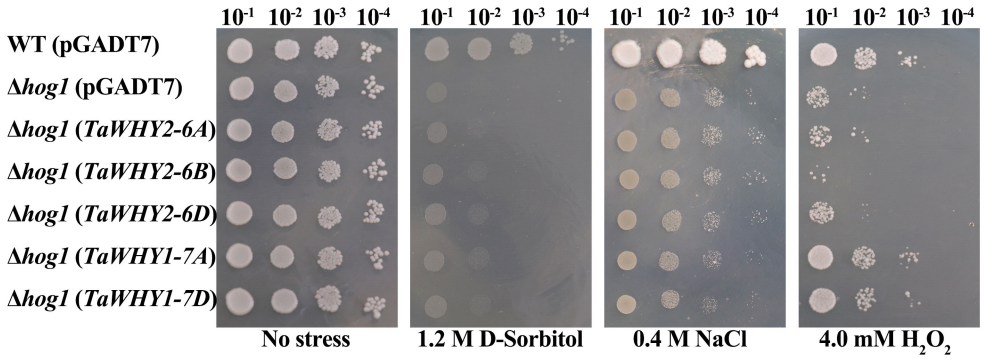


FIGURE 8
The ability of the tolerance in response to 1.2 M D-sorbitol, 0.4 M NaCl, and 4.0 mM H_2O_2 stresses in recombinant yeast cells. For osmotic and oxidative stresses, the yeast cells Δ*hog1* carrying the recombinant vector pGADT7-*TaWHY2-6A*/*TaWHY2-6B*/*TaWHY2-6D*/*TaWHY1-7A*/*TaWHY1-7D* were spotted onto YPD medium plates containing 1.2 M D-sorbitol, 0.4 M NaCl, or 4.0 mM H_2O_2 with serially diluted (10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}) and cultured at 30°C for 3–5 days. The wild-type yeast cells BY4741 and the stress-sensitive mutant Δ*hog1* carrying the empty vector pGADT7 were used as positive and negative controls, respectively.

remained predominantly green. After NaCl treatment for 8 days, both WT and transgenic *Arabidopsis* lines exhibited growth inhibition compared with CK. The growth inhibition was more severe in WT plants compared to transgenic *Arabidopsis*. Thus, the heterologous expression of *TaWHY1-7D* greatly improved drought and salt tolerance in transgenic *Arabidopsis*.

Discussion

Evolutionary relationship of *Whirly* genes in Triticeae species

Whirly genes have been identified in diverse plant species (Desveaux et al., 2005; Janack et al., 2016; Yan et al., 2020; Hu and Shu, 2021). Most plant species have two kinds of Whirly proteins, Whirly1 and Whirly2, whereas *Arabidopsis* and cassava have three Whirly proteins (Cappadocia et al., 2013; Yan et al., 2020). As a heterologous hexaploid species composed of three subgenomes A, B, and D, bread wheat (AABBDD) has undergone two rounds of natural hybridization events (Levy and Feldman, 2022). Therefore, bread wheat has six *Whirly* genes belonging to *Whirly1* and *Whirly2*, and other Triticeae species, including *T. urartu* (AA, diploid), *T. dicoccoides* (AABB, tetraploid), *Ae. tauschii* (DD, diploid), *H. vulgare* (HH, diploid), and *S. cereal* (RR, diploid), have two, four, two, two, and two *Whirly* genes, respectively (Figure 1A; Supplementary Table S2). There was a positive correlation between the number of *Whirly* genes and that of subgenomes in Triticeae species.

The paralogous *Whirly* gene pairs *TaWHY1-4A*/*TaWHY1-7A*/*TaWHY1-7D* and *TaWHY2-6A*/*TaWHY2-6B*/*TaWHY2-6D* were identified in *T. aestivum* genome, which all expanded by WGD or segmental duplication events (Figure 2B; Supplementary Table S4). Interestingly, the paralogous genes of *TaWHY1-7A* and *TaWHY1-7D* were found on chromosome 4A instead of chromosome 7B in *T. aestivum* (Figure 2B). To investigate the origin of *TaWHY1-4A*, a micro-collinear analysis of *TaWHY1-4A* was performed. The results showed that the homologous gene of *TaWHY1-4A* did not exist on subgenome B in other related Triticeae species, but there was homologous gene of *TuWHY1-7A* on chromosome 7A of *T. urartu* and *AetWHY1-7D* on chromosome 7D of *Ae. tauschii* (Figure 3). Similar events also occurred in the SHMT gene family of *T. aestivum* (Hu et al., 2022). Therefore, we speculated that the expansion events of *Whirly1* genes occurred through hybridization and polyploidization, and *TaWHY1-4A* and *TdWHY1-4A* might have originated from *TuWHY1-7A* or *AetWHY1-7D* (Figure 3). However, this speculation still needs further research.

The function of *TaWHY* genes in response to osmotic stress

Whirly proteins are plant-specific transcription factors that regulate plant development and stress resistance in plants (Krupinska et al., 2022; Taylor et al., 2022). Previous studies mainly focused on the function of *Whirly* genes under abiotic stress and biotic stresses, such as drought (Yan et al., 2020), salt (Akbudak and Filiz, 2019), chilling (Zhuang et al., 2020b) or light stresses (Swida-Barteczka et al., 2018). Previous studies indicated that AtWHY1 located in chloroplasts and nucleus (Krause et al., 2005; Ren et al., 2017) could repress the expression of *WRKY53* and delay leaf senescence in *Arabidopsis* (Miao et al., 2013), whereas AtWHY2 was located in the mitochondria and nucleus (Krause et al., 2005; Golin et al., 2020). These were consistent with the higher expression of *TaWHY1* genes (*TaWHY1-4A*, *TaWHY1-7A*, and *TaWHY1-7D*) in leaf sheaths and leaves and higher expression of *TaWHY2* genes (*TaWHY2-6A*, *TaWHY2-6B*, and *TaWHY2-6D*) in roots (Figure 4).

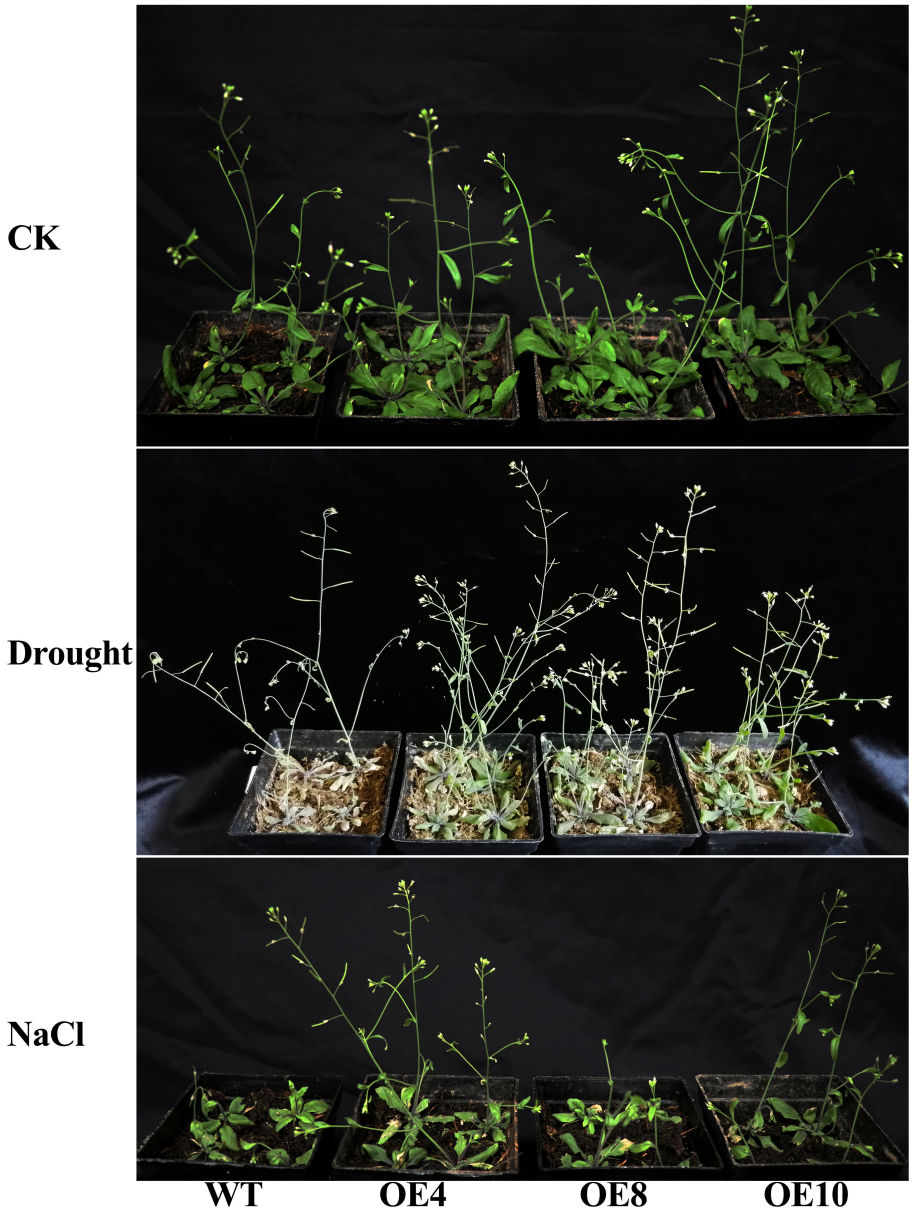


FIGURE 9
The phenotype of the *35S: TaWHY1-7D* transgenic *Arabidopsis* under drought and NaCl stress. Three independent *35S:TaWHY1-7D* transgenic *Arabidopsis* lines (OE4, OE8, and OE10) and wild type (WT) were chosen for functional analysis of *TaWHY1-7D* under normal conditions (CK), drought (water-deficit), and salt (NaCl) stress treatments.

Recently, *Whirly* genes were reported to improve osmotic stress resistance in plants, such as MeWHYs, which could interact with MeCIPK23 to activate ABA biosynthesis and regulate drought resistance in cassava (Yan et al., 2020). In this study, *TaWHY1-7A* and three *TaWHY2* genes were up-regulated under PEG stress, *TaWHY1-7D* was down-regulated, and *TaWHY1-4A* was not significantly changed (Figure 5), suggesting that functional differentiation of *Whirly* genes occurred. All *TaWHYs* were up-regulated under NaCl stress (Figure 5) and improved the resistance of NaCl stress in yeast, respectively (Figure 8). The heterologous expression of *TaWHY1-7D* greatly improved drought and salt tolerance in transgenic *Arabidopsis* (Figure 9). In addition, *Whirly* genes have been reported to regulate ROS homeostasis (Lin et al., 2019), and our results also showed that *TaWHY1-7A* and *TaWHY1-7D* strongly enhanced the oxidative stress tolerance in yeast cells (Figure 8). ROS scavenging also might be an important reason for the improvement of stress resistance in *TaWHY1* genes. However, the growth of Δ*hog1* overexpressing *TaWHY2-6A*, *TaWHY2-6B*, and *TaWHY2-6D* was inhibited under oxidative stress; these were consistent with a previous study that found that

overexpression of *AtWHY2* caused the accumulation of ROS in the plant (Cai et al., 2015). The ROS accumulation might cause cellular stress, thus activating the alternative pathway to reduce ROS levels and eliminate the stress (Cai et al., 2015). GO enrichment analysis also showed that TaWHY1-7D and TaWHY2-6D regulated downstream target genes to respond to H_2O_2 and oxidative stress (Supplementary Figure S4). Based on the above research, we speculate that the *Whirly* genes may play a vital role in plant resistance to osmotic stress. These results provide useful information for further functional studies of *Whirly* genes and lay a foundation to improve wheat yield and quality via molecular breeding under osmotic stress.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding authors.

Author contributions

HL: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. XW: Writing – review & editing. WY: Writing – review & editing. WL: Writing – review & editing. YW: Resources, Writing – review & editing. QW: Resources, Writing – review & editing. YZ: Resources, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Natural Science Foundation of China (Grant No. 32272159).

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.2023.1297228/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Multiple sequence alignment of the conserved domains of *Whirly* genes in Triticeae species. Moitf1-10 and DNA binding domain were marked. The conserved cysteine was marked by red triangle.

SUPPLEMENTARY FIGURE 2
The FPKM values of upstream transcription factors of TaWHYs under drought **(A)** and salt **(B)** stress.

SUPPLEMENTARY FIGURE 3
The FPKM values of downstream target genes of TaWHYs under drought **(A)** and salt **(B)** stress.

SUPPLEMENTARY FIGURE 4
GO enrichment analysis on the downstream target genes of *TaWHY* genes.

SUPPLEMENTARY FIGURE 5
The FPKM values of interacting protein of TaWHYs under drought **(A)** and salt **(B)** stress.

SUPPLEMENTARY FIGURE 6
The PCR detection **(A)** and screening **(B)** of *35S: TaWHY1-7D* transgenic Arabidopsis.

References

Akbudak, M. A., and Filiz, E. (2019). Whirly (Why) transcription factors in tomato (Solanum lycopersicum L.): genome-wide identification and transcriptional profiling under drought and salt stresses. *Mol. Biol. Rep.* 46, 4139–4150. doi: 10.1007/s11033-019-04863-y

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–W208. doi: 10.1093/nar/gkp335

Cai, Q., Guo, L., Shen, Z. R., Wang, D. Y., Zhang, Q., and Sodmergen, (2015). Elevation of pollen mitochondrial DNA copy number by WHIRLY2: altered respiration and pollen tube growth in arabidopsis. *Plant Physiol.* 169, 660–673. doi: 10.1104/pp.15.00437

Cappadocia, L., Parent, J. S., Sygusch, J., and Brisson, N. (2013). A family portrait: structural comparison of the Whirly proteins from Arabidopsis thaliana and Solanum tuberosum. *Acta Crystallogr. Sect F Struct. Biol. Cryst Commun.* 69, 1207–1211. doi: 10.1107/S1744309113028698

Cappadocia, L., Parent, J. S., Zampini, E., Lepage, E., Sygusch, J., and Brisson, N. (2012). A conserved lysine residue of plant Whirly proteins is necessary for higher order protein assembly and protection against DNA damage. *Nucleic Acids Res.* 40, 258–269. doi: 10.1093/nar/gkr740

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020a). TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp.2020.06.009

Chen, Y., Guo, Y., Guan, P., Wang, Y., Wang, X., Wang, Z., et al. (2023). A wheat integrative regulatory network from large-scale complementary functional datasets enables trait-associated gene discovery for crop improvement. *Mol. Plant* 16, 393–414. doi: 10.1016/j.molp.2022.12.019

Chen, Y., Song, W., Xie, X., Wang, Z., Guan, P., Peng, H., et al. (2020b). A collinearity-incorporating homology inference strategy for connecting emerging assemblies in the triticeae tribe as a pilot practice in the plant pangenomic era. *Mol. Plant* 13, 1694–1708. doi: 10.1016/j.molp.2020.09.019

Choudhury, F. K., Rivero, R. M., Blumwald, E., and Mittler, R. (2017). Reactive oxygen species, abiotic stress and stress combination. *Plant J.* 90, 856–867. doi: 10.1111/tpj.13299

CNCB-NGDC Members and Partners (2022). Database resources of the national genomics data center, China national center for bioinformation in 2022. *Nucleic Acids Res.* 50, D27–D38. doi: 10.1093/nar/gkab951

Desveaux, D., Allard, J., Brisson, N., and Sygusch, J. (2002). A new family of plant transcription factors displays a novel ssDNA-binding surface. *Nat. Struct. Biol.* 9, 512–517. doi: 10.1038/nsb814

Desveaux, D., Despres, C., Joyeux, A., Subramaniam, R., and Brisson, N. (2000). PBF-2 is a novel single-stranded DNA binding factor implicated in PR-10a gene activation in potato. *Plant Cell* 12, 1477–1489. doi: 10.1105/tpc.12.8.1477

Desveaux, D., Marechal, A., and Brisson, N. (2005). Whirly transcription factors: defense gene regulation and beyond. *Trends Plant Sci.* 10, 95–102. doi: 10.1016/j.tplants.2004.12.008

Desveaux, D., Subramaniam, R., Despres, C., Mess, J. N., Levesque, C., Fobert, P. R., et al. (2004). A "Whirly" transcription factor is required for salicylic acid-dependent disease resistance in Arabidopsis. *Dev. Cell* 6, 229–240. doi: 10.1016/S1534-5807(04)00028-0

Foyer, C. H., Karpinska, B., and Krupinska, K. (2014). The functions of WHIRLY1 and REDOX-RESPONSIVE TRANSCRIPTION FACTOR 1 in cross tolerance responses in plants: a hypothesis. *Philos. Trans. R Soc. Lond B Biol. Sci.* 369, 20130226. doi: 10.1098/rstb.2013.0226

Golin, S., Negroni, Y. L., Bennewitz, B., Klosgen, R. B., Mulisch, M., La Rocca, N., et al. (2020). WHIRLY2 plays a key role in mitochondria morphology, dynamics, and functionality in Arabidopsis thaliana. *Plant Direct* 4, e00229. doi: 10.1002/pld3.229

Gouet, P., Robert, X., and Courcelle, E. (2003). ESPript/ENDscript: Extracting and rendering sequence and 3D information from atomic structures of proteins. *Nucleic Acids Res.* 31, 3320–3323. doi: 10.1093/nar/gkg556

Gupta, P. K., Balyan, H. S., Sharma, S., and Kumar, R. (2020). Genetics of yield, abiotic stress tolerance and biofortification in wheat (Triticum aestivum L.). *Theor. Appl. Genet.* 133, 1569–1602. doi: 10.1007/s00122-020-03583-3

Hu, Y., and Shu, B. (2021). Identifying strawberry Whirly family transcription factors and their expressions in response to crown rot. *Notulae Botanicae Horti Agrobotanici Cluj-Napoca* 49, 12323. doi: 10.15835/nbha49212323

Hu, P., Song, P., Xu, J., Wei, Q., Tao, Y., Ren, Y., et al. (2022). Genome-wide analysis of serine hydroxymethyltransferase genes in triticeae species reveals that taSHMT3A-1 regulates fusarium head blight resistance in wheat. *Front. Plant Sci.* 13, 847087. doi: 10.3389/fpls.2022.847087

Janack, B., Sosoi, P., Krupinska, K., and Humbeck, K. (2016). Knockdown of WHIRLY1 affects drought stress-induced leaf senescence and histone modifications of the senescence-associated gene HvS40. *Plants-Basel* 5, 37. doi: 10.3390/plants5030037

Krause, K., Kilbienski, I., Mulisch, M., Rodiger, A., Schafer, A., and Krupinska, K. (2005). DNA-binding proteins of the Whirly family in Arabidopsis thaliana are targeted to the organelles. *FEBS Lett.* 579, 3707–3712. doi: 10.1016/j.febslet.2005.05.059

Krupinska, K., Desel, C., Frank, S., and Hensel, G. (2022). WHIRLIES are multifunctional DNA-binding proteins with impact on plant development and stress resistance. *Front. Plant Sci.* 13, 880423. doi: 10.3389/fpls.2022.880423

Letunic, I., Khedkar, S., and Bork, P. (2021). SMART: recent updates, new developments and status in 2020. *Nucleic Acids Res.* 49, D458–D460. doi: 10.1093/nar/gkaa937

Levy, A. A., and Feldman, M. (2022). Evolution and origin of bread wheat. *Plant Cell* 34, 2549–2567. doi: 10.1093/plcell/koac130

Lin, W., Huang, D., Shi, X., Deng, B., Ren, Y., Lin, W., et al. (2019). H2O2 as a feedback signal on dual-located WHIRLY1 associates with leaf senescence in Arabidopsis. *Cells* 8, 1585. doi: 10.3390/cells8121585

Liu, H., Yang, W., Zhao, X., Kang, G., Li, N., and Xu, H. (2022). Genome-wide analysis and functional characterization of CHYR gene family associated with abiotic stress tolerance in bread wheat (Triticum aestivum L.). *BMC Plant Biol.* 22, 204. doi: 10.1186/s12870-022-03589-7

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

Ma, S., Wang, M., Wu, J., Guo, W., Chen, Y., Li, G., et al. (2021). WheatOmics: A platform combining multiple omics data to accelerate functional genomics studies in wheat. *Mol. Plant* 14, 1965–1968. doi: 10.1016/j.molp.2021.10.006

Manh, M. B., Ost, C., Peiter, E., Hause, B., Krupinska, K., and Humbeck, K. (2023). WHIRLY1 acts upstream of ABA-related reprogramming of drought-induced gene expression in barley and affects stress-related histone modifications. *Int. J. Mol. Sci.* 24, 6326. doi: 10.3390/ijms24076326

Marechal, A., Parent, J. S., Sabar, M., Veronneau-Lafortune, F., Abou-Rached, C., and Brisson, N. (2008). Overexpression of mtDNA-associated AtWhy2 compromises mitochondrial function. *BMC Plant Biol.* 8, 42. doi: 10.1186/1471-2229-8-42

Marechal, A., Parent, J. S., Veronneau-Lafortune, F., Joyeux, A., Lang, B. F., and Brisson, N. (2009). Whirly proteins maintain plastid genome stability in Arabidopsis. *Proc. Natl. Acad. Sci. U.S.A.* 106, 14693–14698. doi: 10.1073/pnas.0901710106

Meng, C., Yang, M., Wang, Y., Chen, C., Sui, N., Meng, Q., et al. (2020). SlWHY2 interacts with SlRECA2 to maintain mitochondrial function under drought stress in tomato. *Plant Sci.* 301, 110674. doi: 10.1016/j.plantsci.2020.110674

Miao, Y., Jiang, J., Ren, Y., and Zhao, Z. (2013). The single-stranded DNA-binding protein WHIRLY1 represses WRKY53 expression and delays leaf senescence in a developmental stage-dependent manner in Arabidopsis. *Plant Physiol.* 163, 746–756. doi: 10.1104/pp.113.223412

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., et al. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Res.* 49, D412–D419. doi: 10.1093/nar/gkaa913

Ren, Y., Li, Y., Jiang, Y., Wu, B., and Miao, Y. (2017). Phosphorylation of WHIRLY1 by CIPK14 shifts its localization and dual functions in Arabidopsis. *Mol. Plant* 10, 749–763. doi: 10.1016/j.molp.2017.03.011

Strader, L., Weijers, D., and Wagner, D. (2022). Plant transcription factors - being in the right place with the right company. *Curr. Opin. Plant Biol.* 65, 102136. doi: 10.1016/j.pbi.2021.102136

Subramanian, B., Gao, S., Lercher, M. J., Hu, S., and Chen, W. H. (2019). Evolview v3: a webserver for visualization, annotation, and management of phylogenetic trees. *Nucleic Acids Res.* 47, W270–W275. doi: 10.1093/nar/gkz357

Swida-Barteczka, A., Krieger-Liszkay, A., Bilger, W., Voigt, U., Hensel, G., Szweykowska-Kulinska, Z., et al. (2018). The plastid-nucleus located DNA/RNA binding protein WHIRLY1 regulates microRNA-levels during stress in barley (Hordeum vulgare L.). *RNA Biol.* 15, 886–891. doi: 10.1080/15476286.2018.1481695

Tamura, K., Stecher, G., and Kumar, S. (2021). MEGA11: molecular evolutionary genetics analysis version 11. *Mol. Biol. Evol.* 38, 3022–3027. doi: 10.1093/molbev/msab120

Taylor, R. E., West, C. E., and Foyer, C. H. (2022). WHIRLY protein functions in plants. *Food Energy Secur.* 00, e379. doi: 10.1002/fes3.379

Von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003). STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* 31, 258–261. doi: 10.1093/nar/gkg034

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40, e49. doi: 10.1093/nar/gkr1293

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., et al. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46, W296–W303. doi: 10.1093/nar/gky427

Yan, Y., Liu, W., Wei, Y., and Shi, H. (2020). MeCIPK23 interacts with Whirly transcription factors to activate abscisic acid biosynthesis and regulate drought resistance in cassava. *Plant Biotechnol. J.* 18, 1504–1506. doi: 10.1111/pbi.13321

Yoo, H. H., Kwon, C., Lee, M. M., and Chung, I. K. (2007). Single-stranded DNA binding factor AtWHY1 modulates telomere length homeostasis in Arabidopsis. *Plant J.* 49, 442–451. doi: 10.1111/j.1365-313X.2006.02974.x

Zhuang, K., Gao, Y., Liu, Z., Diao, P., Sui, N., Meng, Q., et al. (2020a). WHIRLY1 regulates HSP21.5A expression to promote thermotolerance in tomato. *Plant Cell Physiol.* 61, 169–177. doi: 10.1093/pcp/pcz189

Zhuang, K., Wang, J., Jiao, B., Chen, C., Zhang, J., Ma, N., et al. (2020b). WHIRLY1 maintains leaf photosynthetic capacity in tomato by regulating the expression of RbcS1 under chilling stress. *J. Exp. Bot.* 71, 3653–3663. doi: 10.1093/jxb/eraa145

frontiers | Frontiers in Plant Science

TYPE Technology and Code
PUBLISHED 03 January 2024
DOI 10.3389/fpls.2023.1290078

OPEN ACCESS

EDITED BY
Xueqiang Wang,
Zhejiang University, China

REVIEWED BY
Zhen Fan,
University of Florida, United States
Waseem Hussain,
International Rice Research Institute (IRRI), Philippines

*CORRESPONDENCE
Santiago Diaz
w.s.diaz@cgiar.org

†PRESENT ADDRESS
Johan Aparicio,
College of Agricultural & Life Sciences, University of Wisconsin-Madison, WI, United States
Daniel Ariza-Suarez,
Molecular Plant Breeding, Institute of Agricultural Sciences, ETH Zurich, Zurich, Switzerland
Bodo Raatz,
Limagrain Vegetable Seed, La Menitré, France
Juan Lobaton,
Department of Evolutionary Biology, National Australian University, Canberra, Australia

RECEIVED 06 September 2023
ACCEPTED 01 December 2023
PUBLISHED 03 January 2024

CITATION
Aparicio J, Gezan SA, Ariza-Suarez D, Raatz B, Diaz S, Heilman-Morales A and Lobaton J (2024) Mr.Bean: a comprehensive statistical and visualization application for modeling agricultural field trials data.
Front. Plant Sci. 14:1290078.
doi: 10.3389/fpls.2023.1290078

Mr.Bean: a comprehensive statistical and visualization application for modeling agricultural field trials data

Johan Aparicio[1†], Salvador A. Gezan[2], Daniel Ariza-Suarez[1†], Bodo Raatz[1†], Santiago Diaz[1*], Ana Heilman-Morales[3] and Juan Lobaton[1†]

[1]Bean Program, Crops for Nutrition and Health, Alliance Bioversity-International Center for Tropical Agriculture (CIAT), Cali, Colombia, [2]Deparment of Statistical Genetics, InternationalVSN, Hemel Hempstead, United Kingdom, [3]Big Data Pipeline Unit, North Dakota State UniversityAES, Fargo, ND, United States

Crop improvement efforts have exploited new methods for modeling spatial trends using the arrangement of the experimental units in the field. These methods have shown improvement in predicting the genetic potential of evaluated genotypes. However, the use of these tools may be limited by the exposure and accessibility to these products. In addition, these new methodologies often require plant scientists to be familiar with the programming environment used to implement them; constraints that limit data analysis efficiency for decision-making. These challenges have led to the development of Mr.Bean, an accessible and user-friendly tool with a comprehensive graphical visualization interface. The application integrates descriptive analysis, measures of dispersion and centralization, linear mixed model fitting, multi-environment trial analysis, factor analytic models, and genomic analysis. All these capabilities are designed to help plant breeders and scientist working with agricultural field trials make informed decisions more quickly. Mr.Bean is available for download at https://github.com/AparicioJohan/MrBeanApp.

KEYWORDS

spatial analysis, experimental designs, multi-environmental analysis, trial, breeding

1 Introduction

The selection of high-yielding and environmentally adapted genotypes in field trials is a fundamental challenge in plant breeding. In these types of trials, multiple genotypes are evaluated to estimate genetic parameters and determine the performance of traits of interest in breeding programs (Mackay et al., 2019). Experimental field design plays a crucial role in plant breeding (Piepho et al., 2022). Two experimental

designs are widely used in traditional breeding field trials: (i) randomized complete block design (RCBD) and (ii) incomplete block design (Alvarado et al., 2020).

Field trials are usually designed to account for spatial heterogeneity, traditionally controlled by blocking. Researchers divide replicates into blocks, as in the so-called incomplete block design. However, spatial variation in trials cannot be fully captured, and has been recognized as a major source of experimental error (Yan, 2021). Spatial heterogeneity in the field can be associated with intrinsic biotic factors such as soil microorganisms, pests, diseases, and weeds. Abiotic factors also drive spatial heterogeneity, including the effects of soil fertility, nutrient concentration, presence of toxic elements, water availability, soil structure, and slope, among others. Agronomic management of the trial can also vary within and across sites (Isik et al., 2017). These conditions promote the generation of localized patterns or microenvironments that differ between experimental units in the field, reducing the overall uniformity of the trial (Bernardeli et al., 2021). For this reason, the experimental designs commonly used in plant breeding aim to separate genotypic information from the environmental variability (non-genetic variation). Separation of genotypic and environmental variability can improve selection accuracy in field trials, reducing the experimental error with increasing genetic gain (Cursi et al., 2021).

To model the genotypic and environmental components in a field trial, researchers use linear mixed models (LMM). These approaches contain a mixture of fixed and random effects to estimate and infer the variance components (Veturi et al., 2012). Some of these procedures can incorporate a component to model the spatial variation in breeding trials (Mao et al., 2020). Understanding spatial variation can improve predictions of the genetic potential of the evaluated genotypes. Towards this end, several approaches have been proposed to correct for spatial heterogeneity in the field (Cullis and Gleeson, 1991; Currie and Durban, 2002; Piepho and Williams, 2010; Robbins et al., 2012). There are two major classes of spatial analysis for field trials in plant breeding: (1) using neighboring plots to adjust the mean of the plot of interest, and (2) predicting the plot values by adding a spatial covariate to the mixed model (Zystro et al., 2018). These approaches can be further classified into those that use spatial variance-covariance structures and those using smoothing techniques to model spatial trends (Rodriguez-Alvarez et al., 2018).

One of the great challenges of in data analysis of plant breeding trials could is requires significant computational resources to process (Harrison and Caccamo, 2022). The complexity of the data and models can make it difficult these analyses. besides, the analysis of this data often involves multiple steps, including modeling, preprocessing, feature selection, and interpretation of results (Xu et al., 2022). Multiple software has been implemented with the aim of solving these problems. However, the implementation of these approaches into end-user tools is limited either by the accessibility of these tools or by the requirements and experience needed to program computer instructions for the models. Intending to help breeders or plant science researchers, this work describes "Mr.Bean", a free R-Shiny application with a friendly and easy-to-use graphical user interface (GUI). This application simplifies the analysis of large-scale plant breeding experiments by using the power and versatility of LMM with or without spatial correction. This application combines the analytical robustness and speed offered by several R packages such as *ASReml-R* (Butler et al., 2017), *SpATS* (Rodriguez-Alvarez et al., 2018), and *lme4* (Bates et al., 2015) with the interactive features and visual power offered by Shiny R (Chang et al., 2023) and plotly (Sievert, 2020). The application also provides a graphical workflow for importing data from the Breeding Management System (BMS) and Breedbase through application programing interfaces (API), that help to identify outliers, and fit field data. Mr.Bean can analyze data from single-location or multi-environmental trials (MET), calculating the best linear unbiased estimator (BLUE), the best linear unbiased predictor (BLUP) (Piepho et al., 2008), and the broad-sense heritabilities. In addition, Mr. Bean offers a module for exploring results from Factor Analytic (FA) MET models using several graphical and multivariate techniques. The application integrates genomic and phenotypic data using the R-package *sommer* (Covarrubias-Pazaran, 2016). It estimates marker effects, variance components with genomic predictions, marker-base heritability, and genomic breeding values (GEBVs).

This application is a convenient and accurate way to analyze agronomic data, visualize field patterns and select genotypes for breeding programs. Mr.Bean aims to help statisticians, quantitative geneticists, and breeders who want to simplify and automate (or semi-automate) routine analysis to accurately predict the genetic potential of genotypes coming out of plant breeding pipelines. Moreover, Mr.Bean offers an alternative way to analyze field data for end-users with no previous experience in R programming language.

2 Methods

2.1 Mr.Bean implementation

Mr.Bean (v2.0.8) was developed in R using the package Shiny (Chang et al., 2023), an elegant and powerful web framework for creating R applications. Shiny supports developers with no previous experience using HTML, CSS, or JavaScript. Our developers improved the application's interactive experience by employing additional extensions like ShinyJS, bs4dash, shinyWidgets, and ShinyBS. Mr.Bean uses a graphical interface designed to work under any web browser or R software as an R-Shiny application, executed in the x86_64-pc-linux-gnu (64-bit) platform. The core component consists of a set of 41 R attached packages, for r-base:4.1.1 or higher. Mr.Bean uses the packages *SpATS* (Rodriguez-Alvarez et al., 2018), *ASReml-R* (Butler et al., 2017), and *lme4* (Bates et al., 2015) for fitting LMM with or without spatial corrections. The *sommer* package (Covarrubias-Pazaran, 2016) within Mr.Bean integrates genomic information to estimate genomic best linear unbiased predictions (GBLUPs).

2.2 Running Mr.Bean

Mr.Bean can be installed through the R software console from GitHub (https://github.com/AparicioJohan/MrBeanApp). It can also be installed and run locally by downloading it directly from the docker hub (https://hub.docker.com/r/johanstevenapa/mrbeanapp). For better understanding and ease in installing the application using GitHub or Docker, a video tutorial that explains the installation step by step is in the following link: https://www.youtube.com/watch?v=YubFj5DEQ2s. The application can be run in a beta version on the internet using any web browser for users without sufficient processing power, which anyone person can access through the following link: https://beanteam.shinyapps.io/MrBean_BETA/ (Figure 1). The beta version is a version that is hosted on a server of the Bioversity-CIAT alliance. The only disadvantage of this Beta version is that the *ASReml*, *Two-Stage analysis*, and *GBLUP* modules are not available and there must be a permanent internet connection. Mr.Bean follows a logical process through data loading, statistical analysis, model development, and results generation (Figure 2).

2.3 Data upload

The Data module allows users to upload their trial data. This module has several ways to import data from the Upload function. Data can be uploaded from your personal computer or via an internet connection to the Breeding (BrAPI) (https://brapi.org/), BMS and BreedBase APIs. The application is prepared to receive datasets with a maximum file size of 100 MB, following the tidy format in which every variable has a single column, and every observation a single assigned row (see Wickham, 2014 for a detailed explanation). Users can upload data in several formats, including comma-separated values (csv), tab-separated values (tsv), plain text (txt), and two different Excel formats ("xlsx" or "xls"). These upload capabilities allow users to identify the missing value character for their dataset.

Once the dataset has been uploaded, the module provides a quick view of the information for navigation (sorting, filtering, and pagination). Additionally, users can create subsets of variables for further analysis. The Descriptives section provides the ability to visually compare different qualitative and quantitative variables using box plots and two-dimensional scatter plots. The Distribution section helps visualize the frequency distribution for each individual trait using a histogram plot, with accompanying summary statistics such as mean, standard deviation, quartiles, and kurtosis, among others. For beginner users, a video tutorial on importing data and making plots in this section is available at the following link: https://www.youtube.com/watch?v=IlahWdDOOzU.

2.4 SpATS module

Here, the user can fit an LMM with spatial correction. SpATS is an attractive alternative to classical analyses of field trials, which model spatial variation as correlated noise (Rodriguez-Alvarez et al., 2018). It uses two-dimensional smoothing surfaces with penalized splines to model the spatial trends within the LMM framework. Hence, the implemented SpATS model is

$$y = \mu + gen + f_{u,v}(col, row) + \epsilon$$

where y is the trait of interest, μ is the overall mean, gen is the effect of the genotype, $f_{u,v}(col,row)$ are the row, column, bilinear polynomial, and smoothing spline effects, and ϵ is the effect of experimental error.

The *Single-Site* function allows the *SpATS* model to be run for experiments in a single location, evaluating one trait at a time. Users can calibrate the model with the *Model Specs* function. This function requires the user to specify the response variable, genotypes, and spatial coordinates for the plots, which are

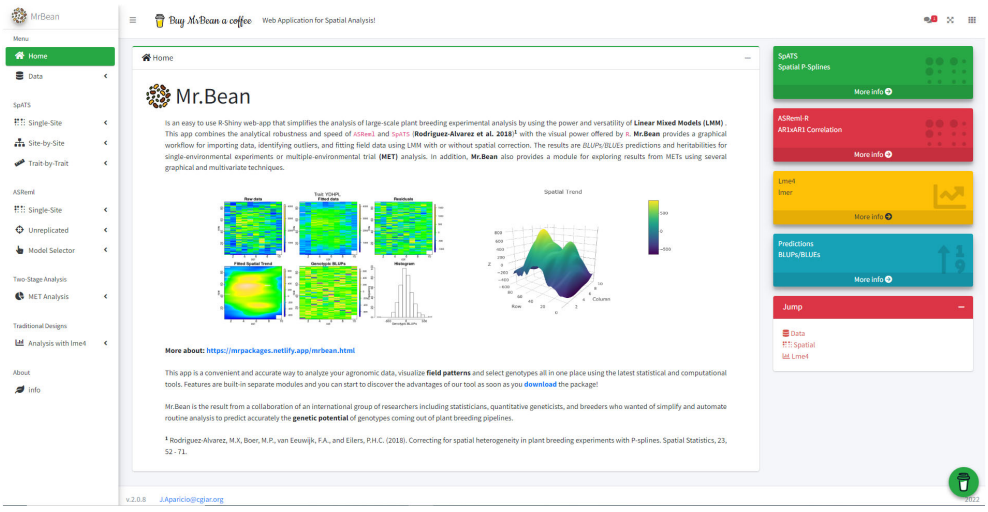


FIGURE 1
Mr.Bean application home page web.

10.3389/fpls.2023.1290078

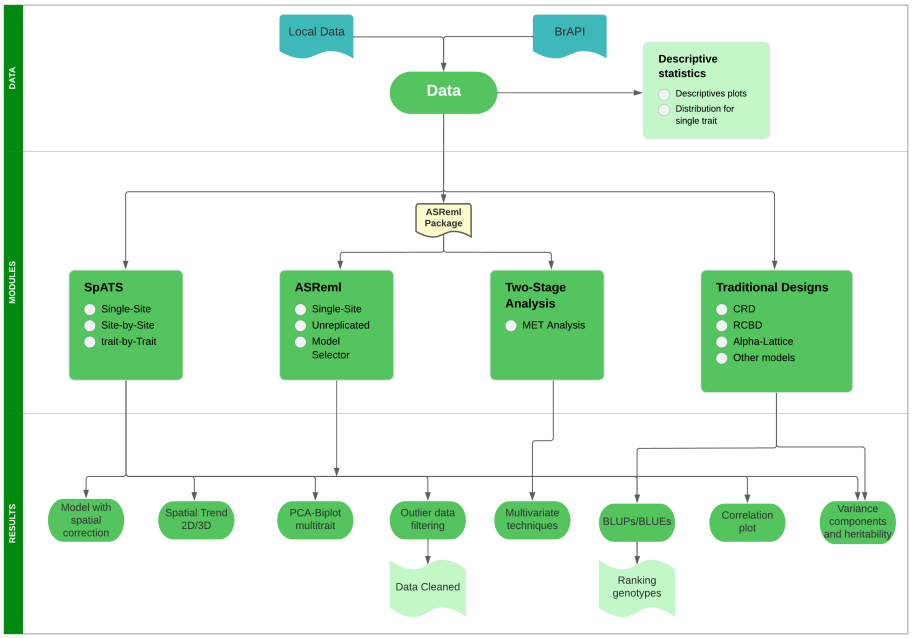


FIGURE 2
Flow diagram showing the logical process that Mr.Bean follows to perform several analyses.

represented in rows and columns. At their discretion, users can select genotype checks for the trial and add additional variables as fixed or random effects, as well as covariates in the LMM. There is a *Help* button for beginner users that guides them step-by-step through each of the parameters required to run the model. The application generates a table with an estimate of the broad-sense heritability, residual standard deviation, R-squared, and coefficient of variation of the fitted model. Users can perform the Least Significant Difference (LSD) test if the genotype factor is selected as a fixed effect in the model. The application also produces tables and graphs summarizing the model's variance components, spatial trends of raw data, fitted data, residuals, and genotype BLUPs with their respective histograms. Moreover, users can visualize spatial trends in the trial plots with two- and three-dimensional graphs.

The *BLUPs/BLUEs* subsection returns the predicted values for each genotype with their respective standard errors, including a histogram showing the distribution. The application also displays an error-bar plot that ranks the genotypic values for the variable of interest. Finally, the *Residuals* subsection provides tools to identify outlier observations from the analysis of residuals. It uses the assumption that residuals from the model follow a normal distribution with a mean of zero, using a 99% confidence interval to identify outlier data that fall beyond the range of ±3 standard deviations from the mean. The application graphs the outliers in field plots, identifying potential outliers or comparing residuals against other traits or factors. These functions contribute to the data cleaning process (quality assurance/quality check), before the user downloads a clean dataset.

The *Site-by-Site* function fits models for experiments evaluated in several locations, one trait at a time. This function also has a *Model Specs* subsection for fitting the model. As with the *Single-Site* function, the user selects the parameters required to run the model (response variable, genotype, spatial coordinates). The *Experiment* parameter allows the user to select sites for evaluation. Users can add other optional parameters (components with random or fixed effect, covariates). In addition, users can visualize the genotypes or lines that are shared between sites or experiments.

The *Results* subsection compares variance components between sites using a bar graph. As with the *Single-Site* function, the application summarizes spatial trends of raw data, fitted data, residuals, and genotype BLUPs with their respective histograms. The application creates a ranked error bar-plot of genotype BLUPs. Between evaluated experiments, the application generates correlation plots of phenotypic coefficients and their significance. Corresponding model components and summaries of each experiment are reported with the heritability estimated using the following equation:

$$H_g^2 = \frac{ED_g}{m_g - 1}$$

where ED_g is the effective dimension for genetic effects, and m_g is the number of genotypes (Rodriguez-Alvarez et al., 2018). As with other parts of this application, users can identify outliers and download clean datasets.

The *Trait-by-Trait* section has only one subsection, *Model Specs*. Users can run the model and observe the results for experiments evaluated at a single site, fitting multiple traits one at a time. This module was designed to compare the quantitative response of different variables. In plant breeding experiments, it is common to compare the behavior of one or more traits in one or more trials. This part of the application generates the same results described in the previous sections – spatial plots for each trait,

summaries, model components, heritability, genotype ranking, outlier identification, etc. It also shows the genetics correlation between traits, offering a graphical display of Pearson's second moment correlation coefficients, a dendrogram plot, and a Principal Component Analysis (PCA) for the traits and genotypes evaluated in the trial.

For beginner users, a video tutorial about *Single-Site*, *Site-by-Site*, and *Trait-by-Trait* analysis in this module is available at the following link: https://www.youtube.com/watch?v=QU_2O2ycZWA&t=303s.

2.5 *ASReml-R* module

Licensed researchers can use the *ASReml-R* and *Two-Stage-Analysis* modules. *ASReml-R* is a statistical software package for fitting linear mixed models using residual maximum likelihood (REML), as reported by Gilmour et al. (1995). The application for spatial analyses, establishes the natural variation in the data as the product of an autoregressive correlation (AR) structure for columns and rows denoted by AR1xAR1. *ASReml-R* is designed to fit the general LMM to moderately large datasets with complex variance models. The package has applications in the analysis of repeated measures data from multivariate analysis of variance and spline-type models, unbalanced design experiments, multi-environment trials, and regular or irregular spatial data (Butler et al., 2017). Many of these features are implemented in Mr.Bean.

Similar to the *SpATS* section, users can run the model for experiments in a single site using the *ASReml-R* function. Using the same interface as in previously described modules, the user selects the parameters of the response variable, genotype, and spatial coordinates with *Model Specs*. Optionally, users can include spatial coordinates (rows and columns) as splines or factors, and other covariates. The application generates spatial trend plots for raw data, fitted data, residuals, environmental variables, and genotype. It also generates a table with goodness-of-fit statistics, such as Akaike information criterion (AIC), Bayesian information criterion (BIC), heritability based on variance components (herit.VC), and heritability based on predictor error variance (herit.PEV), in addition to other statistics. Furthermore, the application generates a summary with the variance components, an ANOVA Wald test, and a 3D empirical variogram for the spatial trend of the residuals. In a *BLUPs/BLUEs* subsection, the *ASReml-R* module generates a table with predicted values and their respective standard errors and weights, a histogram of predicted values, and a ranking of genotypes using error bar plots.

In breeding trials, field experiments often test hundreds of genotypes with few or poor replications, mainly in the early stages of genotype screening. In these cases, checks are used to detect trends and allow the calculation of the residual variance. These trials using local controls assume that checks should have a similar response to the tested genotypes. Typically, augmented designs are the base for unreplicated trials, and their statistical analysis can be based on RCBD or on other spatial configurations (Gezan, 2023). For this reason, the *ASReml-R* module also allows fitting models for single-site unreplicated trials. The *Unreplicated* section presents a similar architecture to the *Single-Site* section by selecting the input parameters and the output results (spatial plots, residuals information, variance component, BLUPs, etc.). This section generates a table with goodness-of-fit statistics (AIC, BIC, herit.PEV, heritVC, A optimality, D optimality) to select the best spatial model by comparing the AR structure for columns, the AR structure for rows, or the AR structure for both spatial coordinates simultaneously.

The *ASReml-R* module can find the best spatial model for the data to be analyzed (*Model Selector* section). Similar to the other parts of this application, the user selects the available parameters. Mr.Bean then generates goodness-of-fit statistics. This section tests all the possible parameters for a model and then internally compares all the models to select the one with the best fit. Models are compared by block, complete blocks, splines, rows and columns, and the residual variance structures.

2.6 Two-stage analysis module

The *MET Analysis* function fits LMMs for multi-environmental trials using *ASReml-R*. This module has its own import data section, in a csv format, and it is independent from the other modules. Similar to the other modules, the user selects the parameters in the *Model Specs* subsection, providing the response variables, genotypes, and experiments, which are the different trials to be analyzed. The user will be able to analyze all trials of the dataset, selecting which trials to evaluate with the subset option. Additionally, there is an option allowing users to include weights in the two-stage analysis. These weights can be calculated by using the standard errors of the BLUEs, or by using the diagonal elements of the inverse of the variance covariance matrix associated with the genotype effect (Smith et al., 2001). In the option *Covariance structure*, the user can select the type of covariance structure to fit the model in the MET analysis. The list of the covariance structures being offered by Mr.Bean are diagonal (diag), uniform correlation (corv), uniform heterogeneous (corh), factor analytic 1 (FA1), factor analytic 2, (FA2), factor analytic 3 (FA3), factor analytic 4 (FA4), and US covariance matrix defined with correlations (corgh). The user can assess the data before running the model, by observing a barplot with the number of genotypes per trial, a heatmap for the shared genotypes between locations, and a barplot for means with standard errors for the selected trait.

The *Results* section shows a correlation matrix and dendrogram between trials evaluated. Also, a covariance matrix for trials is observed. Similar to the outputs of the previous modules, the application generates variance components, a summary of the model, residuals analysis, BLUPs for each genotype in each location, and a PCA biplot for the trials and genotypes (*GxE* option). Moreover, the section has a tool for comparing the model with different covariance structures using the likelihood ratio test (LR-statistic). When the factor analytic has been selected as a covariance matrix to fit the model, the *Factor analytic* section will be enabled. This section displays a bar chart for each factor selected, genotypic variance, and variance explained for each location. In addition, the latent regression can be reviewed for each of the genotypes in each of the selected factors. A dot plot with scores by genotype and a dot plot for loadings by environment is produced for each component selected.

2.7 Traditional designs module

Mr.Bean's *Traditional Designs* module addresses the common lack of information about the spatial arrangement of field plots in trials. The module uses the R package *lme4* (Bates et al., 2015) to fit an LMM without spatial correction. The user must first select the response variable and genotype, before selecting the experimental design. In Mr.Bean have been implemented some traditional experimental designs for plant breeding, such as completely randomized designs (CRD), RCBD, row-column design and alpha-lattice design. Mr.Bean provides these models to analyze data from these designs:

$y_{ij} = \mu + gen_i + \epsilon_{ij}$ for CRD.

$y_{ijk} = \mu + gen_i + rep_j + \epsilon_{ijk}$ for RCBD

$y_{ijk} = \mu + gen_i + rep_j + row_k(rep)_j + col_k(rep)_j + \epsilon_{ijk}$ for row-column design

$y_{ijk} = \mu + gen_i + rep_j + block_k(rep)_j + \epsilon_{ijk}$ for alpha-lattice design.

Where y is the trait of interest, μ is the overall mean, *gen* is the effect of the genotype, *block* is the effect of the block, *rep* is the effect of the replication, *col* and *row* are the effects of the spatial location and ϵ is the effect of the experimental error. Mr.Bean also offers the ability to specify any other model formula using the *lme4* syntax, which is similar to the regular mathematical notation for specifying linear models (Bates et al., 2015).

Like the *SpATS* module, the application provides the significance of the fixed effects in the model using the F statistic, and reports variance components, likelihood-ratio test information, and the broad-sense heritability estimate (Cullis et al., 2006), together with some regularly used information for comparing different fitted models, such as AIC and BIC. The user can also make multiple comparisons when the genotype is taken as a fixed factor. Likewise, as in previously described modules, this module provides an analysis of residuals using a QQplot, a histogram, an analysis of outliers, as well as a list of ranked genotypes.

2.8 *GBLUP* module

The last module implemented in Mr.Bean is the *GBLUP* module. The app allows integrate genomic and phenotypic data with the aim of performing genomic prediction analysis using the R-package *sommer* (Covarrubias-Pazaran, 2016). In the *Genomic Prediction* section, the user only must import the phenotypic data and the genotypic data. The markers genotypic data must be in numerical format (-1, 0, 1), import the genetic map with the physical positions of the markers is also possible. In the same section, the users only have to select the phenotypic variables they want to analyze and the model can be executed. The current method available for this kind of analysis is GBLUP.

Mr.Bean estimates the variance components with genomic predictions, marker-base heritability, and GEBVs for each trait evaluated. Accuracy data and reliability, correlation plots between predicted and observed values of GBLUPs and the estimated squared-marker effect for each physical position similar to the Genome-wide association studies (GWAS) can be observed. Finally, in the *Results* section, the app shows the predictions plot with the fitted and predicted valued for each genotype.

2.9 Testing dataset

The dataset comes from a breeding population (*Vivero Equipo Frijol* or VEF population) of common bean (*Phaseolus vulgaris L.*) developed by the Andean bean breeding program of the Alliance Bioversity-CIAT (Keller et al., 2020). For the single-site analysis, a subset of 260 genotypes of the VEF population was planted in 2022 at the Alliance Bioversity-CIAT's Palmira experimental field station (Colombia, 1,000 m a.s.l. altitude, latitude 3°32′N and longitude 76° 18′W), under drought and irrigation.

For multi-environmental trial analysis, a historical dataset of 1,142 genotypes was planted at the Palmira experiment station, and at two additional sites: Darien, Colombia, with an altitude of 1,600 m a.s.l., (latitude 3°55′N and longitude 76°29′W) and Quilichao, Colombia, with an altitude of 1,000 m a.s.l. (latitude 3°1′N and longitude 76°28′W) over a period of seven years (2013, 2014, 2015, 2016, 2017, 2018, and 2019). For Darien, the trials were planted under three levels of phosphorus concentration – high phosphorus, medium phosphorus, and low phosphorus with optimal precipitation conditions (590 mm) for these trials. For Palmira, the trials were planted under drought and irrigated conditions. In Quilichao, the trials were planted under drought conditions. In total, 13 different trials were conducted (Supplementary Table 1).

The experimental units were row plots of 2.22 m^2 laid out for each replicate of each genotype. The experimental design was an alpha-lattice with two and three replicates. Four traits were evaluated and reported in both datasets. The number of days to flowering (DF) was measured from the planting day to when 50% of the plants in the plot had at least one open flower. Days to physiological maturity (DPM) was measured as the number of days from planting until 50% of plants had at least one pod that had lost its green pigmentation. Yield (YDHA, kg ha^{-1}) was determined for each plot and corrected for seed moisture of 14%. Seed weight (SW100, g 100 $seeds^{-1}$) was obtained from 100 seeds (Diaz et al., 2020).

3 Results

3.1 Single site analysis

Mr.Bean enabled the analysis of the phenotypic distribution of SW100, DPM, DF, and YDHA for 260 lines belonging to the VEF panel dataset, evaluated in Palmira under drought and irrigated conditions in 2022 (Figure 3; Table 1). Water availability conditions (drought and irrigated) affected SW100 and YDHA, two traits that also showed the highest coefficients of variation, 0.24 and 0.14 for drought and 0.29 and 0.13 for irrigation, respectively. The

10.3389/fpls.2023.1290078

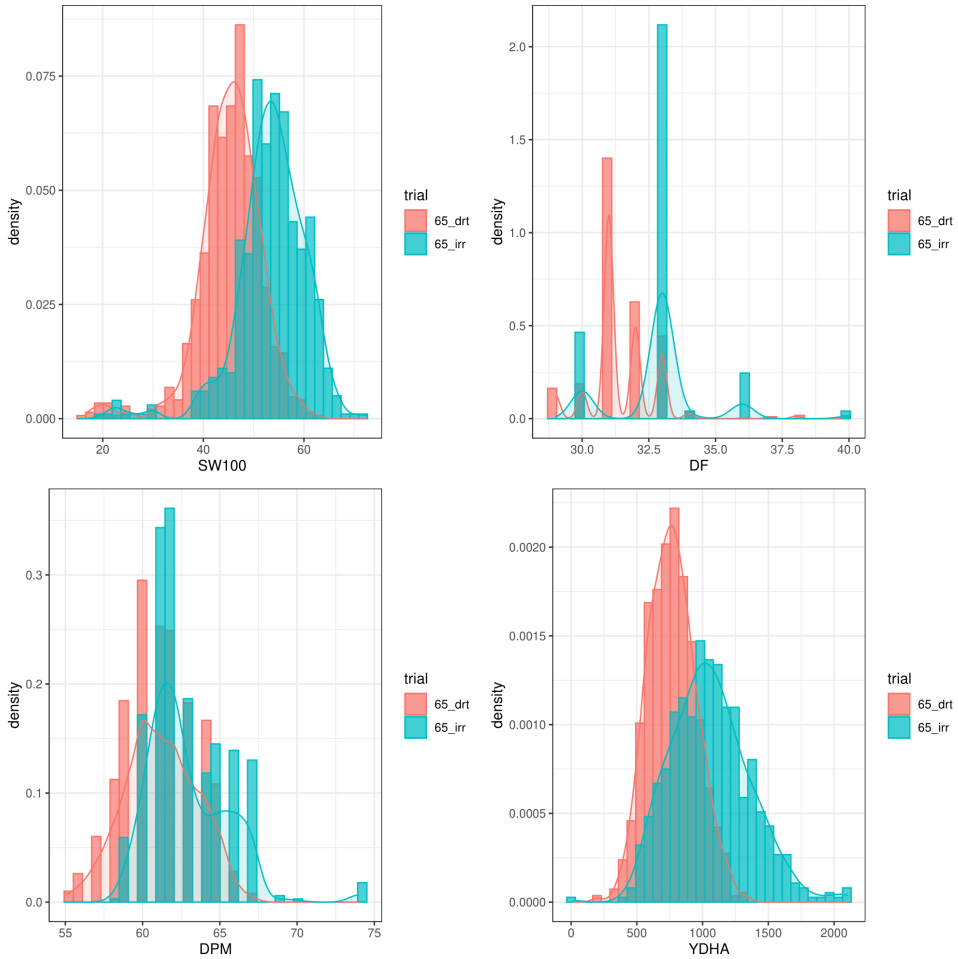


FIGURE 3
Phenotypic distribution of 100 seed weight (SW100), days to physiological maturity (DPM), days to flowering (DF) and yield (YDHA) of 260 lines belonging to VEF evaluated in drought (red plot) and irrigation (blue) conditions in 2022. (Figure generated directly by Mr.Bean).

phenotypic correlation between the traits for the two conditions is shown in the correlation plot (Figure 4). In both conditions, a strong positive correlation was observed between DF and DPM (0.68 – 0.7). On the other hand, a negative correlation was observed between DF and SW100 (-0.35 – 0.5). YDHA was negatively correlated with DF and DPM under drought conditions. However, under irrigated conditions the correlation was positive. Mr.Bean generates a clustering dendrogram from the correlation matrix and a PCA

TABLE 1 Summary statistics for phenotypic response of 100 seed weight (SW100), days to physiological maturity (DPM), days to flowering (DF) and yield (YDHA) of 260 lines belonging to VEF evaluated in drought and irrigation conditions in 2022.

	YDHA (kg ha^{-1})		DF		DPM		SW100 (g)	
	Dro	Irr	Dro	Irr	Dro	Irr	Dro	Irr
Mean	768.21	1057.64	31.51	32.89	61.18	62.87	45.31	53.48
Std. Dev	182.21	308.87	1.38	1.7	2.34	2.54	6.36	7.07
Min	191.83	12.82	29	30	55	58	16.4	20
Median	760.68	1040.48	31	33	61	62	45.6	54
Max	1341.17	2110.06	40	40	67	74	62.8	72.4
CV	0.24	0.29	0.04	0.05	0.04	1.15	0.14	0.13
Skewness	0.19	0.43	2.1	0.77	-0.03	1.15	-1.12	-1.31
Kurtosis	-0.09	0.43	10.47	3.59	-0.44	2.36	3.6	4.17

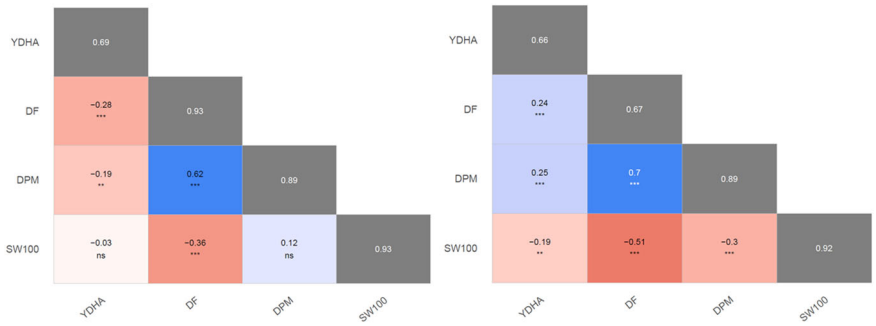


FIGURE 4
Pearson's second moment correlation coefficients and their significances between best linear unbiased estimators (BLUEs) of evaluated traits. The broad-sense heritabilities of the best linear unbiased predictors (BLUPs) are located within the diagonal with the gray background. 100 seed weight (SW100), days to physiological maturity (DPM), days to flowering (DF), and yield (YDHA) of 260 lines belonging to VEF evaluated in drought (left side) and irrigation (right side) conditions in 2022. (Figure generated directly by Mr.Bean) Significance of correlations indicated as ***: $p < .0001$; **: $p < .001$; ns, not significant.

biplot graph for the first two principal components of the distance matrix (Figure 5). The biplot shows the correlation between DF and DPM in both trial conditions (Figures 5A, B). Figure 5 also shows the differences in the performance of the Mesoamerican genotype checks compared to the Andean genotypes.

Model fitting was performed with *SpATS* (Rodriguez-Alvarez et al., 2018) and *ASReml-R* (Butler et al., 2017), using *lme4* under a row-column design (Bates et al., 2015) and considering the genotype effect as random. The heritability and variance components were then calculated. Next, the application calculated the spatial trends for raw data, fitted data, residuals, fitted spatial trend, and genotypic BLUPs for YDHA, using *SpATS* and *ASReml-R* models under drought and irrigated conditions (Figure 6 and Table 2).

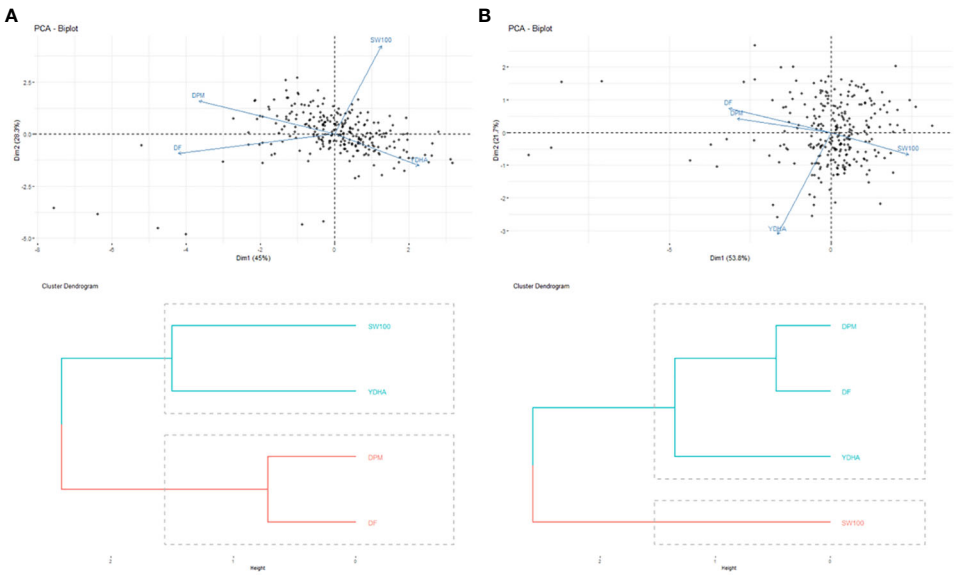


FIGURE 5
Biplot of principal components analysis (top side) and dendrograms (bottom side) of the phenotypic correlation for 100 seed weight (SW100), days to physiological maturity (DPM), days to flowering (DF) and yield (YDHA) of 260 lines (Black points) belonging to VEF population evaluated in: **(A)** drought (left side) and **(B)** irrigation (right side) conditions in 2022. (Figure generated directly by Mr.Bean).

10.3389/fpls.2023.1290078

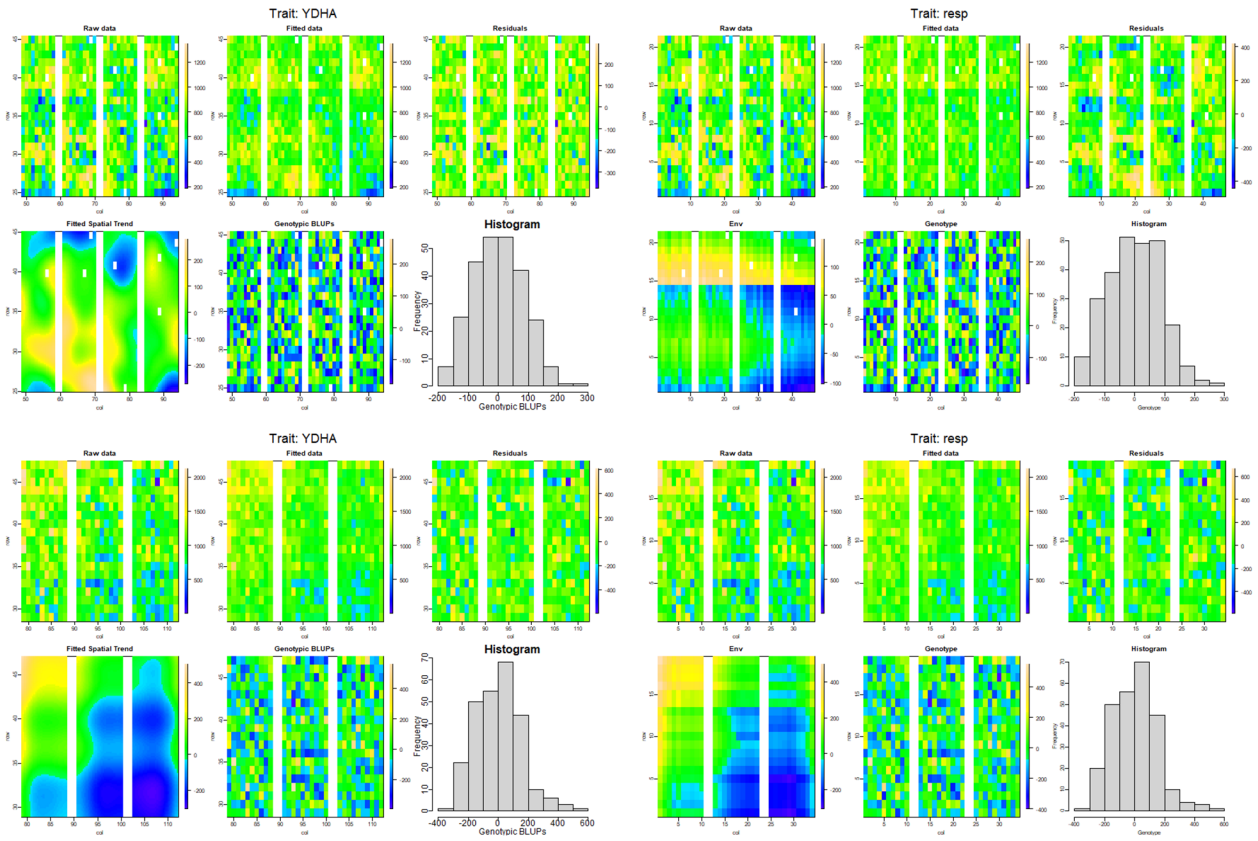


FIGURE 6
Spatial trends plots for raw data, fitted data, residuals, fitted spatial trend, and genotypic BLUPs for YDHA of 260 lines belonging to VEF population evaluated in drought (top side) and irrigation (bottom side) conditions in 2022. The models used for generating the spatial trends were SpATS (left side) and ASReml-R (right side) (Figure generated directly by Mr.Bean).

3.2 Multi-environmental trials analysis

Mr.Bean evaluated the phenotypic distribution of SW100, DPM, DF, and YDHA of the VEF population in 13 trials (Supplementary Figure 1). Similar to a single-site analysis, Figure 7 shows the results for YDHA. The phenotypic correlation for YDHA between trials is shown in the matrix and dendrogram. The trials established in Darien and Quilichao clustered around two representative groups for the 13 trials, in contrast to the trials planted in Palmira (except Pal18A_Irr) (Figure 7B).

Mr.Bean fit the model for MET analysis using *SpATS* and *ASReml-R*, with a two-factor analytic covariance matrix for YDHA. The variance components were then calculated (Table 3). A PCA biplot graph was generated for the first two principal components of the genotype distance matrix (Supplementary Figure 2). A factor analytic structure allowed the generation of scores for each genotype, loadings for each trial evaluated, and weights in the MET model (Supplementary Figure 3). The Mesoamerican check genotypes grouped around a higher positive score for the first component (Supplementary Figure 3A). Similarly, the Darien and Quilichao trials grouped around a higher score for the second component (Supplementary Figure 3B).

4 Discussion

Mr.Bean offers robust analytical tools and visualizations for plant breeders and plant scientist across different disciplines. Mr.Bean was developed by the Bean breeding program from Alliance Bioversity-CIAT in collaboration with other researches from different institutions. Initially thorough to support to Bean breeding program is today a widely sued tool by breeding programs across the world. Some of the crops and breeding programs that have successfully used Mr.Bean include common bean, tropical forages, rice and cassava breeding programs from Alliance Bioversity-CIAT; also barley, spring wheat, soybeans, dry beans breeding programs and research extensions centers at NDSU and UM that used it to analyses data for agronomic research experiments. Evaluations can be focused not only on plant breeding, but can also be applied to research in plant pathology, entomology, physiology, and other fields. The application was developed as an open-source and accessible tool with an easy-to-use graphical interface. Researchers can run Mr.Bean with any web browser.

Mr.Bean was developed in the R language, but no programming experience is required. However, researchers using R can customize

10.3389/fpls.2023.1290078

TABLE 2 Heritability and variance components for yield (YDHA), using *SpATS* (Rodriguez-Alvarez et al., 2018), *ASReml-R* (Butler et al., 2017), and row-columns design with *lme4* (Bates et al., 2015), of 260 lines belonging to the VEF panel dataset, evaluated under drought and irrigated conditions in 2022.

Model		Drought		Irrigation	
	Component	Variance	Std. Dev	Variance	Std. Dev
SpATS	**Genotype**	10260	101.3	34380	185.4
	rep:col_f	798.5	28.26	709.4	26.63
	rep:row_f	1104	33.23	1534	39.17
	f(col)	7813	88.39	17570	132.5
	f(row)	3302	57.46	57360	239.5
	f(col):row	651.5	25.52	17.63	4.199
	col:f(row)	0	0.001	0	0
	f(col):f(row)	59180	243.3	13.86	3.723
	Residual	11910	109.1	32500	180.3
	Heritability	0.69		0.66	
ASReml-R	**spline(row)**	809.769	1057.703	0.001	
	spline(col)	4128.431	4919.915	27206.11	21340.01
	rep:row	1003.913	585.48	4021.367	1801.49
	rep:col	2697.943	801.054	814.545	1134.622
	Genotype	10600.11	1499.37	33178.61	4723.381
	row:col!col	16005.85	1060.085	32862.8	2902.799
	row:col!R	1		1	
	Heritability	0.63		0.65	
Row-Columns design with lme4	**Genotype**	10621	103.06	32505	180.3
	Rep:col_f	3817	61.78	8336	91.3
	Rep:row_f	2024	44.99	12255	110.7
	Residual	16011	126.53	33871	184
	Heritability	0.63		0.62	

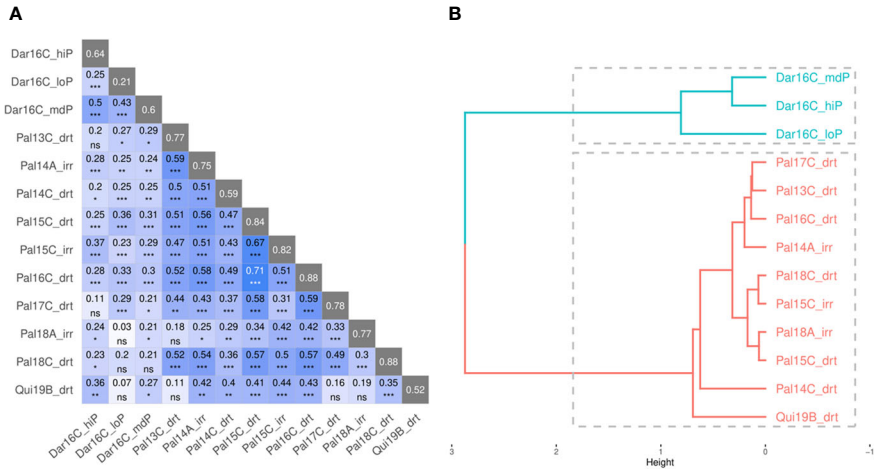


FIGURE 7
Correlation plot **(A)** and dendrogram **(B)** for yield (YDHA) of VEF population evaluated in 13 trials. (Figure generated directly by Mr.Bean). Significance of correlations indicated as ***: p < .0001; **: p < .001; *: p < .01; ns, not significant.

TABLE 3 Variance components for yield (YDHA), using *SpATS* (Rodriguez-Alvarez et al., 2018), and *ASReml-R* with one analytic factor as covariance matrix (Butler et al., 2017) of VEF population evaluated in 13 trials.

Experiment	SpATS		ASReml-R	
	varG	varE	varG	PVE (%)
Dar16C_hiP	53372.31	53845.07	52765.65	24.4
Dar16C_loP	10605.63	35326.7	36767.6	36.9
Dar16C_mdP	30384.25	38648.48	31722.67	31.5
Pal13C_drt	73510.5	57540.4	68551.71	90.1
Pal14A_irr	68039.77	60178.02	49812.17	97.8
Pal14C_drt	39602.25	48152.43	54009.83	70
Pal15C_drt	50911.35	25350.02	29782.21	100
Pal15C_irr	164304.2	94217.9	76122.49	100
Pal16C_drt	115512.9	43428.83	118722.9	89.6
Pal17C_drt	94995.45	65750.95	25156.68	100
Pal18A_irr	220896.3	165857.7	56923.65	100
Pal18C_drt	214225.1	79941.68	109404.3	96.4
Qui19B_drt	82026.66	120853.8	17325.5	100

the open-source application with individual modifications to meet their needs and requirements. Mr.Bean's individual modules are easy to understand and accessible to novice users. The workflow starts with downloading, cleaning, processing, and filtering the raw data for further analysis. The modules can be used for different analyses depending on the nature and purpose of the trials being evaluated. Users can generate graphs and tables with detailed information for future interpretation. Mr.Bean includes several visual tools such as real-time interactive statistical graphs developed in the R Shiny package. These tools support understanding and analyzing the behavior of the raw or processed data.

Mr.Bean models spatial variability – one of the major sources of error in field trials (Singh et al., 2003). The application uses linear mixed models with spatial components of field experiments implemented with *SpATS* and *ASReml-R* packages. The application accommodates traditional experimental designs lacking spatial information, such as randomized complete block designs or alpha-lattice designs, and separates genotypic variance from environmental variance. Ultimately, Mr.Bean facilitates data analysis towards improving genetic gain and making breeding programs more efficient (Covarrubias-Pazaran, 2020).

With single-site and multi-environment trial analysis, Mr.Bean enables breeders to make better use of their data and more robust decisions about genotype performance by calculating BLUEs and BLUPs for every trait and every location, within and across sites. The application estimates the selection response and provides breeders with critical tools to select the best performing genotypes. In addition, Mr.Bean can adjust any variable as a covariate to estimate its effect on the trial. The application allows multi-trait and genetic correlation analysis, allowing the development of a selection index for implementation in breeding programs. Supplementary materials and education videos can be found at github https://github.com/AparicioJohan/MrBeanApp and Youtube https://www.youtube.com/@ndsubigdatapipelineunit5201/.

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

JA: Conceptualization, Investigation, Methodology, Software, Supervision, Validation, Visualization, Writing – original draft. SG: Formal analysis, Methodology, Software, Writing – review & editing. DA-S: Conceptualization, Investigation, Software, Writing – review & editing. BR: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing. SD: Data curation, Formal analysis, Supervision, Writing – original draft, Writing – review & editing. AH-M: Investigation, Methodology, Software, Validation, Writing – review & editing. JL: Funding acquisition, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was funded by the Tropical Legumes III-Improving Livelihoods for

Smallholder Farmers: Enhanced Grain Legume Productivity and Production in Sub-Saharan Africa and South Asia (OPP1114827), and by the AVISA-Accelerated varietal improvement and seed delivery of legumes and cereals in Africa (OPP1198373) projects funded by the Bill and Melinda Gates Foundation. We would like to thank the USAID for their contributions through the CGIAR Research Program on Grain Legumes and Dryland Cereals.

Acknowledgments

We would like to thank the Bean team of the Alliance Bioversity-CIAT for their great support. We also want to thank the AES Big Data Pipeline Unit of the North Dakota State University for their constant help and knowledge in the construction of this application and we appreciate to VSN International for being part of this project.

Conflict of interest

Author SG is employed by VSN International, Author AH-M is employed by AES Big Data Pipeline Unit.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict 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.2023.1290078/full#supplementary-material

SUPPLEMENTARY TABLE 1
Combination of location, year and conditions which established VEF population in each trial.

SUPPLEMENTARY FIGURE 1
Phenotypic distribution of 100 seed weight (SW100), days to physiological maturity (DPM), days to flowering (DF) and yield (YDHA) of VEF population evaluated in 13 trials. (Figure generated directly by Mr.Bean).

SUPPLEMENTARY FIGURE 2
Biplot of the first two principal components of the correlation for yield (YDHA) of 1146 lines (Black points) belonging to the VEF population, evaluated in 13 trials (blue arrows) (Figure generated directly by Mr.Bean).

SUPPLEMENTARY FIGURE 3
Scores of 1,146 lines belonging to VEF population (a) and loading factor of 13 trials by Factor analytic (b) (Figure generated directly by Mr.Bean). The size and color of each individual point correspond to BLUE values for each environment or genotype. big size points and dark blue color correspond to environments or genotypes with higher BLUE values and small size points and yellow color correspond to environments or genotypes with lower BLUE values.

References

Alvarado, G., Rodriguez, F. M., Pacheco, A., Burgueño, J., Crossa, J., Vargas, M., et al. (2020). META-R: A software to analyze data from multi-environment plant breeding trials. *Crop J.* 8, 745–765. doi: 10.1016/j.cj.2020.03.010

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Software* 67 (1), 1–48. doi: 10.18637/jss.v067.i01

Bernardeli, A., Rocha, J. R., Borem, A., Lorenzoni, R., Aguiar, R., Basilio, J. N., et al. (2021). Modeling spatial trends and enhancing genetic selection: An approch to soybean seed composition breeding. *Crop Sci.* 61, 976–988. doi: 10.1002/csc2.20364

Butler, D. G., Cullis, B. R., Gilmour, A. R., Gogel, B. G., and Thompson, R. (2017). *ASReml-R reference manual version 4* (Hemel Hempstead, HP1 1ES, UK: VSN International Ltd).

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., et al. (2023). shiny: Web application framework for R. R package version 1.8.0.9000. Available at: https://github.com/rstudio/shiny, https://shiny.posit.co/.

Covarrubias-Pazaran, G. (2016). Genome-assisted prediction of quantitative traits using R package sommer. *PloS One* 11 (6), e0156744. doi: 10.1371/journal.pone.0156744

Covarrubias-Pazaran, G. (2020)Manual Breeding process assessment: Genetic gain as a high-level key performance indicator. In: *Excellence in breeding platform. Excellenceinbreeding.org/toolbox/tools/eib-breeding-schemeoptimization-manuals* (Accessed March 10, 2023).

Cullis, B. R., and Gleeson, A. C. (1991). Spatial analysis of field experiments-an extension to two dimensions. *Biometrics* 47, 1449–1460. doi: 10.2307/2532398

Cullis, B. R., Smith, A. B., and Coombes, N. E. (2006). On the design of early generation variety trials with correlated data. Journal of Agricultural. *Biological Environ. Stat* 11 (4), 381–393. doi: 10.1198/108571106x154443

Currie, I. D., and Durban, M. (2002). Flexible smoothing with P-splines: a unified approach. *Stat. Modeling* 2 (4), 333–349. doi: 10.1191/1471082x02st039ob

Cursi, D. E., Gazaffi, R., Hoffmann, H. P., Brasco, T. L., do Amaral, L. R., and Neto, D. D. (2021). Novel tools for adjusting spatial variability in the early sugarcane breeding stage. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.749533

Diaz, S., Ariza-Suarez, D., Izquierdo, P., Lobaton, J. D., de la Hoz, J. F., Acevedo, F., et al. (2020). Genetic mapping of agronomic traits in a MAGIC population of common bean (Phaseolus vulgaris L.) under drought conditions. *BMC Genomics* 21, 799. doi: 10.1186/s12864-020-07213-6

Gezan, S. A. (2023) *Unreplicated trials: What can they really do? Part 1*. Available at: https://vsni.co.uk/blogs/unreplicated-trials-part-1 (Accessed March 15, 2023).

Gilmour, A. R., Thompson, R., and Cullis, B. R. (1995). Average information REML, an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* 51, 1440–1450. doi: 10.2307/2533274

Harrison, R. J., and Caccamo, M. (2022). "Managing data in breeding, selection and in practice: A hundred year problem that requires a rapid solution," in *Towards responsible plant data linkage: Data challenges for agricultural research and development*. Eds. H. F. Williamson and S. Leonelli (Springer, Cham), 37–64. doi: 10.1007/978-3-031-13276-6_3

Isik, F., Holland, J., and Malteca, C. (2017). "Spatial analysis," in *Genetic data analysis for plant and animal breeding* (Springer, Cham). doi: 10.1007/978-3-319-55177-7

Keller, B., Ariza-Suarez, D., de la Hoz, J., Aparicio, J. S., Portilla-Benavides, A. E., Buendia, H. F., et al. (2020).). Genomic prediction of agronomic traits in common bean (Phaseolus vulgaris L.) under environmental stress. *Front. Plant Sci.* 711. doi: 10.3389/fpls.2020.01001

Mackay, I., Piepho, H.-P., and Franco, A. A. F. (2019). "Statistical methods for plant breeding," in *Handbook of statistical genomics*. Eds. D. Balding, I. Moltke and Marioni,. J. doi: 10.1002/9781119487845.ch17

Mao, X., Dutta, S., Wong, R. K., and Nettleton, D. (2020). Adjusting for spatial effects in genomic prediction. Journal of Agricultural. *Biol. Environ. Stat* 25, 699–718. doi: 10.1007/s13253-020-00396-1

Piepho, H., Boer, M. P., and Williams, E. R. (2022). Two-dimensional P-splines smoothing for spatial analysis of plant breeding trials. *Biometrical J.* 64, 5. doi: 10.1002/bimj.202100212

Piepho, H., Mohring, J., Melchinger, A., and Buchse, A. (2008). BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica* 161, 209–228. doi: 10.1007/s10681-007-9449-8

Piepho, H., and Williams, E. (2010). Linear variance models for plant breeding trials. *Plant Breed.* 129 1, 1–8. doi: 10.1111/j.1439-0523.2009.01654.x

Robbins, K., Backlund, J., and Schnelle, K. (2012). Spatial corrections of unreplicated trials using a two-dimensional spline. *Crop Sci.* 52 (3), 1138–1144. doi: 10.2135/cropsci2011.08.0417

Rodriguez-Alvarez, M. X., Boer, M. P., van Eeuwijk, F. A., and Eilers, P. H. (2018). Correcting for spatial heterogeneity in plant breeding experiments with P-splines. *Spatial Stat* 23, 52–71. doi: 10.1016/j.spasta.2017.10.003

Sievert, C. (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. (Chapman and Hall/CRC). Available at: https://plotly-r.com.

Singh, M., Malhotra, R. S., Ceccarelli, S., Sarker, A., Grando, S., and Erskine, W. (2003).). Spatial variability models to improve dryland field trials. *Exp. Agric.* 39, 151–160. doi: 10.1017/S0014479702001175

Smith, A., Cullis, B., and Gilmour, A. (2001). Applications: the analysis of crop variety evaluation data in Australia. *Aust. New Z. J. Stat* 43 (2), 129–145. doi: 10.1111/1467-842X.00163

Veturi, Y., Kump, K., Walsh, E., Ott, O., Poland, J., Kolkman, J. M., et al. (2012). Multivariate mixed linear model analysis of longitudinal data: an information-rich statistical technique for analyzing plant disease resistance. *Analytical Theor. Plant Pathol.* 102 (11), 1016–1025. doi: 10.1094/PHYTO-10-11-0268

Wickham, H. (2014). Tidy data. *J. Stat. Software* 59 (10), 1–23. doi: 10.18637/jss.v059.i10

Xu, Y., Zhang, X., Li, H., Zheng, H., Zhang, J., and Olsen, M. (2022). Smart breeding driven by big data, artificial intelligence, and integrated genomic-enviromic prediction. *Mol. Plant* 15 (1), 1664–1695. doi: 10.1016/j.molp.2022.09.001

Yan, W. (2021). A systematic narration of some key concepts and procedures in plant breeding. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.724517

Zystro, J., Colley, M., and Dawson, J. (2018). "Alternative experimental designs for plant breeding," in *Plant breeding reviews*. Ed. I. Goldman. doi: 10.1002/9781119521358.ch3

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 10 January 2024
DOI 10.3389/fpls.2023.1325961

OPEN ACCESS

EDITED BY
Xueqiang Wang,
Zhejiang University, China

REVIEWED BY
Sang-Ho Kang,
National Institute of Agricultural Science, Republic of Korea
Guoxiang Jiang,
Chinese Academy of Sciences (CAS), China

*CORRESPONDENCE
Quan Yang
yangquan@gdpu.edu.cn

RECEIVED 22 October 2023
ACCEPTED 05 December 2023
PUBLISHED 10 January 2024

CITATION
Gao H, Shi M, Zhang H, Shang H and Yang Q (2024) Integrated metabolomic and transcriptomic analyses revealed metabolite variations and regulatory networks in *Cinnamomum cassia* Presl from four growth years.
Front. Plant Sci. 14:1325961.
doi: 10.3389/fpls.2023.1325961

Integrated metabolomic and transcriptomic analyses revealed metabolite variations and regulatory networks in *Cinnamomum cassia* Presl from four growth years

Hongyang Gao[1], Min Shi[1], Huiju Zhang[1], Hongli Shang[1] and Quan Yang[1,2,3,4]*

[1]School of Chinese Materia Medica, Guangdong Pharmaceutical University, Guangzhou, China, [2]Guangdong Provincial Research Center on Good Agricultural Practice & Comprehensive Agricultural Development Engineering Technology of Cantonese Medicinal Materials, Guangzhou, Guangdong, China, [3]Comprehensive Experimental Station of Guangzhou, Chinese Material Medica, China Agriculture Research System (CARS-21-16), Guangzhou, China, [4]Key Laboratory of State Administration of Traditional Chinese Medicine for Production & Development of Cantonese Medicinal Materials, Guangzhou, Guangdong, China

To understand the mechanism of the dynamic accumulation of active ingredients in *Cinnamomum cassia* Presl, metabolomic and transcriptomic analyses of 5~8 years old *C. cassia* were performed. A total of 72 phenylpropanoids, 146 flavonoids, and 130 terpenoids showed marked changes. Most phenylpropanoids and flavonoids showed markedly higher abundances in 6-year-old *C. cassia* than in others, which was related to the higher expression of genes that synthesize and regulate phenylpropanoids and flavonoid. We identified transcription factors (TFs) and genes involved in phenylpropanoids and flavonoids synthesis and regulation through co-expression network analyses. Furthermore, most of the terpenoids in 5-year-old *C. cassia* showed markedly higher abundances than in others, which was due to the differentially expressed genes upstream of the terpenoids pathway. The results of our study provide new insights into the synthesis and accumulation of phenylpropanoid, flavonoids and terpenoids in *C. cassia* at four growth stages.

KEYWORDS

***Cinnamomum cassia* Presl, phenylpropanoids, flavonoids, terpenoids, transcriptome**

1 Introduction

Cinnamomum cassia Presl is a perennial arborous plant of the Lauraceae family, which is an important cash crop in many countries in the world and is widely used in many fields, such as chemical industry, food, and medicine (Jeyaratnam et al., 2016). The bark of *C. cassia*, an important traditional medicinal and edible plants, is often used as a spice to add flavor and aroma to food. It also has anti-inflammatory, hypoglycemic, anti-oxidant, anti-tumor and other pharmacological activities (Koppikar et al., 2010; Shin et al., 2017; Kang and Lee, 2018). In addition, cinnamon essential oil has a broad antibacterial spectrum, can inhibit foodborne pathogens and putrefactive bacteria (Vijayan and Mazumder, 2018).The edible film made by combining with oxidized hydroxypropyl cassava starch has better performance and can be used as packaging materials for fruits and vegetables and food, which can inhibit the pollution of foodborne pathogens and spoilage bacteria and extend the shelf life of food (Zhang et al., 2016; Zhou et al., 2021).

C. cassia has a long growth cycle, which requires at least 4–6 years of growth, sometimes even decades of growth. The bark of *C. cassia* often harvested from 5~8-year-old trees. In recent years, the demand for *C. cassia* in the international market has increased, leading to differences in the harvesting period of *C. cassia* and affecting its quality. The main active component of *C. cassia* is volatile oil, which consists of cinnamaldehyde, cinnamic acid, coumarin, sesquiterpene, and diterpene. In addition, *C. cassia* contains flavonoids, anthocyanins, and other non-volatile components. Among them, cinnamaldehyde, cinnamic acid, coumarin, flavonoids, anthocyanins, and other substances are directly or indirectly produced through phenylpropanoid biosynthesis (Fraser and Chapple, 2011). Terpenoids are synthesized by two different metabolic pathways: the mevalonate (MVA) pathway and the 2-c-methyl-d-erythritol 4-phosphate (MEP) pathway. The composition and content of volatile oil in *C. cassia* are affected by growth years and other factors (Geng et al., 2011). Li et al. (2013) studied the development of oil cells in *C. cassia* leaves of different ages and found that the density of oil cells in leaves of 2-year-old branches was the highest, which directly affected the content of cinnamaldehyde. In addition, Geng et al. (2011) measured the content and composition of volatile oil of *C. cassia* aged from 1 to 12 years and found that the yield and composition fluctuated at each development stage, with the situation first increasing and then decreasing. The cinnamaldehyde content in 6-year-old *C. cassia* is the highest, but its molecular mechanism has not been clarified. In the past, most research on *C. cassia* with different growth years has focused on the differences in chemical components, while research on the synthesis pathway and molecular regulation of effective components in *C. cassia* with different growth years has not been carried out.

At present, integrative analysis of metabolome and transcriptome has been successfully applied to the study of synthesis and regulatory mechanisms of active ingredients in plant. The molecular mechanism of different accumulations of phenylpropanoids, flavonoids, and terpenoids in *Ginkgo biloba* was systematically studied by metabonomics and transcriptomics, and the expression levels of related synthetic genes and regulatory effects of transcription factors (TFs) were analyzed (Meng et al., 2019; Guo et al., 2020). At the same time, researchers have successfully revealed the biological molecular mechanism of effective substance synthesis in *Carthamus tinctorius*, *Dendrobium officinale, Lonicera japonica* Thunb and other plants through integrative analysis of transcriptome and metabolome (Xue et al., 2019; Wang et al., 2021; Li et al., 2022). Although previous studies have used transcriptomics and metabolomics to analyze metabolites and genes in different *C. cassia* tissues. The content differences of active components such as active flavonoids in bark, branches and leaves of *C. cassia* were revealed, and the differentially expressed genes that may affect the synthesis of active components in cinnamon were identified (Gao et al., 2023). However, there is a lack of extensive and comprehensive research on the synthesis of effective substances in *C. cassia* with different growth years.

This study systematically analyzed the differences in gene expression and metabolism between the bark of *C. cassia* aged 5~8 years. Integrative analysis of metabolome and transcriptome were used to study the correlation between these DEGs (differentially expressed genes), TFs and DAMs (differentially accumulated metabolites) in the synthesis pathway of phenylpropanoids, flavonoids, and terpenoids. The results provided theoretical basis for studying the internal mechanism of effective component accumulation and quality formation of *C. cassia* and lay a foundation for efficient cultivation of *C. cassia* and increase the yield of volatile oil of cinnamon.

2 Materials and methods

2.1 Plant materials

C. cassia, aged 5, 6, 7, and 8 years, collected from Sili Village, Tanbin Town, Yunfu City, Guangdong Province (22°50'52"N, 111° 24'35"E), were used in this experiment, and the selected trees had the same cultivation and management conditions. On October 25, 2019, we peeled the bark of 5~8-year-old *C. cassia* about 1 m above the ground, and the collected bark then stored in a −80°C refrigerator for a maximum of a week.

2.2 Metabolite extraction and profiling

Organic reagents were used to extract metabolites from the cinnamon samples. From samples of the same year, 50 μL of filtered extract was mixed as a QC sample. Non-targeted metabonomic analysis based on liquid chromatography-tandem mass spectrometry was used to detect metabolites in 5~8-year-old samples. Six replicates were made for each sample. Chromatographic analysis was performed with ACQUITY UPLC HSS T3 column (Waters). The column temperature was set at 50°C and 5 μL was injected each time. Water containing 0.1% formic acid and methanol containing 0.1% formic acid were used as mobile phase for gradient elution at a flow rate of 0.4ml/min. The products eluted from the chromatographic column were collected using a

mass spectrometer Xevo G2-XS QTOF (Waters, UK) in both positive and negative ion detection modes.

PCA and PLS-DA were used to determine the metabonomic differences among 5~8-year-old samples, and the DAMs were screened based on the conditions of VIP ≥ 1, q-value<0.05 and fold change ≥ 1.2 or ≤ 0.8333. The DAMs between the comparison groups were annotated into the corresponding pathway in the KEGG database, and the significant enrichment pathways of metabolites were screened. TBtools 1.098 was used to create metabolite intensity heatmaps.

2.3 RNA extraction and transcriptome analysis

Using the plant total RNA extraction kit (TIANGEN) to extract RNA. After purification and fragmentation, it was reverse transcribed into cDNA, and then terminal repair was performed. Finally, the poly(A) tail and adaptor was added for PCR amplification and the DNA library was obtained after the amplification product was purified. Deep sequencing of the transcriptome was then carried out on the BGISEQ-500 sequencing platform of the BGI Gene.

SOAPnuke 1.4.0 was used to filter out reads containing low-quality, contaminated joints, and high levels of unknown base N from the raw data obtained from machine sequencing. Then use Bowtie2 2.2.5 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) to compare clean reads to the reference gene sequence. Assembled clean reads using Trinity 2.0.6 software to obtain Unigenes. The Unigenes were compared with seven functional databases NR (ftp://ftp.ncbi.nlm.nih.gov/blast/db), NT (ftp://ftp.ncbi.nlm.nih.gov/blast/db), SwissProt (http://www.expasy.ch/sprot/), KEGG (http://www.genome.jp/kegg), KOG (https://www.ncbi.nlm.nih.gov/COG/), Pfam (http://pfam.xfam.org) and GO (http://geneontology.org) for annotations. RSEM 1.2.8 software was used to calculate the Fragments Per Kilobase Million (FPKM) value of expression, and FPKM > 0.3 was considered differential expression. The difference of gene expression of *C. cassia* in each year was analyzed with 5-year olds as the control. Use DEseq2 to screen for differentially expressed genes, with the screening condition set to q-value ≤ 0.05. Then KEGG enrichment analysis was performed and q-value ≤ 0.05 was considered as significant enrichment. The detailed transcriptome data has been submitted to the NCBI Public Library, with the Sequence Read Archive (SRA) number PRJNA1041972.

2.4 qRT-PCR analysis

Twelve DEGs on the pathway of flavonoids, phenylpropanoids, and terpenoids were selected for qRT-PCR validation. According to the sequence obtained, primers were designed using Primer Quest (Supplementary Table S1), and TB Green® Premix Ex Taq™ II (Takara) was used to conduct qRT-PCR. The thermal cycling conditions were as follows: pre-denaturation at 95°C for 30 s, followed by 40 cycles of 95°C for 5 s and 59°C for 30 s. The melting curve was formed to evaluate the specificity of the expansion product, and the gene expression were calculated by $2^{-\Delta\Delta CT}$.

2.5 Integrative analysis of metabolome and transcriptome

Based on the annotation results of DAMs and DEGs on the KEGG pathway, the gene FPKM values and metabolite intensity in each age group of cinnamon samples were Z-score standardized and a heatmap was drawn. The correlation between DAMs, TFs, and DEGs were calculated using the Pearson correlation coefficient method, with screening conditions of Pearson correlation coefficient >| 0.8 |, P value<0.05. Then Cytoscape 3.7.1 was used to map the network relationship.

3 Results

3.1 Metabonomic analysis of 5~8-year-old C. cassia

C. cassia samples aged 5~8 years were analyzed using UPLC-MS/MS, and ions with RSD ≤ 30% were selected for subsequent analysis. PCA and heatmap cluster analysis showed that there were significant differences between 5~8-year-old samples, indicating that growth years had a greater impact on the accumulation of effective metabolites in *C. cassia* (Figure 1A; Supplementary Figure S1). The 5-year-old samples were used as controls to screen the DAMs. Among them, there were 2,586 metabolites with significant differences in year6-*vs.*-year5, of which 1,045 increased and 1,541 decreased. In year 7-*vs.*-year 5, 1952 DAMs were screened, of which 730 increased and 1,222 decreased. In year8-*vs.*-year5, there were 1,357 different metabolites, of which 685 increased and 672 decreased (Figure 1B). There were 359, 334, 156, 322, 262, and 163 unique DAMs in the year6-*vs.*-year5, year7-*vs.*-year5, year7-*vs.*-year6, year8-*vs.*-year6, and year8-*vs.*-year7 comparison groups, respectively (Figure 1C). The screened DAMs from year6-*vs.*-year5, year7-*vs.*-year5, and year8-*vs.*-year5 were analyzed for KEGG pathway enrichment, and when Qvalue ≤ 0.05 it was considered as significant enrichment. The results indicated that the three groups of DAMs were significantly enriched in biosynthetic pathways of phenylpropanoid, flavonoid, flavone and flavonol, isoflavonoid, monoterpenoid, diterpenoid, sesquiterpenoid and triterpenoid (Figures 1D–F). This indicates that the accumulation of phenylpropanoids, flavonoids, and terpenoids in *C. cassia* changed with the growth years.

3.2 DAMs in 5~8-year-old C. cassia

There were 72, 146, and 130 DAMs related to phenylpropanoids, flavonoids, and terpenoids were screened from 5~8-year-old *C. cassia*. The intensity of different metabolites in 5~8-year-old cinnamon was standardized, and the heatmap was drawn.

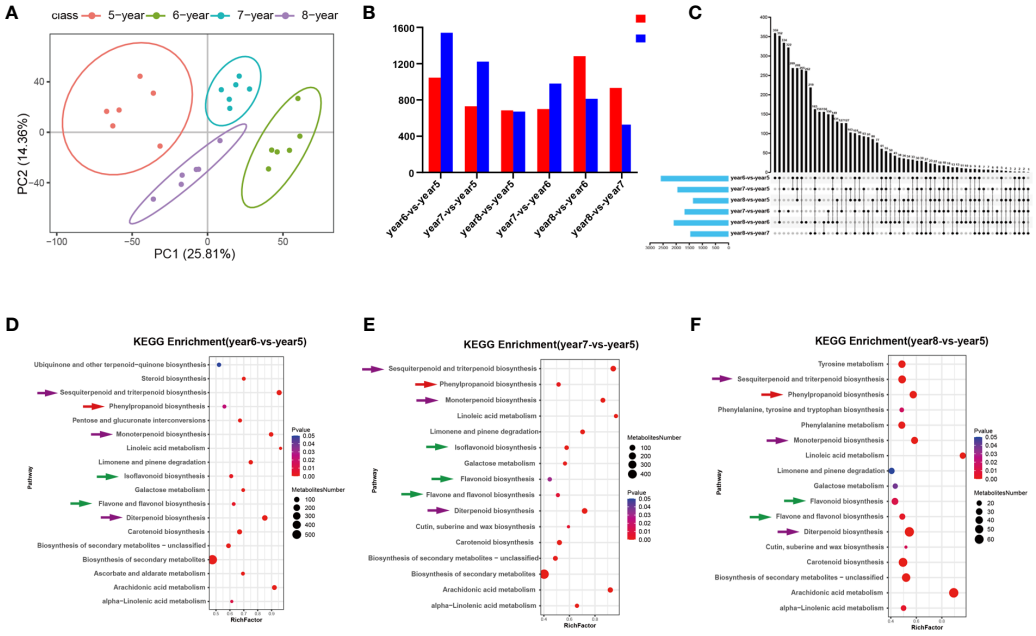


FIGURE 1
Statistical analysis of metabonomic data of 5~8-year-old *C. cassia*. **(A)** Principal component analysis diagram of test sample and quality control sample in negative ion mode. **(B)** Comparing 5~8-year-old *C. cassia* in pairs to obtain the number of different metabolites upregulated and downregulated in each group. **(C)** Upset Plot set diagram of the different metabolites. The left histogram shows the total number of DAMs included in each group comparison. The lower part of the intersection point represents the corresponding comparison group on the left, and the bar graph on the top represents the amount of DAMs shared under the intersection condition. **(D)** KEGG enrichment analysis of DAMs in year6-*vs.*-year5. **(E)** KEGG enrichment analysis of DAMs in year7-*vs.*-year5. **(F)** KEGG enrichment analysis of DAMs in year8-*vs.*-year5.

Among these, the DAMs related to phenylpropanoid substances were divided into five subtypes: phenylpropanoic acids, hydroxycinnamic acids and derivatives, coumarins and derivatives, benzoic acids and derivatives and cinnamaldehydes. Coumarin and hydroxycinnamic acid had the largest accumulation in 8-year-old *C. cassia*, while phenylpropionic acid, benzoic acid, and cinnamaldehyde had the largest accumulation in 6-year-old *C. cassia* (Figure 2A). The DAMs related to flavonoids were divided into flavanol, flavanone, anthocyanidin, flavone, flavonol, and isoflavone, and most of them accumulated in 6-year-old *C. cassia* (Figure 2B). A total of 130 terpenoid-related DAMs were divided into 5 subtypes. Most terpenoids had the largest accumulation in 5-year-old *C. cassia*. All tetraterpenoids had the largest accumulation in 5-year-old *C. cassia* (Figure 2C).

3.3 Transcriptome sequencing analysis

Transcriptome sequencing of *C. cassia* samples from four growth years yielded a total of 526.03 million clean reads. After removing some low-quality sequences, 509.32 million clean reads were obtained, with a total base number of 76.4 Gb. The sequencing data quality evaluation results showed that the Q30 of each sample was ≥ 92.52% (Supplementary Table S2), which indicated that the sequencing results were reliable and could be analyzed in the next step. After assembling clean reads using Trinity software, a total of 131372 Unigenes were obtained and the average length of these Unigenes is 1113nt. Among these unigenes, 31,988 unigenes were 200–300 nt in length, and 99,384 unigenes were longer than 300 nt (Supplementary Figure S2).

3.4 Analysis of DEGs in 5–8-year-old *C. cassia*

A total of 44,455 DEGs were screened by comparing 5~8-year-old samples in pairs. In year6-*vs.*-year5, year7-*vs.*-year5, year8-*vs.*-year5, year7-*vs.*-year6, year8-*vs.*-year6, and year8-*vs.*-year7, 28,837, 22,099, 20,144, 21,794, 14,354, and 12,721 DEGs, respectively, were counted. Among them, except for year8-*vs.*-year7, there were more downregulated genes than upregulated genes in other comparison groups (Figure 3A). The Upstet Plot set diagram directly showed the distribution of DEGs in each comparison group. Among them, there were 411 DEGs in common among the 6 comparison groups, and 3,296, 1,776, 1,177, 1,606, 774, and 466 unique DEGs in the 6 comparison groups. In general, most DEGs were found in the year6-*vs.*-year5 comparison group (Figure 3B). Getorf 6.5.7.0 was used to predict the ORF of unigenes, and hmmsearch 3.0 was used to compare them with the TF protein domain. A total of 2641 TF coding genes belonging to 56 TF families were detected. Among them, the six families with the largest number of TFs were MYB (282), C2H2 (274), bHLH (197), C3H (146), ERF (141), and NAC (120). The 44,455 DEGs were compared with the 2641 TF-coding genes. A total of 53 TF families were identified, including 1,588 differentially expressed TFs. The six families with the most TFs were C2H2, MYB, bHLH, ERF, NAC, and C3H, with 167, 152, 134, 93,

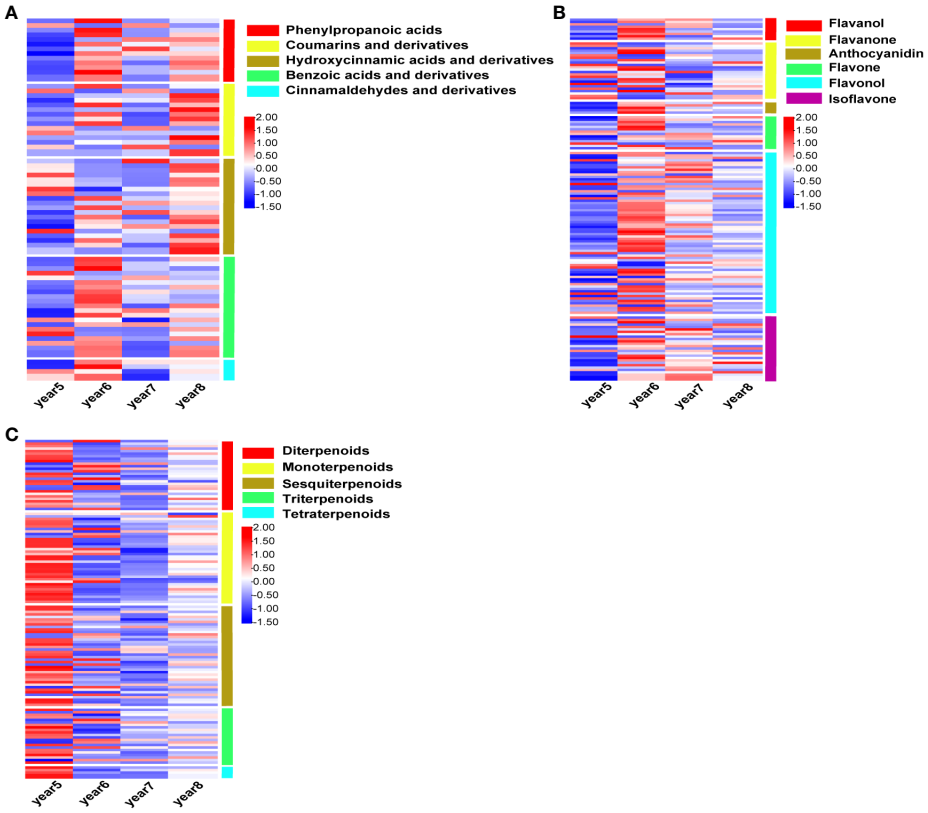


FIGURE 2
Heat map of intensity of phenylpropanoids **(A)**, flavonoids **(B)**, and terpenoids **(C)**, standardized by Z-score in 5~8-year-old *Cinnamomum cassia*. Red indicates high accumulation, blue indicates low accumulation, and the right side of the heat map shows the classification of DAMs.

83, and 81 TFs, respectively (Supplementary Table S3). In order to verify the reliability of the transcriptome data, 12 DEGs related to phenylpropanoid biosynthesis, flavonoid biosynthesis, and terpenoid biosynthesis were selected for RT-qPCR validation. The RT-qPCR results were consistent with the expression patterns in the RNA-seq analysis. This indicated that the results of RNA-seq analysis have high repeatability and reliability (Supplementary Figure S3).

KEGG pathway enrichment analysis was performed on DEGs selected from comparison groups of year6-*vs.*-year5, year7-*vs.*-year5, and year8-*vs.*-year5. The results indicated that DEGs in the three groups were mainly concentrated in three biosynthesis pathways: phenylpropanoid biosynthesis, flavonoid biosynthesis, and terpenoid backbone biosynthesis (Figures 3C–E). KEGG enrichment results of transcriptome data were consistent with KEGG enrichment results of metabolic data.

3.5 Integrative analysis of metabolome and transcriptome of phenylpropanoid biosynthesis

To understand the differences of key genes expression levels and metabolites content in phenylpropanoid biosynthesis, flavonoid biosynthesis, and terpenoid biosynthesis during the development of *C. cassia*, heat maps were used to visually display the expression patterns of metabolites and genes in 5~8-year-old *C. cassia*. In the phenylpropanoid pathway, 30 genes encoding phenylpropanoid biosynthesis-related enzymes were identified. It included 4 cinnamate 4-hydroxylase (*C4H*), 4 phenylalanine ammonia-lyase (*PAL*), 4 peroxidase (*PRX*), 3 caffeoyl-CoA O-methyltransferase (*CCoAOMT*), 3 beta-glucosidase (*BGL*), 3 caffeic acid 3-O-methyltransferase (*COMT*), 3 4-coumarate-CoA ligase (*4CL*), 3 cinnamyl-alcohol dehydrogenase (*CAD*), 2 cinnamoyl-CoA reductase (*CCR*), and 1 ferulic acid-5-hydroxylase (*F5H*). Cinnamaldehyde is the main active component of cinnamon, and the accumulation of cinnamaldehyde reached the highest level in 6-year-old *C. cassia*, which may cause by the high expression of *CCR1*. At the same time, *PAL1* and *PAL2* were highly expressed in 8-year-old *C. cassia*, and the downstream metabolite cinnamic acid also reached maximum accumulation in 8-year-old *C. cassia*. In the lignin synthesis pathway, 6-year-old *C. cassia* had the highest product accumulation, and *C4H4*, *4CL3*, *CCoAOMT2*, *CCoAOMT3*, *CAD1*, *CAD3*, *F5H*, *PRX1*, *PRX2*, *PRX3*, *PRX4*, *COMT1*, and *COMT3* showed similar change patterns (Figure 4A). In the diagram of the regulatory network, 21 TFs, 15 DEGs, and 5 DAMs related to phenylpropanoid biosynthesis were highly correlated. Among them, *CCR1* has a positive regulatory effect on cinnamaldehyde synthesis, and TF *Tify1* was significantly related to most metabolites, genes, and TFs (Figure 4B).

10.3389/fpls.2023.1325961

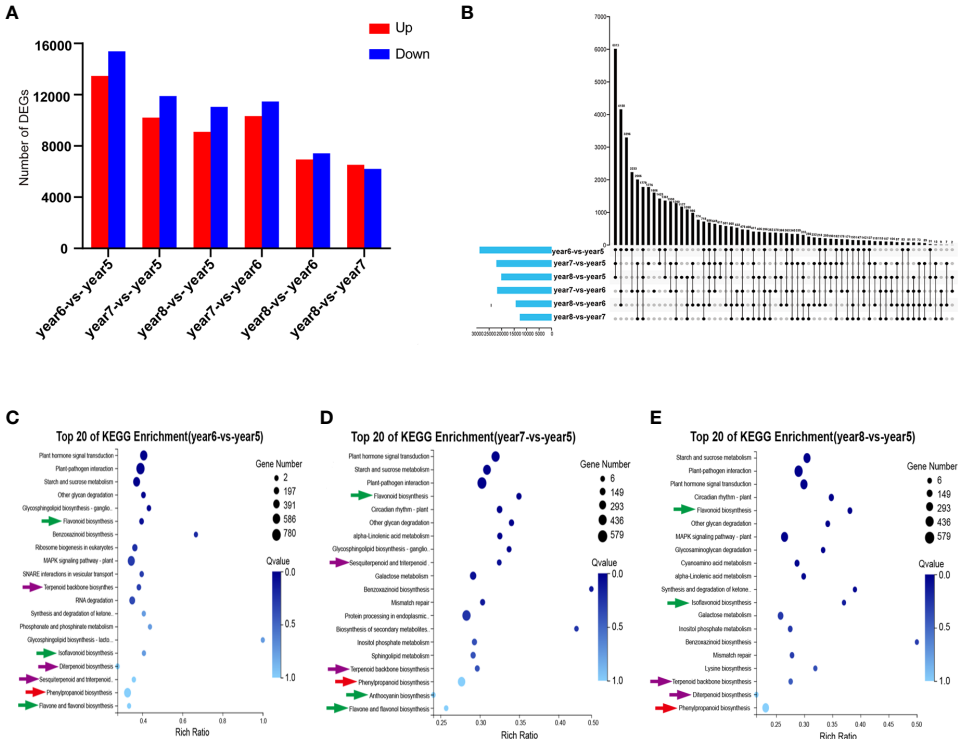


FIGURE 3
Statistical analysis of transcriptome of 5~8-year-old *C. cassia*. **(A)** The number of DEGs upregulated and downregulated in each group after comparing 5~8-year-old *C. cassia* in pairs. **(B)** The Upstet Plot set diagram of DEGs in each group after comparing 5~8-year-old *C. cassia* in pairs. The left histogram shows the total number of DEGs included in each group comparison. The lower part of intersection point represents the corresponding comparison group on the left, and the bar graph on the top represents the amount of DEGs shared under the intersection condition. **(C)** KEGG enrichment analysis of DEG in year6-*vs.*-year5. **(D)** KEGG enrichment analysis of DEGs in year7-*vs.*-year5. **(E)** KEGG enrichment analysis of DEGs in year8-*vs.*-year5.

3.6 Integrative analysis of metabolome and transcriptome of flavonoid biosynthesis

There were 32 key synthetase genes in the flavonoid biosynthesis pathway: 4 Flavanone 3-hydroxylase (*F3H*), 4 flavonol synthase (*FLS*), 3 chalcone synthase (*CHS*), 3 bifunctional dihydroflavonol 4-reductase (*DFR*), 3 *4CL*, 3 leucoanthocyanidin reductase (*LAR*), 4 anthocyanidin reductase (*ANR*), 2 anthocyanidin synthase (*ANS*), 2 chalcone isomerase (*CHI*), 3 flavonol 3-O-glucosyltransferase (*UFGT*) and 1 flavonoid 3'-hydroxylase (*F3'H*). Most flavonoid metabolites were highest in 6-year-old *C. cassia*, which is consistent with the metabolome results and its gene expression pattern (Figure 5A). Among them, the high expression of *F3H* and *FLS* in 6-year-old *C. cassia* made dihydrokaempferol and kaempferol accumulate the highest. In addition, 29 TFs, 20 DEGs, and 7 differentially expressed metabolites constituted a diagram of the regulatory network. *F3H2* and *FLS1* had the strongest correlations with DAMs and TFs. They were positively correlated with metabolites afzelechin, kaempferol, epiafzelechin, and dihydroquercetin and negatively correlated with most TFs. This indicates that *F3H2* and *FLS1* may play a crucial role in regulating the synthesis of flavonoid metabolite. In addition, the correlation between transcription factor *ERF2* and *F3H2* and *FLS1* is high, indicating that *ERF2* may participate in the synthesis of flavonoids by affecting the expression of *F3H2* and *FLS1*, which needs further validation (Figure 5B).

3.7 Integrative analysis of metabolome and transcriptome of terpenoid biosynthesis

29 DEGs were discovered in the terpenoid biosynthesis pathway: 4 isopentenyl-diphosphate Delta-isomerase (*IDI*), 3 1-deoxy-D-xylulose-5-phosphate synthase (*DXS*), 1 mevalonate kinase (*MVK*), 2 farnesyl diphosphate synthase (*FDPS*), 1 2-C-methyl-D-erythritol 2,4-cyclopyrophosphate synthetase (*IspF*), 1 diphosphomevalonate decarboxylase (*MVD*), 1 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase (*IspD*), 2 hydroxymethylglutaryl-CoA synthase (*HMGCS*), 1 1-deoxy-D-xylulose-5-phosphate reductoisomerase (*DXR*), 2 (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase (*GcpE*), 1 phosphomevalonate kinase (*PMVK*), 2 geranylgeranyl diphosphate synthase, type II (*GGPS*), 2 4-Hydroxy-3-methylbut-2-enyl diphosphate reductase (*HDR*), 2 acetyl-CoA C-acetyltransferase (*AACT*), 2 hydroxymethylglutaryl-CoA reductase (*HMGCR*), and 2 geranyl diphosphate synthase (*GPS*). Isoopentenyl diphosphate (IPP) can be synthesized by two routes: MVA pathway and MEP pathway. In the MVA pathway, most of the DEGs were highly expressed in 5-year-old *C. cassia*, which was corresponds to

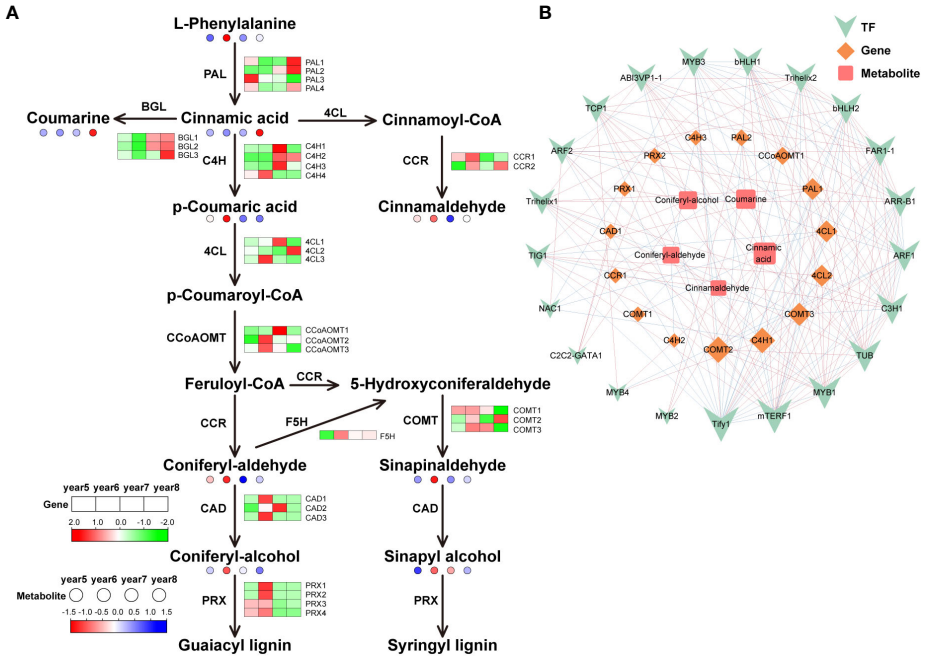


FIGURE 4
Integrative analysis of transcriptome and metabolome of the phenylpropanoid biosynthetic pathway in 5~8-year-old *C. cassia*. **(A)** Phenylpropanoid biosynthesis pathway constructed with DAMs and DEGs. Red and green boxes represent structural genes with upregulated and downregulated expression, respectively, while red and blue dots represent metabolites with upregulated and downregulated accumulation, respectively. **(B)** Correlation network diagram of phenylpropanoids. Among them, the high positive correlation is connected by red lines, and the high negative correlation is connected by blue lines. The size of the icon represents the number of genes, metabolites, and transcription factors that are highly correlated with it. The larger the icon, the more relevant substances.

the results of the metabolomics analysis. In the MEP pathway, *DXS* and *IspD* genes were highly expressed in 6-year-old cinnamon, which regulates the massive accumulation of 4-(Cytidine 5'-diphospho)-2-C-methyl-D-erythritol (CDP-ME) in 6-year-old *C. cassia*. In addition, the high expression of *HDR* and *GGPS2* in 6-year-old *C. cassia* further promoted the accumulation of downstream GGPP (geranylgeranyl diphosphate) (Figure 6A). In the terpenoid biosynthesis pathway, 35 TFs, 13 DEGs, and 2 differentially expressed metabolites together constituted the diagram of the regulatory network. The metabolite Mevalonate-5PP was significantly negatively correlated with *FDPS1* and *AACT2*, which was corresponds to the results in the terpenoid biosynthesis pathway. The transcription factor *Trihelix5* was significantly positively correlated with gene *AACT2*, indicating that Trihelix5 may regulate the synthesis of metabolite Mevalonate 5PP by affecting gene *AACT2*. In conclusion, terpenoid skeleton synthesis, transcription of structural genes, and TF regulation were significantly related (Figure 6B).

4 Discussion

In *C. cassia*, phenylpropanoids, flavonoids, and terpenoids determine its medicinal value and edible quality. In the different growth and development stages of medicinal plants, transcriptional reprogramming and the redirection of metabolic flux occur in a variety of biosynthetic pathways (Liu et al., 2017). Many studies have found that the growth years can significantly affect the accumulation of effective components in plant (Geng et al., 2011; Li et al., 2013). Similar to the results of the metabolome data, DEGs in *C. cassia* were significantly enriched in phenylpropanoid, flavonoid and terpenoid biosynthetic pathways at different growth years (Supplementary Figure S4), indicating that the change in metabolite accumulation patterns was strictly controlled by DEGs.

Phenylpropanoid biosynthesis starts with the early evolution of freshwater algae to terrestrial plants. At present, phenylpropanoid biosynthesis in terrestrial plants has evolved through a variety of branch pathways. PAL is a key enzyme and rate-limiting enzyme connecting primary metabolism and phenylpropanoid biosynthesis, which catalyzes L-phenylalanine to produce trans-cinnamic acid, lignin, coumarin, cinnamaldehyde, and other metabolites (Jiao et al., 2020). As an intermediate product, trans cinnamic acid can be further converted into lignin, coumarin, cinnamaldehyde, and other metabolites. The content of coumarin was highest in 8-year-old *C. cassia*, which is basically consistent with the expression trend of three *BGLs* (Figure 4A), indicating that coumarin synthesis is under the control of these three *BGLs*. Cinnamaldehyde has antibacterial (Vijayan and Mazumder, 2018), anti-tumor (Koppikar et al., 2010), and other activities. Geng et al. (2011) used GC-MS technology to detect and analyze the content of cinnamaldehyde in cinnamon oil extracted from 5~12-year-old cinnamon and found that the content was the highest in 6-year-old *C. cassia*. Gao et al. (2023) analyzed the differences of genes and metabolites in different *C. cassia* tissues through transcriptome and

10.3389/fpls.2023.1325961

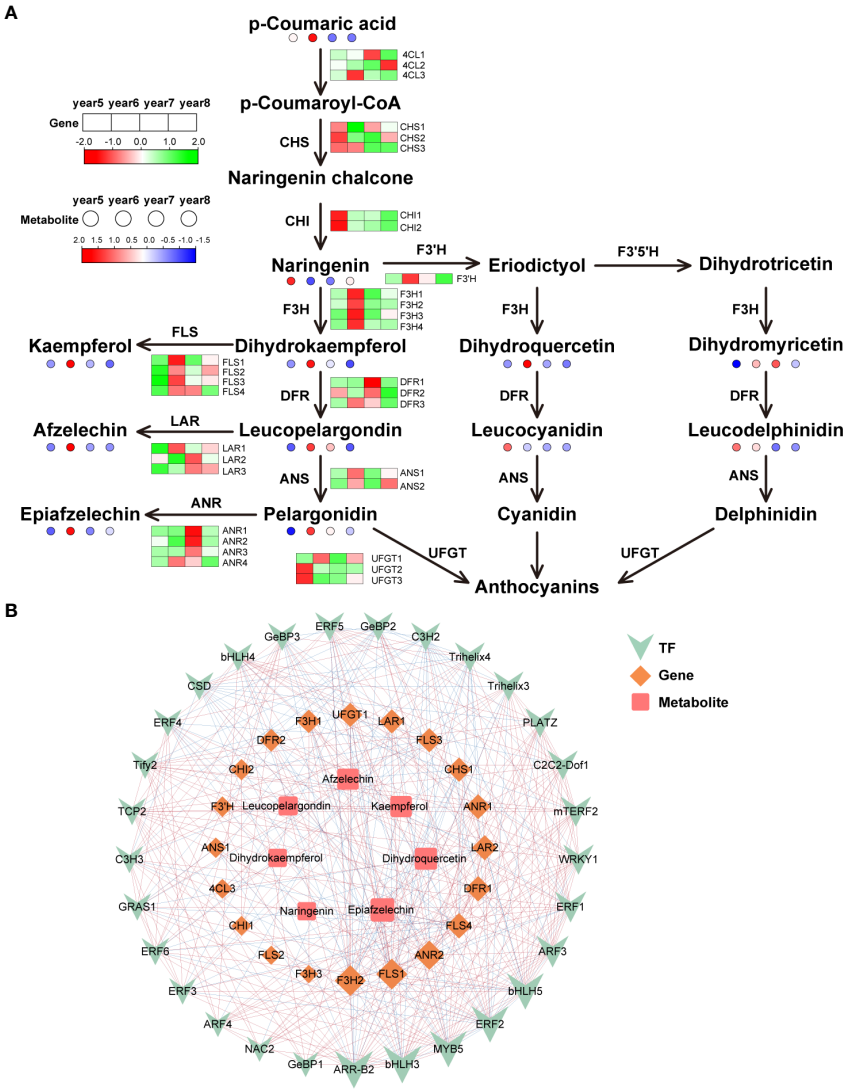


FIGURE 5
Integrative analysis of transcriptome and metabolome of the flavonoids biosynthetic pathway in5~8-year-old *C. cassia*. **(A)** Flavonoids biosynthesis pathway constructed with DAMs and DEGs. **(B)** Correlation network diagram of flavonoids.

metabolomics. They found that cinnamaldehyde content in *C. cassia* bark was higher than that in branches and leaves, and CCR gene content was also higher in *C. cassia* bark, which was corresponded to the results of our study. Therefore, we speculated that this was due to the high expression levels of the *CCR* in 6-year-old *C. cassia* (Figure 4A). In the branching pathway to lignin, the expression of most genes in 6-year-old *C. cassia* was high, and the expression of the *C4H4* gene was the highest, which was consistent with the accumulation of p-coumaric acid. It was inferred that *C4H4* had strong competitiveness for substrates, resulting in the massive production of p-coumaric acid.

The basic structure of flavonoids is C6-C3-C6, and its synthesis pathway is the branch with the most kinds of metabolites in the phenylpropanoid biosynthesis pathway (Stobiecki and Kachlicki, 2006). This pathway is relatively conservative in plant evolution, and the steps of flavonoid synthesis in most plants are the same. According to the RNA-seq map, 32 DEGs related to flavonoid synthesis were identified. *4CL*, *CHS*, *F3H*, *CHI* and *F3'H* regulate the synthesis of early precursors of flavonoids, *DFR*, *ANS*, and *UFGT* regulate anthocyanin synthesis, *FLS* regulates flavonol synthesis, and *LAR* and *ANR* are related to flavanol synthesis. The expression levels of most of these genes was the highest in 6-year-old *C. cassia*, which was consistent with the accumulation of flavonoids and their derivatives (Figure 5A). F3H is the center of the whole flavonoid metabolic pathway, which can catalyze flavanone to generate dihydroflavonol, dihydroquercetin, and dihydromyricetin. These dihydroflavonols are important intermediates in the synthesis of flavonol, flavanol, and anthocyanin (Holton and Cornish, 1995). FLS uses dihydroflavonol as the substrate to form flavonol compounds (Forkmann and Martens, 2001). Xu et al. (2012) cloned the gene *GbFLS* from *Ginkgo biloba L.* into the pET-28a (+). Then, transformed recombinant plasmid into *Escherichia coli* BL21 (DE3). The enzyme activity test results indicated that the recombinant GbFLS protein expressed *in vitro* catalyzes

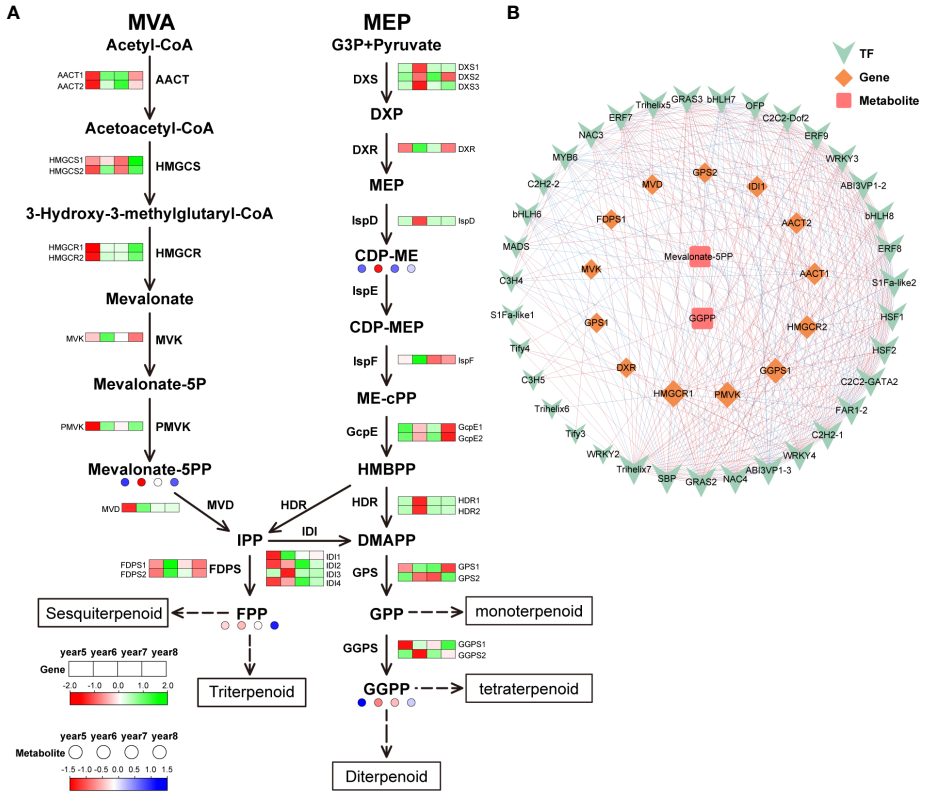


FIGURE 6
Integrative analysis of transcriptome and metabolome of terpenoid biosynthetic pathway in 5~8 years old *C. cassia*. **(A)** Terpenoid biosynthesis pathway constructed with DAMs and DEGs. **(B)** Correlation network diagram of terpenoid.

dihydrokaempferol to generate kaempferol and simultaneously catalyzes naringen to convert kaempferol. This study showed that GbFLS is a multifunctional dioxygenase. In the pathway map, all members of the *F3H* and *FLS* gene families showed the highest expression in 6-year-old *C. cassia* (Figure 5A). It is speculated that the high expression of key structural genes *F3H* and *FLS* in the flavonoid biosynthesis pathway of *C. cassia* led to the mass synthesis of flavonoids.

Terpenoids are important secondary metabolites of *C. cassia*. They are mainly synthesized in two ways: MVA pathway and MEP pathway (Figure 6A). The main difference between the two synthesize pathways is that the synthesis mechanism and final products of the intermediate IPP (isopentenyl pyrophosphate) and the DMAPP (isomer dimethyl allyl pyrophosphate) are different. IPP and DMAPP are common precursors of all terpenoids. The MVA pathway in cytoplasm uses acetyl CoA as a raw material to produce IPP, while the MEP pathway in plastids uses pyruvic acid and glyceraldehyde-3-phosphate as raw materials to form IPP and DMAPP (Vranová et al., 2013). IPP generated by MVA pathways and MEP pathway can pass through the plastid membrane and be used by each other (Zhang et al., 2022). According to the metabolomic data, most terpenoids had the largest accumulation in 5-year-old *C. cassia* (Figure 6A). Therefore, we speculated that the mechanism of IPP formation was different in *C. cassia* with different ages. In the MVA pathway, *MVD*, which regulates IPP synthesis, is highly expressed in 5-year-old *C. cassia*. In the MEP pathway, *HDR* regulating IPP synthesis has the highest expression in 6-year-old *C. cassia*. Through the KEGG pathway annotation results, we screened 29 DEGs related to terpenoid biosynthesis and further analyzed the expression levels of these genes. Most of the structural gene expression patterns on the MVA pathway corresponded to the metabolome data results (Figure 6A), indicating that the biosynthesis of terpenoid compounds in *C. cassia* might be more dependent on the MVA pathway. Schramek et al. (2014) found that the biosynthesis of ginsenoside, a typical terpenoid compound in *Radix Ginseng*, mainly through the MVA pathway by using a $^{13}CO_2$ pulse chase technology. Green *Zanthoxylum armatum* and red *Z. armatum* differ in flavor and aroma because the terpenoids in green *Z. armatum* are synthesized through MVA and MEP, while the terpenoids in red *Z. armatum* are mainly produced through MEP (Fei et al., 2021). In general, terpenoid biosynthesis in plants can depend on a certain pathway, but the MVA pathway and MEP pathway can also compensate for each other to ensure the normal growth of plants.

TFs activate or inhibit the co-expression of multiple genes by specifically binding to the DNA sequence of the regulatory region (Dare et al., 2008). The correlation network between transcriptome and metabolome can be used to clarify functional relationships between genes and metabolites. It can also use to identify key TFs. This study determined the Pearson correlation coefficient of TFs, DEGs, and DAMs related to the synthesis of phenylpropanoids,

flavonoids, and terpenoids and excavated the core regulatory network (Figures 4B, 5B, 6B). A high correlation between specific DEGs, TFs, and metabolites indicates that these structural genes/TFs play an important role in the growth and development of *C. cassia*. The analysis of *C. cassia* transcript libraries in different growth years showed that 1,588 TFs had different expression levels (Supplementary Table S3). During the development of plant, TFs play an important role in regulating the production of effective substances, including positive and negative regulation. *TmMYB3* (Yu et al., 2020), *PpNAC1* (Jin et al., 2022), and *VqWRKY31* (Yin et al., 2022) have been proven to increase substance synthesis by promoting the expression of structural genes. TFs can also be expressed in tissues as repressors to prevent ectopic substances accumulation. Some repressors, such as *PtrMYB57*, can form MBW complexes with other TFs to reduce substance production (Wan et al., 2017). However, some TFs have dual functions, acting as inhibitors and activators (Chen et al., 2021). We identified some TFs highly related to the synthesis of phenylpropanoids, flavonoids, and terpenoids through co-expression network analysis, such as *MYBs*, *ERFs*, *bHLHs*, *NACs*, and *WRKYs* (Figures 4B, 5B, 6B). Previous studies have isolated and identified some TFs that play positive and negative regulatory roles in the of phenylpropanoid and flavonoid in plants. For example, *MYB165* was negatively correlated with various genes in flavonoid and phenylpropanoid biosynthesis pathways in *Populus L.* (Ma et al., 2018). Using yeast hybridization, three *ERF* TF family members have been shown to regulate the synthesis of citrus flavonoids by regulating type IV chalcone isomerase (Zhao et al., 2021). *MsMYB* directly binds to the cis-acting regulatory element of the large subunit of GPP synthetase (*MsGPPS LSU*) and negatively regulates terpenoid biosynthesis (Reddy et al., 2017). The spatiotemporal expression patterns of positive and negative regulators may determine the balance of the accumulation levels of active components in *C. cassia*. This study showed many candidate regulators with active components in *C. cassia*, and the investigators plan to further explore the regulatory mechanisms of these TFs in biosynthesis process of active components.

5 Conclusions

In our study, integrative analysis of metabolome and transcriptome were performed on 5~8-year-old *C. cassia* to understand the dynamic accumulation mechanism of active ingredients. The high levels expression of phenylpropanoid and flavonoid pathway genes in 6-year-old *C. cassia* led to significantly higher content of phenylpropanoids and flavonoids such as cinnamic aldehyde and coumaric acid in 6-year-old than in others. Through co-expression network analysis, genes and TFs were identified that regulate the biosynthesis and regulation of phenylpropanoids and flavonoids, and it was predicted that TFs such as *MYBs*, *bHLHs*, *ERFs*, *NACs*, and *WRKYs* were involved in the regulation of phenylpropanoids and flavonoids. In addition, metabolome analysis showed that the accumulation of terpenoids in 5-year-olds was significantly higher than in others, which was caused by high levels expression upstream genes in the terpenoid synthesis pathway. Together, this study provides new understanding for the accumulation and synthesis of phenylpropanoids, flavonoids, and terpenoids in *C. cassia*, which also lays a solid biological foundation for the breeding of high-quality *C. cassia*.

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

HG: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Writing – original draft, Writing – review & editing. MS: Data curation, Investigation, Visualization, Writing – review & editing. HZ: Investigation, Visualization, Writing – review & editing. HS: Data curation, Visualization, Writing – review & editing. QY: Funding acquisition, Project administration, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was supported by the National Natural Science Foundation (32300316), Basic Research Project of Luoding Cinnamon Industry Development (2018-082) and Youth Innovative Talents Project by Educational Department of Guangdong Province (2019KQNCX057).

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.2023.1325961/full#supplementary-material

References

Chen, W., Zheng, Q., Li, J., Liu, Y., Xu, L., Zhang, Q., et al. (2021). *DkMYB14* is a bifunctional transcription factor that regulates the accumulation of proanthocyanidin in persimmon fruit. *Plant J.* 106, 1708–1727. doi: 10.1111/tpj.15266

Dare, A. P., Schaffer, R. J., Lin-Wang, K., Allan, A. C., and Hellens, R. P. (2008). Identification of a cis-regulatory element by transient analysis of co-ordinately regulated genes. *Plant Methods* 4, 1–10. doi: 10.1186/1746-4811-4-17

Fei, X., Qi, Y., Lei, Y., Wang, S., Hu, H., and Wei, A. (2021). Transcriptome and metabolome dynamics explain aroma differences between green and red prickly ash fruit. *Foods* 10, 391. doi: 10.3390/foods10020391

Forkmann, G., and Martens, S. (2001). Metabolic engineering and applications of flavonoids. *Curr. Opin. Biotechnol.* 12, 155–160. doi: 10.1016/S0958-1669(00)00192-0

Fraser, C. M., and Chapple, C. (2011). The phenylpropanoid pathway in Arabidopsis. *Arabidopsis Book* 9, e0152. doi: 10.1199/tab.0152

Gao, H., Zhang, H., Hu, Y., Xu, D., Zheng, S., Su, S., et al. (2023). *De novo* transcriptome assembly and metabolomic analysis of three tissue types in *Cinnamomum cassia*. *Chin. Herb. Med.* 15, 310–316. doi: 10.1016/j.chmed.2022.06.013

Geng, S., Cui, Z., Huang, X., Chen, Y., Xu, D., and Xiong, P. (2011). Variations in essential oil yield and composition during *Cinnamomum cassia* bark growth. *Ind. Crop Prod.* 33, 248–252. doi: 10.1016/j.indcrop.2010.10.018

Guo, J., Wu, Y., Wang, G., Wang, T., and Cao, F. (2020). Integrated analysis of the transcriptome and metabolome in young and mature leaves of *Ginkgo biloba L. Ind. Crop Prod.* 143, 111906. doi: 10.1016/j.indcrop.2019.111906

Holton, T. A., and Cornish, E. C. (1995). Genetics and biochemistry of anthocyanin biosynthesis. *Plant Cell.* 7, 1071–1083. doi: 10.2307/3870058

Jeyaratnam, N., Nour, A. H., Kanthasamy, R., Nour, A. H., Yuvaraj, A. R., and Akindoyo, J. O. (2016). Essential oil from *Cinnamomum cassia* bark through hydrodistillation and advanced microwave assisted hydrodistillation. *Ind. Crop Prod.* 92, 57–66. doi: 10.1016/j.indcrop.2016.07.049

Jiao, C., Srensen, I., Sun, X., Sun, H., Behar, H., Alseekh, S., et al. (2020). The penium margaritaceum genome: hallmarks of the origins of land plants. *Cell* 181, 1097–1111. doi: 10.1016/j.cell.2020.04.019

Jin, Z., Wang, J., Cao, X., Wei, C., Kuang, J., Chen, K., et al. (2022). Peach fruit *PpNAC1* activates *PpFAD3-1* transcription to provide ω-3 fatty acids for the synthesis of short-chain flavor volatiles. *Hortic. Res.* 4, 9. doi: 10.1093/hr/uhac085

Kang, M. S., and Lee, H. S. (2018). Acaricidal and insecticidal responses of *Cinnamomum cassia* oils and main constituents. *Appl. Biol. Chem.* 61, 653–659. doi: 10.1007/s13765-018-0402-4

Koppikar, S. J., Choudhari, A. S., Suryavanshi, S. A., Kumari, S., Chattopadhyay, S., and Ruchika, K. G. (2010). Aqueous Cinnamon Extract (ACE-c) from the bark of *Cinnamomum cassia* causes apoptosis in human cervical cancer cell line (SiHa) through loss of mitochondrial membrane potential. *BMC Cancer.* 10, 210. doi: 10.1186/1471-2407-10-210

Li, L., Xu, Y., Chen, X., Bao, H., Li, C., Zhang, X., et al. (2022). Weighted gene co-expression network analysis revealed the synthesis of aromatic compounds in Dendrobium catenatum. *Ind. Crop Prod.* 178, 114668. doi: 10.1016/j.indcrop.2022.114668

Li, Y., Kong, D., Huang, R., Liang, H., Xu, C., and Wu, H. (2013). Variations in essential oil yields and compositions of *Cinnamomum cassia* leaves at different developmental stages. *Ind. Crop Prod.* 47, 92–101. doi: 10.1016/j.indcrop.2013.02.031

Liu, G. F., Han, Z. X., Feng, L., Gao, L. P., Gao, M. J., Gruber, M. Y., et al. (2017). Metabolic flux redirection and transcriptomic reprogramming in the albino tea cultivar 'Yu-jin-xiang' with an emphasis on catechin production. *Sci. Rep.* 7, 45062. doi: 10.1038/srep45062

Ma, D., Reichelt, M., Yoshida, K., Gershenzon, J., and Constabel, C. P. (2018). Two R2R3-MYB proteins are broad repressors of flavonoid and phenylpropanoid metabolism in poplar. *Plant J.* 96, 949–965. doi: 10.1111/tpj.14081

Meng, J., Wang, B., He, G., Wang, Y., Tang, X., Wang, S., et al. (2019). Metabolomics integrated with transcriptomics reveals redirection of the phenylpropanoids metabolic flux in ginkgo biloba. *J. Agric. Food Chem.* 67, 3284–3291. doi: 10.1021/acs.jafc.8b06355

Reddy, V. A., Wang, Q., Dhar, N., Kumar, N., Venkatesh, P. N., Rajan, C., et al. (2017). Spearmint R2R3-MYB transcription factor *MsMYB* negatively regulates monoterpene production and suppresses the expression of geranyl diphosphate synthase large subunit (MsGPPS.LSU). *Plant Biotechnol. J.* 15, 1105–1119. doi: 10.1111/pbi.12701

Schramek, N., Huber, C., Schmidt, S., Dvorski, S. E., and Ostrozhenkova, E. (2014). Biosynthesis of ginsenosides in field-grown Panax ginseng. *JSM Biotechnol. BioMed. Eng.* 2, 1033. doi: 10.47739/2333-7117/1033

Shin, W. Y., Shim, D. W., Kim, M. K., Sun, X., Koppula, S., Yu, S. H., et al. (2017). Protective effects of *Cinnamomum cassia* (Lamaceae) against gout and septic responses via attenuation of inflammasome activation in experimental models. *J. Ethnopharmacol.* 205, 173–177. doi: 10.1016/j.jep.2017.03.043

Stobiecki, M., and Kachlicki, P. (2006). "Isolation and identification of flavonoids," in *The Science of Flavonoids*. Ed. E. Grotewold (New York, NY: Springer), 47–69.

Vijayan, V., and Mazumder, A. (2018). *In vitro* inhibition of food borne mutagens induced mutagenicity by cinnamon (*Cinnamomum cassia*) bark extract. *Drug Chem. Toxicol.* 41, 385–393. doi: 10.1080/01480545.2018.1439056

Vranová, E., Coman, D., and Gruissem, W. (2013). Network Analysis of t he MVA and MEP Pathways for Isoprenoid Synthesis. *Annu. Rev. Plant Biol.* 64, 665–700. doi: 10.1146/annurev-arplant-050312-120116

Wan, S., Li, C., Ma, X., and Luo, K. (2017). *PtrMYB57* contributes to the negative regulation of anthocyanin and proanthocyanidin biosynthesis in poplar. *Plant Cell Rep.* 36, 1263–1276. doi: 10.1007/s00299-017-2151-y

Wang, R., Ren, C., Dong, S., Chen, C., Xian, B., Wu, Q., et al. (2021). Integrated metabolomics and transcriptome analysis of flavonoid biosynthesis in safflower (*Carthamus tinctorius L.*) with different colors. *Front. Plant Sci.* 12, 712038. doi: 10.3389/fpls.2021.712038

Xu, F., Li, L., Zhang, W., Cheng, H., Sun, N., Cheng, S., et al. (2012). Isolation, characterization, and function analysis of a flavonol synthase gene from Ginkgo biloba. *Mol. Biol. Rep.* 39, 2285–2296. doi: 10.1007/s11033-011-0978-9

Xue, Q., Fan, H., Yao, F., Cao, X., Liu, M., Sun, J., et al. (2019). Transcriptomics and targeted metabolomics profilings for elucidation of pigmentation in Lonicera japonica flowers at different developmental stages. *Ind. Crop Prod.* 145, 111981. doi: 10.1016/j.indcrop.2019.111981

Yin, W., Wang, X., Liu, H., Wang, Y., Nocker, S., Tu, M., et al. (2022). Overexpression of *VqWRKY31* enhances powdery mildew resistance in grapevine by promoting salicylic acid signaling and specific metabolite synthesis. *Hortic. Res.* 19, 9. doi: 10.1093/hr/uhab064

Yu, C., Luo, X., Zhang, C., Xu, X., Huang, J., Chen, Y., et al. (2020). Tissue-specific study across the stem of Taxus media identifies a phloem-specific TmMYB3 involved in the transcriptional regulation of paclitaxel biosynthesis. *Plant J.* 103, 95–110. doi: 10.1111/tpj.14710

Zhang, Y. Y., Elam, E., Ni, Z. J., Zhang, F., Thakur, K., Wang, S., et al. (2022). LC-MS/MS targeting analysis of terpenoid metabolism in Carya cathayensis at different developmental stages. *Food Chem.* 366, 130583. doi: 10.1016/j.foodchem.2021.130583

Zhang, Y., Liu, Y., Wang, Y., Jiang, P., and Quek, S. Y. (2016). Antibacterial activity and mechanism of cinnamon essential oil against Escherichia coli and Staphylococcus aureus. *Food Control.* 58, 282–289. doi: 10.1016/j.foodcont.2015.05.032

Zhao, C., Liu, X., Gong, Q., Cao, J., Shen, W., Yin, X., et al. (2021). Three AP2/ERF family members modulate flavonoid synthesis by regulating type IV chalcone isomerase in citrus. *Plant Biotechnol. J.* 19, 671–688. doi: 10.1111/pbi.13494

Zhou, Y., Wu, X., Chen, J., and He, J. (2021). Effects of cinnamon essential oil on the physical, mechanical, structural and thermal properties of cassava starch-based edible films. *Int. J. Biol. Macromol.* 184, 574–583. doi: 10.1016/j.ijbiomac.2021.06.067

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 06 March 2024
DOI 10.3389/fpls.2024.1327507

OPEN ACCESS

EDITED BY
Yan Zhao,
Shandong Agricultural University, China

REVIEWED BY
Daisuke Ogawa,
National Agriculture and Food Research Organization (NARO), Japan
Xueqiang Wang,
Zhejiang University, China

*CORRESPONDENCE
Qiang Zhao
zqiang@ncgr.ac.cn
Bin Han
bhan@ncgr.ac.cn

†These authors share first authorship

RECEIVED 25 October 2023
ACCEPTED 19 February 2024
PUBLISHED 06 March 2024

CITATION
Chen R, Lu H, Wang Y, Tian Q, Zhou C, Wang A, Feng Q, Gong S, Zhao Q and Han B (2024) High-throughput UAV-based rice panicle detection and genetic mapping of heading-date-related traits.
Front. Plant Sci. 15:1327507.
doi: 10.3389/fpls.2024.1327507

High-throughput UAV-based rice panicle detection and genetic mapping of heading-date-related traits

Rulei Chen[1,2†], Hengyun Lu[1†], Yongchun Wang[1], Qilin Tian[1], Congcong Zhou[1], Ahong Wang[1], Qi Feng[1], Songfu Gong[3], Qiang Zhao[1*] and Bin Han[1*]

[1]National Center for Gene Research, Key Laboratory of Plant Design/National Key Laboratory of Plant Molecular Genetics, Center for Excellence in Molecular Plant Sciences, Chinese Academy of Sciences, Shanghai, China, [2]University of the Chinese Academy of Sciences, Beijing, China, [3]Center for Excellence in Molecular Plant Sciences, Chinese Academy of Sciences, Shanghai, China

Introduction: Rice (*Oryza sativa*) serves as a vital staple crop that feeds over half the world's population. Optimizing rice breeding for increasing grain yield is critical for global food security. Heading-date-related or Flowering-time-related traits, is a key factor determining yield potential. However, traditional manual phenotyping methods for these traits are time-consuming and labor-intensive.

Method: Here we show that aerial imagery from unmanned aerial vehicles (UAVs), when combined with deep learning-based panicle detection, enables high-throughput phenotyping of heading-date-related traits. We systematically evaluated various state-of-the-art object detectors on rice panicle counting and identified YOLOv8-X as the optimal detector.

Results: Applying YOLOv8-X to UAV time-series images of 294 rice recombinant inbred lines (RILs) allowed accurate quantification of six heading-date-related traits. Utilizing these phenotypes, we identified quantitative trait loci (QTL), including verified loci and novel loci, associated with heading date.

Discussion: Our optimized UAV phenotyping and computer vision pipeline may facilitate scalable molecular identification of heading-date-related genes and guide enhancements in rice yield and adaptation.

KEYWORDS

Oryza sativa, UAV, objective detection, panicle, heading date, QTL

1 Introduction

Oryza sativa is a staple food crop that feeds billions of people worldwide. Optimizing rice yield is critical for global food security, and heading date - the transition from vegetative to reproductive growth - is a key factor determining yield potential. However, traditional manual phenotyping methods for obtaining rice heading-date-related traits are extremely labor-intensive, time-consuming, error-prone, and insufficient for large-scale phenotyping.

Recent advances in computer vision offer transformative potential for fully automatic, high-throughput, and accurate estimation of heading-date-related traits from digital images. Object detection models have proven highly effective for localizing and counting objects in natural images. Leading approaches fall into two main categories: two-stage detectors like Faster R-CNN (Ren et al., 2017) that are accurate but slow, and one-stage detectors such as YOLO (Redmon et al., 2016) that are fast but can struggle with small objects. However, recent advancements in one-stage detectors have narrowed down this accuracy gap, especially in the YOLO family. Newer transformer-based approaches like DETR (Carion et al., 2020) remove hand-designed components like NMS but suffer from convergence issues. Subsequent works have addressed this problem, making the DETR series an attractive model choice overall.

Several studies have already applied these cutting-edge models for analyze rice panicles for traits related to heading date and yield. For instance, Zhou et al. proposed a pipeline using YOLOv5, DeepSORT for tracking identical panicles over time-series images and quantifying the effects of nitrogen on flowering duration and timing (Zhou et al., 2023). The improved Cascade R-CNN is used to detect rice panicles and recognize growth stages from smartphone images under complex field conditions (Tan et al., 2023). The estimated heading dates by counting flowering panicle regions in ground images under an indirectly image classification manner is also performed (Desai et al., 2019). A lightweight model called TinyCCNet for rice panicle segmentation in UAV images is developed, showing potential for agricultural UAVs with limited computing resources (Ramachandran and K.S., 2023). The Res2Net model has been used to classify growth stages and partial least squares regression to estimate heading date from UAV time series images, achieving high accuracy (Lyu et al., 2023). Overall, these studies demonstrate deep learning and computer vision techniques enable accurate, automatic analysis of panicle development from both aerial and ground-based imagery.

However, some obstacles persist in applying off-the-shelf detectors to new specialized domains like panicle counting. Large annotated image datasets are imperative for training high-performing models, but expensive and time-consuming to obtain for niche applications. Different model architectures are often compared only on generic datasets like COCO (Lin et al., 2015), rather than domain-specific tasks like panicle counting. Finally, optimal models for a given application are unclear.

In this paper, we leveraged UAV high-throughput aerial image combined with a semi-automatic annotation workflow to systematically evaluate various state-of-the-art detectors on rice panicle counting. Our comparative analysis identified YOLOv8-X as the top-performing model for our specific application. Subsequently, we utilized YOLOv8-X to extract multiple heading-date-related traits from UAV time-series images with high throughput and accuracy. With these obtained traits, we were able to identify reliable genetic variants using QTL mapping. Some of these variants were consistent with previously published studies, while others facilitated the exploration of novel candidate genes. Our optimized UAV phenotyping and deep learning pipeline helps overcome key limitations, enabling scalable dissection of the genetic basis of rice heading-date-related traits. All relevant code can be accessed at https://github.com/r1cheu/phenocv.

2 Materials and methods

2.1 Rice planting and field image collection

Derived from the crossing of Nipponbare (*Oryza sativa* ssp. *japonica*) and 93–11 (*Oryza sativa* ssp. *indica*), a total of 294 RILs of rice (Huang et al., 2010) were cultivated in Ling Shui, Hainan province at an 18-degree north latitude. The rice was sown in plots measuring 2×1.1m, accommodating 18 plants per plot.

During the rice growth process in 2023, a total of 42 aerial flights were conducted using the DJI Matrice M300 equipped with the ZENMUSE H20 (DJI, Shenzhen, China), which integrated a 20-megapixel zoom camera. Operating at a flight altitude of 18 meters, H20 effectively utilizes its 10x zoom capability to capture clear and detailed imagery of each individual rice panicle within the expansive paddy field.

2.2 Locating plot region

The original images captured by the H20 centered on an individual plot but covered a larger area. Therefore, as a preprocessing step, we extracted the region that only included the central plot from each original image. We first calculated the expected plot width and length based on a known planting density (30cm between plants, 50cm between plots). We used 3800 × 2000 pixels in this work. Next, we binarized the images using OTSU (Otsu, 1979) threshold with the color index of vegetation (CIVE) (Equation 1) (Kataoka et al., 2003). Then, the numbers of white pixels (representing vegetation) per row/column were calculated. The result was smoothed by moving average with a window size of 100. Finally, we defined the row/column that contained the fewest white pixels as the boundary of the plot, since the boundary should contain the minimum number of plant pixels (Equations 2, 3).

The Locating workflow was implemented in Python using the NumPy and OpenCV libraries and is described in Figure 1.

$$0.441 \times R - 0.811 \times G + 0.385 \times B + 18.78754 \quad (1)$$

$$Row\ of\ Plot = min(Row_i + Row_{i+2000}) \quad (2)$$

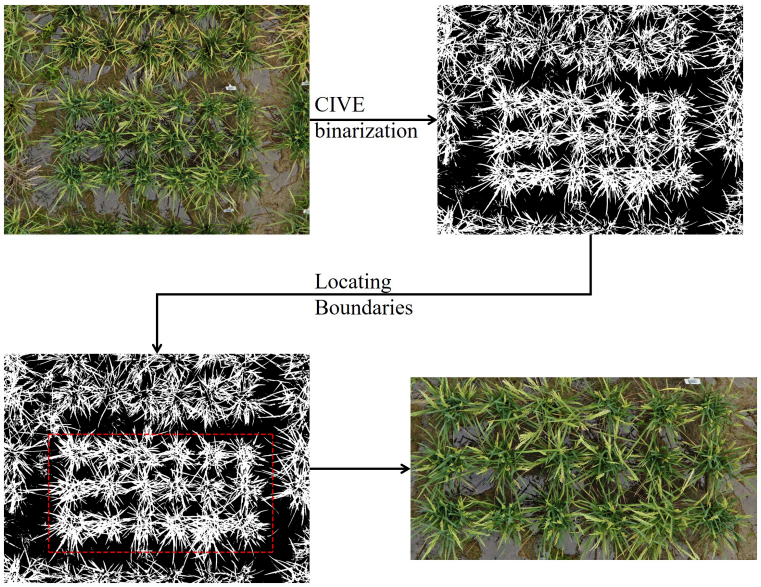


FIGURE 1
Plot extraction workflow. Follow the direction of arrow, the original UAV image (top left) was first binarized using CIVE index and OSTU's thresholding. Next, under a fixed box width of 3800 and height of 2000, the box was moved over the entire image to found the row/column containing the fewest white pixels, thus, locating the boundary. Finally, the plot was cropped from the original image.

$$Column\ of\ Plot = min(Col_i + Col_{i+3800}) \quad (3)$$

Where R, G and B are the pixel values for the corresponding red, green, and blue channels. Row_i denotes the count of white pixels in the i-th row.

2.3 Annotation workflow

In the annotation workflow, to reduce labor costs and accelerate annotations, we utilized the Label Studio interface with the Segment Anything Model (SAM) as the inference backend. SAM can precisely label a panicle using a single-point prompt, thereby allowing for the creation of bounding box around panicle with just one click.

The general annotation workflow is illustrated in Figure 2. Initially, we used a sliding window with the shape of 1000×1000 pixels and a stride of 1000×1000 pixels to divide the 3800×2000 plot images into smaller subimages of 1000 × 1000 pixels. Subsequently, we iterated between model-generated pseudo-labeling, human correction, and model retraining until the dataset was fully labeled or the model's performance met our requirements. This iterative process began with the training of a Faster R-CNN model using approximately 50 labeled images.

In total, we annotated 1852 images and randomly divided them into three datasets with an 8:1:1 ratio. More specifically, we allocated 1530 images for the training set, 161 for the validation set, and another 161 for the test set. Additionally, within the test set, we selected both early-stage and late-stage panicles, creating two subtest sets to ensure a thorough evaluation.

2.4 Prediction workflow

The prediction workflow also commenced from the plot image as depicted in Figure 3. To begin with, each plot image was split into overlapping sub-images with an overlap ratio of 0.25 and window size of 1000 × 1000 pixels. Next, the model detected panicles within

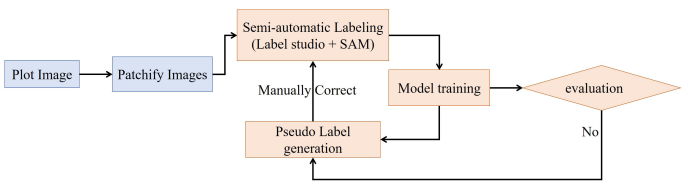


FIGURE 2
Semi-automatic annotation workflow. The workflow begins with plot images which are patchified into smaller sub-images. These patches undergo semi-automatic labeling using Label Studio interfaced with the SAM model for automated suggestions. The labeled sub-images are used to train a model, which is evaluated to determine if performance is sufficient. If not, the model generates pseudolabels on unlabeled data, which re-enters the semi-automatic labeling stage. When the model evaluation is acceptable, the loop breaks and the final model is produced.

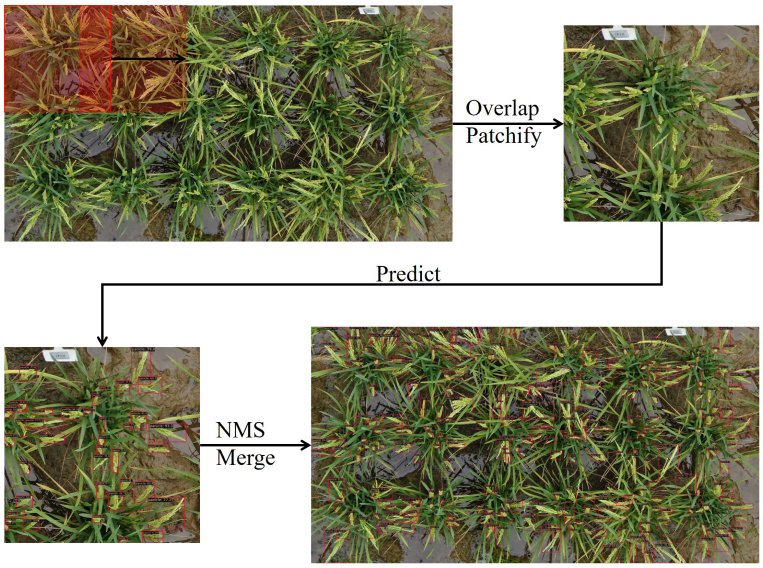


FIGURE 3
Predicting panicle counts from plot images using overlap sliding window approach. Follow the direction of arrow, the plot image was divide into smaller sub-image using a slide window approach. The sub-images were then fed into objective detection model to predict the location of panicles. Subsequently, the predictions from all sub-images were merged using non maximum suppression to remove the redundant prediction.

each sub-image. Lastly, the predictions from the same plot image were merged by employing non-maximum suppression with a threshold of 0.25.

The workflow was implemented in Python using Sahi (Akyon et al., 2022), Pytorch, TorchVision, OpenCV (Bradski, 2000), and NumPy (Harris et al., 2020).

2.5 Model experimental settings

In general, we followed the default training strategies provided by the MMdetection (Chen et al., 2019) and Ultralvtics (Jocher et al., 2023) libraries, which are highly recommended, optimized, and consistently delivered stable performance. The software environments utilized in this paper include Python 3.9, PyTorch 2.0.1, CUDA 11.8, MMdetection v3.10 and Ultralvtics v8.0.158. All the models were trained on 8 NVIDIA A40 GPUs.

2.5.1 Models

We investigated various objective detection models, including Faster R-CNN (Ren et al., 2017), Cascade R-CNN (Cai and Vasconcelos, 2019), YOLO v5 (Jocher, 2020), YOLO v8 (Jocher et al., 2023), RT-DETR (Lv et al., 2023), DINO (Zhang et al., 2023) with different backbones and model sizes, as outlined in Table 1. The implementations of Faster R-CNN, Cascade R-CNN, and DINO utilized the MMdetection library, while the YOLO series and RT-DETR were implemented using the Ultralytics library. All the models were initialized with pretrained weights provided in respective library.

2.5.2 Learning rate scheduling

For Faster R-CNN, Cascade R-CNN with ResNet as backbone, we followed the 2× schedule (He et al., 2019), which entailed fine-tuning for 24 epochs with learning rate drop of 10× at epoch 16 and epoch 22.

However, for Faster R-CNN and Cascade R-CNN with the ConvNext-tiny backbone, we extended the training epoch to 36, and decreased the learning rate at epoch 27 and epoch 33 by a factor of 10×.

As for DINO, it was fine-tuned for 24 epoch, with learning rate decay of 10× at epoch 20.

When it comes to the YOLO series and RT-DETR, we adopted the OneCycle learning rate schedule (Smith and Topin, 2017), which is the default schedule in Ultralytics. We used this schedule for fine-tuning over 100 epochs.

2.5.3 Hyper-parameters

For Faster R-CNN and Cascade R-CNN with ResNet as the backbone, we utilized the SGD optimizer with the following hyperparameters: an initial learning rate of 0.02, 500 steps of linear warm-up, weight decay of 0.0001, and a momentum of 0.9.

For Faster R-CNN with ConvNext-tiny as the backbone, we employed the AdamW optimizer with a learning rate of 0.0001, betas set to (0.9, 0.999), weight decay of 0.05, and a decay rate of 0.95 for layer-wise learning rate decay, with 6 top layers.

For Cascade R-CNN with ConvNext-tiny as the backbone, the learning rate was set to 0.0002, and the decay rate for layer-wise learning rate decay was set to 0.7. Other hyperparameters were consistent with Faster R-CNN using ConvNext-tiny.

As for DINO, we used AdamW with a learning rate of 0.0001 and weight decay of 0.0001, clip gradients with a maximum norm of 0.1 and norm type 2. The learning rate for the backbone was set to 0.00001.

TABLE 1 Performance of detecors on early heading stage, later heading stage, and full test set.

Model	Test				Early		Late	
	$mAP_{50:5:95}$	AP_{50}	R^2	RMSE	R^2	RMSE	R^2	RMSE
Faster RCNN-R50	0.571	0.868	0.907	3.894	0.957	2.687	0.821	4.818
Faster RCNN-R101	0.568	0.865	0.900	4.026	0.952	2.833	0.811	4.950
Faster RCNN-CN-t	0.596	0.887	0.818	5.442	0.921	3.638	0.664	6.797
Cascade RCNN-R50	0.588	0.866	0.880	4.416	0.941	3.152	0.775	5.402
Cascade RCNN-R101	0.588	0.865	0.873	4.545	0.931	3.387	0.769	5.474
Cascade RCNN-CN-t	0.618	0.880	0.805	5.636	0.926	3.507	0.604	7.175
YOLOv5-n	0.613	0.875	0.912	3.794	0.948	2.950	0.845	4.490
YOLOv5-m	0.667	0.895	0.908	3.862	0.966	2.378	0.813	4.930
YOLOv5-x	0.675	0.898	0.906	3.918	**0.966**	**2.401**	0.807	5.006
YOLOv5-n-P6	0.660	0.892	0.920	3.618	0.950	2.889	0.862	4.230
YOLOv5-m-P6	0.673	0.892	0.922	3.574	0.965	2.429	0.848	4.441
YOLOv5-x-P6	0.677	0.899	0.923	3.531	0.962	2.526	0.857	4.316
YOLOv8-n	0.621	0.879	0.918	3.653	0.952	2.846	0.856	4.318
YOLOv8-m	0.666	0.893	0.921	3.590	0.955	2.744	0.859	4.285
YOLOv8-x	0.674	0.897	**0.927**	**3.442**	0.963	2.477	**0.864**	**4.200**
RT-DETR-l	0.630	0.887	-0.389	15.041	0.499	9.156	-1.850	19.245
DINO-R50	0.612	0.885	0.770	6.118	0.910	3.881	0.538	7.751
DINO-Swim-L	**0.677**	**0.914**	0.818	5.545	0.913	3.797	0.655	6.697

Where R50, R101, CN-t, Swim-L, stand for ResNet50 (He et al., 2016), ResNet101, ConvNext-Tiny (Liu et al., 2022), Swim Transformer-Large (Liu et al., 2021). All RCNN models use Feature Pyramid Networks (Ren et al., 2017). P6 represents six stages in the backbone and uses the image size of 1280 × 1280 pixels as inputs, while other YOLO models use the 640 × 640 pixels image as inputs.
Bold indicates that the value is the best metric value in this column.

Regarding the YOLO series and RT-DETR, we utilized the AdamW optimizer with the following hyperparameters: a max learning rate of 0.000714, initial learning rate factors of 0.1, final learning rate factor of 0.0005, weight decay of 0.937, and beta1 of 0.1. The anneal strategy was linear, with 3 warm-up epochs, an initial warm-up momentum of 0.8, and an initial bias learning rate of 0.1.

All the models were trained on 8 GPUs with a mini-batch size of 2 per GPU. During model validation, confidence score thresholds and IoU thresholds for Non-Maximum Suppression (if the model required NMS) were set to 0.05 and 0.5, respectively. For predictions, these thresholds were adjusted to 0.3 and 0.5.

All unmentioned hyperparameters are set to default values in Pytorch.

2.5.4 Data augmentation

To improve model robustness and increase data diversity, we applied various data augmentation techniques, such as vertical and horizontal flipping, HSV color space enhancement, blur, median blur, and CLAHE. For the YOLO series, we also incorporated mosaic and random affine transformations. A detailed configuration is available in Table 2.

2.6 Metrics for evaluation

We employed four metrics to assess count performance, which include the Root Mean Squared Error (RMSE), the Coefficient of Determination (R^2), Mean Average Precision (mAP@50:5:95), and Average Precision at IoU 50 (AP@50). The definitions of RMSE, R^2, mAP@50:5:95, AP@50 are given in Equations 4-10.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$IoU = \frac{area\ \ of\ \ overlap}{area\ \ of\ \ union} \tag{8}$$

TABLE 2 Data augmentation configuration.

Data Aug.	Config(Prob./Frac.)
Horizontal/Vertical Flipping	0.5
HSV-Hue	0.015
HSV-Saturation	0.7
HSV-Value	0.4
RandomAffine	1.0
Mosaic	1 (1-90 epochs), 0(90-100 epochs)
Blur	0.01, limit=(3, 7)
MedianBlur	0.01, limit=(3, 7)
CLAHE	0.01, clip limit=(1, 4), tile grid size=(8, 8)

$$AP@k = \int_0^1 P(R)dR, IoU = k \tag{9}$$

$$mAP@50:5:95 = \frac{1}{9}\sum_{k\in 50,55\ldots 95} AP@k \tag{10}$$

Where n represents the number of test images, y_i denotes the panicle number counted manually, and $\hat{y}_i$ signifies the panicle number derived from the prediction of YOLOv8-X. *TP*, *FP*, and *FN* denote the number of true positives, false positives, and false negatives, respectively. In this study, *TP* refers to bounding boxes that correctly detected rice panicles. *FP* represents bounding boxes that erroneously identified background regions as rice panicles. *FN* signifies ground truth rice panicles that were missed by the detection algorithm.

2.7 Heading-date-related traits extraction

After counting the number of panicles in each plot, we created growth curves represented the panicle count in each plot over time (Figure 4A). These growth curves served as the basis for extracting five static traits and one dynamic trait, as illustrated in Figure 4B. The extraction procedure is described as follows: Firstly, we determined the maximum panicle count. Next, we identified specific developmental stages, which correspond to 10%, 30%, 50%, and 80% of the maximum panicle count. For each of these stages, we used Equation 11 to calculate the date at which each stage was reached. The dynamic trait, the heading stage or heading rate, was defined as the difference between the

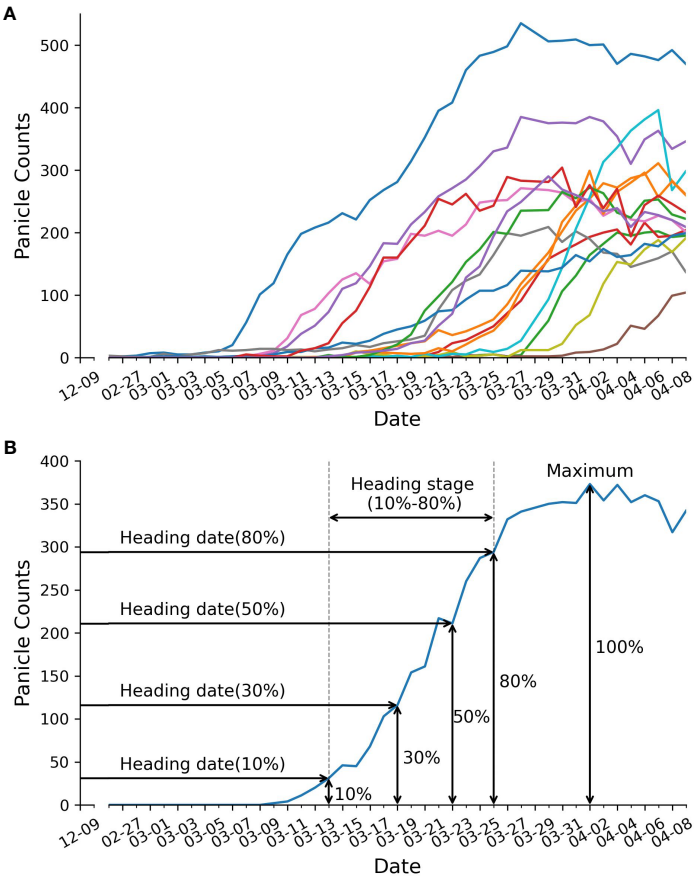


FIGURE 4
Panicle counts over time. **(A)** displays panicle counts over time for 15 RILs. Each colored line represents the panicle counts for a single RIL. **(B)** depicts trait extraction example, including: heading date, duration of heading stage, and maximum panicle counts.

date of reaching 10% of the maximum panicle count and the date of reaching 80% of the maximum panicle count.

$$y \text{ heading date} = \arg\min_{x}(|\text{panicle counts of } x - y \times \text{Maximum}|), \quad y \in \{10\,\%, 30\,\%, 50\,\%, 80\,\%\} \tag{11}$$

Where x denotes the date.

2.8 QTL mapping

The static and dynamic traits were validated through QTL mapping using the UAV-measured heading date-related genetic traits and manually-scored traits collected from RILs. Sequencing and genotyping for the 191 homozygous RILs were conducted using a published pipeline and SEG-MAP (Zhao et al., 2010). Composite interval mapping for QTL analysis was performed using Windows QTL Cartographer version 2.5 (Wang, et al., 2012). The Logarithm of the Odds (LOD) value was calculated to indicate the possibility of QTLs based on likelihood ratio tests.

3 Results

3.1 Collected 2D aerial images

We used the DJI M300 drone, equipped with the H20 camera, to monitor rice experiments from February 26 to April 9, 2023. During this period, we systematically generated 42 series of 2D aerial images for each experimental plot. As a result of all the flight operations, we produced a substantial 160 GB of high-quality 2D imagery.

3.2 Models performance comparison

In order to find the model that best fits panicle detection, we selected several models from three main categories of object detection models.

We trained Faster R-CNN, Cascade R-CNN, YOLOv5, YOLOv8, RT-DETR and DINO with different model sizes and backbones. The performance evaluation was conducted on one main test set and two sub-test sets. These sub-test sets, derived from the main test set, contained early-stage rice panicles and late-stage rice panicles, respectively (refer to Table 1).

Our results indicated that the performance of models aligned with our expectations regarding the Average Precision(AP) metric. Models with more parameters and advanced backbones consistently delivered superior results on this metric. Faster RCNN and Cascade RCNN, which employed ConvNext as their backbone, had higher AP values compared to those using ResNet. Similarly, the AP value of the YOLO series showed an increase as the model size grew. Furthermore, YOLOv5-P6, which employed a larger image resolution as input, performed an additional downsampling, and utilized a higher-level feature map, achieved better performance compared to YOLOv5. The situation in the DETR series mirrored that of the R-CNN and YOLO series, with DINO, which used Swim-L as the backbone, achieving the highest AP value among all models.

The AP metric didn't exhibit a strictly positive correlation with the R^2 and RMSE metrics across various model architectures. This phenomenon was particularly noticeable within the DETR series. For instance, when RT-DETR and DINO-R50 achieved a comparable AP to other models, their R^2 values were significantly lower than those of the YOLO and R-CNN series. DINO-Swim-L, despite attaining the highest AP, only exhibited performance levels on par with the Faster RCNN series in terms of R^2 and RMSE. Surprisingly, RT-DETR-L even yielded a negative R^2 value. After comprehensive consideration of these metrics, our choice for a detector fell on YOLOv8-X. On the test set, early test set, and late test set, its R^2 and RMSE values stood at 0.927, 3.442, 0.963, 2.447, 0.864, and 4.200, respectively. Furthermore, it achieved mAP@50:5:95 and AP@50 values of 0.674 and 0.897 on the test sets.

3.3 Time-series image detection

After a comparative evaluation, we employed the YOLOv8-X model for panicle counting. Following the methodology described in the Methods section, we generated curves illustrating panicle counts over time for 15 out of 294 lines (Figure 4A) and successfully obtained six traits, comprising five static traits and one dynamic trait (Figure 4B). These traits included maximum panicle counts, four heading dates at 10%, 30%, 50% and 80% panicle counts, and the duration of the heading stage (defined as the period between the 80% heading date and the 10% heading date) (Figure 4B). Notably, we were able to capture the dynamic trait of heading stage duration, which was previously unattainable through manual phenotype analysis.

Moreover, we compared the 10% and 30% heading dates with manually recorded heading dates (Figure 5) for validation purposes. The R^2 values for these two developmental stages were 0.9387 and 0.9301, respectively, providing strong support for the validity of our methodology.

3.4 QTL mapping using heading-date-related traits

To assess the biological significance of UAV-evaluated traits in genetic mapping studies, we employed a set of 191 homozygous RILs for genetic linkage analysis. The UAV-based evaluation of heading-daterelated traits was utilized to map QTLs within the population. The genetic distance along the x-axis of 12 chromosomes and the LOD (logarithm of odds) value along the y-axis were used for graphical representation. A threshold value of 3.0 (indicated by the red horizontal line) was employed, and known loci were denoted by red arrows.

Among the traits analyzed, including manual heading date (Figure 6A), UAV-evaluated heading date at 10% panicle counts (Figure 6B), and UAV-evaluated heading date at 30% panicle counts (Figure 6C), we identified three consistent QTLs. Notably, in Figure 6B, the most significant QTL (LOD = 10.26) was located on chromosome 7, approximately 417 kb away from the known gene *Ghd7.1*. This gene, as reported by Yan et al (Yan et al., 2013), plays a crucial role in grain productivity and rice heading.

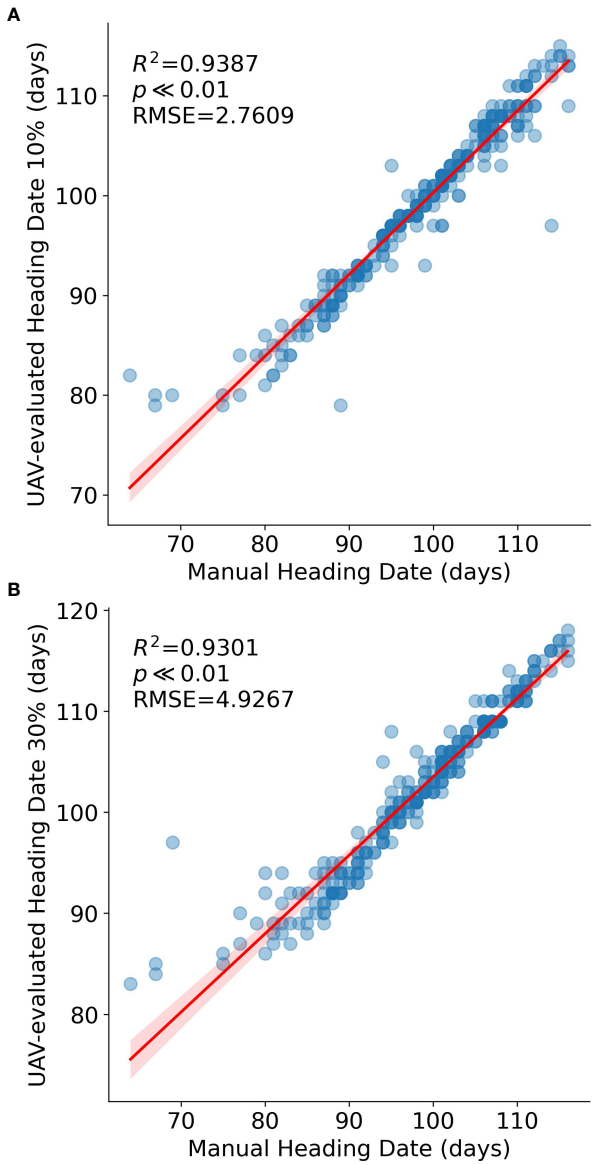


FIGURE 5
Comparison between Manual and UAV-evaluated heading dates (n=294 RILs). Data points represent single RIL measurements. The red line represents the regression line. **(A)** heading date at 10% panicle counts versus Manual heading date, **(B)** heading date at 30% panicle counts versus Manual heading date.

The second highest peak, observed using the heading date (10%) trait, was found on chromosome 3 (LOD = 7.3), approximately 609 kb away from *Hd6* (Ogiso et al., 2010), a gene known to regulate rice flowering and dependent on a functional *Hd1* gene. Furthermore, the third highest peak, identified using the heading date (10%) trait, was situated on chromosome 6 (LOD = 3.84), approximately 263 kb away from *Hd1*, a gene responsible for promoting flowering (Zong et al., 2021). In the trait analysis of UAV-evaluated heading date at 50% panicle count (Figure 6D), we identified two QTLs located on chromosome 3 and 7, as described above. In Figure 6E, we detected a QTL (LOD = 4.58) on chromosome 3, approximately 30 kb away from the *Hd9* gene, which controls rice heading date (Hongxuan et al., 2002).

In addition to static traits, we utilized the dynamic trait, UAV-evaluated heading stage (from 10% panicle counts to 80% panicle counts), to map QTLs, resulting in the identification of two QTLs (Figure 6F). The first QTL was located approximately 140 kb away from *Ghd7* (LOD = 3.99), a gene known to delay heading under long-day conditions while increasing plant height and panicle size (Hu et al., 2020). The second QTL was found approximately 7.1 Mb along chromosome 6 (LOD = 8.24) and was not associated with any known gene. Subsequently, we conducted a QTL mapping using UAV-evaluated panicle count per plant (Figure 6G), we identified a QTL located approximately 706 kb away from the known gene *Ghd7.1*. A comprehensive list of all QTLs identified through QTL mapping is provided in Table 3.

4 Discussion

This study underscores the potential of integrating UAV imagery and object detection models for high throughput, field-based phenotyping of agronomic traits in rice. By harnessing the capabilities of the M300 UAV, equipped with an H20 camera, we are able to swiftly capture images for 294 RILs. This operation, requiring only a single operator, can be completed within a two-hour timeframe. The application of the cutting-edge YOLOv8-X model on UAV-acquired images with a simple image process pipeline, enables the rapid extraction of panicle count data at various developmental timepoints. Additionally, our semi-automatic labeling pipeline reduces the labor cost needed for training a usable object detection model. In summary, our comprehensive approach facilitates cost-effective analysis of six crucial heading-date related traits. Without this approach, a comparable scale of analysis would require a prohibitively extensive investment of time and labor for manual measurements.

Indeed, the application of deep learning to plant phenotyping is becoming increasingly common today. There are several works that focus on panicle detection and heading date estimation using deep learning methods. For instance, in (Zhou et al., 2019), the authors proposed an improved R-FCN for detecting panicles from different stages of rice growth, achieving a precision of 0.868 on their held-out test set. Taking into account the popularity and representativeness of the models, we have not tested the model on our dataset.

Teng integrated several object detection models, such as Faster RCNN and YOLOv5, into a single web platform. These models were used to detect panicles and calculate the panicle number per unit area (PNpM2). They also proposed a tailored YOLOv5 model called Panicle-AI, which has a better AP@.5 of 0.967 than the original YOLOv5 (0.954) on their test set (Teng et al., 2023).In this paper, we not only obtained panicle counts per plant, similar to the panicle number per unit area, but also extracted five additional traits related to heading dates based on time-series images.

Instead of focusing on model modification, some researchers direct their attention to the improvement of NMS, an important part of the objective detection algorithm. This has been proven to perform better in removing redundant bounding boxes under crowded conditions, thereby improving detection accuracy. In our method, we used standard NMS; therefore, there may be an improvement in accuracy when using their method (Wang et al., 2022).

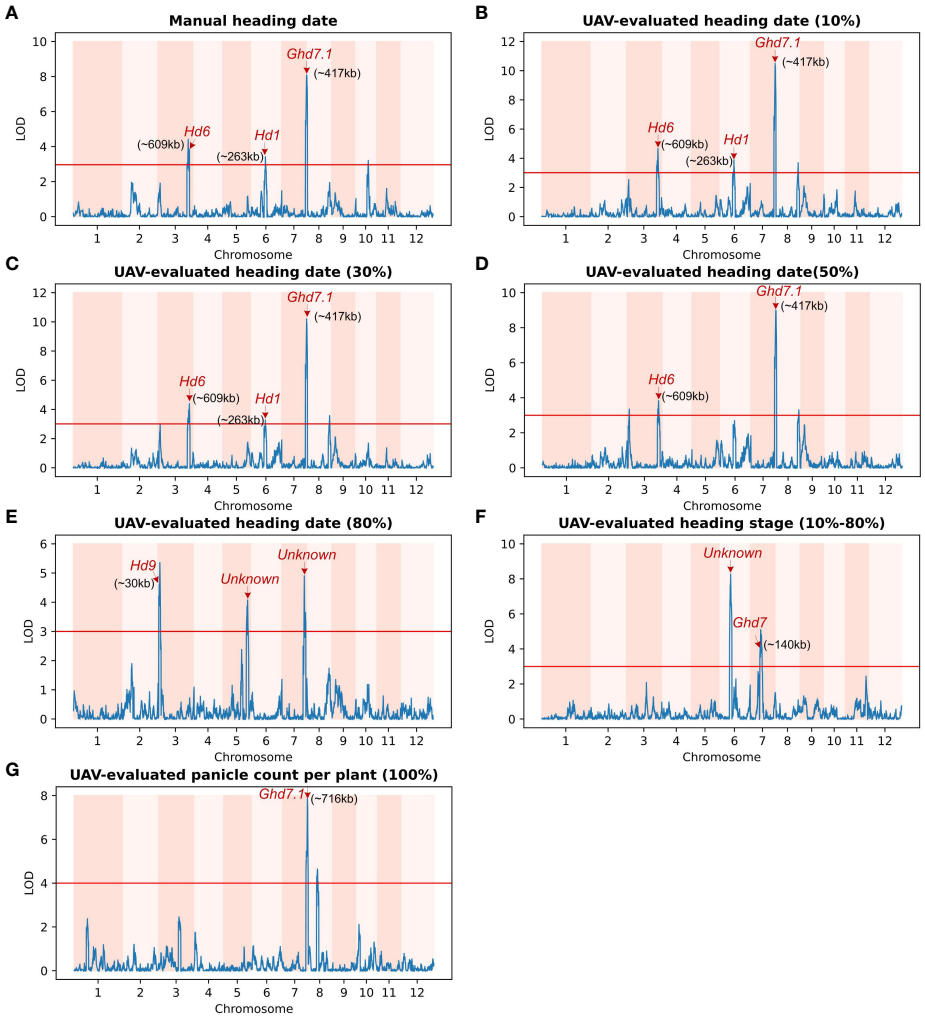


FIGURE 6
Genetic linkage analysis of various UAV-evaluated heading date related traits and manually recording in a population of 191 homozygous recombinant inbred lines (RILs). Red arrows indicate known genes associated with significant single-nucleotide polymorphisms (SNPs).The x-axis represents the genetic distance of the 12 chromosomes, while the y-axis represents the logarithm of the odds (LOD) value. The red horizontal line indicates the significant threshold set at 3.0. **(A)** QTLs identified using Manual heading date. The identified QTLs are close to the *Hd6* gene (chromosome 3), the *Hd1* gene (chromosome 6) and the *Ghd7.1* gene (chromosome 7). **(B)** QTLs identified using UAV-evaluated heading date at 10% panicle counts. **(C)** QTL for UAV-evaluated heading date at 30% panicle counts. Similar to **(A)**, the QTLs identified using UAV-evaluated heading date at 10% and 30% panicle counts are also located in the vicinity of the *Hd6*, *Hd1*, and *Ghd7.1* genes. **(D)** QTL for UAV-evaluated heading date at 50% panicle counts. **(E)** Three loci associated with UAV-evaluated heading date at 80% panicle counts, including one located near *Hd9* (chromosome 3), and another two significant loci on chromosome 5 and 7 that are not associated with any known gene. **(F)** Two QTLs for UAV-evaluated heading stage (date of 80% - date of 10%). The major QTL is not associated with any known gene, while the other is close to the *GHd7* gene. **(G)** QTL for UAV-evaluated panicle counts per plant. The major QTL co-locates with *Ghd7.1* gene.

Another work also focuses on the heading date, but uses a paradigm proposed in 2013 (Girshick et al., 2014), which was no longer used within two years. They concentrate on detecting flowers to estimate the heading date, and their method has not been tested on a large scale population (Desai et al., 2019).

Some other methods do not use object detection, simply employing backbones like ResNet for regression tasks. Guo et al. used a modified DenseNet to directly predict the panicle ratio from images. They achieved an R^2 of 0.992 in their estimation of the heading date. However, their labeling process requires a significant amount of labor to count the number of panicles and the number of tillers via a field survey. Our method requires much less labor, estimating different stages of the heading date based on the panicle number, and further validating through QTL mapping (Guo et al., 2022).

Returning to our result, the high R-squared achieved by the model in panicle counting demonstrates a strong alignment between model predictions and ground truth data. However, it's worth noting that the metric AP, typically employed to assess detection models, exhibits a negative correlation with some models, and it may not comprehensively represent model performance for agricultural tasks such as panicle counting. In the future, adopting metrics like RMSE and R-squared, computed against the ground truth panicle counts for model selection, or devising a tailored loss function that accommodates counting errors, could potentially enhance performance in the panicle counting task (Huang et al., 2016).

TABLE 3 Quantitative trait loci (QTLs) for heading date, heading stage, and panicle count identified in 191 rice RILs using manual and UAV phenotyping.

Traits	Chr	Peak gent. Pos.	IRGSP1.0 (Mb)	LOD	R^2	Add.	Genes
Manual heading date	3	271.51	30.9	3.8	5.8%	2.52	*Hd6*(31.51M)
	6	124.31	9.6	3.4	4.9%	2.47	*Hd1*(9.34M)
	7	202.91	29.2	8.0	12.6%	3.62	*Ghd7.1*(29.62M)
UAV-evaluated heading date(10%)	3	271.51	30.9	4.6	6.4%	2.14	*Hd6*(31.51M)
	6	124.31	9.6	3.8	5.2%	1.94	*Hd1*(9.34M)
	7	202.91	29.2	10.3	15.3%	3.35	*Ghd7.1*(29.62M)
UAV-evaluated heading date(30%)	3	271.51	30.9	4.4	6.0%	2.00	*Hd6*(31.51M)
	6	124.31	9.6	3.2	4.4%	1.74	*Hd1*(9.34M)
	7	202.91	29.2	9.9	14.7%	3.20	*Ghd7.1*(29.62M)
UAV-evaluated heading date(50%)	3	271.51	30.9	3.8	5.5%	1.73	*Hd6*(31.51M) *Ghd7.1*(29.62M)
	7	202.91	29.2	8.9	13.8%	2.78	
UAV-evaluated heading date(80%)	3	8.61	1.3	4.6	7.2%	1.71	*Hd9* (1.27M)
	5	218.01	26.5	4.1	6.8%	1.77	Unknown
	7	182.21	26.2	4.9	7.7%	1.82	Unknown
UAV-evaluated heading stage(10%-80%)	6	97.61	7.1	8.2	13.9%	-1.69	Unknown *Ghd7*(9.15M)
	7	79.61	9.3	4.0	6.5%	1.73	
UAV-evaluated Panicle Count per plant	7	201.31	28.9	7.8	12.1%	-1.85	*Ghd7.1*(29.62M)

Nevertheless, the associations established between the traits extracted from UAV imagery and genetic markers affirm the reliability of our phenotyping methodology. This analysis revealed numerous noteworthy QTLs, encompassing both newly discovered loci and loci corresponding to well-known heading-date genes. Notably, the QTLs identified for the 10% and 30% heading dates coincided with those determined through manual heading date assessment, further validating the effectiveness of our UAV-based phenotyping approach. Particularly, in the later stage (heading date 80%), we unveiled new QTLs. Of significant importance is the successful capture, for the first time, of the dynamic trait—the duration of the heading date, which unveiled previously undiscovered QTLs. These novel QTLs suggest the involvement of additional candidate genes that potentially regulate variations in heading-date-related traits.

To advance this research, ongoing refinement of the detection models is essential to maximize accuracy and generalizability. The semi-automatic annotation workflow introduced in this study has the capacity to streamline the labeling of field images, leading to the creation of more extensive training datasets. This, in turn, holds the promise of progressively boosting model performance in a cost-effective manner. In summary, this study underscores the powerful synergy between UAV and computer vision technologies as a promising framework for expediting genetics research and breeding programs focused on crucial agricultural traits in rice and other crops.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

RC: Writing – original draft, Writing – review & editing, Data curation, Formal Analysis, Methodology. HL: Writing – original draft, Writing – review & editing, Data curation, Formal Analysis, Methodology. YW: Writing – review & editing, Data curation. QT: Writing – review & editing. CZ: Writing – review & editing. AW: Writing – review & editing. QF: Writing – review & editing. SG: Writing – review & editing. QZ: Writing – review & editing, Conceptualization, Funding acquisition, Supervision. BH: Writing – review & editing, Conceptualization, Supervision.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by grants from the Strategic Priority Research Program of the Chinese Academy of Sciences (Precision Seed Design and Breeding, XDA24020205); The National Key Research and Development Program of China (2020YFE0202300); National Natural Science Foundation of China (Grant No. 31871268).

Acknowledgments

The authors would like to thank all members of National Center for Gene Research (NCGR, China) for their assistance during laboratory works and for fruitful discussions. In particular, the authors would like to thank Zhou Ji at the Nanjing Agricultural University (NAU, China) for valuable discussion.

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

Akyon, F. C., Altinuc, S. O., and Temizel, A. (2022). "Slicing aided hyper inference and fine-tuning for small object detection," in *2022 IEEE International Conference on Image Processing (ICIP)*. (New York City, United States: Institute for Electrical and Electronics Engineers (IEEE)),966–970. doi: 10.1109/ICIP46576.2022.9897990

Bradski, G. (2000). The openCV library. *Dr. Dobb's J. Softw. Tools*.

Cai, Z., and Vasconcelos, N. (2019). "Cascade R-CNN: High quality object detection and instance segmentation," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*. (New York City, United States: Institute for Electrical and Electronics Engineers (IEEE)) 43 (5), 1483–1498. doi: 10.1109/TPAMI.2019.2956516

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). "End-to-end object detection with transformers," in *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I* (Springer-Verlag, Berlin, Heidelberg), 213–229. doi: 10.1007/978-3-030-58452-813

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., et al. (2019). MMDetection: Open mmlab detection toolbox and benchmark. *arXiv*. doi: 10.48550/arXiv.1906.07155

Desai, S. V., Balasubramanian, V. N., Fukatsu, T., Ninomiya, S., and Guo, W. (2019). enAutomatic estimation of heading date of paddy rice using deep learning. *Plant Methods* 15, 76. doi: 10.1186/s13007-019-0457-1

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. (New York City, United States: Institute for Electrical and Electronics Engineers (IEEE)), 580–587. doi: 10.1109/CVPR.2014.81

Guo, Z., Yang, C., Yang, W., Chen, G., Jiang, Z., Wang, B., et al. (2022). Panicle Ratio Network: streamlining rice panicle measurement by deep learning with ultra-high-definition aerial images in the field. *J. Exp. Bot.* 73, 6575–6588. doi: 10.1093/jxb/erac294

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature* 585, 357–362. doi: 10.1038/s41586-020-2649-2

He, K., Girshick, R., and Dollar, P. (2019). "Rethinking imagenet pre-training," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE Computer Society, Los Alamitos, CA, USA), 4917–4926. doi: 10.1109/ICCV.2019.00502

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (New York City, United States: Institute for Electrical and Electronics Engineers (IEEE)), 770–778. doi: 10.1109/CVPR.2016.90

Hongxuan, L., Motoyuki, A., Utako, Y., Takuji, S., and Masahiro, Y. (2002). Identification and characterization of a quantitative trait locus, hd9, controlling heading date in rice. *Breed. Sci.* 52, 35–41. doi: 10.1270/jsbbs.52.35

Hu, Y., Song, S., Weng, X., You, A., and Xing, Y. (2020). The heading-date gene ghd7 inhibits seed germination by modulating the balance between abscisic acid and gibberellins. *Crop J.* 9 (2), 297–304. doi: 10.1016/j.cj.2020.09.004

Huang, G., Liu, Z., and Weinberger, K. Q. (2016). "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (New York City, United States: Institute for Electrical and Electronics Engineers (IEEE)), 2261–2269. doi: 10.1109/CVPR.2017.243

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., et al. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* 42, 961–967. doi: 10.1038/ng.695

Jocher, G. (2020). *Ultralytics yolov5*. doi: 10.5281/zenodo.3908559

Jocher, G., Chaurasia, A., and Qiu, J. (2023). *Ultralytics yolov8*. Available at: https://github.com/ultralytics/ultralytics.

Kataoka, T., Kaneko, T., Okamoto, H., and Hata, S. (2003). "Crop growth estimation system using machine vision," in *Proceedings 2003 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*. (New York City, United States: Institute for Electrical and Electronics Engineers (IEEE)) 2, b1079–b1083. doi: 10.1109/AIM.2003.1225492

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., et al. (2015). *Microsoft coco: Common objects in context*. Available at: https://cocodataset.org/.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. (New York City, United States: Institute for Electrical and Electronics Engineers (IEEE)).

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. (New York City, United States: Institute for Electrical and Electronics Engineers (IEEE)).

Lv, W., Zhao, Y., Xu, S., Wei, J., Wang, G., Cui, C., et al. (2023). Detrs beat yolos on real-time object detection. *ArXiv [Preprint]*. doi: 10.48550/arXiv.2304.08069

Lyu, M., Lu, X., Shen, Y., Tan, Y., Wan, L., Shu, Q., et al. (2023). enUAV time-series imagery with novel machine learning to estimate heading dates of rice accessions for breeding. *Agric. For. Meteorol.* 341, 109646. doi: 10.1016/j.agrformet.2023.109646

Ogiso, E., Takahashi, Y., Sasaki, T., Yano, M., and Izawa, T. (2010). The role of casein kinase ii in flowering time regulation has diversified during evolution. *Plant Physiol.* 152, 808–820. doi: 10.1104/pp.109.148908

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Trans. Sys. Man Cybernetics* 9, 62–66. doi: 10.1109/TSMC.1979.4310076

Ramachandran, A., and K.S., S. K. (2023). enTiny Criss-Cross Network for segmenting paddy panicles using aerial images. *Comput. Electrical Eng.* 108, 108728. doi: 10.1016/j.compeleceng.2023.108728

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE Computer Society, Los Alamitos, CA, USA), 779–788. doi: 10.1109/CVPR.2016.91

Ren, S., He, K., Girshick, R., and Sun, J. (2017). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39 (6), 1137–1149. doi: 10.1109/TPAMI.2016.2577031

Smith, L. N., and Topin, N. (2017). Super-convergence: Very fast training of residual networks using large learning rates. *CoRR*. doi: 10.48550/arXiv.1708.07120

Tan, S., Lu, H., Yu, J., Lan, M., Hu, X., Zheng, H., et al. (2023). enIn-field rice panicles detection and growth stages recognition based on RiceRes2Net. *Comput. Electron. Agric.* 206, 107704. doi: 10.1016/j.compag.2023.107704

Teng, Z., Chen, J., Wang, J., Wu, S., Chen, R., Lin, Y., et al. (2023). Panicle-cloud: An open and ai-powered cloud computing platform for quantifying rice panicles from drone-collected imagery to enable the classification of yield production in rice. *Plant Phenomics* 5, 105. doi: 10.34133/plantphenomics.0105

Wang, S., Basten, C. J., and Zeng, Z. -B. (2012). *Windows QTL cartographer 2.5. Department of Statistics*. Raleigh, NC: North Carolina State University. Available at: http://statgen.ncsu.edu/qtlcart/WQTLCart.htm.

Wang, X., Yang, W., Lv, Q., Huang, C., Liang, X., Chen, G., et al. (2022). Field rice panicle detection and counting based on deep learning. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.966495

Yan, W., Liu, H., Zhou, X., Li, Q., Zhang, J., Lu, L., et al. (2013). Natural variation in ghd7. 1 plays an important role in grain yield and adaptation in rice. *Cell Res.* 23, 969–971. doi: 10.1038/cr.2013.43

Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., et al. (2023). DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In. *Eleventh Int. Conf. Learn. Representations*. doi: 10.48550/arXiv.2203.03605

Zhao, Q., Huang, X., Lin, Z., and Han, B. (2010). Seg-map: a novel software for genotype calling and genetic map construction from next-generation sequencing. *Rice* 3, 98–102. doi: 10.1007/s12284-010-9051-x

Zhou, C., Ye, H., Hu, J., Shi, X., Hua, S., Yue, J., et al. (2019). Automated counting of rice panicle by applying deep learning model to images from unmanned aerial vehicle platform. *Sensors* 19. doi: 10.3390/s19143106

Zhou, Q., Guo, W., Chen, N., Wang, Z., Li, G., Ding, Y., et al. (2023). enAnalyzing nitrogen effects on rice panicle development by panicle detection and time-series tracking. *Plant Phenomics* 5, 0048. doi: 10.34133/plantphenomics.0048

Zong, W., Ren, D., Huang, M., Sun, K., Feng, J., Zhao, J., et al. (2021). Strong photoperiod sensitivity is controlled by cooperation and competition among hd1, ghd7 and dth8 in rice heading. *New Phytol.* 229, 1635–1649. doi: 10.1111/nph.16946

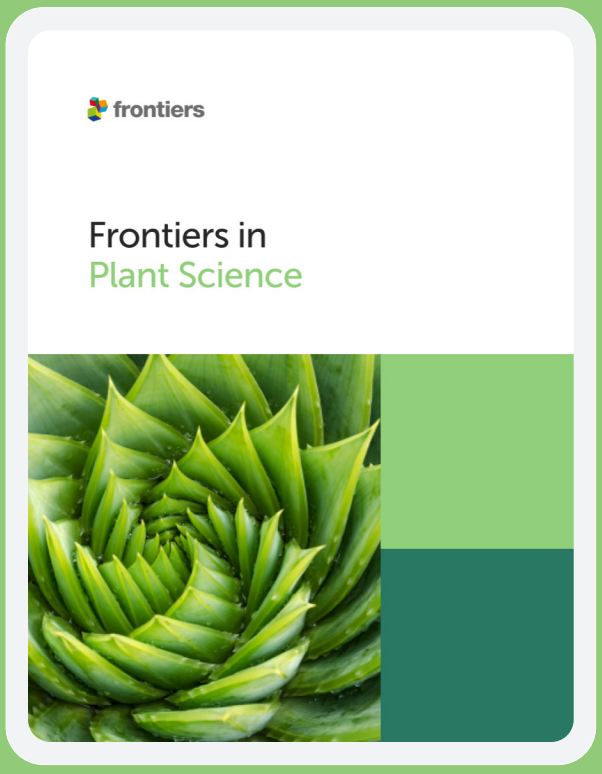
Frontiers in
Plant Science

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