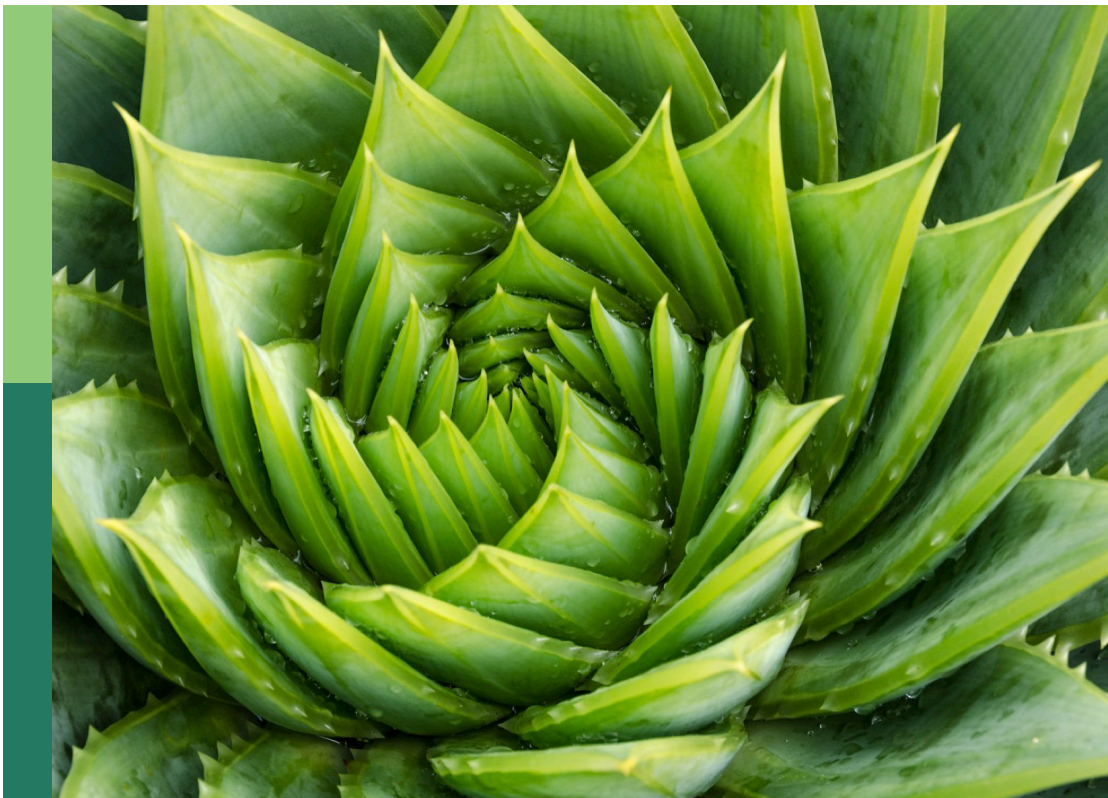
frontiers | Research Topics

Advances in crop biomass production based on multi-omics approach

Edited by
Yin Li, Ramin Yadegari, Jianping Wang, Xingtan Zhang, Shouchuang Wang and Weizhen Liu

Published in
Frontiers in Plant Science



ISSN 1664-8714
ISBN 978-2-8325-2335-3
DOI 10.3389/978-2-8325-2335-3

About Frontiers

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

Frontiers journal series

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

Dedication to quality

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

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Advances in crop biomass production based on multi-omics approach

Topic editors

Yin Li — Huazhong University of Science and Technology, China
Ramin Yadegari — University of Arizona, United States
Jianping Wang — University of Florida, United States
Xingtan Zhang — Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, China
Shouchuang Wang — Hainan University, China
Weizhen Liu — Wuhan University of Technology, China

Citation

Li, Y., Yadegari, R., Wang, J., Zhang, X., Wang, S., Liu, W., eds. (2023). *Advances in crop biomass production based on multi-omics approach*.
Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-2335-3

Table of contents

frontiers | Frontiers in Plant Science

TYPE Editorial
PUBLISHED 19 April 2023
DOI 10.3389/fpls.2023.1155442

OPEN ACCESS

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

*CORRESPONDENCE
Yin Li
yinli2021@hust.edu.cn

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

RECEIVED 31 January 2023
ACCEPTED 11 April 2023
PUBLISHED 19 April 2023

CITATION
Li Y, Liu W, Zhang X, Wang S, Yadegari R and Wang J (2023) Editorial: Advances in crop biomass production based on multi-omics approach. *Front. Plant Sci.* 14:1155442. doi: 10.3389/fpls.2023.1155442

Editorial: Advances in crop biomass production based on multi-omics approach

Yin Li[1]*, Weizhen Liu[2], Xingtan Zhang[3], Shouchuang Wang[4], Ramin Yadegari[5] and Jianping Wang[6]

[1]The Genetic Engineering International Cooperation Base of Chinese Ministry of Science and Technology, The Key Laboratory of Molecular Biophysics of Chinese Ministry of Education, College of Life Science and Technology, Huazhong University of Science & Technology, Wuhan, China, [2]School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, China, [3]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, [4]Hainan Yazhou Bay Seed Laboratory, Sanya Nanfan Research Institute of Hainan University, Sanya, China, [5]School of Plant Sciences, University of Arizona, Tucson, AZ, United States, [6]Agronomy Department, University of Florida, Gainesville, FL, United States

KEYWORDS

genomics, multi-omics, biomass production, biomass related traits, bioinformatics, biomass and bio energy crops

Editorial on the Research Topic
Advances in crop biomass production based on multi-omics approach

Introduction

While as the dominant source of energy during the past century, the detrimental impacts of fossil fuels have become apparent in environmental pollution, unsustainability, and global warming (Sharif et al., 2021). With increasing efforts and capitalization on renewable energy technologies, bioenergy has become one important type of renewable energy. Biomass of plants is an important feedstock of bioenergy production. Plants suitable for biomass production share common characteristics: high yield (of dry matter or a type of biomass, *i.e.*, starch or sugar), low agronomic inputs, and low nutrition requirements. Based on these features, woody species (*e.g.*, willow and poplar), grasses (*e.g.*, sugarcane, switchgrass, and *Miscanthus*), aquatic plants (*e.g.*, algae and duckweed), and oil plants have been considered biomass plants. Additionally, wheat and rice straw are important biomass sources. Biomass has several types according to the source species, the moisture content, and composition of biomass material, such as lignocellulosic biomass from woody plants, biomass from grasses (including cellulosic biomass from grasses or extracted starch/sugar), aquatic plant biomass, and manures (McKendry, 2002). In turn, these biomass types are compatible with different bio-conversion methods, *e.g.*, combustion, fermentation, gasification, pyrolysis, and mechanical extraction of starch or oils. Recently, numerous efforts have been made to convert biomass to high-value chemicals and bio-based materials (Anchan and Dutta, 2021).

Downstream utilizations of biomass (*e.g.*, conversion to biofuels or bio-based chemicals) requires multiple disciplines, such as agricultural science, microbiology, and chemistry. By

contrast, upstream knowledge of biomass, such as the genetic determinants of biomass-related traits and molecular mechanisms of biomass accumulation and composition, relies on plant biology, and agricultural science. Notably, many biomass plants with large and complex genomes (such as sugarcane) have been less studied or have bottlenecks in transformation and traditional genetics (Zhang et al., 2018; Wang et al., 2021; An et al., 2021; Chen et al., 2022). Recently, research on biomass and bioenergy plants has been advanced rapidly due to the development of genomics. For example, state-of-the-art genomic technologies facilitated the successful assembly of reference genomes for sugarcane, *Miscanthus*, and switchgrass (Zhang et al., 2018; Mitros et al., 2020; Lovell et al., 2021). Though huge diversity within and among biomass crops provides invaluable resources for biomass utilization, understanding of biomass production mechanisms is still limited due to shortage of molecular and omic resources and challenges of functional studies. It has become apparent that synergistic integration of multiple omic technologies (*e.g.*, transcriptomics, proteomics, epigenomics, metabolomics, and phenomics) serves as a key approach to circumvent the challenges. This Research Topic includes seven research articles and two reviews, covering several biomass species, including maize, sorghum, sugarcane, rice, and oil plants to reveal the current advances of multi-omics in addressing the mechanisms of biomass production.

Advances in multi-omic technologies and resources facilitate studies on biomass-related traits

This section showcases how omic technologies and resources can facilitate biomass studies. Voelker et al. reported the genome assemblies of 10 sorghum accessions including sweet and non-sweet sorghum genotypes (Boatwright et al.; Kumar et al., 2022). A large number of structural variations (SVs) were identified, which highlighted the SV-related functional difference between sweet and non-sweet sorghum genotypes. Wang et al. developed an image-based phenotypic acquisition method to characterize leaf-sheath traits in detail and applied the method to genome-wide association studies (GWAS), providing a detailed genetic architecture of leaf-sheath morphology. Guo et al. presented an integrative genomic database for oil plants, the Genomic Information Repository for Oil Plants (GROP, www.grop.site), which hosts 22 reference genomes of 18 species with 46 transcriptome datasets (Bayer et al., 2017; Unver et al., 2017; Wang et al., 2018; Song et al., 2020; Sturtevant et al., 2020; Chen et al., 2021). The construction of such an omics repository addresses the need to integrate, share, and analyze the omics data across oil plants for the research community. In addition, Tu et al. reviewed the major applications of regular short-read RNA-seq in plant biology, described a cohort of representative RNA-seq-analysis tools in model plants and major crops, and emphasized that the full utilization of fruitful RNA-seq resources will promote the omic research on under studied species (including biomass crops) to a high level.

Applications of omic approaches provide insights into biomass-related biology

This section collects representative papers using omic technologies to gain insights into biomass-related biological questions. Sugarcane is one of the key biomass and bioenergy crops, providing about 80% of global sugar production and 40% of ethanol production (Zhang et al., 2018). Efforts have been made to investigate the molecular mechanisms of sugar accumulation in sugarcane and in the comparable species sweet sorghum (Li et al., 2018; Li et al., 2019a; Li et al., 2019b), from sugar transportation and physiology to transcriptome and quantitative trait loci mapping (Babu et al., 2009; Liu et al.; Moore, 2005; Aitken et al., 2006; Casu et al., 2007; Zhang et al., 2021). Yuan et al. performed transcriptomic and metabolomic studies on two sugarcane varieties and revealed candidate genes for sucrose metabolism, stem texture, and rind color. While the genes associated with stem sugar accumulation have been identified in sugarcane (Casu et al., 2007; Zhang et al., 2021), epigenetic regulation remains elusive. Xue et al. profiled the DNA methylation in sugarcane (*Saccharum officinarum*) leaves, roots, rinds, and piths, and observed DNA methylation valleys (DMVs) overlapped with transcription factors and sucrose-related genes, indicating the involvement of epigenetic regulation in sucrose metabolism. Liu et al. revealed the link of OsPRR37, a key component of the rice circadian clock, with biomass production through DNA methylation analysis. Overexpression of OsPRR37 in rice led to suppressed growth and lowered biomass likely through the diurnal changes of DNA methylation regulators (such as ROS1A/DNG702) to hypo-methylate a key signal component controlling metabolism, OsHXK1 (Zheng et al., 2021; Zhou et al., 2021). Ain et al. presented a comprehensive review on recent progress in the identification of molecular and genetic factors regulating growth, biomass accumulation, and assimilate partitioning in bioenergy crops. The review highlights a plethora of genes related to cell cycle, cell wall, hormones, and related transcription factors as the targets to improve photosynthesis, carbohydrate allocation, and biomass production in the bioenergy crops. Additionally, this topic also hosts an example of omics-enabled trait association study. Specifically, Wang et al. used comparative RNA-seq to profile seed-specific long-lived mRNA and identify a number of the long-lived mRNA associated with rice seed longevity.

Concluding remarks

This Research Topic exemplifies that multi-omics represent an important route to strengthen the studies of biomass crops, particularly with complex genomes. Importantly, trends emerged from these articles that a combination of multiple omic resources and tools is a powerful approach to gaining new insights into biomass production and related traits. The discoveries will pave the road toward molecular design and breeding biomass crops with tailored bioenergy purposes.

Author contributions

YL, WL, XZ, SW, RY, and JW drafted and revised this editorial based on this Research Topic's contributions. All authors approved the submitted version.

Funding

YL was funded by the National Natural Science Foundation of China (32272126), the Fundamental Research Funds for Central Universities, HUST (2021XXJS070, 3004170157), and Wuhan Knowledge Innovation Project (2022020801010073). WL was funded by the National Natural Science Foundation of China (32200331) and the Major Science and Technology Research Project of Hubei Province (2021AFB002). The project was supported by USDA Research Capacity Fund (Hatch), FLA-AGR-006269 to JW.

Acknowledgments

We thank all authors who submitted their work for this Research Topic, the support of professional editorial staff at Frontiers, and the invaluable time and efforts of reviewers in manuscript evaluation.

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

Aitken, K., Jackson, P., and Mcintyre, C. (2006). Quantitative trait loci identified for sugar related traits in a sugarcane (*Saccharum* spp.) cultivar *Saccharum officinarum* population. *Theor. Appl. Genet.* 112, 1306–1317. doi: 10.1007/s00122-006-0233-2

An, Y., Liu, Y., Liu, Y., Lu, M., Kang, X., Mansfield, S. D., et al. (2021). Opportunities and barriers for biofuel and bioenergy production from poplar. *GCB Bioenergy.* 13, 905–913. doi: 10.1111/gcbb.12829

Anchan, H. N., and Dutta, S. (2021). Recent advances in the production and value addition of selected hydrophobic analogs of biomass-derived 5-(hydroxymethyl) furfural. *Biomass Conv. Bioref.* 13, 2571–2593. doi: 10.1007/s13399-021-01315-1

Babu, C., Koodalingam, K., Natarajan, U., Shanthi, R., and Govindaraj, P. (2009). Assessment of rind hardness in sugarcane (*Sachharum* spp. hybrids) genotypes for development of non lodging erect canes. *Adv. Biol. Res.* 3, 48–52. Available at: http://www.idosi.org/abr/3(1-2)/10.pdf.

Bayer, P. E., Hurgobin, B., Golicz, A. A., Chan, C. K. K., Yuan, Y., Lee, H. T., et al. (2017). Assembly and comparison of two closely related *Brassica napus* genomes. *Plant Biotechnol. J.* 15, 1602–1610. doi: 10.1111/pbi.12742

Casu, R. E., Jarmey, J. M., Bonnett, G. D., and Manners, J. M. (2007). Identification of transcripts associated with cell wall metabolism and development in the stem of sugarcane by affymetrix GeneChip sugarcane genome array expression profiling. *Funct. Integr. Genomics* 7, 153–167. doi: 10.1007/s10142-006-0038-z

Chen, Z., Debernardi, J. M., Dubcovsky, J., and Gallavotti, A. (2022). Recent advances in crop transformation technologies. *Nat. Plants* 8, 1343–1351. doi: 10.1038/s41477-022-01295-8

Chen, X., Tong, C., Zhang, X., Song, A., Hu, M., Dong, W., et al. (2021). A high-quality *Brassica napus* genome reveals expansion of transposable elements, subgenome evolution and disease resistance. *Plant Biotechnol. J.* 19, 615–630. doi: 10.1111/pbi.13493

Kumar, N., Brenton, Z., Myers, M. T., Boyles, R. E., Sapkota, S., Boatwright, J. L., et al. (2022). Registration of the sorghum carbon-partitioning nested association mapping (CP-NAM) population. *J. Plant Regist.* 16, 656–663. doi: 10.1002/plr2.20229

Li, Y., Mehta, R., and Messing, J. (2018). A new high-throughput assay for determining soluble sugar in sorghum internode-extracted juice. *Planta* 248, 785–793. doi: 10.1007/s00425-018-2932-8

Li, Y., Tu, M., Feng, Y., Wang, W., and Messing, J. (2019b). Common metabolic networks contribute to carbon sink strength of sorghum internodes: Implications for bioenergy improvement. *Biotechnol. Biofuels.* 12, 274. doi: 10.1186/s13068-019-1612-7

Li, Y., Wang, W., Feng, Y., Tu, M., Wittich, P. E., Bate, N. J., et al. (2019a). Transcriptome and metabolome reveal distinct carbon allocation patterns during internode sugar accumulation in different sorghum genotypes. *Plant Biotechnol. J.* 17, 472–487. doi: 10.1111/pbi.12991

Lovell, J. T., MacQueen, A. H., Mamidi, S., Bonnette, J., Jenkins, J., Napier, J. D., et al. (2021). Genomic mechanisms of climate adaptation in polyploid bioenergy switchgrass. *Nature* 590, 438–444. doi: 10.1038/s41586-020-03127-1

McKendry, P. (2002). Energy production from biomass (part 1): Overview of biomass. *Bioresour. Technol.* 83, 37–46. doi: 10.1016/s0960-8524(01)00118-3

Mitros, T., Session, A. M., James, B. T., Wu, G. A., Belaffif, M. B., Clark, L. V., et al. (2020). Genome biology of the paleotetraploid perennial biomass crop *Miscanthus*. *Nat. Commun.* 11, 5442. doi: 10.1038/s41467-020-18923-6

Moore, P. H. (2005). Integration of sucrose accumulation processes across hierarchical scales: Towards developing an understanding of the gene-to-crop continuum. *Field Crops Res.* 92, 119–135. doi: 10.1016/j.fcr.2005.01.031

Sharif, A., Bhattacharya, M., Afshan, S., and Shahbaz, M. (2021). Disaggregated renewable energy sources in mitigating CO_2 emissions: new evidence from the USA using quantile regressions. *Environ. Sci. Pollut. Control Ser.* 3, 23–36. doi: 10.1007/s11356-021-13829-2

Song, J. M., Guan, Z., Hu, J., Guo, C., Yang, Z., Wang, S., et al. (2020). Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nat. Plants* 6, 34–45. doi: 10.1038/s41477-019-0577-7

Sturtevant, D., Lu, S., Zhou, Z. W., Shen, Y., Wang, S., Song, J. M., et al. (2020). The genome of jojoba (*Simmondsia chinensis*): A taxonomically isolated species that directs wax ester accumulation in its seeds. *Sci. Adv.* 6, 1–14. doi: 10.1126/sciadv.aay3240

Unver, T., Wu, Z., Sterck, L., Turktas, M., Lohaus, R., Li, Z., et al. (2017). Genome of wild olive and the evolution of oil biosynthesis. *Proc. Natl. Acad. Sci.* 114, E9413–E9422. doi: 10.1073/pnas.1708621114

Wang, C., Kong, Y., Hu, R., and Zhou, G. (2021). *Miscanthus*: A fast-growing crop for environmental remediation and biofuel production. *GCB Bioenergy.* 13, 58–69. doi: 10.1111/gcbb.12761

Wang, B., Wu, Z., Li, Z., Zhang, Q., Hu, J., Xiao, Y., et al. (2018). Dissection of the genetic architecture of three seed-quality traits and consequences for breeding in *Brassica napus*. *Plant Biotechnol. J.* 16, 1336–1348. doi: 10.1111/pbi.12873

Zhang, Q., Hua, X., Liu, H., Yuan, Y., Shi, Y., Wang, Z., et al. (2021). Evolutionary expansion and functional divergence of sugar transporters in saccharum (*S. spontaneum* and *S. officinarum*). *Plant J.* 105, 884–996. doi: 10.1111/tpj.15076

Zhang, J., Zhang, X., Tang, H., Zhang, Q., Hua, X., Ma, X., et al. (2018). Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* l. *Nat. Genet.* 50, 1565–1573. doi: 10.1038/s41588-018-0237-2

Zheng, S., Ye, C., Lu, J., Liufu, J., Lin, L., Dong, Z., et al. (2021). Improving the rice photosynthetic efficiency and yield by editing OsHXK1 *via* CRISPR/Cas9 system. *Int. J. Mol. Sci.* 22, 9554. doi: 10.3390/ijms22179554

Zhou, S., Li, X., Liu, Q., Zhao, Y., Jiang, W., Wu, A., et al. (2021). DNA Demethylases remodel DNA methylation in rice gametes and zygote and are required for reproduction. *Mol. Plant* 14, 1569–1583. doi: 10.1016/j.molp.2021.06.006

frontiers in Plant Science

ORIGINAL RESEARCH
published: 15 February 2022
doi: 10.3389/fpls.2022.839457

CG and CHG Methylation Contribute to the Transcriptional Control of OsPRR37-Output Genes in Rice

Chuan Liu[1], Na Li[1], Zeping Lu[1], Qianxi Sun[1], Xinhan Pang[1], Xudong Xiang[1], Changhao Deng[1], Zhengshuojian Xiong[1], Kunxian Shu[1], Fang Yang[2] and Zhongli Hu[2]*

[1] Chongqing Key Laboratory of Big Data for Bio Intelligence, Chongqing University of Posts and Telecommunications, Chongqing, China, [2] State Key Laboratory of Hybrid Rice, College of Life Sciences, Wuhan University, Wuhan, China

Plant circadian clock coordinates endogenous transcriptional rhythms with diurnal changes of environmental cues. OsPRR37, a negative component in the rice circadian clock, reportedly regulates transcriptome rhythms, and agronomically important traits. However, the underlying regulatory mechanisms of OsPRR37-output genes remain largely unknown. In this study, whole genome bisulfite sequencing and high-throughput RNA sequencing were applied to verify the role of DNA methylation in the transcriptional control of OsPRR37-output genes. We found that the overexpression of *OsPRR37* suppressed rice growth and altered cytosine methylations in CG and CHG sequence contexts in but not the CHH context (H represents A, T, or C). In total, 35 overlapping genes were identified, and 25 of them showed negative correlation between the methylation level and gene expression. The promoter of the hexokinase gene *OsHXK1* was hypomethylated at both CG and CHG sites, and the expression of *OsHXK1* was significantly increased. Meanwhile, the leaf starch content was consistently lower in *OsPRR37* overexpression lines than in the recipient parent Guangluai 4. Further analysis with published data of time-course transcriptomes revealed that most overlapping genes showed peak expression phases from dusk to dawn. The genes involved in DNA methylation, methylation maintenance, and DNA demethylation were found to be actively expressed around dusk. A DNA glycosylase, namely ROS1A/DNG702, was probably the upstream candidate that demethylated the promoter of *OsHXK1*. Taken together, our results revealed that CG and CHG methylation contribute to the transcriptional regulation of OsPRR37-output genes, and hypomethylation of *OsHXK1* leads to decreased starch content and reduced plant growth in rice.

Keywords: rice growth, DNA methylation, RNA-seq, circadian clock, output genes

OPEN ACCESS

Edited by:
Yin Li,
Huazhong University of Science and Technology, China

Reviewed by:
Shi Yan,
Fujian Agriculture and Forestry University, China
Wen Yao,
Henan Agricultural University, China

****Correspondence:***
Chuan Liu
liuchuan@cqupt.edu.cn

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

Received: *20 December 2021*
Accepted: *25 January 2022*
Published: *15 February 2022*

Citation:
Liu C, Li N, Lu Z, Sun Q, Pang X, Xiang X, Deng C, Xiong Z, Shu K, Yang F and Hu Z (2022) CG and CHG Methylation Contribute to the Transcriptional Control of OsPRR37-Output Genes in Rice. Front. Plant Sci. 13:839457. doi: 10.3389/fpls.2022.839457

INTRODUCTION

Plant DNA methylation occurs in the sequence context of CG, CHG, and CHH (where H is A, C, or T) (Zhang et al., 2006). DNA methylation is considered a stable epigenetic mark that can be transmitted across generations (Gehring, 2019). A specific DNA methylation state is also under the dynamic regulation by *de novo* methylation, maintenance of methylation, and active demethylation (Law and Jacobsen, 2010; Zhang et al., 2018). The function of DNA methylation in plant reproductive cells has been extensively studied, and this knowledge has deepened our

understanding of the dynamic DNA methylation patterns during plant development (Zheng et al., 2019; Higo et al., 2020; Kim et al., 2021; Zhou et al., 2021a). In addition, DNA methylation reportedly has roles in regulating plant architecture, plant defense against rice black-streaked dwarf virus, and multiple agronomical traits (Zhang et al., 2015, 2020; Kawakatsu, 2020; Xu et al., 2020). However, whether DNA methylation plays a role in regulating circadian clock output genes remains unclear.

Circadian clock comprises multiple transcription-translation feedback loops, which function to improve plant environmental adaptation. Altered expression of circadian clock genes, such as *CIRCADIAN CLOCK ASSOCIATED1* (*CCA1*), can increase levels of plant growth and fitness (Dodd et al., 2005; Masuda et al., 2020). Meanwhile, expression amplitude of *CCA1* is associated with the CHH methylation level in the promoter region, which determines plant growth vigor (Ng et al., 2014). Recently, it was reported that the circadian clock genes *ZEITLUPE* and *TIMING OF CAB EXPRESSION 1* act downstream of DNA methyltransferases to control circadian rhythm (Tian et al., 2021). These results shed light on DNA methylation-mediated regulation of clock gene expression. *Pseudo-Response Regulators* (*PRR*s) are key components of transcription-translation feedback loops in plants and mediate the circadian regulation of clock output genes (Farre and Liu, 2013), including genes involved in the regulation of growth, flowering, abiotic stress, and yield-related traits (Li C. et al., 2020; Li N. et al., 2020; Sun et al., 2020; Wei et al., 2020; Liang et al., 2021). *OsPRR37* was primarily identified to delay the flowering time and increase grain yield and adaptation in rice (Koo et al., 2013; Liu et al., 2013, 2015; Yan et al., 2013; Gao et al., 2014; Fujino et al., 2019). A recent study further revealed a distinct role of *OsPRR37* in promoting flowering in the japonica variety Zhonghua 11 under natural long-day conditions (Hu et al., 2021). Although several output genes involved in the photoperiodic flowering pathway are used to explain the trait variations (Chen et al., 2021; Zhou et al., 2021c), the underlying mechanism of how *OsPRR37* regulates its output genes and multiple traits remains unclear.

Circadian regulation of plant transcriptome benefits the acute responses of plants to the daily fluctuating environment (Panter et al., 2019). Our previous study confirmed that OsPRR37 protein functions as a transcriptional repressor and confers expanded regulation of transcriptome rhythms (Liu C. et al., 2018). The regulation of circadian-regulated genes by DNA methylation in *Populus trichocarpa* suggested that DNA methylation contributes to the expression levels of clock output genes (Liang et al., 2019). Based on these results, we hypothesize that OsPRR37 uses DNA methylation as a pathway to regulate the transcription of its output genes. In the present study, we sought to determine whether and how DNA methylation regulates OsPRR37-output genes. To this end, whole-genome bisulfide sequencing (WGBS) and high-throughput RNA sequencing (RNA-seq) were applied to identify the overlapping genes, which were considered to be the OsPRR37-output genes regulated by DNA methylation. The available data of time-course transcriptomes were used to confirm the expression change of overlapping genes and to analyze the genes involved in DNA methylation pathways. Our results revealed that DNA methylation was an alternative medium for OsPRR37 to regulate the output genes and plant growth.

MATERIALS AND METHODS

Plant Materials and Growth Conditions

The *OsPRR37* overexpression lines (OE5 and OE9) were described as previously reported (Liu C. et al., 2018). Briefly, *OsPRR37* overexpression lines were generated by overexpressing *OsPRR37* in an elite rice variety, namely Guangluai 4 (GL). NIL-*OsPRR37* is a nearly isogenic line in the GL background and contains the functional allele of *OsPRR37* from the elite variety Teqing. Rice growth phenotypes were obtained from the plants growing under natural long-day conditions in Wuhan University, Wuhan, China (30°54′01″N, 114°37′23″E). For WGBS and RNA sequencing, seeds of GL and OE5 were planted in a growth chamber (PRX-380B, Shanghai Guning Instrument Co., Ltd) for 15 days after germination. The growth chamber was set at 28°C under a 14-h light/10-h dark cycle with the light period of 6:00–20:00. The top most expanded leaves were harvested at 9:00, frozen in liquid nitrogen, and then stored at −80°C for DNA and RNA extraction. The same two biological replicates were applied to WGBS and RNA-seq.

Bisulfite-Seq Library Generation and Sequencing

Briefly, total genomic DNA of rice leaves was extracted using the cetyltrimethylammonium bromide method (Doyle, 1987). The DNA concentration and quality were estimated using NanoDrop 2000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, United States), Qubit 3.0 fluorometer (Life Technologies, Carlsbad, CA, United States), and 1.0% agarose gel electrophoresis. Then, 2-μg genomic DNA spiked with 5-ng unmethylated Lambda DNA (Promega, Madison, WI, United States) was fragmented by sonication to generate fragments measuring 300–500 bp. These fragments were then ligated with 5-methylcytosine-modified adapters and subjected to bisulfide conversion using the ZYMO EZ DNA Methylation-Gold Kit (Zymo Research, Irvine, CA, United States). The bisulfide-converted DNA was purified, recycled, and then amplified by PCR with 10 cycles using KAPA HiFi HotStart Uracil + ReadyMix (Kapa Biosystems, Wilmington, MA, United States) and Illumina 8-bp index primers. The WGBS libraries were analyzed using the Bioanalyzer 2100 system (Agilent Technologies, CA, United States) and sequenced on Illumina NovaSeq 6000 with a paired-end sequencing length of 150 bp (PE150) at Frasergen Bioinformatics Co., Ltd (Wuhan, China). The percentage of cytosines sequenced at cytosine reference positions in the lambda genome was considered to reflect the overall sodium bisulfite non-conversion rate.

Bisulfite-Seq Data Processing and Analysis

Quality control of WGBS data was performed using FastQC (version 0.11.9, Babraham Bioinformatics, United Kingdom).

Sequencing adapters and low-quality reads were removed using Trimmomatic (Bolger et al., 2014). The trimmed reads were then aligned and mapped to the rice reference genome of Nipponbare (MSU_v7.0) using Bismark (Krueger and Andrews, 2011). The percentage methylation level was calculated by mC/(mC + umC), where mC and umC represent the number of methylated and unmethylated reads, respectively. Only the CG/CHG/CHH sites with a read coverage of $\geq$ 5 across all samples were used for differential methylation analyses. Differentially methylated regions (DMRs) were identified using the R package "dmrseq" (Korthauer et al., 2019). Regions with a q-value < 0.05, number of CG/CHG/CHH sites $\geq$ 5 and methylation difference $> 20\%$ were defined as DMRs. DMR distribution on rice chromosomes was plotted using Circos (version 0.69) (Krzywinski et al., 2009). DMRs were annotated using ChIPseeker package (Yu et al., 2015). Methylation status along the genomic regions of DMRs $\pm$ 20 kb was plotted using pyGenomeTracks (version 3.6) (Lopez-Delisle et al., 2021).

RNA Library Generation and Sequencing

Total RNA from rice leaves was extracted using TRIzol Reagent (Invitrogen, CA, United States) for RNA sequencing. RNA purity and integrity were analyzed using a NanoDrop 2000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, United States) and the Bioanalyzer 2100 system (Agilent Technologies, CA, United States). RNA contamination was assessed by 1.5% agarose gel electrophoresis. A total of 1 μg of RNA per sample was used as the input material for library preparation. The mRNA was purified from the total RNA using poly-T oligo-attached magnetic beads. Sequencing libraries were generated from the purified mRNA using the VAHTS Universal V6 RNA-seq Library Kit for MGI (Vazyme, Nanjing, China) following the manufacturer's recommendations with unique index codes. The library quantification and size were assessed using a Qubit 3.0 fluorometer (Life Technologies, Carlsbad, CA, United States) and Bioanalyzer 2100 system (Agilent Technologies, CA, United States). Subsequently, sequencing with a paired-end sequencing length of 150 bp (PE150) was performed on the MGI-SEQ 2000 platform (MGI Tech Co., Ltd. Shenzhen, China) by Frasergen Bioinformatics Co., Ltd (Wuhan, China).

Bioinformatics Analysis of RNA Sequencing and Microarray Data

Sequencing adapters and low-quality reads were removed with fastp (version 0.20.1) (Chen et al., 2018), and the quality of raw reads was evaluated with FastQC (version 0.11.9, Babraham Bioinformatics, United Kingdom). The remaining clean reads were mapped to the rice reference genome of Nipponbare (MSU_v7.0) using Hisat2 (version 2.1.0) (Kim et al., 2019). Mapping statistics were generated using Samtools (version 1.11) (Li H. et al., 2009). TPMCalculator was used to count the reads mapped to individual genes as well as to measure gene expression levels by calculating transcripts per million (TPM) read values (Vera Alvarez et al., 2019). Differentially expressed genes (DEGs) were identified using DESeq2 (Love et al., 2014). Gene Ontology (GO) enrichment analysis was performed using the GO annotation file MSU7.0 gene ID (TIGR) of agriGO v2.0 and clusterProfiler 4.0 (Tian et al., 2017; Wu et al., 2021). KEGG enrichment analysis was performed using KOBAS 3.0, and the output data were plotted using clusterProfiler 4.0 (Bu et al., 2021). Gene symbols with known or unknown function were annotated using MBKbase-rice database[1], funRiceGenes database[2], and China Rice Data Center[3] (Yao et al., 2018; Peng et al., 2020). The raw RNA-seq data of time-course transcriptomes, which were downloaded from NCBI-GEO database (GSE114188), were reanalyzed as per the RNA-seq data processing pipeline in this study. The corresponding time-course samples of GL and OE5 comprise six time points (4:00, 8:00, 12:00, 16:00, 20:00, and 0:00) with three replicates at 45 days growth under natural long-day conditions. Statistical significance of different expressions was evaluated by unpaired Student's t-test at each time point. The microarray data were obtained by GSE19024 on NCBI-GEO (Wang et al., 2010). The corresponding tissue samples of interest were described in **Supplementary Table 1**, which were the subset of samples in a previous study (Wang et al., 2010). Before being used to plot the heatmap, the signal values of biological and technical replicates for the same tissue were averaged.

Quantitative RT-PCR and Starch Content Determination

Quantitative RT-PCR was conducted with the same protocol as previously reported (Liu et al., 2015). The PCR primer sequences were 5′-TGACAAAGCCTAGTACAAATAAGGAGAG-3′ and 5′-CAGTGCTGTGCAGGATGAAATG-3′. Approximately, 0.2 g of fresh leaf samples were weighed before estimating the starch content. Starch content was determined according to previously published protocols (Smith and Zeeman, 2006).

RESULTS

Rice Growth Was Repressed by Overexpressing *OsPRR37*

During the field trails, we observed that rice growth was retarded in *OsPRR37* overexpression lines (OE). To investigate the effects of *OsPRR37* overexpression on rice growth, we record the morphology and dry weight of GL, OE5, OE9, and NIL-*OsPRR37* at 25, 40, and 55 days after sowing the seeds. The growth of OE5 and OE9 was significantly repressed compared to the growth of GL and NIL-*OsPRR37* on these days (**Figures 1A–F**). These results suggest that natural loss-of-function and gain-of-function alleles of *OsPRR37* showed a comparable growth rate during the vegetative growth period. Then, the diurnal expression profile of *OsPRR37* was monitored over a day using quantitative RT-PCR. *OsPRR37* was identified as having similar expression rhythms in GL and NIL-*OsPRR37* as their peak expression phase was around 12:00. Conversely, the expression of *OsPRR37* in OE5 and OE9 was much higher and showed altered rhythms with the peak expression phase around 4:00 (**Figure 1G**). These

[1]http://www.mbkbase.org/rice/

[2]https://funricegenes.github.io/

[3]https://www.ricedata.cn/gene/

results suggested that the overexpression of *OsPRR37* changed its diurnal rhythms and repressed rice growth. In a previous study, it was reported that *OsPRR37* widely regulates output genes and particularly suppresses output genes with phases around 9:00 (Liu C. et al., 2018). To investigate whether DNA methylation associated with OsPRR37 regulates the output genes, samples of GL and OE5 at 9:00 were subjected to WGBS and RNA-seq. The workflow of this study is shown in **Figure 1H**.

Whole-Genome Bisulfide Sequencing and RNA Sequencing Quality Assessment and Alignment

Whole-genome bisulfide sequencing and RNA-seq were used to investigate the role of DNA methylation in the transcriptional regulation of OsPRR37-output genes. WGBS generated 48,752,185 and 63,773,173 raw reads for the two GL replicates and 59,530,161 and 56,527,166 raw reads for the two OE5 replicates. After quality control filtration, 45,246,229 and 59,675,582 clean reads remained for GL, and 56,065,476 and 53,133,046 clean reads remained for OE5. The clean reads ratio ranged from 92.8 to 94.2%. The average percentage of Q30 and GC content for the four sequencing libraries was 92.7 and 22.9%, respectively (**Supplementary Table 2**). Of those clean data, 53.4% (GL-1), 51.7% (GL-2), 54.2% (OE5-1), and 52.2% (OE5-2) were uniquely mapped to the rice genome (**Supplementary Table 3**). Overall, 30,872,222 CG sites, 27,422,379 CHG sites, and 104,533,760 CHH sites were identified with sequencing coverage range from 53.1 to 69.4% (**Supplementary Table 4**). Among these, 10.6%–13.0% CG sites, 7.7–9.7% CHG sites, and 4.0–4.9% CHH sites were methylated (**Supplementary Table 5**). The same samples of WGBS were used in RNA-seq to obtain comparable data. In total, RNA-seq generated 23,559,582 (GL-1), 25,090,638 (GL-2), 25,898,111 (OE5-1), and 26,350,196 (OE5-2) clean read pairs (**Supplementary Table 6**). The percentages of Q30 ranged between 84.9 and 86.1%. Of these clean read pairs, 90.0 to 90.7% were mapped to the reference genome of rice (**Supplementary Table 7**). These data were sufficient and reliable for subsequent differential methylation and expression analysis.

Overexpressing *OsPRR37* Altered Global CG and CHG Methylation

To identify DMRs, three cytosine sequence contexts (CG, CHG, and CHH) were separately applied to differential methylation analysis. Genomic regions with a q-value of < 0.05 and a differential methylation level of $> 20\%$ were considered as DMRs. A total of 321 and 949 DMRs were found in CG (DMR-CG) and CHG (DMR-CHG) sequence contexts, respectively. However, no DMRs were identified in the CHH sequence context. Among DMR-CG, 90 were hypermethylated and 231 were hypomethylated (**Figure 2A**). Conversely, among DMR-CHG, 480 were hypermethylated and 469 were hypomethylated (**Figure 2B**). These data revealed a higher proportion of hypomethylated DMR-CG (72.0%) than DMR-CHG (49.4%). Furthermore, the significance of methylated DMRs across the 12 chromosomes found that DMRs were evenly distributed on the rice genome and were of high significance (**Figures 2C,D**).

To obtain DMR-associated genes (DMGs), DMR-CG and DMR-CHG were both annotated with the R package "ChIPseeker." The result showed that 32.7%, 20.6%, and 15.9% of DMR-CG were located in the promoter region at ≤1 kb, 1–2 kb, and 2–3 kb upstream of transcription start site, respectively (**Figure 3A**). The distal intergenic region accounted for 20.2% of DMR-CG. Conversely, only a small fraction of DMR-CG was annotated within Intron (2.5%), Exon (2.5%), Downstream (≤1 kb: 0.9%, 1–2 kb: 2.2%, and 2–3 kb: 1.2%) and 3′UTR (1.2%). This means that most DMR-CG (69.2%) were located in the promoter region (**Figure 3A**). Similarly, 70.3% of DMR-CHG were located in promoter region (**Figure 3B**). These results supported the notion that cytosine methylation majorly occurred in the promoter sequence. Then, we performed functional enrichment analysis with these DMGs. The network of five most enriched GO ontologies for DMR-CG-associated genes showed that LOC_Os02g03540 and LOC_Os06g41360 enable ribose phosphate diphosphokinase activity and are involved in ribonucleoside monophosphate biosynthetic process (**Figure 3C**). LOC_Os03g45410/OsTBP2 and LOC_Os03g14720 enable obsolete RNA polymerase II transcription factor activity and are involved in transcription initiation from the RNA polymerase II promoter. LOC_Os03g45410/OsTBP2 was reported to be a TATA-binding protein, which interacts with the transcription factor IIB (Zhu et al., 2002). Interestingly, LOC_Os03g14720 is a putative transcription initiation factor IIF. These results highlighted that DMR-CG-associated genes mainly function in gene transcription regulation. However, no GO term was found to be significantly enriched for DMR-CHG-associated genes.

Differentially Expressed Gene Analysis

Although *OsPRR37* overexpression altered the methylation of >1,000 DMGs, the number of transcriptionally regulated DMGs remains unknown. Samples of GL and OE5 were subjected to RNA-seq to profile the genome-wide gene expressions. The overall gene expression level was slightly higher for OE5 than for GL (**Supplementary Figure 1**), whereas the expression correlation between samples was ranged from 0.9804 to 0.9929 (**Supplementary Figure 2**). Genes with low expression (sum of TPM being < 2 in both GL and OE5) were filtered out. A total of 743 DEGs ($|\log_2 FC| > |\log_2 1.5|$, adjusted P-value < 0.05) were identified between GL and OE5 (**Supplementary Figure 3**). Among these DEGs, 286 (38.5%) were downregulated and 457 (61.5%) were upregulated (**Figures 4A,B**). The increment of the mean TPM for upregulated DEGs was larger than the decrement of the mean TPM for downregulated DEGs (**Figure 4A**). The expression of DEGs in the two replicates was similar so that DEGs are robust to be further analyzed (**Figure 4B**).

GO enrichment analysis found that DEGs are instrumental in metal ion binding, transporter activity, and chitinase activity and are mainly involved in carbohydrate metabolic process, chitin catabolic process, defense response to bacterium and fungus, and nitrate assimilation, among other functions (**Figures 4C,D**). The KEGG pathway enrichment analysis showed that upregulated DEGs participate in amino sugar and nucleotide sugar metabolism, carbon metabolism, glycerolipid

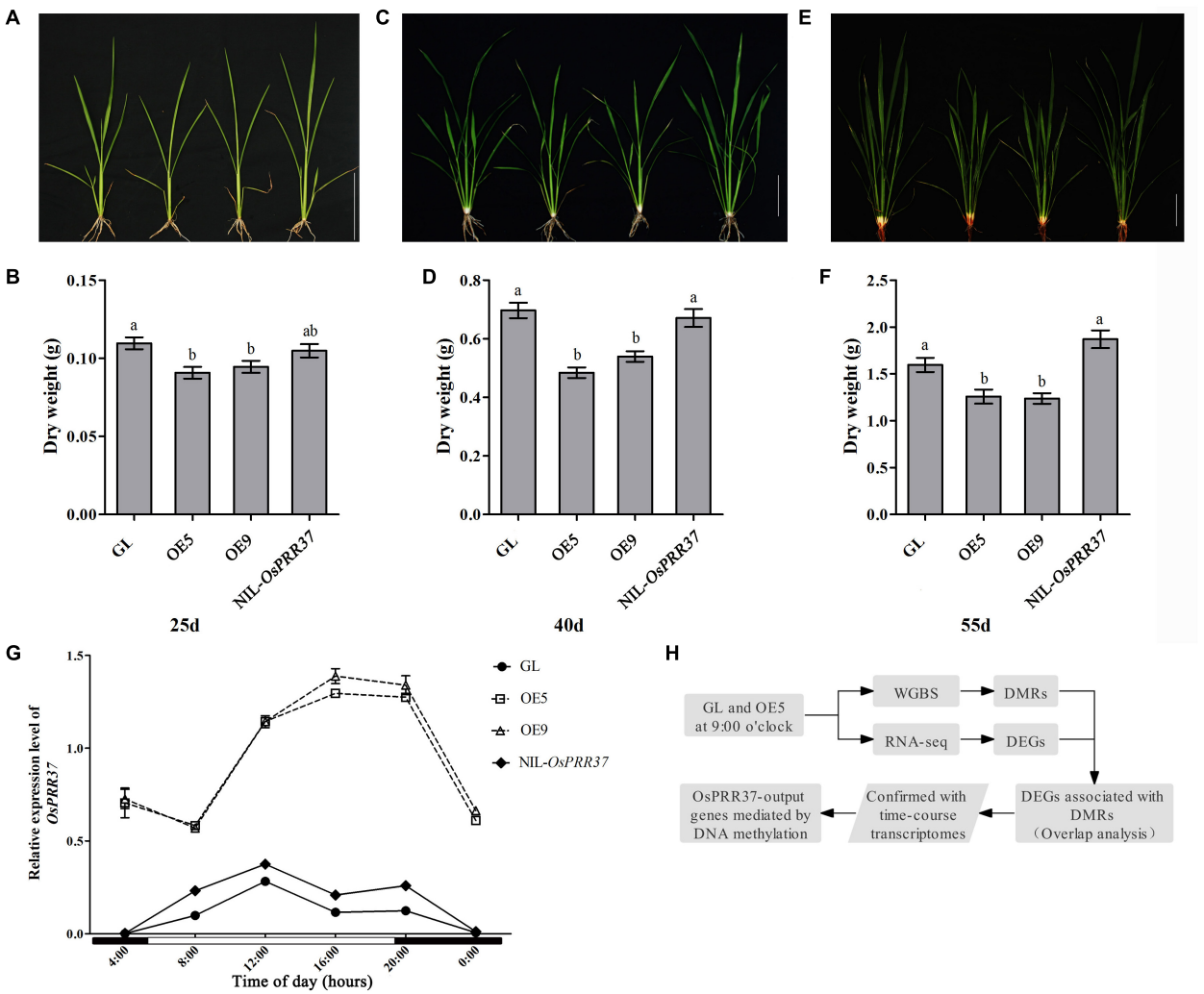


FIGURE 1 | Characterization of rice growth and *OsPRR37* expression rhythms. Rice growth morphology and dry weight were documented at 25 days **(A,B)**, 40 days **(C,D)**, and 55 days **(E,F)** after sowing the seeds. The different lower-case characters above bars represent the significant difference level of $P < 0.05$. **(G)** The expression levels of *OsPRR37* detected by quantitative RT-PCR. The exact time of a natural long-day condition was indicated under the *X* axis. **(H)** Simplified workflow of the present study. OE5 and OE9 are two independent transgenic lines.

metabolism, MAPK signaling pathway, and circadian rhythm, among other roles. Downregulated DEGs are mainly involved in plant–pathogen interaction, nitrogen metabolism, and circadian rhythm (**Figures 4E,F**).

Characterization of Overlapping Genes Between DMR-Associated Genes and Differentially Expressed Genes

Overlap analysis between DMGs and DEGs was performed to identify transcriptionally regulated DMGs. In total, 35 genes were found to be shared between DMG and DEG sets. Among these, five DEGs were found to be differentially methylated in both CG and CHG sequence contexts, whereas 19 DEGs were uniquely shared with DMR-CHG and 11 DEGs were uniquely shared with DMR-CG (**Figure 5A**). The correlation between methylation difference and expression fold-change of overlapping genes was investigated. Consequently, 14 of 16 genes (87.5%) showed negative correlation between expression and CG methylation, and 14 of 16 genes (87.5%) were methylated in promoter regions (**Figure 5B**). In contrast, 16 out of 24 genes (66.7%) showed a negative correlation between expression and CHG methylation, and 15 of 24 genes (62.5%) were methylated in promoter regions (**Figure 5C**). After removing the redundant genes, in total, 25 genes showed negative correlation between expression and cytosine methylation level (**Figures 5B,C**). Functional annotation with the MBKbase, funRiceGenes database and China Rice Data Center identified seven genes with known function: *OsHXK1* (*LOC_Os07g26540*), *OsZIP9* (*LOC_Os05g39540*), *SDT/OsmiR156h* (*LOC_Os06g44034*), *OsMADS18* (*LOC_Os07g41370*), *OsPT11* (*LOC_Os01g46860*), *OsRLCK109/OsBBS1* (*LOC_Os03g24930*), and OsNAS3 (*LOC_Os07g48980*). Among these genes, *OsHXK1* showed the highest negative correlation between methylation difference

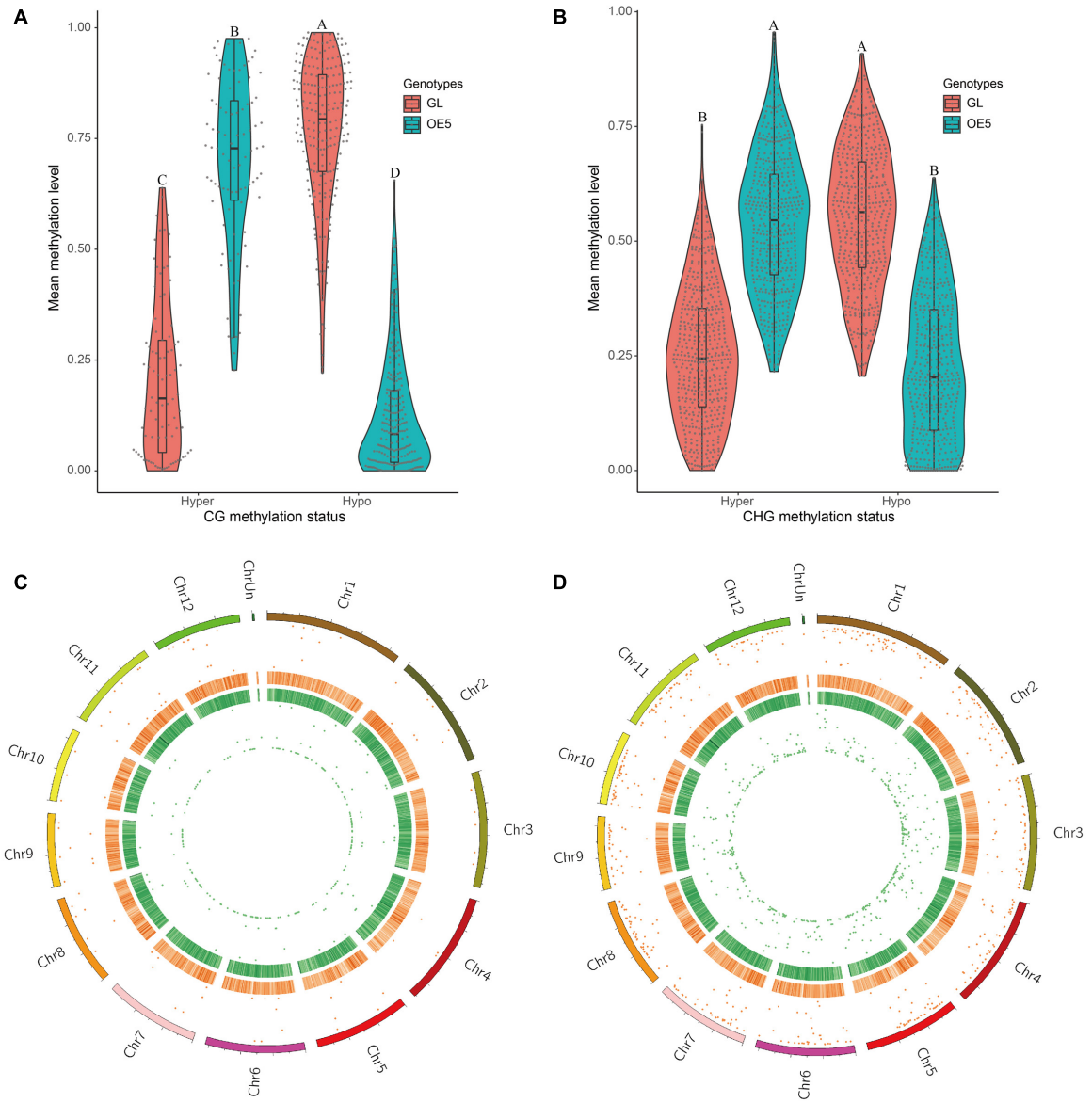


FIGURE 2 | Identification and analysis of differentially methylated regions. Comparisons of hypermethylation and hypomethylation levels in GL and OE5 were plotted for both CG **(A)** and CHG **(B)** sequence contexts. Different letters above violin plots represent significant differences at $P < 0.01$ as revealed by one-way ANOVA analysis (Tukey's multiple comparison test). DMR distribution on rice chromosomes in CG **(C)** and CHG **(D)** sequence contexts is shown. From outer to inner layers, the circular plots represent chromosomes, hyper-DMR distribution (the more outward means higher significance), heatmap of GC content (deeper red colors indicate higher GC content), heatmap of gene density (deeper green colors indicate higher gene density), and hypo-DMR distribution (greater proximity to the center of the circle indicates higher significance).

and expression fold-change (**Figures 5B,C**). Detailed methylation status indicated that the CG and CHG sites in the promoter region of *OsHXK1* were both hypomethylated (**Figure 5D**).

Diurnal Rhythms and Functional Characterization of Overlapping Genes

As OsPRR37 is a key component in the rice circadian clock, diurnal rhythms of overlapping genes were further investigated with the reported time-course RNA-seq data of leaf samples at 45 days growth (Liu C. et al., 2018). Among the 35 overlapping genes, 29 were observed to be differentially expressed with at least one timepoint, which confirmed the identification of DEGs in this study (**Figure 6A** and **Supplementary Figure 4**). We also found that 31 overlapping genes showed diurnal rhythms, and 27 were observed to show a peak expression phase of 16:00–04:00. These data indicated that most differentially methylated DEGs were under circadian control and tended to function from dusk to dawn. Further investigation of the seven reported overlapping genes revealed that four of

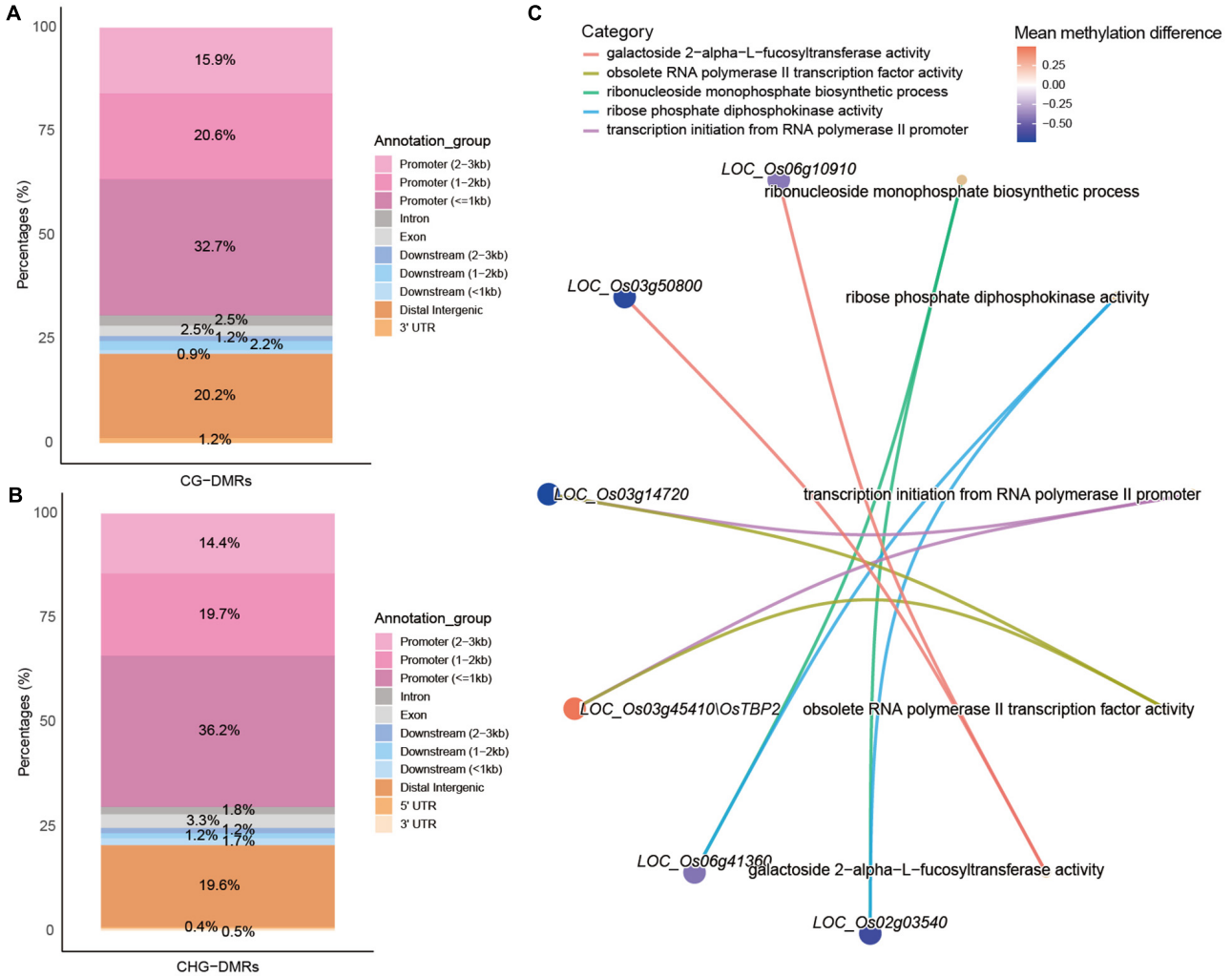


FIGURE 3 | Characterization of DMR-associated genes. The distributions of CG-DMRs **(A)** and CHG-DMRs **(B)** summarized according to the annotated genomic location. The order of the stacking bar charts is consistent with that of the legends on the right. **(C)** The network between genes and enriched GO terms. Genes are indicated as MSU loci or gene symbols. GO categories are displayed in the name of GO terms.

them showed significant differences in their expression across different time points of the day (**Figure 6A**). Interestingly, *OsHXK1*, *SDT*/*OsmiR156h*, *OsMADS18*, and *OsPT11* were diurnally expressed and showed a peak expression phase of 20:00–4:00, indicating that they predominantly function during the night. *SDT*/*OsmiR156h* can modulate the rice yield, plant architecture, and seed dormancy by targeting *Ideal Plant Architecture1* (*IPA1*) (Jiao et al., 2010; Miao et al., 2019). Its continuously high expression suggested the involvement of the *SDT*/*OsmiR156h-IPA1* module in *OsPRR37*-mediated rice growth regulation. Except for *SDT*/*OsmiR156h*, the other six genes were mapped in the microarray data of Zhenshan97 tissues (Wang et al., 2010). *OsMADS18* was widely expressed in the tissues of seedling, leaf, shoot, sheath, stem, and panicle, which is in line with its function in flowering signal transduction (**Figure 6B**; Fornara et al., 2004; Yin et al., 2019). The significant repression of *OsMADS18* can partly explain the delayed growth and flowering (**Figure 6A**). *OsNAS3* encodes a nicotianamine synthase that is important for Fe homeostasis (Aung et al., 2019). Our result showed that *OsNAS3* was widely expressed in germinating seed, plumule, radicle, seedling, leaf, root, shoot, sheath, stem, panicle, and spikelet. *OsRLCK109*/*OsBBS1* was diurnally expressed with a peak phase of 16:00–20:00 and was highly expressed in leaf, root and sheath to regulate leaf senescence and salt stress responses (Zeng et al., 2018). *OsZIP9* was mainly expressed in the root and sheath to uptake zinc for rice growth (Tan et al., 2020; Yang et al., 2020). However, even though *OsPT11* is a rice phosphate transporter that regulates phosphate uptake and transport (Paszkowski et al., 2002; Yang et al., 2012), it was highly expressed in many tissues, such as geminating seed, radicle, root, leaf, sheath, stamen, endosperm, and panicle. The role of *OsZIP9* and *OsPT11* in coordinating ion uptake and rice growth needs to be further confirmed.

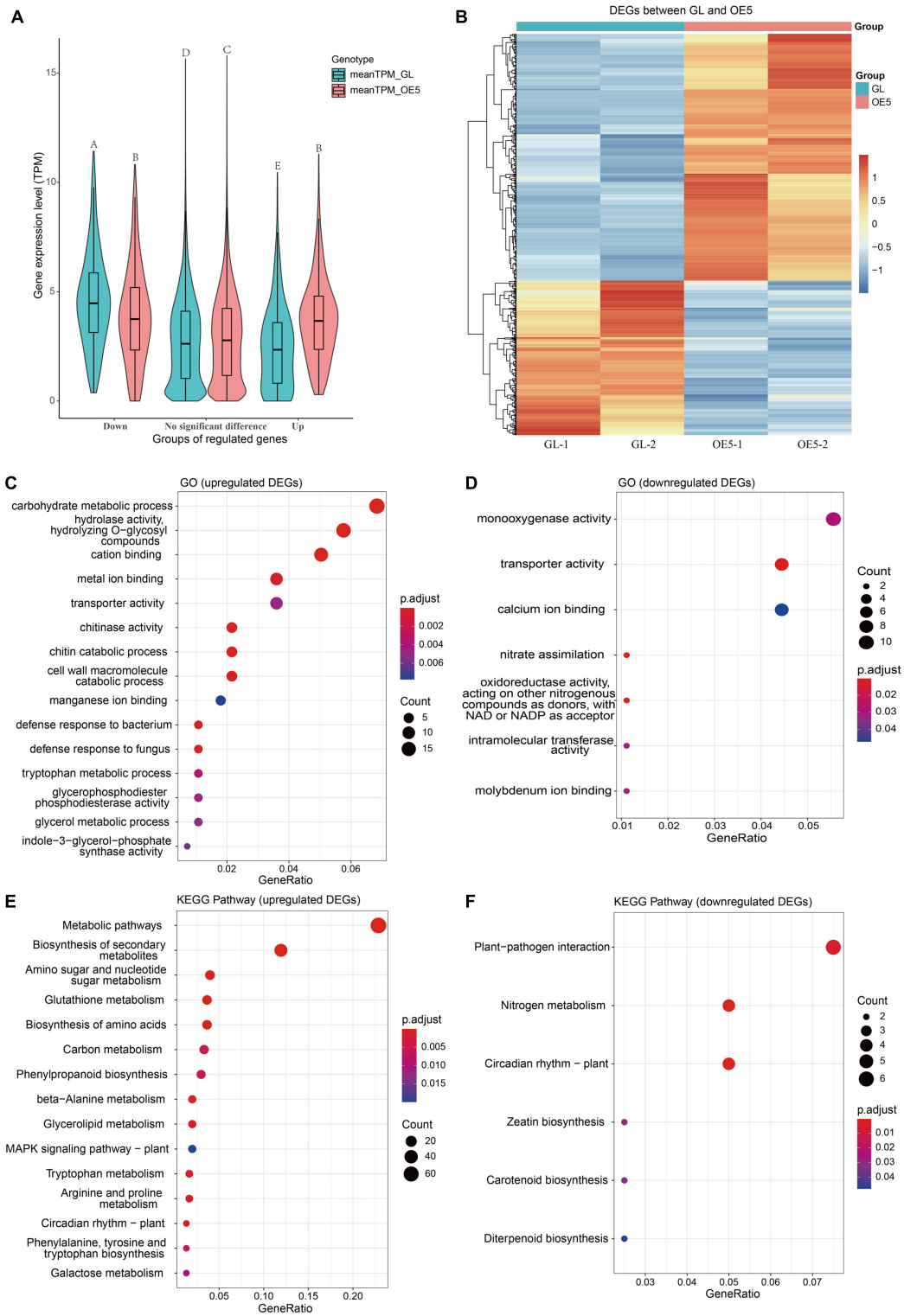


FIGURE 4 | Expression and functional enrichment of differentially expressed genes. **(A)** A comparison of mean expression levels of upregulated, no significant difference, and downregulated genes between GL and OE. Different letters above violin plots represent significant differences at $P < 0.01$ as revealed by one-way ANOVA analysis (Tukey's multiple comparison test). **(B)** The heatmap of 457 upregulated and 286 downregulated DEGs. The TPM value of DEG was scaled by row with the "pheatmap" package in R. Dotplots of significant GO terms for upregulated **(C)** and downregulated **(D)** DEGs. Dotplots of significant KEGG pathways for upregulated **(E)** and downregulated **(F)** DEGs. GO terms and KEGG pathways with adjusted *P*-value < 0.05 were considered as significant, and if the number of significant terms or pathways was > 15, only 15 terms or pathways were plotted.

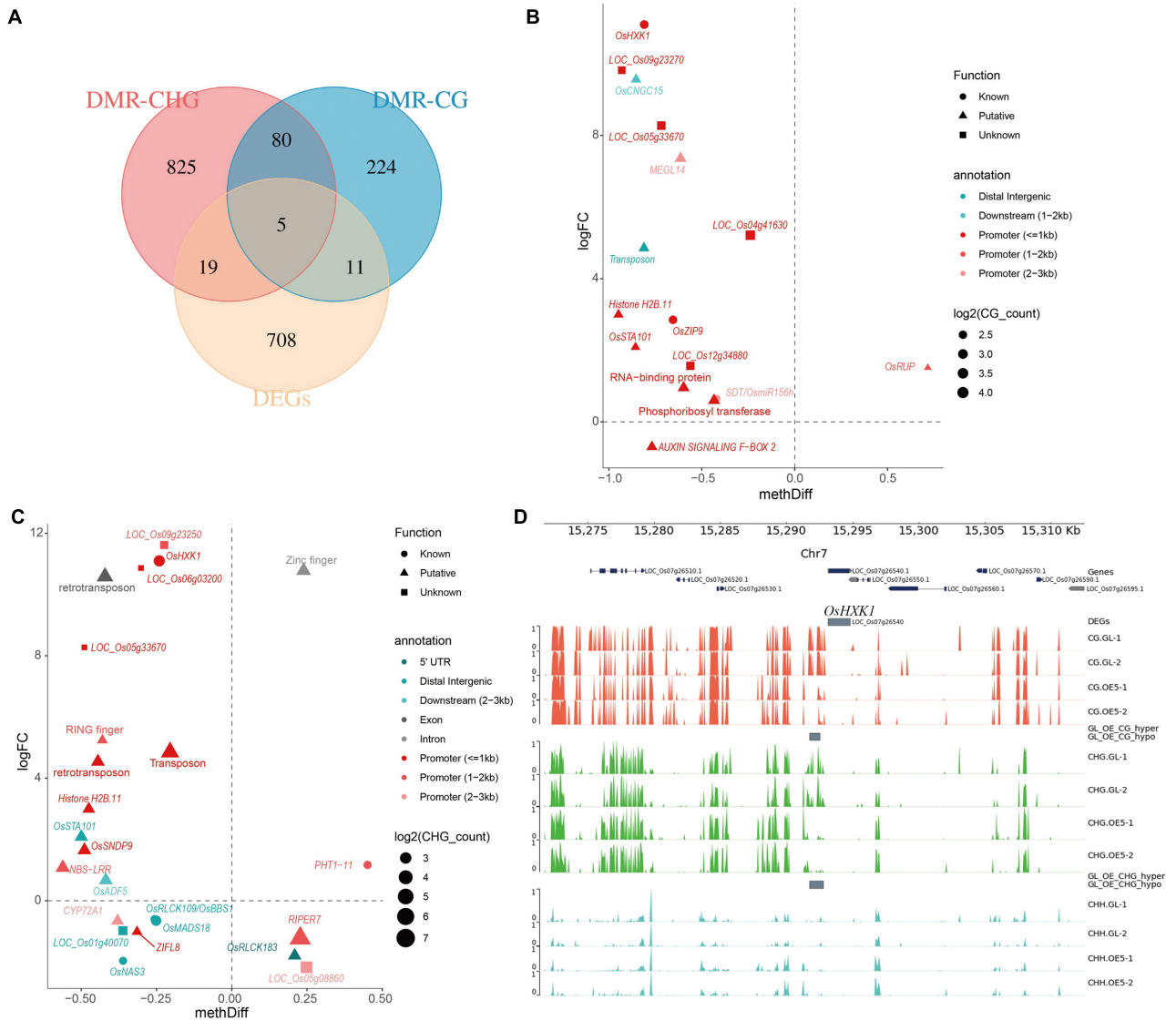


FIGURE 5 | Overlap analysis between DMGs and DEGs. **(A)** The distribution of overlapping genes between DMGs and DEGs. Scatter plots of the correlation between methylation difference and expression fold change for overlapping genes in CG **(B)** and CHG **(C)** contexts. **(D)** Visualization of methylation status for the representative gene *OsHXK1*.

OsHXK1 was identified to be highly expressed in the germinating seed, leaf, stem, stamen, and endosperm (**Figure 6B**). This expression pattern is in close agreement with a previous study wherein *OsHXK1* was reported to regulate reactive oxygen species in rice anthers (Zheng et al., 2019) and knockout of *OsHXK1* improved rice photosynthetic efficiency and yield (Zheng et al., 2021). In plant leaves, starch is accumulated during the day and consumed by respiration at night, and therefore, the starch content can indicate the strength of photosynthesis. To investigate whether the photosynthesis product was altered by the significantly elevated expression of *OsHXK1*, we compared the total starch content in GL, OE5, OE9, and NIL-*OsPRR37* leaves at the ending (20:00) and beginning (4:00) of the day. We found the total starch content to be significantly lower in OE5 and OE9 than in GL and NIL-*OsPRR37* at both time points (**Figure 6C**). The low starch content in OE lines resulted in energy deficit, consequently causing repressed rice growth (**Figure 1**). These results revealed that the enhanced expression of *OsHXK1* by hypomethylation decreased starch content and rice growth, thus suggesting that *OsHXK1* is a key output gene applied by *OsPRR37* to regulate rice growth.

Diurnal Expression Analysis of DNA Methylation Related Genes

DNA methylation patterns are decided by coordinated regulation of DNA methylation and demethylation pathways. To explore

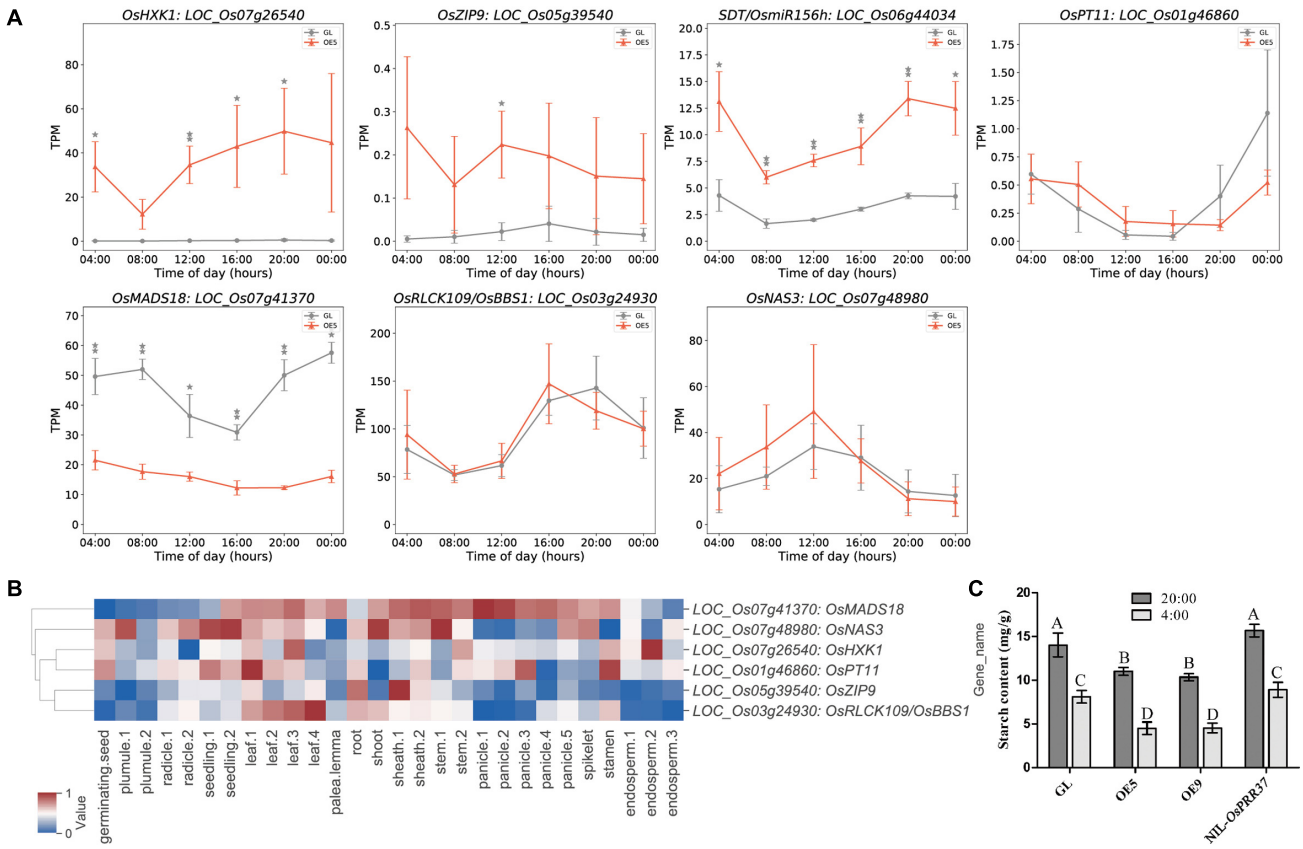


FIGURE 6 | Diurnal and tissue-specific expression analysis of the known overlapping genes. **(A)** Diurnal expression of the known overlapping genes between DMGs and DEGs. The asterisks above curves mean significance difference at $P < 0.05$ (one asterisk) and at $P < 0.01$ (two asterisks). **(B)** Tissue-specific expression analysis of known overlapping genes. **(C)** The measure of the starch content in rice leaves at the end (20:00) and beginning (4:00) of the day. Different letters above bars represent significant differences at $P < 0.01$ as revealed by one-way ANOVA analysis (Bonferroni's multiple comparison test). Values represent the means ± SD of three replicates.

the upstream genes of DMGs, the genes involved in DNA methylation were profiled with time-course RNA-seq data. In total, based on a recent study, 36 genes were grouped into the RNA-directed DNA methylation (RdDM) pathway, methylation maintenance, and DNA demethylation (Sun et al., 2021). Interestingly, most of them (26 genes) exhibited a peak expression phase of 16:00–20:00 (**Figure 7** and **Supplementary Figure 5**), indicating that DNA methylation, methylation maintenance, and DNA demethylation are particularly active around dusk. Among these, several genes encoding Argonaute (AGO) proteins are altered, including the downregulated *AGO1a* and *AGO1d* and the upregulated *AGO1b* and *AGO18* (**Figure 7**). The miR168-*AGO1* module can regulate multiple miRNAs to improve yield, reduce flowering time, and enhance immunity (Wang et al., 2021). Meanwhile, AGO18 sequesters miR168 to alleviate the repression of rice AGO1 (Wu et al., 2015), and a regulation module of *miR168a–OsAGO1/OsAGO18–-miRNAs*-target genes was proposed to regulate agronomically important traits (Zhou et al., 2021a). These results combined with our data suggest a causal link between rice growth repression and the altered module of *miR168a-OsAGO1/OsAGO18*.

A relatively similar expression of methylation maintenance genes was observed between GL and OE5. Conversely, four out of five genes of DNA demethylation showed significantly different expression in GL and OE5 (**Figure 7**). Interestingly, the expression of *ROS1C/DNG701* and *ROS1D/DNG704* was suppressed, whereas that of *ROS1A/DNG702* was increased. These three genes were recently reported to demethylate DNA in the gamete and zygote, which is crucial for zygote gene expression and development (Zhou et al., 2021b). In addition, a mutation in *ROS1A/DNG702* can generally lead to the increase of CG and CHG but not of CHH hypermethylation on genomes of rice endosperms (Liu J. et al., 2018). This result strongly corroborates our data wherein DMRs were only identified in CG and CHG sequence contexts, and because the expression of *ROS1A/DNG702* was increased in OE5, markedly more hypo-DMRs (55.1%) and upregulated DEGs (61.5%) were observed (**Figure 7**). The effect of *ROS1A/DNG702* may be counteracted by the decreased expression of *ROS1C/DNG701* and *ROS1D/DNG704*. Taken together, we believe that ROS1A/DNG702 was the upstream protein that demethylated the promoter regions of *OsHXK1* and

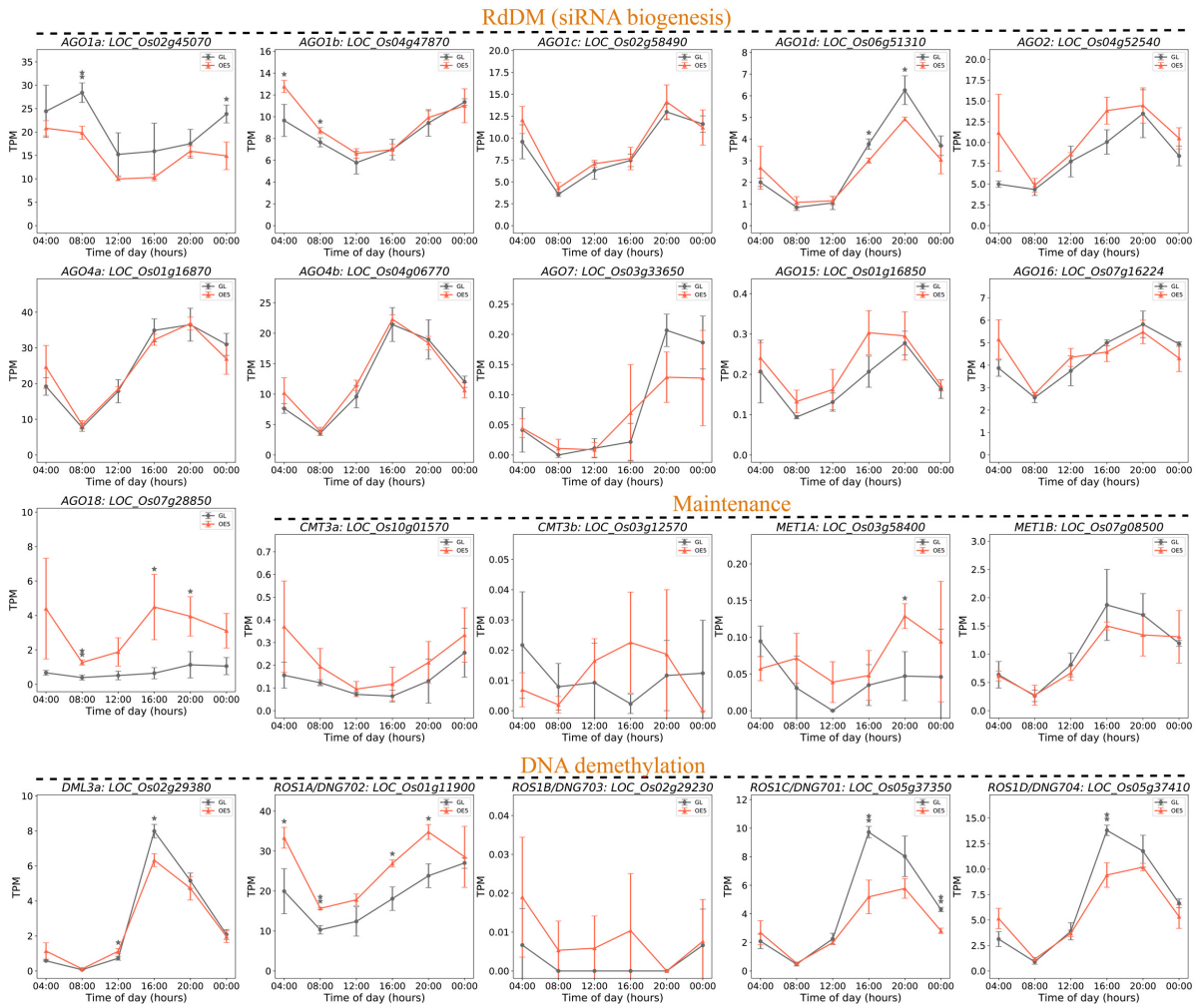


FIGURE 7 | Diurnal expression patterns of DNA methylation-related genes. Eleven *AGO* family members in the RdDM pathway, four genes for methylation maintenance, and five for DNA demethylation were presented with data of time-course transcriptomes. The three groups of genes are arranged under horizontal dash lines with group names of RdDM (siRNA biogenesis), Maintenance, and DNA demethylation. The other five DNA methylation genes and 11 siRNA biogenesis genes in the RdDM pathway are shown in **Supplementary Figure 5**. The asterisks above curves mean significance difference at $P < 0.05$ (one asterisk) and $P < 0.01$ (two asterisks).

enhanced its expression, thus leading to decreased starch content and reduced rice growth.

DISCUSSION

Circadian Rhythm of *OsPRR37* Is Important for Rice Growth

The endogenous expression period of clock genes is crucial for a plant to match the light–dark cycle. If correctly matched, the plant circadian system will enhance photosynthetic carbon fixation and growth (Dodd et al., 2005). Our results found that the circadian expression pattern of *OsPRR37* in OE5 and OE9 was significantly different from that in GL and NIL-*OsPRR37* (**Figure 1**). Although GL contains a loss-of-function allele of *OsPRR37*, the circadian rhythm and plant growth observed were similar between GL and NIL-*OsPRR37* (**Figure 1**). Moreover, GL and NIL-*OsPRR37* showed no significant difference in the starch content (**Figure 6C**). These results confirmed that disturbing circadian rhythm of *OsPRR37* decreased starch content and plant growth.

Input and Output Pathways for *OsPRR37*

The regulatory network of transcription-translation feedback loops in the core circadian clock is well drawn based on exciting results of research on clock genes (Nakamichi, 2020). However, the inputs and outputs of the circadian clock remain unclear. The photosynthetic endogenous sugar levels provide metabolic entrainment to the circadian clock system through the morning-phased gene *PRR7*, the homolog of *OsPRR37* in Arabidopsis (Haydon et al., 2013). A recent study reported that PRR7 mediates the circadian input to the promoter of CCA1 in

the shoots (Nimmo and Laird, 2021). These results indicate an entrainment route of sugar–PRR7–CCA1. In rice, sugars suppress *OsCCA1* expression while OsCCA1 regulates *IPA1* expression to mediate panicle and grain development (Wang et al., 2020). Our results suggested that the *SDT*/*OsmiR156h*-IPA1 module was involved in modulating OsPRR37-mediated rice growth. Based on these results, whether and how sugar–*OsPRR37*–*OsCCA1*–*SDT*/*OsmiR156h*–*IPA1* comprises an integrated pathway need more evidence in the future. Furthermore, the enrichment analysis found that upregulated genes were enriched in carbohydrate metabolic process (**Figure 4C**), amino sugar and nucleotide sugar metabolism, and carbon metabolism pathways (**Figure 4E**). These results suggested that sugar and carbon metabolism pathways are altered by *OsPRR37* overexpression. Meanwhile, several downregulated genes are enriched in nitrate assimilation (**Figure 4D**) and nitrogen metabolism (**Figure 4F**), suggesting that nitrate assimilation and metabolism would be other pathways coordinated by OsPRR37 to affect plant growth.

The Role of Differentially Methylated OsPRR37-Output Genes

Epigenetic modifications are closely associated with alterations in chromatin structure, such as histone modification and DNA methylation. Rhythmic transcription of Arabidopsis clock genes was considered to be regulated by rhythmic histone modification (Song and Noh, 2012). However, to our knowledge, there has been no research on the role of DNA methylation in regulating clock output genes. OsPRR37 was believed to repress morning-phased output genes and indirectly activate evening-phase output genes (Liu C. et al., 2018). In the present study, as we focused on samples in the morning (9:00), and our primary goal was to identify the key overlapping genes that were downregulated by OsPRR37. In this process, we hoped to get some insight into how OsPRR37 is associated with DNA methylation pathways so as to directly repress the morning-phased output genes. However, the results showed that 25 out of 35 overlapping genes were upregulated, and the expression levels of 22 genes were negatively correlated with methylation levels (**Figures 5B,C**). These results supported our hypothesis that DNA methylation contributed to the regulation of OsPRR37-output genes, but the dynamic methylation of these output genes is probably under an indirect regulation of OsPRR37. In other words, the differentially methylated output genes are in the most downstream of OsPRR37, such as *OsHXK1*, *SDT*/*OsmiR156h*, and *OsMADS18*, which are more directly to regulate rice growth, flowering, and yield.

The Hierarchical Regulation Network of *OsPRR37*

Different members of PRRs are supposed to function at their specific times of the day to repress clock output genes (Farre and Liu, 2013). Accumulating evidence has indicated that PRRs interact with other proteins to regulate the transcription of output genes. The B-box (BBX)-containing proteins BBX19 and BBX18 can physically interact with PRR9, PRR7, and PRR5 in a precise temporal order from dawn to dusk, thus cooperatively regulating the output genes (Yuan et al., 2021). OsPRR73 interacts with histone deacetylase 10 (HDAC10) to co-repress *OsHKT2;1*, a plasma membrane-localized Na(+) transporter, and confers salt stress tolerance to rice (Wei et al., 2020). The OsPRR37 protein can interact with Ghd8 and NF-YCs, which form an alternative OsNF-Y heterotrimer to affect Hd1-mediated regulation of *Hd3a* and flowering (Goretti et al., 2017). The distinct role of *OsPRR37* in the ZH11 background indicated that OsPRR37 can associate with different partners to perform different functions (Hu et al., 2021). Moreover, the 35 identified differentially methylated DEGs accounted for only a small proportion of DEGs (**Figure 5A**). These results draw a map of the hierarchical regulation network for OsPRR37 and thus put forward an interesting question about the partners of OsPRR37 with which it regulates the large amount of remaining DEGs. Nevertheless, differentially methylated DEGs are the key candidates to regulate rice growth.

Epigenetic marks that modulate the expression of genes behind the traits of interest have potential applications in crop enhancement (Kakoulidou et al., 2021). With the development of multi-omics technologies and related data processing pipelines (Feng et al., 2021; Iqbal et al., 2021), the hierarchical regulation network of the circadian clock will be gradually parsed and applied to improve rice traits. Recently, the representative role of OsPRR37 in the control of photoperiodic flowering was systematically reviewed (Chen et al., 2021; Zhou et al., 2021c). However, the underlying mechanism of how OsPRR37 regulates its output genes to affect multiple agronomic traits remains unclear. By integrative analysis of WGBS and RNA-seq data, our results revealed that DNA methylation contributes to the regulation of OsPRR37-output genes, which provides an alternative strategy to improve plant growth through epigenetic modulation of OsPRR37-output genes.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The name of the repository and accession number can be found below: GEO, NCBI: GSE192416.

AUTHOR CONTRIBUTIONS

CL designed the study, carried out most of the data analysis, and wrote the manuscript. NL performed the qRT-PCR and determined the starch content. ZL, QS, XP, XX, CD, and ZX performed the data analysis of time-course transcriptomes and tissue-specific microarray data. KS, FY, and ZH provided insightful suggestions on data analysis and writing of manuscript. All authors have read and agreed to the submitted version of the manuscript.

FUNDING

This work was funded by Chongqing Postdoctoral Science Foundation (2018LY10), Chongqing Natural Science Foundation

(cstc2019jcyj-msxmX0274 and cstc2020jcyj-msxmX0746), and Scientific and Technological Research Program of Chongqing Municipal Education Commission (KJQN202100641).

ACKNOWLEDGMENTS

We wish to thank Daichang Yang of the Wuhan University for providing the field area to plant transgenic rice. We also wish to thank Jianming Zeng of the University of Macau and his biotrainee team for providing bioinformatics courses.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.839457/full#supplementary-material

REFERENCES

Aung, M. S., Masuda, H., Nozoye, T., Kobayashi, T., Jeon, J. S., An, G., et al. (2019). Nicotianamine Synthesis by OsNAS3 Is Important for Mitigating Iron Excess Stress in Rice. *Front. Plant Sci.* 10:660. doi: 10.3389/fpls.2019.00660

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

Bu, D., Luo, H., Huo, P., Wang, Z., Zhang, S., He, Z., et al. (2021). KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic Acids Res.* 49, W317–W325. doi: 10.1093/nar/gkab447

Chen, R., Deng, Y., Ding, Y., Guo, J., Qiu, J., Wang, B., et al. (2021). Rice functional genomics: decades' efforts and roads ahead. *Sci. China Life Sci.* 65, 33–92. doi: 10.1007/s11427-021-2024-0

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890. doi: 10.1093/bioinformatics/bty560

Dodd, A. N., Salathia, N., Hall, A., Kevei, E., Toth, R., Nagy, F., et al. (2005). Plant circadian clocks increase photosynthesis, growth, survival, and competitive advantage. *Science* 309, 630–633. doi: 10.1126/science.1115581

Doyle, J. J. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* 19, 11–15. doi: 10.1016/0031-9422(80)85004-7

Farre, E. M., and Liu, T. (2013). The PRR family of transcriptional regulators reflects the complexity and evolution of plant circadian clocks. *Curr. Opin. Plant Biol.* 16, 621–629. doi: 10.1016/j.pbi.2013.06.015

Feng, J. W., Lu, Y., Shao, L., Zhang, J., Li, H., and Chen, L. L. (2021). Phasing analysis of the transcriptome and epigenome in a rice hybrid reveals the inheritance and difference in DNA methylation and allelic transcription regulation. *Plant Commun.* 2:100185. doi: 10.1016/j.xplc.2021.100185

Fornara, F., Parenicova, L., Falasca, G., Pelucchi, N., Masiero, S., Ciannamea, S., et al. (2004). Functional characterization of OsMADS18, a member of the AP1/SQUA subfamily of MADS box genes. *Plant Physiol.* 135, 2207–2219. doi: 10.1104/pp.104.045039

Fujino, K., Yamanouchi, U., Nonoue, Y., Obara, M., and Yano, M. (2019). Switching genetic effects of the flowering time gene Hd1 in LD conditions by Ghd7 and OsPRR37 in rice. *Breed Sci.* 69, 127–132. doi: 10.1270/jsbbs.18060

Gao, H., Jin, M., Zheng, X. M., Chen, J., Yuan, D., Xin, Y., et al. (2014). Days to heading 7, a major quantitative locus determining photoperiod sensitivity and regional adaptation in rice. *Proc. Natl. Acad. Sci. U. S. A.* 111, 16337–16342. doi: 10.1073/pnas.1418204111

Gehring, M. (2019). Epigenetic dynamics during flowering plant reproduction: evidence for reprogramming? *New Phytol.* 224, 91–96. doi: 10.1111/nph.15856

Goretti, D., Martignago, D., Landini, M., Brambilla, V., Gomez-Ariza, J., Gnesutta, N., et al. (2017). Transcriptional and post-transcriptional mechanisms limit heading date 1 (Hd1) function to adapt rice to high latitudes. *PLoS Genet.* 13:e1006530. doi: 10.1371/journal.pgen.1006530

Haydon, M. J., Mielczarek, O., Robertson, F. C., Hubbard, K. E., and Webb, A. A. (2013). Photosynthetic entrainment of the Arabidopsis thaliana circadian clock. *Nature* 502, 689–692. doi: 10.1038/nature12603

Higo, A., Saihara, N., Miura, F., Higashi, Y., Yamada, M., Tamaki, S., et al. (2020). DNA methylation is reconfigured at the onset of reproduction in rice shoot apical meristem. *Nat. Commun.* 11:4079. doi: 10.1038/s41467-020-17963-2

Hu, Y., Zhou, X., Zhang, B., Li, S., Fan, X., Zhao, H., et al. (2021). OsPRR37 alternatively promotes heading date through suppressing the expression of Ghd7 in the Japonica variety zhonghua 11 under natural long-day conditions. *Rice* 14, 20. doi: 10.1186/s12284-021-00464-1

Iqbal, Z., Iqbal, M. S., Khan, M. I. R., and Ansari, M. I. (2021). Toward integrated multi-omics intervention: rice trait improvement and stress management. *Front. Plant Sci.* 12:741419. doi: 10.3389/fpls.2021.741419

Jiao, Y., Wang, Y., Xue, D., Wang, J., Yan, M., Liu, G., et al. (2010). Regulation of OsSPL14 by OsmiR156 defines ideal plant architecture in rice. *Nat. Genet.* 42, 541–544. doi: 10.1038/ng.591

Kakoulidou, I., Avramidou, E. V., Baranek, M., Brunel-Muguet, S., Farrona, S., Johannes, F., et al. (2021). Epigenetics for crop improvement in times of global change. *Biology* 10:766. doi: 10.3390/biology10080766

Kawakatsu, T. (2020). RNA-directed DNA methylation links viral disease and plant architecture in rice. *Mol. Plant* 13, 814–816. doi: 10.1016/j.molp.2020.03.013

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37, 907–915. doi: 10.1038/s41587-019-0201-4

Kim, S., Park, J. S., Lee, J., Lee, K. K., Park, O. S., Choi, H. S., et al. (2021). The DME demethylase regulates sporophyte gene expression, cell proliferation, differentiation, and meristem resurrection. *Proc. Natl. Acad. Sci. U. S. A.* 118:e2026806118. doi: 10.1073/pnas.2026806118

Koo, B. H., Yoo, S. C., Park, J. W., Kwon, C. T., Lee, B. D., An, G., et al. (2013). Natural variation in OsPRR37 regulates heading date and contributes to rice cultivation at a wide range of latitudes. *Mol. Plant* 6, 1877–1888. doi: 10.1093/mp/sst088

Korthauer, K., Chakraborty, S., Benjamini, Y., and Irizarry, R. A. (2019). Detection and accurate false discovery rate control of differentially methylated regions from whole genome bisulfite sequencing. *Biostatistics* 20, 367–383. doi: 10.1093/biostatistics/kxy007

Krueger, F., and Andrews, S. R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 27, 1571–1572. doi: 10.1093/bioinformatics/btr167

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, 1639–1645. doi: 10.1101/gr.092759.109

Law, J. A., and Jacobsen, S. E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* 11, 204–220. doi: 10.1038/nrg2719

Li, C., Li, Y. H., Li, Y., Lu, H., Hong, H., Tian, Y., et al. (2020). A domestication-associated gene GmPRR3b regulates the circadian clock and flowering time in soybean. *Mol. Plant* 13, 745–759. doi: 10.1016/j.molp.2020.01.014

Li, N., Zhang, Y., He, Y., Wang, Y., and Wang, L. (2020). Pseudo response regulators regulate photoperiodic hypocotyl growth by repressing PIF4/5 transcription. *Plant Physiol.* 183, 686–699. doi: 10.1104/pp.19.01599

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Liang, L., Chang, Y., Lu, J., Wu, X., Liu, Q., Zhang, W., et al. (2019). Global methylomic and transcriptomic analyses reveal the broad participation of DNA methylation in daily gene expression regulation of Populus trichocarpa. *Front. Plant Sci.* 10:243. doi: 10.3389/fpls.2019.00243

Liang, L., Zhang, Z., Cheng, N., Liu, H., Song, S., Hu, Y., et al. (2021). The transcriptional repressor OsPRR73 links circadian clock and photoperiod pathway to control heading date in rice. *Plant Cell Environ.* 44, 842–855. doi: 10.1111/pce.13987

Liu, C., Qu, X., Zhou, Y., Song, G., Abiri, N., Xiao, Y., et al. (2018). OsPRR37 confers an expanded regulation of the diurnal rhythms of the transcriptome and photoperiodic flowering pathways in rice. *Plant Cell Environ.* 41, 630–645. doi: 10.1111/pce.13135

Liu, J., Wu, X., Yao, X., Yu, R., Larkin, P. J., and Liu, C. M. (2018). Mutations in the DNA demethylase OsROS1 result in a thickened aleurone and improved nutritional value in rice grains. *Proc. Natl. Acad. Sci. U. S. A.* 115, 11327–11332. doi: 10.1073/pnas.1806304115

Liu, C., Song, G., Zhou, Y., Qu, X., Guo, Z., Liu, Z., et al. (2015). OsPRR37 and Ghd7 are the major genes for general combining ability of DTH, PH and SPP in rice. *Sci. Rep.* 5:12803. doi: 10.1038/srep12803

Liu, T., Liu, H., Zhang, H., and Xing, Y. (2013). Validation and characterization of Ghd7.1, a major quantitative trait locus with pleiotropic effects on spikelets per panicle, plant height, and heading date in rice (Oryza sativa L.). *J. Integr. Plant Biol.* 55, 917–927. doi: 10.1111/jipb.12070

Lopez-Delisle, L., Rabbani, L., Wolff, J., Bhardwaj, V., Backofen, R., Gruning, B., et al. (2021). pyGenomeTracks: reproducible plots for multivariate genomic datasets. *Bioinformatics* 37, 422–423. doi: 10.1093/bioinformatics/btaa692

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

Masuda, K., Yamada, T., Kagawa, Y., and Fukuda, H. (2020). Time lag between light and geat diurnal cycles modulates CIRCADIAN CLOCK ASSOCIATION 1 rhythm and growth in Arabidopsis thaliana. *Front. Plant Sci.* 11:614360. doi: 10.3389/fpls.2020.614360

Miao, C., Wang, Z., Zhang, L., Yao, J., Hua, K., Liu, X., et al. (2019). The grain yield modulator miR156 regulates seed dormancy through the gibberellin pathway in rice. *Nat. Commun.* 10:3822. doi: 10.1038/s41467-019-11830-5

Nakamichi, N. (2020). The transcriptional network in the Arabidopsis circadian clock system. *Genes* 11:1284. doi: 10.3390/genes11111284

Ng, D. W., Miller, M., Yu, H. H., Huang, T. Y., Kim, E. D., Lu, J., et al. (2014). A role for CHH methylation in the parent-of-origin effect on altered circadian rhythms and biomass heterosis in Arabidopsis intraspecific hybrids. *Plant Cell* 26, 2430–2440. doi: 10.1105/tpc.113.115980

Nimmo, H. G., and Laird, J. (2021). Arabidopsis thaliana PRR7 provides circadian input to the CCA1 promoter in shoots but not roots. *Front. Plant Sci.* 12:750367. doi: 10.3389/fpls.2021.750367

Panter, P. E., Muranaka, T., Cuitun-Coronado, D., Graham, C. A., Yochikawa, A., Kudoh, H., et al. (2019). Circadian regulation of the plant transcriptome under natural conditions. *Front. Genet.* 10:1239. doi: 10.3389/fgene.2019.01239

Paszkowski, U., Kroken, S., Roux, C., and Briggs, S. P. (2002). Rice phosphate transporters include an evolutionarily divergent gene specifically activated in arbuscular mycorrhizal symbiosis. *Proc. Natl. Acad. Sci. U. S. A.* 99, 13324–13329. doi: 10.1073/pnas.202474599

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, D1085–D1092. doi: 10.1093/nar/gkz921

Smith, A. M., and Zeeman, S. C. (2006). Quantification of starch in plant tissues. *Nat. Protoc.* 1, 1342–1345. doi: 10.1038/nprot.2006.232

Song, H. R., and Noh, Y. S. (2012). Rhythmic oscillation of histone acetylation and methylation at the Arabidopsis central clock loci. *Mol. Cells* 34, 279–287. doi: 10.1007/s10059-012-0103-5

Sun, H., Zhang, W., Wu, Y., Gao, L., Cui, F., Zhao, C., et al. (2020). The circadian clock gene, TaPRR1, is associated with yield-related traits in wheat (Triticum aestivum L.). *Front. Plant Sci.* 11:285. doi: 10.3389/fpls.2020.00285

Sun, S., Zhu, J., Guo, R., Whelan, J., and Shou, H. (2021). DNA methylation is involved in acclimation to iron-deficiency in rice (Oryza sativa). *Plant J.* 107, 727–739. doi: 10.1111/tpj.15318

Tan, L., Qu, M., Zhu, Y., Peng, C., Wang, J., Gao, D., et al. (2020). ZINC TRANSPORTER5 and ZINC TRANSPORTER9 Function Synergistically in Zinc/Cadmium Uptake. *Plant Physiol.* 183, 1235–1249. doi: 10.1104/pp.19.01569

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

Tian, W., Wang, R., Bo, C., Yu, Y., Zhang, Y., Shin, G. I., et al. (2021). SDC mediates DNA methylation-controlled clock pace by interacting with ZTL in Arabidopsis. *Nucleic Acids Res.* 49, 3764–3780. doi: 10.1093/nar/gkab128

Vera Alvarez, R., Pongor, L. S., Marino-Ramirez, L., and Landsman, D. (2019). TPMCalculator: one-step software to quantify mRNA abundance of genomic features. *Bioinformatics* 35, 1960–1962. doi: 10.1093/bioinformatics/bty896

Wang, F., Han, T., Song, Q., Ye, W., Song, X., Chu, J., et al. (2020). The rice circadian clock regulates tiller growth and panicle development through strigolactone signaling and sugar sensing. *Plant Cell* 32, 3124–3138. doi: 10.1105/tpc.20.00289

Wang, H., Li, Y., Chern, M., Zhu, Y., Zhang, L. L., Lu, J. H., et al. (2021). Suppression of rice miR168 improves yield, flowering time and immunity. *Nat. Plants* 7, 129–136. doi: 10.1038/s41477-021-00852-x

Wang, L., Xie, W., Chen, Y., Tang, W., Yang, J., Ye, R., et al. (2010). A dynamic gene expression atlas covering the entire life cycle of rice. *Plant J.* 61, 752–766. doi: 10.1111/j.1365-313X.2009.04100.x

Wei, H., Wang, X., He, Y., Xu, H., and Wang, L. (2020). Clock component OsPRR73 positively regulates rice salt tolerance by modulating OsHKT2;1-mediated sodium homeostasis. *EMBO J.* 40, e105086. doi: 10.15252/embj.2020105086

Wu, J., Yang, Z., Wang, Y., Zheng, L., Ye, R., Ji, Y., et al. (2015). Viral-inducible Argonaute18 confers broad-spectrum virus resistance in rice by sequestering a host microRNA. *Elife* 4:e05733. doi: 10.7554/eLife.05733

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* 2:100141. doi: 10.1016/j.xinn.2021.100141

Xu, L., Yuan, K., Yuan, M., Meng, X., Chen, M., Wu, J., et al. (2020). Regulation of rice tillering by RNA-directed dna methylation at miniature inverted-repeat transposable elements. *Mol. Plant* 13, 851–863. doi: 10.1016/j.molp.2020.02.009

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

Yang, M., Li, Y., Liu, Z., Tian, J., Liang, L., Qiu, Y., et al. (2020). A high activity zinc transporter OsZIP9 mediates zinc uptake in rice. *Plant J.* 103, 1695–1709. doi: 10.1111/tpj.14855

Yang, S. Y., Gronlund, M., Jakobsen, I., Grotemeyer, M. S., Rentsch, D., Miyao, A., et al. (2012). Nonredundant regulation of rice arbuscular mycorrhizal symbiosis by two members of the phosphate transporter1 gene family. *Plant Cell* 24, 4236–4251. doi: 10.1105/tpc.112.104901

Yao, W., Li, G., Yu, Y., and Ouyang, Y. (2018). funRiceGenes dataset for comprehensive understanding and application of rice functional genes. *Gigascience* 7, 1–9. doi: 10.1093/gigascience/gix119

Yin, X., Liu, X., Xu, B., Lu, P., Dong, T., Yang, D., et al. (2019). OsMADS18, a membrane-bound MADS-box transcription factor, modulates plant architecture and the abscisic acid response in rice. *J. Exp. Bot.* 70, 3895–3909. doi: 10.1093/jxb/erz198

Yu, G., Wang, L. G., and He, Q. Y. (2015). ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* 31, 2382–2383. doi: 10.1093/bioinformatics/btv145

Yuan, L., Yu, Y., Liu, M., Song, Y., Li, H., Sun, J., et al. (2021). BBX19 fine-tunes the circadian rhythm by interacting with PSEUDO-RESPONSE REGULATOR proteins to facilitate their repressive effect on morning-phased clock genes. *Plant Cell* 33, 2602–2617. doi: 10.1093/plcell/koab133

Zeng, D. D., Yang, C. C., Qin, R., Alamin, M., Yue, E. K., Jin, X. L., et al. (2018). A guanine insert in OsBBS1 leads to early leaf senescence and salt stress sensitivity in rice (Oryza sativa L.). *Plant Cell Rep.* 37, 933–946. doi: 10.1007/s00299-018-2280-y

Zhang, C., Wei, Y., Xu, L., Wu, K. C., Yang, L., Shi, C. N., et al. (2020). A Bunyavirus-inducible ubiquitin ligase targets RNA polymerase IV for degradation during viral pathogenesis in rice. *Mol. Plant* 13, 836–850. doi: 10.1016/j.molp.2020.02.010

Zhang, H., Lang, Z., and Zhu, J. K. (2018). Dynamics and function of DNA methylation in plants. *Nat. Rev. Mol. Cell Biol.* 19, 489–506. doi: 10.1038/s41580-018-0016-z

Zhang, X., Sun, J., Cao, X., and Song, X. (2015). Epigenetic mutation of RAV6 affects leaf angle and seed size in rice. *Plant Physiol.* 169, 2118–2128. doi: 10.1104/pp.15.00836

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S. W., Chen, H., et al. (2006). Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. *Cell* 126, 1189–1201. doi: 10.1016/j.cell.2006.08.003

Zheng, S., Li, J., Ma, L., Wang, H., Zhou, H., Ni, E., et al. (2019). OsAGO2 controls ROS production and the initiation of tapetal PCD by epigenetically regulating OsHXK1 expression in rice anthers. *Proc. Natl. Acad. Sci. U. S. A.* 116, 7549–7558. doi: 10.1073/pnas.1817675116

Zheng, S., Ye, C., Lu, J., Liufu, J., Lin, L., Dong, Z., et al. (2021). Improving the rice photosynthetic efficiency and yield by editing OsHXK1 via CRISPR/Cas9 system. *Int. J. Mol. Sci.* 22:9554. doi: 10.3390/ijms22179554

Zhou, J., Zhang, R., Jia, X., Tang, X., Guo, Y., Yang, H., et al. (2021a). CRISPR-Cas9 mediated OsMIR168a knockout reveals its pleiotropy in rice. *Plant Biotechnol. J.* 20, 310–322. doi: 10.1111/pbi.13713

Zhou, S., Li, X., Liu, Q., Zhao, Y., Jiang, W., Wu, A., et al. (2021b). DNA demethylases remodel DNA methylation in rice gametes and zygote and are required for reproduction. *Mol. Plant* 14, 1569–1583. doi: 10.1016/j.molp.2021.06.006

Zhou, S., Zhu, S., Cui, S., Hou, H., Wu, H., Hao, B., et al. (2021c). Transcriptional and post-transcriptional regulation of heading date in rice. *New Phytol.* 230, 943–956. doi: 10.1111/nph.17158

Zhu, Q., Ordiz, M. I., Dabi, T., Beachy, R. N., and Lamb, C. (2002). Rice TATA binding protein interacts functionally with transcription factor IIB and the RF2a bZIP transcriptional activator in an enhanced plant in vitro transcription system. *Plant Cell.* 14, 795–803. doi: 10.1105/tpc.010364

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

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

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 16 May 2022
doi: 10.3389/fpls.2022.857390

Transcriptome-Wide Characterization of Seed Aging in Rice: Identification of Specific Long-Lived mRNAs for Seed Longevity

Bingqian Wang[1†], Songyang Wang[1†], Yuqin Tang[2], Lingli Jiang[1], Wei He[2], Qinlu Lin[2], Feng Yu[1] and Long Wang[1,3*]*

[1] State Key Laboratory of Chemo/Biosensing and Chemometrics, Hunan Province Key Laboratory of Plant Functional Genomics and Developmental Regulation, College of Biology, Hunan University, Changsha, China, [2] National Engineering Laboratory for Rice and By-Product Deep Processing, Central South University of Forestry and Technology, Changsha, China, [3] Longping Agricultural Science and Technology Huangpu Research Institute, Guangzhou, China

Various long-lived mRNAs are stored in seeds, some of which are required for the initial phase of germination and are critical to seed longevity. However, the seed-specific long-lived mRNAs involved in seed longevity remain poorly understood in rice. To identify these mRNAs in seeds, we first performed aging experiment with 14 rice varieties, and categorized them as higher longevity (HL) and lower longevity (LL) rice varieties in conventional rice and hybrid rice, respectively. Second, RNA-seq analysis showed that most genes showed similar tendency of expression changes during natural and artificial aging, suggesting that the effects of these two aging methods on transcription are comparable. In addition, some differentially expressed genes (DEGs) in the HL and LL varieties differed after natural aging. Furthermore, several specific long-lived mRNAs were identified through a comparative analysis of HL and LL varieties after natural aging, and similar sequence features were also identified in the promoter of some specific long-lived mRNAs. Overall, we identified several specific long-lived mRNAs in rice, including gibberellin receptor gene *GID1*, which may be associated with seed longevity.

Keywords: seed longevity, RNA-seq analysis, rice varieties, artificial aging, natural aging

OPEN ACCESS

Edited by:
Yin Li,
Huazhong University of Science and Technology, China

Reviewed by:
Bing Bai,
University of Copenhagen, Denmark
Zhiyong Li,
Southern University of Science and Technology, China
X. Deng,
Chinese Academy of Tropical Agricultural Sciences, China

****Correspondence:***
Feng Yu
feng_yu@hnu.edu.cn
orcid.org/0000-0002-5221-281X
Long Wang
wanglong8591@hnu.edu.cn
orcid.org/0000-0002-3424-8181

[†]These authors share first authorship

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

Received: *18 January 2022*
Accepted: *14 April 2022*
Published: *16 May 2022*

Citation:
Wang B, Wang S, Tang Y, Jiang L, He W, Lin Q, Yu F and Wang L (2022) Transcriptome-Wide Characterization of Seed Aging in Rice: Identification of Specific Long-Lived mRNAs for Seed Longevity. Front. Plant Sci. 13:857390. doi: 10.3389/fpls.2022.857390

INTRODUCTION

The seed is the carrier of biological genetic information and the basis of agricultural production. Seed longevity, the period over which seeds remain viable, is an important agronomical trait that determines its viability, storability, and quality (Zhao et al., 2021). Typically, seed longevity is measured using the final germination percentage and the indices of seedling percentage after aging (Zhao et al., 2021). Generally, seeds with high seed vigor germinate and emerge more quickly, are more resistant to stress and have the potential for high yield and quality in agricultural practice. Seed aging refers to the reduction in seed viability, loss of vitality and irreversible changes that result in the inability to germinate. Aging is a process that occurs along with the prolonged storage of seeds. The degree of seed aging is compounded by improper storage conditions, especially high

temperature and high humidity. During the storage period, a series of harmful events will occur inside the seed, such as cell membrane damage, DNA damage and mutation, and long-lived mRNA degradation. Thus, the reduction in seed longevity is often associated with damage to nucleic acids and proteins.

Seed longevity is determined by genetic and physiological storage potential of the seeds (Qun et al., 2007; Bewley et al., 2012) and by their interaction with environmental factors and events causing deterioration during storage. Several genes controlling seed longevity in rice have been identified. For example, the transcription factor ABSCISIC ACID-INSENSITIVE3 (ABI3) plays a central role in seed longevity (Sano et al., 2016), as the *abi3* mutant is intolerant to desiccation and exhibits rapid viability loss during dry storage. Additionally, the indole-3-acetic acid (IAA)-amido synthetase gene GRETCHEN HAGEN3-2 (OsGH3-2) acts as a negative regulator of seed viability by regulating many genes related to the abscisic acid (ABA) pathway, subsequently regulating the accumulation level of ABA (Yuan et al., 2021). Gibberellin (GA) is another well-known phytohormone that control seed dormancy and germination, in a manner different from ABA. GIBBERELLIN INSENSITIVE DWARF1 (GID1) encodes a soluble GA receptor, plays important role in seed germination (Ge and Steber, 2018). The *Arabidopsis* contains 3 GID1 orthologs, named *AtGID1a*, *AtGID1b,* and *AtGID1c*, while rice contains only a single GID1 (Ueguchi-Tanaka et al., 2005; Nakajima et al., 2006). It has been shown that Gibberellic acid (GA3)-treated seeds or those of the quintuple DELLA mutant (with constitutive GA signaling) had higher artificial aging resistance, indicating that GA might play a positive role in seed longevity (Bueso et al., 2014). In addition, several QTLs controlling seed longevity in rice have been identified, and using 299 indica accessions, it was shown that eight major loci related to sugar metabolism, DNA repair and transcription, reactive oxygen species (ROS) and embryonic/root development were associated with seed longevity (Lee et al., 2019). To date, proteomic analyses revealed that changes in the regulation of posttranslational modifications, protein synthesis, and protein turnover play crucial roles in seed longevity, and that proteins associated with metabolism, energy, and protein synthesis were enriched after the artificial aging of seeds (Zhang et al., 2016).

Dry seeds accumulate various mRNAs, called long-lived mRNAs, that are thought to be translated after the onset of imbibition and to function during the early stage of imbibition (Bai et al., 2020). More than 12,000 different long-lived mRNAs have been identified in *Arabidopsis* dry seeds, and some of them are essential for seed longevity (Nakabayashi et al., 2005). Abscisic acid-responsive elements (ABREs) containing the core motif ACGT were overrepresented in the promoters of highly expressed genes in dry seeds (Nakabayashi et al., 2005). *De novo* protein synthesis during the initial phase of seed germination occurs from long-lived mRNAs stored in mature dry seeds without *de novo* transcription (Kimura and Nambara, 2010), and 17% of long-lived mRNAs that are specifically associated with monosomes are translationally upregulated during seed germination (Bai et al., 2020); thus, the translational capacity of dry seeds is important for seed vigor (Rajjou et al., 2007). High-throughput sequencing aid to identify potential seed longevity-related genes through transcriptome sequencing. For instance, several genes involved in ABA biosynthetic processes and the DNA damage response pathway has been identified through RNA-seq (Qu et al., 2020). However, more seed longevity-related genes need exploration.

Since natural aging too long, the aging process must be artificially accelerated for seed longevity research. The controlled deterioration treatment (CDT) was applied to accelerate seed aging for a short period (Rajjou and Debeaujon, 2008). It has been shown that similar molecular events accompany CDT and natural aging at the proteome level in the model plant *Arabidopsis thaliana* (Rajjou and Debeaujon, 2008). Several other aging methods, such as the artificial aging method (AA, aging at high temperature and high relative humidity) and the elevated partial pressure of oxygen (EPPO) method (Groot et al., 2012; Buijs et al., 2020), have been successfully used for seed aging study. The results of different aging methods are affected by different loci in the genome (Buijs et al., 2020; Fenollosa et al., 2020). At present, artificial aging treatment is widely used by seed companies as a vigor assay for numerous seed species to determine the mechanisms of seed vigor loss during storage (Li et al., 2017; Min et al., 2017). However, it is unknown whether natural and artificial aging are distinct on the transcriptional level.

Here, we selected 14 conventional and hybrid rice varieties and identified them as higher longevity (HL) and lower longevity (LL) varieties. RNA-seq analysis showed that most differentially expressed gene changes after naturing aging were similar to that of after artificial aging, indicating that the effects of these two aging methods on the transcription level are similar. Lastly, we identified several specific long-lived mRNAs through a comparative analysis of DEGs in HL and LL varieties after aging.

MATERIALS AND METHODS

Seed Material and Growth Conditions

Seven conventional rice varieties and seven hybrid rice varieties were used for the follow-up experiments (**Supplementary Table 1**). Conventional varieties (YZX, XW13, YC, HM, YH988, XEH, NX32) were purchased from Zhangjiajie Farm (Hunan Province, China), and these seeds were planted at Changsha Observation and Research Station for Agriculture Ecosystems, Chinese Academy of Sciences (Xiangfeng Village, Jinjing Town, Changsha) under the same fertilization and management conditions, harvested in September and stored at −20°C for later analysis. The hybrid varieties LLY1353, LLYHZ, LLY1988, SLY5814, JLY1212, HR2, and LLY534 were planted in the same field and were purchased from Hunan Yahua Seed Industry Co., Ltd.

Determination of the Initial Water Content in Rice

Prior to starting the aging tests, all seeds were dried under a constant weight with initial moisture content, the initial moisture content of the rice seeds was measured by a halogen moisture analyzer. Seeds with a moisture content higher than 15% were dried at a constant temperature of 30°C, and the water content was measured every 12 h until the moisture content was less than

15%. When the moisture content of all varieties dropped below 15% and was basically the same, drying was stopped. Seeds were then used for the aging experiment.

Natural and Artificial Aging Treatment

The natural aging treatment was performed as follows: approximately 100 g of rice seeds of each variety that had been dried to a consistent moisture content (approximately 14%) was used, and the seeds were stored at room temperature (5–33°C) and 60–80% relative humidity in a laboratory in Changsha for 1 year.

Artificial accelerated aging treatment was performed as described by Qu et al. (2020). One hundred g seeds were wrapped in nylon bags, with 6 nylon bags for each variety, and marked as artificial aging for 0, 10, 15, 20, 25, and 30 d. The seeds in each bag were evenly placed in an artificial climate chamber (42°C, and humidity 87%) for 10–30 days.

Germination Rate Determination and ID_{50}

For each variety, approximately 10 g seeds treated with aging for different days were used for the germination experiment. The seeds were immersed in a 400-fold diluted "84" solution for 10 min and then washed with distilled water to remove floating seeds. Each sample was set three biological repetitions, 100 seeds for each repetition. These seeds were placed in a petri dish impregnated with moist filter paper. After that, the seeds were placed in an artificial climate chamber at 30°C, and water evaporation was observed every day and water was replenished if needed. After 8 days of germination, the number of germinated seeds was counted and recorded. ID_{50} refers to the time required for the seed germination rate to drop to half of the initial germination rate.

Conductivity Measurement

The rice seeds were shelled with a small shelling machine, and 25 health rice grains were selected. After being rinsed three times with distilled water, the samples were dried with filter paper. The rice grains were placed in a 50 mL beaker, and then 20 mL of distilled water was added and soaked for 12 h at 25°C, resulting in three blank controls. Measurements were carried out using a DDS-11A digital display conductivity meter. First, the electrode was placed in distilled water for calibration before measurement, and then the conductivity values of the sample (B) and the blank control (A) were measured. The conductivity of the sample was calculated as follows: conductivity = value B − value A.

RNA-Seq

Seeds of four varieties (LLY534, JLY1212, YZX, and NX32) that were subjected to 10 days of artificial aging or 1 year of natural aging, and those from untreated controls, were collected and immediately treated with liquid nitrogen on ice using a small-scale gluten washing machine and finally stored on dry ice. Each treatment was set two biological replicates and all samples were sent to Hangzhou Lianchuan Biotechnology Co., Ltd. for RNA-seq.

RNA-seq libraries were constructed and paired-end sequenced by Hangzhou Lianchuan Biotechnology Co., Ltd. RNA-seq analysis was performed according to Qu et al. (2020). Briefly, sequenced reads were screened, and quality-controlled sequences were mapped using HISAT2 v2.1.1 (Pertea et al., 2016). Transcript splicing and merging were conducted with StringTie 1.3.0. Normalized expression values were calculated with Ballgown. We defined genes as differentially expressed when they had a $p < 0.05$ and $|\log_2 FC| > 1$. The sequencing data reported in this paper are summarized in **Supplementary Table 2** and have been deposited in the GSA database (Genome Sequence Archive in the BIG Data Center, Chinese Academy of Sciences; PRJCA006248) (Members, 2018).

Bioinformatics Analysis

For Gene Ontology (GO) enrichment analysis, GENEONTOLOGY[1] was used to assess the detected DEGs according to Biological Process, Molecular Function, and Cellular Component ontologies. TBtools and Venny (version 2.1.0) were used for some gene screening work (Oliveros, 2007–2015; Chen et al., 2020), and TBtools and R software (version 3.5.1) were used for graphing.

Motif Analysis

The sequences of rice were extracted from the Rice Genome Annotation Project,[2] and TBtools (GXF sequences extract function) was used to extract the 5′UTR, 3′UTR and promoter sequences of each gene. DNA motif analyses were performed using the MEME suite (Bailey and Elkan, 1994), the FIMO was used for identified motif. Firstly, motif was entered in the input motif box. The 5′UTR or promoter sequences were entered in the input the sequences box. Advanced options were set $p < 1.0E\text{-}4$ and start search. Then frequencies of the background genes (DEGs in NX32-natural aging vs. NX32-0d) were also calculated.

The MEME was used to identified the enriched motif in 5′UTR, 3′UTR and promoter sequences. Briefly, the classical mode was selected for motif discovery. Sequences were uploaded into the primary sequence box. Motif width was set to 6–9 bp.

Coexpression Regulatory Network

The network reconstruction was performed using the STRING application in Cytoscape (Shannon et al., 2003). Pearson's correlation coefficient between AP2 transcription factor and targeted gene of > 0.7 (positive regulation) or < 0.7 (negative regulation) were used as a threshold and visualized using Cytoscape.

RT-qPCR Analysis

For the mRNA expression analyses, total RNA was extracted from rice seeds using Trizol (Takara 9109, Japan). cDNA was synthesized by using Maxima H Minus First Strand cDNA Synthesis Kit (Thermo Fisher Scientific K1682, United States) following the manufacturer's protocol. qPCR was performed using Bio-Rad CFX96 with SYBR Premix Ex Taq II (Innovagene

[1] http://geneontology.org/

[2] http://rice.uga.edu/

SQ101-01, China). The primers used for the qPCR analysis are listed in **Supplementary Table 4**, and *OsACTIN* was used as an internal reference. The cDNAs were amplified following denaturation using 42-cycle programs (95°C, 15 s; 60°C, 20 s per cycle).

Statistics

Significant differences in the data were analyzed by Student's *t*-test or by multivariate comparison (one-way ANOVA) using SPSS (version 17.0) software. The significant differences of the changes during the aging between HL and LL were analyzed by multivariate comparison (two-way ANOVA).

RESULTS

Classification of Rice Varieties by Seed Longevity After Aging

To obtain rice varieties with higher longevity (HL) and lower longevity (LL), we selected 14 rice varieties, including 7 conventional and 7 hybrid rice varieties (**Supplementary Table 1**). Then, we carried out artificial aging experiment and assessed seed longevity with a germination assay. Artificial aging led to a rapid decline in the germination rate of all rice varieties. Prior to aging, the germination rates of the YC and HM seeds were 39.3 and 82.1%, respectively, which were lower than those of other rice varieties (**Figure 1A**). The germination rates of YH998 and NX32 were significantly higher than those of the other varieties after aging, while the germination rate of YZX, XEH, and YC were relatively low (**Figure 1A**). In terms of the germination rate of hybrid rice varieties after aging, LLY534 and LLYHZ had higher germination rates, while JLY1212, SLY5814, and HR2 had lower germination rates (**Figure 1B**). Given that different rice varieties have different initial germination rate before aging, it is not accurate to use only the germination rate of seeds to evaluate seed longevity. The half inhibitory time (ID_{50}) refers to the time that the seed germination rate is reduced to half of the non-aged germination rate. The larger the ID_{50} value is, the slower the seed germination rate decreases with aging, which reflects the higher longevity of the seeds. According to the results, YH998 did not reach half maximal inhibition even after 30 days of artificial aging, and NX32 had the highest ID_{50} (25.3 days). Since YH988 is a red rice that is rich in anthocyanins, which might have a role in anti-oxidation (Zhu, 2018), we chose NX32 as the HL seed variety. The ID_{50} values of YZX and YC were 9.8 and 5.2 days, respectively (**Figure 1C**). However, the initial germination rate of YC was much lower than that of the other varieties; therefore, we chose YZX as the LL seed variety among the conventional rice varieties. In hybrid rice, the ID_{50} value of LLY534 was the highest (27.2 days; **Figure 1D**), while the ID_{50} values of JLY1212, SLY5814, and HR2 were 16.0, 13.6, and 12.3 days, respectively (**Figure 1D**). In addition, the initial germination rate of SLY5814 was lower than that of the other varieties. Although the germination rate of HR2 were similar to JLY1212 after aging, the fatty acid content and eating quality of JLY1212 were worse after aging (data not shown), and JLY1212 having a wider planting area in China. Thus, LLY534 and JLY1212 were chosen for further research as the HL and LL seed varieties among the hybrid rice varieties. In summary, we obtained rice varieties with higher or lower longevity in both conventional rice and hybrid rice.

Comparison of the Effects of Natural and Artificial Aging on Seed Longevity

Aging is a natural process. A major drawback of natural aging is that it takes a long time, often approximately 1–2 years. Artificial aging, also known as the accelerated aging of seeds, costs a shorter time span, inducing the desired phenotypic changes in seeds (Hay et al., 2019). However, the effect of these two aging methods on the germination rate is still unclear in rice. To further compare the germination rates of NX32, YZX, LLY534, and JLY1212 under natural aging and artificial aging, we chose another batch of seeds for an additional experiment. NX32 had the highest seed longevity, and YZX had the lowest seed longevity (**Figures 2A,B**). Concerning the hybrid rice varieties, LLY534 had higher seed longevity, and its germination rate remained at approximately 48%, even after 30 days of artificial aging. Moreover, JLY1212 had lower seed longevity, and its seed vigor decreased rapidly compared with LLY534 after artificial aging (**Figures 2A,B**). In addition, the germination rates of NX32, YZX, LLY534, and JLY1212 after 1 year of natural aging were 94.1, 82.3, 95.6, and 64.3%, respectively (**Figure 2C**). We analyzed the correlation between the germination rate of seeds after 1 year of natural aging and 10 days of artificial aging and found that the correlation coefficient was high (Pearson's $R = 0.91$; **Figure 2D**), suggesting that the effect of artificial aging treatment for 10 days was similar to the effect of 1 year of natural aging.

The cell membrane of rice seeds is often damaged during aging, and cytosolic solutes can flow into intercellular spaces, leading to an increase in the conductivity of the seed soaking solution (Panobianco et al., 2007). We then tested the electrical conductivity to evaluate the vigor of the seeds. Compared with NX32 rice seeds, those of YZX had a higher electrical conductivity increase after the artificial aging treatment (**Figure 2E**), and the electrical conductivity of JLY1212 was higher than that of LLY534 before and after aging (**Figure 2E**). These data indicated that electrical conductivity could be used as an indicator for evaluating seed longevity and that aging treatment might had a greater impact on the membrane integrity of LL rice varieties than that of HL rice varieties.

Transcriptomic Analysis of Rice Varieties After Natural and Artificial Aging

The germination rate of rice seeds after artificial aging for 10 days was similar to that of seeds after natural aging for 1 year, suggesting that the effect of an appropriate artificial aging time could mimic the effect of natural aging for 1 year. To investigate the difference between natural and artificial aging at the transcriptional level, RNA-seq experiments were performed for rice seeds with 1-year natural aging or 10-days artificial aging. Regarding conventional rice varieties, NX32 treated with natural aging had 607 differentially expressed genes (DEGs) compared with the mock treatment (stored at −20°C for 1 year),

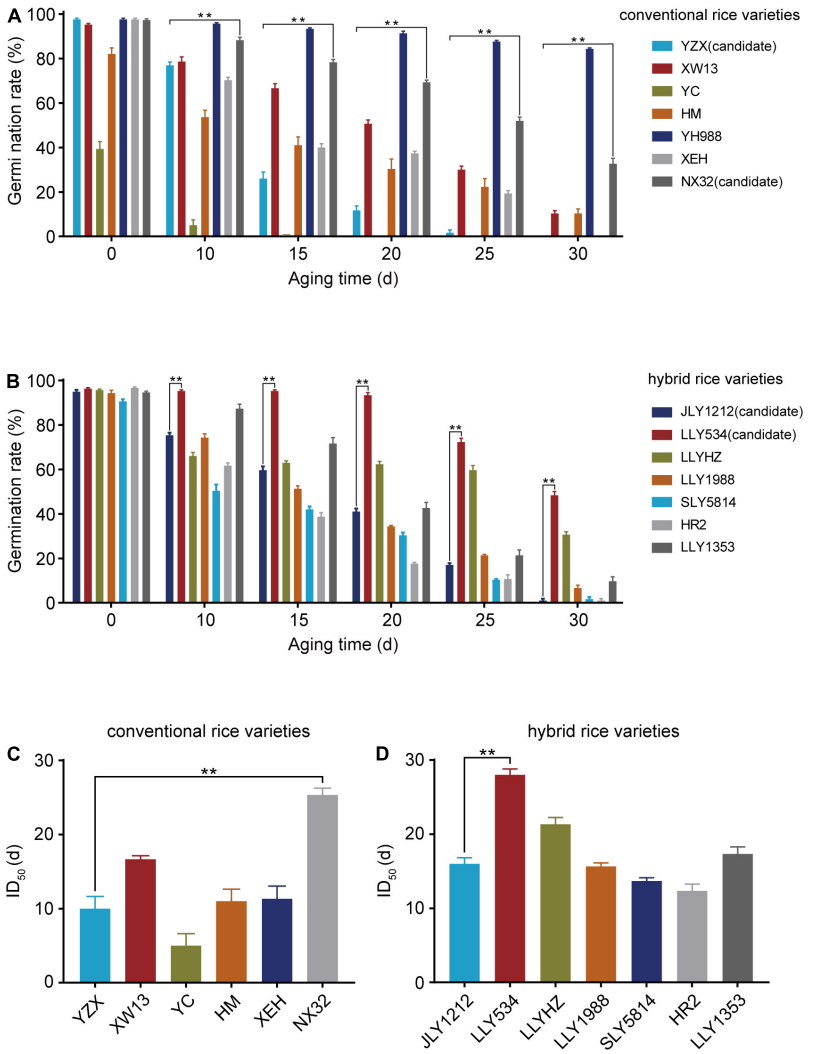


FIGURE 1 | Germination rate and ID_{50} of conventional and hybrid rice seeds. **(A,B)** Germination rates of seven conventional **(A)** and seven hybrid rice varieties **(B)** after artificial aging treatment. The germination rates were recorded after the seeds germinated for 8 days. The experiments were repeated three times, and the error bars represent the SDs of three biological replicates (**$P < 0.01$, one-way ANOVA with Tukey's test). **(C,D)** Half maximal inhibitory days (ID_{50}) of six conventional **(C)** and seven hybrid varieties **(D)**. YH988 is not shown because it did not reach ID_{50} after 30 days artificial aging. Data are the means ± SDs based on three biological replicates (**$P < 0.01$, one-way ANOVA with Tukey's test).

of which 307 were upregulated and 300 were downregulated. In addition, 371 upregulated genes and 327 downregulated genes were identified in NX32 treated with 10-d artificial aging (**Figures 3A,B** and **Supplementary Table 2**; $|\log_2 FC| \geq 1$; $p < 0.05$). For the YZX rice variety treated with natural aging, there were 600 upregulated genes and 254 downregulated genes. For YZX treated with 10-d artificial aging, there were 254 upregulated genes and 277 downregulated genes compared with the mock treatment (**Figures 3A,B** and **Supplementary Table 2**). In hybrid seed varieties, 498 upregulated genes and 183 downregulated genes were identified in LLY534 treated with natural aging, and 447 upregulated genes and 345 downregulated genes were identified in LLY534 treated with 10-d artificial aging (**Figures 3A,B** and **Supplementary Table 2**). In addition, 380 upregulated genes and 317 downregulated genes were detected in JLY1212 treated with natural aging, and 581 upregulated genes and 433 downregulated genes were detected in JLY1212 treated with 10-d artificial aging (**Figures 3A,B** and **Supplementary Table 2**). Heatmap analysis indicated that most gene expression changes ($p < 0.05$ for artificial aging or natural aging) in NX32, YZX, and LLY534 in natural aging and artificial aging were correlated and changed in the same direction (**Figures 3C–E**). The overlapping gene changes ($p < 0.05$ for artificial aging or natural aging) between artificial aging and natural aging were in the same direction, and the values were moderately consistent in NX32 ($r = 0.53$, $p < 0.05$; **Figure 3F**), YZX ($r = 0.49$, $p < 0.05$; **Figure 3G**), and LLY534 ($r = 0.47$, $p < 0.05$; **Figure 3H**).

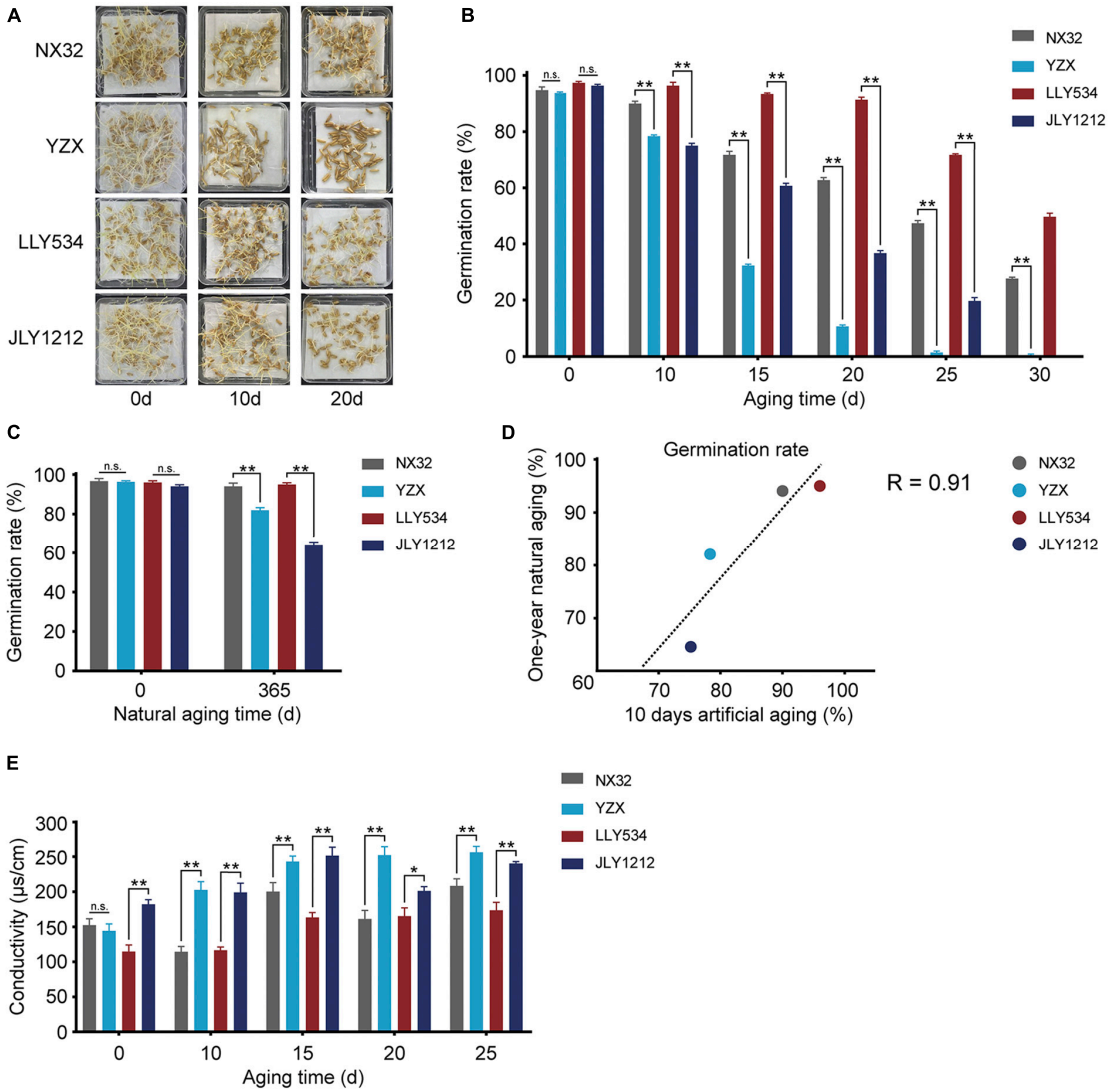


FIGURE 2 | Germination rate and conductivity of seeds of the four rice varieties. **(A,B)** Germination rates of two conventional rice varieties (HL rice variety NX32 and LL rice variety YZX) and two hybrid rice varieties (HL rice variety LLY534 and LL rice JLY1212) under artificial aging conditions (0, 10, 15, 20, 25, and 30 days) (n.s., not significant; $^{*}P < 0.05$, $^{**}P < 0.01$, one-way ANOVA with Tukey's test). **(C)** Germination rates of two conventional varieties (HL rice variety NX32 and LL rice variety YZX) and two hybrid varieties (HL rice variety LLY534 and LL rice variety JLY1212) under natural aging for 1 year ($^{*}P < 0.05$, $^{**}P < 0.01$). **(D)** Correlation analysis between natural aging (1 year) and artificial aging (10 d) of two conventional varieties (HL rice variety NX32 and LL rice variety YZX) and two hybrid varieties (HL rice variety LLY534 and LL rice variety JLY1212) (Pearson's $R = 0.91$). **(E)** Seed conductivity of two conventional varieties (HL rice variety NX32 and LL rice variety YZX) and two hybrid varieties (HL rice variety LLY534 and LL rice variety JLY1212) under artificial aging conditions (0, 10, 15, 20, 25, and 30 d) (n.s., not significant; $^{*}P < 0.05$, $^{**}P < 0.01$, one-way ANOVA with Tukey's test).

These results suggested that natural and artificial aging showed a similar effect on the transcription in rice seeds.

Comparison of mRNA Expression Levels in Higher Longevity and Lower Longevity Rice Varieties After Natural or Artificial Aging

To test whether there is a difference in the expression of long-lived mRNAs between HL and LL varieties, we made a Venn diagram for the long-lived mRNA of these varieties under aging conditions. The results showed that the number of overlapping genes was relatively small across HL and LL varieties under both artificial and natural aging conditions. There were only 39 overlapping genes in NX32 and YZX under natural aging conditions (**Figure 4A**), 75 overlapping genes in LLY534 and JLY1212 under natural aging conditions (**Figure 4B**), 18 overlapping genes in NX32 and YZX under artificial aging conditions (**Figure 4C**) and 30 overlapping genes in LLY534 and JLY1212 under artificial aging conditions (**Figure 4D**). The overall expression trend of HL and LL rice varieties was determined based on the heatmap, which showed that some

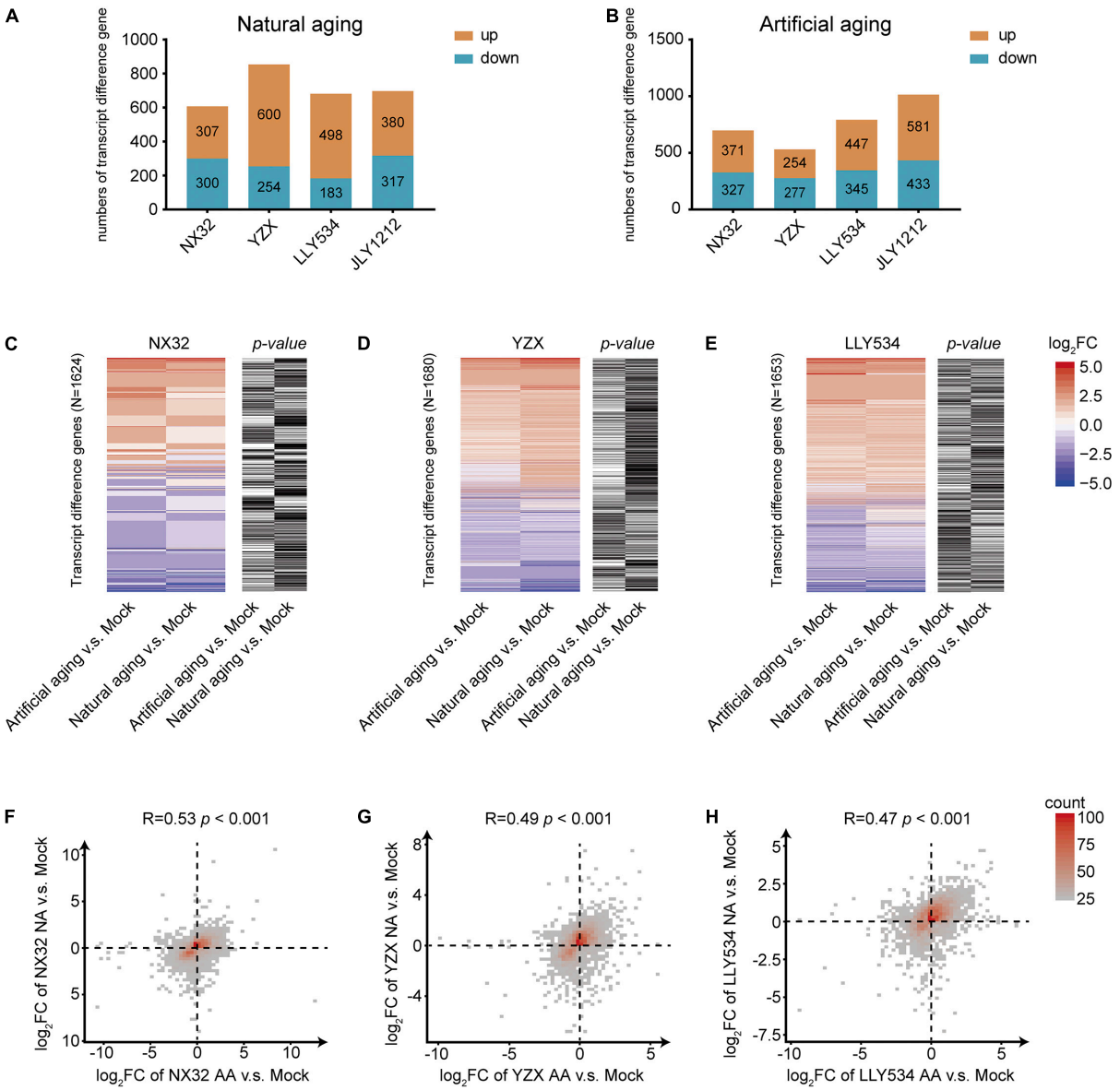


FIGURE 3 | Comparison of gene expression between artificial and natural aging of HL and LL varieties. **(A,B)** Number of differentially expressed transcripts ($|\log_2 FC| > 1$ and $p < 0.05$) in the four rice varieties after natural aging **(A)** and artificial aging **(B)** compared with mock conditions. **(C–E)** Heatmap of changes in gene expression in NX32 **(C)**, YZX **(D)** and LLY534 **(E)** after artificial aging and natural aging compared with mock treatment. Blue–red represents mRNA expression levels, where red represents higher mRNA expression levels, and blue represents lower mRNA expression levels. Black–white represents the *p*-value of the expression level of the transcripts. If the *p*-value is less than 0.05, it is black; otherwise, it is white. **(F–H)** Density figures showing the difference in NX32 **(F)**, YZX **(G)** and LLY534 **(H)** transcript changes ($|\log_2 FC| > 1$ or $p < 0.05$) between artificial aging and natural aging. The x- and y-axes represent changes in the transcription level under artificial aging and natural aging, respectively. Spearman correlation was used for correlation analysis. NA represents natural aging, AA represents artificial aging.

of the genes in NX32 and YZX were different under natural aging conditions ($|\log_2 FC| \geq 1$ for NX32 or YZX; $p < 0.05$ for NX32 or YZX; **Figure 4E**), while the same tendency was found in the comparison between LLY534 and JLY1212 with natural aging (**Figure 4F**) and NX32 and YZX with artificial aging (**Figure 4G**). In particular, the degree of mRNA changes is the most obvious between JLY1212 vs. Mock and LLY534 vs. Mock after artificial aging (**Figure 4H**), which is consistent with the lowest germination rate of JLY1212 after aging. These results indicated that there are certain differences in the transcription levels between HL and LL rice varieties after aging.

Comparison of Gene Ontology Terms in Higher Longevity and Lower Longevity Rice Varieties After Natural Aging

To further analyze the difference in biological pathways between HL and LL rice varieties, we compared the Gene Ontology

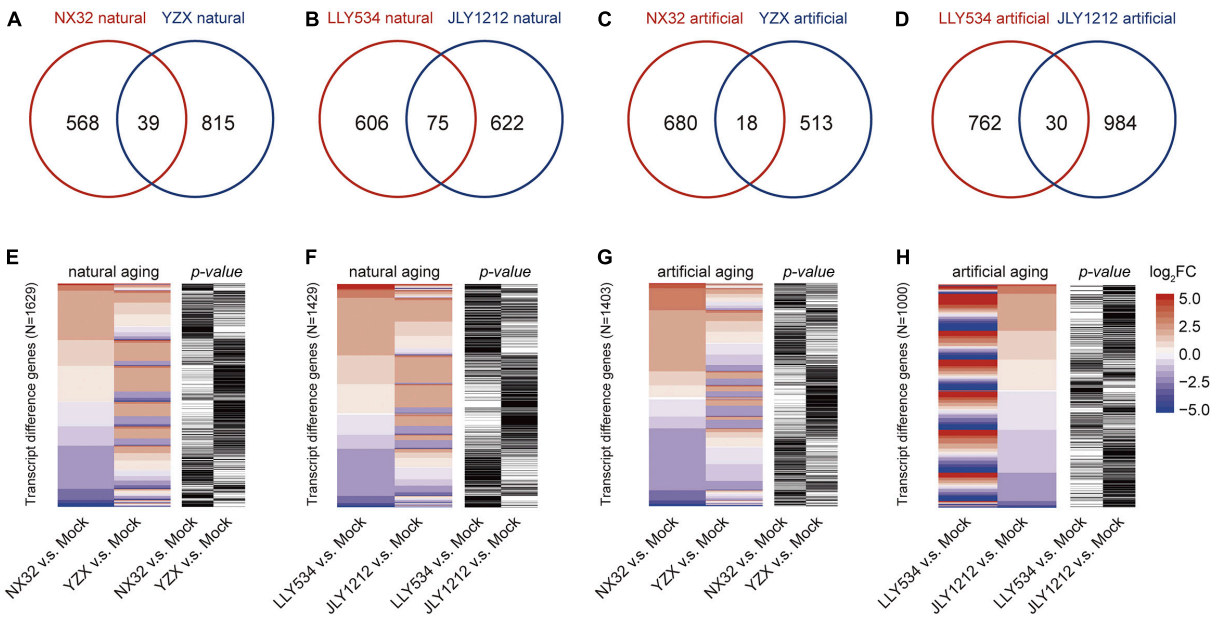


FIGURE 4 | Comparison of significantly differentially expressed genes between four rice varieties with natural and artificial aging. **(A,B)** Venn diagram depicting the overlap of significant transcript differences ($|\log_2 FC| > 1$ and $p < 0.05$) between NX32 and YZX **(A)** and between LLY534 and JLY1212 **(B)** under natural aging conditions. **(C,D)** Venn diagram depicting the overlap of significant transcript differences ($|\log_2 FC| > 1$ and $p < 0.05$) between NX32 and YZX **(C)** and between LLY534 and JLY1212 **(D)** under artificial aging conditions. **(E,F)** Heatmap of differentially expressed genes in conventional varieties NX32 and YZX **(E)** and in hybrid varieties JLY1212 and LLY534 **(F)** when comparing natural aging with mock aging. Blue–red represents mRNA expression levels, and black–white represents the *p*-value of the expression level of the transcripts. If the *p*-value is less than 0.05, it is black; otherwise, it is white. **(G,H)** Heatmap of differentially expressed genes in conventional varieties NX32 and YZX **(G)** and in hybrid varieties JLY1212 and LLY534 **(H)** when comparing artificial aging with mock aging. Blue–red represents mRNA expression levels, and black–white represents the *p*-value of the expression level of the transcripts. If the *p*-value is less than 0.05, it is black; otherwise, it is white.

(GO) terms for DEGs in HL and LL rice varieties under different aging conditions. In conventional rice varieties, GO analysis showed that DEGs in NX32 after natural aging were involved in lipid storage, seed oil body biogenesis, negative regulation of the seed dormancy process, release of seeds from dormancy and positive regulation of seed germination. In addition, GO terms related to stress hormones were also enriched, such as response to positive regulation of the gibberellic acid-mediated signaling pathway (**Figure 5A**). Moreover, DEGs in YZX after natural aging functioned in seed maturation, cellular response to abscisic acid stimulus, response to abscisic acid, seed germination, cellular water homeostasis and response to desiccation (**Figure 5B**), especially the enrichment of seed maturation and cellular response to abscisic acid stimulus. The DEGs of the two conventional rice varieties after aging were mostly related to the processes of seed vigor, dormancy, and germination. Besides, there were also certain differences, the most enriched GO terms of the YZX variety were seed maturation and cellular response to abscisic acid stimulus (**Figure 5B**), while lipid storage and seed oil body biogenesis was enriched in NX32 (**Figure 5A**). In hybrid rice varieties, GO term analysis showed that DEGs in LLY534 after natural aging for 1 year were involved in translation, ribosome biogenesis, response to abscisic acid and seed germination (**Figure 5C**). However, DEGs in JLY1212 after natural aging for 1 year functioned in cytoplasmic translation, regulation of seed dormancy process, lipid storage and negative regulation of gibberellic acid mediated signaling pathway (**Figure 5D**). Similarly, the main signaling pathways enriched in LLY534 and JLY1212 also showed certain differences, which also coincided with the greater difference between the DEGs in the HL and LL varieties. In summary, the main enriched biological pathways of HL and LL rice varieties after 1 year of aging have certain differences, which may be one of the reasons for the difference in seed longevity.

Analysis of the Specific Long-Lived RNAs for Seed Longevity

Previously, it has been reported that the stability of embryonic RNAs required for germination is related to seed longevity (Saighani et al., 2021). These long-lived mRNAs play important roles in the process of protein synthesis during the initial phase of seed germination. Because most transcripts were degraded during aging, we selected transcripts that were down regulated in HL varieties but had a slower degradation rate than that of LL varieties [$\log_2 FC$ HL < 0 and $\log_2 FC$ LL < 0 and ($\log_2 FC$ HL-$\log_2 FC$ LL) > 0] as the long-lived mRNAs ($p < 0.05$ for HL or LL) (**Figures 6A,B**). By two-way ANOVA, we screened out these special long-lived mRNAs that degrade significantly slower ($p < 0.05$ for the changes during the aging between HL and LL) in HL varieties than in LL varieties. In conventional rice, 174 long-lived mRNAs were identified in NX32 v.s. YZX

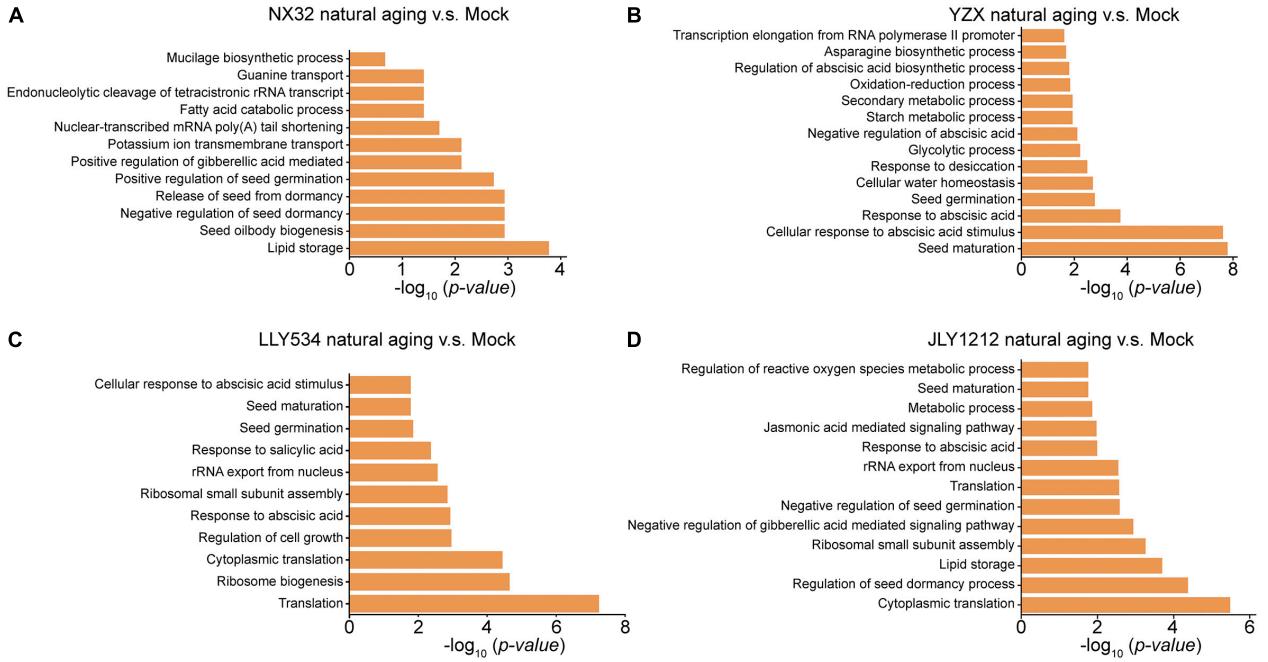


FIGURE 5 | Gene Ontology (GO) enrichment of genes with significant expression changes after natural aging treatment compared with mock treatment in four rice varieties. **(A–D)** GO terms of NX32 **(A)**, YZX **(B)**, LLY534 **(C)** and JLY1212 **(D)** after natural aging. The y-axis represents the functional subcategory, and the x-axis represents the value of the $-\log_{10}$(*p*-value), which indicates the significance level.

after natural aging (**Figure 6B** and **Supplementary Table 3**). In hybrid rice, 305 long-lived mRNAs were identified in LLY534 v.s. JLY1212 after natural aging (**Figure 6B** and **Supplementary Table 3**). The degradation rate of these long-lived mRNAs is slower after aging in HL varieties, and the degradation rate of these long-lived mRNAs is more rapid after aging in LL varieties. To identify more reliable long-lived mRNAs that participate in the regulation of seed vigor in both conventional rice and hybrid rice, we used Venn analysis to identify the overlapping genes in the NX32 v.s. YZX and LLY534 v.s. JLY1212 comparisons (**Supplementary Figures 1A,B**) and 14 overlapping genes were identified under natural aging conditions (**Figures 6C,D**, **Supplementary Figure 2**, and **Supplementary Table 5**). Of them, GID1 (LOC_Os05g33730) is gibberellin receptor, indicating that the GA pathway may be related to seed vitality. In addition, LOC_Os04g33460, a starch branching enzyme IIa (OsBEIIa), was also identified. Further, we analyzed the expression of *GID1* and *OsBEIIa* in HL and LL varieties seeds with or without natural aging using qPCR. The results were consistent with the RNA-seq data (**Figure 6E**), indicating the reliability of the RNA-seq data. In addition, in conventional rice, 168 long-lived mRNAs were identified in NX32 v.s. YZX after artificial aging (**Supplementary Figure 1B** and **Supplementary Table 3**). In hybrid rice, 210 long-lived mRNAs were identified in LLY534 v.s. JLY1212 after artificial aging (**Supplementary Figure 1B** and **Supplementary Table 3**). One overlapping genes (LOC_Os02g10180) was identified in the comparison of NX32 vs. YZX and LLY534 v.s. JLY1212 after artificial aging. And only a few genes overlapped between artificial aging and natural aging in NX32 v.s. YZX or LLY534 v.s. JLY1212 comparisons (**Supplementary Figure 1A**). Since the number of overlapping long-lived mRNA identified under naturing aging is more abundant, we used them in the follow-up analysis.

It has been suggested that the mRNA stored in the mature seed is related to the ribonucleic acid protein complex, indicating that they are translated during seed germination. It has identified a conserved motif, GAAGAAGAA, which is significantly enriched at the 5′UTR and present at low levels in general seed ribosome-associated transcripts (Bueso et al., 2013). However, we did not find this motif enriched in the 14 overlapping long-lived mRNA. We analyzed whether these 14 overlapping long-lived mRNAs have similar sequence features in the promoter, 5′UTR or 3′UTR. It showed that three repeats of the sequence GGCGGCGGC was enriched in the promoter (p = 1.3E-3, percentage = 83.3%, background percentage = 51.3%; **Supplementary Figure 3** and **Supplementary Table 6**). In addition, this *cis*-element was recognized by the AP2/EREBP transcription factors family (Castro-Mondragon et al., 2022). The AP2/DREBP transcription factor family plays a crucial role in seed development, seed storage metabolism and seed longevity (Okamuro et al., 1997; Cernac and Benning, 2004; Pereira Lima et al., 2017). The mRNA expression levels of AP2/EREBP transcription factor members in naturing aging were analyzed and their DEGs data were used to build a possible transcriptional regulation pathway on rice longevity regulation mediated by AP2/EREBPs (**Supplementary Figures 4A–D**). In summary, we identified 14 specific long-lived mRNAs that might be important to seed longevity. The

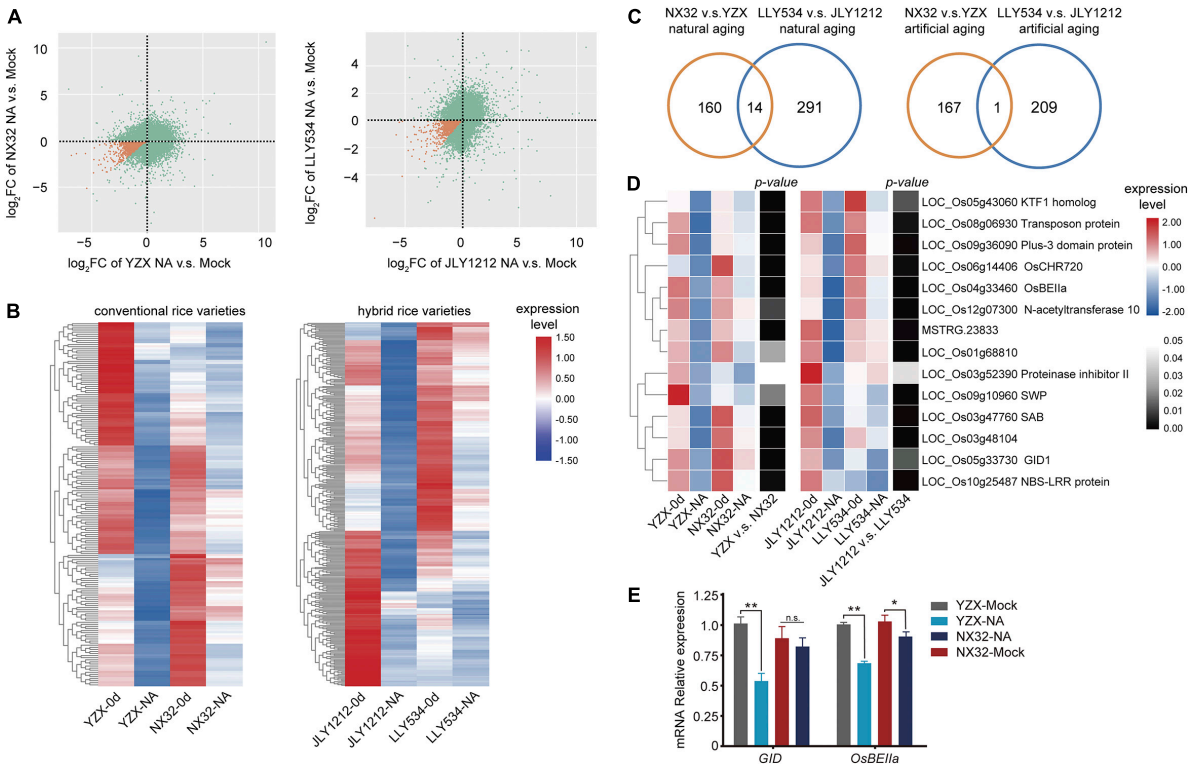


FIGURE 6 | Heatmap analysis of long-lived mRNAs in HL and LL varieties. **(A)** Figure showing the log_2FC in NX32, YZX, LLY534, and JLY1212 when comparing natural aging with mock treatment ($p < 0.05$). The orange dots represent the specific long-lived mRNA. NA represent natural aging. **(B)** Heatmap of specific long-lived mRNAs that degrade significantly slower in HL varieties NX32 and LLY534 than in LL varieties YZX and JLY1212 when comparing naturing aging with mock treatment ($p < 0.05$). **(C)** Venn diagram depicting the overlap of specific long-lived mRNAs that degrade significantly slower in HL varieties than in LL varieties ($p < 0.05$) between NX32 vs. YZX and LLY534 vs. JLY1212 under natural and artificial aging conditions. **(D)** Heatmap of overlapping specific long-lived mRNAs that degrade significantly slower in HL varieties than in LL varieties ($p < 0.05$) between conventional rice varieties (NX32 and YZX) and hybrid rice varieties (LLY534 and JLY1212) under natural aging conditions. NA represents natural aging. Black–white represents the *p*-value of the changes during the aging between HL and LL. **(E)** qRT-PCR analysis of *GID1* and *OsBEIIa* expression in NX32 and YZX with or without natural aging. *OsACTIN* was used as the internal control. The data are presented as the mean ± SD ($n = 3$) (n.s., not significant; $^*P < 0.05$, $^{**}P < 0.01$, one-way ANOVA with Tukey's test).

gibberellin receptor gene GID1 suggested that is GA pathway may be involved in seed vigor.

DISCUSSION

Long-lived mRNAs are very important for seed vigor, and their degradation has been detected alongside viability loss in seeds (Fleming et al., 2018). It has been reported that RNA is more vulnerable to oxidation by ROS than DNA due to its single-strandedness (Kong and Lin, 2010), the oxidation of mRNA is not random but selective (Bazin et al., 2011), and damaged mRNA cannot be translated, which will lead to a loss of seed longevity. Dry seeds often serve as the final point of seed development or the initial step in the seed germination series during transcriptomic analysis (Bai et al., 2017; Pereira Lima et al., 2017). It lacks of study on the changes in the transcriptome during seed storage and the identification of specific mRNAs associated with longevity. In this study, we firstly identified 2 HL rice varieties and 2 LL rice varieties by screening 14 rice varieties (7 conventional and 7 hybrid varieties) with artificial aging, and analyzed the effect of artificial and natural aging on transcriptional events. We found that most gene expression changes in the HL and LL varieties under natural and artificial aging were correlated, indicating that artificial and natural aging have similar effects on transcription events. In addition, our results suggested that the degradation of some transcripts occurred specifically during aging, which is consistent with the highly selective nature of RNA oxidation, as some mRNAs are more susceptible to oxidative damage or targeted oxidation (Shan et al., 2003; Chang et al., 2008; Bazin et al., 2011). However, this result differs from a previous result showing that transcripts were degraded non-specifically (Fleming et al., 2018). Previous studies have shown that artificial aging (CDT method) and natural aging have similar effects at the level of protein abundance changes (Rajjou et al., 2008). However, the similarity between CDT and artificial aging during seed longevity is controversial (Nguyen et al., 2012; Buijs et al., 2018, 2020), as different QTLs are involved in seed longevity depending on the seed aging protocol used (Nagel et al., 2011, 2015; Arif et al., 2017). Different aging methods have different main effects on seeds, which might be one of the reasons for this controversy. Changsha has a subtropical monsoon climate, the air

is humid all year, and the temperature in summer is higher than that during the rest of the year; thus, natural aging occurs under conditions of high humidity, which may be one of the reasons why the transcriptomes are similar under both natural aging and artificial aging (high temperature and high humidity). Therefore, the artificial aging method in this research could mimic the natural aging method in high-humidity areas. Moreover, we found RNA is much more prone to oxidative modifications than DNA, even during anhydrobiosis, which would lead to the abundance of oxidized transcripts changing during the after-ripening dry period (Bazin et al., 2011). During rice seed storage, the embryo still has a certain level of activity due to the high humidity in the environment, which leads to the expression of genes related to DNA repair or RNA processing. These might the reason that some mRNAs increased after aging.

During seed maturation, long-lived mRNAs required for the initial stage of germination are synthesized and then stored in the seeds until they are required. Long-lived mRNA will be degraded in the process of seed storage, which inevitably affects the reduction of seed vigor. To identify the long-lived mRNAs that play an important role in seed vigor, we mainly compared the DEGs in HL and LL varieties after natural aging. The heatmap of HL and LL varieties showed that there were certain differences in the expression of some genes in these varieties after aging, and this difference was more obvious in JLY1212 under artificial aging, which might indicate that JLY1212 was more intolerant to storage. At the same time, there were some differences in the GO term enrichment of the HL and LL varieties after natural aging, especially regarding seed maturation, seed dormancy and lipid storage. The degree of enrichment of signaling pathways, such as response to ABA, response to salicylic acid and regulation of GA, also differs between HL and LL varieties; these pathways are associated with seed vigor (Zhao et al., 2021). In addition, we identified 14 special long-lived mRNAs, and their expression levels were significantly different in HL and LL varieties after aging. A motif involved in the initial process of seed germination was enriched in the promoter of GID1. It has been reported that gibberellin can promote seed germination (Yamaguchi and Kamiya, 2001), and gibberellin has an inhibitory effect on seed deterioration (Bueso et al., 2016); seeds treated with GA are more tolerant to aging. The GA 20-oxidase (AtGA20ox) and GA 3-oxidase (AtGA3ox) catalyzed successive steps in the synthesis of bioactive GAs, which had highly lower transcript levels in AtGID1-overexpressing plants than in wild-type plants. Overexpression of AtGID1 increased the sensitivity of *Arabidopsis* to GA (Ju et al., 2018), suggesting a potential role of GID1 in seed longevity. In addition, three AtGID1 receptors have partially specialized functions in seed germination in *Arabidopsis*, AtGID1c play positive regulator of seed germination, whereas AtGID1b negatively regulate germination in dormant seeds in the dark (Ge and Steber, 2018). There are several putative GA receptor genes in rice (Miao et al., 2019), therefore, different GID1 homologous genes may play different roles in rice seed longevity. The search for genes in the GA signaling network may be important for the study of seed longevity (Bueso et al., 2014).

ABA is the other major phytohormones in seed development and seed vigor regulations. It reported that OsHIPL1 protein may modulate endogenous ABA levels and altering OsABIs expression and interacts directly with OsPIP1;1 to affect seed vigor in rice (He et al., 2022), we analyzed the expression of OsHIPL1 and OsPIP1;1 in the HL and LL varieties (**Supplementary Figures 5A,B**), unfortunately, their expression levels have no difference between two varieties, suggesting the differences between the reverse genetic method and the forward genetic method (e.g., transcriptomic analysis). We also identified several long-lived mRNAs with unknown functions, which might be the missed or omitted regulators in rice longevity. The identification of specific long-lived mRNAs in seeds would help to design genetic approaches for using mutants of these mRNAs to understand the mechanisms of genes involved in seed storability regulation in the future. Further work will validate the role of the characterized genes in seed longevity and explore the mechanism by which they are regulated by transcription factor AP2/EREBP.

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: GSA database, https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA006248.

AUTHOR CONTRIBUTIONS

LW conceived the project and designed the research. BW, YT, LJ, and WH performed the research. FY and QL contributed to new reagents and analytical tools. BW and SW analyzed the RNA-seq data. BW and LW wrote the manuscript. All authors reviewed and approved the manuscript for publication.

FUNDING

This work was supported by the National Natural Science Foundation of China (NSFC-32000208), the Science and Technology Innovation Program of Hunan Province (2021RC3044, 2021JJ40056, and 2022WK2007), and the Changsha Municipal Natural Science Foundation (kq2014039).

ACKNOWLEDGMENTS

We thank Qian Liu (Hunan University) for his help with bioinformatic analysis, and thanks to Fan Xu for helping us modify the language grammar.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.857390/full#supplementary-material

REFERENCES

Arif, M. A. R., Nagel, M., Lohwasser, U., and Börner, A. (2017). Genetic architecture of seed longevity in bread wheat (*Triticum aestivum* L.). *J. Biosci.* 42, 81–89. doi: 10.1007/s12038-016-9661-6

Bai, B., Shi, B., Hou, N., Cao, Y., Meng, Y., Bian, H., et al. (2017). microRNAs participate in gene expression regulation and phytohormone cross-talk in barley embryo during seed development and germination. *BMC Plant Biol.* 17:150. doi: 10.1186/s12870-017-1095-2

Bai, B., van der Horst, S., Cordewener, J. H. G., America, T., Hanson, J., and Bentsink, L. (2020). Seed-stored mRNAs that are specifically associated to monosomes are translationally regulated during germination. *Plant Physiol.* 182, 378–392. doi: 10.1104/pp.19.00644

Bailey, T. L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in bipolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2, 28–36.

Bazin, J., Langlade, N., Vincourt, P., Arribat, S., Balzergue, S., El-Maarouf-Bouteau, H., et al. (2011). Targeted mRNA oxidation regulates sunflower seed dormancy alleviation during dry after-ripening. *Plant Cell* 23, 2196–2208. doi: 10.1105/tpc.111.086694

Bewley, J. D., Bradford, K., and Hilhorst, H. (2012). *Seeds: Physiology of Development, Germination and Dormancy*. Berlin: Springer Science and Business Media.

Bueso, E., Muñoz-Bertomeu, J., Campos, F., Brunaud, V., Martínez, L., Sayas, E., et al. (2013). *Arabidopsis thaliana* HOMEOBOX25 uncovers a role for gibberellins in seed longevity. *Plant Physiol.* 164, 999–1010. doi: 10.1104/pp.113.232223

Bueso, E., Muñoz-Bertomeu, J., Campos, F., Brunaud, V., Martínez, L., Sayas, E., et al. (2014). *Arabidopsis thaliana* homeobox25 uncovers a role for gibberellins in seed longevity. *Plant Physiol.* 164, 999–1010.

Bueso, E., Muñoz-Bertomeu, J., Campos, F., Martínez, C., Tello, C., Martínez-Almonacid, I., et al. (2016). *Arabidopsis* COGWHEEL 1 links light perception and gibberellins with seed tolerance to deterioration. *Plant J.* 87, 583–596. doi: 10.1111/tpj.13220

Buijs, G., Kodde, J., Groot, S. P., and Bentsink, L. (2018). Seed dormancy release accelerated by elevated partial pressure of oxygen is associated with DOG loci. *J. Exp. Bot.* 69, 3601–3608. doi: 10.1093/jxb/ery156

Buijs, G., Willems, L. A. J., Kodde, J., Groot, S. P. C., and Bentsink, L. (2020). Evaluating the EPPO method for seed longevity analyses in *Arabidopsis*. *Plant Sci.* 301:110644. doi: 10.1016/j.plantsci.2020.110644

Castro-Mondragon, J. A., Riudavets-Puig, R., Rauluseviciute, I., Berhanu Lemma, R., Turchi, L., Blanc-Mathieu, R., et al. (2022). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 50, D165–D173. doi: 10.1093/nar/gkab1113

Cernac, A., and Benning, C. (2004). WRINKLED1 encodes an AP2/EREB domain protein involved in the control of storage compound biosynthesis in *Arabidopsis*. *Plant J.* 40, 575–585. doi: 10.1111/j.1365-313X.2004.02235.x

Chang, Y., Kong, Q., Shan, X., Tian, G., Ilieva, H., Cleveland, D. W., et al. (2008). Messenger RNA oxidation occurs early in disease pathogenesis and promotes motor neuron degeneration in ALS. *PLoS One* 3:e2849. doi: 10.1371/journal.pone.0002849

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.2020.06.009

Fenollosa, E., Jené, L., and Munné-Bosch, S. (2020). A rapid and sensitive method to assess seed longevity through accelerated aging in an invasive plant species. *Plant Methods* 16, 1–11. doi: 10.1186/s13007-020-00607-3

Fleming, M. B., Patterson, E. L., Reeves, P. A., Richards, C. M., Gaines, T. A., and Walters, C. (2018). Exploring the fate of mRNA in aging seeds: protection, destruction, or slow decay? *J. Exp. Bot.* 69, 4309–4321. doi: 10.1093/jxb/ery215

Ge, W., and Steber, C. M. (2018). Positive and negative regulation of seed germination by the *Arabidopsis* GA hormone receptors, GID 1a, b, and c. *Plant Direct* 2:e00083. doi: 10.1002/pld3.83

Groot, S. P. C., Surki, A. A., de Vos, R. C. H., and Kodde, J. (2012). Seed storage at elevated partial pressure of oxygen, a fast method for analysing seed ageing under dry conditions. *Ann. Bot.* 110, 1149–1159. doi: 10.1093/aob/mcs198

Hay, F. R., Valdez, R., Lee, J. S., and Sta Cruz, P. C. (2019). Seed longevity phenotyping: recommendations on research methodology. *J. Exp. Bot.* 70, 425–434. doi: 10.1093/jxb/ery358

He, Y., Chen, S., Liu, K., Chen, Y., Cheng, Y., Zeng, P., et al. (2022). OsHIPL1, a hedgehog-interacting protein-like 1 protein, increases seed vigor in rice. *Plant Biotechnol. J.* doi: 10.1111/pbi.13812

Ju, Y., Feng, L., Wu, J., Ye, Y., Zheng, T., Cai, M., et al. (2018). Transcriptome analysis of the genes regulating phytohormone and cellular patterning in *Lagerstroemia* plant architecture. *Sci. Rep.* 8, 1–14. doi: 10.1038/s41598-018-33506-8

Kimura, M., and Nambara, E. (2010). Stored and neosynthesized mRNA in *Arabidopsis* seeds: effects of cycloheximide and controlled deterioration treatment on the resumption of transcription during imbibition. *Plant Mol. Biol.* 73, 119–129. doi: 10.1007/s11103-010-9603-x

Kong, Q., and Lin, C. L. (2010). Oxidative damage to RNA: mechanisms, consequences, and diseases. *Cell. Mol. Life Sci.* 67, 1817–1829. doi: 10.1007/s00018-010-0277-y

Lee, J.-S., Velasco-Punzalan, M., Pacleb, M., Valdez, R., Kretzschmar, T., McNally, K. L., et al. (2019). Variation in seed longevity among diverse indica rice varieties. *Ann. Bot.* 124, 447–460. doi: 10.1093/aob/mcz093

Li, T., Zhang, Y., Wang, D., Liu, Y., Dirk, L. M., Goodman, J., et al. (2017). Regulation of seed vigor by manipulation of raffinose family oligosaccharides in maize and *Arabidopsis thaliana*. *Mol. Plant* 10, 1540–1555. doi: 10.1016/j.molp.2017.10.014

Members, B. D. C. (2018). Database resources of the BIG data center in 2018. *Nucleic Acids Res.* 46, D14–D20. doi: 10.1093/nar/gkx897

Miao, C., Wang, Z., Zhang, L., Yao, J., Hua, K., Liu, X., et al. (2019). The grain yield modulator miR156 regulates seed dormancy through the gibberellin pathway in rice. *Nat. Commun.* 10, 1–12. doi: 10.1038/s41467-019-11830-5

Min, C. W., Lee, S. H., Cheon, Y. E., Han, W. Y., Ko, J. M., Kang, H. W., et al. (2017). In-depth proteomic analysis of Glycine max seeds during controlled deterioration treatment reveals a shift in seed metabolism. *J. Proteomics* 169, 125–135. doi: 10.1016/j.jprot.2017.06.022

Nagel, M., Kranner, I., Neumann, K., Rolletschek, H., Seal, C. E., Colville, L., et al. (2015). Genome-wide association mapping and biochemical markers reveal that seed ageing and longevity are intricately affected by genetic background and developmental and environmental conditions in barley. *Plant Cell Environ.* 38, 1011–1022. doi: 10.1111/pce.12474

Nagel, M., Rosenhauer, M., Willner, E., Snowdon, R. J., Friedt, W., and Börner, A. (2011). Seed longevity in oilseed rape (*Brassica napus* L.)–genetic variation and QTL mapping. *Plant Genet. Resour.* 9, 260–263. doi: 10.1017/S1479262111000372

Nakabayashi, K., Okamoto, M., Koshiba, T., Kamiya, Y., and Nambara, E. (2005). Genome-wide profiling of stored mRNA in *Arabidopsis thaliana* seed germination: epigenetic and genetic regulation of transcription in seed. *Plant J.* 41, 697–709. doi: 10.1111/j.1365-313X.2005.02337.x

Nakajima, M., Shimada, A., Takashi, Y., Kim, Y. C., Park, S. H., Ueguchi-Tanaka, M., et al. (2006). Identification and characterization of *Arabidopsis* gibberellin receptors. *Plant J.* 46, 880–889. doi: 10.1111/j.1365-313X.2006.02748.x

Nguyen, T.-P., Keizer, P., van Eeuwijk, F., Smeekens, S., and Bentsink, L. (2012). Natural variation for seed longevity and seed dormancy are negatively correlated in *Arabidopsis*. *Plant Physiol.* 160, 2083–2092. doi: 10.1104/pp.112.206649

Okamuro, J. K., Caster, B., Villarroel, R., Van Montagu, M., and Jofuku, K. D. (1997). The AP2 domain of APETALA2 defines a large new family of DNA binding proteins in *Arabidopsis*. *Proc. Natl. Acad. Sci. U.S.A.* 94, 7076–7081. doi: 10.1073/pnas.94.13.7076

Oliveros, J. (2007–2015). Venny. An Interactive Tool for Comparing Lists with Venn's Diagrams. Available online at: https://bioinfogp.cnb.csic.es/tools/venny/index.html

Panobianco, M., Vieira, R. D., and Perecin, D. (2007). Electrical conductivity as an indicator of pea seed aging of stored at different temperatures. *Sci. Agric.* 64, 119–124. doi: 10.1590/S0103-90162007000200003

Pereira Lima, J. J., Buitink, J., Lalanne, D., Rossi, R. F., Pelletier, S., Da Silva, E. A. A., et al. (2017). Molecular characterization of the acquisition of longevity during seed maturation in soybean. *PLoS One* 12:e0180282. doi: 10.1371/journal.pone.0180282

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

Qu, J., Wang, M., Liu, Z., Jiang, S., Xia, X., Cao, J., et al. (2020). Preliminary study on quality and storability of giant hybrid rice grain. *J. Cereal Sci.* 95:103078. doi: 10.1016/j.jcs.2020.103078

Qun, S., Wang, J.-H., and Sun, B.-Q. (2007). Advances on seed vigor physiological and genetic mechanisms. *Agr. Sci. China* 6, 1060–1066. doi: 10.1016/S1671-2927(07)60147-3

Rajjou, L., and Debeaujon, I. (2008). Seed longevity: survival and maintenance of high germination ability of dry seeds. *C. R. Biol.* 331, 796–805. doi: 10.1016/j.crvi.2008.07.021

Rajjou, L., Lovigny, Y., Groot, S. P., Belghazi, M., Job, C., and Job, D. (2008). Proteome-wide characterization of seed aging in *Arabidopsis*: a comparison between artificial and natural aging protocols. *Plant Physiol.* 148, 620–641. doi: 10.1104/pp.108.123141

Rajjou, L., Lovigny, Y., Job, C., Belghazi, M., Groot, S., and Job, D. (2007). *Seed Quality and Germination*. Wallingford: Centre for Agriculture and Bioscience International. doi: 10.1079/9781845931971.0324

Saighani, K., Kondo, D., Sano, N., Murata, K., Yamada, T., and Kanekatsu, M. (2021). Correlation between seed longevity and RNA integrity in the embryos of rice seeds. *Plant Biotechnol.* 38, 277–283. doi: 10.5511/plantbiotechnology.21.0422a

Sano, N., Rajjou, L., North, H. M., Debeaujon, I., Marion-Poll, A., and Seo, M. (2016). Staying alive: molecular aspects of seed longevity. *Plant Cell Physiol.* 57, 660–674. doi: 10.1093/pcp/pcv186

Shan, X., Tashiro, H., and Lin, C. L. (2003). The identification and characterization of oxidized RNAs in Alzheimer's disease. *J. Neurosci.* 23, 4913–4921. doi: 10.1523/JNEUROSCI.23-12-04913.2003

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

Ueguchi-Tanaka, M., Ashikari, M., Nakajima, M., Itoh, H., Katoh, E., Kobayashi, M., et al. (2005). Gibberellin insensitive dwarf1 encodes a soluble receptor for gibberellin. *Nature* 437, 693–698. doi: 10.1038/nature04028

Yamaguchi, S., and Kamiya, Y. (2001). Gibberellins and light-stimulated seed germination. *J. Plant Growth Regul.* 20, 369–376. doi: 10.1007/s003440010035

Yuan, Z., Fan, K., Wang, Y., Tian, L., Zhang, C., Sun, W., et al. (2021). OsGRETCHENHAGEN3-2 modulates rice seed storability via accumulation of abscisic acid and protective substances. *Plant Physiol.* 186, 469–482. doi: 10.1093/plphys/kiab059

Zhang, Y.-X., Xu, H.-H., Liu, S.-J., Li, N., Wang, W.-Q., Møller, I. M., et al. (2016). Proteomic analysis reveals different involvement of embryo and endosperm proteins during aging of yliangyou 2 hybrid rice seeds. *Front. Plant Sci.* 7:1394. doi: 10.3389/fpls.2016.01394

Zhao, J., He, Y., Huang, S., and Wang, Z. (2021). Advances in the identification of quantitative trait loci and genes involved in seed vigor in rice. *Front. Plant Sci.* 12:659307. doi: 10.3389/fpls.2021.659307

Zhu, F. (2018). Anthocyanins in cereals: composition and health effects. *Food Res. Int.* 109, 232–249. doi: 10.1016/j.foodres.2018.04.015

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

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

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 16 June 2022
doi: 10.3389/fpls.2022.921536

Integrated Metabolomics and Transcriptome Analyses Unveil Pathways Involved in Sugar Content and Rind Color of Two Sugarcane Varieties

Zhaonian Yuan[1,2,3†], Fei Dong[4,5], Ziqin Pang[1,2†], Nyumah Fallah[1,2], Yongmei Zhou[2,5†], Zhi Li[1,2] and Chaohua Hu[1,2]*

[1] Key Laboratory of Sugarcane Biology and Genetic Breeding, Ministry of Agriculture, Fujian Agriculture and Forestry University, Fuzhou, China, [2] College of Agricultural, Fujian Agriculture and Forestry University, Fuzhou, China, [3] Province and Ministry Co-sponsored Collaborative Innovation Center of Sugar Industry, Nanning, China, [4] College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou, China, [5] Center for Genomics and Biotechnology, Fujian Agriculture and Forestry University, Fuzhou, China

OPEN ACCESS

Edited by:
Xingtan Zhang,
Agricultural Genomics Institute at Shenzhen (CAAS), China

Reviewed by:
Jiangfeng Huang,
Guangxi University, China
Weilong Kong,
Wuhan University, China

***Correspondence:**
Zhaonian Yuan
yuanzn05@163.com

†ORCID:
Zhaonian Yuan
orcid.org/0000-0003-1502-3291
Ziqin Pang
orcid.org/0000-0002-0943-1224
Yongmei Zhou
orcid.org/0000-0002-3340-2810

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

Received: 16 April 2022
Accepted: 18 May 2022
Published: 16 June 2022

Citation:
Yuan Z, Dong F, Pang Z, Fallah N, Zhou Y, Li Z and Hu C (2022) Integrated Metabolomics and Transcriptome Analyses Unveil Pathways Involved in Sugar Content and Rind Color of Two Sugarcane Varieties. Front. Plant Sci. 13:921536. doi: 10.3389/fpls.2022.921536

Metabolic composition can have potential impact on several vital agronomic traits, and metabolomics, which represents the bioactive compounds in plant tissues, is widely considered as a powerful approach for linking phenotype–genotype interactions. However, metabolites related to cane traits such as sugar content, rind color, and texture differences in different sugarcane cultivars using metabolome integrated with transcriptome remain largely inconclusive. In this study, metabolome integrated with transcriptome analyses were performed to identify and quantify metabolites composition, and have better insight into the molecular mechanisms underpinning the different cane traits, namely, brix, rind color, and textures in the stems (S) and leaves (L) of sugarcane varieties FN41 and 165402. We also identified metabolites and associated genes in the phenylpropanoid and flavonoid biosynthesis pathways, starch and sucrose metabolism. A total of 512 metabolites from 11 classes, with the vast majority (122) belonging to flavonoids were identified. Moreover, the relatively high amount of D-fructose 6-p, D-glucose6-p and glucose1-p detected in FN41L may have been transported and distributed by source and sink of the cane, and a majority of them reached the stem of sugarcane FN41L, thereby promoting the high accumulation of sugar in FN41S. Observations also revealed that genes such as C4H, CHS, F3H, F3'H, DFR, and FG2 in phenylpropanoid and flavonoid biosynthesis pathways were the major factors impacting the rind color and contrasting texture of FN41 and 165204. Further analysis revealed that weighted gene co-expression network analysis (WGCNA) hub genes and six transcription factors, namely, Tify and NAC, MYB-related, C2C2-Dof, WRKY, and bHLH play a key role in phenylpropanoid biosynthesis, flavone and flavonol biosynthesis, starch and sucrose metabolism. Additionally, metabolites such as L-phenylalanine, tyrosine, sinapaldehyde, pinobanksin, kaempferin, and nictoflorin

were the potential drivers of phenotypic differences. Our finding also demonstrated that genes and metabolites in the starch and sucrose metabolism had a significant effect on cane sugar content. Overall, this study provided valuable insight into the molecular mechanisms underpinning high sugar accumulation and rind color in sugarcane, which we believe is important for future sugarcane breeding programs and the selection of high biomass varieties.

Keywords: sugarcane, metabolome and transcriptome, flavonoids, sugar metabolism, WGCNA

INTRODUCTION

Sugarcane (*Saccharum* spp.) is a perennial C4 Graminaceous plant mainly cultivated in tropical and subtropical areas. It is widely known as a main sugar and biofuel feedstock crop that accounts for about 80% of global sugar production and 40% of ethanol production (Zhang et al., 2018). A number of studies have shown that a large quantity of sucrose is stored in the sugarcane stem, accounting for about 650 mM per kg (Welbaum and Meinzer, 1990). Furthermore, it has been reported that the quality of cane cultivar is contingent upon various conditions such as rind hardness, sucrose percent in juice and purity. High sugar content is one of the main objectives of sugarcane breeding, and increasing sugar content is economically important for the development of the sugarcane industry (Thirugnanasambandam et al., 2017). Bearing in mind the increasing demand for sugar by the world growing population, it is essential to produce cultivars with high sugar content. Cane varieties are known to have a close association with sugar yield and its yield-related parameters, namely, brix, stalk diameter, stalk weight, stalk number, stalk height, and fiber (Mancini et al., 2012). The vast majority of sugar-related agronomic traits such as HR brix, sucrose percent, number of green leaves, leaf area and internode length have demonstrated a significant relationship with rind hardness (Babu et al., 2009). The mechanisms underpinning sucrose accumulation have been investigated at various levels including identifying and characterizing individual metabolites (Glassop et al., 2007; Lin et al., 2022), transcriptome (Aitken et al., 2006), and genes in the sucrose pathway and movement of sucrose within plants (Moore, 2005), localization of genes (Rae et al., 2005), genomic maps advancement and quantitative trait loci localization (Casu et al., 2007).

In general, plants have developed different mechanisms to adapt to changing environmental conditions, for instance, the development of foliar trichomes, glandular hairs and a wax layer, and the production of metabolites (Granados-Sánchez et al., 2008). Previous studies showed that secondary metabolites including flavonoids, terpenoids, phenolics, proanthocyanidins, carotenoids, etc., are antioxidant agents (Rao et al., 2019), they also inhibit and deters oviposition and feeding. These metabolites can also protect plants against predators and pathogens (Ikonen et al., 2001), impede insect growth, attract pollinators (Agati and Tattini, 2010), and act as allelopathic agents (Sarker and Oba, 2018). Moreover, metabolites play crucial roles in protecting plant against fungi, bacteria, and viruses (Sun et al., 2008), and protect against ultraviolet radiation and high light (Rao et al., 2019). They can, besides, have potential impacts on other aspects of plant growth, development, and nutritional quality that are important in sugarcane production as well as different species and differ among plants of the same species, between diverse plant tissues (e.g., new and mature leaves, root, stem, fruit, etc.) (Rao et al., 2021). Therefore, metabolites provide immense potential in molecular breeding program. To sum up, it is of essence to investigate metabolites and their fundamental regulatory mechanisms from a more macroscopic standpoint such as metabolome.

Many thanks to the development of a high-throughput metabolite identification tool for sugarcane (Schaker et al., 2017), and the identification and quantification of all metabolites in biological samples (Patti et al., 2012). For instance, the metabolomics tool was employed to compare and quantify metabolites and their antioxidant activities in young and mature leaves of 12 different sugarcane varieties. It was revealed that the mature leaves of sugarcane varieties Taitang172 and ROC22 contained a significant amount of flavonoid, and these varieties exhibited high antioxidant activities among the 12 sugarcane varieties (Rao et al., 2021). In related study, Chen et al. (2018) detected 68 metabolites belonging to 11 metabolite classes, which varied considerably among the different tissues of Tieguanyin Tea cultivar using untargeted metabolomics. Wijma et al. (2021) also identified a high quantity of biosynthesis of secondary metabolites, amino acid metabolism, xenobiotics biodegradation and metabolism in different tissues of sugarcane. In sugarcane plants, several studies have been conducted using metabolomic analysis to study different biological problems. For instance, a previous study identified co-expression and specific metabolites associated with metabolic pathways correlated with Brix and fiber content using metabolite profiling (Perlo et al., 2020). In another study, targeted metabolomics tool was also employed to quantify 16 phenolamide and 90 flavonoid metabolites in the seedlings of different rice tissues (Dong et al., 2015). However, most of these previous studies only focused on metabolomics tool to investigate metabolites in different tissues of the plant. Whereas the integration of metabolomics with transcriptomics to investigate the different bioactive compounds and different potential transcriptional regulations in sugarcane, which is essential in tracking the changes of metabolites and

Abbreviations: PAL, phenylalanine ammonia lyase; PTAL, phenylalanine/tyrosine ammonia-lyase; C4H, cinnamate4-hydroxylase; 4CL, 4-coumarate CoA ligase; HCT, hydroxycinnamoyl CoA shikimate/quinate hydroxycinnamoyl transferase; C3H, *P*-coumarate 3-hydroxylase; COMT, caffeic acid *O*-methyltransferase; CCoAOMT, caffeoyl-CoA *O*-methyltransferase; CHI, chalcone isomerase; CCR, cinnamoyl-CoA reductase; CAD, cinnamyl alcohol dehydrogenase; F5H, ferulate5-hydroxylase; F3'5'H, flavonoid 3',5'-hydroxylase; POD, peroxidase; FLS, flavonol synthase; DFR, dihydroflavonol 4-reductase.

their corresponding regulatory genes within specific cane tissues remains largely inconclusive.

Recently, metabolomics integrated with transcriptomics has been widely used to investigate the metabolites and related genes involved in biological pathways such as color variation and quality formation in many plants. For example, a combined transcriptomic and metabolomic analyses were adopted to identify the carbohydrate and organic acid metabolism genes associated with brix in of two types of tomato fruits. The study revealed that L-malic acid, citric acid, and genes involved in CHO metabolism were significantly associated with sugar content in tomato fruits (Li et al., 2021). Moreover, integrated transcriptome and metabolome tools were used to establish a global map of metabolite accumulation and gene regulation during fruit development in wild and cultivated watermelons (Gong et al., 2021). The analysis of metabolite and transcriptome profiles during the storage of two peach cultivars revealed the molecular mechanisms underlying different fruit textures in peach (Wang et al., 2018). Metabolic and proteomic analyses were also employed to identify potential proteins and pathways involved in sugarcane resistance (Wang et al., 2020). A previous study identified seven candidate genes involved in anthocyanin biosynthesis by transcriptomic and metabolomic analyses in three sugarcane cultivars of different colors. These authors identified some candidate genes associated with anthocyanin biosynthesis using transcriptomic and metabolomic analyses. They also found key flavonoids and anthocyanins that caused color difference, and the key candidate genes that regulated these metabolites (Ni et al., 2021). However, metabolites related to cane traits such as sugar content, rind color, and texture differences in different sugarcane cultivars using metabolome integrated with transcriptome remain largely elusive.

In the present study, we employed integrated metabolomic and transcriptomic analyses to detect and quantify the composition of metabolites in two distinct sugarcane cultivars (165204 and FN41), and to beer understand their relationship with cane traits such as sugar content, rind color, and texture. This study also aimed at identifying metabolites and associated genes in the phenylpropanoid and flavonoid biosynthesis pathways, and starch and sucrose metabolism. The results from this study will offer new insights on sugarcane stem growth and sugar accumulation and provide a theoretical basis for further research such as the validation of gene function and the genetic improvement of sugarcane cultivars.

MATERIALS AND METHODS

Plant Materials and Growth Condition

Two sugarcane cultivars ("165204" and "FN41") were cultivated in a randomized field plot according to standard agricultural practices in a field at the Baisha Town, Fuzhou City, Fujian Province, China (E 119°14', N 26°16') in 2019. The region has a subtropical monsoon climate with an altitude of 123 m, an average annual temperature of 17–20° and an annual rainfall of 1,200–2,100 mm. The site was previously used for sugarcane monoculture cropping system using a conventional approach. The following basic soil properties were measured: OM = 28.73 g/kg, total nitrogen (TN) = 1.22 g/kg, total phosphorus (TP) = 0.71 g/kg, and total potassium (TK) = 9.19 g/kg, this environment is suitable for sugarcane growth. The sugarcane cv. 165204 cultivated contained a green rind with a brittle texture, while cv. FN41 consisted of a purple rind with a hard texture. The treatments included (i) sugarcane monoculture with 165204 and (ii) sugarcane monoculture with FN41. Two varieties of sugarcane were cultivated on March 7, 2019, after the soil was plowed (40 cm depth) using rotary tillage. Sugarcane monoculture was cultivated with a line spacing of 1.2 and a planting density of 85,000 buds/hm^2. The experiment was set in a randomized block design with two treatments and three replicates constituting a total of six plots, with each covering an area of 144.0 m^2 (24.0 × 6.0 m). All plots were fertilized with the traditional local fertilizer application of 250 kg/hm^2 of urea, 100 kg/hm^2 of K_2O, and 450 kg/hm^2 of calcium superphosphate per season. Forty and sixty percent of the total fertilizer application were applied at the seedling and elongation stages of sugarcane, respectively. Sugarcane agronomic traits were investigated at the sugarcane maturation stage on 2 January 2020. The fresh stems and leaves of the two sugarcane varieties were collected on the same day, specifically, the cane stems of the seventh (middle) node of sugarcane, and the first fully expanded leaf of sugarcane as the material. Three biological replicates were collected for each tissue (stems and leaves) in two sugarcane cultivars (165204 and FN41), and a total of 12 samples were collected. "FN41L" and "165204L" represent the leaf tissues, while "FN41S" and "165204S" represent the stem tissues of sugarcane varieties FN41 and 165204, respectively. All the flesh samples were washed with DEPC water and 75% ethanol, wrapped in tin foil and labeled, then immediately placed in liquid nitrogen and stored at –80°C until further analysis.

Analysis of the Properties of Sugarcane

To measure the stalk diameter and height of the plants, 30 sugarcane plants were randomly selected from each bed and measured with a tape and Vernier caliper. Extech Portable Sucrose Brix Refractometer (Mid-State Instruments, San Luis Obispo, CA, United States) was used to determine sucrose content and calculated through using the formula: sucrose (%) = Brix (%) × 1.0825 – 7.703. To understand the brittleness and stiffness of the stems of two sugarcane cultivars, we determined the mechanical properties of sugarcane stems by the method of testing in tensile strength perpendicular to the grain of wood. The tensile strength was measured according to the standard GB/T14017-2009. The cane stems of FN41 and 165204 cut into dumbbell shape and were tested using the UTM4304X electronic universal testing machine with a jig adapted to the tensile strength of sugarcane (model: JDSB104B). Test operation steps: In brief, we applied tensile force at a uniform speed along the main stem of sugarcane through the jig of the testing machine in the direction of the main stem until the stem was destructed (**Supplementary Figure 1**), and the tensile strength of sugarcane was calculated by adopting Elastic modulus (E) = (F/S) × $(dL/L)^{-1}$. F represents the tensile strength, S stands the cross-sectional area of the sugarcane, dL denotes

the elongation of the sugarcane, while L represents the original length of the sugarcane. The elastic modulus can be regarded as an indicator of the ease of producing elastic deformation of a material. The larger the value, the greater the stiffness of the material; the smaller the value, the more brittle the material.

Sample Preparation and Extraction for Metabolomic Analysis

Firstly, the plant samples were freeze-dried in a lyophilizer (Scientz-100F, Ningbo, China), then ground to powder using a grinding instrument (MM 400, Retsch) for 1.5 min. Next, 100 mg of the powder was weighed and dissolved in 1.2 ml of 70% methanol extraction solution. Then, the dissolved samples were placed in a refrigerator at 4°C overnight and vortexed six times to improve the extraction rate. After overnight incubation, the mixture was centrifuged at 10,000 *g* for 10 min and the supernatant was filtered with a microporous membrane (SCAA-104, 0.22 μm pore size; ANPEL, Shanghai, China). The samples were stored in a sample injection bottle for UPLC-MS/MS analysis. Finally, a quality-control sample (mix) was prepared by mixing an equal amount of all samples to monitor the stability of the analytical conditions for assay analysis.

Ultra Performance Liquid Chromatography and ESI-Q TRAP-MS/MS Conditions

Metabolite profiling was performed using an UPLC-ESI-MS/MS system [UPLC (Ultra Performance Liquid Chromatography), Shim-pack UFLC SHIMADZU CBM30A system[1]; MS/MS (Tandem mass spectrometry), Applied Biosystems 6500 Q TRAP]. The analytical conditions were as follow, UPLC: column, Waters ACQUITY UPLC HSS T3 C18 (1.8 μm, 2.1 mm*100 mm). The mobile phase consisted of solvent A, pure water with 0.04% acetic acid, and solvent B, acetonitrile with 0.04% acetic acid. Sample measurements were performed with a gradient program with the starting conditions of 95% A, 5% B. Within 10 min, a linear gradient to 5% A, 95% B was programmed, and a composition of 5% A, 95% B was kept for 1 min. Subsequently, a composition of 95% A and 5.0% B were adjusted within 0.10 min and kept for 2.9 min. The column oven was set to 40°C and volume of 2 μl. The effluent was alternatively connected to an ESI-triple quadrupole-linear ion trap (Q TRAP)-MS.

Linear ion trap (LIT) and triple quadrupole (QQQ) scans were acquired on a triple quadrupole-linear ion trap mass spectrometer (Q TRAP), API 6500 Q TRAP UPLC/MS/MS System, equipped with an ESI Turbo Ion-Spray interface, operating in positive and negative ion mode and controlled by Analyst 1.6.3 software (AB Sciex). The ESI source operation parameters were as follow: ion source, turbo spray; source temperature 550°C; ion spray voltage (IS) 5500 V (positive ion mode)/-4500 V (negative ion mode); ion source gas I (GSI), gas II(GSII) and curtain gas (CUR) were set at 50, 60, and 30.0 psi with a high collision gas (CAD), respectively. Instrument tuning and mass calibration were performed with 10 and 100 μmol/L polypropylene glycol solutions in QQQ and LIT modes, respectively. QQQ scans were acquired as MRM experiments with collision gas (nitrogen) set to 5 psi. DP and CE for individual MRM transitions were done with further DP and CE optimization. A specific set of MRM transitions were monitored accordingly for each period according to the metabolites eluted (Fraga et al., 2010).

Metabolite Quantification and Data Analysis

Qualitative analysis of metabolites was performed according to the secondary spectrum information based on the self-built Metware Database (MWDB) of Metware Biotechnology Co., Ltd. (Wuhan, China) and other public databases of metabolite information including MassBank[2], KNAPSAcK[3], HMDB[4], and METLIN[5] (Zhu et al., 2013). Metabolite quantification was carried out with data acquired in the multiple reaction monitoring (MRM) mode of QQQ mass spectrometry. Mass spectrometry data were then analyzed and quantified using Analyst software v1.6.3 and Multiquant Software v3.0.2.

The data of metabolites profiling were pre-processed using unit variance (UV) scaling before multivariate analysis. Principal component analysis (PCA) was executed using the prcomp function in R software (version 3.0.3). Pearson's correlation coefficient between samples was calculated in R using the cor function. Hierarchical cluster analysis (HCA) was performed using R package pheatmap based on the Euclidean distance coefficient. Further, orthogonal signal correction and Partial Least Squares-Discriminant Analysis (OPLS-DA) were executed after log2 transformation and Mean Centering of raw data by the MetaboAnalyst package in R software. The differentially expressed metabolites were screened based on OPLS-DA analysis by the following criteria: (1) Metabolites with fold change ≥ 2 or fold change ≤ 0.5; (2) Based on the above, the metabolites with VIP (variable importance in project) ≥ 1 were selected. We conducted a combine analysis between the metabolome and transcriptome datasets, the mean of all biological replicates of differential metabolites in the metabolome data and the mean value of expression of differential transcripts in the transcriptome data were examined. Later, we transformed the log2 datasets using the 'cor' package from the R software[6]. The Pearson correlation (r) was then employed between metabolites and transcripts in phenylpropanoid and flavonoid biosynthesis pathway, followed by starch and sucrose metabolism pathway was represented by network diagrams, and the genes and metabolites were selected when $R^2 > 0.8$ (Cho et al., 2016). Metabolome and transcriptome relationships were visualized using the Cytoscape software version 3.6.1 (Su et al., 2014).

[1]https://www.shimadzu.com.cn/

[2]http://www.massbank.jp/

[3]http://kanaya.naist.jp/KNApSAcK/

[4]http://www.hmdb.ca/

[5]http://metlin.scripps.edu/index.php

[6]www.r-project.org/

Transcriptome Sequencing and Data Analysis

Total RNA was extracted from sugarcane samples using TRIzol reagent (Invitrogen, CA, United States) according to the manufacturer's instructions. The isolated RNA was further treated with RNase-Free DNase (Promega, Madison, WI, United States) to remove possible genomic DNA. Qubit 2.0 fluorometer (Life Technologies, Carlsbad, CA, United States) and Agilent Bioanalyzer 2100 (Agilent Technologies, Palo Alto, CA, United States) were used to estimate the concentration and purification of the RNA, and its quality was confirmed using 1% agarose gel electrophoresis. High-quality RNA was used for further library construction. Library construction, library clustering and high-throughput sequencing were carried out by adopting Metware Biotechnology Co., Ltd (Wuhan, China) with an Illumina HiSeq^{TM} 2500 platform (Illumina Inc., San Diego, CA, United States). Subsequently, the clean reads were obtained by removing the adaptors, reads with N greater than 10%, and whose base number with low-quality bases ($Q < 20$) were greater than 50%. The error rate, Q20, Q30, and GC content of the clean data were recorded to evaluate the RNA-seq quality. The raw RNA-seq read data were deposited in the Short Read Archive[7] and can be accessed using the BioProject ID: PRJNA805530.

The clean reads were mapped to the reference sugarcane genome sequence using HISAT2 v2.1.0. Novel genes and transcripts were also predicted using StringTie v1.3.3b (Pertea et al., 2016). Subsequently, the gene expression levels of the samples were estimated as fragments per kilobase of exon model per million mapped fragments (FPKM) using featureCounts v1.6.1 (Liao et al., 2014). The differentially expressed genes (DEGs) were identified using DESeq2 v1.22.2 with $|\log2\text{Fold Change}| \geq 1$ and false discovery rate (FDR) < 0.05. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway functional enrichment analyses were performed via the Gene Ontology Database and KEGG Database[8], respectively (Ashburner et al., 2000; Kanehisa et al., 2004). Fisher's exact test was used to select the significant GO categories and KEGG pathways with the threshold of FDR < 0.05. Besides, we also annotated gene functions based on the following databases: NR (NCBI non-redundant protein sequences), KOG (euKaryotic Orthologous Groups) (Koonin et al., 2004), COG (Clusters of Orthologous Groups of proteins) (Tatusov et al., 2000), Pfam (Protein family) (Finn et al., 2014), Trembl (Translated EMBL Nucleotide Sequence Data Library) and Swiss-Prot (a manually annotated and reviewed protein sequence database) (Apweiler et al., 2004) with the BLAST program (Evalue $\leq$ 1e-5) (Altschul et al., 1997). Transcription factors (TFs) among the DEGs were predicted using iTAK online program[9]. Alternative splicing (AS) events were detected using rMATS v4.0.2. The SNP (Single Nucleotide Polymorphism) and indel (Insertion-Deletion) variants were called using GATK v3.8 and then annotated using ANNOVAR[10].

[7]http://www.ncbi.nlm.nih.gov/sra/

[8]https://www.genome.jp/kegg

[9]http://itak.feilab.net/cgi-bin/itak/index.cgi

[10]http://www.openbioinformatics.org/annovar/

Co-expression Analysis

We used R package WGCNA (Langfelder and Horvath, 2008) to construct the gene co-expression network, and the genes with average gene expression greater than 10 were selected. After filtering, we obtained a total of 15,652 genes to construct the module. Some parameters are as follows: the soft thresholding power of the correlation network was set at 20, the deepSplit value was 2, the minimum gene module size was equal to 100, and the modules whose distance was less than 0.15 were merged and the total of 20 modules were generated. Later, Pearson correlation analysis showed that the module was co-expressed with the abundance of 24 metabolites related to phenylpropanoid biosynthesis (ko00940), flavonoid biosynthesis (ko00941), flavone and flavonol biosynthesis (ko00944), and starch and sucrose metabolism (ko00500). Finally, Cytoscape 3.6.1 was used to visualize the core genes in the core co-expression module (Kohl et al., 2011).

Quantitative RT-PCR Validation

The expression level of genes was validated using Quantitative RT-PCR (qRT-PCR) according to the instructions of TransStart® Top Green qPCR SuperMix (Transgen Biotech, Beijing, China). A total of 26 genes were selected and verified using qRT-PCR. The Gene-specific primers for qRT-PCR were designed with NCBI primer-blast tool[11] and listed in **Supplementary Table 1**. The RNA samples used for qRT-PCR analysis were aliquots of the samples used in the RNA-seq experiments. Each qPCR reaction was performed using three biological replicates and three technical replicates. The PCR reaction conditions were as follows: 95°; for 10 min followed by 40 cycles of 95° for 30 s and 60° for 1 min. Reactions were performed using an Applied Biosystems 7500 Real-Time PCR system. The actin gene was used as the internal reference gene for normalization of expression, and relative expression was calculated using the delta-delta Ct method ($2^{-\Delta\Delta Ct}$ method).

RESULTS

Phenotype and Quality Traits Description of "FN41" and "165204"

Sugar content and texture of sugarcane are some of the most important indexes of sugarcane quality and have some significant relationship. Sugarcane with harder texture tend to produce more and sweeter sugar content, which are ideal for sugarcane squeezing, thus significantly reducing the production cost of sucrose and providing huge economic benefits for the sugar industries. On the other hand, the color difference of sugarcane has certain ornamental and economic value in the sugarcane industry. The sugarcane variety "FN41" has hard texture and purple color, with a higher sugar content of 17.22%, and "165204" cultivar texture consisted of a crisp and green ring color and 12.85% of sugar content. These two cultivars are deemed excellent materials for studying the mechanism of sugar

[11]https://www.ncbi.nlm.nih.gov/tools/primer-blast/

content, texture and color difference of sugarcane. The findings of this study provide new insights into the molecular mechanism underpinning the accumulation of high sugar, hardness and color difference.

Although the two sugarcane cultivars were grown simultaneously in the same field and under the same conditions, the morphology of the tissues within the stalks and stem color were distinct (**Figure 1A**). FN41 had a purple rind, with fibers compressed fibers, while 165204 consisted of green rind and a loose connective fibrous tissue. During harvest, several important traits including the quality and yield of the two cultivars were evaluated (**Figure 1B**). Sugar content showed a significant differences were observed between FN41 and 165204 (percentage concentration of sugar, *p*-value = 2.80E-08) and stem height (*p*-value = 0.014). FN41 had a higher brix and stem height compared to 165204. While the other qualitative trait evaluated in our study were not significantly different between the two cultivars. Therefore, we speculated that the two cultivars of sugarcane have certain differences in sugar content, rind color, and pith texture.

Metabolome Profiling and Identification of the Differentially Accumulated Metabolites Between FN41 and 165204

To quantify the total metabolites in the stems and leaves of the two varieties we adopted a metabolomics tool. A total of 512 metabolites grouped into 11 classes were identified from the 12 samples. Among them, there were 122 flavonoids, 89 phenolic acids, 67 amino acids and derivatives, 58 lipids, 44 organic acids, 35 nucleotides and derivatives, followed by 21 alkaloids, 7 lignans and coumarins, 3 tannins, 3 terpenoids, and 63 other metabolites (**Supplementary Table 2**). One QC sample was inserted for every 10 analyzed samples to monitor the reproducibility of the instrument's analytical process. The overlay of the TIC plots between different quality control (QC)

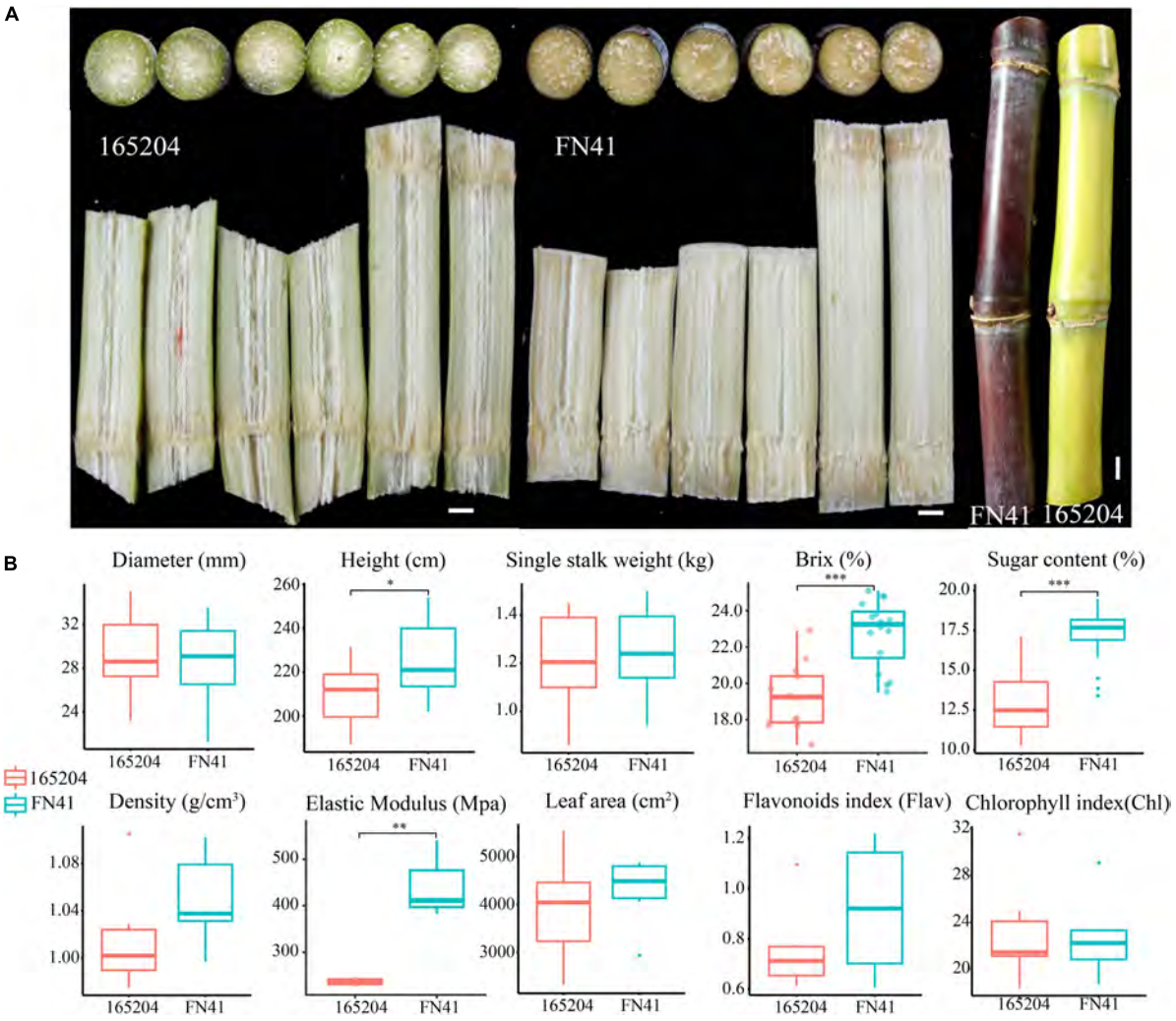


FIGURE 1 | Physiological characteristics of FN41 and 165204 sugarcane cultivars, bar = 1 cm **(A)**; Comparison of agronomic traits between FN41 and 165204 **(B)**. $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

samples demonstrated the high repeatability and reliability of the data in this study (**Supplementary Figure 2**). We also performed PCA to visualize the metabolites composition in the samples. It was observed that metabolites composition in samples along PC1 (*x*-axis) accounted for 48.5% and PC2 (*y*-axis) represented 12.6% of variability, respectively. The analysis also revealed that metabolites composition in the stems of both varieties were densely clustered together, while metabolites composition in the leaf of both varieties leaf samples exhibited the opposite (**Figure 2**). The hierarchical cluster analysis (HCA) further revealed that the trends of metabolites composition were distinctly different between FN41 and 165204 (**Supplementary Figure 3**).

To pinpoint the significant differentially expressed dead-end metabolites (DEMs) associated with phenotype, the VIP (variable importance in project) $\geq$ 1.0 together with fold change $\geq$ 2 or $\leq$ 0.5 were set as the thresholds. We identified a total of 364 DEMs among the four groups compared (165204S_vs_FN41S, 165204L_vs_FN41L, 165204S_vs_165204L, and FN41S_vs_FN41L), including 74, 92, 280 and 250 DEMs in 165204S_vs_FN41S, 165204L_vs_FN41L, 165204S_vs_165204L, and FN41S_vs_FN41L, respectively (**Supplementary Table 3**), with most showing significantly high accumulation. In addition, only 26 DEMs were shared between the 165204L_vs_FN41L and 165204S_vs_FN41S, while 66 and 48 DEMs were exclusively associated with 165204L_vs_FN41L and 165204S_vs_FN41S, respectively. Nevertheless, the number of DEMs common in 165204S_vs_165204L and FN41S_vs_FN41L were 193, much larger than the 87 and 57 DEMs unique to 165204S_vs_165204L and FN41S_vs_FN41L (**Figure 3**). We also noticed that these DEMs were from distinct classes and mainly constituted flavonoids, phenolic acids, and amino acids and derivatives, suggesting that there were a variety of primary and secondary metabolites involved in different tissue and dissimilarity between species. KEGG pathway analysis among the DEMs revealed that KEGG pathways, including carbon metabolism, flavone and flavonol biosynthesis, flavonoid biosynthesis, phenylalanine metabolism, and phenylpropanoid biosynthesis were significantly enriched in the compared groups (**Figure 4** and **Supplementary Table 4**). These results implied that the DEMs related to the flavone and flavonol biosynthesis, flavonoids biosynthesis, and phenylpropanoids biosynthesis are likely to play important roles in the different cultivars.

Transcriptome Sequencing Revealed Differentially Expressed Genes in the Different Cultivars

To better understand the molecular basis of the metabolic differences detected in the different cultivars, transcriptome sequencing was performed using the stem and leaf tissues. A total of 104.52 Gb of clean reads was generated from the 12 libraries after removing the adaptor sequences and low-quality reads. The percentage of the high-quality score (Q30) was more than 93.92%, GC contents varied from 51.02 to 55.12%, and the successfully mapped ratio was more than 86.03% (**Supplementary Table 5**). The correlation coefficients between the biological replicates of the same tissues were greater than 0.88 (**Supplementary Figure 4**). These results indicated the high quality of the sequencing data. We carried out an evaluation of differentially expressed genes (DEGs) via the four pair-wise comparison groups (165204S_vs_FN41S, 165204L_vs_FN41L, 165204S_vs_165204L, and FN41S_vs_FN41L). The analysis revealed that 165204S_vs_165204L had the largest number of DEGs, consisting of 11,575, of which 6,628 were up-regulated and 4,947 were down-regulated (**Figure 5**). Whereas a lower

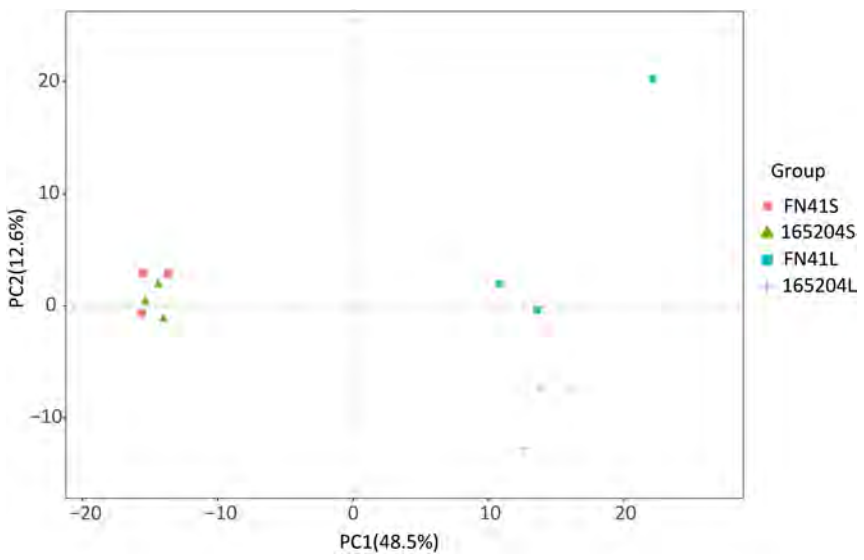


FIGURE 2 | Principal component analysis (PCA) of the metabolites detected in the sugarcane stem and leaf samples with three biological replicates. "FN41L" and "165204L" represent the leaf tissues of sugarcane varieties FN41 and 165204, respectively; "FN41S" and "165204S" represent the stem tissues of sugarcane varieties FN41 and 165204, respectively, similarly hereinafter.

number of DEGs was identified in 165204S_vs_FN41S, with 2,837, of which 1,938 were up-regulated and 899 were down-regulated. The comparison of FN41S_vs_FN41L resulted in 10,757 DEGs, including 6,419 up-regulated and 4,338 down-regulated. The comparison of 165204L_vs_FN41L revealed a total of 9,597 DEGs, among which 6,115 were up-regulated and 3,482 DEGs were down-regulated. The detailed information about the diversity of DEGs is available in **Supplementary Table 6**. The results of DEGs between different comparison groups indicated that the gene expression profiles varied significantly between these two different sugarcane species. To confirm the transcriptome data from RNA-Seq, 26 DEGs were selected

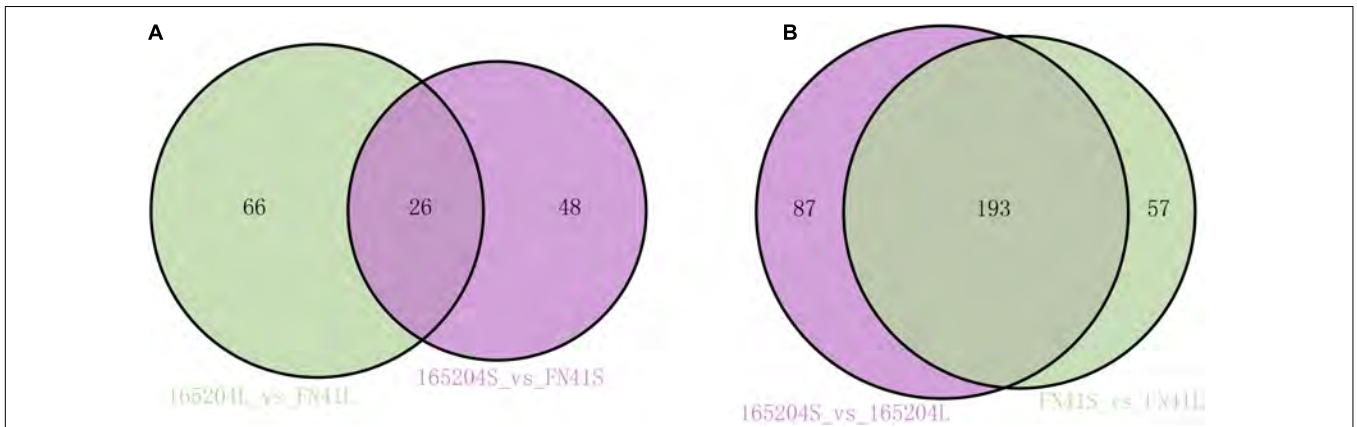


FIGURE 3 | Venn diagrams illustrating differential metabolism between two varieties in the same compartments **(A)**, and in the stems and leaves of the same variety **(B)**.

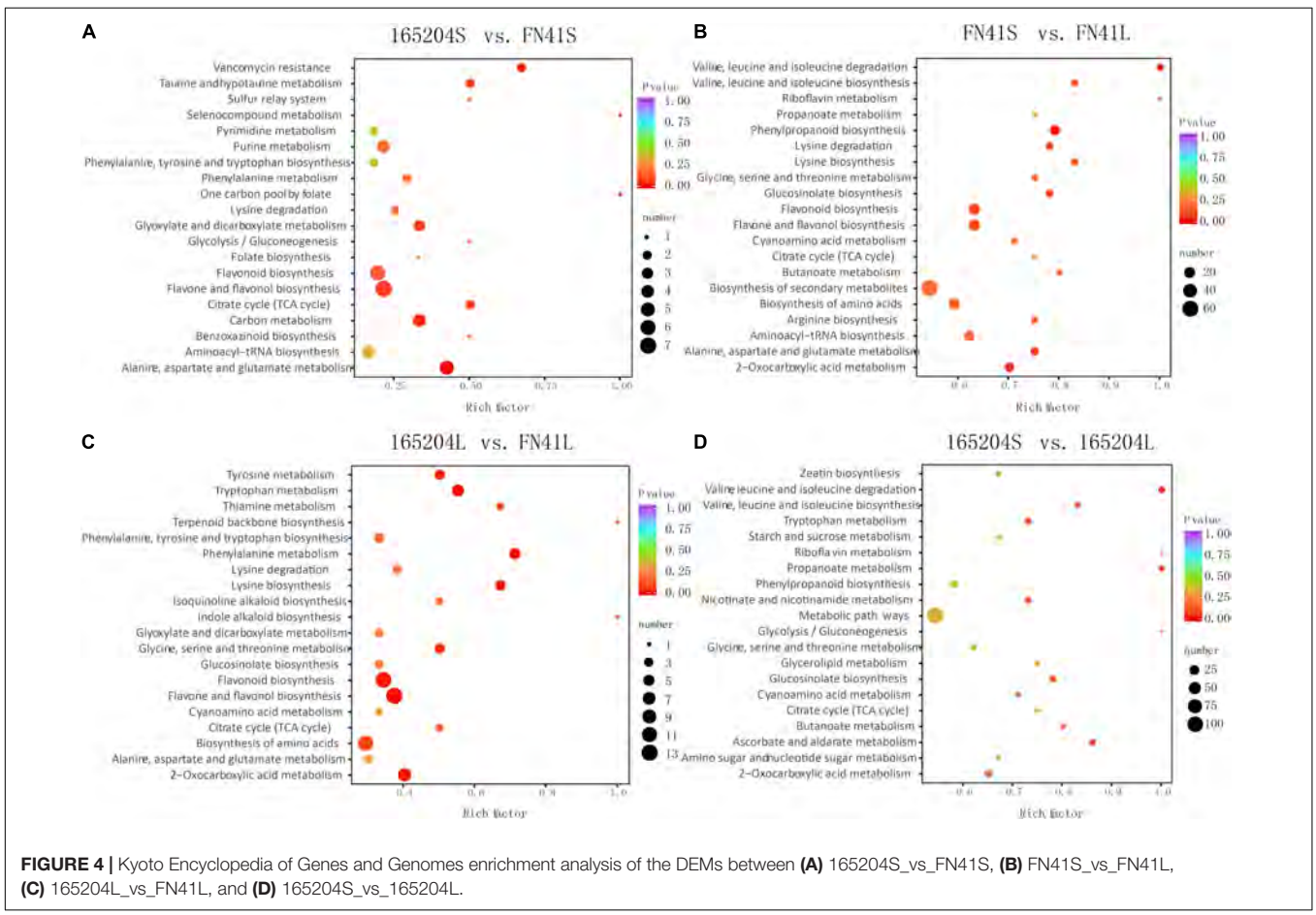


FIGURE 4 | Kyoto Encyclopedia of Genes and Genomes enrichment analysis of the DEMs between **(A)** 165204S_vs_FN41S, **(B)** FN41S_vs_FN41L, **(C)** 165204L_vs_FN41L, and **(D)** 165204S_vs_165204L.

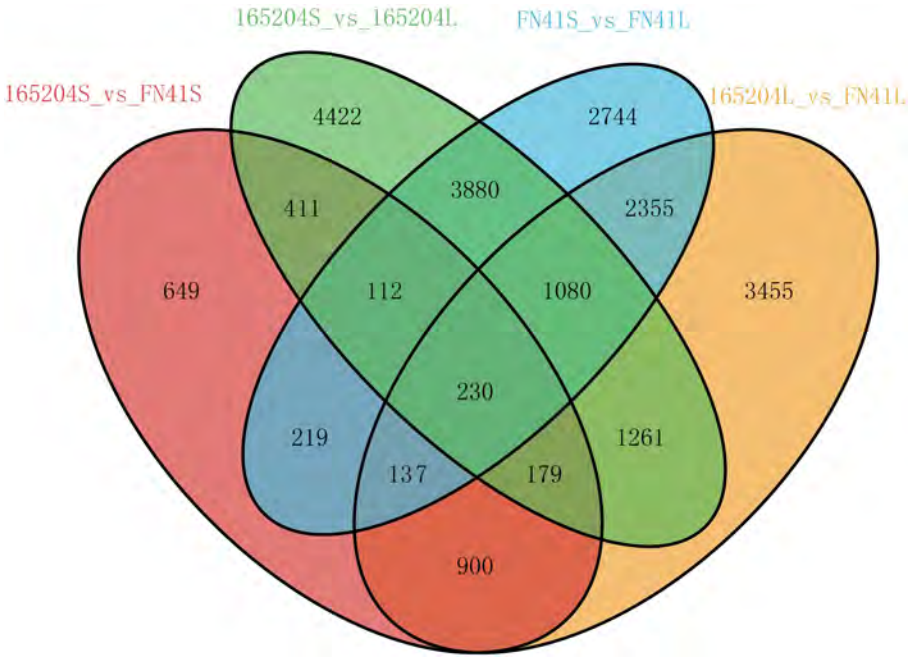


FIGURE 5 | Venn diagram of the DEGs of the four comparison groups.

randomly for qRT-PCR analysis (**Figure 6**). The qRT-PCR results were similar to the gene expression profiles in the transcriptome data, suggesting the transcriptome results were reliable.

Association of Metabolic Analysis and Transcriptomic Analysis

To minimize the false positives single-omics analysis, the integrated analysis of KEGG pathway enrichment, functional analysis and correlation analyses were performed between the transcriptome and metabolome. All the DEMs and DEGs were mapped to the KEGG pathway database to identify the main biological pathways to better understand the relationship between genes and metabolites. KEGG enrichment analysis showed that pathways enriched in at least one omics data were 5, 15, 29, and 27 KEGG pathways (p-value < 0.05) in the 165204S_vs_FN41S, 165204L_vs_FN41L, 165204S_vs_165204L, and FN41S_vs_FN41L pair-wise comparison groups, respectively. The DEGs and DEMs were mainly enriched in phenylalanine metabolism, flavonoid biosynthesis, flavone and flavonol biosynthesis, starch and sucrose metabolism, glycine, serine and threonine metabolism, carbon metabolism, and citrate cycle (TCA cycle). Later, four pathways including phenylpropanoid biosynthesis (ko00940), flavonoid biosynthesis (ko00941), flavone and flavonol biosynthesis (ko00944), and starch and sucrose metabolism (ko00500) were selected for subsequent analysis to explore the potential links between the metabolome and the transcriptome data.

To further identify modules related to phenylpropanoid biosynthesis (ko00940), flavonoid biosynthesis (ko00941), flavone and flavonol biosynthesis (ko00944), and starch and sucrose metabolism (ko00500), the significantly changed phenolic acids, flavonoids and saccharides were combined with RNA-seq data to construct a co-expression network (**Figure 7A** and **Supplementary Table 7**). Twenty modules (labeled in different colors) were identified in the dendrogram, where the gray module represents genes that were not assigned to specific modules. Remarkably, the purple module showed a significant correlation with the accumulation pattern of phenylpropanoid biosynthesis ($r > 0.85$ or $r < -0.85$, $p < 0.001$), while the pink module showed a significant correlation with the accumulation pattern of flavone and flavonol biosynthesis ($r > 0.9$ or $r < -0.9$, $p < 0.001$). Whereas the yellow module showed a significant correlation with the accumulation pattern of starch and sucrose metabolism ($r > 0.89$ or $r < -0.79$, $p < 0.001$) (**Figure 7B**). Among these genes, 490 genes of the purple module were positively related to p-coumaraldehyde, sinapic acid, caffeic acid and coniferyl alcohol. We also noticed that 1,002 genes of the yellow module were negatively related to D-fructose-6P, D-glucose-6p, and α-D-glucose-1P. 595 genes of the pink module were positively related to kaempferin, nicotiflorin, and vitexin 2"-*O*-rhamnoside.

Based on the number of connections between genes in the co-expression network, the top 50 node genes in the purple, pink and yellow modules were selected to generate the co-expression subnetwork (**Figure 7C** and **Supplementary Table 7**). Among these hub genes, we found six transcription factors in the three modules, namely, Tify (Sspon.05G0031500-1C) and NAC (Sspon.06G0028920-1C), in the purple module, followed MYB-related (Sspon.01G0014260-1T) and C2C2-Dof (Sspon.04G0023660-4P), in the yellow module, and WRKY (Sspon.03G0003750-3C) and bHLH (Sspon.06G0010740-1A), in

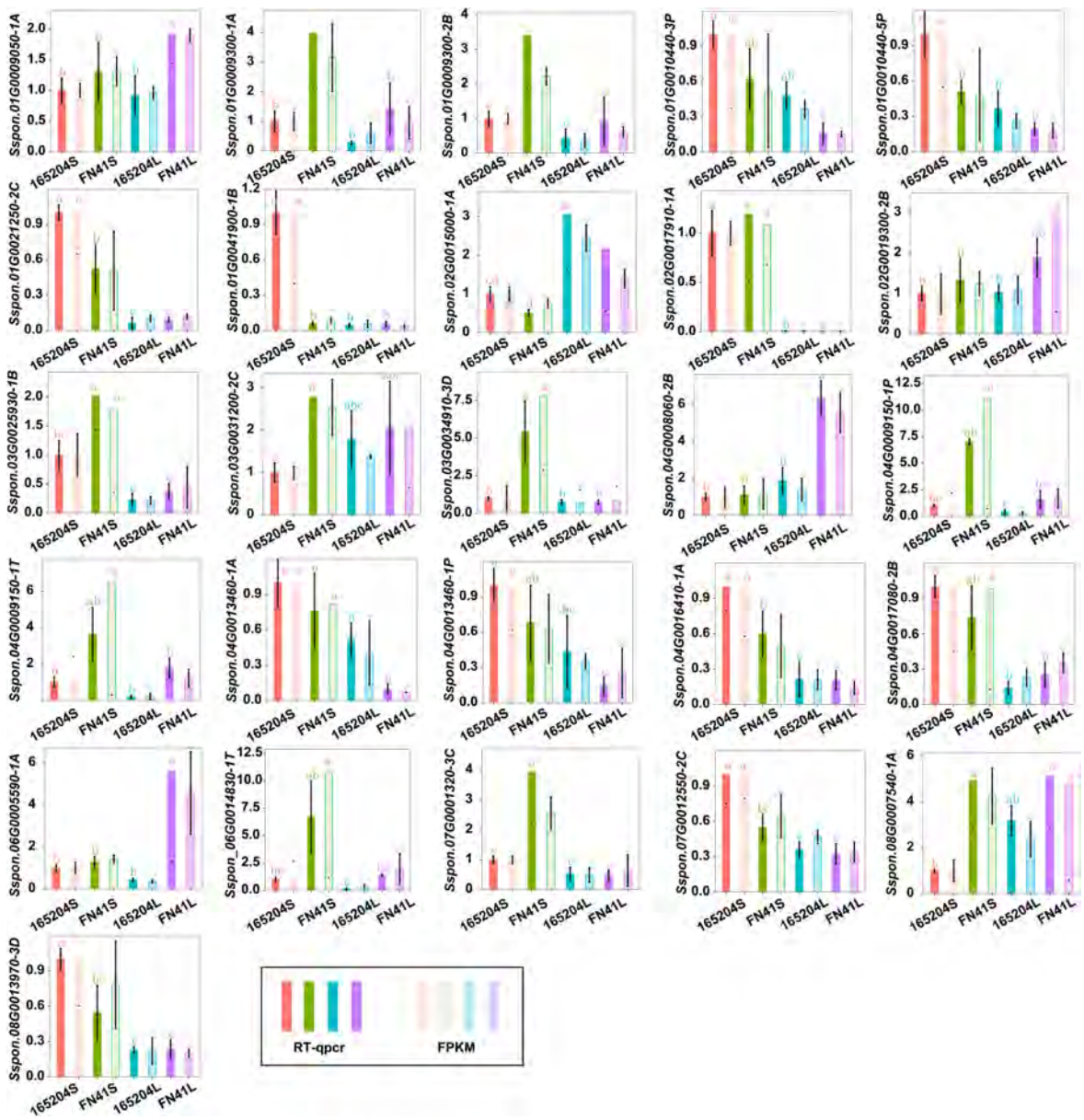


FIGURE 6 | RT-qPCR analysis showing the significance level of various genes identified in the different tissues of cane.

the pink module. These transcription factors play a key role in phenylpropanoid biosynthesis, flavone and flavonol biosynthesis, and starch and sucrose metabolism.

Integrating Related Genes and Metabolites in the Phenylpropanoid and Flavonoid Biosynthesis Pathway

To elucidate the differences in the production of phenylpropanoids and flavonoids metabolism between the two sugarcane species, we identified and mapped the DEGs and DEMs that were predicted to be involved in the phenylpropanoid and flavonoid biosynthesis (**Figure 8**). We observed that a total of 16 DEMs were mapped to these pathways, including eight phenylpropanoid biosynthesis, four flavonoid biosynthesis four flavone and flavonol biosynthesis. The profiles of DEMs between the two tissues showed the content of *p*-coumaraldehyde, caffeoyl quinic acid, coniferaldehyde, coniferyl alcohol, sinapyl alcohol, pinobanksin, naringin, vitexin, nicotiflorin, kaempferin, and vitexin were more evident in the leaves than the stems. While the accumulation of L-phenylalanine, tyrosine, sinapinaldehyde, and neohesperidin in the stems were higher than that in the leaves. We also compared the DEMs between the two cultivars, it was observed that the precursors of the phenylpropanoid biosynthesis pathway (L-phenylalanine and Tyrosine) were more abundant in FN41S and FN41L than that in 165204S

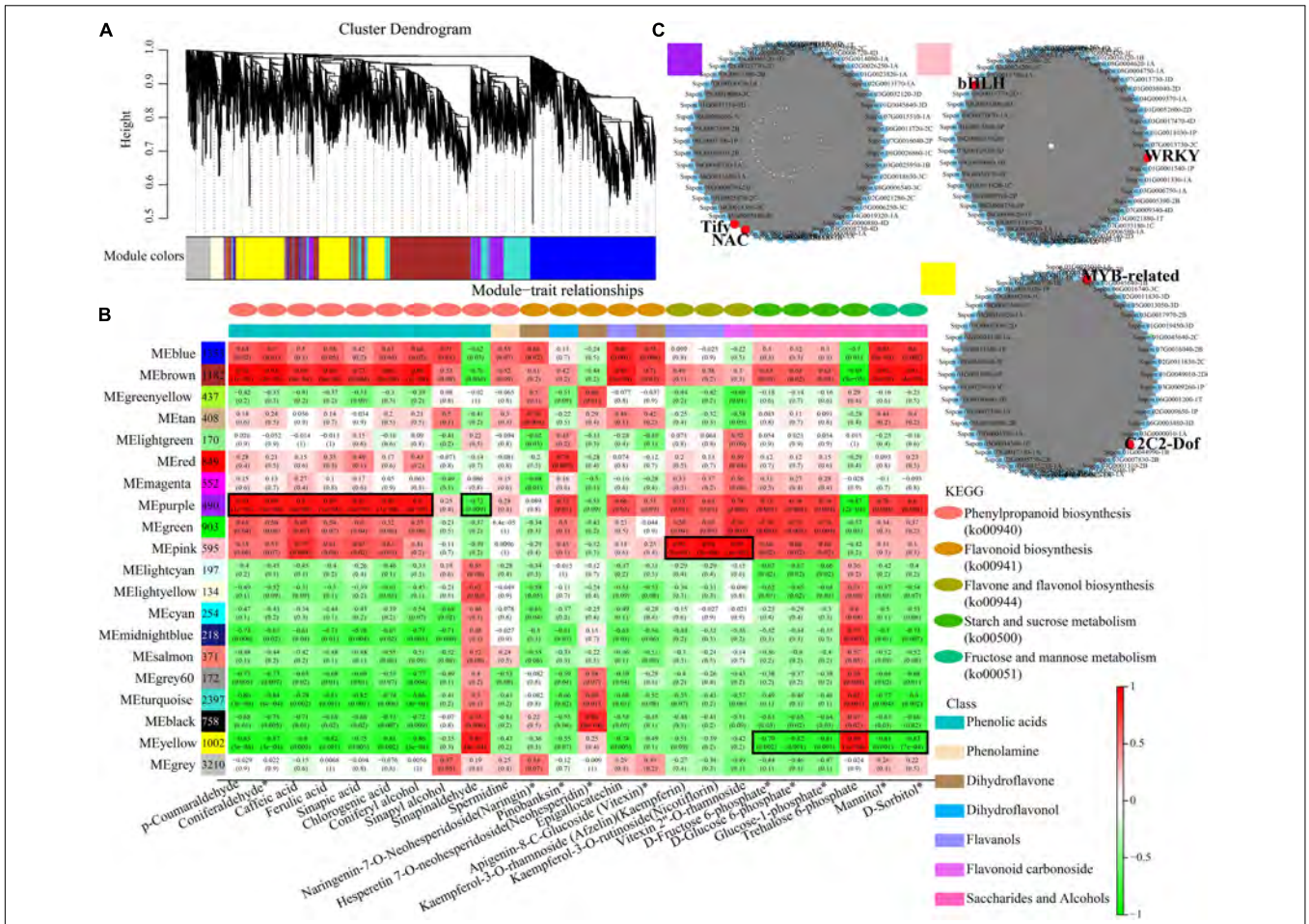


FIGURE 7 | Co-expression network analysis. **(A)** Hierarchical cluster tree showing 20 modules obtained by weighted gene co-expression network analysis (WGCNA). The gray modules represent genes that are not divided into specific modules. Each branch in the tree points to a gene. **(B)** Matrix of module-metabolite associations. Combining the gene expression profile data of stem and leaf tissues of different sugarcane varieties and the change patterns of phenylpropanoid biosynthesis, flavone and flavonol biosynthesis, starch and sucrose metabolism, displayed by the WGCNA analysis. The number of genes in each module is shown in the left box, followed by correlation coefficient and *p*-value between modules and metabolites, which are displayed at the intersection of rows and columns. **(C)** Co-expression sub-network analysis of purple, pink, and yellow modules related to the accumulation of phenylpropanoid biosynthesis, flavone and flavonol biosynthesis, starch and sucrose metabolism. The first 50 nodes of purple, pink, and yellow modules to build the network were selected, and transcription factors are shown in red.

and 165204L, while sinapaldehyde, pinobanksin, and kaempferin, and nictoflorin followed the same trend. The expression of pathway genes was also affected across different tissues. The majority of DEGs were observed to be both up-regulated and down-regulated, such as PAL, 4CL, HCT, CCR, CAD, C3H, POD, and CHI. However, some DEGs exhibited unique expression profiles in a specific species or tissue (specific expression data is shown in **Supplementary Table 8**).

A Pearson's correlation coefficients (PCCs) analysis was performed to measure the degree of correlation between DEGs and DEMs. In the phenylpropanoid and flavonoid biosynthesis pathways, the results of the PCC calculation showed that forty-three DEGs were significantly associated with eleven DEMs, with the vast majority demonstrating positive association. Specifically, 43 pairs were significantly and positively correlated (PCC value $>$ 0.8), whereas 27 pairs revealed significant and negative correlations (PCC value $<$ −0.8). Among them, the number of differential genes associated with the coniferaldehyde were more (23 DEGs), followed by epigallocatechin (18 DEGs) (**Figure 9**). Besides, Sspon.03G0012160-2B was associated with the metabolites, namely, coniferaldehyde and spermidine, and demonstrated a highly positive correlation, followed by sinapic acid, *p*-Coumaraldehyde and chlorogenic acid, exhibiting a significant and positive correlation with Sspon.08G0002670-2B. Whereas Sspon.01G0001310-3P and Sspon.03G0020600-2B were significantly and positively correlated with epigallocatechin and ferulic acid, respectively.

Integrating Related Genes and Metabolites in the Starch and Sucrose Metabolism

The DEGs and DEMs of starch and sucrose metabolism in the FN41 and 165204 of the two tissues were investigated

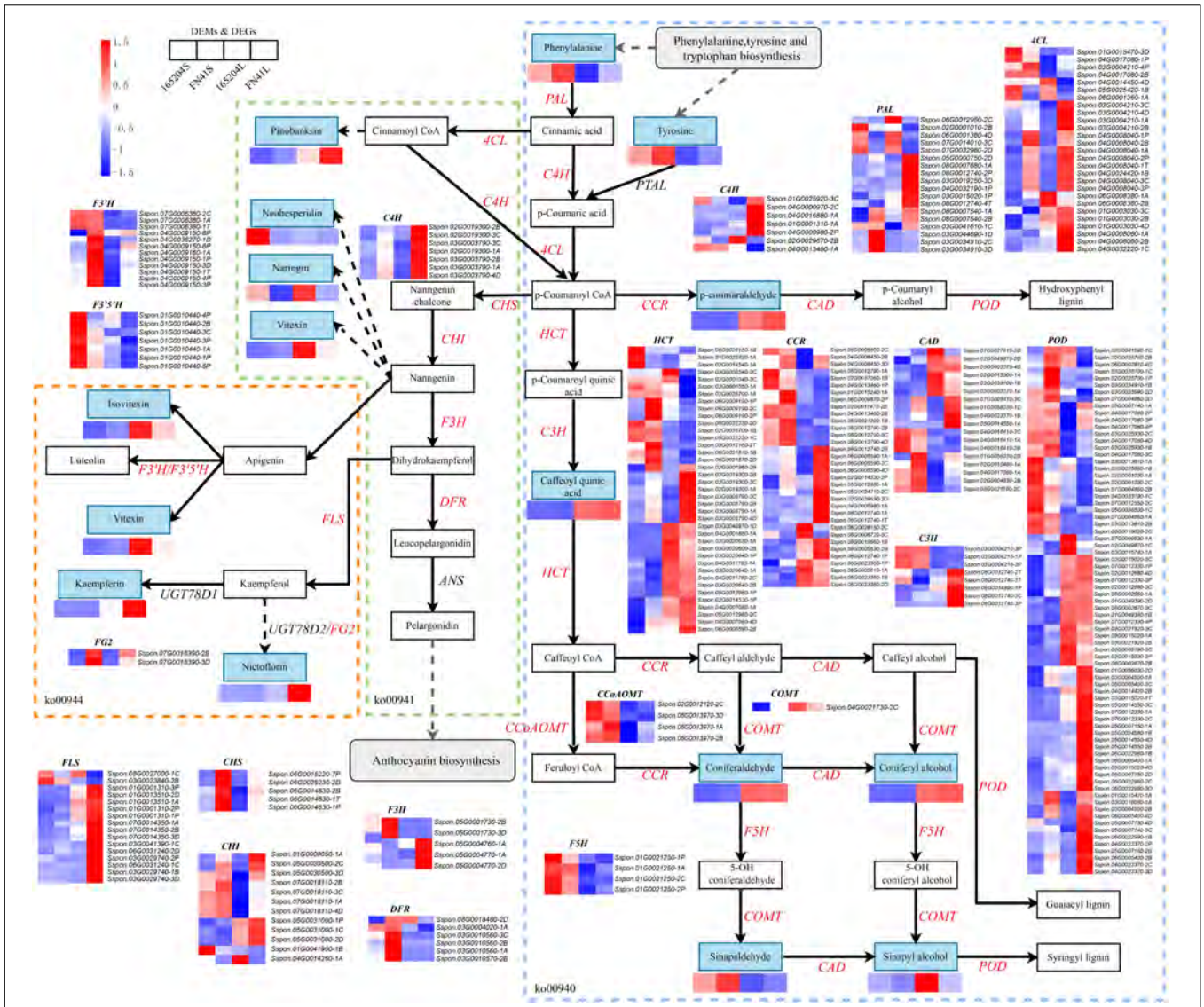


FIGURE 8 | Diagram of phenylpropanoid and flavonoid biosynthesis pathways with their related DEGs and DEMs. PAL, phenylalanine ammonia lyase; PTAL, phenylalanine/tyrosine ammonia-lyase; C4H, cinnamate4-hydroxylase; 4CL, 4-coumarate CoA ligase; HCT, hydroxycinnamoyl CoA shikimate/quinate hydroxycinnamoyl transferase; C3H, P-coumarate 3-hydroxylase; COMT, caffeic acid *O*-methyltransferase; CCoAOMT, caffeoyl-CoA *O*-methyltransferase; CHI, chalcone isomerase; CCR, cinnamoyl-CoA reductase; CAD, cinnamyl alcohol dehydrogenase; F5H, ferulate5-hydroxylase; F3′5′H, flavonoid 3′,5′-hydroxylase; POD, peroxidase; FLS, flavonol synthase; DFR, dihydroflavonol 4-reductase; Non-significant DEGs are shown in black. The solid line indicates the metabolic reactions in only one step. The dash line presents more than one step of the metabolic reaction.

using heat maps (**Figure 10**). Four DEMs were mapped including D-Fructose-6p, D-Glucose-6p, α-D-Glucose-1p and trehalose-6p. D-Fructose-6p, D-Glucose-6p and α-D-Glucose-1p were more pronounced in the leaves than the stems, while trehalose-6p exhibited the contrary. Interestingly, the content of D-fructose-6p, D-glucose-6p, and α-D-glucose-1p were more pronounced in FN41L and FN41S, respectively. Furthermore, we identified 315 DEGs classified into 26 gene families with diverse regulation patterns, highlighting the complex regulation of sugar metabolism in sugarcane (specific expression data is shown in **Supplementary Table 9**). We identified eighteen INV DEGs, the majority of which were significantly up-regulated in FN41 as compared to 165204 cultivar. Besides, two malZ genes (Sspon.03G0016860-1A&-2B) were found up-regulated in FN41, while the other five malZ genes (Sspon.08G0003790-3C&-1P, Sspon.08G0018720-1B&-2D, and Sspon.08G0025470-1C) were significantly down-regulated in FN41 In this study, we identified 11 SuS DEGs, all were up-regulated in the stems of the two cultivars. Among them, four SuS genes (Sspon.01G0009300-1A&-2B&-3C and Sspon.01G0011850-4D) were up-regulated in FN41 as compared to 165204, while the remaining SuS genes were down-regulated in FN41, indicating that SuS was negatively correlated with sugar accumulation. Three SPS DEGs and four SPP DEGs were also identified, of which two SPS

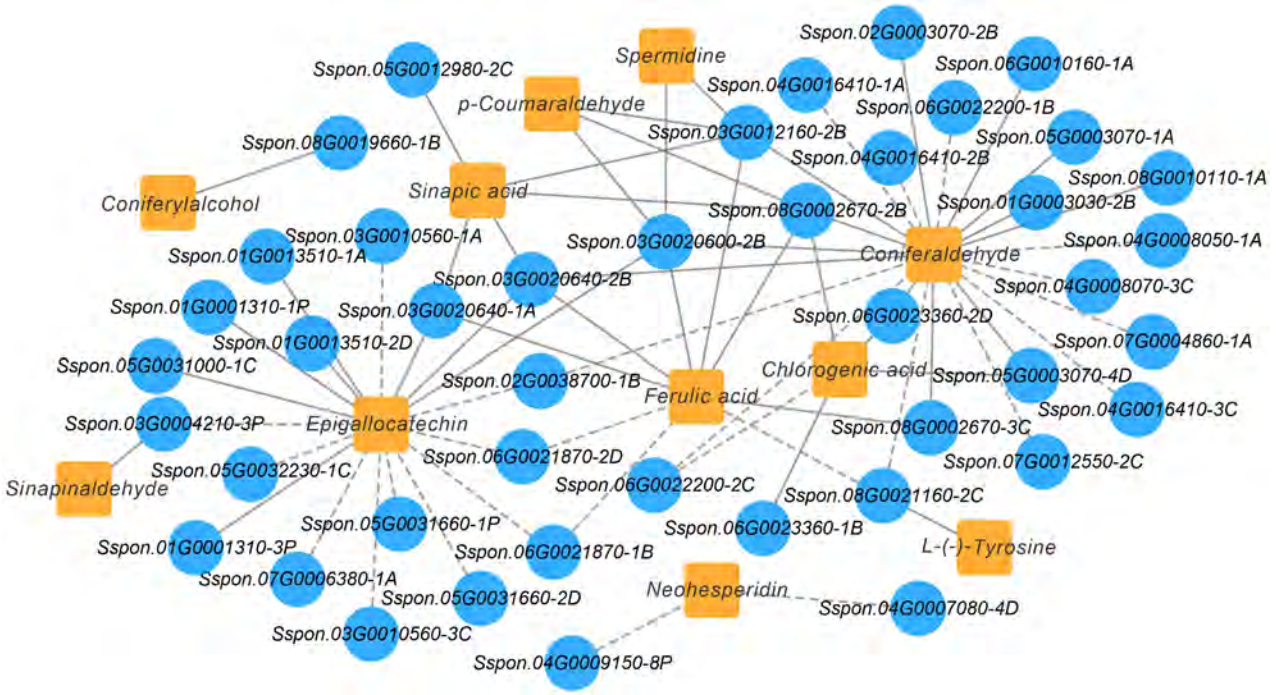


FIGURE 9 | Co-expression analysis illustrating genes and metabolites in phenylpropanoid and flavonoid biosynthesis pathway. Nodes represent genes or metabolites, and edges represent relationships between any two genes. Edges with solid and dashed lines represent positive and negative correlations, respectively, as determined by a Pearson correlation coefficient > 0.8 or < –0.8, respectively.

(Sspon.03G0028140-2C&-1P) and two SPP (Sspon.07G0026460-1B and Sspon.04G0011960-1A) were upregulated in FN41, which indicate a higher sucrose formation in FN41.

It is worth noting that the results of PCC analysis between genes and metabolites showed that in the starch and sucrose metabolism pathway, trehalose 6-phosphate was the only metabolite significantly associated with 47 DEGs, of which 31 pairs were significantly and positively correlated, while 16 pairs were significantly and negatively correlated (**Supplementary Figure 5**). Among them, Sspon.02G0017170-2C, Sspon.02G0015810-2B, and Sspon.02G0017910-1A showed a significant and positive correlation with trehalose 6-phosphate.

DISCUSSION

Metabolomics is an effective tool for measuring metabolite composition of various plant tissues. Targeted and untargeted metabolomics techniques have also been used to identify and quantify metabolites present in different organs of plant species during different development stages (Wang et al., 2018; Xiao et al., 2021). In this study, a targeted metabolomic approach was adopted to investigate the metabolic changes in sugarcane stems and leaves of two contrasting cultivars, FN41 and 165204. A total of 512 metabolites from 11 classes were detected in the sugarcane stems and leaves. This finding was confirmed with the study conducted by Glassop et al. (2007) in which they identified 121 and 71 metabolites in cultivar TMS and TBS using metabolomics tool, respectively. Studies have revealed that flavonoids play a vital role in plant tissues color formation, plant development and food quality (Xu et al., 2015). Flavonoids are also the largest and the most studied group of plant phenols with variable phenolic structures which could be further divided into flavones, isoflavones, flavonols, flavanols, flavanones, and anthocyanins (Panche et al., 2016). For instance, Yang et al. (2022) revealed that flavonoids in vegetative tissues of rice were observed to be abundant. Correspondingly, we found that flavonoids were the most dominant metabolites in the two sugarcane cultivars, exhibiting distinct distribution patterns in the various plant tissues. KEGG pathway enrichment analysis further showed that a large number of metabolites, namely, phenylpropanoid biosynthesis, flavonoid biosynthesis, and flavone and flavonol biosynthesis pathways were enriched in the DEMs. This finding is in consonance with a previous study, wherein it documented that metabolites such as phospholipids, amino acids and most lipids and fatty acids were enriched during rice seed germination (Yang et al., 2022), suggesting that flavonoids play an important role in distinguishing the two phenotypes of sugarcane.

Sucrose synthesis takes place in the cytoplasm of the leaf pulp of sugarcane, and the main rate-limiting enzyme is SPS. SPS catalyzes the irreversible conversion of uridine diphosphate glucose (UDPG) to fructose-6P to form sucrose-6'p which has been immediately catalyzed by SPP to form sucrose. SPS is also known to be the key enzyme in resynthesizing sucrose from the hexoses in the sink tissue (Huber and Huber, 1996). In this study, we noticed that two SPS genes (Sspon.03G0028140-2C&-1P)

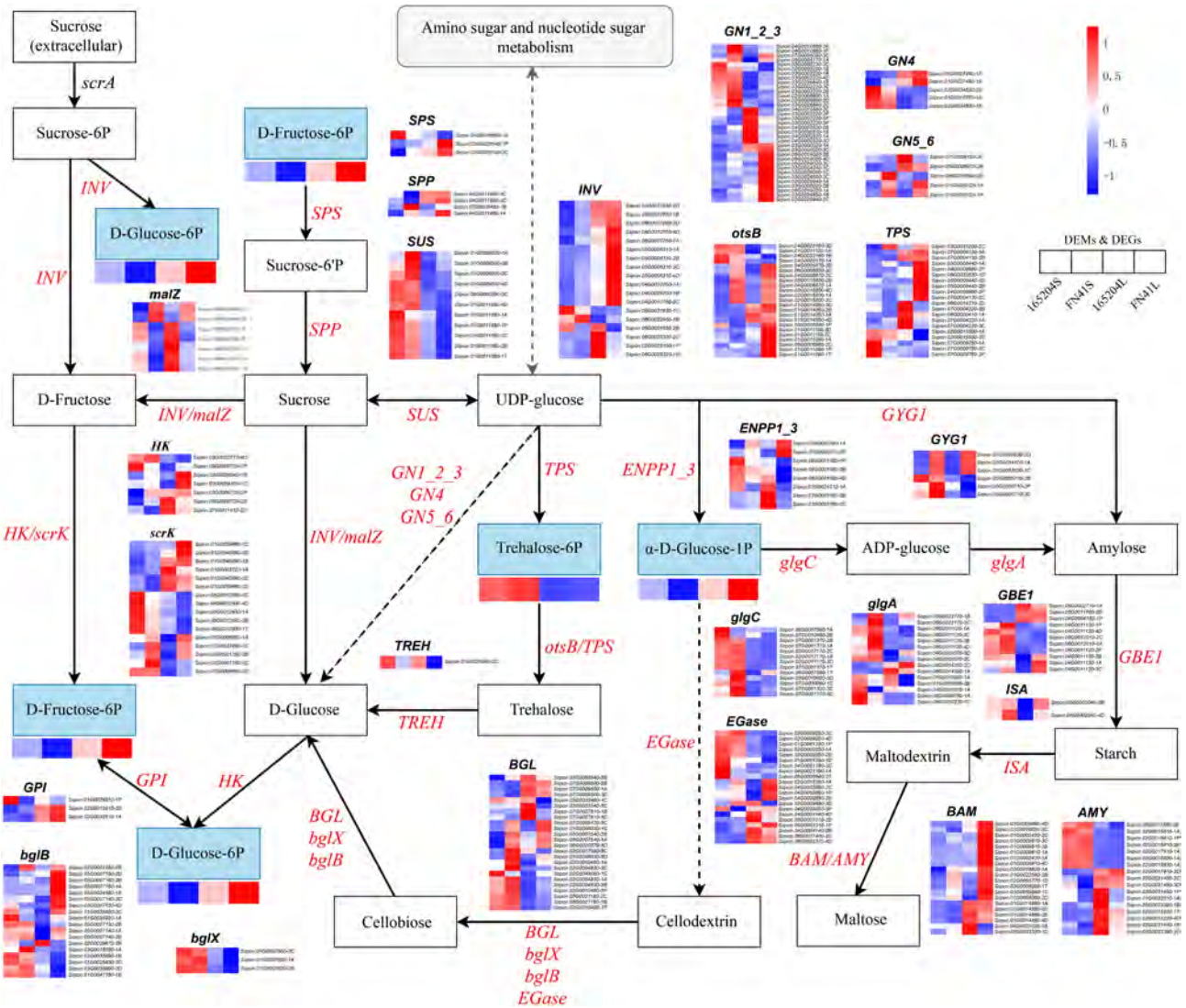


FIGURE 10 | Diagram of starch and sucrose metabolism pathway with their related DEGs and DEMs. SPS, sucrose phosphate synthase; SPS, sucrose phosphatase; SUS, sucrose synthase; INV, invertase; HK, hexokinase; scrK, phosphofructokinase; non-significant DEGs are shown in black. The solid line indicates the metabolic reactions in only one step. The dash line presents more than one step of the metabolic reaction.

and two SPP (Sspon.07G0026460-1B and Sspon.04G0011960-1A) that were up-regulated in FN41, which agreed with the finding reported by Nguyen-Quoc et al. (1999), wherein the overexpression of SPS gene in tomatoes results in increased sucrose loading and transport rate. Our result is also in line with the study conducted by Park et al. (2008). The authors mentioned that over-expression of an Arabidopsis SPS gene resulted in a considerable improvement of sink sucrose concentrations in tobacco (*Nicotiana tabacum* cv. Xanthi) plants. We therefore, postulated that the differences in sugar content among two sugarcane cultivars may largely be associated with the up-regulation of these genes in FN41.

Previous study has shown that the sugar content of the watermelon fruit is mainly determined by three enzyme families, sucrose synthase (SUS), SPS and insoluble acid convertase (IAI) (Liu et al., 2013). SUS can catalyze both sucrose synthesis and sucrose catabolism, but mainly convert UDP-glucose into sucrose (Schmölzer et al., 2016). The overexpression of a potato sucrose synthase gene in cotton enhances fiber production and sucrose supply by expanding the plant leaves (Xu et al., 2012). In tomato fruit, SUS contributes to the accumulation of glucose and fructose (Li et al., 2021). Nevertheless, in the present study 7 of the 11 SuS genes were down-regulated in FN41 and SuS was negatively correlated with sugar accumulation. This is probably due to the dual role of SUS, which also known to perform the catabolic function during sugarcane growth and development.

In plants, sucrose is irreversibly hydrolyzed by invertase (INV) into glucose and fructose (Koch, 2004). INV can be classified into cell-wall invertase (CWIN), vacuolar invertase (VIN), and cytoplasmic invertase (CIN) according to the cell

location (Wan et al., 2018). CWIN catalyzes cytoplasmic sucrose hydrolysis, which is involved in sucrose cytoplasmic unloading and hexose supply for development; VIN has an important function in hexose accumulation, cell expansion; and CIN contains cytoplasmic sugar homeostasis. In a previous research, 14 INV members were cloned in sugarcane, and VIN was induced by fructose treatment (Wang et al., 2017). In the present study, the majority of INV was upregulated in FN41 and four INV genes (Sspon.02G0025100-1P, Sspon.05G0001850-2B, and Sspon.06G0025320-1B&-2C) were downregulated in FN41. This may be due to the fact that different INVs play versatile roles in sugar metabolism and signaling in sugarcane. Hexokinase (HK) is a fructose and glucose phosphorylating enzyme, and also act as a sugar sensor that may regulate sugar-dependent gene repression or activation (Jang et al., 1997). HK catalyzes the first committed step of glucose metabolism by converting glucose to D-glucose-6p (Dai et al., 1995). The biosynthesis of trehalose is accomplished through trehalose 6-phosphate synthase (TPS) and trehalose 6-phosphate phosphatase (otsB), and trehalose plays a protective role against stress in plants (Paul, 2007). Metabolites including D-fructose-6P, D-glucose-6p, α-D-glucose-1P, and trehalose-6P were involved in the starch and sucrose metabolism of sugarcane. The distributions of metabolites suggested that FN41 synthesizes more monosaccharides in photosynthetic organs (source tissues) to convert these into other forms of carbohydrates and transport them for storage in heterotrophic cells (sink tissues). Therefore, we inferred the metabolites in starch and sucrose metabolism could be important, and the manipulation of these metabolites-related genes could provide prospects for increasing sugar content in sugarcane.

The phenylpropanoid pathway not only gives rise to flavonoids, but also converts them into lignin and various other aromatic metabolites such as coumarins, phenolic volatiles, or hydrolyzable tannins (Vogt, 2010). Phenylalanine and tyrosine are aromatic amino acids (AAAs) that are used for the synthesis of proteins. In plants, high carbon flux is committed to the biosynthesis of phenylalanine and tyrosine because they serve as precursors of numerous natural products, such as pigments, alkaloids, hormones, and cell wall components (Maeda and Dudareva, 2012). We noticed that phenylalanine and tyrosine were significantly up-regulated in FN41, implying that FN41 has more metabolic substrates for subsequent metabolic synthesis and may produce more energy in the sugarcane stems. Lignin is one of the most important secondary metabolites and one of the main components of the plant cell wall that play an important role in plant development such as enhancing the overall mechanical strength of plants, promoting transportation through the vascular bundles (Boerjan et al., 2003). Phenylpropanoids such as sinapyl alcohol, coniferyl alcohol, and coumaryl alcohol act as important precursors of lignin biosynthesis (Liu et al., 2018). In apples, it was shown that the reduced levels of sinapaldehyde and *p*-coumaryl alcohol ultimately led to significant lignin loss and growth retardation (Zhou et al., 2019). Recent studies have also shown that the cellulose content decreased while lignin content increased during pigmentation of winter jujube, and guaiacyl-syringyl (G-S) lignin was the main lignin type in the pericarp (Zhang et al., 2021). A precursor of *S*-lignin and sinapaldehyde, was found to be significantly up-regulated in the expression of FN41 stems in this study, which we believed had effects on the color and fiber composition of the stems, thereby promoting the synthesis of *S*-lignin. Cinnamate 4-hydroxylase (C4H), a cytochrome P450-dependent monooxygenase, catalyzes the first oxidative step of the phenylpropanoid pathway in higher plants by transforming *trans*-cinnamate into *p*-coumarate, which is a key substrate required for the formation of all flavonoids (Ayabe and Akashi, 2006). Plant growth and lignin accumulation were inhibited in the Arabidopsis C4H mutant (Schilmiller et al., 2009).

A number of studies have revealed that the leaf and stem of sugarcane are have strong relationship with source and sink (Roopendra et al., 2018). In this study, the leaf area, flavonoid index and chlorophyll index in FN41L were more pronounced than that of 165204L, which is the key to the significant difference of sugar accumulation in stems of FN41 and 165204. We also observed that the sugar content of FN41S was significantly higher than that of 165204S. This finding is in agreement with the study conducted by Roopendra et al. (2018), wherein it was revealed that the sucrose content of cane culm, possibly influenced by source–sink variation in sugarcane tissue. We believed that the relatively high amount of D-fructose 6-p, D-glucose6-p, and glucose1-p detected in FN41L may have been transported and distributed by source and sink of the plant, and a majority of them reached the stem of sugarcane FN41L, thereby promoting the high accumulation of sugar in FN41L.

Differential gene analysis provides us with a correlation of the possible gene functions at developmental stages based on the changes in the expression levels of DEGs. However, each gene in the differential analysis is isolated, whereas, in reality, genes and gene products are composed of regulatory networks to perform functions. WGCNA is a widely used systems biology method for describing the correlation patterns among genes across different samples that could be used to effectively screen specific modules of interest with highly related genes (Langfelder and Horvath, 2008). In this study, three modules related to phenylpropanoid biosynthesis (ko00940), flavonoid biosynthesis (ko00941), flavone and flavonol biosynthesis (ko00944), and starch and sucrose metabolism (ko00500) were identified, including hub genes and six transcription factors. In a previous study, C2C2-Dof zinc finger family were found differentially expressed between immature and mature tissues in the high-fiber sugarcane only. They were also influenced cellulose and lignin metabolism as well as the prominent players in carbon metabolism (Lakshmi et al., 2018). In maize, *ZmDOF36* acted as a critical regulatory factor in starch synthesis, and could help devise strategies for modulating starch production in maize endosperm (Wu et al., 2019). The abnormal expression of *bHLH3* disrupts the balance of the network and redirects flavonoid metabolic flux in pale-colored fruits, resulting in differences in the levels and proportions of anthocyanins, flavones, and flavonols among differently colored mulberry fruits (Li et al., 2020). NAC domain-containing protein could be involved in many biological processes such as secondary wall biosynthesis and abiotic stress response (Olsen et al., 2005).

The color and texture of sugarcane stems are not only important quality indicators but also the critical parameters that affect the consumer acceptance of fresh sugarcane products. Moreover, flavonoids are not only the main compounds that determine the color of flowers, fruits and leaves but also play an important role in plant growth, development, and plant adaptation to environment. Flavonoid metabolites and their associated genes in several plants have been comprehensively studied (Albert et al., 2014). Many species of plants start to change color after the activation of the flavonoid-related enzymes (Vogt, 2010). Vitexin, isovitexin, and pinobanksin are active components of many medicinal plants and have received increased attention as their wide range of pharmacological effects, such as antioxidant, antivirus, and antibacterial effects (He et al., 2016). The detection of these metabolites in sugarcane in the present study demonstrated the feasibility of extracting antioxidant substances from sugarcane leaves and stems. Naringin and neohesperidin have been reported to be responsible for the bitterness of citrus and are mainly influenced by its sugar content (Wang et al., 2016). In this study, the contents of naringin and neohesperidin were significantly downregulated in FN41, implying that these two metabolites may have played significantly contributed to the sweet taste of sugarcane. We also detected four anthocyanins in this study, while metabolites did not differ significantly between the samples. This may account for that the sugarcane stem rind and stem pith being sampled in a mixed sample, resulting in a non-significant difference in anthocyanin content. Chalcone synthase (CHS) was the first enzyme to be identified in flavonoid biosynthesis and located at the upstream point of the flavonoid biosynthesis pathway (Kreuzaler and Hahlbrock, 1972). Previous studies revealed that expression of MdCHS3 from apple in poplar resulted in reduced total lignin content and increased cell wall carbohydrate content in transgenic poplar cell walls (Mahon et al., 2021). Furthermore, silencing of the CHS gene could shift the anthocyanin pathway to the synthesis of chlorogenic acid and its complexes, and CHS is a key regulatory protein for anthocyanin biosynthesis in red and nectarine peaches (Rahim et al., 2014). In this study, all of the CHS genes were significantly up-regulated in FN41, demonstrating that CHS may play an important role in the synthesis of anthocyanin and lignin of two sugarcane cultivars. Flavanone 3-hydroxylase (F3H) converts naringen into dihydrokaempferol, and Dihydroflavonol-4-reductase (DFR) reduces dihydrokaempferol to leucoanthocyanidin, followed by oxidation of colorless leucoanthocyanidin to the precursor of anthocyanidins catalyzed by anthocyanin synthase (ANS) (Springob et al., 2003). These three enzymes are key enzymes in the synthesis of flavonol and anthocyanin. It was demonstrated in orange carnation that pigments synthesis is restricted when F3H expression was inhibited (Zuker et al., 2002). In grape berries, sugar-induced anthocyanin accumulation and F3H expression (Zheng et al., 2009). The purple leaf trait of ornamental kale was controlled by a gene BoPr encoding a DFR (Liu et al., 2017). Flavonoid 3′,5′ hydroxylases (F3′5′H) and (F3′H) are required for the biosynthesis of flavones, flavanones, flavonols, and anthocyanins, and has the potential to determine the pattern of flavonoid B-ring hydroxylation (Ayabe and Akashi, 2006). Our study revealed that F3′H and F3′5′H were upregulated in FN41 and 165402, respectively, triggering competition for substrates between F3′5′H and F3′H. In the biosynthesis of flavonoids and flavanols, the UGT78D family catalyzes glycosylation and occurs at the O-3 or O-7 position (Kc et al., 2018). We observed that flavonol-3-*O*-glucoside L-rhamnosyltransferase (FG2) was upregulated in FN41.

CONCLUSION

To conclude, we explored the molecular mechanism of differential sugar accumulation, rind color, and texture in two sugarcane cultivars. High sugar content was observed in FN41 as compared to 165204. Comparison of the differences in the level of metabolites and gene expression was performed. The analysis identified the metabolites and genes that have the potential to regulate sugar content, rind color, and texture in sugarcane. The results also suggested that genes such as C4H, CHS, F3H, F3′H, DFR, and FG2 in phenylpropanoid and flavonoid biosynthesis pathways may be a major factor impacting the rind color and contrasting texture of FN41 and 165204 sugarcane stems. Moreover, metabolites including L-phenylalanine, tyrosine, sinapaldehyde, pinobanksin, kaempferin, and nictoflorin were the potential drivers of phenotypic differences. Our findings also indicated that genes and metabolites in the starch and sucrose metabolism may have an important effect on sugar content in sugarcane. Overall, this study revealed molecular mechanisms underpinning the accumulation of sugar content, rind color, and texture of two sugarcane verities, which we believed is important for future sugarcane breeding programs and the selection of high biomass varieties. Up-regulated genes in FN41, namely, F3H, DFR, F3'H, and FG2 should be addressed in future studies to probe the specific mechanism.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The raw RNA-seq read data were deposited in the Short Read Archive (http://www.ncbi.nlm.nih.gov/sra/) and can be accessed using the BioProject ID: PRJNA805530.

AUTHOR CONTRIBUTIONS

All authors contributed to intellectual input and provided assistance to this study and manuscript preparation. ZY and ZP designed the research and conducted the experiments. FD analyzed the data and wrote the manuscript. YZ designed the qPCR primer and made RT-qPCR experiments. ZL, NF, and CH reviewed the manuscript. ZY supervised the work and approved the manuscript for publication.

FUNDING

This research was supported by China Agriculture Research System of MOF and MARA (CARS-170208), the Natural

Science Foundation of Fujian Province (2017J01456), the Special Foundation for Scientific and Technological Innovation of Fujian Agriculture and Forestry University (KFA17172A and KFA17528A), and the National Natural Science Foundation of China (31771723).

ACKNOWLEDGMENTS

The authors thank the staff of Wuhan Metware Biotechnology Co., Ltd. (Wuhan, China) for their support during metabolomic data analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.921536/full#supplementary-material

Supplementary Figure 1 | Diagram of the method of testing in tensile strength perpendicular to grain of wood.

Supplementary Figure 2 | Total ion chromatograms (TIC) under positive **(A)** and negative mode **(B)**.

Supplementary Figure 3 | Hierarchical clustering analysis of all metabolites detected in this study. The abscissa indicates three biological replicates of FN41stems (FN41S1, FN41S2, and FN41S3), 165204 stems (165204S1, 165204S2, and 165204S3), FN41 leaves (FN41L1, FN41L2, and FN41L3), and 165204 leaves (165204L1, 165204L2, and 165204L3), and the ordinate indicates the metabolites detected in this study. The red segments indicate a relatively high content of metabolites, while the blue segments indicate a relatively low content of metabolites.

Supplementary Figure 4 | Heat map depicting correlation between biological replicate.

Supplementary Figure 5 | Co-expression analysis of genes and metabolites in starch and sucrose metabolism pathway. Nodes represent genes or metabolites, and edges represent relationships between any two genes. Edges with solid and dashed lines represent positive and negative correlations, respectively, as determined by a Pearson correlation coefficient > 0.8 or <-0.8, respectively.

REFERENCES

Agati, G., and Tattini, M. (2010). Multiple functional roles of flavonoids in photoprotection. *New Phytol.* 186, 786–793.

Aitken, K., Jackson, P., and Mcintyre, C. (2006). Quantitative trait loci identified for sugar related traits in a sugarcane (*Saccharum* spp.) cultivar× *Saccharum officinarum* population. *Theor. Appl. Genet.* 112, 1306–1317. doi: 10.1007/s00122-006-0233-2

Albert, N. W., Davies, K. M., and Schwinn, K. E. (2014). Gene regulation networks generate diverse pigmentation patterns in plants. *Plant Signal. Behav.* 9:e29526. doi: 10.4161/psb.29526

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., et al. (2004). UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 32, D115–D119.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29.

Ayabe, S. I., and Akashi, T. (2006). Cytochrome P450s in flavonoid metabolism. *Phytochem. Rev.* 5, 271–282.

Babu, C., Koodalingam, K., Natarajan, U., Shanthi, R., and Govindaraj, P. (2009). Assessment of rind hardness in sugarcane (*Sachharum* spp. hybrids) genotypes for development of non lodging erect canes. *Adv. Biol. Res.* 3, 48–52.

Boerjan, W., Ralph, J., and Baucher, M. (2003). Lignin biosynthesis. *Annu. Rev. Plant Biol.* 54, 519–546.

Casu, R. E., Jarmey, J. M., Bonnett, G. D., and Manners, J. M. (2007). Identification of transcripts associated with cell wall metabolism and development in the stem of sugarcane by Affymetrix GeneChip Sugarcane Genome Array expression profiling. *Funct. Integr. Genomics* 7, 153–167. doi: 10.1007/s10142-006-0038-z

Chen, S., Lin, J., Liu, H., Gong, Z., Wang, X., Li, M., et al. (2018). Insights into tissue-specific specialized metabolism in Tieguanyin tea cultivar by untargeted metabolomics. *Molecules* 23:1817. doi: 10.3390/molecules23071817

Cho, K., Cho, K.-S., Sohn, H.-B., Ha, I. J., Hong, S.-Y., Lee, H., et al. (2016). Network analysis of the metabolome and transcriptome reveals novel regulation of potato pigmentation. *J. Exp. Bot.* 67, 1519–1533. doi: 10.1093/jxb/erv549

Dai, N., Schaffer, A. A., Petreikov, M., and Granot, D. (1995). *Arabidopsis thaliana* hexokinase cDNA isolated by complementation of yeast cells. *Plant Physiol.* 108, 879–880. doi: 10.1104/pp.108.2.879

Dong, X., Gao, Y., Chen, W., Wang, W., Gong, L., Liu, X., et al. (2015). Spatiotemporal distribution of phenolamides and the genetics of natural variation of hydroxycinnamoyl spermidine in rice. *Mol. Plant* 8, 111–121.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42, D222–D230.

Fraga, C. G., Clowers, B. H., Moore, R. J., and Zink, E. M. (2010). Signature-discovery approach for sample matching of a nerve-agent precursor using liquid chromatography- mass spectrometry, XCMS, and chemometrics. *Anal. Chem.* 82, 4165–4173. doi: 10.1021/ac1003568

Glassop, D., Roessner, U., Bacic, A., and Bonnett, G. D. (2007). Changes in the sugarcane metabolome with stem development. Are they related to sucrose accumulation? *Plant Cell Physiol.* 48, 573–584. doi: 10.1093/pcp/pcm027

Gong, C., Zhu, H., Lu, X., Yang, D., Zhao, S., Umer, M. J., et al. (2021). An integrated transcriptome and metabolome approach reveals the accumulation of taste-related metabolites and gene regulatory networks during watermelon fruit development. *Planta* 254:35. doi: 10.1007/s00425-021-03680-7

Granados-Sánchez, D., Ruíz-Puga, P., and Barrera-Escorcia, H. (2008). Ecología de la herbivoría. *Rev. Chapingo Ser. Cienc. For. Ambiente* 14, 51–63.

He, M., Min, J.-W., Kong, W.-L., He, X.-H., Li, J.-X., and Peng, B.-W. (2016). A review on the pharmacological effects of vitexin and isovitexin. *Fitoterapia* 115, 74–85. doi: 10.1016/j.fitote.2016.09.011

Huber, S. C., and Huber, J. L. (1996). Role and regulation of sucrose-phosphate synthase in higher plants. *Ann. Rev. Plant Biol.* 47, 431–444. doi: 10.1146/annurev.arplant.47.1.431

Ikonen, A., Tahvanainen, J., and Roininen, H. (2001). Chlorogenic acid as an antiherbivore defence of willows against leaf beetles. *Entomol. Exp. Appl.* 99, 47–54.

Jang, J. C., León, P., and Li, Z. (1997). Hexokinase as a sugar sensor in higher plants. *Plant Cell Online* 9, 5–19. doi: 10.1105/tpc.9.1.5

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32, D277–D280. doi: 10.1093/nar/gkh063

Kc, S., Liu, M., Zhang, Q., Fan, K., Shi, Y., and Ruan, J. (2018). Metabolic changes of amino acids and flavonoids in tea plants in response to inorganic phosphate limitation. *Int. J. Mol. Sci.* 19:3683. doi: 10.3390/ijms19113683

Koch, K. (2004). Sucrose metabolism: regulatory mechanisms and pivotal roles in sugar sensing and plant development. *Curr. Opin. Plant Biol.* 7, 235–246. doi: 10.1016/j.pbi.2004.03.014

Kohl, M., Wiese, S., and Warscheid, B. (2011). Cytoscape: software for visualization and analysis of biological networks. *Methods Mol. Biol.* 696, 291–303. doi: 10.1007/978-1-60761-987-1_18

Koonin, E. V., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Krylov, D. M., Makarova, K. S., et al. (2004). A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.* 5:R7. doi: 10.1186/gb-2004-5-2-r7

Kreuzaler, F., and Hahlbrock, K. (1972). Enzymatic synthesis of aromatic compounds in higher plants: formation of naringenin (5, 7, 4′-trihydroxyflavanone) from p-coumaroyl coenzyme A and malonyl coenzyme A. *FEBS Lett.* 28, 69–72. doi: 10.1016/0014-5793(72)80679-3

Lakshmi, K., Hoang, N. V., Agnelo, F., Botha, F. C., and Henry, R. J. (2018). Transcriptome analysis highlights key differentially expressed genes involved in cellulose and lignin biosynthesis of sugarcane genotypes varying in fiber content. *Sci. Rep.* 8:11612. doi: 10.1038/s41598-018-30033-4

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

Li, H., Yang, Z., Zeng, Q., Wang, S., Luo, Y., Huang, Y., et al. (2020). Abnormal expression of *bHLH3* disrupts a flavonoid homeostasis network, causing differences in pigment composition among mulberry fruits. *Hortic. Res.* 7:83. doi: 10.1038/s41438-020-0302-8

Li, N., Wang, J., Wang, B., Huang, S., Hu, J., Yang, T., et al. (2021). Identification of the carbohydrate and organic acid metabolism genes responsible for brix in tomato fruit by transcriptome and metabolome analysis. *Front. Genet.* 12:714942. doi: 10.3389/fgene.2021.714942

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

Lin, F., Rensing, C., Pang, Z., Zou, J., Lin, S., Letuma, P., et al. (2022). Metabolomic analysis reveals differential metabolites and pathways involved in grain chalkiness improvement under rice ratooning. *Field Crops Res.* 283:108521.

Liu, J., Guo, S., He, H., Zhang, H., Gong, G., Ren, Y., et al. (2013). Dynamic characteristics of sugar accumulation and related enzyme activities in sweet and non-sweet watermelon fruits. *Acta Physiol. Plant.* 35, 3213–3222.

Liu, Q., Luo, L., and Zheng, L. (2018). Lignins: biosynthesis and biological functions in plants. *Int. J. Mol. Sci.* 19:335. doi: 10.3390/ijms19020335

Liu, X.-P., Gao, B.-Z., Han, F.-Q., Fang, Z.-Y., Yang, L.-M., Zhuang, M., et al. (2017). Genetics and fine mapping of a purple leaf gene, BoPr, in ornamental kale (*Brassica oleracea* L. var. acephala). *BMC Genomics* 18:230. doi: 10.1186/s12864-017-3613-x

Maeda, H., and Dudareva, N. (2012). The shikimate pathway and aromatic amino Acid biosynthesis in plants. *Annu. Rev. Plant Biol.* 63, 75–105. doi: 10.1146/annurev-arplant-042811-105439

Mahon, E. L., De Vries, L., Jang, S.-K., Middar, S., Kim, H., Unda, F., et al. (2021). Exogenous chalcone synthase expression in developing poplar xylem incorporates naringenin into lignins. *Plant Physiol.* 188, 984–996. doi: 10.1093/plphys/kiab499

Mancini, M., Leite, D., Perecin, D., Bidóia, M., Xavier, M., Landell, M., et al. (2012). Characterization of the genetic variability of a sugarcane commercial cross through yield components and quality parameters. *Sugar Tech.* 14, 119–125.

Moore, P. H. (2005). Integration of sucrose accumulation processes across hierarchical scales: towards developing an understanding of the gene-to-crop continuum. *Field Crops Res.* 92, 119–135.

Nguyen-Quoc, B., N'tchobo, H., Foyer, C. H., and Yelle, S. (1999). Overexpression of sucrose phosphate synthase increases sucrose unloading in transformed tomato fruit. *J. Exp. Bot.* 50, 785–791.

Ni, Y., Chen, H., Liu, D., Zeng, L., Chen, P., and Liu, C. (2021). Discovery of genes involved in anthocyanin biosynthesis from the rind and pith of three sugarcane varieties using integrated metabolic profiling and RNA-seq analysis. *BMC Plant Biol.* 21:214. doi: 10.1186/s12870-021-02986-8

Olsen, A. N., Ernst, H. A., Leggio, L. L., and Skriver, K. (2005). NAC transcription factors: structurally distinct, functionally diverse. *Trends Plant Sci* 10, 79–87. doi: 10.1016/j.tplants.2004.12.010

Panche, A., Diwan, A., and Chandra, S. (2016). Flavonoids: an overview. *J. Nutr. Sci.* 5:e47

Park, J.-Y., Canam, T., Kang, K.-Y., Ellis, D. D., and Mansfield, S. D. (2008). Over-expression of an *Arabidopsis* family A sucrose phosphate synthase (SPS) gene alters plant growth and fibre development. *Transgen. Res.* 17, 181–192. doi: 10.1007/s11248-007-9090-2

Patti, G. J., Yanes, O., and Siuzdak, G. (2012). Metabolomics: the apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol.* 13, 263–269. doi: 10.1038/nrm3314

Paul, M. (2007). Trehalose 6-phosphate. *Curr. Opin. Plant Biol.* 10, 303–309.

Perlo, V., Botha, F. C., Furtado, A., Hodgson-Kratky, K., and Henry, R. J. (2020). Metabolic changes in the developing sugarcane culm associated with high yield and early high sugar content. *Plant Direct* 4:e00276. doi: 10.1002/pld3.276

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

Rae, A. L., Grof, C. P., Casu, R. E., and Bonnett, G. D. (2005). Sucrose accumulation in the sugarcane stem: pathways and control points for transport and compartmentation. *Field Crops Res.* 92, 159–168.

Rahim, M. A., Busatto, N., and Trainotti, L. (2014). Regulation of anthocyanin biosynthesis in peach fruits. *Planta* 240, 913–929. doi: 10.1007/s00425-014-2078-2

Rao, M. J., Ahmed, U., Ahmed, M. H., Duan, M., Wang, J., Wang, Y., et al. (2021). Comparison and quantification of metabolites and their antioxidant activities in young and mature leaves of sugarcane. *ACS Food Sci. Technol.* 1, 362–373.

Rao, M. J., Xu, Y., Huang, Y., Tang, X., Deng, X., and Xu, Q. (2019). Ectopic expression of citrus UDP-GLUCOSYL TRANSFERASE gene enhances anthocyanin and proanthocyanidins contents and confers high light tolerance in Arabidopsis. *BMC Plant Biol.* 19:603. doi: 10.1186/s12870-019-2212-1

Roopendra, K., Sharma, A., Chandra, A., and Saxena, S. (2018). Gibberellin-induced perturbation of source-sink communication promotes sucrose accumulation in sugarcane. *3 Biotech* 8:418. doi: 10.1007/s13205-018-1429-2

Sarker, U., and Oba, S. (2018). Augmentation of leaf color parameters, pigments, vitamins, phenolic acids, flavonoids and antioxidant activity in selected Amaranthus tricolor under salinity stress. *Sci. Rep.* 8:12349. doi: 10.1038/s41598-018-30897-6

Schaker, P. D., Peters, L. P., Cataldi, T. R., Labate, C. A., Caldana, C., and Monteiro-Vitorello, C. B. (2017). Metabolome dynamics of smutted sugarcane reveals mechanisms involved in disease progression and whip emission. *Front. Plant Sci.* 8:882. doi: 10.3389/fpls.2017.00882

Schilmiller, A. L., Stout, J., Weng, J.-K., Humphreys, J., Ruegger, M. O., and Chapple, C. (2009). Mutations in the cinnamate 4-hydroxylase gene impact metabolism, growth and development in Arabidopsis. *Plant J.* 60, 771–782. doi: 10.1111/j.1365-313X.2009.03996.x

Schmölzer, K., Gutmann, A., Diricks, M., Desmet, T., and Nidetzky, B. (2016). Sucrose synthase: a unique glycosyltransferase for biocatalytic glycosylation process development. *Biotechnol. Adv.* 34, 88–111. doi: 10.1016/j.biotechadv.2015.11.003

Springob, K., Nakajima, J.-I., Yamazaki, M., and Saito, K. (2003). Recent advances in the biosynthesis and accumulation of anthocyanins. *Nat. Prod. Rep.* 20, 288–303. doi: 10.1039/b109542k

Su, G., Morris, J. H., Demchak, B., and Bader, G. D. (2014). Biological network exploration with Cytoscape 3. *Curr. Protoc. Bioinformatics* 47, 8.13.11–18.13.24. doi: 10.1002/0471250953.bi0813s47

Sun, J.-Y., Gaudet, D. A., Lu, Z.-X., Frick, M., Puchalski, B., and Laroche, A. (2008). Characterization and antifungal properties of wheat nonspecific lipid transfer proteins. *Mol. Plant Microbe Interact.* 21, 346–360. doi: 10.1094/MPMI-21-3-0346

Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 28, 33–36. doi: 10.1093/nar/28.1.33

Thirugnanasambandam, P. P., Hoang, N. V., Furtado, A., Botha, F. C., and Henry, R. J. (2017). Association of variation in the sugarcane transcriptome with sugar content. *BMC Genomics* 18:909. doi: 10.1186/s12864-017-4302-5

Vogt, T. (2010). Phenylpropanoid biosynthesis. *Mol. Plant* 3, 2–20.

Wan, H., Wu, L., Yang, Y., Zhou, G., and Ruan, Y.-L. (2018). Evolution of sucrose metabolism: the dichotomy of invertases and beyond. *Trends Plant Sci.* 23, 163–177. doi: 10.1016/j.tplants.2017.11.001

Wang, L., Zheng, Y., Ding, S., Zhang, Q., Chen, Y., and Zhang, J. (2017). Molecular cloning, structure, phylogeny and expression analysis of the invertase gene family in sugarcane. *BMC Plant Biol.* 17:109. doi: 10.1186/s12870-017-1052-0

Wang, S., Tu, H., Wan, J., Chen, W., Liu, X., Luo, J., et al. (2016). Spatio-temporal distribution and natural variation of metabolites in citrus fruits. *Food Chem.* 199, 8–17. doi: 10.1016/j.foodchem.2015.11.113

Wang, Y., Zhang, X., Yang, S., and Yuan, Y. (2018). Lignin involvement in programmed changes in peach-fruit texture indicated by metabolite and transcriptome analyses. *J. Agric. Food Chem.* 66, 12627–12640. doi: 10.1021/acs.jafc.8b04284

Wang, Z., Song, Q., Shuai, L., Htun, R., Malviya, M. K., Li, Y., et al. (2020). Metabolic and proteomic analysis of nitrogen metabolism mechanisms involved in the sugarcane–Fusarium verticillioides interaction. *J. Plant Physiol.* 251:153207. doi: 10.1016/j.jplph.2020.153207

Welbaum, G. E., and Meinzer, F. C. (1990). Compartmentation of solutes and water in developing sugarcane stalk tissue. *Plant Physiol.* 93, 1147–1153. doi: 10.1104/pp.93.3.1147

Wijma, M., Lembke, C. G., Diniz, A. L., Santini, L., Zambotti-Villela, L., Colepicolo, P., et al. (2021). Planting season impacts sugarcane stem development, secondary metabolite levels, and natural antisense transcription. *Cells* 10:3451. doi: 10.3390/cells10123451

Wu, J., Chen, L., Chen, M., Zhou, W., Dong, Q., Jiang, H., et al. (2019). The DOF-domain transcription factor *ZmDOF36* positively regulates starch synthesis in transgenic maize. *Front. Plant Sci.* 10:465. doi: 10.3389/fpls.2019.00465

Xiao, L., Cao, S., Shang, X., Xie, X., Zeng, W., Lu, L., et al. (2021). Metabolomic and transcriptomic profiling reveals distinct nutritional properties of cassavas with different flesh colors. *Food Chem.* 2:100016. doi: 10.1016/j.fochms.2021.100016

Xu, S.-M., Brill, E., Llewellyn, D. J., Furbank, R. T., and Ruan, Y.-L. (2012). Overexpression of a potato sucrose synthase gene in cotton accelerates leaf expansion, reduces seed abortion, and enhances fiber production. *Molecular Plant* 5, 430–441. doi: 10.1093/mp/ssr090

Xu, W., Dubos, C., and Lepiniec, L. (2015). Transcriptional control of flavonoid biosynthesis by MYB–bHLH–WDR complexes. *Trends Plant Sci.* 20, 176–185. doi: 10.1016/j.tplants.2014.12.001

Yang, C., Shen, S., Zhou, S., Li, Y., Mao, Y., Zhou, J., et al. (2022). Rice metabolic regulatory network spanning the entire life cycle. *Mol. Plant* 15, 258–275. doi: 10.1016/j.molp.2021.10.005

Zhang, J., Zhang, X., Tang, H., Zhang, Q., Hua, X., Ma, X., et al. (2018). Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. *Nat. Genet.* 50, 1565–1573.

Zhang, Q., Wang, L., Wang, Z., Zhang, R., Liu, P., Liu, M., et al. (2021). The regulation of cell wall lignification and lignin biosynthesis during pigmentation of winter jujube. *Hortic. Res.* 8:238. doi: 10.1038/s41438-021-00670-4

Zheng, Y., Li, T., Liu, H., Pan, Q., Zhan, J., and Huang, W. (2009). Sugars induce anthocyanin accumulation and flavanone 3-hydroxylase expression in grape berries. *Plant Growth Regul.* 58, 251–260.

Zhou, K., Hu, L., Li, Y., Chen, X., Zhang, Z., Liu, B., et al. (2019). MdUGT88F1-mediated phloridzin biosynthesis regulates apple development and canker resistance. *Plant Physiol.* 180, 2290–2305.

Zhu, Z.-J., Schultz, A. W., Wang, J., Johnson, C. H., Yannone, S. M., Patti, G. J., et al. (2013). Liquid chromatography quadrupole time-of-flight mass spectrometry characterization of metabolites guided by the METLIN database. *Nat. Protoc.* 8, 451–460. doi: 10.1038/nprot.2013.004

Zuker, A., Tzfira, T., Ben-Meir, H., Ovadis, M., Shklarman, E., Itzhaki, H., et al. (2002). Modification of flower color and fragrance by antisense suppression of the flavanone 3-hydroxylase gene. *Mol. Breed.* 9, 33–41.

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

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

frontiers | Frontiers in Plant Science

REVIEW
published: 23 June 2022
doi: 10.3389/fpls.2022.839588

Genetic Determinants of Biomass in C_4 Crops: Molecular and Agronomic Approaches to Increase Biomass for Biofuels

*Noor-ul- Ain[1], Fasih Ullah Haider[2], Mahpara Fatima[1], Habiba[3], Yongmei Zhou[1] and Ray Ming[4]**

[1] Fujian Provincial Key Laboratory of Haixia Applied Plant Systems, FAFU and UIUC-SIB Joint Center for Genomics and Biotechnology, College of Crop Sciences, Fujian Agriculture and Forestry University, Fuzhou, China, [2] College of Resources and Environmental Sciences, Gansu Agricultural University, Lanzhou, China, [3] Fujian Provincial Key Laboratory of Plant Functional Biology, College of Life Science, Fujian and Agriculture and Forestry University, Fujian, China, [4] Department of Plant Biology, The University of Illinois at Champaign-Urbana, Champaign, IL, United States

Bio-based fuels have become popular being efficient, cost-effective, and eco-friendly alternatives to fossil fuels. Among plant sources exploited as feedstocks, C_4 grasses, such as sugarcane, maize, sorghum, and miscanthus, are highly resourceful in converting solar energy into chemical energy. For a sustainable and reliable supply of feedstocks for biofuels, we expect dedicated bioenergy crops to produce high biomass using minimum input resources. In recent years, molecular and genetic advancements identified various factors regulating growth, biomass accumulation, and assimilate partitioning. Here, we reviewed important genes involved in cell cycle regulation, hormone dynamics, and cell wall biosynthesis. A number of important transcription factors and miRNAs aid in activation of important genes responsible for cell wall growth and re-construction. Also, environmental components interacting with genetic controls modulate plant biomass by modifying gene expression in multiple interacting pathways. Finally, we discussed recent progress using hybridization and genome editing techniques to improve biomass yield in C_4 grasses. This review summarizes genes and environmental factors contributing biomass yield in C_4 biofuel crops which can help to discover and design bioenergy crops adapting to changing climate conditions.

Keywords: biomass accumulation, C_4 crops, hormone dynamics, cell wall growth, circadian rhythms, fossil fuels

OPEN ACCESS

Edited by:
Xingtan Zhang, Agricultural Genomics Institute at Shenzhen (CAAS), China

Reviewed by:
Sajid Fiaz, The University of Haripur, Pakistan
Xiaomin Feng, Guangdong Academy of Sciences, China

***Correspondence:**
Ray Ming
rayming@illinois.edu

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

Received: *20 December 2021*
Accepted: *17 January 2022*
Published: *23 June 2022*

Citation:
Ain N-u, Haider FU, Fatima M, Habiba, Zhou Y and Ming R (2022) Genetic Determinants of Biomass in C_4 Crops: Molecular and Agronomic Approaches to Increase Biomass for Biofuels. Front. Plant Sci. 13:839588. doi: 10.3389/fpls.2022.839588

INTRODUCTION

The world population is expected to reach 9 billion by 2050, which was only ∼1.9 billion in the twentieth century (Mullet, 2017). Rapid population expansion and rise in energy demands have heightened research interests to build more sustainable, cost-effective, and eco-friendly energy sources. The gradual effects of using rapidly depleting finite fossil fuels turned out in global

Abbreviations: ASEAN, Association of South East Asian Nations; eCO_2, Equivalent CO_2; GHGs, Greenhouse Gases; mg/h, megagrams per hectare; PEP, Phosphoenolpyruvate; RUBISCO, Ribulose Bisphosphate Carboxylase/Oxygenase; SQUAMOSA, SQUAMOSA Promoter-Binding Protein-Like (SPL); TFs, Transcription factors.

warming (greenhouse effect) and subsequent climate change that is ongoing. Owing to intensive industrialization and urbanization, the level of greenhouse gases (GHGs) has increased by 50 times in the atmosphere, which is among the primary drivers of climate change (Subramaniam et al., 2020). International Energy Agency predicted that from 2020 to 2025, the mean annual near-surface temperature would increase by 1°C with a range of 0.9–1.8°C in comparison with pre-industrial time-temperature level (time spanning from 1850 to 1900) along with the onset of frequent tropical cyclones (IEA, 2018). To minimize the disastrous effects of climate change and fulfill energy demands, International Energy Agency (IEA) has devised to exploit renewable energy sources. Major renewable energy resources, such as wind, solar, plant biomass, ocean, and hydropower being sustainable sources, can also aid in reducing the use of fossil fuels and consequently GHGs. In the year 2019, a 12.5% increase in biofuel production was observed from 142.6 to 160.9 million liters, whereas in 2020 due to the global pandemic situation, compromised prices of fossil crude oil have made transport biofuels less competitive. However, future predictions suggest that the average global output of bioethanol will further rise to 182 billion liters which was 160 million liters in 2019, whereas the United States and ASEAN region will contribute for biodiesel and Hydro-treated vegetable oil (HVO) (Lauf et al., 2021).

Biofuels are the fuels derived from biological substances (e.g., agricultural wastes, animal matter, algae, forest vegetation, and energy crops) (Koçar and Civaş, 2013). In this review, we will primarily focus on the high biomass yielding energy crops which are commercially used for energy generation. Presently, regarding biofuel crops, there is a concern of conflict about their use for food, feed, and energy generation. However, several plant species are efficient in accumulating high biomass with minimal inputs. Among them, ryegrass (*Lolium perenne*), bamboo (*Bambusa vulgaris*), poplar (*Populus deltoides*), and willow (*Salix*) are from C_3 category of photosynthesis while giant miscanthus (*Miscanthus giganteus*), sweet sorghum (*Sorghum bicolor*), pearl millet (*Pennisetum glaucum*), napier grass (*Pennisetum purpureum*), maize (*Zea mays*), sugarcane (*Saccharum*), and switchgrass (*Panicum virgatum*) are ideal C_4 energy crops (Byrt et al., 2011). C_4 grasses take more advantage of biomass accumulation, owing to higher energy-conserving photosynthetic machinery, stress tolerance, water, and nitrogen use efficiency (Somerville et al., 2010). C_4 photosynthesis system possesses specialized biochemistry and anatomical modifications that protect the oxygenation of RUBISCO. Whereas, the PEP enzyme, instead of RUBISCO serves as the first substrate of CO_2 in mesophyll cells that reduces the energy losses caused by photorespiration (van der Weijde et al., 2013). This structural collaboration of mesophyll (M) and bundle sheath (BS) enables C_4 plants to harvest more solar energy with improved water use efficiency (WUE) and nitrogen use efficiency (NUE).

In a couple of decades, the increasing trend of biofuel use in developed countries motivated many researchers to focus their interests on biofuel crops and their bio-products. Among C_4 crops, Miscanthus and switchgrass have been extensively studied for this purpose in the United States and Europe. Sugarcane is contributing a major share of biofuels in Brazil. Sorghum is also a promising competitor as a bio-energy crop owing to drought tolerance and genetic diversity in sweet and grain sorghum (Byrt et al., 2011). Advancement in recent technologies of saccharification and lignocellulose digestion, cultivation of sugarcane is widely practiced as a biofuel crop (Carvalho-Netto et al., 2014). Scientists are more concerned with introducing specialized hybrids or transgene as energy crops to meet the objectives of sustainable energy by biofuels. Plant biomass depends on genetic, physiological, edaphic and environmental factors. Few publications encompassed basis of biomass accumulation at genetic level but a comprehensive study of genetic, physiological and environmental factors constituting to biomass, was lacking (Endo et al., 2009; Byrt et al., 2011; Lima et al., 2017). Therefore, besides biological and genetic basis we addressed environmental modulations-based plant biomass accumulation. Considering the general overview of the need for bioenergy and the importance of biomass crops with special concern for C_4 crops, this review aims at identifying the genes involved in different growth-related processes and how growth patterns are modified in changed climatic conditions. Furthermore, possible tools and strategies are discussed that are effective in opting for increasing biomass in C_4 crops.

GENETIC BASIS FOR BIOMASS

In plants, several genes belonging to different functional and structural categories are involved in vegetative development. Owning to advancement in molecular and genetic studies, many genes and transcription factors have been identified in contributing growth from juvenile to vegetative stage as previously covered in the reviews (Demura and Ye, 2010; Lima et al., 2017; Kandel et al., 2018). Following sections supplicated the existing literature about genetic aspects of plant growth regulation.

Growth and Developmental Regulation

Cell Cycle Machinery

Plant organ size control is a central component of biomass productivity. In many animals and plants, overall organ growth rate and size are associated with cell number that is controlled by strict action of cell division together with cell expansion (Vernoux et al., 2000). As in other eukaryotes, the cell cycle in plants consists of DNA-replication (S-phase), and mitosis (M-phase), which are separated by postmitotic (G1) and pre-mitotic interphase (G2) gap phases, respectively (Scofield et al., 2014). Cell cycle machinery is strongly modulated at different points to verify the fidelity of chromosome duplication and cell division. Highly conserved control mechanisms, the checkpoints G1/S and G2/M transitions confirm that either cell cycle process has been precisely accomplished at each phase before entering the next phase or not (Barnum and O'Connell, 2014). Different core cell cycle protein groups including CYCLINS (CYCs) complexed with CYCLIN-DEPENDENT KINASES (CDKs), the E2F/DIMERISATION PROTEIN (DP) transcriptional regulatory proteins, KIP-RELATED PROTEIN/INTERACTOR

OF CDKs (KRP/ICK), RETINOBLASTOMA-RELATED (RBR), SIAMESE/SIAMESE-RELATED (SIM/SMR), proteins and the multi-subunit E3 ubiquitin ligase ANAPHASE-PROMOTING COMPLEX/CYCLOSOME (APC/C) control the progression of events involved different phases (Inz, 2006; Inagaki and Umeda, 2011). Genetic modulation of these proteins to enhance biomass has been reported in model plant species (*Arabidopsis*, tobacco, rice), with few C_4 plants (sorghum and maize). Elevated growth rate in tobacco has been resulted from overexpressing the cyclin D-type (*CycD2*) gene from *Arabidopsis*. These plants were found to have normal cell and meristem size but taller stem overall, showing increased growth rate from seedling to maturity (Boucheron et al., 2005). Defective shoot and root formation, as well as a reduction in endoreduplication, were noticed in tobacco ectopically expressing *CycA3; 2* (Yu et al., 2003). Transcriptome analysis identified elevated levels of cell cycle (i.e., cyclins) genes in bioenergy sorghum immature internodes which shows that their initial increase in size is due to cell-division coupled growth (Kebrom et al., 2017). Further, quadruple (*ick1/krp1*, *ick2/krp2*, *ick6/krp3*, *ick7/krp4*) and quintuple (*ick1/krp1*, *ick2/krp2*, *ick6/krp3*, *ick7/krp4*, *ick5/krp7*) mutants of CDK negative regulator *ICK/KRP* genes reported to have stimulated CDK activity and cell proliferation that resulted in increased fresh and dry weights; larger cotyledons; leaves; petals and seeds (Cheng et al., 2013). Overexpression of novel *Arabidopsis* ABAP1 protein decreased cell proliferation by limiting mitotic DNA replication in negative feedback loop during leaf development, by repressing transcription of pre-replication complex (pre-RC) genes (Masuda et al., 2008).

Another essential component of cell cycle machinery is the anaphase-promoting complex (APC), a multi-subunit E3 ligase that modulates cyclins (Cyc) and CDKs activity in checkpoints to ensure the maintenance of cell division rate (Bodrug et al., 2021). APC/C-subunits remain conserved throughout evolution, however, gene duplication of different subunits has been observed in some plants (de Lima et al., 2010). Overexpression of *Arabidopsis CDC27a/APC3a* in tobacco was associated with apical meristem restructuration, altered cell-cycle marker expression, and accelerated plant growth up to 30% at the flowering time leading to increased biomass production (Rojas et al., 2009). While *APC10* overexpression in *Arabidopsis* causes CYCB1;1 protein degradation, thereby accelerating the transition through mitosis (Eloy et al., 2011). Transgenic tobacco plants overexpressing the *APC10* gene are taller with larger leaves, produce more seed capsules, and have an augmented biomass accumulation. Furthermore, a cross between *APC3a*- and *APC10*- overexpressing tobacco T1 plants have enhanced growth phenotype compared to the overexpression of single *APC/C* subunits (de Freitas Lima et al., 2013). Down-regulation of rice *OsCCS52A, an* APC/C subunit resulted in reduced plant height and smaller seeds with an endosperm defect in endoreduplication (Su'udi et al., 2012). Semi-dwarfism and reduced leaf size are also observed in *CCS52A* ortholog rice *tillering and dwarf 1* (*tad1*) mutant (Xu et al., 2012). SAMBA negatively modulates cell proliferation through APC/C interaction. In maize, *samba-1* and *samba-3* mutants showed developmental defects, involving short plant height, reduced leaf size due to an altered cell expansion and cell division rate (Gong et al., 2021). In addition, several DRP family members like DRP1A, DRP1E, DRP2A, DRP2B, and DRP5B are regarded as SAMBA interactors. All of these proteins, except DRP5B, are localized to the cell plate and mutations in *DRP1E* and *DRP1A* resulted in defective cell plate assembly and cytokinesis, as well as defects in cell expansion (Hong et al., 2003; Kang et al., 2003; Fujimoto et al., 2008).

Long Vegetative Duration

The plant life cycle is divided into vegetative, transition, and reproductive developmental phases. The vegetative phase is associated with meristems producing stems and leaves. The transition phase is related to an elevation of the apical meristem, while the reproductive phase centers on meristems capable of producing flowers or reproductive organs. The vegetative phase starts at germination and continues through tillering, the meristems actively produce a stem, buds, internodes, and leaves. The long vegetative phase establishes continuous leaf development needed to capture sunlight for photosynthesis that supplies nutrients for expansion of roots for anchoring, storage, and uptake of minerals for increased biomass production. Some high-yielding C_4 crops are *Miscanthus x giganteus*, *Sorghum bicolor*, *Pennisetum,* and sugarcane (*Saccharum* spp.) genotypes, characterized by enriched canopies, taller stems, and longer growing seasons (Somerville et al., 2010; Mullet et al., 2014). The biomass yield of *Miscanthus* × *giganteus* in some mid-west United States locations reached 44–61 Mg/ha at peak biomass accumulation (Heaton et al., 2008). *Pennisetum purpureum* and *Pennisetum typhoides* reached their record yields of ∼88 and 80 Mg/ha, respectively during longer growing seasons (Somerville et al., 2010). Similarly, high-biomass first-generation sorghum hybrids accumulated ∼40–50 dry Mg/ha during ∼180 days growing season when grown in the south-central United States (Olson et al., 2012). Biomass varies among types of C_4 crops for various reasons like stem sink strength, shoot/root partitioning, and season length. For example, *Miscanthus* × *giganteus* produced higher biomass than switchgrass and maize when these were grown in similar regions in the United States due to differences in shoot/root biomass partitioning (Anderson et al., 2011). High-biomass sorghum hybrids with long growing seasons produce twice shoot biomass (∼40–50 Mg/ha) when compared to grain sorghum hybrids in optimum growth conditions and ∼30% more biomass in rain-fed conditions when both hybrids are grown in similar regions in the south-central United States (Olson et al., 2012; Truong et al., 2017). The increased biomass yield of high-biomass sorghum hybrids was because of delayed flowering initiation resulting in prolonged vegetative growth duration that increased total light energy capture, improved radiation interception and use efficiency, and elevated biomass partitioning.

Further, delayed flowering in long days concomitant with an extensive period of vegetative growth resulted in increased biomass yield as observed in several C_4 grasses. Photoperiod regulated flowering in sorghum is extensively studied by modulating flowering regulators florigen related genes (*CN8, CN12, CN15*), upstream activators (*CONSTANS*

(CO) and *EARLY HEADING DATE 1* (*EHD1*)), and repressors *(PRR37* and *GHD7)* of these genes that are regulated by photoperiod and output from the circadian clock, once sorghum exits in juvenile phase (Murphy et al., 2011, 2014; Dong et al., 2012). Prolonged vegetative meristem activity with increased biomass yield was observed in several plants by overexpressing flowering-time genes (Demura and Ye, 2010). Indeed, the activation of the flowering promoting factor-like1, flowering locus T1, C-like MADS-box protein, early flowering 3 as well as embryo flowering 1-like protein in tomato IL2-6 cultivar, supported late-maturing performance (Caruso et al., 2016). Overexpression of gibberellin 20-oxidase-1 and ARGOS also resulted in an extended growing period by delaying the flowering time to give rise to larger organ size and taller plants (Hu et al., 2003; Voorend et al., 2016). Regulation switch from vegetative to reproductive phase can be controlled by manipulating genes from several developmental pathways, for example, gibberellin, circadian, and flowering-related genes.

Hormones Dynamics and Primary Growth

Plant hormones are a diverse group of chemical substances controlling growth and development-related events in plants by regulating meristematic cell division and cell elongation. These chemical signals modulate microtubule and cell plate formation, cell wall constituent deposition, and remodeling which are key factors of growth and thus biomass accumulation. During the green revolution, scientists exploited the traditional plant breeding approaches for the selection of short stature, higher grain yield producing cultivars, which were low in levels of endogenous hormones like gibberellin (Sánchez-Rodríguez et al., 2010), auxin (Vanneste and Friml, 2009), and brassinolide (Müssig, 2005) that are important as growth regulators and performing growth regulatory functions from cellular to developmental levels (**Table 1**).

Gibberellin

In Gibberellin (GA) signaling pathways, manipulation of both positive and negative regulators employed positive effects on growth and biomass accumulation. One control point of biomass can be the substantial increase of GA rate-limiting enzyme GIBBERELLIN 20-OXIDASE (GA 20-OXIDASE) which is involved in the last steps of GA biosynthesis in the cytoplasm. One of the primitive functional evidence of *GA20ox* in *Arabidopsis* highlights the accelerated shoot growth, elongated hypocotyls, and onset of early flowering (Coles et al., 1999). In a potential biofuel crop, switchgrass, ectopic expression of *ZmGA20ox* resulted in increased growth and biomass-related traits (Do et al., 2016). Recently, nine genes of *GA20-ox1* were identified in sweet sorghum (bioenergy sorghum) and showed differential spatiotemporal patterns of expression while *SbGA20ox1* was predominantly related to increased stem biomass and assimilates partitioning (Wang et al., 2020). "Green revolution" gene *GA20-oxidase* is involved in the synthesis of principal biopolymer in the cell wall, i.e., cellulose in sorghum, whereas *dwarf1-1* cellulose deficient and male gametophyte-dysfunctional mutant showed ablation of GA and altered expression of three *CESA* genes generating cellulose deficient and dwarf phenotypes (Petti et al., 2015). Similarly, plants overexpressing *ZmGA20ox* displayed longer internodes and leaves, more tillers, and twofold increase in maize biomass (Voorend et al., 2016). Secondly, GA-INSENSITIVE DWARF1 (*GID1*) is the first receptor of bioactive GA in the signaling pathway, which shows the highest affinity for GA4 (bioactive form). Overexpressing the *GID1* gene shows a substantial increase in shoot elongation and growth in *Arabidopsis*, rice, and poplar (Sakamoto et al., 2004; Hirano et al., 2008; Mauriat and Moritz, 2009). The third main player in the gibberellin signaling pathway is the DELLA repressor gene which hinders the transcription of GA receptor, i.e., GID1. DELLA proteins act as a feedback regulatory control in the GA signaling pathway and are implicated in dwarf phenotype in maize with shifts in flowering time (Lawit et al., 2010). A recently conducted study on sugarcane affirmed the similar inhibitory roles of DELLA proteins. DELLA proteins interact with PIF4 and elements in the ethylene signaling pathway *ScEIN3/ScEIL1*, moreover, DELLA silenced lines showed changes in carbon allocation in storage and structural molecules and increased culm growth (Garcia Tavares et al., 2018).

Auxin

Auxin is a very important hormone involved in the growth process and cell wall architecture. Numerous mutants related to auxin synthesis, transport, and signaling showed overall dwarf phenotypes, defects in tropisms, and altered organ morphology (Vanneste and Friml, 2009). Auxin influx facilitator AUXIN1/LIKE-AUX1 (*AUX/LAX*) is involved in inflorescence development and root gravitropism. It is reported that mutations in homologs of AUX1 genes in maize (*ZmAUX1*) and *Setaria viridis* (*SvAUX1)* resulted in defective branch development of inflorescence, reduced plant height, increased panicle length, and sparse panicle phenotype (Huang et al., 2017). Aux/IAA homolog ROOTLESS WITH UNDETECTABLE MERISTEMS 1 (*RUM1*) in maize is involved in seminal and lateral roots formation. Transcriptome analysis of *rum1* showed down-regulated expression of like-auxin1 *(lax1),* the plethora genes plethora 1 *(plt1),* baby boom 1 *(bbm1)*, and heat shock complementing factor 1 *(hscf1*), and the auxin response factors *arf8 and arf37* (Zhang et al., 2015). In maize, *ARF5* (MONOPTEROS) is involved in vascular cells differentiation and *rum1* showed defective xylem organization and more lignin deposition in root cells (Zhang et al., 2014).

Brassinosteroid

Brassinosteroids (BR) comprise an important group of steroidal hormones originally isolated from the pollen of brassica plants (Rehman et al., 2022a). Brassinolide (BL) is the biologically active BR that is synthesized from compound campesterol with the aid of a cytochrome P450-mediated pathway (Bishop, 2007). This was a relatively novel and less studied hormone in the past, but recently it has gained attention as an active growth-promoting hormone owing to its involvement in many physiological functions (Fridman and Savaldi-Goldstein, 2013; Rehman et al., 2022b). Several genes are involved in the signaling pathways of brassinolide from BL perception

TABLE 1 | List of C_4 crops genes as candidates for exploiting biomass-related traits.

Crop species	Gene/Enzyme manipulated	Description	Comments	References
Cell cycle machinery				
Tobacco	*CycA3;3*	A-type cyclins	Antisense expression led the formation of defective embryo and impaired callus formation	Yu et al., 2003
Tobacco	*Arath-CYCD2 or Arath-CYCD3*	D-type cyclins (G_1-specific cyclins)	OE transgenics exhibited increased cell number but not cell size with higher leaf initiation rates	Boucheron et al., 2005
Maize	*Samba1* and *samba 3*	SAMBA	CRISPR/Cas9 mutant lines accelerated cell cycle, erect and shortened foliage upper top leaf length, ligule formation and internode elongation	Gong et al., 2021
Hormone related genes				
Arabidopsis	*GA20ox*	GA20-oxidase	OE produced 25% taller plants, accelerated shoot growth and early flowering	Coles et al., 1999
Switchgrass	*GA20ox*	GA20-oxidase	OE lines showed 2 folds biomass increase due to more tillers, leaf size and elevated bioactive GAs	Do et al., 2016
Sorghum	*GA20ox1*	GA20-oxidase	Higher expression in bioenergy sorghum culms. Moreover, sweet sorghum had higher GA levels and biomass.	Wang et al., 2020
Sugarcane	*GAI*	DELLA repressors	OE lines showing the stunted culm growth and development and modulation of shoot-to-root ratio in sugarcane	Garcia Tavares et al., 2018
Maize	*AUX1*	(AUX/LAX) Auxin influx facilitators	Mutant showed inflorescence development and root gravitropism	Huang et al., 2017
Green foxtail	*AUX1*	(AUX/LAX)	Mutants led to defective inflorescence, reduced plant height, increased panicle length and sparse panicle phenotype	Huang et al., 2017
Maize	*Aux/IAA*	RUM1 (ROOTLESS WITH UNDETECTABLE MERISTEMS 1)	Mutant *rum1* showed defective xylem organization and more lignin deposition in root cells	Zhang et al., 2015
Maize	*ARF5*	(MONOPTEROS)	Mutant showed root altered patterning of vascular cells differentiation, thick cell walls with higher lignin contents	Zhang et al., 2014
Maize	*D11*	Biosynthesis of BL	Higher expression in young ears and seeds, Improve seed quantity and quality.	Sun et al., 2021
Maize	*BRI1*	BRASSINOSTEROID INSENSITIVE 1	Mutant displayed overall dwarf stature, shortened internodes, folded dark green leaves, decreased auricle formation and feminization of female flowers	Kir et al., 2015
Cell wall biosynthesis related genes				
Arabidopsis	*xxt1 xxt2*	Xyloglucan transferase	Double mutant showed aberrant root hairs and modified mechanical properties	Cavalier et al., 2008
Arabidopsis	*TED6 and TED7*	Tracheary Element (TE) Differentiation-Related 6 and 7	RNAi showed delay in TE differentiation, abnormality in SCW and cellulose synthesis	Endo et al., 2009
Arabidopsis	*LAC4 & LAC17*	Laccases	T-DNA insertion showed LAC17 effected the deposition of G lignin units in the interfascicular fiber. *lac4-2 lac17* double mutant resulted in 40% reduced lignin.	Berthet et al., 2011
Plant species	**Gene/Enzyme manipulated**	**Description**	**Comments**	**References**
Sugarcane	*LAC*	Laccases	Complementation in *Arabidopsis* lac17 (–19% lignin) mutant restored lignin content	Cesarino et al., 2013
Sorghum	*(bmr) bmr2*, *bmr6*, and *bmr12*	Brown midrib mutant	2 years-based field study of EMS *bmr* mutants displayed decreased levels of lignin	Sattler et al., 2014
Maize	(bm3) *COMT*	Brown-midrib-3, lacking caffeic acid O-methyltransferase (*COMT*)	Antisense (AS225), and bm3 maize plants resulted in disturbed cell wall assembly.	Guillaumie et al., 2008
Sorghum	*CslF6*	Cellulose synthase-like F6 (CslF6) Glucan biosynthesis	Chimeric cDNA construct modifies the fine structure of (1,3;1,4)-β-glucan polysaccharide chain	Dimitroff et al., 2016

(Continued)

TABLE 1 | (Continued)

Crop species	Gene/Enzyme manipulated	Description	Comments	References
Transcription factors and MicroRNA				
Switch grass	*ERF001*	SHINE "SHINE/WAX INDUCER" (SHN/WIN) TF (AP2/ERF) superfamily	Increased biomass, and efficient saccharification process	Wuddineh et al., 2015
Sugarcane	*SHN1*	SHINE "SHINE/WAX INDUCER" (SHN/WIN) TF e factor (AP2/ERF) superfamily	OE results modified cell walls and increase in biomass by (91–340%),	Martins et al., 2018
Maize	*MYB46/83*	MYB (myeloblastosis)	Synthesis and thickening of cell wall	Zhong et al., 2011
Switchgrass	*R2R3-MYB*	MYB (myeloblastosis)	OE lines showed an increased biomass up to ∼ 63% and reduced lignin content around 50%	Shen et al., 2012
Maize	*MYB31* and *MYB42*	MYB (myeloblastosis)	Redirected phenylpropanoid and lignin biosynthesis in *Arabidopsis*, reduced S/G ratio (S, syringl units; G, guaiacyl units)	Sonbol et al., 2009; Vélez-Bermúdez et al., 2015
Sorghum	*Myb60*	Myeloblastosis (*MYB*)	Overexpressed lines displayed enhanced lignification in leaf midribs and increased phenolics	Scully et al., 2016
Finger millet	*bHLH57*	*(BASIC* HELIX-LOOP-HELIX)	Over-expressing depicted resistance to salinity stress with enhanced photosynthetic efficiency and increased biomass	Babitha et al., 2015
Maize	*Dof 1*	Zinc finger protein	Increased NUE in transgenic sorghum and wheat. Activation of carbon skeleton metabolism, i.e., PEPC activity	Peña et al., 2017
Maize	miR156, AtSPL9, *MIR172*	*SQUAMOSA Promoter-Binding Protein-Like* (*SPL*) miRNA	Delays reproductive phase leading the prolonged vegetative stage	Lauter et al., 2005; Chuck et al., 2007
Switchgrass	GAUT4-KD, miRNA156-OE, MYB4-OE, COMT-KD and FPGS-KD).	Myeloblastosis (*MYB*) miRNA	Increased contents of carbohydrates by 12% and ethanol yields by 21%	Dumitrache et al., 2017

to activation of responsive genes for example receptor-like kinase BRASSINOSTEROID-INSENSITIVE 1 (*BRI1*), BRI1-ASSOCIATED RECEPTOR KINASE 1, SOMATIC EMBROGENESIS RECEPTOR KINASE 1, and a repressor gene GSK3-like kinase BIN2 (BRASSINOSTEROID-INSENSITIVE 2) (Sánchez-Rodríguez et al., 2010). After the discovery of BL, mutant analysis in *Arabidopsis* revealed that plants deficient in the genes related to BL biosynthesis or signaling pathways showed dwarf phenotype, compromised male fertility, delay in flowering time, altered patterns of vascular development, and impaired photomorphogenesis (Feldmann et al., 1989). In a very recent study on maize, an endoplasmic reticulum localized gene, i.e., *ZmD11* related to the biosynthesis of BL rescued the panicle architecture and plant height in *cpb1* mutant in maize and rice. *ZmD11* increased seed length, seed weight, and both seed starch and protein contents in rice and maize crops (Sun et al., 2021). brassinosteroid-deficient dwarf1 *(brd1)* gene encoding brassinosteroid C-6 oxidase having a maize *lilliputian1* allele *(lil-1)* caused alteration in gravitropic response of root, leaf cell density, and more wax deposition conferring the adaptive mechanism to stress (Castorina et al., 2018). BR receptor, i.e., BRASSINOSTEROID INSENSITIVE1 (*BRI1*) RNA interference (RNAi) knock-out mutants in maize showed overall dwarf stature, shortened internodes, folded dark green leaves, decreased auricle formation, and feminization of male flowers (Kir et al., 2015). Similarly, in sorghum, the nuclear localization of BRASSINOSTEROID INSENSITIVE 2 (*BIN2*) was inhibited by DW1 indicating its role in BR signaling. Sorghum lines harboring mutated Dw1 (dw1) showed impaired skotomorphogenesis, lamina joint bending, and insensitive to BR gene regulation and feedback (Hirano et al., 2017) (**Figure 1A**).

This cluster of genes involved in biosynthesis and signaling of the important growth-promoting hormones highlights the connections with the cell wall, carbohydrates, and photosynthesis-related pathways. Further studies need to elucidate growth patterns of double or triple mutants from multiple hormone pathways at transcriptional and biochemical levels for efficient biomass response.

Cell Wall-Related Genes

The cell wall is a non-living protective cell layer that comprises 70% of the world's plant biomass (Poorter and Villar, 1997; Pauly and Keegstra, 2008). Second-generation cellulosic biofuel (bioethanol, biohydrogen, and biomethanol) produced from the plant biomass mostly comes from the cell wall. During plant growth and cell extensibility, several processes are involved among which cell wall loosening and rearrangement strongly contribute in plant biomass traits. In plant species, numerous gene families related to cell wall biogenesis, membrane trafficking, remodeling, secondary cell wall synthesis, and signaling comprise ∼10% of plant genomes (Lauter et al., 2005; Yong et al., 2005; McCann and Carpita, 2007).

Many studies involving plant biomass engineering techniques showed a strong effect on cell wall-related genes on growth and biomass accumulation processes in C_4 biofuel crops e.g., miscanthus (van der Weijde et al., 2013), sorghum (Scully et al.,

2016; Xia et al., 2018; Tetreault et al., 2020), and sugarcane (Jung et al., 2012; Bottcher et al., 2013). Genes responsible for cellulose synthesis mainly include members of cellulose synthase (*CesA*) and cellulose synthase-like (*Csl*) families. A recently published comparative study have identified 77, 35, and 109 CesA/Csl genes in *Miscanthus floridulus*, *S. bicolor*, and *S. spontaneum*, respectively. Among the 10 groups of CesA genes classified by phylogenetic approaches, a new group was identified in *Miscanthus floridulus*, i.e., *CesAX* which was not present in C_3 rice. Higher expression of CesA genes and their duplicates mainly followed by WGD (Whole Genome Duplication) showed the additive effects in gene expression levels resulting in more cellulose accumulation (Zhang et al., 2021). Silencing or mutations of *CesA* genes in *Arabidopsis* and certain other monocots have resulted in certain functional deformities but there is no authentic evidence that over-expression of *CesA* will certainly increase the cellulose content of the cell wall. In *Miscanthus* × *giganteus*, cloning of six *MgCesA* genes showed the involvement of MgCesA2, 3, 4, 7, and 8 in primary cell wall biosynthesis and rest in (MgCesA10, 11, 12) secondary cell wall synthesis and formation of cellulose synthase complex (Zeng et al., 2020).

Two cellulose interacting genes *TED6* and *TED7* enhanced the cellulose synthesis in xylem vessel elements, and lack of function mutants resulted in the failed secondary cell wall formation in *Arabidopsis* (Endo et al., 2009). Similarly, expression profiling of interspecific sugarcane hybrids showed upregulation of *CesA*, laccases, and callose synthase-related genes in high biomass extreme F_2 segregants (Wai et al., 2017). Lignin has been an undesirable component of the cell walls in terms of bioenergy generation, although it accounts for ∼30% terrestrial organic carbon fixation in the biosphere (Battle et al., 2000). In sorghum, brown midrib mutant (*bmr*) showed decreased levels of lignin and the formation of an altered subunit. The *Bmr* gene is a biosynthesis gene in monolignol units, which yield the hydroxycinnamic subunits of lignin. Another class of genes, laccases, are involved in the oxidation of monolignol units before the incorporation into cell wall polymers (Bonawitz and Chapple, 2010). *In vitro* oxidation of lignin precursors (Liang et al., 2006) and localization in lignin synthesizing tissues (Ranocha et al., 2002; Caparrós-Ruiz et al., 2006) have been experimentally proved by laccases in plants. In *Arabidopsis*, T-DNA insertion lines exhibited the reduced lignin content in single mutants of *AtLAC4* and *AtLAC17* whereas, double mutants displayed 40% reduced lignin but with irregular xylem tissues (Berthet et al., 2011). Sugarcane, a benchmark of biomass-derived biofuels showed the strong interactions of *SofLAC* genes with phenylpropanoid biosynthesis genes in a co-expression network. For the confirmation of monolignols oxidation, complementation of the *SofLAC* gene under the native promoter *AtLAC17* was performed in *Arabidopsis lac17* mutants having reduced lignin. SofLAC repaired the lignin content in *Arabidopsis* but lignin composition was altered in complemented *lac17* mutant lines (Cesarino et al., 2013). Xyloglucans (XyG) comprise a major class in hemicellulose proportion of cell wall and are extensively found in primary cell walls of eudicots and non-gramineous monocots. In *Arabidopsis*, double mutant *xxt1 xxt2* displayed a considerable reduction in detectable xyloglucan and altered mechanical properties (Cavalier et al., 2008). For an ideal bioenergy crop, higher lignin content is an undesirable trait due to its recalcitrance to degradation, whereas higher crystalline cellulose content is favored due to its digestibility. Conversely, hemicelluloses are crosslinking both lignin and cellulose causing a decrease in cellulose crystallinity, but a reduced level of hemicellulose branching ensures easy separation of cell wall components (Torres et al., 2015). However, molecular alteration of hemicellulose is handicapped due to its vague and complex biosynthesis and subsequent pathways. Research advances to this field are nevertheless confined at molecular levels in model species whereas, application of this knowledge in bioenergy-related species is the main goal (**Figure 1B**).

Transcriptional Regulation and miRNA Role

Transcription factors are the genes encoding proteins (besides RNA polymerase) that are essentially required for transcription. Owing to the regulatory role in transcription activity, they are capable of controlling the expression of many downstream key genes related to growth and development. In plants, DNA transcription involves more than 1,500 TFs to regulate target genes by binding with cis-regulatory elements in the promoter region (Singh et al., 2002). Secondary cell wall formed between the primary cell wall and cell membrane strongly contributes to the development and is an important attribute for the biofuel industry. Biosynthesis and remodeling of cell wall components are achieved through an orchestrated action of TFs and downstream genes. Therefore, detailed knowledge of transcription factors controlling secondary cell wall initiation genes, polysaccharides synthesis, lignification process, and a parallel process of programmed cell death (PCD) of xylem cells (Ohashi-Ito et al., 2010) is important to dissect for biomass regulation (**Table 1**).

Transcription factors of NAC (NAM—No Apical Meristem, ATAF, CUC—CUP/SHAPED Cotyledon) family activates a nexus of downstream transcription factors for example *MYBR2B3,* and act as master switches by binding with cell wall biosynthetic genes (Martins et al., 2018). SHINE "SHINE/WAX INDUCER" (*SHN/WIN*) TF is a member of ETHYLENE RESPONSIVE FACTOR (*ERF*) that functions as a regulator of cell wall biosynthesis genes, resulting in increased cellulose and decreased lignin contents (Ambavaram et al., 2011). In switchgrass (*Panicum virgatum*) *PvERF001* gene which is the homolog of *AtSHN2* conferred activation of cell wall synthesis and accumulation of biomass (Wuddineh et al., 2015). Likewise, sugarcane transcription factor *ShSHN1* overexpressed in rice induced changes in cell wall composition and increase in biomass by (91–340%), pectin (26–209%), cellulose content (10–22%), and saccharification efficiency (5–53%) in rice transgenic plants (Martins et al., 2018). McCarthy reported that *AtMYB46* and its paralog *AtMYB83* are found to function as activators of the secondary cell walls and are expressed in xylem fibers and vessels during secondary cell wall development (McCarthy et al., 2009). In *Arabidopsis myb46/83* double mutant, maize orthologs of *AtMYB46/83* successfully complemented the secondary cell wall synthesis and thickening after rescuing the defected walls

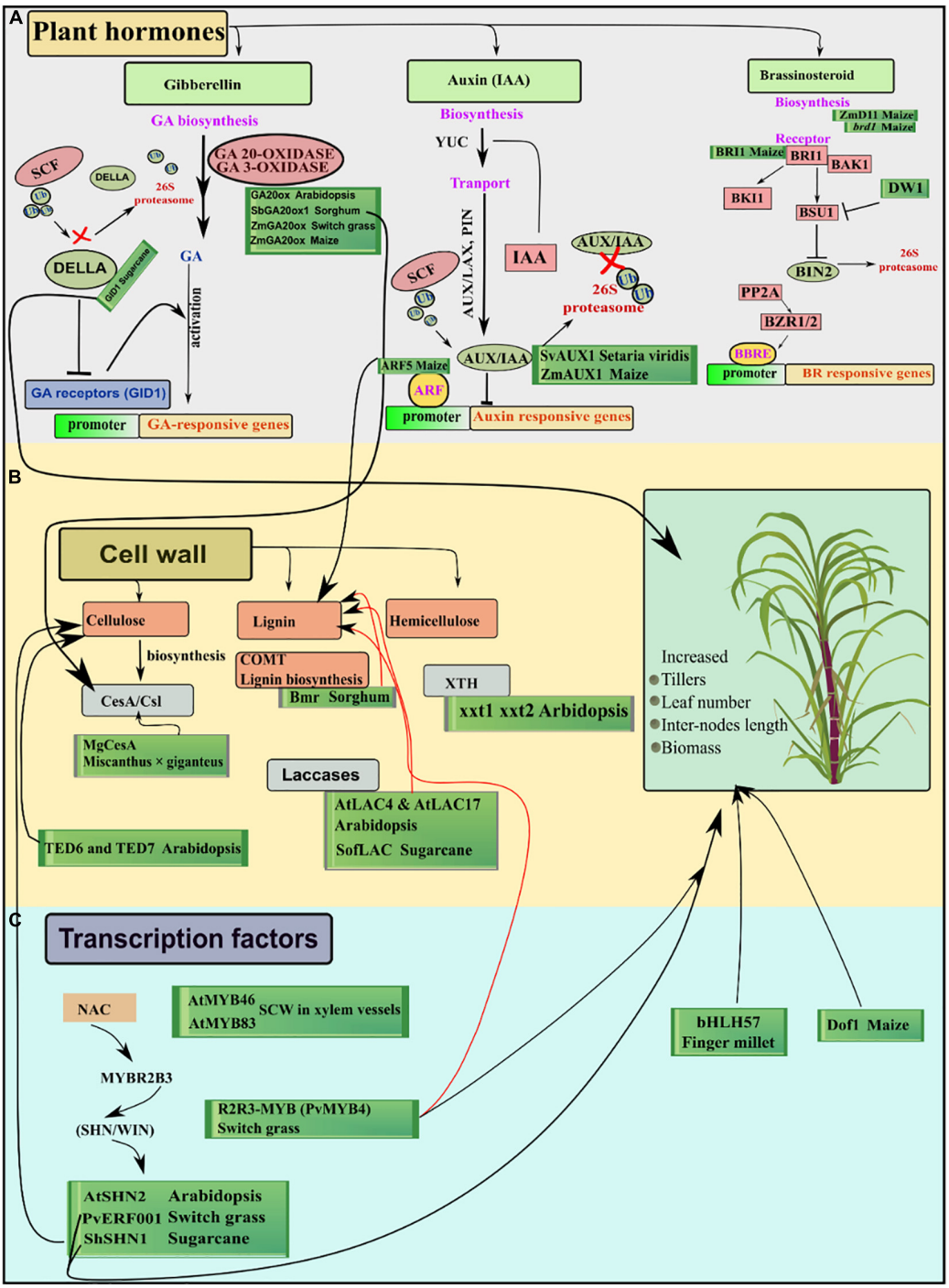


FIGURE 1 | Overview of the genes involved in different pathways in important C_4 biomass plants. **(A)** Highlights the hormone genes directly or indirectly related to growth and cell wall functioning. **(B)** Shows multiple genes involved in biosynthesis and remodeling of different cell wall-related components. **(C)** Transcription factors regulate many genes involved in different pathways of secondary cell wall synthesis leading to modified cell wall components improving saccharification efficiency. Black arrows show positive, red arrows show negative growth impacts and T lines show inhibitory influence on other genes. Whereas, green boxes enclose the functionally characterized genes in their respective pathways.

(Zhong et al., 2011). Similarly, in switchgrass, overexpression transgenic lines of R2R3-MYB (*PvMYB4*) showed increased biomass up to ∼63% and reduced lignin content around 50% (Shen et al., 2012). Some MYB genes in grasses, e.g., maize and switchgrass function in lineage-specific fashion regarding lignin biosynthesis regulation (Agarwal et al., 2016). For example,

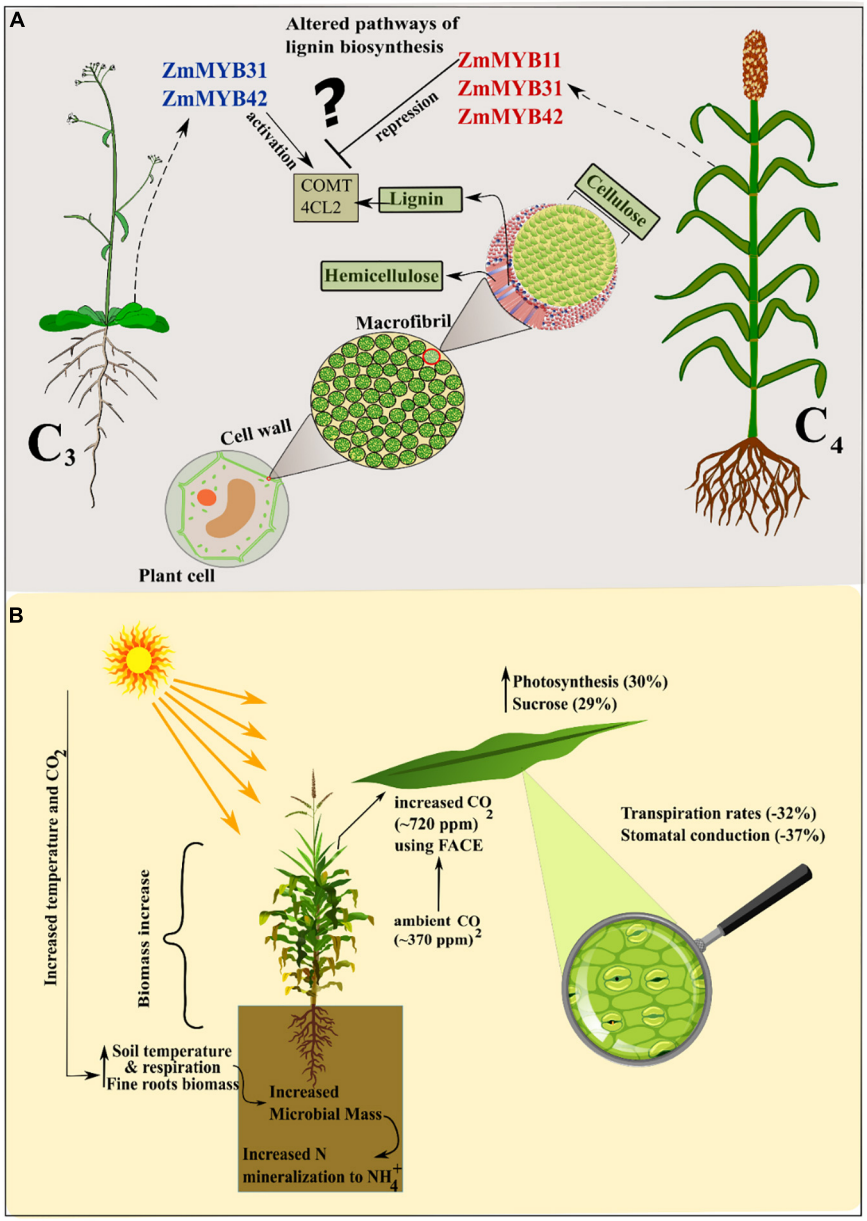


FIGURE 2 | (A) Contrasting transcriptional regulation patterns of *MYB* TFs in C_3 and C_4 plants. In C_3 (*Arabidopsis*) ectopic expression of maize TFs enhance lignin biosynthesis whereas in C_4 (Maize) lignin content is reduced. This gives rise to lineage-specific transcription in C_4 plants. **(B)** Increase in temperature and CO_2 enhances photosynthesis and altered control of stomatal aperture enhancing WUE. Soil microbiota mass in the rhizosphere is also increasingly responsible for modifications in the nutrient pool.

co-immunoprecipitation and ChIP-seq assays (co-IP) assays showed that *ZmMYB11*, *ZmMYB31*, and *ZmMYB42* induced reduction in expression of lignin biosynthesis-related genes *COMT* (caffeic acid-O-methyltransferase) and *4CL2* in maize (Vélez-Bermúdez et al., 2015). ZmMYB31 and ZmMYB42 in exogenous expression in *Arabidopsis* redirected phenylpropanoid and lignin-related genes in *Arabidopsis* contrary to maize. Moreover, *ZmMYB31* and *ZmMYB42* didn't down regulate the *ZmF5H* (ferulate-5-hydroxylase) gene in maize as compared to *Arabidopsis*, leading to a decrease in S/G ratio (S, syringl units; G, guaiacyl units) (Sonbol et al., 2009; Fornalé et al., 2010; Vélez-Bermúdez et al., 2015). Ectopic expression of transcription factor *SbMyb60* in sorghum showed involvement in monolignol biosynthesis pathways and increased lignin concentration and plant biomass. Constitutively overexpressing *SbMyb60* displayed enhanced lignification in leaf midribs and soluble phenolic compounds in plant biomass (Scully et al., 2016). This suggests that in monocot grasses the route of MYB TFs

and their regulatory pathways are more diverse and need to be investigated considering the models from grasses. Furthermore, the differences in C_3 and C_4 photosynthetic regulatory pathways should be studied in detail to increase the cellulose and hemicellulose contents and decreased contents of recalcitrant, i.e., lignin. (**Figure 2B**). Finger millet transgenic plants over-expressing *bHLH57 (BASIC* HELIX-LOOP-HELIX) depicted resistance to salinity stress with enhanced photosynthetic efficiency and increased biomass (Babitha et al., 2015). *AHL* (AT-HOOK MOTIF NUCLEAR LOCALIZED) family of transcription factors in *Arabidopsis* controls the petiole and rosette growth and architecture by antagonizing the role of growth-promoting PHYTOCHROMEINTERACTING FACTORS (*PIFs*) (Favero et al., 2020). Maize zinc finger protein *Dof1* transcription factor increased the nitrogen use efficiency in transgenic sorghum and wheat. Tissue-specific expression of *ZmDof 1* under rbcS1(maize) promoter increased growth by activation of carbon skeleton metabolism, i.e., *PEPC* activity (Peña et al., 2017) (**Figure 1C**).

SQUAMOSA Promoter-Binding Protein-Like (*SPL*) transcription factors are regulated by microRNAs (miRNA), i.e., miR156 and AtSPL9 which positively regulate another miRNA miR172 (Chen et al., 2010). *AtSPL9* TF binds to the promoter region of *the MIR172* gene and induces the transcription activity of downstream genes and repress adult-related characteristics in *Arabidopsis*. At the later plant stage, *the miR156-AtSPL9-miR172* regulatory pathway progresses with the decrease in *miR156* levels and increase in *miR172* leading to repression of FLOWERING LOCUS T (FT) gene. This event allows entering the plant to the reproductive phase and the same regulatory pathway is conserved in maize where *miR172* represses *Glossy15*, an *AP2-like* TF (Lauter et al., 2005; Chuck et al., 2007). This interactive regulatory role of transcription factor and microRNA is an effective molecular tool to extend the vegetative phase for enhanced sink capacity and biomass accumulation. Dumitrache et al. (2017) developed independently overexpressing (OE) and silenced (KD) transgenic switchgrass lines of specific genes and *miRNA GAUT4-KD*, *miRNA156-OE*, *MYB4-OE*, *COMT-KD*, and FPGS-KD (Dumitrache et al., 2017). Continuous monitoring of 2-year ratoon transgenes showed increased contents of carbohydrates by 12% and ethanol yields by 21% as compared to controlled conditions. In *Arabidopsis*, the use of zinc finger artificial transcription factor (*ZF-ATF*)-mediated interrogation lines helped in understanding growth and biomass-related characteristics. Introgression lines of two *Arabidopsis* genomes harboring 3F-EAR encoding T-DNA and 3F-VP16 encoding T-DNA constructs showed substantially large phenotypes. Whereas, 3F-EAR is *Arabidopsis* based ERF-associated Amphiphilic Repression (*EAR*) motif and acts as a dominant repressor evident from the previous studies and VP16 protein originated from the herpes simplex virus as a transcriptional activator (Ohta et al., 2001; Hiratsu et al., 2003; McCarthy et al., 2009). Further research is needed to elucidate the differences in transcriptional regulation of *Arabidopsis* and C_4 grasses, involving the promoter analysis to identify important cis-elements and the spatio-temporal regulation of TFs, affecting development-related genes.

Hybridity and Polyploidy

Hybrids and polyploids are common in plants. Hybridization between and within species is a natural process and is estimated to occur in ∼25% of plant species (Mallet, 2005). Hybrid vigor is a common consequence of hybridization and refers to superior hybrid performance in yield, biomass, or other agronomic parameters. Polyploidy refers to a cell or organism having two or more sets of basic chromosomes. An autopolyploid is derived from genome duplication within the same species, such as alfalfa (*Medicago sativa*), sugarcane (*Saccharum*), and potato (*Solanum tuberosum*), while allopolyploids are formed by chromosome doubling following hybridization between species. Allopolyploid is a "doubled interspecific hybrid," which leads to heterozygosity and hybrid vigor fixation. Many crops like cotton (*Gossypium hirsutum*), bread wheat (*Triticum aestivum*), and oilseed rape (*Brassica napus*) are cultivated as allopolyploids while rice (*Oryza sativa*), and maize (*Zea mays*) are mainly grown as hybrids. Both ploidy and hybridity affect growth vigor and cell size which are directly associated with plant biomass production (Chen, 2010).

In maize, increased ploidy had a detrimental effect on plant size which increases from haploid to triploid, but reduces in tetraploid (Riddle et al., 2006), and is consistent with smaller haploid *Arabidopsis* plants than diploids (Ravi and Chan, 2010). Induced polyploidy in hybrids may facilitate improving yield components, diminish hybrid vigor breaks in subsequent generations and restore inter-subspecific hybrids fertility (Miller et al., 2012). In *sorghum,* the colchicine-treated polyploidy induced plants showed high biomass with longer leaf length and stronger root system (Ardabili et al., 2015). A triploid *Miscanthus* × *giganteus*, C_4 grass is considered an excellent bioenergy crop due to its capacity to capture greenhouse gases by sequestering carbon in underground rhizomes and high biomass production when compared to diploid Miscanthus species (Chae et al., 2013). The modern sugarcanes are polyploid interspecific hybrids combining disease resistance, hardiness, and ratooning of *Saccharum spontaneum* and high sugar content from *Saccharum officinarum*. Genome restructuring and gene expression modifications in these cultivars due to polyploidy provide a selective advantage for a wider geographical adaptation, increased vigor, sucrose, and fiber content (Ming et al., 2001; Hoang et al., 2015). There is increased global demand for alternative fuel sources, and sugarcane is gaining importance as a biofuel crop with its high biomass production potential, besides being a major sugar crop.

ENVIRONMENTAL CUES INFLUENCING BIOMASS ACCUMULATION

Plants being sessile in nature are exposed to constantly changing environment enveloping around them. In the presence of judicious input resources and ideal genotypes, still plants are exposed to fluctuating environment in terms of CO_2 concentration, irradiance, and temperature which are function of plant growth. Here we reviewed that how fluctuations in external environment alter gene expression of important pathways and eventually biomass accumulation patterns.

Ambient CO_2 Fluctuations

With the gradual increase of GHGs in the atmosphere, the extent of CO_2 is also increasing in the air as CO_2 also comes under the category of greenhouse gases. This increase in GHGs tends to rise the global temperature which seriously changes the genetic and physiological attributes affecting the growing patterns of plants. This global warming accompanied by climate change has increased the variability of precipitation and a continuous increase in ambient CO_2 concentration up to 490–1,260 ppm by the end of the twenty-first century (IPCC, 2007). Considering the global changes, fluctuations in CO_2, light, and temperature are having both direct and indirect effects on the growth and biomass production of C_3, C_4, and CAM photosynthesis plants (Ainsworth and Long, 2005). Owing to anatomical and functional modifications in C_4 species, it was assumed that C_4 plants will be less affected by CO_2 rise as ambient CO_2 already meets the maximum saturation due to bundle sheath cells, as for C_3 plants CO_2 is a limiting factor for maximum photosynthesis. However, previous researches suggest different reasons for the increased response in terms of growth and productivity against elevated CO_2 in C_4 plants as (1) direct effect on Rubisco as CO_2 saturation point increases (Ziska and Bunce, 1997), (2) leakiness of bundle sheath cells (Watling et al., 2000), (3) young leaves are supposed to undergo preliminary C_3 photosynthesis system (Cousins and Bloom, 2003), (4) reduced stomatal aperture to enhance WUE or maintaining inner optimal temperature (Ghannoum et al., 2000).

Dieleman et al. (2012) published a meta-analysis of multifactorial experiments in which all treatments showed that increased CO_2 and warming increased plant biomass and soil respiration. An increase in only CO_2 treatment elicited more biomass of fine roots, soil respiration, and a decrease in foliar nitrogen (Dieleman et al., 2012). In another study De Souza et al. (2008) compared the effects of ambient (∼370 ppm) and increased (∼720 ppm) CO_2 concentration on sugarcane growth and biomass. Elevated CO_2 led to an increase of 17% in plant height, 30% in photosynthesis, 29% in sucrose contents, and 40% more accumulation of plant biomass (De Souza et al., 2008). This is thought to be achieved by physiological modifications regarding WUE as transpiration rates and stomatal conduction was reduced by –32 and –37%. An experiment involving free-air carbon enrichment (FACE) showed an elevated photosynthetic rate in young leaves and increased biomass and leaf number in sorghum and maize, respectively (Maroco et al., 1999; Cousins and Bloom, 2003). Recently, a study focused on the membrane properties and photosystem II activity of maize and pearl millet under elevated CO_2 and temperature. The results showed that maize outperformed in biomass accumulation in the presence of high CO_2 while pearl millet was more responsive in high temperature (Bordignon et al., 2019). Conversely, a comparative study conducted on two weedy species of C_3 (*Chenopodium album*) and C_4 (*Setaria viridis*) plants showed the decreased biomass at elevated temperature alone but a dramatic increase in biomass and seed yield by 33.9 and 114.4%, respectively, at increased temperature and CO_2 concentrations in *Chenopodium album*. On the other hand, *Setaria viridis* showed 1.6- and 1.3-fold more biomass in an increased temperature and CO_2 conditions as compared to control and only increased temperature (Lee, 2011). Experiments conducted on wheat and maize as the representatives of C_3 and C_4 plants showed that high CO_2 in C_3 (wheat) helps in ammonium NH_4^+ assimilation which was very less in ambient CO_2 and showed declined rate of photosynthesis. Overall results showed that cellular and chloroplast CO_2 enhanced electron flux in wheat as compared to maize. Several experiments showed that C_3 plants are more responsive toward elevated CO_2 as they reduce the extent of photorespiration and benefit more from enhanced CO_2. However, recently reported from a 20-year continuous investigation using FACE experiment on 88 grassland plots, that for 12 years C_3 plants showed increased biomass owing to higher eCO_2 (Equivalent CO_2) (Reich et al., 2018). Whereas, in subsequent 8 years a shift in this trend was seen from C_3 to C_4 plants, which resulted in enhanced biomass and soil nitrogen mineralization in C_4 plants. These results are challenging to current concepts regarding the C_3-C_4 eCO_2 paradigm (**Figure 2A**).

Circadian Rhythm Modulations by Environmental Factors

The instabilities of external cues of the environment, such as light, temperature, and nutrition, evoke a well-developed endogenous time-keeping mechanism in plants which is called the circadian cycle that allow in the modulation of energy and developmental metabolism. For example, fluctuations in the diurnal rhythm of light duration, temperature, and nutrition level modify intricate transcriptional and post-transcriptional loops in plants similar to animals (Harmer et al., 2001; Inoue et al., 2017). In *Arabidopsis*, the circadian clock starts by the mutualistic interaction of two transcription factors circadian clock associated 1 (*CCA1*) and late elongated hypocotyl (*LHY*) (Mizoguchi et al., 2002) during the morning hours. These two MYB TFs (*CCA1, LHY*) repress the transcription of TIMING OF CAB EXPRESSION 1 (*TOC1*) also known as PSEUDORESPONSE REGULATOR 1 (*PRR1*), which acts as a regulator of many downstream genes. *CCA1* Hiking Expedition (*CHE*) (Pruneda-Paz et al., 2009), and GIGANTEA (*GI*) (Strayer et al., 2000) also acts as a positive regulator in circadian loops, whereas, *CCA1* and *LHY* mRNA decrease during the mid-day by *TOC1* homologs (*PRR9* and *PRR7*) in feedback mechanism (Farré et al., 2005). During the evening, Lux Arrythmo (LUX) mediated transcriptional repression of PRR9 takes place by early flowering 4 (*ELF4*) and (*ELF3*) (McClung, 2014).

In grasses, growth activity is mediated by a rib zone called shoot apical meristem and is strongly influenced by light and shade conditions. In densely grown sorghum populations, leaves experience more shade conditions and inhibit shoot branching, more stem elongation, and early flowering. A combination of these responses is called Shade Avoidance Syndrome (SAS) (Casal, 2013), which is a survival strategy in plants for the quest for more sunlight and resources. Plant photoreceptors act to monitor the light environment and perception, and they work in synchrony with the circadian clock to regulate growth and

development in plants (Devlin and Kay, 2001). Fluctuations in the light intensity change the expression of 24 circadian genes in bioenergy sorghum, among which CCA1 and LHY showed 12.7-fold lower and 5.9-fold higher expression in internodes of shade-treated plants in comparison to control. CCA1 and LHY regulate the expression of downstream genes for example in sorghum, the homolog of Arabidopsis Granule Bound Starch Synthase1 (*GBSS1*) gene which functions in starch biosynthesis was also down regulated by 6.6-fold. Other clock-related genes for example evening core clock genes were upregulated in shade-exposed plants. For example, the expression of *BT2*, and THIAMIN C SYNTHASE (*THIC*) function in important pathways related to hormones, sugars, and thianmin synthesis are regulated by circadian cycle genes (Mandadi et al., 2009). Similarly, 6 days of exposure to extended darkness resulted in malfunctioning in photosynthetic pigments, reduction in photoassimilates, and total soluble and insoluble carbohydrates in maize. However, CO_2 exchange was not disturbed but transient carbon pools were largely consumed by elevated levels of nocturnal respiration rather than transport toward the sink. Underlying processes may involve signals by trehalose-6-phosphate and circadian rhythms which are controlling stress response in multi-dynamic pathways (Graf et al., 2010; Wingler et al., 2012). Light quality also mediates growth responses by genetic and physiological modifications. Transcriptome and metabolome profile of blue light exposed maize plants showed genes and metabolites related to stomatal, carotenoid, photosynthesis, and circadian cycle-related genes. *CCA1* gene was upregulated in presence of blue light, this upregulation is consistent with the expression of many downstream genes related to photosynthesis, starch synthesis, and stomatal development processes (Liu and Zhang, 2021). In maize, the early stage of light exposure stimulates the circadian rhythm genes to increase light absorbance by regulation of photosynthetic genes (Khan et al., 2010). As C_4 plants possess well-developed Kranz anatomy for the specialized storage of CO_2 and water. In maize plant *ZmPIP4c*, a water transport aquaporin gene showed high expression profiles in bundle sheath cells during diurnal change and is potentially responsible for the transport of water in mesophyll cells. The synchronization in the expression of NAD-malic enzyme gene (*NAD-ME*), phosphoenolpyruvate carboxylase gene (*PEPC*), and carbonic anhydrase gene (*CA*) from base to tip is consistent with the circadian rhythm regulating cycle for efficient WUE in both light and dark conditions (Xiang et al., 2020). Moreover, as already discussed hybridization leads to heterosis, which is the outcome of enhanced photosynthesis and metabolism possibly influenced by CIRCADIAN CLOCK ASSOCIATED1 (*CCA1*). Two homologs of maize, *ZmCCA1a,* and *ZmCCA1b* are diurnally upregulated in *Arabidopsis*, *cca1 Arabidopsis* mutant was complemented by *ZmCCA1*. Whereas, *ZmCCA1b* showed disruption in circadian rhythms leading to reduced heterosis and plant height in the greenhouse and slightly compensated in field conditions upon light exposure. In hybrids, the temporal shift of *ZmCCA1*-binding targets suggest the activated photosynthesis and growth vigor genes in the morning phase relative to the inbred lines (Ko et al., 2016). The behavior of circadian genes is a good indicator to enhance our understanding of environment-influenced genetic modulations.

TOOLS AND STRATEGIES TO ENHANCE BIOMASS

Hybridization and Molecular Techniques

Until now, some plant species have been given attention for the improvement of biomass which includes miscanthus, switchgrass, willow, poplar, and eucalyptus, with their improvement history dating back to the second half of the twentieth century (Allwright and Taylor, 2016; Clifton-Brown et al., 2019). Their improvement relied on hybridization or breeding methods (crossing of different strains, species, or lines) leading to heterosis or hybrid vigor of the F_1 heterozygotes with higher fitness in the population. Heterotic fitness refers to superior growth, stature, fertility, and biomass in offsprings. Several factors, for example, transcriptional regulation and epigenetic changes drive improved characters in hybrids. We have already reviewed that breeding techniques are employed mostly in C_4 grasses for achieving enhanced biomass. However, the approval and release of commercial cultivars take a long period of time, which factually delays the cultivation in agriculture systems and slows down the progress of conventional breeding (Clifton-Brown et al., 2019). Conventional breeding techniques advance our understanding toward marker-assisted selection for biomass-related traits, stress tolerance, and scarification in biofuel grasses and woody plants. In switchgrass, marker-assisted breeding enabled the understanding of substitution of cell wall hemicellulose polymers backbone and remodeling (De Souza et al., 2015) whereas identification of specific loci was identified from potential markers for high ethanol generation from switchgrass populations (Chen et al., 2016). Prairie cordgrass is a potential C_4 bioenergy crop, and two clones of prairie cordgrass were crossed and developed SSR markers for marker-assisted selection of biomass traits (Gedye et al., 2012). Another study comprising 28 sugarcane genotypes identified simple sequence repeat (SSR) markers associated with stalk number and stalk volume (Bilal et al., 2015). In another experiment, 40 putative quantitative trait alleles (QTAs) were identified from a self-crossed (295) population of "R570," with each allele contributing to phenotypic variation by 3–7% in sugarcane (Hoarau et al., 2002). Likewise, in sorghum, four QTLs were identified that control tiller number and formation (Hart et al., 2001). Recently research on the genotypes of *M. sinensis* indicated the genetic diversity of cell wall constituents and concluded that a higher ratio of para-coumaric acid to lignin contents and trans-ferulic acid (TFA) increased the saccharification efficacy (van der Weijde et al., 2017a). In sugarcane, stalk number is influenced by genes and their alleles with additive and non-additive effects or their interactions (Hongkai et al., 2009; Carvalho-Netto et al., 2014). Considering the use of genome editing techniques for bioenergy crops, the use of Transcription activator-like effector nucleases (TALENs) have been employed for targeted modification in the genome. High lignin content is an undesirable character for

TABLE 2 | Few examples of genome editing techniques, engineering the C_4 plants for biofuels.

Crops	Targeted genes	Technique	Improved traits	Associated pathway	References
Sugarcane	*COMT*	Transcription activator-like effector nucleases (TALENs)	11–32% reduced lignin Increased hemicellulose contents	Methyltransferase cell wall	Jung and Altpeter, 2016
Sugarcane	*COMT*	RNAi	12% reduction in lignin and improved scarification by 32%	Cell wall	Jung et al., 2013
Switchgrass	*Pv4CL1*	CRISPR/Cas9	8–30% reduced lignin 7–11% and 23–32% increase in glucose and xylose release	Lignin synthesis pathways	Park et al., 2017
Sorghum	BIOMASS YIELD 1 BY1	CRISPR/Cas9	Displayed reduced plant height, narrow stems, erect and narrow leaves, and abnormal floral organs.	Shikimate pathway	Chen et al., 2020

biofuel crops as in sugarcane, TALEN induced mutation in caffeic acid O-methyltransferase (*COMT*) sequences modified cell wall compositions. Pyrosequencing showed mutation frequencies up to 99% and revealed 29–32% reduced lignin and elevated hemicellulose contents (Jung and Altpeter, 2016). Similarly, RNAi-derived *COMT* silenced sugarcane callus-derived plants showed a 12% reduction in lignin and 32% improved scarification but compromised agronomic performance (Jung et al., 2013). CRISPR/Cas9 is a recent genome-editing technique expanding its applications to construct desirable genetic circuits (Khakhar et al., 2018). CRISPR/Cas9 system was employed in switchgrass for the reduction of lignin contents (Park et al., 2017). Knock-out mutant of the *Pv4CL1* gene encoding 4-coumarate: coenzyme A ligase (4CL) displayed increased scarification as compared to wild. Tiller formation is one of the indices of biomass, two genes grassy tiller1 (*gt1*) and teosinte branched1 (*tb1*) control tiller formation in maize (Whipple et al., 2011). BRANCHED1 (BRC1) gene in sorghum is the homolog of *tb1* gene, ectopic expression of *tb1* gene in *Arabidopsis* promoted axillary buds formation *Arabidopsis* (Kebrom et al., 2006). Genome editing-based mutagenesis using CRISPR/Cas9 showed proliferated tillers in switchgrass as compared to wild plants (Clifton-Brown et al., 2019). In sorghum, *by1* mutant obtained by knocking out BIOMASS YIELD 1 gene using CRISPR/Cas9 displayed reduced plant height narrow stem length, erect and narrow leaves, and abnormal floral organs. BIOMASS YIELD 1 gene translates into an enzymatic protein catalyzing the first step of the shikimate pathway (Chen et al., 2020). *BY1* gene showed its role in primary metabolism and secondary metabolites for example flavonoids. In *Arabidopsis* hormone activated Cas9-based repressor (*HACRs*) showed significant results that can be utilized to achieve high grass biomass and economic yield (Khakhar et al., 2018). Similarly, in sugarcane, using transgenic and molecular techniques, an ortholog of the *SLR1/D8/RHT1/GAI* gene showed substantial stem growth and structural modifications in storage organs by regulating source-sink allocation changes (Garcia Tavares et al., 2018). Although several transgenic lines with enhanced features as bioenergy crops have been developed, there is a need for a suitable selection process and quality evaluation. However, due to cross-pollination in many grass species, transgenic lines pose a threat of seed contamination. Introgression and hybridization require labor-intensive and time-consuming efforts with uncertain outcomes. Therefore, recent genome editing techniques for example synthetic genetic circuits (SGC) or CRISPR offer sophisticated and foreseeable mutation induction in first-generation mutant lines (Scheben and Edwards, 2018). Moreover, Near-infrared spectroscopy (NIRS) and thermal aerial imaging technologies aid the phenotyping of constituents and high-throughput options to screen abiotic stress tolerance, respectively (van der Weijde et al., 2017b). Few examples in **Table 2** highlight the use of integrated phenotyping and molecular technologies for biomass related research.

Nitrogen Management

Plant biomass of crops including C_4 plants is influenced by a variety of variables, i.e., plant genotype, photoperiod, solar radiation, soil temperature, soil humidity, and many more. Soil nutrient availability is one of the most significant variables determining the crop biomass in C_4 crops. So, by regulating the optimal amounts of nutrient availability in soil, growers may optimize the biomass output for biofuels and of course economic gain (grain production). Soil degradation and low soil fertility status significantly minimize the nutrient availability in the soil to plants (Chatzistathis and Therios, 2013; Tanveer et al., 2019). Optimum fertilization appears to be the most common method chosen by farmers in instances of restricted nutrient availability in soils to improve the biomass of cultivated C4 crops. Macronutrients, i.e., nitrogen, phosphorous, magnesium, potassium, sulfur, and calcium, and micronutrients, i.e., copper, zinc, manganese, iron, chlorine, and molybdenum are classified as important nutrients for improving the biomass of C_4 crops. Nutrient scarcity has a detrimental impact on biomass productivity (Anjum et al., 2019). Vegetation flushes of C_4 crops are severely hampered by nitrogen shortage as nitrogen is an integral component of chlorophyll, pyrimidines, purines, amino acids, proteins, and nucleic acids in C_4 crops. Borges et al. (2019) reported that the appropriate method of nitrogen application at the appropriate time (early season) significantly improved the biomass of sugarcane (*Saccharum officinarum*) by 30% as compared to traditional fertilizers practices. In addition, climate prediction models guided nitrogen management methods can be opted, as the climate is a significant determinant of crop growth, nitrogen demand, and nitrogen losses processes. Seasonal climatic projections might be used to establish nitrogen management plans for "dry" and "wet" years, directing application rate, timing, and frequency of nitrogen fertilizer application, as well as the advantages of employing different types of nitrogen fertilizer in C_4 crops

like sugarcane, maize (*Zea mays*), sorghum (*Sorghum bicolor*), pearl millet (*Pennisetum glaucum*), and Napier grass (*Pennisetum purpureum*) (Anjum et al., 2019). Moreover, Khan Khyber et al. (2010) reported that integrated application of 50% urea with 50% poultry manure significantly enhanced the grain yield of maize by 57.14%, respectively, as compared to plots having 0% nitrogen application. Although much effort has previously been done to maximize yields while maintaining high nutrient utilization efficiencies, more integrated approach results are still needed to minimize the nitrogen inputs while maintaining optimum usage efficiency in C_4 crops (Noor, 2017). To attain the maximum production of biomass in C_4 crops, maintaining the optimum levels of all the necessary soil nutrients should always be taken care of. However, determining the best nutrient prescription in terms of boosting productivity especially biomass production in C_4 crops while also guaranteeing food security and environmental friendliness is a difficult task, and still needed further studies and detailed analysis.

Silicon Foliar Application

Silicon is a chemical element having atomic number 14 and is represented with the symbol Si. In plants application of silicon significantly enhances the crop biomass; improves the tolerance to biotic and abiotic stresses, and aid plant stability and protection (Zargar et al., 2019). In connection to the enhancement of cell wall elasticity and stiffness, silicon that is firmly linked to the cell walls is naturally present as a structural material. When the quantity of monosilicic acid in the xylem sap is high, it becomes a significant osmolyte, increasing the plant's water and osmotic potential. Furthermore, in terms of structural material and osmolytes, Si consumes less energy than biomolecules like proline and lignin. As a result, for a cheap cost, silicon can enhance the homeostasis of C_4 plants' tolerance to a variety of biotic and abiotic stressors in terrestrial environments. C_4 plant biomass recovery mediated by silicon is thought to have a bell-shaped response curve to abiotic stressors and an S-shaped response curve to biotic stresses. Silicon treatment to abiotic and biotic stressed crops can boost averaged plant biomass carbon and crop productivity by 35 and 24%, respectively. The efficacy of silicon-mediated restoration, on the other hand, varies substantially depending on the plant species and cultivars, the severity of abiotic and biotic stressors, and the amount of bio-available silicon. Ashraf et al. reported that the application of silicon significantly improved the biomass production in sugarcane by 77% under salinity stress (Ashraf et al., 2009). Similarly, the application of calcium silicate improved the crop biomass of sugarcane and enhanced the resistance in sugarcane against stem borer (Keeping and Meyer, 2002; Meyer and Keeping, 2005). A study conducted on maize reported that the application of silicon under water stress conditions significantly enhanced the crop biomass and nutrient uptake (Kaya et al., 2007). It was reported that the application of silicon significantly improved the plant biomass under agricultural soil contaminated with heavy metals like cadmium (Liang et al., 2005; Lukačová et al., 2013), and in arsenic (Ullah et al., 2016). Application of silicon is a significant option for improving the crop biomass of C_4 crops, but still more research and detailed analysis should be done as most of the silicon application trials have so far been done in pots, field-scale to eco-system-scale investigation is needed. Moreover, several issues, such as the coupling relations between Si and plant essential elements, the efficiencies of Si-mediated plant biomass carbon restoration among plant species and stress intensities, and the relationship between the biogeochemical Si cycle and the resilience of terrestrial ecosystems, all require more research, particularly in fragmented landscapes.

Foliar Application of Plant Growth Regulators/Growth Hormones

Exogenous application of plant growth regulators/growth hormones (PGRs) has been found to improve plant stress tolerance and increase growth processes (Liu et al., 2019). Growth hormones are identified as playing a critical role in maintaining the plant morphology, flower blooming, development, photosynthetic activity, and stomatal closure in terms of physiological functions in C_4 plants (Sharma et al., 2020). Exogenous growth hormones were also used to control seed germination, root elongation, cell development, and tiller formation in C_4 plants cultivated under trace-metal contaminated soils (Maghsoudi et al., 2019). Similarly, in cereals like maize foliar application of plant growth hormones under abiotic stresses considerably improved the leaf area, plant growth, dry biomass, and stem diameter of C_4 plants (Tran and Popova, 2013; Qandeel et al., 2020). It was reported that exogenous application of various types of brassinosteroids, i.e., 28-homobrassinolide, and 24-epibrassinolide significantly enhanced the biomass and productivity of maize, sugarcane, and sorghum grown under abiotic stresses, i.e., drought, salinity, and trace-metal contaminated soils (Tanveer et al., 2018, 2019). In another study, it was reported that application of 1-amino-cyclopropane-1-carboxylic-acid (ACC-deaminase), humic acid, and oxalic acid considerably enhanced bacterial community development in the rhizosphere, facilitating the remediation of organic pollutants, and improved the plant biomass, which had previously been hindered by the presence of organic contaminants in soil (Ping et al., 2006; Wen-Jie et al., 2011). Similarly, plant growth regulators like melatonin and indole acetic acid are also reported to significantly improve the plant biomass under various abiotic stresses (Rostami et al., 2021). Likewise, the application of abscisic acid, salicylic acid, auxins, cytokinin, methyl jasmonate, and ethylene are also documented to significantly improve the plant biomass under abiotic stresses (Hasan et al., 2019). Yet, to explain precise processes related to the impacts of growth hormones on plant biomass, integrative studies combining conventional and sequencing techniques are required.

CONCLUDING REMARKS AND PROSPECTS

Biofuels being an alternative to fossil fuels are considered an integral part of sustainable energy generation systems. To develop the biofuel industry on a sustainable basis, increasing plant biomass is a prime goal as feedstock in the biofuels industry.

Biomass accumulation is a complex biological trait. However, advancements in genetics and biotechnology have deciphered that a plethora of genes are controlling growth and development starting from the cell cycle to the juvenile, vegetative, and reproductive maturity phase. Most of the pathways discussed highlight the important genes which have been exploited to tailor the bioenergy crops. Among them, genes involved in the cell cycle, cell wall, and hormone and related transcriptional factors considerably modify the carbohydrate allocation and improve photosynthetic efficiency. But the real challenge is the successful introduction of bioenergy specialized crops in fields on a sustainable basis. Moreover, we pointed out altered regulatory patterns of transcription activity of MYB TFs in C_3 and C_4 crops which indicate the lineage-specific carbohydrate storage biopolymer incorporation in both. Therefore, research focuses should be directed on C_4 crops considering only C_4 models in terms of biomass accumulation and later on energy generation.

Regarding environmental factors which are acting upon biomass accumulation, CO_2, light, and temperature are among unavoidable stresses to threaten the growth process. For this, architecture for the maximum light interception, nutrient absorption traits, and certain anatomical changes can be engineered in wild plants to enable the cultivation of bioenergy and orphan lignocellulose crops on marginal lands (resource-deprived). Furthermore, optimization of locality-based bioenergy crops and cultural practices to enhance biomass is critical but not yet elaborated. The above-mentioned tools and practices including breeding, molecular methods (DNA-free genome editing method CRISPR/Cas9), high throughput sequencing, and cultural practices can be opted for engineering and validation of multiple genes from different pathways to generate climate-smart energy crops. The afore-mentioned strategies will only be realistic if they are part of an integrated approach to agriculture that is developed collaboratively with agronomists, engineers, and farmers to contribute to a bio-based economy.

AUTHOR CONTRIBUTIONS

NA and RM conceptualized the review. NA wrote the review with the assistance of FH and MF. Habiba and NA designed the figures and tables. YZ improved and revised. All authors finally revised and approved the manuscript.

FUNDING

This work was supported by a startup fund from the Fujian Agriculture and Forestry University.

REFERENCES

Agarwal, T., Grotewold, E., Doseff, A. I., and Gray, J. (2016). MYB31/MYB42 syntelogs exhibit divergent regulation of phenylpropanoid genes in maize, sorghum and rice. *Sci. Rep.* 6:28502. doi: 10.1038/srep28502

Ainsworth, E. A., and Long, S. P. (2005). What have we learned from 15 years of free-air CO_2 enrichment (FACE)? A meta-analytic review of the responses of photosynthesis, canopy properties and plant production to rising CO_2. *New Phytol.* 165, 351–372. doi: 10.1111/J.1469-8137.2004.01224.X

Allwright, M. R., and Taylor, G. (2016). Molecular breeding for improved second generation bioenergy crops. *Trends Plant Sci.* 21, 43–54. doi: 10.1016/j.tplants.2015.10.002

Ambavaram, M. M. R., Krishnan, A., Trijatmiko, K. R., and Pereira, A. (2011). Coordinated activation of cellulose and repression of lignin biosynthesis pathways in rice. *Plant Physiol.* 155, 916–931. doi: 10.1104/pp.110.168641

Anderson, E., Arundale, R., Maughan, M., Oladeinde, A., Wycislo, A., and Voigt, T. (2011). Growth and agronomy of *Miscanthus* × *giganteus* for biomass production. *Biofuels* 2, 167–183. doi: 10.4155/bfs.10.80

Anjum, K., Cheema, A., Farooq, M., Haider, F. U., Cheema, S. A., and Ur Rehman, H. (2019). Article citation: exploring the potential of selenium (Se) and *Moringa* (*Moringa oleifera* L.) leaf extract on the production and performance of *Triticum aestivum* L. Introduction. *J. Res. Ecol.* 7, 2390–2402.

Ardabili, G. S., Zakaria, R. A., and Zare, N. (2015). In vitro induction of polyploidy in *Sorghum bicolor* L. *Cytologia* 80, 495–503. doi: 10.1508/CYTOLOGIA.80.495

Ashraf, M., Rahmatullah, Afzal, M., Ahmed, R., Mujeeb, F., Sarwar, A., et al. (2009). Alleviation of detrimental effects of NaCl by silicon nutrition in salt-sensitive and salt-tolerant genotypes of sugarcane (*Saccharum officinarum* L.). *Plant Soil* 326, 381–391. doi: 10.1007/S11104-009-0019-9

Babitha, K. C., Vemanna, R. S., Nataraja, K. N., and Udayakumar, M. (2015). Overexpression of EcbHLH57 transcription factor from *Eleusine coracana* L. in tobacco confers tolerance to salt, oxidative and drought stress. *PLoS One* 10:e0137098. doi: 10.1371/JOURNAL.PONE.0137098

Barnum, K. J., and O'Connell, M. J. (2014). Cell cycle regulation by checkpoints. *Methods Mol. Biol.* 1170, 29–40. doi: 10.1007/978-1-4939-0888-2_2

Battle, M., Bender, M. L., Tans, P. P., White, J. W. C., Ellis, J. T., Conway, T., et al. (2000). Global carbon sinks and their variability inferred from atmospheric O_2 and δ13C. *Science* 287, 2467–2470. doi: 10.1126/SCIENCE.287.5462.2467

Berthet, S., Demont-Caulet, N., Pollet, B., Bidzinski, P., Cézard, L., Le Bris, P., et al. (2011). Disruption of LACCASE4 and 17 results in tissue-specific alterations to lignification of *Arabidopsis thaliana* stems. *Plant Cell* 23, 1124–1137. doi: 10.1105/TPC.110.082792

Bilal, M., Saeed, M., Nasir, I. A., Tabassum, B., Zameer, M., Khan, A., et al. (2015). Association mapping of cane weight and tillers per plant in sugarcane. *Biotechnol. Biotechnol. Equip.* 29, 617–623. doi: 10.1080/13102818.2015.1008203

Bishop, G. J. (2007). Refining the plant steroid hormone biosynthesis pathway. *Trends Plant Sci.* 12, 377–380. doi: 10.1016/J.TPLANTS.2007.07.001

Bodrug, T., Welsh, K. A., Hinkle, M., Emanuele, M. J., and Brown, N. G. (2021). Intricate Regulatory mechanisms of the anaphase-promoting complex/cyclosome and its role in chromatin regulation. *Front. Cell Dev. Biol.* 9:687515. doi: 10.3389/FCELL.2021.687515

Bonawitz, N. D., and Chapple, C. (2010). The genetics of lignin biosynthesis: connecting genotype to phenotype. *Annu. Rev. Genet.* 44, 337–363. doi: 10.1146/ANNUREV-GENET-102209-163508

Bordignon, L., Faria, A. P., França, M. G. C., and Fernandes, G. W. (2019). Osmotic stress at membrane level and photosystem II activity in two C_4 plants after growth in elevated CO_2 and temperature. *Ann. Appl. Biol.* 174, 113–122. doi: 10.1111/aab.12483

Borges, C. D., Carvalho, J. L. N., Kölln, O. T., Sanches, G. M., Silva, M. J., Castro, S. G. Q., et al. (2019). Can alternative N-fertilization methods influence GHG emissions and biomass production in sugarcane fields? *Biomass Bioenergy* 120, 21–27. doi: 10.1016/J.BIOMBIOE.2018.10.017

Bottcher, A., Cesarino, I., Brombini dos Santos, A., Vicentini, R., Mayer, J. L. S., Vanholme, R., et al. (2013). Lignification in sugarcane: biochemical characterization, gene discovery, and expression analysis in two genotypes contrasting for lignin content. *Plant Physiol.* 163, 1539–1557. doi: 10.1104/PP.113.225250

Boucheron, E., Healy, J. H. S., Bajon, C., Sauvanet, A., Rembur, J., Noin, M., et al. (2005). Ectopic expression of *Arabidopsis* CYCD2 and CYCD3 in tobacco has

distinct effects on the structural organization of the shoot apical meristem. *J. Exp. Bot.* 56, 123–134. doi: 10.1093/JXB/ERI001

Byrt, C. S., Grof, C. P. L., and Furbank, R. T. (2011). C_4 plants as biofuel feedstocks: optimising biomass production and feedstock quality from a lignocellulosic perspective. *J. Integr. Plant Biol.* 53, 120–135. doi: 10.1111/j.1744-7909.2010.01023.x

Caparrós-Ruiz, D., Fornalé, S., Civardi, L., Puigdomènech, P., and Rigau, J. (2006). Isolation and characterisation of a family of Laccases in maize. *Plant Sci.* 171, 217–225. doi: 10.1016/J.PLANTSCI.2006.03.007

Caruso, G., Gomez, L. D., Ferriello, F., Andolfi, A., Borgonuovo, C., Evidente, A., et al. (2016). Exploring tomato *Solanum pennellii* introgression lines for residual biomass and enzymatic digestibility traits. *BMC Genet.* 17:56. doi: 10.1186/s12863-016-0362-9

Carvalho-Netto, O. V., Bressiani, J. A., Soriano, H. L., Fiori, C. S., Santos, J. M., Barbosa, G. V., et al. (2014). The potential of the energy cane as the main biomass crop for the cellulosic industry. *Chem. Biol. Technol. Agric.* 1:20. doi: 10.1186/S40538-014-0020-2

Casal, J. J. (2013). Photoreceptor signaling networks in plant responses to shade. *Annu. Rev. Plant Biol.* 64, 403–427. doi: 10.1146/ANNUREV-ARPLANT-050312-120221

Castorina, G., Persico, M., Zilio, M., Sangiorgio, S., Carabelli, L., and Consonni, G. (2018). The maize lilliputian1 (lil1) gene, encoding a brassinosteroid cytochrome P450 C-6 oxidase, is involved in plant growth and drought response. *Ann. Bot.* 122, 227–238. doi: 10.1093/AOB/MCY047

Cavalier, D. M., Lerouxel, O., Neumetzler, L., Yamauchi, K., Reinecke, A., Freshour, G., et al. (2008). Disrupting two *Arabidopsis thaliana* xylosyltransferase genes results in plants deficient in xyloglucan, a major primary cell wall component. *Plant Cell* 20, 1519–1537. doi: 10.1105/TPC.108.059873

Cesarino, I., Araújo, P., Sampaio Mayer, J. L., Vicentini, R., Berthet, S., Demedts, B., et al. (2013). Expression of SofLAC, a new laccase in sugarcane, restores lignin content but not S:G ratio of *Arabidopsis* lac17 mutant. *J. Exp. Bot.* 64, 1769–1781. doi: 10.1093/jxb/ert045

Chae, W. B., Hong, S. J., Gifford, J. M., Lane Rayburn, A., Widholm, J. M., and Juvik, J. A. (2013). Synthetic polyploid production of *Miscanthus sacchariflorus*, *Miscanthus sinensis*, and *Miscanthus x giganteus*. *GCB Bioenergy* 5, 338–350. doi: 10.1111/j.1757-1707.2012.01206.x

Chatzistathis, T., and Therios, I. (2013). "How soil nutrient availability influences plant biomass and how biomass stimulation alleviates heavy metal toxicity in soils: the cases of nutrient use efficient genotypes and phytoremediators, respectively," in *Biomass Now – Cultivation and Utilization*, ed. M. D. Matovic (London: IntechOpen), 428–448. doi: 10.5772/53594

Chen, J., Zhu, M., Liu, R., Zhang, M., Lv, Y., Liu, Y., et al. (2020). BIOMASS YIELD 1 regulates sorghum biomass and grain yield via the shikimate pathway. *J. Exp. Bot.* 71, 5506–5520. doi: 10.1093/JXB/ERAA275

Chen, S., Kaeppler, S. M., Vogel, K. P., and Casler, M. D. (2016). Selection signatures in four lignin genes from switchgrass populations divergently selected for in vitro dry matter digestibility. *PLoS One* 11:e0167005. doi: 10.1371/JOURNAL.PONE.0167005

Chen, X., Zhang, Z., Liu, D., Zhang, K., Li, A., and Mao, L. (2010). SQUAMOSA promoter-binding protein-like transcription factors: star players for plant growth and development. *J. Integr. Plant Biol.* 52, 946–951. doi: 10.1111/j.1744-7909.2010.00987.x

Chen, Z. J. (2010). Molecular mechanisms of polyploidy and hybrid vigor. *Trends Plant Sci.* 15, 57–71. doi: 10.1016/J.TPLANTS.2009.12.003

Cheng, Y., Cao, L., Wang, S., Li, Y., Shi, X., Liu, H., et al. (2013). Downregulation of multiple CDK inhibitor ICK/KRP genes upregulates the E2F pathway and increases cell proliferation, and organ and seed sizes in *Arabidopsis*. *Plant J.* 75, 642–655. doi: 10.1111/TPJ.12228

Chuck, G., Meeley, R., Irish, E., Sakai, H., and Hake, S. (2007). The maize tasselseed4 microRNA controls sex determination and meristem cell fate by targeting Tasselseed6/indeterminate spikelet1. *Nat. Genet.* 39, 1517–1521. doi: 10.1038/ng.2007.20

Clifton-Brown, J., Harfouche, A., Casler, M. D., Dylan Jones, H., Macalpine, W. J., Murphy-Bokern, D., et al. (2019). Breeding progress and preparedness for mass-scale deployment of perennial lignocellulosic biomass crops switchgrass, miscanthus, willow and poplar. *GCB Bioenergy* 11, 118–151. doi: 10.1111/gcbb.12566

Coles, J. P., Phillips, A. L., Croker, S. J., García-Lepe, R., Lewis, M. J., and Hedden, P. (1999). Modification of gibberellin production and plant development in *Arabidopsis* by sense and antisense expression of gibberellin 20-oxidase genes. *Plant J.* 17, 547–556. doi: 10.1046/J.1365-313X.1999.00410.X

Cousins, A. B., and Bloom, A. J. (2003). Influence of elevated CO_2 and nitrogen nutrition on photosynthesis and nitrate photo-assimilation in maize (*Zea mays* L.). *Plant Cell Environ.* 26, 1525–1530. doi: 10.1046/J.1365-3040.2003.01075.X

de Freitas Lima, M., Eloy, N. B., Bottino, M. C., Hemerly, A. S., and Ferreira, P. C. G. (2013). Overexpression of the anaphase-promoting complex (APC) genes in *Nicotiana tabacum* promotes increasing biomass accumulation. *Mol. Biol. Rep.* 40, 7093–7102. doi: 10.1007/s11033-013-2832-8

de Lima, M. F., Eloy, N. B., Pegoraro, C., Sagit, R., Rojas, C., Bretz, T., et al. (2010). Genomic evolution and complexity of the anaphase-promoting complex (APC) in land plants. *BMC Plant Biol.* 10:254. doi: 10.1186/1471-2229-10-254

De Souza, A. P., Gaspar, M., Da Silva, E. A., Ulian, E. C., Waclawovsky, A. J., Nishiyama, M. Y., et al. (2008). Elevated CO_2 increases photosynthesis, biomass and productivity, and modifies gene expression in sugarcane. *Plant Cell Environ.* 31, 1116–1127. doi: 10.1111/j.1365-3040.2008.01822.x

De Souza, A. P., Kamei, C. L. A., Torres, A. F., Pattathil, S., Hahn, M. G., Trindade, L. M., et al. (2015). How cell wall complexity influences saccharification efficiency in *Miscanthus sinensis*. *J. Exp. Bot.* 66, 4351–4365. doi: 10.1093/JXB/ERV183

Demura, T., and Ye, Z. H. (2010). Regulation of plant biomass production. *Curr. Opin. Plant Biol.* 13, 298–303. doi: 10.1016/j.pbi.2010.03.002

Devlin, P. F., and Kay, S. A. (2001). Circadian photoperception. *Annu. Rev. Physiol.* 63, 677–694. doi: 10.1146/ANNUREV.PHYSIOL.63.1.677

Dieleman, W. I. J., Vicca, S., Dijkstra, F. A., Hagedorn, F., Hovenden, M. J., Larsen, K. S., et al. (2012). Simple additive effects are rare: a quantitative review of plant biomass and soil process responses to combined manipulations of CO_2 and temperature. *Glob. Chang. Biol.* 18, 2681–2693. doi: 10.1111/j.1365-2486.2012.02745.x

Dimitroff, G., Little, A., Lahnstein, J., Schwerdt, J. G., Srivastava, V., Bulone, V., et al. (2016). (1,3;1,4)-β-Glucan biosynthesis by the CSLF6 enzyme: position and flexibility of catalytic residues influence product fine structure. *Biochemistry* 55, 2054–2061. doi: 10.1021/ACS.BIOCHEM.5B01384

Do, P. T., De Tar, J. R., Lee, H., Folta, M. K., and Zhang, Z. J. (2016). Expression of ZmGA20ox cDNA alters plant morphology and increases biomass production of switchgrass (*Panicum virgatum* L.). *Plant Biotechnol. J.* 14, 1532–1540. doi: 10.1111/pbi.12514

Dong, Z., Danilevskaya, O., Abadie, T., Messina, C., Coles, N., and Cooper, M. (2012). A gene regulatory network model for floral transition of the shoot apex in maize and its dynamic modeling. *PLoS One* 7:e43450. doi: 10.1371/JOURNAL.PONE.0043450

Dumitrache, A., Natzke, J., Rodriguez, M., Yee, K. L., Thompson, O. A., Poovaiah, C. R., et al. (2017). Transgenic switchgrass (*Panicum virgatum* L.) targeted for reduced recalcitrance to bioconversion: a 2-year comparative analysis of field-grown lines modified for target gene or genetic element expression. *Plant Biotechnol. J.* 15, 688–697. doi: 10.1111/PBI.12666

Eloy, N. B., Lima, M. D. F., Damme, V., Vanhaeren, H., and Gonzalez, N. (2011). The APC/C subunit 10 plays an essential role in cell proliferation during leaf development. *Plant J.* 68, 351–363. doi: 10.1111/j.1365-313X.2011.04691.x

Endo, S., Pesquet, E., Yamaguchi, M., Tashiro, G., Sato, M., Toyooka, K., et al. (2009). Identifying new components participating in the secondary cell wall formation of vessel elements in *Zinnia* and *Arabidopsis*. *Plant Cell* 21, 1155–1165. doi: 10.1105/TPC.108.059154

Farré, E. M., Harmer, S. L., Harmon, F. G., Yanovsky, M. J., and Kay, S. A. (2005). Overlapping and distinct roles of PRR7 and PRR9 in the *Arabidopsis* circadian clock. *Curr. Biol.* 15, 47–54. doi: 10.1016/J.CUB.2004.12.067

Favero, D. S., Kawamura, A., Shibata, M., Takebayashi, A., Jung, J. H., Suzuki, T., et al. (2020). AT-hook transcription factors restrict petiole growth by antagonizing PIFs. *Curr. Biol.* 30, 1454–1466.e6. doi: 10.1016/J.CUB.2020.02.017

Feldmann, K. A., Marks, M. D., Christianson, M. L., and Quatrano, R. S. (1989). A dwarf mutant of *Arabidopsis* generated by T-DNA insertion mutagenesis. *Science* 243, 1351–1354. doi: 10.1126/SCIENCE.243.4896.1351

Fornalé, S., Shi, X., Chai, C., Encina, A., Irar, S., Capellades, M., et al. (2010). ZmMYB31 directly represses maize lignin genes and redirects the

phenylpropanoid metabolic flux. *Plant J.* 64, 633–644. doi: 10.1111/J.1365-313X.2010.04363.X

Fridman, Y., and Savaldi-Goldstein, S. (2013). Brassinosteroids in growth control: how, when and where. *Plant Sci.* 209, 24–31. doi: 10.1016/j.plantsci.2013.04.002

Fujimoto, M., Arimura, S. I., Nakazono, M., and Tsutsumi, N. (2008). *Arabidopsis* dynamin-related protein DRP2B is co-localized with DRP1A on the leading edge of the forming cell plate. *Plant Cell Rep.* 27, 1581–1586. doi: 10.1007/S00299-008-0583-0/FIGURES/3

Garcia Tavares, R., Lakshmanan, P., Peiter, E., O'Connell, A., Caldana, C., Vicentini, R., et al. (2018). ScGAI is a key regulator of culm development in sugarcane. *J. Exp. Bot.* 69, 3823–3837. doi: 10.1093/JXB/ERY180

Gedye, K. R., Gonzalez-Hernandez, J. L., Owens, V., and Boe, A. (2012). Advances towards a marker-assisted selection breeding program in prairie cordgrass, a biomass crop. *Int. J. Plant Genomics* 2012:313545. doi: 10.1155/2012/313545

Ghannoum, O., von Caemmerer, S., Ziska, L. H., and Conroy, J. P. (2000). The growth response of C_4 plants to rising atmospheric CO_2 partial pressure: a reassessment. *Plant Cell Environ.* 23, 931–942. doi: 10.1046/J.1365-3040.2000.00609.X

Gong, P., Bontinck, M., Demuynck, K., de Block, J., Gevaert, K., Eeckhout, D., et al. (2021). SAMBA controls the rate of cell division in maize development through APC/C interaction. *bioRxiv* [Preprint]. doi: 10.1101/2021.04.22.440954

Graf, A., Schlereth, A., Stitt, M., and Smith, A. M. (2010). Circadian control of carbohydrate availability for growth in *Arabidopsis* plants at night. *Proc. Natl. Acad. Sci. U.S.A.* 107, 9458–9463. doi: 10.1073/PNAS.0914299107

Guillaumie, S., Goffner, D., Barbier, O., Martinant, J.-P., Pichon, M., and Barrière, Y. (2008). Expression of cell wall related genes in basal and ear internodes of silking brown-midrib-3, caffeic acid O-methyltransferase (COMT) down-regulated, and normal maize plants. *BMC Plant Biol.* 8:71. doi: 10.1186/1471-2229-8-71

Harmer, S. L., Panda, S., and Kay, S. A. (2001). Molecular bases of circadian rhythms. *Annu. Rev. Cell Dev. Biol.* 17, 215–253. doi: 10.1146/ANNUREV.CELLBIO.17.1.215

Hart, G. E., Schertz, K. F., Peng, Y., and Syed, N. H. (2001). Genetic mapping of *Sorghum bicolor* (L.) Moench QTLs that control variation in tillering and other morphological characters. *Theor. Appl. Genet.* 103, 1232–1242. doi: 10.1007/S001220100582

Hasan, M. N., Hasan, M. R., Foysal, S. H., Hoque, H., Khan, M. F., Bhuiyan, M. F. H., et al. (2019). In-vitro regeneration of *Citrus sinensis* (L.) Osbeck from mature seed derived embryogenic callus on different solid basal media. *Am. J. Plant Sci.* 10, 285–297. doi: 10.4236/AJPS.2019.102022

Heaton, E. A., Dohleman, F. G., and Long, S. P. (2008). Meeting US biofuel goals with less land: the potential of *Miscanthus*. *Glob. Chang. Biol.* 14, 2000–2014. doi: 10.1111/J.1365-2486.2008.01662.X

Hirano, K., Kawamura, M., Araki-Nakamura, S., Fujimoto, H., Ohmae-Shinohara, K., Yamaguchi, M., et al. (2017). Sorghum DW1 positively regulates brassinosteroid signaling by inhibiting the nuclear localization of BRASSINOSTEROID INSENSITIVE 2. *Sci. Rep.* 7:126. doi: 10.1038/s41598-017-00096-w

Hirano, K., Ueguchi-Tanaka, M., and Matsuoka, M. (2008). GID1-mediated gibberellin signaling in plants. *Trends Plant Sci.* 13, 192–199. doi: 10.1016/J.TPLANTS.2008.02.005

Hiratsu, K., Matsui, K., Koyama, T., and Ohme-Takagi, M. (2003). Dominant repression of target genes by chimeric repressors that include the EAR motif, a repression domain, in *Arabidopsis*. *Plant J.* 34, 733–739. doi: 10.1046/J.1365-313X.2003.01759.X

Hoang, N. V., Furtado, A., Botha, F. C., Simmons, B. A., and Henry, R. J. (2015). Potential for genetic improvement of sugarcane as a source of biomass for biofuels. *Front. Bioeng. Biotechnol.* 3:182. doi: 10.3389/fbioe.2015.00182

Hoarau, J.-Y., Grivet, L., Offmann, B., Raboin, L.-M., Diorflar, J.-P., Payet, J., et al. (2002). Genetic dissection of a modern sugarcane cultivar (*Saccharum* spp.).II. Detection of QTLs for yield components. *Theor. Appl. Genet.* 105, 1027–1037. doi: 10.1007/S00122-002-1047-5

Hong, Z., Geisler-Lee, C. J., Zhang, Z., and Verma, D. P. S. (2003). Phragmoplastin dynamics: multiple forms, microtubule association and their roles in cell plate formation in plants. *Plant Mol. Biol.* 53, 297–312. doi: 10.1023/B:PLAN.0000006936.50532.3A

Hongkai, Z., Guifu, L., Jiannong, L., and Juemin, H. (2009). Genetic analysis of sugarcane biomass yield and its component traits using ADAA model. *J. Trop. Agric.* 47, 70–73.

Hu, Y., Xie, Q., and Chua, N. H. (2003). The *Arabidopsis* auxin-inducible gene ARGOS controls lateral organ size. *Plant Cell* 15, 1951–1961. doi: 10.1105/tpc.013557

Huang, P., Jiang, H., Zhu, C., Barry, K., Jenkins, J., Sandor, L., et al. (2017). Sparse panicle1 is required for inflorescence development in *Setaria viridis* and maize. *Nat. Plants* 3:17054. doi: 10.1038/nplants.2017.54

IEA (2018). *CO2 Status Report 2017*. Paris: International Energy agency.

Inagaki, S., and Umeda, M. (2011). Cell-cycle control and plant development. *Int. Rev. Cell Mol. Biol.* 291, 227–261. doi: 10.1016/B978-0-12-386035-4.00007-0

Inoue, K., Araki, T., and Endo, M. (2017). Integration of input signals into the gene network in the plant circadian clock. *Plant Cell Physiol.* 58, 977–982. doi: 10.1093/PCP/PCX066

Inz, D. (2006). Cell cycle regulation in plant development. *Annu. Rev. Genet.* 40, 77–105. doi: 10.1146/annurev.genet.40.110405.090431

IPCC (2007). *Climate Change 2007: Synthesis Report. Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Core Writing Team, eds R. K. Pachauri, and A. Reisinger (Geneva: IPCC), 104.

Jung, J. H., and Altpeter, F. (2016). TALEN mediated targeted mutagenesis of the caffeic acid O-methyltransferase in highly polyploid sugarcane improves cell wall composition for production of bioethanol. *Plant Mol. Biol.* 92, 131–142. doi: 10.1007/S11103-016-0499-Y

Jung, J. H., Fouad, W. M., Vermerris, W., Gallo, M., and Altpeter, F. (2012). RNAi suppression of lignin biosynthesis in sugarcane reduces recalcitrance for biofuel production from lignocellulosic biomass. *Plant Biotechnol. J.* 10, 1067–1076. doi: 10.1111/J.1467-7652.2012.00734.X

Jung, J. H., Vermerris, W., Gallo, M., Fedenko, J. R., Erickson, J. E., and Altpeter, F. (2013). RNA interference suppression of lignin biosynthesis increases fermentable sugar yields for biofuel production from field-grown sugarcane. *Plant Biotechnol. J.* 11, 709–716. doi: 10.1111/PBI.12061

Kandel, R., Yang, X., Song, J., and Wang, J. (2018). Potentials, challenges, and genetic and genomic resources for sugarcane biomass improvement. *Front. Plant Sci.* 9, 1–14. doi: 10.3389/fpls.2018.00151

Kang, B.-H., Busse, J. S., and Bednarek, S. Y. (2003). Members of the *Arabidopsis* dynamin-like gene family, ADL1, are essential for plant cytokinesis and polarized cell growth. *Plant Cell* 15, 899–913. doi: 10.1105/TPC.009670

Kaya, C., Tuna, L., and Higgs, D. (2007). Effect of silicon on plant growth and mineral nutrition of maize grown under water-stress conditions. *J. Plant Nutr.* 29, 1469–1480. doi: 10.1080/01904160600837238

Kebrom, T. H., Burson, B. L., and Finlayson, S. A. (2006). Phytochrome B represses teosinte branched1 expression and induces sorghum axillary bud outgrowth in response to light signals. *Plant Physiol.* 140, 1109–1117. doi: 10.1104/PP.105.074856

Kebrom, T. H., McKinley, B., and Mullet, J. E. (2017). Dynamics of gene expression during development and expansion of vegetative stem internodes of bioenergy sorghum. *Biotechnol. Biofuels* 10:159. doi: 10.1186/S13068-017-0848-3

Keeping, M. G., and Meyer, J. H. (2002). Calcium silicate enhances resistance of sugarcane to the African stalk borer *Eldana saccharina* walker (Lepidoptera: Pyralidae). *Agric. For. Entomol.* 4, 265–274. doi: 10.1046/J.1461-9563.2002.00150.X

Khakhar, A., Leydon, A. R., Lemmex, A. C., Klavins, E., and Nemhauser, J. L. (2018). Synthetic hormone-responsive transcription factors can monitor and reprogram plant development. *Elife* 7:e34702. doi: 10.7554/ELIFE.34702

Khan, S., Rowe, S. C., and Harmon, F. G. (2010). Coordination of the maize transcriptome by a conserved circadian clock. *BMC Plant Biol.* 10:126. doi: 10.1186/1471-2229-10-126

Khan Khyber, H., Wahid, M. A., Rasul, F., and Mohkum Hammad, H. (2010). Nitrogen management strategies for sustainable maize production assessing the impact of climate change on wheat & cotton grown under different types of soils in various agro-environmental conditions of southern Punjab-Pakistan using crop simulation models. View project effect of nitrogen on seed production view project. *Crop Environ.* 1, 49–52.

Kir, G., Ye, H., Nelissen, H., Neelakandan, A. K., Kusnandar, A. S., Luo, A., et al. (2015). RNA interference knockdown of BRASSINOSTEROID INSENSITIVE1

in maize reveals novel functions for brassinosteroid signaling in controlling plant architecture. *Plant Physiol.* 169, 826–839. doi: 10.1104/PP.15.00367

Ko, D. K., Rohozinski, D., Song, Q., Taylor, S. H., Juenger, T. E., Harmon, F. G., et al. (2016). Temporal shift of circadian-mediated gene expression and carbon fixation contributes to biomass heterosis in maize hybrids. *PLoS Genet.* 12:e1006197. doi: 10.1371/journal.pgen.1006197

Koçar, G., and Civaş, N. (2013). An overview of biofuels from energy crops: current status and future prospects. *Renew. Sustain. Energy Rev.* 28, 900–916. doi: 10.1016/J.RSER.2013.08.022

Lauf, T., Memmler, M., and Schneider, S. (2021). *Emissions Balance of Renewable Energy Sources Determination of Avoided Emissions in 2020 (Emissionsbilanz Erneuerbarer Energieträger Bestimmung Der Vermiedenen Emissionen Im Jahr 2020)*. Available online at: http://inis.iaea.org/search/search.aspx?orig_q=RN:53029699

Lauter, N., Kampani, A., Carlson, S., Goebel, M., and Moose, S. P. (2005). microRNA172 down-regulates glossy15 to promote vegetative phase change in maize. *Proc. Natl. Acad. Sci. U.S.A.* 102, 9412–9417. doi: 10.1073/PNAS.0503927102

Lawit, S. J., Wych, H. M., Xu, D., Kundu, S., and Tomes, D. T. (2010). Maize DELLA proteins dwarf plant8 and dwarf plant9 as modulators of plant development. *Plant Cell Physiol.* 51, 1854–1868. doi: 10.1093/PCP/PCQ153

Lee, J. S. (2011). Combined effect of elevated CO_2 and temperature on the growth and phenology of two annual C_3 and C_4 weedy species. *Agric. Ecosyst. Environ.* 140, 484–491. doi: 10.1016/j.agee.2011.01.013

Liang, M., Davis, E., Gardner, D., Cai, X., and Wu, Y. (2006). Involvement of AtLAC15 in lignin synthesis in seeds and in root elongation of *Arabidopsis*. *Planta* 224, 1185–1196. doi: 10.1007/S00425-006-0300-6

Liang, Y., Wong, J. W. C., Wei, L., Liang, Y., Wong, J. W. C., and Wei, L. (2005). Silicon-mediated enhancement of cadmium tolerance in maize (*Zea mays* L.) grown in cadmium contaminated soil. *Chemosphere* 58, 475–483. doi: 10.1016/J.CHEMOSPHERE.2004.09.034

Lima, M. d. F., Eloy, N. B., Siqueira, J. A. B. d., Inzé, D., Hemerly, A. S., and Ferreira, P. C. G. (2017). Molecular mechanisms of biomass increase in plants. *Biotechnol. Res. Innov.* 1, 14–25. doi: 10.1016/j.biori.2017.08.001

Liu, T., and Zhang, X. (2021). Transcriptome and metabolomic analyses reveal regulatory networks controlling maize stomatal development in response to blue light. *Int. J. Mol. Sci.* 22:5393. doi: 10.3390/ijms22105393

Liu, Y., Chen, X., Wang, X., Fang, Y., Zhang, Y., Huang, M., et al. (2019). The influence of different plant hormones on biomass and starch accumulation of duckweed: a renewable feedstock for bioethanol production. *Renew. Energy* 138, 659–665. doi: 10.1016/J.RENENE.2019.01.128

Lukačová, Z., Švubová, R., Kohanová, J., and Lux, A. (2013). Silicon mitigates the Cd toxicity in maize in relation to cadmium translocation, cell distribution, antioxidant enzymes stimulation and enhanced endodermal apoplasmic barrier development. *Plant Growth Regul.* 70, 89–103. doi: 10.1007/S10725-012-9781-4

Maghsoudi, K., Arvin, M. J., and Ashraf, M. (2019). Mitigation of arsenic toxicity in wheat by the exogenously applied salicylic acid, 24-Epi-brassinolide and silicon. *J. Soil Sci. Plant Nutr.* 20, 577–588. doi: 10.1007/S42729-019-00147-3

Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends Ecol. Evol.* 20, 229–237. doi: 10.1016/j.tree.2005.02.010

Mandadi, K. K., Misra, A., Ren, S., and McKnight, T. D. (2009). BT2, a BTB protein, mediates multiple responses to nutrients, stresses, and hormones in *Arabidopsis*. *Plant Physiol.* 150, 1930–1939. doi: 10.1104/PP.109.139220

Maroco, J. P., Edwards, G. E., and Ku, M. S. B. (1999). Photosynthetic acclimation of maize to growth under elevated levels of carbon dioxide. *Planta* 210, 115–125. doi: 10.1007/S004250050660

Martins, A. P. B., Brito, M. S., Mayer, J. L. S., Llerena, J. P. P., Oliveira, J. F., Takahashi, N. G., et al. (2018). Ectopic expression of sugarcane SHINE changes cell wall and improves biomass in rice. *Biomass Bioenergy* 119, 322–334. doi: 10.1016/j.biombioe.2018.09.036

Masuda, H. P., Cabral, L. M., de Veylder, L., Tanurdzic, M., Engler, J. A., Geelen, D., et al. (2008). ABAP1 is a novel plant armadillo BTB protein involved in DNA replication and transcription. *EMBO J.* 27, 2746–2756. doi: 10.1038/EMBOJ.2008.191

Mauriat, M., and Moritz, T. (2009). Analyses of GA20ox- and GID1-over-expressing aspen suggest that gibberellins play two distinct roles in wood formation. *Plant J.* 58, 989–1003. doi: 10.1111/J.1365-313X.2009.03836.X

McCann, M., and Carpita, N. (2007). Looking for invisible phenotypes in cell wall mutants of *Arabidopsis thaliana*. *Plant Biosyst.* 139, 80–83. doi: 10.1080/11263500500059801

McCarthy, R. L., Zhong, R., and Ye, Z.-H. (2009). MYB83 is a direct target of SND1 and acts redundantly with MYB46 in the regulation of secondary cell wall biosynthesis in *Arabidopsis*. *Plant Cell Physiol.* 50, 1950–1964. doi: 10.1093/PCP/PCP139

McClung, C. R. (2014). Wheels within wheels: new transcriptional feedback loops in the *Arabidopsis* circadian clock. *F1000Prime Rep.* 6:2. doi: 10.12703/P6-2

Meyer, J. H., and Keeping, M. G. (2005). Impact of silicon in alleviating biotic stress in sugarcane in South Africa. *Proc. S. Afr. Sugar Technol. Assoc.* 23, 14–18.

Miller, M., Zhang, C., and Chen, Z. J. (2012). Ploidy and hybridity effects on growth vigor and gene expression in *Arabidopsis thaliana* hybrids and their parents. *G3 Genes Genomes Genet.* 2, 505–513. doi: 10.1534/G3.112.002162

Ming, R., Liu, S. C., Moore, P. H., Irvine, J. E., and Paterson, A. H. (2001). QTL analysis in a complex autopolyploid: genetic control of sugar content in sugarcane. *Genome Res.* 11, 2075–2084. doi: 10.1101/gr.198801

Mizoguchi, T., Wheatley, K., Hanzawa, Y., Wright, L., Mizoguchi, M., Song, H.-R., et al. (2002). LHY and CCA1 are partially redundant genes required to maintain circadian rhythms in *Arabidopsis*. *Dev. Cell* 2, 629–641. doi: 10.1016/s1534-5807(02)00170-3

Mullet, J., Morishige, D., McCormick, R., Truong, S., Hilley, J., McKinley, B., et al. (2014). Energy sorghum—a genetic model for the design of C_4 grass bioenergy crops. *J. Exp. Bot.* 65, 3479–3489. doi: 10.1093/JXB/ERU229

Mullet, J. E. (2017). High-biomass C_4 grasses—filling the yield gap. *Plant Sci.* 261, 10–17. doi: 10.1016/j.plantsci.2017.05.003

Murphy, R. L., Klein, R. R., Morishige, D. T., Brady, J. A., Rooney, W. L., Miller, F. R., et al. (2011). Coincident light and clock regulation of pseudoresponse regulator protein 37 (PRR37) controls photoperiodic flowering in sorghum. *Proc. Natl. Acad. Sci. U.S.A.* 108, 16469–16474. doi: 10.1073/PNAS.1106212108

Murphy, R. L., Morishige, D. T., Brady, J. A., Rooney, W. L., Yang, S., Klein, P. E., et al. (2014). Ghd7 (Ma6) represses sorghum flowering in long days: Ghd7 alleles enhance biomass accumulation and grain production. *Plant Genome* 7:plantgenome2013.11.0040. doi: 10.3835/PLANTGENOME2013.11.0040

Müssig, C. (2005). Brassinosteroid-promoted growth. *Plant Biol.* 7, 110–117. doi: 10.1055/S-2005-837493

Noor, M. A. (2017). Nitrogen management and regulation for optimum NUE in maize – a mini review. *Cogent Food Agric.* 3:1348214. doi: 10.1080/23311932.2017.1348214

Ohashi-Ito, K., Oda, Y., and Fukuda, H. (2010). *Arabidopsis* VASCULAR-RELATED NAC-DOMAIN6 directly regulates the genes that govern programmed cell death and secondary wall formation during xylem differentiation. *Plant Cell* 22, 3461–3473. doi: 10.1105/TPC.110.075036

Ohta, M., Matsui, K., Hiratsu, K., Shinshi, H., and Ohme-Takagi, M. (2001). Repression domains of class II ERF transcriptional repressors share an essential motif for active repression. *Plant Cell* 13, 1959–1968. doi: 10.1105/TPC.010127

Olson, S. N., Ritter, K., Rooney, W., Kemanian, A., McCarl, B. A., Zhang, Y., et al. (2012). High biomass yield energy sorghum: developing a genetic model for C_4 grass bioenergy crops. *Biofuels Bioprod. Biorefin.* 6, 640–655. doi: 10.1002/BBB.1357

Park, J.-J., Yoo, C. G., Flanagan, A., Pu, Y., Debnath, S., Ge, Y., et al. (2017). Defined tetra-allelic gene disruption of the 4-coumarate:coenzyme A ligase 1 (Pv4CL1) gene by CRISPR/Cas9 in switchgrass results in lignin reduction and improved sugar release. *Biotechnol. Biofuels* 10:284. doi: 10.1186/S13068-017-0972-0

Pauly, M., and Keegstra, K. (2008). Cell-wall carbohydrates and their modification as a resource for biofuels. *Plant J.* 54, 559–568. doi: 10.1111/J.1365-313X.2008.03463.X

Peña, P. A., Quach, T., Sato, S., Ge, Z., Nersesian, N., Changa, T., et al. (2017). Expression of the maize dof1 transcription factor in wheat and sorghum. *Front. Plant Sci.* 8:434. doi: 10.3389/fpls.2017.00434

Petti, C., Hirano, K., Stork, J., and DeBolt, S. (2015). Mapping of a cellulose-deficient mutant named dwarf1-1 in *Sorghum bicolor* to the green revolution gene gibberellin20-oxidase reveals a positive regulatory association between gibberellin and cellulose biosynthesis. *Plant Physiol.* 169, 705–716. doi: 10.1104/pp.15.00928

Ping, L., Luo, Y., Wu, L., Qian, W., Song, J., and Christie, P. (2006). Phenanthrene adsorption by soils treated with humic substances under different pH and

temperature conditions. *Environ. Geochem. Health* 28, 189–195. doi: 10.1007/s10653-005-9030-0

Poorter, H., and Villar, R. (1997). *The Fate of Acquired Carbon in Plants: Chemical Composition and Construction Costs*. Available online at: https://agris.fao.org/agris-search/search.do?recordID=US1997062597 (accessed September 6, 2021).

Pruneda-Paz, J. L., Breton, G., Para, A., and Kay, S. A. (2009). A functional genomics approach reveals CHE as a component of the *Arabidopsis* circadian clock. *Science* 323, 1481–1485. doi: 10.1126/SCIENCE.1167206

Qandeel, M., Jabbar, A., Haider, F. U., Virk, A. L., and Ain, N. U. (2020). Effects of plant growth regulators and dates planting on spring maize production under agro-climatic conditions of Faisalabad, Pakistan. *Pak. J. Agric. Agric. Eng. Vet. Sci.* 36, 120–128. doi: 10.47432/2020.36.2.5

Ranocha, P., Chabannes, M., Chamayou, S., Danoun, S., Jauneau, A., Boudet, A.-M., et al. (2002). Laccase down-regulation causes alterations in phenolic metabolism and cell wall structure in poplar. *Plant Physiol.* 129, 145–155. doi: 10.1104/PP.010988

Ravi, M., and Chan, S. W. L. (2010). Haploid plants produced by centromere-mediated genome elimination. *Nature* 464, 615–618. doi: 10.1038/nature08842

Rehman, A., Shahzad, B., Haider, F. U., Ibraheem Ahmed, H. A., Lee, D.-J., Im, S. Y., et al. (2022a). "Chapter 1 - An introduction to brassinosteroids: history, biosynthesis, and chemical diversity," in *Brassinosteroids in Plant Developmental Biology and Stress Tolerance*, eds J. Q. Yu, G. J. Ahammed, and P. Krishna (Amsterdam: Elsevier), 1–14. doi: 10.1016/B978-0-12-813227-2.00006-0

Rehman, A., Shahzad, B., Haider, F. U., Moeen-ud-din, M., Ullah, A., and Khan, I. (2022b). "Chapter 8 - Brassinosteroids in plant response to high temperature stress," in *Brassinosteroids in Plant Developmental Biology and Stress Tolerance*, ed. G. J. A. P. K. Jing Quan Yu (Amsterdam: Elsevier), 173–187. doi: 10.1016/B978-0-12-813227-2.00014-X

Reich, P. B., Hobbie, S. E., Lee, T. D., and Pastore, M. A. (2018). Response to comment on "Unexpected reversal of C_3 versus C_4 grass response to elevated CO_2 during a 20-year field experiment". *Science* 361, 317–320. doi: 10.1126/science.aau8982

Riddle, N. C., Kato, A., and Birchler, J. A. (2006). Genetic variation for the response to ploidy change in *Zea mays* L. *Theor. Appl. Genet.* 114, 101–111. doi: 10.1007/s00122-006-0414-z

Rojas, C. A., Eloy, N. B., De Freitas Lima, M., Rodrigues, R. L., Franco, L. O., Himanen, K., et al. (2009). Overexpression of the *Arabidopsis* anaphase promoting complex subunit CDC27a increases growth rate and organ size. *Plant Mol. Biol.* 71, 307–318. doi: 10.1007/S11103-009-9525-7

Rostami, S., Azhdarpoor, A., Ali, M., Dehghani, M., Reza, M., Jaskulak, M., et al. (2021). The effects of exogenous application of melatonin on the degradation of polycyclic aromatic hydrocarbons in the rhizosphere of Festuca *. *Environ. Pollut.* 274:116559. doi: 10.1016/j.envpol.2021.116559

Sakamoto, T., Miura, K., Itoh, H., Tatsumi, T., Ueguchi-Tanaka, M., Ishiyama, K., et al. (2004). An overview of gibberellin metabolism enzyme genes and their related mutants in rice. *Plant Physiol.* 134, 1642–1653. doi: 10.1104/PP.103.033696

Sánchez-Rodríguez, C., Rubio-Somoza, I., Sibout, R., and Persson, S. (2010). Phytohormones and the cell wall in *Arabidopsis* during seedling growth. *Trends Plant Sci.* 15, 291–301. doi: 10.1016/j.tplants.2010.03.002

Sattler, S. E., Saballos, A., Xin, Z., Funnell-Harris, D. L., Vermerris, W., and Pedersen, J. F. (2014). Characterization of novel sorghum brown midrib mutants from an EMS-mutagenized population. *G3 Genes Genomes Genet.* 4, 2115–2124. doi: 10.1534/G3.114.014001

Scheben, A., and Edwards, D. (2018). Towards a more predictable plant breeding pipeline with CRISPR/Cas-induced allelic series to optimize quantitative and qualitative traits. *Curr. Opin. Plant Biol.* 45, 218–225. doi: 10.1016/J.PBI.2018.04.013

Scofield, S., Jones, A., and Murray, J. A. H. (2014). The plant cell cycle in context. *J. Exp. Bot.* 65, 2557–2562. doi: 10.1093/JXB/ERU188

Scully, E. D., Gries, T., Sarath, G., Palmer, N. A., Baird, L., Serapiglia, M. J., et al. (2016). Overexpression of SbMyb60 impacts phenylpropanoid biosynthesis and alters secondary cell wall composition in *Sorghum bicolor*. *Plant J.* 85, 378–395. doi: 10.1111/tpj.13112

Sharma, A., Sidhu, G. P. S., Araniti, F., Bali, A. S., Shahzad, B., Tripathi, D. K., et al. (2020). The role of salicylic acid in plants exposed to heavy metals. *Molecules* 25:540. doi: 10.3390/molecules25030540

Shen, H., He, X., Poovaiah, C. R., Wuddineh, W. A., Ma, J., Mann, D. G. J., et al. (2012). Functional characterization of the switchgrass (*Panicum virgatum*) R2R3-MYB transcription factor PvMYB4 for improvement of lignocellulosic feedstocks. *New Phytol.* 193, 121–136. doi: 10.1111/J.1469-8137.2011.03922.X

Singh, K. B., Foley, R. C., and Oñate-Sánchez, L. (2002). Transcription factors in plant defense and stress responses. *Curr. Opin. Plant Biol.* 5, 430–436. doi: 10.1016/S1369-5266(02)00289-3

Somerville, C., Youngs, H., Taylor, C., Davis, S. C., and Long, S. P. (2010). Feedstocks for lignocellulosic biofuels. *Science* 329, 790–792. doi: 10.1126/SCIENCE.1189268

Sonbol, F.-M., Fornalé, S., Capellades, M., Encina, A., Touriño, S., Torres, J.-L., et al. (2009). The maize ZmMYB42 represses the phenylpropanoid pathway and affects the cell wall structure, composition and degradability in *Arabidopsis thaliana*. *Plant Mol. Biol.* 70, 283–296. doi: 10.1007/S11103-009-9473-2

Strayer, C., Oyama, T., Schultz, T. F., Raman, R., Somers, D. E., Mas, P., et al. (2000). Cloning of the *Arabidopsis* clock gene TOC1, an autoregulatory response regulator homolog. *Science* 289, 768–771. doi: 10.1126/SCIENCE.289.5480.768

Subramaniam, Y., Masron, T. A., and Azman, N. H. N. (2020). Biofuels, environmental sustainability, and food security: a review of 51 countries. *Energy Res. Soc. Sci.* 68:101549. doi: 10.1016/j.erss.2020.101549

Sun, H., Xu, H., Li, B., Shang, Y., Wei, M., Zhang, S., et al. (2021). The brassinosteroid biosynthesis gene, ZmD11, increases seed size and quality in rice and maize. *Plant Physiol. Biochem.* 160, 281–293. doi: 10.1016/J.PLAPHY.2021.01.031

Su'udi, M., Cha, J. Y., Jung, M. H., Ermawati, N., Han, C. D., Kim, M. G., et al. (2012). Potential role of the rice OsCCS52A gene in endoreduplication. *Planta* 235, 387–397. doi: 10.1007/S00425-011-1515-8

Tanveer, M., Shahzad, B., Sharma, A., Biju, S., and Bhardwaj, R. (2018). 24-Epibrassinolide; an active brassinolide and its role in salt stress tolerance in plants: a review. *Plant Physiol. Biochem.* 130, 69–79. doi: 10.1016/j.plaphy.2018.06.035

Tanveer, M., Shahzad, B., Sharma, A., and Khan, E. A. (2019). 24-Epibrassinolide application in plants: an implication for improving drought stress tolerance in plants. *Plant Physiol. Biochem.* 135, 295–303. doi: 10.1016/J.PLAPHY.2018.12.013

Tetreault, H. M., O'Neill, P., Toy, J., Gries, T., Funnell-Harris, D. L., and Sattler, S. E. (2020). Field evaluation of sorghum (*Sorghum bicolor*) lines that overexpress two monolignol-related genes that alter cell wall composition. *Bioenergy Res.* 14, 1–12. doi: 10.1007/s12155-020-10218-4

Torres, A. F., Visser, R. G. F., and Trindade, L. M. (2015). Bioethanol from maize cell walls: genes, molecular tools, and breeding prospects. *GCB Bioenergy* 7, 591–607. doi: 10.1111/gcbb.12164

Tran, T. A., and Popova, L. P. (2013). Functions and toxicity of cadmium in plants: recent advances and future prospects. *Turk. J. Bot.* 37, 1–13.

Truong, S. K., McCormick, R. F., and Mullet, J. E. (2017). Bioenergy sorghum crop model predicts VPD-limited transpiration traits enhance biomass yield in water-limited environments. *Front. Plant Sci.* 8:335. doi: 10.3389/FPLS.2017.00335

Ullah, A., Rolf Richter, P., Ahmed, H., Pratap Singh, V., Mohan Prasad, S., Kumar Chauhan, D., et al. (2016). Silicon nanoparticles more efficiently alleviate arsenate toxicity than silicon in maize cultiver and hybrid differing in arsenate tolerance. *Front. Environ. Sci.* 4:46. doi: 10.3389/fenvs.2016.00046

van der Weijde, T., Alvim Kamei, C. L., Torres, A. F., Vermerris, W., Dolstra, O., Visser, R. G. F., et al. (2013). The potential of C_4 grasses for cellulosic biofuel production. *Front. Plant Sci.* 4:107. doi: 10.3389/fpls.2013.00107

van der Weijde, T., Kamei, C. L. A., Severing, E. I., Torres, A. F., Gomez, L. D., Dolstra, O., et al. (2017a). Genetic complexity of miscanthus cell wall composition and biomass quality for biofuels. *BMC Genomics* 18:406. doi: 10.1186/s12864-017-3802-7

van der Weijde, T., Kiesel, A., Iqbal, Y., Muylle, H., Dolstra, O., Visser, R. G. F., et al. (2017b). Evaluation of *Miscanthus sinensis* biomass quality as feedstock for conversion into different bioenergy products. *GCB Bioenergy* 9, 176–190. doi: 10.1111/GCBB.12355

Vanneste, S., and Friml, J. (2009). Auxin: a trigger for change in plant development. *Cell* 136, 1005–1016. doi: 10.1016/J.CELL.2009.03.001

Vélez-Bermúdez, I.-C., Salazar-Henao, J. E., Fornalé, S., López-Vidriero, I., Franco-Zorrilla, J.-M., Grotewold, E., et al. (2015). A MYB/ZML complex regulates

wound-induced lignin genes in maize. *Plant Cell* 27, 3245–3259. doi: 10.1105/TPC.15.00545

Vernoux, T., Autran, D., and Traas, J. (2000). Developmental control of cell division patterns in the shoot apex. *Plant Mol. Biol.* 43, 569–581. doi: 10.1023/A:1006464430936

Voorend, W., Nelissen, H., Vanholme, R., De Vliegher, A., Van Breusegem, F., Boerjan, W., et al. (2016). Overexpression of GA20-OXIDASE1 impacts plant height, biomass allocation and saccharification efficiency in maize. *Plant Biotechnol. J.* 14, 997–1007. doi: 10.1111/pbi.12458

Wai, C. M., Zhang, J., Jones, T. C., Nagai, C., and Ming, R. (2017). Cell wall metabolism and hexose allocation contribute to biomass accumulation in high yielding extreme segregants of a *Saccharum* interspecific F2 population. *BMC Genomics* 18:773. doi: 10.1186/s12864-017-4158-8

Wang, Y., Sun, J., Ali, S. S., Gao, L., Ni, X., Li, X., et al. (2020). Identification and expression analysis of *Sorghum bicolor* gibberellin oxidase genes with varied gibberellin levels involved in regulation of stem biomass. *Ind. Crops Prod.* 145:111951. doi: 10.1016/j.indcrop.2019.111951

Watling, J. R., Press, M. C., and Quick, W. P. (2000). Elevated CO_2 induces biochemical and ultrastructural changes in leaves of the C_4 cereal sorghum. *Plant Physiol.* 123, 1143–1152. doi: 10.1104/PP.123.3.1143

Wen-Jie, W., Ling, Q., Yuan-Gang, Z., Dong-Xue, S., Jing, A., Hong-Yan, W., et al. (2011). Changes in soil organic carbon, nitrogen, pH and bulk density with the development of larch (*Larix gmelinii*) plantations in China. *Glob. Chang. Biol.* 17, 2657–2676. doi: 10.1111/j.1365-2486.2011.02447.x

Whipple, C. J., Kebrom, T. H., Weber, A. L., Yang, F., Hall, D., Meeley, R., et al. (2011). grassy tillers1 promotes apical dominance in maize and responds to shade signals in the grasses. *Proc. Natl. Acad. Sci. U.S.A.* 108, E506–E512. doi: 10.1073/PNAS.1102819108

Wingler, A., Delatte, T. L., O'Hara, L. E., Primavesi, L. F., Jhurreea, D., Paul, M. J., et al. (2012). Trehalose 6-phosphate is required for the onset of leaf senescence associated with high carbon availability. *Plant Physiol.* 158, 1241–1251. doi: 10.1104/PP.111.191908

Wuddineh, W. A., Mazarei, M., Turner, G. B., Sykes, R. W., Decker, S. R., Davis, M. F., et al. (2015). Identification and molecular characterization of the switchgrass AP2/ERF transcription factor superfamily, and overexpression of PvERF001 for improvement of biomass characteristics for biofuel. *Front. Bioeng. Biotechnol.* 3:101. doi: 10.3389/FBIOE.2015.00101

Xia, J., Zhao, Y., Burks, P., Pauly, M., and Brown, P. J. (2018). A sorghum NAC gene is associated with variation in biomass properties and yield potential. *Plant Direct* 2:e00070. doi: 10.1002/pld3.70

Xiang, L., Yuling, L., Zhiying, M., and Zheng, L. (2020). Identification of maize aquaporin gene ZmPIP4c as a signature of C_4 traits. *Mol. Plant Breed.* 11, 2819–2825. doi: 10.5376/mgg.2020.11.0002

Xu, C., Wang, Y., Yu, Y., Duan, J., Liao, Z., Xiong, G., et al. (2012). Degradation of MONOCULM 1 by APC/CTAD1 regulates rice tillering. *Nat. Commun.* 3:750. doi: 10.1038/NCOMMS1743

Yong, W., Link, B., O'Malley, R., Tewari, J., Hunter, C. T., Lu, C.-A., et al. (2005). Genomics of plant cell wall biogenesis. *Planta* 221, 747–751. doi: 10.1007/S00425-005-1563-Z

Yu, Y., Steinmetz, A., Meyer, D., Brown, S., and Shen, W.-H. (2003). The tobacco A-type cyclin, Nicta;CYCA3;2, at the nexus of cell division and differentiation. *Plant Cell* 15, 2763–2777. doi: 10.1105/TPC.015990

Zargar, S. M., Mahajan, R., Bhat, J. A., Nazir, M., and Deshmukh, R. (2019). Role of silicon in plant stress tolerance: opportunities to achieve a sustainable cropping system. *3 Biotech* 9:73. doi: 10.1007/S13205-019-1613-Z

Zeng, X., Sheng, J., Zhu, F., Zhao, L., Hu, X., Zheng, X., et al. (2020). Differential expression patterns reveal the roles of cellulose synthase genes (CesAs) in primary and secondary cell wall biosynthesis in *Miscanthus* × *giganteus*. *Ind. Crops Prod.* 145:112129. doi: 10.1016/J.INDCROP.2020.112129

Zhang, G., Ge, C., Xu, P., Wang, S., Cheng, S., Han, Y., et al. (2021). The reference genome of *Miscanthus floridulus* illuminates the evolution of *Saccharinae*. *Nat. Plants* 7, 608–618. doi: 10.1038/s41477-021-00908-y

Zhang, Y., Paschold, A., Marcon, C., Liu, S., Tai, H., Nestler, J., et al. (2014). The Aux/IAA gene rum1 involved in seminal and lateral root formation controls vascular patterning in maize (*Zea mays* L.) primary roots. *J. Exp. Bot.* 65, 4919–4930. doi: 10.1093/jxb/eru249

Zhang, Y., von Behrens, I., Zimmermann, R., Ludwig, Y., Hey, S., and Hochholdinger, F. (2015). LATERAL ROOT PRIMORDIA 1 of maize acts as a transcriptional activator in auxin signalling downstream of the Aux/IAA gene rootless with undetectable meristem 1. *J. Exp. Bot.* 66, 3855–3863. doi: 10.1093/JXB/ERV187

Zhong, R., Lee, C., McCarthy, R. L., Reeves, C. K., Jones, E. G., and Ye, Z.-H. (2011). Transcriptional activation of secondary wall biosynthesis by rice and maize NAC and MYB transcription factors. *Plant Cell Physiol.* 52, 1856–1871. doi: 10.1093/PCP/PCR123

Ziska, L. H., and Bunce, J. A. (1997). Influence of increasing carbon dioxide concentration on the photosynthetic and growth stimulation of selected C_4 crops and weeds. *Photosynth. Res.* 54, 199–208. doi: 10.1023/A:1005947802161

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

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

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 28 June 2022
doi: 10.3389/fpls.2022.826875

Dissecting the Genetic Structure of Maize Leaf Sheaths at Seedling Stage by Image-Based High-Throughput Phenotypic Acquisition and Characterization

Jinglu Wang[1,2†], Chuanyu Wang[1,2†], Xianju Lu[1,2], Ying Zhang[1,2], Yanxin Zhao[3], Weiliang Wen[1,2], Wei Song[4] and Xinyu Guo[1,2]**

[1] Beijing Key Lab of Digital Plant, Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing, China, [2] National Engineering Research Center for Information Technology in Agriculture, Beijing Academy of Agriculture and Forestry Sciences, Beijing, China, [3] Beijing Key Laboratory of Maize DNA Fingerprinting and Molecular Breeding, Maize Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing, China, [4] Key Laboratory of Crop Genetics and Breeding of Hebei Province, Institute of Cereal and Oil Crops, Hebei Academy of Agriculture and Forestry Sciences, Shijiazhuang, China

The rapid development of high-throughput phenotypic detection techniques makes it possible to obtain a large number of crop phenotypic information quickly, efficiently, and accurately. Among them, image-based phenotypic acquisition method has been widely used in crop phenotypic identification and characteristic research due to its characteristics of automation, non-invasive, non-destructive and high throughput. In this study, we proposed a method to define and analyze the traits related to leaf sheaths including morphology-related, color-related and biomass-related traits at V6 stage. Next, we analyzed the phenotypic variation of leaf sheaths of 418 maize inbred lines based on 87 leaf sheath-related phenotypic traits. In order to further analyze the mechanism of leaf sheath phenotype formation, 25 key traits (2 biomass-related, 19 morphology-related and 4 color-related traits) with heritability greater than 0.3 were analyzed by genome-wide association studies (GWAS). And 1816 candidate genes of 17 whole plant leaf sheath traits and 1,297 candidate genes of 8 sixth leaf sheath traits were obtained, respectively. Among them, 46 genes with clear functional descriptions were annotated by single nucleotide polymorphism (SNPs) that both Top1 and multi-method validated. Functional enrichment analysis results showed that candidate genes of leaf sheath traits were enriched into multiple pathways related to cellular component assembly and organization, cell proliferation and epidermal cell differentiation, and response to hunger, nutrition and extracellular stimulation. The results presented here are helpful to further understand phenotypic traits of maize leaf sheath and provide a reference for revealing the genetic mechanism of maize leaf sheath phenotype formation.

Keywords: maize, leaf sheath, image-based traits, GWAS, pathways

OPEN ACCESS

Edited by:
Weizhen Liu,
Wuhan University of Technology, China

Reviewed by:
Chenglong Huang,
Huazhong Agricultural University, China
Huihui LI,
Institute of Crop Sciences (CAAS), China

****Correspondence:***
Wei Song
sw1717@126.com
Xinyu Guo
guoxy73@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: *01 December 2021*
Accepted: *17 February 2022*
Published: *28 June 2022*

Citation:
Wang J, Wang C, Lu X, Zhang Y, Zhao Y, Wen W, Song W and Guo X (2022) Dissecting the Genetic Structure of Maize Leaf Sheaths at Seedling Stage by Image-Based High-Throughput Phenotypic Acquisition and Characterization. Front. Plant Sci. 13:826875. doi: 10.3389/fpls.2022.826875

INTRODUCTION

Maize leaf sheath is located at the base of leaf and wraps around the stem node. It plays a role of protecting and supporting the leaf. At the same time, it can protect the young and tender intermediate meristems and young buds on the stem, and enhance the mechanical support of the stem (Dong et al., 2019). In the sink-source relationship, the leaf sheath can be used as a nutrient storage organ in the early stage, namely, "sink." And it can be also used as an organ for the production or export of assimilates in the later stage of growth, that is, "source." It is well known that leaf sheaths usually have elongation zones. As a result of intercellular growth, the cells elongate in two separate directions, above and below, and differentiate into longitudinally parallel vascular bundles (Russell and Evert, 1985). Hence, maize leaf sheaths can also be used as part of the "flow." In summary, the role of maize leaf sheaths in the plant is very important and deserves more attention and in-depth study. In addition, maize purple plant pigments are anthocyanin pigments. A large number of domestic and foreign studies have shown that purple-red anthocyanin pigments have anti-oxidation, anti-aging, immune enhancement and tumor prevention functions (Zhang et al., 2014; Li et al., 2020; Peniche-Pavia and Tiessen, 2020; Chatham and Juvik, 2021). Therefore, it is of great theoretical and practical importance to study the phenotypic characteristics of maize leaf sheaths and to analyze their genetic structure.

Maize has a rich diversity due to its long planting history and wide geographical span. Among them, the leaf sheath phenotype of maize also varies between populations, particularly in color. As a consequence, traditional studies of maize leaf sheaths are usually based on a qualitative description or classification of leaf sheath color. The leaf sheaths of maize commonly come in purple and green. It has been pointed out that these color changes are usually related to pigments (Fan et al., 2008). Li et al. (2018) conducted genetic analysis and gene localization for the purple leaf sheath trait using a recombinant inbred line population of maize and found that the gene *GRMZM5G822829* was highly significantly differentially expressed between the purple and green leaf sheath parents. Yang S. et al. (2014) used the maize white sheath inbred line K10 as the research material, and conducted preliminary genetic mechanism and gene mapping of the white sheath traits. The results showed that the white leaf sheath trait has nothing to do with cytoplasmic inheritance, but was controlled by recessive nuclear genes and was under polygenic control.

Nowadays, with the rapid development of high-throughput phenotyping technology, it has become possible to obtain massive crop phenotypic information quickly, efficiently and accurately (Zhao et al., 2019). Among them, image-based phenotype acquisition methods have been widely used in crop phenotype identification and characterization due to their automatic, non-invasive, non-destructive, and high-throughput characteristics (Green et al., 2012; Tanabata et al., 2012; Bucksch et al., 2014; Gage et al., 2017; Chopin et al., 2018; Zhang et al., 2020; Zhou et al., 2021). Based on image data, a variety of phenotypes can be analyzed, which can break through the limitations of subjective cognition and carry out deeper research (Mir et al., 2019). For example, the web-based tool PhenoPhyte is a flexible affordable method to quantify 2D phenotypes from imagery. And it can distinguish different experimental Settings through experimental database management and calculate the phenotypic parameters related to leaf area in phenotypic images (Green et al., 2012). SmartGrain, as a high-throughput phenotyping software for measuring seed shape through image analysis, using a new image analysis method to reduce the time taken in the preparation of seeds and in image capture (Tanabata et al., 2012). TIPS is a system for automated image-based phenotyping of maize tassels, and it allows morphological features of maize tassels to be quantified automatically, with minimal disturbance, at a scale that supports population-level studies. And it is expected to accelerate the discovery of associations between genetic loci and tassel morphology characteristics (Gage et al., 2017). Another maize image analysis software is Maize-IAS, which is an integrated application supporting one-click analysis of maize phenotype, embedding multiple functions, with a high efficiency and potential capability to image-based plant research (Zhou et al., 2021). Thus, image technology has become a high-throughput means to obtain and analyze the phenotypic information of large populations of crops. The phenotypic information can be used for quantitative trait loci mapping and genome-wide association studies. It is helpful to break the gap between crop traits and genetic markers and promote the study of crop phenotypic-genotype association (Zhao et al., 2019; Yang et al., 2020; Song et al., 2021).

Genome-wide association study (GWAS) as an analytical method for identifying the relationship between a target trait and a genetic marker or candidate gene within a group of individuals, provides a powerful tool for researchers concerned with and exploring the genetic mechanisms of phenotype formation across multiple individuals (Xiao et al., 2016; Liu and Yan, 2019). In particular, the mixed linear model (MLM) methods have proven useful in controlling for population structure and relatedness within GWAS. In the MLM-based methods, population structure is fitted as a fixed effect, while kinship among individuals is incorporated into the variance-covariance structure of individual random effects (Zhang et al., 2010). Since the publication of maize B73 reference genome (Schnable et al., 2009), GWAS has been widely used in maize genetics research, and has played a great role in the analysis of genetic mechanisms such as traditional maize agronomic traits (Wallace et al., 2016; Zhou et al., 2016; Li et al., 2017; Dai et al., 2018; Du et al., 2018; Zhou et al., 2018; Owens et al., 2019), key phenotypes (Cui et al., 2016; Li et al., 2016; Liu et al., 2016; Zhang et al., 2016a; Sanchez et al., 2018; Guo et al., 2019; Mazaheri et al., 2019) and stress resistance (Zhang et al., 2016b; Shi et al., 2018; Cooper et al., 2019; Wang et al., 2019; Xie et al., 2019). However, there are few genetic studies on the phenotype of maize leaf sheath.

Through image acquisition, image segmentation, feature extraction and manual measurement of leaf sheaths of 418 maize inbred lines at V6 stage, this study proposed a method to define and analyze the shape, size, color and other phenotypes related to leaf sheaths, and developed a pipeline for image-based traits with phenotypic data analysis and genetic

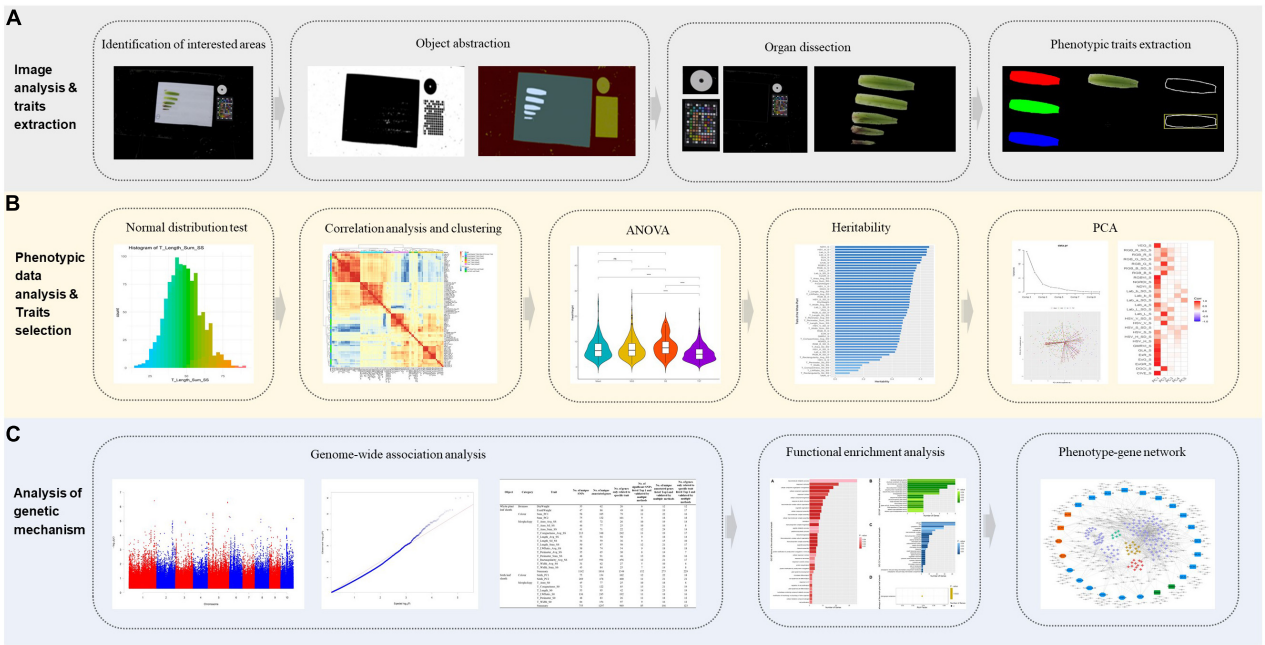


FIGURE 1 | The flowchart of image-based high-throughput phenotypic analysis and genetic mechanism analysis of maize leaf sheaths at seedling stage. **(A)** Image analysis and phenotypic traits extraction. **(B)** Phenotypic data analysis and key traits selection for genetic analysis. **(C)** Genetic mechanism analysis of key traits.

mechanism analysis (**Figure 1**). In addition, 87 leaf sheath-related phenotypic traits including morphology, color and biomass were obtained. Based on these phenotypic traits, leaf sheath characteristics of maize association analysis population were analyzed. In order to further analyze the mechanism of leaf sheath phenotype formation, 25 key traits of maize were analyzed by GWAS, and 1,816 candidate genes of 17 whole plant leaf sheath traits and 1,297 candidate genes of 8 sixth leaf sheath traits were obtained, respectively. This study has achieved high-throughput acquisition of the phenotype from maize leaf sheath. And it also can provide a reference for revealing the genetic mechanism of maize leaf sheath phenotype formation.

MATERIALS AND METHODS

Plant Materials, Growth Conditions, and Sample Collection

418 inbred lines used in this study were from the maize association mapping panel published by Yang et al. (2011); **Supplementary Table 1**, which were classified into four subpopulations: Non-stiff stalk (NSS) with 124 lines, Stiff stalk (SS) with 31 lines, Tropical-subtropical (TST) with 164 lines, and 99 mixed lines (Mixed). The plants were grown in the Beijing Academy of Agriculture and Foresting Science in Beijing, China. Maize seeds were planted manually at a depth of 5 cm on 17 May 2019. Each inbred line was planted in 4 rows with 7 plants per row. Planting density and water and fertilizer management were based on local field production (Lu et al., 2020).

Image Acquisition, Analysis, and Feature Extraction

Maize plants were grown to the V6 stage, and three plants were sampled from each inbred line population. The leaf sheaths were spread out on a white soft background plate and fixed with pins. Blade images were captured by an image acquisition device (Canon EOS 5D Mark III) with a resolution of 5,760 × 3,840 pixels. The image processing program (**Figure 1A**) is developed by Visual Studio Express 2015, using the open-source image processing library OpenCV 2.3.

The image processing and feature extraction methods were summarized as follows: (**a**) **Identification of interested areas.** The original color image was converted into a grayscale image, an adaptive thresholding algorithm was used to segment foreground and background. (**b**) **Object abstraction.** The foreground contained a circular marker, a color checker board and leaf sheaths, these components were separated according to shape, inner composition pattern and chromatic property. (**c**) **Organ dissection.** The largest contour was considered as the sixth leaf sheath, the rest contours were labeled as the fifth leaf sheath candidate regions. First, found out bounding box of these candidate regions, and calculated length/width ratio, if the ratio value was more than 3.0, then we argued that belonged to leaf sheath candidate. Chosen the largest leaf sheath candidate regions to compute centroid coordinate which denote by C_{cdit}, and centroid of sixth leaf sheath denoted by C_{six}, the Euclidean distance of above two point is D_{cent}, if the D_{cent} was less than 1/2 length of sixth leaf sheath bounding box, the candidate regions was labeled as the fifth leaf sheath, repeated the procedure for the remaining candidates, until the last one was tested. (**d**)

Phenotypic traits calculation. Phenotypic measurement and image-based feature extraction were performed on the whole plant leaf sheath and the sixth leaf sheath of maize at the V6 stage, respectively. The specific calculation and definitions for each trait are detailed in **Supplementary Table 2** and **Supplementary Note 1**.

Together with the two biomass traits, dry weight (DryWeight) and fresh weight (FreshWeight) of the whole plant, measured manually by electronic balances, totally 87 leaf sheaths related traits that covering three types (morphology, color and biomass) and two objects (the whole plant leaf sheath and the sixth leaf sheath) were obtained in this study (**Supplementary Table 2**).

Statistical Analysis of Phenotypic Data

The "lm" function in R (Version 3.6.3) software[1] was used to carry out linear regression analysis on the leaf sheath area extracted by the image-based method and the dry/fresh weight measured by manual. The R^2 obtained from the model represent the accuracy of the software algorithm.

Analysis of variance (ANOVA) and descriptive statistical analysis were conducted via R (Version 3.6.3) software to determine whether each phenotype is different between different subpopulations. Pearson correlation analysis was used to calculate the correlation coefficients among phenotypic traits. And *pamk*, a function of R package "FPC," was used to perform unsupervised hierarchical cluster analysis (HCA) using Pearson correlation coefficient as distance measure, and then 87 traits were grouped based on clustering.

Broad sense heritability (H^2) usually means the percentage of genetic variation (V_A) to the total variation of a phenotype. It can be used to compare the relationship between genetic (σ_A^2) and environmental (σ_e^2) factors for a specific phenotypic variation (V_P). Heritability (H^2) was calculated for each trait as follows:

$$H^2 = \frac{V_A}{V_P} = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_e^2}$$

where σ_A^2 is the genetic variance, σ_e^2 is the environmental variance. The analysis was performed in ASReml-R v.4.0 by using the "asreml" function of R package asreml (Butler, 2009).

Genome-Wide Association Study

Genotypic data of maize association mapping panel were obtained from Maizego.[2] Firstly, the genotypic data of 418 inbred lines needed in this study were extracted, and 794,722 SNPs with minimum allele frequency (MAF) greater than 0.05 and call rate greater than 0.9 filtered by PLINK 1.09 software were used in GWAS. For GWAS, a multi-locus random-SNP-effect mixed linear model tool (R package "mrMLM" version 4.0) (Zhang et al., 2019) including six multi-locus GWAS methods (mrMLM, FASTmrMLM, FASTmrEMMA, ISIS EM-BLASSO, pLARmEB, and pKWmEB) was used on each leaf sheath related phenotypic traits separately to test the statistical association between phenotypes and genotypes. In addition, population structure estimated by STRUCTURE program version 2.3.4 (Hubisz et al., 2009) and relative kinship calculated by TASSEL 5 (Bradbury et al., 2007) with 794,722 SNPs were brought into the model. These six Multi-locus GWAS methods were processed in two steps. First, each SNP on the genome was filtered with a P-value $\leq$ 0.5/N, N is the total number of genome-wide SNPs. Then, all the SNPs that are potentially associated with the trait were included in a multi-locus genetic model further screened with a defeat P-value = 0.0002 to declare a significance of SNPs that associated with a given trait. The results obtained by the six multi-locus GWAS methods were regarded as significant SNPs associated with phenotypic traits. Furthermore, SNPs with the highest significance obtained by each method were regarded as Top 1, and SNPs identified by multiple methods were considered to be more reliable results. All candidate genes were annotated by ANNOVAR software according to the latest maize B73 reference genome (B73 RefGen_v4) available in EnsemblPlants[3] and NCBI Gene database.[4]

Functional and Network Analysis

The biological functions of candidate genes with high confidence for each phenotypic trait (Top1 SNP annotation or multiple GWAS validation) were explored by pathway enrichment analysis. Enrichment analysis of Gene Ontology (GO) (Ashburner et al., 2000) was conducted using PlantRegMap (Jin et al., 2015). And KOBAS V3.0 (Bu et al., 2021) was used to enrich Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa, 2002) pathway. Among them, GO terms and KEGG pathways with the P-value less than 0.05 were considered to be significantly enriched results.

In order to have a better view of the relationship between each trait and its candidate genes, an open-source software platform (Cytoscape v3.7.2) (Shannon et al., 2003) was used to visualize the complex trait-candidate gene-pathway network and integrate the input data by their attribute information.

RESULTS

Phenotypic Extraction of Leaf Sheath

In this study, image analysis was used to replace the traditional leaf sheath phenotype acquisition methods. In addition to conventional traits such as length, width, and surface area of leaf sheaths, many traits such as leaf sheaths morphology and color were also extracted based on image, realizing high-throughput acquisition of phenotypic of maize leaf sheaths. After processing the original image, a total of 1,116 valid image samples were obtained, covering 418 inbred lines. According to these leaf sheath images, the characteristics of the whole plant leaf sheath and the sixth leaf sheath of maize at V6 stage were extracted, and a total of 85 2D leaf sheath-related traits were obtained. Together with two biomass traits obtained by measuring the dry and fresh weight of the whole plant leaf sheath, totally 87 traits covering morphology, color and biomass these three

[1] https://cran.r-project.org/

[2] www.maizego.org/Resources.html

[3] http://plants.ensembl.org/Zea_mays/Info/Traits

[4] https://www.ncbi.nlm.nih.gov/gene

types (**Supplementary Table 2** and **Supplementary Note 1**) were analyzed in this study. Of these, there were 50 whole plant leaf sheath traits, including two biomass traits, 18 morphological traits and 30 color traits. And 37 traits of the sixth leaf sheath, including 7 morphological traits and 30 color traits.

The measurement accuracy of the image-based phenotypic acquisition method was valued by the linear regression analysis on the leaf sheath area extracted by the image-based method and the dry/fresh weight measured by manual, and the R^2 obtained from the model represent the accuracy of the software algorithm. As shown in **Figure 2**, the R^2 of two models were 0.77 and 0.81, respectively. The R^2 of both models were close to 1, indicating that the measurement accuracy of the image-based phenotypic acquisition method is high, and the traits could be used for subsequent analysis.

Phenotypic Characteristics of Leaf Sheath

The basic statistical analysis results (**Supplementary Table 3**) of 87 leaf sheath traits showed that the phenotypic traits of inbred lines in maize association analysis population had extensive continuous variation, with the variation coefficient ranging from −0.67 to 21.49. Furthermore, it can be seen from the data histogram that the phenotypic traits data were normally distributed, indicating that all traits were quantitative traits.

Pearson correlation analysis was performed on 87 leaf sheath phenotypes, and clustering was performed based on Pearson correlation coefficient, as shown in **Figure 3**. Cluster analysis results showed that 87 phenotypes of the three types could be divided into 6 groups, and each group had clear characteristics (marked with different colors in **Figure 3**). The Morphological characteristics of leaf sheath can be divided into three groups. Group I (Morphological Traits_Basic): 16 basic morphological traits describing leaf sheath length, width and area, etc. Group II (Morphological Traits_Shape1): 4 traits were used to describe the morphological shape of leaf sheath. And Group III (Morphological Traits_Shape2): 5 traits to characterize the variation of morphological type of leaf sheath. There was no significant correlation between the 9 traits describing leaf sheath shape in the two groups and other traits, indicating that leaf sheath shape was basically unrelated to leaf sheath size, area and color. The 16 basic morphological traits of leaf sheath had a significant positive correlation with DryWeight and FreshWeight (P-value < 0.05), and clustered into the same group (Morphological Traits_Basic and Biomass Traits). This result is consistent with prior knowledge, which indicates the reliability of data and the significance of obtaining various traits from images. The leaf sheath Color Traits were also divided into three groups. The first group (Color Traits_Subset1) consisted mainly of comprehensive color traits, the second group (Color Traits_Subset2) of traits were mostly the variation degree of the single-channel color values, and the third group (Color Traits_Subset3) was composed of single-channel color traits and four comprehensive color traits. The 24 comprehensive color traits were separated into two groups, because CIVE, CIVE_S, ExR, and ExR_S mainly represent red, while the other comprehensive traits mainly represent blue and green, indicating the accuracy of data extraction.

87 phenotypic traits were analyzed among different subpopulations in turn. The results showed that the inbred lines of TST subpopulation had distinct characteristics and were significantly different from at least one subpopulation in 84 traits (96.55%) (P-value < 0.05). Among them, 64 traits (73.56%) showed significant differences between TST and all other three subpopulations (P-value < 0.05) (**Supplementary Figure 1**). The traits with significant differences between TST and other subpopulations covered all three types of traits, indicating that the leaf sheaths of tropical and subtropical maize inbred lines (TST) were different from those of other climate zone maize inbred lines in terms of morphology, color and biomass. In order to further explore which traits had the greatest difference between TST and other subpopulations, excluding the two biomass traits, the other 62 phenotypic traits were divided into four groups according to trait types and research objects. Consequently, the four groups were 12 leaf sheath morphological traits of whole plant, 21 leaf sheath color traits of whole plant, 4 leaf sheath morphological traits and 25 leaf sheath color traits of the sixth leaf, respectively. Then principal component analysis (PCA) was carried out for each group of traits, and the results showed that the samples analyzed in each group were divided into two categories (**Figure 4A** and **Supplementary Figure 2**). However, the Average Silhouette Width is the highest after clustering according to 12 morphological traits of the whole plant leaf sheath, which is 0.54 (**Figure 4B**). In addition, 10 of the 12 traits (T_Area_Avg_SS, T_Area_Sd_SS, T_Area_Sum_SS, T_Compactness_Avg_SS, T_Length_Avg_SS, T_Width_Avg_SS, T_Length_Sd_SS, T_Width_Sd_SS, T_LWRatio_Avg_SS, T_Width_Sum_SS, T_Perimeter_Avg_SS, T_Perimeter_Sd_SS) were basic morphological traits of leaf sheath morphology, suggesting that the main differences between TST and other subpopulations were manifested in the conventional phenotypic traits such as leaf sheath length, width and area.

Heritability Analysis

Heritability analysis was performed on 87 leaf sheath phenotypic traits extracted from 2D images, and the results are shown in **Figures 5A,B**. For the whole plant leaf sheath traits, the heritability of these 50 traits ranged from 1.01E-07 to 0.6601. Among them, the heritability of DryWeight and FreshWeight was 0.5234 and 0.5429, respectively. And the heritability of 18 morphological traits ranged from 0.1029 to 0.5470, and 13 (72.22%) of these traits had a heritability greater than 0.3. Except VARI-S, the heritability of the other 29 color traits ranged from 0.3142 to 0.6601, and 29 (96.67%) of these color traits had a heritability greater than 0.3. For the sixth leaf sheath traits, the heritability of these 37 traits ranged from 0.1594 to 0.5754. And the heritability of 7 morphological traits ranged from 0.2683 to 0.5754, of which 6 (85.71%) had heritability greater than 0.3. The heritability of 30 color traits ranged from 0.1594 to 0.5955, and 27 (90.00%) of them had heritability greater than 0.3. To further investigate the genetic mechanism of phenotypic traits related to maize leaf sheaths, traits with heritability greater than 0.3 were screened for further genetic analysis in this study.

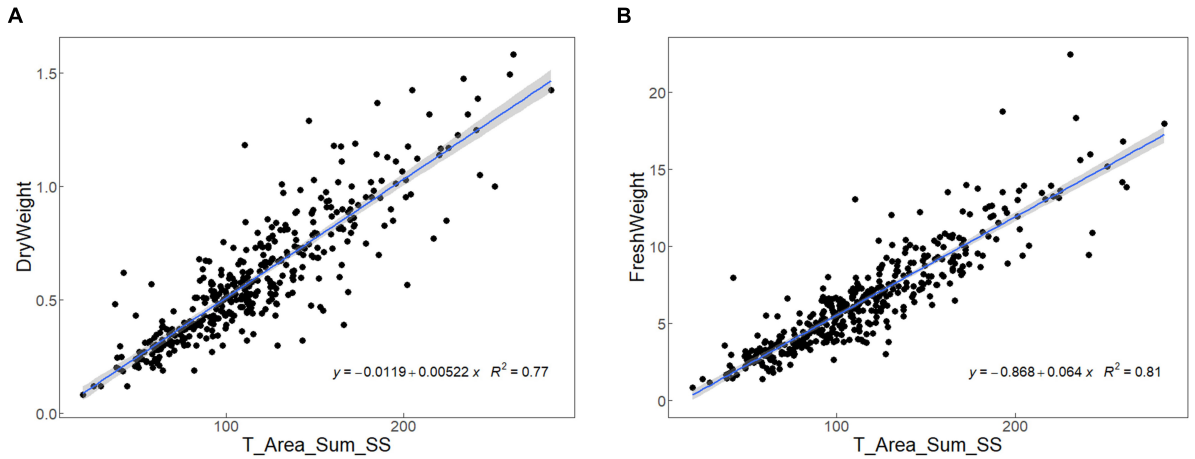


FIGURE 2 | The R^2 obtained from the linear regression analysis on the leaf sheath area extracted by the image-based method and the dry **(A)**/fresh **(B)** weight measured by manual.

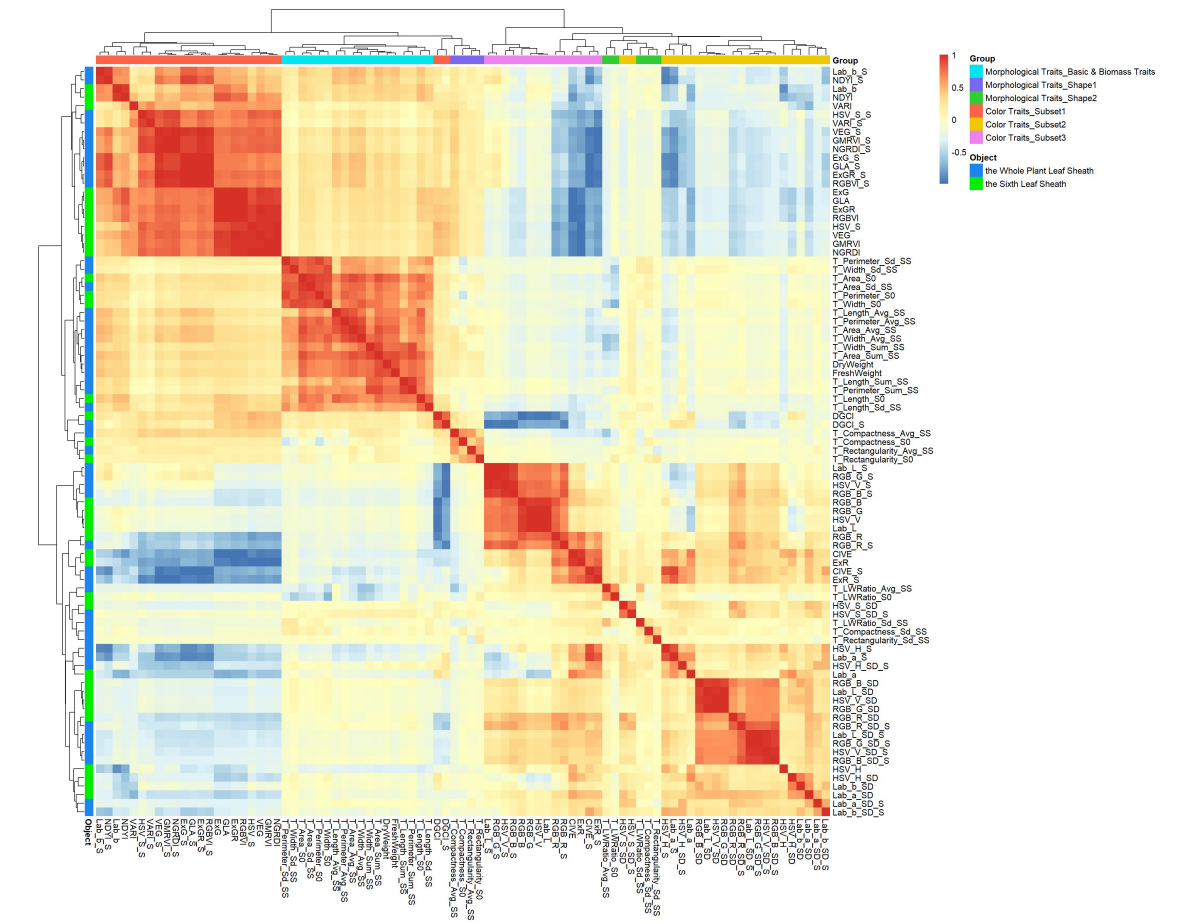


FIGURE 3 | Correlation analysis and clustering of 87 leaf sheath traits. 87 traits were obtained from two objects: the whole plant leaf sheath and the sixth leaf sheath. After clustering, 87 traits were divided into six groups and marked with different colors: Morphological Traits_Basic and Biomass Traits, Morphological Traits_Shape1, Morphological Traits_Shape2, Color Traits_Subset1, Color Traits_Subset2 and Color Traits_Subset3.

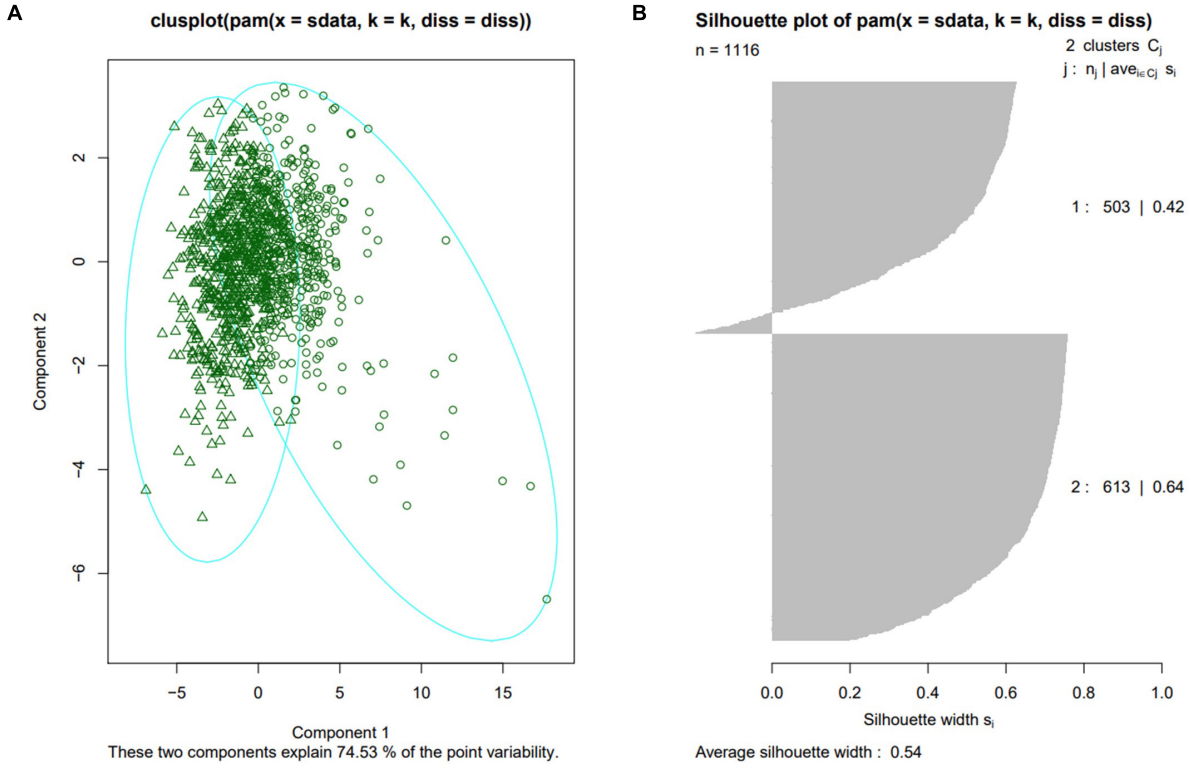


FIGURE 4 | Sample grouping results based on 12 morphological traits of the whole plant leaf sheath. **(A)** Clusplot of the first two principal components of 12 morphological traits of the whole plant leaf sheath. **(B)** Silhouette plot of samples characterized by 12 morphological traits of the whole plant leaf sheath.

However, due to the large number of color traits with heritability greater than 0.3, PCA was applied to the color traits of the whole plant and the sixth leaf sheath separately to accomplish the dimensionality reduction and key feature extraction. The results showed that for the color traits of these two objects, the first and second principal components (PCs) were strongly correlated with most color variables, and the cumulative contribution value of the first two principal components was 0.72 and 0.66, respectively (**Figures 5C,D**). Therefore, 4 traits that consisted of the first two PCs of the two objects color traits were selected for subsequent GWAS. Adding to the 21 non-color traits with heritability greater than 0.3, totally 25 key traits (2 biomass-related, 19 morphology-related and 4 color-related traits) with high heritability was used to explore the genetic mechanisms by GWAS.

Significant Single Nucleotide Polymorphism Obtained by Genome-Wide Association Study

In conclusion, the multi-locus random-SNP effect mixed linear model in R software package "mrMLM" (version 4.0) (Zhang et al., 2019) was used for GWAS analysis of biomass traits, morphological traits and color principal components related to 2D leaf sheaths, including 17 whole plant leaf sheath traits and 8 sixth leaf sheath traits. Finally, 1142 SNPs significantly related to 17 whole plant leaf sheath traits and 755 SNPs significantly related to 8 sixth leaf sheath traits were identified (P-value $<$ 6.4e-07) (**Table 1**). Additionally, among the results of the 6 GWAS methods, the most significant (Top1) SNP obtained by each method and the SNPs verified by two or more methods were considered to be highly reliable results. As a consequence, 152 SNPs significantly associated with 17 whole plant leaf sheath traits and 85 SNPs significantly associated with 8 sixth leaf sheath traits were obtained. These highly significant or multi-method verification results will be reported as the key findings of this study.

Identification and Annotation of Candidate Genes

Gene annotation was performed on 1142 SNPs significantly related to 17 whole plant leaf sheath traits and 755 SNPs significantly related to 8 sixth leaf sheath traits by using the latest maize B73 reference genome (B73 RefGen_v4) available in EnsemblPlants and NCBI Gene databases. Finally, 1,816 candidate genes of 17 whole plant leaf sheath traits and 1,297 candidate genes of 8 sixth leaf sheath traits were obtained, respectively. Among them, 275 genes of 17 whole plant leaf sheath traits and 146 genes of 8 sixth leaf sheath traits were derived from the most significant SNP (Top1) obtained by each method and SNPs annotations verified by multiple methods (**Table 1**). Genes annotated by SNPs with the highest significance or multi-method validation were

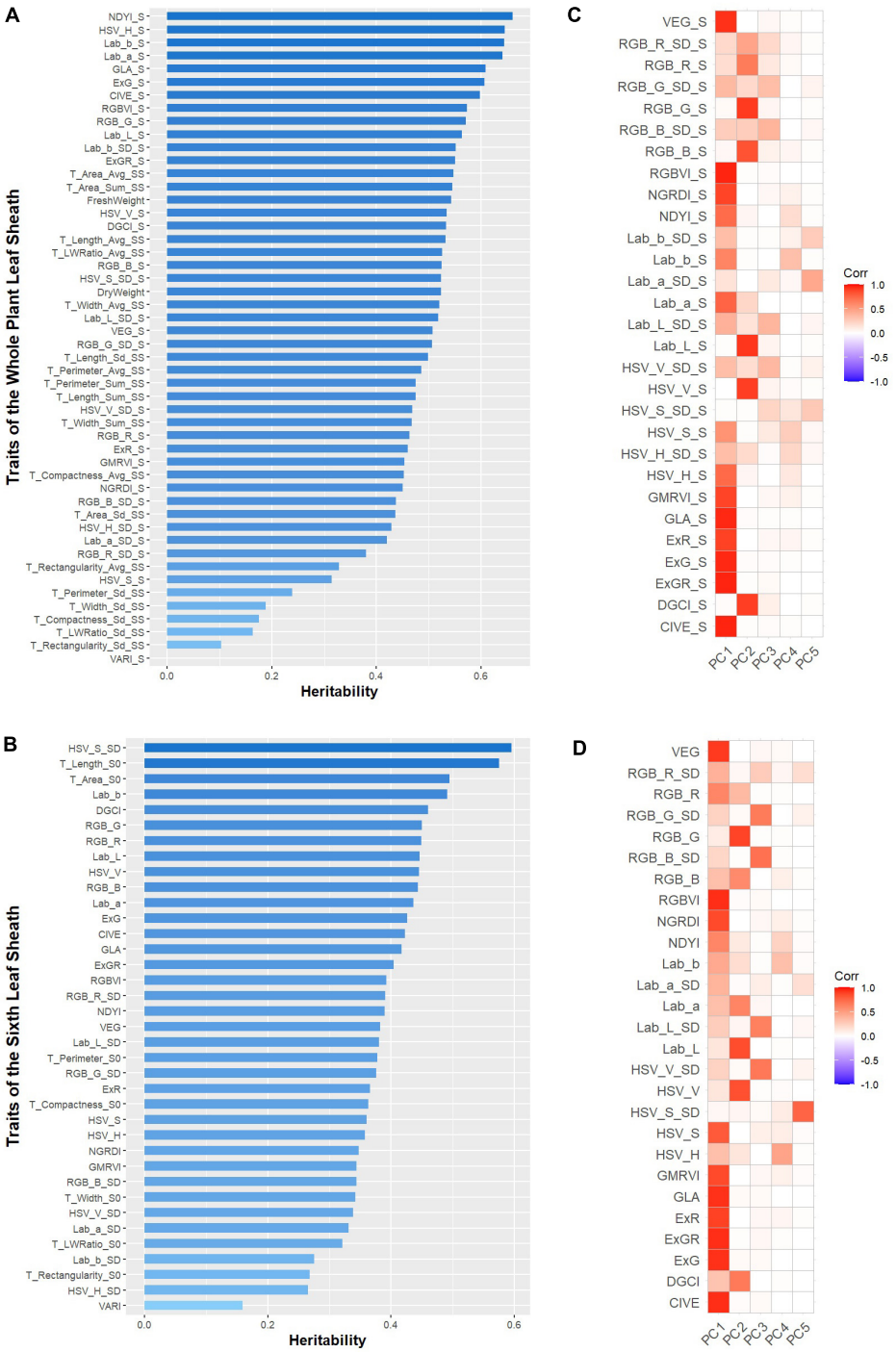


FIGURE 5 | The broad-sense heritability (H^2) of the investigated 87 phenotypic traits and principal component analysis (PCA) of color traits with heritability greater than 0.3. **(A)** The broad-sense heritability (H^2) of the 50 phenotypic traits of the whole plant leaf sheath. **(B)** The broad-sense heritability (H^2) of the 37 phenotypic traits of the sixth leaf sheath. **(C)** The first five principal components for color traits of the whole plant leaf sheath. **(D)** The first five principal components for color traits of the sixth leaf sheath.

further retrieved in NCBI Gene database, and 270 candidate genes of 25 key traits for leaf sheath phenotype had detailed functional descriptions (**Supplementary Table 4**). Among them, a total of 46 genes with clear functional descriptions were annotated by SNPs that both Top1 and multi-method validated (**Table 2**).

Hence, it was obvious that each leaf sheath-related trait had its own specific candidate gene, whether it was from the whole

TABLE 1 | Summary of significant loci from genome-wide association study.

Object	Category	Trait	No. of unique SNPs	No. of unique annotated genes	No. of genes only related to specific trait	No. of significant SNPs listed Top 1* and validated by multiple methods	No. of unique annotated genes listed Top1 and validated by multiple methods	No. of genes only related to specific trait listed Top 1 and validated by multiple methods
Whole plant leaf sheath	Biomass	DryWeight	35	62	26	6	12	12
		Freshweight	47	86	49	10	19	17
	Color	Sum_PC1	58	105	79	11	22	16
		Sum_PC2	75	132	101	9	18	18
	Morphology	T_Area_avg_SS	43	72	26	10	19	14
		T_Area_Sd_SS	46	77	25	10	16	8
		T_Area_sum_SS	41	71	22	12	24	14
		T_Compactness_Avg_SS	213	348	242	11	19	17
		T_Length_Avg_SS	53	94	50	9	18	14
		T_Length_Sd_SS	34	59	36	9	15	10
		T_Length_Sum_SS	50	87	50	12	21	17
		T_LWRatio_Avg_SS	38	74	54	9	18	18
		T_Perimeter_Avg_SS	35	65	30	8	14	7
		T_Perimeter_Sum_SS	48	91	46	13	26	16
		T_Rectangularity_Avg_SS	347	591	456	12	21	17
		T_Width_Avg_SS	31	62	27	5	10	6
		T_Width_Sum_SS	45	84	25	7	14	8
		Summary	**1,142**	**1,816**	**1,344**	**152**	**275**	**229**
Sixth leaf sheath	Color	Sixth_PC1	75	134	102	12	22	18
		Sixth_PC2	269	478	400	11	21	21
	Morphology	T_Area_S0	45	77	25	10	18	6
		T_Compactness_S0	72	122	85	13	21	19
		T_Length_S0	53	95	42	14	23	19
		T_LWRatio_S0	136	245	192	11	18	16
		T_Perimeter_S0	48	83	26	11	18	12
		T_Width_S0	84	154	97	9	16	12
		Summary	**755**	**1,297**	**969**	**85**	**146**	**123**

**Top1: the most significant SNP obtained by each GWAS method.*

plant or the sixth leaf alone. In addition to the common traits that could be extracted in previous studies, the 2D leaf sheath-related traits proposed in this study also identified significant loci and candidate genes. Consequently, it is necessary to subdivide and refine the phenotype of plants at maize seedling stage (**Table 1**). In addition, some of these traits had overlapped genes in the whole plant and the sixth leaf sheath (**Table 1**), indicating that these traits were genetically related to a certain degree. If the study on the sixth leaf sheath can be used instead of the whole plant study at V6 stage, it will greatly save the cost of phenotype acquisition.

Pathways Enriched by Functional Enrichment Analysis

In order to further explore the function of candidate genes, we used functional enrichment analysis to enrich the candidate genes annotated by the most significant SNP (Top1) and verified by multiple methods in the whole plant and the sixth leaf sheath, respectively. For the whole plant leaf sheath traits, a total of 81 GO terms and 1 KEGG pathways ($P < 0.05$) were obtained by enrichment of candidate genes for leaf sheath phenotype, among which 37 GO terms belonged to GO BP (biological process) (**Figures 6A–D**). In GO BP terms, the pathways with the highest significance were related to cellular component assembly and organization. For instance, "ribosome assembly" (GO:0042255, *P*-value = 2.70E-09), "organelle assembly" (GO:0070925, *P*-value = 3.50E-09), "ribonucleoprotein complex assembly" (GO:0022618, *P*-value = 6.90E-08), "cellular macromolecular complex assembly" (GO:0034622, *P*-value = 1.30E-05), "cellular component assembly" (GO:0022607, *P*-value = 5.80E-05) and "cellular component organization or biogenesis" (GO:0071840, *P*-value = 0.00016). Notably, several pathways related to cell proliferation and epidermal cell differentiation were identified by GO analysis: "regulation of cell proliferation" (GO:0042127, *P*-value = 0.00834), "cell proliferation" (GO:0008283,

TABLE 2 | Detailed functional descriptions of 46 genes annotated by both Top1 and multi-method validated SNPs.

Gene	Description	Chromosome	Genomic_ nucleotide_ accession.version	Start_position_ on_the_genomic_ accession	End_position_on _ the_genomic_ accession	Trait	Object
GRMZM2G073826	Transcription factor MYB3R-5	5	NC_050100.1	137,986,909	138,015,364	FreshWeight	Whole plant leaf sheath
GRMZM2G418206	Proteinaceous RNase P 1, chloroplastic/ mitochondrial	5	NC_050100.1	137,893,158	137,908,282	FreshWeight	Whole plant leaf sheath
GRMZM2G040452	Catalytic/protein phosphatase type 2C	4	NC_050099.1	237,957,682	237,960,493	Sum_PC1	Whole plant leaf sheath
GRMZM2G085945	Zinc finger protein	5	NC_050100.1	219,927,636	219,928,886	Sum_PC2	Whole plant leaf sheath
Zm00001d009690	RNA cytidine acetyltransferase 1	8	NC_050103.1	76,264,531	76,273,801	Sum_PC2	Whole plant leaf sheath
GRMZM2G103721	Phosphatidylinositol 3-kinase, root isoform	4	NC_050099.1	74,538,889	74,548,829	T_Area_Avg_SS	Whole plant leaf sheath
GRMZM2G134248	Long chain base biosynthesis protein 1a	4	NC_050099.1	74,702,902	74,704,363	T_Area_Avg_SS	Whole plant leaf sheath
GRMZM2G156238	C2 Domain-containing protein At1g53590	4	NC_050099.1	226,930,981	226,946,730	T_Area_Avg_SS	Whole plant leaf sheath
GRMZM2G126860	Protein SUPPRESSOR OF K(+) TRANSPORT GROWTH DEFECT 1	8	NC_050103.1	14,003,169	14,008,552	T_Area_Avg_SS	Whole plant leaf sheath
GRMZM2G126956	DNA damage-binding protein 2	8	NC_050103.1	14,023,075	14,027,906	T_Area_Avg_SS	Whole plant leaf sheath
GRMZM2G161169	Taxane 10-beta-hydroxylase	4	NC_050099.1	6,214,120	6,216,528	T_Area_Sum_SS	Whole plant leaf sheath
GRMZM2G065496	B3 Domain-containing protein	1	NC_050096.1	168,954,911	168,957,922	T_Compactness_ Avg_SS	Whole plant leaf sheath
zma-MIR169i	MicroRNA MIR169i	4	NC_050099.1	49,606,834	49,607,024	T_Compactness_ Avg_SS	Whole plant leaf sheath
FHA9	Myosin-9	1	NC_050096.1	5,773,058	5,778,341	T_Length_Avg_SS	Whole plant leaf sheath
GRMZM2G371137	Probable LRR receptor-like serine/threonine-protein kinase At1g12460	1	NC_050096.1	5,836,567	5,841,667	T_Length_Avg_SS	Whole plant leaf sheath
GRMZM2G047715	Homeobox-leucine zipper protein HOX7	4	NC_050099.1	126,441,242	126,446,006	T_Length_Avg_SS	Whole plant leaf sheath
GRMZM2G105933	Putative protein kinase superfamily protein	4	NC_050099.1	126,356,761	126,359,205	T_Length_Avg_SS	Whole plant leaf sheath
GRMZM2G097605	DNA repair helicase UVH6	10	NC_050105.1	91,197,531	91,202,542	T_Length_Sd_SS	Whole plant leaf sheath
GRMZM2G135770	Putative regulator of chromosome condensation (RCC1) family protein	4	NC_050099.1	84,989,713	84,995,057	T_Length_Sum_SS	Whole plant leaf sheath
GRMZM2G419305	Agenet domain-containing protein/bromo-adjacent homology (BAH) domain-containing protein	4	NC_050099.1	85,143,298	85,148,546	T_Length_Sum_SS	Whole plant leaf sheath
GRMZM2G030839	Phosphomevalonate kinase	9	NC_050104.1	148,330,626	148,338,890	T_LWRatio_ Avg_SS	Whole plant leaf sheath
GRMZM2G094592	IRK-interacting protein	7	NC_050102.1	138,421,602	138,423,876	T_Perimeter_ Avg_SS	Whole plant leaf sheath
GRMZM2G143160	Serine/threonine-protein kinase MPS1	1	NC_050096.1	271,894,088	271,899,733	T_Perimeter_ Sum_SS	Whole plant leaf sheath
GRMZM2G147332	Oxysterol-binding protein-related protein 1C	1	NC_050096.1	271,997,558	272,015,271	T_Perimeter_ Sum_SS	Whole plant leaf sheath
GRMZM2G319357	Low molecular weight protein-tyrosine-phosphatase slr0328	1	NC_050096.1	209,876,521	209,881,331	T_Perimeter_ Sum_SS	Whole plant leaf sheath

(Continued)

TABLE 2 | (Continued)

Gene	Description	Chromosome	Genomic_ nucleotide_ accession.version	Start_position_ on_the_genomic_ accession	End_position_on _ the_genomic_ accession	Trait	Object
GRMZM2G022926	OSJNBa0070C17.17-like protein	10	NC_050105.1	144,286,874	144,289,145	T_Perimeter_ Sum_SS	Whole plant leaf sheath
GRMZM2G028676	Vacuolar ATPase assembly integral membrane protein VMA21-like domain	10	NC_050105.1	144,223,404	144,224,474	T_Perimeter_ Sum_SS	Whole plant leaf sheath
GRMZM2G128248	dnaJ protein	8	NC_050103.1	172,072,905	172,075,250	T_Rectangularity_ Avg_SS	Whole plant leaf sheath
GRMZM2G123537	Pumilio homolog 3	4	NC_050099.1	176,122,607	176,128,541	T_Width_Sum_SS	Whole plant leaf sheath
zma-MIR172c	MicroRNA MIR172c	4	NC_050099.1	176,265,726	176,265,848	T_Width_Sum_SS	Whole plant leaf sheath
GRMZM2G309025	S-domain class receptor-like kinase 3	7	NC_050102.1	165,446,107	165,448,937	T_Width_Sum_SS	Whole plant leaf sheath
GRMZM2G339645	CSLF3—cellulose synthase-like family F	7	NC_050102.1	165,345,171	165,348,432	T_Width_Sum_SS	Whole plant leaf sheath
GRMZM2G066997	Remorin	5	NC_050100.1	193,077,446	193,080,950	T_Width_Sum_SS, T_Area_S0, T_LWRatio_S0	Whole plant and Sixth leaf sheath
GRMZM2G477314	CF9	1	NC_050096.1	288,099,406	288,101,023	Sixth_PC1	Sixth leaf sheath
GRMZM2G091303	Xyloglucan endotransglucosylase/ hydrolase protein 24	10	NC_050105.1	143,472,231	143,474,095	Sixth_PC1	Sixth leaf sheath
CKX10	Cytokinin dehydrogenase 10	1	NC_050096.1	21,236,2957	212,366,306	Sixth_PC2	Sixth leaf sheath
GRMZM2G122126	6-Phosphogluconolactonase	1	NC_050096.1	212,272,609	212,274,483	Sixth_PC2	Sixth leaf sheath
GRMZM2G138355	Nudix hydrolase 13	10	NC_050105.1	114,632,712	114,636,142	Sixth_PC2	Sixth leaf sheath
Zm00001d027570	Putative protein phosphatase 2C 48	1	NC_050096.1	8,216,152	8,220,401	T_Area_S0	Sixth leaf sheath
GRMZM6G207008	Characterized LOC100272314	4	NC_050099.1	172,883,765	172,884,420	T_Compactness_ S0	Sixth leaf sheath
Zm00001d051817	DNA topoisomerase 2	4	NC_050099.1	172,603,028	172,604,370	T_Compactness_ S0	Sixth leaf sheath
GRMZM2G339907	NDR1/HIN1-like protein 26	7	NC_050102.1	163,429,440	163,430,393	T_LWRatio_S0	Sixth leaf sheath
GRMZM2G039811	Transmembrane 9 superfamily member 9	2	NC_050097.1	204,934,314	204,938,233	T_Perimeter_S0	Sixth leaf sheath
GRMZM2G153369	Hydrophobic protein RCI2B	2	NC_050097.1	205,006,561	205,007,668	T_Perimeter_S0	Sixth leaf sheath
GRMZM2G007122	Putative ubiquitin-conjugating enzyme family	6	NC_050101.1	175,737,238	175,740,387	T_Perimeter_S0	Sixth leaf sheath
GRMZM2G155686	Gibberellin 2-oxidase8	6	NC_050101.1	175,763,619	175,765,545	T_Perimeter_S0	Sixth leaf sheath

P-value = 0.02853), "root epidermal cell differentiation" (GO:0010053, *P*-value = 0.02308), "plant epidermal cell differentiation" (GO:0090627, *P*-value = 0.02853) and "plant epidermis development" (GO:0090558, *P*-value = 0.02884). In addition, the one KEGG pathway was "Sphingolipid metabolism" (zma00600, *P*-value = 0.02218).

For the sixth leaf sheath traits, a total of 57 GO terms and 4 KEGG pathways (*P*-value < 0.05) were enriched in the sixth leaf sheath phenotype candidate genes, among which 31 GO terms belonged to GO BP (**Figures 6E–H**). In GO BP terms, several pathways related to response to hunger, nutrition and extracellular stimulation were enriched by genes *GRMZM2G147450* and *GRMZM2G059121*: "cellular response to phosphate starvation" (GO:0016036, *P*-value = 0.00245), "cellular response to starvation" (GO:0009267, *P*-value = 0.00643), "disaccharide metabolic process" (GO:0005984, *P*-value = 0.00779), "response to starvation" (GO:0042594, *P*-value = 0.00779), "cellular

response to nutrient levels" (GO:0031669, *P*-value = 0.00842), "response to nutrient levels" (GO:0031667, *P*-value = 0.01184), "cellular response to extracellular stimulus" (GO:0031668, *P*-value = 0.01184) and "cellular response to external stimulus" (GO:0071496, *P*-value = 0.01284). In addition, candidate genes for the sixth leaf sheath traits were also enriched in multiple pathways related to cell proliferation and epidermis development. For example, "plant epidermis morphogenesis" (GO:0090626, *P*-value = 0.00519), "cell proliferation" (GO:0008283, *P*-value = 0.00606) and "plant epidermis development" (GO:0090558, *P*-value = 0.03493). The most striking result of KEGG is "Alanine, aspartate and glutamate metabolism" (zma00250, *P*-value = 0.01283). And the other three pathways are "Pyrimidine metabolism" (zma00240, *P*-value = 0.01379), "Metabolic pathways" (zma01100, *P*-value = 0.01382) and "Phagosome" (zma04145, *P*-value = 0.03899).

Trait-Candidate Gene-Pathway Network Visualization

Cytoscape V3.7.2 was used to draw the trait-candidate gene-pathway network of 2D maize leaf sheath traits at seedling stage, and to show the relationship between 270 candidate genes and 25 key traits, and between candidate genes and their enriched pathways. The whole network consisted of 444 nodes and 1,144 edges (**Figure 7**). In the network, there were 25 traits (the largest nodes), including 17 whole plant leaf sheath traits (round rectangle nodes) and 8 sixth leaf sheath traits (octagon nodes). And the types of traits—morphology, color and biomass—were also marked in blue, orange and green, respectively. In addition, the candidate genes were marked with small gray circular nodes, and the pathways were marked with small diamond. Among them, pathways related to cellular component assembly and organization were marked in earthy yellow, pathways related to cell proliferation and epidermal cell differentiation were marked in grass green, and pathways related to response to hunger, nutrition and extracellular stimulation were marked in red.

DISCUSSION

Maize leaf sheaths wrap stem to provide structural support and protect developing leaves, which is of great biological significance. This study broke the traditional method of phenotypic acquisition of maize leaf sheath, and proposed an image-based high-throughput acquisition and data analysis scheme for phenotypic traits of maize leaf sheath from image acquisition, image phenotypic analysis and leaf sheath phenotypic data analysis. Firstly, a simple and reliable environment for maize leaf sheath image acquisition was established, and the acquisition time of a single sample image was less than 10s. Then, a maize leaf sheath phenotypic image analysis software with friendly interactive interface was developed based on open-source software development tools. Based on the image analysis, 85 leaf sheath phenotypic traits including shape and color can be analyzed, and the calculation time for a single image was less than 60s. Finally, phenotypic traits were extracted and analyzed from leaf sheath images of 418 maize inbred lines, and the statistical description results of leaf sheath phenotypic traits of large maize populations were obtained. It is time-consuming and laborious to obtain the traditional traits such as length and width of leaf sheath manually, but the image-based phenotypic acquisition method can quickly obtain the length and width of leaf sheath in less than 1 min. Besides, more than 80 phenotypic traits can also be extracted. Thus, efficient and high-throughput acquisition of leaf sheath phenotypes was achieved. Moreover, this method is suitable for large populations and can help to obtain leaf sheath phenotype in maize association analysis population.

A large number of traits can be extracted from plant images, and a variety of new traits can be determined from different dimensions. However, the interpretability of the traits still needs further study. In this study, correlation analysis, cluster analysis and PCA were performed on 87 leaf sheath-related phenotypic traits of maize association analysis population. The results showed that there were differences in morphological characteristics and color traits of leaf sheath, with correlation coefficients less than 0.5. In the morphological characteristics of leaf sheath, it can be divided into three groups with definite significance due to the different features described. Color traits can be subdivided into three subsets with distinctive features. Therefore, although some traits cannot explain their biological significance by themselves, combined with trait grouping and its highly correlated traits, the phenotypic traits with less clear meanings can be characterized.

In order to verify the reliability of phenotypic acquisition from leaf sheath images, correlation analysis was conducted between dry and fresh weight of maize leaf sheath measured manually and leaf sheath morphological traits obtained from images. The results showed that 16 morphological characteristics of leaf sheath had a significant positive correlation with DryWeight and FreshWeight (p-value < 0.05), and clustered into the same group (Morphological Traits_Basic and Biomass Traits). This result was consistent with the prior knowledge, revealing the reliability of the data, and demonstrating that the various traits obtained from the image were meaningful. Moreover, among the color traits extracted from the image, 24 comprehensive color traits were divided into two groups, CIVE, CIVE_S, ExR and ExR_S mainly represent red, while the remaining comprehensive traits mainly represent blue and green. The clustering results based on the phenotypic data were consistent with the trait characteristics, which also showed the accuracy of the data extraction.

It can be seen from the results of this study that image-based high-throughput phenotypic acquisition techniques can obtain novel traits that breeders cannot evaluate through traditional methods, such as geometric and color traits described quantitatively. In this study, 88.51% (77/87) of leaf sheath-related phenotypic traits had heritability greater than 0.3, indicating that the formation of these phenotypes was influenced by genetic factors. To further dissect the genetic mechanisms underlying these phenotypes with heritability greater than 0.3, GWAS was used to analyze the 25 key leaf sheath-related traits, and totally 3,113 candidate genes for leaf sheath-related traits were obtained. The candidate genes with high significance or verified by multiple methods were considered as high reliability results, which

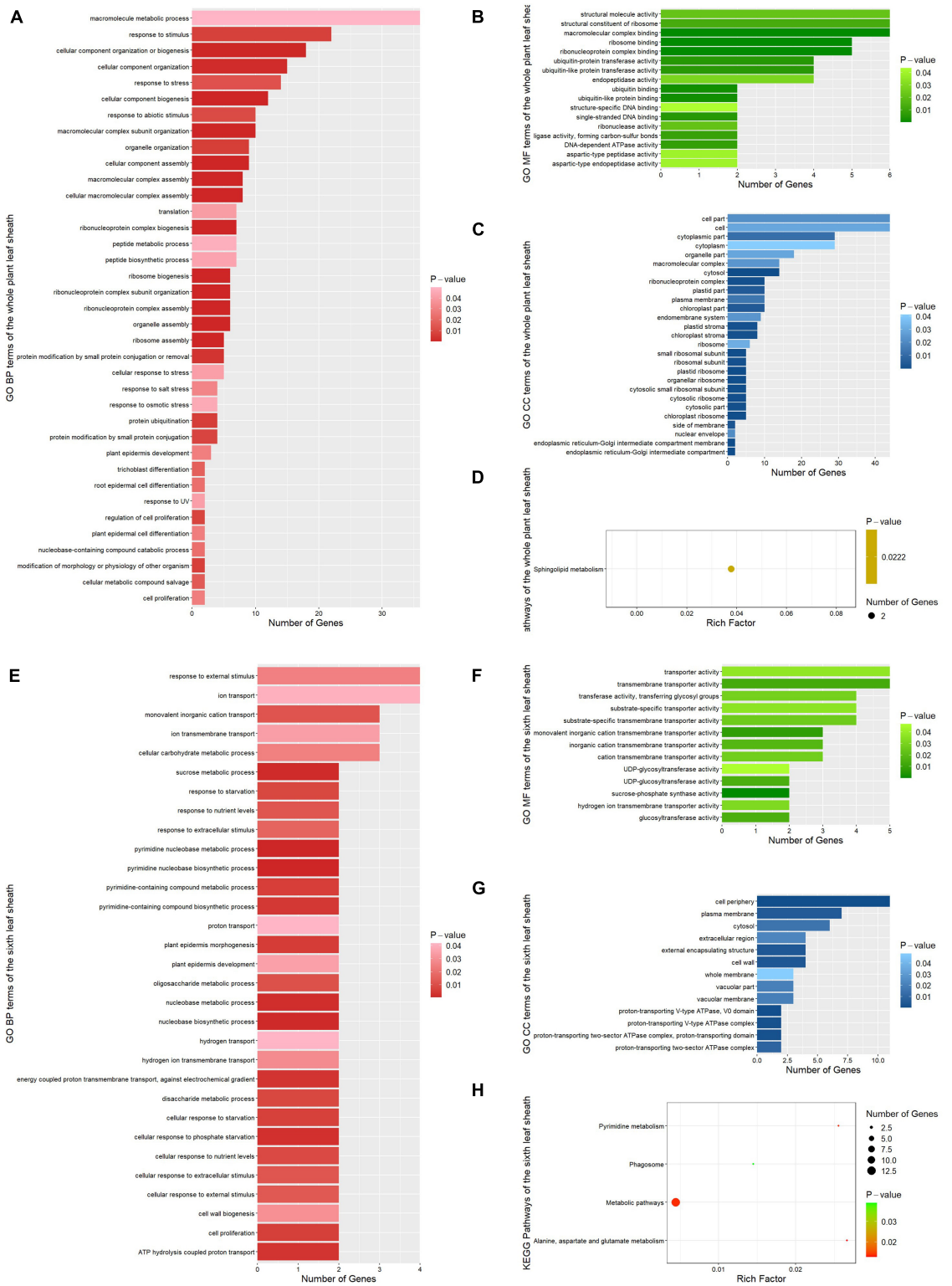


FIGURE 6 | Functional enrichment results of all candidate genes associated with phenotypic traits. **(A)** GO BP (biological process) terms enriched by the whole plant leaf sheath candidate genes. **(B)** GO MF (molecular function) terms enriched by the whole plant leaf sheath candidate genes. **(C)** GO CC (cellular components) terms enriched by the whole plant leaf sheath candidate genes. **(D)** KEGG pathways enriched by the whole plant leaf sheath candidate genes. **(E)** GO BP terms enriched by the sixth leaf sheath candidate genes. **(F)** GO MF terms enriched by the sixth leaf sheath candidate genes. **(G)** GO CC terms enriched by the sixth leaf sheath candidate genes. **(H)** KEGG pathways enriched by the sixth leaf sheath candidate genes.

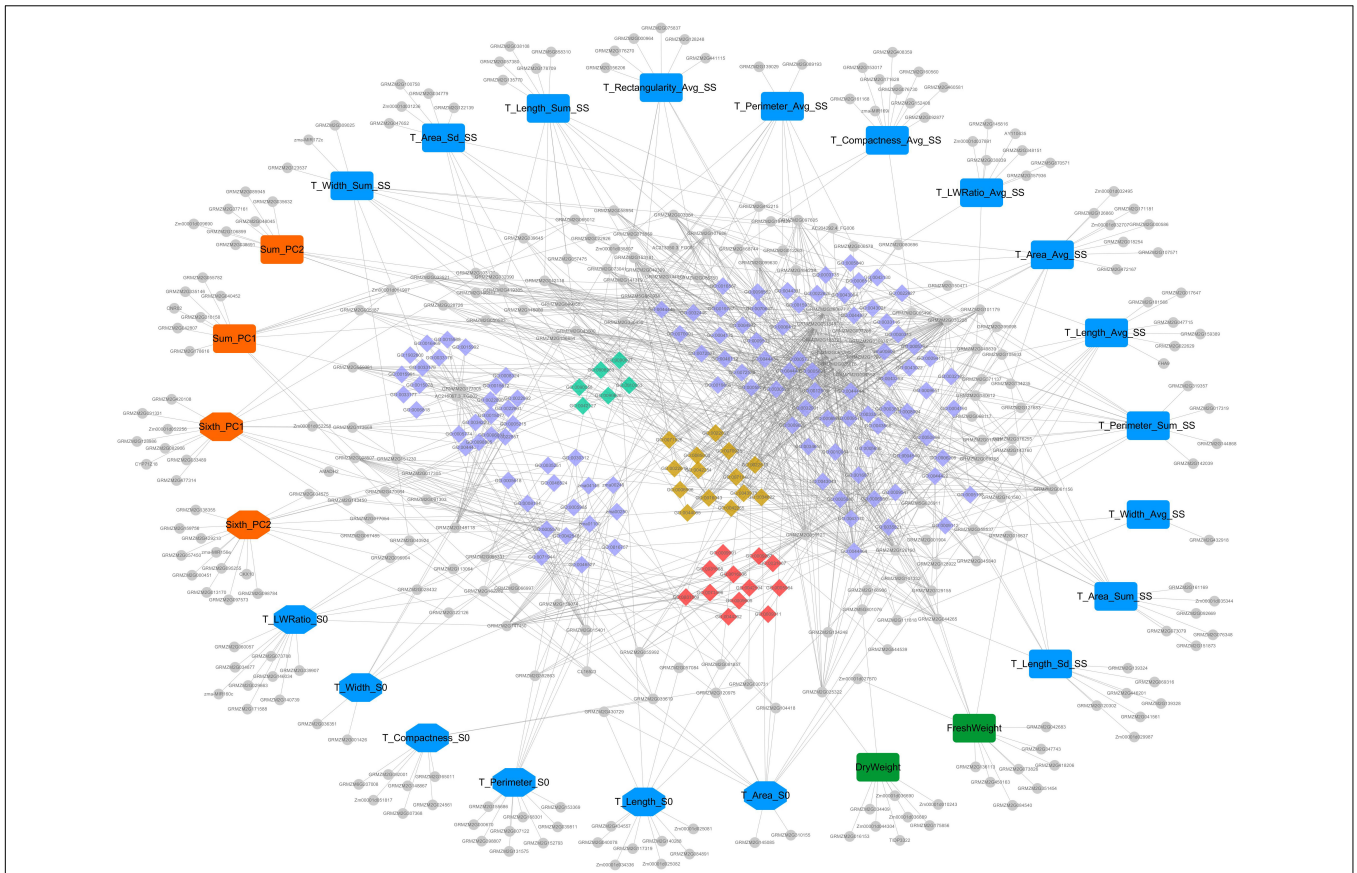


FIGURE 7 | The "trait-gene-pathway" network constructed by 25 key traits and their candidate genes and pathways. Traits, genes and pathways (GO terms and KEGG pathways) are shown in different shapes and sizes. Of the 25 large nodes, 17 round rectangle nodes represent the whole plant leaf sheath traits, and 8 octagon nodes represent the sixth leaf sheath traits. And different color represents different type of traits (blue- morphology, orange- color and green- biomass). The colorful small diamonds represent GO terms and KEGG pathways enriched by candidate genes. Among them, pathways related to cellular component assembly and organization were marked in earthy yellow, pathways related to cell proliferation and epidermal cell differentiation were marked in grass green, and pathways related to response to hunger, nutrition and extracellular stimulation were marked in red. Candidate genes are represented by the small gray circular nodes.

would provide reference for subsequent functional verification of maize leaf sheath candidate genes. For example, cytokinin dehydrogenase 10 (*CKX10*) is a candidate gene for major component traits of the color of the sixth leaf sheath (Sixth_PC2). Meanwhile, it has been reported that *CKX10* plays an important role in dry matter accumulation in V6 stage leaves (Lu et al., 2020). *CKX10* is a member of the CKX family, and a great deal of work has been done on this gene family in gramineae (Mameaux et al., 2012), including some studies on maize. In transcriptome analysis of maize, *CKX10* has also been reported as one of the DEGs of KEGG pathways associated with hormone metabolism (Zheng et al., 2020). Therefore, we speculate that *CKX10* plays an important role in the formation of leaf sheath color in maize V6 stage. It is worth noting that some loci of these high confidence results had a high explanatory power (PVE > 5%) for phenotypic variation. For example, *GRMZM2G135770*, putative regulator of chromosome condensation (RCC1) family protein, was annotated by chr4.S_84970911 on chromosome 4, which was significantly associated with the trait T_Length_Sum_SS, and explained 6.54% of the phenotypic variance. *GRMZM2G156238*, C2 domain-containing protein At1g53590, which has been proved to be tissue-specific (Stelpflug et al., 2016). It was reported in the study of organ-specific and stress-induced gene expression mapping of maize (Hoopes et al., 2019). In this study, it was annotated by chr4.S_224037650, also located on chromosome 4, which was significantly associated with the trait T_Area_Avg_SS, explaining 5.17% of the phenotypic variance. And *GRMZM2G156238* was also associated with the other two leaf sheath morphological traits (T_Area_S0 and T_Perimeter_Avg_SS). The above results proved the reliability of the phenotype-genotype association analysis process and results of this study. At the same time, it also reflects the significance of trait refinement for the research of crop phenotypic genetics, that is, the more refined the trait, the stronger the phenotype interpretability of the obtained locus.

Pigment plays an important role in plant reproduction and adaptability, and the research on plant pigment has always been a hot topic. In this study, phenotypic traits of leaf sheath color of maize inbred lines from four subpopulations with different environmental adaptability were analyzed. The results showed

that there were significant differences in 48 leaf sheath color traits between tropical and subtropical maize inbred lines (TST) and maize inbred lines from other climatic zones (P-value < 0.05), which showed that the color of maize leaf sheath was closely related to the ecological adaptability and evolution of maize. In addition, the changes of pigment deposition, distribution and shade among different kinds of maize are of great value to the study of maize functional genome and the application of maize genetics and breeding. Leaf sheath color is also an important morphological marker to guide maize breeding. It can be used for more intuitive selection and more directly genetic research of related special traits. In this study, a total of 60 leaf sheath color traits were extracted based on images, including 30 for the whole plant leaf sheath and 30 for the sixth leaf sheath. In addition to simple single-channel color traits, a number of novel comprehensive color traits were also extracted. The results of heritability analysis showed that the heritability of color trait was generally high, so it was necessary to conduct GWAS analysis to explore the genetic factors behind these traits. In our study, PCA was used to reduce the dimensionality of the color traits with heritability greater than 0.3, and then the first two principal components were selected for GWAS. As a consequence, more than 800 candidate genes related to color traits were identified (**Table 1**). These results greatly enrich the existing research results on maize leaf sheath genetics and provide a theoretical basis for better explaining the mechanism of maize leaf sheath phenotype formation.

In recent years, phenomics has emerged as a rapidly growing data-intensive discipline. The rapid development of phenomics-related technologies and research tools has brought about a huge amount of phenotypic information at multiple scales and data diversity, such as RGB, hyperspectral, near-infrared, thermal and fluorescence imaging and other image data, as well as data on various physiological traits during plant growth (Kim et al., 2017). Crop life activity is a dynamic process under the combined action of genes and environment. As high-throughput sequencing technologies continue to develop and improve, single-omics studies are becoming increasingly sophisticated. And the integration of multi-omics data to study crops is on the rise. Genomic studies combining genomic and phenotypic data have been conducted in many crops and have rapidly decoded the functions of a large number of unknown genes. In 2014, 13 traditional agronomic traits of rice were combined with two newly defined traits and 141 related loci were identified using GWAS (Yang W. et al., 2014). In 2015, 29 leaf phenotypic traits at three key fertility stages were resolved using high-throughput leaf phenotype acquisition (HLS) and subjected to GWAS analysis, and 73 loci regulating leaf size, 123 loci regulating leaf color and 177 new loci regulating leaf shape (Yang et al., 2015). In 2021, 48 maize stem micro-phenotypic traits were automatically extracted by micro-CT image processing pipeline and 1,562 significant SNPs were identified for 30 key traits by GWAS (Zhang et al., 2021). It is clear that combining high-throughput phenotyping techniques with large-scale QTL or GWAS analysis not only greatly expands our understanding of the dynamic developmental processes in crops, but also provides a new tool for plant genomics, gene characterization and breeding research.

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

XG and WS conceived and supervised the project and agreed to serve as the author responsible for contact and ensure communication. XL, YZ, YaZ, and WW conducted the experiment and collected the data. JW and CW analyzed data, prepared figures, wrote the manuscript, and drafted and revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was financially supported by the Construction of Collaborative Innovation Center of Beijing Academy of Agricultural and Forestry Sciences (KJCX201917), the National Natural Science Foundation of China (31871519), the Science and Technology Innovation Team of Maize Modern Seed Industry in Hebei (21326319D), and the China Agriculture Research System of MOF and MARA.

ACKNOWLEDGMENTS

We thank the Maize Research Center Department of the Beijing Academy of Agriculture and Forestry Sciences for preparing the seed for the trial. We also thank Prof. Jianbing Yan, from Huazhong Agricultural University, Xiaohong Yang, from the National Maize Improvement Center of China, China Agricultural University, for providing seeds of the maize inbred lines.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.826875/full#supplementary-material

Supplementary Figure 1 | The trait variation between subpopulations (TST, NSS, SS, and mixed) for the 87 phenotypic traits (divided into three types).

Supplementary Figure 2 | Sample grouping results based on **(A)** 21 leaf sheath color traits of whole plant, **(B)** 25 leaf sheath color traits of the sixth leaf and **(C)** 4 leaf sheath morphological traits, respectively.

Supplementary Table 3 | Descriptive statistical analysis results of 87 maize leaf sheath phenotypic traits.

Supplementary Table 4 | 270 candidate genes of 25 key traits for leaf sheath phenotype with detailed functional descriptions from NCBI Gene database.

REFERENCES

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

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

Bu, D., Luo, H., Huo, P., Wang, Z., Zhang, S., He, Z., et al. (2021). KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic Acids Res.* 49, W317–W325. doi: 10.1093/nar/gkab447

Bucksch, A., Burridge, J., York, L. M., Das, A., Nord, E., Weitz, J. S., et al. (2014). Image-based high-throughput field phenotyping of crop roots. *Plant Physiol.* 166, 470–486. doi: 10.1104/pp.114.243519

Butler, D. (2009). *asreml: asreml() fits the linear mixed model. R package version 3.0*. Available online at: www.vsni.co.uk (accessed October, 2021).

Chatham, L. A., and Juvik, J. A. (2021). Linking anthocyanin diversity, hue, and genetics in purple corn. *G3* 11:jkaa062. doi: 10.1093/g3journal/jkaa062

Chopin, J., Kumar, P., and Miklavcic, S. J. (2018). Land-based crop phenotyping by image analysis: consistent canopy characterization from inconsistent field illumination. *Plant Methods* 14:39. doi: 10.1186/s13007-018-0308-5

Cooper, J. S., Rice, B. R., Shenstone, E. M., Lipka, A. E., and Jamann, T. M. (2019). Genome-Wide Analysis and Prediction of Resistance to Goss's Wilt in Maize. *Plant Genome.* 12:180045. doi: 10.3835/plantgenome2018.06.0045

Cui, Z., Luo, J., Qi, C., Ruan, Y., Li, J., Zhang, A., et al. (2016). Genome-wide association study (GWAS) reveals the genetic architecture of four husk traits in maize. *Bmc Genom.* 17:946. doi: 10.1186/s12864-016-3229-6

Dai, L., Wu, L., Dong, Q., Yan, G., Qu, J., and Wang, P. (2018). Genome-wide association analysis of maize kernel length. *J. Northwest A F Univ.* 46, 20–28.

Dong, L., Qin, L., Dai, X., Ding, Z., Bi, R., Liu, P., et al. (2019). Transcriptomic Analysis of Leaf Sheath Maturation in Maize. *Int. J. Mol. Sci.* 20:2472. doi: 10.3390/ijms20102472

Du, Y., Yang, S., Zhu, L., Zhao, Y., Huang, Y., Chen, J., et al. (2018). Genome-wide Association Analysis of Tassel Branch Number under High and Low Plant Densities in Maize (*Zea mays* L.). *Mol. Plant Breed.* 16, 5970–5977.

Fan, F., Fan, Y., Du, J., and Zhuang, J. (2008). Fine Mapping of C (Chromogen for Anthocyanin) Gene in Rice. *Rice Sci.* 15, 1–6. doi: 10.1016/s1672-6308(08)60012-8

Gage, J. L., Miller, N. D., Spalding, E. P., Kaeppler, S. M., and de Leon, N. (2017). TIPS: a system for automated image-based phenotyping of maize tassels. *Plant Methods* 13:21. doi: 10.1186/s13007-017-0172-8

Green, J. M., Appel, H., Rehrig, E. M., Harnsomburana, J., Chang, J. F., Balint-Kurti, P., et al. (2012). PhenoPhyte: a flexible affordable method to quantify 2D phenotypes from imagery. *Plant Methods* 8:45. doi: 10.1186/1746-4811-8-45

Guo, J., Liu, W., Zheng, Y., Liu, H., Zhao, Y., Zhu, L., et al. (2019). Genome-wide Association Analysis of Maize (*Zea mays*) Grain Quality Related Traits Based on Four Test Cross Populations. *J. Agric. Biotechnol.* 27, 809–824.

Hoopes, G., Hamilton, J., Wood, J., Esteban, E., Pasha, A., Vaillancourt, B., et al. (2019). An updated gene atlas for maize reveals organ-specific and stress-induced genes. *Plant J.* 97, 1154–1167. doi: 10.1111/tpj.14184

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

Jin, J., He, K., Tang, X., Li, Z., Lv, L., Zhao, Y., et al. (2015). An Arabidopsis Transcriptional Regulatory Map Reveals Distinct Functional and Evolutionary Features of Novel Transcription Factors. *Mol. Biol. Evol.* 32, 1767–1773. doi: 10.1093/molbev/msv058

Kanehisa, M. (2002). The KEGG database. *Novartis Found Symp.* 247, 91–101..

Kim, S. L., Solehati, N., Choi, I. C., Kim, K. H., and Kwon, T. R. (2017). Data management for plant phenomics. *J. Plant Biol. Volum.* 60, 285–297. doi: 10.1007/s12374-017-0027-x

Li, B., Varkani, K. N., Sun, L., Zhou, B., Wang, X., Guo, L., et al. (2020). Protective role of maize purple plant pigment against oxidative stress in fluorosis rat brain. *Transl. Neurosci.* 11, 89–95. doi: 10.1515/tnsci-2020-0055

Li, K., Wang, H., Hu, X., Liu, Z., Wu, Y., and Huang, C. (2016). Genome-Wide Association Study Reveals the Genetic Basis of Stalk Cell Wall Components in Maize. *PLoS One* 11:e0158906. doi: 10.1371/journal.pone.0158906

Li, K., Zhang, X., Guan, Z., Shen, Y., and Pan, G. (2017). Genome-wide Association Analysis of Plant Height and Ear Height in Maize. *J. Maize Sci.* 25, 1–7.

Li, P., Du, C., Zhang, Y., Yin, S., Zhang, E., Fang, H., et al. (2018). Combined bulked segregant sequencing and traditional linkage analysis for identification of candidate gene for purple leaf sheath in maize. *PLoS One* 13:e0190670. doi: 10.1371/journal.pone.0190670

Liu, H., and Yan, J. (2019). Crop genome-wide association study: a harvest of biological relevance. *Plant J.* 97, 8–18. doi: 10.1111/tpj.14139

Liu, N., Xue, Y., Guo, Z., Li, W., and Tang, J. (2016). Genome-Wide Association Study Identifies Candidate Genes for Starch Content Regulation in Maize Kernels. *Front. Plant Sci.* 7:1046. doi: 10.3389/fpls.2016.01046

Lu, X., Wang, J., Wang, Y., Wen, W., Zhang, Y., Du, J., et al. (2020). Genome-Wide Association Study of Maize Aboveground Dry Matter Accumulation at Seedling Stage. *Front. Genet.* 11:571236. doi: 10.3389/fgene.2020.571236

Mameaux, S., Cockram, J., Thiel, T., Steuernagel, B., Stein, N., Taudien, S., et al. (2012). Molecular, phylogenetic and comparative genomic analysis of the cytokinin oxidase/dehydrogenase gene family in the Poaceae. *Plant Biotechnol. J.* 10, 67–82. doi: 10.1111/j.1467-7652.2011.00645.x

Mazaheri, M., Heckwolf, M., Vaillancourt, B., Gage, J. L., Burdo, B., Heckwolf, S., et al. (2019). Genome-wide association analysis of stalk biomass and anatomical traits in maize. *Bmc Plant Biol.* 19:45. doi: 10.1186/s12870-019-1653-x

Mir, R. R., Reynolds, M., Pinto, F., Khan, M. A., and Bhat, M. A. (2019). High-throughput phenotyping for crop improvement in the genomics era. *Plant Sci.* 282, 60–72. doi: 10.1016/j.plantsci.2019.01.007

Owens, B. F., Mathew, D., Diepenbrock, C. H., Tiede, T., Wu, D., Mateos-Hernandez, M., et al. (2019). Genome-Wide Association Study and Pathway-Level Analysis of Kernel Color in Maize. *G3* 9, 1945–1955. doi: 10.1534/g3.119.400040

Peniche-Pavia, H. A., and Tiessen, A. (2020). Anthocyanin Profiling of Maize Grains Using DIESI-MSQD Reveals That Cyanidin-Based Derivatives Predominate in Purple Corn, whereas Pelargonidin-Based Molecules Occur in Red-Pink Varieties from Mexico. *J. Agric. Food Chem.* 68, 5980–5994. doi: 10.1021/acs.jafc.9b06336

Russell, S. H., and Evert, R. F. (1985). Leaf vasculature in *Zea mays* L. *Planta* 164, 448–458. doi: 10.1007/BF00395960

Sanchez, D. L., Liu, S., Ibrahim, R., Blanco, M., and Luebberstedt, T. (2018). Genome-wide association studies of doubled haploid exotic introgression lines for root system architecture traits in maize (*Zea mays* L.). *Plant Sci.* 268, 30–38. doi: 10.1016/j.plantsci.2017.12.004

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science* 326, 1112–1115. doi: 10.1126/science.1178534

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

Shi, T., Zhang, M., Wu, N., and Wang, P. (2018). Genome-wide Association Study of Drought Resistance at Maize Seeding Stage. *J. Maize Sci.* 26, 44–48.

Song, P., Wang, J., Guo, X., Yang, W., and Zhao, C. (2021). High-throughput phenotyping: Breaking through the bottleneck in future crop breeding. *Crop J.* 9, 633–645. doi: 10.1016/j.cj.2021.03.015

Stelpflug, S., Sekhon, R., Vaillancourt, B., Hirsch, C., Buell, C., deLeon, N., et al. (2016). An expanded maize gene expression atlas based on RNA sequencing and its use to explore root development. *Plant Genome 9*. doi: 10.3835/plantgenome2015.04.0025

Tanabata, T., Shibaya, T., Hori, K., Ebana, K., and Yano, M. (2012). SmartGrain: high-throughput phenotyping software for measuring seed shape through image analysis. *Plant Physiol.* 160, 1871–1880. doi: 10.1104/pp.112.205120

Wallace, J. G., Zhang, X., Beyene, Y., Semagn, K., Olsen, M., Prasanna, B. M., et al. (2016). Genome-wide Association for Plant Height and Flowering Time across 15 Tropical Maize Populations under Managed Drought Stress and Well-Watered Conditions in Sub-Saharan Africa. *Crop Sci.* 56, 2365–2378. doi: 10.2135/cropsci2015.10.0632

Wang, N., Liu, B., Liang, X., Zhou, Y., Song, J., Yang, J., et al. (2019). Genome-wide association study and genomic prediction analyses of drought stress tolerance

in China in a collection of off-PVP maize inbred lines. *Mol. Breed.* 39:113. doi: 10.1007/s11032-019-1013-4

Xiao, Y., Tong, H., Yang, X., Xu, S., Pan, Q., Qiao, F., et al. (2016). Genome-wide dissection of the maize ear genetic architecture using multiple populations. *New Phytol.* 210, 1095–1106. doi: 10.1111/nph.13814

Xie, Y., Feng, Y., Chen, Q., Zhao, F., Zhou, S., Ding, Y., et al. (2019). Genome-wide association analysis of salt tolerance QTLs with SNP markers in maize (*Zea mays* L.). *Genes Genom.* 41, 1135–1145. doi: 10.1007/s13258-019-00842-6

Yang, S., Wang, J., Li, Q., Liu, L., Zhang, Z., Pan, G., et al. (2014). Genetic Analysis and Preliminary Mapping of a White Sheath Gene in Maize Inbred Line K10. *J. Plant Genet. Res.* 15, 1167–1172.

Yang, W., Feng, H., Zhang, X., Zhang, J., Doonan, J. H., Batchelor, W. D., et al. (2020). Crop Phenomics and High-Throughput Phenotyping: Past Decades, Current Challenges, and Future Perspectives. *Mol. Plant* 13, 187–214. doi: 10.1016/j.molp.2020.01.008

Yang, W., Guo, Z., Huang, C., Duan, L., Chen, G., Jiang, N., et al. (2014). Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice. *Nat. Commun.* 5:5087. doi: 10.1038/ncomms6087

Yang, W., Guo, Z., Huang, C., Wang, K., Jiang, N., Feng, H., et al. (2015). Genome-wide association study of rice (*Oryza sativa* L.) leaf traits with a high-throughput leaf scorer. *J. Exp. Bot.* 66, 5605–5615. doi: 10.1093/jxb/erv100

Yang, X., Gao, S., Xu, S., Zhang, Z., Prasanna, B. M., Li, L., et al. (2011). Characterization of a global germplasm collection and its potential utilization for analysis of complex quantitative traits in maize. *Mol. Breed.* 28, 511–526. doi: 10.1007/s11032-010-9500-7

Zhang, C., Craine, W. A., McGee, R. J., Vandemark, G. J., Davis, J. B., Brown, J., et al. (2020). Image-Based Phenotyping of Flowering Intensity in Cool-Season Crops. *Sensors* 20:1450. doi: 10.3390/s20051450

Zhang, J., Guo, J., Liu, Y., Zhang, D., Zhao, Y., Zhu, L., et al. (2016a). Genome-wide association study identifies genetic factors for grain filling rate and grain drying rate in maize. *Euphytica* 212, 201–212. doi: 10.1007/s10681-016-1756-5

Zhang, X., Warburton, M. L., Setter, T., Liu, H., Xue, Y., Yang, N., et al. (2016b). Genome-wide association studies of drought-related metabolic changes in maize using an enlarged SNP panel. *Oretic. Appl. Genet.* 129, 1449–1463. doi: 10.1007/s00122-016-2716-0

Zhang, Y., Li, P., Ren, W., Ni, Y., and Zhang, Y. (2019). *mrMLM: Multi-Locus Random-SNP-Effect Mixed Linear Model Tools for Genome-Wide Association Study. R package version 4.0.*

Zhang, Y., Wang, J., Du, J., Zhao, Y., Lu, X., Wen, W., et al. (2021). Dissecting the phenotypic components and genetic architecture of maize stem vascular bundles using high-throughput phenotypic analysis. *Plant Biotechnol. J.* 19, 35–50. doi: 10.1111/pbi.13437

Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42, 355–360. doi: 10.1038/ng.546

Zhang, Z., Zhou, B., Wang, H., Wang, F., Song, Y., Liu, S., et al. (2014). Maize purple plant pigment protects against fluoride-induced oxidative damage of liver and kidney in rats. *Int. J. Environ. Res. Public Health* 11, 1020–1033. doi: 10.3390/ijerph110101020

Zhao, C., Zhang, Y., Du, J., Guo, X., Wen, W., Gu, S., et al. (2019). Crop Phenomics: Current Status and Perspectives. *Front. Plant Sci.* 10:714. doi: 10.3389/fpls.2019.00714

Zheng, H., Yang, Z., Wang, W., Guo, S., Li, Z., Liu, K., et al. (2020). Transcriptome analysis of maize inbred lines differing in drought tolerance provides novel insights into the molecular mechanisms of drought responses in roots. *Plant Physiol. Biochem.* 149, 11–26. doi: 10.1016/j.plaphy.2020.01.027

Zhou, G., Hao, D., Chen, G., Lu, H., Shi, M., Mao, Y., et al. (2016). Genome-wide association study of the husk number and weight in maize (*Zea mays* L.). *Euphytica* 210, 195–205. doi: 10.1007/s10681-016-1698-y

Zhou, G., Hao, D., Xue, L., Chen, G., Lu, H., Zhang, Z., et al. (2018). Genome-wide association study of kernel moisture content at harvest stage in maize. *Breed. Sci.* 68, 622–628. doi: 10.1270/jsbbs.18102

Zhou, S., Chai, X., Yang, Z., Wang, H., Yang, C., and Sun, T. (2021). Maize-IAS: a maize image analysis software using deep learning for high-throughput plant phenotyping. *Plant Methods* 17:48. doi: 10.1186/s13007-021-00747-0

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

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

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 06 October 2022
DOI 10.3389/fpls.2022.1023938

OPEN ACCESS

EDITED BY
Xingtan Zhang,
Agricultural Genomics Institute at Shenzhen (CAAS), China

REVIEWED BY
Lei Zhang,
Chinese Academy of Agricultural Sciences (CAAS), China
Hui Guo,
Bionano Genomics, United States

*CORRESPONDENCE
Zhengjia Wang
wzhj21@163.com;
wzj@zafu.edu.cn
Kean-Jin Lim
keanjin.lim@zafu.edu.cn

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

RECEIVED 20 August 2022
ACCEPTED 22 September 2022
PUBLISHED 06 October 2022

CITATION
Guo W, Jin H, Chen J, Huang J, Zheng D, Cheng Z, Liu X, Yang Z, Chen F, Lim K-J and Wang Z (2022) GROP: A genomic information repository for oil plants. *Front. Plant Sci.* 13:1023938. doi: 10.3389/fpls.2022.1023938

GROP: A genomic information repository for oilplants

Wenlei Guo[1], Hongmiao Jin[1], Junhao Chen[1,2], Jianqin Huang[1], Dingwei Zheng[1], Zhitao Cheng[1], Xinyao Liu[1], Zhengfu Yang[1], Fei Chen[3,4], Kean-Jin Lim[1*] and Zhengjia Wang[1*]

[1]State Key Laboratory of Subtropical Silviculture, College of Forestry and Biotechnology, Zhejiang A&F University, Hangzhou, China, [2]Department of Biology, Saint Louis University, St. Louis, MO, United States, [3]College of Tropical Crops, Sanya Nanfan Research Institute, Hainan University, Haikou, China, [4]Hainan Yazhou Bay Seed Laboratory, Sanya Nanfan Research Institute, Hainan University, Sanya, China

Biomass energy is an essential component of the agriculture economy and represents an important and particularly significant renewable energy source in the fight against fossil fuel depletion and global warming. The recognition that many plants naturally synthesize hydrocarbons makes these oil plants indispensable resources for biomass energy, and the advancement of next-generation sequencing technology in recent years has now made available mountains of data on plants that synthesize oil. We have utilized a combination of bioinformatic protocols to acquire key information from this massive amount of genomic data and to assemble it into an oil plant genomic information repository, built through website technology, including Django, Bootstrap, and echarts, to create the Genomic Information Repository for Oil Plants (GROP) portal (http://grop.site/) for genomics research on oil plants. The current version of GROP integrates the coding sequences, protein sequences, genome structure, functional annotation information, and other information from 18 species, 22 genome assemblies, and 46 transcriptomes. GROP also provides BLAST, genome browser, functional enrichment, and search tools. The integration of the massive amounts of oil plant genomic data with key bioinformatics tools in a database with a user-friendly interface allows GROP to serve as a central information repository to facilitate studies on oil plants by researchers worldwide.

KEYWORDS

genomic data, oil plants, bioinformatics, information repository, transcriptomic data

Introduction

The advancement of human society has required enormous amounts of energy in many forms, beginning with firewood and expanding to coal and then to petroleum, natural gas, and kerogen shale. Fossil fuels began to occupy a crucial position when civilization became industrialized with the introduction of steam power. Now, however,

the excessive dependence on fossil fuels has become a potential threat to civilization. Among the 125 questions about exploration and discovery published in *Science* (Levine et al., 2021), three are related to energy: (1) Can we stop global climate change? (2) Where do we put all the excess carbon dioxide? and (3) Could we live in a fossil fuel–free world? The key answer to all three of these questions is probably biomass energy.

One significant contributor to biomass energy is oil plants—plants that naturally synthesize hydrocarbons, predominantly lipids, *in vivo*. The oil plants comprise a large group of herbs, shrubs, and trees, with commercial species that include oil palm (*Elaeis guineensis*), oilseed rape (*Brassica napus*), peanut (*Arachis hypogaea*), and soybean (*Glycine max*). In these plants, the lipids are mainly stocked in their seeds, although other species store hydrocarbons in their leaves, fruits, or stems. Plant lipids can be classified into five broad categories: (1) fatty acids with 16 to 18 carbons; (2) very long chain fatty acids with over 18 carbons; (3) polyunsaturated fatty acids; (4) hydroxy fatty acids; and (5) wax esters. These valuable lipids play vital roles in the food, paint, lubricant, feed, and medical industries due to their unique chemical properties (Bates et al., 2013; Belayneh et al., 2018). Some of these plant lipids are now recognized as having great medicinal value. For instance, omega-3 fatty acids, such as eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA), are essential in the human diet, as they cannot be synthesized *in vivo*, and they are now known to reduce heart attack risk(Ruxton et al., 2004) and to be effective in treating neurodegenerative and neurological disorders(Dyall, 2015), cancer(Moloudizargari et al., 2018), fetal development disorders (Dunstan et al., 2008), and cardiovascular disease (Bouwens et al., 2009). However, oil plants are now increasingly being recognized for their biomass value and development potential.

The growing desire to capitalize on the significant industrial and ecological value of oil plants has led to a multitude of scientific projects aimed at improving the yield and quality of these plants. Thanks to the rapid development of sequencing technology, an enormous collection of genome sequencing and transcriptomic data has been generated for many oil plants in recent years. The first wild olive (*Olea europaea*) genome (≈1.48G) was completed and published in 2017, and two Oleaceae-specific paleopolyploidization events were identified, leading to expansion and new functionalization of several gene families (*ACPTE*, *EAR*, *FAD2*, and *FAD2*) involved in lipid synthesis (Unver et al., 2017). For oil plants used in cosmetics and as lubricants, researchers have utilized a combination of multiple sequencing techniques, such as PacBio, Illumina, and Hi-C, to assemble a high-quality chromosome-level jojoba (*Simmondsia chinensis*) genome (≈887Mb, 2n=26) (Sturtevant et al., 2020). Oilseed rape (*Brassica napus*), the plant that produces the well-known canola oil that accounts for approximately 13–16% of all globally consumed vegetable oils (Wang et al., 2018), was "recently" formed about 7500 years ago by natural hybridization and polyploidization (Chalhoub et al., 2014), so the complex genome assembly of different accessions has been sequenced and improved many times (Chalhoub et al., 2014; Bayer et al., 2017; Sun et al., 2017; Zou et al., 2019; Song et al., 2020; Chen et al., 2021).

These advances have revealed important information about stress responses, domestication, and lipid synthesis in oil plants at the genomic level. The rapid accumulation of genomic data has also accelerated the process of plant molecular breeding by identifying and locating precise gene targets. Thus, constructing a database that can integrate, share, and visualize genomic and transcriptomic data for oil plants has become a necessity for the research community. We have addressed this need by constructing a public repository, the Genomic Repository of Oil Plants (GROP, www.grop.site), that stores and shares the genomic and transcriptomic data of oil plants.

GROP is the first digital resource library to store a range of oil plant genomic data, including genes, genome sequences, genome features, gene annotations, and transcriptome profiles. GROP also provides a batch of search tools and data visualization functions, including gene, gene family, transcription factor, protein kinase, and keyword searches, which allow users to retrieve relevant information quickly from the large collection of genomic data. A Basic Local Alignment Search Tool (BLAST) server has been deployed in GROP to integrate genomic, gene nucleotide, and protein sequences. A genome browser was also embedded in GROP for the integrative visualization of genomic sequences, annotation data, and functional genomic data. GROP also includes tools for Gene Ontology (GO) enrichment, pathway enrichment, and expression visualization that allow users to perform functional analyses of their gene sets. We expect that GROP will serve as an efficient genomic data center for the research community interested in oil plants. We plan to continuously update GROP to include newly generated genomic data.

Material and methods

Genomic and transcriptome data resource acquisition

Genomic data includes genome sequences, coding sequences (CDS), protein sequences (PEP), and genome structure annotation files (GFF) that are essential for in-depth exploration of oil production by plants. The original genomic data sources in GROP are composed of two parts, with one part coming from major public bioinformatics databases and the second part being a new version of the pecan (*Carya illinoinensis*) genome assembly contributed by our research

team. The raw sequencing data of the 46 oil plant transcriptomes in the database were downloaded from the NCBI Sequence Read Archive (SRA) database, which comprises RNA-seq from different tissues at different developmental stages and from plants in different environments. The RNA-seq data were downloaded using the SRA toolkit. All data can be traced in the database (Supplementary Table S1).

Gene annotation

Gene function annotation compares a gene sequence or protein sequence with bioinformatics databases to predict the function of the gene. In GROP, the latest InterPro (Finn et al., 2016) protein resource package (v86.0) and Panther classification data were collected. We then used InterProScan v5.52 software to perform functional predictions against protein sequences with 30 threads and a GO term output mode. These annotation results provide comprehensive prospects for potential functions. The KEGG online analysis service BlastKOALA (Kanehisa et al., 2016) was utilized to identify vital genes in the KEGG pathways.

Gene family or protein domain classification was conducted using Pfam (El-Gebali et al., 2019). We first obtained the conserved domain feature resources included in the Pfam database, and we then used the HMMER software (Finn et al., 2011) to search all the oil plant protein sequences against the HMM seeds. Transcriptional factors and protein kinases were identified using the iTAK software (Zheng et al., 2016) with the default parameters.

Expression profile calculations

For the 46 oil plant transcriptome datasets, a standard and optimized pipeline was set up for automated calculations: (1) fastq-dump command with split-3 was executed to transform SRA data to the fastq format data using the SRA toolkit; (2) fastp (Chen et al., 2018) was used for quality controlling, filtering, adapter trimming, and per-read quality pruning; (3) a genome index was built and the raw sequencing data mapped to the corresponding genome to obtain a BAM file using the STAR software (Dobin et al., 2013); and (4) the gene expression value was calculated and normalized to FPKM (fragments per kilobase of exon per million mapped fragments) and joint different samples to a complete matrix using the RSEM software (Li et al., 2009). These final expression profiles were ultimately uploaded to the cloud server.

BLAST, genome browser, and FTP server implementation

We provide clear and easy use of similar sequence search services for oil plant researchers by deploying an online BLAST service based on the coding, protein, and genome sequences of 22 oil plant genome assemblies. This BLAST service is set up using SequenceServer (Priyam et al., 2019) and optimized with the jstree module, which splits the sequence library into three parts (coding, protein, and genome sequences) and classifies different assemblies that belong to the same species into one node.

The genome browser is deployed on the Linux server so that users can access the genomic information intuitively and interactively using multiple web browsers (Chrome, Firefox, IE, Safari, etc.). We configured an extendable genome browser using Jbrowser (Buels et al., 2016) to integrate genomic information about sequences, gene structure, RNA-seq, mutation sites, etc.

We have provided a rapid data acquisition channel for worldwide researchers by setting up a File Transfer Protocol (FTP) download site utilizing the vsftp technique. In this case, complete genomic sequences, genome structure annotation data, multifold functional annotation data, and 46 RNA-seq expression matrices have been cached on the FTP site.

Data model and website implementation

The data structure and relationship of the four most significant data models (Species, Genome assembly, Transcriptome profile, Gene) are shown in Figure 1. These four data models are instrumental in the data retrieval and assay procedures within the whole dataset (Table 1). In general, GROP was established using a series of modern website development techniques, including HTML, CSS, Bootstrap, MySQL, and Django. Django was the key element connecting the front webpage and genomic information, and it facilitates the development process. The Django framework also allows us to decorate a data management system that will facilitate future data updates. The whole GROP project has been deployed on a cloud server using a mixture pipeline of Gunicorn and Nginx, as we described previously (Guo et al., 2020). Ultimately, a user-friendly and usage-flexible public genomic platform has been built for the oil plant scientific community.

Results

The GROP homepage

The homepage of the Genomic Repository of Oil Plants (Figure 2) consists of three parts, from top to bottom: a navigation bar, the main content, and footer content. The homepage also offers links to several important bioinformatic resources and services. The navigation bar includes multiple dropdown menus that link to important information and online tools, such as “Species,” “Genome assembly,” and “Tools.” The

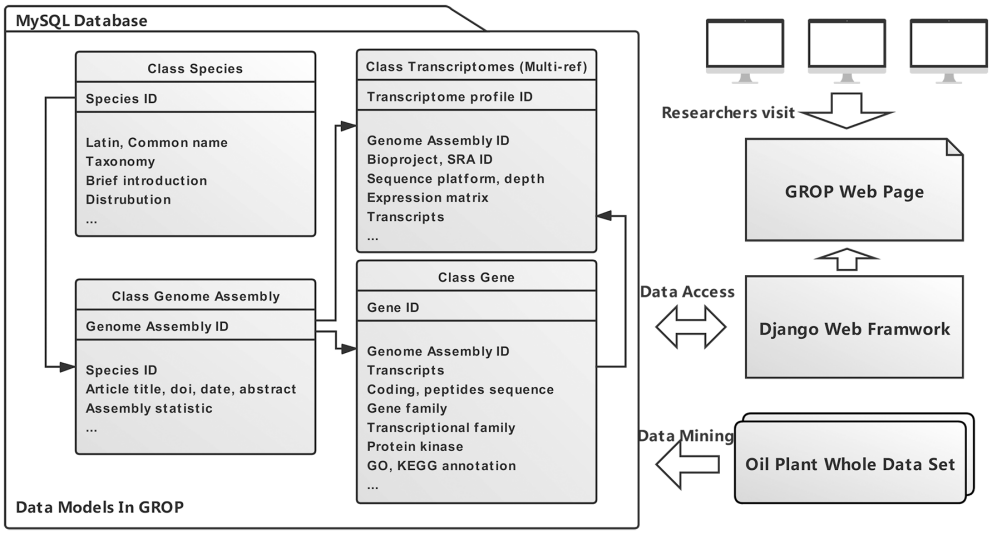


FIGURE 1
An overview of the GROP complete architecture. In general, four data models (species, genome assembly, transcriptome profile, and gene) were built through data mining from an oil plant whole dataset. Researchers are able to visit GROP *via* the Django web framework.

"Species" dropdown menu in the navigation bar (Figure 2A) lists the 18 species currently included in GROP. Clicking on a species name redirects users to its description page. The "Genome assembly" dropdown menu currently lists 22 genome assemblies for 18 oil plants and links to a new page that includes the statistical information of the genome assembly and the abstract of its publication. The "Tool" contains links to all the tools, which are described in detail in the sections below. GROP also allows users to download all raw data with an FTP protocol, and the link is also included in the navigation bar. The main content (Figure 2B) is divided into three sections, from top to bottom: the top section contains the four most used tools in GROP; the middle section shows pictures and links to the 18 oil plants in the GROP; and the bottom section includes an introduction, links, toolbox, and data statistics and recent updates. At the bottom of the homepage, the footer section (Figure 2C) shows the logos of the techniques used for the development of GROP and provides the author contact information.

TABLE 1 Data statictis of dataset in genomics respository of oil plant.

Data model	Item count
Species	18
Genome assembly	22
Transciptome profile	46
Gene	413,509
Gene family	586,165
Transcriptional factor	33,485
Protein kinase	16,405
GO annotation	3,649,008
KEGG annotation	409,358

BLAST and genome browser for genomic data

GROP provides a BLAST tool for sequence similarity searches. The gray box at the top of the BLAST interface (Figure 3A) is the input box for the query sequences. In the middle of the page, the user can select a target sequence database that includes coding, protein, and genomic sequences for each species (Figure 3B). The optional advanced parameters, such as the E-value threshold and the number of alignments to be shown, can be chosen in the input box at the bottom of the page (Figure 3C). The returned results are divided into three parts: the top left part is a graphical representation of the Blast hits found (Figure 3D) and provides a quick overview of the query sequence and the resulting hit sequences. The bottom left part shows the BLAST table and alignments of the BLAST results (Figure 3E). The download links are provided in the right half (Figure 3F), allowing users to download BLAST results in either the tab-separated or XML formats.

The basic interface of the genome browser (Figure 4) provided by GROP includes three parts: the selection bar on the left (Figure 4A), the navigation bar on the right (Figure 4B), and the main display of the Genome Browser (Figure 4C). Users can select an oil plant species from a list in the navigation bar. Users can also upload various types of genomic annotation

FIGURE 2
Homepage of the GROP database. **(A)** Navigation bar of GROP, including the species introduction, tool links, etc. **(B)** Main content, including a quick link for the four most useful tools, species list, repository introduction, etc. **(C)** Website footer of GROP.

information to be displayed in the main panel. Available tracks also provide users with annotated information of the gene structure and other desired genomic information. For example, BAM files can be uploaded to the genome browser for visualization of aligned reads to the reference genome and to detect base mismatches, insertions, deletions, and other variation information. With these functions, users can visually analyze their own genomic data of interest.

Single gene search

The single gene search is one of the most important search tools in GROP. The gene search tool can be implemented by entering a gene identifier, which can be obtained in various ways (BLAST, JBrowse, etc.). The gene search returns the source organism, gene structure, sequence length, gene family, coding, and protein sequences, and expression level. The functional

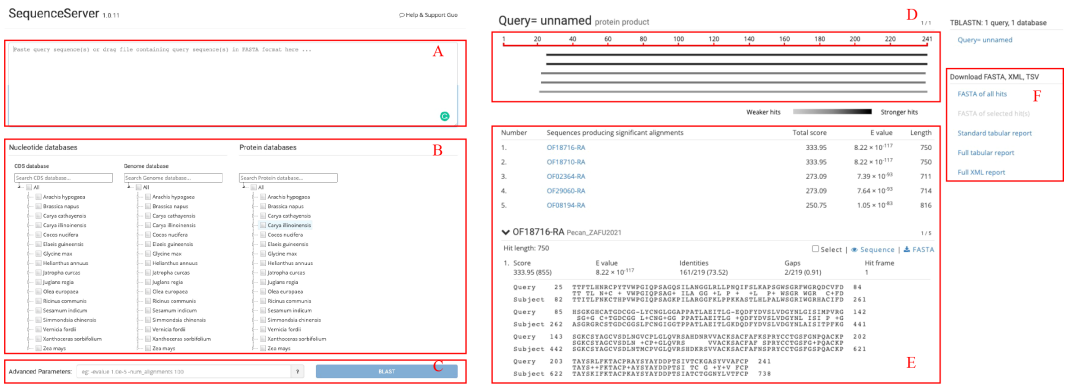


FIGURE 3
BLAST interface of GROP. **(A)** Query sequence input box for the BLAST program. **(B)** Search databases containing nucleotide and protein sequences. **(C)** BLAST advanced parameter input box. **(D)** Visualized BLAST alignment map. **(E)** Detailed BLAST result information. **(F)** BLAST results for download.

annotation information from various bioinformatic databases, such as Amigo, CDD, Gene3D, InterProScan, MobiDBLite, Pfam, and Swiss-prot, is also provided in the gene research results.

Bulk search and download

GROP provides multiple search methods to find clusters of conserved genes, including gene families, transcription factors, and protein kinases. Entering the Pfam ID of a gene family and genome assembly into the gene family search function will return a list of genes. Several methods to manipulate sequences in bulk are also supported at the top of the gene list page. GROP also provides a keyword search tool, based on an SQL fuzzy search for the entire database and including gene coding, protein, and genome sequences, as well as InterProScan, KEGG metabolic pathway, and Pfam annotation data.

An FTP download site was built with the complete data resources for oil plants. This download site stores the nucleic acid sequences of coding proteins, protein sequences, genome sequences, genome structural features files for each genome, functional annotation information about protein domains, GO terms, and KEGG metabolic terms. We have also integrated the gene expression data in the FTP site. Through the site, researchers can download the entire dataset in the repository.

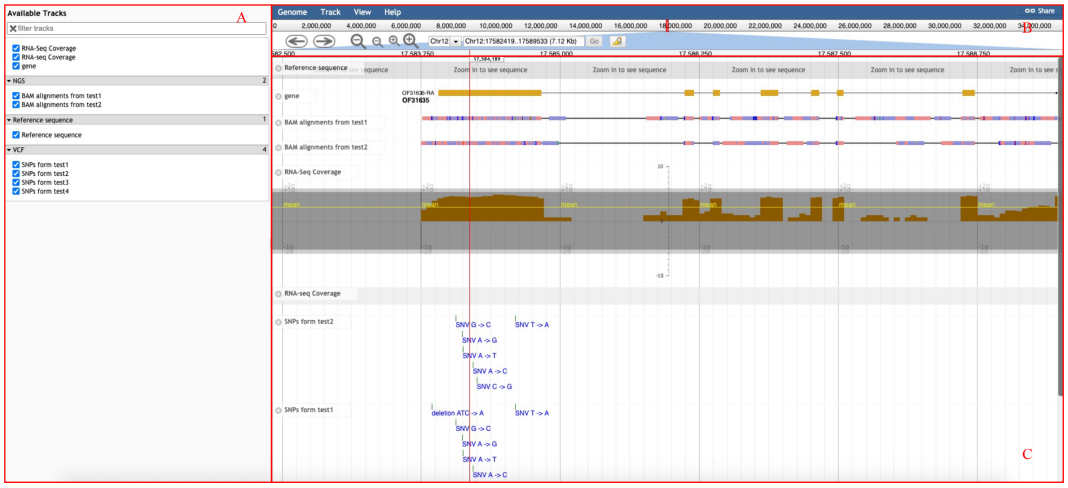


FIGURE 4
Genome browser interface of GROP. **(A)** Selection bar for genomic information in the browser. **(B)** Navigation bar for chromosome position, zoom in and out. **(C)** Visualization interface for selected genomic information.

Enrichment tool for GO and pathway terms

On the GROP site, GO enrichment analysis can be performed with a simple workflow: select the genome version of the oil plant; select an ontology category (Biological Process, Cellular Component, Molecular Function, or All); and provide a list of gene IDs and the threshold p-value (Figure 5A). The result of the enrichment analysis returned by GROP is a bar plot of the enrichment analysis, where each bar represents an enriched GO term, the length of the bar indicates the number of genes included in the GO term, and the shade of the bar indicates the p-value, with a color closer to red indicating a smaller p-value with higher confidence (Figure 5B). A table of the results of the enrichment analysis, presented beneath the bar graph, lists each enriched GO ID and the gene frequency, p-value, link to the gene list, GO function description, and annotation link. With this tool, researchers can conveniently perform GO enrichment and KEGG enrichment analysis on a specific list of genes in the 18 oil plants, thereby eliminating the need to use the R package or other programs.

Expression visualizer

The heatmap is a popular visualization of gene expression data that shows the expression levels of multiple genes or transcripts in different environments, developmental stages, or tissues. Researchers can easily comprehend the gene expression pattern under different backgrounds.

A gene expression heatmap is generated in the Expression Visualizer of GROP when the users select a RNA-seq Bioproject from the dropdown list and then provide gene IDs of the query. (Figure 6A). The gene expression visualizer returns an expression heatmap of the input genes of the corresponding transcriptome (Figure 6B). The horizontal axis and vertical axes of the heatmap represent different genes and different samples in the transcriptome, respectively. The heatmap is available in different scales and can be downloaded. Gene expression data are also provided as a data matrix at the bottom of the web page for further analysis (Figure 6C).

Discussion

The agricultural development and utilization of biomass is an important direction in developing renewable energy that can be used to replace traditional fuel oil or coal energy. Oil plants, such as hickory, olive, pecan, soybean, and walnut, are important sources of healthy edible oil as well as important sources of economic income for the food industry and agroforestry. By contrast, jatropha, jojoba, oil palm, and tung oil trees are widely used in manufacturing industries for the production of plastics, lacquers, artificial rubber, printing inks, etc. The oil produced by these oil plants is environmentally friendly and pollution-free and has high economic and environmental value. Therefore, the investigation of germplasm resources, genetics, and molecular regulation of fatty acid synthesis traits has been the main focus of oil crop research.

The rapid development of high-throughput sequencing has allowed the assembly and annotation of the whole genome

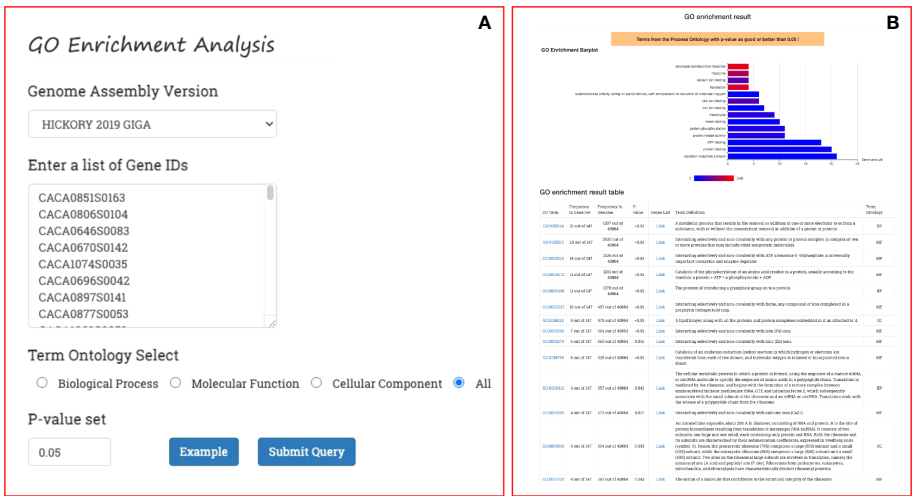


FIGURE 5
GO and KEGG enrichment analysis process in GROP. **(A)** Submission form for genome version, gene set, ontology term, and p-value. **(B)** Returned enrichment result, including a histogram and a summary table.

all_stages	accumulation_stage_1	accumulation_stage_2	accumulation_stage_3	seeds_initiate_stage_1	seeds_initiate_stage_2	seeds_initiate_stage_3
91.0	70.0	53.0	73.0	97.0	119.0	102.0
357.0	256.0	267.0	263.0	912.0	658.0	671.0
200.0	134.0	141.0	153.0	532.0	263.0	261.0
48.0	38.0	43.0	28.0	82.0	41.0	44.0
7.0	4.0	1.0	2.0	13.0	8.0	10.0
193.0	104.0	104.0	99.0	396.0	161.0	198.0
163.0	125.0	111.0	115.0	456.0	312.0	277.0

FIGURE 6
Gene expression visualizer in GROP. **(A)** Input panel of the transcriptome version and gene set. **(B)** Gene expression heatmap for different gene sets, environments, and developmental stages. **(C)** Gene expression data matrix.

sequences of many important model plants and field crops; however, the genomes of oil plants are difficult to decipher due to the large proportion of repetitive sequences, high heterozygosity, and large genome size. Nevertheless, in recent years, the reduction in the cost of second-generation sequencing technology and the development of third-generation long-read sequencing has led to breakthroughs and the accumulation of a huge amount of genomic and transcriptomic data for oil plants. Therefore, the current issue is how to effectively analyze, integrate, and share the genomic and transcriptomic data of oil plants in the post-genomic era.

This study is the first to construct a genomic database for oil plants. The work has analyzed and integrated the genomic data for 22 oil plant genomes and 46 transcriptomic datasets using various bioinformatics software, such as HMMER, InterProScan, iTAK, and STAR, and has stored this information in a MySQL database. A user-friendly web platform was also established using Django. The resulting repository provides gene, gene family, transcription factor, protein kinase, and keyword searches, with efficient retrieval. The repository also provides gene ontology (GO) and metabolic pathway (KEGG) enrichment analysis tools to resolve significantly distributed functional or metabolic pathways in gene sets. Researchers are also able to find the expression data for gene sets from 46 transcriptomic gene expression datasets in the Expression Visualizer. The BLAST and the JBrowse browser tools are available in the repository, allowing researchers to search for homologous sequences, browse the location and structure of genes, and view the variation and expression abundance of genes in different species in combination with VCF and BAM files.

Currently, we still anticipate expanding the amount of data and adding omics data, such as metabolomes. Persons interested in further development of GROP are welcomed to share data or to participate in any other kind of collaboration. The construction of the Genomic Repository of Oil Plants will fuel genomics and molecular biology research while enriching our future understanding of oil plants. We believe that GROP will become the data center for oil plant studies, and that efforts in GROP will contribute substantially to oil plant research.

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

ZW, WG, K-JL conceived and designed the research; WG, FC obtained, analyzed the data, organized, and constructed the architecture of website. JH, JC, HJ, DZ, ZC and XL collect the information about genome assembly and transcriptomes. WG, K-JL designed the layout of web pages. WG, HJ wrote, K-JL revised the manuscript. ZW and YZ acquired the funding. All authors contributed to the article and approved the submitted version.

Funding

This work is supported by the grant from the National Key R&D Program of China (2018YFD1000604); the Zhejiang Agriculture New Variety Breeding Major Science and Technology Special (2021C02066-12); the National Natural Science Foundation of China (32001335).

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.2022.1023938/full#supplementary-material

References

Bates, P. D., Stymne, S., and Ohlrogge, J. (2013). Biochemical pathways in seed oil synthesis. *Curr. Opin. Plant Biol.* 16, 358–364. doi: 10.1016/j.pbi.2013.02.015

Bayer, P. E., Hurgobin, B., Golicz, A. A., Chan, C. K. K., Yuan, Y., Lee, H. T., et al. (2017). Assembly and comparison of two closely related brassica napus genomes. *Plant Biotechnol. J.* 15, 1602–1610. doi: 10.1111/pbi.12742

Belayneh, H. D., Wehling, R. L., Cahoon, E., and Ciftci, O. N. (2018). Lipid composition and emulsifying properties of camelina sativa seed lecithin. *Food Chem.* 242, 139–146. doi: 10.1016/j.foodchem.2017.08.082

Bouwens, M., Van De Rest, O., Dellschaft, N., Bromhaar, M. G., De Groot, L. C. P. G. M., Geleijnse, J. M., et al. (2009). Fish-oil supplementation induces antiinflammatory gene expression profiles in human blood mononuclear cells. *Am. J. Clin. Nutr.* 90, 415–424. doi: 10.3945/ajcn.2009.27680

Buels, R., Yao, E., Diesh, C. M., Hayes, R. D., Munoz-Torres, M., Helt, G., et al. (2016). JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.* 17, 66. doi: 10.1186/s13059-016-0924-1

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I. A. P., Tang, H., Wang, X., et al. (2014). Early allopolyploid evolution in the post-neolithic brassica napus oilseed genome. *Sci. (1979)* 345, 950–953. doi: 10.1126/science.1253435

Chen, X., Tong, C., Zhang, X., Song, A., Hu, M., Dong, W., et al. (2021). A high-quality brassica napus genome reveals expansion of transposable elements, subgenome evolution and disease resistance. *Plant Biotechnol. J.* 19, 615–630. doi: 10.1111/pbi.13493

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890. doi: 10.1093/bioinformatics/bty560

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Dunstan, J. A., Simmer, K., Dixon, G., and Prescott, S. L. (2008). Cognitive assessment of children at age 2 1/2 years after maternal fish oil supplementation in pregnancy: A randomised controlled trial. *Arch. Dis. Child Fetal Neonatal Ed* 93, F45-50. doi: 10.1136/adc.2006.099085

Dyall, S. C. (2015). Long-chain omega-3 fatty acids and the brain: A review of the independent and shared effects of EPA, DPA and DHA. *Front. Aging Neurosci.* 7. doi: 10.3389/fnagi.2015.00052

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

Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., et al. (2016). InterPro in 2017–beyond protein family and domain annotations. *Nucleic Acids Res.* 45, D190–D199. doi: 10.1093/nar/gkw1107

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* 39, 29–37. doi: 10.1093/nar/gkr367

Guo, W., Chen, J., Li, J., Huang, J., Wang, Z., and Lim, K. J. (2020). Portal of juglandaceae: A comprehensive platform for juglandaceae study. *Hortic. Res.* 7, 1-8. doi: 10.1038/s41438-020-0256-x

Kanehisa, M., Sato, Y., and Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.* 428, 726–731. doi: 10.1016/j.jmb.2015.11.006

Levine, G. A., French, B., and Sanders, S. eds. (2021). *125 questions: Exploration and Discovery*. (Shanghai, China: Science/AAAS).

Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., and Dewey, C. N. (2009). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* 26, 493–500. doi: 10.1093/bioinformatics/btp692

Moloudizargari, M., Mortaz, E., Asghari, M. H., Adcock, I. M., Redegeld, F. A., and Garssen, J. (2018). Effects of the polyunsaturated fatty acids, EPA and DHA, on hematological malignancies: A systematic review. *Oncotarget* 9, 11858–11875. doi: 10.18632/oncotarget.24405

Priyam, A., Woodcroft, B. J., Rai, V., Moghul, I., Munagala, A., Ter, F., et al. (2019). Sequenceserver: A modern graphical user interface for custom BLAST databases. *Mol. Biol. Evol.* 36 (12), 2922–2924. doi: 10.1093/molbev/msz185

Ruxton, C. H. S., Reed, S. C., Simpson, M. J. A., and Millington, K. J. (2004). The health benefits of omega-3 polyunsaturated fatty acids: A review of the evidence. *J. Hum. Nutr. Dietetics* 17, 449–459. doi: 10.1111/j.1365-277X.2004.00552.x

Song, J. M., Guan, Z., Hu, J., Guo, C., Yang, Z., Wang, S., et al. (2020). Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of brassica napus. *Nat. Plants* 6, 34–45. doi: 10.1038/s41477-019-0577-7

Sturtevant, D., Lu, S., Zhou, Z. W., Shen, Y., Wang, S., Song, J. M., et al. (2020). The genome of jojoba (Simmondsia chinensis): A taxonomically isolated species that directs wax ester accumulation in its seeds. *Sci. Adv.* 6, 1–14. doi: 10.1126/sciadv.aay3240

Sun, F., Fan, G., Hu, Q., Zhou, Y., Guan, M., Tong, C., et al. (2017). The high-quality genome of brassica napus cultivar 'ZS11' reveals the introgression history in semi-winter morphotype. *Plant J.* 92, 452–468. doi: 10.1111/tpj.13669

Unver, T., Wu, Z., Sterck, L., Turktas, M., Lohaus, R., Li, Z., et al. (2017). Genome of wild olive and the evolution of oil biosynthesis. *Proc. Natl. Acad. Sci.* 114, E9413–E9422. doi: 10.1073/pnas.1708621114

Wang, B., Wu, Z., Li, Z., Zhang, Q., Hu, J., Xiao, Y., et al. (2018). Dissection of the genetic architecture of three seed-quality traits and consequences for breeding in brassica napus. *Plant Biotechnol. J.* 16, 1336–1348. doi: 10.1111/pbi.12873

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

Zou, J., Mao, L., Qiu, J., Wang, M., Jia, L., Wu, D., et al. (2019). Genome-wide selection footprints and deleterious variations in young Asian allotetraploid rapeseed. *Plant Biotechnol. J.* 17, 1998–2010. doi: 10.1111/pbi.13115

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 13 October 2022
DOI 10.3389/fpls.2022.1036764

OPEN ACCESS

EDITED BY
Yin Li,
Huazhong University of Science and Technology, China

REVIEWED BY
Chuang Ma,
Northwest A&F University, China
Enhua Xia,
Anhui Agriculture University, China

*CORRESPONDENCE
Haifeng Wang
haifengwang@gxu.edu.cn
Baoshan Chen
chenbs2008@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 05 September 2022
ACCEPTED 28 September 2022
PUBLISHED 13 October 2022

CITATION
Xue Y, Zou C, Zhang C, Yu H, Chen B and Wang H (2022) Dynamic DNA methylation changes reveal tissue-specific gene expression in sugarcane. *Front. Plant Sci.* 13:1036764. doi: 10.3389/fpls.2022.1036764

Dynamic DNA methylation changes reveal tissue-specific gene expression in sugarcane

Yajie Xue[1,2†], Chengwu Zou[1,2†], Chao Zhang[1,2], Hang Yu[1,2], Baoshan Chen[1*] and Haifeng Wang[1,2*]

[1]State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources, College of Agriculture, Guangxi University, Nanning, China, [2]Guangxi Colleges and Universities Key Laboratory of Crop Cultivation and Tillage, Guangxi University, Nanning, China

DNA methylation is an important mechanism for the dynamic regulation of gene expression and silencing of transposons during plant developmental processes. Here, we analyzed genome-wide methylation patterns in sugarcane (*Saccharum officinarum*) leaves, roots, rinds, and piths at single-base resolution. DNA methylation patterns were similar among the different sugarcane tissues, whereas DNA methylation levels differed. We also found that DNA methylation in different genic regions or sequence contexts plays different roles in gene expression. Differences in methylation among tissues resulted in many differentially methylated regions (DMRs) between tissues, particularly CHH DMRs. Genes overlapping with DMRs tended to be differentially expressed (DEGs) between tissues, and these DMR-associated DEGs were enriched in biological pathways related to tissue function, such as photosynthesis, sucrose synthesis, stress response, transport, and metabolism. Moreover, we observed many DNA methylation valleys (DMVs), which always overlapped with transcription factors (TFs) and sucrose-related genes, such as *WRKY*, *bZIP*, *WOX*, *SPS*, and *FBPase*. Collectively, these findings provide significant insights into the complicated interplay between DNA methylation and gene expression and shed light on the epigenetic regulation of sucrose-related genes in sugarcane.

KEYWORDS

sugarcane, DNA methylation, differentially methylated regions, DNA methylation valleys, epigenetics

Introduction

DNA methylation is among the most common epigenetic modifications in eukaryotic genomes and is involved in regulating gene transcription and transposon silencing (Law and Jacobsen, 2010; Zhang et al., 2018a). In animals, DNA methylation mainly occurs at CG sites, whereas in plants, it occurs at CG, CHG, and CHH sites (H represents A, T, or C) (Law

and Jacobsen, 2010). Information on DNA methylation in plants is mainly derived from model plants, such as *Arabidopsis thaliana* and rice (*Oryza sativa*). Methylation in three different contexts is established and maintained by different pathways. CG methylation is mainly catalyzed by methyltransferase 1 (MET1) (Kankel et al., 2003), while chromomethylase 3 (CMT3) is responsible for maintaining CHG methylation. Recent studies have shown that CMT2 is also involved in the maintenance of CHG methylation (Lindroth et al., 2001; Stroud et al., 2014) and plays a major role in maintaining asymmetric CHH methylation. CHH methylation maintained by CMT2 always occurs at long transposable elements (TEs), which are often located in peri-centromeric regions (Zemach et al., 2013; Stroud et al., 2014; Gouil and Baulcombe, 2016). In all three contexts, cytosines can be *de novo* methylated by the RNA-directed DNA methylation (RdDM) pathway, which also involves domains rearranged methyltransferase (DRM2) and several other proteins (Law and Jacobsen, 2010; Kawashima and Berger, 2014; Cuerda-Gil and Slotkin, 2016). DNA methylation is dynamically regulated by methylases and demethylases, and four DNA demethylases have been identified in *A. thaliana*, including ROS1, DME, DML2, and DML3 (Choi et al., 2002; Gong et al., 2002; Morales-Ruiz et al., 2006; Ortega-Galisteo et al., 2008).

Recently, extensive studies have shown that DNA methylation plays an important role in plant growth, development, and stress response (Zhang et al., 2018a; Chang et al., 2020). For example, deficient non-CG methylation levels in Arabidopsis resulted in a twisted leaf shape, shorter stature, and partial sterility phenotypic defects (Chan et al., 2006). In addition, 70% of drought-induced methylation changes in rice were recovered after irrigation resumed (Wang et al., 2011). Salt stress inhibits DNA methylation in the promoter region of *OsMYB91*, promoting its expression and increasing salt tolerance in rice (Zhu et al., 2015). Although extensive studies on plant DNA methylation have been reported, most have focused on models or economically important crops, such as rice, soybean (*Glycine max*), sorghum (*Sorghum bicolor*), cassava (*Manihot esculenta*), and tomato (*Solanum lycopersicum*) (Li et al., 2012; Song et al., 2013; Wang et al., 2015; Turco et al., 2017; Wang et al., 2018). These genomes are relatively small and have low complexity, and very few DNA methylation studies have been conducted on species with large genomes and high genome complexity, such as bread wheat (*Triticum aestivum*), Norway spruce (*Picea abies*), and Chinese pine (*Pinus tabuliformis*) (Ausin et al., 2016; Li et al., 2019; Niu et al., 2022). Owing to the complexity of the sugarcane (*Saccharum officinarum*) genome (large genome size and polyploidy), its reference genome has only recently been released, providing an unprecedented opportunity to investigate the role of DNA methylation in sugarcane growth.

Here, we explored genome-wide DNA methylation profiles in four different sugarcane tissues using whole-genome bisulfite sequencing (WGBS). Combined with transcriptome data, we investigated the association between DNA methylation changes and expression divergence among four tissues (leaf, rind, pith, and root). Moreover, comparative multi-omics analysis revealed the regulatory role of DNA methylation variation in the different sugarcane tissues, especially in genes related to important agronomic traits. Thus, our study provides a unique insight into the role of DNA methylation in sugarcane research.

Materials and methods

Plant materials and tissue collection

Sugarcane cultivar Zhongzhe No. 1 was grown at the Fusui planting base of Guangxi University (22°17'N, 107°31'E). We selected sugarcane at the mature stage for sampling, in which root, leaf +1, rind and pith of 10th stalk were collected.

Whole-genome bisulfite sequencing and analysis

The whole-genome bisulfite sequencing (WGBS) library was constructed as described by Wang (Wang et al., 2015). WGBS libraries were sequenced on an Illumina NovaSeq 6000 system (Illumina, San Diego, CA, USA) to obtain pair-end 150-bp reads.

Raw 150-bp paired-end reads were subjected to quality control filters using FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and trimmed using Trimmomatic v0.39 (Bolger et al., 2014). The clean reads were aligned to the sugarcane reference genome (Zhang et al., 2018b) using BSMAP v2.90 (Xi and Li, 2009), and up to 10 base mismatches were allowed. Only uniquely mapped reads were used to estimate the methylation ratios. The methylation ratio was calculated from the number of sequenced cytosines divided by the total read depth [mC/(mC + non-mC)], and visual analysis was conducted using ViewBS v0.1.9 (Huang et al., 2018).

Reproducibility between replicates of BS-seq was calculated as methylation levels in 100-kb regions in both replicates, and Pearson correlation coefficients between replicates were calculated.

The differentially methylated regions (DMRs) between different tissues were calculated using the methylKit R package (Akalin et al., 2012); the genome was divided into 100bp bins and mC sites covered by more than 3 reads were used for subsequent analysis. The methylation differences between all sequence contexts were as follows: CG difference was greater than 0.4, CHG difference was greater than 0.2, and CHH difference was greater than 0.1.

Transcriptome sequencing and analysis

Total RNA was isolated from the same tissues used in the WGBS library using TRIzol reagent (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's instructions. The RNA-

seq library was constructed following the Illumina kit's recommendation and sequenced using Illumina NovaSeq 6000 (Illumina) with paired-end reads of 150 bp.

FASTQC was used for initial read quality control. Clean reads were mapped to the sugarcane reference genome (Zhang et al., 2018b) using hisat2 V2.1.0 with default settings (Kim et al., 2015). We used Stringtie v2.1.4 to calculate the gene expression levels (Pertea et al., 2015). Differentially expressed genes (DEGs) were identified using DESeq2 v1.32.0 (Sahraeian et al., 2017) with a 4-fold change and FDR < 0.05.

Gene ontology enrichment analysis

Gene functions were annotated using eggNOG-mapper (Huerta-Cepas et al., 2019), and Gene Ontology (GO) enrichment analysis was performed using GOATOOLS with false-discovery rate correction (<0.05) (Klopfenstein et al., 2018).

Identification and characterization of sugarcane DMVs

The DNA methylation valleys (DMVs) in sugarcane were identified as previously described (Lin et al., 2017; Chen et al., 2018; Li et al., 2018). Briefly, the genome was first divided into 1-kb bins, and we calculated the DNA methylation levels in each bin. The DMV is the bin where the methylation levels of all sequence contexts are less than 5% in all tissues. Next, all overlapping DMVs were merged (Figure 6B). Finally, genes (gene body and flanking 1 kb) located in the DMVs were defined as DMV genes.

Results

Characterization of DNA methylation patterns among different sugarcane tissues

To investigate the DNA methylation patterns in sugarcane, we used WGBS to examine cytosine methylation in four sugarcane tissues: leaf, root, rind, and pith. Each sample was sequenced in two biological replicates, and approximately 70% of the reads were aligned to the reference genome, except for one biological replicate of the roots (Table S1). Pearson's correlation coefficients between different biological replicates were greater than 0.95, except in the roots, indicating the high reproducibility and accuracy of our sequencing data (Figure S1). Next, we merged the two replicates because their data were highly correlated. There were 1,137 million cytosines that could be methylated in sugarcane, accounting for 39.2% of the sugarcane genome; approximately 87% of the total cytosines were covered by at least one read (Figure S2).

From the distribution of global DNA methylation, we found that gene-enriched regions showed low CG and CHG methylation levels, while transposable element (TE)-enriched regions had high methylation levels (Figure 1A). This result is consistent with previous studies on other plants (Song et al., 2013; Wang et al., 2015; Song et al., 2015). In addition, we found that CHH methylation was slightly enriched in the gene-enriched regions compared with that in regions with dense CG and CHG methylation (Figure 1A). The distribution of CHH methylation in sugarcane is consistent with that in maize (*Zea mays*) (Gent et al., 2013). We also found a negative correlation between gene and TE densities (R=-0.68, p < 2.2e-16) (Figures 1A and Figure S3). To better understand the relationship between DNA methylation levels and gene and TE densities, we calculated their correlation coefficients. We found that both mCG and mCHG methylation negatively correlated with gene density, indicating that these two DNA methylation contexts were mostly located in gene-poor heterochromatic regions. However, leaf tissue showed no correlation, and the other three tissues showed positive correlations between gene density and mCHH levels (Figure S4). This result was consistent with findings in rice, sorghum (*Sorghum bicolor*), and maize (Gent et al., 2013; Niederhuth et al., 2016). As expected, TE density positively correlated with CG and CHG methylation levels (R > 0.7) but showed a weak correlation with CHH methylation (|R|<0.3) (Figure S5).

To investigate the relationship between TE methylation and the distance between TEs and adjacent genes, we calculated the methylation levels of TEs. We found that higher TE CHH methylation levels in all tissues positively correlated with the closer distances of TEs to the gene, but this phenomenon was not observed in CG and CHG methylation (Figure S6). Altogether, these results suggest that gene and transposon densities and methylation levels correlate, and the distribution of genes and transposons in the genome jointly shapes the landscape of DNA methylation in different regions of the genome.

Genome-wide distribution and global DNA methylation levels showed obvious DNA methylation changes among the four tissues (Figures 1A, B). Pith tissue showed the highest DNA methylation levels, followed by the rind, root, and leaf. In contrast to methylation levels, we found no significant differences in the proportion of methylcytosines among the four tissues, with CHH methylcytosine being the most abundant (>67%), followed by CG and CHG methylcytosines (Figure 1C). This finding is consistent with other plant studies (Wang et al., 2015; Xu et al., 2018).

DNA methylation patterns of gene and TE regions

Genome-wide DNA methylation analysis revealed substantial differences in methylation levels among the four

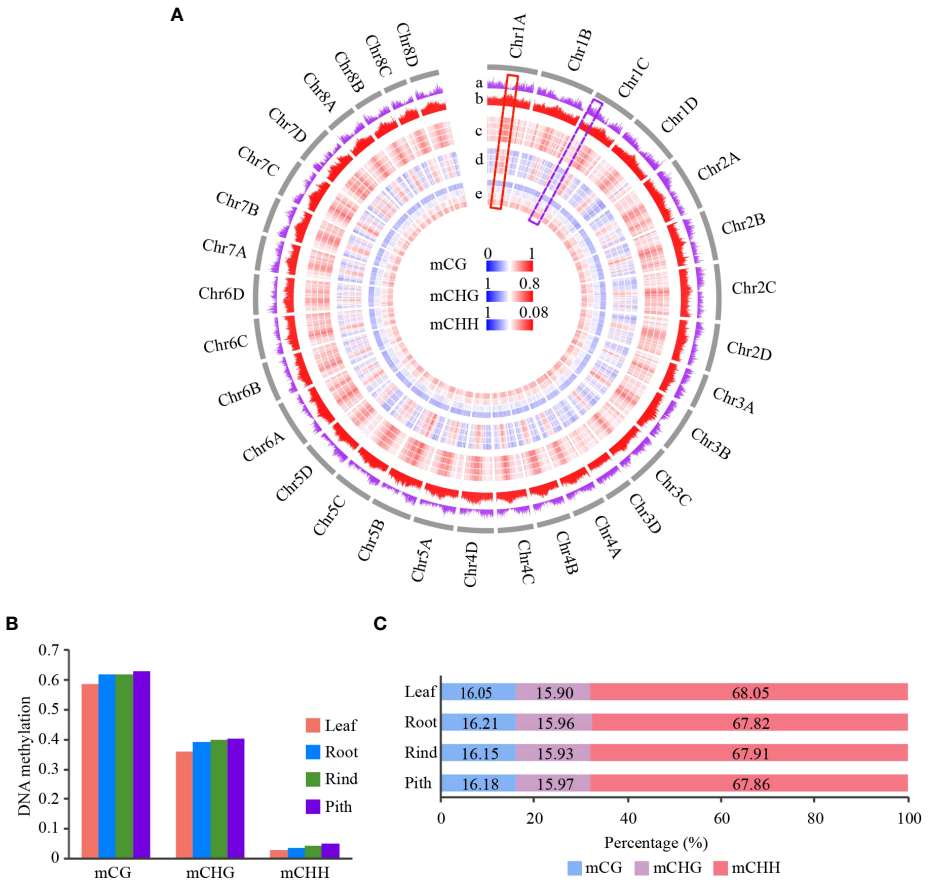


FIGURE 1
Genome-wide DNA methylation profile of different tissues in sugarcane. **(A)** Circle plot of gene and TE densities and methylation level of CG, CHG, and CHH across 32 homologous chromosomes in sugarcane. DNA methylation level is represented in a heatmap; blue and red indicate low and high methylation levels, respectively. Gene and TE density are represented in a histogram. Average DNA methylation level and gene/TE density are calculated using a 500-kb window. The gray circle indicates chromosomes. From the outer circle to the inner circle: a, gene density; b, TE density (TE density is the ratio of TE length to window length); c, CG methylation; d, CHG methylation; e, CHH methylation. For the DNA methylation circle, the order from outer to inner is Leaf, Root, Rind, and Pith. **(B)** Global average DNA methylation levels of CG, CHG, and CHH across different tissues in sugarcane. **(C)** Relative proportion of methyl-cytosines in the three sequence contexts across different tissues.

sugarcane tissues. Next, we analyzed DNA methylation levels in the gene and transposon regions of the four tissues. The results of the meta-analysis of gene and TE regions were consistent with those of the genome-wide methylation analysis, i.e., leaf and pith tissues showed the lowest and highest methylation levels, respectively (Figures 2A, B). This trend was also consistent between the gene body and TE regions (Figure 2). Strikingly, gene body regions showed relatively high CHG and CHH methylation levels, in addition to dense CG methylation (Figure 2A), differing from many other plant species, such as Arabidopsis and rice (Cokus et al., 2008; Li et al., 2012). In the sugarcane genome, more than 58.7% of the sequences consisted of repetitive elements (Zhang et al., 2018b), such as TEs, and 42.3% of protein-coding genes contained TE sequences in the gene body regions, particularly in intron regions (Figure S7A). After excluding genes with intronic TE insertions, we found that the methylation levels of gene body regions were notably reduced in all three sequence contexts, especially non-CG methylation levels. However, methylation levels of the flanking regions were slightly reduced (Figures S7B, C). This result suggested that most of the non-CG methylation of gene body regions was determined by intronic TE insertions, which have been reported in maize and other plant genomes with abundant TEs (Wang et al., 2015; Wang et al., 2021; Niu et al., 2022).

Next, we compared DNA methylation between different types of transposons, including Class I and II transposons. Class I transposons showed higher levels of CG and CHG methylation than Class II transposons, both in the transposon body and flanking regions. However, CHH methylation was higher in Class II transposons than that in Class I transposons (Figure 2C). Long terminal repeat (LTR)-type transposons mainly include Gypsy and Copia LTRs, whereas DNA

transposons contain several types of transposons (Figure S8A). The Gypsy and Copia LTRs showed very similar DNA methylation patterns (Figures S8B–E). However, different types of DNA transposons exhibit substantially different methylation patterns. For example, the CHH methylation level of the PIF-Harbinger transposon was significantly higher than that of the other types of transposons (Figure S9).

Active demethylase is associated with reduced DNA methylation among different tissues

DNA methylation levels are dynamically regulated by DNA methylases and demethylases. The decrease in DNA methylation levels can be attributed to the low expression of DNA methylase or high demethylase expression. To investigate the changes in DNA methylation levels among the four tissues, we searched for and annotated the DNA methylase and demethylase genes in the sugarcane genome (Table S2). As the sugarcane genome was assembled and annotated into four sets of homologous chromosomes, we identified more homologous genes in the sugarcane genome than in Arabidopsis and other plants. We first examined the expression levels of DNA methylase across the four tissues, and we did not observe a gradual increase in expression levels of these genes from the leaves to the roots and stem (rinds and piths) (Figures 3A, B). In addition, we found that only a few genes were differentially expressed in the RdDM pathway (Figures 3B). These results suggest that increased DNA methylation levels from the leaves to the piths were not attributed to the increased expression of DNA methylases and genes involved in the RdDM pathway.

We also examined the expression of putative DNA demethylase genes. Consistent with the changes in DNA methylation levels across the four tissues, we found that several demethylated genes, such as *ROS* and *DME*, were expressed at lower levels in pith tissue than in other tissues (Figure 3C). This result suggests that the DNA demethylation pathway plays a critical role in methylation level changes across the four tissues.

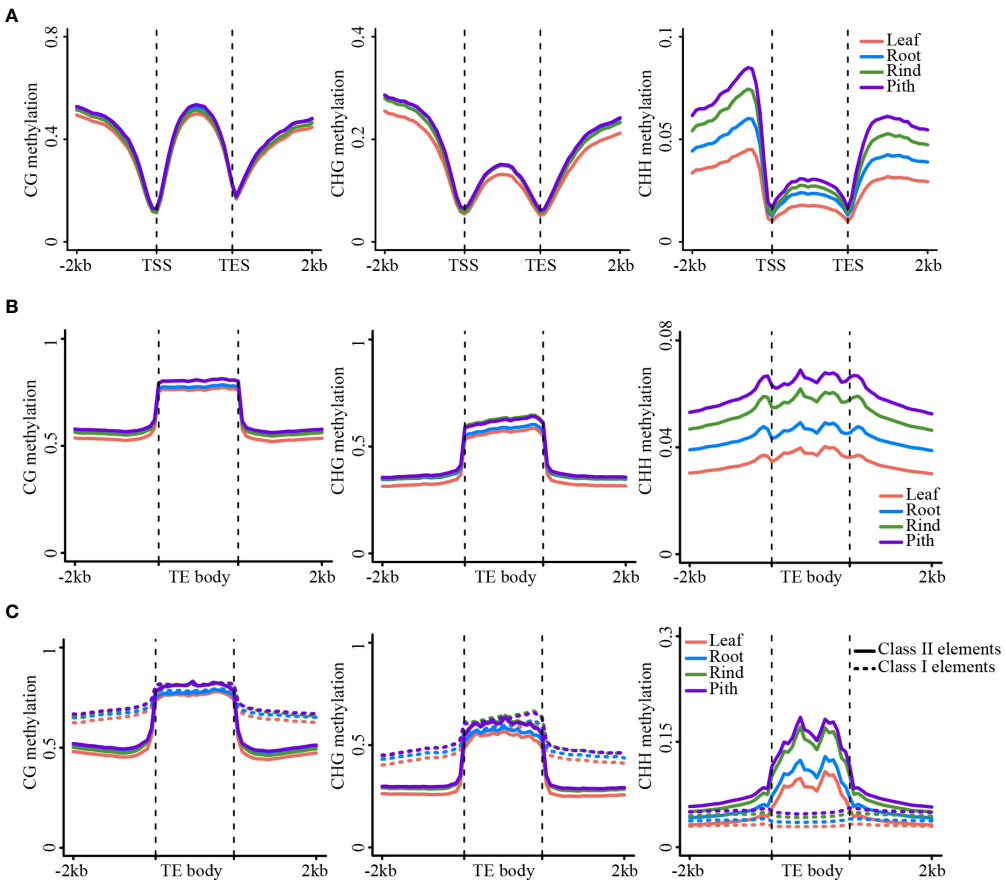


FIGURE 2
DNA methylation patterns in gene/TE and flanking regions. **(A)** The metaplot of the gene body and the flanking region. **(B)** The metaplot of TE and the flanking region. **(C)** The metaplot of the Class II TE and Class I TE. CG (left), CHG (middle), CHH (right).

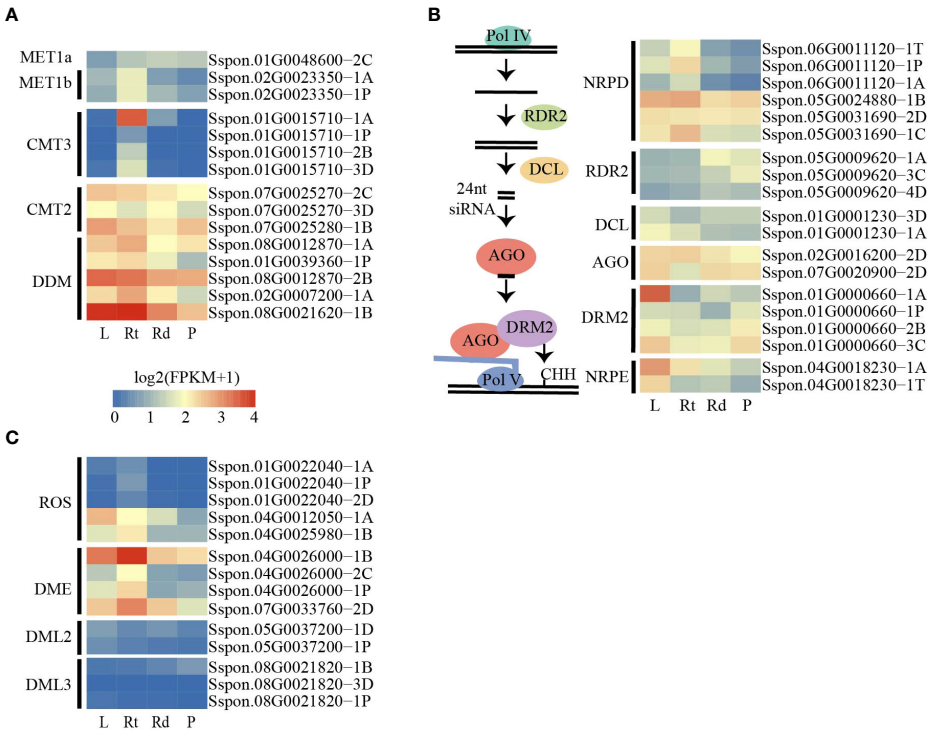


FIGURE 3
DNA methylated and demethylated genes are active in sugarcane tissues. **(A)** Heatmap showing the expression patterns of methylation-related genes among tissues. **(B)** Schematic diagram of the canonical RdDM pathway (left). Heatmap showing the expression patterns of canonical RdDM pathway genes among tissues. **(C)** Heatmap showing the expression patterns of demethylated genes. L, leaf; Rt, root; Rd, rind; P, pith.

The association between DNA methylation and gene activity

Cumulative evidence has shown that methylation of the gene body and flanking regions is involved in regulating gene expression (Wang et al., 2015; Zhang et al., 2018a; Xu et al., 2018; Wang et al., 2019; Cai et al., 2021). CG methylation of gene body regions is always positively correlated with gene expression, whereas non-CG methylation of gene body regions negatively correlates with gene expression (Wang et al., 2015; Xu et al., 2018; Wang et al., 2019; Cai et al., 2021). In addition, recent studies have shown that CHH methylation of promoter regions could promote adjacent gene expression (Gent et al., 2013; Xu et al., 2018; Cai et al., 2021). To explore the relationship between DNA methylation and gene expression in sugarcane, we first performed RNA-seq of the same tissues for DNA methylation analysis. We found that approximately 75% of clean reads were aligned to the sugarcane genome, and the Pearson correlation coefficients between different biological replicates of RNA-seq were between 0.88 to 0.96 (Table S3), indicating the high reproducibility of our RNA-seq data. All genes were divided into expressed (FPKM ≥ 1, 38,750 genes) and unexpressed (FPKM < 1, 45,976 genes) groups and their methylation levels were calculated separately (Figure 4A). Compared with unexpressed genes, CG gene body methylation levels of expressed genes were higher than those of unexpressed genes, whereas non-CG methylation was lower in expressed gene body regions than that in unexpressed genes. Consistent with previous studies (Xu et al., 2018; Wang et al., 2019; Cai et al., 2021), DNA methylation levels at the transcription start site (TSS) and transcription end site (TES) were significantly reduced in expressed genes compared with those in unexpressed genes. Additionally, a significant increase in CHH promoter methylation was observed in the expressed genes. A similar phenomenon was observed in the other three tissues (Figure S10).

Next, all expressed genes were divided into four groups according to their expression levels from low to high [FPKM = 0 (Cluster1), 0 < FPKM ≤ 2 (Cluster 2), 2 < FPKM ≤ 10 (Cluster 3), and FPKM > 10 (Cluster 4)], and the methylation level of each group of genes was calculated (Figures 4B and Figure S11). For all three methylation contexts, the methylation levels near TSS and TES sites decreased as the expression level increased, and the methylation level was lowest when the expression was highest. In the gene body regions, CG methylation was positively correlated with gene expression, and the genes with the medium high

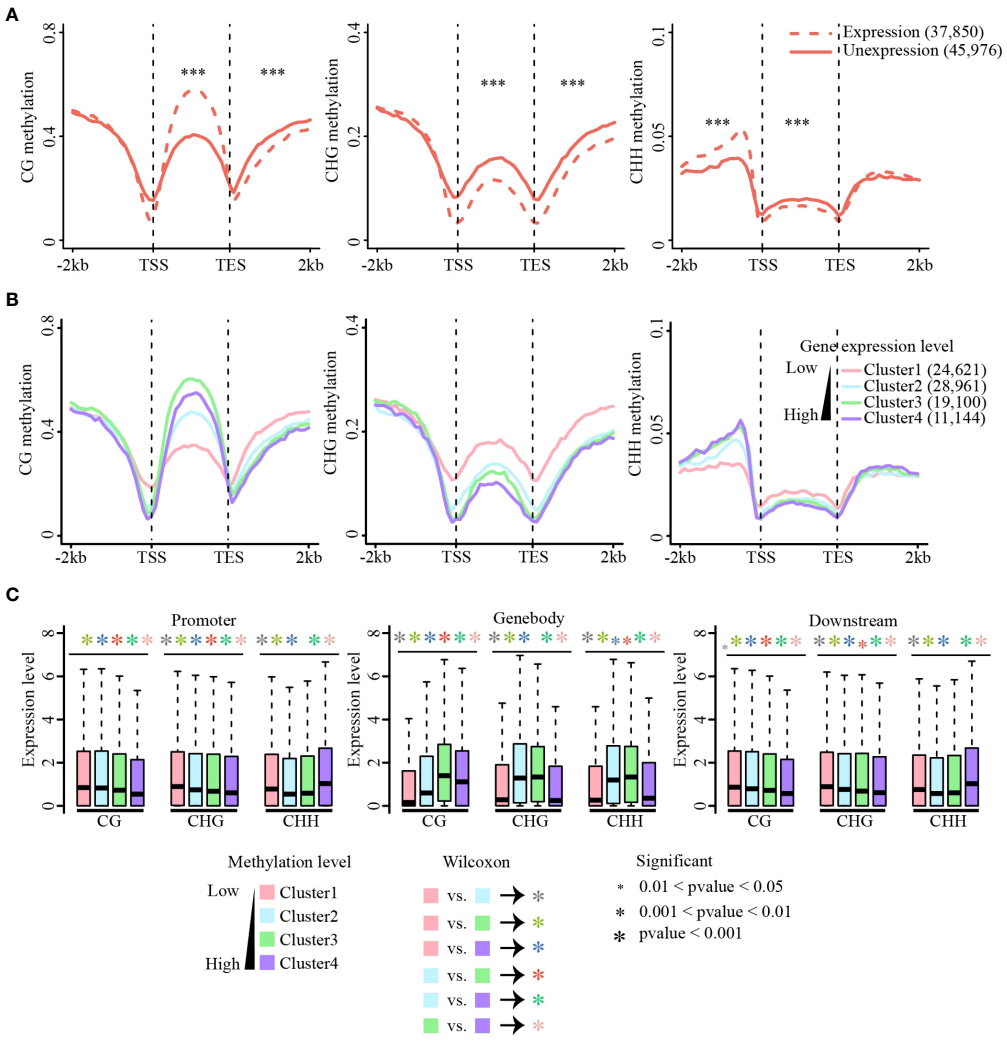


FIGURE 4
Association of methylation and gene expression. **(A)** Methylation level changes between the expressed and unexpressed genes in CG, CHG, and CHH sequence contexts. Methylation differences between expressed and unexpressed genes were tested using the Wilcoxon rank-sum test. ***p-value < 0.001. **(B)** Correlations between methylation levels (CG, CHG, and CHH) and gene expression across gene body and flanking regions. Methylation level of each gene group [Cluster1 (FPKM = 0), Cluster2 (0 < FPKM ≤ 2), Cluster3 (2 < FPKM ≤ 10), Cluster4 (FPKM ≥ 10)] were calculated. CG (left), CHG (middle), and CHH (right). **(C)** Expression levels of methylated genes in the gene body and flanking regions. Genes were divided into four quartiles based on methylation levels, from the first quartile (the lowliest methylated 25% of genes) to the fourth quartile (the most highly methylated 25% of genes). Expression differences between different clusters were tested using the Wilcoxon rank-sum test. The colors of the asterisk (*) represent the comparison between different clusters, and the size of the asterisk (*) indicates the criterion of significance. **(A–C)** are data in leaf tissue.

expression showed the highest CG methylation. This is consistent with the phenomenon observed in studies on many other plants (Wang et al., 2015; Xu et al., 2018; Wang et al., 2019). Non-CG methylation levels were significantly negatively correlated with gene expression, except for CHH methylation in the promoter regions (Figure 4B).

To further examine the relationship between gene expression and DNA methylation in different contexts (CG, CHG, and CHH) and genic regions (i.e., upstream, gene body, and downstream regions), we sorted all genes according to their methylation levels from low to high and divided them into four equal groups (Clusters 1 to 4). Consistent with the above analysis, CG methylation of gene body regions promoted gene expression, but CG methylation at either the upstream or downstream regions always inhibited gene expression. CHG and CHH methylation mostly repressed gene expression, except for upstream CHH methylation (Figures 4C and Figure S12). Collectively, this relationship between DNA methylation and gene expression is conserved in most of the studied plant species. Our findings indicate that DNA methylation is involved

in gene expression regulation and DNA methylation of different genic regions and sequence contexts plays different roles in gene expression.

Extensive changes in gene expression among different tissues in sugarcane

From the above analysis, we found that DNA methylation levels are associated with gene expression in sugarcane. For example, DNA methylation at different genic regions or sequence contexts affects gene expression differently (Figure 4). To further explore gene expression changes across different sugarcane tissues, we examined the expression dynamics across different tissues (leaf, root, rind, and pith) and identified 21,460 DEGs between different tissues (Figure S13). To search for functional signatures of different tissues, we performed GO enrichment analysis to characterize DEGs from the comparisons between different tissues. We found that upregulated genes in leaves were enriched in photosynthesis and monosaccharide catabolic processes. However, upregulated genes in roots were enriched in response to abiotic and biotic stimuli; upregulated genes in the rind were enriched in pathways related to transport, such as intercellular and carbohydrate transport. Compared with leaves and roots, upregulated genes in the pith were involved in carbohydrate transport and organic substance metabolic and biosynthetic processes. Additionally, upregulated genes in the pith relative to those in the rind were enriched in terms associated with fructose export from the vacuole to the cytoplasm, regulation of the syringal lignin biosynthetic process, plant-type cell wall organization or biogenesis. These results confirm that DEGs from different tissues are involved in biological pathways related to tissue-specific physiological functions.

Identification of differentially methylated regions among different tissues

To characterize methylation changes among different tissues in sugarcane, we defined DMRs in each sequence context according to the method of Akalin (Akalin et al., 2012). A total of 113,536 CG-DMRs, 396,224 CHG-DMRs, and 1,146,516 CHH-DMRs were identified. Compared with CG and CHG DMRs, CHH DMRs were the most abundant across different comparisons among the four tissues. Meanwhile, compared with hypo-DMRs (lower DNA methylation in the right comparison), hyper-DMRs (higher methylation in the left comparison) were dominant (Figure 5A), consistent with the increased DNA methylation levels from leaf to root and then to rind and pith in the above analysis. Next, we examined the distribution of DMRs in different genomic features such as TE, intergenic, upstream, downstream, intron, and exon regions. As shown in Figure 5B, TE, intergenic, and genetic (upstream, downstream, intron, and exon) regions account for 53.52%, 27.45%, and 19.31% of the sugarcane genome, respectively. We found that CG DMRs are mainly located in the intergenic regions; Non-CG DMRs are mainly enriched in the TE regions, especially CHH methylation. This may indicate that CHH methylation changes mainly occur in the TE and intergenic regions (Figure 5B). Moreover, we found that many DMRs (~20%) were located in genic regions, including upstream, exon, intron, and downstream regions. Therefore, we hypothesized that DMRs adjacent to the gene regions might affect gene expression.

Differential expression genes are associated with differentially methylated regions

We found substantial differences in gene expression and DNA methylation levels across different sugarcane tissues. In particular, many DMRs occur in the gene body and/or proximal regions, and these DMRs might contribute to changes in the expression of adjacent genes. From the above analysis, we found a large number of DMRs, including hyper- and hypo-methylated DMRs, in the gene body and flanking regions. Except for CG-DMRs, both CHG and CHH DMRs showed distinct distributions of hyper- and hypo-DMRs across the gene regions (Figures 5C and Figure S14). Strikingly, we observed that DMR-overlapped genes were more likely to be differentially expressed than DMR-non-overlapping genes, which was consistent across the comparisons between tissues (Table S4). These results indicate that changes in DNA methylation are associated with DEGs.

More than 40% of the DEGs contained DMRs across all six comparisons of the four tissues (Figure 5D). To understand how DMR-associated genes were associated with tissue divergence, we performed GO enrichment analysis of DMR-associated up- and down-regulated DEGs. Compared with the other three tissues, DMR-associated highly expressed genes in the roots were mainly involved in response to stress and root morphogenesis (Figure S15). For example, DMR-associated genes encoding phosphoinositide-specific phospholipase C (PI-PLC, *Sspon.05G0021570-2P*), class III peroxidase (PRX, *Sspon.01G0012950-1A*), and MYB (*Sspon.01G0019490-1A*) were highly expressed in roots, and their homologous genes in Arabidopsis were involved in growth, response to stresses, and lignin synthesis (Meijer and Munnik, 2003; Shigeto and Tsutsumi, 2016; Chezem et al., 2017). We also found that highly expressed DMR-associated DEGs in the leaves were significantly enriched in photosynthesis and sucrose-related pathways (Figure S16), such as photosynthesis, pigment metabolic process, and sucrose biosynthetic process. Furthermore, many biological processes related to sugar biosynthesis and metabolism were enriched in the DMR-

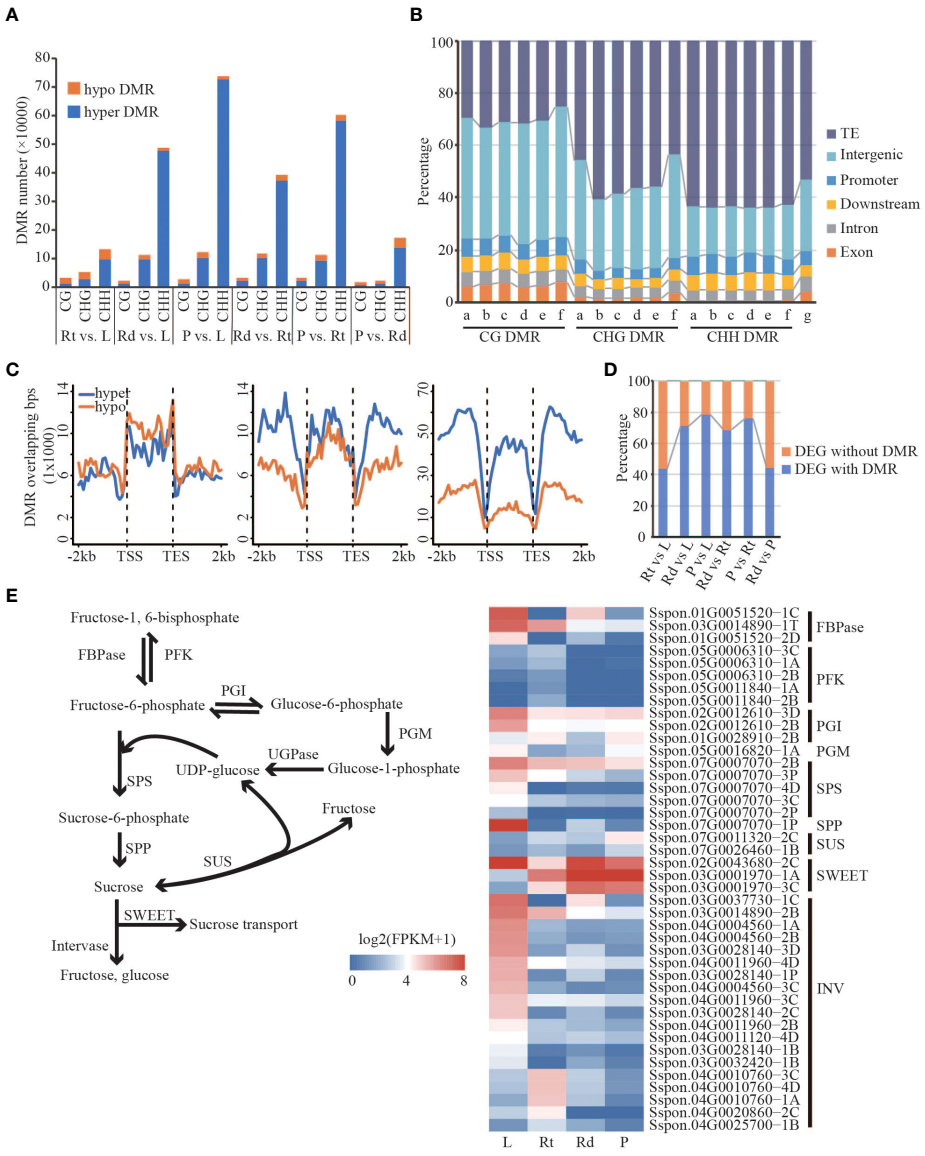


FIGURE 5
Differentially expressed genes are associated with differential methylation. **(A)** Barplot of hyper/hypo DMR. **(B)** The distribution of DMR in different regions of the genome. a, Rt vs. L; b, Rd vs. L; c, P vs. L; d, Rd vs. Rt; e, P vs. L; f, P vs. Rd; g, genome. **(C)** Distribution of DMR in the gene body and flanking region (Rt vs. L). **(D)** The proportion of DEG with DMR/without DMR. **(E)** Sucrose synthesis and hydrolysis pathways and the expression pattern of DMR-DEGs related to sucrose synthesis and hydrolysis pathways. FBPase, fructose-1,6-bisphosphatase; PFK, phosphofructokinase; PGI, phosphoglucose; PGM, phosphoglucomutase; SPS, sucrose phosphate synthase; SPP, sucrose-6F-phosphate phosphohydrolase; SUS, sucrose synthase; SWEET, sugars will eventually be exported transporters; INV, invertase; L, leaf; Rt, root; Rd, rind; P, pith; DEG, differentially expressed gene; DMR, differentially methylated region.

associated DEGs (Figure S16). For example, *Sspon.02G0012860-2B* (Figure S16), which encodes NAD oxidoreductase, was upregulated in leaves. A recent study showed that NAD oxidoreductase was functional downstream of the photosynthetic electron transport chain and participated in the Calvin cycle, pigment synthesis (Pierella Karlusich and Carrillo, 2017), and is a key enzyme linking the light reaction of photosynthesis to carbon metabolism. Gene encoding inorganic pyrophosphatase (PPi, *Sspon.04G0005360-3D*) was highly expressed in leaves and its homologous genes in Arabidopsis are key enzymes in sucrose synthesis (Farré et al., 2000). Unlike leaves and roots, upregulated DMR-associated genes in the stem (rind and pith) were enriched in transport-related pathways such as sucrose and intracellular transport, cellular metabolic process, and hydrocarbon metabolic process. For example, *Sspon.04G0012730-4D* (Sugars Will Eventually be

Exported Transporters; SWEET), *Sspon.01G005290-1A* (polyol/monosaccharide transporter 5), and *Sspon.01G0026170-1A* (Mfs transporter) encoding sugar transporters (Figures S15 and Figure S16) were upregulated DMR-associated genes in the rinds. Sugar transporters function in sugar transport, distribution, and utilization in the phloem, as well as maintaining the balance between source and sink (Julius et al., 2017). *Sspon.02G0017170-1A* and *Sspon.02G0017170-3D* (Figure S15) were highly expressed DMR-associated genes in the piths encoding ADP-glucose pyrophosphorylase (AGPase); their homologous genes in Arabidopsis catalyze ADP glucose synthesis and release pyrophosphate, and are the key enzymes determining starch synthesis (Tetlow et al., 2004). *Sspon.02G0019390-3C* (Figure S15), encoding phosphoglucomutase, was upregulated in the DMR-associated genes in piths. In Arabidopsis, its homologous gene catalyzes the mutual conversion of glucose-1-phosphate and glucose-6-phosphate, key steps in sucrose metabolism and synthesis (Streb et al., 2009). We found that genes with many DMRs were highly expressed in the stem. In conclusion, DMR-associated DEGs in different sugarcane tissues are involved in essential biological pathways and have tissue-specific physiological functions that are closely related to photosynthesis, sugar metabolism, growth, and sugarcane development.

High sucrose accumulation is a characteristic feature of sugarcane. We found that DMR-associated DEGs were enriched in essential biological pathways (Figures S15, S16), such as sucrose synthesis, carbohydrate metabolism, and stress response. To investigate how DMR-associated genes contribute to the regulation of sucrose accumulation, we focused on the sucrose synthesis and hydrolysis pathways (Figure 5E). We observed that genes involved in the sucrose synthesis pathway, including *FBPase*, *PGI*, *SPS*, and *SPP*, were highly expressed in the leaves. However, in contrast to the other three tissues, genes encoding sucrose synthase (SUS) showed lower expression in leaves. These results suggest that sucrose synthesis in leaves mainly depends on the SPS-mediated pathway, consistent with previous studies (Buczynski et al., 1993; Verma et al., 2011). Moreover, *SWEETs* involved in sucrose transport were highly expressed in the leaves and stems, suggesting that the remaining sucrose was transported into sink tissues for consumption and storage, except for consumption in the leaves. Interestingly, we found that all *invertases* (INVs) involved in sucrose hydrolysis (Figure 5E) had a lower expression in stem tissue (rind and pith) than that in leaf and root tissue, indicating that sucrose transported to the stem was mainly used for storage, confirming our suspicion. Taken together, efficient sucrose synthesis in leaves, intense sucrose transport from leaves to stem, and low INV activity in the stem might be responsible for the high sucrose accumulation in sugarcane, indicating that DNA methylation-regulated genes function in high sucrose accumulation in sugarcane.

Transcription factor genes are enriched in sugarcane DNA methylation valleys

Recent studies have shown that there are always lowly methylated or unmethylated regions in the genome, also known as DNA methylation valleys (DMVs) (Stadler et al., 2011; Lin et al., 2017; Chen et al., 2018; Li et al., 2018; Crisp et al., 2020). During soybean seed development, genes contained in DMVs tend to be enriched in tissue-specific biological pathways such as protein storage and fatty acid metabolism (Lin et al., 2017; Chen et al., 2018). Next, we scanned DNA methylome data from the four tissues for regions with <5% bulk methylation in all three cytosine sequence contexts as described in (Chen et al., 2018) and identified 28,531, 26,445, 25,311, and 24,666 DMVs in leaves, roots, rinds, and piths, respectively. Among these DMVs, 17,208 (2.9%), which were hypomethylated, were common to all four tissues and did not change significantly across different tissues. There were 8,704 non-redundant DMVs, accounting for 1.8% (51.7 Mb) of the genome length, which was significantly lower than the DMV ratio in other plants, implying species-specific DMV distribution (Figure 6A). For example, a 6.3-kb DMV exhibited low levels of all DNA methylation contexts across the four tissues and contained two protein-coding genes (Figure 6B).

We further examined the DMV regions and identified DMV genes if the gene body or flanking 1-kb regions overlapped with the DMV. We identified 1,734 genes located in DMVs, and transcription factors (TFs) (13.1%) were significantly enriched in these DMV genes ($p < 2.2e\text{-}16$, chi-squared test) (Figure 6C). GO enrichment analysis showed that these DMV genes were involved in regulating gene expression, developmental, stimulus-related, and saccharide-related processes (Figure 6D). In addition, we found that many TFs played important roles in these processes. For example, 488 DMV genes were associated with sucrose metabolism, of which 136 (28%) were TF encoding genes. Notably, many of these DMV TF genes were significantly differentially expressed in the four tissues (Figure S17). For example, the *bZIP* TF gene, *Sspon.04G0028570-1P*, was highly expressed in root tissue relative to the other tissues, and its homologous genes play an important role in biotic and abiotic stress in Arabidopsis (Droge-Laser et al., 2018) (Figure 6E). Moreover, *Sspon.06G0007540-2C* (Figure S17) encoding bZIP2 was higher in the piths than in the other three tissues, and the co-expression of its homologous genes *AtbZIP2* and *KIN10* in Arabidopsis activates *DIN6-LUC* to inhibit respiration (Baena-Gonzalez et al., 2007), suggesting that *Sspon.06G0007540-2C* might inhibit cellular respiration in the piths, reducing the consumption of sugar and facilitating sugar accumulation in the piths. In addition, some genes (non-TFs) located in the DMV regions were involved in the saccharide pathway. For example, *Sspon.03G0028140-3D* (Figure 6F) encoding SPS, a key gene regulating the conversion of photosynthetic products into sucrose and starch, was highly expressed in leaves (Verma

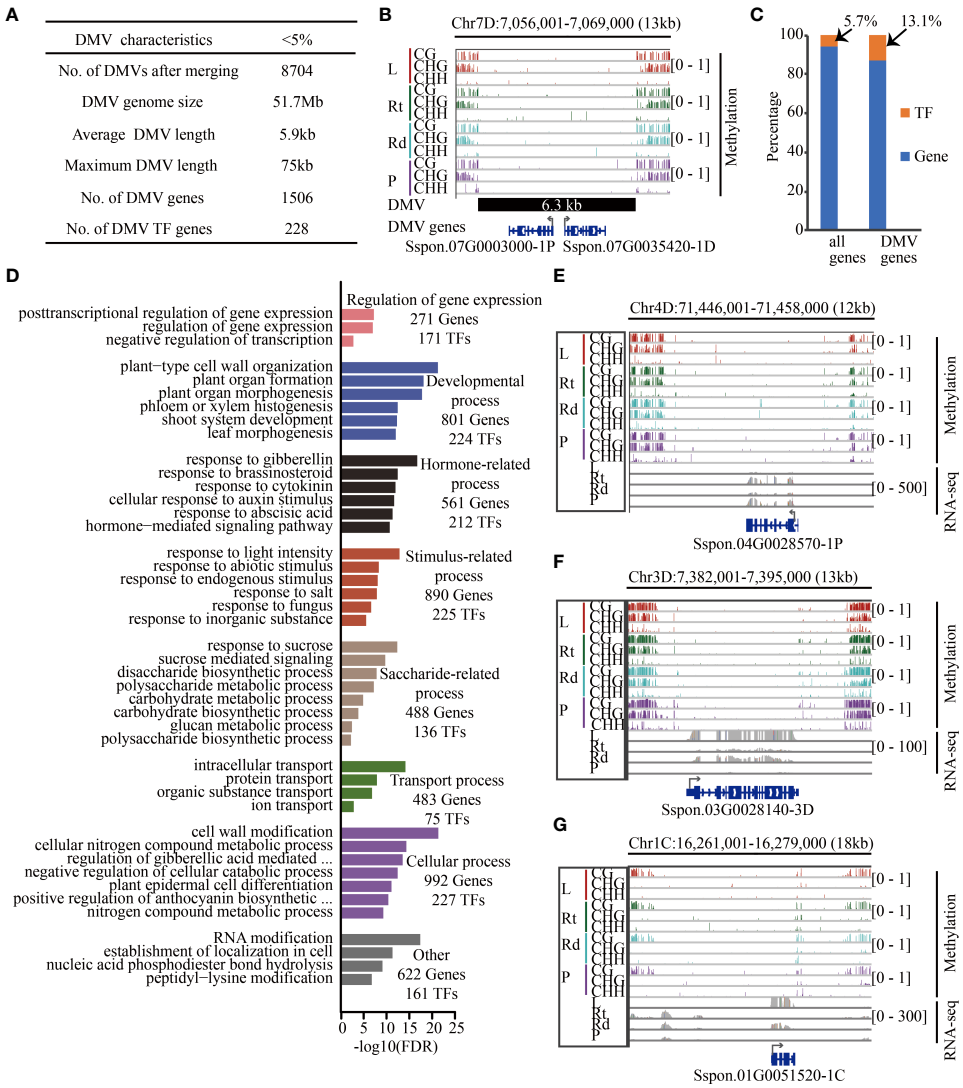


FIGURE 6
Transcription factors are enriched in sugarcane DMVs. **(A)** Summary of sugarcane DMV characteristics. **(B)** IGV of an 18-kb DMV located on chromosome 7D. Genes in blue color (Sspon.07G0003000-1P and Sspon.07G0035420-1D encoding the pectin lyase-like superfamily are involved in carbohydrate metabolic process) are located within this DMV, including 1 kb of 5′ and 3′ flanking regions. **(C)** Proportion of TF in genome and DMV regions. **(D)** Enriched GO terms with an FDR < 0.05. **(E–G)** Methylome and RNA-Seq genome browser views of three genes between at least two tissues. Red, green, cyan, and purple bars indicate leaf, root, rind, and pith, respectively. Gray collapsed bars indicate expression level. L, leaf; Rt, root; Rd, rind; P, pith; DMV, DNA methylation valley; TF, transcription factor; IGV, integrative genomics viewer.

et al., 2011). Furthermore, *Sspon.01G0051520-1C* (Figure 6G), which encodes FBPase involved in sucrose synthesis, was highly expressed in leaves—decreased *FBPase* expression inhibits sucrose synthesis (Strand et al., 2000; Lee et al., 2008). Therefore, *Sspon.03G0028140-3D* (Figure 6F) and *Sspon.01G0051520-1C* (Figure 6G) were highly expressed in leaves, suggesting that they played a role in transforming photosynthetic products and sucrose synthesis. Taken together, these data show that TFs and genes located in DMVs play essential roles in sugarcane development, stress response, and sucrose synthesis.

Discussion

Publication of the sugarcane genome provided us with an unprecedented opportunity to investigate the role of DNA methylation in sugarcane. In the present study, we analyzed the dynamics of DNA methylation among tissues in sugarcane and the relationship between DNA methylation and gene expression, which will enhance knowledge in sugarcane epigenetics. DNA methylation levels are dynamically regulated by DNA methylases and demethylases (Law and Jacobsen, 2010; Zhang et al., 2018a). We observed that DNA methylation levels

differed among the tissues (Figure 1). Furthermore, as shown in Figure 3, the expression patterns of *DME* and *ROS* in tissues are consistent with those of *MET1b* and *CMT3*, and the expression pattern of *Sspon.04G0012050-1A* (*ROS*) is consistent with those of *CMT2* and *DRM2* (*Sspon.01G0000660-1A*). Based on the fact that DNA demethylases can eliminate the mC of all sequence contexts (Choi et al., 2002; Gong et al., 2002; Morales-Ruiz et al., 2006; Ortega-Galisteo et al., 2008), we suggested that the DNA demethylation pathway plays a critical role in changes in methylation levels across the four tissues.

Cumulative evidence has shown that methylation of the gene body and flanking regions is involved in regulating gene expression (Wang et al., 2015; Xu et al., 2018; Zhang et al., 2018a; Wang et al., 2019; Cai et al., 2021). In tea plant (*Camellia sinensis*) (Tong et al., 2021), the methylation levels in all three sequence contexts of unexpressed genes were higher than those of expressed genes, the methylation levels of CHH in flanking regions of unexpressed genes were lower. However, CHH methylation patterns in sugarcane gene body and upstream regions were consistent with tea plant, whereas CG methylation pattern in gene body was opposite to that in the tea plant (Tong et al., 2021); the difference in CG methylation patterns in gene body between sugarcane and tea plant may be related to species specificity, such as genome size, and TE content. Moreover, CG methylation of gene body regions is always positively correlated with gene expression (Wang et al., 2015; Xu et al., 2018; Wang et al., 2019; Cai et al., 2021), but we observed that the highest-expressed genes did not have the highest CG methylation levels in the gene body (Figures 4 and Figure S11). Methylation of the gene body can quantitatively impede transcript elongation in Arabidopsis (Zilberman et al., 2007). This may lead to the highest expression of genes without the highest CG methylation levels in the gene body. CG, CHG, and CHH methylation levels near the TSS were negatively correlated with gene expression (Figures 4B and Figure S11), similar to results for rice, soybean, apple (*Malus*), tea (*Camellia sinensis*), wild barley (*Hordeum vulgare*), Arabidopsis, and human (Zilberman et al., 2007; Laurent et al., 2010; Li et al., 2012; Song et al., 2013; Xu et al., 2018; Wang et al., 2019; Cai et al., 2021), demonstrating that methylation near the TSS is a common mechanism to suppress gene expression in eukaryotes. Additionally, highly expressed genes were correlated with higher CHH methylation levels in the promoter region (200–2,000 bp) close to the TSS (Figures 4B and Figure S11); a similar observation was made in soybean, maize, apple, and wild barley (Gent et al., 2013; Song et al., 2013; Xu et al., 2018; Cai et al., 2021). Taken together, the relationship between DNA methylation and gene expression is conserved in most of the studied plant species.

Genes regulated by DNA methylation are involved in several important biological pathways (Wang et al., 2015; Cheng et al., 2018; Wang et al., 2018; Xu et al., 2018; Wang et al., 2019). For example, highly expressed genes affected by DNA methylation in cassava are involved in carbohydrate metabolism, including hexose and glucose metabolism (Wang et al., 2015). Furthermore, upregulated genes regulated by DNA methylation during strawberry (*Fragaria* × *ananassa*) ripening are involved in fruit ripening-related processes, such as cytokinin and abscisic acid biosynthesis (Cheng et al., 2018). We found that DMR-DEGs in sugarcane were significantly enriched in biological pathways of tissue-specific physiological functions. For example, DMR-associated DEGs with higher expression in roots were significantly enriched in stress response and root morphogenesis (Figures S15, S16). Genes upregulated in leaves regulated by DNA methylation were involved in photosynthesis, hydrocarbon biosynthesis, and metabolic processes (Figure S16). DMR-associated DEGs that were highly expressed in the stem (rind and pith) were significantly enriched in transport-related pathways and metabolism-related processes, such as sucrose transport and hydrocarbon metabolic process (Figures S15, S16). In conclusion, DMR-associated DEGs between different tissues are involved in the biological pathways of tissue-specific physiological functions, which are essential for plant growth and development.

We observed that DMR-associated DEGs were enriched in important biological pathways (Figures S15, S16), such as sucrose synthesis, carbohydrate metabolism, and stress response. The high sucrose accumulation in sugarcane has attracted our attention to sucrose synthesis and hydrolysis pathways. As shown in Figure 5E, sucrose in leaves is mainly derived from the SPS-mediated sucrose synthesis pathway, and genes involved in sucrose transport are more highly expressed in stems than in leaves and roots. Moreover, INV involved in sucrose hydrolysis showed lower expression in the stem. Sugarcane has a universal source-sink system; except for consumption during leaf growth, the sucrose synthesized in leaves is exported to sink tissues and used for consumption and storage (Buczynski et al., 1993; Verma et al., 2011; Julius et al., 2017). Previous studies have indicated that SPS activity is a biochemical marker of high sucrose content in sugarcane (Verma et al., 2011). Collectively, we suggest that efficient sucrose synthesis in the leaves, intense sucrose transport to the stem, and low INV activity in the stem may be responsible for the high sucrose accumulation in sugarcane, indicating that DNA methylation plays an important role in sucrose accumulation in sugarcane.

Recent studies have shown that lowly methylated and unmethylated regions contain functional regulatory elements (Stadler et al., 2011; Lin et al., 2017; Chen et al., 2018; Li et al., 2018; Crisp et al., 2020). For instance, genes located in the DMVs of human embryonic stem cells or vertebrates, such as *Foxa1*, *Wnt1*, *GATA*, and *SOX2* (Stadler et al., 2011; Xie et al., 2013; Li

et al., 2018), are involved in development and TF activity. DMVs during seed formation are enriched in TFs and development-related genes such as *WOX*, *PLETHORA*, *PIN1*, and *YUCCA4* (Lin et al., 2017; Chen et al., 2018). We also found many DMVs in sugarcane, which always overlapped with TFs, development, and sucrose-related genes such as *WRKY*, *bZIP*, *WOX*, *SPS*, and *FBPase* (Figures 6D–G and Figure S17), which function in sugarcane growth, morphogenesis, stress response, and carbohydrate metabolism, indicating that DMVs are common and essential for growth and development. Furthermore, approximately 40% of the genes (670 genes) located in the DMVs were differentially expressed between at least two tissues. Recent studies have shown that genes located in DMVs are enriched in H3K27me3 and H3K4me3 (Xie et al., 2013; Chen et al., 2018). Therefore, we hypothesized that DEGs located in sugarcane DMVs might be regulated by histone modification and TF regulation.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: NCBI, PRJNA730638.

Author contributions

HW and BC conceived the study and supervised all parts of the project. CZo and YX collected samples and performed sequencing. YX, HY performed DNA methylation. YX and CZh performed transcriptome analysis and comparative analysis. YX and HW wrote the manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (No. 32160142) and Sugarcane Research Foundation of Guangxi University (Grant No. 2022GZA002) to HW and BC is supported by grant from Department of Science and Technology of Guangxi Zhuang Autonomous Region (AD17129002). YX is supported by Innovation Project of Guangxi Graduate Education (YCBZ2021005).

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.2022.1036764/full#supplementary-material

References

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Melnick, A., et al. (2012). methylKit: A comprehensive r package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.* 13, R87. doi: 10.1186/gb-2012-13-10-r87

Ausin, I., Feng, S., Yu, C., Liu, W., Kuo, H. Y., Jacobsen, E. L., et al. (2016). DNA Methylome of the 20-gigabase Norway spruce genome. *Proc. Natl. Acad. Sci. U.S.A.* 113, E8106–E8113. doi: 10.1073/pnas.1618019113

Baena-Gonzalez, E., Rolland, F., Thevelein, J. M., and Sheen, J. (2007). A central integrator of transcription networks in plant stress and energy signalling. *Nature* 448, 938–942. doi: 10.1038/nature06069

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

Buczynski, S. R., Thom, M., Chourey, P., and Maretzki, A. (1993). Tissue distribution and characterization of sucrose synthase isozymes in sugarcane. *J. Plant Physiol.* 142, 641–646. doi: 10.1016/S0176-1617(11)80895-3

Cai, S., Shen, Q., Huang, Y., Han, Z., Wu, D., Chen, Z. H., et al. (2021). Multi-omics analysis reveals the mechanism underlying the edaphic adaptation in wild barley at evolution slope (Tabigha). *Adv. Sci. (Weinh)* 8, e2101374. doi: 10.1002/advs.202101374

Chang, Y. N., Zhu, C., Jiang, J., Zhang, H., Zhu, J. K., and Duan, C. G. (2020). Epigenetic regulation in plant abiotic stress responses. *J. Integr. Plant Biol.* 62, 563–580. doi: 10.1111/jipb.12901

Chan, S. W., Henderson, I. R., Zhang, X., Shah, G., Chien, J. S., and Jacobsen, S. E. (2006). RNAi, DRD1, and histone methylation actively target developmentally important non-CG DNA methylation in arabidopsis. *PloS Genet.* 2, e83. doi: 10.1371/journal.pgen.0020083

Cheng, J., Niu, Q., Zhang, B., Chen, K., Yang, R., Zhu, J. K., et al. (2018). Downregulation of RdDM during strawberry fruit ripening. *Genome Biol.* 19, 212. doi: 10.1186/s13059-018-1587-x

Chen, M., Lin, J.-Y., Hur, J., Pelletier, J. M., Baden, R., Pellegrini, M., et al. (2018). Seed genome hypomethylated regions are enriched in transcription factor genes. *Proc. Natl. Acad. Sci.* 115, E8315. doi: 10.1073/pnas.1811017115

Chezem, W. R., Memon, A., Li, F. S., Weng, J. K., and Clay, N. K. (2017). SG2-type R2R3-MYB transcription factor MYB15 controls defense-induced lignification and basal immunity in arabidopsis. *Plant Cell* 29, 1907–1926. doi: 10.1105/tpc.16.00954

Choi, Y., Gehring, M., Johnson, L., Hannon, M., Harada, J. J., Goldberg, R. B., et al. (2002). DEMETER, a DNA glycosylase domain protein, is required for endosperm gene imprinting and seed viability in arabidopsis. *Cell* 110, 33–42. doi: 10.1016/s0092-8674(02)00807-3

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., et al. (2008). Shotgun bisulphite sequencing of the arabidopsis genome reveals DNA methylation patterning. *Nature* 452, 215–219. doi: 10.1038/nature06745

Crisp, P. A., Marand, A. P., Noshay, J. M., Zhou, P., Lu, Z., Schmitz, R. J., et al. (2020). Stable unmethylated DNA demarcates expressed genes and their cis-regulatory space in plant genomes. *Proc. Natl. Acad. Sci. U.S.A.* 117, 23991–24000. doi: 10.1073/pnas.2010250117

Cuerda-Gil, D., and Slotkin, R. K. (2016). Non-canonical RNA-directed DNA methylation. *Nat. Plants* 2, 16163. doi: 10.1038/nplants.2016.163

Droge-Laser, W., Snoek, B. L., Snel, B., and Weiste, C. (2018). The arabidopsis bZIP transcription factor family-an update. *Curr. Opin. Plant Biol.* 45, 36–49. doi: 10.1016/j.pbi.2018.05.001

Farré, E. M., Geigenberger, P., Willmitzer, L., and Trethewey, R. N. (2000). A possible role for pyrophosphate in the coordination of cytosolic and plastidial carbon metabolism within the potato tuber. *Plant Physiol.* 123, 681–688. doi: 10.1104/pp.123.2.681

Gent, J. I., Ellis, N. A., Guo, L., Harkess, A. E., Yao, Y., Zhang, X., et al. (2013). CHH islands: *de novo* DNA methylation in near-gene chromatin regulation in maize. *Genome Res.* 23, 628–637. doi: 10.1101/gr.146985.112

Gong, Z., Morales-Ruiz, T., Ariza, R. R., Roldán-Arjona, T., David, L., and Zhu, J. K. (2002). ROS1, a repressor of transcriptional gene silencing in arabidopsis, encodes a DNA glycosylase/lyase. *Cell* 111, 803–814. doi: 10.1016/s0092-8674(02)01133-9

Gouil, Q., and Baulcombe, D. C. (2016). DNA Methylation signatures of the plant chromomethyltransferases. *PloS Genet.* 12, e1006526. doi: 10.1371/journal.pgen.1006526

Huang, X., Zhang, S., Li, K., Thimmapuram, J., Xie, S., and Wren, J. (2018). ViewBS: A powerful toolkit for visualization of high-throughput bisulfite sequencing data. *Bioinformatics* 34, 708–709. doi: 10.1093/bioinformatics/btx633

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., et al. (2019). eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 47, D309–d314. doi: 10.1093/nar/gky1085

Julius, B. T., Leach, K. A., Tran, T. M., Mertz, R. A., and Braun, D. M. (2017). Sugar transporters in plants: New insights and discoveries. *Plant Cell Physiol.* 58, 1442–1460. doi: 10.1093/pcp/pcx090

Kankel, M. W., Ramsey, D. E., Stokes, T. L., Flowers, S. K., Haag, J. R., Jeddeloh, J. A., et al. (2003). Arabidopsis MET1 cytosine methyltransferase mutants. *Genetics* 163, 1109–1122. doi: 10.1093/genetics/163.3.1109

Kawashima, T., and Berger, F. (2014). Epigenetic reprogramming in plant sexual reproduction. *Nat. Rev. Genet.* 15, 613–624. doi: 10.1038/nrg3685

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

Klopfenstein, D. V., Zhang, L., Pedersen, B. S., Ramírez, F., Warwick Vesztrocy, A., Naldi, A., et al. (2018). GOATOOLS: A Python library for gene ontology analyses. *Sci. Rep.* 8, 10872. doi: 10.1038/s41598-018-28948-z

Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C. T., et al. (2010). Dynamic changes in the human methylome during differentiation. *Genome Res.* 20, 320–331. doi: 10.1101/gr.101907.109

Law, J. A., and Jacobsen, S. E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* 11, 204–220. doi: 10.1038/nrg2719

Lee, S. K., Jeon, J. S., Bornke, F., Voll, L., Cho, J. I., Goh, C. H., et al. (2008). Loss of cytosolic fructose-1,6-bisphosphatase limits photosynthetic sucrose synthesis and causes severe growth retardations in rice (Oryza sativa). *Plant Cell Environ.* 31, 1851–1863. doi: 10.1111/j.1365-3040.2008.01890.x

Lindroth, A. M., Cao, X., Jackson, J. P., Zilberman, D., Mccallum, C. M., Henikoff, S., et al. (2001). Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation. *Science* 292, 2077–2080. doi: 10.1126/science.1059745

Lin, J. Y., Le, B. H., Chen, M., Henry, K. F., Hur, J., Hsieh, T. F., et al. (2017). Similarity between soybean and arabidopsis seed methylomes and loss of non-CG methylation does not affect seed development. *Proc. Natl. Acad. Sci. U.S.A.* 114, E9730–e9739. doi: 10.1073/pnas.1716758114

Li, Z., Wang, M., Lin, K., Xie, Y., Guo, J., Ye, L., et al. (2019). The bread wheat epigenomic map reveals distinct chromatin architectural and evolutionary features of functional genetic elements. *Genome Biol.* 20, 139. doi: 10.1186/s13059-019-1746-8

Li, Y., Zheng, H., Wang, Q., Zhou, C., Wei, L., Liu, X., et al. (2018). Genome-wide analyses reveal a role of polycomb in promoting hypomethylation of DNA methylation valleys. *Genome Biol.* 19, 18. doi: 10.1186/s13059-018-1390-8

Li, X., Zhu, J., Hu, F., Ge, S., Ye, M., Xiang, H., et al. (2012). Single-base resolution maps of cultivated and wild rice methylomes and regulatory roles of DNA methylation in plant gene expression. *BMC Genomics* 13, 300. doi: 10.1186/1471-2164-13-300

Meijer, H. J., and Munnik, T. (2003). Phospholipid-based signaling in plants. *Annu. Rev. Plant Biol.* 54, 265–306. doi: 10.1146/annurev.arplant.54.031902.134748

Morales-Ruiz, T., Ortega-Galisteo, A. P., Ponferrada-Marín, M. I., Martínez-Macías, M. I., Ariza, R. R., and Roldán-Arjona, T. (2006). DEMETER and REPRESSOR OF SILENCING 1 encode 5-methylcytosine DNA glycosylases. *Proc. Natl. Acad. Sci. U.S.A.* 103, 6853–6858. doi: 10.1073/pnas.0601109103

Niederhuth, C. E., Bewick, A. J., Ji, L., Alabady, M. S., Kim, K. D., Li, Q., et al. (2016). Widespread natural variation of DNA methylation within angiosperms. *Genome Biol.* 17, 194. doi: 10.1186/s13059-016-1059-0

Niu, S., Li, J., Bo, W., Yang, W., Zuccolo, A., Giacomello, S., et al. (2022). The Chinese pine genome and methylome unveil key features of conifer evolution. *Cell* 185, 204–217.e14. doi: 10.1016/j.cell.2021.12.006

Ortega-Galisteo, A. P., Morales-Ruiz, T., Ariza, R. R., and Roldán-Arjona, T. (2008). Arabidopsis DEMETER-LIKE proteins DML2 and DML3 are required for appropriate distribution of DNA methylation marks. *Plant Mol. Biol.* 67, 671–681. doi: 10.1007/s11103-008-9346-0

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295. doi: 10.1038/nbt.3122

Pierella Karlusich, J. J., and Carrillo, N. (2017). Evolution of the acceptor side of photosystem I: ferredoxin, flavodoxin, and ferredoxin-NADP(+) oxidoreductase. *Photosynth Res.* 134, 235–250. doi: 10.1007/s11120-017-0338-2

Sahraeian, S. M. E., Mohiyuddin, M., Sebra, R., Tilgner, H., Afshar, P. T., Au, K. F., et al. (2017). Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nat. Commun.* 8, 59–59. doi: 10.1038/s41467-017-00050-4

Shigeto, J., and Tsutsumi, Y. (2016). Diverse functions and reactions of class III peroxidases. *New Phytol.* 209, 1395–1402. doi: 10.1111/nph.13738

Song, Q., Guan, X., and Chen, Z. J. (2015). Dynamic roles for small RNAs and DNA methylation during ovule and fiber development in allotetraploid cotton. *PloS Genet.* 11, e1005724. doi: 10.1371/journal.pgen.1005724

Song, Q. X., Lu, X., Li, Q. T., Chen, H., Hu, X. Y., Ma, B., et al. (2013). Genome-wide analysis of DNA methylation in soybean. *Mol. Plant* 6, 1961–1974. doi: 10.1093/mp/sst123

Stadler, M. B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., et al. (2011). DNA-Binding factors shape the mouse methylome at distal regulatory regions. *Nature* 480, 490–495. doi: 10.1038/nature10716

Strand, A., Zrenner, R., Trevanion, S., Stitt, M., Gustafsson, P., and Gardeström, P. (2000). Decreased expression of two key enzymes in the sucrose biosynthesis pathway, cytosolic fructose-1,6-bisphosphatase and sucrose phosphate synthase, has remarkably different consequences for photosynthetic carbon metabolism in transgenic arabidopsis thaliana. *Plant J.* 23, 759–770. doi: 10.1046/j.1365-313x.2000.00847.x

Streb, S., Egli, B., Eicke, S., and Zeeman, S. C. (2009). The debate on the pathway of starch synthesis: A closer look at low-starch mutants lacking plastidial phosphoglucomutase supports the chloroplast-localized pathway. *Plant Physiol.* 151, 1769–1772. doi: 10.1104/pp.109.144931

Stroud, H., Do, T., Du, J., Zhong, X., Feng, S., Johnson, L., et al. (2014). Non-CG methylation patterns shape the epigenetic landscape in arabidopsis. *Nat. Struct. Mol. Biol.* 21, 64–72. doi: 10.1038/nsmb.2735

Tetlow, I. J., Morell, M. K., and Emes, M. J. (2004). Recent developments in understanding the regulation of starch metabolism in higher plants. *J. Exp. Bot.* 55, 2131–2145. doi: 10.1093/jxb/erh248

Tong, W., Li, R., Huang, J., Zhao, H., Ge, R., Wu, Q., et al. (2021). Divergent DNA methylation contributes to duplicated gene evolution and chilling response in tea plants. *Plant J.* 106, 1312–1327. doi: 10.1111/tpj.15237

Turco, G. M., Kajala, K., Kunde-Ramamoorthy, G., Ngan, C. Y., Olson, A., Deshphande, S., et al. (2017). DNA Methylation and gene expression regulation associated with vascularization in sorghum bicolor. *New Phytol.* 214, 1213–1229. doi: 10.1111/nph.14448

Verma, A. K., Upadhyay, S. K., Verma, P. C., Solomon, S., and Singh, S. B. (2011). Functional analysis of sucrose phosphate synthase (SPS) and sucrose synthase (SS) in sugarcane (Saccharum) cultivars. *Plant Biol. (Stuttg)* 13, 325–332. doi: 10.1111/j.1438-8677.2010.00379.x

Wang, H., Beyene, G., Zhai, J., Feng, S., Fahlgren, N., Taylor, N. J., et al. (2015). CG gene body DNA methylation changes and evolution of duplicated genes in cassava. *Proc. Natl. Acad. Sci. U.S.A.* 112, 13729–13734. doi: 10.1073/pnas.1519067112

Wang, W. S., Pan, Y. J., Zhao, X. Q., Dwivedi, D., Zhu, L. H., Ali, J., et al. (2011). Drought-induced site-specific DNA methylation and its association with drought

tolerance in rice (Oryza sativa l.). *J. Exp. Bot.* 62, 1951–1960. doi: 10.1093/jxb/erq391

Wang, L., Shi, Y., Chang, X., Jing, S., Zhang, Q., You, C., et al. (2019). DNA Methylome analysis provides evidence that the expansion of the tea genome is linked to TE bursts. *Plant Biotechnol. J.* 17, 826–835. doi: 10.1111/pbi.13018

Wang, L., Xie, J., Hu, J., Lan, B., You, C., Li, F., et al. (2018). Comparative epigenomics reveals evolution of duplicated genes in potato and tomato. *Plant J.* 93, 460–471. doi: 10.1111/tpj.13790

Wang, Q., Xu, J., Pu, X., Lv, H., Liu, Y., Ma, H., et al. (2021). Maize DNA methylation in response to drought stress is involved in target gene expression and alternative splicing. *Int. J. Mol. Sci.* 22, 8285–8303. doi: 10.3390/ijms22158285

Xie, W., Schultz, M. D., Lister, R., Hou, Z., Rajagopal, N., Ray, P., et al. (2013). Epigenomic Analysis of Multilineage Differentiation of Human Embryonic Stem Cells. *Cell* 153, 1134–1148. doi: 10.1016/j.cell.2013.04.022

Xi, Y., and Li, W. (2009). BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinf.* 10, 232. doi: 10.1186/1471-2105-10-232

Xu, J., Zhou, S., Gong, X., Song, Y., Van Nocker, S., Ma, F., et al. (2018). Single-base methylome analysis reveals dynamic epigenomic differences associated with water deficit in apple. *Plant Biotechnol. J.* 16, 672–687. doi: 10.1111/pbi.12820

Zemach, A., Kim, M. Y., Hsieh, P. H., Coleman-Derr, D., Eshed-Williams, L., Thao, K., et al. (2013). The arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. *Cell* 153, 193–205. doi: 10.1016/j.cell.2013.02.033

Zhang, H., Lang, Z., and Zhu, J. K. (2018a). Dynamics and function of DNA methylation in plants. *Nat. Rev. Mol. Cell Biol.* 19, 489–506. doi: 10.1038/s41580-018-0016-z

Zhang, J., Zhang, X., Tang, H., Zhang, Q., Hua, X., Ma, X., et al. (2018b). Allele-defined genome of the autopolyploid sugarcane saccharum spontaneum l. *Nat. Genet.* 50, 1565–1573. doi: 10.1038/s41588-018-0237-2

Zhu, N., Cheng, S., Liu, X., Du, H., Dai, M., Zhou, D. X., et al. (2015). The R2R3-type MYB gene OsMYB91 has a function in coordinating plant growth and salt stress tolerance in rice. *Plant Sci.* 236, 146–156. doi: 10.1016/j.plantsci.2015.03.023

Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T., and Henikoff, S. (2007). Genome-wide analysis of arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nat. Genet.* 39, 61–69. doi: 10.1038/ng1929

frontiers | Frontiers in Plant Science

TYPE Review
PUBLISHED 08 December 2022
DOI 10.3389/fpls.2022.1038109

OPEN ACCESS

EDITED BY
Weizhen Liu,
Wuhan University of Technology, China

REVIEWED BY
Yubin Li,
Qingdao Agricultural University, China
Atsushi Fukushima,
Kyoto Prefectural University, Japan
Chongjing Xia,
Southwest University of Science and Technology, China

*CORRESPONDENCE
Min Tu
719378705@qq.com;
12739@whpu.edu.cn
Guangsen Song
1697446119@qq.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 06 September 2022
ACCEPTED 21 November 2022
PUBLISHED 08 December 2022

CITATION
Tu M, Zeng J, Zhang J, Fan G and Song G (2022) Unleashing the power within short-read RNA-seq for plant research: Beyond differential expression analysis and toward regulomics. *Front. Plant Sci.* 13:1038109. doi: 10.3389/fpls.2022.1038109

Unleashing the power within short-read RNA-seq for plant research: Beyond differential expression analysis and toward regulomics

Min Tu[1*†], Jian Zeng[2†], Juntao Zhang[1], Guozhi Fan[1] and Guangsen Song[1*]

[1]School of Chemical and Environmental Engineering, Wuhan Polytechnic University, Wuhan, China, [2]Guangdong Provincial Key Laboratory of Utilization and Conservation of Food and Medicinal Resources in Northern Region, Shaoguan University, Shaoguan, Guangdong, China

RNA-seq has become a state-of-the-art technique for transcriptomic studies. Advances in both RNA-seq techniques and the corresponding analysis tools and pipelines have unprecedently shaped our understanding in almost every aspects of plant sciences. Notably, the integration of huge amount of RNA-seq with other omic data sets in the model plants and major crop species have facilitated plant regulomics, while the RNA-seq analysis has still been primarily used for differential expression analysis in many less-studied plant species. To unleash the analytical power of RNA-seq in plant species, especially less-studied species and biomass crops, we summarize recent achievements of RNA-seq analysis in the major plant species and representative tools in the four types of application: (1) transcriptome assembly, (2) construction of expression atlas, (3) network analysis, and (4) structural alteration. We emphasize the importance of expression atlas, coexpression networks and predictions of gene regulatory relationships in moving plant transcriptomes toward regulomics, an omic view of genome-wide transcription regulation. We highlight what can be achieved in plant research with RNA-seq by introducing a list of representative RNA-seq analysis tools and resources that are developed for certain minor species or suitable for the analysis without species limitation. In summary, we provide an updated digest on RNA-seq tools, resources and the diverse applications for plant research, and our perspective on the power and challenges of short-read RNA-seq analysis from a regulomic point view. A full utilization of these fruitful RNA-seq resources will promote plant omic research to a higher level, especially in those less studied species.

KEYWORDS

plant transcriptomics, RNA-seq data analysis, alternative splicing, alternative polyadenylation, coexpression network, gene regulatory network, regulomics

Introduction

RNA-seq and its-derived techniques have been commercially available and routinely used by biological scientists, largely owing to the rapidly increased outputs of major sequencing platforms, improved sequencing accuracy and ever reduced costs (Stark et al., 2019). RNA-seq has shaped nearly every aspects of our understanding in plant research, from plant development and phytohorome signaling to plant metabolism and stress tolerance.

RNA-seq can be divided into the short-read (Nagalakshmi et al., 2008) and long-read RNA-seq technologies (Sharon et al., 2013). In short-read RNA-seq, Illumina sequencing platform has been dominant, while other platforms, such as Thermo Scientific platforms (e.g., Ion PGM and Ion S5) or the BGI Genomics platforms (e.g., DNBSEQ), have been frequently used in certain circumstances or been gaining attentions recently (Patterson et al., 2019; Foox et al., 2021). A short-read RNA-seq library is typically sequenced to a read depth of 10~30 million reads per sample with a read length varied from 50 to 200 bp. By contrast, a number of approaches (*e.g.*, Pacific Bioscience, PacBio and Oxford Nanopore, ONT) provide long, uninterrupted sequencing of a single RNA or DNA molecules, constituting the third generation of real-time fluorescence sequencing paradigm (Sharon et al., 2013; Cartolano et al., 2016; Oikonomopoulos et al., 2016). A typical long-read RNA-seq produces 500,000 to 10 million reads per run with a read length ranging from 1,000 to 50,000 bp depending on the technologies and platforms (Stark et al., 2019). The long-read sequencing platforms are particularly suited for *de novo* transcriptome assembly and identification of novel transcripts and isoforms, as these approaches overcome some intrinsic issues related to short-read sequencing.

While the rise of the long-read RNA-seq, the short-read RNA-seq still is dominating the current utilizations in plant sciences and has provided the majority of the data sets deposited in public sequencing databases. With the recent advancement of tools developed for analyzing short-read sequencing data, the RNA-seq technology can be used for various applications, including but not limited to: (1) *de novo* assembly of transcriptome with or without a reference genome; (2) detection of new transcripts or correction of existing gene structures based on RNA-seq evidence; (3) to obtaining the expression profiles at gene or transcript levels and to construct the expression atlas covering a range of conditions and tissue types; (4) to identify alternative splicing and alternative 5' or 3' untranslated regions (5'UTR or 3'UTR, respectively); (5) to construct gene co-expression networks (GCNs) and predict gene regulatory relationships in a large scale (also known as gene regulatory networks, GRN). Here, GCN stand for a network that can be constructed from a large set of RNA-seq data and includes multiple clusters or modules. The module represents a group of genes determined statistically with high correlation in their expression profiles and usually associations in their functions (reviewed in Gupta and Pereira, 2019). Notably, many genes within the same module do not represent the direct targets of their upstream regulators. Thus, to further disentangle the direct regulator-targets pairs from the indirectly regulated or co-expressed genes, prediction of GRNs is another important task in RNA-seq data analysis. Identification of GRNs can be achieved by harnessing the following resources: (1) identifying transcription factors (TFs) from co-expressed modules; (2) identifying a group of co-expressed genes with the statistically enriched cis-regulatory elements from a certain family of TF; (3) leveraging the information of direct TF targets by using existing results from chromatin immunoprecipitation sequencing (ChIP-seq) or DNA affinity purification sequencing (DAP-seq) experiments (O'Malley et al., 2016; Galli et al., 2020); (4) applying the well-established algorithms for GRN inference. While the many utilizations of RNA-seq, the differential gene expression (DGE) is still the most often used analysis in many plant researches, especially those carried on in crop species.

Here, we highlight typical examples of the tools and applications that have been used in the model plants (Arabidopsis and rice) and other major crops (*e.g.*, tomato, wheat, maize and soybean) (Table 1). These applications demonstrate the power and comprehensiveness of short-read, bulk RNA-seq analyses. Meanwhile, it is worth noting that DGE has long been the primary analysis in the RNA-seq studies of other less-studied plant species. In fact, many species, especially those minor crops, biomass crops or orphan crops, are key to provide sustainable agriculture and to reach global food and energy security. Particularly, major biomass crops, such as sorghum, sugarcane, *Miscanthus*, and switchgrass, have large yield of biomass and stress tolerance (Mullet et al., 2014; Boyles et al., 2019), justifying the significance for researching on gene expression and regulation associated with biomass composition and production.

The limited utilization of RNA-seq in the minor plant species has been partly due to: (1) the limited genomic resources; (2) lacking bioinformatic tools that are user friendly, with a graphical user interface, or well adapted to the omics data of various species. In this context, we summarize a variety of bioinformatic tools covering the diverse applications of bulk RNA-seq analysis to facilitate the full use of short-read RNA-seq data, and to help unleash the power of bulk RNA-seq in studies of plants, especially in the minor and under-utilized crops (Table 1; Figure 1). Notably, there have been several excellent reviews regarding the development of RNA-seq technologies, comprehensive summary of RNA-seq tools and calculation of GCNs and GRNs in plant sciences (Van Verk et al., 2013; Conesa et al., 2016; Proost and Mutwil, 2016; Gaudinier and Brady, 2016; Sahraeian et al., 2017; Saelens et al., 2018; Haque et al., 2018; Stark et al., 2019; Gupta and Pereira, 2019). We aim

TABLE 1 Summary of the representative resources and tools for analyzing the short-read RNA-seq data in plants.

Name	Reference	URL	Implementation	Classification[1]
Plant Reactome	Nathani et al., 2017 & Nathani et al., 2020	http://plantreactome.gramene.org	Web Page	Annotation
Strawberry	Liu and Dickerson, 2017	https://github.com/ruolin/strawberry	Stand Alone	Annotation
iDEP	Ge et al., 2018	http://ge-lab.org/idep/	R Package	Annotation
TransFlow	Seoane et al., 2018	https://github.com/seoanezonjic/TransFlow.	Stand Alone	Annotation
MorphDB	Zwaenepoel et al., 2018	http://bioinformatics.psb.ugent.be/webtools/morphdb/morphDB/index/.	Web Page	Annotation
PISO	Feng et al., 2019	http://cbi.hzau.edu.cn/piso/.	Web Page	Annotation
MapMan 4/Mercator4	Schwacke et al., 2019	https://www.plabipd.de/portal/legacy-mercator4	Web Page	Annotation
PlantCircBase	Chu et al.,2018	http://ibi.zju.edu.cn/plantcircbase/	Web Page	Annotation & Expr.
Gramene	Tello-Ruiz et al., 2018	http://www.gramene.org	Web Page	Annotation & Expr.
LeGOO	Carrere et al., 2020	https://www.legoo.org	Web Page	Annotation & Expr.
ZEAMAP	Gui et al., 2020	http://www.zeamap.com	Web Page	Annotation & Expr.
BarleyNet	Lee et al., 2020	http://www.inetbio.org/barleynet	Web Page	Annotation & Expr.
SAT-Assembler	Zhang et al., 2014	https://sourceforge.net/projects/sat-assembler/	Stand Alone	Assembler
BinPacker	Liu et al., 2016	http://sourceforge.net/projects/transcriptomeassembly/files/BinPacker_1.0.tar.gz/download	Stand Alone	Assembler
Rascaf	Song et al., 2016	https://github.com/mourisl/Rascaf.	Stand Alone	Assembler
IGB	Freese et al., 2016	http://bioviz.org/igb.	Web Page	Browser
eFP-Seq Browser	Sullivan et al., 2019	https://bar.utoronto.ca/eFP-Seq_Browser/	Web Page	Browser
RNAprof	Tran et al., 2016	http://rna.igmors.u-psud.fr/Software/rnaprof.php	Stand Alone	AS/APA
Apatrap	Ye et al., 2018	https://apatrap.sourceforge.io.	Stand Alone	AS/APA
Name	**Citation**	**URL**	**Implementation**	**Classification**
priUTR	Tu and Li, 2020	https://github.com/mint1234/3UTR-	Stand Alone	AS/APA
3D RNA-Seq	Guo et al., 2021	https://ics.hutton.ac.uk/3drnaseq	R Package	AS/APA
TEtranscripts	Jin et al., 2015	http://hammelllab.labsites.cshl.edu/software	Stand Alone	Expression
expVIP	Borrill et al., 2016	www.wheat-expression.com	Web Page	Expression
OryzaExpress	Kudo et al., 2017	http://plantomics.mind.meiji.ac.jp/OryzaExpress/	Web Page	Expression
BAR	Waese and Provart, 2016	http://bar.utoronto.ca	Web Page	Expression
DPMIND	Fei et al.2018	http://202.195.246.60/DPMIND/	Web Page	Expression
PEATmoss	Fernandez-Pizo et al., 2020	https://peatmoss.online.uni-marburg.de	Web Page	Expression
ASmir	Wang et al., 2019	http://forestry.fafu.edu.cn/bioinfor/db/ASmiR	Web Page	Expression
Soybean Expression Atlas	Machado et al.,2020	http://venanciogroup.uenf.br/resources/	Web Page	Expression
Grape-RNA	Wang et al., 2020	http://www.grapeworld.cn/gt/2	Web Page	Expression
CORNET	Van Bel and Coppens, 2017	http://bioinformatics.psb.ugent.be/cornet/	Web Page	Expr. & Coexp.
NaDH	Brockmoller et al., 2017	http://nadh.ice.mpg.de/	Web Page	Expr. & Coexp.
NorWood	Jokipii-Lukkari et al., 2017	http://norwood.congenie.org	Web Page	Expr. & Coexp.
AspWood	Sundell et al., 2017	http://aspwood.popgenie.org	Web Page	Expr. & Coexp.
RED	Xia et al., 2017	http://expression.ic4r.org	Web Page	Expr. & Coexp.
EXPath	Zheng et al.2017	http://expathtool.itps.ncku.edu.tw/	Web Page	Expr. & Coexp.
TomExpress	Zouine et al., 2017	http://tomexpress.toulouse.inra.fr	Web Page	Expr. & Coexp.
Maize eFP Brower	Hoopes et al., 2019	bar.utoronto.ca/efp_maize	Web Page	Expr. & Coexp.
ATTED	Obayashi et al., 2018	http://atted.jp	Web Page	Expr. & Coexp.
MCENet	Tian et al., 2018	http://bioinformatics.cau.edu.cn/MCENet/	Web Page	Expr. & Coexp.

(Continued)

TABLE 1 Continued

Name	Reference	URL	Implementation	Classification[1]
AppleMDO	Da et al., 2019	http://bioinformatics.cau.edu.cn/AppleMDO/	Web Page	Expr. & Coexp.
Name	**Citation**	**URL**	**Implementation**	**Classification**
Melonet-DB	Yano et al., 2018	http://melonet-db.agbi.tsukuba.ac.jp/	Web Page	Expr. & Coexp. & Anno.
TPIA	Xia et al., 2019	http://tpia.teaplant.org	Web Page	Expr. & Coexp. & Anno.
Plant Regulomics	Ran et al., 2020	http://bioinfo.sibs.ac.cn/plant-regulomics.	Web Page	Expr. & Coexp. & Anno.
CSI	Penfold et al., 2015a & Penfold et al., 2015b	http://go.warwick.ac.uk/systemsbiology/software.	Stand Alone	Network construction
RSAT-Plants	Contreras-Moreira et al., 2016	http://plants.rsat.eu	Web Page	Network construction
tcgsaseq	Agniel and Hejblum, 2017	https://cran.r-project.org/web/packages/tcgsaseq.	R Package	Network construction
SeqEnrich	Becker et al., 2017	http://www.belmontelab.com	Stand Alone	Network construction
ExRANGES	Desai et al., 2017	http://github.com/DohertyLab/ExRANGES	R Package	Network construction
LSTrAP	Proost et al., 2017	https://github.molgen.mpg.de/proost/LSTrAP	Stand Alone	Network construction
RSAT	Nguyen et al., 2018	http://www.rsat.eu/	Stand Alone	Network construction
NetMiner	Yu et al., 2018	https://github.com/czllab/NetMiner.	Stand Alone	Network construction
ExpressWeb	Savelli et al., 2019	http://polebio.lrsv.upstlse.fr/ExpressWeb/	R Package	Network construction
HTRgene	Ahn et al., 2019	http://biohealth.snu.ac.kr/software/HTRgene.	R Package	Network construction
Compare Transcriptome Analysis	Lee et al., 2019	https://github.com/LiLabAtVT/CompareTranscriptome.git).	R Package	Network construction
JASPAR	Fornes et al., 2020	http://jaspar.genereg.net	Web Page	Network construction
GENIE3	Harrington et al., 2020	https://github.com/Uauy-Lab/GENIE3_scripts/	Stand Alone	Network construction
LSTrAP-Cloud	Tan et al., 2020	https://github.com/tqiaowen/LSTrAP-Cloud	Stand Alone	Network construction
RSAT	Ksouri et al., 2021	https://github.com/RSAT-doc/motif_discovery_clusters	Web Page	Network construction

1. The RNA-seq resources and tools are classified by their functions, including annotation, expression atlas (expression, or abbreviated as 'Expr.'), co-expression analysis (abbreviated as 'Coexp.'), alternative splicing and alternative polyadenylation (abbreviated as 'AS/APA'), and network construction (tools for calculating coexpression networks or gene regulatory networks). These resources and tools are first sorted by classification and then by publication years.

at neither comprehensively cataloguing the RNA-seq analysis tools for plant research, nor summarizing the achievements that RNA-seq have been reached in plant research. We emphasize that recent advancements in RNA-seq analysis tools allow to fully unleash the power of short-read, bulk RNA-seq in many plant species like biomass crops, to provide deep insights into gene regulation at multiple levels and to go toward regulomics, an analogous term to other omics that portraits transcription control in a genome-wide manner (Werner, 2003; Werner, 2004). Particularly, regulomics refers to the omic-scale study of gene expression regulation happened at transcriptional or post-transcriptional levels (Werner, 2004), such as the regulation between transcription factors/coregulators and their targets and the interaction between non-coding RNAs (e.g., miRNAs anf lncRNAs) and mRNAs.

The applications of the short-read, bulk RNA-seq in plant sciences

The short-read RNA-seq technique includes several core steps, from RNA extraction, cDNA synthesis, adapter ligation, PCR amplification, to the sequencing of library and data analysis. Four key stages are required for the RNA-seq data

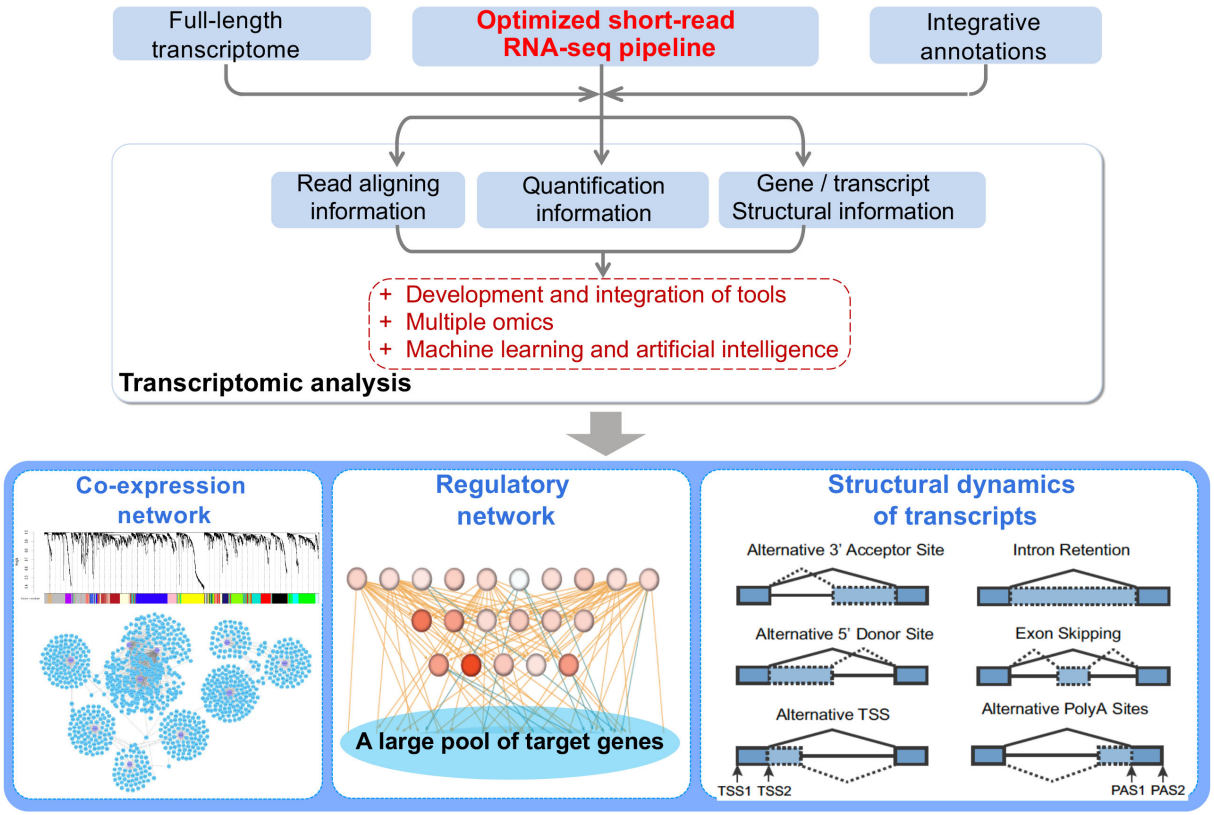


FIGURE 1
The power of short-read, bulk RNA-seq can be unleashed by integrating the following resources and tools related to RNA-seq analysis: **(1)** Full-length transcriptome can be achieved by full-length cDNA sequencing, PacBio Iso-seq or the Oxford Nanopore sequencing technologies, and these full-length transcriptomes can help to better annotate gene structures and serve as the basis for expression profiling at the transcript-level. **(2)** For many less-studied plant species, multiple functional annotation resources can be applied to provide a comprehensive annotation, facilitating biological interpretation of sets of DEG or gene networks. **(3)** Through application of the tools introduced here and in previous reviews, high-quality GCNs and GRNs can be made to prioritize hub genes or key regulators involved in the certain biological process or phenotypes.

analysis: (1) The first stage takes the raw sequencing reads to quality control and maps the quality-controlled reads to the transcriptome, which can be obtained from a reference genome or be assembled from transcriptomic data; (2) The second stage quantifies the number of reads mapped to each gene or transcript, producing an expression matrix; (3) The third stage modifies the expression matrix by normalization between samples, accounting for technical differences, and removing lowly expressed genes/transcripts; (4) The last stage calculates differentially expressed genes or transcripts by statistical models. Particularly, the number of computational tools for analyzing RNA-seq data has been increased dramatically in the recent decade (Stark et al., 2019). As such, substantial influences can be generated on the biological conclusions drawn from the RNA-seq data due to several aspects: differences in the computational approaches used, software parameters or statistical models selected and distinct combinations of the tools in a pipeline (Conesa et al., 2016). The optimal set of computational approaches for RNA-seq depends on the experimental setup, the biological questions being addressed and other factors, and is beyond the scope of our mini-review (Conesa et al., 2016; Sahraeian et al., 2017). However, several sets of RNA-seq tools are well recognized, representing the classic pipelines (Trapnell et al., 2012; Grabherr et al., 2012; Pertea et al., 2017). These includes five main components: (1) the splice-aware aligners (*e.g.*, TopHat, STAR, HISAT and HISAT 2; Kim et al., 2019) to map RNA-seq reads to the reference genome; (2) the tools for reads extraction [*e.g.*, HTSeq (Anders et al., 2014) and featureCount (Liao et al., 2014)]; (3) the tools for transcript construction (e.g., CuffLinks, StringTie) (Trapnell et al., 2012; Pertea et al., 2017); (4) the tools for estimates gene/transcript abundance [e.g., CuffDiff2, Ballgown and RSEM (Li and Dewey, 2011)]; and (5) the tools to identify differentially expressed genes or transcripts based on statistical analyses (such as edgeR (Robinson et al., 2010), DESeq2 (Love et al., 2014), Ballgown and CuffDiff2). The majority of the applications and

computational tools summarized in the follow are compatible with these classic RNA-seq pipelines.

RNA-seq data enhance transcriptome assembly

The number of plant species with at least one reference genome have multiplied dramatically over the past few years, with 798 land plant species having genome assemblies (as of Jan. 2021) (Marks et al., 2021). While these genomic resources greatly ease the RNA-seq analysis, still the complexity in plant genomes and transcriptomes presents major challenges in RNA-seq analysis. Many plant species feature large genomes (for example, the median sizes of currently sequenced monocots and eudicots respectively are more than 500 Mb) or complex auto- or allo- polyploid genomes with some hybridization and introgressions (Zhang et al., 2018; Zhao et al., 2021; Sun et al., 2022). Many genomes are expanded by repetitive sequences (such as transposons), making it difficult to achieve complete and accurate annotation of multi-exonic genes. Besides, alternative splicing (AS) and alternative polyadenylation (APA) further enhance transcriptome complexity. In addition, gene families commonly seen in the plant genomes are shaped by whole genome duplication, segmental duplication and tandem duplication. The members within a gene family or the homo-/homoeo-logous alleles (in polyploid) usually share high sequence similarity between each other, thus posing ad-ditional challenges in accurate quantification of the expression levels by using RNA-seq data.

To overcome these challenges, two strategies have been evolved when a reference genome is available: **(1)** to assembly transcripts first and then to quantify expression; **(2)** to simultaneously construct transcripts and to quantify expression. For the genome-guided transcriptome analysis, multiple pipelines have been established that differ in the algorithms used and the speed and computational resources required, including the classic TopHat-Cufflink-Cuffdiff pipeline (Trapnell et al., 2012) and HISAT-StringTie-Ballgown pipeline (Pertea et al., 2017), as well as the new "Strawberry" tool (Liu and Dickerson, 2017). By contrast, when a reference genome and gene annotations do not exist, a transcriptome needs to be firstly *de novo* assembled to facilitate expression quantification. However, *de novo* assembly based on short-read RNA-seq data usually leads to fractured and incomplete view of transcriptome, complicating downstream analysis (Malik et al., 2018). Several tools for *de novo* assembling full-length transcripts have become popular with different algorithms and features, such as Trinity (Haas et al., 2013), Oasis (Schulz et al., 2012), Trans-AbySS (Robertson et al., 2010), SOAPdenovo-Trans (Xie et al., 2014), Corset (Garber et al., 2011) and BinPacker (Liu et al., 2016). More recently, Grouper provides a complete pipeline for processing *de novo* transcriptomic analysis by using a new method for clustering assembled contigs (Malik et al., 2018). TransFlow provides a versatile workflow to enhance *de novo* transcriptome analyses and to annotate transcript structures more accurately by combining short-read and long-read sequencing data (Seoane et al., 2018).

RNA-seq data empower the construction of expression atlas

Rapid accumulation of immense sets of RNA-seq data allows the establishment of expression atlantes. An expression atlas collects a large number of RNA-seq data from a certain species and re-analyzes these data using standardized, open-source pipelines to remove potential batch effects and any influences caused by other factors, such as different research groups, sequencing platforms and experiments (Papatheodorou et al., 2018). Establishing expression atlas has been proved very valuable in model organisms to promote not only omics studies but more importantly our understanding in gene functions, as clues to gene function can often be inferred by examining when and where a gene is expressed in the organism (Alberts et al., 2002). In model plants and major crops, such expression atlantes have served as key resources to the research community. For example, the information hub of Arabidopsis (TAIR; Berardini et al., 2015) and maize (MaizeGDB; Lawrence et al., 2008) have implemented with the expression atlas for each species. Maize expression atlas websites have been updated or built separately by multiple groups to integrate more RNA-seq data, other omics data sets or visualizations (Sekhon et al., 2013; Stelpflug et al., 2015; Tian et al., 2018; Hoopes et al., 2019; Gui et al., 2020). Similarly, the rice expression atlas has been updated from microarray to RNA-seq data sets and established by several groups respectively (Sato et al., 2013; Kudo et al., 2017; Xia et al., 2017). Recently, the expression atlantes have also been built for other important crops, such as tomato (TomExpress, Zouine et al., 2017), soybean (Machado et al., 2020), wheat (Borrill et al., 2016), barley (BarleyNet, Lee et al., 2020) and sorghum (Makita et al., 2015). The trend of building RNA-seq-based expression atlas has been spread to many less-studied plant species, for example, *Picea abies* (the Norwood database, Jokipii-Lukkari et al., 2017), *Populus tremula* (the Aspwood database, Sundell et al., 2017), chickpea (Kudapa et al., 2018), *Physcomitrella Paten* (Perroud et al., 2018; Fernandez-Pizo et al., 2020), tabacco (NaDH- Brockmoller et al., 2017), water melon (Melonet-DB - Yano et al., 2018), apple (AppleMDO- Da et al., 2019), tea (TPIA - Xia et al., 2019), grape (Wang et al., 2020), and *Medicago truncatula* (LeGOO- Carrere et al., 2020).

Notably, two types of the integrative websites are particularly valuable in facilitating comparative functional genomics and molecular breeding. (1) The expression atlas website includes a number of useful functions, from the visualization, comparison

and functional enrichment of the omics data to comprehensive annotations of genes or gene families and useful functions such as primer design, BLAST and ortholog identification. (2) The RNA-seq data are further utilized to construct co-expression modules and integrated with other types of omics data, for example epigenomic data sets. In addition, major plant genomics websites (for instance, the Phytozome (Goodstein et al., 2011) Ensembl Plants (Bolster et al., 2017), and Gramene (Tello-Ruiz et al., 2018)) serve as the central data hub to link numerous plant genomes to those of the model species, which are well characterized and annotated. These iconic plant genomic hubs lay a solid foundation for transferring and comparing the omic information from model plants to less-studied species.

RNA-seq data capture large-scale co-expression networks

One major cornerstone of the data-driven biological interpretation of large-scale RNA-seq data is to transform expression data into networks and modules. Among the network representation methods, co-expression network is the one that has been widely applied and successful in many species (Farber and Lusis, 2008). In a co-expression network, genes are connected by edges that quantify the similarity between gene expression patterns, and the genes expressed similarly are grouped together forming a co-expression module. Co-expression network can be calculated by different approaches, from correlation-based methods like Pearson Correlation Coefficiency (PCC) (D'haeseleer et al., 2000) and weighted gene co-expression network analysis (WGCNA) (Langfelder and Horvath, 2008; Langfelder and Horvath, 2012), to linear modelling (Vasilevski et al., 2012) and mutual information methods (Daub et al., 2004). Through the "guilt-by-association" principle, genes in a co-expression module possibly indicate similar functions and modes of transcriptional regulation (Wolfe et al., 2005), or similar cellular compartments of the protein products (Ryngajllo et al., 2011).

Over the past decade, high-quality co-expression networks and their hosting data hubs have served as a valuable resource to facilitate the gene functional studies in model plant species and many major crops, including Arabidopsis (Van Bel et al., 2017; Obayashi et al., 2018), rice (Xia et al., 2017), maize (Miao et al., 2017; Tian et al., 2018; Hoopes et al., 2019), and tomato (Zouine et al., 2017). More recently, co-expression networks have been built in other plant species (Kudapa et al., 2018), including some forest species with biomass purposes (Jokipii-Lukkari et al., 2017; Sundell et al., 2017), demonstrating the power of network representation in providing molecular functional insights into biomass production. Nonetheless, the biologists who work on less-studied plant species might neither have the bioinformatic skills nor afford the computational resources that are required to integrate large-scale RNA-seq data sets and to construct high-quality networks. Thus, user-friendly online or offline tools have been developed to lower the bar for co-expression-based analysis, such as the Kallisto-based LSTrAP pipeline (Proost et al., 2017), the LSTrAP-Cloud (Tan et al., 2020) and the ExpressWeb (Savelli et al., 2019). Besides, computational methods have been reported to improve the quality of co-expression network identification (NetMiner, Yu et al., 2018; PCC-HRR Liesecke et al., 2018). These tools aim toward paving the way to perform co-expression analysis in plant species without limitations.

Leveraging these resources related to network analysis can enhance our understanding in biomass production in different plant species. On one hand, several expression atlas or co-expression resources contains a number of samples from the grass species (*i.e.*, rice, wheat and maize) across stem elongation, thus making possible to identify co-expressed modules associated with stem growth or straw biomass accumulation (Borrill et al., 2016; Kudo et al., 2017; Hoopes et al., 2019; Obayashi et al., 2018). On the other hand, valuable web resources (the AspWood and NorWood database for Populus tremula and Picea abies, respectively) demonstrate the power for generating insights into wood formation and cell wall biosynthesis (Jokipii-Lukkari et al., 2017; Sundell et al., 2017). Moreover, AspWood exemplifies comparative analysis between the coexpression networks from two species, highlighting that conserved coexpression patterns are detected for many processes during wood formation (e.g., cambial growth, secondary cell wall deposition and xylem maturation). In addition, many of the cell wall metabolic regulators identified by coexpression analysis still maintain relatively conserved functions in biomass accumulation in other grasses, such as sorghum (Hennet et al., 2020). To facilitate such comparative analysis between model and non-model species, ATTED and Plant Regulomics have laid foundation for cross-study and cross-species comparisons and retrieving upstream regulators of certain genes of interest (Obayashi et al., 2018; Ran et al., 2020).

While the efforts made in co-expression analyses, three types of challenges remain in: (1) analysis of time-course expression data, (2) inference of gene regulatory networks (GRNs) from the co-expression data, and (3) comparison of co-expression modules between plant species.

First, clustering or co-expression analysis particularly for time-course data emphasizes on capturing the nonstationary time dependence in the data, for which multivariate clustering algorithms or nonlinear regression modelling methods usually perform better than the traditional clustering approaches (Heard et al., 2005). Thus, computational tools such as Smoothing spline clustering (SSClust) (Ma et al., 2006) or tcgsaSeq (Agniel and Hejblum, 2017) have been developed to identify gene clusters from time-course expression data.

Second, new computational approaches have also been available to predict gene regulatory cascade from large-scale

RNA-seq data, *e.g.* the nonparametric Bayesian and Markov clustering methods (Penfold et al., 2015a; Penfold et al., 2015b; Desai et al., 2017; Yu et al., 2019). Successful examples have been shown in crops, *i.e.* Harrington et al. (2020) report the GRNs in wheat built with the GENIE3 software. Another group develops the tool HTRgene to specifically extract stress-responsive regulatory network, highlighting the value of GRNs in underpinning particular biological questions (Ahn et al., 2019). Another key to infer GRNs is to identify overrepresented known *cis*-regulatory motifs in the gene promoters that are possibly functional in the regulation of gene expression. Computational search of *cis*-motifs in the promoter region can be readily conducted by using online websites, such as PlantCARE (Lescot et al., 2002), PlantPAN (Chow et al., 2019), or Jaspar (Fornes et al., 2020). Recently, identification of the overrepresented *cis*-motifs has been achieved by the Regulatory Sequence Analysis Tools (RSAT; Nguyen et al., 2018; Ksouri et al., 2021) and its plant-adopted version RSAT-plant (Contreras-Moreira et al., 2016; Ksouri et al., 2021). Lately, resources for visualization and efficient deployment of gene regulatory omics data (ChIP-seq, for instance) have been also available at ChIP-Hub (Fu et al., 2022) and Connec-TF (Brooks et al., 2021), making possible for transferring the TF-target regulatory relationship from the model plants to non-model species.

Last, for the comparison of coexpression networks between species, successful examples have been reported in Brassicaceae (Becker et al., 2017). ATTED-II (Obayashi et al., 2018) is a database hosting 16 co-expression platforms from nine species, allowing the comparison of co-expression modules between the species. In particular, as the resources and tools to move RNA-seq analysis toward regulomics have become mature, the Plant Regulomics database has been built, hosting a huge volume of transcriptomic and epigenomic data sets for six representative species (*i.e.*, Arabidopsis, rice, maize, soybean, tomato and wheat) and enabling the query of upstream regulators of genes (Ran et al., 2020). The Plant Regulomics database sets a nice example for future RNA-seq-centered web interface and analysis direction for other plant species.

RNA-seq data identify alternative splicing and alternative polyadenylation

While the expression atlas and co-expression analysis are based mainly on gene expression levels, RNA-seq data can also capture structural changes in the transcripts, presenting another layer of regulatory information with biological significance. Two major structural alterations are frequently detected in the transcriptome: (1) Alternative splicing (AS), a phenomenon in which particular exons of a gene may be included or excluded from the processed messenger RNA (mRNA), leading to multiple proteins encoded from a single gene; (2) Alternative polyadenylation (APA), a phenomenon in which a transcript is processed to produce multiple isoforms differing in their untranslated regions (UTRs), in most of the cases, 3'UTRs. Both AS and APA greatly increase the complexity of transcriptome or the repertoire of proteins, and are involved in the molecular, physiological and developmental pathways (Seo et al., 2013; Srivastava et al., 2018). In human, Arabidopsis and maize, respectively, ~95%, 61% and 57% of multi-exonic genes are alternatively spliced, respectively (Pan et al., 2008; Reddy et al., 2013; Wang et al., 2016). In parallel, over 80% and 75% of the genes in human and Arabidopsis respectively can produce multiple mRNA isoforms through APA (Mayr, 2016; Guo et al., 2016). The 3'UTR regions harbor *cis*-acting elements, which regulate various mRNA properties, including RNA stability, transportation, subcellular movement and translation efficiency (Srivastava et al., 2018).

Currently, computational methods for identifying differential AS have been achieved with different quantification schemas, such as those using count-based models (*i.e.*, DEXSeq (Anders et al., 2012), DSGseq (Wang et al., 2013), SpliceCompass (Aschoff et al., 2013), rMATS (Shen et al., 2012), rDiff (Drewe et al., 2013) and RNAprof (Tran et al., 2016)), and those modelling isoform ratios (*i.e.*, Cufflinks and DiffSplice) (Hu et al., 2013). Notably, some new genome assemblies of plants might not have the standard gene annotations as those of human or mouse, and not be readily compatible with some AS quantification tools or need considerable bioinformatic customizations. This issue presents somewhat a technical bar to identify and quantify AS in any plant species, even though identification of differential AS events can be done in major plant species with rMATS and CuffDiff (Liu et al., 2014). Also, new tools for identify intron retention, a particular type of AS frequently seen in plants, has been reported (Mao et al., 2017), enriching the toolbox for AS analysis.

For alternative polyadenylation, user-friendly tools compatible with the genomes of non-model plant species are relatively limited, whereas major efforts have been made to capture 3'UTRs by specific experimental protocols, such as PAT-seq (Harrison et al., 2015), 3'READs (Hoque et al., 2013), and mTAIL-seq (Lim et al., 2016). Only a handful of tools have been reported to identify 3'UTR variations and to calculate differential 3'UTRs using short-read RNA-seq data from plants. The priUTR pipeline detects differential 3'UTR events from Cufflink-derived, genome-guided transcriptome assemblies, discovering the link between 3'UTR and m6A epitranscriptomic modification (Tu and Li, 2020). APAtrap is one of the tools providing flexible and highly efficient APA detection for plant RNA-seq data (Ye et al., 2018). In addition, RNAprof detect both AS and APA events in plant RNA-seq data sets (Tran et al., 2016), while 3D RNA-seq provides three-way differential analysis: differential expression (DE), differential alternative splicing (DAS) and differential transcript usage

(DTU) of RNA-seq data (Guo et al., 2021). These recent methods promise the identification of differential AS and APA events as a regular analysis of plant RNA-seq data.

Discussion and concluding remarks

Many of the short-read, bulk RNA-seq data accumulated today from less-studied plants may be under utilized. Thus, making full use of these data by integrating RNA-seq tools presents an exciting yet challenging prospect. Still, improvements can be made in the following aspects: (1) to integrate with the long-read RNA-seq data; (2) to develop tools or optimize the current pipelines to adapt to complex plant genomes.

PacBio isoform sequencing (Iso-seq) has been the main choice for identifying full-length transcripts. Besides, high-quality full-length isoform sequencing has greatly expanded our understanding in genome annotation, isoform phasing, detection of fusion transcript and alternative splicing and alternative polyadenylation (APA). For example, automated annotation pipelines have been developed to combine the advantages of different annotation methods, including *ab initio* and protein evidence-based prediction and long-read sequencing data (Cook et al., 2018; Tardaguila et al., 2018). However, limited by the medium throughput, Iso-seq-based transcript quantification is far from affordable, especially for the project with a tight budget or a large number of samples. Thus, combining the Iso-seq-derived transcriptome and short-read RNA-seq represents an affordable strategy to both accurately capture a large number of transcripts and to quantify them (Figure 1). On another hand, ONT technology has demonstrated its potential in detection of poly(A) tail length and RNA modifications. Therefore, combination of ONT RNA-seq technologies and short-read RNA-seq results will enable novel insights into epitranscriptomic regulation. It is worth to note that while full-length transcriptomes based on the long-read sequencing technologies are apparently advantageous over the short-read RNA-seq in identification of alternative splicing and polyadenylation, tools analyzing short-read sequencing data for these purposes (such as rMATS, rDiff, RNAProf, APAtrap and priUTR) still have their particular niches because short-read RNA-seq are still dominant in the less-studied plant species and are cost affordably for most of the labs, even in high sequencing depth.

In addition, expression quantification may be complicated by other difficulties associated with plant genomes. Polyploid, including both allopolyploid and autopolyploid, are widespread in land plants. Polyploid species are frequent in biomass crops, such as the allopolyploid Miscanthus species (Mitros et al., 2020) and autopolyploid sugarcane species (Zhang et al., 2018). High levels of sequence similarity between the homo-/homoeologous alleles or gene members pose many challenges to the alignment of short reads and subsequent expression quantification. Thus, tools for the RNA-seq analysis of polyploid species or the pipelines tuned for such expression quantification are necessary (Kuo et al., 2018; Paya-Milans et al., 2018), as polyploid species have begun to be assembled recently.

Notably, short-read RNA-seq also has major merits in other plant-related research areas, especially single-cell/single nuclear RNA-seq and meta-transcriptome analysis, owing to the compatibility and cost affordability. Short-read RNA-seq facilitates meta-transcriptome characterization, profiling gene expression in a microbial community and providing a snapshot for functional exploration (Turner et al., 2013; Salazar et al., 2019). In particular, deep RNA-seq can be used to profile the gene expression from both the host and pathogens to obtain insights into plant-microbial interactions (Rudd et al., 2015).

More recently, short-read RNA-seq has been pushed to single-cell resolution due to a series of technological advancements, including robotics, microfluidics and hydrogel droplets (Zhang et al., 2019). In a few years, efforts in single-cell RNA-seq (scRNA-seq) or single-nuclei RNA-seq (snRNA-seq) have expanded from model plants (Arabidopsis, tomato and rice) to non-model species (e.g., maize and poplar), from organ development and cell differentiation to wood formation (Gutzat et al., 2020; Xu et al., 2020; Li et al., 2021; Kajala et al., 2021; Chen et al., 2021a; Wang et al., 2021; Bezrutczyk et al., 2021; Liu et al., 2022). Undoubtedly, single-cell transcriptomics are leading the fore frontier of plant single-cell biology and playing an ever-increasing role in plant research and breeding. Excellent reviews and public database on plant scRNA-seq datasets are available (Shaw et al., 2021; Chen et al., 2021b; Shahan et al., 2021). Due to the differences in several aspects of the wet- and dry-lab parts between the single-cell and bulk RNA-seq experiments, the merits of short-read RNA-seq in single-cell plant biology is beyond the scope of this review and can be found elsewhere (Shaw et al., 2021).

In summary, our work discusses a representative collection of RNA-seq analysis tools covering gene annotation, construction of expression atlas, gene regulation and alternative splicing. We emphasize that the integration of these tools will unleash the power within RNA-seq analysis, uncover the gene regulatory complexity for many less-studied plant species, and, ultimately, promote the functional genomics of these species.

Author contributions

MT and JiZ developed the conceptual outline and drafted the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (31901537), the start-up funding for young talents at Wuhan Polytechnic University (No. 53210052172 to

M.T.) and the Opening fund of Hubei Key Laboratory of Bioinorganic Chemistry & Materia Medica (No. BCMM202205 to M.T.).

Acknowledgments

We thank the invaluable time and efforts of reviewers in manuscript evaluation.

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/fpls.2022.1038109/full#supplementary-material

References

Agniel, D., and Hejblum, B. P. (2017). Variance component score test for time-course gene set analysis of longitudinal RNA-seq data. *Biostatistics* 18, 589–604. doi: 10.1093/biostatistics/kxx005

Ahn, H., Jung, I., Chae, H., Kang, D., Jung, W., and Kim, S. (2019). HTRgene: a computational method to perform the integrated analysis of multiple heterogeneous time-series data: case analysis of cold and heat stress response signaling genes in arabidopsis. *BMC Bioinf.* 20 (Suppl16), 588. doi: 10.1186/s12859-019-3072-2

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P.. (2002). *Molecular biology of the cell. 4th edition* (New York: Garland Science). Available at: https://www.ncbi.nlm.nih.gov/books/NBK26818.

Anders, S., Pyl, P. T., and Huber, W. (2014). HTSeq — a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169. doi: 10.1093/bioinformatics/btu638

Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons fromRNA-seq data. *Genome Res.* 22, 2008–2017. doi: 10.1101/gr.133744.111

Aschoff, M., Hotz-Wagenblatt, A., Glatting, K. H., Fischer, M., Eils, R., and Kdonig, R. (2013). SplicingCompass: differential splicing detection using RNA-seq data. *Bioinformatics* 29, 1141–1148. doi: 10.1093/bioinformatics/btt101

Becker, M. G., Walker, P. L., Pulgar-Vidal, N. C., and Belmonte, M. F. (2017). SeqEnrich: A tool to predict transcription factor networks from co-expressed arabidopsis and *Brassica napus* gene sets. *PloS One* 12, e0178256. doi: 10.1371/journal.pone.0178256

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., et al. (2015). The arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis* 53, 474–485. doi: 10.1002/dvg.22877

Bezrutczyk, M., Zollnet, N. R., Kruse, C. P. S., Hartwig, T., Lautwein, T., Kohrer, K., et al. (2021). Evidence for phloem loading *via* the abaxial bundle sheath cells in maize leaves. *Plant Cell.* 33, 531–547.

Bolster, D. M., Staines, D. M., Perry, E., and Kersey, P. J. (2017). Ensembl plants: Integrating tools for visualizing, mining, and analyzing plant genomic data. *Methods Mol. Biol.* 1533, 1–31. doi: 10.1007/978-1-4939-6658-5_1

Borrill, P., Ramirez-Gonzalez, R., and Uauy, C. (2016). expVIP: a customizable RNA-seq data analysis and visualiza-tion platform. *Plant Physiol.* 170, 2172–2186. doi: 10.1104/pp.15.01667

Boyles, R. E., Brenton, Z. W., and Kresovich, S. (2019). Genetic and genomic resources of sorghum to connect genotype with phenotype in contrasting environments. *Plant J.* 97, 19–39. doi: 10.1111/tpj.14113

Brockmoller, T., Ling, Z., Li, D., Gaquerel, E., Baldwin, I. T., and Xu, S. (2017). *Nicotiana attenuata* data hub (NaDH): an integrative platform for exploring genomic, transcriptomic and metabolomic data in wild tobacco. *BMC Genomics* 18, 79. doi: 10.1186/s12864-016-3465-9

Brooks, M. D., Juang, C. L., Katari, M. S., Alvarez, J. M., Pasquino, A., Shih, H. J., et al. (2021). ConnecTF: A platform to integrate transcription factor–gene interactions and validate regulatory networks. *Plant Physiol.* 185, 49–66. doi: 10.1093/plphys/kiaa012

Carrere, S., Verdenaud, M., Gough, C., Gouzy, J., and Gamas, P. (2020). LeGOO: An expertized knowledge database for the model legume *Medicago truncatula*. *Plant Cell Physiol.* 61 (1), 203–211. doi: 10.1093/pcp/pcz177

Cartolano, M., Huettel, B., Hartwig, B., Reinhardt, R., and Schneeberger, K. (2016). cDNA library enrichment of full length transcripts for SMRT long read sequencing. *PloS One* 11, e0157779. doi: 10.1371/journal.pone.0157779

Chen, Y., Tong, S., Jiang, Y., Ai, F., Feng, Y., Zhang, J., et al. (2021a). Transcriptional landscape of highly lignified poplar stems at single-cell resolution. *Genome Biol.* 22, 319.

Chen, H., Yin, X., Guo, L., Yao, J., Ding, Y., Xu, X., et al. (2021b). PlantscRNAdb: A database for plant single-cell RNA analysis. *Mol. Plant* 14, 855–857.

Chow, C. N., Lee, T. Y., Hung, C. H., Li, G. Z., Tseng, K. C., Liu, Y. H., et al. (2019). PlantPAN3.0: a new and updated resource for reconstructing transcriptional regulatory networks from ChIP-seq experiments in plants. *Nucleic Acids Res.* 47, D1155–D1163. doi: 10.1093/nar/gky1081

Chu, Q., Bai, P., Zhu, X., Zhang, X., Mao, L., Zhu, Q., et al. (2018). Characteristics of plant circular RNAs. *Brief Bioinform.* 21, 135–143. doi: 10.1093/bib/bby111

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol.* 17, 13. doi: 10.1186/s13059-016-0881-8

Contreras-Moreira, B., Castro-Mondragon, J. A., Rioualen, C., Cantalapiedra, C. P., and van Helden, J. (2016). RSAT::Plants: Motif discovery within clusters of upstream sequences in plant genomes. *Methods Mol. Biol.* 1482, 279–295. doi: 10.1007/978-1-4939-6396-6_18

Cook, D., Valle-Inclan, J. E., Pajoro, A., Rovenich, H., Thomma, B., and Faino, L. (2018). Long read annotation (LoReAn): automated eukaryotic genome annotation based on long-read cDNA sequencing. *Plant Physiol.* 179, 38–54. doi: 10.1104/pp.18.00848

Da, L., Liu, Y., Yang, J., Tian, T., She, J., Ma, X., et al. (2019). AppleMDO: A multi-dimensional omics database for apple co-expression networks and chromatin states. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.01333

Daub, C. O., Steuer, R., Selbig, J., and Kloska, S. (2004). Estimating mutual information using b-spline functions–an improved similarity measure for analysing gene expression data. *BMC Bioinf.* 5, 118. doi: 10.1186/1471-2105-5-118

Desai, J. S., Sarto, R. C., Lawas, L. M., Jagadish, S. V. K., and Doherty, C. J. (2017). Improving gene regulatory network inference by incorporating rates of transcriptional changes. *Sci. Rep.* 7, 17244. doi: 10.1038/s41598-017-17143-1

D'haeseleer, P., Liang, S., and Somogyi, R. (2000). Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16, 707–726. doi: 10.1093/bioinformatics/16.8.707

Drewe, P., Stegle, O., Hartmann, L., Kahles, A., Bohnert, R., Wachter, A., et al. (2013). Accurate detection of differential RNA processing. *Nucleic Acids Res.* 41, 5189–5198. doi: 10.1093/nar/gkt211

Farber, C. R., and Lusis, A. J. (2008). Integrating global gene expression analysis and genetics. *Adv. Genet.* 60, 571–601. doi: 10.1016/S0065-2660(07)00420-8

Fei, Y., Wang, R., Li, H., Liu, S., Zhang, H., and Huang, J. (2018) DPMIND: Degradome-based plant MiRNA-target interaction and network database. *Bioinformatics* 34, 1618–1620. doi: 10.1093/bioinformatics/btx824

Feng, J., Huang, S., Guo, Y., Liu, D., Song, J., Gao, J., et al. (2019). Plant ISOform sequencing database (PISO): a comprehensive repertory of full-length transcripts in plants. *Plant Biotechnol. J.* 17, 1001–1003. doi: 10.1111/pbi.13076

Fernandez-Pizo, N., Hass, F. B., Meyberg, R., Ullrich, K. K., Hiss, M., Perroud, P., et al. (2020). PEATmoss (Physcomitrella expression atlas tool): a unified gene expression atlas for the model plant physcomitrella patens. *Plant J.* 102, 165–177. doi: 10.1111/tpj.14607

Foox, J., Tighe, S. W., Nicolet, C. M., Zook, M., Byrska-Bishop, M., Clarke, W. E., et al. (2021). Performance assessment of DNA sequencing platforms in the ABRF next-generation sequencing study. *Nat. Biotechnol.* 39, 1129–1140.

Fornes, O., Castro-Mondragon, J. A., Khan, A., van der Lee, R., Zhang, X., Richmond, P. A., et al. (2020). JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 48, D87–D92. doi: 10.1093/nar/gkz1001

Freese, N. H., Norris, D. C., and Loraine, A. E. (2016). Integrated genome browser: visual analytics platform for genomics. *Bioinformatics* 32, 2089–2095. doi: 10.1093/bioinformatics/btw069

Fu, L., Zhu, T., Zhou, X., Yu, R., He, Z., Zhang, P., et al. (2022). ChIP-hub provides an integrative platform for exploring plant regulome. *Nat. Commun.* 13, 3413. doi: 10.1038/s41467-022-30770-1

Galli, M., Feng, F., and Gallavotti, A. (2020). Mapping regulatory determinants in plants. *Front. Genet.* 11, 591194.

Garber, M., Grabherr, M. G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods* 8, 469–477. doi: 10.1038/nmeth.1613

Gaudinier, A., and Brady, S. M. (2016). Mapping transcriptional networks in plants: Data-driven discovery of novel biological mechanisms. *Annu. Rev. Plant Biol.* 67, 575–594.

Ge, S. X., Son, E. W., and Yao, R. (2018). iDEP: an integrated web application for differential expression and pathway analysis of RNA-seq data. *BMC Bioinform.* 19, 534. doi: 10.1186/s12859-018-2486-6

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2011). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40, D1178–D1186. doi: 10.1093/nar/gkr944

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2012). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

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, 101241. doi: 10.1016/j.isci.2020.101241

Guo, C., Spinelli, M., Liu, M., Li, Q., and Liang, C. (2016). A genome-wide study of "non-3UTR" polyadenylation sites in *Arabidopsis thaliana*. *Sci. Rep.* 6, 28060. doi: 10.1038/srep28060

Guo, W., Tzioutziou, N. A., Stephen, G., Milne, I., Cailxtom, C. P., Waugh, R., et al. (2021). 3D RNA-seq: a powerful and flexible tool for rapid and accurate differential expression and alternative splicing analysis of RNA-seq data for biologists. *RNA Biol.* 18, 1574–1587. doi: 10.1080/15476286.2020.1858253

Gupta, C., and Pereira, A. (2019). Recent advances in gene function prediction using context-specific coexpression networks in plants. *F1000Research* 2019, 8.

Gutzat, R., Rembart, K., Nussbaumer, T., Hofmann, F., Pisuparti, R., Bradamante, G., et al. (2020). Arabidopsis shoot stem cells display dynamic transcription and DNA methylation patterns. *EMBO J.* 39, e103667.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., et al. (2013). *De novo* transcript sequence reconstruction from rna-seq: reference generation and analysis with trinity. *Nat. Protoc.* 8, 1494–1512. doi: 10.1038/nprot.2013.084

Haque, S., Ahmad, J. S., Clark, N. M., Williams, C. M., and Sozzani, R. (2018). Computational prediction of gene regulatory networks in plant growth and development. *Curr. Opin. Plant Biol.* 47, 96–105.

Harrington, S. A., Backhaus, A. E., Singh, A., Hassani-Pak, K., and Uauy, C. (2020). The wheat GENIE3 network provides biologically-relevant information in polyploid wheat. *G3* 10, 3675. doi: 10.1534/g3.120.401436

Harrison, P. F., Powell, D. R., Clancy, J. L., Presis, T., Boag, P. R., Traven, A., et al. (2015). PAT-seq: a method to study the integration of 3'-UTR dynamics with gene expression in the eukaryotic transcriptome. *RNA* 21, 1502–1510. doi: 10.1261/rna.048355.114

Heard, N. A., Holmes, C. C., and Stephens, D. A. (2005). A quantitative study of gene regulation involved in the immune response of *Anopheline mosquitoes*. *J. Am. Stat. Assoc.* 101, 18–29. doi: 10.1198/016214505000000187

Hennet, L., Berer, A., Trabanco, N., Ricciuti, E., Dufayard, J. F., Bocs, S., et al. (2020). Transcriptional regulation of sorghum stem composition: Key players identified through Co-expression gene network and comparative genomics analyses. *Front. Plant Sci.* 11, 224.

Hoopes, G. M., Hamilton, J. P., Wood, J. C., Esteban, E., Pasha, A., Vaillancourt, B., et al. (2019). An updated gene atlas for maize reveals organ-specific and stress-induced genes. *Plant J.* 97, 1154–1167. doi: 10.1111/tpj.14184

Hoque, M., Ji, Z., Zheng, D., Luo, W., Li, W., You, B., et al. (2013). Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. *Nat. Methods* 10, 133–139. doi: 10.1038/nmeth.2288

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, 39. doi: 10.1093/nar/gks1026

Jin, Y., Tam, O. H., Paniagua, E., and Hammell, M. (2015). TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* 31, 3593–3599. doi: 10.1093/bioinformatics/btv422

Jokipii-Lukkari, S., Sundell, D., Nilsson, O., Hvidsten, T. R., Street, N. R., and Tuominen, H. (2017). NorWood: a gene expression resource for evo-devo studies of conifer wood development. *New Phyto.* 216, 482–494. doi: 10.1111/nph.14458

Kajala, K., Gouran, M., Shaar-Moshe, L., Mason, G. A., Rodriguez-Medina, J., Kawa, D., et al. (2021). Innovation, conservation, and repurposing of gene function in root cell type development. *Cell.* 184, 3333–3348.

Kim, D., Paggi, J. M., Park, C., Park, C., and Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37, 907–915. doi: 10.1038/s41587-019-0201-4

Ksouri, N., Castro-Mondragon, J. A., Montardit-Tarda, F., van Helden, J., Contreras-Moreira, B., and Gogorcena, Y. (2021). Tuning promoter boundaries improves regulatory motif discovery in nonmodel plants: the peach example. *Plant Physiol.* 185, 1242–1258. doi: 10.1093/plphys/kiaa091

Kudapa, H., Garg, V., Chitikineni, A., and Varshney, R. K. (2018). The RNA-seq-based high resolution gene expression atlas of chickpea (*Cicer arietinum* l.) reveals dynamic spatio-temporal changes associated with growth and development. *Plant Cell Environ.* 41, 2209–2225. doi: 10.1111/pce.13210

Kudo, T., Terashima, S., Takaki, Y., Nakamura, Y., Kobayashi, M., and Yano, K. (2017). Practical utilization of OryzaExpress and plant omics data center databases to explore gene expression networks in oryza sativa and other plant species. *Methods Mol. Biol.* 1533, 229–240. doi: 10.1007/978-1-4939-6658-5_13

Kuo, T. C. Y., Hatakeyama, M., Tameshige, T., Shimizu, K. K., and Sese, J. (2018). Homeolog expression quantification methods for allopolyploids. *Brief. Bioinfo.* 21, 395–407. doi: 10.1093/bib/bby121

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

Langfelder, P., and Horvath, S. (2012). Fast r functions for robust correlations and hierarchical clustering. *J. Stat. Software* 46, 11.

Lawrence, C. J., Harper, L. C., Schaeffer, M. L., Sen, T. Z., Seigfried, T.E., and Campbell, D.A.. (2008). MaizeGDB: the maize model organism database for basic, translational, and applied research. *Intl. J. Plant Genomics* 2008, 496957. doi: 10.1155/2008/496957

Lee, J., Heath, L. S., Grene, R., and Li, S. (2019). Comparing time series transcriptome data between plants using a network module finding algorithm. *Plant Methods* 15, 61. doi: 10.1186/s13007-019-0440-x

Lee, S., Lee, T., Yang, S., and Lee, I. (2020). BarleyNet: A network-based functional omics analysis server for cultivated barley, *Hordeum vulgare* l. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.00098

Lescot, M., D'hais, P., Thijs, G., Marchal, K., Moreau, Y., Van de Peer, Y., et al. (2002). PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res.* 30, 325–327. doi: 10.1093/nar/30.1.325

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

Li, H., Dai, X., Huang, X., Xu, M., Wang, Q., Yan, X., et al. (2021). Single-cell RNA sequencing reveals a high-resolution cell atlas of xylem in populus. *J. Inte. Plant Biol.* 63, 1906–1921.

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

Liesecke, F., Daudu, D., Duge de Bernonville, R., Besseau, S., Clastre, M., Courdavault, V., et al. (2018). Ranking genome-wide correlation measurements improves microarray and RNA-seq based global and targeted co-expression networks. *Sci. Rep.* 8, 10885. doi: 10.1038/s41598-018-29077-3

Lim, J., Lee, M., Son, A., Chang, H., and Kim, V. N. (2016). mTAIL-seq reveals dynamic poly(A) tail regulation in oocyte-to-embryo development. *Genes Dev.* 30, 1671–1682. doi: 10.1101/gad.284802.116

Liu, R., and Dickerson, J. (2017). Strawberry: Fast and accurate genomeguided transcript reconstruction and quantification from RNA-seq. *PloS Comput. Biol.* 13, e1005851. doi: 10.1371/journal.pcbi.1005851

Liu, J., Li, G., Chang, Z., Yu, T., Liu, B., McMullen, R., et al. (2016). BinPacker: Packing-based *de novo* transcriptome assembly from RNA-seq data. *PloS Comput. Biol.* 12, e1004772. doi: 10.1371/journal.pcbi.1004772

Liu, G., Li, J., Li, J., Chen, Z., Yuan, P., Chen, R., et al. (2022). Single-cell transcriptome reveals the redifferentiation trajectories of the early stage of *de novo* shoot regeneration in arabidopsis thaliana. *bioRxiv*. doi: 10.1101/2022.01.01.474510

Liu, R., Loraine, A. E., and Dickerson, J. A. (2014). Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems. *BMC Bioinform.* 15, 364.

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

Ma, P., Castillo-Davis, C. I., Zhong, W., and Liu, J. S. (2006). A data-driven clustering method for time course gene expression data. *Nucleic Acids Res.* 34, 1261–1269. doi: 10.1093/nar/gkl013

Machado, F. B., Moharana, K. C., Almeida-Silva, F., Gazara, R. K., Pedrosa-Silva, F., Coelho, F. S., et al. (2020). Systematic analysis of 1,298 RNA-seq samples and construction of a comprehensive soybean (*Glycine max*) expression atlas. *Plant J.* 103, 1894–1909. doi: 10.1101/2019.12.23.886853

Makita, Y., Shimada, S., Kawashima, M., Kondou-Kuriyama, T., Toyoda, T., and Matsui, M. (2015). MOROKOSHI: transcriptome database in *Sorghum bicolor*. *Plant Cell Physiol.* 56, e6. doi: 10.1093/pcp/pcu187

Malik, L., Almodaresi, F., and Patro, R. (2018). Grouper: graph-based clustering and annotation for improved *de novo* transcriptome analysis. *Bioinformatics* 34, 3265–3272. doi: 10.1093/bioinformatics/bty378

Mao, R., Liang, C., Zhang, Y., Hao, X., and Li, J. (2017). 50/50 expressional odds of retention signifies the distinction between retained introns and constitutively spliced introns in arabidopsis thaliana. *Front. Plant Sci.* 8, 1728. doi: 10.3389/fpls.2017.01728

Marks, R. A., Hotaling, S., Frandsen, P. B., and VanBuren, R. (2021). Representation and participation across 20 years of plant genome sequencing. *Nat. Plants* 7, 1571–1578. doi: 10.1038/s41477-021-01031-8

Mayr, C. (2016). Evolution and biological roles of alternative 3'UTRs. *Trends Cell Biol.* 26, 227–237. doi: 10.1016/j.tcb.2015.10.012

Miao, Z., Han, Z., Zhang, T., Chen, S., and Ma, C. (2017). A systems approach to a spatiotemporal understanding of the drought stress response in maize. *Sci. Rep.* 7, 6590. doi: 10.1038/s41598-017-06929-y

Mitros, T., Session, A. M., James, B. T., Wu, G., Belaffif, M. B., Clark, L. V., et al. (2020). Genome biology of the paleotetraploid perennial biomass crop miscanthus. *Nat. Commun.* 11, 5442. doi: 10.1038/s41467-020-18923-6

Mullet, J., Morishige, D., McCormick, R., Truong, S., Hilley, J., McKinley, B., et al. (2014). Energy sorghum- a genetic model for the design of C4 grass bioenergy crops. *J. Exp. Bot.* 65, 3479–3489. doi: 10.1093/jxb/eru229

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., GerstEIN, M., et al. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320, 1344–1350. doi: 10.1126/science.1158441

Naithani, S., Gupta, P., Preece, J., D'Eustachio, P., Elser, J. L., Garg, P., et al. (2020). Plant reactome: a knowledgebase and resource for comparative pathway analysis. *Nucleic Acids Res.* 48, D1093–D1103. doi: 10.1093/nar/gkz996

Naithani, S., Preece, J., D'Eustachio, P., Gupta, P., Amarasinghe, V., Dharmawardhana, P. D., et al. (2017). Plant reactome: a resource for plant pathways and comparative analysis. *Nucl. Acid Res.* 45, D1029–D1039. doi: 10.1093/nar/gkw932

Nguyen, N. T. T., Contreras-Moreira, B., Castro-Mondragon, J. A., Santana-Garcia, W., Ossio, R., Robles-Espinoza, C. D., et al. (2018). RSAT 2018: regulatory sequence analysis tools 20th anniversary. *Nucleic Acids Res.* 46, W209–W216. doi: 10.1093/nar/gky317

Obayashi, T., Aoki, Y., Tadaka, S., Kagaya, Y., and Kinoshita, K. (2018). ATTED-II in 2018: A plant coexpression database based on investigation of the statistical property of the mutual rank index. *Plant Cell Physiol.* 59, e3(1–e37). doi: 10.1093/pcp/pcx191

Oikonomopoulos, S., Wang, Y. C., Djambazian, H., Badescu, D., and Ragoussis, J. (2016). Benchmarking of the Oxford nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. *Sci. Rep.* 6, 31602. doi: 10.1038/srep31602

O'Malley, R. C., Huang, S. C., Song, L., Lewsey, M. G., Bartlett, A., Nery, J. R., et al. (2016). Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell.* 165, 1280–1292.

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* 40, 1413–1415. doi: 10.1038/ng.259

Papatheodorou, I., Fonseca, N. A., Keays, M., Tang, Y. A., Barrera, E., Bazant, W., et al. (2018). Expression atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res.* 46, D246–D251. doi: 10.1093/nar/gkx1158

Patterson, J., Carpenter, E. J., Zhu, Z., An, D., Liang, X., Geng, C., et al. (2019). Impact of sequencing depth and technology on *de novo* RNA-seq assembly. *BMC Genomics* 20, 604. doi: 10.1186/s12864-019-5965-x

Paya-Milans, M., Olmstead, J. W., Nunez, G., Rinehart, T. A., and Staton, M. (2018). Comprehensive evaluation of RNA-seq analysis pipelines in diploid and polyploid species. *GigaScience* 7, giy132. doi: 10.1093/gigascience/giy132

Penfold, C. A., Millar, J. B., and Wild, D. L. (2015b). Inferring orthologous gene regulatory networks using interspecies data fusion. *Bioinformatics* 31, i97–i105. doi: 10.1093/bioinformatics/btv267

Penfold, C. A., Shifaz, A., Brown, P. E., Nicholson, A., and Wild, D. L. (2015a). CSI: a nonparametric Bayesian approach to network inference from multiple perturbed time series gene expression data. *Stat. Appl. Genet. Mol. Biol.* 14, 307–310. doi: 10.1515/sagmb-2014-0082

Perroud, P., Haas, F. B., Hiss, M., Ullrich, K. K., Alboresi, A., Amirebrahimi, M., et al. (2018). The physcomitrella patens gene atlas project: large-scale RNA-seq based expression data. *Plant J.* 95, 168–182. doi: 10.1111/tpj.13940

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., and Salzberg, S. I. (2017). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and ballgown. *Nat. Protoc.* 11, 1650–1667. doi: 10.1038/nprot.2016.095

Proost, S., Krawczyk, A., and Mutwil, M. (2017). LSTrAP: efficiently combining RNA sequencing data into co-expression networks. *BMC Bioinfo.* 18, 444. doi: 10.1186/s12859-017-1861-z

Proost, S., and Mutwil, M. (2016). Tools of the trade: studying molecular networks in plants. *Cur Opn Plant Sci.* 30, 143–150.

Ran, X., Zhao, F., Wang, Y., Liu, J., Zhang, Y., Ye, L., et al. (2020). Plant regulomics: a data-driven interface for retrieving upstream regulators from plant multi-omics data. *Plant J.* 101, 237–248. doi: 10.1111/tpj.14526

Reddy, A. S., Marquez, Y., Kalyna, M., and Barta, A. (2013). Complexity of the alternative splicing landscape in plants. *Plant Cell* 25, 3657–3683. doi: 10.1105/tpc.113.117523

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., et al. (2010). *De novo* assembly and analysis of RNA-seq data. *Nat. Methods* 7, 909–912. doi: 10.1038/nmeth.1517

Robinson, M. D., McCarthy, D. J., and Smyth, G. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Rudd, J. J., Kanyuka, K., Hassani-Pak, K., Derbyshire, M., Andongabo, A., and Devonshire, J. (2015). Transcriptome and metabolite profiling of the infection cycle of zymoseptoria tritici on wheat reveals a biphasic interaction with plant immunity involving differential pathogen chromosomal contributions and a variation on the hemibiotrophic lifestyle definition. *Plant Physiol.* 167, 1158–1185.

Ryngajllo, M., Childs, L., Lohse, M., Giorgi, F. M., Lude, A., Selbig, J., et al. (2011). Slocx: predicting subcellular localization of arabidopsis proteins leveraging gene expression data. *Front. Plant Sci.* 2. doi: 10.3389/fpls.2011.00043

Saelens, W., Cannoodt, R., and Saeys, Y. (2018). A comprehensive evaluation of module detection methods for gene expression data. *Nat. Commun.* 9, 1090. doi: 10.1038/s41467-018-03424-4

Sahraeian, S. M., Mohiyuddin, M., Sbra, R., Tilgner, H., Afshar, P. T., Au, K. F., et al. (2017). Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nat. Commun.* 8, 59. doi: 10.1038/s41467-017-00050-4

Salazar, G., Paoli, L., Alberti, A., Huerta-Cepas, J., Ruscheweyh, H. J., Cuenca, M., et al. (2019). Gene expression changes and community turnover differentially shape the global ocean metatranscriptome. *Cell.* 179, 1068–1083.

Sato, Y., Takehisa, H., Kamatsuki, K., Minami, H., Namiki, N., Ikawa, H., et al. (2013). RiceXPro version 3.0: Expanding the informatics resource for rice transcriptome. *Nucleic Acids Res.* 41, D1206–D1213. doi: 10.1093/nar/gks1125

Savelli, B., Picard, S., Roux, C., and Dunand, C. (2019). ExpressWeb: A web application for clustering and visualization of expression data. *bioRxiv*. doi: 10.1101/625939

Schulz, M. H., Zerbino, D. R., Vingron, M., and Birney, E. (2012). Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 28, 1086–1092. doi: 10.1093/bioinformatics/bts094

Schwacke, R., Ponce-Soto, G. Y., Krause, K., Bolger, A. M., Arsova, B., Hallab, A., et al. (2019). MapMan4: A refined protein classification and annotation framework applicable to multi-omics data analysis. *Mol. Plant* 12, 879–892. doi: 10.1016/j.molp.2019.01.003

Sekhon, R. S., Briskine, R., Hirsch, C. N., Myers, C. L., Springer, N. M., Buell, C. R., et al. (2013). Maize gene atlas developed by RNA sequencing and comparative evaluation of transcriptomes based on RNA sequencing and microarrays. *PloS One* 8, e61005. doi: 10.1371/journal.pone.0061005

Seoane, P., Espigares, M., Carmona, R., Polonio, A., Quintana, J., Cretazzo, E., et al. (2018). TransFlow: a modular framework for assembling and assessing accurate *de novo* transcriptomes in non-model organisms. *BMC Genomics* 19 (Suppl 14), 416. doi: 10.1186/s12859-018-2384-y

Seo, P. J., Park, M. J., and Park, C. M. (2013). Alternative splicing of transcription factors in plant responses to low temperature stress: mechanisms and functions. *Planta* 237, 1415–1424. doi: 10.1007/s00425-013-1882-4

Shahan, R., Nolan, T. M., and Benfey, P. N. (2021). Single-cell analysis of cell identity in the arabidopsis root apical meristem: insights and opportunities. *J. Exp. Botany*. 72, 6679–6686.

Sharon, D., Tilgner, H., Grubert, F., and Snyder, M. (2013). A single- molecule long- read survey of the human transcriptome. *Nat. Biotechnol.* 31, 1009–1014. doi: 10.1038/nbt.2705

Shaw, R., Tian, X., and Xu, J. (2021). Single-cell transcriptome analysis in plants: Advances and challenges. *Mol. Plant* 14, 115–126.

Shen, S., Park, J. W., Huang, J., Dittmar, K. A., Lu, Z. X., Zhou, Q., et al. (2012). MATS: a Bayesian framework for flexible detection of differential alternative splicing fromRNA-seq data. *Nucleic Acids Res.* 40, 61. doi: 10.1093/nar/gkr1291

Srivastava, A. K., Lu, Y., Zinta, G., Lang, Z., and Zhu, J. K. (2018). UTR-dependent control of gene expression in plants. *Trends Plant Sci.* 23, 248–259. doi: 10.1016/j.tplants.2017.11.003

Song, L., Shankar, D. S., and Florea, L. (2016). Rascaf: Improving genome assembly with RNA sequencing data. *Plant Genome* 9, 1–12. doi: 10.3835/plantgenome2016.03.0027

Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA Sequencing: the teenage years. *Nat. Rev. Genet.* 20, 631–656. doi: 10.1038/s41576-019-0150-2

Stelpflug, S. C., Sekhon, R. S., Vaillancourt, B., Hirsch, C. N., Buell, C. R., de Leon, N., et al. (2015). An expanded maize gene expression atlas based on RNA sequencing and its use to explore root development. *Plant Genome* 9, 1–16. doi: 10.3835/plantgenome2015.04.0025

Sullivan, A., Purohit, P. K., Freese, N. H., Pasha, A., Ewaese, J., et al. (2019). An 'eFP-seq browser' for visualizing and exploring RNA sequencing data. *Plant J.* 100, 641–654. doi: 10.1111/tpj.14468

Sun, Y., Shang, L., Zhu, Q., and Fan, L. (2022). Twenty years of plant genome sequencing: achievements and challenges. *Trends Plant Sci.* 27, 391–401. doi: 10.1016/j.tplants.2021.10.006

Sundell, D., Street, N. R., Kumar, M., Mellerowicz, E. J., Kucukoglu, M., Johnsson, C., et al. (2017). AspWood: high-spatial-resolution transcriptome profiles reveal uncharacterized modularity of wood formation in *Populus tremula*. *Plant Cell.* 29, 1585–1604. doi: 10.1105/tpc.17.00153

Tan, Q., Goh, W., and Mutwil, M. (2020). LSTrAP-cloud: A user-friendly cloud computing pipeline to infer coexpression networks. *Genes* 11, 428. doi: 10.3390/genes11040428

Tardaguila, M., de la Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F. J., Del Risco, H., et al. (2018). SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res.* 28, 396–411. doi: 10.1101/gr.222976.117

Tello-Ruiz, M. K., Naithani, S., Stein, J. C., Gupta, P., Campbell, M., Olson, A., et al. (2018). Gramene 2018: unifying comparative genomics and pathway resources for plant research. *Nucleic Acids Res.* 46, D1181–D1189. doi: 10.1093/nar/gkx1111

Tian, T., You, Q., Yan, H., Xu, W., and Su, Z. (2018). MCENet: A database for maize conditional co-expression network and network characterization collaborated with multi-dimensional omics levels. *J. Genet. Genomics* 45, 351–360. doi: 10.1016/j.jgg.2018.05.007

Tran, V. D., Souiai, O., Romero-Barrios, N., Crespi, M., and Gautheret, D. (2016). Detection of generic differential RNA processing events from RNA-seq data. *RNA Biol.* 13, 59–67. doi: 10.1080/15476286.2015.1118604

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. *Nat. Protoc.* 7, 562–578. doi: 10.1038/nprot.2012.016

Tu, M., and Li, Y. (2020). Profiling alternative 3'untranslated regions in sorghum using RNA-seq data. *Front. Genet.* 11. doi: 10.3389/fgene.2020.556749

Turner, T. R., Ramakrishnan, K., Walshaw, J., Heavens, D., Alston, M., Swarbreck, D., et al. (2013). Comparative metatranscriptomics reveals kingdom level changes in the rhizosphere microbiome of plants. *ISME J.* 7, 2248–2258.

Van Bel, M., and Coppens, F. (2017). Exploring plant co-expression and gene-gene interactions with CORNET 3.0. *Methods Mol. Biol.* 1533, 201–212. doi: 10.1007/978-1-4939-6658-5_11

Van Verk, M. C., Hickman, R., Pieterse, C. M., and Van Wees, S. C. (2013). RNA-Seq: revelation of the messengers. *Trend Plant Sci.* 18, 175–179. doi: 10.1016/j.tplants.2013.02.001

Vasilevski, A., Giorgi, F. M., Bertinetti, L., and Usadel, B. (2012). LASSO modeling of the arabidopsis thaliana seed/seedling transcriptome: a model case for detection of novel mucilage and pectin metabolism genes. *Mol. Biosyst.* 8, 2566–2574. doi: 10.1039/C2MB25096A

Waese, J., and Provart, N. J. (2016). The bio-analytic resource: Data visualization and analytic tools formultiple levels of plant biology. *Curr. Plant Biol.* 7-8, 2–5. doi: 10.1016/j.cpb.2016.12.001

Wang, Y., Huan, Q., Li, K., and Qian, W. (2021). Single-cell transcriptome atlas of the leaf and root of rice seedlings. *J. Genet. Genomics* 48, 881–898.

Wang, B., Kumar, V., Olson, A., and Ware, D. (2019). Reviving the transcriptome studies: An insight into the emergence of single-molecule transcriptome sequencing. *Front. Genet.* 10. doi: 10.3389/fgene.2019.00384

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, 164–170. doi: 10.1016/j.gene.2012.11.045

Wang, B., Tseng, E., Regulski, M., Clark, T. A., Hon, T., Jiao, Y., et al. (2016). Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat. Commun.* 7, 11708. doi: 10.1038/ncomms11708

Wang, Y., Zhang, R., Liang, Z., and Li, S. (2020). Grape-RNA: A database for the collection, evaluation, treatment, and data sharing of grape RNA-seq datasets. *Genes* 11, 315. doi: 10.3390/genes11030315

Werner, T. (2003). Promoters can contribute to the elucidation of protein function. *Trends Biotechnol.* 21, 9–13. doi: 10.1016/s0167-7799(02)00003-3

Werner, T. (2004). Proteomics and regulomics: the yin and yang of functional genomics. *Mass Spectrom. Rev.* 23, 25–33. doi: 10.1002/mas.10067

Wolfe, C. J., Kohane, I. S., and Butte, A. J. (2005). Systematic survey reveals general applicability of 'guilt-by-association' within gene coexpression networks. *BMC Bioinf.* 6, 227. doi: 10.1186/1471-2105-6-227

Xia, E., Li, F., Tong, W., Li, P., Wu, Q., Zhao, H., et al. (2019). Tea plant information archive: a comprehensive genomics and bioinformatics platform for tea plant. *Plant Biotechnol. J.* 17, 1938–1953. doi: 10.1111/pbi.13111

Xia, L., Zou, D., Sang, J., Xu, X., Yin, H., Li, M., et al. (2017). Rice expression database (RED): An integrated RNA-seq-derived gene expression database for rice. *J. Genet. Genomics* 44, 235–241. doi: 10.1016/j.jgg.2017.05.003

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., et al. (2014). SOAPdenovo-trans: *De novo* transcriptome assembly with short RNA-seq reads. *Bioinformatics* 30, 1660–1666. doi: 10.1093/bioinformatics/btu077

Xu, X., Crow, M., Rice, B. R., Li, F., Harris, B., Liu, L., et al. (2020). Single-cell RNA sequencing of developing maize ears facilitates functional analysis and trait candidate gene discovery. *Dev. Cell.* 56, 1–12.

Yano, R., Nonaka, S., and Ezura, H. (2018). Melonet-DB, a grand RNA-seq gene expression atlas in melon (*Cucumis melo* l.). *Plant Cell Physiol.* 59, e4(1–e15). doi: 10.1093/pcp/pcx193

Ye, C., Long, Y., Ji, G., Li, Q. S., and Wu, X. (2018). APAtrap: identification and quantification of alternative polyadenylation sites from RNA-seq data. *Bioinformatics* 34, 1841–1849.

Yu, H., Jiao, B., Lu, L., Wang, P., Chen, S., Liang, C., et al. (2018). NetMiner-an ensemble pipeline for building genome-wide and high-quality gene coexpression network using massive-scale RNAseq samples. *PloS One* 13, e0192613. doi: 10.1371/journal.pone.0192613

Yu, H., Lu, L., Jiao, B., and Liang, C. (2019). Systematic discovery of novel and valuable plant gene modules by large-scale RNA-seq samples. *Bioinformatics* 35, 361–364. doi: 10.1093/bioinformatics/bty642

Zhang, Y., Sun, Y., and Cole, J. R. (2014). A scalable and accurate targeted gene assembly tool (SAT-assembler) for next-generation sequencing data. *PloS Comput. Biol.* 10, e1003737. doi: 10.1371/journal.pcbi.1003737

Zheng, H., Wu, N., Chow, C. N., Tseng, K. C., Chien, C. H., Hung, Y. C., et al. (2017). EXPath tool–a systemfor comprehensively analyzing regulatory pathways and coexpression networks from high-throughput transcriptome data. *DNA Res.* 24, 371–375. doi: 10.1093/dnares/dsx009

Zhang, T. Q., Xu, Z. G., Shang, G. D., and Wang, J. W. (2019). A single-cell RNA sequencing profiles the developmental landscape of arabidopsis root. *Mol. Plant* 12, 648–660.

Zhang, J., Zhang, X., Tang, H., Zhang, Q., Hua, X., Ma, X., et al. (2018). Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* l. *Nat. Genet.* 50, 1565–1573. doi: 10.1038/s41588-018-0237-2

Zhao, X., Fu, X., Yin, C., and Lu, F. (2021). Wheat speciation and adaptation: perspectives from reticulate evolution. *aBIOTECH* 2, 386–402. doi: 10.1007/s42994-021-00047-0

Zouine, M., Maza, E., Djari, A., Lauvernier, M., Frasse, P., Smouni, A., et al. (2017). TomExpress, a unified tomato RNA-seq platform for visualization of expression data, clustering and correlation networks. *Plant J.* 92, 727–735. doi: 10.1111/tpj.13711

Zwaenepoel, A., Diels, T., Amar, D., Van Parys, T., Shamir, A., Van de Peer, Y., et al. (2018). MorphDB: Prioritizing genes for specialized metabolism pathways and gene ontology categories in plants. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00352

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 04 January 2023
DOI 10.3389/fpls.2022.1040909

OPEN ACCESS

EDITED BY
Jianping Wang,
University of Florida, United States

REVIEWED BY
Ana Pontaroli,
BioLumic Limited, New Zealand
Kranthi Varala,
Purdue University, United States

*CORRESPONDENCE
William G. Voelker
wvoelker@uncc.edu

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

RECEIVED 09 September 2022
ACCEPTED 09 December 2022
PUBLISHED 04 January 2023

CITATION
Voelker WG, Krishnan K, Chougule K, Alexander LC Jr., Lu Z, Olson A, Ware D, Songsomboon K, Ponce C, Brenton ZW, Boatwright JL and Cooper EA (2023) Ten new high-quality genome assemblies for diverse bioenergy sorghum genotypes. *Front. Plant Sci.* 13:1040909. doi: 10.3389/fpls.2022.1040909

Ten new high-quality genome assemblies for diverse bioenergy sorghum genotypes

William G. Voelker[1,2]*, Krittika Krishnan[1,2], Kapeel Chougule[3], Louie C. Alexander Jr.[1,2], Zhenyuan Lu[3], Andrew Olson[3], Doreen Ware[3,4], Kittikun Songsomboon[1,2], Cristian Ponce[1,2], Zachary W. Brenton[5,6], J. Lucas Boatwright[6,7] and Elizabeth A. Cooper[1,2]

[1]Dept. of Bioinformatics & Genomics, University of North Carolina at Charlotte, Charlotte, NC, United States, [2]North Carolina Research Campus, Kannapolis, NC, United States, [3]Cold Spring Harbor Research Laboratory, Cold Spring Harbor, NY, United States, [4]United States Department of Agriculture - Agricultural Research Service in the North Atlantic Area (USDA-ARS NAA), Robert W. Holley Center for Agriculture and Health, Ithaca, NY, United States, [5]Carolina Seed Systems, Darlington, SC, United States, [6]Advanced Plant Technology, Clemson University, Clemson, SC, United States, [7]Dept. of Plant and Environmental Sciences, Clemson University, Clemson, SC, United States

Introduction: Sorghum (*Sorghum bicolor* (L.) Moench) is an agriculturally and economically important staple crop that has immense potential as a bioenergy feedstock due to its relatively high productivity on marginal lands. To capitalize on and further improve sorghum as a potential source of sustainable biofuel, it is essential to understand the genomic mechanisms underlying complex traits related to yield, composition, and environmental adaptations.

Methods: Expanding on a recently developed mapping population, we generated *de novo* genome assemblies for 10 parental genotypes from this population and identified a comprehensive set of over 24 thousand large structural variants (SVs) and over 10.5 million single nucleotide polymorphisms (SNPs).

Results: We show that SVs and nonsynonymous SNPs are enriched in different gene categories, emphasizing the need for long read sequencing in crop species to identify novel variation. Furthermore, we highlight SVs and SNPs occurring in genes and pathways with known associations to critical bioenergy-related phenotypes and characterize the landscape of genetic differences between sweet and cellulosic genotypes.

Discussion: These resources can be integrated into both ongoing and future mapping and trait discovery for sorghum and its myriad uses including food, feed, bioenergy, and increasingly as a carbon dioxide removal mechanism.

KEYWORDS

sorghum, genome assembly and annotations, pangenomics, bioenergy, structural variation

Abbreviations: CP-NAM, Carbon Partitioning Nested Association Mapping; SV, Structural Variant; SNP, Single Nucleotide Polymorphism; TE, Transposable Element; LTR, Long Terminal Repeat; GO, Gene Ontology.

Introduction

Sorghum (*Sorghum bicolor* (L.) Moench) is a versatile, adaptable, and widely grown cereal crop that is valued for its efficiency, drought tolerance, and ability to grow in marginalized soils (Wayne Smith and Frederiksen, 2000). Present-day genotypes exhibit extensive genetic, phenotypic, morphological, and physiological diversity which stems both from their historical spread and modern breeding efforts aimed at optimizing sorghum for different end uses. With its wealth of naturally occurring genetic diversity and advantageous traits, sorghum has enormous value as a sustainable, fast-growing, and high-yielding bioenergy crop (Calviño and Messing, 2012).

Currently, sorghum is classified into four major ideotypes: grain, sweet, cellulosic, and forage. All of these types can be used in different bioenergy production methods (Wu et al., 2010), but to fully capitalize on their potential, it is essential to gain a better understanding of the genomic changes driving traits related to yield, carbon partitioning, and local adaptation. However, these types of traits are often difficult to dissect due to the nature of their underlying genetic architecture (Brachi et al., 2011), which can involve hundreds to thousands of genes and complex mutations that are not easily captured by short-read sequencing.

Structural genomic mutations are an important source of variation in many species, and can play key roles in phenotypic diversification and evolution. Advances in sequencing technology, especially the advent of high-throughput long-read sequencing, have made the detection of structural variants feasible in many plant species where these types of changes were previously uncharacterized. More recently, there has also been a surge in the generation of pan-genomic data for a number of important crop species, which has offered exciting new insights into the extensive diversity of these plants and the potential influence of complex structural mutations on agronomically important phenotypes (2022; Golicz et al., 2016; Zhang et al., 2019; Danilevicz et al., 2020; Zhou et al., 2020; Della Coletta et al., 2021; Hufford et al., 2021; Li et al., 2021).

Previous genomic work in sorghum has linked structural mutations to a number of key traits including dwarfing (Multani et al., 2003), juicy stalks (Zhang et al., 2018), chilling tolerance (Wu et al., 2019), and flowering time (Li et al., 2018). A whole-genome comparison of the sweet sorghum genotype 'Rio' with 'BTx623,' (a short-statured, early maturing grain sorghum) found hundreds of gene presence/absence variations (PAVs), several of which occurred among known sucrose transporters (Cooper et al., 2019). Furthermore, a genome-wide association study (GWAS) exploring the genetic architecture of bioenergy-related traits found that a large deletion in a sorghum-specific iron transporter was linked to stalk sugar accumulation (2020; Brenton et al., 2016). Most recently, we undertook a broad survey of genome-wide deletions in a panel of nearly 350 diverse sorghum accessions, and found large deletions in multiple genes related to biotic and abiotic stress responses that were unique to particular geographic origins, and appeared to play a role in local adaptation (Songsomboon et al., 2021).

Taken together, these results suggest that unraveling complex traits in sorghum and other crops will require a comprehensive picture of both structural and single nucleotide mutations. In this study, we have expanded on the recently published Carbon-Partitioning Nested Association Mapping (CP-NAM) population that was developed and publicly released as a key genetic resource for the characterization and improvement of sorghum for multiple different end uses (2022; Boatwright et al., 2021; Kumar et al., 2022). We generated high-quality *de novo* genome assemblies for 10 of the CP-NAM parents and used these genomes to identify millions of novel variants, including a number of large structural variants (SVs) occurring in genes or pathways that could be essential for optimizing sorghum as a bioenergy feedstock.

Materials and methods

Sample collection and sequencing

Seeds for each genotype were ordered from the U.S. Department of Agriculture's Germplasm Resource Information Network (GRIN)(https://www.ars-grin.gov/) and grown in the greenhouses at the North Carolina Research Campus (NCRC) in Kannapolis, NC. High-molecular-weight DNA was extracted from each sample using a modified high-salt CTAB extraction protocol (Inglis et al., 2018). Purified DNA was sent to the David H. Murdock Research Institute (DHRMI) for quality control, library preparation, and sequencing on a PacBio Sequel I system.

De novo assembly

Raw subreads for each genotype were combined and converted to FASTQ format using the bam2fastx toolkit from PacBio. Reads were then corrected, trimmed, and assembled using Canu(v2.1.1) (Koren et al., 2017). For one of the genotypes, 'Grassl', Canu failed to produce contigs due to reduced read coverage after trimming, so the final assembly was instead produced using Flye(v2.9) with the Canu corrected reads (Kolmogorov et al., 2019).

The resulting contigs for all genotypes were scaffolded into chromosomes using RagTag (v2.1.0) (Alonge et al., 2021) and the parameters '-r -g 1 -m 10000000'. Contigs were ordered based on their alignment to the BTx623 v3.1 reference genome (Paterson et al., 2009) with minimap2 (Li, 2018). RagTag was run *without* the correction step to avoid unnecessary fragmentation of the contigs and unplaced contigs were discarded. Assembled genome metrics were assessed both

before and after scaffolding using QUAST(5.2.0) (Gurevich et al., 2013).

Annotation

Protein and non-coding genes were annotated by building a pan-gene working set using representative pan-gene models selected from a comparative analysis of gene family trees from 18 Sorghum genomes (McCormick et al., 2018; Deschamps et al., 2018; Cooper et al., 2019; Wang et al., 2021; Tao et al., 2021) sourced from SorghumBase(https://www.sorghumbase.org/). This pan-gene representative was propagated onto the 10 sorghum genome assemblies using Liftoff (v1.6.3) (Shumate and Salzberg, 2021) with parameters(-a 0.95 -s 0.95 -p 20 -copies -cds -polish). The gene structures were updated with available transcriptome evidence from Btx623 using PASA (v2.4.1) (Haas et al., 2003). Additional improvements to structural annotations were done in PASA using full length sequenced cDNAs and sorghum ESTs downloaded from NCBI using the query (EST[Keyword]) AND sorghum[Organism]. The working set was assigned Annotation Edit Distance(AED) scores using MAKER-P (v3.0) (Campbell et al., 2014) and transcripts with AED score < 1 were classified as protein coding. Those with AED=1 were further filtered to keep any non-BTx623 based models with a minimum protein length of 50 amino acids and a complete CDS as protein coding. The remaining models with AED=1 were classified as non-coding. Gene ID assignment was made as per the existing nomenclature schema established for Sorghum reference genomes (McCormick et al., 2018).

On average, approximately 55 thousand working sets of models were generated for each sorghum line, out of which an average of 41 thousand were coding and roughly 13 thousand were non-coding (Supplementary Table 1). More than half (61%) of the protein coding models mapped to a BTx623 reference gene, along with 23% of the non-coding models (Supplementary Figure 1A). On average ~42% single exon genes come from the reference BTx623 genome, while ~52% come from non-BTX623 lines. ~92% of the single exon genes that are not found in non-sorghum reference genomes, are found in two or more sorghum accessions. ~29% of these have a supporting AED score of less than 1 (Supplementary Figure 1B). Functional domain identification was completed with InterProScan (v5.38-76.0) (Jones et al., 2014). TRaCE (Olson and Ware, 2020) was used to assign canonical transcripts based on domain coverage, protein length, and similarity to transcripts assembled by Stringtie. Finally, the protein coding annotations were imported to Ensembl core databases, verified, and validated for translation using the Ensembl API (Stabenau et al., 2004).

In order to assign gene ages, protein sequences were aligned to the canonical translations of gene models from *Zea mays*, *Oryza sativa*, *Brachypodium distachyon*, and *Arabidopsis thaliana* obtained from Gramene release 62 (Tello-Ruiz et al., 2020) using USEARCH v11.0.667_i86linux32 (Edgar, 2010). If there was a hit with minimum sequence identity of 50% (-id 0.5) to an *Arabidopsis* protein, the gene was classified as being from Viridiplanteae, if there was a hit to rice the gene was classified as Poaceae, and if a hit was to maize the gene was classified as Andropogoneae. If there were no hits then the gene was classified as sorghum specific.

Repeat analysis

Transposable elements (TEs) were identified and annotated in each genome using EDTA (Ou et al., 2019). TE-greedy-nester (Lexa et al., 2020) was used to further annotate both complete and fragmented Long Terminal Repeat (LTR) retrotransposons. Sequence divergence in the LTR regions was used to estimate retrotransposon age (SanMiguel et al., 1998; Jedlicka et al., 2020). The left and right LTR sequences were extracted from the assembled genomes using the coordinates reported by TE-greedy-nester and the getfasta tool from the BEDTools package(v2.29.0) (Quinlan and Hall, 2010). For each TE, the two LTR sequences were aligned using Clustal-W (Thompson et al., 1994) as implemented in the R package msa (Bodenhofer et al., 2015). Genetic distance was calculated based on the K80 model using the dist.dna function in the R package phangorn (Schliep, 2011). The time of divergence was calculated based on the equation T=K/(2 * r) (Bowen and McDonald, 2001), where T is the time of divergence, K is the genetic distance, and r is the substitution rate. A value of 0.013 mutations per million years was used for r, consistent with the molecular clock rate for LTRs estimated in rice (Ma and Bennetzen, 2004). To determine if any of the shell genes across all the genotypes had overlaps with TEs, a custom python script was used to match the annotated shell gene coordinates with TE coordinates identified by TE-greedy-nester (Lexa et al., 2020). A flanking sequence of 1000bp upstream and downstream was considered. In order to find the overlaps, only the contigs that were placed into chromosomes by RagTag(v2.1.0) (Alonge et al., 2021) were included since the unplaced contig sequences were not a part of TE-greedy-nester analysis.

Variant calling

Filtered and scaffolded reads were realigned to the BTx623 reference genome using the nucmer program from the MUMmer(v4.0) package (Delcher et al., 2003; Marçais et al., 2018) with the following parameters '-c 100 -b 500 -l 50'. Alignments were filtered using the delta-filter program from the MUMmer package with the parameters '-m -i 90 -l 100' and converted to coordinate files using show-coords with the

parameters '-THrd'. Variants were then called using Syri(v1.6) (Goel et al., 2019).

Individual Syri VCF files were split by variant type (SNPs, Deletions, Insertions, Inversions, and Translocations) resulting in separate files for each variant type for each genotype. Insertions or deletions smaller than 50 bp were classified as small indels while those equal to or larger than 50 bp were classified as SVs. More complex SV types that could not be validated with raw reads were not considered for further analysis.

The Syri program produces a nonstandard VCF format which includes information on variants from overlapping syntenic blocks. This can result in duplicated variants and fragmented insertions that must be addressed before subsequent analysis with downstream tools. Duplicates of existing variants were removed for all variant types, and fragmented insertions were combined into single variants (Supplementary Figure 2). These processed variant files were then zipped and indexed using bgzip and tabix (Li et al., 2009) and then merged across genotypes using the merge function from the bcftools package with the parameters '-0 -I 'ChrB:join, Parent:join,DupType:join,modified:join' -O v'. This resulted in one variant file for each type of variant that included the genotypes for all individuals. Insertions, deletions, and SNPs were then annotated using SIFT (v2.4) (Vaser et al., 2016) and the BTx623 version 3.1.1 annotation to identify overlap with genes for insertions and deletions and missense prediction for single nucleotide variants.

Phylogeny

Gene PAVs was called from pan-gene lift-off annotation information using custom python scripts. As per default liftoff parameters, gene presence was identified with a threshold of 95 percent similarity. PAVs for each genotype were encoded as a binary vector (with 0 indicating gene absence, and 1 indicating presence). Distance between genotypes was then calculated using the dist() function from the stats(v3.6.2) package in R using the Jaccard distance, and a phylogenetic tree was constructed using the NJ() function from the phangorn package. The SNP phylogeny used to confirm the PAV phylogeny was created using SNPs called from the program Syri. Similar to the PAV tree, this phylogeny was built based on a presence/absence binary matrix of SNPs. Genetic distance was calculated using the dist() function and the NJ() function in R.

Gene ontology analysis

Gene ontology (GO) terms for genes affected by large insertions and deletions or nonsynonymous SNPs were curated from the publicly available annotation information file associated with BTx623 v3.1.1 in phytozome (https://phytozome-next.jgi.doe.gov/). GO enrichment analysis was performed using the R package topGO(v1.0) (Alexa and Rahnenfuhrer, 2016). The classic Fisher's Test was used to assess significance of enriched terms, and terms with a p-value <0.05 were considered significant and kept for further analysis. Redundant and highly similar GO terms were defined and reduced based on semantic similarity using the R packages AnnotationForge (Carlson and Pages, 2022) and rrvgo (Sayols, 2020).

Results

Assembly quality and characteristics

To capture the genetic diversity of bioenergy sorghum, we sequenced the parents of the previously established CP-NAM population, which included globally diverse genotypes representative of sweet, cellulosic, grain and forage type bioenergy sorghums (Boatwright et al., 2021) (Table 1). The initial contig-level assemblies showed a range of N50 values, with

TABLE 1 Genotype origins, races, and types.

Name	Alternate ID	Race	Origin	Type
Grassl	PI 154844	Caudatum	Uganda	Sweet & Cellulosic
PI 329311	IS 11069	Durra	Ethiopia	Cellulosic
PI 506069	Mbonou	Guinea-bicolor	Togo	Cellulosic
PI 510757	AP79-714	Durra	Cameroon	Cellulosic
Chinese Amber	PI 22913	Bicolor	China	Sweet
Rio	PI 563295	Durra-caudatum	USA	Sweet
Leoti	PI 586454	Kafir-bicolor	Hungary	Sweet
PI 229841	IS 2382	Kafir	South Africa	Grain

(Continued)

10.3389/fpls.2022.1040909

TABLE 1 Continued

Name	Alternate ID	Race	Origin	Type
PI 297155	IS 13633	Kafir	Uganda	Grain, Forage
PI 655972	Pink Kafir	Kafir	USA	Forage

Information adapted from GRIN and (Boatwright et al., 2021).

the lowest being 176 kb and the highest at over 3 Mbp (Supplementary Table 2). The three sweet genotypes in particular had a higher number of raw reads and more contiguous assemblies than the other types (Figures 1A, B), most likely as a result of differences in the effectiveness of the extraction protocol. After scaffolding and filtering unplaced contigs, all 10 genotypes showed similar levels of high contiguity, with final assembly sizes that were 90-98% the size of the BTx623 reference genome and over 90% of known BTx623 genes contained within the scaffolds (Figures 1C, D).

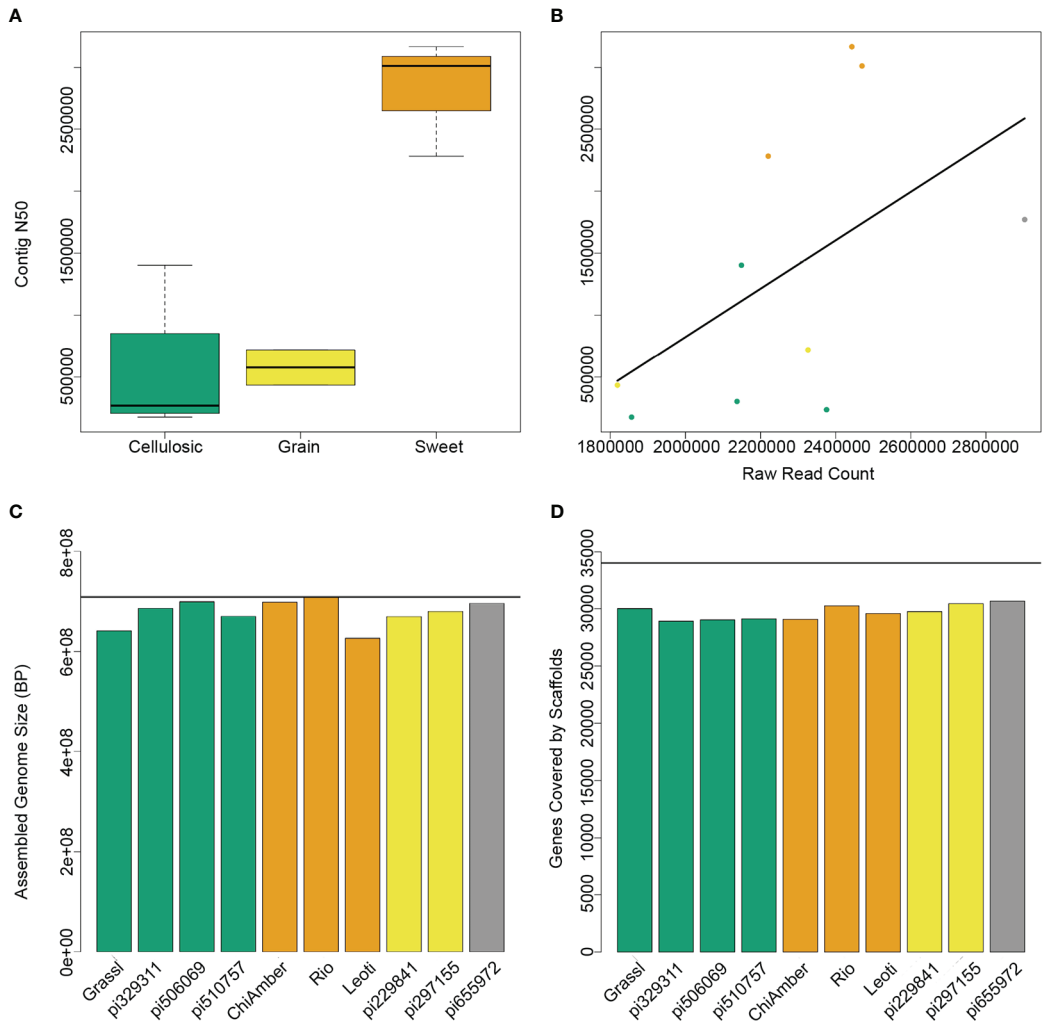


FIGURE 1
Assembly metrics for 10 sorghum genotypes. **(A)** Contig N50 levels for different ideotypes show higher contiguity for sweet genotypes. **(B)** Raw read counts prior to assembly are highly correlated with contig N50, and sweet genotypes (orange) have higher read counts than cellulosic (green) or grain (yellow) genotypes. **(C)** Assembled genome size after scaffolding and filtering for each genotype shows that despite differences in mean contig size, the final assemblies for both sweet and non-sweet types are very close to the expected reference genome size (horizontal black line). **(D)** The number of BTx623 genes contained within the final scaffolds is very similar across all genotypes regardless of type.

Gene annotation

Genes shared across deeper evolutionary time scales were more conserved than sorghum-specific genes (Figure 2). The sweet genotypes show slightly more conserved genes when compared to other genotypes (Figure 2). Out of 62,044 genes annotated in the pan-genome, around 36.69 percent(22,762 genes) were found to be core to all genotypes, 50.32 percent (31,218 genes) were shell genes (present in more than one genome, but not all of the genomes), and 12.99 percent(8,064 genes) were found to be cloud genes (unique to a single genome) (Supplementary Figure 3A). The majority of shell genes were present in 9 of 10 genomes, with the second largest proportion of shell genes being present in 2 of 10 genomes (Supplementary Figure 3B). Of shell genes identified, 44 and 45 were identified to be exclusive to all sweet and all non-sweet genotypes respectively. Only 1-2 percent of shell genes in each genotype overlapped with or were flanked by LTRs, indicating that transposable element activity was not mediating the majority of observed gene content variation (Supplementary Table 3).

Genomic landscape of variation

Over 10.5 million single nucleotide variants were called across the 10 genomes, as well as over 7.4 million small indels and over 24 thousand large structural variants (insertions and deletions ≥ 50 bp) (Figure 3, Tables 2, 3). Well over half (~65%) of these variants were defined as cloud variation (Table 3), while the remaining variants were mostly shell. Only a small handful of core variants were present in all of the genotypes except the BTx623 reference. Phylogenetic relationships were inferred using gene presence/absence to estimate genetic distance (Supplementary Figure 4A), demonstrating that sweet, cellulosic, and grain genotypes come from separate clades within the category of bioenergy-type sorghum. These results were confirmed by SNP phylogeny (Supplementary Figure 4B).

Genes affected by structural variants and SNPs

There was a total of 171,000 SNPs that were found to be both located in genic regions and encoding nonsynonymous variants, and more than 2.5 thousand large SVs present in genic regions. GO enrichment analyses of affected genes revealed that SNPs and SVs tended to impact distinct categories of genes (Figure 4), with protein phosphorylation being the only significant category to appear in both datasets.

In addition to protein phosphorylation, genes impacted by large insertions or deletions showed enrichment in GO categories related to Golgi vesicle transport, photosynthesis, nucleoside metabolism, protein modifications, and programmed cell death (Figure 4B). Nonsynonymous SNPs, on the other hand, were enriched in genes involved in pollen-pistil interactions, cell wall biogenesis, cell proliferation, posttranscriptional regulation and polysaccharide metabolism (Figure 4A).

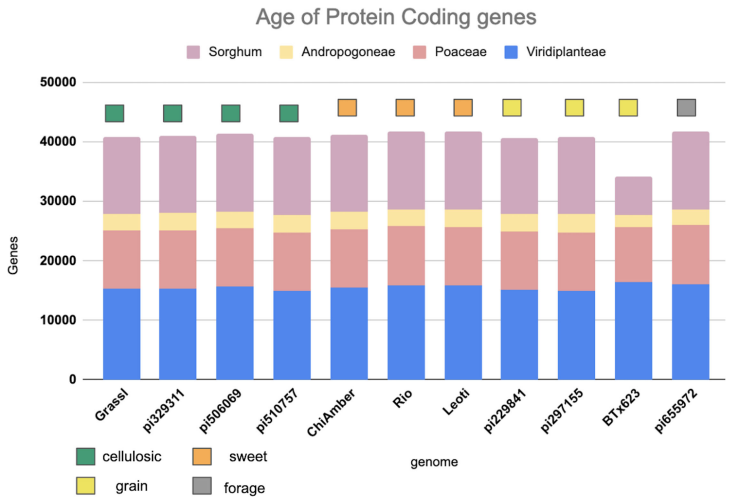


FIGURE 2
Age of protein coding genes among the sorghum lines based on minimum sequence identity. Bar color indicates the level of phylogenetic conservation, with blue indicating genes conserved across monocots and dicots; peach indicating the proportion of genes shared among the grasses; yellow indicating the proportion of genes shared between sorghum and maize, and light purple representing the proportion of sorghum-specific genes.

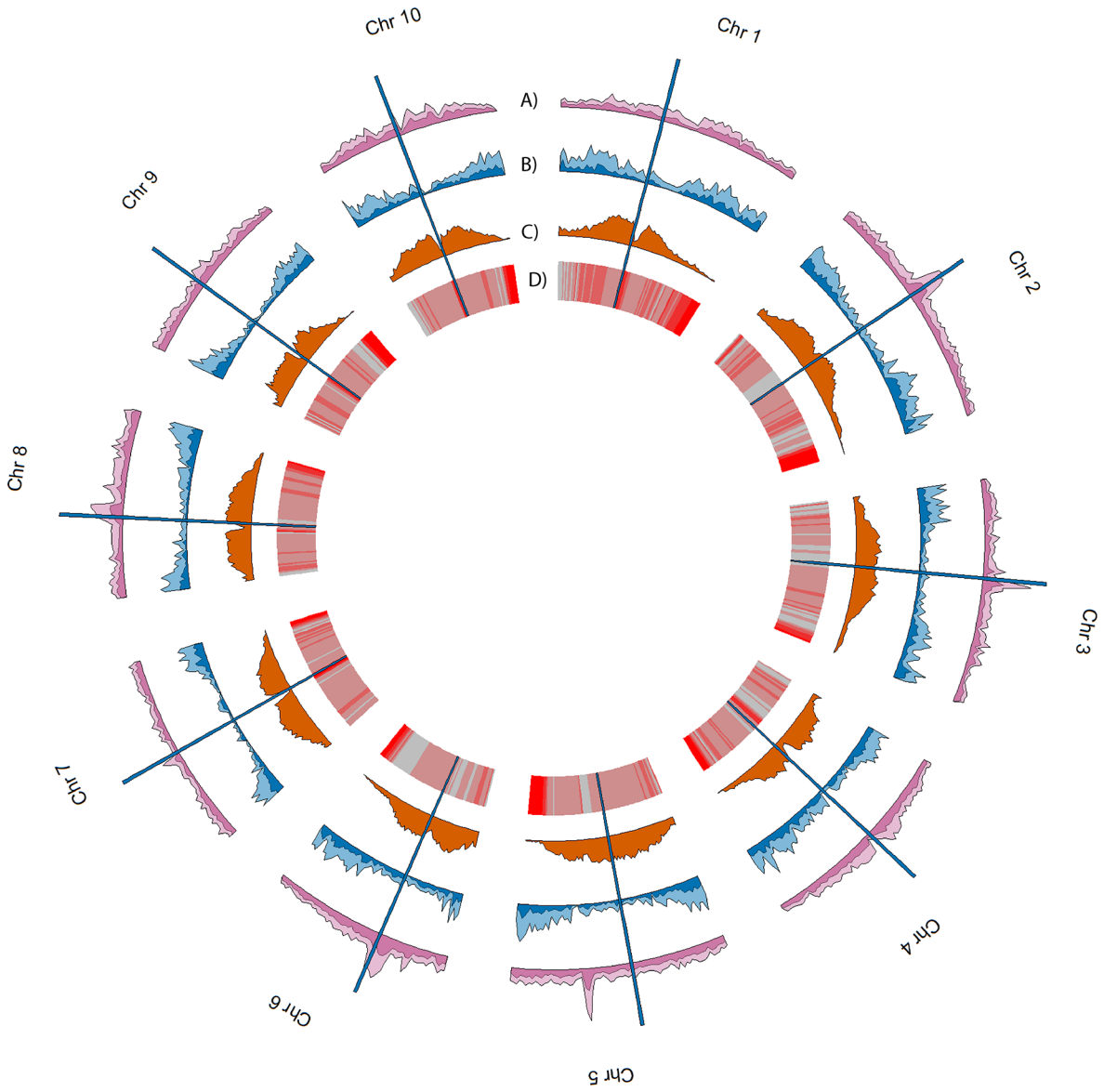


FIGURE 3
Genomic landscape of variation averaged across the 10 genomes. Density estimates in tracks A-C were performed in 1Mb non-overlapping sliding windows. **(A)** and **(B)** respectively show average SNP density and average SV density, with lighter colors indicating cloud variants and darker colors indicating shell and core variants. **(C)** shows the average TE density, and **(D)** shows TE age averaged across 1Mb sliding windows. Red indicates younger TEs while gray indicates older. Vertical blue bars spanning all tracks indicate the approximate position of the centromeres of each chromosome.

TABLE 2 Variants found in each NAM parent genotype.

Genotype	Deletions (bp>=50)	Insertions (bp>=50)	Indels (bp<50)	SNPs	Nonsynonymous
Grassl	2,721	1,714	976,703	2,659,850	37,265
PI 329311	3,560	1,956	1,319,281	3,321,035	47,482
PI 506069	3,531	1,865	888,425	3,003,469	47,555

(Continued)

TABLE 2 Continued

Genotype	Deletions (bp>=50)	Insertions (bp>=50)	Indels (bp<50)	SNPs	Nonsynonymous
PI 510757	2,952	1,919	1,593,228	2,859,852	44,168
Chinese Amber	3,560	1,744	994,023	2,975,137	48,780
Rio	2,563	1,791	717,304	2,119,637	35,714
Leoti	3,279	1,435	785,360	2,790,452	43,473
PI 229841	2,830	1,490	1,447,030	2,546,090	41,679
PI 297155	2,412	1,335	1,151,594	2,052,203	34,863
PI 655972	2,401	1,113	631,705	1,953,106	32,758

TABLE 3 Core vs. Shell vs. Cloud variants.

Type	Deletions	Insertions	Total SVs	Indels	SNPs
Core	34	28	62	12,231	103,065
Shell	6,306	2,250	8,556	1,246,552	5,245,181
Cloud	7,855	8,232	16,087	6,195,713	5,416,344
Total	**14,195**	**10,510**	**24,705**	**7,454,496**	**10,764,590**

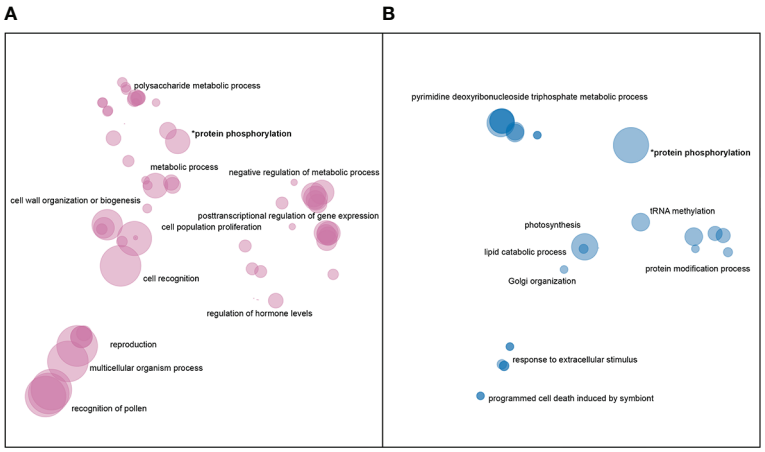


FIGURE 4
Enriched GO terms for genes impacted by **(A)** nonsynonymous SNPs and **(B)** large SVs. GO terms in each dataset were clustered and plotted based on semantic similarity as described in the Materials and Methods. Circle size is proportional to p-value, with larger circles indicating more significant terms.

Repeat analysis

Overall, the TE composition was highly similar across all 10 genotypes (Figures 5, 3), with the LTR-Gypsy superfamily comprising the majority of elements. The age analysis revealed an abundance of younger TEs, with a mean age of 1.28 million years old along with a high frequency of very young TEs approximately 0.1 million years old and very few old TEs (6-8 million years) (Figure 5; Supplementary Figure 5). Most (97.5%) of the TEs were non-nested, with TE-greedy-nester reporting the presence of only a handful (2.5%) of nested TEs. The overall distribution of TE age followed a similar pattern across all of the

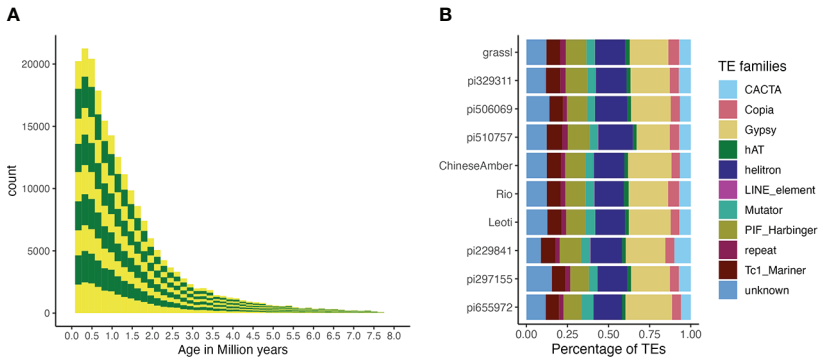


FIGURE 5
TE age and composition. **(A)** A stacked bar plot describing the distribution of TE counts by age across all genotypes. Alternating colors indicate different genotypes, and distributions are stacked in the order of the labels in figure 5b (i.e., the bottom yellow distribution shows the TE age frequencies for pi655972, while the top shows the distribution for Grassl)., T he Y-axis is number of TEs and the X-axis is their age in millions of years. **(B)** The proportion of superfamilies of TEs based on average counts of each superfamily across all genomes.

genotypes, with younger TEs being randomly distributed throughout the genome (Figure 3, Supplementary Figure 6A-J) as previously observed by (Paterson et al., 2009).

Differences in sweet and non-sweet genotypes

Structural variants that were present in all three sweet genotypes (Leoti, ChineseAmber, and Rio) but either absent from or rare among non-sweet genotypes, were significantly enriched among genes with functions related to metal ion transport, in particular iron ion transport, as well as genes involved in oxidative stress response, cell cycle arrest, and phosphatidylserine biosynthetic processes. Conversely, variants found only in all of the non-sweet genotypes tended to impact very different categories of genes, such as those involved in glycolytic processes, cytochrome assembly, and both RNA and DNA regulation (Figure 6).

Discussion

Unraveling the molecular mechanisms controlling complex traits such as carbon partitioning, yield, and stress response is an essential step for crop improvement efforts aimed at creating effective and sustainable bioenergy feedstocks for the future. However, not only do these types of traits often involve changes in large numbers of genes, but an ever-increasing number of pan-genomics studies in crop plants have demonstrated that these changes can encompass complex structural mutations in addition to SNPs (2022; Cooper et al., 2019; Zhang et al., 2019; Brenton et al., 2020; Zhou et al., 2020; Hufford et al., 2021; Songsomboon et al., 2021). Therefore, the development of multiple reference-quality genomes within crop species is critical to the exploration of complex genetic architectures and has clear benefits when compared to a single reference genome, especially in the case of larger structural variants (Della Coletta et al., 2021). By *de novo* assembling 10 new high-quality genomes for the parents of the CP-NAM population (Boatwright et al., 2022), we have been able to uncover millions of novel variants, including thousands of large insertions and deletions.

Importantly, we found that SVs within coding regions impacted different types of genes compared to SNPs, highlighting the importance of incorporating both into future trait mapping studies. Many nonsynonymous SNPs that were segregating among the genotypes occurred in gene categories that have previously been linked to carbon allocation in sorghum and other closely related species. For instance, protein phosphorylation induces key signaling cascades in plants that control a variety of processes, and protein kinases have been shown to be highly differentially expressed in both sweet sorghum (Cooper et al., 2019) and sugarcane (Waclawovsky et al., 2010) during stem sugar accumulation. Similarly, genes involved in the regulation of plant hormones such as auxin were also enriched for non-coding SNPs, and these pathways are known to be essential for vegetative plant growth and stem elongation, both of which are key phenotypes for biomass accumulation (Kebrom et al., 2017).

Like SNPs, gene-impacting SVs were also found to affect many genes related to protein phosphorylation; in fact, this was the top category among genes containing large variants. But other categories enriched for high-impact insertions and deletions were distinct from the SNP dataset, and contained many genes involved in pathways related to both abiotic and biotic stress responses, which has been observed before in diverse bioenergy sorghums (Songsomboon et al., 2021).

10.3389/fpls.2022.1040909

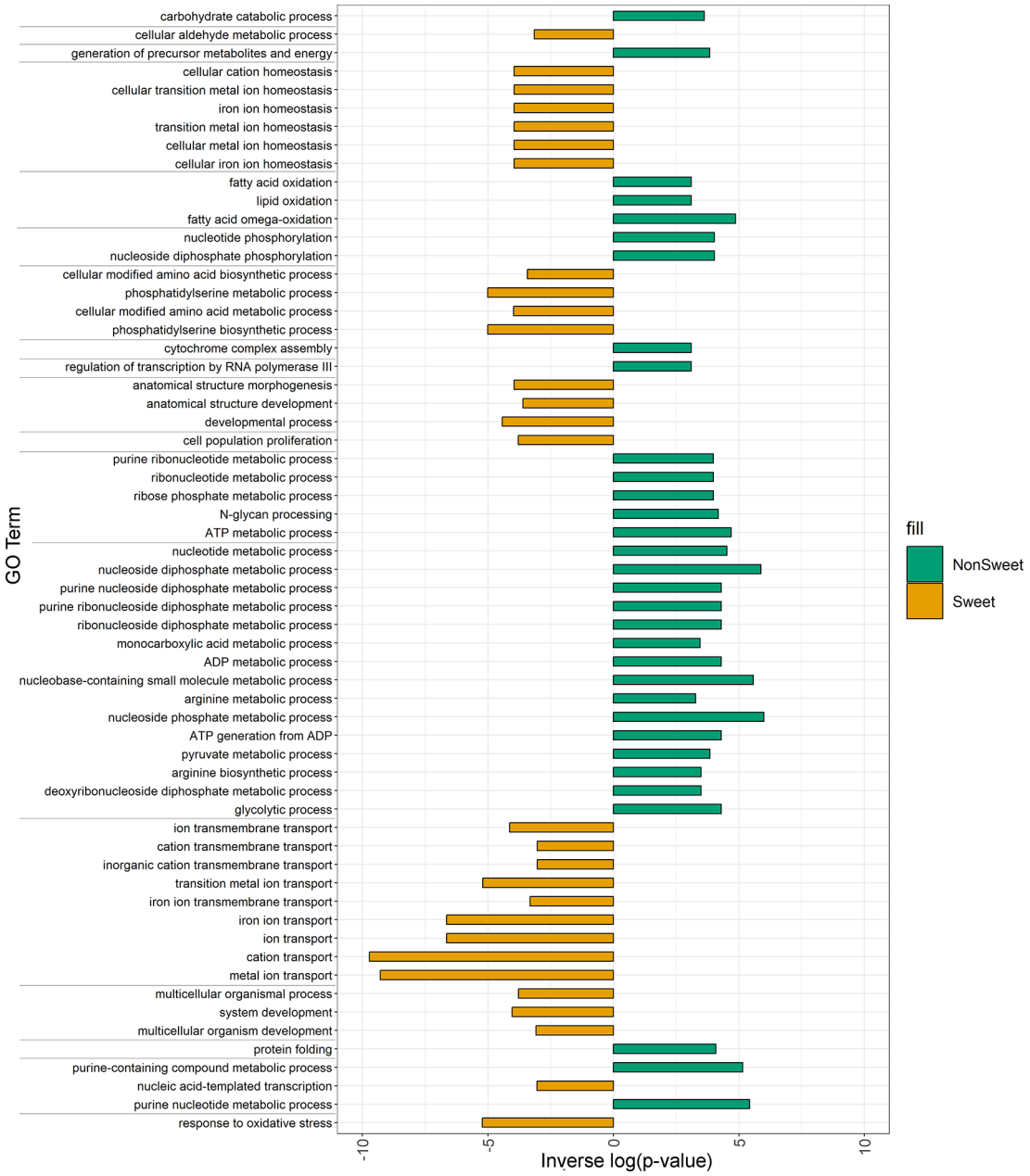


FIGURE 6
Enriched GO terms for genes impacted by SVs and Indels in both Non-Sweet and Sweet Genotypes. Orange bars indicated gene categories in sweet genotypes that were significantly impacted ($p<0.05$). Green bars indicated gene categories in non-Sweet genotypes that were significantly impacted ($p<0.05$). The length of each bar corresponds to significance (-log(p-value)). Terms have been clustered and sorted based on semantic similarity.

Additionally our study identified structural variants affecting genes involved in tRNA nucleoside modifications, programmed cell death in response to symbionts, and photosynthetic light response, all of which were previously identified by other studies as GO terms of interest in relation to sorghum stress response (Ortiz et al., 2017; Wang et al., 2017).

SVs strictly occurring in either sweet or non-sweet genotypes also offer unique insights into the differences between these types that could be key to dissecting differences in carbon allocation in sorghum. Of particular interest is the fact that SVs restricted to sweet sorghum genotypes affected many genes related to metal metabolism and iron transport. This connection between iron

transport and sugar accumulation has been observed in other comparative genomic studies of sorghum (2020; Brenton et al., 2016; Cooper et al., 2019), and appears to be a key factor distinguishing sweet sorghums from both cellulosic and grain types.

Over a third of protein coding genes and over 75 percent of noncoding genes annotated in this study did not map back to the Btx623 reference genome. With a growing number of studies illustrating the importance of noncoding DNA and RNA as potential regulatory elements (Waititu et al., 2020), it is evident that large pan-genome annotations are vital in quickly identifying and annotating potential regulatory 'pseudo-genes' as well as protein coding genes that are divergent from the common reference. Previous pan-genome studies in sorghum and maize have identified high levels of gene content variation, with 53-64 percent of genes identified as non-core (Tao et al., 2021; Ruperao et al., 2021; Hufford et al., 2021). We corroborate these findings with about 63 percent of our genes being identified as either shell or cloud to our population, despite this particular population lacking wild representation, indicating relatively high amounts of latent variation, even among domesticated varieties of sorghum.

Taken together, our results demonstrate the value of exploring genome-wide patterns of both SNPs and larger structural variants to gain new insights into the genetic architectures of complex and agronomically important traits. To advance both sorghum breeding efforts and our understanding of crop plant evolution, we have generated this new extensive dataset that is publicly available through SorghumBase (Gladman et al., 2022) and which can be readily integrated into an already valuable genetic resource for future mapping 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 below: https://www.ebi.ac.uk/ena, PRJEB55613 https://ftp.sorghumbase.org/Voelker_et_al_2022/, N/A.

Author contributions

WV: Writing, variant analysis, created figures and tables, performed scaffolding. KK: Performed TE analysis, alignments, variant calling and wrote corresponding methods sections. LA: Aided in scripting of figure creation and filtering of variants. KS: Growing and DNA Extraction of plant material. CP: Aided in genome assembly. KC, ZL, AO: Gene and transposable element annotations. DW: Experimental design, writing. ZB: designed CP-NAM population, provided genetic materials JB: development and release of CP-NAMs. EC: Writing, created figures, conceived the project, advised, and helped direct analysis. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by startup funds from UNC Charlotte and the United States Department of Agriculture grant USDA-ARS 8062-21000-041-00D.

Acknowledgments

The authors would like to thank S. Kresovich and M. Myers for providing plant materials, J. Lotito and N.C. State for providing and overseeing the greenhouse and growth chambers facilities at the N.C. Research Campus, and the DHMRI Genomics Core for providing sequencing services. The authors would also like to acknowledge the University Research Computing team at UNC Charlotte and S. Blanchard for providing essential IT support and resources. Finally, the authors thank two reviewers for their insightful comments and suggestions.

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.2022.1040909/full#supplementary-material

References

Alexa, A., and Rahnenfuhrer, J. (2016). topGO: Enrichment analysis for gene ontology. *R package version 2.50.0.* doi: 10.18129/B9.bioc.topGO

Alonge, M., Lebeigle, L., Kirsche, M., Aganezov, S., Wang, X. Z., Lippman, B., et al. (2021). Automated assembly scaffolding elevates a new tomato system for high-throughput genome editing. *bioRxiv*. doi: 10.1101/2021.11.18.469135

Boatwright, J. L., Brenton, Z. W., Boyles, R. E., Sapkota, S., Myers, M. T., Jordan, K. E., et al. (2021). Genetic characterization of a sorghum bicolor multiparent mapping population emphasizing carbon-partitioning dynamics. *G3* 11 (4): jkab060. doi: 10.1093/g3journal/jkab060

Boatwright, J. L., Sapkota, S., Myers, M., Kumar, N., Cox, A., Jordan, K. E., et al. (2022). Dissecting the genetic architecture of carbon partitioning in sorghum using multiscale phenotypes. *Front. Plant Sci.* 13, 790005. doi: 10.3389/fpls.2022.790005

Bodenhofer, U., Bonatesta, E., Horejš-Kainrath, C., and Hochreiter, S. (2015). Msa: an r package for multiple sequence alignment. *Bioinformatics* 31 (24), 3997–3999. doi: 10.1093/bioinformatics/btv494

Bowen, N. J., and McDonald, J. F. (2001). Drosophila euchromatic LTR retrotransposons are much younger than the host species in which they reside. *Genome Res.* 11 (9), 1527–1540. doi: 10.1101/gr.164201

Brachi, B., Morris, G. P., and Borevitz, J. O. (2011). Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biol.* 12 (10), 232. doi: 10.1186/gb-2011-12-10-232

Brenton, Z. W., Cooper, E. A., Myers, M. T., Boyles, R. E., Shakoor, N., Zielinski, K. J., et al. (2016). A genomic resource for the development, improvement, and exploitation of sorghum for bioenergy. *Genetics* 204 (1), 21–33. doi: 10.1534/genetics.115.183947

Brenton, Z. W., Juengst, B. T., Cooper, E. A., Myers, M. T., Jordan, K. E., S. M., et al. (2020). Species-specific duplication event associated with elevated levels of nonstructural carbohydrates in sorghum bicolor. *G3* 10 (5), 1511–1520. doi: 10.1534/g3.119.400921

Calviño, M., and Messing, J. (2012). Sweet sorghum as a model system for bioenergy crops. *Curr. Opin. Biotechnol.* 23 (3), 323–329. doi: 10.1016/j.copbio.2011.12.002

Campbell, M. S., Holt, C., Moore, B., and Yandell, M. (2014). Genome annotation and curation using MAKER and MAKER-p. *Curr. Protoc. Bioinf.* 48:4.11, 1–39. doi: 10.1002/0471250953.bi0411s48

Carlson, M., and Pages, H. (2022). AnnotationForge: code for building annotation database packages. *R package version 1.40.0*. doi: 10.18129/B9.bioc.AnnotationForge

Cooper, E. A., Brenton, Z. W., Flinn, B. S., Jenkins, J., Shu, S., Flowers, D., et al. (2019). A new reference genome for sorghum bicolor reveals high levels of sequence similarity between sweet and grain genotypes: implications for the genetics of sugar metabolism. *BMC Genomics* 20 (1), 420. doi: 10.1186/s12864-019-5734-x

Danilevicz, M. F., Fernandez, C. G. T., Marsh, J. I., Bayer, P. E., and Edwards, D. (2020). Plant pangenomics: approaches, applications and advancements. *Curr. Opin. Plant Biol.* 54, 18–25. doi: 10.1016/j.pbi.2019.12.005

Delcher, A. L., Salzberg, S. L., and Phillippy, A. M. (2003). Using MUMmer to identify similar regions in large sequence sets. *Curr. Protoc. Bioinf.* Chapter 10, 10.3. doi: 10.1002/0471250953.bi1003s00

Della Coletta, R., Qiu, Y., Ou, S., M., Hufford, B., and Hirsch, C. N. (2021). How the pan-genome is changing crop genomics and improvement. *Genome Biol.* 22 (1), 3. doi: 10.1186/s13059-020-02224-8

Deschamps, S., Zhang, Y., Llaca, V., Ye, L., Sanyal, A., King, M., et al. (2018). A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. *Bioinformatics* 9 (1), 2460–2461. doi: 10.1038/s41467-018-07271-1

Edgar, RC. (2010). Search and clustering orders of magnitude faster than BLAST. *Nat. Commun.* 9 (1), 4844. doi: 10.1093/bioinformatics/btq461

Gladman, N., Olson, A., Wei, S., Chougule, K., Lu, Z., Tello-Ruiz, M., et al. (2022). SorghumBase: a web-based portal for sorghum genetic information and community advancement. *Planta* 255 (2), 35. doi: 10.1007/s00425-022-03821-6

Goel, M., Sun, H., Jiao, W.-B., and Schneeberger, K. (2019). SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* 20 (1), 277. doi: 10.1186/s13059-019-1911-0

Golicz, A. A., Batley, J., and Edwards, D. (2016). Towards plant pangenomics. *Plant Biotechnol. J.* 14 (4), 1099–1105. doi: 10.1111/pbi.12499

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29 (8), 1072–1075. doi: 10.1093/bioinformatics/btt086

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K. Jr., Hannick, L. I., et al (2003). Improving the arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 31 (19), 5654–5666. doi: 10.1093/nar/gkg770

Hufford, M. B., Seetharam, A. S., Woodhouse, M. R., Chougule, K. M., Ou, S., Liu, J., et al. (2021). *De novo* assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* 373 (6555), 655–662. doi: 10.1126/science.abg528

Inglis, P. W., Castro, M., Pappas, R., Resende, L.V., and Grattapaglia, D. (2018). Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for high-throughput SNP genotyping and sequencing applications. *PloS One* 13 (10), e0206085. doi: 10.1371/journal.pone.0206085

Jedlicka, P., Lexa, M., and Kejnovsky, E. (2020). What can long terminal repeats tell us about the age of LTR retrotransposons, gene conversion and ectopic recombination? *Front. Plant Sci.* 11, 644. doi: 10.3389/fpls.2020.00644

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics* 30 (9), 1236–1240. doi: 10.1093/bioinformatics/btu031

Kebrom, T. H., McKinley, B., and Mullet, J. E. (2017). Dynamics of gene expression during development and expansion of vegetative stem internodes of bioenergy sorghum. *Biotechnol. Biofuels* 10, 159. doi: 10.1186/s13068-017-0848-3

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* 37 (5), 540–546. doi: 10.1038/s41587-019-0072-8

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., Phillippy, A. M., et al. (2017). Canu: scalable and accurate long-read assembly *via* adaptive k-mer weighting and repeat separation. *Genome Res.* 27 (5), 722–736. doi: 10.1101/gr.215087.116

Kumar, N., Brenton, Z., Myers, M. T., Boyles, R. E., Sapkota, S., Boatwright, J. L., et al. (2022). Registration of the sorghum carbon-partitioning nested association mapping (CP-NAM) population. *J. Plant Regist.* 16 (3), 656-663. doi: 10.1002/plr2.20229

Lexa, M., Jedlicka, P., Vanat, I., Cervenansky, M., and Kejnovsky, E. (2020). TE-greedy-nester: structure-based detection of LTR retrotransposons and their nesting. *Bioinformatics* 36 (20), 4991–4999. doi: 10.1093/bioinformatics/btaa632

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34 (18), 3094–3100. doi: 10.1093/bioinformatics/bty191

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence Alignment/Map format and SAMtools. *Bioinformatics* 25 (16), 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, X., Guo, T., Mu, Q., Li, X., and Yu, J. (2018). Genomic and environmental determinants and their interplay underlying phenotypic plasticity. *Proc. Natl. Acad. Sci. United States America* 115 (26), 6679–6684. doi: 10.1073/pnas.1718326115

Li, J., Yuan, D., Wang, P., Wang, Q., Sun, M., Liu, Z., et al. (2021). Cotton pan-genome retrieves the lost sequences and genes during domestication and selection. *Genome Biol.* 22 (1), 119. doi: 10.1186/s13059-021-02351-w

Ma, J., and Bennetzen, J. L. (2004). Rapid recent growth and divergence of rice nuclear genomes. *Proceedings of the national academy of sciences of the united states of America* 101, 34, 12404–12410. doi: 10.1073/pnas.0403715101

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., Zimin, A., et al. (2018). MUMmer4: A fast and versatile genome alignment system. *PloS Comput. Biol.* 14 (1), e1005944. doi: 10.1371/journal.pcbi.1005944

McCormick, R. F., Truong, S. K., Sreedasyam, A., Jenkins, J., Shu, S., Sims, D., et al. (2018). The sorghum bicolor reference genome: Improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *Plant J.: For Cell Mol. Biol.* 93 (2), 338–354. doi: 10.1111/tpj.13781

Multani, D. S., Briggs, S. P., Chamberlin, M. A., Blakeslee, J. J., Murphy, A. S., Johal, G. S., et al. (2003). Loss of an MDR transporter in compact stalks of maize br2 and sorghum dw3 mutants. *Science* 302 (5642), 81–84. doi: 10.1126/science.1086072

Olson, A. J., and Ware, D. (2020). Ranked choice voting for representative transcripts with TraCE. *Cold Spring Harbor Lab* 38 (1), 261-264. doi: 10.1101/2020.12.15.422742

Ortiz, D., Hu, J., and Salas Fernandez, M. G. (2017). Genetic architecture of photosynthesis in sorghum bicolor under non-stress and cold stress conditions. *J. Exp. Bot.* 68 (16), 4545–4557. doi: 10.1093/jxb/erx276

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R. A., Hellinga, A. J., et al. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* 20 (1), 275. doi: 10.1186/s13059-019-1905-y

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., et al. (2009). The sorghum bicolor genome and the diversification of grasses. *Nature* 457 (7229), 551–556. doi: 10.1038/nature07723

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26 (6), 841–842. doi: 10.1093/bioinformatics/btq033

Ruperao, P., Thirunavukkarasu, N., Gandham, P., Selvanayagam, S., Govindaraj, M., Nebie, B., et al. (2021). Sorghum pan-genome explores the functional utility for genomic-assisted breeding to accelerate the genetic gain. *Front. Plant Sci.* 12 (June), 666342. doi: 10.3389/fpls.2021.666342

SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y., and Bennetzen, J. L. (1998). The paleontology of intergene retrotransposons of maize. *Nat. Genet.* 20 (1), 43–45. doi: 10.1038/1695

Sayols, S. (2020). *Rrvgo: a bioconductor package to reduce and visualize gene ontology terms*. doi: 10.18129/B9.bioc.rrvgo

Schliep, K. P. (2011). Phangorn: phylogenetic analysis in r. *Bioinformatics* 27 (4), 592–593. doi: 10.1093/bioinformatics/btq706

Shumate, A., and Salzberg, S. L. (2021). Liftoff: Accurate mapping of gene annotations. *Bioinformatics* 37 (12), 1639–1643. doi: 10.1093/bioinformatics/btaa1016

Songsomboon, K., Brenton, Z., Heuser, J., Kresovich, S., Shakoor, N., Mockler, T., et al. (2021). Genomic patterns of structural variation among diverse genotypes of sorghum bicolor and a potential role for deletions in local adaptation. *G3* 11 (7). doi: 10.1093/g3journal/jkab154

Stabenau, A., McVicker, G., Melsopp, C., Proctor, G., Clamp, M., and Birney, E. (2004). The ensembl core software libraries. *Genome Res.* 14 (5), 929–933. doi: 10.1101/gr.1857204

Tao, Y., Luo, H., Xu, J., Cruickshank, A., Zhao, X., Teng, F., et al. (2021). Extensive variation within the pan-genome of cultivated and wild sorghum. *Nat. Plants* 7 (6), 766–773. doi: 10.1038/s41477-021-00925-x

Tello-Ruiz, M. K., Naithani, S., Gupta, P., Olson, A., Wei, S., Preece, J., et al. (2020). Gramene 2021: harnessing the power of comparative genomics and pathways for plant research. *Nucleic Acids Res.* 49 (d1), D1452–D1463. doi: 10.1093/nar/gkaa979

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22 (22), 4673–4680. doi: 10.1093/nar/22.22.4673

Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M., and Ng, P. C. (2016). SIFT missense predictions for genomes. *Nat. Protoc.* 11 (1), 1–9. doi: 10.1038/nprot.2015.123

Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* 19 (2), 327–335. doi: 10.1101/gr.073585.107

Waclawovsky, A. J., Sato, P. M., Lembke, C. G., Moore, P. H., and Souza, G. M. (2010). Sugarcane for bioenergy production: an assessment of yield and regulation of sucrose content. *Plant Biotechnol. J.* 8 (3), 263–276. doi: 10.1111/j.1467-7652.2009.00491.x

Waititu, J. K., Zhang, C., Liu, J., and Wang, H. (2020). Plant non-coding RNAs: Origin, biogenesis, mode of action and their roles in abiotic stress. *Int. J. Mol. Sci.* 21 (21). doi: 10.3390/ijms21218401

Wang, Y., Pang, C., Li, X., Hu, Z., Lv, Z., Zheng, B., Chen, P., et al, et al. (2017). Identification of tRNA nucleoside modification genes critical for stress response and development in rice and arabidopsis. *BMC Plant Biol.* 17 (1), 261. doi: 10.1186/s12870-017-1206-0

Wang, B., Jiao, Y., Chougule, K., Olson, A. J., Huang, J., Llaca, V., et al (2021). Pan-genome analysis in sorghum highlights the extent of genomic variation and sugarcane aphid resistance genes. *bioRxiv*. doi: 10.1101/2021.01.03.424980

Wayne Smith, C., and Frederiksen, R. A. (2000). *Sorghum: Origin, history, technology, and production* (John Wiley & Sons).

Wu, X., Staggenborg, S., Propheter, J. L., Rooney, W. L., Yu, J., Wang, D., et al. (2010). Features of sweet sorghum juice and their performance in ethanol fermentation. *Ind. Crops Prod.* 31 (1), 164–170. doi: 10.1016/j.indcrop.2009.10.006

Wu, Y., Guo, T., Mu, Q., Wang, J., Li, X., Wu, Y., et al. (2019). Allelochemicals targeted to balance competing selections in African agroecosystems. *Nat. Plants* 5 (12), 1229–1236. doi: 10.1038/s41477-019-0563-0

Zhang, L.-M., Leng, C.-Y., Luo, H., Wu, X.-Y., Liu, Z.-Q., Zhang, Y.-M., et al. (2018). Sweet sorghum originated through selection of dry, a plant-specific NAC transcription factor gene. *Plant Cell* 30 (10), 2286–2307. doi: 10.1105/tpc.18.00313

Zhang, B., Zhu, W., Diao, S., Wu, X., Lu, J., Ding, C., et al. (2019). The poplar pangenome provides insights into the evolutionary history of the genus. *Commun. Biol.* 2, 215. doi: 10.1038/s42003-019-0474-7

Zhou, Y., Chebotarov, D., Kudrna, D., Llaca, V., Lee, S., Rajasekar, S., et al. (2020). A platinum standard pan-genome resource that represents the population structure of Asian rice. *Sci. Data* 7 (1), 113. doi: 10.1038/s41597-020-0438-2

Zhou, Y., Zhang, Z., Bao, Z., Li, H., Lyu, Y., Zan, Y., et al. (2022). Graph pangenome captures missing heritability and empowers tomato breeding. *Nature* 606 (7914), 527–534. doi: 10.1038/s41586-022-04808-9

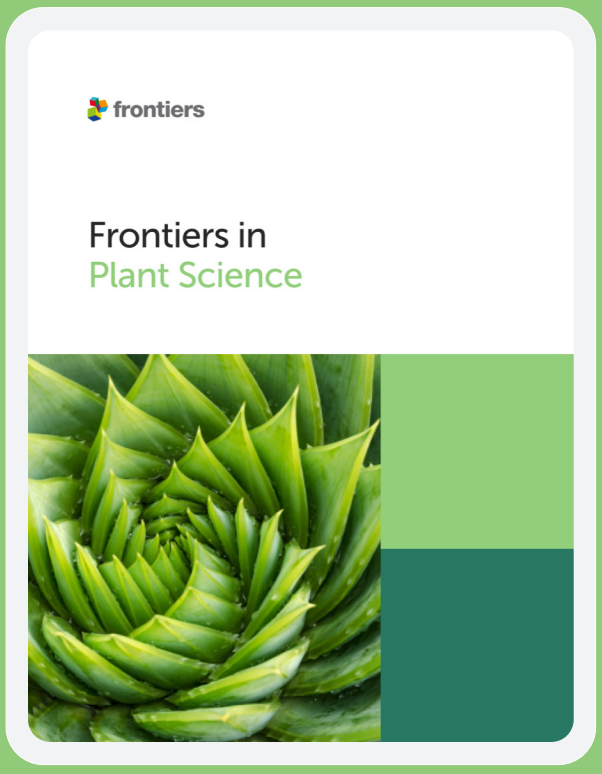
Frontiers in Plant Science

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