COMPARATIVE GENOMICS AND FUNCTIONAL GENOMICS ANALYSES IN PLANTS

EDITED BY: Xiaoming Song, Wei Chen, Weike Duan, Zhiyong Liu, Hongjian Wan and Rong Zhou

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics



frontiers

Frontiers eBook Copyright Statement

The copyright in the text of individual articles in this eBook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this eBook is the property of Frontiers.

Each article within this eBook, and the eBook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this eBook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or eBook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-88966-972-1
DOI 10.3389/978-2-88966-972-1

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

COMPARATIVE GENOMICS AND FUNCTIONAL GENOMICS ANALYSES IN PLANTS

Topic Editors:
Xiaoming Song, North China University of Science and Technology, China
Wei Chen, Chengdu University of Traditional Chinese Medicine, China
Weike Duan, Huaiyin Institute of Technology, China
Zhiyong Liu, Shenyang Agricultural University, China
Hongjian Wan, Zhejiang Academy of Agricultural Sciences, China
Rong Zhou, Aarhus University, Denmark

Citation: Song, X., Chen, W., Duan, W., Liu, Z., Wan, H., Zhou, R., eds. (2021). Comparative Genomics and Functional Genomics Analyses in Plants. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88966-972-1

Table of Contents

07 ***Editorial: Comparative Genomics and Functional Genomics Analyses in Plants***
Xiaoming Song, Zhiyong Liu, Hongjian Wan, Wei Chen, Rong Zhou and Weike Duan

10 ***Genome-Wide Analyses of the Genetic Screening of C_2H_2-Type Zinc Finger Transcription Factors and Abiotic and Biotic Stress Responses in Tomato (*Solanum lycopersicum*) Based on RNA-Seq Data***
Tingting Zhao, Tairu Wu, Jia Zhang, Ziyu Wang, Tong Pei, Huanhuan Yang, Jingfu Li and Xiangyang Xu

27 ***Micro-Evolution Analysis Reveals Diverged Patterns of Polyol Transporters in Seven Gramineae Crops***
Weilong Kong, Tong Sun, Chenhao Zhang, Yalin Qiang and Yangsheng Li

41 ***Chromosome Level Genome Assembly of* Andrographis paniculata**
Ying Liang, Shanshan Chen, Kunhua Wei, Zijiang Yang, Shengchang Duan, Yuan Du, Peng Qu, Jianhua Miao, Wei Chen and Yang Dong

55 ***The Complete Chloroplast Genome Sequences of 14* Curcuma *Species: Insights Into Genome Evolution and Phylogenetic Relationships Within Zingiberales***
Heng Liang, Yan Zhang, Jiabin Deng, Gang Gao, Chunbang Ding, Li Zhang and Ruiwu Yang

72 ***Molecular Characterization, Expression Pattern and Function Analysis of Glycine-Rich Protein Genes Under Stresses in Chinese Cabbage (*Brassica rapa *L. ssp.* pekinensis*)***
Xiaonan Lu, Yaxiong Cheng, Ming Gao, Meilan Li and Xiaoyong Xu

86 ***The Genome Sequence of Alpine* Megacarpaea delavayi *Identifies Species-Specific Whole-Genome Duplication***
Qiao Yang, Hao Bi, Wenjie Yang, Ting Li, Jiebei Jiang, Lei Zhang, Jianquan Liu and Quanjun Hu

95 ***Correlation Analysis of Expression Profile and Quantitative iTRAQ-LC-MS/MS Proteomics Reveals Resistance Mechanism Against TuMV in Chinese Cabbage (*Brassica rapa *ssp.* pekinensis*)***
Shanwu Lyu, Liwei Gao, Rujia Zhang, Changwei Zhang and Xilin Hou

109 ***Similarities and Differences of Photosynthesis Establishment Related mRNAs and Novel lncRNAs in Early Seedlings (Coleoptile/Cotyledon vs. True Leaf) of Rice and* Arabidopsis**
Yafei Shi, Jian Chen and Xin Hou

127 ***Comprehensive Analysis of Respiratory Burst Oxidase Homologs (Rboh) Gene Family and Function of* GbRboh5/18 *on Verticillium Wilt Resistance in* Gossypium barbadense**
Ying Chang, Bo Li, Qian Shi, Rui Geng, Shuaipeng Geng, Jinlei Liu, Yuanyuan Zhang and Yingfan Cai

141 ***Transcriptomic Analysis Reveals Important Roles of Lignin and Flavonoid Biosynthetic Pathways in Rice Thermotolerance During Reproductive Stage***
Zhenzhen Cai, Fengyu He, Xin Feng, Tong Liang, Hongwei Wang Shuangcheng Ding and Xiaohai Tian

156 ***Genome-Wide Identification and Characterization of the bHLH Transcription Factor Family in Pepper (*Capsicum annuum *L.)***
Zhishuo Zhang, Juan Chen, Chengliang Liang, Feng Liu, Xilin Hou and Xuexiao Zou

170 ***Genome-Wide Identification, Structural, and Gene Expression Analysis of BRI1-EMS-Suppressor 1 Transcription Factor Family in* Cucumis sativus**
Si Ma, Tingting Ji, Meiting Liang, Shihui Li, Yongqiang Tian and Lihong Gao

182 ***Genome-Wide Identification, Expression Profiling, and Evolution of Phosphate Transporter Gene Family in Green Algae***
Long Wang, Liang Xiao, Haiyan Yang, Guanglei Chen, Houqing Zeng, Hongyu Zhao and Yiyong Zhu

194 ***Recurrent Drought Conditions Enhance the Induction of Drought Stress Memory Genes in* Glycine max *L.***
Yeon-Ki Kim, Songhwa Chae, Nam-Iee Oh, Nguyen Hoai Nguyen and Jong-Joo Cheong

203 ***Identification of Proteins of Tobacco Mosaic Virus by Using a Method of Feature Extraction***
Yu-Miao Chen, Xin-Ping Zu and Dan Li

214 ***Integrating Sugar Metabolism With Transport: Elevation of Endogenous Cell Wall Invertase Activity Up-Regulates* SlHT2 *and* SlSWEET12c *Expression for Early Fruit Development in Tomato***
Lei Ru, Yong He, Zhujun Zhu, John W. Patrick and Yong-Ling Ruan

223 ***HpeNet: Co-expression Network Database for* de novo *Transcriptome Assembly of* Paeonia lactiflora *Pall***
Minghao Sheng, Jiajie She, Wenying Xu, Yan Hong, Zhen Su and Xiaodong Zhang

236 ***Characterization of Chloroplast Genomes From Two* Salvia *Medicinal Plants and Gene Transfer Among Their Mitochondrial and Chloroplast Genomes***
Chengwen Gao, Chuanhong Wu, Qian Zhang, Xia Zhao, Mingxuan Wu, Ruirui Chen, Yalin Zhao and Zhiqiang Li

246 ***First Draft Genome Assembly of the Seaweed* Sargassum fusiforme**
Shengqin Wang, Lidong Lin, Yijian Shi, Weiguo Qian, Nan Li, Xiufeng Yan, Huixi Zou and Mingjiang Wu

251 ***Genome-Wide Identification and Expression Profile of the SNAT Gene Family in Tobacco (*Nicotiana tabacum*)***
Jiemei Zhang, Zhengping Yao, Renjun Zhang, Zongmin Mou, Honghui Yin, Tianyang Xu, Dake Zhao and Suiyun Chen

261 ***Genome-Wide Identification of Circular RNAs in Response to Low-Temperature Stress in Tomato Leaves***
Xuedong Yang, Yahui Liu, Hui Zhang, Jinyu Wang, Gaurav Zinta, Shangbo Xie, Weimin Zhu and Wen-Feng Nie

273 ***Transcriptional Analysis of Metabolic Pathways and Regulatory Mechanisms of Essential Oil Biosynthesis in the Leaves of* Cinnamomum camphora *(L.) Presl***
Jiexi Hou, Jie Zhang, Beihong Zhang, Xiaofang Jin, Haiyan Zhang and Zhinong Jin

283 *Improved* de novo *Assembly of the Achlorophyllous Orchid* Gastrodia elata
Shanshan Chen, Xiao Wang, Yangzi Wang, Guanghui Zhang, Wanling Song, Xiao Dong, Michael L. Arnold, Wen Wang, Jianhua Miao, Wei Chen and Yang Dong

293 *Transcriptomes of Different Tissues of Flax (*Linum usitatissimum *L.) Cultivars With Diverse Characteristics*
Alexey A. Dmitriev, Roman O. Novakovskiy, Elena N. Pushkova, Tatiana A. Rozhmina, Alexander A. Zhuchenko, Nadezhda L. Bolsheva, Artemy D. Beniaminov, Vladimir A. Mitkevich, Liubov V. Povkhova, Ekaterina M. Dvorianinova, Anastasiya V. Snezhkina, Anna V. Kudryavtseva, George S. Krasnov and Nataliya V. Melnikova

299 *Genome-Wide Identification and Expression Analysis of the Class III Peroxidase Gene Family in Potato (*Solanum tuberosum *L.)*
Xuanshong Yang, Jiazheng Yuan, Wenbin Luo, Mingyue Qin, Jiahan Yang, Weiren Wu and Xiaofang Xie

314 *Identification and Expression Analysis of the* CsMYB *Gene Family in Root Knot Nematode-Resistant and Susceptible Cucumbers*
Chunyan Cheng, Qingrong Li, Xing Wang, Ying Li, Chuntao Qian, Ji Li, Qunfeng Lou, Molly Jahn and Jinfeng Chen

331 *Comparative Transcriptome Reveals the Genes' Adaption to Herkogamy of* Lumnitzera littorea *(Jack) Voigt*
Ying Zhang, Yukai Chen, Yan Zhou, Jingwen Zhang, He Bai and Chunfang Zheng

341 *Initial Complete Chloroplast Genomes of* Alchemilla *(Rosaceae): Comparative Analysis and Phylogenetic Relationships*
Peninah Cheptoo Rono, Xiang Dong, Jia-Xin Yang, Fredrick Munyao Mutie, Millicent A. Oulo, Itambo Malombe, Paul M. Kirika, Guang-Wan Hu and Qing-Feng Wang

357 *UDP-Glucose Dehydrogenases: Identification, Expression, and Function Analyses in Upland Cotton (*Gossypium hirsutum*)*
Tingting Jia, Qun Ge, Shuya Zhang, Zhen Zhang, Aiying Liu, Senmiao Fan, Xiao Jiang, Yulong Feng, Lipeng Zhang, Doudou Niu, Shen Huang, Wankui Gong, Youlu Yuan and Haihong Shang

370 *Genome-Wide Identification and Comparative Analysis of* ARF *Family Genes in Three Apiaceae Species*
Qiaoying Pei, Nan Li, Qihang Yang, Tong Wu, Shuyan Feng, Xuehuan Feng, Zange Jing, Rong Zhou, Ke Gong, Tong Yu, Zhen Wang and Xiaoming Song

380 *Integration of Transcriptional and Post-transcriptional Analysis Revealed the Early Response Mechanism of Sugarcane to Cold Stress*
Xing Huang, Yongsheng Liang, Baoqing Zhang, Xiupeng Song, Yangrui Li, Zhengqiang Qin, Dewei Li, Rongfa Chen, Zhongfeng Zhou, Yuchi Deng, Jiguang Wei and Jianming Wu

393 *Profiling of MicroRNAs and Their Targets in Roots and Shoots Reveals a Potential MiRNA-Mediated Interaction Network in Response to Phosphate Deficiency in the Forestry Tree* Betula luminifera
Junhong Zhang, Yan Lin, Fangmin Wu, Yuting Zhang, Longjun Cheng, Menghui Huang and Zaikang Tong

410 ***Transcriptomics of Biostimulation of Plants Under Abiotic Stress***
Susana González-Morales, Susana Solís-Gaona, Marin Virgilio Valdés-Caballero, Antonio Juárez-Maldonado, Araceli Loredo-Treviño and Adalberto Benavides-Mendoza

434 ***Genome-Wide Characterization of HSP90 Gene Family in Cucumber and Their Potential Roles in Response to Abiotic and Biotic Stresses***
Kaijing Zhang, Shuaishuai He, Yihu Sui, Qinghai Gao, Shuangshuang Jia, Xiaomin Lu and Li Jia

450 ***Comparative Transcriptome Analysis of Early- and Late-Bolting Traits in Chinese Cabbage (*Brassica rapa*)***
Xiaochun Wei, Md. Abdur Rahim, Yanyan Zhao, Shuangjuan Yang, Zhiyong Wang, Henan Su, Lin Li, Liujing Niu, Md. Harun-Ur-Rashid, Yuxiang Yuan and Xiaowei Zhang

465 ***Identify of Fast-Growing Related Genes Especially in Height Growth by Combining QTL Analysis and Transcriptome in* Salix matsudana *(Koidz)***
Guoyuan Liu, Qingshan Yang, Junfeng Gao, Yuwei Wu, Zhicong Feng, Jingke Huang, Hang Zou, Xingzhao Zhu, Yanhong Chen, Chunmei Yu, Bolin Lian, Fei Zhong and Jian Zhang

frontiers in Genetics

EDITORIAL
published: 11 May 2021
doi: 10.3389/fgene.2021.687966

Editorial: Comparative Genomics and Functional Genomics Analyses in Plants

Xiaoming Song[1], Zhiyong Liu[2]*, Hongjian Wan[3]*, Wei Chen[1]*, Rong Zhou[4]* and Weike Duan[5]**

[1] Center for Genomics and Bio-Computing, School of Life Sciences, North China University of Science and Technology, Tangshan, China, [2] College of Horticulture, Shenyang Agricultural University, Shenyang, China, [3] Institute of Vegetables, Zhejiang Academy of Agricultural Sciences, Hangzhou, China, [4] Department of Food Science, Aarhus University, Aarhus, Denmark, [5] College of Life Sciences and Food Engineering, Huaiyin Institute of Technology, Huai'an, China

Keywords: plant evolution, genome sequencing, molecular marker development, database construction, gene family

OPEN ACCESS

Edited and reviewed by:
TingFung Chan, The Chinese University of Hong Kong, China

****Correspondence:***
Xiaoming Song
songxm@ncst.edu.cn
Zhiyong Liu
lzyky99@163.com
Hongjian Wan
wanhongjian@sina.com
Wei Chen
greatchen@ncst.edu.cn
Rong Zhou
rong.zhou@food.au.dk
Weike Duan
weikeduan@126.com

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *30 March 2021*
Accepted: *06 April 2021*
Published: *11 May 2021*

Citation:
Song X, Liu Z, Wan H, Chen W, Zhou R and Duan W (2021) Editorial: Comparative Genomics and Functional Genomics Analyses in Plants. Front. Genet. 12:687966. doi: 10.3389/fgene.2021.687966

Editorial on the Research Topic

Comparative Genomics and Functional Genomics Analyses in Plants

INTRODUCTION

With rapidly developing sequencing technology and lower sequencing costs, over 500 plant genomes have been sequenced since the first release of the *Arabidopsis thaliana* genome in 2000. Meanwhile, more and more omics datasets have been released utilizing techniques such as genome re-sequencing, pan-genomics, RNA-Seq, metabolomics, and proteomics. The release of these datasets provides us with an excellent opportunity and rich data resources to study the evolution of plant genes. In recent years, in addition to the rapid development of bioinformatics, many types of software and tools have also become available for comparative genomics analyses. However, in the face of more and more kinds of omics data, we still need more efficient, powerful, and user-friendly tools, software, pipelines, websites, and databases.

Together with the rich omics datasets and advanced bioinformatics tools and methods, we hope to make more progress in plant comparative genomics and functional genomics. Utilizing the different types of omics data will help us to better understand the evolutionary history of plants, provide resources for molecular studies on important agronomic and economic traits, and develop new gene function verification methods. Therefore, we organized this Research Topic about "Comparative Genomics and Functional Genomics Analyses in Plants."

This Research Topic covered a broad range of comparative and functional genomic studies. Comparative genomics analyses to better understand genes or species evolution utilizing bioinformatics, including gene family analyses and plant evolutionary analyses. Functional genomics analyses to better understand plant genes or genome evolution utilizing omics datasets and molecular biology methods, such as including *de novo* genome sequencing and pan-genomic analyses, genome re-sequencing and GWAS analyses, RNA-seq and metabolomics analyses, and characterization of novel genes.

Finally, a total of 36 outstanding works was presented in the issue of this Research Topic. All of these articles were divided into five types, including 30 Original Research, 2 Brief Research Report, 2 Data Report, 1 Review, and 1 Methods. These articles covered a wide range of fields, including gene family analyses, transcriptome, *de novo* genome, chloroplast genome, and others. Here are some representative articles selected from each of these categories.

GENE FAMILY ANALYSES

Among 36 articles, 13 were involved in the gene family analyses in different species, including Apiaceae crops (ARF), Chinese cabbage (GRP), cotton (UGD and Rboh), cucumber (HSP90, BES1, and MYB), Gramineae crops (PLTs), green algae (PHT), pepper (bHLH), potato (PRXs), tobacco (SNAT), and tomato (C2H2-ZFP) (**Supplementary Table 1**). Most of these studies conducted comprehensive analyses of the gene family, such as gene family identification, gene structure, conserved motif, phylogenetic relationship, orthologous and paralogous genes, collinearity analysis, and expression pattern analyses. All of these studies provide rich resources for the comparative analysis of the related gene families in plants.

TRANSCRIPTOME

In the Research Topic, there were about 12 articles that belonged to the transcriptomic studies (**Supplementary Table 2**). Among which, most RNA-seq experiments were designed as various biotic and abiotic stresses, including cold (Huang et al.; Yang et al.), heat (Cai et al.), drought (Kim et al.), phosphate deficiency (Zhang et al.), turnip mosaic virus infection (Lyu et al.). Combining with RNA-Seq, several other technologies were also used to explore the gene expression or gene regulatory pathway. For example, Lyu et al. used the RNA-Seq and quantitative iTRAQ-LC-MS/MS to reveal the resistance mechanism against TuMV in Chinese Cabbage. Liu et al. explored the height growth-related genes by combining QTL analysis and transcriptome in *Salix matsudana*. In addition to mRNA expression analysis, some studies have also focused on the expression of non-coding RNA, such as microRNAs (Zhang et al.) and circular RNAs (Yang et al.). All of these transcriptome datasets can be downloaded using the accession number list in the **Supplementary Table 2** from NCBI website (https://www.ncbi.nlm.nih.gov). These RNA-Seq datasets will lay the foundation for gene function analysis in the future.

DE NOVO GENOME

In this topic, four studies reported and released the plant genomes, which provided the basic and rich resources for comparative and functional genomics analysis in plants (**Table 1**).

Liang et al. reported a high-quality reference genome for *Andrographis paniculata* (Chuanxinlian) using the PacBio and Illumina sequencing. The assembly genome was 284 Mb, and contig N50 was 5.14 Mb. The contigs were further assembled into 24 pseudo-chromosomes using Hi-C technique. This high-quality genome builds the foundation for exploring the biosynthetic pathways of various medicinal compounds in the future.

Yang et al. reported a draft genome for *Megacarpaea delavayi* (Brassicaceae), a plant that lived in the high mountains of southwest China at high altitudes. The assembly genome was 883 Mb, and contained a total of 41,114 protein-coding genes. They found that *M. delavayi* experienced an independent whole-genome duplication (WGD), paralleling to those WGDs

TABLE 1 | The list of chloroplast genome and *de novo* genome studies in this Research Topic.

Category	Species	Database	Accession No.	Links
Chloroplast genome	*Salvia yangii*	GenBank	MT012421	Gao et al.
	Salvia miltiorrhiza f. alba		MT012420	
	Salvia miltiorrhiza		NC_020431	
	Salvia officinalis		NC_038165	
	Salvia przewalskii		NC_041091	
	Salvia japonica		NC_035233	
	Salvia chanryoenica		NC_040121	
	Salvia bulleyana		NC_041092	
	Salvia rosmarinus		NC_027259	
	Curcuma rosesana	GenBank	MT395655	Liang et al.
	Curcuma amarissima		MT395645	
	Curcuma sichuanensis		MT395646	
	Curcuma elata		MT395649	
	Curcuma zanthorrhiza		MT395651	
	Curcuma longa		MT395644	
	Curcuma yunnanensis		MT395657	
	Curcuma sp.1		MT395653	
	Curcuma sp.2		MT395648	
	Curcuma wenyujin		MT395650	
	Curcuma phaeocaulis		MT395652	
	Curcuma aromatica		MT395654	
	Curcuma alismatifolia		MT395647	
	Curcuma flaviflora		MT395656	
	Alchemilla argyrophylla	GenBank	MT382661	Rono et al.
	Alchemilla pedata		MT382662	
	Dasiphora fruticosa		MF683841	
	Fragaria iinumae		KC507759	
	Fragaria nipponica		KY769125	
	Fragaria pentaphylla		KY434061	
	Fragaria orientalis		KY769126	
	Fragaria chiloensis		JN884816	
	Fragaria virginiana		JN884817	
	Fragaria mandshurica		KC507760	
	Fragaria vesca		KC507760	
	Rosa multiflora		NC_039989	
	Rosa odorata		KF753637	
	Hagenia abyssinica		KX008604	
De novo genome	*Gastrodia elata*	NCBI	PRJNA394702	Chen et al.

(Continued)

TABLE 1 | Continued

Category	Species	Database	Accession No.	Links
	Megacarpaea delavayi	NCBI	PRJNA630110	Yang et al.
			PRJCA002887	
	Andrographis paniculata	NCBI	PRJNA549104	Liang et al.
	Sargassum fusiforme	NCBI	PRJNA597239	Wang et al.
			SRR12124361	

of Iberis, Biscutella, and Anastatica in the early Miocene. *M. delavayi* specific and fast-evolving genes were mainly involved in "DNA repair" and "response to UV-B radiation," which might help it survive in high-altitude environments. This genome will provide valuable resources for studying the adaptation of plants to high-altitude habitats.

Chen et al. reported a genome of *Gastrodia elata*, which is an achlorophyllous orchid plant that displays a distinctive evolutionary strategy to adapt to the non-photosynthetic lifestyle. The assembly genome of *G. elata* was 1.12 Gb with a contig N50 size of 110 kb and scaffold N50 size of 1.64 Mb. The genes related to photosynthesis, leaf development, and plastid division pathways were found to be lost or under relaxed selection during the evolution. Thus, this genome provides a good resource for investigating the evolution of orchids and other achlorophyllous plants.

Wang et al. reported the first draft genome of the Seaweed *Sargassum fusiforme* using PacBio and Illumina technology. The assembled length of *S. fusiforme* genome was ∼394.4 Mb with a contig N50 value of ∼142.1 Kb, and a total of 20,222 putative genes were detected. This genome can be used as a genomic reference for the discovery of functional products and evolutionary studies in the Chromista kingdom.

All of these genome datasets can be downloaded using the accession number list in the **Table 1** from NCBI website.

CHLOROPLAST GENOME

In addition to the *de novo* genome, there were three studies about the chloroplast genome in this Research Topic (**Table 1**).

Gao et al. reported two whole chloroplast genomes of Salvia medicinal plants and compared them with seven other Salvia chloroplast genomes. Salvia species have been widely used as medicinal plants, which played important roles in the treatment and recovery of individuals with COVID-19. Gao et al. identified ten mutation hotspots as candidate DNA barcodes for Salvia species. Especially, they detected the transfer of nine large-sized chloroplast genome fragments into the mitochondrial genome.

Liang et al. reported the complete chloroplast genome of fourteen Curcuma species and constructed the phylogenetic tree of the 25 Zingiberaceae species. The 25 complete chloroplast genomes were ranging from 155,890 to 164,101 bp, and six divergent hotspots regions were detected. The phylogenetic trees showed that Musaceae was the basal group in Zingiberales, and Curcuma had a close relationship with Stahlianthu. This study provided the high-quality chloroplast genomes for Zingiberaceae study in the future.

Rono et al. reported the complete chloroplast genomes of two Alchemilla (Rosaceae) species *Alchemilla pedata* and *A. argyrophylla*. The chloroplast genomes were a typical circular quadripartite structure with the length of 152, 438 and 152,427 bp for *A. pedata* and *A. argyrophylla*, respectively. The high variation in five regions located in the intergenic spacers was detected through the comparative analysis with eight other Rosaceae species. Based on 26 chloroplast genomes of Rosaceae species, phylogenetic analysis revealed a monophyletic clustering of Alchemilla nested within subfamily Rosoideae. These complete chloroplast genomes will contribute to species delineation and further evolutionary studies in Rosaceae species.

All of these chloroplast genome datasets can be downloaded using the accession number list in the **Table 1** from NCBI website.

AUTHOR CONTRIBUTIONS

XS compiled the contributions from all authors. All authors contributed to the article and approved the final version of the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (31801856), Hebei Province Higher Education Youth Talents Program (BJ2018016), and China Postdoctoral Science Foundation (2020M673188).

ACKNOWLEDGMENTS

We greatly appreciate the contributions from all authors to our issue and all the reviewers for their time and efforts. We also would like to thank the Chief Editor and Editorial Office of Frontiers in Genetics for their support during the whole process of manuscript processing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.687966/full#supplementary-material

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.

Copyright © 2021 Song, Liu, Wan, Chen, Zhou and Duan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 28 May 2020
doi: 10.3389/fgene.2020.00540

Genome-Wide Analyses of the Genetic Screening of C_2H_2-Type Zinc Finger Transcription Factors and Abiotic and Biotic Stress Responses in Tomato (*Solanum lycopersicum*) Based on RNA-Seq Data

Tingting Zhao[1†], Tairu Wu[1†], Jia Zhang[1], Ziyu Wang[1], Tong Pei[1], Huanhuan Yang[1], Jingfu Li[1] and Xiangyang Xu[1,2]*

[1] *Laboratory of Genetic Breeding in Tomato, College of Horticulture and Landscape Architecture, Northeast Agricultural University, Harbin, China,* [2] *Key Laboratory of Biology and Genetic Improvement of Horticultural Crops (Northeast Region), Ministry of Agriculture and Rural Affairs, Northeast Agricultural University, Harbin, China*

OPEN ACCESS

Edited by:
Hongjian Wan, ZheJiang Academy of Agricultural Sciences, China

Reviewed by:
Ren Li, Chinese Academy of Agricultural Sciences, China
Xiaohui Zhou, Jiangsu Academy of Agricultural Sciences (JAAS), China
Ji Li, Nanjing Agricultural University, China

***Correspondence:**
Xiangyang Xu
xxy709@126.com

† These authors have contributed equally to this work

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

Received: *12 March 2020*
Accepted: *04 May 2020*
Published: *28 May 2020*

Citation:
Zhao T, Wu T, Zhang J, Wang Z, Pei T, Yang H, Li J and Xu X (2020) Genome-Wide Analyses of the Genetic Screening of C_2H_2-Type Zinc Finger Transcription Factors and Abiotic and Biotic Stress Responses in Tomato (Solanum lycopersicum) Based on RNA-Seq Data. Front. Genet. 11:540. doi: 10.3389/fgene.2020.00540

C_2H_2-type zinc finger proteins are classic and extensively studied members of the zinc finger family. C_2H_2-type zinc finger proteins participate in plant growth, development and stress responses. In this study, 99 C_2H_2-type zinc finger protein genes were identified and classified into four groups, and many functionally related *cis*-elements were identified. Differential C_2H_2-ZFP gene expression and specific responses were analyzed under drought, cold, salt, and pathogen stresses based on RNA-Seq data. Thirty-two C_2H_2 genes were identified in response to multiple stresses. Seven, 3, 5, and 8 genes were specifically expressed under drought, cold, salt, and pathogenic stresses, respectively. Five glycometabolism and sphingolipid-related pathways and the endocytosis pathway were enriched by KEGG analysis. The results of this study represent a foundation for further study of the function of C_2H_2-type zinc finger proteins and will provide us with genetic resources for stress tolerance breeding.

Keywords: C_2H_2-type zinc finger gene family, biotic stress, abiotic stress, transcription factor, tomato

INTRODUCTION

Tomato (*Solanum lycopersicum*) is one of the most important vegetable crops of Solanaceae (Mueller et al., 2005). However, the yield and quality of tomato are greatly affected by various biotic and abiotic stresses such as pathogen infection, low temperature, salt, and drought when the plants are exposed to complex environmental conditions. Upon stress perception, transcription factors (TFs) bind to their target genes to regulate their expression and orchestrate biochemical and physiological modifications critical for stress tolerance and the adaptation of plant growth (Hichri et al., 2014).

C_2H_2-type zinc finger proteins (C_2H_2-ZFPs), which are members of an important TF family, are also called TFIIIA-type ZFPs or classical ZFPs and are widely distributed in eukaryotic genomes (Huang et al., 2005). There are 176, 189, 211, and 109 C_2H_2-ZFPs in Arabidopsis, rice, maize, and poplar, respectively (Englbrecht et al., 2004; Agarwal et al., 2007; Liu et al., 2015;

Wei et al., 2015). The C_2H_2-ZFPs of eukaryotes generally have a specific conserved sequence consisting of 25–30 amino acids: X-X-C-X(1-5)-C-X(12)-H-X(3-6)-H (X: any amino acid; number: the number of amino acids). The two C (Cys) and two H (His) residues in the sequence form a coordination bond with a zinc ion and then form a tetrahedral structure composed of a two-stranded antiparallel β-sheet and an α-helix (Carl et al., 2001). EPF1 (later renamed ZPT2-1) was identified in Petunia as the first plant-specific ZFP. Takatsuji discovered that EPF1 interacts with the promoter region of the 5-enolpyruvylshikimate-3-phosphate synthase gene (EPSPS) and that the expression of EPF1 parallels the expression of EPSPS (Hiroshi et al., 1994). More zinc finger TFs have been subsequently identified in other plants and have been found to play crucial roles in the regulation of development and responses to biotic and abiotic stresses (Sang et al., 2004; Jiang and Pan, 2012).

Many C_2H_2-ZFPs genes involved in biotic and abiotic stresses have been studied in detail. GmZF1 and GmZFP3, two soybean C_2H_2-ZFPs, positively regulate the cold response and negatively regulate the drought response, respectively. Both of these proteins might be involved in the ABA-dependent pathway during the stress response (Zhang et al., 2016a). *OsMSR15* contains two C_2H_2-ZFP motifs, and its expression is strongly upregulated by cold, drought, and heat stresses in different rice tissues at different developmental stages (Zhang et al., 2016b). The expression of a novel ZFP gene, *StZFP1*, cloned from potato, is increased after salt stress as well as after infection by *Phytophthora infestans* and exogenous ABA (Tian et al., 2010). Overexpression of the *CAZFP1* gene, a zinc-finger protein gene isolated from pepper leaves, enhances resistance against infection by the pathogens *Xanthomonas campestris* and *Colletotrichum coccodes* in transgenic Arabidopsis plants (Sang et al., 2004). In tobacco, *ZFT1* functions as a transcription repressor and binds to the EP1S sequence. Overexpression of *ZFT1* renders tobacco plants more tolerant to TMV (Uehara et al., 2005). In this study, C_2H_2-ZFP genes are identified from the tomato genome by a bioinformatic analysis method and are analyzed for many aspects, including their phylogenetic relationships, gene structures, conserved protein motifs, chromosomal locations and promoter *cis*-elements. The abiotic and biotic stress response genes of the C_2H_2-ZFP family are screened under drought, salt, cold, and pathogen infection stresses based on corresponding RNA-Seq data. The results of the present work will provide us with comprehensive genome-wide knowledge of the tomato C_2H_2-ZFP TF family and will also provide us with potential gene resources for biotic and abiotic stress tolerance breeding.

MATERIALS AND METHODS

Identification and Phylogenetic Analysis of the C_2H_2-ZFP Gene Family

Tomato genome sequence data were obtained from the Solanaceae Genomics Network (SGN)[1] (Tomato Genome, 2012). The tomato genome version was GCF_000188115.3_SL2.50. Arabidopsis genes were obtained from the Arabidopsis Information Resource (TAIR[2]). The C_2H_2-ZFPs of tomato (SlZFs) were predicted using the HLH hidden Markov model (HMM) profile obtained from Pfam[3] (PF00096) and analyzed manually using the SMART[4] database to confirm the presence of C_2H_2-ZFP domains (SM000355) [33,34]. Finally, the obtained genes were compared with members of the C_2H_2-ZFP gene family in PlnTFDB v3.0[5] (Riano-Pachon et al., 2007). The ExPaSy site[6] (Elisabeth et al., 1999) was used to calculate the molecular weights and isoelectric points (pI) of the deduced polypeptides. Multiple sequence alignment was performed using ClustalX 1.83 (Julie et al., 1997). Phylogenetic analyses were performed using the neighbor-joining method in MEGA 5.0 (Tamura et al., 2011) and one unrooted neighbor-joining tree was constructed with 1,000 bootstrap replications.

Exon/Intron Structure Analysis and Identification of Conserved Motifs

The exon/intron arrangement of the ZFP genes was generated by the GSDS (Gene structure display server)[7] using both DNA sequences and their corresponding coding sequences (Hu et al., 2015). Additionally, online MEME (Multiple Expectation Maximization for Motif Elicitation)[8] was performed to search for conserved motifs for each C_2H_2-ZFP gene with the following parameters: the maximum number of motifs was set to 15, and the optimum motif width was set to 6 to 50 residues (Bailey et al., 2009). Each structural motif annotation was performed using the Pfam and SMART tools.

Chromosomal Location

The chromosomal location data of the C_2H_2-ZFP genes were obtained from SGN, and a map was generated via MapInspect software[9]. Different colors were used to represent different groups of C_2H_2-ZFP genes. Pink, green, blue, and orange represent the I, II, III, and IV groups, respectively. This approach allowed different groups of genes to be more easily distinguished in the distribution of the chromosome.

Promoter *cis*-Element Analysis

The transcription start site was designated + 1. The promoter sequences (from −2 kb to + 1 bp) of all C_2H_2-ZFP genes were obtained from Phytozome and analyzed using the program PlantCARE online[10]. The *cis*-elements were predicted and located.

[1] http://solgenomics.net

[2] http://www.arabidopsis.org/index.jsp

[3] http://pfam.xfam.org

[4] http://smart.embl-heidelberg.de/

[5] http://plntfdb.bio.uni-potsdam.de/v3.0/

[6] http://web.expasy.org/protparam/

[7] http://gsds.cbi.pku.edu.cn

[8] http://meme.nbcr.net/meme/

[9] http://mapinspect.software.informer.com

[10] http://bioinformatics.psb.ugent.be/webtools/plantcare/html/

Plant Materials and Stress Treatments

Tomato Micro-Tom and CGN18423 (which carries the Cf-19 gene that confers resistance to *Cladosporium fulvum*) were from the Tomato Research Institute of Northeast Agricultural University and were grown in a growth chamber (Conviron, Canada) with a light intensity of 120 μM photons m^{-2} s^{-1} (photoperiod 16 h, day/night temperature 22/18°C). Next, 100 healthy Micro-Tom seedlings were selected for subsequent analyses (10 seedlings used for organ analysis and 90 for drought, salt, cold stress treatments). The tomato seedlings used for drought, salt, and cold stress treatments were transferred to hydroponics and grown for 48 h when they were in the raising period of the four-leaf stage. Then, 30 seedlings were treated with 15% PEG 6000 to simulate drought stress. Young leaves were collected and frozen in liquid nitrogen at 0, 3 and 6 h after drought treatment. Thirty seedlings were transferred to a 5°C growth chamber for cold treatment. Young leaves were collected and frozen in liquid nitrogen at different time points (0, 4, and 12 h) after treatment. Additionally, 30 tomato seedlings were exposed to 200 mM sodium chloride (NaCl). Young leaves were collected and frozen in liquid nitrogen at different time points (0, 2, and 8 h) after treatment. Three biological replicates were performed for each time point. The materials and methods for pathogen stress are shown in Zhao et al. (2019).

RNA Isolation and cDNA Synthesis

Total RNA was extracted from leaf samples using a plant RNA mini kit (Watson, China) according to the manufacturer's handbook. The first-strand cDNA was transcribed with the TransScript® One-Step gDNA Removal and cDNA Synthesis SuperMix Kit according to the manufacturer's instructions.

Primer Design and qRT-PCR Verification for Different Tissues and Organs

Because of the large number of genes in the gene family and to rule out absolute tissue-specific preferences for subsequent analysis using transcriptome data (all sequencing materials were leaves), 25 genes that distribute in four groups were selected randomly for analysis to determine the tissue – specific general distribution of the family genes. 21 genes were successfully amplified by the primers we designed. The primers were designed by Primer 5.0 software. The primers are shown in **Supplementary Table S1**. All the primers were checked for primer pair specificity using the NCBI (National Center for Biotechnology Information) tool[11]. The *Solanum lycopersicum* actin (Tom41) gene (GenBank ID: U60480.1) was used as a reference gene. The primer efficiencies are listed in **Supplementary Table S1**. SYBR® Green Master Mix was used for quantitative PCR in an iQ5 real-time PCR detection system. The $2^{-\Delta\Delta CT}$ method was used to calculate the gene expression (Livak and Schmittgen, 2001). Three replicates were performed for each abiotic stress treatment, and the standard errors of the three replicates were also calculated.

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

cDNA Library Construction and Sequencing

The analysis process of the disease-resistant response transcriptome is described in Zhao et al. (2019). The samples collected under the NaCl, PEG and low temperature stresses mentioned above were used for RNA-Seq analysis, which was completed by the Beijing Genomics Institute (BGI). First, mRNA with a polyA tail was enriched with oligo-dT magnetic beads. Subsequently, rRNA was hybridized with DNA probes, and then, the DNA/RNA hybrid chains were selectively digested by RNaseH. Finally, the DNA probes were digested with DNaseI to purify RNA. After purification, the desired RNA was obtained. The enriched mRNA was broken into short fragments and then reverse transcribed into double stranded cDNA using random hexamer primers. Finally the end of the synthetic double stranded cDNA was smoothed, and the synthetic double stranded cDNA was amplified by PCR. The PCR product was thermally denatured into a single strand, and the single strand DNA was cycled to obtain a single stranded circular DNA library and then sequenced using the Illumina HiSeq platform.

Reads Mapping, Expression Levels, Gene Annotation, and Enrichment Analysis of C_2H_2-Type Zinc Finger TFs

Raw reads obtained from the Illumina HiSeq platform were processed to filter out adapters, shorter reads and low quality reads. The resulting reads (clean reads) were mapped onto the tomato reference genome using HISAT (Kim et al., 2015). To measure the gene expression levels, the total number of fragments per kilobases per million reads (FPKM) of each gene was calculated based on the length of the gene and the counts of reads mapped to the gene. We used getorf (EMBOSS:6.5.7.0) to detect the ORFs of genes. Then, hmmsearch was used to match each ORF to the transcription factor protein domain. Genes were identified according to the transcription factor family characteristics described by PlantTFDB. Then, we classified the C_2H_2-type Zinc Finger TFs functionally and used the phyper function in R software for GO and KEGG pathway enrichment analysis.

Real-Time PCR Validation

The mRNA that remained after cDNA library construction was transcribed using the method mentioned above with the cDNA Synthesis SuperMix Kit. To validate the accuracy of the RNA-Seq assay, 5 genes under abiotic and biotic stresses were selected randomly to confirm their relative expression by qRT-PCR. The primers were designed on NCBI. The primers are shown in **Supplementary Table S2**. qRT-PCR was applied using the method mentioned above.

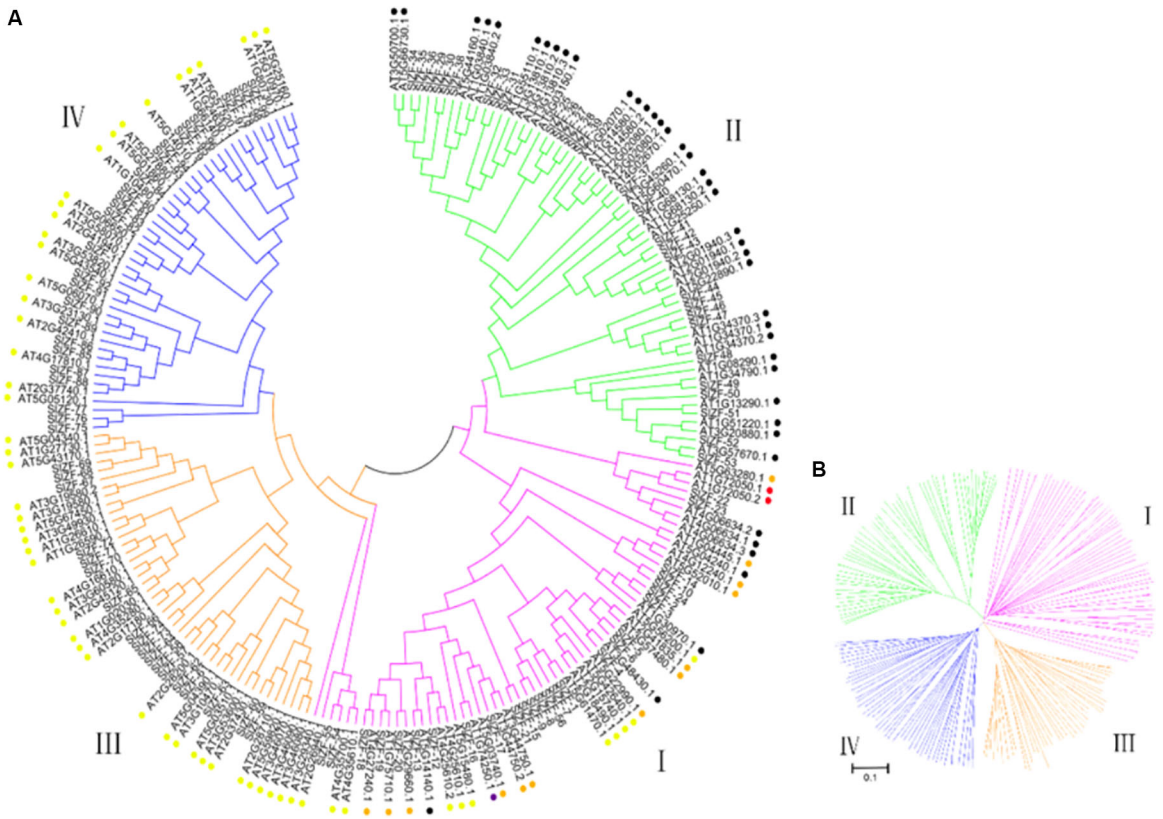


FIGURE 1 | Phylogenetic tree of the C_2H_2-ZFPs from tomato and Arabidopsis. **(A)** A phylogenetic tree made from 99 tomato and 116 Arabidopsis ZFPs was constructed using MEGA 5.0 by the neighbor-joining (NJ) method with 1000 bootstrap replicates. The four groups are indicated in four distinct colors. The black spots behind the names of the C_2H_2-ZFPs in Arabidopsis represent subset A, the red spots represent subset B, the yellow spots represent subset C1, the orange spots represent subset C2, and the purple spots represent subset C3. **(B)** A phylogenetic tree was constructed to clearly show the four groups of all genes.

RESULTS

Identification and Phylogenetic Analyses of C_2H_2-ZFP Genes in Tomato

After analysis of the entire genome of the C_2H_2-ZFP TF family, 99 C_2H_2-ZFPs were identified and named SlZF-1 to SlZF-99. **Supplementary Table S3** shows that the molecular masses of the tomato C_2H_2-ZFPs range from 11.014 (*SlZF-88*) to 139.814 kDa (*SlZF-4*). The pI values of the predicted proteins varied from 3.5367 (*SlZF-54*) to 10.615 (*SlZF-79*). The lengths of proteins varied from 96 (**SlZF-88**) to 1,253 (*SlZF-4*) amino acids, almost consistent with the molecular masses.

The phylogenetic relationship of the C_2H_2-ZFPs was determined using the neighbor-joining (NJ) method with 1,000 bootstrap replicates to construct an unrooted phylogenetic tree from alignments of the full-length C_2H_2-ZFP sequences of the 99 proteins from tomato and 116 proteins from Arabidopsis (**Figure 1A**). **Figure 1B** shows that all the genes were clearly classified into four groups, labeled from I to IV. Group II, which consists of 29 C_2H_2-ZFP genes (*SlZF-25* to *SlZF-53*), is the largest group, followed by groups I and IV, containing 26 C_2H_2-ZFP genes each (group I: *SlZF-1* to *SlZF-10*, *SlZF-12* to *SlZF-24*, and *SlZF-54* to *SlZF-56*; group IV: *SlZF-11* and *SlZF-75* to *SlZF-99*), and the smallest group is group III, containing 18 C_2H_2-ZFP genes (*SlZF-57* to *SlZF-74*). The C_2H_2-ZFP family genes of tomato and Arabidopsis are relatively evenly interspersed and distributed among the four groups, indicating that the C_2H_2-ZFP genes are closely related in Arabidopsis and tomato.

Gene Structure and Conserved Motifs of the C_2H_2-ZFP Genes in Tomato

To examine the structural evolution of the C_2H_2-ZFP genes in tomato, the organization of their introns and exons were analyzed using GSDS (**Figure 2**). Genes in the same group generally had similar numbers of introns. As the schematic structures suggest, the C_2H_2-ZFP group III and IV genes (other than *SlZF-82* and *SlZF-83*, which had one intron each) had zero introns. Most of the C_2H_2-ZFP genes (69%) in group II had two or three introns. In contrast, the gene structure appeared to be more diverse among the members of group I, with striking differences in their numbers of introns (varying from 0 to 9).

In total, 15 distinct motifs were discovered (**Supplementary Table S4**). The logos of the 15 motifs are listed in **Supplementary Table S5**, and the motif distribution in the genes is shown in **Figure 2**. Most motifs in the same group were distributed similarly. In group II, the linked motifs 5, 1, 11, 4, 3, 2, 6, and

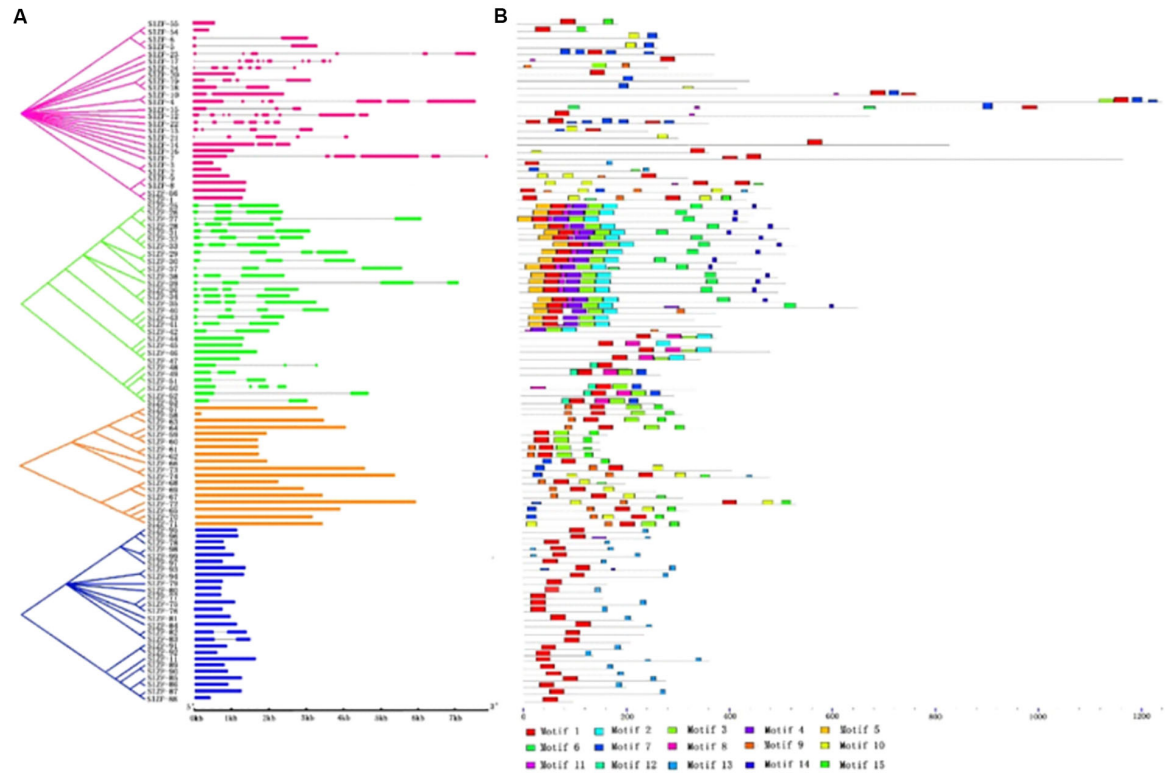


FIGURE 2 | Phylogenetic relations, gene structure and motif compositions of the tomato C_2H_2-ZFP genes. The 99 genes were divided into four groups. The four phylogenetic subgroups are marked with different colored backgrounds. **(A)** The exons and introns of each subgroup are represented by particular colored boxes and gray lines, respectively. **(B)** Each colored box represents a motif, and the black lines represent non-conserved sequences.

14 and the linked motifs 12, 1, 8, 3, and 7 were exhibited by the groups SlZF-25 to SlZF-39 and SlZF-49 to SlZF-53, respectively. Additionally, all of the C_2H_2-ZFPs in group III (16/18) except *SlZF-69* and *SlZF-73* had the linked motifs 1 and 15; those two exceptions both had the linked motifs 9, 1, and 10. Nearly half of the C_2H_2-ZFPs in group III (7/18) had the linked motifs 9, 1, 3, and 15. Most of the C_2H_2-ZFPs in group IV (21/26) had the linked motifs 1 and 13.

Chromosomal Locations of Tomato C_2H_2-ZFP Genes

To unravel the evolutionary patterns of this C_2H_2-ZFP gene family, we analyzed the locations of the 99 tomato C_2H_2-ZFP genes, finding that combinations of the 96 were distributed across all 12 chromosomes (**Figure 3**). The other three genes, *SlZF-14*, *SlZF-79* and *SlZF-88*, were not be mapped due to the lack of complete annotation in SGN (Sol Genomics Network).

Promoter *cis*-Element Analysis

Here, analysis of the Plant CARE database revealed multiple *cis*-acting elements in the promoter regions between −1 and 2000 bp upstream of the transcription start sites of all the putative tomato C_2H_2-ZFP genes. **Figure 4** and **Supplementary Table S6** show the presence of multiple *cis*-acting elements related to the phytohormone and environmental stress signal responsiveness of the C_2H_2-ZFP family promoters such as the C-repeat/DRE (cold- and dehydration-responsive element), HSE (heat stress-responsive element), LTR (low-temperature-responsive element), ABRE and CE3 (abscisic acid-responsive elements), CGTCA-motif and TGACG (MeJA-responsive elements), GARE-motif and TATC-box (gibberellin-responsive elements), SARE and TCA-element (salicylic acid-responsive elements), and ERE (ethylene-responsive element). The HSE and TCA-element were found in 63 and 58 promoters of the C_2H_2-ZFP genes, respectively. Several *cis*-acting elements involved in tissue-specific expression, e.g., an as-2-box (involved in shoot-specific expression), a GCN4_motif (involved in endosperm expression), an HD-Zip 2 (involved in the control of leaf morphology development), and an RY-element (involved in seed-specific regulation), were also found. Several *cis*-elements were unique to individual C_2H_2-ZFP genes; for instance, the C-repeat/DRE, CE3, and SARE motifs were present only in SlZF-13, SlZF-45, and SlZF-89, respectively.

Expression Profiles of the Tomato C_2H_2-ZFP Genes in Different Tissues and Organs

To further understand the tissue specificity of the C_2H_2-ZFP genes, the expression patterns of 21 genes across a variety of tomato tissues, including roots, stems, leaves, flowers, green

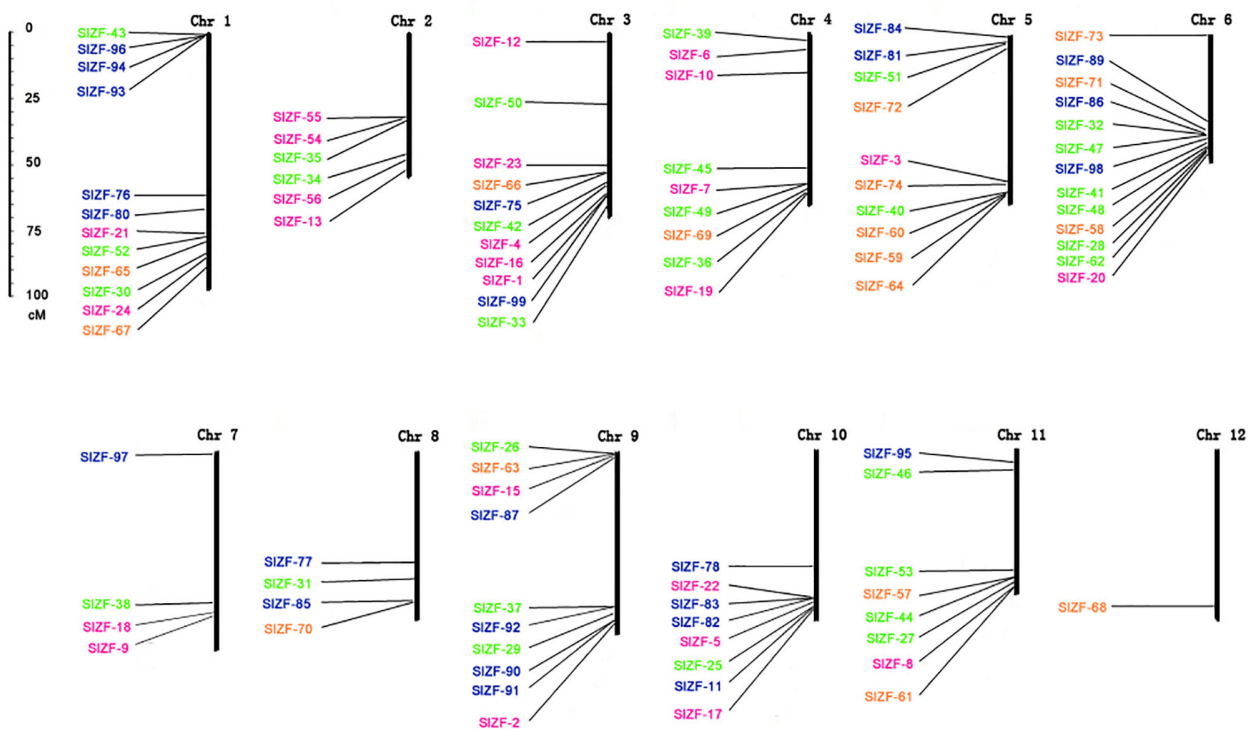


FIGURE 3 | Chromosomal locations of the tomato C_2H_2-ZFP genes. Ninety-six out of genes of 99 genes were located on a chromosome. The numbers belonging to same group are marked by the same color. The scale represents centimorgans (cM).

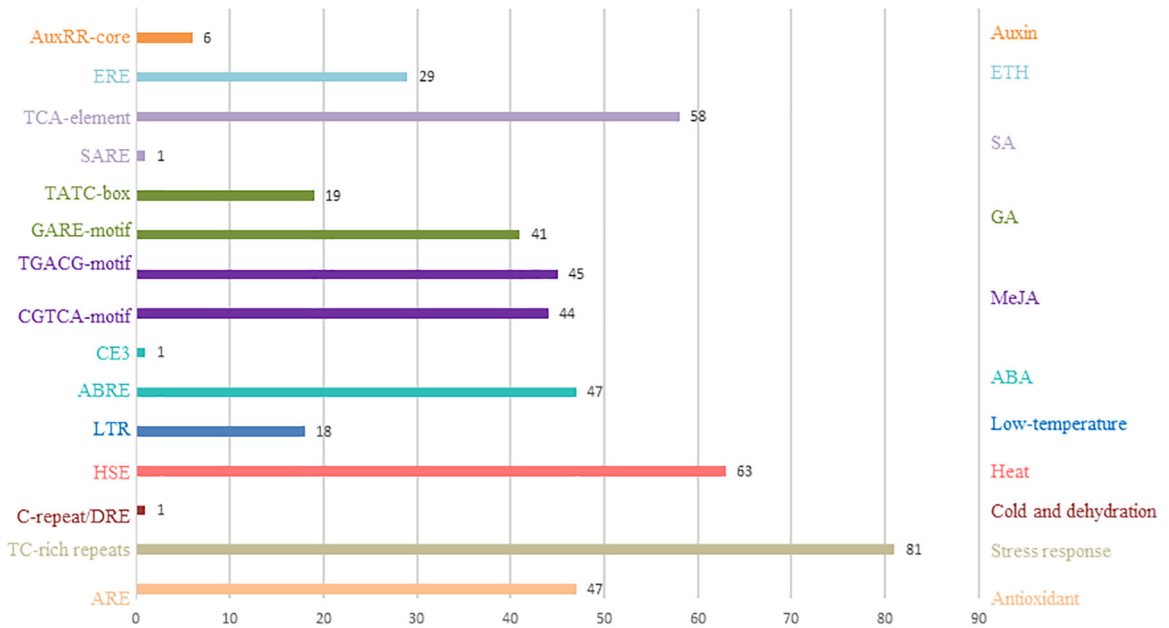


FIGURE 4 | Number of C_2H_2-ZFP genes containing various *cis*-acting elements.

fruits, mellow fruits and sepals, were analyzed by the qRT-PCR method. As shown in **Figure 5**, two genes (SlZF-24 and SlZF-28) were significantly highly expressed in root, and two genes (*SlZF-6* and *SlZF-13*) were significantly highly expressed in flower. Furthermore, 10 genes (*SlZF-6, SlZF-13, SlZF-15, SlZF-17, SlZF-18, SlZF-20, SlZF-22, SlZF-24, SlZF-28,* and *SlZF-30*) and

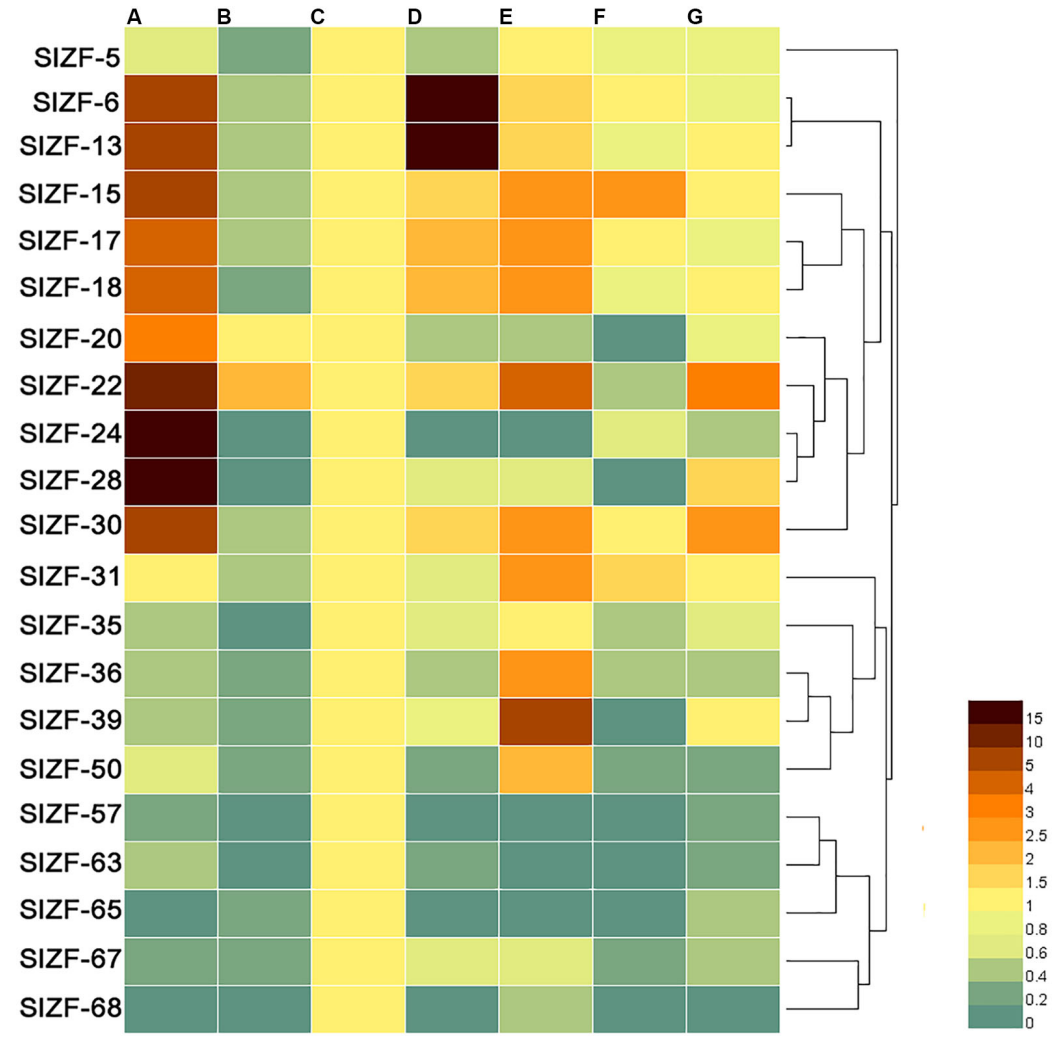


FIGURE 5 | Expression patterns of the C_2H_2-ZFP genes in different tomato tissues and organs. **(A)** Whole root; **(B)** young stems; **(C)** mature green leaves; **(D)** flowers at anthesis; **(E)** 20 days post-anthesis fruit; **(F)** breaker stage ripening fruit; **(G)** sepals. All tissues or organs were collected 4 months after sowing. Brown indicates higher expression levels, and bottle green indicates lower expression levels. Genes with similar profiles across arrays are grouped on the right by a hierarchical clustering method.

eight genes (*SlZF-15, SlZF-17, SlZF-18, SlZF-22, SlZF-30, SlZF-31, SlZF-36,* and *SlZF-39*) were highly expressed in root and green fruit, respectively. *SlZF-22* was expressed in many tissues and organs. Interestingly, all of the 21 genes were expressed similarly in leaves. Genes with similar profiles across arrays were grouped on the right by a hierarchical clustering method.

C_2H_2-ZFP Gene Expression Pattern Analysis Based on RNA-Seq Data

Four groups of RNA-Seq data, including tomato drought, salt, cold, and pathogen infection response transcriptome data, were used for this analysis. The statistics of the pathogen infection response transcriptome data information are shown in Zhao et al. (2019), and around 6.4 G of data were generated per sample. The details of raw reads, clean reads and read quality of each sample are shown in **Supplementary Table S7**. Through RNA-Seq data analysis, 50 C_2H_2-ZFP genes were screened under drought stress (**Figure 6**), 49 C_2H_2-ZFP genes were screened under cold stress (**Figure 7**), 50 C_2H_2-ZFP genes were screened under salt stress (**Figure 8**) and 51 genes were screened under pathogen infection (**Figure 9**). The FPKM of each C_2H_2-ZFP genes in three replicates at different time points under the four stresses are shown in **Supplementary Tables S8–S11**. We screened and identified genes whose absolute value of $\log_2$ (FPKM of second time point/FPKM of CK0) was greater than 1 as differentially expressed genes (DEGs) and obtained 28 drought-related genes, 31 salt-related genes, 13 cold-related genes and 24 pathogen-related genes. As shown in **Figure 10A** (a detailed gene list is shown in **Supplementary Table S12**), a small number of genes was involved in only one stress response, and most genes were involved in more than one stress response. Three genes (*SlZF-58*,

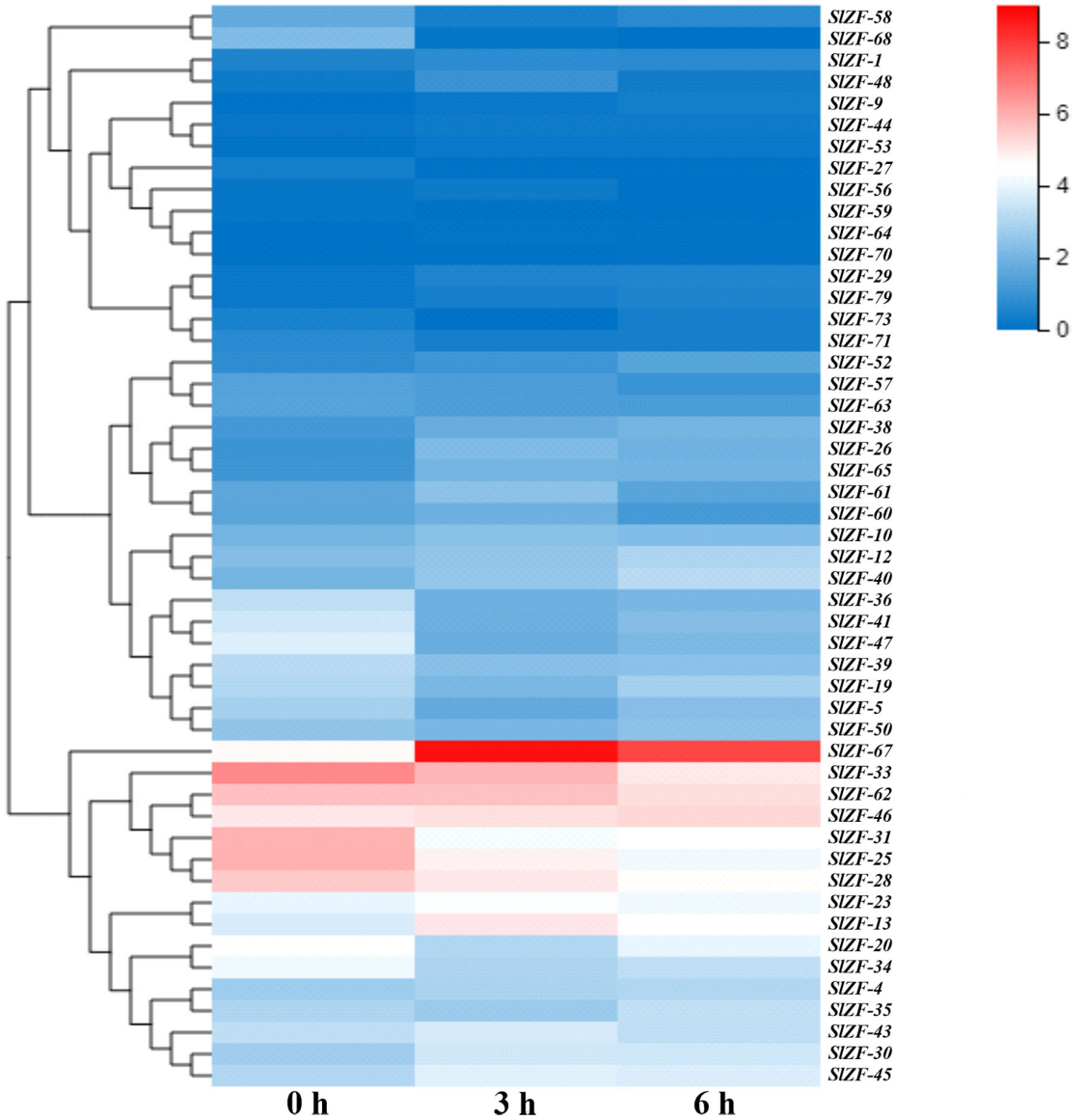


FIGURE 6 | C_2H_2-ZFP gene expression under drought stress determined by RNA-Seq. Heatmap of C_2H_2-ZFP gene expression under drought stress after 0, 3, and 6 h.

SlZF-61, *SlZF-5*) participated in all four stress responses. The numbers of upregulated genes and downregulated genes involved in the four stresses were similar. The numbers of upregulated and downregulated genes in the three abiotic stresses were similar, but the number of upregulated genes was significantly higher than the number of downregulated genes under biotic stress (**Figure 10A**). Further analysis shows (**Figure 10B**) that the highest proportion of the same response genes was 34% under drought and salt stresses, followed by pathogenic and salt stresses at 33%. Additionally, the lowest proportion of the same response genes was 15% under drought and cold stresses. The statistics of the genes that specifically respond to only one of the four kinds of stresses (including differentially expressed genes under only one stress and genes whose expression patterns were different from those under the other three stresses) were evaluated. The results are shown in **Figure 10C**. A total of 23 genes were screened out. Seven genes were specifically related to drought stress, five genes were specifically related to salt stress, three genes were specifically related to cold stress, and eight genes were specifically related to pathogen stress.

GO Enrichment Analysis of the C_2H_2-ZFP Genes

GO enrichment of all the C_2H_2-ZFP genes was performed. All 53 C_2H_2-ZFP genes identified through transcriptome data were involved in nucleic acid binding, organic cyclic compound binding, heterocyclic compound binding and binding. These four

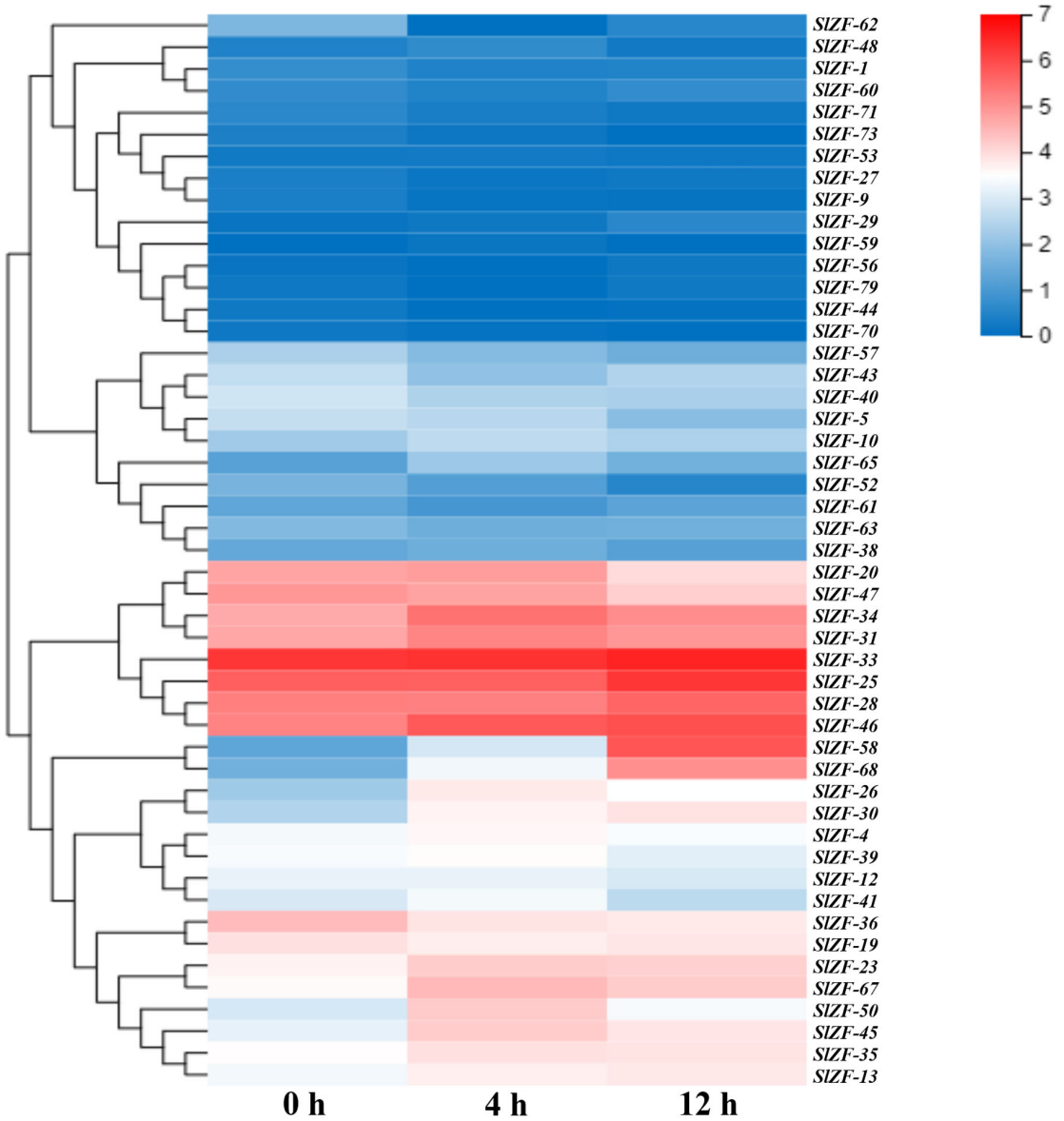


FIGURE 7 | C_2H_2-ZFP gene expression under heat stress determined by RNA-Seq. Heatmap of C_2H_2-ZFP gene expression under cold stress after 0, 4, and 12 h.

terms all belong to the molecular function category. Twenty C_2H_2-ZFP genes were involved in the nucleus, intracellular membrane-bounded organelle, membrane-bounded organelle, organelle and intracellular organelle, which belong to the cellular component category. Additionally, a few C_2H_2-ZFP genes were enriched in other GO terms (**Figure 11**).

Analysis of the KEGG Metabolic Pathway Enrichment and Expression Patterns in Different Pathways of the C_2H_2-ZFP Genes

The C_2H_2-ZFP genes were enriched into 6 KEGG pathways (Q-value $\leq$ 0.05), including glycosphingolipid biosynthesis, glycosaminoglycan degradation, sphingolipid metabolism, other glycan degradation, galactose metabolism, and endocytosis (**Figure 12**). Twenty-three C_2H_2-ZFP genes were enriched in the first five pathways. Seven C_2H_2-ZFP genes were enriched in the endocytosis pathway. Therefore, the genes enriched in these six pathways could be divided into two groups: the genes in the first five pathways were in one group (Group 1), and the genes in the endocytosis pathway were in another group (Group 2). The expression patterns of these genes were statistically analyzed under different stresses, and the results are shown in **Figure 13A**. The expression patterns of genes in the same group were different under different stresses, and the proportions of upregulated genes and downregulated genes were significantly different. In the first group, there were fewer upregulated genes than downregulated

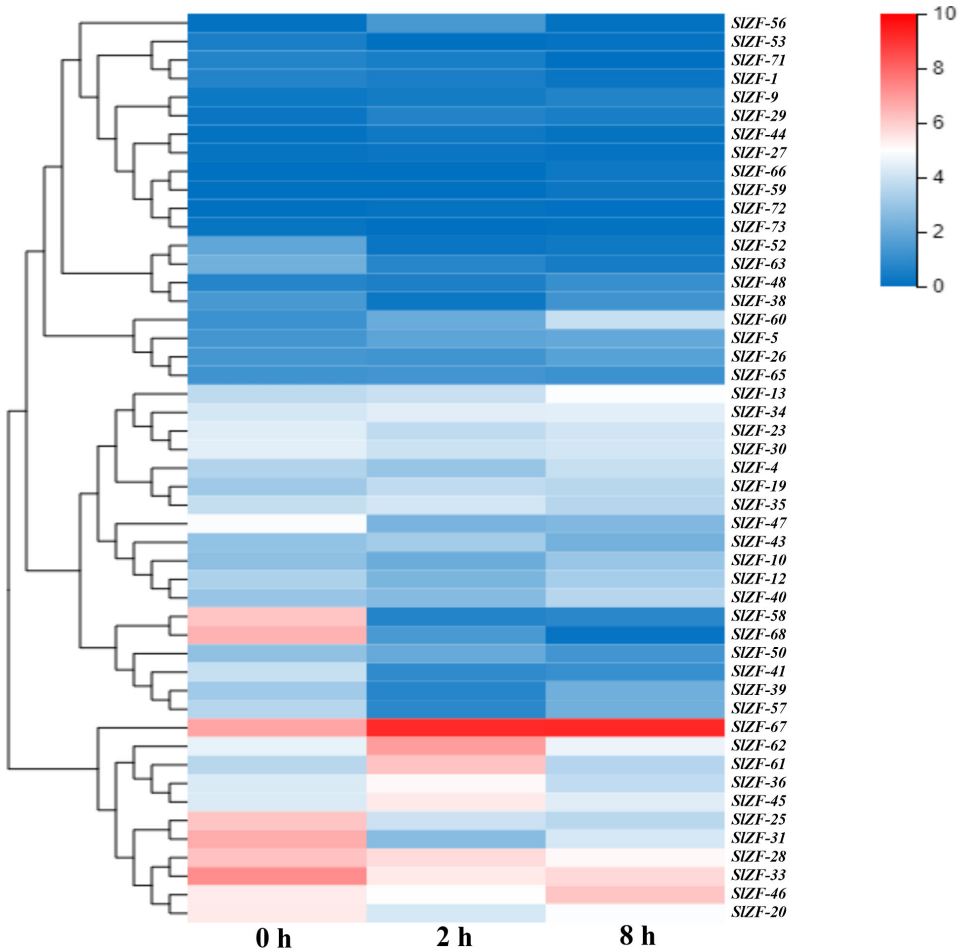


FIGURE 8 | C_2H_2-ZFP gene expression under salt stress determined by RNA-Seq. Heatmap of C_2H_2-ZFP gene expression under salt stress after 0, 2, and 8 h.

genes under the three abiotic stresses, while the upregulated genes were more numerous than the downregulated genes under biotic stress. In the second group, there were fewer upregulated genes under drought stress and salt stresses, but more upregulated genes under cold stress and pathogen stresses. We further analyzed the distribution of genes that were specifically expressed under each stress in the metabolic pathways. As shown in **Figure 13B**, 4 genes with specific expression under drought stress belonged to Group 1, and were all upregulated. Two genes were specifically expressed under both salt stress and cold stress; two genes belonging to Group 1 were downregulated under salt stress, and two genes belonging to Group 2 were upregulated under cold stress. Three genes were specifically expressed under pathogen stress in group 1, and 1 gene was specifically expressed under pathogen stress in Group 2, all of which were upregulated except for one gene in Group 1.

Validation of Gene Expression of Genes via qRT-PCR

To validate the expression data obtained from RNA-Seq, SlZF-58, SlZF-68, SlZF-30, SlZF-33, and SlZF-47 were chosen for confirmation by qRT-PCR under the four kinds of stresses. The expression levels of the five genes were basically consistent with the results of RNA-Seq (**Figure 14**). These results indicate that the RNA-Seq results in this study are valid.

DISCUSSION

The Evolution of the C_2H_2-ZFPs Family Among Different Species Is Conservative

C_2H_2-ZFPs are related to plant growth and development, abiotic stress, biotic stress, and other phenomena [14,18,19], showing that C_2H_2-ZFPs play an important role in plant life processes. In this study, we identified 99 potential C_2H_2-ZFPs in tomato, and the full-length C_2H_2-ZFP sequences of these 99 proteins were classified into four groups, labeled from I to IV, through an unrooted phylogenetic tree. We found that the phylogenetic relationship between the C_2H_2-ZFPs belonging to set A of the Arabidopsis C_2H_2-ZFPs and the group II C_2H_2-ZFPs of tomato was close. The phylogenetic relationship between the

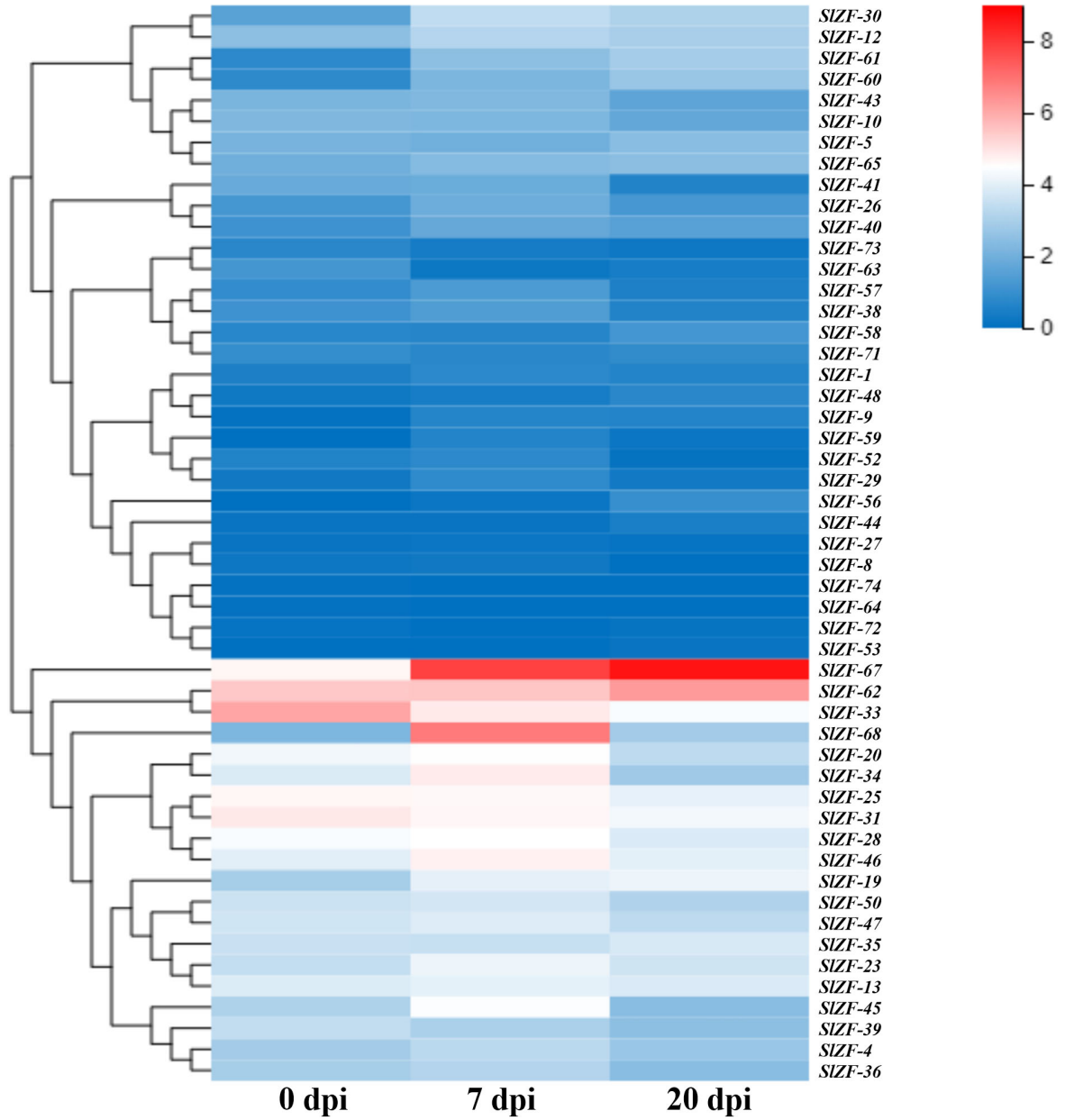


FIGURE 9 | C_2H_2-ZFP gene expression under pathogen infection stress. Heatmap of C_2H_2-ZFP gene expression under pathogen infection stress at 0, 7, and 20 days.

C_2H_2-ZFPs belonging to set C1 of the Arabidopsis C_2H_2-ZFPs and groups III and IV C_2H_2-ZFPs of tomato was close, though all of the Arabidopsis C_2H_2-ZFPs are in group I, which may demonstrate that different plants have conserved different duplicates. Furthermore, the physical parameters of C_2H_2-ZFPs in different plants have been shown to be very similar. For example, the molecular masses of the tomato C_2H_2-ZFP genes range from 11.014 (SlZF-88) to 139.814 kDa (SlZF-4), the molecular masses of the maize C_2H_2-ZFP genes range from 10.535 to 179.433 kDa and the molecular masses of poplar C_2H_2-ZFP genes range from 17.626 to 191.195 kDa (Liu et al., 2015; Wei et al., 2015). This similarity suggests that the C_2H_2-ZFP family is relatively conserved in evolution. The protein-coding sequences of eukaryotes are interrupted by introns [20]. To further study the distribution of the introns in the genome of tomato, we investigated C_2H_2-ZFP TFs and the relationship between tomato C_2H_2-ZFP TFs and those of other species. The number of introns within the C_2H_2-ZFP groups III and IV genes was either 0 or 1, while gene structure of the numbers within the C_2H_2-ZFP groups I and II appeared to be more variable and irregular in terms of intron numbers. These values are very similar to the numbers

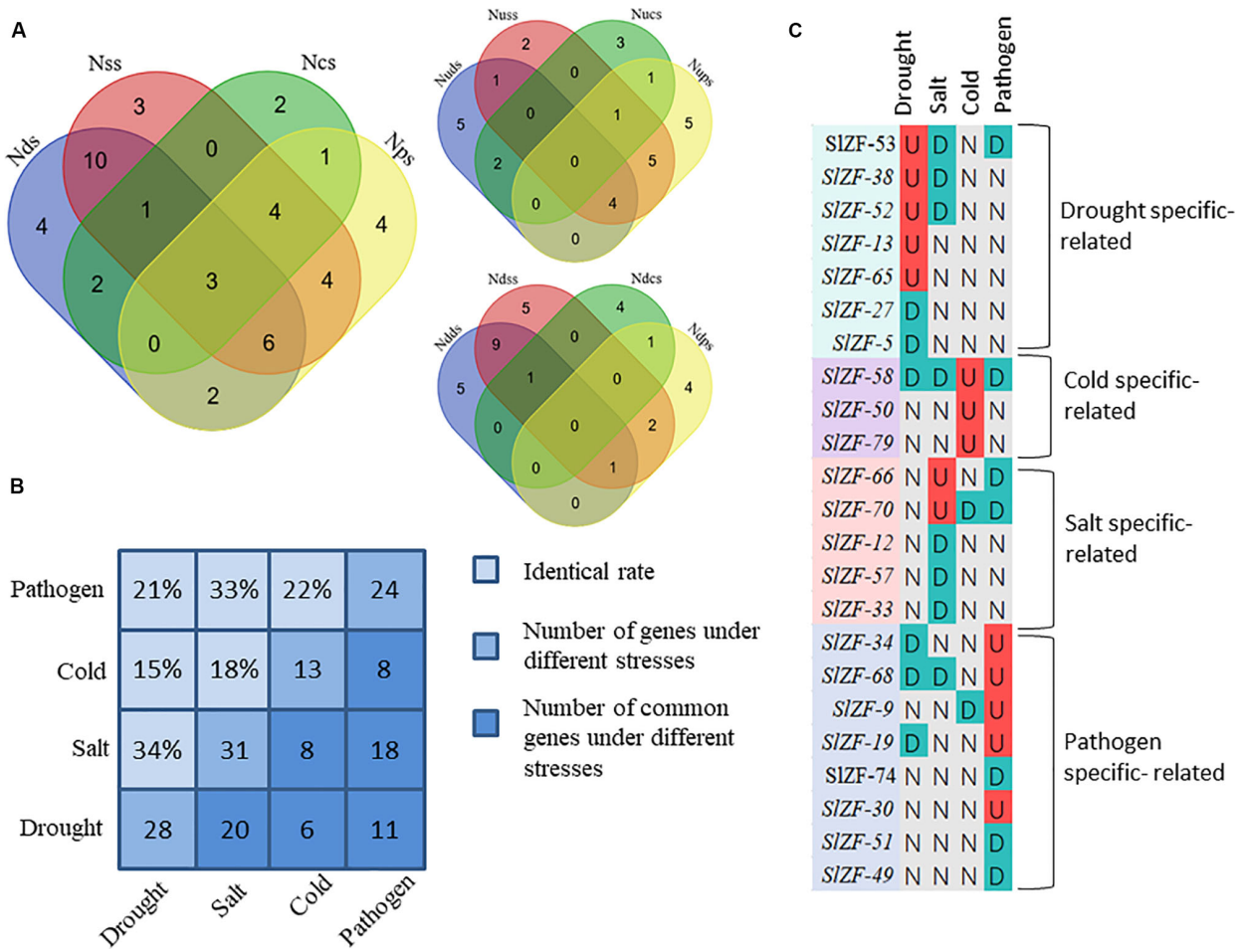


FIGURE 10 | Comparisons of differentially expressed genes and genes with different response patterns under the four stresses. **(A)** Venn diagrams showing the numbers the up- and down-regulated genes in the four comparisons. Nd(s/c/p)s: the number of differentially expressed genes under drought (salt/cold/pathogen) stress; Nud(s/c/p)s: the number of upregulated genes under drought (salt/cold/pathogen) stress; Ndd(s/c/p)s: the number of downregulated genes under drought (salt/cold/pathogen) stress. **(B)** Proportions of the identical DEGs under any two stresses. **(C)** A representation of a gene specifically expressed; "U" means upregulated; "D" means downregulated; "N" means no change.

of Maize and Poplar C_2H_2-ZFP introns. Most maize ZmZFP genes in their group C had 0 or 1 intron with the exception of a few members. In contrast, the gene intron–exon structures are more variable in group A and group B, which had more intron numbers. Poplar C_2H_2-ZFP genes in their group I and II possessed zero to three introns except for one gene. By contrast, the numbers of introns of genes in groups III and IV are large and irregular (Liu et al., 2015; Wei et al., 2015). Ninety-six of all 99 tomato C_2H_2-ZFP genes were unevenly distributed across all 12 chromosomes. Gene duplication plays an important role in genomic expansion (Vision et al., 2000), and the uneven distribution of the locations of the C_2H_2-ZFP genes on the chromosomes may be related to the duplication of the tomato genome (Ma et al., 2015). Nevertheless, all four groups of the C_2H_2-ZFP gene family were found on chromosomes 1, 3, 5, 6, 9, and 11, suggesting that members of different C_2H_2-ZFP gene family groups may interact via a network of proteins.

C_2H_2-ZFPs Gene Families May Have Multiple Functions

Fifteen motifs were found in this study. Most C_2H_2-ZFPs in tomato have motif 1, whose best possible match sequence contains the QALGGH amino acid sequence. Many plant ZFPs that are plant specific contain this motif (Agarwal et al., 2007; Liu et al., 2015; Wei et al., 2015). High conservation of the QALGGH sequence is critical for DNA-binding activity (Ken-ichi et al., 1998). In addition, 'DLELRL' (motif 13), a hexapeptide motif, has been found in 24 C_2H_2-ZFPs. This hexapeptide alone is sufficient to confer the ability to repress transcription to a DNA-binding domain in Arabidopsis (Hiratsu et al., 2004). Another repressor is the ERF-associated amphiphilic repression (EAR) motif, or DLN-box, which has been found in 22 C_2H_2-ZFPs. Stress-associated EAR-repressors may control the initiation of stress-activated gene expression under various stresses such as cold, salt, drought, and oxidative stress (Kazan, 2006; Ciftci-Yilmaz et al., 2007; Kielbowicz-Matuk, 2012). Motif 9 is a short, leucine-rich region

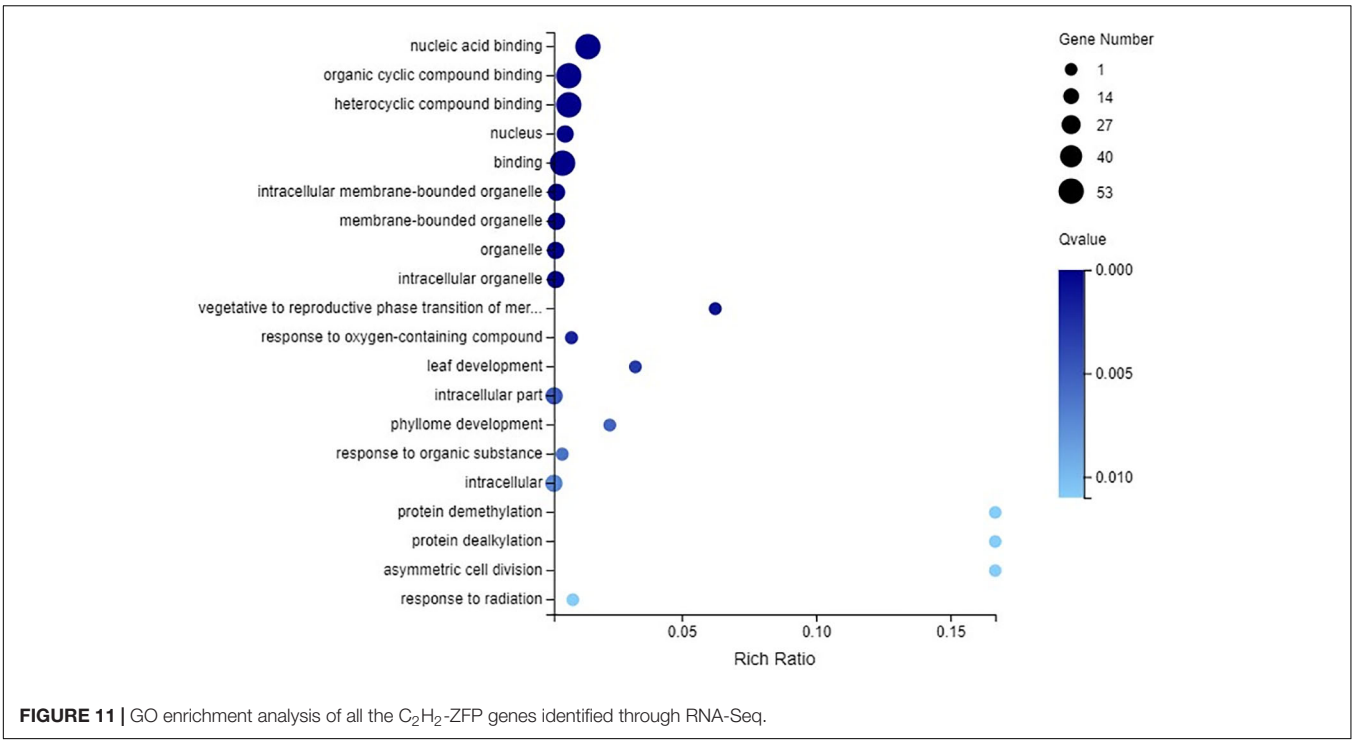


FIGURE 11 | GO enrichment analysis of all the C_2H_2-ZFP genes identified through RNA-Seq.

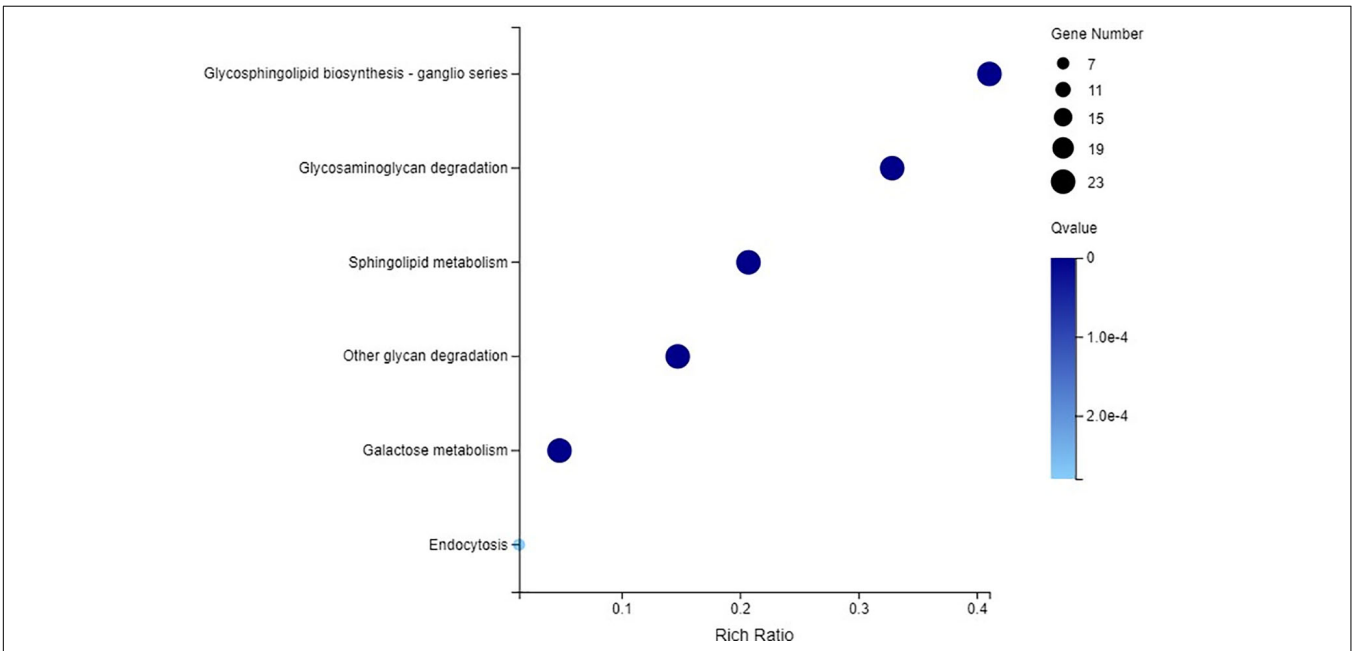


FIGURE 12 | KEGG pathway enrichment analysis of all the C_2H_2-ZFP genes identified through RNA-Seq.

with the core sequence EXEXXAXCLXXL (L-box) and is thought to play roles in protein–protein interactions or in maintaining folded structures (Kielbowicz-Matuk, 2012). We also discovered many motifs that may be related to DNA-binding or protein–protein interactions.

Most of the C_2H_2-ZFPs genes in tomato have abiotic stress-related and stress response *cis*-elements, including C-repeat/DRE, HSE, LTR, and TC-rich repeats. In addition, many C_2H_2-ZFPs genes have more than one phytohormone-related *cis*-elements (ABRE, CE3, CGTCA-motif, TGACG, GARE-motif, TATC-box, SARE, TCA-element, and ERE). Phytohormones such as salicylic acid (SA), jasmonic acid (JA), ethylene (ET), and abscisic acid (ABA) play important roles in plant responses to biotic and abiotic stresses and in the regulation

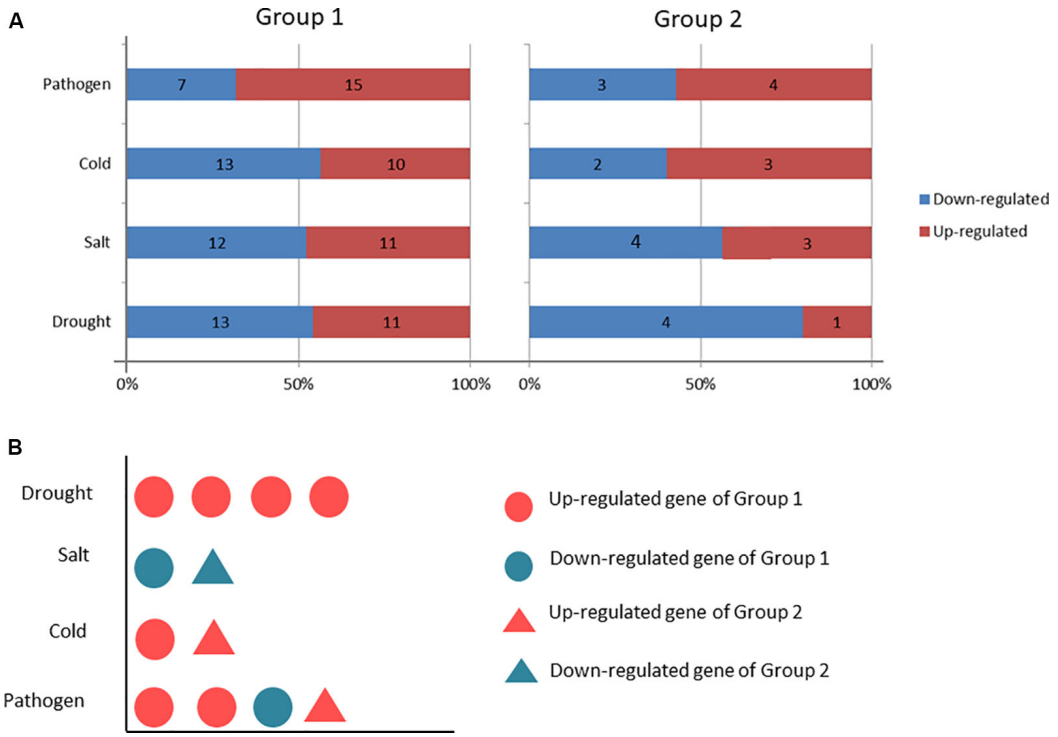


FIGURE 13 | Comparisons of the upregulated and downregulated genes between the two groups. Group1 includes genes from the first five pathways. Group2 contains genes from the sixth pathway. **(A)** Comparison of the upregulated and downregulated gene numbers under different stresses in each group. **(B)** Distribution of the number of genes specifically expressed under the different stresses in the two groups.

of developmental processes (Vision et al., 2000). These findings suggest that C_2H_2-ZFPs genes may modulate the biotic and abiotic stress responses directly or indirectly in tomato.

In this study, transcriptome data analysis showed that some of the genes in this family had significant changes in expression under four stresses, indicating that these genes were involved in the stress resistance and disease resistance responses of tomato plants. A C_2H_2-ZFP encoded by *OsMSR15* has been previously shown to have positive effects on cold, drought, and heat stress in different tissues of rice at different stages of development (Zhang et al., 2016a). *StZFP1*, which was cloned from potato, can be induced under abiotic stresses, such as salt, drought, and exogenous ABA (Tian et al., 2010). *GmZF1* in soybean regulates the expression of cold-regulated genes in transgenic *Arabidopsis thaliana* (*A. thaliana*) and enhances the anti-cold ability of *A. thaliana* (Yu et al., 2014). GO enrichment results showed that all the C_2H_2-ZFP genes screened through RNA-Seq were enriched in the molecular function category, indicating that all the C_2H_2-ZFP genes have multiple functions.

In addition, in the analysis of tissue-specific expression, we found that the expression of many genes was obviously tissue-specific. Previous studies have revealed that the C_2H_2-ZFP family of TFs have broad expression patterns under normal growth conditions in many organs and tissues. *RABBIT EARS* (*RBE*), a C_2H_2-ZFP gene, plays an important role in the development of petals in *A. thaliana* (Takeda et al., 2004). The product of *RBE* inhibits the expression of *TCP4* in *A. thaliana* at the early stage of petal development. The expression of *TCP4* in the flower of rbe-1 mutant plants is significantly higher than that in the flower of wild-type plants, which fully indicates that the expression of *RBE* is related to *TCP4* (Li et al., 2016). In this study, the results showed that *SlZF-24* and *SlZF-28* were significantly highly expressed in root and that *SlZF-6* and *SlZF-13* were preferentially expressed in flower, providing some clues about the functions of these genes, which may be related to root and flower development. Previous studies have revealed that root and flower development are influenced by genes such as *SUPERMAN*, a C_2H_2-ZFP in Arabidopsis that is a regulator of floral homeotic genes, and *TaZFP34*, which is a transcriptional repressor in wheat and is involved in modulating the root-to-shoot ratio (Chang et al., 2016).

Many C_2H_2-ZFP Genes of Tomato Respond to Multiple Stresses

Thirty- two genes were involved in differential expression under at least two stresses, while only 13 DEGs were involved in one type of stress, proving that most genes respond to multiple stresses. Previous studies have shown that some of the C_2H_2-ZFP genes are indeed associated with multiple stresses. *IbZFP1*, which is a Cys2/His2 zinc finger protein gene from sweet potato, is induced by NaCl and PEG. Overexpression of *IbZFP1* significantly enhances salt and drought tolerance in transgenic Arabidopsis plants (Wang et al., 2015). *ZFP245* is a cold- and drought-responsive gene that encodes a C_2H_2-ZFP in rice. Overexpression of *ZFP245* results in enhanced tolerance under

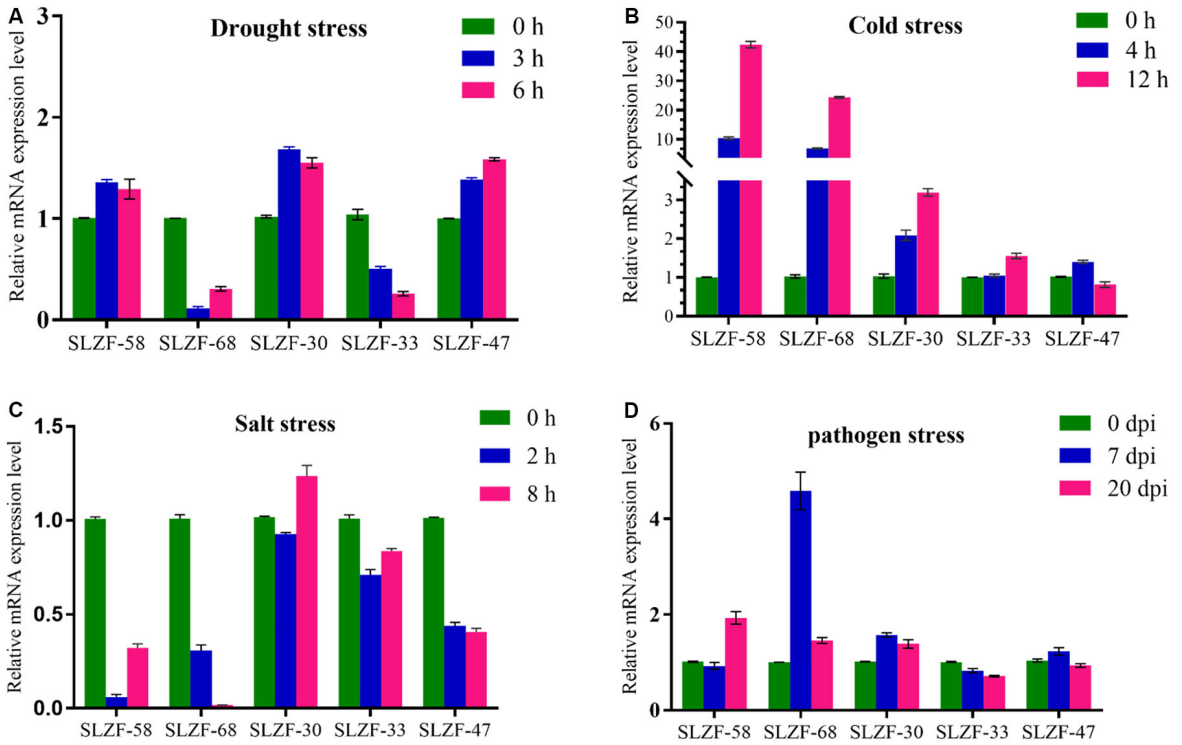


FIGURE 14 | Real-time quantitative PCR validation of five randomly selected genes under the four stresses. **(A)** The expression of the five genes under drought stress measured by qRT-PCR. **(B)** The expression of the five genes under cold stress measured by qRT-PCR. **(C)** The expression of the five genes under salt stress measured by qRT-PCR. **(D)** The expression of the five genes under pathogen stress measured by qRT-PCR. The data were analyzed by three independent repeats, and standard deviations are shown with error bars.

drought and cold stresses (Huang et al., 2009). A total of 11 DEGs were involved in the three stresses, and 7 genes had the same expression patterns under all three stresses, indicating that most of the DEGs involved in the three stresses had the same expression pattern. Six DEGs were identified under salt, pathogen and drought stresses, among which 4 genes were all upregulated and 1 gene was downregulated under these three stresses. However, only 1 DEG was the same under cold, drought and salt stresses (*SLZF-73*), and was downregulated under all three stresses. Under the four stresses, the cold-related DEGs were the fewest in number, and the cold-specific response genes were also the fewest in number. Indicating that the C_2H_2-ZFP genes have the weakest response to cold stress.

Differences May Exist Among the Regulation of C_2H_2-ZFP Genes Under Different Stresses

KEGG analysis showed that the genes of this family were mainly enriched in six pathways including glycosphingolipid biosynthesis, glycosaminoglycan degradation, sphingolipid metabolism, other glycan degradation, galactose metabolism, and endocytosis. The first five pathways are all related to the function of glycometabolism and sphingolipid, and the endocytosis pathway is related to membrane function, all of these functions are closely related to the resistance of plants to environmental stress (Jae et al., 2008; Ali et al., 2018; Couchoud et al., 2019). In this study, many C_2H_2-ZFP genes were differentially expressed in these pathways, suggesting that this family of genes may play a role in stress tolerance regulation through these pathways. In the pathways related to glycometabolism and sphingolipid, the proportion of upregulated C_2H_2-ZFP genes under biotic stress was higher than that of the proportion of genes upregulated under abiotic stresses, indicating that the sensitivity of the regulatory points of this family gene in the glycometabolism and sphingolipid pathways to biotic and abiotic stresses is different. In the endocytosis pathway, the proportion of upregulated genes under drought stress was the lowest, which was significantly different from the other three stresses. The proportions of upregulated genes under pathogen stress and cold stress were similar, and the data distribution was very different from that in the glycometabolism and sphingolipid -related pathways, indicating that the regulatory points of C_2H_2-ZFP genes in the two pathways are relatively independent. In the analysis of stress-specific response genes, we found that all the drought stress-specific response genes were enriched in the glycometabolism and sphingolipid pathways, and that all of them were upregulated, indicating that the C_2H_2-ZFP family has drought-specific response factors and that these response factors regulate the drought stress response only by enhancing processes related to glycometabolism and sphingolipid. There are specific

response genes in both types of pathways under other stresses, the most notable of which is salt stress. Additionally, the specific response genes are all downregulated under salt stress. Therefore, it is speculated that under salt, cold and pathogen stresses, although some members of the C_2H_2-ZFP family regulate plant' resistance to environmental stress through the glycometabolism and sphingolipid-related pathways and the endocytosis pathway, the regulation process is more specific under salt stress compared to the processes under cold and pathogen stresses.

CONCLUSION

Ninety-nine C_2H_2-type zinc finger protein genes were identified and classified into four groups based on phylogenetic analysis. These genes were distributed across all 12 chromosomes and were expressed in different tissues or organs including roots, stems, leaves, flowers and fruits, some of which showed tissue-specific expression patterns. Thirty-two C_2H_2-ZFP genes were identified in response to multiple stresses. Seven, 3, 5, and 8 genes were specifically expressed under drought, cold, salt, and pathogen stresses, respectively. KEGG analysis revealed that the genes in the C_2H_2-ZFP family mainly exert anti-stress regulation through the glycosphingolipid biosynthesis, glycosaminoglycan degradation, sphingolipid metabolism, other glycan degradation, galactose metabolism, and endocytosis pathways. The results of this study are the foundation for further studies on the function of C_2H_2-type zinc finger proteins and will provide a basis for stress tolerance breeding.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the NCBI GEO accession GSE148530.

AUTHOR CONTRIBUTIONS

XX and TZ designed the study and supervised the project and revised the manuscript. TZ, TW, JZ, TP, and ZW implemented the study and acquired the data. TW and TZ drafted the manuscript. HY and JL helped to drafted the manuscript. All authors read and approved the final manuscript.

FUNDING

This study was supported by the National Key R&D Program of China (2017YFD0101900), the China Agriculture Research System (CARS-23-A-16), the China postdoctoral science foundation grant (2018 M630333), and the University Nursing Program for Young Scholars with Creative Talents in Heilongjiang Province (UNPYSCT-2018169).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00540/full#supplementary-material

REFERENCES

Agarwal, P., Arora, R., Ray, S., Singh, A. K., Singh, V. P., Takatsuji, H., et al. (2007). Genome-wide identification of C2H2 zinc-finger gene family in rice and their phylogeny and expression analysis. *Plant. Mol. Biol.* 65, 467–485. doi: 10.1007/s11103-007-9199-y

Ali, U., Li, H., Wang, X., and Guo, L. (2018). Emerging roles of sphingolipid signaling in plant response to biotic and abiotic stresses. *Mol. Plant.* 11, 1328–1343. doi: 10.1016/j.molp.2018.10.001

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–W208. doi: 10.1093/nar/gkp335

Carl, O. P., Ezra, P., and Robert, A. J. (2001). Design and selection of novel Cys2his2 Zinc Finger Proteins. *Annu. Rev. Biochem.* 70, 313–340. doi: 10.1146/annurev.biochem.70.1.313

Chang, H., Chen, D., Kam, J., Richardson, T., Drenth, J., Guo, X., et al. (2016). Abiotic stress upregulated TaZFP34 represses the expression of type-B response regulator and SHY2 genes and enhances root to shoot ratio in wheat. *Plant. Sci.* 252, 88–102. doi: 10.1016/j.plantsci.2016.07.011

Ciftci-Yilmaz, S., Morsy, M. R., Song, L., Coutu, A., Krizek, B. A., Lewis, M. W., et al. (2007). The EAR-motif of the Cys2/His2-type zinc finger protein Zat7 plays a key role in the defense response of *Arabidopsis* to salinity stress. *J. Biol. Chem.* 282, 9260–9268. doi: 10.1074/jbc.M611093200

Couchoud, M., Der, C., Girodet, S., Vernoud, V., Prudent, M., and Leborgne-Castel, N. (2019). Drought stress stimulates endocytosis and modifies membrane lipid order of rhizodermal cells of Medicago truncatula in a genotype-dependent manner. *BMC Plant Biol.* 19:221. doi: 10.1186/s12870-019-1814-y

Elisabeth, G., Christine, H., Alexandre, G. D., Marc, R. W., Ron, D. A., and Amos, B. (1999). "Protein identification and analysis tools on the ExPASy server," in *The Proteomics Protocols Handbook*, ed. J. M. Walker (Totowa, NJ: Humana Press), 571–607. doi: 10.1385/1-59259-584-7:531

Englbrecht, C. C., Schoof, H., and Bohm, S. (2004). Conservation, diversification and expansion of C2H2 zinc finger proteins in the *Arabidopsis* thaliana genome. *BMC Genomics* 5:39. doi: 10.1186/1471-2164-5-39

Hichri, I., Muhovski, Y., Zizkova, E., Dobrev, P. I., Franco-Zorrilla, J. M., Solano, R., et al. (2014). The *Solanum lycopersicum* zinc finger2 cysteine-2/histidine-2 repressor-like transcription factor regulates development and tolerance to salinity in tomato and *Arabidopsis*. *Plant Physiol.* 164, 1967–1990. doi: 10.1104/pp.113.225920

Hiratsu, K., Mitsuda, N., Matsui, K., and Ohme-Takagi, M. (2004). Identification of the minimal repression domain of SUPERMAN shows that the DLELRL hexapeptide is both necessary and sufficient for repression of transcription in *Arabidopsis*. *Biochem. Biophys. Res. Commun.* 321, 172–178. doi: 10.1016/j.bbrc.2004.06.115

Hiroshi, T., Norihisa, N., and Yukihisa, K. (1994). A new family of zinc finger proteins in petunia: structure, DNA sequence recognition, and floral organ-specific expression. *Plant Cell* 6, 947–958. doi: 10.1105/tpc.6.7.947

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

Huang, J., Sun, S. J., Xu, D. Q., Yang, X., Bao, Y. M., Wang, Z. F., et al. (2009). Increased tolerance of rice to cold, drought and oxidative stresses mediated by the overexpression of a gene that encodes the zinc finger protein ZFP245. *Biochem. Biophys. Res. Commun.* 389, 556–561. doi: 10.1016/j.bbrc.2009.09.032

Huang, J., Wang, J., and Zhang, H. (2005). Rice ZFP15 gene encoding for a novel C2H2-type zinc finger protein lacking DLN box, is regulated by spike development but not by abiotic stresses. *Mol. Biol. Rep.* 32, 177–183. doi: 10.1007/s11033-005-2338-2330

Jae, S. K., Julia, F., Chang, H. K., Hiroyuki, K., Meenu, V., Akihiro, U., et al. (2008). Salt tolerance of *Arabidopsis* thaliana requires maturation of N-glycosylated proteins in the Golgi apparatus. *Proc. Natl. Acad. Sci. U S A.* 105, 5933–5938. doi: 10.1073/pnas.0800237105

Jiang, L., and Pan, L. J. (2012). Identification and expression of C2H2 transcription factor genes in *Carica papaya* under abiotic and biotic stresses. *Mol. Biol. Rep.* 39, 7105–7115. doi: 10.1007/s11033-012-1542-y

Julie, D. T., Toby, J. G., Frédéric, P., François, J., and Desmond, G. H. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25, 4876–4882. doi: 10.1093/nar/25.24.4876

Kazan, K. (2006). Negative regulation of defence and stress genes by EAR-motif-containing repressors. *Trends Plant Sci.* 11, 109–112. doi: 10.1016/j.tplants.2006.01.004

Ken-ichi, K., Ayako, S., Akira, K., Zbigniew, R., Yoshiaki, K., Hitoshi, N., et al. (1998). Cys2/His2 zinc-finger protein family of petunia: evolution and general mechanism of target-sequence recognition. *Nucleic Acids Res.* 26, 608–615. doi: 10.1093/nar/26.2.608

Kielbowicz-Matuk, A. (2012). Involvement of plant C(2)H(2)-type zinc finger transcription factors in stress responses. *Plant Sci.* 18, 78–85. doi: 10.1016/j.plantsci.2011.11.015

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

Li, J., Yanzhi, W., Yongxia, Z., Weiyao, W., Vivian, F. I., and Tengbo, H. (2016). RABBIT EARS regulates the transcription of TCP4 during petal development in *Arabidopsis*. *J. Exp. Bot.* 67, 6473–6480. doi: 10.1093/jxb/erw419

Liu, Q., Wang, Z., Xu, X., Zhang, H., and Li, C. (2015). Genome-Wide analysis of C2H2 Zinc-Finger family transcription factors and their responses to abiotic stresses in poplar (Populus trichocarpa). *PLoS One* 10:e0134753. doi: 10.1371/journal.pone.0134753

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

Ma, J., Li, M. Y., Wang, F., Tang, J., and Xiong, A. S. (2015). Genome-wide analysis of Dof family transcription factors and their responses to abiotic stresses in Chinese cabbage. *BMC Genomics* 16:33. doi: 10.1186/s12864-015-1242-1249

Mueller, L. A., Tanksley, S. D., Giovannoni, J. J., van Eck, J., Stack, S., Choi, D., et al. (2005). The tomato sequencing project, the first cornerstone of the international Solanaceae Project (SOL). *Comp. Funct. Genomics* 6, 153–158. doi: 10.1002/cfg.468

Riano-Pachon, D. M., Ruzicic, S., Dreyer, I., and Mueller-Roeber, B. (2007). PlnTFDB: an integrative plant transcription factor database. *BMC Bioinformatics* 8:42. doi: 10.1186/1471-2105-8-42

Sang, H. K., Jeum, K. H., Sung, C. L., Kee, H. S., Ho, W. J., and Byung, K. H. (2004). CAZFP1, Cys2/His2-type zinc-finger transcription factor gene functions as a pathogen-induced early-defense gene in *Capsicum annuum*. *Plant Mol. Biol. Rep.* 55, 883–904. doi: 10.1007/s11103-005-2151-2150

Takeda, S., Matsumoto, N., and Okada, K. (2004). RABBIT EARS, encoding a SUPERMAN-like zinc finger protein, regulates petal development in *Arabidopsis thaliana*. *Development* 131, 425–434. doi: 10.1242/dev.00938

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28, 2731–2739. doi: 10.1093/molbev/msr121

Tian, Z. D., Zhang, Y., Liu, J., and Xie, C. H. (2010). Novel potato C2H2-type zinc finger protein gene, StZFP1, which responds to biotic and abiotic stress, plays a role in salt tolerance. *Plant Biol.* 12, 689–697. doi: 10.1111/j.1438-8677.2009.00276.x

Tomato Genome, C. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485, 635–641. doi: 10.1038/nature11119

Uehara, Y., Takahashi, Y., Berberich, T., Miyazaki, A., Takahashi, H., Matsui, K., et al. (2005). Tobacco ZFT1, a transcriptional repressor with a Cys2/His2 type zinc finger motif that functions in spermine-signaling pathway. *Plant Mol. Biol.* 59, 435–448. doi: 10.1007/s11103-005-0272-270

Vision, T. J., Daniel, G. B., and Steven, D. T. (2000). The Origins of Genomic Duplications in *Arabidopsis*. *Science* 290, 2114–2117.

Wang, F., Tong, W., Zhu, H., Kong, W., Peng, R., Liu, Q., et al. (2015). A novel Cys2/His2 zinc finger protein gene from sweetpotato, IbZFP1, is involved in salt and drought tolerance in transgenic *Arabidopsis*. *Planta* 243, 783–797. doi: 10.1007/s00425-015-2443-2449

Wei, K., Pan, S., and Li, Y. (2015). Functional characterization of maize C2H2 Zinc-Finger gene family. *Plant Mol. Biol. Rep.* 34, 761–776. doi: 10.1007/s11105-015-0958-957

Yu, G. H., Jiang, L. L., Ma, X. F., Xu, Z. S., Liu, M. M., Shan, S. G., et al. (2014). A soybean C2H2-type zinc finger gene GmZF1 enhanced cold tolerance in transgenic *Arabidopsis*. *PLoS One* 9:e109399. doi: 10.1371/journal.pone.0109399

Zhang, D., Tong, J., Xu, Z., Wei, P., Xu, L., Wan, Q., et al. (2016a). Soybean C2H2-Type Zinc Finger Protein GmZFP3 with Conserved QALGGH motif negatively regulates drought responses in transgenic *Arabidopsis*. *Front. Plant Sci.* 7:325. doi: 10.3389/fpls.2016.00325

Zhang, X., Zhang, B., Li, M. J., Yin, X. M., Huang, L. F., Cui, Y. C., et al. (2016b). OsMSR15 encoding a rice C2H2-type zinc finger protein confers enhanced drought tolerance in transgenic *Arabidopsis*. *J. Plant Biol.* 59, 271–281. doi: 10.1007/s12374-016-0539-539

Zhao, T., Liu, W., Zhao, Z., Yang, H., Bao, Y., Zhang, D., et al. (2019). Transcriptome profiling reveals the response process of tomato carrying Cf-19 and *Cladosporium fulvum* interaction. *BMC Plant Biol.* 19:572. doi: 10.1186/s12870-019-2150-y

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.

Copyright © 2020 Zhao, Wu, Zhang, Wang, Pei, Yang, Li and Xu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 19 June 2020
doi: 10.3389/fgene.2020.00565

Micro-Evolution Analysis Reveals Diverged Patterns of Polyol Transporters in Seven Gramineae Crops

*Weilong Kong, Tong Sun, Chenhao Zhang, Yalin Qiang and Yangsheng Li**

State Key Laboratory of Hybrid Rice, College of Life Sciences, Wuhan University, Wuhan, China

OPEN ACCESS

Edited by:
Hongjian Wan,
ZheJiang Academy of Agricultural Sciences, China

Reviewed by:
Ertugrul Filiz,
Duzce University, Turkey
Mehanathan Muthamilarasan,
University of Hyderabad, India
Venkata Suresh Bonthala,
Helmholtz Zentrum München, Germany
Yan Xiang,
Anhui Agricultural University, China

***Correspondence:**
Yangsheng Li
lysh2001@whu.edu.cn

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

Received: *03 March 2020*
Accepted: *11 May 2020*
Published: *19 June 2020*

Citation:
Kong W, Sun T, Zhang C, Qiang Y and Li Y (2020) Micro-Evolution Analysis Reveals Diverged Patterns of Polyol Transporters in Seven Gramineae Crops. Front. Genet. 11:565. doi: 10.3389/fgene.2020.00565

Polyol transporters (PLTs), also called polyol/monosaccharide transporters, is of significance in determining plant development and sugar transportation. However, the diverged evolutionary patterns of the *PLT* gene family in Gramineae crops are still unclear. Here a micro-evolution analysis was performed among the seven Gramineae representative crops using whole-genome sequences, i.e., *Brachypodium distachyon* (Bd), *Hordeum vulgare* (Hv), *Oryza rufipogon* (Or), *Oryza sativa* (Os), *Sorghum bicolor* (Sb), *Setaria italica* (Si), and *Zea mays* (Zm), leading to the identification of 12, 11, 12, 15, 20, 24, and 20 *PLT* genes, respectively. In this study, all *PLT* genes were divided into nine orthogroups (OGs). However, the number of *PLT* genes and the distribution of *PLT* OGs were not the same in these seven Gramineae species, and different OGs were also subject to different purification selection pressures. These results indicated that the *PLT* OGs of the *PLT* gene family have been expanded or lost unevenly in all tested species. Then, our results of gene duplication events confirmed that gene duplication events promoted the expansion of the *PLT* gene family in some Gramineous plants, namely, Bd, Or, Os, Si, Sb, and Zm, but the degree of gene family expansion, the type of *PLT* gene duplication, and the differentiation time of duplicate gene pairs varied greatly among these species. In addition, the sequence alignment and the internal repeat analysis of all PLTs protein sequences implied that the PLT protein sequences may originate from an internal repeat duplication of an ancestral six transmembrane helical units. Besides that, the protein motifs result highlighted that the PLT protein sequences were highly conserved, whereas the functional differentiation of the *PLT* genes was characterized by different gene structures, upstream elements, as well as co-expression analysis. The gene expression analysis of rice and maize showed that the *PLT* genes have a wide range of expression patterns, suggesting diverse biological functions. Taken together, our finding provided a perspective on the evolution differences and the functional characterizations of *PLT* genes in Gramineae representative crops.

Keywords: polyol transporters, Gramineae crops, orthologous gene pairs, orthogroups, duplicate gene pairs, functional differentiation

INTRODUCTION

Sugar is the main product of plant photosynthesis. Sugar (monosaccharides, sucrose, and polyols) plays an important role in the entire life cycle of plants (Deng et al., 2019; Kong et al., 2019a). It can not only provide energy for the growth and the development of plants but also is used for storage and transportation (Kühn and Grof, 2010). In various plant metabolic pathways, sugar can be used as a signal molecule (Wingenter et al., 2010; Kong et al., 2019a). Plant sugar transporters mainly mediate the transportation of sugars, participate in the loading and unloading of various sugars between "source and sink" tissues, and affect the distribution of plant carbohydrates (Kong et al., 2019a; Misra et al., 2019; Patil et al., 2019). To date, more and more sugar transporters have been identified and experimentally verified in various plant species, namely, monosaccharide transporters (MSTs), sucrose transporters, and sugars will eventually be exported transporters (Jeena et al., 2019; Kong et al., 2019a; Misra et al., 2019).

As we all know, polyol transporters (PLTs) belong to a subfamily of the MST superfamily (Deng et al., 2019; Kong et al., 2019a). Previous studies have demonstrated that PLTs can transport polyols and monosaccharides, also known as polyol/monosaccharide transporters (PMTs) (Büttner, 2007; Klepek et al., 2010). Besides that, PLTs reportedly loaded polyols into the phloem or unloaded sugars from the phloem for accumulation in sink tissues (Noiraud et al., 2001; Watari et al., 2004; Conde et al., 2007; Juchaux-Cachau et al., 2007). The first PLT was identified in celery (*Apium graveolens*), named AgMaT1, which had a high affinity for mannitol (Km = 275 μM) and showed a high expression level in mature leaves (Noiraud et al., 2001). Next, AgMaT2 was identified as an H(+)/mannitol co-transporter, which was less selective for other polyol molecules in celery and played a role in defense mechanisms (Juchaux-Cachau et al., 2007). What is more, Arabidopsis PMT5 (AtPMT5) transported a broad range of sugars, including cyclic and linear polyols, hexoses, and pentoses (Klepek, 2005; Reinders et al., 2005; Klepek et al., 2010). A tissue expression analysis showed that *AtPMT1* and *AtPMT2* were relatively highly expressed in some immature tissues, such as young germinating pollen and pollen tubes. Notably, they had a high ability to transport fructose (Klepek, 2005; Klepek et al., 2010). In *Lotus japonicus*, 14 putative *PLT* genes were identified and three genes (*LjPLT1*, *LjPLT3*, and *LjPLT9*) among them were found to respond to salinity and/or osmotic stresses (Lu et al., 2017). Another example is the *VvPLT1* gene in grapevine, which was up-regulated under salt- and water-deficit stresses, as well as exogenous salicylic acid and abscisic acid treatments, and had a high affinity for sorbitol and mannitol (Conde et al., 2015). Also, *OeMaT1* in *Olea europaea* was significantly induced by salinity and drought stresses (Conde et al., 2011). Another study reported that *Hevea brasiliensis HbPLT2* expressing yeast displayed active absorption of xylitol but a marginal level of absorption of inositol, mannitol, and sorbitol (Dusotoit-Coucaud et al., 2010).

Previous studies have focused on the transport substrate, tissue-specific expression, and subcellular localization of PLTs (Kühn and Grof, 2010). However, the evolutionary dynamics of *PLT* genes in plants are still unclear. Given the strong transportation capacity of PLTs, they are of great significance to the growth and the yield increase of Gramineae crops. On the other hand, several Gramineae crops with high-quality genomes are ideal model systems to study the short-term evolutionary dynamics of gene families and their consequences (Kong et al., 2019a,b,c). Here seven Gramineae crops, namely, *Brachypodium distachyon* (Bd), *Hordeum vulgare* (Hv), *Setaria italica* (Si), *Sorghum bicolor* (Sb), *Zea mays* (Zm), *Oryza rufipogon* (Or), and *Oryza sativa* (Os), were selected for a systematical study of Gramineae *PLT* gene family, including evolutionary analysis, topology, 3D structure, genetic structure, conserved domains, gene duplication events, selection pressures, orthogroups (OGs), co-expression analysis, upstream elements in promoters, and tissue expression analysis. Our study will facilitate further molecular evolution and function study of Gramineae PLTs.

MATERIALS AND METHODS

Plant Materials and Treatments

Seedlings of rice cultivar "9311" (*O. sativa* ssp. *indica*) were grown in containers with sponges as supporting materials in Yoshida solution in a Wuhan University greenhouse, with 16/8-h light/dark photoperiod at 26°C and 60% relative humidity (Deng et al., 2019; Kong et al., 2019a). The 12-day-old seedlings were changed to 40 and 4°C for the hot and the cold treatments. The leaves were collected at 0, 1, 3, 6, and 12 h after the hot or the cold treatment (Kong et al., 2019a).

In this study, three biological replicates of every treatment were generated, and each biological replicate was collected and pooled together from over 12 plants. To ensure the accuracy of the results, three technical replicates were set for each biological replicate in quantitative real-time PCR (qRT-PCR) (Kong et al., 2019d).

Identification and Phylogenetic Tree Construction of PLT Genes

Firstly, the genome databases of Bd, Hv, Si, Sb, Zm, and Or were obtained from Ensembl Plants release 41[1]. A genome database of Os (MSU 7.0) was downloaded from TIGR Database[2]. The hidden Markov model profile of the Sugar_tr (PF00083) was downloaded from the Pfam website[3]. Then, all these candidate PTL protein sequences were separately obtained by HMMER 3.2.1 with default parameters (Finn et al., 2011) and BlastP homology search with E-value cutoff e^{-5} (Kong et al., 2017; Wu et al., 2019). Only the longest splice variant of each gene locus was adopted in this study. After removing the redundant sequences, all candidate PLT protein sequences were submitted to SMART[4], Pfam[5], and National Center for Biotechnology

[1]http://plants.ensembl.org/index.html
[2]http://rice.plantbiology.msu.edu
[3]http://pfam.xfam.org/
[4]http://smart.embl-heidelberg.de/
[5]http://pfam.xfam.org/search/sequence

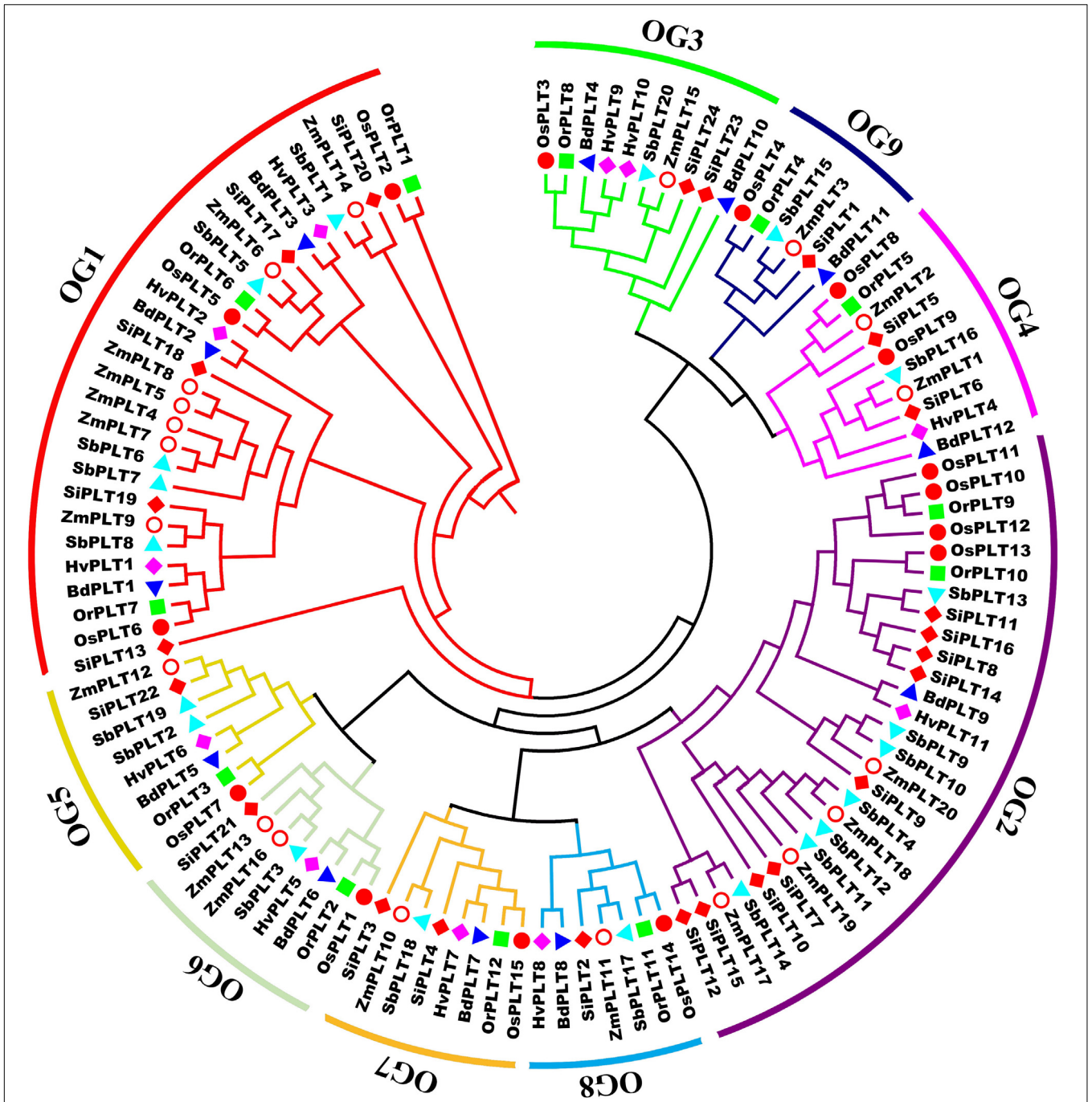
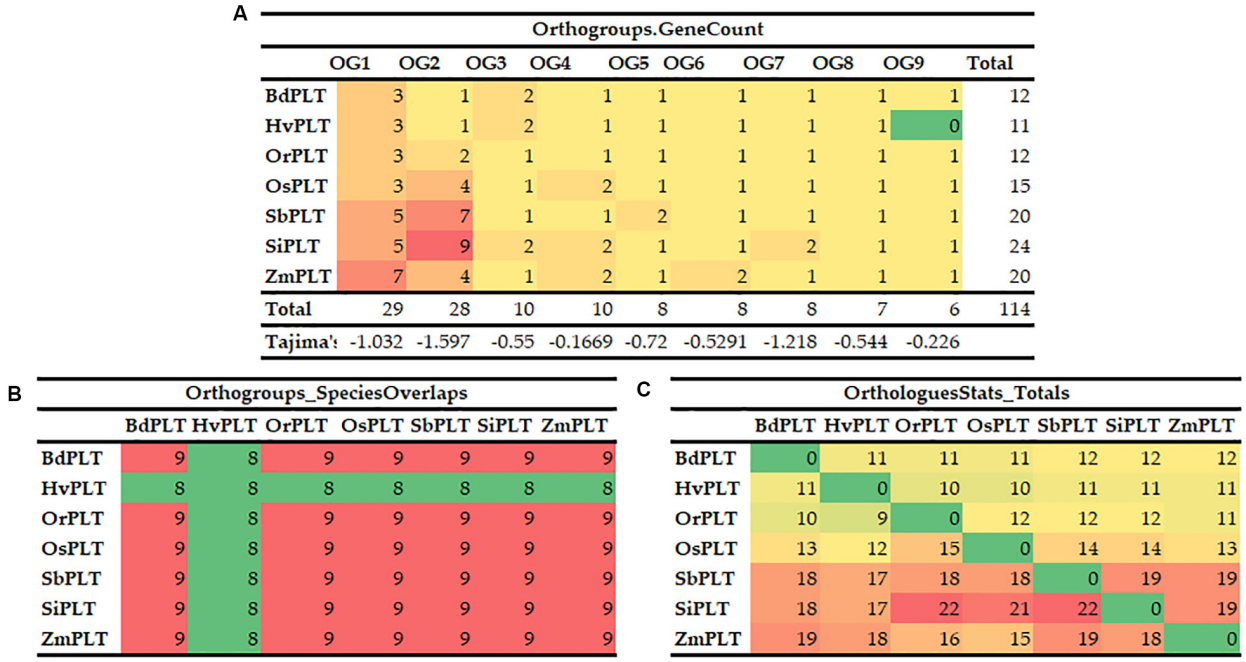


FIGURE 1 | A maximum likelihood phylogeny tree of the polyol transporter protein sequences from seven Gramineae crops, namely, *Brachypodium distachyon* (Bd), *Hordeum vulgare* (Hv), *Setaria italica* (Si), *Sorghum bicolor* (Sb), *Zea mays* (Zm), *Oryza rufipogon* (Or), and *Oryza sativa* ssp. *japonica* (Os). This tree was established by IQ-tree software with a 1,000-replicate bootstrap test and a 1,000-random-addition-replicate SH-aLRT test. The colors in the outside circles and the inside branches represent different orthogroups (OG), including OG1, OG2, OG3, OG4, OG5, OG6, OG7, OG8, and OG9. Markers of different shapes markers indicate the different species in the phylogeny tree. The orthogroup analysis was performed using OrthoFinder v2.0 software with default parameters.

Information Protein Basic Local Alignment Search Tool[6] to confirm the highly conserved Sugar_tr domain (Kong et al., 2017; Kong et al., 2019d).

Multiple sequence alignment of all full-length PLT protein sequences was performed using MUSCLE with default parameters (Edgar, 2004). We used ModelFinder software to determine that the best substitution model of the aligned PLT protein sequences was JTT + F + G4

[6]https://blast.ncbi.nlm.nih.gov/



A Orthogroups.GeneCount

	OG1	OG2	OG3	OG4	OG5	OG6	OG7	OG8	OG9	Total
BdPLT	3	1	2	1	1	1	1	1	1	12
HvPLT	3	1	2	1	1	1	1	1	0	11
OrPLT	3	2	1	1	1	1	1	1	1	12
OsPLT	3	4	1	2	1	1	1	1	1	15
SbPLT	5	7	1	1	2	1	1	1	1	20
SiPLT	5	9	2	2	1	1	2	1	1	24
ZmPLT	7	4	1	2	1	2	1	1	1	20
Total	29	28	10	10	8	8	8	7	6	114
Tajima's	-1.032	-1.597	-0.55	-0.1669	-0.72	-0.5291	-1.218	-0.544	-0.226	

B Orthogroups_SpeciesOverlaps

	BdPLT	HvPLT	OrPLT	OsPLT	SbPLT	SiPLT	ZmPLT
BdPLT	9	8	9	9	9	9	9
HvPLT	8	8	8	8	8	8	8
OrPLT	9	8	9	9	9	9	9
OsPLT	9	8	9	9	9	9	9
SbPLT	9	8	9	9	9	9	9
SiPLT	9	8	9	9	9	9	9
ZmPLT	9	8	9	9	9	9	9

C OrthologuesStats_Totals

	BdPLT	HvPLT	OrPLT	OsPLT	SbPLT	SiPLT	ZmPLT
BdPLT	0	11	11	11	12	12	12
HvPLT	11	0	10	10	11	11	11
OrPLT	10	9	0	12	12	12	11
OsPLT	13	12	15	0	14	14	13
SbPLT	18	17	18	18	0	19	19
SiPLT	18	17	22	21	22	0	19
ZmPLT	19	18	16	15	19	18	0

FIGURE 2 | Orthogroups of *PLT* genes among the seven Gramineae crops, namely, BdPLT, HvPLT, OrPLT, OsPLT, SbPLT, SiPLT, and ZmPLT. Orthogroup overlaps **(A)**, statistical results of orthologous genes between different crops **(B)**, and gene counts and Tajima's values of different orthogroups in these seven crops **(C)**.

(Kalyaanamoorthy et al., 2017). Subsequently, the tree reconstruction was conducted by the IQ-tree software with a bootstrap test for 1,000 replicates and an SH-aLRT test for 1,000 random addition replicates (Nguyen et al., 2015).

Orthogroup Analysis of PLTs, Gene Duplicate Events, and Selective Stress Analysis

PLT OGs were identified using OrthoFinder software with default parameters after an all-*vs*-all BlastP search of PLT protein sequences with E-value of 1 e^{-3} (Emms and Kelly, 2015; Kong et al., 2019d). Additionally, the selective forces on OGs were evaluated by Tajima's *D*-values using DnaSP 5.0 (Librado and Rozas, 2009; Kalyaanamoorthy et al., 2017).

In order to explore the expansion mechanism of the *PLT* gene family in Gramineous crops, the gene duplication events of *PLT* genes were analyzed with default parameters using the "duplicate_gene_classifier" script in Multiple Collinearity Scan toolkit X version (MCScanX) after a BlastP search of intraspecies PLT protein sequences with an E-value of 1 e^{-5} (Wang et al., 2012; Kong et al., 2019a). Next, the physical gene locations and the gene duplication events of *PLT* genes were visualized by Circos (Krzywinski et al., 2009).

In this study, we used DnaSP v5.0 software to determine non-synonymous (Ka) to synonymous (Ks) substitution rates (Ka/Ks) of duplicate gene pairs in seven tested species (Librado and Rozas, 2009; Kong et al., 2019b). Furthermore, the divergence time of the duplicate gene pairs was predicted using the formula: $T = Ks/(2 \times 9.1 \times 10^{-9}) \times 10^{-6}$ Mya (Deng et al., 2019; Kong et al., 2019c).

Internal Repeat Analyses, Transmembrane Helical Domains, 3D Structure Analysis, Conserved Motifs, and Gene Structures

The multiple alignment result of all PLT protein sequences was generated by the Clustal X software (Thompson et al., 1997) for internal repeat analyses for exploring the generation mechanism of PLT proteins. HHrepID was used to generate the internal repeats with these parameters: maximal number of multiple sequence alignment generation steps was one, score secondary structure was no, repeat family *P*-value threshold was 1 e^{-2}, self-alignment *P*-value threshold was 1 e^{-1}, merge rounds was three, and domain boundary detection was no (Biegert and Söding, 2008). At the same time, we also predicted the transmembrane helical (TMH) domains of all PLT protein alignment results in TMHMM Severv.2.0[7] (Kong et al., 2019b). On the other hand, the 3D structure models of all rice PLT protein sequences were analyzed in the SWISS-MODEL website[8] with default parameters, and the best template result was chosen for each PLT protein sequence (Waterhouse et al., 2018).

The MEME online tool[9] was used to search the conserved motifs of all PLTs protein sequences, with a motif width of 6–100

[7]http://www.cbs.dtu.dk/services/TMHMM/

[8]https://swissmodel.expasy.org/

[9]http://meme-suite.org/tools/meme

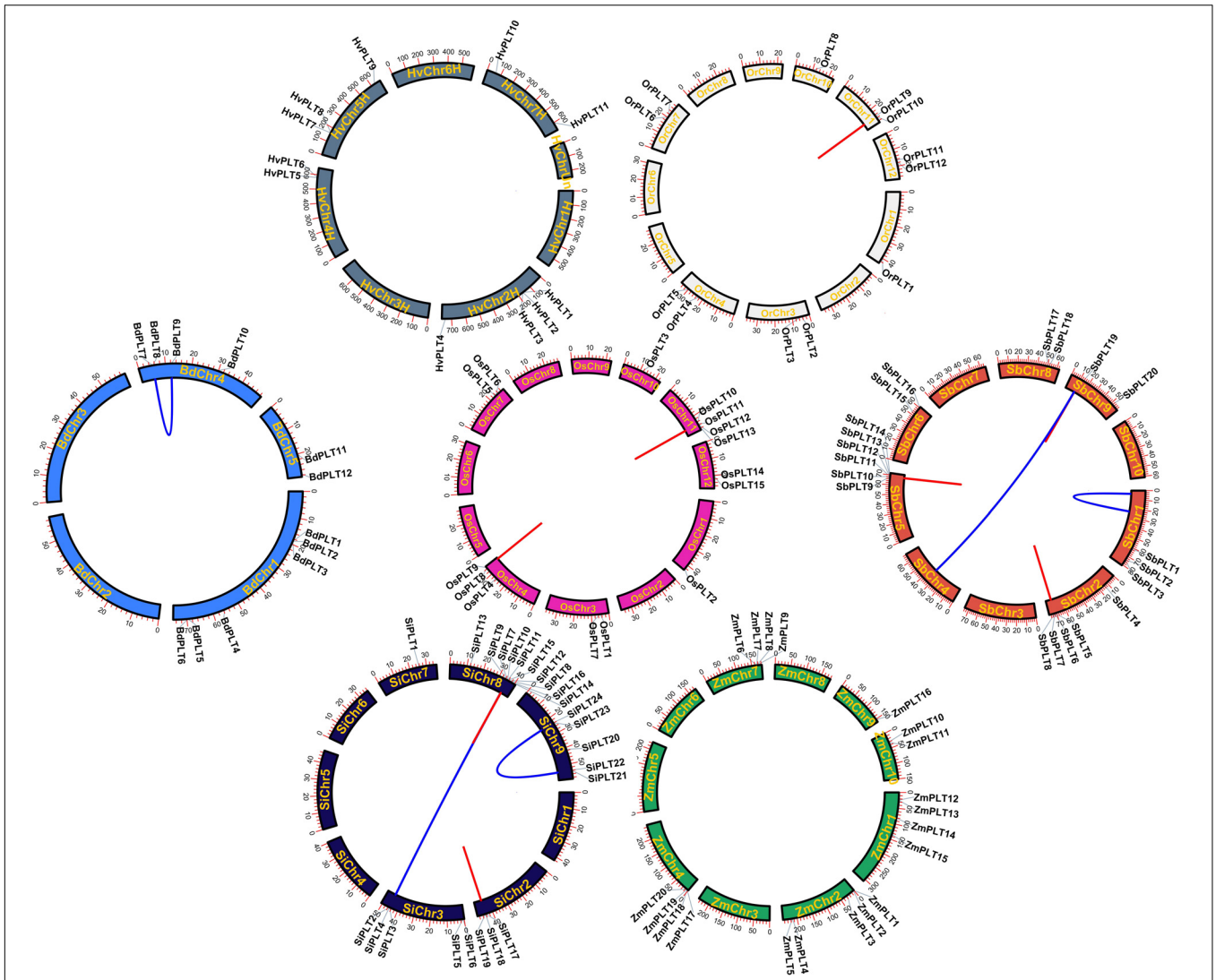


FIGURE 3 | The chromosome locations and gene duplication events of *PLT* genes in seven species, namely, *Brachypodium distachyon* (Bd), *Hordeum vulgare* (Hv), *Setaria italica* (Si), *Sorghum bicolor* (Sb), *Zea mays* (Zm), *Oryza rufipogon* (Or), and *Oryza sativa* ssp. *japonica* (Os). The red line meant tandem duplication and the blue line meant whole-genome duplication/segmental duplication and tandem duplication.

amino acids, a limit of 20 motifs, and all other default parameters (Bailey et al., 2015). A gene exon–intron structure analysis of all *PLT* genes was performed on the Gene Structure Display Server 2.0 (GSDS 2.0[10]) (Hu et al., 2015).

Quantitative Analysis of Rice *PLT* Genes Under Heat and Cold Stresses

All RNAs were extracted by the TRIzol reagent (Invitrogen, Beijing, China) and reversed to cDNAs using HiScript III 1st Strand cDNA Synthesis Kit (Vazyme, Shanghai, China). The qRT-PCR reactions (10 μl) were formulated using 2× SYBR Green qPCR Master Mix (US Everbright® Inc., Suzhou, China), following the manufacturer's instructions. The qRT-PCR reactions were detected on a CFX96 Touch™ Real-Time PCR Detection System (Bio-Rad, Hercules, CA, United States). The actin gene was used as an internal control (An et al., 2017; Kong et al., 2019b), and relative expression values were calculated by $2^{-\Delta\Delta CT}$ method based on three biological replicates, with each replicate having three technical replicates (Kong et al., 2019d).

Cis-Elements, Expression Analyses, and Co-expression Analysis

Two-kbp upstream sequences from the transcription start site of all *PLT* genes in these seven tested species were regarded as promoter sequences and extracted from gff3 and genetic fasta using Gtf/Gff3 sequences extractor tool in TBtools (Chen et al., 2018). Then, all upstream promoter sequences were submitted to the PLANTCARE website for the cis-elements identification (Lescot et al., 2002). To determine the expression patterns of *PLT*

[10]http://gsds.cbi.pku.edu.cn/

TABLE 1 | Ka/Ks values and the divergence times of duplicate gene pairs in seven tested species.

Seq_1	Seq_2	Orthogroup	Duplication type	Ka	Ks	Ka_Ks	The divergence time (Mya)
SiPLT15	SiPLT12	OG2	TD	0.0407	0.11324	0.35943	6.2219856
OsPLT11	OsPLT12	OG2	TD	0.0777	0.15527	0.50042	8.5314223
SbPLT9	SbPLT10	OG2	TD	0.04197	0.16748	0.25057	9.2021418
SbPLT11	SbPLT12	OG2	TD	0.06961	0.19956	0.34881	10.964933
ZmPLT7	ZmPLT8	OG1	TD	0.06642	0.22753	0.2919	12.501534
OrPLT9	OrPLT10	OG2	TD	0.21296	0.31901	0.66758	17.528014
SbPLT6	SbPLT7	OG1	TD	0.08471	0.31966	0.265	17.56382
OsPLT8	OsPLT9	OG4	TD	0.1421	0.3455	0.4113	18.983323
SiPLT18	SiPLT19	OG1	TD	0.12165	0.40981	0.29684	22.517235
ZmPLT12	ZmPLT16	OG5	WGD/SD	0.20089	0.43838	0.45825	24.086673
ZmPLT4	ZmPLT6	OG1	WGD/SD	0.15838	0.45722	0.34639	25.122193
SiPLT2	SiPLT11	OG8	WGD/SD	0.41746	0.5403	0.77264	29.687065
BdPLT7	BdPLT9	OG7	WGD/SD	0.43375	0.60859	0.71272	33.439094
SiPLT3	SiPLT8	OG7	WGD/SD	0.37861	0.62256	0.60815	34.206506
SiPLT24	SiPLT22	OG3	WGD/SD	0.39954	0.70733	0.56485	38.864047

genes in rice and maize, the RNA-seq data of various tissues from rize and maize were obtained from the MBKbase database[11] (Peng et al., 2020) and the maize eFP Browser[12] (data source: Hoopes et al., 2019), respectively. The fragments per kilobase of transcript per million mapped reads values were used to represent the gene expression levels in this study, and all heatmaps were made by R package (pheatmap) (Kolde, 2012).

To understand the potential molecular regulation and gene network, the correlation coefficient between *PLT* genes was calculated based on Pearson coefficient. On the other hand, the co-expression network of rice *PLT* genes was constructed based on massive expression data (including 9 tissues, 24 projects, and 284 experiments) using a co-expression tool in the Rice

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

[12]http://bar.utoronto.ca/efp_maize/cgi-bin/efpWeb.cgi

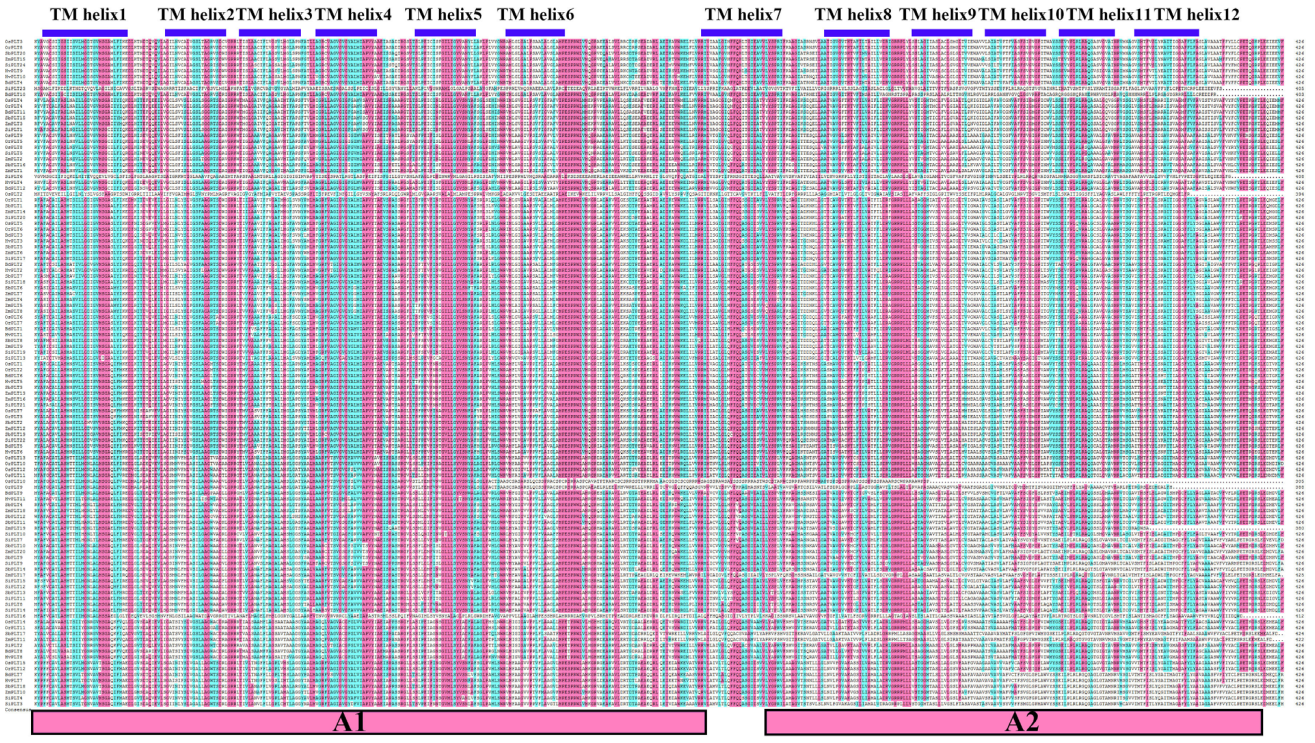


FIGURE 4 | Sequence alignment result of polyol transporter protein sequences. The residues in red background represent the highly conserved residues. Twelve transmembrane helices were marked by blue boxes and the two functional duplicated regions are shown in red boxes at the bottom, namely, A1 and A2.

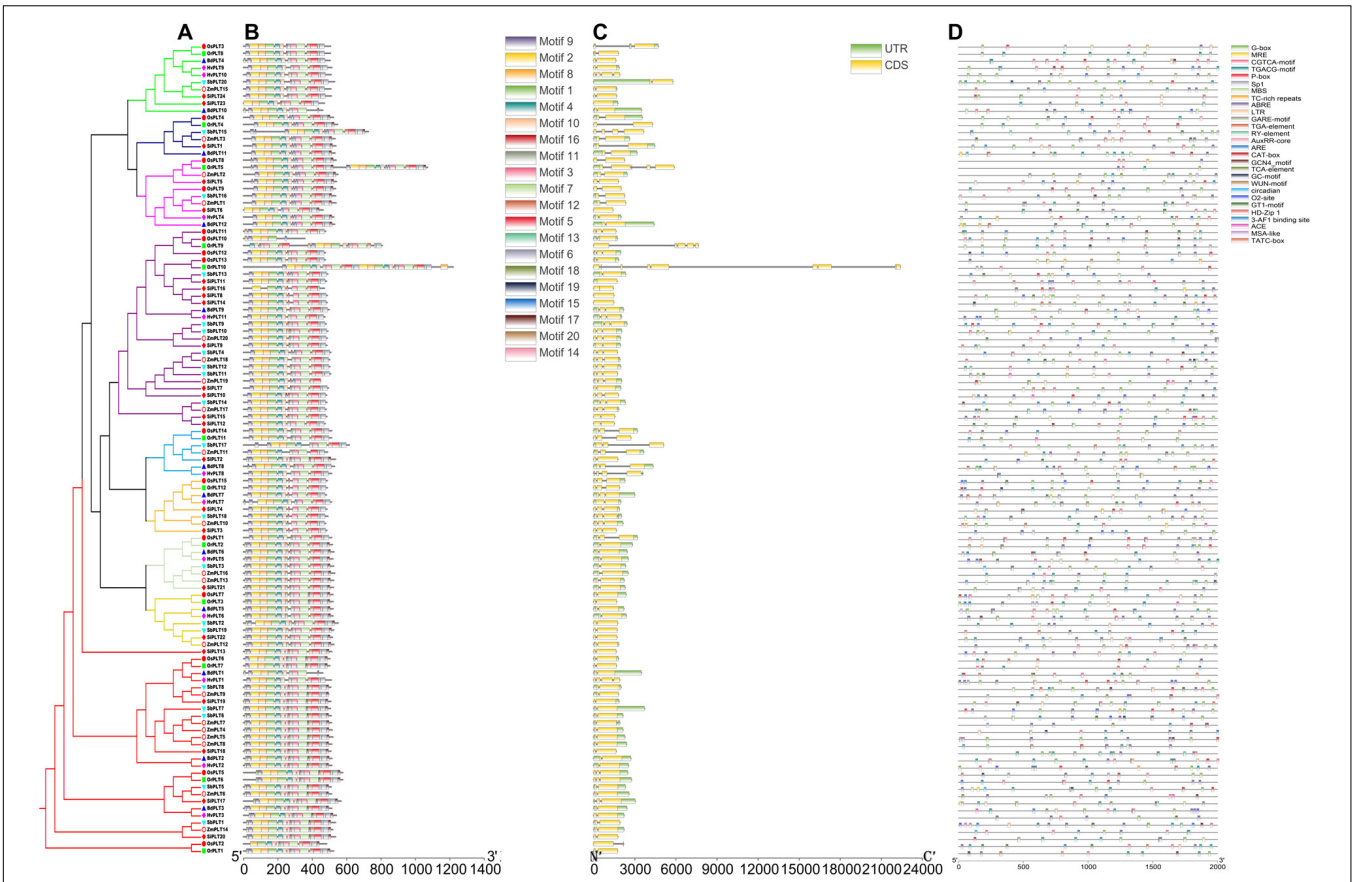


FIGURE 5 | The phylogenetic tree **(A)**, conserved motif **(B)**, exon/intron compositions **(C)** and cis-elements **(D)** in the upstream promoters of the *PLT* genes in these seven species. The different colors of the branches represent different orthogroups and the different shapes of the markers mean different species in the phylogeny tree. The relative lengths of genes and proteins are shown by the width of the gray bars. The different motifs/cis-elements are distinguished by different colored boxes. The exons and introns are displayed by yellow boxes and gray lines, respectively.

Expression Database (RED[13]). Here the MSU 7.0 gene IDs of the *PTL* genes were input IDs, and the Pearson's *r*-value was 0.85. The co-expressed gene network was visualized with Cytoscape V 3.6.1 (Shannon et al., 2003). Next, a gene ontology (GO) enrichment analysis of the obtained co-expressed genes was implemented by the GOseq R packages based on Wallenius non-central hyper-geometric distribution (Young et al., 2010).

RESULTS

Identification and OG Classification of *PLT* Genes

A total of 114 *PLT* genes were identified in these seven tested Gramineae crops, namely, 12 in Bd, 11 in Hv, 12 in Or, 15 in Os, 20 in Sb, 24 in Si, and 20 in Zm (**Figures 1**, **2A**, **5A** and **Supplementary Table S1**). We found that Sb, Si, and Zm had more *PLT* genes than Bd, Hv, Or, and Os (**Figures 1**, **2C**). In order to further analyze the reasons for the quantitative differences of the *PLT* genes in these tested species, a phylogenetic tree and OGs classification were conducted. As a result, all PLT proteins were clearly divided into nine OGs based on OrthoFinder analysis, namely, OG1, OG2, OG3, OG4, OG5, OG6, OG7, OG8, and OG9 (**Figures 1**, **2**). The *PTL* genes in the same OGs were clustered together in the phylogenetic tree (**Figures 1**, **2** and **Supplementary Tables S1, S2**). We thus inferred that the number difference of the *PLT* gene among these tested species may be caused by the unequal loss and expansion of OGs in these different Gramineae crops (**Figure 2** and **Supplementary Table S2**). For example, Hv had the smallest number of *PLT* genes, probably due to its loss of OG9 compared to other species in this study (**Figure 2** and **Supplementary Table S2**). Another reason was that OG1 and OG2 expanded more obviously in Sb, Si, and Zm than in Bd, Hv, Or, and Os, which resulted in more *PLT* genes in Sb, Si, and Zm than in Bd, Hv, Or, and Os (**Figure 2A** and **Supplementary Table S2**). In addition, the specific expansion of OGs in the individual tested species also contributed to the difference in the number of *PLT* genes between species, namely, the specific expansion of OG5 in Sb, OG6 in Zm, and OG7 in Si (**Figure 2A** and **Supplementary Table S2**). Interestingly, OG8 was a single OG, which implied that the *PLT* gene in OG8 was very conservative during the evolution of Gramineae

[13]http://expression.ic4r.org/

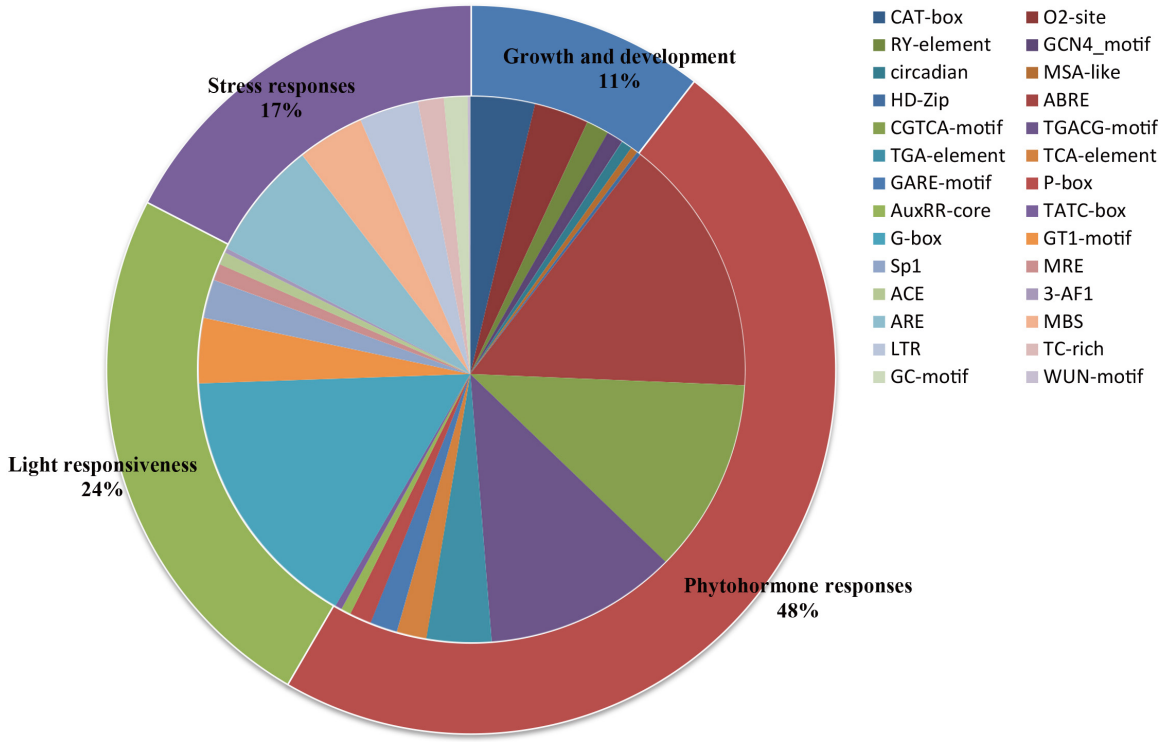


FIGURE 6 | Cis-acting elements of all *PLT* genes in seven Gramineae crops. The ratios of primary categories/cis-elements are displayed by different sizes in the pie charts.

crops (**Figure 2C** and **Supplementary Table S2**). Furthermore, the orthologs were counted and summarized based on the OrthoFinder analysis result (**Figure 2C** and **Supplementary Tables S3, S4**). As expected, Sb, Si, and Zm had more orthologs than Bd, Hv, Or, and Os (**Figure 2B**). All orthologs could be divided into four types: many to many, one to one, many to one, and one to many (**Supplementary Tables S3, S4**). Finally, we evaluated the selection pressure of these nine OGs according to Tajima's *D*-values using DnaSP 5.0 software. The result showed that the Tajima's *D*-values of these nine OGs were less than 0, and from −0.1669 to −1.597 (**Figure 2C**), which denoted that these nine OGs were subjected to different intensities of purification selection pressure.

Gene Duplicate Events of OGs and Selective Pressure of Duplicate Gene Pairs

In order to learn the specific expansion mechanism of OGs, duplicate gene pairs were identified and classified by MCScanX software (Wang et al., 2012). The result showed that a total of nine duplicate gene pairs originated from tandem duplication events, namely, one in Or, two in Os, two in Si, three in Sb, one in Zm, and zero in Hv and Bd (**Figure 3** and **Table 1**). Besides that, six duplicate gene pairs from whole-genome duplication (WGD) or segmental duplication (SD) events were discovered in Bd (one), Zm (two), and Si (three). These results proved that gene duplications played significant roles in the OG expansion of the *PLT* gene family in these important Gramineae crops. Subsequently, the selective stress and the divergence time of these duplicate gene pairs were calculated to reveal the expansion differences of the *PLT* genes in these Gramineae crops. All Ka/Ks values of *PLT* duplicate gene pairs ranged from 0.25 to 0.77, suggesting that these duplicate gene pairs were mainly subject to purification selection pressure. Notably, the divergence times of the tandem duplication gene pairs were both later (<23 MYa) than that of the WGD/SD gene pairs (>24 MYa). These results indicated that the earlier expansion mode of all the Gramineous tested crops was WGD/SD and the later expansion mode was TD.

TMH Prediction, Internal Repeat Analyses, Conserved Motifs, and Gene Structure Analysis

Our results of the protein sequence alignment and the TMH prediction revealed that all PLT protein sequences were strictly conserved and both had 12 TMHs, which supported the previous conclusion that PLT proteins consisted of 12 TMHs (**Figure 4**; Johnson et al., 2006). Interestingly, we also noticed that the first six TMHs were very similar to the next six TMHs (**Figure 4**). Therefore, we used the HHrepID program to perform the internal repeat analysis for all PLT protein sequences. As expected, we found that TMH1–TMH6 and TMH7–TMH12 were distributed in two duplicated regions, i.e., A1 and A2 (**Figure 4** and **Supplementary Figure S1**), which implied that PLT protein sequences may originate from an internal repeat duplication

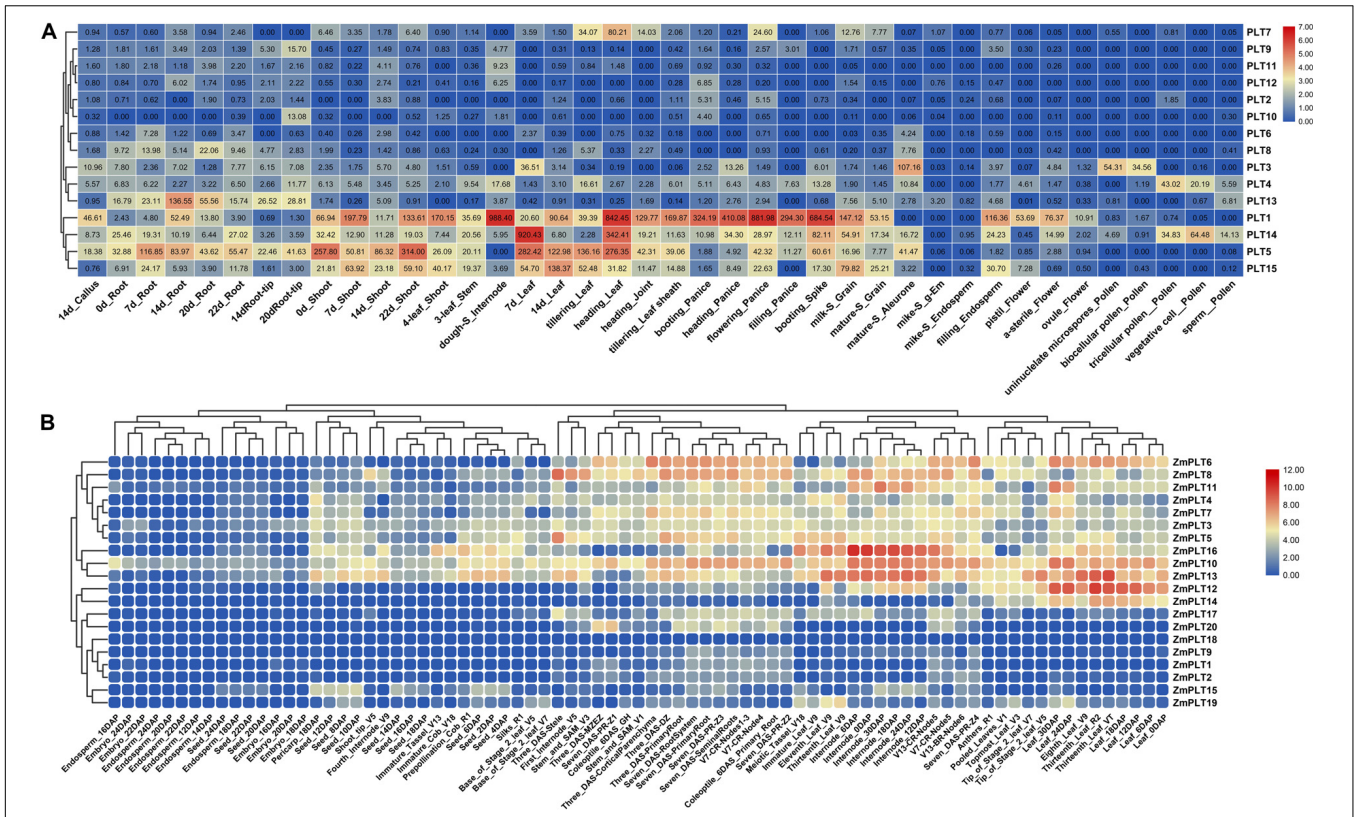


FIGURE 7 | Expression patterns of *PLT* genes in different tissues from rice **(A)** and maize **(B)**. The absolute expression values of rice *PLT* genes were obtained from MBKbase (http://www.mbkbase.org/rice). The absolute expression values of maize *PLT* genes were obtained from maize eFP Browser (http://bar.utoronto.ca/efp_maize/cgi-bin/efpWeb.cgi, data source: Hoopes et al., 2019).

of an ancestral six-TMH unit. Our findings were in agreement with the original hypotheses and provided a theoretical basis (Saier, 2000; Johnson et al., 2006). What is more, the 3D-structure models and ligands were predicted by the SWISS-MODEL website. We found that all PLT proteins belonged to two templates, namely, OsPLT4, OsPLT5, OsPLT6, OsPLT10, and OsPLT14 in 6h7d.1.A (description: sugar transport protein 10) and OsPLT1, OsPLT2, OsPLT3, OsPLT7, OsPLT8, OsPLT9, OsPLT11, OsPLT12, OsPLT13, and OsPLT15 in 6n3i.1.A (description: D-xylose transporter) (**Supplementary Figure S2** and **Supplementary Table S5**). Among them, OsPLT4 had a ligand; the ligand was 1× OLC: (2R)-2,3-dihydroxypropyl(9Z)-octadec-9-enoate (**Supplementary Table S5**).

To further explore the conservation and the sequence variation of the *PLT* genes in Gramineous crops, we separately utilized the MEME suite and GSDS 2.0 to predict the distribution of conserved domains and the intron–exon composition of all the *PLT* genes. Interestingly, all PLT protein sequences showed almost the same motif distribution, suggesting that the PLT proteins were very conserved in all tested species (**Figure 5B**). However, the *PLT* genes from the same OGs showed different intron–exon compositions among these seven tested species (**Figure 5C**). It was speculated that genetic sequence loss and exon fusion may occur during the evolution of the *PLT* gene family.

Cis-Elements in Upstream Promoters and Gene Expression Analysis of Rice and Maize *PLT* Genes

The analysis of upstream elements and gene expression can provide a perspective for gene function speculation and functional differentiation research (Cao et al., 2019; Kong et al., 2019a; Zhang et al., 2020). We thus investigated the upstream elements of all the *PLT* genes in Gramineous crops and the *PLT* gene expression in various tissues of rice and maize at different developmental stages. Additionally, the expression changes of rice *PLT* genes was also detected by qRT-PCR.

A total of 2,670 cis-elements responsible for growth and development, phytohormone responses, light responsiveness, and stress responses were identified and unevenly distributed on the promoters of all the *PLT* genes (**Figure 5D** and **Supplementary Table S6**). The statistics of all cis-elements showed that 48% of them belonged to phytohormone responses and were involved in abscisic acid responsiveness, auxin responsiveness (AuxRR-core and TGA-element), gibberellin responsiveness (GARE-motif, P-box, and TATC-box), MeJA responsiveness (TGACG-motif), and salicylic acid responsiveness (TCA-element) (**Figure 6** and **Supplementary Table S6**). A total of 24% of the elements were light responsiveness elements, including, G-box, GT1-motif, Sp1,

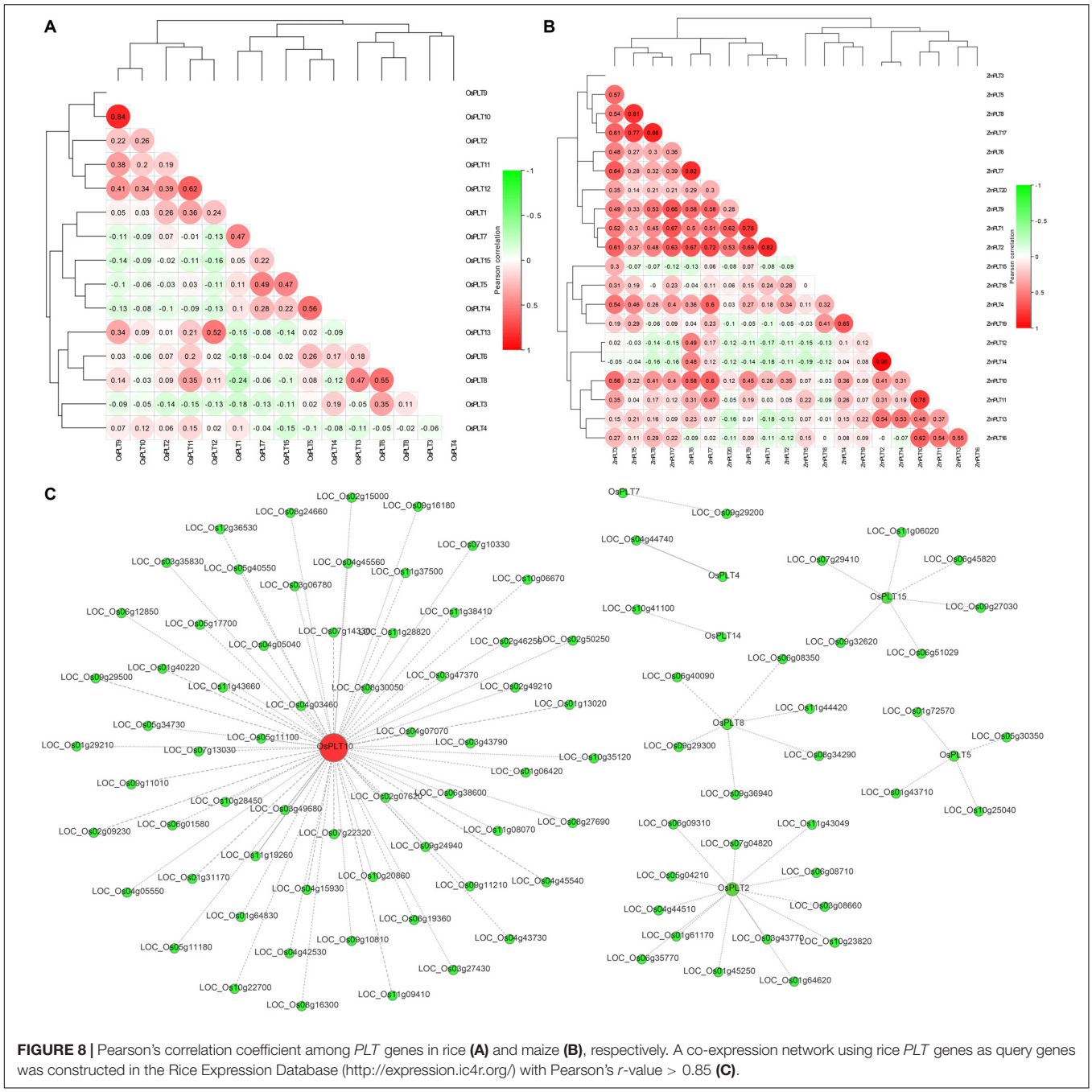


FIGURE 8 | Pearson's correlation coefficient among *PLT* genes in rice **(A)** and maize **(B)**, respectively. A co-expression network using rice *PLT* genes as query genes was constructed in the Rice Expression Database (http://expression.ic4r.org/) with Pearson's *r*-value > 0.85 **(C)**.

MRE, ACE, and 3-AF1 (**Figure 6** and **Supplementary Table S6**). The cis-elements associated with plant growth and development (11%) were CAT-box, responsible for meristem expression; O2, site responsible for zein metabolism regulation; RY, element responsible for seed-specific regulation; GCN4_motif, responsible for endosperm expression; circadian, responsible for circadian control; MSA-like element, responsible for cell cycle regulation; as well as HD-Zip, responsible for the differentiation of the palisade mesophyll cells (**Figure 6** and **Supplementary Table S6**). In addition, ARE, MBS, LTR, TC-rich, GC-motif, and WUN-motif cis-elements were involved in anaerobic induction, drought inducibility, low temperature responsiveness, defense and stress responsiveness, anoxic-specific inducibility, and wound responsiveness, respectively (**Figure 6** and **Supplementary Table S6**).

The expression analyses of the *PLT* genes in rice and maize showed that the *PLT* gene has different expression profiles in various tissues, which indicated that the *PLT* gene may have multiple biological functions and participate in the growth and the development of multiple organs (**Figure 7**). Several *PLT* genes showed tissue-specific expression, suggesting that they play a role in specific tissues (**Figure 7**). For instance, *OsPLT13*

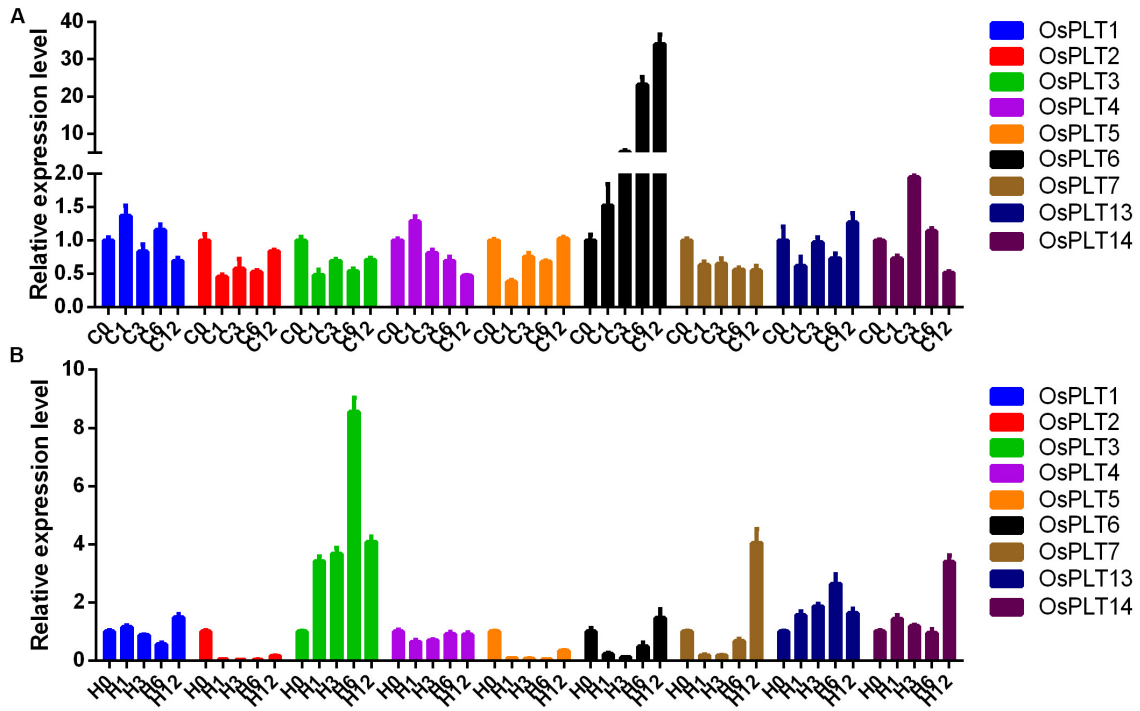


FIGURE 9 | Expression changes of rice PLT genes under cold **(A)** and heat **(B)** stresses by qRT-PCR. The numbers 0, 1, 3, 6, and 12 represent 0, 1, 3, 6, and 12 h after treatments.

gene was expressed in a root-specific manner (**Figure 7A**). Similarly, *OsPLT3* and *OsPLT4* were specifically expressed in pollen at different developmental stages. In order to further explore the functional differentiation of the *PLT* genes in rice and maize, a correlation coefficient between the *PLT* genes was calculated through Pearson coefficient (**Figure 8**). Most of the correlation coefficients between the *PLT* genes in rice were less than 0.8 (**Figure 8A**) and the maize species had a similar situation (**Figure 8B**). However, the correlation coefficients between several *PLT* gene pairs were greater than 0.8, namely, *OsPLT9* and *OsPLT10*, Zm*PLT5* and *ZmPLT8*, *ZmPLT8* and *ZmPLT15*, *ZmPLT6* and *ZmPLT7*, *ZmPLT1* and *ZmPLT2*, as well as *ZmPLT12* and *ZmPLT14* (**Figures 8A,B**). These results revealed that the *PLT* gene family in rice and maize have undergone significant functional differentiation, but some *PLT* genes may be functionally redundant. Notably, the correlation coefficients of five *PLT* duplicate gene pairs in rice and maize were less than 0.8, namely, *OsPLT11* and *OsPLT12* (0.62), *OsPLT8* and *OsPLT9* (0.14), *ZmPLT7* and *ZmPLT8* (0.32), *ZmPLT12* and *ZmPLT16* (-0), as well as *ZmPLT4* and *ZmPLT6* (0.36), which demonstrated that these *PLT* duplicate gene pairs had obvious subfunctionalization.

Co-expression Network, GO Annotation, and qRT-PCR of Rice *PLT* Gene Under Heat and Cold Stresses

In this study, a co-expression network using rice *PLT* genes as query genes was constructed on the RED website (Pearson's *r*-value > 0.85) to find the functional partners of rice *PLT* genes. A total of 96 genes were identified in this co-expression network (**Figure 8C** and **Supplementary Table S6**). *OsPLT10* was a hub gene and had 64 partners, while *OsPLT2*, *OsPLT8*, *OsPLT15*, and *OsPLT5* had 13, 6, 6, and 4 partners, respectively (**Figure 8C**). A GO annotation was performed to analyze the biological processes involved in this co-expression network (**Supplementary Table S7**). The results showed that the genes in this co-expression network can be annotated to 269 GO terms, suggesting that the co-expression network may be involved in multiple biological processes of rice growth and development (**Supplementary Table S7**).

The expression changes of nine rice *PLT* genes under cold and heat stresses were detected. The majority of the rice *PLT* genes were down-regulated under heat stress and cold stress (**Figure 9**). However, the expression level of *OsPLT6* was induced by cold stress. *OsPLT3* and *OsPLT13* had obvious up-regulated expressions under heat stress (**Figure 9**).

DISCUSSION

Gene Duplication and Orthogroup Expansion Promoted the Expansion of the *PLT* Gene Family in Gramineae Crops

Previous gene family studies have demonstrated that gene duplication events play a crucial role in family gene expansion (Büttner, 2007; Van Holle and Van Damme, 2015;

Kong et al., 2019b,d). Our earlier studies of MSTs also found that duplication events led to the expansion of the rice STP and ERD gene families (Deng et al., 2019). In this study, we found that gene duplication events also affected the expansion of PTL gene families in some Gramineous crops and ultimately led to differences in the total number of *PLT* genes in these seven tested species. We also found that some OGs in the Gramineous crops had an obvious expansion phenomenon, especially OG1 and OG2. The expansion of OGs had species bias, and the expansion was larger in Sb, Si, and Zm than in Bd, Hv, Or, and Os. On the other hand, the unequal loss of OGs also affected the total number of *PLT* genes in each species. For example, Hv had the least *PLT* gene compared with Bd, Or, Os, Si, Sb, and Zm, which was due to the fact that Hv did not have any *PLT* gene duplication events, that the expansion of OGs was not obvious, and that OG9 was lost. Our results supported the theoretical models of gene family evolution, such that gene families continuously undergo stochastic gain and loss events (Zimmer et al., 1980; Reed and Hughes, 2004). On the other hand, the types and the numbers of *PLT* gene duplication events were not the same in the tested crops, which indicated that the evolution and the expansion patterns of the *PLT* gene family were different in each species. Thus, the Ka/Ks values and the divergence times of the duplicate gene pairs were calculated. All Ka/Ks values of duplicate gene pairs from different types of duplication events were less than 0. Interestingly, the divergence times of the TD duplicate gene pairs (<23 MYa) were later than that of the WGD/SD gene pairs (>24 MYa). This result indicated that different types of gene duplication events were responsible for the expansion of the *PLT* gene family in different evolutionary periods of Gramineous crops, but the driving force and the cause of this phenomenon were still unclear.

Gene Structure Variations and Different Selection Pressures Promoted the Functional Differentiation of PLT

Previous studies reported that PLT proteins had 12 transmembrane domains and high consistency at the N- and C-termini (Baud et al., 2005; Johnson et al., 2006; Klepek et al., 2010). Our results of the conservative motifs and the transmembrane domain of all PLT protein sequences revealed that the PLTs were very conserved, with 12 transmembrane domains. Internal repeat duplication analysis confirmed that the PLTs protein sequence may originate from an internal repeat duplication of an ancestral six-TMHs unit (Saier, 2000; Johnson et al., 2006). In this study, we found that nine OGs had different purification selection pressures, which provided evolutionary motivation for the functional differentiation of the members of the *PLT* gene family. Significant differences in upstream promoter elements, genomic intron–exon structure, and gene expression among *PLT* genes indeed confirmed that the *PLT* genes were functionally differentiated in rice and maize. The theoretical models of gene duplication proposed that duplicate gene pairs continuously underwent functional differentiation (subfunctionalization), gained new functions (neofunctionalization), or lost original functions (pseudogenization) after the gene duplication events (Lynch and Force, 2000; He and Zhang, 2005; Innan and Kondrashov, 2010). Gene expression data and co-expression analysis could provide a perspective on the functional fates of duplicate gene pairs (Reinders et al., 2005; Johnson et al., 2006; Kong et al., 2019b). The correlation coefficients of all duplicate gene pairs were less than 0.8, suggesting that these gene pairs had significant functional differentiation. It should be noted that the correlation coefficients between some *PLT* genes in rice and maize are greater than 0.8, which indicates that the *PLT* genes have functional redundancy.

PLT Genes Were Widely Involved in Plant Growth and Development and Stress Responses

The RNA data analysis results from a large number of tissues in rice and maize reflected that the *PLT* genes have a wide range of expression patterns. In this study, we used rice *PLT* genes as the query gene to construct a co-expression network. The GO annotation results of the co-expression network genes showed that multiple GO terms were involved, and *OsPLT10* was the key core gene of the co-expression network. These results indicated that the *PLT* genes have multiple biological functions in plants. Cold and heat stresses frequently severely affect the growth, development, and food yield of crops in many countries (Mittler, 2006; Zhao et al., 2014; Kong et al., 2019b). Previous studies of PLT proteins have focused on transporting monosaccharide types, salt stress response, and drought stress response (Noiraud et al., 2001; Watari et al., 2004; Klepek, 2005; Reinders et al., 2005; Conde et al., 2007, 2011, 2015; Juchaux-Cachau et al., 2007; Dusotoit-Coucaud et al., 2010; Lu et al., 2017). However, few studies have been conducted on the effects of heat or cold stress on the *PLT* genes. In this study, we found that *OsPLT6* was identified as a cold stress-inducible gene, whereas *OsPLT3* and *OsPLT13* were heat stress-inducible genes. These results suggested that the *PLT* genes may also be involved in cold and heat stresses. Although the co-expression network and the qRT-PCR data provided us with clues to the function prediction of rice *PLT* genes and candidate gene screening for genetic breeding, these remain to be experimentally verified in future studies.

DATA AVAILABILITY STATEMENT

All datasets generated/analyzed for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

WK performed all of the experiments, analyzed the data, prepared the figures and the tables, and wrote the manuscript. WK and YL conceived and designed the experiments. TS, CZ, and YQ prepared parts of the figures and the tables. All the authors read and approved the final version of the manuscript.

FUNDING

The National Special Key Project for Transgenic Breeding (Grant No. 2016ZX08001001) and the National Key Research and Development Program of China (2016YFD0100400) funded this study.

ACKNOWLEDGMENTS

We thank the reviewers and the editor for their careful reading and helpful comments on this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00565/full#supplementary-material

FIGURE S1 | An internal repeat of *PLT* genes. The proteins themselves are represented by the dark diagonal lines, and the duplicated regions are represented by the above the dark diagonal lines.

FIGURE S2 | 2D models of 6h7d.1.A and 6n3i.1.A.

FIGURE S3 | Sketch map of maize tissues.

TABLE S1 | Detailed information on the *PLT* genes in seven tested Gramineous crops.

TABLE S2 | Nine orthogroups among the seven tested Gramineous crops.

TABLE S3 | Statistics of the four types of orthologs between different Gramineous crops.

TABLE S4 | Orthologs list among different Gramineous crops.

TABLE S5 | 3D structure prediction results of rice PLTs proteins.

TABLE S6 | Genes and Pearson's *r*-values in the co-expression network.

TABLE S7 | GO enrichment of genes in the co-expression network.

REFERENCES

An, B., Jie, L., Xiaolong, D., Silan, C., Chao, O., Huiyun, S., et al. (2017). Silencing of D-lactate dehydrogenase impedes glyoxalase system and leads to methylglyoxal accumulation and growth inhibition in rice. *Front. Plant Sci.* 8:2071. doi: 10.3389/fpls.2017.02071

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

Baud, S., Wuillème, S., Lemoine, R., Kronenberger, J., Caboche, M., Lepiniec, L. C., et al. (2005). The AtSUC5 sucrose transporter specifically expressed in the endosperm is involved in early seed development in *Arabidopsis*. *Plant J.* 43, 824–836. doi: 10.1111/j.1365-313x.2005.02496.x

Biegert, A., and Söding, J. (2008). HHrepID: de novo protein repeat identification by probabilistic consistency. *Bioinformatics* 24, 807–814. doi: 10.1093/bioinformatics/btn039

Büttner, M. (2007). The monosaccharide transporter(-like) gene family in *Arabidopsis*. *FEBS Lett.* 581, 2318–2324. doi: 10.1016/j.febslet.2007.03.016

Cao, Y., Liu, W., Zhao, Q., Long, H., Li, Z., Liu, M., et al. (2019). Integrative analysis reveals evolutionary patterns and potential functions of SWEET transporters in Euphorbiaceae. *Int. J. Biol. Macromol.* 139, 1–11. doi: 10.1016/j.ijbiomac.2019.07.102

Chen, C., Chen, H., He, Y., and Xia, R. (2018). TBtools, a toolkit for biologists integrating various biological data handling tools with a user-friendly interface. *BioRxiv* [Preprint] doi: 10.1101/289660

Conde, A., Regalado, A., Rodrigues, D., Costa, J. M., Blumwald, E., Chaves, M. M., et al. (2015). Polyols in grape berry: transport and metabolic adjustments as a physiological strategy for water-deficit stress tolerance in grapevine. *J. Exp. Bot.* 66, 889–906. doi: 10.1093/jxb/eru446

Conde, A., Silva, P., Agasse, A., Conde, C., and Geros, H. (2011). Mannitol transport and mannitol dehydrogenase activities are coordinated in *Olea europaea* under salt and osmotic stresses. *Plant Cell Physiol.* 52, 1766–1775. doi: 10.1093/pcp/pcr121

Conde, C., Silva, P., Agasse, A., Lemoine, R., Delrot, S., Tavares, R., et al. (2007). Utilization and transport of mannitol in *Olea europaea* and implications for salt stress tolerance. *Plant Cell Physiol.* 48, 42–53. doi: 10.1093/pcp/pcl035

Deng, X., An, B., Zhong, H., Yang, J., Kong, W., and Li, Y. (2019). A novel insight into functional divergence of the MST gene family in rice based on comprehensive expression patterns. *Genes* 10:239. doi: 10.3390/genes10030239

Dusotoit-Coucaud, A., Porcheron, B., Brunel, N., Kongsawadworakul, P., Franchel, J., Viboonjun, U., et al. (2010). Cloning and characterization of a new polyol transporter (HbPLT2) in *Hevea brasiliensis*. *Plant Cell Physiol.* 51, 1878–1888. doi: 10.1093/pcp/pcq151

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Emms, D. M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16:157.

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39, W29–W37.

He, X., and Zhang, J. (2005). Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* 169, 1157–1164. doi: 10.1534/genetics.104.037051

Hoopes, G. M., Hamilton, J. P., Wood, J. C., Esteban, E., Pasha, A., Vaillancourt, B., et al. (2019). An updated gene atlas for maize reveals organ-specific and stress-induced genes. *Plant J.* 97, 1154–1167. doi: 10.1111/tpj.14184

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

Innan, H., and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* 11, 97–108. doi: 10.1038/nrg2689

Jeena, G. S., Kumar, S., and Shukla, R. K. (2019). Structure, evolution and diverse physiological roles of SWEET sugar transporters in plants. *Plant Mol. Biol.* 100, 351–365. doi: 10.1007/s11103-019-00872-4

Johnson, D. A., Hill, J. P., and Thomas, M. A. (2006). The monosaccharide transporter gene family in land plants is ancient and shows differential subfamily expression and expansion across lineages. *BMC Evol. Biol.* 6:64. doi: 10.1186/1471-2148-6-64

Juchaux-Cachau, M., Landouar-Arsivaud, L., Pichaut, J.-P., Campion, C., Porcheron, B., Jeauffre, J., et al. (2007). Characterization of AgMaT2, a plasma membrane mannitol transporter from celery, expressed in phloem cells, including phloem parenchyma cells. *Plant Physiol.* 145, 62–74. doi: 10.1104/pp.107.103143

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Na. Methods* 14, 587–589. doi: 10.1038/nmeth.4285

Klepek, Y. S. (2005). *Arabidopsis* polyol transporter5, a new member of the monosaccharide transporter-like superfamily, mediates H+-symport of numerous substrates, including myo-inositol, glycerol, and ribose. *Plant Cell* 17, 204–218. doi: 10.1105/tpc.104.026641

Klepek, Y. S., Volke, M., Konrad, K. R., Wippel, K., Hoth, S., Hedrich, R., et al. (2010). *Arabidopsis thaliana* polyol/monosaccharide transporters 1 and 2: fructose and xylitol/H+ symporters in pollen and young xylem cells. *J. Exp. Bot.* 61, 537–550. doi: 10.1093/jxb/erp322

Kolde, R. (2012). *Pheatmap: Pretty Heatmaps. R Package Version 61, 617.*

Kong, W., An, B., Zhang, Y., Yang, J., Li, S., Sun, T., et al. (2019a). Sugar transporter proteins (STPs) in Gramineae crops: comparative analysis,

phylogeny, evolution, and expression profiling. *Cells* 8:560. doi: 10.3390/cells8060560

Kong, W., Gong, Z., Zhong, H., Zhang, Y., Zhao, G., Gautam, M., et al. (2019b). Expansion and evolutionary patterns of glycosyltransferase family 8 in gramineae crop genomes and their expression under salt and cold stresses in *Oryza sativa* ssp. *japonica*. *Biomolecules* 9:188. doi: 10.3390/biom9050188

Kong, W., Zhang, Y., Deng, X., Zhang, C., and Li, Y. (2019c). Comparative genomic and transcriptomic analyses suggests the evolutionary dynamic of GH3 genes in Gramineae crops. *Front. Plant Sci.* 10:1297. doi: 10.3389/fpls.2019.01297

Kong, W., Zhong, H., Deng, X., Gautam, M., Gong, Z., Zhang, Y., et al. (2019d). Evolutionary analysis of GH3 genes in six Oryza species/subspecies and their expression under salinity stress in *Oryza sativa* ssp. *japonica*. *Plants* 8:30. doi: 10.3390/plants8020030

Kong, W., Yang, S., Wang, Y., Bendahmane, M., and Fu, X. (2017). Genome-wide identification and characterization of aquaporin gene family in *Beta vulgaris*. *PeerJ*. 5:e3747. doi: 10.7717/peerj.3747

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

Kühn, C., and Grof, C. P. (2010). Sucrose transporters of higher plants. *Curr. Opin. Plant Biol.* 13, 287–297. doi: 10.1016/j.pbi.2010.02.001

Lescot, M., Déhais, 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

Librado, P., and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25, 1451–1452. doi: 10.1093/bioinformatics/btp187

Lu, T., Leru, L., Yehu, Y., Mingchao, H., Yanbo, C., Xinlan, X., et al. (2017). Heterogeneity in the expression and subcellular localization of polyol/monosaccharide transporter genes in *Lotus japonicus*. *PloS One* 12:e0185269. doi: 10.1371/journal.pone.0185269

Lynch, M., and Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154, 459–473.

Misra, V. A., Wafula, E. K., Wang, Y., dePamphilis, C. W., and Timko, M. P. (2019). Genome-wide identification of MST, SUT and SWEET family sugar transporters in root parasitic angiosperms and analysis of their expression during Host parasitism. *BMC Plant Biol.* 19:196. doi: 10.1186/s12870-019-1786-y

Mittler, R. (2006). Abiotic stress, the field environment and stress combination. *Trends Plant Sci.* 11, 15–19. doi: 10.1016/j.tplants.2005.11.002

Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Noiraud, N., Maurousset, L., and Lemoine, R. (2001). Identification of a mannitol transporter, AgMaT1, in celery phloem. *Plant Cell* 13, 695–705. doi: 10.1105/tpc.13.3.695

Patil, S. S., Prashant, R., Kadoo, N. Y., Upadhyay, A., and Gupta, V. S. (2019). Global study of MFS superfamily transporters in arabidopsis and grapes reveals their functional diversity in plants. *Plant Gene* 18:100179. doi: 10.1016/j.plgene.2019.100179

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.

Reed, W. J., and Hughes, B. D. (2004). A model explaining the size distribution of gene and protein families. *Math. Biosci.* 189, 97–102. doi: 10.1016/j.mbs.2003.11.002

Reinders, A., Panshyshyn, J. A., and Ward, J. M. (2005). Analysis of transport activity of *Arabidopsis* sugar alcohol permease homolog AtPLT5. *J. Biol. Chem.* 280, 1594–1602. doi: 10.1074/jbc.m410831200

Saier, M. H. Jr. (2000). Families of transmembrane sugar transport proteins: microreview. *Mol. Microbiol.* 35, 699–710. doi: 10.1046/j.1365-2958.2000.01759.x

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

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25, 4876–4882. doi: 10.1093/nar/25.24.4876

Van Holle, S., and Van Damme, E. J. M. (2015). Distribution and evolution of the lectin family in soybean (*Glycine max*). *Molecules* 20, 2868–2891. doi: 10.3390/molecules20022868

Wang, Y., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40, e49–e49. doi: 10.1093/nar/gkr1293

Watari, J., Kobae, Y., Yamaki, S., Yamada, K., and Shiratake, K. (2004). Identification of sorbitol transporters expressed in the phloem of apple source leaves. *Plant Cell Physiol.* 45, 1032–1041. doi: 10.1093/pcp/pch121

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., et al. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46, W296–W303.

Wingenter, K., Schulz, A., Wormit, A., Wic, S., Trentmann, O., Hoermiller, I. I., et al. (2010). Increased activity of the vacuolar monosaccharide transporter TMT1 alters cellular sugar partitioning, sugar signaling, and seed yield in *Arabidopsis*. *Plant Physiol.* 154, 665–677. doi: 10.1104/pp.110.162040

Wu, G.-Q., Wang, J.-L., and Li, S.-J. (2019). Genome-wide identification of Na+/H+ antiporter (NHX) genes in sugar beet (*Beta vulgaris* L.) and their regulated expression under salt stress. *Genes* 10:401. doi: 10.3390/genes10050401

Young, M. D., Wakefield, M. J., Smyth, G. K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* 11:R14.

Zhao, X., Wensheng, W., Fan, Z., Jianli, D., Zhikang, L., Binying, F., et al. (2014). Comparative metabolite profiling of two rice genotypes with contrasting salt stress tolerance at the seedling stage. *PLoS One*. 9:e108020. doi: 10.1371/journal.pone.0108020

Zhang, L., Jiao, C., Cao, Y., Cheng, X., Wang, J., Jin, Q., et al. (2020). Comparative analysis and expression patterns of the PLP_deC genes in *Dendrobium officinale*. *Int. J. Mol. Sci.* 21:54. doi: 10.3390/ijms21010054

Zimmer, E. A., Martin, S. L., Beverley, S. M., Kan, Y. W., and Wilson, A. C. (1980). Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. *Proc. Natl. Acad. Sci. U. S. A.* 77, 2158–2162. doi: 10.1073/pnas.77.4.2158

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.

Copyright © 2020 Kong, Sun, Zhang, Qiang and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 30 June 2020
doi: 10.3389/fgene.2020.00701

Chromosome Level Genome Assembly of *Andrographis paniculata*

Ying Liang[1†], Shanshan Chen[2†], Kunhua Wei[1†], Zijiang Yang[3], Shengchang Duan[4], Yuan Du[4], Peng Qu[1], Jianhua Miao[1], Wei Chen[5,6,7*] and Yang Dong[1,3,5,7*]*

[1] Guangxi Key Laboratory of Medicinal Resources Protection and Genetic Improvement, Guangxi Botanical Garden of Medicinal Plants, Nanning, China, [2] BGI College, Zhengzhou University, Zhengzhou, China, [3] National and Local Joint Engineering Research Center on Germplasm Innovation and Utilization of Chinese Medicinal Materials in Southwest China, Yunnan Agricultural University, Kunming, China, [4] NowBio Biotechnology Company, Kunming, China, [5] State Key Laboratory for Conservation and Utilization of Bio-Resources in Yunnan, Yunnan Agricultural University, Kunming, China, [6] College of Agronomy and Biotechnology, Yunnan Agricultural University, Kunming, China, [7] Yunnan Research Institute for Local Plateau Agriculture and Industry, Kunming, China

OPEN ACCESS

Edited by:
Xiaoming Song,
North China University of Science and Technology, China

Reviewed by:
Marcio Resende,
University of Florida, United States
Ergude Bao,
Beijing Jiaotong University, China

***Correspondence:**
Jianhua Miao
mjh1962@vip.163.com
Wei Chen
wchenntr@gmail.com
Yang Dong
loyalyang@163.com

[†] These authors have contributed equally to this work

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

Received: 20 February 2020
Accepted: 09 June 2020
Published: 30 June 2020

Citation:
Liang Y, Chen S, Wei K, Yang Z, Duan S, Du Y, Qu P, Miao J, Chen W and Dong Y (2020) Chromosome Level Genome Assembly of Andrographis paniculata. Front. Genet. 11:701. doi: 10.3389/fgene.2020.00701

Andrographis paniculata (Chinese name: Chuanxinlian) is an annual dicotyledonous medicinal plant widely grown in China and Southeast Asia. The dried plant has a highly acclaimed usage in the traditional Chinese medicine for its antipyretic, anti-inflammatory, and analgesic effects. In order to help delineate the biosynthetic pathways of various secondary metabolites, we report in this study a high-quality reference genome for *A. paniculata*. With the help of both PacBio single molecule real time sequencing and Illumina sequencing reads for error correction, the *A. paniculata* genome was assembled into a total size of 284 Mb with a contig N50 size of 5.14 Mb. The contigs were further assembled into 24 pseudo-chromosomes by the Hi-C technique. We also analyzed the gene families (e.g., *KSL*, and *CYP450*) whose protein products are essential for synthesizing bioactive compounds in *A. paniculata.* In conclusion, the high-quality *A. paniculata* genome assembly builds the foundation for decoding the biosynthetic pathways of various medicinal compounds.

Keywords: PacBio sequencing, Hi-C, genome assembly, medicinal plant, *Andrographis paniculata*

INTRODUCTION

During the course of human history, all civilizations have tried to explore various plants for medicinal purposes and they formed unique empirical knowledge about how these herbal plants could be used to treat various diseases. Even though much of this knowledge has gradually given way to modern medicine, WHO found that the popularity of herbal medicines increased in almost all parts of the world in recent years (Fitzsimons, 2013). For this reason, herbal plants not only raise the enthusiasm from the general public, but also remain a rich source for discovering novel drug candidates among the researchers.

Abbreviations: BUSCO, Benchmarking Universal Single-Copy Orthologs; CPS, copalyl diphosphate synthase; CYP, cytochrome P450; DMAPP, dimethylallyl diphosphate; GGPPS, geranylgeranyl pyrophosphate synthase; GO, gene ontology; IPP, isopentenyl diphosphate; KSL, kaurene synthase-like protein; TE, transposable element.

With the emergence and development in high-throughput sequencing technology, the genome sequences of more than 200 plants have been reported (Lin et al., 2011). Particularly, genome sequencing is a powerful tool for studying various aspects of physiology and genetics in non-model plants, many of which are traditional herbal plants (Zerikly and Challis, 2009; De Luca et al., 2012; Chae et al., 2014). For example, the reference genome of *Scutellaria baicalensis* (Zerikly and Challis, 2009; De Luca et al., 2012; Chae et al., 2014; Zhao et al., 2019), *Panax ginseng* (Xu et al., 2017), mint (Vining et al., 2017), and opium poppy (Guo et al., 2018) provided insights into the genes involved in the biosynthesis of unique flavonoids, terpenes, alkaloids and many other secondary metabolites. Additionally, the reference genome of *Salvia splendens* was very valuable for helping marker-assisted breeding, genome editing, and molecular genetics (Dong et al., 2018). As one of the participants of the Herbal Plant Genomics Initiative, our team has reported the genomes of many Chinese herbal plants in past years, including *Salvia miltiorrhiza Bunge* (Zhang et al., 2015), *Dendrobium officinale* (Yan et al., 2015), maca (Zhang et al., 2016), *Panax notoginseng* (Chen et al., 2017), and fleabane (Yang et al., 2017). The high-quality genome assembly of *Andrographis paniculata* is presented in this manuscript as a continuum of the bigger research project.

Andrographis paniculata (**Figure 1A**) is a dicotyledonous medicinal plant widely distributed and used in tropical and subtropical regions of Asia, including India, China, Thailand, and Malaysia (Lim et al., 2012). This annual plant belongs to the family of Acanthaceae in the order of Lamiales. The dried plant has a highly acclaimed usage in the traditional Chinese medicine for its antipyretic, anti-inflammatory, and analgesic effects (Sun et al., 2019). Previous pharmacological research identified andrographolide and neoandrographolide as the main therapeutic constituents in *A. paniculata* (Srivastava and Akhila, 2010). Andrographolide is a labdane-related diterpenoid and it exhibits anti-cancer (Luo et al., 2014), anti-virus (Chen et al., 2009), antimicrobial and anti-inflammatory activities (Chua, 2014), suggesting potential pharmaceutical values. The leaves of *A. paniculata* contain major amounts of diterpene lactone compounds, including about 0.1% of deoxyandrographolide, and about 0.2% of neoandrographolide (Srivastava and Akhila, 2010). Even though the biosynthesis of andrographolide and neoandrographolide is achieved by the combination of various isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP) (Chen et al., 2011), the complete profile of *CYTOCHROME P450* genes (*CYPs*), *COPALYL DIPHOSPHATE SYNTHASE* genes (*CPSs*), and *KAURENE SYNTHASE-LIKE PROTEIN* genes (*KSLs*) has not been fully investigated in the *A. paniculata* genome.

The genome of *A. paniculata* is highly heterozygous and contains many repetitive sequences. These characteristics pose a big challenge in terms of acquiring a high-quality whole-genome reference assembly. A previous research effort reported an *A. paniculata* genome assembly of ∼269 Mb in total size with a contig N50 of 388 Kb (Sun et al., 2019). These benchmarks suggest the existence of relatively large un-assembled genome and gaps among contigs. Therefore, it is valuable to improve the genome assembly of *A. paniculata* using better raw data and assembly pipeline. Herein, we reported a reference genome of *A. paniculata* obtained from the PacBio single-molecule real time sequencing data in the size of 284 Mb. The contig N50 size was improved to 5.14 Mb, which is more than 12-fold longer than before. The resultant contigs were further assembled into 24 pseudo-chromosomes by the Hi-C technology, thereby yielding a high-quality *A. paniculata* genome assembly.

MATERIALS AND METHODS

DNA Extraction, Library Construction, and Sequencing

A single *A. paniculata* individual was obtained from Guangxi Medicinal Botanical Garden. Plant DNA was extracted from young leaves with the Novel Plant Genomic DNA Rapid Extraction Kit (Genenode Biotech, Beijing, China) according to the product manual.

A total of 15 μg *A. paniculata* genomic DNA was used to construct eleven PacBio libraries (mean size of 20 kb) with the SMRT Template Prep Kit (Pacific Biosciences, United States). These libraries were sequenced on a PacBio Sequel platform with recommended protocols from the manufacturer (**Supplementary Table 1**). Sequence reads with a quality score lower than 0.8 were removed.

In order to perform genome survey and genomic base correction, we also obtained Illumina reads for *A. paniculata*. In brief, 2 μg of genomic DNA was used to construct each library. The genomic DNA was sheared and insert sizes of 241, 313, 424, and 533 bp were selected for the four libraries. All libraries were sequenced on an Illumina HiSeq X Ten platform. The Illumina raw data were processed by fq_filter_V1.5 to remove the low-quality reads, and the parameters were set as follows: -q 33 -t 20 -ta 5 -tb 10 -tc 5 -td 10. We filtered out low-quality reads as specified by the following criteria: (1) filter a read if more than 5% of bases were N or poly-A, (2) filter a read if more than 30 bases were low quality, (3) if the read was contaminated with adaptor sequence, (4) if the size of a read was too small, and (5) if two copies of the paired-end reads had identical sequence (remove both copies). The resultant reads were then corrected by the SOAPec_v2.0.1 package with default settings.

Genome Size Estimation

The genome size of *A. paniculata* was assessed by both flow cytometry and *k*-mer analysis. For the flow cytometry approach, 20 mg of plant tissue was placed in 1.0 ml ice-cold nuclei isolation buffer in a Petri dish. The plant tissue was minced in the buffer with a new razor blade. The homogenate was filtered through a 42-mm nylon mesh into a labeled sample tube. Propidium iodide was added to a final concentration of 50 mg/ml simultaneously with RNase (50 mg/ml) and the sample was incubated on ice for 20 min before measurement. *Zea mays* (B73) was used as internal standard, the internal reference maize genome size is 2.3 Gb (Schnable et al., 2009).

For the *k-mer* analysis, Jellyfish v.2.2.5 (Marcais and Kingsford, 2011) was used to perform *k-mer* analysis on the Illumina sequencing error-corrected data. First,

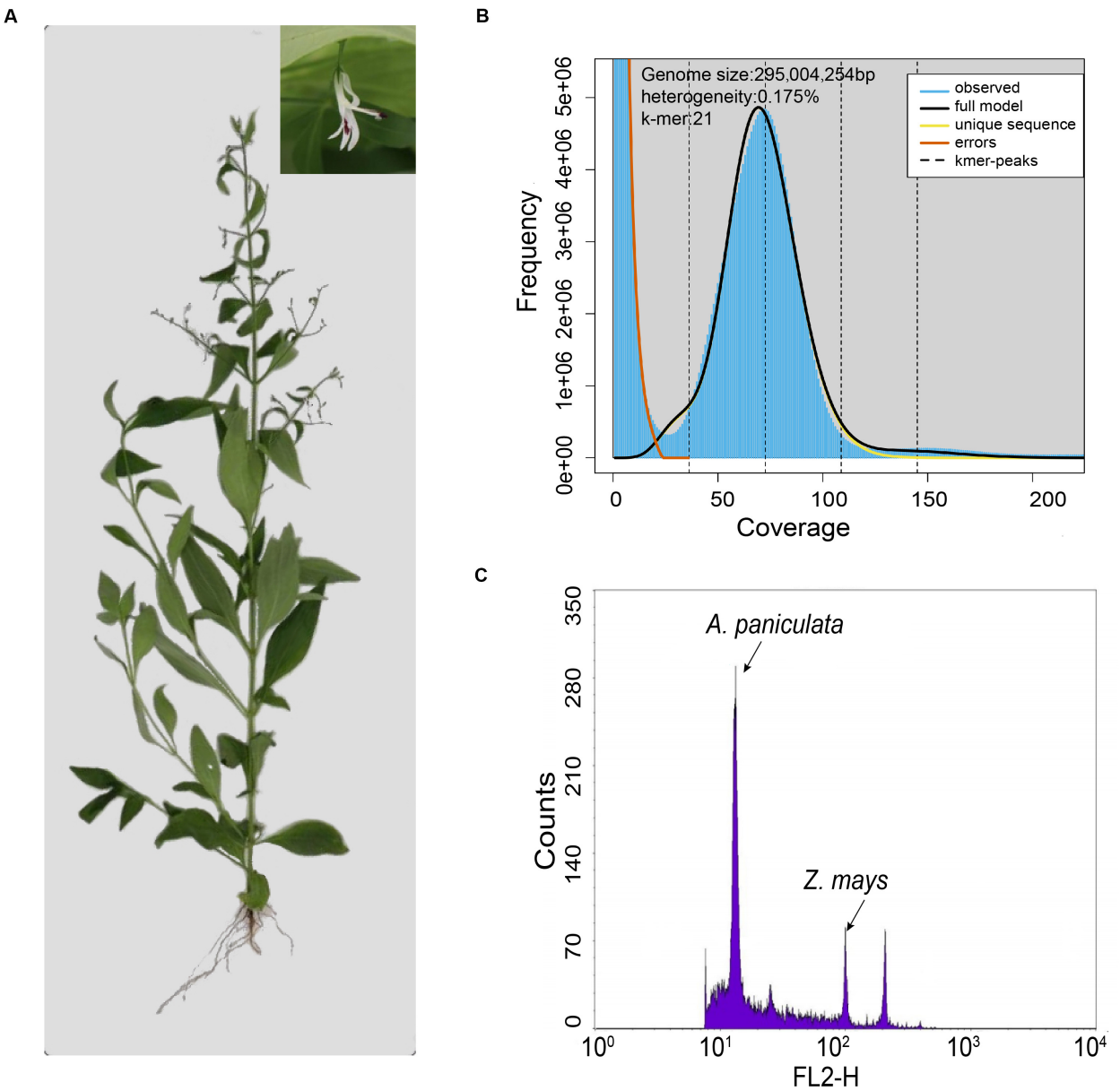


FIGURE 1 | Evaluation of the genome size of *A. paniculata*. **(A)** Photo of a single *A. paniculata* plant with close-up image of the flower. **(B)** The 21-*mer* analysis of the *A. paniculata* genome using GenomeScope. **(C)** Flow cytometry analysis of *A. paniculata* genome size comparing with *Zea mays*.

FastQC software was used to identify the quality of the input sequencing data for the Illumina sequencing data. The first 500,000,000 lines of the lane2 fastq file were extracted, and the extracted reads were used to identify heterozygosity with the '–m 21' option. The graph was generated by GenomeScope (Vurture et al., 2017).

Genome Assembly

A total of 53.98 Gb of PacBio data were used in the *de novo* assembly of the *A. paniculata* genome according to an assembly pipeline named HERA (Du and Liang, 2019). In brief, CANU v1.8 (Koren et al., 2017) was used to correct the raw PacBio data and assemble contigs. The resultant contigs were further improved with HERA. Next, the pipeline used the assembled genome to create an index file using bwa-mem (Li and Durbin, 2010). The processed next-generation sequencing data were aligned to the reference genome. Samtools (Li et al., 2009) was used to sort the resulting bam files. Finally, the bam files were used to polish the contig twice with Pilon (Walker et al., 2014). In this study, a commercial service provider from the Institute of Genetics and Developmental Biology in Beijing was recruited to assemble the *A. paniculata* genome using a professional version of HERA (Du and Liang, 2019), which is said to run faster and have better performance resolving repetitive sequences.

Hi-C Library Construction and Pseudomolecule Clustering

Three gram of *A. paniculata* leaves were harvested and crosslinked in a 2% formaldehyde solution for 15min at room temperature. Crosslinking was quenched by adding glycine to a final concentration of 250 mM. The fixed plant tissue was then ground in liquid nitrogen and suspended in extraction buffer for nuclei isolation. After the nuclei were separated, chromatin was solubilized in 0.1% (m/v) SDS at 65°C for 10 min. After SDS was quenched by Triton X-100 (final concentration of 1%), solubilized chromatin was digested by 400 units of *Dpn*II (New England Biolabs, MA, United States) at 37°C overnight. The following steps included biotin labeling of the DNA and blunt-end ligation of DNA fragments. After cross-linking was reversed by the treatment with proteinase K, DNA was purified so that biotin labels could be removed from non-ligated fragment ends. DNA fragments were sonicated into sizes of 400 bp so that paired-end libraries could be obtained. These libraries were sequenced on a NovaSeq 6000 platform (Illumina, United States) to acquire the Hi-C data.

Low-quality Hi-C reads were removed according to the following two criteria: (1) filter a read if more than 10% of bases were N, (2) filter a read if more than 50% bases were low quality ($Q \leq 5$). Clean Hi-C reads were mapped to the draft assembly with Juicer (Juicer, juicer_tools.1.7.6_jcuda.0.8.jar) (Durand et al., 2016a). A candidate chromosome-length assembly was generated automatically using the 3d-DNA pipeline to correct misjoins, order, orientation, and then anchor contigs from the draft assembly (Dudchenko et al., 2017). Manual review and refinement of the candidate assembly was performed in Juicebox Assembly Tools (Version 1.9.1) (Durand et al., 2016b) for quality control and interactive correction. And then the genome was finalized using the "run-asm-pipeline-post-review.sh -s finalize –sort-output –bulid-gapped-map" in 3d-DNA with manually adjusted assembly as input (Dudchenko et al., 2017).

Genome Annotation

The repetitive sequences were identified via sequence alignment and *de novo* prediction. RepeatMasker (Chen, 2004) was used to compare the assembled genome with the RepBase database (Release16.10)[1] using default settings (Bao et al., 2015). Repeatproteinmask searches[2] were used for prediction of homologs using default settings. For *de novo* annotation of repetitive elements, LTR_finder (Xu and Wang, 2007)[3], Piler (Edgar and Myers, 2005)[4], RepeatScout (Price et al., 2005)[5], and RepeatModeler[6] were used to construct the *de novo* library, and annotation was carried out with Repeatmasker (cutf 100, cpu 100, run qsub, -nolow,-no, -norna,-parallel 1). Tandem repeats were identified across the genome with Tandem Repeats Finder (cutf 100, cpu 100, period_size 2000, run qsub Match 2 Mismatch 7 Delta 7 PM 80 PI 10 Minscore 50 MaxPeriod 2000).

According to their characteristics and redundancy, the repeat consensus sequences were first classified using TEsort (Zhang et al., 2019) with REXdb database[7]. For *Copia* and *Gyspy* superfamilies, complete elements were identified based on the presence and order of conserved domains including capsid protein, aspartic proteinase, integrase, reverse transcriptase and RNase H as described in Wicker (Wicker and Keller, 2007). We extracted all reverse transcriptase and multiple sequence alignment of the extracted RT were then conducted by MAFFT (Nakamura et al., 2018) and the phylogenetic tree was constructed with IQTREE (Jana et al., 2016). Itol[8] was used for the visualization and edit of the tree. Finally, density of the TE consensus copies according to their lineages were computed along pseudomolecules and visualized using R.

GlimmerHMM (Aggarwal and Ramaswamy, 2002), SNAP (Korf, 2004), GenScan (Aggarwal and Ramaswamy, 2002), and Augustus (Stanke et al., 2006) were used for *ab initio* prediction of protein-coding genes with default settings. The homology-based prediction utilized reference protein sequence from *Arabidopsis thaliana* (Michael et al., 2018), *Sesamum indicum* (Wang et al., 2014), *Solanum lycopersicum* (Consortium, 2012), and *Vitis vinifera* (Girollet et al., 2019) according to an established protocol. RNA-seq data sets for *A. paniculata* leaf and root tissues were obtained from the National Center for Biotechnology Information (NCBI) database (SRX652837, SRX655521), and subsequently used for *de novo* assembly of the transcriptome. We aligned all RNA reads to the *A. paniculata* genome using TopHat (Trapnell et al., 2009), assembled the transcripts with Cufflinks (Trapnell et al., 2013) using default parameters, and predicted the open reading frames to obtain reliable transcripts with hidden Markov model (HMM)-based training parameters. Finally, the above 3 gene structure models were compiled by Evidence Modeler tool (Haas et al., 2008) with the following weights: transcripts-set > homology-set > *ab initio*-set and redundant genes were removed.

The t-RNAscan-SE tool (Chan and Lowe, 2019) was used to predict tRNA in the genome sequence with *E*-value set to 1e-5. Plant RNA sequences from Rfam database were selected as reference to predict the rRNA by BLASTN with E-value set to 1e-5. The miRNA and snRNA genes were also predicted by BLASTN against the Rfam database with *E*-value set to e-1.

Gene function was annotated by performing BLASTP (*E*-value $\leq$ 1e-5) against the protein databases. SwissProt[9], TrEMBL (see footnote 9), KEGG[10], and InterPro[11] were used for screening the functional domains of the proteins. Gene Ontology (GO) terms for each gene were extracted from the corresponding InterPro entries.

[1] http://www.girinst.org/repbase/index.html
[2] http://www.repeatmasker.org/
[3] https://github.com/xzhub/LTR_Finder
[4] http://www.drive5.com/piler/
[5] http://www.repeatmasker.org/
[6] http://www.repeatmasker.org/RepeatModeler.html
[7] http://repeatexplorer.org/
[8] https://itol.embl.de/
[9] http://www.uniprot.org/
[10] http://www.genome.jp/kegg/
[11] https://www.ebi.ac.uk/interpro/

Evolutionary and Phylogenetic Analyses

Protein sequences of *A. paniculata*, *Sesamum indicum*, *Salvia miltiorrhiza*, *Oryza sativa*, *Catharanthus roseus, Arabidopsis thaliana, Solanum lycopersicum*, and *Helianthus annuus* were downloaded from JGI. A full alignment protein search using BLASTP with the parameter *E*-value = 1e-5 was performed to verify the gene family clusters in these species and *A. paniculata*. Ortholog clustering and gene family clustering analyses were performed using OrthoMCL (Li et al., 2003). Venn diagram format was drawn using a web tool[12] (Zhang et al., 2016).

An all-against-all BLASTP comparison with a cutoff *E*-value of 1e-5 was preformed, and the results were clustered into groups of homologous proteins using Markov chain clustering with the default inflation parameter. All 1212 single-copy orthologous genes identified in the gene family cluster analysis from the aforementioned species were used to construct a phylogenetic tree. Multiple sequence alignments were performed for each gene using MUSCLE v.3.7[13] with default settings (Edgar, 2004).

The MCMCTREE program within the PAML package (Yang, 2007) was used to estimate divergence time of *A. paniculata*, *S. indicum*, *S. miltiorrhiza*, *O. sativa*, *C. s roseus, A. thaliana, S. lycopersicum*, and *H. annuus*. The HKY85 model (model = 4) and independent rates molecular clock (clock = 2) were used for calculation.

CAFE v1.7 (De Bie et al., 2006) is a tool for analyzing the evolution of gene family size based on the stochastic birth and death model. With the calculated phylogeny and the divergence time, this software was applied to identify gene families that had undergone expansion and/or contraction in the aforementioned species with the parameters: -filter -cpu 10 -lrt -simunum 1000.

Synteny Analysis of Two Genome Assemblies

The fasta and hic.gff files of the published genome assembly (Sun et al., 2019) were downloaded from NCBI. They were combined with the fasta and contig.gff files of our genome assembly by makeblastdb. BLASTP was used to align these sequence, and MCScanX (Wang et al., 2012) was used to perform synteny analysis between two genome assemblies.

Analysis of Key Gene Families in the *A. paniculata* Genome

We used hmmsearch to perform a preliminary screening of the gene family (*CYP450* and *terpene sythases, TPSs*) in *A. paniculata* and the gene ID was intercepted with an *E*-value $\leq$ 1e-59. The corresponding protein sequence was used as a query for TBLASTN (*E* = 1e-5) with both versions of the assembled *A. paniculata* genome sequence. The *CYP450* genes from other species were downloaded from the Cytochrome *CYP450* homepage[14] (Nelson, 2009) and the *TPS* genes from other species were acquired from a previous publication (Chen et al., 2011). Multiple sequence alignment was carried out with MUSCLE v3.7 (Edgar, 2004) using default parameters. The maximum likelihood (ML) phylogenetic tree was constructed using MEGA7 (Nelson, 2009) with 1,000 bootstraps.

The RNA-Seq data of the roots and leaves of *A. paniculata* were downloaded from the NCBI database (SRX652837, SRX655521). FPKM value was calculated for each protein-coding gene by Cufflinks (v. 2.1.1) (Trapnell et al., 2013). The heatmap was made with the pheatmap package.

[12]http://bioinfogp.cnb.csic.es/tools/venny/index.html

[13]http://www.drive5.com/muscle

[14]http://drnelson.uthsc.edu/CytochromeP450.html

RESULTS

Genomic Sequencing and High-Quality Genome Assembly

Genome survey with Illumina reads showed that the estimated genome size of *A. paniculata* with *21-mer* analysis was about 295 Mb (**Figure 1B**). This number was slightly smaller than the estimated genome size of 310 Mb from flow cytometry analysis (**Figure 1C**), but larger than the previous estimate of 280 Mb (Sun et al., 2019). The difference probably reflects the between-individual variation of the *A. paniculata* plant. Moreover, the heterozygosity of this sequenced genome was estimated to be 0.175% (**Figure 1B**).

In order to obtain a high-quality *A. paniculata* genome assembly, we constructed 11 PacBio SMRT sequencing libraries, which produced 53.96 Gb clean data (**Supplementary Table 1**). They covered about 183-fold of the estimated genome. We also generated about 238.6 Gb of Illumina sequencing data to polish and correct the error reads that are associated with PacBio sequencing. A combination of these data through a assembly pipeline (see the section on genome assembly methods for details) yielded a draft genome assembly of ∼284.3 Mb with 270 contigs (**Table 1**). This assembly represented about 91.6 – 96.3% of the estimated genome size. The longest contig was 9.30 Mb in size and the contig N50 was about 5.15 Mb. This benchmark is more than 12-fold longer than that of a previously reported assembly (Sun et al., 2019).

TABLE 1 | The library information and data statistics for the *A. paniculata* assembly.

Estimated genome size (Mb)	**295 - 310**
Assembly statistics	
Assembly size (Mb) Number of N50 contig	284.3 22
N50 contig length (bp)	5,149,272
Number of N90 contigs	51
N90 contig length (bp)	2,610,185
Transposable elements content	57.35%
Gene annotation statistics	
Total number of protein-coding genes	24,015
Total exon number	136,156
Average exon number per gene	5.67
Average exon size (bp)	227.14
Total intro length (bp)	453,323
Total number of non-protein-coding genes	6,591

We assessed the completeness of the genome assembly of *A. paniculata* by using the Benchmarking Universal Single-Copy Orthologs (BUSCO) approach (Simao et al., 2015). The result showed that 91.0% of the plant BUSCO genes could be recovered in the genome assembly and 3% of the plant BUSCO genes had partial matches (**Supplementary Table 2**). We also mapped cleaned Illumina paired reads back to the genome assembly using BWA mem (Li and Durbin, 2010). We found that 97% of the reads could be mapped to the genome assembly and 94% of the reads were found to be properly paired. The high mapping rate indicate high completeness of the assembly. These two benchmarks were also higher than those reported in the previously assembly (Sun et al., 2019). In addition, our contig assembly shared high collinearity with its counterpart from a previous report (**Figure 2**). These data collectively suggest that the genome assembly of *A. paniculata* in this study has high quality and could be used for subsequent analyses.

Finally, we obtained about 43.33 Gb (∼152 × coverage) clean Hi-C sequencing data, with which 246,855,874 bp (86.8% of all bases) of the genome assembly were organized into 24 pseudo-chromosomes (**Figure 3** and **Supplementary Table 3**). As expected, the Hi-C interaction decreases as the physical distance between two sequences increases.

Repeat Annotation

Transposable elements (TEs) accounted for about 163.1 Mb or 57.35% of the *A. paniculata* genome (**Supplementary Table 4**). Breakdown of the TE statistics showed that DNA retrotransposons and long terminal repeat (LTR) retrotransposons were major subtypes in the *A. paniculata*

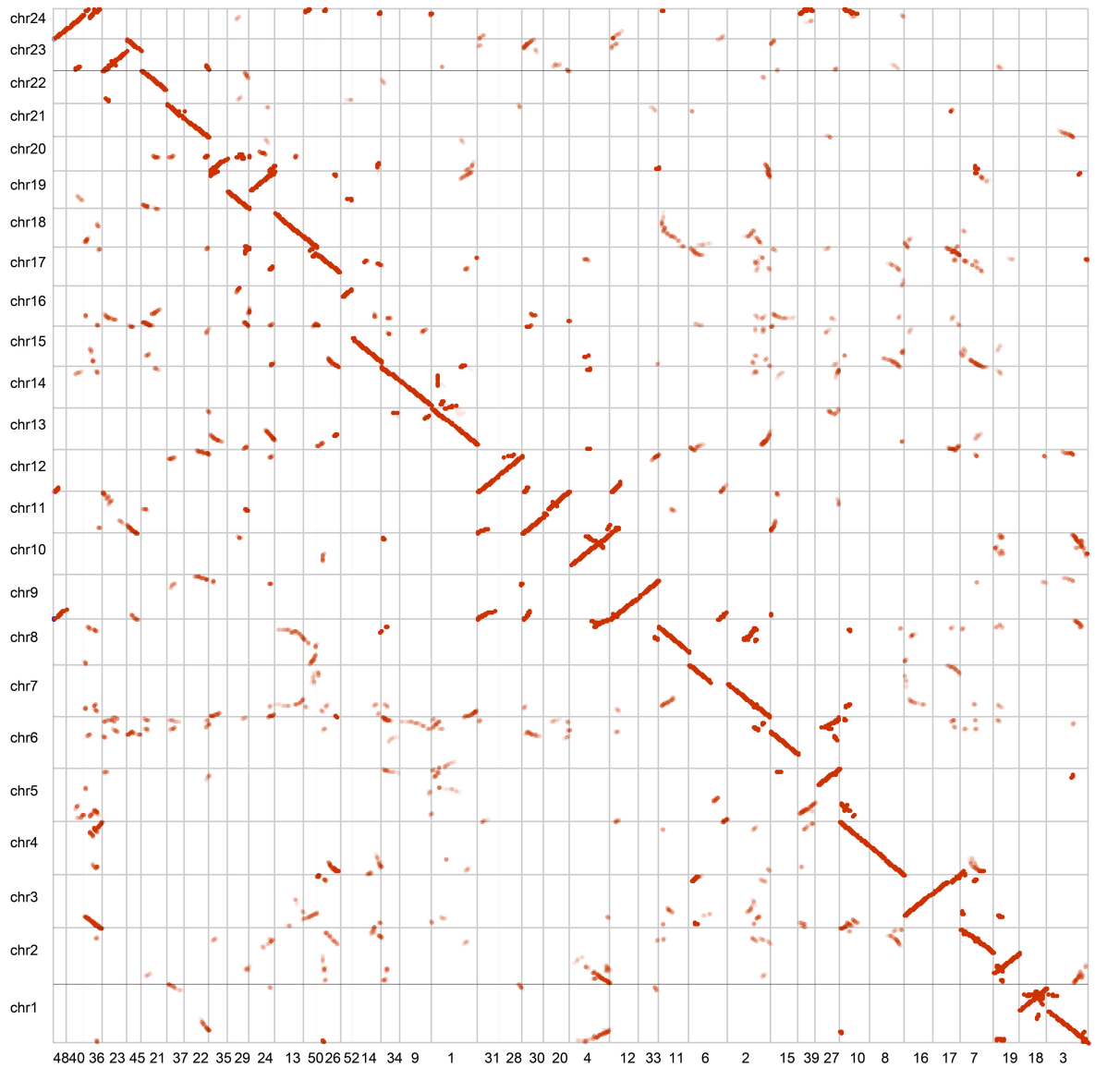


FIGURE 2 | Collinear analysis of the assemblies of *A. paniculata* genome.

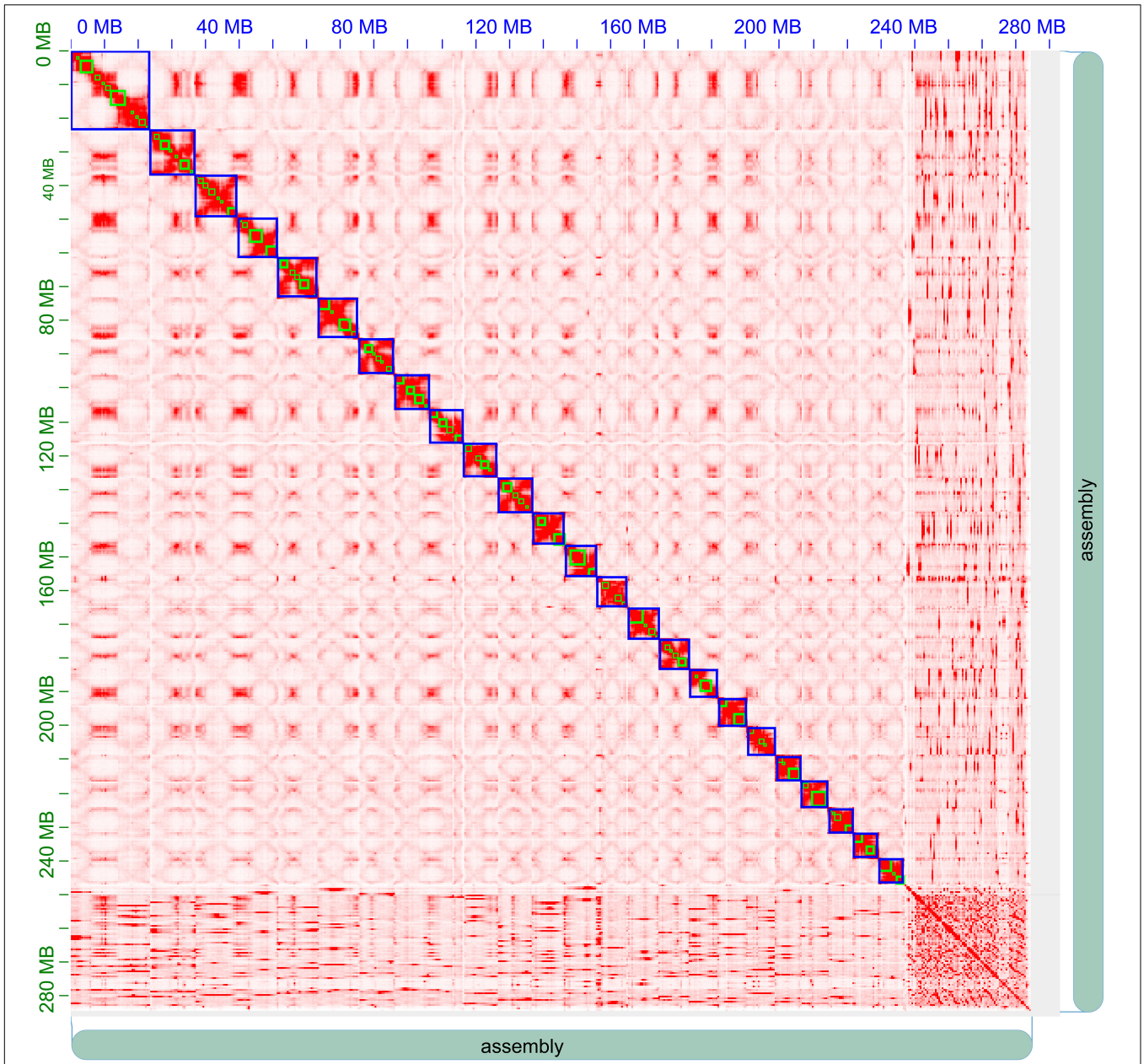


FIGURE 3 | Hi-C clustering heatmap for pseudomolecule construction in *A. paniculata*. In all, 24 pseudomolecules are assembled and indicated by blue boxes.

genome (**Supplementary Table 5**). TEs constitute an import part in plant genomes. We analyzed the evolutionary history of various *Ty3-gypsies* and *Ty1-copia* retroposon elements in the *A. paniculata* genome and identified unique dynamics of invasion patterns for different TE lineages (**Figures 4A,B**). For example, Ivana, Ogre, and Ale elements are all relatively young, suggesting an intense and recent burst of insertion or a strong selection against these TE elements. In contrast, Angela elements are the most ancient ones and the bimodal distribution suggests that the burst of insertion occurred twice in the evolutionary history of *A. paniculata* (**Figure 4B**). In addition, TE family distribution varied across the genome (**Figure 4C**). Ogre elements tend to cluster closely, thereby yielding prominent hotspot regions in the genome of *A. paniculata*.

Protein-Coding Gene Annotation

A combination of *ab initio* based, homology based, and RNA-Seq based methods were used to predict 24,015 protein-coding genes in the *A. paniculata* genome (**Supplementary Table 6**). The predicted mRNA was on average 3,175 bp in length, containing about 5.67 exons with an average CDS length of 1,287 bp. The number of predicted protein-coding genes was comparable to that of *S. indicum*, but much smaller than that of *O. sativa* and *H. annuus* (**Figure 5A**). Orthologous clustering analysis

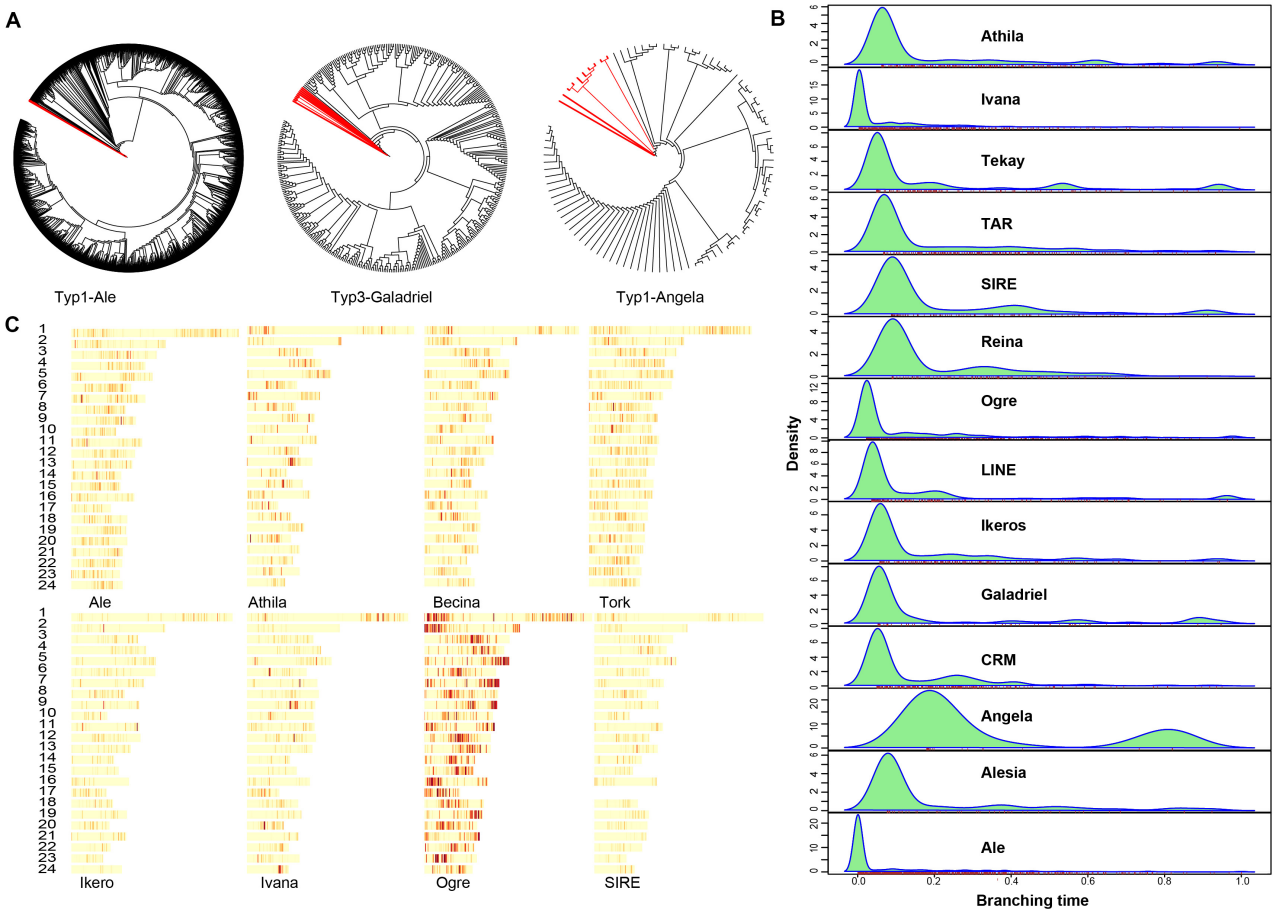


FIGURE 4 | Repeat analysis **(A)** Neighbor-joining (NJ) trees were built from RT domain sequence similarities among different lineage-specific copies identified in the *A. paniculata* genome. Deep branching revealed ancient expansion while flat branching is consistent with a recent burst of insertion activity. Red branches correspond to outgroup sequences. **(B)** The average age of TEs was revealed for the different lineages by the branching distribution in the NJ trees built from RT (light blue). **(C)** The density of different TE lineages inferred from the detection of their protein-coding domains along pseudomolecules.

showed that the *A. paniculata* genome contained 4,449 single-copy orthologs, 6,731 multiple-copy orthologs, 1,234 unique paralogs, 7,165 other orthologs, and 4,436 unclustered genes (**Supplementary Table 7**). In addition, Venn diagram showed that 7,798 gene families were shared by *A. paniculata*, *S. indicum, S. miltiorrhiza,* and *O. sativa*. A total of 525 gene families were unique to *A. paniculata*. This number was lower than that of the other three plant species (**Figure 5B**). Among the 24,015 protein-coding genes, a total of 19,824 predicted genes are supported by the RNA-seq expression data (FPKM > 0.05). Functional annotation of predicted protein-coding genes showed that 91.5% could obtain TrEMBL annotation; 62.2% could obtain GO annotation; 77.2% could obtain KEGG annotation; and 81.0% could obtain InterPro annotation (**Supplementary Table 8**).

Non-protein-Coding Gene Annotation

A combination of homolog and *ab initio* based methods identified a total of 6,591 non-protein-coding genes (**Supplementary Table 9**). These predicted genes comprised of 93 microRNA genes, 524 transfer RNA (tRNA) genes, 5,785 ribosomal RNA (rRNA) genes, and 189 small nuclear RNA genes (**Supplementary Table 9**).

Phylogenetic Analysis of *A. paniculata*

Phylogenetic analysis *A. paniculata* and seven other plant species showed that *A. paniculata* shared a common ancestor with *S. miltiorrhiza* approximately 57.0 Myr ago by calculated by r8s (**Figure 5C**). This estimate corresponds to the report from a previous study (Sun et al., 2019). During the course of evolution, a total of 1,064 and 1,396 gene families in *A. paniculata* were found to undergo expansion and contraction, respectively (**Figure 5D**).

Genomic Analysis of Key Genes in the Terpenoid Biosynthetic Pathway

Terpenoids are the largest group of plant secondary metabolites that are key targets for pharmaceutical screening and design (Srivastava and Akhila, 2010). Despite the structural diversity, these compounds share a common biosynthetic pathway (Yazaki et al., 2017). Terpenoids are derived from two five-carbon chemicals: IPP and dimethylallyl diphosphate

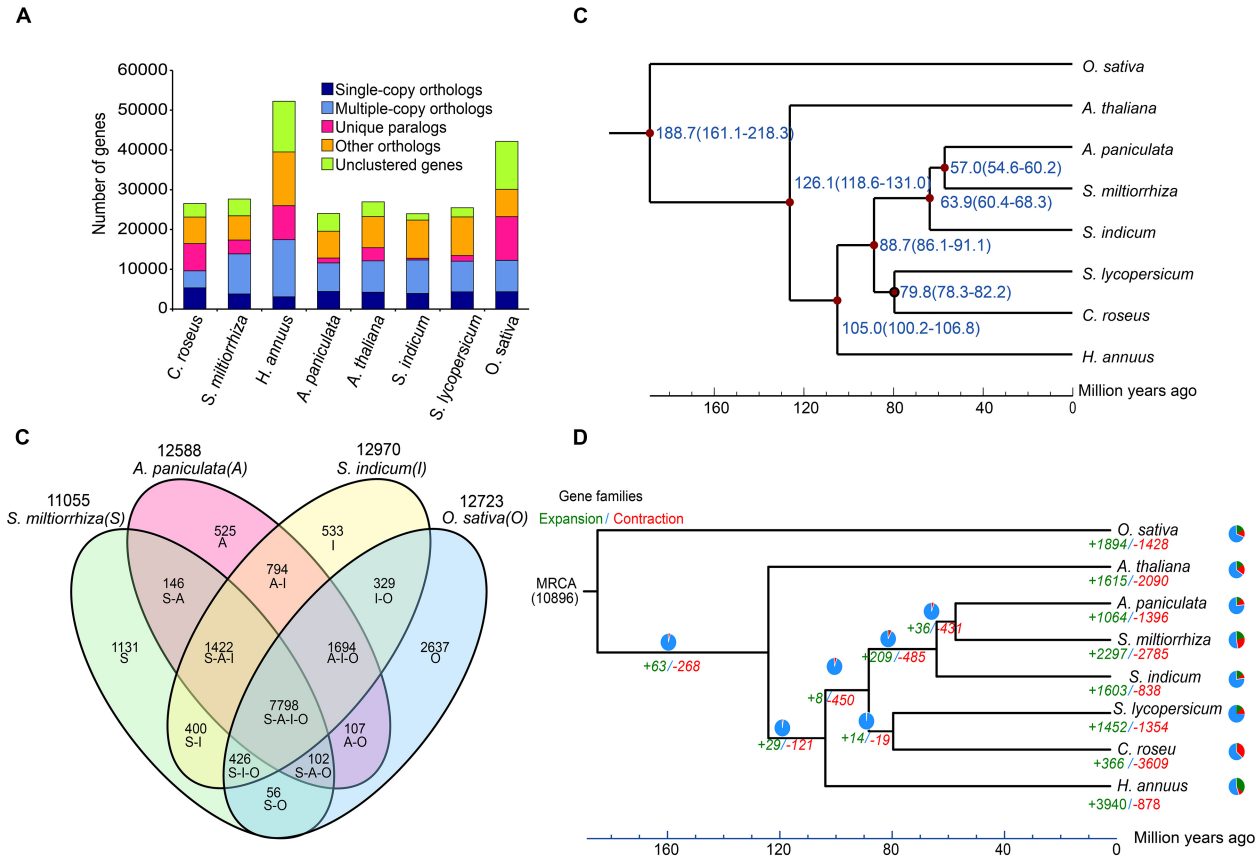


FIGURE 5 | Comparative analyses of the *A. paniculata* genome. **(A)** Major groups of orthologous genes in eight plant genomes. **(B)** Venn diagram of shared orthologous gene families among *A. paniculata*, *S. indicum*, *S. miltiorrhiza*, and *O. sativa* genomes. **(C)** Estimation of the time points of divergence (time range shown in parentheses) between *A. paniculata* and seven other plant species based on orthologous single-copy gene pairs. **(D)** Expansion and contraction of gene families in eight plant genomes. Pie diagram on each branch and node corresponds to combined change across lineages.

(DMAPP) (Lange et al., 2001). Their biosynthesis involve the classical acetate/mevalonate pathway in the cytosol and the pyruvate/glyceraldehyde-3-phosphate pathway in the plastids (**Figure 6A**) (Chen et al., 2011). Eventually, the condensation of IPP and DMAPP in various combinations will give rise to the countless terpenoids in plants (Srivastava and Akhila, 2010). In the *A. paniculata* genome, according to the KEGG annotation results, we identified 41 putative genes that were involved in the terpenoid backbone biosynthesis (**Supplementary Table 10**). This number is larger than that found in the *Panax notoginseng* genome (Chen et al., 2017). Additionally, almost all these putative genes exhibited certain levels of expression in the leaf and root tissue of *A. paniculata*. A previous report showed that the *GERANYLGERANYL PYROPHOSPHATE SYNTHASE* (*GGPPS*) group of genes in *A. paniculata* might be involved in the biosynthesis of andrographolide (Wang et al., 2019). In our genome assembly, a total number of 13 putative *GGPPS* genes were identified, and 11 out of 13 putative genes were expressed (**Supplementary Table 10**).

Besides *GGPPS* genes, previous studies suggest that *COPALYL DIPHOSPHATE SYNTHASE* (*CPS*) genes are implicated in the biosynthetic pathway of andrographolide and neoandrographolide (Garg et al., 2015; Shen et al., 2016a,b). The CPS enzymes are very similar in protein sequence to kaurene synthase (KS) and kaurene synthase-like (KSL) proteins (Zi et al., 2014). Despite their distinct catalytic activity, CPSs and KSLs belong to the c-type and e-type subfamilies of the terpene synthase (TPS) enzymes, respectively (Chen et al., 2011). In the *A. paniculata* genome, we identified a total of 53 putative *TPS* genes (**Supplementary Table 11**). Phylogenetic analysis of the TPS protein sequences from *A. paniculata* and eight other species showed that there were 24 putative *TPS-a1* genes, 10 putative *TPS-b* genes, 4 putative *TPS-c* genes, 8 putative *TPS-e* genes, 5 putative *TPS-f* genes, and 2 putative *TPS-g* genes (**Figure 6B**). This suggest that the *A. paniculata* genome may have four *CPS* genes and eight *KS*/*KSL* genes, all of which show certain levels of expression in the leaf and root tissues of the plant (**Figures 6C,D**).

Genomic Analysis of the Cytochrome P450 Gene Family in *A. paniculata*

Various biosynthetic pathways in plants rely on members of the *CYTOCHROME P450* (*CYP450*) gene family to accomplish

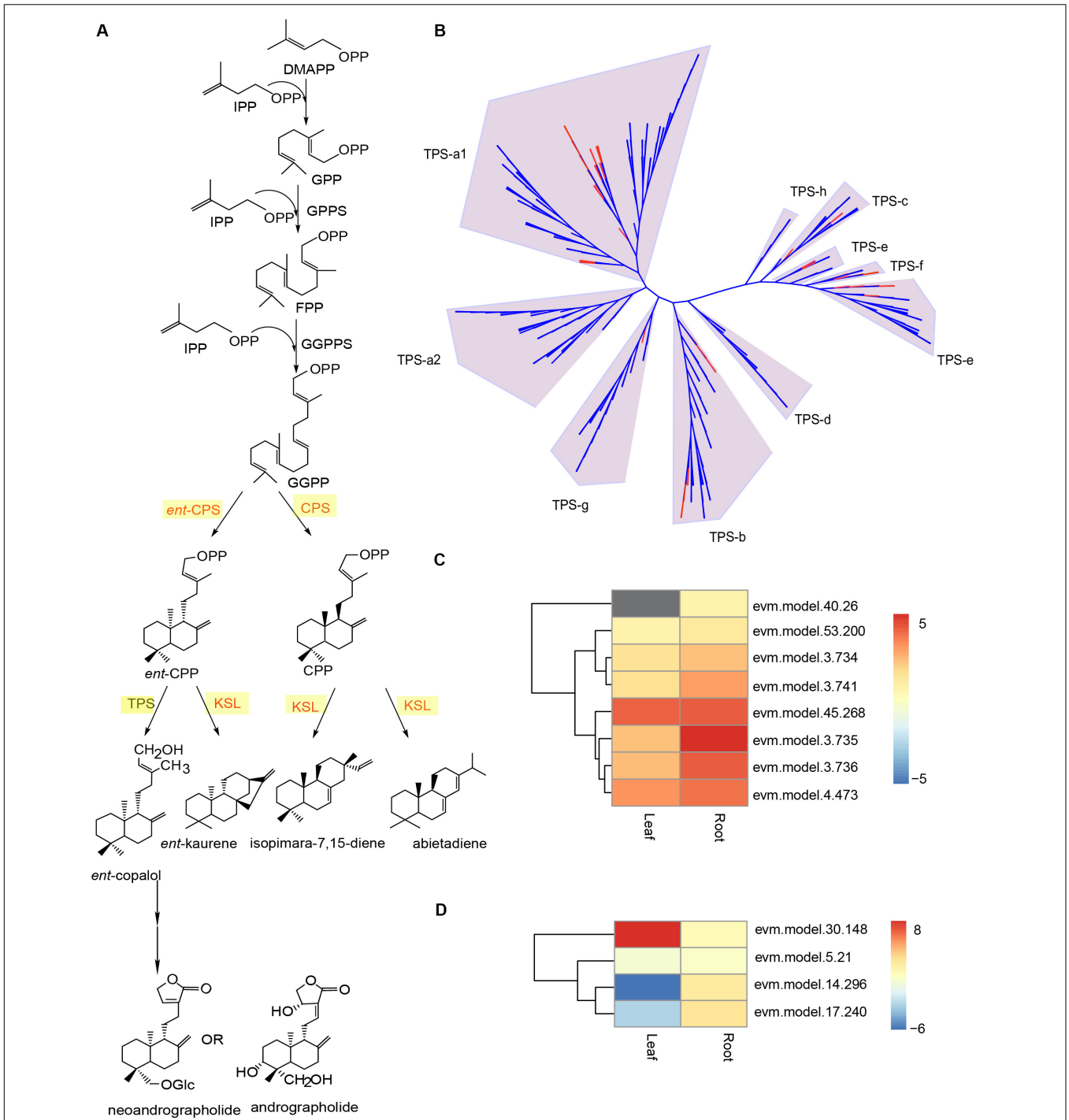


FIGURE 6 | Genomic analysis of putative *TPS* genes in the *A. paniculata* genome. **(A)** Schematic biosynthetic pathway of andrographolide and neoandrographolide. **(B)** Phylogenetic analysis of *TPS* genes in *A. paniculata* and other plants. Red branches indicate putative *TPS* genes in *A. paniculata*. **(C)** Expression levels of putative *KS/KSL* genes in the leaf and root tissue of *A. paniculata*. **(D)** Expression levels of putative *CPS* genes in the leaf and root tissue of *A. paniculata*.

chemical modification (Schuler, 1996; Mizutani and Ohta, 1998; Morant et al., 2003). For this reason, we investigated the putative *CYP450* genes in *A. paniculata*, and found a total of 205 candidates (**Supplementary Table 12**). The phylogenetic tree of the putative CYP450 protein sequences exhibited nine major clans (**Figure 7A**). The majority of candidate genes belonged to the clan 71 (104 genes), clan 72 (33 genes), clan 85 (33 genes), and clan 86 (25 genes), respectively. We also found that some *CYP450* genes appeared in a cluster in the contigs of the assembly (**Figure 7B**). This result agrees with the findings that it is fairly

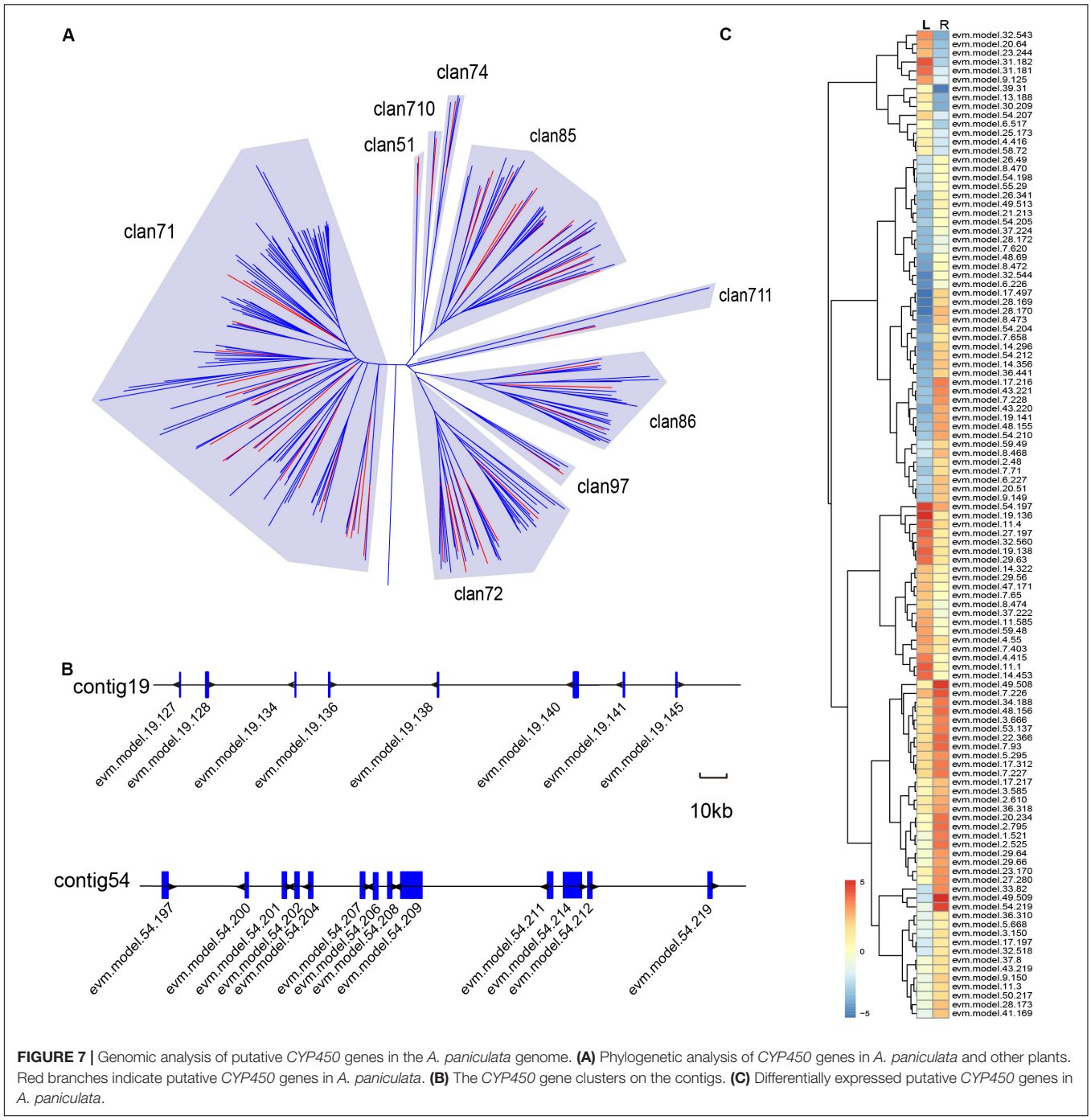


FIGURE 7 | Genomic analysis of putative *CYP450* genes in the *A. paniculata* genome. **(A)** Phylogenetic analysis of *CYP450* genes in *A. paniculata* and other plants. Red branches indicate putative *CYP450* genes in *A. paniculata*. **(B)** The *CYP450* gene clusters on the contigs. **(C)** Differentially expressed putative *CYP450* genes in *A. paniculata*.

common to have gene clusters for specific biosynthetic pathways in plant genomes (Nutzmann and Osbourn, 2014). Among the 205 candidate genes, we identified 111 putative *CYP450* genes that were differentially expressed in the root and leaf of *A. paniculata* (**Figure 7C**). Additionally, the expression level of evm.49.509 in the *CYP450* family was the highest in the root of *A. paniculata*, followed by evm.model.49.208, evm.model.7.226, and evm.model.54.219. In comparison, evm.model.19.136 was the highest expressed gene in the leaves of *A. paniculata*, followed by evm.model.54.197 and evm.model.11.4.

DISCUSSION

With the advancement and cost reduction of genome sequencing technologies, more and more plant species have revealed their near-complete genetic composition at a single-base resolution. Because most medicinal plant have highly repetitive and/or heterozygous genomes, a high-quality draft assembly is usually difficult and costly to secure. Particularly, lower contiguity of the genome assembly at the contig level will impede genomic analysis of functional genes, delineation

of biosynthetic pathways, and the development of novel pharmaceutical candidate.

We reported a chromosome-level reference genome for *A. paniculata* with improved benchmark values. Our updated *A. paniculata* genome was 284 Mb with a contig N50 size of 5.14 M. This number was more than 12-fold longer than that of the previous report (Sun et al., 2019). The BUSCO value reached 91.7%, which demonstrated the high completeness of the genome assembly.

Repetitive elements account for a large percentage of plant genomes. For example, the genome of lettuce contains about 74.2% of the sequence as repeats and the genome of sunflower includes more than 75% LTRs (Badouin et al., 2017; Reyes-Chin-Wo et al., 2017). We analyzed the repetitive element sequences in the *A. paniculata* genome and found that the repeats of the *A. paniculata* genome is as high as 57%, among which long terminal repeat retrotransposons (LTR-RTs) are predominant. Phylogenetic analyses of various *copia* and *gypsies* RT subclasses showed that the burst of invasion of many RT were recent events except Angela elements. This is in contrast to the pea genome where Angela elements drives the invasion fairly recently (Kreplak et al., 2019). Given these results, it is worth noting that TE annotation remains a challenging task for plant genomes. For instance, about 21.8% of *de novo* identified TEs in the *A. paniculata* genome could not be assigned to a particular category. Moreover, TEs tend to insert into the structures of existing TE elements, creating nested TEs in the genome. With our data, it would be interesting to use newly developed pipelines (e.g., Extensitve *de novo* TE Annotator) (Ou et al., 2019) to deconvolute nested TEs in the genome of *A. paniculata.*

In the present study, we analyzed the *Cyp450* gene family in *A. paniculata*, and identified 205 putative *CYP450* genes with conserved motifs. The results showed that all major classes of *Cyp450* reported by the Nutzmann and Osbourn (2014) could be found in the *A. paniculata* genome. The number of *Cyp450* genes with high expressions is larger in roots than in leaf. Furthermore, terpene synthesis (TPSs) is one of the main drivers of terpene diversification. The terpene synthase (TPS) family genes are generally categorized into seven clades (Shen et al., 2018). In the *A. paniculata* genome, we identified a total of 53 putative *TPS* genes, most of which belong to the *TPS-a* and *TPS-b* subfamilies. This is in line with the fact that *TPS-a* and *TPS-b* subfamilies represent the angiosperm-specific genes that have diverged from the other *TPS* genes (Shen et al., 2018).

CONCLUSION

The high-quality genome of *A. paniculata* will not only lay out the foundation for investigating the genetic basis for secondary metabolite biosynthesis, but also serve as an important resource for the study of other plant species in the *Andrographis* genus.

DATA AVAILABILITY STATEMENT

All raw sequence reads and the genome assembly have been deposited at NCBI under the BioProject accession number PRJNA549104.

AUTHOR CONTRIBUTIONS

YL, KW, and PQ collected the samples and performed the experiments. SC, ZY, YDu, and SD completed the data analysis. YDo, WC, and JM edited and modified the manuscript. All authors read and approved the manuscript.

FUNDING

This study was supported by the National Key R&D Program of China (2019YFC1711100), Yunnan Provincial Key Programs of Yunnan Eco-friendly Food International Cooperation Research Center Project (2019ZG00908), the Guangxi Innovation-Driven Development Project (GuiKe AA18242040), National Natural Science Foundation of China (81460582 and 81473309), China Agriculture Research System (CARS-21), and Guangxi Science and Technology Project (GuiKe AD17129044).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00701/full#supplementary-material

REFERENCES

Aggarwal, G., and Ramaswamy, R. (2002). *Ab initio* gene identification: prokaryote genome annotation with GeneScan and GLIMMER. *J. Biosci.* 27, 7–14. doi: 10.1007/BF02703679

Badouin, H., Gouzy, J., Grassa, C. J., Murat, F., Staton, S. E., Cottret, L., et al. (2017). The sunflower genome provides insights into oil metabolism, flowering and asterid evolution. *Nature* 546, 148–152. doi: 10.1038/nature22380

Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* 6:11. doi: 10.1186/s13100-015-0041-9

Chae, L., Kim, T., Nilo-Poyanco, R., and Rhee, S. Y. (2014). Genomic signatures of specialized metabolism in plants. *Science* 344, 510–513. doi: 10.1126/science.1252076

Chan, P. P., and Lowe, T. M. (2019). TRNAscan-SE: searching for tRNA genes in genomic sequences. *Methods Mol. Biol.* 1962, 1–14. doi: 10.1007/978-1-4939-9173-0_1

Chen, F., Tholl, D., Bohlmann, J., and Pichersky, E. (2011). The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant J.* 66, 212–229. doi: 10.1111/j.1365-313X.2011.04520.x

Chen, J. X., Xue, H. J., Ye, W. C., Fang, B. H., Liu, Y. H., Yuan, S.-H., et al. (2009). Activity of andrographolide and its derivatives against influenza virus in vivo and in vitro. *Biol. Pharm. Bull.* 32, 1385–1391. doi: 10.1248/bpb.32.1385

Chen, N. (2004). Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* 25, 4.10.1–4.10.14. doi: 10.1002/0471250953.bi0410s05

Chen, W., Kui, L., Zhang, G., Zhu, S., Zhang, J., Wang, X., et al. (2017). Whole-genome sequencing and analysis of the Chinese herbal plant *Panax notoginseng*. *Mol. Plant* 10, 899–902. doi: 10.1016/j.molp.2017.02.010

Chua, L. S. (2014). Review on liver inflammation and antiinflammatory activity of *Andrographis paniculata* for hepatoprotection. *Phytother. Res.* 28, 1589–1598. doi: 10.1002/ptr.5193

Consortium, T. G. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485, 635–641. doi: 10.1038/nature11119

De Bie, T., Cristianini, N., Demuth, J. P., and Hahn, M. W. (2006). CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22, 1269–1271. doi: 10.1093/bioinformatics/btl097

De Luca, V., Salim, V., Atsumi, S. M., and Yu, F. (2012). Mining the biodiversity of plants: a revolution in the making. *Science* 336, 1658–1661. doi: 10.1126/science.1217410

Dong, A. X., Xin, H. B., Li, Z. J., Liu, H., Sun, Y. Q., Nie, S., et al. (2018). High-quality assembly of the reference genome for scarlet sage, *Salvia splendens*, an economically important ornamental plant. *Gigascience* 7:giy068. doi: 10.1093/gigascience/giy068

Du, H., and Liang, C. (2019). Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads. *Nat. Commun.* 10:5360. doi: 10.1038/s41467-019-13355-3

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., et al. (2017). De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* 356, 92–95. doi: 10.1126/science.aal3327

Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., et al. (2016a). Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* 3, 99–101. doi: 10.1016/j.cels.2015.07.012

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., et al. (2016b). Juicer provides a one-click system for analyzing loop-resolution hi-C experiments. *Cell Syst.* 3, 95–98. doi: 10.1016/j.cels.2016.07.002

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Edgar, R. C., and Myers, E. W. (2005). PILER: identification and classification of genomic repeats. *Bioinformatics* 21(Suppl. 1), i152–i158. doi: 10.1093/bioinformatics/bti1003

Fitzsimons, D. W. (2013). World health organization. *Acta Med. Port.* 26, 186–187.

Garg, A., Agrawal, L., Misra, R. C., Sharma, S., and Ghosh, S. (2015). *Andrographis paniculata* transcriptome provides molecular insights into tissue-specific accumulation of medicinal diterpenes. *BMC Genomics* 16:659. doi: 10.1186/s12864-015-1864-y

Girollet, N., Rubio, B., and Bert, P. F. (2019). De novo phased assembly of the *Vitis riparia* grape genome. *Sci. Data* 6:127. doi: 10.1038/s41597-019-0133-3

Guo, L., Winzer, T., Yang, X., Li, Y., Ning, Z., He, Z., et al. (2018). The opium poppy genome and morphinan production. *Science* 362, 343–347. doi: 10.1126/science.aat4096

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* 9:R7. doi: 10.1186/gb-2008-9-1-r7

Jana, T., Lam-Tung, N., Arndt, V. H., and Quang, M. B. (2016). W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res.* 44, W232–W235. doi: 10.1093/nar/gkw256

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, 722–736. doi: 10.1101/gr.215087.116

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics* 5:59. doi: 10.1186/1471-2105-5-59

Kreplak, J., Madoui, M., Cápal, P., Novák, P., and Burstin, J. (2019). A reference genome for pea provides insight into legume genome evolution. *Nat. Genet.* 51, 1411–1422. doi: 10.1038/s41588-019-0480-1

Lange, B. M., Ketchum, R. E., and Croteau, R. B. (2001). Isoprenoid biosynthesis. Metabolite profiling of peppermint oil gland secretory cells and application to herbicide target analysis. *Plant Physiol.* 127, 305–314. doi: 10.1104/pp.127.1.305

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics* 26, 589–595. doi: 10.1093/bioinformatics/btp698

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

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189. doi: 10.1101/gr.1224503

Lim, J. C., Chan, T. K., Ng, D. S., Sagineedu, S. R., Stanslas, J., Fred Wong, W. S., et al. (2012). Andrographolide and its analogues: versatile bioactive molecules for combating inflammation and cancer. *Clin. Exp. Pharmacol. Physiol.* 39, 300–310. doi: 10.1111/j.1440-1681.2011.05633.x

Lin, Y., Li, J., Shen, H., Zhang, L., Papasian, C. J., Deng, H. W., et al. (2011). Comparative studies of de novo assembly tools for next-generation sequencing technologies. *Bioinformatics* 27, 2031–2037. doi: 10.1093/bioinformatics/btr319

Luo, X., Luo, W., Lin, C., Zhang, L., and Li, Y. (2014). Andrographolide inhibits proliferation of human lung cancer cells and the related mechanisms. *Int. J. Clin. Exp. Med.* 7, 4220–4225.

Marcais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* 27, 764–770. doi: 10.1093/bioinformatics/btr011

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

Mizutani, M., and Ohta, D. (1998). Two isoforms of NADPH:cytochrome P450 reductase in *Arabidopsis thaliana*. Gene structure, heterologous expression in insect cells, and differential regulation. *Plant Physiol.* 116, 357–367. doi: 10.1104/pp.116.1.357

Morant, M., Bak, S., Moller, B. L., and Werck-Reichhart, D. (2003). Plant cytochromes P450: tools for pharmacology, plant protection and phytoremediation. *Curr. Opin. Biotechnol.* 14, 151–162. doi: 10.1016/s0958-1669(03)00024-7

Nakamura, T., Yamada, K. D., Tomii, K., Katoh, K., and Hancock, J. (2018). Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* 34, 2490–2492. doi: 10.1093/bioinformatics/bty121

Nelson, D. R. (2009). The cytochrome p450 homepage. *Hum. Genomics* 4, 59–65.

Nutzmann, H. W., and Osbourn, A. (2014). Gene clustering in plant specialized metabolism. *Curr. Opin. Biotechnol.* 26, 91–99. doi: 10.1016/j.copbio.2013.10.009

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J., Hellinga, A. J., et al. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* 20:275. doi: 10.1186/s13059-019-1905-y

Price, A. L., Jones, N. C., and Pevzner, P. A. (2005). De novo identification of repeat families in large genomes. *Bioinformatics* 21(Suppl. 1), i351–i358. doi: 10.1093/bioinformatics/bti1018

Reyes-Chin-Wo, S., Wang, Z., Yang, X., Kozik, A., Arikit, S., Song, C., et al. (2017). Genome assembly with *in vitro* proximity ligation data and whole-genome triplication in lettuce. *Nat. Commun.* 8:14953. doi: 10.1038/ncomms14953

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

Schuler, M. A. (1996). The role of cytochrome P450 monooxygenases in plant-insect interactions. *Plant Physiol.* 112, 1411–1419. doi: 10.1104/pp.112.4.1411

Shen, Q., Li, L., Jiang, Y., and Wang, Q. (2016a). Functional characterization of ent-copalyl diphosphate synthase from *Andrographis paniculata* with putative involvement in andrographolides biosynthesis. *Biotechnol. Lett.* 38, 131–137. doi: 10.1007/s10529-015-1961-7

Shen, Q., Liu, Q., Congcong, L. I., Yuping, F. U., and Wang, Q. (2016b). Functional characterization of ApCPS involved in andrographolides biosynthesis by virus-induced gene silencing. *Acta Bot. Boreali* 36, 17–22.

Shen, Q., Zhang, L., Liao, Z., Wang, S., Yan, T., Shi, P., et al. (2018). The genome of *Artemisia annua* provides insight into the evolution of asteraceae family and artemisinin biosynthesis. *Mol. Plant* 11, 776–788. doi: 10.1016/j.molp.2018.03.015

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Srivastava, N., and Akhila, A. (2010). Biosynthesis of andrographolide in *Andrographis paniculata*. *Phytochemistry* 71, 1298–1304. doi: 10.1016/j.phytochem.2010.05.022

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., Morgenstern, B., et al. (2006). AUGUSTUS: *Ab initio* prediction of alternative transcripts. *Nucleic Acids Res.* 34, W435–W439. doi: 10.1093/nar/gkl200

Sun, W., Leng, L., Yin, Q., Xu, M., Huang, M., Xu, Z., et al. (2019). The genome of the medicinal plant *Andrographis paniculata* provides insight into the biosynthesis of the bioactive diterpenoid neoandrographolide. *Plant J.* 97, 841–857. doi: 10.1111/tpj.14162

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., Pachter, L., et al. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* 31, 46–53. doi: 10.1038/nbt.2450

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–1111. doi: 10.1093/bioinformatics/btp120

Vining, K. J., Johnson, S. R., Ahkami, A., Lange, I., Parrish, A. N., Trapp, S. C., et al. (2017). Draft genome sequence of *Mentha longifolia* and development of resources for mint cultivar improvement. *Mol. Plant* 10, 323–339. doi: 10.1016/j.molp.2016.10.018

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., et al. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* 33, 2202–2204. doi: 10.1093/bioinformatics/btx153

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9:e112963. doi: 10.1371/journal.pone.0112963

Wang, J., Lin, H. X., Su, P., Chen, T., Guo, J., Gao, W., et al. (2019). Molecular cloning and functional characterization of multiple geranylgeranyl pyrophosphate synthases (ApGGPPS) from *Andrographis paniculata*. *Plant Cell Rep.* 38, 117–128. doi: 10.1007/s00299-018-2353-y

Wang, L., Yu, S., Tong, C., Zhao, Y., Liu, Y., Song, C., et al. (2014). Genome sequencing of the high oil crop sesame provides insight into oil biosynthesis. *Genome Biol.* 15:R39. doi: 10.1186/gb-2014-15-2-r39

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). *MCScanX*: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40:e49. doi: 10.1093/nar/gkr1293

Wicker, T., and Keller, B. (2007). Genome-wide comparative analysis of Copia retrotransposons in Triticeae, rice, and *Arabidopsis* reveals conserved ancient evolutionary lineages and distinct dynamics of individual *copia* families. *Genome Res.* 17, 1072–1081. doi: 10.1101/gr.6214107

Xu, J., Chu, Y., Liao, B., Xiao, S., Yin, Q., Bai, R., et al. (2017). Panax ginseng genome examination for ginsenoside biosynthesis. *Gigascience* 6, 1–15. doi: 10.1093/gigascience/gix093

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 35, W265–W268. doi: 10.1093/nar/gkm286

Yan, L., Wang, X., Liu, H., Tian, Y., Lian, J., Yang, R., et al. (2015). The genome of *Dendrobium officinale* illuminates the biology of the important traditional Chinese orchid herb. *Mol. Plant* 8, 922–934. doi: 10.1016/j.molp.2014.12.011

Yang, J., Zhang, G., Zhang, J., Liu, H., Chen, W., Wang, X., et al. (2017). Hybrid de novo genome assembly of the Chinese herbal fleabane *Erigeron breviscapus*. *Gigascience* 6, 1–7. doi: 10.1093/gigascience/gix028

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Yazaki, K., Arimura, G. I., and Ohnishi, T. (2017). 'Hidden' terpenoids in plants: their biosynthesis, localization and ecological roles. *Plant Cell Physiol.* 58, 1615–1621. doi: 10.1093/pcp/pcx123

Zerikly, M., and Challis, G. L. (2009). Strategies for the discovery of new natural products by genome mining. *Chembiochem* 10, 625–633. doi: 10.1002/cbic.200800389

Zhang, G., Tian, Y., Zhang, J., Shu, L., Yang, S., Wang, W., et al. (2015). Hybrid de novo genome assembly of the Chinese herbal plant danshen (*Salvia miltiorrhiza* Bunge). *Gigascience* 4:62. doi: 10.1186/s13742-015-0104-3

Zhang, J., Tian, Y., Yan, L., Zhang, G., Wang, X., Zeng, Y., et al. (2016). Genome of plant maca (*Lepidium meyenii*) illuminates genomic basis for High-Altitude adaptation in the central andes. *Mol. Plant* 9, 1066–1077. doi: 10.1016/j.molp.2016.04.016

Zhang, R., Wang, Z., Ou, S., and Li, G. (2019). TEsorter: lineage-level classification of transposable elements using conserved protein domains. *bioRxiv [Preprint]* doi: 10.1101/800177

Zhao, Q., Yang, J., Cui, M. Y., Liu, J., Fang, Y., et al. (2019). The reference genome sequence of *Scutellaria baicalensis* provides insights into the evolution of wogonin biosynthesis. *Mol. Plant* 12, 935–950. doi: 10.1016/j.molp.2019.04.002

Zi, J., Matsuba, Y., Hong, Y. J., Jackson, A. J., Tantillo, D. J., Pichersky, E., et al. (2014). Biosynthesis of lycosantalonol, a cis-prenyl derived diterpenoid. *J. Am. Chem. Soc.* 136, 16951–16953. doi: 10.1021/ja508477e

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.

Copyright © 2020 Liang, Chen, Wei, Yang, Duan, Du, Qu, Miao, Chen and Dong. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 23 July 2020
doi: 10.3389/fgene.2020.00802

The Complete Chloroplast Genome Sequences of 14 *Curcuma* Species: Insights Into Genome Evolution and Phylogenetic Relationships Within Zingiberales

*Heng Liang[1], Yan Zhang[1], Jiabin Deng[2], Gang Gao[3], Chunbang Ding[1], Li Zhang[4] and Ruiwu Yang[1]**

[1] College of Life Science, Sichuan Agricultural University, Yaan, China, [2] School of Geography and Tourism, Guizhou Education University, Guiyang, China, [3] College of Life Sciences and Food Engineering, Yibin University, Yibin, China, [4] College of Science, Sichuan Agricultural University, Yaan, China

OPEN ACCESS

Edited by:
Hongjian Wan, Zhejiang Academy of Agricultural Sciences, China

Reviewed by:
Hui Yao, Chinese Academy of Medical Sciences, China
Inkyu Park, Korea Institute of Oriental Medicine (KIOM), South Korea

***Correspondence:**
Ruiwu Yang
yrwu@sicau.edu.cn

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

Received: *14 May 2020*
Accepted: *06 July 2020*
Published: *23 July 2020*

Citation:
Liang H, Zhang Y, Deng J, Gao G, Ding C, Zhang L and Yang R (2020) The Complete Chloroplast Genome Sequences of 14 Curcuma Species: Insights Into Genome Evolution and Phylogenetic Relationships Within Zingiberales. Front. Genet. 11:802. doi: 10.3389/fgene.2020.00802

Zingiberaceae is taxonomically complex family where species are perennial herb. However, lack of chloroplast genomic information severely hinders our understanding of Zingiberaceae species in the research of evolution and phylogenetic relationships. In this study, the complete chloroplast (cp) genomes of fourteen *Curcuma* species were assembled and characterized using next-generation sequencing. We compared the genome features, repeat sequences, sequence divergence, and constructed the phylogenetic relationships of the 25 Zingiberaceae species. In each Zingiberaceae species, the 25 complete chloroplast genomes ranging from 155,890 bp (*Zingiber spectabile*) to 164,101 bp (*Lanxangia tsaoko*) contained 111 genes consisting of 77 protein coding genes, 4 ribosomal RNAs and 30 transfer RNAs. These chloroplast genomes are similar to most angiosperm that consisted of a four-part circular DNA molecules. Moreover, the characteristics of the long repeats sequences and simple sequence repeats (SSRs) were found. Six divergent hotspots regions (*matK-trnk*, *Rps16-trnQ*, *petN-psbM*, *rpl32*, *ndhA,* and *ycf1*) were identified in the 25 Zingiberaceae chloroplast genomes, which could be potential molecular markers. In addition to *Wurfbainia longiligularis*, the ψ*ycf1* was discovered among the 25 Zingiberaceae species. The shared protein coding genes from 52 Zingiberales plants and four other family species as out groups were used to construct phylogenetic trees distinguished by maximum parsimony (MP), maximum likelihood (ML) and Bayesian inference (BI) and showed that Musaceae was the basal group in Zingiberales, and *Curcuma* had a close relationship with *Stahlianthu*. Besides this, *Curcuma flaviflora* was clustered together with *Zingiber*. Its distribution area (Southeast Asia) overlaps with the latter. Maybe hybridization occur in related groups within the same region. This may explain

why Zingiberaceae species have a complex phylogeny, and more samples and genetic data were necessary to confirm their relationship. This study provide the reliable information and high-quality chloroplast genomes and genome resources for future Zingiberaceae research.

Keywords: Zingiberaceae, *Curcuma*, chloroplast genome, comparative analysis, phylogenetic analysis

INTRODUCTION

Over the past decade, due to the linkage of genetic material existed in the plastome, the whole plastid genomes have been used to assess the phylogenetic relationship and taxonomic significance in some areas. The chloroplast as one of crucial and essential organelle plays an irreplaceable role in several vital biochemical processes and photosynthesis of plant (Neuhaus and Emes, 2000). Chloroplast DNA has the characteristics of haploid inheritance, relatively small genome size and slow mutation rate in plants (Palmer et al., 1988). These chloroplast DNA features were used by scientists through comparing with chloroplast DNA phylogenies, proving to be very useful in understanding of plant phylogenetic studies and clearer taxonomic levels (Alwadani et al., 2019; Gonçalves et al., 2019). Based on previous research results, whole chloroplast genome was acceptably considered circular and its genes evolving was a single unit (Gitzendanner et al., 2018; Gu et al., 2019). Many recent studies have shown that the entire chloroplasts or protein-coding genes can provide stronger and greater phylogeography and phylogeny evidence (Shaw et al., 2014). The chloroplast DNA genome in complex genome plants has been proved comparatively inexpensive and easily sequenced (Guo et al., 2017; Saarela et al., 2018).

Although the chloroplast DNA genome data as a powerful and simple strategy has been used to analyze the evolutionary relationships of plant, their findings are discrepant and incongruous with the evidence of DNA barcoding makers and morphological character (Huang et al., 2014). Intergeneric or interspecific hybridization, gene flow, heterogeneity of evolutionary rates and sampling error, etc were the main sources of the discordance of different datasets (Rieseberg and Soltis, 1991). While these factors are not always a single one to cause discordance in the progress of phylogenies. Meanwhile, these factors also have different manifestations at different stages in the evolutionary processes operating therein (Joly et al., 2009; Kobayashi et al., 2017). The discrepant and incongruous evidence of DNA barcoding makers and morphological character have been widely used in explaining the problems of phylogeny (McKinnon et al., 1999; Meng and Kubatko, 2009). Thus, it is also a challenge to make sure the mechanism of inconformity. Different from the nuclear genes of amphilepsis, the chloroplast DNA is matrilineal inheritance and lacks recombination in most angiosperms. Therefore, chloroplast DNA datasets were used to reflect and analyze seed-mediated gene flow and reduce the influence of genes transmitted by pollen to confirm the female parent (Ruhlman et al., 2017). Compared to nuclear gene flow, chloroplast gene flow usually is easily identified (Du et al., 2009). Thus, the chloroplast gene is a useful evidence of interspecific gene flow.

Zingiberaceae, belonging to the order Zingiberales, is a controversial and complex (the classification and taxonomic status) flowering family of medicinal, edible and horticultural importance which widely occur in the regions of subtropical and tropical Asia (Soltis et al., 2000; Wood et al., 2000; Kress et al., 2002; Kanlayanapaphon and Newman, 2003). Due to lack of differences in morphological characters and a reliable reference genome, the phylogenetic relationships and compellent taxonomy of Zingiberaceae plants are still a huge challenge and intractable (Ye et al., 2008; Chen and Xia, 2010). Taking the genus *Curcuma* as an example, there are at least 120 species in the world, and about 10 *Curcuma* species are distributed in China (Li et al., 2001; Leong-Škorničková et al., 2007). *Curcuma* is a young genus and the hybridization and introgression are ubiquitous in its closely related species (Mallet, 2005). On account of the flower parts, shape and color of bracts, rhizome color and position of inflorescences in *Curcuma* plant are neither unique nor universal to all species; the traditional diagnostic characters identification of *Curcuma* is poor and difficult (Kress et al., 2002). In recent years, in order to deepen understanding of the taxonomy and phylogeny of *Curcuma*, the method based on modern branch was used by taxonomist and molecular phylogenetics (Deng et al., 2018).

However, use of only one or several certain DNA fragments usually leads to any additional evidence, incomplete conclusions and more confusion in Zingiberaceae species. Therefore, complete chloroplast genomes will be a better sequencing solutions to identifying species and reconstructing phylogeny relationship (Rogalski et al., 2015). Unfortunately, there are few publicly available chloroplast genomes data of Zingiberaceae species at present.

Similar studies were also performed on family or genus with taxonomic difficulties such as *Stahlianthus* (Li et al., 2019), *Alpinia* (Li et al., 2020), *Curcuma* (Gui et al., 2020), numerous previous endeavors have provided chloroplast genomes of 8 Zingiberaceae species and illuminated further insights into the phylogeny and taxonomy of the Zingiberaceae species (*Stahlianthus involucratus* clustered with *Curcuma longa* and *Curcuma roscoeana*, and *Alpinia* was a sister branch to *Amomum*), but still have not reached a satisfied resolution. Due to the incomplete cognition on Zingiberaceae classification, only few *Curcuma* species have often been described by different taxonomists. Here, we reported 14 new complete chloroplast genomes of *Curcuma* including seven firstly sequenced chloroplast genomes (*Curcuma zanthorrhiza, Curcuma elata, Curcuma yunnanensis, Curcuma alismatifolia, Curcuma amarissima, Curcuma sichuanensis,* and *Curcuma rosesana*) and combined with previously published Zingiberaceae complete chloroplast genomes data to visualize and test the

genome organization, evolutionary rates, and phylogenetic relationships. All samples, assembled using a *de novo* approach, showed high similarities to previously published chloroplast genomes from *Curcuma*, including oligonucleotide repeats, microsatellites, RNA editing sites, amino acids frequencies, codon usage, synonymous and non-synonymous substitutions, GC content, gene synteny and genome sizes. Our goals were: (1) To observe the differences of the chloroplast genome in 25 Zingiberaceae species; (2) to find the highly polymorphic variable regions in chloroplast genomes and reliable markers for Zingiberaceae; (3) to infer the phylogenetic relationships within Zingiberaceae and discover the status of Zingiberaceae in Zingiberales based on the chloroplast genome.

RESULTS

The Length and Features of Chloroplast Genome

The complete chloroplast genomes in the 25 Zingiberaceae species ranged in size from 155,890 bp (*Zingiber spectabile*) to 164,101 bp (*Lanxangia tsaoko*) (**Table 1** and **Supplementary Table S1**). Similar to other studies of chloroplast genome, All the complete chloroplast genomes had a circular molecule conjoined and four conjoined structures (Xiang et al., 2016; Zhou et al., 2017). The quadripartite structure included the large single copy (LSC) region and small single copy (SSC) region separated by two inverted repeat (IR) regions (**Figure 1**). The LSC of the chloroplast genomes accounted for 53.58–58.27% of the total size and it ranged in size from 86,921 bp (*Curcuma wenyujin*) to 90,832 bp (*Zingiber spectabile*); the SSC of the chloroplast genomes accounted for 9.34–11.87% of the total size and it ranged in size from 152,88 bp (*Wurfbainia longiligularis*) to 185,08 bp (*Zingiber spectabile*). The IR of the chloroplast genomes accounted for 14.93–18.44% of the total size and it ranged in size from 30,119 bp (*Stahlianthus involucratus*) to 23,277 bp (*Zingiber spectabile*). In the genus *Curcuma*, the LSC of the chloroplast genomes accounted for 53.58–58.27% of the total size, ranging in size from 86,921bp (*Curcuma wenyujin*) to 87,856 bp (*Curcuma flaviflora*); the SSC of the chloroplast genomes accounted for 9.53–9.68% of the total size, ranging in size from 15,372 bp (*Curcuma alismatifolia*) to 15,744 bp (*Curcuma flaviflora*). The IR of the chloroplast genomes accounted for 18.26–18.37% of the total size and ranging in size from 29,718 bp (*Curcuma wenyujin*) to 29,897 bp (*Curcuma alismatifolia*). The GC contents of the chloroplast genomes in the family Zingiberaceae were similar.

Among the 111 functional genes, there are 77 protein-coding genes, 30 tRNA genes and 4 rRNA genes in the Zingiberaceae plastomes encoded (**Table 2**), quite similar to previously published Zingiberaceae species. Among these chloroplast genes of 25 Zingiberaceae species, 54 gene fragments were connected with self-replication, 4 genes were connected with DNA dependent RNA polymerase, 8 genes were connected with large subunit of ribosome, 12 genes were connected with small subunit of ribosome, 44 genes were connected with photosynthesis, 5 genes were connected with encoded the subunits of photo system I, 6 genes were connected with ATP synthesis, 14 genes were connected with encoded the subunits of photo system II, 11 genes were connected with subunits of NADH-dehydrogenase, 6 genes were connected with subunits of cytochrome, and only one gene was connected with subunits of rubisco. The *infA* gene was related to the translational initiation factor. The protein coding exons accounted for 47.29–52.18% of the total length and regions ranged from 77,372 to 84,187 bp, the rRNA accounted for 5.52–5.81% of the total length and regions ranged from 9,054 to 9,102 bp, the tRNA accounted for 1.41–1.87% of the total length and regions ranged from 2,311 to 3,029 bp, the intergenic regions accounted for 29.58–38.11% of the total length and ranged from 47,943 to 62,355 bp, the intronic regions accounted for 7.59–13.87% of the total length and ranged from 12,416 to 22,757 bp (**Supplementary Table S2**). In the genus *Curcuma*, the protein coding exons accounted for 50.83–51.19% of the total length and regions ranged from 82,930 to 83,042 bp, the rRNA accounted for 5.55–5.59% of the total length and regions ranged from 9,054 to 9,058 bp, the tRNA accounted for 1.79–1.80% of the total length and regions were 2,920 bp, the intergenic regions accounted for 29.58%–30.00% of the total length and ranged from 47,943 to 48,949 bp, the intronic regions ranged accounted for 11.82–11.90% of the total length and ranged from 19,164 to 19,364 bp (**Table 3**). Introns were more easy to lose its function than excons during evolution (Mattick, 1994; Kelchner, 2002; Chorev and Carmel, 2012). Among the 111 distinct genes, 18 genes were connected to contained introns. The genes of *ycf3* and *clpP* contained two introns; a total of 14 genes contained one intron, including 8 coding genes and 6 tRNA genes. In most Zingiberaceae species, a total of 18 genes were related to introns. Among the 18 genes, there were13 genes in the LSC regions, 4 genes in the IR regions and only one gene in the SSC regions (**Supplementary Table S3**).

Codon Usage

The codon usage frequency of 64 coding genes for 25 Zingiberaceae species were estimated. All protein-coding genes were encoded by 37,160 (*Amomum krervanh*) to 52,160 (*Wurfbainia villosa*). UGA, UAG, and UAA were considered as the termination codons. For these Zingiberaceae species (**Table 4** and **Supplementary Table S4**), we found that the UUA encoded leucine had the highest RSCU (Relative Synonymous Codon Usage) value at approximately 1.92, and the GCG encoded alanine had the lowest RSCU value at approximately 0.33. The AAA encoded lysine was the most frequent amino acid and had more than 2,200 frequencies in most of the 25 Zingiberaceae species. The results of RSCU in the 25 Zingiberaceae species showed G or C was biased toward a lower nucleotide frequency than A or T at the third codon position and the result was similar to other angiosperms chloroplast genomes researches.

Comparative Chloroplast Genomic Analysis Within 25 Zingiberaceae Species

To characterize genome divergence, we used the program mVISTA to perform sequence alignments of 25 Zingiberaceae species and took the annotation of *Curcuma sichuanensis* as

TABLE 1 | Complete chloroplast genomes for 14 Curcuma species.

	C. rosesana	*C. amarissima*	*C. sichuanensis*	*C. elata*	*C. zanthorrhiza*	*C. longa*	*C. yunnanensis*
Total length (bp)	162157	162158	162133	162171	162192	162220	162133
GC (%)	36.21	36.2	36.22	36.22	36.2	36.19	36.22
LSC							
Length (bp)	87043	86982	87013	87037	86994	87041	87013
GC (%)	33.98	34	34	34.01	33.99	33.99	34
Length (%)	53.68	53.64	53.67	53.67	53.64	53.66	53.67
SSC							
Length (bp)	15616	15678	15622	15634	15700	15681	15622
GC (%)	29.75	29.65	29.77	29.76	29.66	29.65	29.77
Length (%)	9.63	9.67	9.64	9.64	9.68	9.67	9.64
IR							
Length (bp)	29751	29751	29751	29752	29751	29751	29751
GC (%)	41.14	41.15	41.15	41.15	41.15	41.14	41.15
Length (%)	18.35	18.35	18.35	18.35	18.34	18.34	18.35
	C. sp.1	***C. sp.2***	***C. wenyujin***	***C. phaeocaulis***	***C. aromatica***	***C. alismatifolia***	***C. flaviflora***
Total length (bp)	162155	162133	162024	162031	162243	162715	163141
GC (%)	36.21	36.22	36.22	36.22	36.21	36.15	36.09
LSC							
Length (bp)	87041	87013	86921	86923	87109	87416	87856
GC (%)	33.99	34	34.03	34.03	33.99	33.91	33.85
Length (%)	53.68	53.67	53.65	53.65	53.69	53.72	53.85
SSC							
Length (bp)	15616	15622	15671	15676	15634	15509	15744
GC (%)	29.75	29.77	29.72	29.71	29.76	29.82	29.52
Length (%)	9.63	9.64	9.67	9.67	9.64	9.53	9.65
IR							
Length (bp)	29751	29751	29718	29718	29752	29897	29784
GC (%)	41.14	41.15	41.14	41.14	41.15	41.05	41.12
Length (%)	18.35	18.35	18.34	18.34	18.34	18.37	18.26

GC, guanine-cytosine; LSC, large single-copy region; SSC, short single-copy region; IRs, inverted repeats.

a reference. The comparison indicated that the 25 chloroplast genomes were highly similar (**Figure 2**). The characterize genome divergence was low among the 14 *Curcuma* species. The results indicated the divergence in IR regions was lower than in LSC and SSC regions. Moreover, the variation of divergence in coding regions was less than that in non-coding. Among the coding genes, the highly divergent regions included *matK*, *rpoC2*, *rpoB*, *ndhF*, *ycf1,* and *ycf2*. In addition, some of the highly divergent regions, such as *psbK-psbI*, *atpH-atplI*, *rpoB-petN*, *trnS-psbM*, *trnT-psbD*, *ndhC-trnR*, *petA-psbJ*, and *ndhF-rpl32,* were among the intergenic regions.

The number and sequence of genes in the IR regions were more conserved than that in the LSC and SSC regions within the 25 chloroplast genomes. The IR amplification and contraction were the causes of the differences of chloroplast genome size among the 25 Zingiberaceae species. Among the analysis of the genes across 25 chloroplast genomes, it is obvious that the evolutionary rates had some differences. In summary, the average number of Ka/Ks was less than 0.5 for 92.54% genes. The 20 genes showed Ka/Ks > 1 in at least one species. There were two genes (*ndhc* and *rps8*) showing a high rates for more than half species, and the genes of *ndhc* and *rps8* may be under positive selection. Most of genes associated with photosynthesis showed the lowest rates of evolution, and seven of them (*atpH*, *petL*, *petN*, *psaC*, *psaJ*, *psbM,* and *psbT*) showed Ka/Ks were 0. The Ka/Ks of *atpH*, *petL*, *petN*, *psaC*, *psaJ*, *psbM*, *psbT*, *rps2,* and *rps7* were low, and the main reason was no non-synonymous substitutions (**Supplementary Table S5**). We also inferred three genes with the most variable Ka/Ks, *atpF* (Subunits of ATP synthase), *rps8* (Small subunit of ribosome) and *ndhC* (Subunits of NADH dehydrogenase). For the *atpF* gene, the values of Ka/Ks were 0.2397 to 1.3544 (aside from a few Ka were 0), the *rps8* gene ranged from 0.1290 to 1.0608 and the *ndhC* gene ranged from 0.1593 to 1.3253. The result indicated that the genes of *ndhC* and *rps8* evolved slower than *atpF* gene.

The Differences of Genome Size Among the 25 Zingiberaceae Species

Among the 25 Zingiberaceae species, the shortest genome was *Zingiber spectabile* (155,890 bp), and the longest was *Lanxangia tsaoko* (164,101). In the 14 *Curcuma* species, the genome

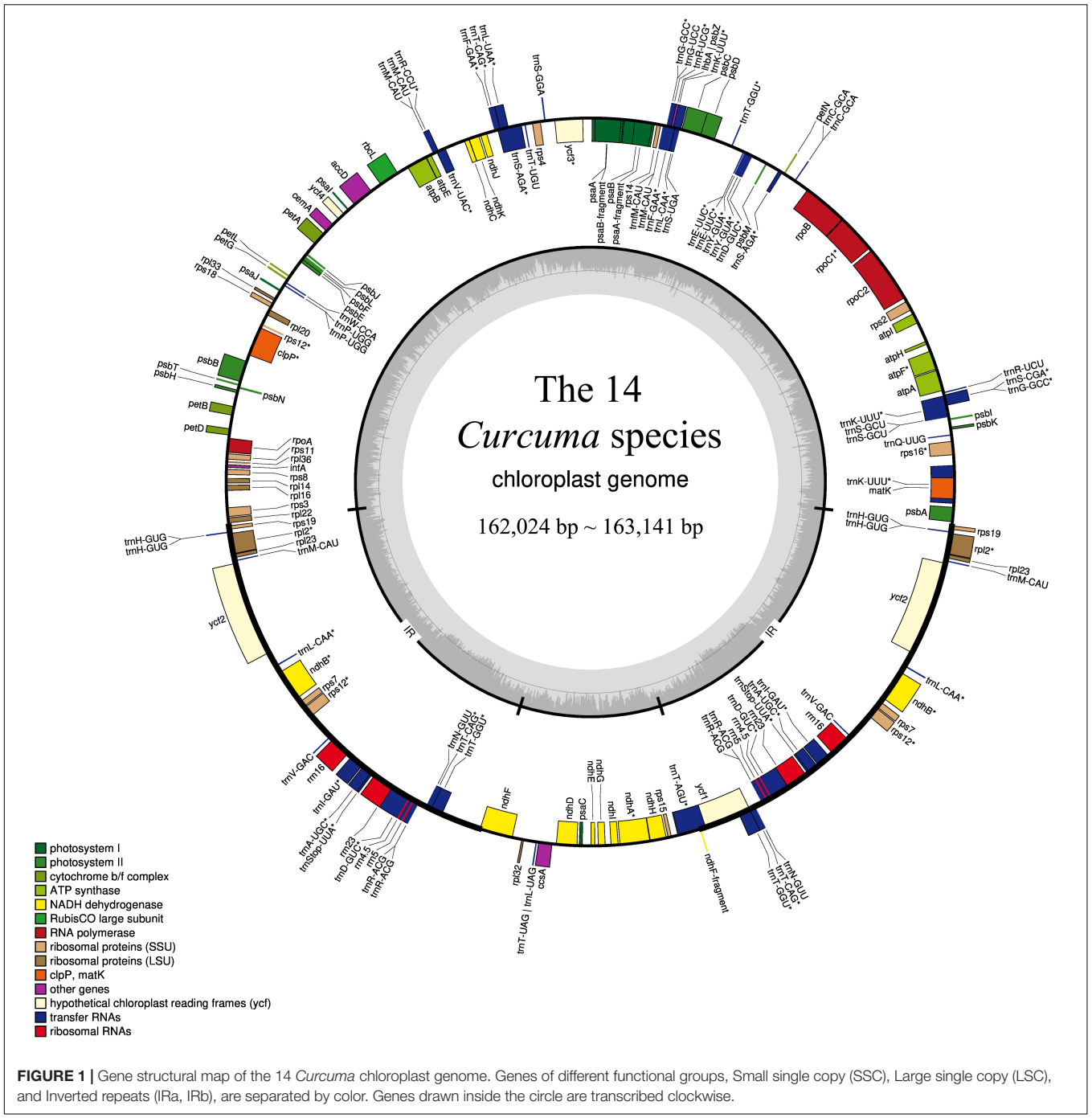


FIGURE 1 | Gene structural map of the 14 *Curcuma* chloroplast genome. Genes of different functional groups, Small single copy (SSC), Large single copy (LSC), and Inverted repeats (IRa, IRb), are separated by color. Genes drawn inside the circle are transcribed clockwise.

length of *Curcuma flaviflora* was the longest (163,141 bp) while that of *Curcuma wenyujin* was the shortest (162,024 bp). Except for *Curcuma flaviflora* and *Curcuma alismatifolia*, the sizes of the chloroplast genomes of *Curcuma* species were between 162,024 bp and 162,243 bp (**Table 1**). Except *Zingiber spectabile*, lengths of other genera of Zingiberaceae species were greater than 160,000 bp (**Supplementary Table S1**). In a word, the sizes of the chloroplast genomes of 14 *Curcuma* species were significantly similar to those of other Zingiberaceae species. In the 25 Zingiberaceae species, the sizes of the intergenic regions (IGS) ranged from 47,943 to 62,355 bp, and the sizes of the IGS ranged from 47,943 to 48,949 bp among the 14 *Curcuma* species, similar to the sizes of the complete chloroplast genomes (**Table 3** and **Supplementary Table S2**). As most same angiosperm plants, the difference of the variation in genome sizes was caused by the size of IGS. In the 25 Zingiberaceae species, the percentage of GC content was 36.00–36.29% with an average of 36.04%. In the 14 *Curcuma* species, the percentage of GC content was 36.09–36.22% (an average of 36.20%), which was higher than that of

TABLE 2 | Genes contained in the chloroplast genome sequence of 25 Zingiberaceae species.

Gene category	Groups of genes	Name of genes
Self-replication	Ribosomal RNAs	*rrn16*[b];*rrn23*[b];*rrn4.5*[b];*rrn5*[b]
	Transfer RNAs	*trnA-UGC*[a,b]; *trnC-GCA*; *trnD-GUC*; *trnE-UUC*; *trnF-GAA*; *trnfM-CAU trnG-UCC*[a]; *trnG-GCC*; *trnH-GUG*; *trnI-CAU*[b] *trnI-GAU*[a,b]; *trnK-UUU*[a] *trnL-CAA*[b]; *trnL-UAA*[a]; *trnL-UAG*; *trnM-CAU*; *trnN-GUU*[b]; *trnP-UGG trnQ-UUG*; *trnR-ACG*[b]; *trnR-UCU*; *trnS-GCU*; *trnS-GGA*; *trnS-UGA trnT-GGU*; *trnT-UGU*; *trnV-GAC*; *trnV-UAC*[a]; *trnW-CCA*;*trnY-GUA*
	Small subunit of ribosome	*rps2*; *rps3*; *rps4*; *rps7*[b]; *rps8*; *rps11*; *rps12*[a,b]; *rps14*; *rps15*; *rps16*[a]; *rps18*; *rps19*
	Large subunit of ribosome	*rpl2*[a,b]; *rpl14*; *rpl16*[a]; *rpl20*; *rpl23*[b]; *rpl32*; *rpl33*; *rpl36*
	DNA dependent RNA polymerase	*rpoA*; *rpoB*; *rpoC1*[a]; *rpoC2*
Photosynthesis	Subunits of photosystem I	*psaA*; *psaB*; *psaC*; *psaI*; *psaJ*
	Subunits of photosystem II	*psbA*; *psbB*; *psbC*; *psbD*; *psbE*; *psbF*; *psbI; psbJ*; *psbK; psbL*; *psbM*; *psbN*; *psbT*; *psbZ*
	Subunits of cytochrome	*petA*; *petB*[a]; *petD*; *petG*; *petL*; *petN*
	Subunits of ATP synthase	*atpA*; *atpB*; *atpE*; *atpF*[a]; *atpH*; *atpI*
	ATP-dependent protease subunit p gene	*clpP*[a]
	Large subunit of Rubisco	*rbcL*
	Subunits of NADH dehydrogenase	*ndhA*[a]; *ndhB*[a,b]; *ndhC*; *ndhD*; *ndhE*; *ndhF*; *ndhG*; *ndhH*; *ndhI*; *ndhJ*; *ndhK*
Other genes	Maturase	*matK*
	Envelop membrane protein	*cemA*
	Acetyl-CoAcarboxylase	*accD*
	c-type cytochrom synthesis gene	*ccsA*
	Translational initiation factor	*infA*
Genes of unknown function	Conserved open reading frames	*ycf1*[b]; *ycf2*[b]; *ycf3*[a]; *ycf4*

[a]*Intron-containing genes.* [b]*Genes located in the IR regions.*

the other genera of Zingiberaceae species (36.09%) (**Table 1** and **Supplementary Table S1**).

Contraction and Expansion of Inverted Repeats (IRs) Among the 25 Zingiberaceae Species

Of the 25 Zingiberaceae species, the genomic structure was highly conserved, including the number and sequence of genes. But, the IRa and IRb boundaries changed in structure (**Figure 3**). Although the sizes of the IR region were more highly conserved than the other regions, and the contraction and expansion of IR boundaries were accounted as playing a valuable role in genome size (Gu et al., 2017).

Among the 25 Zingiberaceae species, the sizes of the IRs of *Stahlianthus involucratus* was the longest (30,119 bp) and that of *Zingiber spectabile* was the shortest (23,277 bp). In the 14 *Curcuma* species, the sizes of the IRs of *Curcuma flaviflora* was the longest (29,784 bp) and that of *Curcuma elata* was the shortest (29,752 bp). Within the 25 Zingiberaceae species, the *rpl22-rps19* genes were located in the boundaries of LSC/IRb regions except *Zingiber spectabile*, in which there were *trnM-ycf2* genes and the *rpl22-rps19* gene was missing in the junctions of the LSC/IRb regions. The *ycf1-ndhF* genes were located at the IRb/SSC boundary of 25 Zingiberaceae species. Except *Alpinia oxyphylla*, IRb/SSC boundary was embedded in the *ndhF* genes. The IRb/SSC and SSC/IRa regions were variable. Except *Amomum krervanh*, *Wurfbainia longiligularis* and *Alpinia oxyphylla*, the SSC/IRa junctions in the chloroplast genomes of Zingiberaceae species was embedded in the *ycf1* gene, with 929–4,238 bp in the IRa region. The *rps19-psbA* genes were located in the junctions of IRa/LSC regions in the 24 Zingiberaceae species except *Zingiber spectabile*, in which the *ycf2* genes was located in the junctions of IRa/LSC regions, while the *rps19-psbA* genes could not be found in the junctions of IRa/LSC regions. For the 14 *Curcuma* species, *psbA* was located on the right side of IRa/LSC regions with the same distance of 25 bp. For the 10 Zingiberaceae species (*Amomum krervanh*, *Amomum compactum*, *Wurfbainia longiligularis*, *Wurfbainia villosa*, *Kaempferia elegans*, *Kaempferia galangal*, *Alpinia oxyphylla*, *Stahlianthus Involucratus*, *Lanxangia tsaoko,* and *Zingiber officinale*), *psbA* was located on the right side of IRa/LSC regions with the distance of 97–146 bp. In the IRb region of the 25 Zingiberaceae species, we found the *rps19* was pseudogene, and the reason was that the *rps19* coding gene existed in the boundaries of IRa-LSC. In summary, the contraction and expansion of inverted repeats were detected in the 25 Zingiberaceae species.

Repeat Structure Analysis

Of the 25 Zingiberaceae species, there were a total of 1,177 long repeats of two types (**Supplementary Table S6**). Among these Zingiberaceae species, all of them had forward and palindromic repeats. Except *Curcuma flaviflora*, the other species of the *Curcuma* had the same number of long repeats. For the other Zingiberaceae species, the number of repeats were 32–49. Most of Zingiberaceae species had 49 long repeats. *Zingiber spectabile* which had the smallest number of repeats had 19 palindromic and 13 forward repeats (**Figure 4A**). On the whole, the range of copy

TABLE 3 | Distribution of genes and intergenic regions for 14 species in *Curcuma*.

	C. rosesana	*C. amarissima*	*C. sichuanensis*	*C. elata*	*C. zanthorrhiza*	*C. longa*	*C. yunnanensis*
Protein Coding Genes Length (bp)	82975	82975	82975	82975	82975	82975	82975
GC (%)	36.91	36.92	36.92	36.92	36.92	36.91	36.92
Length (%)	51.17	51.17	51.18	51.17	51.16	51.15	51.18
rRNA							
Length (bp)	9054	9054	9054	9054	9054	9054	9054
GC (%)	55.15	55.15	55.15	55.15	55.15	55.15	55.15
Length (%)	5.58	5.58	5.58	5.58	5.58	5.58	5.58
tRNA							
Length (bp)	2920	2920	2920	2920	2920	2920	2920
GC (%)	52.91	52.91	52.91	52.91	52.91	52.91	52.91
Length (%)	1.8	1.8	1.8	1.8	1.8	1.8	1.8
Intergenic Regions							
Length (bp)	47992	47997	47967	47979	48019	48055	47967
GC (%)	30.19	30.17	30.22	30.23	30.15	30.16	30.22
Length (%)	29.6	29.6	29.58	29.59	29.61	29.62	29.58
Intron							
Length (bp)	19216	19212	19217	19243	19224	19216	19217
GC (%)	19216	19212	19217	19243	19224	19216	19217
Length (%)	11.85	11.85	11.85	11.87	11.85	11.85	11.85
	C. sp.1	***C. sp.2***	***C. wenyujin***	***C. phaeocaulis***	***C. aromatica***	***C. alismatifolia***	***C. flaviflora***
Protein Coding Genes Length (bp)	82975	82975	82943	82943	82975	83042	82930
GC (%)	36.91	36.92	36.94	36.94	36.92	36.9	36.89
Length (%)	51.17	51.18	51.19	51.19	51.14	51.04	50.83
rRNA							
Length (bp)	9054	9054	9054	9054	9054	9058	9054
GC (%)	55.15	55.15	55.15	55.15	55.15	55.17	55.15
Length (%)	5.58	5.58	5.59	5.59	5.58	5.57	5.55
tRNA							
Length (bp)	2920	2920	2920	2920	2920	2920	2920
GC (%)	52.91	52.91	52.91	52.91	52.91	52.88	52.95
Length (%)	1.8	1.8	1.8	1.8	1.8	1.79	1.79
Intergenic Regions							
Length (bp)	47990	47967	47943	47947	48051	48331	48949
GC (%)	30.19	30.22	30.22	30.22	30.21	30.11	29.97
Length (%)	29.6	29.58	29.59	29.59	29.62	29.7	30
Intron							
Length (bp)	19216	19217	19164	19167	19243	19364	19288
GC (%)	19216	19217	19164	19167	19243	19364	19288
Length (%)	11.85	11.85	11.83	11.83	11.86	11.9	11.82

GC, guanine-cytosine; LSC, large single-copy region; SSC, short single-copy region; IRs, inverted repeats.

length was 30–91 bp. The numbers of most repeat sequences in the total length were 30, 31, 32, and 46 (**Figure 4B**).

Simple Sequence Repeat (SSR) Analysis

Simple sequence repeats (SSRs), also called microsatellites, are made up of short and tandem repeat nucleotide sequences with sizes of 1–6 bp (Powell et al., 1995). SSRs are widely distributed in the chloroplast genome, and play a critical role in species authentication and are applied to molecular markers (Doorduin et al., 2011; Park et al., 2016; Yan et al., 2019). Here, we found that the distribution of 33–108 SSRs in the Zingiberaceae species that ranged in size from 10 to 20 bp (**Figure 5** and **Supplementary Table S7**). There were 6 kinds of SSRs that were discovered. Among these SSRs, only the chloroplast genomes of *Curcuma flaviflora*, *Curcuma zanthorrhiza*, *Wurfbainia villosa*, *Wurfbainia longiligularis*, *Amomum compactum*, *Amomum krervanh* and *Kaempferia galangal* had the hexa-nucleotide repeats, and the highest number of hexa-nucleotide repeats was detected in *Curcuma flaviflora*. Among 25 Zingiberaceae species, the numbers of mononucleotide repeats ranged from 27 to 90, followed by trinucleotide ranging from 3 to 9; pentanucleotide ranging from 6 to 1; hexa-nucleotide ranging from 0 to 7; dinucleotide ranging from 0 to 6 and tetranucleotide ranging from 0 to 3 (**Figure 5A**). Therefore, mononucleotide repeats

TABLE 4 | Codon content of amino acid and stop codon of 8 species.

		C. sichuanensis	*A. compactum*	*A. oxyphylla*	*W. villosa*	*Z. officinale*	*K. galanga*	*S. involucratus*	*L. tsaoko*
Amino acid	**Codon**				**RSCU[a]**				
Ala	GCA	1.31	1.29	1.29	1.29	1.25	1.29	1.32	1.29
Ala	GCC	0.56	0.54	0.55	0.55	0.58	0.56	0.55	0.56
Ala	GCG	0.32	0.34	0.31	0.34	0.32	0.32	0.32	0.33
Ala	GCU	1.81	1.83	1.85	1.83	1.86	1.83	1.81	1.82
Cys	UGC	0.54	0.55	0.53	0.55	0.5	0.54	0.56	0.55
Cys	UGU	1.45	1.45	1.47	1.45	1.5	1.46	1.44	1.45
Asp	GAC	0.35	0.37	0.38	0.37	0.36	0.37	0.36	0.38
Asp	GAU	1.65	1.63	1.62	1.63	1.64	1.63	1.64	1.62
Glu	GAA	1.49	1.5	1.52	1.5	1.49	1.51	1.5	1.51
Glu	GAG	0.51	0.5	0.48	0.5	0.51	0.49	0.51	0.49
Phe	UUC	0.71	0.7	0.68	0.71	0.71	0.69	0.71	0.7
Phe	UUU	1.29	1.3	1.32	1.29	1.29	1.31	1.29	1.3
Gly	GGA	1.59	1.61	1.6	1.59	1.59	1.6	1.6	1.59
Gly	GGC	0.34	0.33	0.33	0.34	0.36	0.36	0.34	0.34
Gly	GGG	0.67	0.65	0.63	0.67	0.61	0.62	0.67	0.68
Gly	GGU	1.4	1.41	1.44	1.4	1.43	1.42	1.39	1.39
His	CAC	0.43	0.42	0.41	0.42	0.45	0.44	0.44	0.43
His	CAU	1.57	1.58	1.59	1.58	1.55	1.56	1.56	1.57
Ile	AUA	0.97	0.96	0.98	0.97	0.96	0.98	0.97	0.99
Ile	AUC	0.55	0.55	0.53	0.55	0.53	0.53	0.56	0.54
Ile	AUU	1.48	1.48	1.49	1.48	1.51	1.49	1.47	1.48
Lys	AAA	1.47	1.47	1.47	1.47	1.5	1.5	1.47	1.43
Lys	AAG	0.53	0.53	0.53	0.53	0.5	0.5	0.53	0.57
Leu	CUA	0.8	0.79	0.8	0.8	0.76	0.8	0.8	0.84
Leu	CUC	0.41	0.41	0.41	0.41	0.4	0.39	0.4	0.4
Leu	CUG	0.38	0.38	0.36	0.37	0.37	0.36	0.4	0.38
Leu	CUU	1.24	1.24	1.24	1.26	1.23	1.23	1.24	1.27
Leu	UUA	1.86	1.89	1.93	1.88	1.94	1.95	1.84	1.83
Leu	UUG	1.31	1.3	1.26	1.29	1.3	1.25	1.31	1.28
Met	AUG	1	1	1	1	1	1	1	1
Asn	AAC	0.47	0.46	0.46	0.46	0.45	0.47	0.48	0.48
Asn	AAU	1.53	1.54	1.54	1.54	1.55	1.53	1.52	1.52
Pro	CCA	1.18	1.2	1.19	1.18	1.18	1.15	1.18	1.19
Pro	CCC	0.73	0.75	0.77	0.74	0.77	0.76	0.73	0.73
Pro	CCG	0.48	0.47	0.44	0.47	0.48	0.49	0.48	0.48
Pro	CCU	1.62	1.58	1.6	1.6	1.57	1.6	1.61	1.59
Gln	CAA	1.52	1.51	1.52	1.51	1.51	1.52	1.52	1.52
Gln	CAG	0.48	0.49	0.48	0.49	0.49	0.48	0.48	0.48
Arg	AGA	1.9	1.88	1.89	1.9	1.94	1.88	1.87	1.93
Arg	AGG	0.75	0.74	0.7	0.73	0.71	0.7	0.78	0.75
Arg	CGA	1.25	1.27	1.27	1.27	1.24	1.29	1.24	1.24
Arg	CGC	0.35	0.34	0.33	0.34	0.34	0.35	0.34	0.35
Arg	CGG	0.46	0.47	0.42	0.47	0.41	0.42	0.47	0.45
Arg	CGU	1.3	1.3	1.38	1.28	1.36	1.36	1.3	1.28
Ser	AGC	0.35	0.35	0.36	0.35	0.35	0.35	0.36	0.36
Ser	AGU	1.23	1.22	1.3	1.22	1.34	1.31	1.22	1.2
Ser	UCA	1.18	1.15	1.15	1.16	1.16	1.17	1.17	1.19
Ser	UCC	0.99	0.99	0.95	0.97	0.97	0.95	1	0.97
Ser	UCG	0.55	0.58	0.54	0.57	0.54	0.53	0.56	0.59
Ser	UCU	1.69	1.71	1.7	1.72	1.64	1.69	1.69	1.7

(Continued)

TABLE 4 | Continued

		C. sichuanensis	*A. compactum*	*A. oxyphylla*	*W. villosa*	*Z. officinale*	*K. galanga*	*S. involucratus*	*L. tsaoko*
Amino acid	**Codon**				**RSCU[a]**				
Thr	ACA	1.28	1.28	1.28	1.28	1.25	1.29	1.27	1.28
Thr	ACC	0.72	0.69	0.69	0.7	0.67	0.69	0.71	0.7
Thr	ACG	0.47	0.46	0.45	0.47	0.43	0.44	0.46	0.46
Thr	ACU	1.54	1.58	1.58	1.55	1.64	1.58	1.56	1.57
Val	GUA	1.48	1.51	1.49	1.49	1.5	1.48	1.5	1.47
Val	GUC	0.49	0.49	0.45	0.49	0.48	0.47	0.5	0.48
Val	GUG	0.57	0.55	0.57	0.57	0.56	0.56	0.57	0.58
Val	GUU	1.46	1.45	1.49	1.45	1.46	1.49	1.44	1.46
Trp	UGG	1	1	1	1	1	1	1	1
Tyr	UAC	0.44	0.43	0.43	0.43	0.43	0.44	0.43	0.44
Tyr	UAU	1.56	1.57	1.57	1.57	1.57	1.56	1.57	1.56
Stop*	UAA	1.28	1.23	1.3	1.28	1.43	1.33	1.26	1.22
Stop*	UAG	0.91	0.94	0.91	0.93	0.83	0.92	0.9	1
Stop*	UGA	0.81	0.83	0.79	0.8	0.75	0.76	0.83	0.78

[a]*Relative synonymous codon usage; *Stop codon.*

may play a more significant role than the other SSRs in genetic variation (Jiao et al., 2012).

Among these SSRs in the 25 Zingiberaceae species, the percentages of A/T mononucleotide repeats were 43.12 and 54.35% respectively. The percentages of C/G mononucleotide repeats were 2.1 and 0.43% respectively. The combination of A/T in most the other SSRs may be the reason for bringing about the high AT content in the whole chloroplast genomes within the 25 Zingiberaceae species (**Figure 5B**). Among all SSR types, previous studies found that A/T always richest continually used bases (Wu et al., 2017). In the 14 *Curcuma* species, the number of A mononucleotide repeats ranged from 24 to 30, with T mononucleotide repeats ranging from 28 to 41, except in *Curcuma flaviflora*. In addition, in the same genus, we can infer that the number of A/T repeats may be similarity. However, in *Curcuma flaviflora* was 15/20, which were much lower than those of the other 13 *Curcuma* species. This may be the main reason why the lower intergenic spacers were in *Curcuma flaviflora*.

Among the 25 Zingiberaceae species (**Figure 5C**), SSRs were located in intergenic spacers, with an average of 41. The next SSRs were distributed in coding genes with an average of 20. The least number of SSRs were found in the introns, with an average of 10 (**Figure 5D**). The SSR loci were located in 5 coding genes (*rpoC2*, *rps14*, *petA*, *rps18,* and *ycf1)* and 32 intergenic regions of the 25 Zingiberaceae species. Therefore, these results suggest that the availability of SSRs can be greatly applied to develop useful molecular markers for identifying genetic diversity, evolutionary studies and phylogeny.

Divergence Hotspots

In chloroplast genomes research, divergent hotspots are usually used as important evidence for species authentication and provide information about phylogeny (Yang et al., 2013; Du et al., 2017; Jiang et al., 2017; Niu et al., 2017; Zhou et al., 2018). To determine divergent hotspots region in 25 Zingiberaceae species, we compared the Pi values of intergenic regions and coding regions in DnaSP 5.1. The sliding window analysis indicated that the average value of the IR regions was lower than that in the LSC and SSC regions, which showed that most of the variation in LSC and SSC regions (**Figure 6**). Among the 25 Zingiberaceae species, the average value of nucleotide diversity (Pi) was 0.01257. In LSC and SSC regions, there were five mutational hotspots including *Rps16-trnQ*, *ycf1*, *ndhA*, *psaC,* and *rps12* which exhibited remarkably higher Pi values (>0.03). But in the IR regions, there was no one mutational hotspot that exhibited remarkably higher Pi values (>0.01). The average value of nucleotide diversity (Pi) was 0.00267 among the 14 *Curcuma* species, where there were five mutational hotspots including *matK-trnk*, *Rps16-trnQ*, *petN-psbM*, *rpl32*, *ndhA* and *ycf1* exhibited remarkably higher Pi values (>0.006). By contrast, the Pi values of 25 Zingiberaceae species were higher than those of the 14 *Curcuma* species (**Figure 6** and **Supplementary Table S8**). The mutational hotspots might be adopted as appropriate loci for population genetic and phylogeographic studies.

Phylogenetic Analyses of the Zingiberaceae Species Based on Chloroplast Genome Sequence Within Zingiberales

To discuss the phylogenetic positions of 56 chloroplast genomes (The species name and GenBank accession numbers could be find in **Supplementary Table S9**) based on the 79 shared protein coding genes, we used phylogenetic tree to infer the phylogenetic relationships. These chloroplast genomes from seven families within Zingiberales included 35 Zingiberaceae species, four Musaceae species, two Strelitziaceae species, one Lowiaceae species, one Cannaceae species, seven Marantaceae species, two Costaceae species and four other family species as out groups. The topologies of the three dataset (MP, ML, and BI) yielded a similar structure. The seven families of Zingiberales can be classified into seven monophyletic clades (Musaceae,

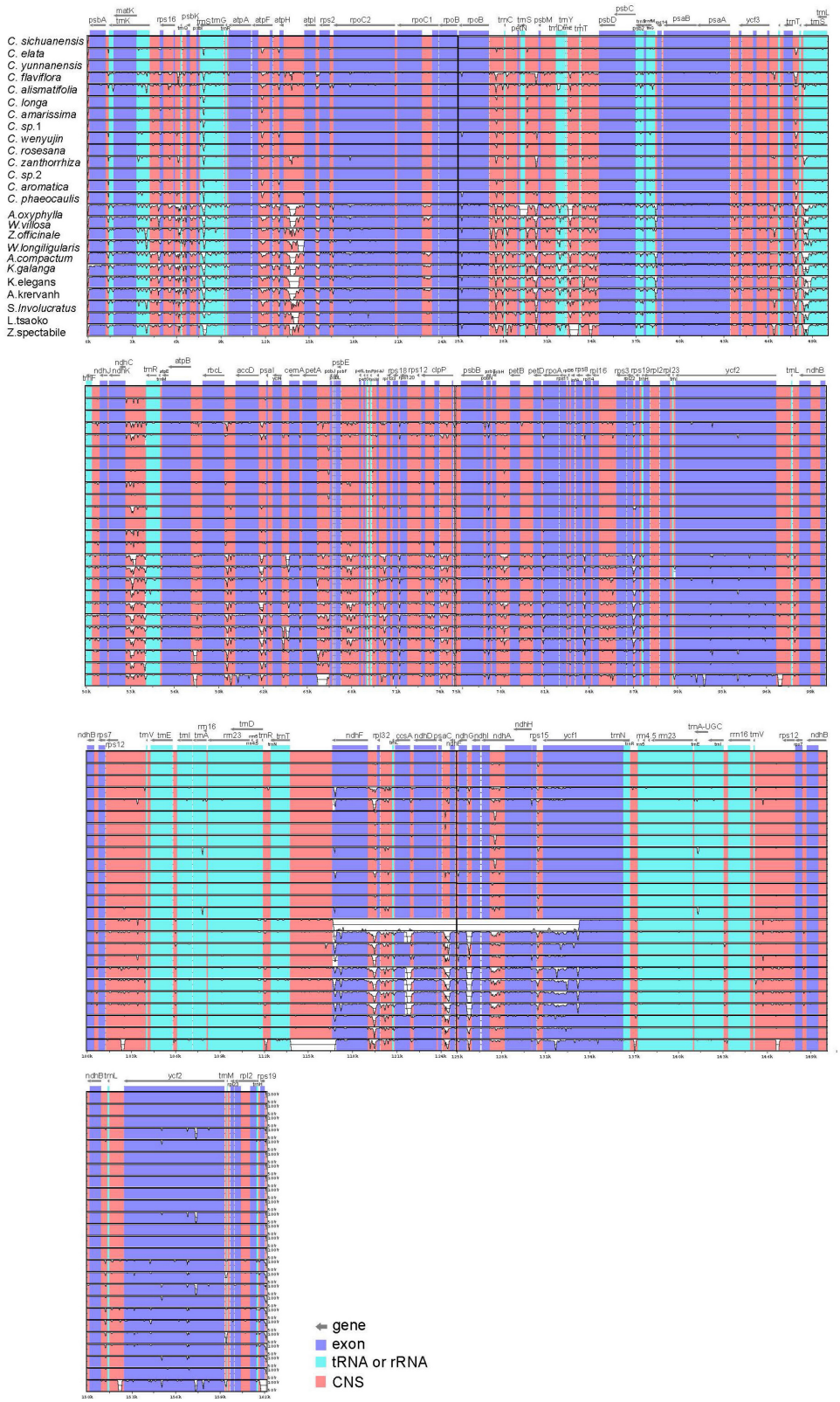


FIGURE 2 | Sequence alignment of 25 Zingiberaceae chloroplast genome with *C. sichuanensis* as a reference by using in mVISTA. The Y-scale represents the percentage of identity ranging from 50 to 100%.

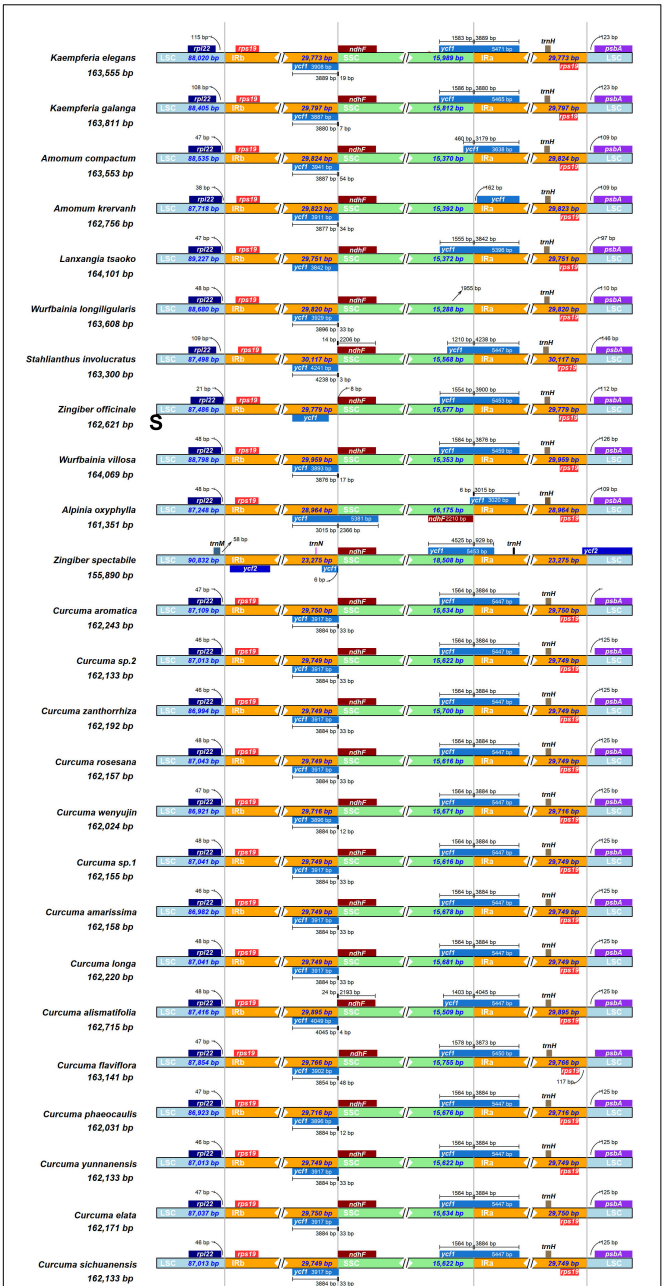


FIGURE 3 | Comparison of the junctions between the LSC, SSC and IR regions among 25 Zingiberaceae chloroplast genome. The figure is not to scale LSC, SSC, and IR.

Strelitziaceae, Lowiaceae, Cannaceae, Marantaceae Costaceae, and Zingiberaceae). In addition, Musaceae was the basal group in Zingiberales. Cannaceae was sister to Marantaceae. The Zingiberaceae were gathered into one clade, and *Siphonochilus kirkii* was the first to be isolated (basal taxa). Most species from the same genus were clustered together. The genetic relationship of most of *Wurfbainia* and *Amomum* were very close. Besides that, *Amomum paratsaoko* and *Lanxangia tsaoko* formed a closely relationship. The 14 *Curcuma* species were gathered into one clade except *Curcuma alismatifolia* and *Curcuma flaviflora*. *Curcuma alismatifolia* was initially clustered together with *Stahlianthus Involucratus* and then with other 12 *Curcuma* species. *Curcuma flaviflora* was clustered together with *Zingiber*, and the similar clustering result was measured in previous studies (Cui et al., 2019; Gui et al., 2020). In the ML tree, under 90% were only three nodes with bootstrap values and four nodes in the MP tree (**Figure 7**). The Musaceae family was the earliest branch and it was sister to the other families in Zingiberales. The 35 Zingiberaceae species clustered for one clade, which was sister to two Costaceae species (*Chamaecostus acaulis* and *Costus dubius*). The tree was cleared to reveal the phylogenetic relationships among Zingiberales species.

DISCUSSION

In this study, we compared the complete chloroplast genome of 25 Zingiberaceae species, which exhibited a typical quadripartite structures and included 107–116 particular genes composed of 78–85 coding genes, 24–31 tRNAs and 4 rRNAs. In these chloroplast genomes, 155,890–164,101 bp were generated in length, the overall GC contents were between 36.00 and 36.29%. It was obvious that the structure, genome length and organization had a high conservation. Through comparative study, it was found that the intergenic areas was the largest variable regions among the 25 Zingiberaceae chloroplast genomes, and it is a frequent phenomenon in chloroplast genomes (Wicke et al., 2011).

The analysis found that the Ka/Ks and evolutionary rate were low in these Zingiberaceae species as expected. As a common phenomenon in photosynthetic plants, the evolutionary rates of photosynthesis genes (*atpH*, *petL*, *petN*, *psaC*, *psaJ*, *psbM*, *psbT*, *rps2,* and *rps7*) were low. The three genes evolved faster including *atpF* (Subunits of ATP synthase), *rps8* (Small subunit of ribosome) and *ndhC* (Subunits of NADH dehydrogenase). In most cases, the rates of evolution in some genes are species-specific, *clpP* gene is highly conserved plastid-encoding gene, but in some angiosperms is the most variablegene. The plastid gene *Clp* protease complex was sharply different in evolutionary rate (Williams et al., 2019). Unlike most plants, the results indicate that the mean of Ka/Ks of the *clpP* gene in 25 Zingiberaceae species was 0.02. We found two introns were contained in the *clpP* gene, which may lead to the low Ka/Ks. Genes under positive selection usually caused lots of repeating amino acid sequence insertions in varying degrees and may also be closely tied to a near increase in the rate of diversification after they adapted to fit a new ecological environment (Erixon and Oxelman, 2008; Piot et al., 2018).

The interval of the four chloroplast genome regions played an important role in some plant species evolution. In most cases, contraction and expansion of the IR region could result in the production of pseudogenes such as ψ*ycf1*. In addition to *Wurfbainia longiligularis*, the ψ*ycf1* was discovered among the Zingiberaceae species.

Complete chloroplast genomes provide sufficient informative sites for resolving phylogenetic relationships of plant, and have

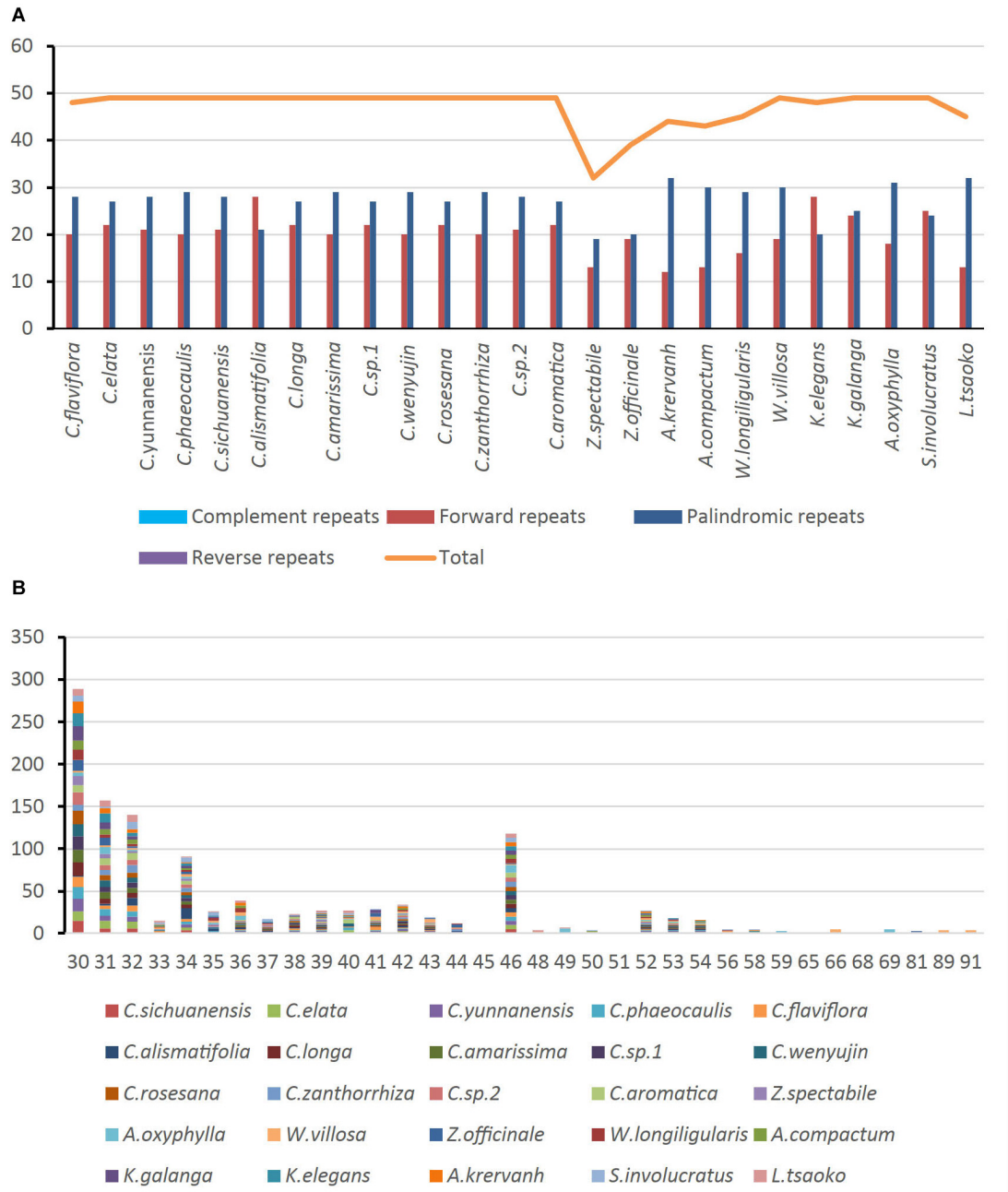


FIGURE 4 | Number of long repetitive repeats on the complete chloroplast genome sequence of 25 Zingiberaceae species; **(A)** frequency of repeat type; **(B)** frequency of the repeats.

been examined to be effective in the ability of differentiation in lower taxonomic levels (He et al., 2012; Zhang et al., 2016). According to DNA sequences of the nuclear internal transcribed spacer (ITS) and plastid *matK* regions, *Siphonochilus* and *Tamijia* are the basal taxa of the Zingiberaceae family (Kress et al., 2002). But in morphology show that *Alpinieae* is the basal taxa of the Zingiberaceae family (Wood et al., 2000). In our study, we can only speculate that *Siphonochilus* is the basal taxa in Zingiberaceae because of lack of reliable chloroplast genome of *Tamijia*. Besides that, *Curcuma* is paraphyletic with *Stahlianthu* on the basis of the evidence from morphological study. Due to lack of published chloroplast genomes of *Stahlianthu*, our results suggested *Curcuma* had a close relationship with *Stahlianthu* only. The findings were similar to those in previous studies where *Curcuma flaviflora* had a close relation with *Zingiber* species, and they infer that the genes of genetic variations in *Curcuma* has evolved (forest in the wild and heliophyte) in ordet to adapt to the variety of environment (Gui et al., 2020). Besides,

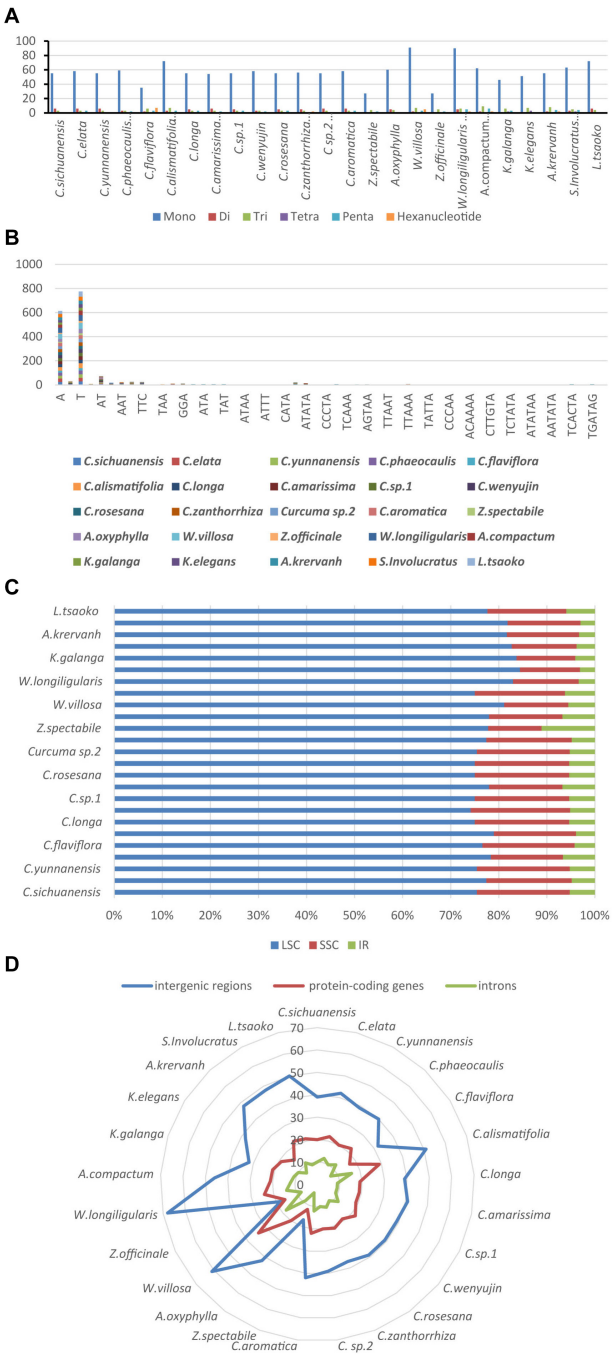


FIGURE 5 | The comparison of SSR distribution in 25 chloroplast genomes; **(A)** number of different SSR types; **(B)** frequency of common motifs; **(C)** frequency of SSRs in the LSC, IR, SSC region; **(D)** frequency of SSRs in the intergenic regions, protein-coding genes and introns.

Curcuma lies dormant in winter, rhizomes fleshy with tuber-bearing roots, and blooms at the rainy season, similar to the drought-enduring plants, and the origination and evolution of *Curcuma* is relate to the drought-enduring plants or not? It is still necessary to need a lot more work to tease out. The MP, ML and BI tree showed a very high level of support in our study. In morphological classification of *Curcuma*, flowers with funnel-shaped corollas, lateral staminodes petaloid, lobes with oblong and ovate and base adnate to labellum and filament. The characteristics of *Zingiber* were valgus lip with 3-lobed, lateral staminodes adnate to labellum, central lobe retuse or cleft at apex and the style is slender, extending beyond anther locules. The morphology of *Curcuma flaviflora* was consistent with the genus of *Curcuma*, although the morphology of *Curcuma flaviflora* and *Zingiber* was different. The distribution range of them overlapped, and the introgressive hybridization may exist extensively. The results showed that *Curcuma* and even Zingiberaceae species may have a complex phylogeny, and more samples and DNA data were necessary to confirm their relationship. To understand their evolutionary history would provide us more useful information to study how to adapt to the environment and raise productivity in plant.

MATERIALS AND METHODS

Plant Material and DNA Extraction

In the study, the fresh leaves of fourteen species of *Curcuma* within Zingiberaceae were provided by Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, and identified by Mr. Jianping Wang. All fresh leaves from 14 samples were desiccated in silica gel and deposited at the herbarium of Sichuan Agriculture University (SICAU). A modified CTAB protocol was used to extract total genomic DNA (Zhang et al., 2017), and the concentration and quality of the extracted DNA were respectively determined by gel electrophoresis and Nanodrop 2000C spectrophotometry.

Chloroplast Genome Sequencing, Assembly, Mapping, Annotation, and Structure

Following the instruction of library, the purified DNA was used to sequence library construction. Total DNA was sequenced in Illumina HiSeq 2000, Paired-end libraries was read of 150 bp (Gu et al., 2018). Raw sequencing data from the samples were trimmed and filtered by NGSQC Toolkit version 2.3.3 (Patel and Jain, 2012). The SPAdes v3.6.0 was used to *de novo* assembly the matched paired-end reads, and the PRICE (Paired-Read Iterative Contig Extension) was used to repeat the combination of contig and uncompleted *de novo* assembly (Bankevich et al., 2012). In order to check the *de novo* assembly, reads were aligned against the assembled genome using BWA (Li and Durbin, 2009) with default parameters. 15.6% of reads were successfully aligned, but this was sufficient to give a mean coverage of more than 5000×. This mapping enabled us to resolve some short regions of ambiguous sequence in the assembly. The automatic annotator DOGMA was used to annotate the chloroplast genome with BLAST searches against the chloroplast genomes of sibling species. Open reading frames (ORFs) was matched with annotated chloroplast genome, and then the remaining lacking protein evidence ORFs were disregarded

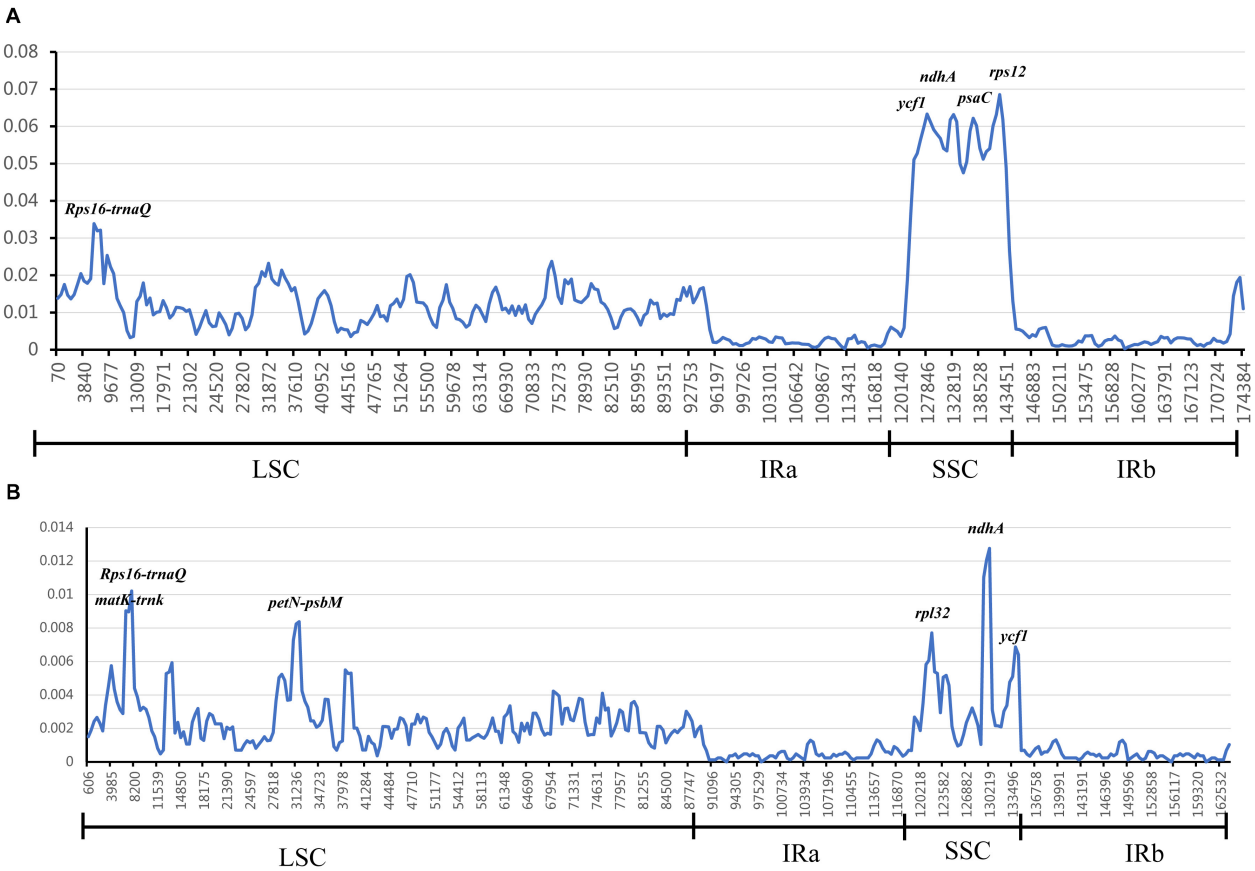


FIGURE 6 | Sliding window analysis of 25 Zingiberaceae chloroplast genomes. The nucleotide variability (Pi) value in: **(A)** 25 Zingiberaceae species; **(B)** 14 *Curcuma* species.

(Wyman et al., 2004). The genes were considered potential pseudogenes which contained one or more frameshift mutations or premature stop codons. The online program OGDRAW was used to draw the circular chloroplast genome map and then manually edited (Conant and Wolfe, 2008).

Codon Usage Analysis

The chloroplast genome of 25 Zingiberaceae species were analyzed for the relative synonymous codon usage (RSCU). When the RSCU value > 1.00, it means the use of a codon is more frequent than expected, and vice versa. The software DAMBE5 was used to obtain the RSCU (Xia, 2013).

Genome Comparison and Molecular Marker Identification

In order to compare the chloroplast genomes of 25 Zingiberaceae species, we took the reference sequence from *Curcuma sichuanensis*, and used mVISTA in LAGAN mode in pairwise alignments (Frazer et al., 2004). The rates of Ka/Ks was calculated using DnaSP v6.0 (Rozas et al., 2017). Seventy seven protein coding gene from 25 Zingiberaceae were used to calculate evolutionary rate variation (Librado and Rozas, 2009). The nucleotide diversity (Pi) values and sequence polymorphism of Zingiberaceae species were evaluated and the regions of LSC, SSC, and IR were also calculated in DnaSP 5.1 (Rozas et al., 2003).

Long Repetitive Sequences and Simple Sequence Repeats (SSRs) Analysis

We calculated two types of long repeats: forward (F) and palindromic (P) in the chloroplast genomes. The REPuter was used to detect the size and location of the long repeats (Kurtz et al., 2001). The positions of repeats were identified according to the parameters set as the following conditions: (1) the sequence identity was >90%; (2) the minimum repeat was 30 bp; (3) the Hamming distance was 1. SSRs were detected using the Msatcommander 0.8.2.0 with the following settings: ≥10 for mono-; ≥8 for di-; ≥4 for tri-, tetra-; ≥3 for penta- and hexanucleotide SSRs.

Phylogenetic Analysis

To determine the phylogenetic relationships and examining the phylogenetic status of Zingiberaceae within Zingiberales, we used MAFFT version 7.0 software to align the 79 shared protein coding gene DNA sequences in default parameter settings (Katoh and Standley, 2013). The Zingiberales species included Zingiberaceae, Musaceae, Strelitziacea, Lowiaceae

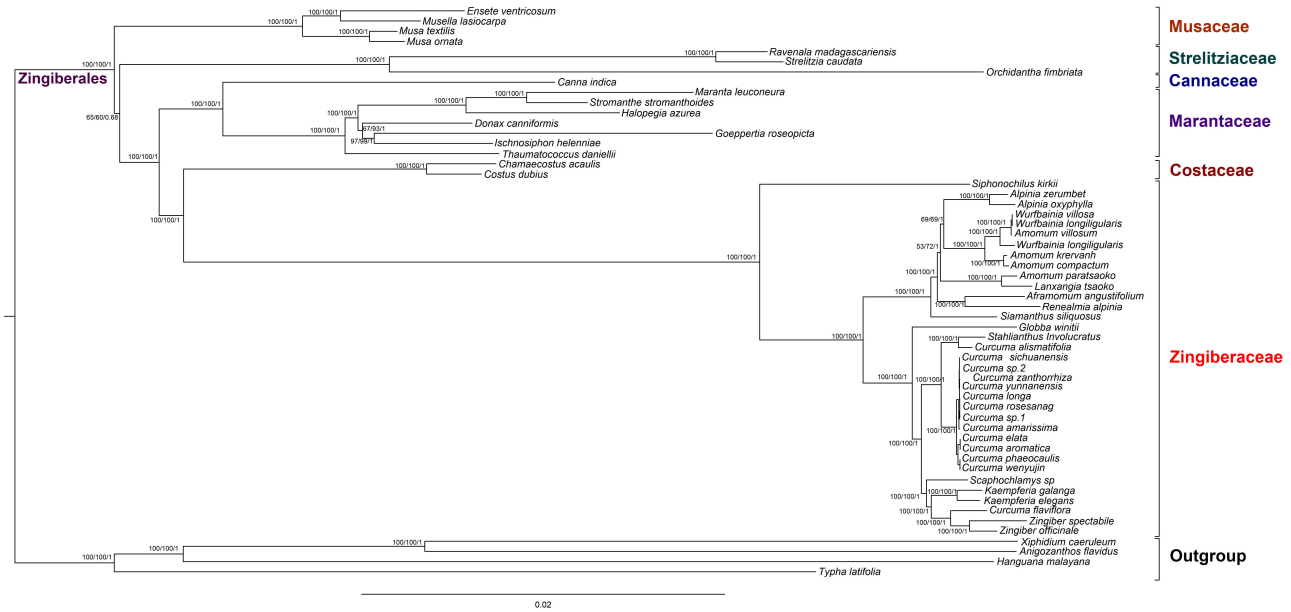


FIGURE 7 | Phylogenetic relationships of 52 Zingiberales species taxa inferred from Maximum Parsimony (MP), Bayesian Inference (BI), and Maximum Likelihood (ML) analyses of protein-coding genes. *Xiphidium caeruleum*, *Anigozanthos flavidus*, *Hanguana malayana,* and *Typha latifolia* was used as the outgroup. Numbers above nodes are support values with MP bootstrap values on the left, ML bootstrap values in the middle, and Bayesian posterior probabilities (PP) values on the right.

species, Cannaceae, Marantaceae and Costaceae species with four other species as outgroups. The maximum parsimony (MP) and maximum likelihood (ML) tree were inferred in the software PUAP* (Swofford and Swofford, 2002). We used Modeltest version 3.7 (Posada and Buckley, 2004) to calculate the best-fit model based on Akaike's information criterion (AIC). The Bayesian inference (BI) tree was analyzed in MrBayes version 3.1.2 (Ronquist and Huelsenbeck, 2003). A random tree and run for 3,000,000 generations started in Markov chain Monte Carlo (MCMC). Sampling in every 100 generations and 25% of our samples (which the first 2500 trees) were discarded in burnin. The Bayesian posterior probabilities of the nodal supports were inferred and the 50% majority-rule consensus tree was constructed based on the rest of trees.

CONCLUSION

In this study, 14 complete chloroplast genome of *Curcuma* were reported, including seven firstly sequenced chloroplast genomes (*C. zanthorrhiza, C. elata, C. yunnanensis, C. alismatifolia, C. amarissima, C. sichuanensis,* and *C. rosesana*) comparatively analyzed with other published genus in the family of Zingiberaceae for the first time. The structure and gene content of chloroplast genome of 25 Zingiberaceae were similarity, and showed highly conserved. The MP, ML, and BI tree showed that 52 Zingiberales species were completely clustered into seven families with high bootstrap values. This study provide the availability of chloroplast genome sequences for improving species identification and phylogenetic research in further study within Zingiberales.

DATA AVAILABILITY STATEMENT

The assembled chloroplast genome of 14 *Curcuma* was deposited in GenBank: MT395655, MT395645, MT395646, MT395649, MT395651, MT395644, MT395657, MT395653, MT395648, MT395650, MT395652, MT395654, MT395647, and MT395656.

AUTHOR CONTRIBUTIONS

RY designed the research and revised the manuscript. HL, YZ, JD, and GG performed the experiments and analyzed the data. HL wrote the manuscript. CD and LZ contributed to the materials and revised the manuscript. All the authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by Sichuan Science and Technology Program (2018TJPT0013 and 2019YFS0021).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00802/full#supplementary-material

TABLE S1 | Complete chloroplast genomes for 11 Zingiberaceae specie.

TABLE S2 | Distribution of genes and Intergenic regions for 11 species in Zingiberaceae.

TABLE S3 | The genes having intron in the 25 Zingiberaceae chloroplast genomes.

TABLE S4 | Codon content of amino acid and stop codon of 25 Zingiberaceae species.

TABLE S5 | The rates of Ka, Ks.

TABLE S6 | The comparison of Long repeats among 25 Zingiberaceae species.

TABLE S7 | The comparison of SSRs among 25 Zingiberaceae species.

TABLE S8 | The Pi value of 14 *Curcuma* species.

TABLE S9 | The GenBank accession numbers of 56 species using in phylogenetic analysis.

REFERENCES

Alwadani, K. G., Janes, J. K., and Andrew, R. L. (2019). Chloroplast genome analysis of box-ironbark Eucalyptus. *Mol. Phylogenet. Evol.* 136, 76–86. doi: 10.1016/j.ympev.2019.04.001

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Chen, J., and Xia, N. H. (2010). Chromosome cytology, leaf epidermal morphology and palynology of *Curcuma rubrobracteata* (Zingiberaceae). *Nordic J. Bot.* 28, 212–215. doi: 10.1111/j.1756-1051.2009.00536.x

Chorev, M., and Carmel, L. (2012). The function of introns. *Front. Genet.* 3:55. doi: 10.3389/fgene.2012.00055

Conant, G. C., and Wolfe, K. H. (2008). GenomeVx: simple web-based creation of editable circular chromosome maps. *Bioinformatics* 24, 861–862. doi: 10.1093/bioinformatics/btm598

Cui, Y., Chen, X., Nie, L., Sun, W., Hu, H., Lin, Y., et al. (2019). Comparison and phylogenetic analysis of chloroplast genomes of three medicinal and edible Amomum Species. *Int. J. Mol. Sci.* 20:4040. doi: 10.3390/ijms20164040

Deng, J., Gao, G., Ahmad, K., Luo, X., Zhang, F., Li, S., et al. (2018). Evaluation on genetic relationships among China's endemic Curcuma L. herbs by mtDNA. *Phyton Int. J. Exp. Bot.* 87, 156–161.

Doorduin, L., Gravendeel, B., Lammers, Y., Ariyurek, Y., Chin-A-Woeng, T., and Vrieling, K. (2011). The complete chloroplast genome of 17 individuals of pest species *Jacobaea vulgaris*: SNPs, microsatellites and barcoding markers for population and phylogenetic studies. *DNA Res.* 18, 93–105. doi: 10.1093/dnares/dsr002

Du, F. K., Petit, R. J., and Liu, J. Q. (2009). More introgression with less gene flow: chloroplast vs. mitochondrial DNA in the *Picea asperata* complex in China, and comparison with other conifers. *Mol. Ecol.* 18, 1396–1407. doi: 10.1111/j.1365-294x.2009.04107.x

Du, Y.-P., Bi, Y., Yang, F.-P., Zhang, M.-F., Chen, X.-Q., Xue, J., et al. (2017). Complete chloroplast genome sequences of Lilium: insights into evolutionary dynamics and phylogenetic analyses. *Sci. Rep.* 7, 1–10.

Erixon, P., and Oxelman, B. (2008). Whole-gene positive selection, elevated synonymous substitution rates, duplication, and indel evolution of the chloroplast clpP1 gene. *PLoS One* 3:e1386. doi: 10.1371/journal.pone.0001386

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., and Dubchak, I. (2004). VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* 32, W273–W279.

Gitzendanner, M. A., Soltis, P. S., Wong, G. K. S., Ruhfel, B. R., and Soltis, D. E. (2018). Plastid phylogenomic analysis of green plants: a billion years of evolutionary history. *Am. J. Bot.* 105, 291–301. doi: 10.1002/ajb2.1048

Gonçalves, D. J., Simpson, B. B., Ortiz, E. M., Shimizu, G. H., and Jansen, R. K. (2019). Incongruence between gene trees and species trees and phylogenetic signal variation in plastid genes. *Mol. Phylogenet. Evol.* 138, 219–232. doi: 10.1016/j.ympev.2019.05.022

Gu, C., Dong, B., Xu, L., Tembrock, L. R., Zheng, S., and Wu, Z. (2018). The complete chloroplast genome of Heimia myrtifolia and comparative analysis within myrtales. *Molecules* 23:846. doi: 10.3390/molecules23040846

Gu, C., Ma, L., Wu, Z., Chen, K., and Wang, Y. (2019). Comparative analyses of chloroplast genomes from 22 Lythraceae species: inferences for phylogenetic relationships and genome evolution within Myrtales. *BMC Plant Biol.* 19:281. doi: 10.1186/s12870-019-1870-3

Gu, C., Tembrock, L. R., Zhang, D., and Wu, Z. (2017). Characterize the complete chloroplast genome of *Lagerstroemia floribunda* (Lythraceae), a narrow endemic crape myrtle native to Southeast Asia. *Conserv. Genet. Resour.* 9, 91–94. doi: 10.1007/s12686-016-0628-6

Gui, L., Jiang, S., Xie, D., Yu, L., Huang, Y., Zhang, Z., et al. (2020). Analysis of complete chloroplast genomes of Curcuma and the contribution to phylogeny and adaptive evolution. *Gene. Mar.* 30:144355. doi: 10.1016/j.gene.2020.144355

Guo, X., Liu, J., Hao, G., Zhang, L., Mao, K., Wang, X., et al. (2017). Plastome phylogeny and early diversification of Brassicaceae. *BMC Genomics* 18:176. doi: 10.1186/s12864-017-3555-3

He, S., Wang, Y., Volis, S., Li, D., and Yi, T. (2012). Genetic diversity and population structure: implications for conservation of wild soybean (Glycine soja Sieb. et Zucc) based on nuclear and chloroplast microsatellite variation. *Int. J. Mol. Sci.* 13, 12608–12628. doi: 10.3390/ijms131012608

Huang, D. I., Hefer, C. A., Kolosova, N., Douglas, C. J., and Cronk, Q. C. (2014). Whole plastome sequencing reveals deep plastid divergence and cytonuclear discordance between closely related balsam poplars, *P opulus* balsamifera and *P. trichocarpa* (*S alicaceae*). *New Phytol.* 204, 693–703. doi: 10.1111/nph.12956

Jiang, D., Zhao, Z., Zhang, T., Zhong, W., Liu, C., Yuan, Q., et al. (2017). The chloroplast genome sequence of *Scutellaria baicalensis* provides insight into intraspecific and interspecific chloroplast genome diversity in *Scutellaria*. *Genes* 8:227. doi: 10.3390/genes8090227

Jiao, Y., Jia, H.-M., Li, X.-W., Chai, M.-L., Jia, H.-J., Chen, Z., et al. (2012). Development of simple sequence repeat (SSR) markers from a genome survey of Chinese bayberry (*Myrica rubra*). *BMC Genomics* 13:201. doi: 10.1186/1471-2164-13-201

Joly, S., McLenachan, P. A., and Lockhart, P. J. (2009). A statistical approach for distinguishing hybridization and incomplete lineage sorting. *Am. Nat.* 174, E54–E70.

Kanlayanapaphon, C., and Newman, M. (2003). The phylogeny of tribe Zingibereae (Zingiberaceae) based on ITS (nrDNA) and trnL-F (cpDNA) sequences. *Edinburgh J. Bot.* 60, 483–507. doi: 10.1017/s0960428603000362

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kelchner, S. A. (2002). Group II introns as phylogenetic tools: structure, function, and evolutionary constraints. *Am. J. Bot.* 89, 1651–1669. doi: 10.3732/ajb.89.10.1651

Kobayashi, Y., Misumi, O., Odahara, M., Ishibashi, K., Hirono, M., Hidaka, K., et al. (2017). Holliday junction resolvases mediate chloroplast nucleoid segregation. *Science* 356, 631–634. doi: 10.1126/science.aan0038

Kress, W. J., Prince, L. M., and Williams, K. J. (2002). The phylogeny and a new classification of the gingers (Zingiberaceae): evidence from molecular data. *Am. J. Bot.* 89, 1682–1696. doi: 10.3732/ajb.89.10.1682

Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* 29, 4633–4642. doi: 10.1093/nar/29.22.4633

Leong-Škorničková, J., Šída, O., Jarolímová, V., Sabu, M., Fér, T., Trávníček, P., et al. (2007). Chromosome numbers and genome size variation in Indian species of Curcuma (Zingiberaceae). *Ann. Bot.* 100, 505–526. doi: 10.1093/aob/mcm144

Li, D. M., Xu, Y. C., and Zhu, G. F. (2019). Complete chloroplast genome of the plant *Stahlianthus Involucratus* (Zingiberaceae). *Mitochondrial DNA Part B Resou.* 4, 2702–2703. doi: 10.1080/23802359.2019.1644227

Li, D. M., Zhu, G. F., Xu, Y. C., Ye, Y. J., and Liu, J. M. (2020). Complete chloroplast genomes of three medicinal *alpinia* Species: Genome organization, comparative analyses and phylogenetic relationships in family zingiberaceae. *Plants-Basel* 9:286. doi: 10.3390/plants9020286

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, J., Zhang, D., and Gao, L. (2001). The overview research of chinese radix curcumae. *Nei Mongol J. Tradit. Chinese Med.* 1, 37–38.

Librado, P., and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25, 1451–1452. doi: 10.1093/bioinformatics/btp187

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

Mattick, J. S. (1994). Introns: evolution and function. *Curr. Opin. Genet. Dev.* 4, 823–831. doi: 10.1016/0959-437x(94)90066-3

McKinnon, G. E., Steane, D. A., Potts, B. M., and Vaillancourt, R. E. (1999). Incongruence between chloroplast and species phylogenies in *Eucalyptus subgenus* Monocalyptus (Myrtaceae). *Am. J. Bot.* 86, 1038–1046. doi: 10.2307/2656621

Meng, C., and Kubatko, L. S. (2009). Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. *Theor. Population Biol.* 75, 35–45. doi: 10.1016/j.tpb.2008.10.004

Neuhaus, H., and Emes, M. (2000). Nonphotosynthetic metabolism in plastids. *Annu. Rev. Plant Biol.* 51, 111–140.

Niu, Z., Xue, Q., Zhu, S., Sun, J., Liu, W., and Ding, X. (2017). The complete plastome sequences of four orchid species: insights into the evolution of the Orchidaceae and the utility of plastomic mutational hotspots. *Front. Plant Sci.* 8:715. doi: 10.3389/fpls.2017.00715

Palmer, J. D., Jansen, R. K., Michaels, H. J., Chase, M. W., and Manhart, J. R. (1988). Chloroplast DNA variation and plant phylogeny. *Ann. Missouri Bot. Garden* 75, 1180–1206.

Park, H., Kim, C., Lee, Y. M., and Kim, J. H. (2016). Development of chloroplast microsatellite markers for the endangered *Maianthemum bicolor* (Asparagaceae sl). *Appl. Plant Sci.* 4:1600032. doi: 10.3732/apps.1600032

Patel, R. K., and Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One* 7:e30619. doi: 10.1371/journal.pone.0030619

Piot, A., Hackel, J., Christin, P.-A., and Besnard, G. (2018). One-third of the plastid genes evolved under positive selection in PACMAD grasses. *Planta* 247, 255–266. doi: 10.1007/s00425-017-2781-x

Posada, D., and Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: advantages of the AIC and Bayesian approaches over likelihood ratio tests. *Syst. Biol.* 53, 793–808. doi: 10.1080/10635150490522304

Powell, W., Morgante, M., McDevitt, R., Vendramin, G., and Rafalski, J. (1995). Polymorphic simple sequence repeat regions in chloroplast genomes: applications to the population genetics of pines. *Proc. Natl. Acad. Sci. U.S.A.* 92, 7759–7763. doi: 10.1073/pnas.92.17.7759

Rieseberg, L. H., and Soltis, D. (1991). Phylogenetic consequences of cytoplasmic gene flow in plants. *Evol. Trends Plants* 5, 65–84.

Rogalski, M., do Nascimento Vieira, L., Fraga, H. P., and Guerra, M. P. (2015). Plastid genomics in horticultural species: importance and applications for plant population genetics, evolution, and biotechnology. *Front. Plant Sci.* 6:586. doi: 10.3389/fpls.2015.00586

Ronquist, F., and Huelsenbeck, J. P. (2003). MrBayes 3: bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572–1574. doi: 10.1093/bioinformatics/btg180

Rozas, J., Ferrer-Mata, A., Sanchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., et al. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol. Biol. Evol.* 1, 3299–3302. doi: 10.1093/molbev/msx248

Rozas, J., Sánchez-DelBarrio, J. C., Messeguer, X., and Rozas, R. (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19, 2496–2497. doi: 10.1093/bioinformatics/btg359

Ruhlman, T. A., Zhang, J., Blazier, J. C., Sabir, J. S., and Jansen, R. K. (2017). Recombination-dependent replication and gene conversion homogenize repeat sequences and diversify plastid genome structure. *Am. J. Bot.* 104, 559–572. doi: 10.3732/ajb.1600453

Saarela, J. M., Burke, S. V., Wysocki, W. P., Barrett, M. D., Clark, L. G., Craine, J. M., et al. (2018). A 250 plastome phylogeny of the grass family (Poaceae): topological support under different data partitions. *PeerJ* 6:e4299. doi: 10.7717/peerj.4299

Shaw, J., Shafer, H. L., Leonard, O. R., Kovach, M. J., Schorr, M., and Morris, A. B. (2014). Chloroplast DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in angiosperms: the tortoise and the hare IV. *Am. J. Bot.* 101, 1987–2004. doi: 10.3732/ajb.1400398

Soltis, D. E., Soltis, P. S., Chase, M. W., Mort, M. E., Albach, D. C., Zanis, M., et al. (2000). Angiosperm phylogeny inferred from 18S rDNA, rbcL, and atpB sequences. *Bot. J. Linn. Soc.* 133, 381–461. doi: 10.1006/bojl.2000.0380

Swofford, D., and Swofford, D. (2002). *PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4.0 a149.*

Wicke, S., Schneeweiss, G. M., Depamphilis, C. W., Müller, K. F., and Quandt, D. (2011). The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Mol. Biol.* 76, 273–297. doi: 10.1007/s11103-011-9762-4

Williams, A. M., Friso, G., van Wijk, K. J., and Sloan, D. B. (2019). Extreme variation in rates of evolution in the plastid Clp protease complex. *Plant J.* 98, 243–259. doi: 10.1111/tpj.14208

Wood, T. H., Whitten, W. M., and Williams, N. H. (2000). Phylogeny of Hedychium and Related genera(Zingiberaceae) Based on ITS sequence data. *Edinburgh J. Bot.* 57, 261–270. doi: 10.1017/s0960428600000196

Wu, M., Li, Q., Hu, Z., Li, X., and Chen, S. (2017). The complete Amomum kravanh chloroplast genome sequence and phylogenetic analysis of the commelinids. *Molecules* 22:1875. doi: 10.3390/molecules22111875

Wyman, S. K., Jansen, R. K., and Boore, J. L. (2004). Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20, 3252–3255. doi: 10.1093/bioinformatics/bth352

Xia, X. (2013). DAMBE5: a comprehensive software package for data analysis in molecular biology and evolution. *Mol. Biol. Evol.* 30, 1720–1728. doi: 10.1093/molbev/mst064

Xiang, B., Li, X., Qian, J., Wang, L., Ma, L., Tian, X., et al. (2016). The complete chloroplast genome sequence of the medicinal plant Swertia mussotii using the PacBio RS II platform. *Molecules* 21:1029. doi: 10.3390/molecules21081029

Yan, Y. D., Li, X. Y., Worth, J. R., Lin, X. Y., Ruhsam, M., Chen, L., et al. (2019). Development of chloroplast microsatellite markers for *Glyptostrobus pensilis* (Cupressaceae). *Appl. Plant Sci.* 7:e11277. doi: 10.1002/aps3.11277

Yang, J.-B., Tang, M., Li, H.-T., Zhang, Z.-R., and Li, D.-Z. (2013). Complete chloroplast genome of the genus Cymbidium: lights into the species identification, phylogenetic implications and population genetic analyses. *BMC Evol. Biol.* 13:84. doi: 10.1186/1471-2148-13-84

Ye, X.-B., Chen, J., and Liu, N. (2008). Curcuma nankunshanensis (Zingiberaceae), a new species from China. *J. Trop. Subtrop. Bot.* 16, 472–476.

Zhang, X., Zhou, T., Kanwal, N., Zhao, Y., Bai, G., and Zhao, G. (2017). Completion of eight Gynostemma BL.(Cucurbitaceae) chloroplast genomes: characterization, comparative analysis, and phylogenetic relationships. *Front. Plant Sci.* 8:1583. doi: 10.3389/fpls.2017.01583

Zhang, Y., Du, L., Liu, A., Chen, J., Wu, L., Hu, W., et al. (2016). The complete chloroplast genome sequences of five Epimedium species: lights into phylogenetic and taxonomic analyses. *Front. Plant Sci.* 7:306. doi: 10.3389/fpls.2016.00306

Zhou, J., Chen, X., Cui, Y., Sun, W., Li, Y., Wang, Y., et al. (2017). Molecular structure and phylogenetic analyses of complete chloroplast genomes of two Aristolochia medicinal species. *Int. J. Mol. Sci.* 18:1839. doi: 10.3390/ijms18091839

Zhou, Y., Nie, J., Xiao, L., Hu, Z., and Wang, B. (2018). Comparative chloroplast genome analysis of rhubarb botanical origins and the development of specific identification markers. *Molecules* 23:2811. doi: 10.3390/molecules23112811

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.

Copyright © 2020 Liang, Zhang, Deng, Gao, Ding, Zhang and Yang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 23 July 2020
doi: 10.3389/fgene.2020.00774

Molecular Characterization, Expression Pattern and Function Analysis of Glycine-Rich Protein Genes Under Stresses in Chinese Cabbage (*Brassica rapa* L. ssp. *pekinensis*)

*Xiaonan Lu, Yaxiong Cheng, Ming Gao, Meilan Li and Xiaoyong Xu**

College of Horticulture, Shanxi Agricultural University; and Collaborative Innovation Center for Improving Quality and Increasing Profits of Protected Vegetables in Shanxi, Taigu, China

OPEN ACCESS

Edited by:
Xiaoming Song,
North China University of Science and Technology, China

Reviewed by:
Tongkun Liu,
Nanjing Agricultural University, China
Md Sariful Islam,
Agricultural Research Service (USDA), United States
Jun Tang,
Jiangsu Academy of Agricultural Sciences, China

***Correspondence:**
Xiaoyong Xu
xuxy7926@163.com

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

Received: *02 April 2020*
Accepted: *30 June 2020*
Published: *23 July 2020*

Citation:
Lu X, Cheng Y, Gao M, Li M and Xu X (2020) Molecular Characterization, Expression Pattern and Function Analysis of Glycine-Rich Protein Genes Under Stresses in Chinese Cabbage (Brassica rapa L. ssp. pekinensis). Front. Genet. 11:774. doi: 10.3389/fgene.2020.00774

Plant Glycine-rich proteins (GRP), a superfamily with a glycine-rich domain, play an important role in various stresses such as high or low temperature stress and drought stress. *GRP* genes have been studied in many plants, but seldom in Chinese cabbage (*Brassica rapa* L. ssp. *pekinensis*). In this study, a total of 64 GRP genes were identified in Chinese cabbage by homology comparative analysis. The physical and chemical characteristics predicted by ProtParam tool revealed that 62.5% of BrGRPs were alkaline, 53.1% were stable, and 79.7% were hydrophilic. Conserved domain analysis by MEME and TBtools showed that 64 BrGRPs contained 20 of the same conserved motifs, based on which BrGRPs were classified into five main classes and four subclasses in class IV to clarify their evolutionary relationship. Our results demonstrated that The *BrGRP* genes were located on ten chromosomes and in three different subgenomes of Chinese cabbage, and 43 pairs of orthologous *GRP* genes were found between Chinese cabbage and *Arabidopsis*. According to the transcriptome data, 64 *BrGRP* genes showed abnormal expression under high temperature stress, 52 under low temperature stress, 39 under drought stress, and 23 responses to soft rot. A large number of stress-related *cis*-acting elements, such as DRE, MYC, MYB, and ABRE were found in their promoter regions by PlantCare, which corresponded with differential expressions. Two *BrGRP* genes-*w546* (*Bra030284*) and *w1409* (*Bra014000*), both belonging to the subfamily Subclass IVa *RBP-GRP* (RNA binding protein-glycine rich protein), were up-regulated under 150 mmol·L^{-1} NaCl stress in Chinese cabbage. However, the overexpressed *w546* gene could significantly inhibit seed germination, while *w1409* significantly accelerated seed germination under 100 mmol·L^{-1} NaCl or 300 mmol·L^{-1} mannitol stresses. In short, most *BrGRP* genes showed abnormal expression under adversity stress, and some were involved in multiple stress responses, suggesting a potential capacity to resist multiple biotic and abiotic stresses, which is

worthy of further study. Our study provides a systematic investigation of the molecular characteristics and expression patterns of *BrGRP* genes and promotes for further work on improving stress resistance of Chinese cabbage.

Keywords: Chinese cabbage, *GRP* gene family, expression analysis, biotic and abiotic stress, overexpression, *RBP-GRP*, resistance

INTRODUCTION

Plant glycine-rich protein (GRP) is a class of proteins consisting of glycine-rich repeat sequences. The first glycine-rich cell wall protein, PtGRP1, was isolated from *Petunia hybrida* in 1986 (Condit and Meagher, 1986; Chen et al., 2020) and similar proteins have since been found in almost all plants, such as *Zea mays*, *Oryza sativa*, *Arabidopsis thaliana*, and *Nicotiana tabacum* (Mangeon et al., 2009).

Plant GRPs can be divided into five main classes based on their primary structure, the conserved domain, and the arrangement of glycine repeats (Fusaro et al., 2001; Bocca et al., 2005). The class I family contains a typical structural feature of the GRP protein in which the glycine-rich repeats mostly appear as (Gly)n-X, where n is generally an odd number and X can be any amino acid. At the same time, the N-terminus contains or does not contain a signal peptide sequence. For example, PtGRP1 has 67% of glycine in glycine repeat (Gly)n-X (Mangeon et al., 2009). Class II of GRPs are similar to class I at the N-terminus, and contains a cysteine-rich polypeptide chain at the end of the C-terminal glycine repeat region that may play an important role in the pathogen-related process, such as *AtGRP3* (Park et al., 2001). The glycine content of class III is lower than that of the first two classes, but the oleosin domain is the unique motif for this class. For example, oleosin-GRPs have an oleosin conserved area and are located in an oil-rich cell structure of tapetum, which mainly plays a role in stabilizing the triglyceride and phospholipid bilayer in the membrane structure (Ferreira et al., 1997; Murphy et al., 2001). Class IV GRPs, also named RNA-binding proteins (RBPs), have no signal peptide at their N-terminus. These proteins have other structures besides the glycine-rich sequences, such as the RNA-recognition motif (RRM), cold-shock domain (CSD), and CCHC (CCHC = C-X_2-C-X_4-H-X_4-C; C = Cys, H = His, X = variable amino acid) zinc finger structure. Class IV can be divided into four subgroups based on its domains: Subclass IVa (single RRM), Subclass IVb (single RRM and CCHC zinc-finger motif), Subclass IVc (cold shock domain and two or even more zinc-fingers), and Subclass IVd (two RRM motifs) (Mangeon et al., 2010). These RNA recognition domains can usually recognize each other or interact with proteins without any RNA recognition domain to bind to their target (Dreyfuss et al., 2002; Lorkovic and Barta, 2002). Class V GRPs are similar to the members of class III as the glycine repeats are arranged differently and presented with the mixed mode of (GGX)n and (GXGX)n (Bocca et al., 2005).

Plant *GRP* genes are often specifically expressed and play different roles in different development stages and tissues (Czolpinska and Rurek, 2018). The GRP proteins of classes I and II are active components of cell walls, which play a positive regulatory role in plant cell division and organ differentiation (Czolpinska and Rurek, 2018). The AtGRPs of class II can interact with the cell wall-associated receptor protein kinase AtWak1 to participate in the signal transduction to prevent viruses from invading plants (Park et al., 2001). The GRPs of class III regulate pollen development and hydration, while class IV GRPs with an RNA recognition domain and are involved in molecular processes such as alternative splicing or transcriptional regulation; the latter also play an important role in stomatal regulation and seed and stamen development (Winter et al., 2007; Czolpinska and Rurek, 2018). An oleosin-*GRP BrGRP17* gene from Chinese cabbage showed higher expression levels in the flower buds of male fertile plants than in sterile ones (Xu et al., 2013). The deletion mutant of RNA-binding glycine-rich protein five (*Atrbg5*) also yielded shorter roots, smaller leaves, and shorter flower axes, but overexpressed *AtRBG5* promoted cell elongation and tissue growth in *Arabidopsis* (Mangeon et al., 2009). In summary, the function of the plant *GRP* gene family varies across plant growth and development.

Besides that, plant GRPs are also involved in responses to various abiotic stress such as salt, drought, and temperature, and may play an important role in resisting adversity stress. A notable increase of *RB-GRP* (*RNA-Binding GRP*) was observed in *S. Bicolor* seedlings subjected to NaCl treatments with 1 M and 500 mM (Aneeta et al., 2002). Overexpressing *AtRZ-1a*, a zinc finger-containing GRP, lead to retarded germination and seedling growth under salt or dehydration stress conditions in transgenic *Arabidopsis* (Kim et al., 2005) while the loss-of-function mutants of *AtRZ-1a* germinated earlier and grew faster than the wild-type plants under the same conditions (Kim Y.O. et al., 2007). *NtGRP-1a* was up-regulated under drought stress and could be maintained for 3 to 6 days (Chen et al., 2010). The ryegrass *LPGRP1* gene was up-regulated under cold stress treatment (Shinozuka et al., 2006). The expression level of the *AtRBG7* gene was higher at low temperatures than high ones (32°C) (Wienkoop et al., 2008). Overexpression of the *AtRBG7* gene increased cold tolerance but inhibited seed germination and plant growth under drought stress (Kim et al., 2008), while overexpression of *AtRBG2* also increased cold tolerance and lead to a higher germination rate under salt stress in *Arabidopsis* (Kim J.Y. et al., 2007). Ectopic expression of *AtRBG2* and *AtRBG7* in rice could also increase crop yield under drought stress (Yang et al., 2014). In other words, different *GRP* genes may play different functions in plant responses to various stresses, and expression dynamics may vary under the same stress.

Up to date, the *GRP* gene family has been studied in many plants (Kar et al., 2012; Zhang et al., 2014; Krishnamurthy et al., 2015; Lu et al., 2019), however, the *BrGRP* gene family has not been reported on Chinese cabbage. In this study, 64 *GRP* genes in Chinese cabbage were identified based on the

existing BRAD genome database[1] (Cheng et al., 2011) and their phyletic evolution, module prediction, and chromosomal localization were further investigated. Moreover, the expression patterns of these *BrGRP* genes were also detected in different tissues and various abiotic stresses based on open transcriptome databases, their functions, and evolutions in development and stress response were also discussed. Furthermore, the function of two *BrGRP* genes screened from the normalized cDNA library of male sterile bud from Chinese cabbage flower (Li et al., 2018) were further characterized under salt treatment. These results provide valuable information for further exploration into the function of *BrGRP* genes in Chinese cabbage.

MATERIALS AND METHODS

Identification and Sequence Conservation of *GRP* Genes in Chinese Cabbage

The *GRP* genes in *Arabidopsis thaliana* (TAIR database[2]) (Lamesch et al., 2012) were employed as a query to search against Chinese cabbage genome database (BRAD v1.5[1]) (Cheng et al., 2011). A total of 64 genes in Chinese cabbage genome were identified as possible members of *BrGRP* genefamily using the Blastn program. All putative protein sequences of *GRP* genes in *Arabidopsis thaliana* and Chinese cabbage were identified for the GRP conserved domain using the NCBI Conserved Domain Database[3], and Pfam[4]. The sequences without glycine-rich protein domains were removed. Finally, the nucleotide and deduced amino acid sequences of *BrGRP* genes were confirmed for further analysis.

Characterization Analysis and Subcellular Localization Prediction of BrGRPs

Physical and chemical characteristics of the BrGRP sequences – i.e., the molecular weight (MW), theoretical point (pI), instability index, aliphatic index, and grand average of hydropathicity (GRAVY) – were further analyzed using the ProtParam tool in ExPASy[5]. The subcellular localization of these BrGRPs was predicted by the ProtComp tool on Softberry[6].

Multiple Sequence Alignment and Phylogenetic Analysis

Multiple sequence alignments of the published protein sequences were performed by Clustal X with default parameters, including 13 OsGRPs (Krishnamurthy et al., 2015), 18 MaGRPs (Zhang et al., 2014; Krishnamurthy et al., 2015), 32 GrGRPs (Yang et al., 2019), 37 GaGRPs (Yang et al., 2019) and 9 ItGRPs (Lu et al., 2019) from rice[7], maize[8], *Gossypium raimondii*[9], *Gossypium arboreum*[9], and sweet potato[10], respectively. The phylogenetic tree was constructed by the Neighbor-Joining method (NJ) and Maximum likelihood (ML) on MEGA X (Kumar et al., 2018) and the check parameter bootstrap value was set to 1000 times.

Analysis of the Conserved Domain and Gene Structure of *BrGRP* Genes

The structures of the coding/non-coding region of *BrGRP* genes were mapped by the software TBtools (Chen et al., 2020). Next, the conservative motifs were analyzed by MEME (version 5.0.3[11]), with the number of motifs set to 20 and the other parameters set to default values. The LOGO of conservative motifs was listed, and TBtools was used to export the corresponding Scalable Vector Graphics (SVGs) (Chen et al., 2020).

Identification of the Orthologous *BrGRP* Genes and Syntenic Analysis in Chinese Cabbage

According to the genomic and chromosome database (v2.5) of Chinese cabbage (Cheng et al., 2011, 2013) the identified *BrGRP* genes were located on ten chromosomes in three fractionated subgenomes, and the locations of the *BrGRP* genes on chromosome were visualized using MapChart (Voorrips, 2002). Syntenic relationships between BrGRP homologs of Chinese cabbage and *Arabidopsis thaliana* was defined in BRAD database (Cheng et al., 2011) and the corresponding circos were drawn out on TBtools (Chen et al., 2020).

Expression Pattern of *BrGRP* Genes

To analyze the expression pattern of *BrGRP* genes in Chinese cabbage, transcriptome data from *B. rapa* "Chiifu" (Tong et al., 2013) and the inbred line "Fushanbaotou," a typical heading Chinese cabbage (Wang et al., 2012) was used for gene expression profiling in eight tissues: callus, root, stem, leaf, flower, silique, rosette, and folding leaves.

The differences in *BrGRP* genes expression under biotic and abiotic stress were also analyzed based on the transcriptome data of Chinese cabbage under high temperature at 45°C (Dong et al., 2015), low temperature at 4°C (Zhang et al., 2016), drought (Guo et al., 2017), and soft rot stress (Liu et al., 2019). The expression levels of *BrGRP* genes were calculated with Fragments Per kb per Million read (FPKM) values (Mortazavi et al., 2008), and analyzed by STEM (Short Time-series Expression Miner) (Ernst and Bar-Joseph, 2006). Venn diagram and heat map were generated by TBtools (Chen et al., 2020) according to the differentially expressed data.

Analysis of *Cis*-Acting Elements on *BrGRPs* Promoter

To further identify the *cis*-acting elements on the promoter regions of *BrGRP* genes, a 2-kb fragment upstream of

[1]http://brassicadb.org/brad/
[2]http://www.arabidopsis.org/
[3]http://www.ncbi.nlm.nih.gov/cdd
[4]http://pfam.xfam.org/
[5]http://web.expasy.org/protparam/
[6]http://linux1.softberry.com/
[7]http://rice.plantbiology.msu.edu/
[8]https://maizegdb.org/
[9]https://www.cottongen.org/
[10]http://sweetpotato-garden.kazusa.or.jp/
[11]http://meme-suite.org/tools/meme

the start codon was extracted by TBtools (Chen et al., 2020) and further identified by PlantCare (Lescot et al., 2002)[12] with the adversity related *cis*-acting elements MYC (CANNTG), MYB (C/TAACNA/G), ABRE (ABA-responsive element, ACGT), LTR (low-temperature-responsive element, CCG AAA), DRE (CCGAC), W box (TTGACC), and TC rich repeats (GTTTTCTTAC). A Venn diagram was constructed using TBtools (Chen et al., 2020) based on the types of *cis*-acting element.

Plant Material and qRT-PCR

The plump seeds of heading Chinese cabbage 18c901 (a homozygous inbred line) were planted in a tray filled with substrate and maintained in an intelligent chamber at 25°C, 2000 l× of light intensity, 16/8 h of light-dark period of, and 70% humidity. Plants with flower buds were treated with 150 mmol·L^{-1} NaCl solution. After 7 days' treatment, the leaves, the whole flowers and roots were collected and quickly frozen by liquid nitrogen for quantitative PCR analysis.

Total RNA was extracted using Total RNA Kit (TIANGEN, China). All RNA was analyzed by agarose gel electrophoresis and then quantified with a Nanodrop ND-1000 spectrophotometer. DNA-free RNA was used to synthesize the first strand of cDNA using PrimeScript^{TM} RT Master Mix (Perfect Real Time) (Takara, Japan). The quantitative RT-PCR was run on ABI 7500 system using SYBR Green PreMix (Takara, Japan). The Chinese cabbage GAPDH (*AF536826.1*) was used as an internal control. The reactions were carried out by the following program: 94°C for 5 min; followed by 40 cycles of 94°C for 30 s, 55°C for 30 s, 72°C for 30 s; and 72°C for 5 min. Each reaction was performed in biological triplicates, and the data from real-time PCR amplification were analyzed using the $2^{-\Delta\Delta CT}$ method. The primers sequences were shown in **Supplementary Table S1**.

Generation of the Transgenic *Arabidopsis* Plants

The overexpression vector PHZM27-*w546*/*w1409* was constructed by inserting *w546*/*w1409* under the CaMV35s promoter and NOS terminator of PHZM27, and then transferred into the *Agrobacterium tumefaciens* strain GV3101 by electroporation. The flower-dip method was applied to transform *Arabidopsis* (Columbia-0, WT) according to the protocol described by Clough and Bent (1998). Transgenic *Arabidopsis* plants were determined on MS medium with by kanamycin screening assay (Park et al., 1998) and PCR analysis with 35S forward and *w546*/*w1409*-specific reverse primers (**Supplementary Table S1**).

NaCl and Mannitol Stress in Transgenic *Arabidopsis* Plants

The homozygous seeds of T_3 generation transgenic *Arabidopsis* were inoculated on 1/2 MS medium with 100 mmol·L^{-1} NaCl and 300 mmol·L^{-1} mannitol, respectively. The number of germination individuals were calculated every day. Each treatment contained 30 individuals and repeated three times. The cultivation environment was set as follows: temperature 22°C, light intensity 2000 l×, light-dark cycle 16/8 h, and humidity about 70%.

[12]http://bioinformatics.psb.ugent.be/webtools/plantcare/html/

RESULTS

Identification and Characterization of *BrGRP* Genes in Chinese Cabbage

To identify the *BrGRP* genes in Chinese cabbage, we first screened the *GRP* genes of *Arabidopsis* in TAIR and NCBI databases. A total of 41 *AtGRP* genes were obtained, of which 10 had no orthologous genes; a total of 64 *BrGRP* genes in BRAD database were identified from the remaining 31 *AtGRP* genes (**Supplementary Table S2**). Further prediction analysis revealed that the protein characteristics of these *BrGRP* genes showed great differences in molecular weight, theoretical point, etc. The length of amino acids ranged from 52 (Bra027145) to 2038 aa (Bra028693), and the theoretical point ranged from 4.59 (Bra008020) to 11.56 (Bra028691), with 62.5% BrGRPs with a theoretical point over 7. More detailed information including instability index, aliphatic index, and grand average of hydropathicity were also predicted. The results demonstrated that 46.9% of proteins showed an aliphatic index over 40, while 79.7% proteins showed a grand average of hydropathicity with a negative value. All these results suggested that these BrGRPs mainly existed on the stable alkaline hydrophilic protein.

Subcellular localization showed that most BrGRPs (36 of 65) were secreted into the extracellular matrix; 18 BrGRPs were localized in nuclear region, six in the mitochondria, two BrGRPs including Bra037177 and Bra031809 in the cytoplasm, and Bra031159 and Bra005945 in plasma membrane and chloroplast, respectively (**Supplementary Table S2**).

Sequences Analysis of BrGRP Genes and Phylogenetic Relationship

To gain insight into conserved domain of these BrGRPs, a total of 64 BrGRP genes showed similar conserved domains and 20 motifs were identified using MEME tool, including two glycine-rich motifs (motif 6, 7), two cysteine-rich domains (motif 9, 14), two RNA recognition motifs (motif 1, 5), one CCHC-Zinc finger structural motifs (motif 4), one cold shock motif (CSD) (motif 10), one oleosin lipid motif (motif 3), and one signal peptide motif (motif 8) (**Figure 1C** and **Supplementary Figure S1**). According to motif types, 64 BrGRPs were constructed a phylogenetic tree using the NJ method (**Figure 1A**) and were divided into five classes (**Figure 1C**). Class I contained six members, the N-termini with a typical glycine repeat structure followed by a signal peptide sequence or no signal peptide sequence. Class II was composed of 12 members, which comprised a cysteine-rich polypeptide chain at the end of C-terminal with a repeated glycine. Eleven BrGRPs were classed into Class III, whose glycine content was lower than that of the first two classes, and contained an oleosin domain. Class V (five members) had the lowest glycine content, and contained repeats of the pattern GGX/GX (X represents

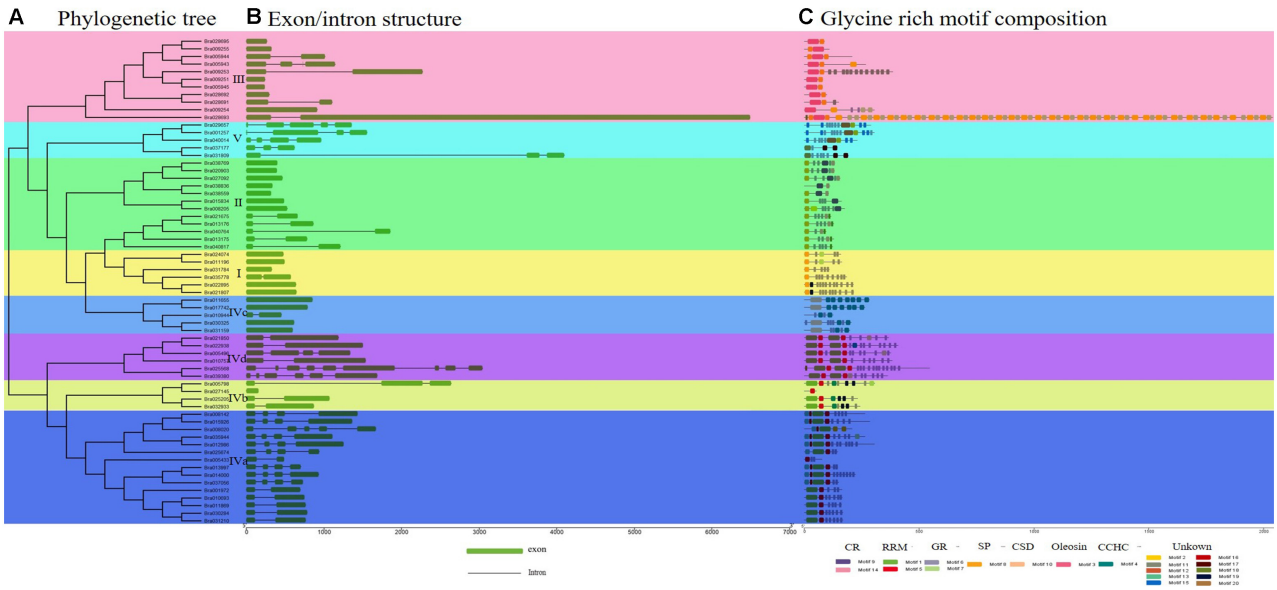


FIGURE 1 | Phylogenetic relationship, gene structure, and conserved structural composition of 64 *BrGRP* genes in Chinese cabbage. **(A)** Phylogenetic tree created using Clustal X to align the amino acid sequences and MEGA X to generate the phylogenetic tree by a contiguous method with a calibration parameter of 1000. **(B)** Sequence structure distribution of *BrGRP* genes; the green and black boxes represent the exon and intron, respectively. Scale indicates 1.0 kb. **(C)** Schematic diagram of the conserved *GRP* protein motif in Chinese cabbage predicted by MEME. Colored boxes indicate different motifs; CSD, the cold-shock domain; RRM, RNA-recognition motif; oleosin, oleosin-conserved domain; CR, cysteine-rich domain; CCHC, zinc-finger; GR, glycine-rich domain. Scale bars represent 500 aa.

any amino acid), which coincided with class V of *Arabidopsis* demonstrated by Gilberto Sachetto (personal communicated).

In addition to the glycine-rich region, class IV (30 members) also had other structures, including one RNA recognition domain, cold shock domain, and CCHC-zinc finger domain. Based on these different conserved domains, class IV can be further divided into four subgroups (Subclass IVa, IVb, IVc, and IVd). Subclass IVa (15 members) contained a single RRM motif, while Subclass IVb (four members) included a single RRM and CCHC zinc-finger motif. Subclass IVc (five members) contained a cold shock domain and two or more zinc-fingers. Subclass IVd (six members) contained two RRM motifs. These five subclasses in *BrGRPs* were highly similar to those of previous studies in *Arabidopsis*, Rice and Maize (Mangeon et al., 2010; Krishnamurthy et al., 2015; Czolpinska and Rurek, 2018). The genes in the same subgroup shared a close phylogenetic relationship, high sequence similarity, and similar gene structures, revealing evolutionary conservation in the *GRP* gene family.

The distribution of the exon-intron structure in the 64 *GRP* genomic sequences was exhibited as **Figure 1B**. Twenty-three *BrGRP* genes had no introns, while the remained 41 *BrGRP* genes contained two or more exons, of which *Bra025568* contained nine exons, the most. For exon numbers, class I and II of *BrGRPs* had the least exons (one or two), classes III and V had 1–3 and 3–4 exons, respectively, while class IV appeared variety ranging from1 to 9 exons.

The phylogenetic relationships of 64 BrGRPs, 41 AtGRPsand other reported GRPs from cotton, rice, maize, and sweet potato were evaluated using NJ and ML method (**Figure 1A** and **Supplementary Figure S2**). The topological structure of the GRP gene phylogenetic tree constructed by the NJ method (**Supplementary Figure S2A**) and the ML method (**Supplementary Figure S2B**) is basically the same, and both have a high degree of confidence. The results showed that these BrGRPs were also classified into five classes (class I-V). Most GRPs belonged to class IV (30 BrGRPs, 18 AtGRPs 13 OsGRPs, 18 MaGRPs, 32 GrGRPs, 37 GaGRPs, and 9 ItGRPs), followed by class II (12 BrGRPs and 7 AtGRPs) and class III (11 BrGRPs and 7 AtGRPs), with several genes in class I (6 BrGRPs and 4 AtGRPs) and class V (5 BrGRPs and 5 AtGRPs). Interest, the reported GRPs from other species were all separated into the class IV which may suggest their most important function among five classes.

Chromosome Localization and Orthologous Gene Analysis of BrGRPs in Chinese Cabbage

To examine their chromosomal distributions, the 64 *BrGRP* genes were mapped onto the chromosomes and three fractionated subgenomes of Chinese cabbage based on the *B. rapa* genome database (chromosome v1.5). Three *BrGRP* genes (*Bra040764*, *Bra040817*, *Bra039380*) could not be assigned to any chromosome, but the other 61 genes were successfully identified in 10 chromosomes; 13 *BrGRP* genes on Chromosome A03, while only one on Chromosome A06. Furthermore, 23 *BrGRP* genes were anchored on the least fractionated (MF1) subgenome, 18 genes on the medium fractionated (MF2) subgenome, and 20 genes

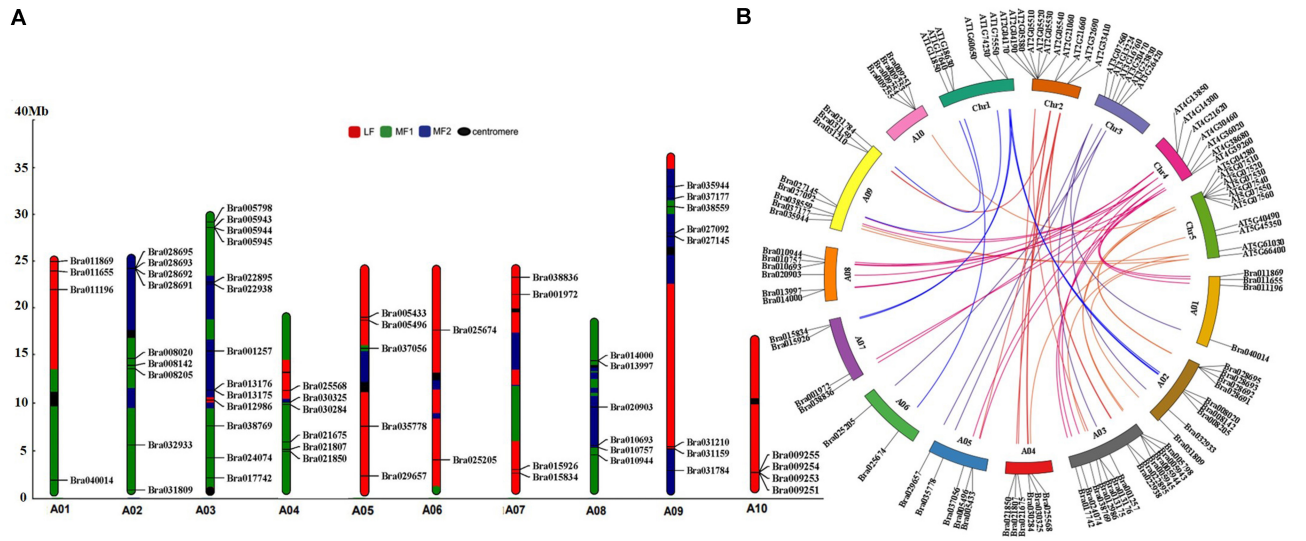


FIGURE 2 | Chromosome distribution and syntenic analysis of GRP genes in Chinese cabbage. **(A)** MF1, the least fractionated subgenome; MF2, the medium fractionated subgenome; LF, the most fractionated subgenome; centromere, the position of the chromosome centromere. Syntenic analysis of GRP genes in Chinese cabbage and *Arabidopsis thaliana*. **(B)** The orthologous and paralogous GRP genes are localized on the chromosomes of Chinese cabbage (A01–A10) and *Arabidopsis thaliana* (Chr1–Chr5).

on the most fractionated (LF) subgenome (**Figure 2A** and **Supplementary Table S3**).

Moreover, 43 pairs of *BrGRP* gene syntenic paralogs were found on different subgenomes of Chinese cabbage (**Figure 2A** and **Supplementary Tables S3,S4**). For example, *Bra022895* and *Bra021807* with the highest sequence similarities to *AT2G32690*, both located in the MF2 and MF1 subgenome, respectively. To further understand the duplication of the *BrGRP* genes during the whole genome duplication in Chinese cabbage, the orthologous analysis of GRP homologous genes was also compared between Chinese cabbage and *Arabidopsis thaliana* (**Figure 2B** and **Supplementary Table S4**). A total of 31 *AtGRP* genes were found to be orthologous in Chinese cabbage, and most of *GRP* genes in *Arabidopsis* had 1–6 orthologous genes in Chinese cabbage, but 10 *AtGRP* genes had no orthologous genes in Chinese cabbage. The results indicated that *BrGRP* genes had evolved during whole-genome duplication in Chinese cabbage, providing a valuable reference for uncovering the functions of *BrGRP* genes in Chinese cabbage.

Expression Profiling of *BrGRPs* in Chinese Cabbage

Based on the published transcriptomic data, the expression of *BrGRP* genes was further analyzed in Chinese cabbage in different tissues and during two development stages under various abiotic and biotic stresses, and a set of *BrGRP* genes was identified to be abnormally expressed.

Different Tissues

The tissue-specific expression profiling of 64 *BrGRP* genes were detected in different tissues based on the transcriptome data from the Chinese cabbage line "Chiifu" (Tong et al., 2013; **Figure 3A** and **Supplementary Table S5**). *BrGRP* gene expressions were different among the root, stem, leaf, flower, silique and callus of Chinese cabbage, but 11 genes were found no expression in any tissue. Among them, *Bra031210* had the highest expression in four tissues, including the root, leaf, stem, and silique, while *Bra038559* and *Bra040764* had the highest expression in the flower and callus, respectively. Thirty-six *BrGRP* genes all were expressed in six tissues, whereas 7 *BrGRP* genes were specifically expressed in one or few tissues (**Figure 3B**). In detail, *Bra008020* was only expressed in flower, *Bra038836* in flower and silique, and *Bra028693* in callus and root. *BrGRP* genes had various expression levels in different tissues, suggesting that they played various roles in organ development and other relevant biological processes.

In addition, *BrGRP* genes were also differentially expressed in two development stages. Ten *BrGRP* genes from classes I, II, IVa, or IVc, were found in the rosette and folding leaves; nine *BrGRP* genes (*Bra035944*, *Bra031784, Bra022895*, *Bra020903*, *Bra014000*, *Bra013176*, *Bra011655*, and *Bra008205*) showed higher expression levels in rosette leaves than in folding leaves, whereas *Bra030284* showed higher expression only in folding leaves (**Figure 3C** and **Supplementary Table S6**). Three *BrGRP* genes (*Bra035944*, *Bra014000* and *Bra030284*) from Subclass IVa – showed differential expression patterns in rosette and folding leaves. Therefore, the above results suggested that *BrGRP* genes might play different roles during the two development stages.

Temperature Stress

The expression levels of 64 *BrGRP* genes were compared in "Chiifu" seedlings at 0, 0.5, 1, 2, 3, and 4 h after 45°C, respectively

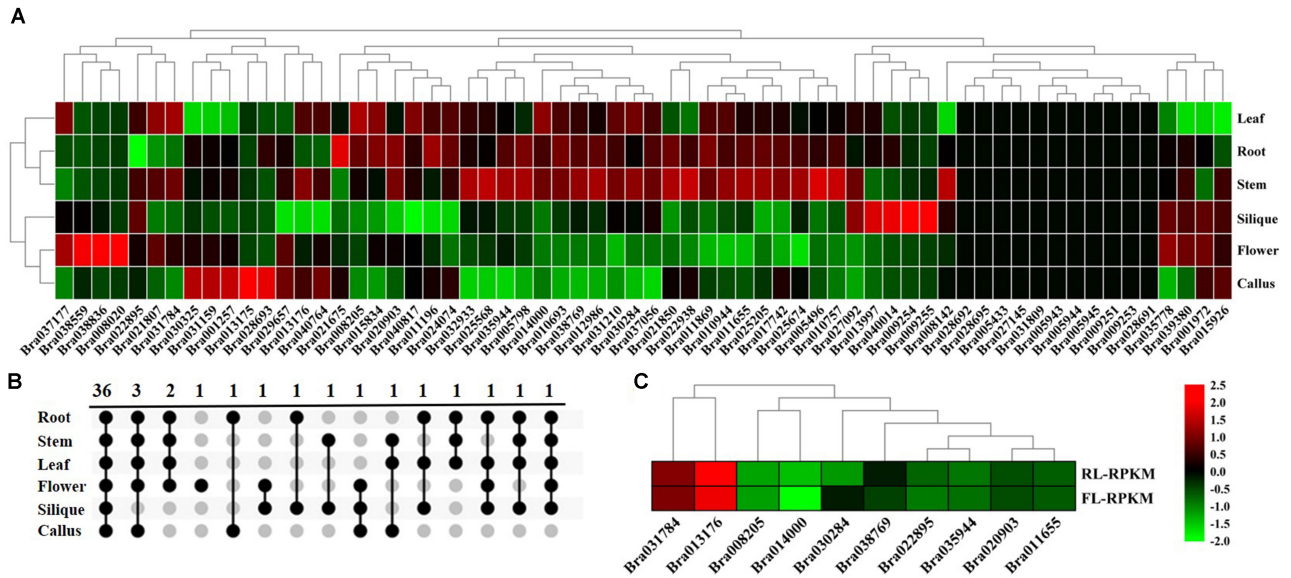


FIGURE 3 | Expression analysis of *BrGRP* gene expression in Chinese cabbage. **(A)** Differential expression pattern of the *BrGRP* gene in different tissues of Chinese cabbage. **(B)** Venn diagram of *BrGRP* expression in six tissues of Chinese cabbage. Numbers represent the number of genes. **(C)** Differential expression pattern of the *BrGRP* gene of Chinese cabbage in the rosette and folding leaves.

(Dong et al., 2015; **Figure 4A** and **Supplementary Table S7**), and were divided into 31 expression patterns by STEM software (**Supplementary Table S8**). Among them, seven expression patterns (profiles 24, 39, 43, 44, 46, 47, and 49) with a total of 20 *BrGRP* genes were up-regulated at five time points under heat stress. Eight *BrGRP* genes in profile 49 were significantly up-regulated five times under heat stress. Fourteen *BrGRP* genes with four expression patterns (profiles 0, 1, 8, and 10) were down-regulated five times under heat stress. Another 20 expression patterns showed disordered fluctuating. For example, *Bra035944* of profile 15 was down-regulated at 0.5, 3, and 4 h, and up-regulated at 1 and 2 h of heat stress. *Bra025568* of profile 30 was down-regulated at 0.5, 2, and 4 h, and up-regulated at 1 and 3 h. *Bra027092* of profile 34 was down-regulated at 1 and 4 h, and up-regulated at 0.5, 2, and 3 h.

A total of 52 *BrGRP* genes were found to express in two true leaves based on the transcriptome data of 29-day-old leaves under 4°C (Zhang et al., 2016). Among them, the expression of 26 genes was significantly different with the low-temperature treatment than with the control treatment (25°C) [log2fold-change (FC)>1]: 17 genes were up-regulated and the remaining nine were down-regulated (**Figure 4B** and **Supplementary Table S9**).

Drought Stress

The root transcriptome data of Chinese cabbage (CR2355 and ATC92037) were also analyzed to study the expression pattern of *BrGRP* genes under drought stress (Guo et al., 2017). CR2355 is a drought-tolerant (DT) material that can keep the required biomass for mature plants when suffering from transient drought stress during the reproductive phase, while ATC92037 is a drought-sensitive (DS) material which showed a significant reduction in biomass after transient drought stress. After simulating drought stress with 2.5% PEG 6000 in seedlings, 39 *BrGRP* genes were found to differentially express at 4, 8, and 12 h between the DT and the DS materials, and then divided into 11 and 14 profiles, respectively (**Figure 4C** and **Supplementary Tables S10,S11**). Among them, for little difference in the expression levels among *BrGRP* genes, there were 6 genes (*Bra040764*, *Bra038769*, *Bra021850*, *Bra040817*, *Bra024074*, and *Bra022895*) – in ATC92037 (DS), while *Bra020903* and *Bra011196* in CR2355 (DT). Certainly, some genes showed the same expression pattern between DS and DT, e.g., nine *BrGRP* genes (*Bra001257*, *Bra021807*, *Bra010757*, *Bra031210*, *Bra015926*, *Bra013176*, *Bra035944*, *Bra011809*, and *Bra031784*) shared profiles 1, 4, 5, and 6. Profile 4 represented continuous down-regulation at 4, 8, and 12 h and profiles 1, 5, and 6 showed continuous down-regulation during the first 8 h of drought stress, after which it was up-regulated until 12 h. Five *BrGRP* genes (*Bra001972*, *Bra025205*, *Bra022938*, *Bra005798*, and *Bra030325*) showed different expressions between DT and DS materials. *Bra005798* was down-regulated or its expression level was not significantly changed in DS material, but it was up-regulated in DT material after drought treatment. *Bra011196* was up-regulated in DS plants but down-regulated or did not change in DT plants after drought treatment. The different expression patterns of these *BrGRP* genes in two materials under drought stress may play different roles in the response of drought tolerance in Chinese cabbage.

Soft Rot

To determine the expression of *BrGRP* genes response to soft rot, the transcriptome data from soft rot-resistant mutant (*sr*) and wild control (inbred line "A03")were used at 0, 6, 12, and

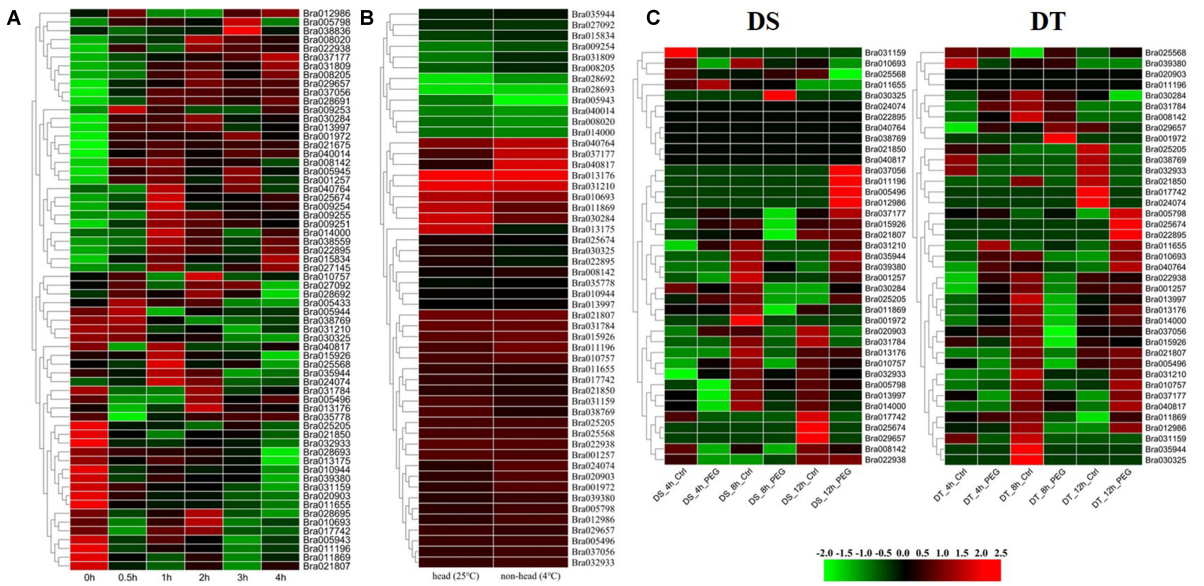


FIGURE 4 | Expression analysis of *BrGRP* genes in Chinese cabbage. **(A)** Differential expression pattern of *BrGRP* gene in Chinese cabbage under 45°C heat stress for 0.5, 1, 2, 3, and 4 h. **(B)** Differential expression pattern of the *BrGRP* gene in Chinese cabbage under low-temperature stress at 25 and 4°C. **(C)** Differential expression profiles of *BrGRP* genes in drought-sensitive (DS) and drought-tolerant (DT) cultivars after drought treatment 4, 8, and 12 h. Ctrl represents the control group without treatment; PEG represents the treatment group under drought stress.

24 h after inoculation with soft rot, respectively (Liu et al., 2019). As **Figure 5A** and **Supplementary Table S12** shown, a total of 23 *BrGRP* genes were identified, which were divided into 12 expression patterns (**Supplementary Table S13**). Most of *BrGRP* genes were down-regulated within 12 h after inoculation, and up-regulated at 12–24 h (profiles 14, 18, 21, and 25). Four *BrGRP* genes (*Bra011196*, *Bra013997*, *Bra035944*, and *Bra011655*) were up-regulated four times after inoculation in *sr* mutant. These results suggested that *BrGRP* genes had a special response to pathogens.

In summary, Venn diagram revealed the number of *BrGRP* genes involved in various stress responses (**Figure 5B** and **Supplementary Table S14**). Twenty *BrGRP* genes (e.g., *Bra025674*, *Bra011869*, and *Bra015926*) showed differential expressions under four types of stresses. Ten *BrGRP* genes (e.g., *Bra028693* and *Bra010944*) responded to heat and cold stresses at different expression levels. Only three *BrGRP* genes (*Bra035778*, *Bra028692*, and *Bra013175*) were induced simultaneously by r heat, cold and soft rot stresses. Furthermore, 12 *BrGRP* genes were only differentially expressed under heat stress. Meanwhile, we found that the expression patterns of 64 *BrGRP* genes were not completely consistent under different abiotic stresses. For example, *Bra030284* was up-regulated under both high- and low-temperature stresses, but showed up-regulated expression in drought-stressed DS plants and down-regulation in DT plants. *Bra010693* was up-regulated under low temperature and drought in DT plants, but down-regulated under high temperature and drought in DS plants. *Bra011869* was down-regulated in low temperature, high temperature, and drought in DS material, but up-regulated under drought in DT material and soft rot. The differential expression of these *BrGRP* genes under various stresses suggested their different functional dissimilation, but needs further validation.

Analysis of *Cis*-Acting Elements in the Promoter Region of *BrGRPs*

Most *BrGRP* genes were identified under various stresses, so the stress-related *cis*-acting elements were further analyzed. From **Figure 5C** and **Supplementary Table S15**), many stress-related *cis*-acting elements were found in the promoter regions of *BrGRP* genes, including DRE, MYB, MYC, ABRE, LTRE, W-box, and TC-rich repeats. DRE has been identified as a *cis*-acting element involved in drought, high salt, and low-temperature stresses, and MYB involved in drought, cold, and salt stresses (Dai et al., 2007). MYC is involved in drought and ABA (Onishi et al., 2006) and LTRE in low-temperature stress (Maestrini et al., 2009). W-box can be combined by WRKY transcription factors to participate in plant responses to stresses such as diseases, drought, and ABA (Wang et al., 2009). TC-rich repeats participate in plant defense and stress response (Banerjee et al., 2013).

The number of stress-related *cis*-acting elements in the promoter region of the 64 *BrGRP* genes was varied from four (*Bra032933*) to 34 (*Bra009254*) and the promoter of each *BrGRP* gene distributed 3–6 types of *cis*-acting elements. The promoter region of 98.4% (63/64) *BrGRP* genes contained MYC element and MYB element. ABRE element was regarded to involve in ABA and drought stress (Manavella et al., 2008), which was located on 89.1% (57/64) of *BrGRP* gene promoters. LTRE element accounted for 50% (32/64) *BrGRP* gene promoters, moreover, DRE, W-box, and TC-rich repeats for 31.3% (20/64), 46.9% (30/64), and 34.4% (22/64) of *BrGRP* gene promoters,

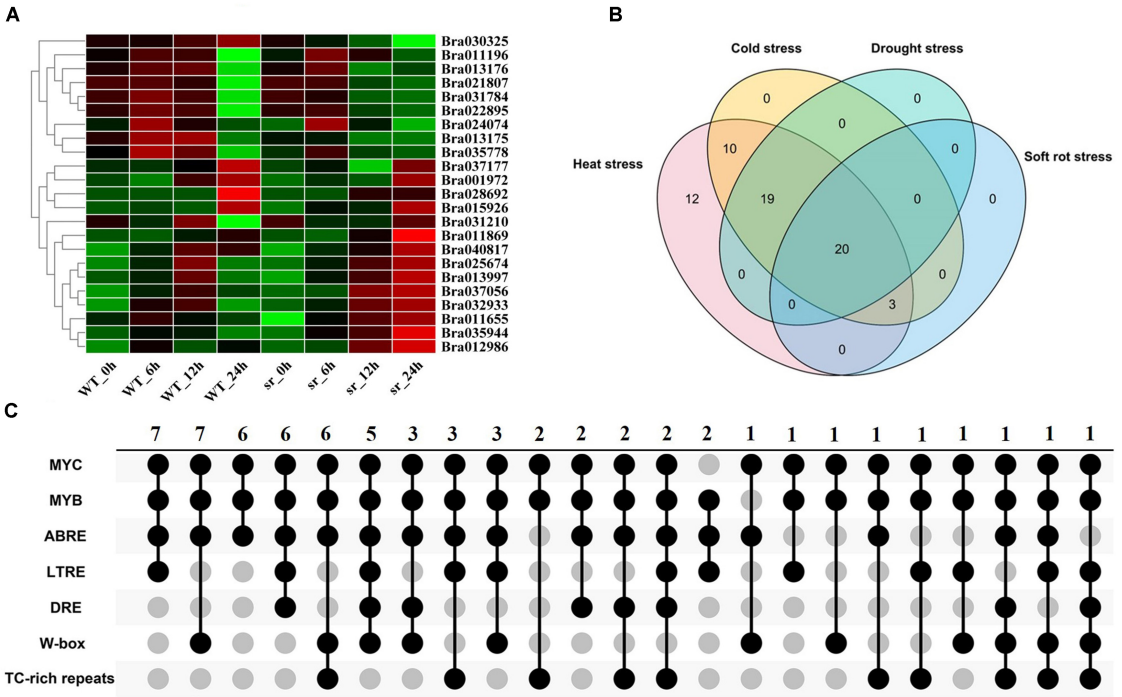


FIGURE 5 | Expression analysis of *BrGRPs* under soft rot stress **(A)**. The numbers of *BrGRPs* involved in various stresses showed by a Venn diagram **(B)**. The numbers of the stress-related *cis*-acting elements in the promoter regions of *BrGRPs* showed by the Venn diagram **(C)**. MYC (CANNTG), MYB (C/TAACNA/G), ABRE (ABA-responsive element, ACGT), LTR (low-temperature-responsive element, CCG AAA), DRE (CCGAC), W box (TTGACC), TC rich repeats (GTTTTCTTAC). Numbers represent the number of involved genes.

respectively. The *BrGRP* genes whose promoters contained the *cis*-acting elements played an important role in the response to adversity stresses.

The promoter regions of each *BrGRP* gene contained several types of stress-related *cis*-acting elements. For example, *Bra031159*, *Bra037056,* and five other genes only had ABRE, LTRE, MYB, and MYC. *Bra028691*, *Bra028692*, and five other genes only had ABRE, MYB, MYC, and W-box. Differences in *cis*-acting element types s on the *BrGRP* gene promoter regions may be associated with different expression patterns among *BrGRP* genes under biotic and abiotic stresses. However, the number of stress response *cis*-acting elements in the promoter region is not completely consistent with the expression patterns of *BrGRP* genes under stress. For example, ABRE element is related to drought (Manavella et al., 2008) but *Bra028691* containing 11 ABRE elements were induced by high-temperature stress. Consequently, to elaborated the gene function need to further explore the regulation mechanism of transcription factors.

Two *BrGRP* Genes Participated in NaCl Stress

In our study, *w546* (*Bra030284*) was mainly expressed in the leaves, but clearly lower expression in the roots and flowers, accounting for 3–7% of the leaves (**Figure 6A**). After 7 days under salt stress, the expression of *w546* was all up-regulated in three tissues (up to 114 times in roots) of Chinese cabbage.

Although the expression of *w1409* was the highest among 64 BrGRP genes in leaves, but without clear expression differences in the three tissues (**Figure 6B**). Under salt stress, the expression of *w1409* was highly up-regulated in the roots (nine times higher than in the control roots), but was slightly repressed in the leaves and flowers. The results revealed that the two *BrGRP* genes were induced under salt stress. Although, *w546* and *w1409* were both from Subclass IVa, their expression levels were still different, suggesting that they respond differently to abiotic stress.

To further clarify the role of the two *BrGRP* genes in responses to NaCl and mannitol stress, seed germination was measured in three independent lines of transgenic *Arabidopsis* plants. The results showed that there was no clear different in germination rate among three transgenic *w546* lines (w546ox.3, w546ox.5, and w546ox.7) and the control wild-type (WT) in a normal environment or without any stress, although the seed germination potential was a little lower than that of the control WT (**Figures 6C,F**). However, 100 mmol·L^{-1} NaCl significantly inhibited the seed germination of w546ox transgenic lines with, less than 66%, significantly lower than that of the WT (100%); the germination potential was less than 16%, far lower than that of the WT (73.33%) (**Figures 6D,G**). For 300 mmol·L^{-1} mannitol treatment, the final germination rate of three transgenic *w546* lines was not much different from that of the control, but the seed germination potentials of the w546ox lines were much less than 1/3 of the WT (which was about 78.67%); even the germination potentials of w546ox.5 and w546ox.3 were 0 (**Figures 6E,H**).

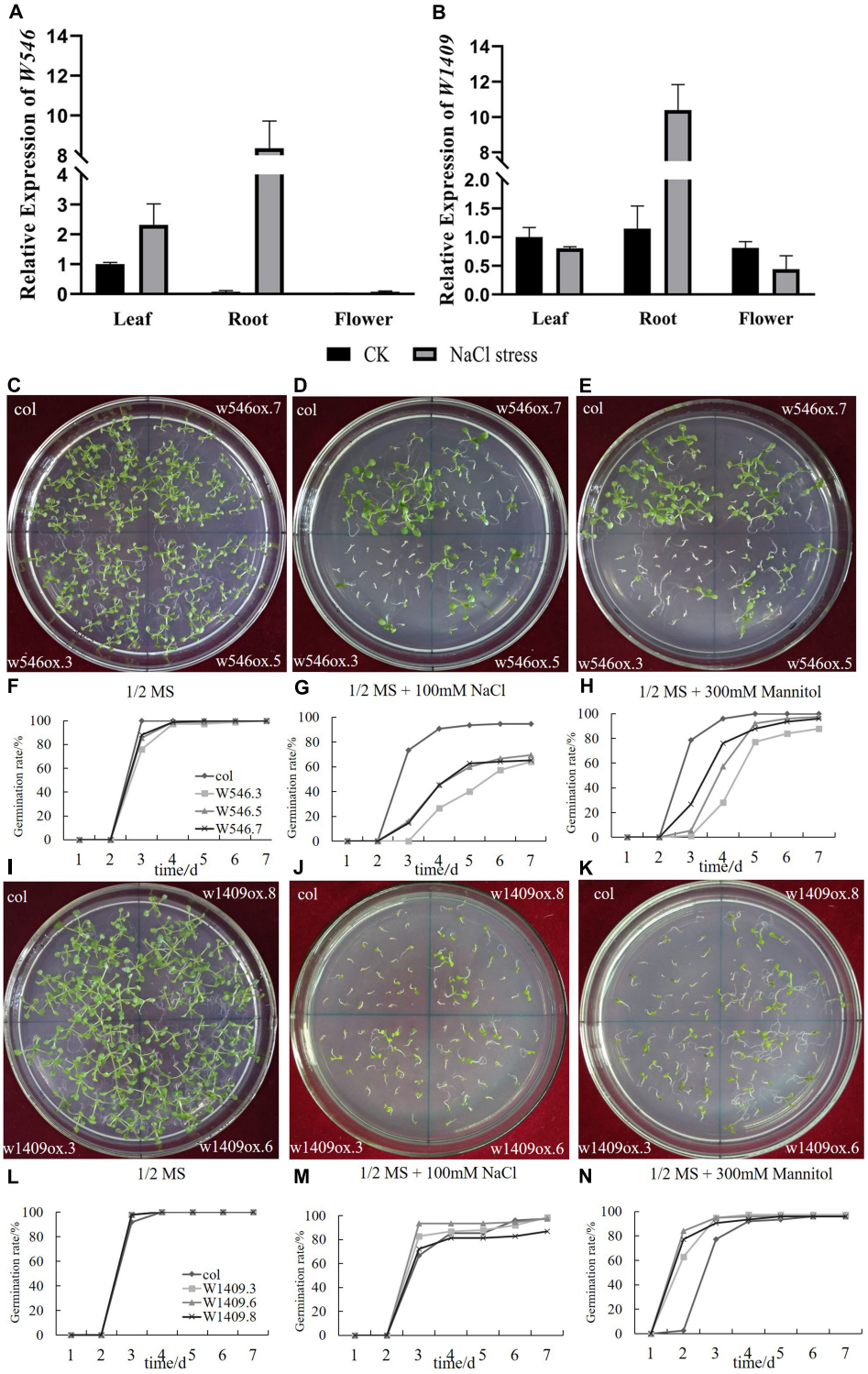


FIGURE 6 | Two *BrGRP* genes involved in NaCl stress. **(A,B)** Expression analysis of two *BrGRP* genes *w546* (*Bra030284*) and *w1409* (*Bra014000*) in three tissues, leaf root and flower, of Chinese cabbage under salt stress by qRT-PCR analysis. Data were normalized with the *GAPDH* gene and vertical bars indicate standard deviation. **(C–E)** the photo of the seed germination of *w546* -overexpressed *Arabidopsis* under control, NaCl and mannitol treatments. **(F–H)** The statistic results of **(C–E)**. **(I–K)** the photo of the seed germination of *w1409* -overexpressed *Arabidopsis* under control, NaCl and mannitol treatments. **(L–N)** The statistic results of **(I–K)**. WT: wild type. w546ox.3, w546ox.5, w546ox.7 are three individual transgenic lines with the *w546* overexpressed; w1409ox.3, w1409ox.6, and w1409ox.8 are three individual transgenic lines with *w1409* overexpressed.

Therefore, an overexpressed *w546* gene could significantly inhibit seed germination in transgenic *Arabidopsis* under salt stress, and the different phenomena between NaCl and mannitol stress suggested that Na^+ damage may occur by osmotic stress.

Like with *w546*, the seed germination rate of w1409ox lines were also not significantly different from that of the WT without stress (**Figures 6I,L**). Under the 100 mmol·L^{-1} NaCl treatment, the germination potentials of w1409ox.8 were not significantly different from that of the control, but the germination potential of w1409ox.3 and w1409ox.6 were 82.67% and 93.33%, respectively, which were higher than WT (66.67%). However, there was no significant difference in germination rate between three transgenic lines and the WT in the end, except for w1409ox.3 (**Figures 6J,M**). Under the 300 mmol·L^{-1} mannitol treatment, the germination potential of w1409ox.3 and w1409ox.6 was both 94.67%, and that of W1409ox.8 reached 90.67%, which were all higher than that of the WT (77.33%), although the final germination rate was not significantly different (**Figures 6K,N**), which was similar to the result under NaCl stress. Therefore, overexpressed *w1409* may significantly accelerate seed germination rate in transgenic lines under NaCl and mannitol stress treatments.

In summary, overexpression of two *BrGRP* genes had the opposite effects on seed germination under 100 mmol·L^{-1} NaCl and 300 mmol·L^{-1} mannitol stresses, although *w546* and *w1409* both belonged to the same subfamily of Subclass IVa *RBP-GRP*, which suggested different functions between w546 and *w1409*.

DISCUSSION

The plant *GRP* gene family is a superfamily with a glycine repeat (Gly) n-X domain. However, due to the diversification of their protein domains, gene expression patterns, and subcellular localization, these *GRP* genes are sometimes not considered a gene superfamily, but a group of proteins with some repeating structural motifs (Kar et al., 2012). This may change the classification of the *GRP* gene family. Fifteen, 22, 12, and 18 glycine-rich RNA-binding proteins (RBGs) have been identified in *Arabidopsis*, Chinese cabbage, rice and maize genomes, respectively (Zhang et al., 2014; Krishnamurthy et al., 2015). Furthermore, 9 and 51 *GRP* genes have been identified in sweet potato and *Curcuma longa L.*, respectively (Kar et al., 2012; Lu et al., 2019). In our study, 64 *BrGRP* genes were preliminarily identified in Chinese cabbage based on BRAD genome sequence, and were divided into five classes according to their conserved domains, which was similar to the classifications in previous studies (Fusaro et al., 2001; Bocca et al., 2005).

The Chinese cabbage genome not only underwent three genome-wide replication events, which also occurred in other cruciferous plants (Thomas et al., 2006; Franzke et al., 2011) but can also be divided into three subgenomic groups, LF, MF1, and MF2, according to the number of genes lost (Wang et al., 2011; Cheng et al., 2013). In this study, the syntenic analysis of *GRP* genes between Chinese cabbage and *Arabidopsis thaliana* also verified this genome-wide replication event, confirming that Chinese cabbage originated from a hexaploid ancestor, and underwent rearrangement to become diploid after chromosome fusion (Wang et al., 2011). Meanwhile, 64 *BrGRP* genes were found in Chinese cabbage, including 20 LF genes, 23 MF1 genes, and 18 MF2 genes, which were not in a three-fold relationship with the 41 *GRP* genes of *Arabidopsis*. This may suggest that *BrGRP* genes in Chinese cabbage still evolved after the genome-wide replication event, and a larger-scale loss-of-function event occurred, preventing functional redundancy.

GRP gene expression in *Arabidopsis thaliana* is tissue- or organ-specific. *GRP* genes in classes I and II are mainly expressed in seeds, siliques, roots, and leaves; class III has the highest expression in shoot tips, rosettes, seeds, and flowers; class IV is highly expressed in seeds, siliques, rosettes, and flowers; and Subclass V is up-regulated only in inflorescences (Vivek et al., 2015; Czolpinska and Rurek, 2018). This was not completely consistent with the expression of *BrGRP* genes in various tissues of Chinese cabbage, class I was mainly expressed in roots and leaves; class II was mainly expressed in root, stem, leaves and flowers; class IV had higher expression levels in various tissues; and classes III and V had lower or no expression in all tissues. *GRP* gene classes I and II may act as the active components of plant cell walls and play a crucial role in plant cell growth and organ differentiation (Park et al., 2001). *BrGRP* gene class IV has strong RNA recognition and binding abilities, and may participate in the molecular process of plant growth and development by activating splicing or regulating transcription (Czolpinska and Rurek, 2018). In the study, we found that the *BrGRP* genes that were differentially expressed at various developmental stages of Chinese cabbage belonged to classes I, II, and IV, which is similar to previous studies (Winter et al., 2007; Vivek et al., 2015) and might act similar functions in terms of growth and development.

Plant *GRP* genes that can be induced by various stresses may play another role in plant resistance. Among 8 glycine-rich RNA-binding protein genes (*AtGR-RBP1–AtGR-RBP8*) reported in *Arabidopsis thaliana,* all except for *AtGR-RBP5* and *AtGR-RBP6* were strongly expressed by low-temperature stress (Kim et al., 2005; Kwak et al., 2005). Under drought and salt stresses, *AtGR-RBP1* expression increased, but the expression of *AtGR-RBP4* and *AtGR-RBP7* gradually decreased, whereas that of *AtGR-RBP5* and *AtGR-RBP6* did not change (Kwak et al., 2005). In tobacco, the expression of *NtGRP1* was induced and continuously increased during the first 24 h of waterlogging stress, and then decreased; it was present at low levels under high or low temperature, drought, high salt, and ABA stress (Lee et al., 2009; Chen et al., 2010). *NtRGP2* and *NtRGP3* were also expressed by waterlogging and high and low-temperature stresses, but were not affected by ABA treatment with 100 μmol·L^{-1} (Chen et al., 2010). Among the four glycine-rich RNA-binding protein genes (*OsGR-RBP1-OsGR-RBP4*) in rice, only *OsGR-RBP4* was expressed by high temperature, high salt, and drought stresses (Sahi et al., 2007). In this study, a total of 64 *BrGRP* genes were identified as being differentially expressed under high temperature, low temperature and drought stresses, and soft rot. Although the *BrGRP*

genes expressed in biotic and abiotic stresses were different, and a total of 64, 52, 39, and 23 of *BrGRP* genes were induced under high temperature, low temperature, drought, and soft rot treatment, respectively.

The same *BrGRP* gene also showed different expression dynamics under different stresses, suggested that they had multiple expression patterns and different functions under adversity stress in Chinese cabbage. *AtRZ-1* was strongly expressed by low temperature and freezing stress, but was negatively regulated during seed germination and seedling growth under drought and high salt stress (Kim et al., 2005; Kim Y.O. et al., 2007; Kim and Kang, 2006). However, overexpression of *AtRZ-1B* and *AtRZ-1C* or loss of function mutations do not affect *Arabidopsis* seed germination or seedling growth under these same stress conditions (Kim et al., 2010). *AtRZ-1B/1C* regulate RNA splicing, gene expression, and many key aspects of plant development via interactions with proteins, such as SR (Wu et al., 2016). Here, most *BrGRPs* were identified to take part in stress responses, and a large number of stress-related elements were identified in their promoter region, suggesting an inextricable link and different response mechanisms between *BrGRP* genes and abiotic stress. Furthermore, two genes in Subclass Iva, *w546* (*Bra028063*) and *w1409* (*Bra014000*), were significantly up-regulated under salt stress, and their expression levels were significantly different during seed germination in transgenic *Arabidopsis* plants under salt and mannitol stresses, suggesting that *BrGRP* genes had different molecular responses to various stresses. It is very interesting that there was no clear growth or developmental defect in the two types of transgenic plants, suggesting that they can tolerate stress well.

CONCLUSION

In this study, 64 *BrGRP* genes were identified in the Chinese cabbage genome based on the 41 *AtGRP* genes in *Arabidopsis*. The *BrGRP* genes in Chinese cabbage were mainly composed of alkaline hydrophilic stable proteins and are secreted outside the cell membrane and nucleus, with only a few found in organelles such as mitochondria and chloroplast. *BrGRP* genes were divided into five classes, and shared close relationships with their homologs in *Arabidopsis*. Chromosomal localization of these *BrGRP* genes and syntenic analysis with *Arabidopsis thaliana* strongly confirmed that Chinese cabbage did undergo a genome-wide triple duplication event during its evolution. The specific expression of these genes was evaluated under various stresses, and 3–6 types of response stress *cis*-acting elements in the promoter region of these *BrGRP* genes were also identified, suggesting that they have potential roles in plant stress responses. Based on the effects of *BrGRP* gene expression in various tissues of Chinese cabbage and the germination of *Arabidopsis* strains overexpressing two *BrGRP* genes under abiotic stress, we found that the expression of *BrGRP* genes in Chinese cabbage is induced by abiotic stress. *Arabidopsis* lines overexpressing *BrGRP* genes accelerated or inhibited seed germination under abiotic stress, but which *BrGRP* genes might play more important regulating mechanism remained unclear in response to biotic and abiotic stresses.

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

XL and XX conceived and designed the experiments. XL and YC wrote the manuscript. XL, YC, and MG were responsible for data analysis. ML provided helpful advice on data analysis. ML and XX revised the manuscript and supervised the research. All authors read and approved the final manuscript.

FUNDING

This research was supported by the National Natural Science Foundation of China (31401885), the Natural Science Foundation of Shanxi Province (201801D121248), and Shanxi Province Key Research and Development Program Key Projects (201703D211001-04-01).

ACKNOWLEDGMENTS

We would like to thank Dr. Gilberto Sachetto (Universidade Federal de Viçosa) for providing the information of class V *AtGRP* genes and associate professor Jiangli Shi from Henan Agricultural University for English language editing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00774/full#supplementary-material

REFERENCES

Aneeta, R., Sanan-Mishra, N., Tuteja, N., Kumar, P., and Sopory, S. (2002). Salinity- and ABA-induced up-regulation and light-mediated modulation of mRNA encoding glycine-rich RNA-binding protein from *Sorghum bicolor*. *Biochem. Biophys. Res. Commun.* 296, 1063–1068. doi: 10.1016/s0006-291x(02)02050-8

Banerjee, J., Sahoo, D. K., Dey, N., Houtz, R. L., and Maiti, I. B. (2013). An intergenic region shared by At4g35985 and At4g35987 in *Arabidopsis thaliana* is a tissue specific and stress inducible bidirectional promoter analyzed in transgenic *Arabidopsis* and tobacco plants. *PLoS One* 8:e79622. doi: 10.1371/journal.pone.0079622

Bocca, S. N., Magioli, C., Mangeon, A., Junqueira, R. M., Cardeal, V., Margis, R., et al. (2005). Survey of glycine-rich proteins (GRPs) in the Eucalyptus expressed

sequence tag database (ForEST). *Genet. Mol. Biol.* 28, 608–624. doi: 10.1590/S1415-47572005000400016

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* S1674–2052(20)30187–8. doi: 10.1016/j.molp.2020.06.009

Chen, X., Zeng, Q.-C., Lu, X.-P., Yu, D.-Q., and Li, W.-Z. (2010). Characterization and expression analysis of four glycine-rich RNA-binding proteins involved in osmotic response in tobacco (*Nicotiana tabacum* cv. Xanthi). *Agric. Sci. China* 9, 1577–1587. doi: 10.1016/s1671-2927(09)60254-6

Cheng, F., Liu, S., Wu, J., Fang, L., Sun, S., Liu, B., et al. (2011). BRAD, the genetics and genomics database for *Brassica plants*. *BMC Plant Biol.* 11:136. doi: 10.1186/1471-2229-11-136

Cheng, F., Mandakova, T., Wu, J., Xie, Q., Lysak, M. A., and Wang, X. (2013). Deciphering the diploid ancestral genome of the Mesohexaploid *Brassica rapa*. *Plant Cell* 25, 1541–1554. doi: 10.1105/tpc.113.110486

Clough, S. J., and Bent, A. F. (1998). Floral dip: a simplified method for Agrobacterium-mediated transformation of *Arabidopsis thaliana*. *Plant J.* 16, 735–743. doi: 10.1046/j.1365-313x.1998.00343.x

Condit, C. M., and Meagher, R. B. (1986). A gene encoding a novel glycine-rich structural protein of petunia. *Nature* 323, 178–181. doi: 10.1038/323178a0

Czolpinska, M., and Rurek, M. (2018). Plant glycine-rich proteins in stress response: an emerging, still prospective story. *Front. Plant Sci.* 9:302. doi: 10.3389/fpls.2018.00302

Dai, X., Xu, Y., Ma, Q., Xu, W., Wang, T., Xue, Y., et al. (2007). Overexpression of an R1R2R3 MYB gene, OsMYB3R-2, increases tolerance to freezing, drought, and salt stress in transgenic *Arabidopsis*. *Plant Physiol.* 143, 1739–1751. doi: 10.1104/pp.106.094532

Dong, X., Yi, H., Lee, J., Nou, I. S., Han, C. T., and Hur, Y. (2015). Global gene-expression analysis to identify differentially expressed genes critical for the heat stress response in *Brassica rapa*. *PLoS One* 10:e0130451. doi: 10.1371/journal.pone.0130451

Dreyfuss, G., Kim, V. N., and Kataoka, N. (2002). Messenger-RNA-binding proteins and the messages they carry. *Nat. Rev. Mol. Cell Biol.* 3, 195–205. doi: 10.1038/nrm760

Ernst, J., and Bar-Joseph, Z. (2006). STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics* 7:191. doi: 10.1186/1471-2105-7-191

Ferreira, M. A., Engler, J. D. A., Miguens, F. C., Montagu, M. V., Engler, G., and Oliveira, D. E. D. (1997). Oleosin gene expression in *Arabidopsis thaliana* tapetum coincides with accumulation of lipids in plastids and cytoplasmic bodies. *Plant Physiol. Biochem.* 35, 729–739. doi: 10.1104/pp.115.1.283

Franzke, A., Lysak, M. A., Al-Shehbaz, I. A., Koch, M. A., and Mummenhoff, K. (2011). Cabbage family affairs: the evolutionary history of *Brassicaceae*. *Trends Plant Sci.* 16, 108–116. doi: 10.1016/j.tplants.2010.11.005

Fusaro, A., Mangeon, A., Junqueira, R. M., Rocha, C. A. B., Coutinho, T. C., Margis, R., et al. (2001). Classification, expression pattern and comparative analysis of sugarcane expressed sequences tags (ESTs) encoding glycine-rich proteins (GRPs). *Genet. Mol. Biol.* 24, 263–273. doi: 10.1590/s1415-47572001000100035

Guo, Y. M., Samans, B., Chen, S., Kibret, K. B., Hatzig, S., Turner, N. C., et al. (2017). Drought-tolerant *Brassica rapa* shows rapid expression of gene networks for general stress responses and programmed cell death under simulated drought stress. *Plant Mol. Biol. Rep.* 35, 416–430. doi: 10.1007/s11105-017-1032-4

Kar, B., Nayak, S., and Joshi, R. K. (2012). Classification and comparative analysis of *Curcuma longa* L. expressed sequences tags (ESTs) encoding glycine-rich proteins (GRPs). *Bioinformation* 8, 142–146. doi: 10.6026/97320630008142

Kim, J. S., Jung, H. J., Lee, H. J., Kim, K. A., Goh, C. H., Woo, Y., et al. (2008). Glycine-rich RNA-binding protein 7 affects abiotic stress responses by regulating stomata opening and closing in *Arabidopsis thaliana*. *Plant J.* 55, 455–466. doi: 10.1111/j.1365-313X.2008.03518.x

Kim, J. Y., Park, S. J., Jang, B., Jung, C. H., Ahn, S. J., Goh, C. H., et al. (2007). Functional characterization of a glycine-rich RNA-binding protein 2 in *Arabidopsis thaliana* under abiotic stress conditions. *Plant J.* 50, 439–451. doi: 10.1111/j.1365-313X.2007.03057.x

Kim, W. Y., Kim, J. Y., Jung, H. J., Oh, S. H., Han, Y. S., and Kang, H. (2010). Comparative analysis of *Arabidopsis* zinc finger-containing glycine-rich RNA-binding proteins during cold adaptation. *Plant Physiol. Biochem.* 48, 866–872. doi: 10.1016/j.plaphy.2010.08.013

Kim, Y. O., and Kang, H. (2006). The role of a zinc finger-containing glycine-rich RNA-binding protein during the cold adaptation process in *Arabidopsis thaliana*. *Plant Cell Physiol.* 47, 793–798. doi: 10.1093/pcp/pcj047

Kim, Y. O., Kim, J. S., and Kang, H. (2005). Cold-inducible zinc finger-containing glycine-rich RNA-binding protein contributes to the enhancement of freezing tolerance in *Arabidopsis thaliana*. *Plant J.* 42, 890–900. doi: 10.1111/j.1365-313X.2005.02420.x

Kim, Y. O., Pan, S., Jung, C. H., and Kang, H. (2007). A zinc finger-containing glycine-rich RNA-binding protein, atRZ-1a, has a negative impact on seed germination and seedling growth of *Arabidopsis thaliana* under salt or drought stress conditions. *Plant Cell Physiol.* 48, 1170–1181. doi: 10.1093/pcp/pcm087

Krishnamurthy, P., Kim, J. A., Jeong, M., Kang, C. H., and Lee, S. I. (2015). Defining the RNA-binding glycine-rich (RBG) gene superfamily: new insights into nomenclature, phylogeny, and evolutionary trends obtained by genome-wide comparative analysis of *Arabidopsis*, Chinese cabbage, rice and maize genomes. *Mol. Genet. Genomics* 290, 2279–2295. doi: 10.1007/s00438-015-1080-0

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096

Kwak, K. J., Kim, Y. O., and Kang, H. (2005). Characterization of transgenic *Arabidopsis* plants overexpressing GR-RBP4 under high salinity, dehydration, or cold stress. *J. Exp. Bot.* 56, 3007–3016. doi: 10.1093/jxb/eri298

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., et al. (2012). The *Arabidopsis* information resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* 40, D1202–D1210. doi: 10.1093/nar/gkr1090

Lee, M. O., Kim, K. P., Kim, B. G., Hahn, J. S., and Hong, C. B. (2009). Flooding stress-induced glycine-rich RNA-binding protein from *Nicotiana tabacum*. *Mol. Cells* 27, 47–54. doi: 10.1007/s10059-009-0004-4

Lescot, M., Dehais, 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

Li, Z.-Z., Wu, P.-Q., Cheng, Y.-X., Song, H.-X., and Xu, X.-Y. (2018). Construction and evaluation of flower bud normalized cDNA library of male sterile of Chinese cabbage (*Brassica rapa* L. pekinensis). *Plant Physiol. J.* 54, 291–296. doi: 10.13592/j.cnki.ppj.2017.0497

Liu, M., Wu, F., Wang, S., Lu, Y., Chen, X., Wang, Y., et al. (2019). Comparative transcriptome analysis reveals defense responses against soft rot in Chinese cabbage. *Hortic. Res.* 6:68. doi: 10.1038/s41438-019-0149-z

Lorkovic, Z. J., and Barta, A. (2002). Genome analysis: RNA recognition motif (RRM) and K homology (KH) domain RNA-binding proteins from the flowering plant *Arabidopsis thaliana*. *Nucleic Acids Res.* 30, 623–635. doi: 10.1093/nar/30.3.623

Lu, Y., Sun, J., Yang, Z., Zhao, C., Zhu, M., Ma, D., et al. (2019). Genome-wide identification and expression analysis of glycine-rich RNA-binding protein family in sweet potato wild relative *Ipomoea trifida*. *Gene* 686, 177–186. doi: 10.1016/j.gene.2018.11.044

Maestrini, P., Cavallini, A., Rizzo, M., Giordani, T., Bernardi, R., Durante, M., et al. (2009). Isolation and expression analysis of low temperature-induced genes in white poplar (*Populus alba*). *J. Plant Physiol.* 166, 1544–1556. doi: 10.1016/j.jplph.2009.03.014

Manavella, P. A., Dezar, C. A., Ariel, F. D., and Chan, R. L. (2008). Two ABREs, two redundant root-specific and one W-box cis-acting elements are functional in the sunflower HAHB4 promoter. *Plant Physiol. Biochem.* 46, 860–867. doi: 10.1016/j.plaphy.2008.05.003

Mangeon, A., Junqueira, R. M., and Sachetto-Martins, G. (2010). Functional diversity of the plant glycine-rich proteins superfamily. *Plant Signal. Behav.* 5, 99–104. doi: 10.4161/psb.5.2.10336

Mangeon, A., Magioli, C., Menezes-Salgueiro, A. D., Cardeal, V., de Oliveira, C., Galvao, V. C., et al. (2009). AtGRP5, a vacuole-located glycine-rich protein involved in cell elongation. *Planta* 230, 253–265. doi: 10.1007/s00425-009-0940-4

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628. doi: 10.1038/nmeth.1226

Murphy, D. J., Hernández-Pinzón, I., and Patel, K. (2001). Role of lipid bodies and lipid-body proteins in seeds and other tissues. *J. Plant Physiol.* 158, 471–478. doi: 10.1078/0176-1617-00359

Onishi, M., Tachi, H., Kojima, T., Shiraiwa, M., and Takahara, H. (2006). Molecular cloning and characterization of a novel salt-inducible gene encoding an acidic isoform of PR-5 protein in soybean (*Glycine max* [L.] Merr.). *Plant Physiol. Biochem.* 44, 574–580. doi: 10.1016/j.plaphy.2006.09.009

Park, A. R., Cho, S. K., Yun, U. J., Jin, M. Y., Lee, S. H., Sachetto-Martins, G., et al. (2001). Interaction of the *Arabidopsis* receptor protein kinase Wak1 with a glycine-rich protein, AtGRP-3. *J. Biol. Chem.* 276, 26688–26693. doi: 10.1074/jbc.M101283200

Park, S. K., Howden, R., and Twell, D. (1998). The *Arabidopsis thaliana* gametophytic mutation gemini pollen1 disrupts microspore polarity, division asymmetry and pollen cell fate. *Development* 125, 3789–3799. doi: 10.1016/j.brainresbull.2012.05.011

Sahi, C., Agarwal, M., Singh, A., and Grover, A. (2007). Molecular characterization of a novel isoform of rice (*Oryza sativa* L.) glycine rich-RNA binding protein and evidence for its involvement in high temperature stress response. *Plant Sci.* 173, 144–155. doi: 10.1016/j.plantsci.2007.04.010

Shinozuka, H., Hisano, H., Yoneyama, S., Shimamoto, Y., Jones, E. S., Forster, J. W., et al. (2006). Gene expression and genetic mapping analyses of a perennial ryegrass glycine-rich RNA-binding protein gene suggest a role in cold adaptation. *Mol. Genet. Genomics* 275, 399–408. doi: 10.1007/s00438-005-0095-3

Thomas, B. C., Pedersen, B., and Freeling, M. (2006). Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res.* 16, 934–946. doi: 10.1101/gr.4708406

Tong, C., Wang, X., Yu, J., Wu, J., Li, W., Huang, J., et al. (2013). Comprehensive analysis of RNA-seq data reveals the complexity of the transcriptome in *Brassica rapa*. *BMC Genomics* 14:689. doi: 10.1186/1471-2164-14-689

Vivek, K., Hanlon, M. R., Sergio, C., Ferlanti, E. S., Svetlana, K., Maria, K., et al. (2015). Araport: the *Arabidopsis* information portal. *Nucleic Acids Res.* 43, 1003–1009. doi: 10.1093/nar/gku1200

Voorrips, R. E. (2002). MapChart: software for the graphical presentation of linkage maps and QTLs. *J. Hered.* 93, 77–78. doi: 10.1093/jhered/93.1.77

Wang, F., Li, L., Li, H., Liu, L., Zhang, Y., Gao, J., et al. (2012). Transcriptome analysis of rosette and folding leaves in Chinese cabbage using high-throughput RNA sequencing. *Genomics* 99, 299–307. doi: 10.1016/j.ygeno.2012.02.005

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., et al. (2011). The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* 43, 1035–1039. doi: 10.1038/ng.919

Wang, Z., Zhu, Y., Wang, L., Liu, X., Liu, Y., Phillips, J., et al. (2009). A WRKY transcription factor participates in dehydration tolerance in Boea hygrometrica by binding to the W-box elements of the galactinol synthase (BhGolS1) promoter. *Planta* 230, 1155–1166. doi: 10.1007/s00425-009-1014-3

Wienkoop, S., Morgenthal, K., Wolschin, F., Scholz, M., Selbig, J., and Weckwerth, W. (2008). Integration of metabolomic and proteomic phenotypes: analysis of data covariance dissects starch and RFO metabolism from low and high temperature compensation response in *Arabidopsis thaliana*. *Mol. Cell. Proteomics* 7, 1725–1736. doi: 10.1074/mcp.M700273-MCP200

Winter, D., Vinegar, B., Nahal, H., Ammar, R., Wilson, G. V., and Provart, N. J. (2007). An "Electronic fluorescent pictograph" browser for exploring and analyzing large-scale biological data sets. *PLoS One* 2:e718. doi: 10.1371/journal.pone.0000718

Wu, Z., Zhu, D., Lin, X., Miao, J., Gu, L., Deng, X., et al. (2016). RNA binding proteins RZ-1B and RZ-1C play critical roles in regulating Pre-mRNA splicing and gene expression during development in *Arabidopsis*. *Plant Cell* 28, 55–73. doi: 10.1105/tpc.15.00949

Xu, X. Y., Sun, X. L., Zhang, J., Huang, W. W., Zhang, L. G., and Fang, Z. Y. (2013). Identification of candidate genes associated with male sterility in CMS7311 of heading Chinese cabbage (*Brassica campestris* L. ssp pekinensis). *Acta Physiol. Plant.* 35, 3265–3270. doi: 10.1007/s11738-013-1346-2

Yang, D. H., Kwak, K. J., Kim, M. K., Park, S. J., Yang, K. Y., and Kang, H. (2014). Expression of *Arabidopsis* glycine-rich RNA-binding protein AtGRP2 or AtGRP7 improves grain yield of rice (*Oryza sativa*) under drought stress conditions. *Plant Sci.* 214, 106–112. doi: 10.1016/j.plantsci.2013.10.006

Yang, W., Yu, M., Zou, C., Lu, C., Yu, D., Cheng, H., et al. (2019). Genome-wide comparative analysis of RNA-binding Glycine-rich protein family genes between *Gossypium arboreum* and *Gossypium raimondii*. *PLoS One* 14:e0218938. doi: 10.1371/journal.pone.0218938

Zhang, C. W., Wei, Y. P., Xiao, D., Gao, L. W., Lyu, S. W., Hou, X. L., et al. (2016). Transcriptomic and proteomic analyses provide new insights into the regulation mechanism of low-temperature-induced leafy head formation in Chinese cabbage. *J. Proteomics* 144, 1–10. doi: 10.1016/j.jprot.2016.05.022

Zhang, J., Zhao, Y., Xiao, H., Zheng, Y., and Yue, B. (2014). Genome-wide identification, evolution, and expression analysis of RNA-binding glycine-rich protein family in maize. *J. Integr. Plant Biol.* 56, 1020–1031. doi: 10.1111/jipb.12210

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.

Copyright © 2020 Lu, Cheng, Gao, Li and Xu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 03 August 2020
doi: 10.3389/fgene.2020.00812

The Genome Sequence of Alpine *Megacarpaea delavayi* Identifies Species-Specific Whole-Genome Duplication

Qiao Yang[1†], Hao Bi[1†], Wenjie Yang[1], Ting Li[1], Jiebei Jiang[1], Lei Zhang[1], Jianquan Liu[1,2] and Quanjun Hu[1]*

[1] *Key Laboratory of Bio-Resource and Eco-Environment of Ministry of Education, State Key Laboratory of Hydraulics and Mountain River Engineering, College of Life Sciences, Sichuan University, Chengdu, China,* [2] *State Key Laboratory of Grassland Agro-Ecosystem, Institute of Innovation Ecology, Lanzhou University, Lanzhou, China*

OPEN ACCESS

Edited by:
Xiaoming Song, North China University of Science and Technology, China

Reviewed by:
Michael Eric Schranz, Wageningen University and Research, Netherlands
Jingyin Yu, Boyce Thompson Institute, United States

***Correspondence:**
Quanjun Hu
huquanjun@gmail.com

† These authors have contributed equally to this work

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

Received: *30 March 2020*
Accepted: *06 July 2020*
Published: *03 August 2020*

Citation:
Yang Q, Bi H, Yang W, Li T, Jiang J, Zhang L, Liu J and Hu Q (2020) The Genome Sequence of Alpine Megacarpaea delavayi Identifies Species-Specific Whole-Genome Duplication. Front. Genet. 11:812. doi: 10.3389/fgene.2020.00812

Megacarpaea delavayi (Brassicaceae), a plant found the high mountains of southwest China at high altitudes (3000–4800 m), is used as a vegetable or medicine. Here, we report a draft genome for this species. The assembly genome of *M. delavayi* is 883 Mb, and 61.59% of the genome is composed of repeat sequences. Annotation of the genome identified a total of 41,114 protein-coding genes. We found that *M. delavayi* experienced an independent whole-genome duplication (WGD), paralleling those independent WGDs in *Iberis*, *Biscutella*, and *Anastatica* in the early Miocene. Phylogenetic analyses based on the single-copy genes confirmed the position of the genus *Megacarpaea* within the expanded lineage II of the family and resolved its basal divergence to a subclade consisting of *Anastatica, Iberis*, and *Biscutella.* Species-specific and fast-evolving genes in *M. delavayi* are mainly involved in "DNA repair" and "response to UV-B radiation." These genetic changes may together help this species survive in high-altitude environments. The reference genome reported here provides a valuable resource for studying adaptation of this and other alpine plants to the high-altitude habitats.

Keywords: *Megacarpaea delavayi*, genome sequence, alpine adaptation, whole-genome duplication, Brassicaceae

INTRODUCTION

Polyploidy (whole-genome duplication, WGD), which occurs frequently through evolutionary histories of plants (Wu et al., 2020), contributes greatly to both species diversification and colonization of the new niches (Soltis and Soltis, 2016). Numerous independent WGDs were identified for angiosperm families, such as Asteraceae (Huang et al., 2016), Poaceae (McKain et al., 2016), and Brassicaceae (Edger et al., 2015). The Brassicaceae-specific WGD was named At-α WGDs (Bowers et al., 2003), which occurred 23–43 million years ago (Mya) (Mandáková et al., 2010). In addition, more independent WGDs were revealed to be specific to lineages or species within the family (Mandáková et al., 2017). It is increasingly clear that subsequent lineage-specific or species-specific WGD events laid the foundation for species diversification, environmental adaptability, and stress tolerance of the Brassicaceae species (Kagale et al., 2014).

In this study, we aimed to examine whether WGD occurred in one alpine herb, *Megacarpaea delavayi* (2*n* = 18) of Brassicaceae through sequencing its draft genome. This species grows on swampy meadows, steep grassy slopes, and open thickets of the high mountains (Hengduan Mountains) in southwest China at elevations of 3000–4800 m (Cheo et al., 2001). It has been collected as a wild vegetable and medicine for years by the local inhabitants in the high-altitude regions (Zhong et al., 2015). The dried plants of *M. delavayi* are used to treat dysentery, lung cough, and disordered indigestion by Bai and Tibetan people (Lei et al., 2009; Shen et al., 2009). Other species of the genus are distributed in the high-elevation regions from the Hengduan Mountains to Himalaya and central Asia (Cheo et al., 2001). However, phylogenetic relationships of the genus *Megacarpae* in the family Brassicaceae remain unclear although both recent studies suggested its likely position in the expanded lineage II of the family Brassicaceae (Guo et al., 2017; Nikolov et al., 2019). Maximum likelihood analyses based on targeted enrichment sequence data suggested the close relationship between *Megacarpaea*, *Iberis*, *Cochlearia* and others although this received little support according to coalescent analyses of these data (Nikolov et al., 2019). In addition, in these analyses, it remained unsolved whether *Biscutella* should be placed in this clade (Nikolov et al., 2019) although another study based on genome-scale single-copy genes suggested the well-supported close relationship between *Biscutella* and *Iberis* (Kiefer et al., 2019). Independent WGDs, which might have led to incorrect gene orthology alignments (Walker et al., 2017), seem to account for these conflicting phylogenies (Nikolov et al., 2019).

Here, we report the assembly and comparative genomic analysis of the *M. delavayi* genome. We revealed that one independent WGD occurred in this species in the early Miocene, paralleling those WGDs in other genera. We determined the phylogenetic relationship of the genus *Megacarpaea* based on genome-scale single-copy genes. We further found that numerous species-specific and fast-evolving genes existed in this species, which may be beneficial for its survival in the alpine habitat.

MATERIALS AND METHODS

Plant Materials and Genomic DNA Extraction

One wild *M. delavayi* (2*n* = 18) individual (**Figure 1A**) was collected from Cangshan Mountain (3181 m, N25.659, E100.117) in Yunnan Province, China. Fresh and healthy leaves were immediately frozen at −80°C for DNA extraction. High-quality genomic DNA from leaf tissue was extracted with the CTAB method (Liu et al., 2009). We used 1% agarose gel electrophoresis to check the quality of the high-molecular-weight DNA. High-quality genomic DNA with an effective concentration of more than 2 nM was used to construct the library.

Genome Sequencing and Assembly

We constructed Illumina paired-end libraries with small (230, 500, and 800 bp) and large (2, 5, 10, and 20 kb) insert sizes and read lengths of 150 bp (**Supplementary Table S1**). We sequenced them using the Illumina HiSeq 2500 platform (Illumina, San Diego, CA, United States) at Novogene (Tianjin, China) following the manufacturer protocols. Short reads were first subjected to quality filtering with Trimmomatic v0.36 (Bolger et al., 2014), error correction with BFC v1.8 (Li, 2015), and mate-pair data deduplication with FastUniq v1.1 (Xu et al., 2012). The genome was initially assembled into scaffolds with Platanus v1.2.4 (Kajitani et al., 2014) with the first round of gap closing. An additional gap closing procedure was performed with GapCloser v1.12 (Luo et al., 2012). Finally, we evaluated genome assemblies for completeness using BUSCO v3.0.2 (Waterhouse et al., 2018) with "embryophyta_odb9" (**Supplementary Table S2**).

Genome Size Estimation

The sequencing reads were used to estimate genome size of *M. delavayi* based on k-mer frequencies. We used quality-filtered Illumina short reads and 17-mer frequency distribution to estimate the genome size with Jellyfish v2.2.9 (Marçais and Kingsford, 2011). The highest peak value of k-mer distribution was used to estimate the sequencing depth. We plotted the distribution of k-mer depth against the frequency with the main peak occurring at a depth of 19 (**Supplementary Figure S1**). A Perl script[1] was used to calculate the size of the *M. delavayi* genome.

Repetitive Identification

We used a combination of homology-based and *de novo* search methods to identify repeat sequences in the *M. delavayi* genome. In terms of homology-based prediction, RepeatMasker v4.0.7 (Tarailo-Graovac and Chen, 2009) was used to find repeat elements at the DNA level with the Repbase library. *De novo* repeat annotation of the *M. delavayi* genome was performed with RepeatModeler v1.0.11 (Smit and Hubley, 2008).

We identified intact long-terminal repeat retrotransposons (LTR-RTs) by searching the *M. delavayi* genome with LTRharvest v1.5.10 (Ellinghaus et al., 2008; -motif tgca -motifmis 1) and LTR_Finder v1.06 (Xu and Wang, 2007; -D 20000 -d 1000 -L 5000 -l 100). Then, LTR_retriever v1.9 (Ou and Jiang, 2018) was used to integrate the results of LTR_Finder and LTRharvest (**Supplementary Table S3**). Using a substitution rate (r) of 7×10^{-9} substitutions per site per year (Ossowski et al., 2010), we calculated the insertion time (T) for each LTR retrotransposon as $T = K/(2r)$, where K is genetic distance and r is the rate of nucleotide substitution per site per year ($r = 7 \times 10^{-9}$).

Gene Annotation

A combination of *de novo*, homology-based, and transcript-based approaches were used to predict protein-coding genes in the *M. delavayi* genome. Before the transcriptome could be aligned, RNA-Seq reads needed to be assembled into transcripts. We first filtered RNA-Seq reads with potential low-quality regions using Trimmomatic v0.36 (Bolger et al., 2014). After quality control was performed, all clean reads were assembled into *de novo* transcripts with Trinity v2.8.4 (Haas et al., 2013). Then,

[1]https://github.com/josephryan/estimate_genome_size.pl

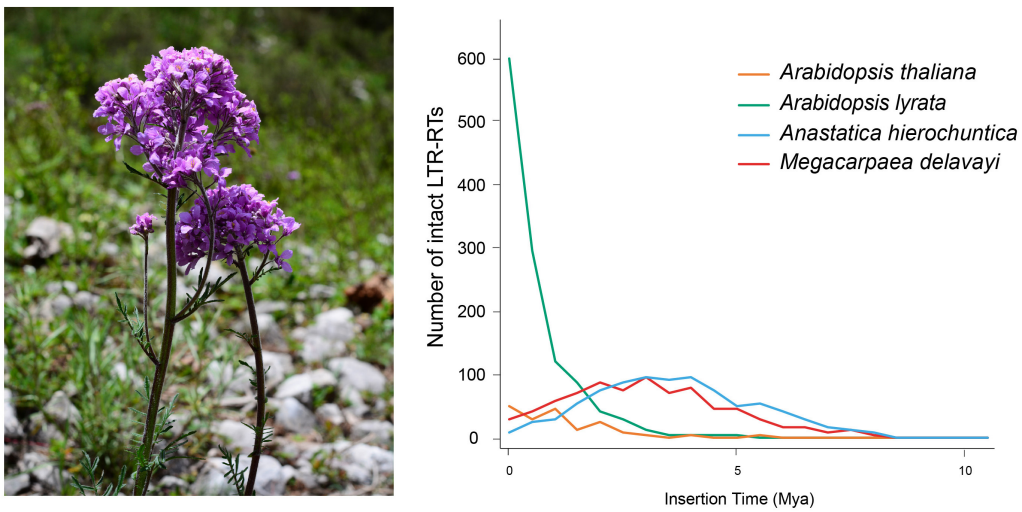


FIGURE 1 | (A) *M. delavayi* in flowering stage of growth. Photo taken by Lei Zhang in Pingwu County (3,077 m a.s.l.), Sichuan Province, China. **(B)** Insertion time distribution of intact LTR-RTs of *M. delavayi* and other Brassicaceae species.

we used PASA v2.1.0 (Haas et al., 2003) to obtain information on the gene structure annotation by aligning the assembled transcripts with the genomes. Protein sequences of *Aethionema arabicum*, *Arabidopsis lyrata*, *Arabidopsis thaliana*, *Brassica rapa*, and *Capsella rubella*, were obtained for homology-based gene annotation. GlimmerHMM v3.0.4 (Majoros et al., 2004) was used to predict the gene structure in each protein-coding region. We performed *de novo* prediction with AUGUSTUS v3.2.3 (Stanke et al., 2006) to annotate protein-coding genes. The gene model parameters were trained from *A. thaliana* and our transcriptome data set. The above three gene prediction results were merged with EVidenceModeler v1.1.1 (Haas et al., 2008) to form a comprehensive and non-redundant reference gene list. Weights of evidence for gene models were defined as follows: *de novo* prediction weight (Augustus) = 1, homology-based prediction weight (GlimmerHMM) = 5, transcript-based prediction weight (PASA) = 10. The EVM merged result was updated with an additional round of PASA annotation to add UTRs and provide information on alternative splicing variants to gene models.

To obtain functional annotation of protein-coding genes, we used Blast2GO v2.5 (Conesa and Götz, 2008) for gene ontology (GO) annotation based on the NCBI-NR database. The predicted genes were mapped to KEGG pathways using KAAS (Moriya et al., 2007) to obtain the KEGG annotation. For Swiss-Prot annotations, we employed BLAST + v2.2.31 (Camacho et al., 2009; Blastp with the E-value cutoff 1×10^{-5}) to align proteins to the Swiss-Prot databases. InterProScan v5.31-70 was used to determine the domains/motifs (Jones et al., 2014; **Supplementary Table S4**).

Gene Family Identification

We downloaded the protein-coding genes of *Arabis alpina*, *Boechera stricta*, *Crucihimalaya himalaica*, *Eutrema heterophyllum*, and *Lepidium meyenii* together with *Megacarpaea delavayi* to identify orthologous groups (**Supplementary Table S5**). To remove redundancy caused by alternative splicing variations, we retained only the gene models at each gene locus that encoded the longest protein sequence, and putative fragmented genes that encoded protein sequences shorter than 50 aa and stop codon ratios greater than 20% were filtered out. Then, we used Diamond v0.9.22 (Buchfink et al., 2014; E-value cutoff 1×10^{-5}) to compare all filtered protein sequences and used OrthoMCL v2.0.9 (Li et al., 2003) to cluster genes into orthologous groups. Genes that could not be clustered into any gene family and for which only one species existed were considered species specific. Finally, we summarized the gene family cluster results for six species in Venn diagram format.

Phylogenetic Analyses

We downed the reported genomes of all related species from two main lineages of Brassiaceae, including *Aethionema arabicum*, *Anastatica hierochuntica*, *Arabidopsis lyrate*, *Arabidopsis thaliana*, *Biscutella auriculate, Biscutella laevigata, Brassica rapa*, *Boechera stricta, Crucihimalaya himalaica, Eutrema heterophyllum, Iberis amara, Kernera saxatilis, Lepidium meyenii, Macropodium nivale*, and *Noccaea caerulescens* (**Supplementary Table S5**). We used the single-copy orthologous genes identified in the gene family cluster analyses from these species and *M. delavayi* to construct phylogenetic tree. Multiple sequence alignments were performed for the protein sequence of each single-copy orthologous gene with MAFFT v7.313 (Katoh and Standley, 2013). Then, the alignments were concatenated to generate a super alignment matrix, which was used to generate a maximum likelihood tree with the PROTGAMMAILGX model in RAxML v8.2.11 (Stamatakis, 2014).

The divergence time between these species was estimated with the MCMCtree program in PAML v4.9 (Yang, 2007). The F84 model (model = 4) and independent rates molecular clock

(clock = 2) were used for calculations in MCMCtree. The MCMC process was run for 1,500,000 iterations, with a sample frequency of 150, after a burn-in of 500,000 iterations. We ran the program twice for each data type to confirm that the results were convergent between runs. We looked up three calibration points in the TimeTree database (Kumar et al., 2017) to estimate the Brassicaceae divergence time: divergence time for *Ae. arabicum* and other Brassicaceae plants was 32–43 Mya, divergence time for Lineages II+ expanded lineage II and lineage I was 23.4–33.5 Mya, divergence time for *L. meyenii* and other lineage I plants was 11.9–20.6 Mya. The phylogenetic analyses for Brassicaceae was visualized with FigTree v1.4.3[2].

Gene Family Expansion and Contraction

CAFÉ (De Bie et al., 2006) is a tool for analyzing evolutionary changes of the gene families. This software uses the stochastic birth and death process to model gene gain and loss over a phylogeny. Based on the results of phylogeny and divergence time, we applied CAFE v4.2 to identify gene families that had undergone expansion and contraction on the phylogeny tree (*p*-value cutoff 0.05). For each significantly expanded and contracted gene family in *M. delavayi*, we inferred functional information based on its functional annotations.

WGD and Positively Selected Genes

We used MCScanX (Wang et al., 2012) to detect syntenic blocks (regions with at least five collinear genes) and duplication levels (depth) for four species: *A. hierochuntica*, *A. thaliana*, *B. rapa*, and *M. delavayi*. To recover the WGD event, we calculated synonymous substitution rates (Ks) for syntenic genes using codeml in PAML. To further examine whether the recent WGDs were shared by *M. delavayi* and the closely related species, we extracted homologous gene groups to construct phylogenetic trees of the orthologous genes. We used phylogenetic relationships of the homologous genes identified the likely WGD nodes.

To identify fast-evolving genes, we used MCScanX to search for syntenic blocks. Similar to previous research (Wu et al., 2019), we calculated non-synonymous substitution (Ka) and synonymous substitution (Ks) for the collinear orthologous gene pairs using the Perl script "add_ka_and_ks_to_collinearity.pl" in MCScanX. The ratio of Ka to Ks is a commonly used indicator of selective pressure acting on protein-coding genes with a ratio >1 representing positive selection.

RESULTS

Genome Size Estimation

The distribution of short subsequence (k-mer) frequency, also known as the k-mer spectrum, is widely used to estimate genome size (Li et al., 2013; Zhang et al., 2015). A k-mer depth distribution was obtained from Jellyfish (Marçais and Kingsford, 2011) analyses, and the peak depth was clearly visible from the distribution data (**Supplementary Figure S1**). The genome size of *M. delavayi* calculated by the aforementioned Perl script was estimated to be approximately 899 Mb (**Table 1**).

Genome Assembly and Annotation

To sequence the whole genome of *M. delavayi*, we generated 133 Gb of paired-end and mate-pair clean reads (150 × assembled sequence coverage) with different insert sizes using an Illumina HiSeq 2500 platform (**Supplementary Table S1**). The final assembly genome of *M. delavayi* (883.81 Mb) consisted of 763,815 contigs (contig N50, 65.48 kb; longest contig, 794.13 kb; **Table 1**). Completeness of the genome regions was further assessed with BUSCO. Of a core set of 1440 single-copy ortholog genes of the Embryophyta lineage, 94.5% were complete in the genome, which contained 85.6% of the plant single-copy orthologs and 8.9% of the plant duplicate orthologs (**Supplementary Table S2**), which suggests that the genome of *M. delavayi* was well-assembled with high completeness and accuracy.

De novo, homology-based, and transcript-based approaches were combined to annotate protein-coding sequences (Chen et al., 2020; Kang et al., 2020; Zhang et al., 2020). In total, 41,114 genes were predicted with an average gene length and number of exons of 1853 bp and 6.57, respectively (**Table 1**). Moreover, we annotated functions of the predicted genes with Swiss-Prot, InterProScan, GO, and KEGG databases. The results showed that 85.78% of all the protein-coding genes were successfully annotated by at least one database (**Supplementary Table S4**).

Repetitive Elements Analysis

Through combination of *de novo* searches and homology-based methods, we identified nearly 536 Mb repetitive elements, representing 61.59% of the *M. delavayi* genome (**Table 2**). Retrotransposons (LTR, SINEs, and LINEs) were the most abundant, accounting for 44.06% of the genome. LTR-RTs represented 39.25% of the genome, and Ty3/Gypsy (24.71%) made up major elements of LTR-RTs (**Table 2**).

A high proportion of repetitive elements in the *M. delavayi* genome were LTR-RTs, and the proliferation of retrotransposons might have been responsible for genome expansion (Zhang et al., 2019). To estimate insertion time of the LTR-RTs, we identified complete LTR-RTs in four Brassicaceae species (*A.*

TABLE 1 | Statistics for the *M. delavayi* genome assembly and annotation.

Feature	Value
Estimated size (Mb)	899
Assembly size (Mb)	883.81
Number of contigs	763,815
Maximum contig length (kb)	794.13
Contig N50 (kb)	65.48
Number of contigs at least N50	3,402
GC content (%)	31.64
Repeat content (%)	61.59
Number of protein-coding genes	41,114
Average gene length (bp)	1,852.54
Average exon length (bp)	266.06
Average number of exons per gene	6.57

[2]http://tree.bio.ed.ac.uk/

TABLE 2 | Repeat content (subtypes) of *M. delavayi* genome.

Repeat type	Repeat size (bp)	Percentage of genome (%)
DNA transposons	103,382,966	11.70
LTR retroelements:	346,878,665	39.25
Ty1/Copia	20,388,180	2.31
Ty3/Gypsy	218,392,577	24.71
SINE retroelements	1,084,358	0.12
LINE retroelements	41,410,409	4.69
TRF	27,943,342	3.16
Simple repeats	6,422,772	0.73
Low complexity	718,144	0.08
Unclassified	9,975,172	1.13
Unknown	6,485,326	0.73
Total	536,335,637	61.59

hierochuntica, *A. thaliana*, *A. lyrate*, and *M. delavayi*). We identified 814 complete LTR-RTs in the *M. delavayi* genome and 176 in *A. thaliana* and 1227 in *A. lyrate* (**Supplementary Table S3**). *A. thaliana* had more microdeletions in transposons than *M. delavayi* (Hu et al., 2011), and *A. lyrata* had a comparatively higher proportion of recent insertions (Slotte et al., 2013), consistent with our study. The insertion time distribution showed that *M. delavayi* LTR-RTs expanded within the past 5 million years (based on r of 7×10^{-9} substitutions per site per year; **Figure 1B**). In general, recent expansion of repeat sequences may have played a key role in increasing the genome size of *M. delavayi*.

Phylogenetic Analyses

Sixteen species were selected to identify orthologous groups, and they were clustered into 51,589 orthologous groups. A total of 361 single-copy gene families were identified and used to construct the maximum likelihood phylogenetic tree. Three major lineages were identified: lineages I, and traditionally recognized lineages II and those added to the expanded lineage II. Phylogenetic analysis confirmed the phylogenetic position of *M. delavayi* in the expanded lineage II of Brassicaceae (**Figure 2A**), consistent with previous published research (Guo et al., 2017; Kiefer et al., 2019; Nikolov et al., 2019). Within this lineage, *M. delavayi* diverged early as one subclade, and another comprised *Anastatica*, *Iberis* and *Biscutella* as suggested before based on the genome-scale single-copy genes. The close relationships between the latter three genera agree with phylogenetic analyses similarly based on the genome-scale, single-copy genes (Kiefer et al., 2019). Based the fossil-calibrated phylogeny, *M. delavayi* diverged from

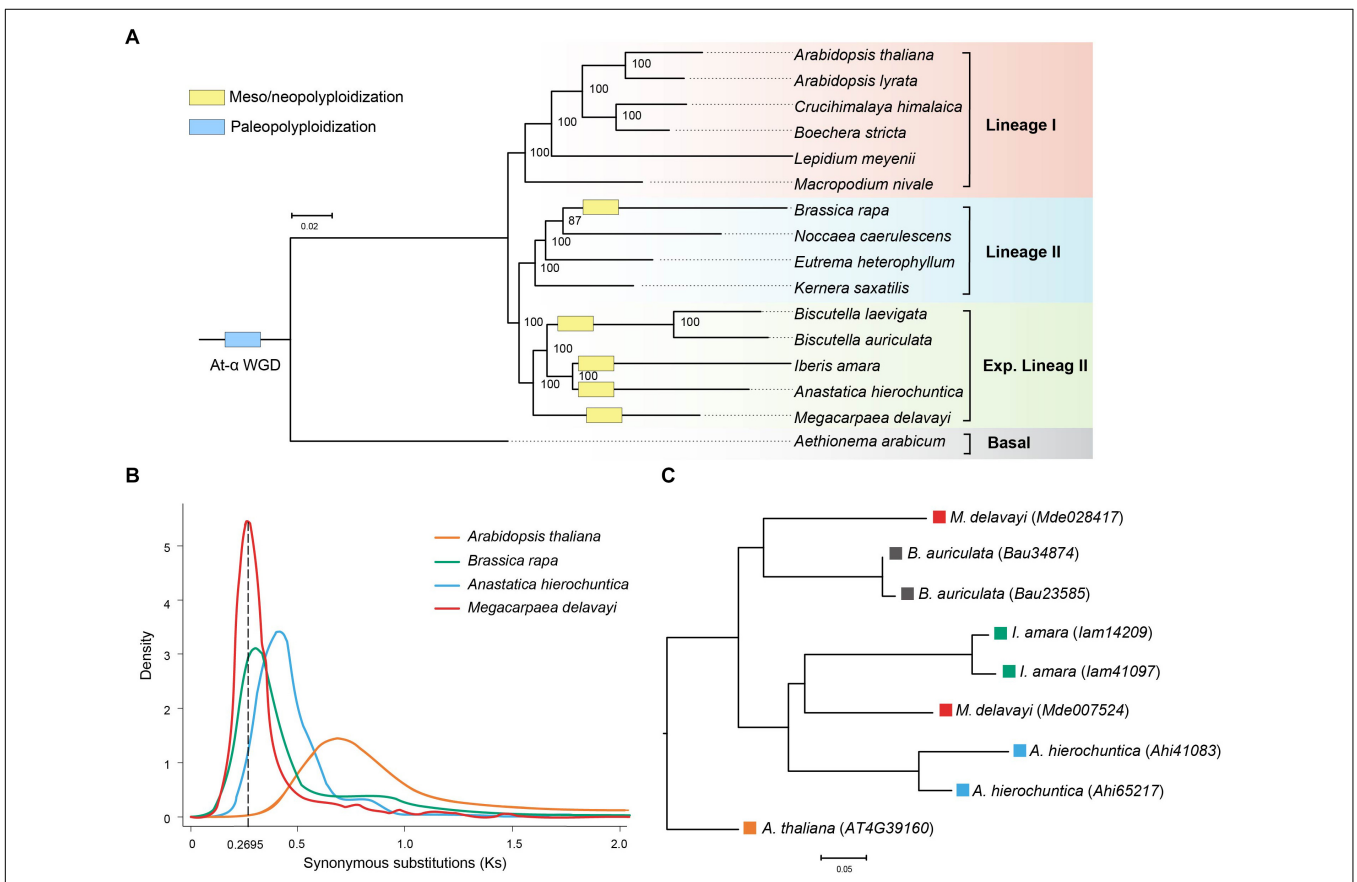


FIGURE 2 | (A) Phylogenetic relationship of *M. delavayi* and other Brassicaceae species. Numbers on nodes represent bootstrap values. **(B)** Synonymous substitutions (Ks) estimated from *M. delavayi*-specific WGD. **(C)** A phylogenetic tree of the homologous genes from the expanded lineage II.

other three genera of the expanded lineage II around 20.53 Mya (**Supplementary Figure S2**).

Whole-Genome Duplication

The distribution of Ks was analyzed to uncover and assess the frequency of WGD events in the Brassicaceae (Kagale et al., 2014). Using syntenic orthologs within each genome to construct the distribution of Ks, we found *M. delavayi* had undergone a further more species-specific WGD (**Figure 2B**) after the well-known ancient At-α (23–47 Mya) paleopolyploid WGD. Although the Brassicaceae diverged from other closely related eudicots at the beginning of the Cenozoic era, the rapid species diversification of the family occurred only within the Miocene (<23 Mya; Kagale et al., 2014). Importantly, the lineage- or species-specific polyploid or WGD events seemed to have promoted species diversification during this recent stage (Kagale et al., 2014; Kiefer et al., 2019). Four genera, *Brassica*, *Anastatica*, *Iberis*, and *Biscutella*, of the expanded lineage II were suggested to experience independent WGDs (Wang et al., 2011; Kiefer et al., 2019). We examined Ks distributions of these species. We confirmed these WGDs and found that they occurred between 16 and 23 Mya almost at the same stage after their divergences (**Figure 2B**). These WGDs seem to occur independently based on the Ks distributions despite the slight differences between them. It remains interesting to further examine whether *Megacarpaea* shared a WGD with the closely related *Anastatica*, *Iberis*, and *Biscutella* based the homologous genes. After the WGD, the duplicated gene may have been randomly lost in the derived lineages, and therefore, it is difficult to identify all paralogous genes between different lineages. We, therefore, used *Arabidopsis* without further WGD as an out-group. We extracted a set of homologous gene groups: *Arabidopsis* (1): *Megacarpaea* (2): *Anastatica* (2): *Iberis* (2): *Biscutella* (2). We recovered 24 groups of homologous genes, but phylogenetic analyses suggested that only one group could be used to construct a gene tree with most subclades statistically supported (**Figure 2C**). On this tree, two genes from *Megacarpaea* did not cluster together, and their clustering with other subclades failed to receive statistical support. Because two genes from each of the other three genera comprised a monophyletic clade, respectively, it is highly likely that WGDs occurred in these genera independently.

Gene Family Expansion and Contraction

Gene loss and gain are the primary reasons for functional changes (Xing et al., 2019). To better understand the relationships between the gene families of *M. delavayi* and other crucifer, we performed a systematic comparison of genes among different species. Phylogenetic analyses indicated that *M. delavayi* was phylogenetically categorized into expanded lineage II. Further comparisons of these species revealed 2673 expanded and 3600 contracted gene families in the *M. delavayi* genome. Significant expansion or contraction in the size of particular gene families is often associated with the adaptive divergence of related species (Zhang et al., 2016). Also, a total of 41 gene families showed significant expansion ($P < 0.05$), and 37 gene families showed significant contraction ($P < 0.05$) in *M. delavayi*. The significantly expanded gene families contained 312 genes, which are mainly involved in "response to light," "response to salt stress," "response to water deprivation," and "calcium-mediated signaling" (**Supplementary Table S6**). This also agrees with the previous predication that some species from the expanded lineage II of Brassicaceae are salt-tolerant (Monihan et al., 2020).

We further examined the shared and species-specific gene families between *M. delavayi* and other alpine crucifers with genomes available: *A. alpine* (Willing et al., 2015), *B. stricta*, *C. himalaica* (Zhang et al., 2019), *E. heterophyllum* (Guo et al., 2018), and *L. meyenii* (Zhang et al., 2016). These alpine crucifers were found to develop more species-specific genes to adapt to alpine habitats. We next examined whether *M. delavayi* developed more such genes in addition to those shared with other alpine crucifers. We identified a total of 28,835 homologous gene families, 13,117 of which were shared by six alpine crucifers. We identified 1383 gene families specific to the *M. delavayi* genome (**Figure 3A**). GO enrichments of these genes in species-specific gene families revealed that they were mainly involved in "response to UV-B," "response to cold," "DNA repair," and "cellular response to salt stress" (**Figure 3B** and **Supplementary Table S7**).

Fast-Evolving Genes in *M. delavayi*

In plants, tolerance to UV-B radiation and cold are critical for surviving at high altitudes (Zhang et al., 2016). Increased Ka relative to Ks in certain genes may explain the adaptive evolution of organisms at the molecular level (Qiu et al., 2012). We identified 1203 syntenic gene blocks containing 20,442 collinear gene pairs in the *M. delavayi* and *A. thaliana* genomes. A total of 327 genes under positive selection had a Ka/Ks ratio greater than 1.0. The Swiss-Port functional classification revealed that the fast-evolving genes with putative functions were related to "DNA repair," "response to UV-B radiation," "defense response," and "response to cold" (**Supplementary Table S8**). In particular, we found that *DRT101* (*Mde002296.11*) related to the DNA repair from UV damage evolved quickly in the *M. delavayi* genome (**Supplementary Table S8**). In the maca genome, genes related to DNA repair (*DRT102*) have also been found to evolve rapidly (Zhang et al., 2016). Both *DRT101* and *DRT102* belong to the DNA-damage-repair/toleration (*DRT*) genes (**Supplementary Table S9**), and they may encode UV-specific excision repair activities (Pang et al., 1993; Hays and Pang, 1994). The accelerated evolution of DRT and other genes may help *M. delavayi* adapt to the high-altitude environment.

DISCUSSION

In this study, we performed *de novo* assembly of the *M. delavayi* genome based on an Illumina HiSeq 2500 platform. *M. delavayi* is the first sequenced species of the genus *Megacarpaea*. This reference genome provides a basis for further studying speciation based on genomic data for the genus and comparative genomics studies in the family Brassicaceae. The genome of *M. delavayi* was estimated at 899 Mb, and the final assembly genome was

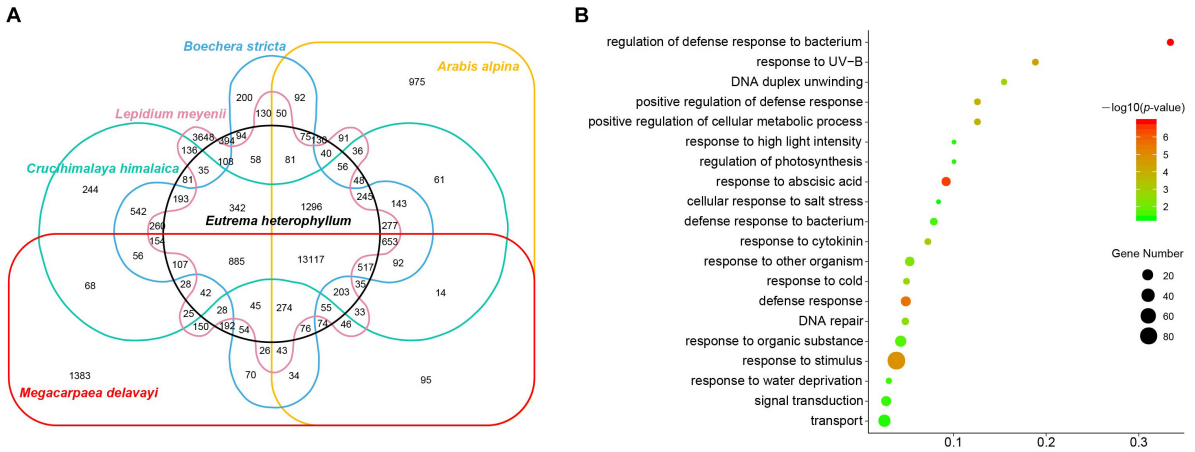


FIGURE 3 | (A) Venn diagram showing shared and unique gene families among six species of Brassicaceae. **(B)** Gene ontology (GO) enrichment of *M. delavayi*–specific genes.

883.81 Mb, representing about 98% of the estimated genome size (**Table 1**). WGD and expansion of repetitive elements in the *M. delavayi* genome might have led to its larger size than other species (**Figure 1B** and **Supplementary Table S3**). Based on the genome-scale single copy genes, our phylogenetic analysis clearly shows that *Megacarpaea* was placed in the expanded lineage II of Brassiaceae, and it was closely related to *Anastatica*, *Iberis,* and *Biscutella*. This finding resolved the ambiguous phylogenetic position of the genus *Megacarpaea* in the previous study (Nikolov et al., 2019) because of the difficulties in gene orthologous alignments (Walker et al., 2017).

Our genomic analyses suggested one species-specific WGD event in the *M. delavayi* genome. This WGD paralleled independently to those that occurred in the closely related species, *Anastatica*, *Biscutella,* and *Iberis.* These WGDs were estimated to occur within the Miocene shortly after the radiative divergences of the sampled genera. Such repeated WGDs accompanying lineage divergences together drove species diversification of the family (Kagale et al., 2014). In addition, WGDs should have also played an important role for crucifers to colonize the arid habitats because of the obvious advantages of the polyploids under selective pressure (Dong et al., 2019).

Compared with the published alpine crucifers with available genomes (Guo et al., 2018; Zeng et al., 2019; Zhang et al., 2019), we also found that *M. delavayi* retained numerous species-specific genes, which are involved in "response to UV-B," "response to cold," "DNA repair," and "cellular response to salt stress" (**Figure 3B** and **Supplementary Table S7**). These genes derived from WGD or other ways may also play an important role for *M. delavayi* to adapt to a cold and UV-B stressed habitat at high altitude (Cheviron and Brumfield, 2012). In addition, fast-evolving genes in *M. delavayi* were also found to be involved in "DNA repair" and abiotic stresses (**Supplementary Table S8**). All these findings suggest that *M. delavayi* had developed obvious genomic changes to adapt to alpine habitats. The reference genome presented here provides an important resource for further studying molecular adaptation of this and other alpine plants to the highlands.

DATA AVAILABILITY STATEMENT

The genomic sequence data of *Megacarpaea delavayi* in this study have been deposited in the NCBI under BioProject PRJNA630110. The assembled genome and genome annotation information have been deposited in the National Genomics Data Center (https://bigd.big.ac.cn/?lang=zh) under BioProject PRJCA002887.

AUTHOR CONTRIBUTIONS

QH and JL designed the research. LZ and HB collected the materials and performed the genome sequencing and assembly. QY, WY, HB, TL, JJ, and LZ performed the genome annotation and evolution analyses. QY, QH, and JL wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported equally by the National Key Research and Development Program of China (2017YFC0505203) and the Fundamental Research Funds for the Central Universities (2018CDDY-S02-SCU and SCU2019D013), the National High-Level Talents Special Support Plans (10 Thousand of People Plan), and 985 and 211 Projects of the Sichuan University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00812/full#supplementary-material

FIGURE S1 | 17-k-mer frequency distribution of sequencing reads from *M. delavayi*. The size of the *M. delavayi* genome was estimated at 899 Mb.

FIGURE S2 | Phylogenetic tree with divergence time and expanded/contracted gene family for *M. delavayi* and other Brassicaceae species. The numbers above the branches are the predicted divergence times.

TABLE S1 | Sequencing statistics of the *M. delavayi* genome.

TABLE S2 | Evaluation *M. delavayi* genome assembly with BUSCO.

TABLE S3 | Numbers of annotated protein-coding genes in *M. delavayi* genome.

TABLE S4 | Intact LTR-RTs in *M. delavayi* and other Brassicaceae species.

TABLE S5 | Protein data sets used for comparative genomic analyses.

TABLE S6 | Function of significantly expanded genes in *M. delavayi* genome.

TABLE S7 | GO enrichment of *M. delavayi* specific genes.

TABLE S8 | Function of fast-evolving genes in *M. delavayi* genome.

TABLE S9 | DRT genes copy number in *M. delavayi* and *L. meyenii* genomes (identity > 80%).

REFERENCES

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

Bowers, J. E., Chapman, B. A., Rong, J., and Paterson, A. H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422, 433–438. doi: 10.1038/nature01521

Buchfink, B., Xie, C., and Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421

Chen, Y., Ma, T., Zhang, L., Kang, M., Zhang, Z., Zheng, Z., et al. (2020). Genomic analyses of a "living fossil": the endangered dove-tree. *Mol. Ecol. Resour.* 20, 1–14.

Cheo, T. Y., Lu, L., Yang, G., Al-Shehbaz, I., and Dorofeev, V. (2001). Megacarpaea. *Flora China* 8, 39–40.

Cheviron, Z. A., and Brumfield, R. T. (2012). Genomic insights into adaptation to high-altitude environments. *Heredity* 108, 354–361. doi: 10.1038/hdy.2011.85

Conesa, A., and Götz, S. (2008). Blast2GO: a comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genomics* 2008:619832.

De Bie, T., Cristianini, N., Demuth, J. P., and Hahn, M. W. (2006). CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22, 1269–1271. doi: 10.1093/bioinformatics/btl097

Dong, Y., Gupta, S., Sievers, R., Wargent, J. J., Wheeler, D., Putterill, J., et al. (2019). Genome draft of the *Arabidopsis* relative *Pachycladon cheesemanii* reveals novel strategies to tolerate New Zealand's high ultraviolet B radiation environment. *BMC Genomics* 20:838. doi: 10.1186/s12864-019-6084-4

Edger, P. P., Heidel-Fischer, H. M., Bekaert, M., Rota, J., Glöckner, G., Platts, A. E., et al. (2015). The butterfly plant arms-race escalated by gene and genome duplications. *Proc. Natl. Acad. Sci. U.S.A.* 112, 8362–8366. doi: 10.1073/pnas.1503926112

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* 9:18. doi: 10.1186/1471-2105-9-18

Guo, X., Hu, Q., Hao, G., Wang, X., Zhang, D., Ma, T., et al. (2018). The genomes of two *Eutrema* species provide insight into plant adaptation to high altitudes. *DNA Res.* 25, 307–315. doi: 10.1093/dnares/dsy003

Guo, X., Liu, J., Hao, G., Zhang, L., Mao, K., Wang, X., et al. (2017). Plastome phylogeny and early diversification of Brassicaceae. *BMC Genomics* 18:176. doi: 10.1186/s12864-017-3555-3

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K., Hannick, L. I., et al. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 31, 5654–5666. doi: 10.1093/nar/gkg770

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 using the Trinity platform for reference generation and analysis. *Nat. Protoc.* 8, 1494–1512. doi: 10.1038/nprot.2013.084

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* 9:R7.

Hays, J. B., and Pang, Q. (1994). *UV-B-Inducible and Constitutive Genes that Mediate Repair and Toleration Of UV-Damaged DNA in the Plant Arabidopsis Thaliana*. Berlin: Springer.

Hu, T. T., Pattyn, P., Bakker, E. G., Cao, J., Cheng, J. F., Clark, R. M., et al. (2011). The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.* 43, 476–483.

Huang, C.-H., Zhang, C., Liu, M., Hu, Y., Gao, T., Qi, J., et al. (2016). Multiple polyploidization events across asteraceae with two nested events in the early history revealed by nuclear phylogenomics. *Mol. Biol. Evol.* 33, 2820–2835. doi: 10.1093/molbev/msw157

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, 1236–1240. doi: 10.1093/bioinformatics/btu031

Kagale, S., Robinson, S. J., Nixon, J., Xiao, R., Huebert, T., Condie, J., et al. (2014). Polyploid evolution of the Brassicaceae during the *Cenozoic era*. *Plant Cell* 26, 2777–2791. doi: 10.1105/tpc.114.126391

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., et al. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* 24, 1384–1395. doi: 10.1101/gr.170720.113

Kang, M., Wu, H., Yang, Q., Huang, L., Hu, Q., Ma, T., et al. (2020). A chromosome-scale genome assembly of *Isatis indigotica*, an important medicinal plant used in traditional Chinese medicine: an Isatis genome. *Hortic. Res.* 7:18.

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kiefer, C., Willing, E.-M., Jiao, W.-B., Sun, H., Piednoël, M., Hümann, U., et al. (2019). Interspecies association mapping links reduced CG to TG substitution rates to the loss of gene-body methylation. *Nat. plants* 5, 846–855. doi: 10.1038/s41477-019-0486-9

Kumar, S., Stecher, G., Suleski, M., and Hedges, S. B. (2017). TimeTree: a resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* 34, 1812–1819. doi: 10.1093/molbev/msx116

Lei, S., Xiaobo, L., Guirong, S., Yun, Y., and Bin, L. (2009). Effect of *Megacarpaea delavayi* Franch on digestive juice in rat with heat due to food stagnation. *Chinese J. Ethnomed. Ethnopharm.* 18, 1–3.

Li, H. (2015). BFC: correcting Illumina sequencing errors. *Bioinformatics* 31, 2885–2887. doi: 10.1093/bioinformatics/btv290

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189. doi: 10.1101/gr.1224503

Li, M., Tian, S., Jin, L., Zhou, G., Li, Y., Zhang, Y., et al. (2013). Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. *Nat. Genet.* 45, 1431–1438. doi: 10.1038/ng.2811

Liu, W. S., Wei, W., and Dong, M. (2009). Clonal and genetic diversity of Carex moorcroftii on the Qinghai-Tibet plateau. *Biochem. Syst. Ecol.* 37, 370–377. doi: 10.1016/j.bse.2009.07.003

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1:18.

Majoros, W. H., Pertea, M., and Salzberg, S. L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* 20, 2878–2879. doi: 10.1093/bioinformatics/bth315

Mandáková, T., Joly, S., Krzywinski, M., Mummenhoff, K., and Lysaka, M. A. (2010). Fast diploidization in close mesopolyploid relatives of *Arabidopsis*. *Plant Cell* 22, 2277–2290. doi: 10.1105/tpc.110.074526

Mandáková, T., Li, Z., Barker, M. S., and Lysak, M. A. (2017). Diverse genome organization following 13 independent mesopolyploid events in Brassicaceae

contrasts with convergent patterns of gene retention. *Plant J.* 91, 3–21. doi: 10.1111/tpj.13553

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770. doi: 10.1093/bioinformatics/btr011

McKain, M. R., Tang, H., McNeal, J. R., Ayyampalayam, S., Davis, J. I., DePamphilis, C. W., et al. (2016). A phylogenomic assessment of ancient polyploidy and genome evolution across the Poales. *Genome Biol. Evol.* 8, 1150–1164.

Monihan, S. M., Magness, C. A., Ryu, C.-H., McMahon, M. M., Beilstein, M. A., and Schumaker, K. S. (2020). Duplication and functional divergence of a calcium sensor in the Brassicaceae. *J. Exp. Bot.* 71, 2782–2795. doi: 10.1093/jxb/eraa031

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* 35, W182–W185.

Nikolov, L. A., Shushkov, P., Nevado, B., Gan, X., Al-Shehbaz, I. A., Filatov, D., et al. (2019). Resolving the backbone of the Brassicaceae phylogeny for investigating trait diversity. *New Phytol.* 222, 1638–1651. doi: 10.1111/nph.15732

Ossowski, S., Schneeberger, K., Lucas-Lledó, J. I., Warthmann, N., Clark, R. M., Shaw, R. G., et al. (2010). The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* 327, 92–94. doi: 10.1126/science.1180677

Ou, S., and Jiang, N. (2018). LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* 176, 1410–1422. doi: 10.1104/pp.17.01310

Pang, Q., Hays, J. B., Rajagopal, I., and Schaefer, T. S. (1993). Selection of *Arabidopsis* cDNAs that partially correct phenotypes of *Escherichia coli* DNA-damage-sensitive mutants and analysis of two plant cDNAs that appear to express UV-specific dark repari activities. *Plant Mol. Biol.* 22, 411–426. doi: 10.1007/bf00015972

Qiu, Q., Zhang, G., Ma, T., Qian, W., Ye, Z., Cao, C., et al. (2012). The yak genome and adaptation to life at high altitude. *Nat. Genet.* 44, 946–949.

Shen, L., Fang, C., Wang, J., Sa, L., and Yang, L. (2009). The effect of *megacarpaea delavayi* franch on intestinal propulsive and gastric emptying function in mice. *J. Dali Univ.* 8, 8–10.

Slotte, T., Hazzouri, K. M., Ågren, J. A., Koenig, D., Maumus, F., Guo, Y. L., et al. (2013). The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nat. Genet.* 45, 831–835. doi: 10.1038/ng.2669

Smit, A. F. A., and Hubley, R. (2008). *RepeatModeler Open-1.0*. Available online at: http://www.repeatmasker.org/ (accessed November 24, 2017).

Soltis, P. S., and Soltis, D. E. (2016). Ancient WGD events as drivers of key innovations in angiosperms. *Curr. Opin. Plant Biol.* 30, 159–165. doi: 10.1016/j.pbi.2016.03.015

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: *A b* initio prediction of alternative transcripts. *Nucleic Acids Res.* 34, W435–W439.

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform.* 25, 4–10.

Walker, J. F., Yang, Y., Moore, M. J., Mikenas, J., Timoneda, A., Brockington, S. F., et al. (2017). Widespread paleopolyploidy, gene tree conflict, and recalcitrant relationships among the carnivorous Caryophyllales. *Am. J. Bot.* 104, 858–867. doi: 10.3732/ajb.1700083

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., et al. (2011). The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* 43, 1035–1040.

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40:e49. doi: 10.1093/nar/gkr1293

Waterhouse, R. M., Seppey, M., Simao, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., et al. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* 35, 543–548. doi: 10.1093/molbev/msx319

Willing, E. M., Rawat, V., Mandáková, T., Maumus, F., James, G. V., Nordström, K. J. V., et al. (2015). Genome expansion of *Arabis alpina* linked with retrotransposition and reduced symmetric DNA methylation. *Nat. Plants* 1, 1–7.

Wu, H., Ma, T., Kang, M., Ai, F., Zhang, J., Dong, G., et al. (2019). A high-quality *Actinidia chinensis* (kiwifruit) genome. *Hortic. Res.* 6, 1–9.

Wu, S., Han, B., and Jiao, Y. (2020). Genetic contribution of paleopolyploidy to adaptive evolution in angiosperms. *Mol. Plant* 13, 59–71. doi: 10.1016/j.molp.2019.10.012

Xing, Y., Liu, Y., Zhang, Q., Nie, X., Sun, Y., Zhang, Z., et al. (2019). Hybrid de novo genome assembly of Chinese chestnut (*Castanea mollissima*). *Gigascience* 8:giz112.

Xu, H., Luo, X., Qian, J., Pang, X., Song, J., Qian, G., et al. (2012). FastUniq: a fast de novo duplicates removal tool for paired short reads. *PLoS One* 7:e52249. doi: 10.1371/journal.pone.0052249

Xu, Z., and Wang, H. (2007). LTR-FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 35, W265–W268.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Zeng, X., Yuan, H., Dong, X., Peng, M., Jing, X., Xu, Q., et al. (2019). Genome-wide Dissection of Co-selected UV-B Responsive Pathways in the UV-B Adaptation of Qingke. *Mol. Plant* 13, 112–127. doi: 10.1016/j.molp.2019.10.009

Zhang, J., Tian, Y., Yan, L., Zhang, G., Wang, X., Zeng, Y., et al. (2016). Genome of plant maca (*Lepidium meyenii*) illuminates genomic basis for high-altitude adaptation in the central andes. *Mol. Plant* 9, 1066–1077. doi: 10.1016/j.molp.2016.04.016

Zhang, T., Hu, Y., Jiang, W., Fang, L., Guan, X., Chen, J., et al. (2015). Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. *Nat. Biotechnol.* 33, 531–537.

Zhang, T., Qiao, Q., Novikova, P. Y., Wang, Q., Yue, J., Guan, Y., et al. (2019). Genome of *Crucihimalaya himalaica*, a close relative of *Arabidopsis*, shows ecological adaptation to high altitude. *Proc. Natl. Acad. Sci. U.S.A.* 116, 7137–7146. doi: 10.1073/pnas.1817580116

Zhang, Z., Chen, Y., Zhang, J., Ma, X., Li, Y., Li, M., et al. (2020). Improved genome assembly provides new insights into genome evolution in a desert poplar (*Populus euphratica*). *Mol. Ecol. Resour.* 20, 1–14.

Zhong, L., Zhang, L. Q., and Lan, M. (2015). Indigenous vegetables in Yunnan Province, China. *Acta Hortic.* 1102, 89–92. doi: 10.17660/actahortic.2015.1102.10

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.

Copyright © 2020 Yang, Bi, Yang, Li, Jiang, Zhang, Liu and Hu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 August 2020
doi: 10.3389/fgene.2020.00963

Correlation Analysis of Expression Profile and Quantitative iTRAQ-LC-MS/MS Proteomics Reveals Resistance Mechanism Against TuMV in Chinese Cabbage (*Brassica rapa* ssp. *pekinensis*)

Shanwu Lyu[1,2], Liwei Gao[1,2], Rujia Zhang[1,2], Changwei Zhang[1,2] and Xilin Hou[1,2]**

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

OPEN ACCESS

Edited by:
Zhiyong Liu,
Shenyang Agricultural University, China

Reviewed by:
Zhang De Shuang,
Beijing Vegetable Research Center, China
Arif Ashraf,
University of Massachusetts Amherst, United States

***Correspondence:**
Changwei Zhang
changweizh@njau.edu.cn
Xilin Hou
hxl@njau.edu.cn

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

Received: *30 April 2020*
Accepted: *31 July 2020*
Published: *20 August 2020*

Citation:
Lyu S, Gao L, Zhang R, Zhang C and Hou X (2020) Correlation Analysis of Expression Profile and Quantitative iTRAQ-LC-MS/MS Proteomics Reveals Resistance Mechanism Against TuMV in Chinese Cabbage (Brassica rapa ssp. pekinensis). Front. Genet. 11:963. doi: 10.3389/fgene.2020.00963

The arms race between plants and viruses never ceases. Chinese cabbage, an important type of *Brassica* vegetable crop, is vulnerable to plant virus infection, especially to *Turnip mosaic virus* (TuMV). To better examine the molecular mechanisms behind the virus infection, we conducted the correlation analysis of RNA-Seq and quantitative iTRAQ-LC-MS/MS in TuMV-infected and in healthy Chinese cabbage leaves. There were 757 differentially expressed genes and 75 differentially expressed proteins that were screened in Chinese cabbage plants infected with TuMV. These genes were enriched in many pathways, and among them, the plant hormone signal transduction, plant-pathogen interaction, and protein processing in the endoplasmic reticulum pathways were suggested to be closely related pathways. The correlation analysis between RNA-Seq and quantitative iTRAQ-LC-MS/MS was then further explored. Finally, we obtained a preliminary network of several candidate genes associated with TuMV infection, and we found that they mainly belonged to calcium signaling pathways, heat shock proteins, WRKY transcription factors, and non-specific lipid transfer proteins. These results may lead to a better understanding of antiviral mechanisms and of disease-resistant breeding.

Keywords: transcriptome (RNA-Seq), proteome, correlation analysis, Chinese cabbage, *Turnip mosaic virus* (TuMV) resistance

INTRODUCTION

Brassica vegetable crops not only provide multiple components of nutrition such as proteins, vitamins, trace elements, and inorganic salts, but also serve as important research materials. Among *Brassica* vegetable crops, Chinese cabbage (*Brassica rapa* ssp. *pekinensis*) has the largest cultivated area and yield, and the completion of its genome sequence in 2011 has made its transcriptomic and proteomic analysis more feasible (Cao et al., 2006; Cheng et al., 2011).

However, several biotic stresses have been constant threats to both crop quality and yield. *Turnip mosaic virus* (TuMV), a member of the *Potyvirus* genus, has emerged as an important pathogen to *Brassica* crops (Walsh and Jenner, 2002). It ranks only second to *Cucumber mosaic*

virus (CMV) and is the most prevalent viral pathogen infecting Chinese cabbage, especially in Asia, North America, and Europe (Tomlinson, 1987). It infects 318 kinds of dicotyledons and several monocotyledons through artificial inoculation (Walsh and Jenner, 2002; Shattuck, 2010). Moreover, it can develop conspicuous symptoms such as systemic vein clearing, necrosis, chlorosis, and withering; internal necrosis would occur following further cold storage (Li et al., 2019). On account of its broad parasitifer and its non-persistent transmission, many means of prevention against TuMV are ineffective, and some of them can cause agricultural chemical pollution (Walsh and Jenner, 2002). Furthermore, climate change also enhances the difficulties in controlling damage caused by plant viruses (Jones and Barbetti, 2012). As such, investigation of natural plant genes that confer resistance against viruses may provide effective and environmentally friendly ways for the reduction or restriction of viral infections (Fraser, 1990; Hughes et al., 2002; Walsh and Jenner, 2002; Li et al., 2019).

Plant viruses are a type of obligate intracellular parasite, so their infection cycles are intimately related to and rely completely on the host cell component, including host factors. Host factors encompass proteins, nucleic acids, carbohydrates, lipids, and metabolites which can be used for fighting off pathogens (Laliberte and Sanfacon, 2010). Accordingly, the discovery of new TuMV host factors and their applications in production using modern molecular biology techniques are a trend of modern genetic engineering to confer plant disease resistance. Up to now, several TuMV resistance loci or genes have been mapped or cloned in *Brassica* vegetable crops. In *Brassica napus*, *TuRB01* and *TuRB02* were initially mapped (Walsh et al., 1999), afterward, *TuRB04* and *TuRB05* were identified by Jenner et al. (2002, 2003), followed by *TuRB03* (Hughes et al., 2003). In *Brassica rapa*, *ConTR01*, *retr01* (Rushholme, 2000; Walsh and Jenner, 2002; Rusholme et al., 2007), *BcTuR3* (Ma et al., 2010), *Rnt1-1* (Fujiwara et al., 2011), *TuRBCH01* (Xinhua et al., 2011), *retr02* (Qian et al., 2013), and *TuRB01b* (Lydiate et al., 2014), have been identified, as well as *retr03* (Shopan et al., 2017) in *Brassica juncea*. Moreover, numerous disease resistance (*R*) genes also have been identified including the NBS-LRR gene family. Several genes in this family have been reported such as *HRT*, *N*, *RCY1*, *Rx1*, *Rx2*, and *Sw5* (Kang et al., 2005). Aside from the *R* genes, the translation initiation factors *eIF4E* and *eIF4G* were also isolated as recessive resistance genes in *Arabidopsis thaliana* and *B. rapa* (Sanfaçon, 2015; Li et al., 2018). Recently, *retr02* has been suggested to be an allele of *eIF(iso)4E* (Qian et al., 2013), and that eIF4E-related resistance has been reported as the more effective type (Yeam et al., 2007; Mazier et al., 2011; Rodríguez-Hernández et al., 2012). Chung et al. (2014) constructed a genetic map based on high-throughput SNP and discovered a novel dominant TuMV resistance locus (*TuMV-R*), containing four CC-NBS-LRR resistance genes and two pathogenesis-related-1 genes. Recently, Wang et al. (2015) identified TuMV-responsive as well as new types of miRNA in non-heading Chinese cabbage through high-throughput sequencing.

Despite these excellent signs of progress, functional data remain scarce, especially in terms of finding new host factors involved in virus infection, replication, or transport. To this end, we conducted the correlation analysis by combing RNA-Seq and quantitative iTRAQ-LC-MS/MS on Chinese cabbage infected with TuMV. Most of the differentially expressed genes (DEGs) and differentially expressed proteins (DEPs) belong to the plant hormone signal transduction, plant-pathogen interaction, and protein processing in the endoplasmic reticulum pathways. Subsequently, we found many candidate resistance genes associated with TuMV infection, mainly focusing on the calcium signaling pathways, heat shock proteins, WRKY transcription factors, and lipid transfer proteins. These results can serve as an important source in research involving antiviral mechanisms and disease-resistant breeding.

MATERIALS AND METHODS

Plant Material

The Chinese cabbage (*chiifu-401-42*) used in this study is a typically sequenced cultivar and medium sensitive to TuMV infection. Seeds were pre-germinated for 2 days at 25°C and were transferred into an incubator set at 20–25°C with a 16-h light/8-h dark photoperiod and a constant 85% relative humidity. Mechanical inoculation with TuMV was performed based on a previous report (Lv et al., 2015). Briefly, the five-leaf stage seedlings were rubbed with the TuMV inoculum (pathotype C4) by grounding the TuMV-infected Chinese cabbage leaves with phosphate buffer (pH 7.4). The plants inoculated only with buffer served as control. At 21 days post-inoculation (dpi), >2.0 g each of the infected and the control leaves were collected for RNA-Seq, proteomics, and qRT-PCR verification. After collection, all samples were immediately frozen using liquid nitrogen and were stored at −80°C for further experiments. Three biological replicates were conducted for each of the subsequent analysis.

RNA Isolation and Sequencing

For Illumina sequencing, total RNA was extracted from the TuMV-infected and control leaves using TRIzol reagent (Takara Bio Inc., Otsu, Japan) following the manufacturer's protocol. RNA quantity and quality assessment were performed using the Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA, United States) and gel electrophoresis. After treatment with DNase I and enrichment using the oligo (dT) magnetic beads (for eukaryotes), the mRNA was fragmented into short fragments (about 200 bp). Subsequently, double-stranded cDNA was synthesized, and the sequence adaptors were ligated to the fragments. Finally, the Agilent 2100 Bioanalyzer and ABI StepOnePlus Real-Time PCR System were used to qualify and quantify the sample libraries. These libraries were sequenced at the Beijing Genomics Institute (Shenzhen, China[1]) using the Illumina HiSeq™ 2000 (San Diego, CA, United States).

Quantitative Real-Time PCR Validation and RT-PCR Detection

The RNA of each sample was extracted as previously described and was reverse-transcribed using the PrimeScript™ RT reagent

[1]http://www.genomics.cn/index.php

Kit with gDNA Eraser (Perfect Real Time Takara Bio Inc., Otsu, Japan). For RT-PCR, the amplification procedure used was as follows: initial denaturation at 94°C for 4 min, followed by 30 cycles of 94°C for 30 s, 55°C for 30 s, and 72°C for 30 s. Final extension was performed at 72°C for 10 min. The length of the coat protein (*CP*; GenBank: EF028235) gene segment is 204 bp. To verify the reliability of Illumina sequencing, we randomly selected 25 genes from the 757 DEGs. Their transcript levels were quantified using the $2^{-\Delta\Delta CT}$ method (Livak and Schmittgen, 2001). qRT-PCR was performed in three biological replicates following protocols used in the previous study (Lv et al., 2015). All of the specific primers for the 25 genes were designed using Beacon Designer v 7.9, and Chinese cabbage *Actin* (*Bra028615*) was used as the reference gene for quantitative expression analysis (**Supplementary Table S1**). Correlation test of RNA-Seq and qRT-PCR was analyzed by using ggpubr (v 0.4.0) R package[2].

Protein Extraction and Peptide iTRAQ Labeling

Total proteins were extracted from the TuMV-infected and control leaves. Samples were homogenized in lysis buffer solution (7 M Urea, 1 mM PMSF, 2 mM EDTA, 2 M Thiourea, 10 mM DTT, and 4% CHAPS) and were precipitated by acetone. Then, 100 μg digested proteins of each sample were labeled with different isobaric tags, and the iTRAQ tags used were as follows: CK_1-114 isobaric tag, CK_2-116 isobaric tag, CK_3-118 isobaric tag, TuMV_1-117, TuMV_2-119, and TuMV_3-121. These samples were blended and lyophilized before conducting strong cation exchange chromatography (SCX) using the Shimadzu LC-20AB HPLC system (Shimadzu, Kyoto, Japan). Each of the above components was fractionated using a Prominence LC-20AD Nano HPLC (Shimadzu, Kyoto, Japan). Finally, Q-EXACTIVE (Thermo Fisher Scientific, San Jose, CA, United States), which involves nanoelectrospray ionization followed by tandem mass spectrometry (MS/MS), was used for data acquisition. The resolution used in the Orbitrap was 70,000, and the higher-energy collision dissociation (HCD) mode (27 ± 12% collision energy) was used for MS/MS.

Bioinformatics Analysis

The Brassicaceae Genome Data v 1.5[3] was used as the reference genome. After removing data corresponding to impurities using the Trimmomatic program (v 0.33) (Kim et al., 2018), clean reads from the raw data were mapped to the reference genome using a short read alignment software SOAPaligner/SOAP2 (Li et al., 2009) with two mismatches for mapping the reads to the reference genome. The result mapped to the genome was recorded for further analysis. Expression levels of each unigene were calculated using the reads kilobases per million reads (RPKM) method (Mortazavi et al., 2008). Afterward, we used the Pearson correlation coefficients (PCC) to assess repeatability within replications. The longest transcript was used for the calculation of gene expression and coverage if there were several transcripts in a unigene. We used a false discovery rate (FDR) of ≤0.001 and the absolute value of $\log_2$Ratio ≥ 1 to screen for DEGs. Mascot v2.3.02 was used for protein identification against the database containing Bra_Chromosome_V1.5 (40796 sequences)[4]. Ratios with *p*-values < 0.05 and fold changes > 1.2 were considered as significant. Functional annotations were conducted using the Blast2GO program against the non-redundant protein database (NR; NCBI), KEGG database[5] and the COG database[6].

RESULTS

Symptom and Molecular Identification of TuMV-Infected Chinese Cabbage

Before sampling, we first determined whether all the inoculated Chinese cabbage were infected by TuMV. The phenotypes and transcription levels were used as evaluating indicators. Results showed that all the inoculated Chinese cabbage presented severe symptoms such as severe systemic vein clearing, necrosis, and stunting (**Figure 1A**). Upon obtaining the RT-PCR and qRT-PCR results, the 204 bp amplicons were observed to be quite clear in the infected samples, compared with the healthy ones, and TuMV was observed to have largely proliferated (average relative expression level = 1226.30, **Figure 1B**). The results indicated the presence of the TuMV virion in inoculated samples and showed that the appropriate time for sampling is 21 dpi.

RNA Sequencing, Mapping, and DEGs Screening

The expression levels of the transcriptome and proteome were detected by RNA-Seq and iTRAQ according to the workflow illustrated in **Figure 2A**. The results of transcriptome, proteome, and correlation analysis are illustrated in **Table 1**. A standard bioinformatics analysis was performed as follows: first, after deleting low-quality reads and adaptor sequences, we obtained 7,195,399 (CK_1, 59.83%), 7,610,342 (CK_2, 64.63%), 8,119,719 (CK_3, 65.55%), 6,486,095 (TuMV_1, 55.35%), 7,563,676 (TuMV_2, 62.89%), and 7,503,701 (TuMV_3, 62.25%) mapped reads in total, compared with the total reads, corresponding to 12,025,649, 11,775,794, 12,387,122, 11,718,227, 12,026,610, and 12,054,525 based on the Brassicaceae Genome Data v 1.5 (**Supplementary Table S2**). The average value of unique matches was 58.12% with a 17.70% mismatch (≤2 bp). Then, we performed a quality assessment of reads, and the average clean reads accounted for 99.43%. Standard deviations (SD) between the biological replicates were 0.1% (CK) and 0.05% (TuMV), respectively (**Supplementary Table S3** and **Supplementary Figure S1**), indicating that the sequencing results were of high quality and were reproducible. Next, to determine sequencing depth, our analysis results showed that the number of identified genes became saturated when the number of reads reached 12 million. In our study, the

[2]https://rpkgs.datanovia.com/ggpubr/

[3]http://brassicadb.org/brad/datasets/pub/Genomes/Brassica_rapa/V1.0/V1.5/

[4]http://brassicadb.org/brad/datasets/pub/Genomes/Brassica_rapa/V1.0/V1.5/Brapa_sequence_v1.5.fa.gz

[5]http://www.genome.jp/kegg/

[6]http://www.ncbi.nlm.nih.gov/COG/

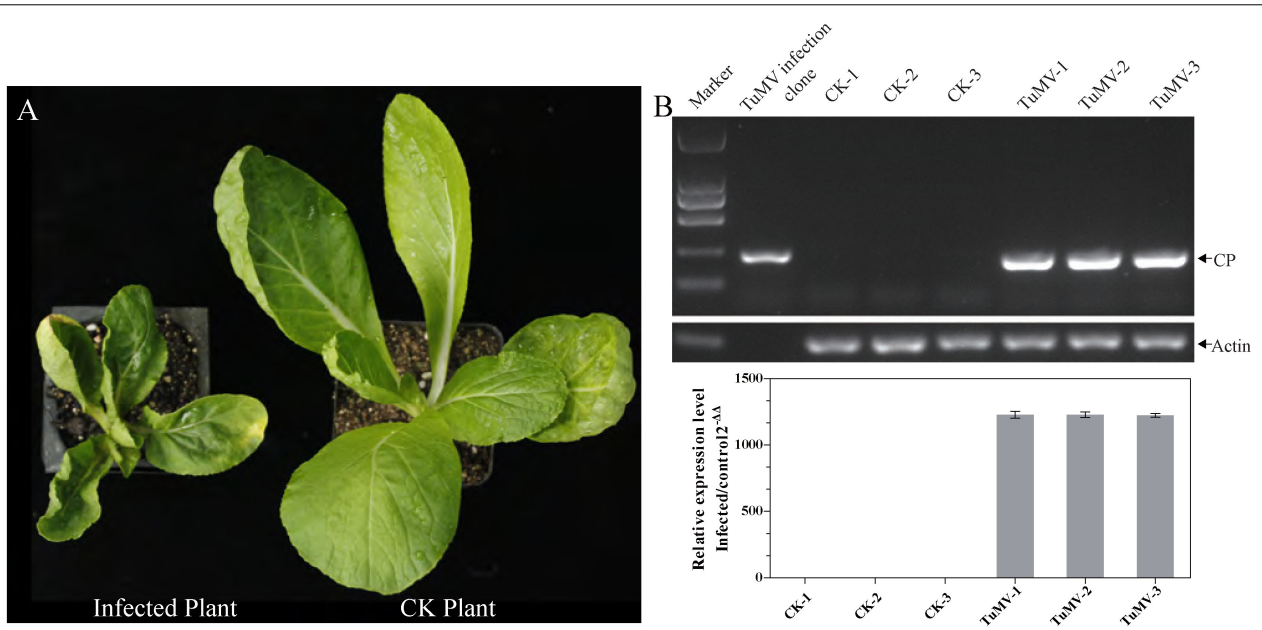


FIGURE 1 | The identification of morbidity in inoculated Chinese cabbage. **(A)** The left plant was photographed at 21 dpi, and it showed severe symptoms such as systemic vein clearing, necrosis, and stunting. The right plant served as control. **(B)** For RT-PCR detection, RNA extraction and reverse transcription were performed on each sample. *CP* expression level was quantified by qRT-PCR, and the length of *CP* segment was 204 bp. *Actin 7* (*Bra028615*) was used as an internal reference gene.

number of each sample exceeded 6 million (**Supplementary Figure S2**). Sequencing libraries were built according to the method reported by Wang et al. (2009), which led to an increased randomness (**Supplementary Figure S3**). Gene coverage is also an important assessment index of sequencing quality. Here, more than 50% of genes corresponded to over 80% of the gene coverage (**Supplementary Figure S4**). Finally, in the statistical analysis of gene expression levels, we applied the RPKM method to estimate transcript abundance, as well as all the other information, including gene length, RPKM, gene coverage, $\log_2$ratio (TuMV/CK), and probability; these are shown in **Supplementary Table S4**.

Moreover, to screen DEGs, we used the NOIseq method (Tarazona et al., 2011) according to its good true positive and false-positive rates. Finally, we obtained 757 remarkable DEGs following a filtering condition of fold change $\geq$ 2 and probability $\geq$ 0.8 (**Figure 2B** and **Supplementary Table S5**). Among them, 322 genes were found to be upregulated and 435 genes were downregulated.

qRT-PCR Based Validation

For assessment of the RNA-Seq data, RNA-Seq libraries were subjected to qRT-PCR. First, 25 genes were selected randomly from the DEGs which were related to the AP2-EREBP transcription factors, basic helix-loop-helix (bHLH) transcription factors, chloroplast and mitochondria gene families, cytochrome P450, MAP kinase (MAPK) families, MYB gene families, NAC transcription factor families, and WRKY transcription factor families. The qRT-PCR results of these metabolism and disease resistance genes coincided with the RNA-Seq data (r = 0.84, p-value = 2e-07), demonstrating the high reliability of RNA-Seq data (**Figure 3**).

Annotation and Functional Classification

For an improved understanding of the gene functions, all the identified unigenes and proteins, DEGs, and DEPs were annotated by aligning the Nr database of NCBI to GO terms using BLAST2GO (default parameters) (Conesa et al., 2005). Enrichment was performed by WEOGO (Ye et al., 2006). Genome-wide transcriptome or proteome expression levels were used as background.

The results of the DEG GO reassignments showed that all unigenes were categorized into 44 GO terms in three ontologies, consisting of biological process (23 subcategories), cellular component (9 subcategories), and molecular function (12 subcategories) (**Figure 4A** and **Supplementary Table S6**). In biological process, 'percentages of genes in the rhythmic process,' 'response to stimulus process,' 'multi-organism process,' 'metabolic process,' 'immune system process,' 'biological process,' and 'cellular process' were higher than the genome-wide transcriptome expression levels. In the classification of biological process, it was found that this category exerted repercussions on TuMV infection. In molecular function, 'molecular function transporter activity' and 'binding' were the two predominant subcategories, and the proportion of 'auxiliary transport protein' in the DEGs was significantly higher than that of the background. This suggested that more auxiliary transport proteins were involved in the process of TuMV infection. In the cellular component, the 'extracellular' and 'extracellular region' were the dominant subcategories. There

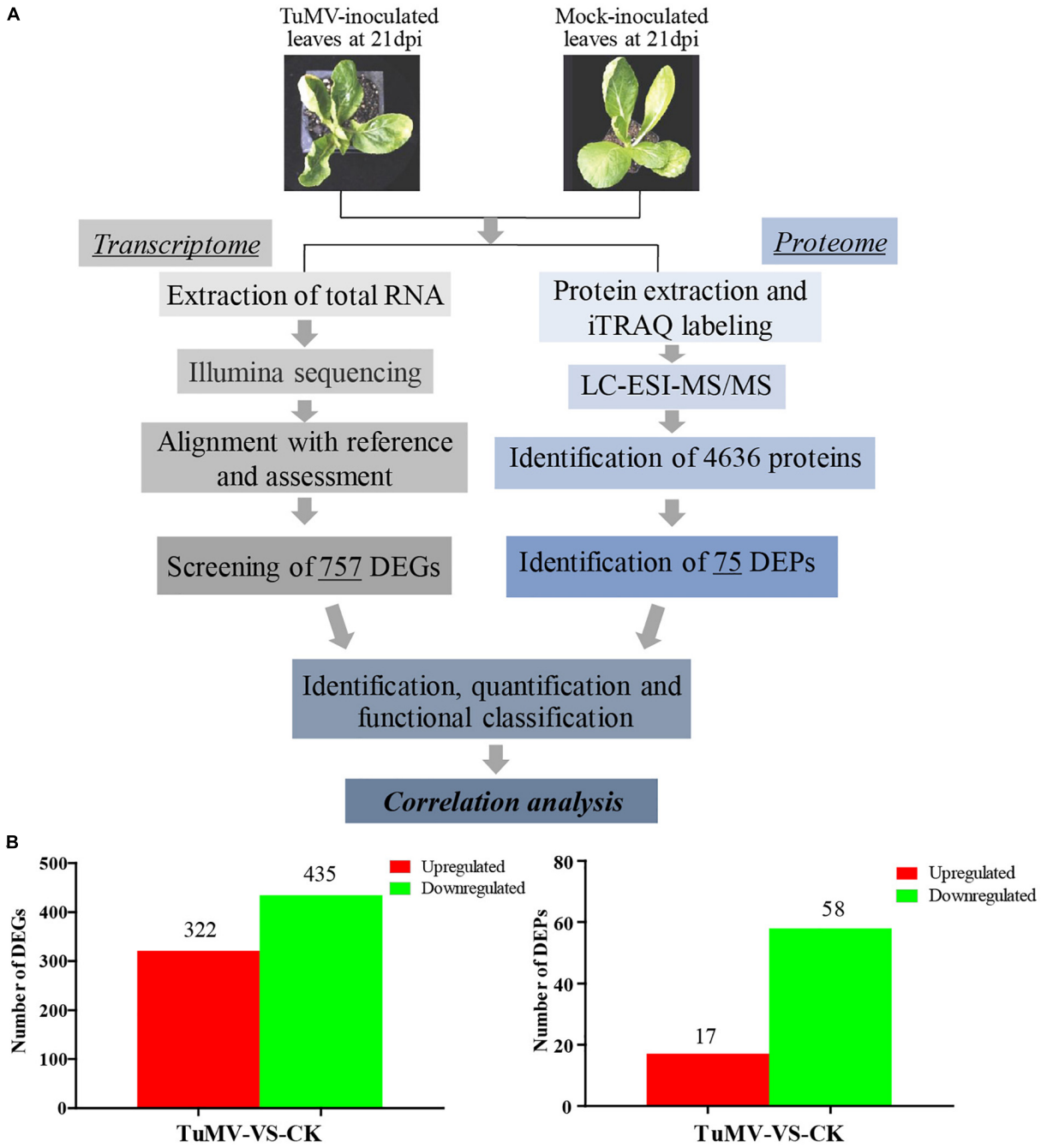


FIGURE 2 | Experimental workflow and summary of the present study. **(A)** Transcriptomic and proteomic methods that were used to analyze correlation and to find viral disease-related genes. **(B)** Significant differentially expressed genes screened at the transcription and proteome levels.

was no significant difference between DEGs and genome-wide genes in this part.

To identify possible pathways related to TuMV infection, all the unigenes were annotated by the Kyoto Encyclopedia of Genes and Genomes (KEGG) database[5] (Kanehisa et al., 2007) (**Supplementary Table S7**), and the pathways with *Q* value $\leq$ 0.05 were considered as significantly enriched. The top 20 enrichment terms are shown in **Figure 5**. The majority of these pathways are metabolic pathways (125/4990 genes), biosynthesis of secondary metabolites (81/2757 genes), plant hormone signal transduction (56/1555 genes), circadian rhythm (36/269 genes), and plant-pathogen interaction (34/1633 genes) pathways. In addition, protein processing in the endoplasmic reticulum (10/610 genes), regulation of autophagy (6/156 genes), and endocytosis (4/265 genes) may be involved in TuMV assembly and transport.

TABLE 1 | Summary of results of the transcriptome, proteome, and correlation analysis.

Group name	Type	Number of proteins	Number of genes	Number of correlations
TuMV-VS-CK	Identification	4636	31513	4592
TuMV-VS-CK	Quantitation	3498	31513	3476
TuMV-VS-CK	Differential Expression	75	757	12

Proteomics Characterized by iTRAQ and Data Analysis

The Mascot search engine (Matrix Science, London, United Kingdom; version 2.3.02) was used in protein identification against the Bra_Chromosome_V1.5 database containing 40,796 sequences. The quality error distribution of spectrogram matching was conducted to assess the identifying quality (**Supplementary Figure S5A**). 92,152 spectra out of 366,867 total spectra were matched after quality control, and among them, 60,115 were considered as unique spectrums. A total of 13,348 unique peptides were identified, and 4,636 proteins were obtained (**Supplementary Figure S5B**). The protein mass distribution ranged from 7.608 to 592.995 kDa, corresponding to *Bra005963* and *Bra013017*, and the percentage of >20 kDa accounted for 93% (4316/4636) (**Supplementary Figure S5C**). Since stable results can be obtained based on the stability of the mass spectrum experiment process, we conducted repeated analysis (**Supplementary Figure S5D**). The obtained mean coefficient of variation (CV) was 9.6%, suggesting that the data were reliable for further analysis. For protein abundance, at least two unique spectra contained in a protein was set as the standard. Proteins with p-values < 0.05 and fold changes of > 1.2 were considered as DEPs (**Supplementary Figure S5E**). Under these thresholds, a total of 75 DEPs were obtained, while 17 proteins were upregulated and 58 were downregulated (**Figure 2B** and **Supplementary Table S8**).

The results of GO reassignments showed that proteins were also categorized into three ontologies (**Figure 4B** and **Supplementary Table S9**); however, compared with the all identified proteins, ratios of 'response to stimulus,' 'multi-organism process,' 'immune system process,' and 'death' in 'biological process' were slightly higher. Compared with the DEG GO results, these were all high in 'response to stimulus,' 'multi-organism process,' and 'immune system process.' In the molecular function component, 'binding and catalytic' were high, but no gene was detected in 'auxiliary transport protein.' In the cellular component, 'extracellular region' was the only term in which the significant protein ratios were higher than that of all identified proteins. All the DEPs were also annotated to 45 pathways by retrieving the KEGG database (**Supplementary Table S10**). There were seven pathways that contained greater than or equal to four proteins, namely, metabolic pathways (17, 32.08%), biosynthesis of secondary metabolites (11, 20.75%), nitrogen metabolism (5,

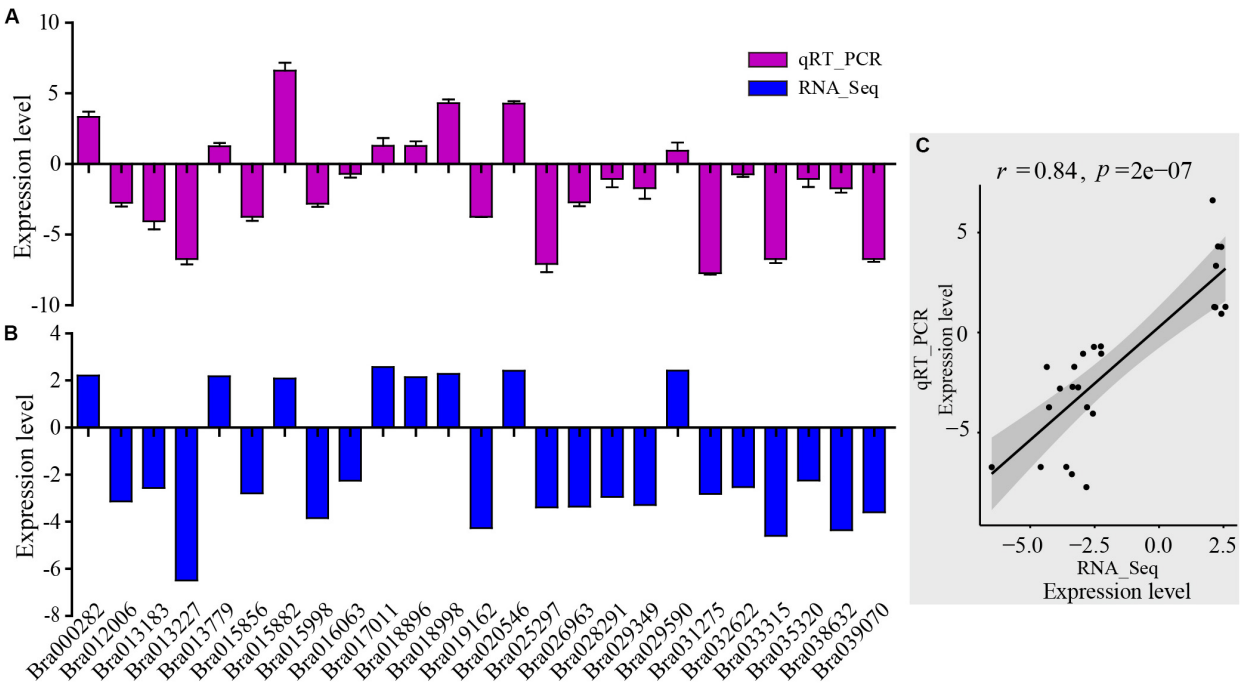


FIGURE 3 | Validation of RNA-Seq results by qRT-PCR. **(A)** qRT-PCR expression levels were determined based on the blue bar. **(B)** RNA-Seq expression levels were determined based on the purple bar. **(C)** Scatter plots of correlation test between qRT-PCR and RNA-Seq results. r is Pearson correlation coefficient. The p-value is the significance level of the correlation. There were 25 genes randomly selected from the DEGs related to AP2-EREBP transcription factors, basic helix-loop-helix (bHLH) transcription factors, chloroplast and mitochondria gene families, cytochrome P450, MAP kinase (MAPK) families, MYB gene families, NAC transcription factor family, and WRKY transcription factor family.

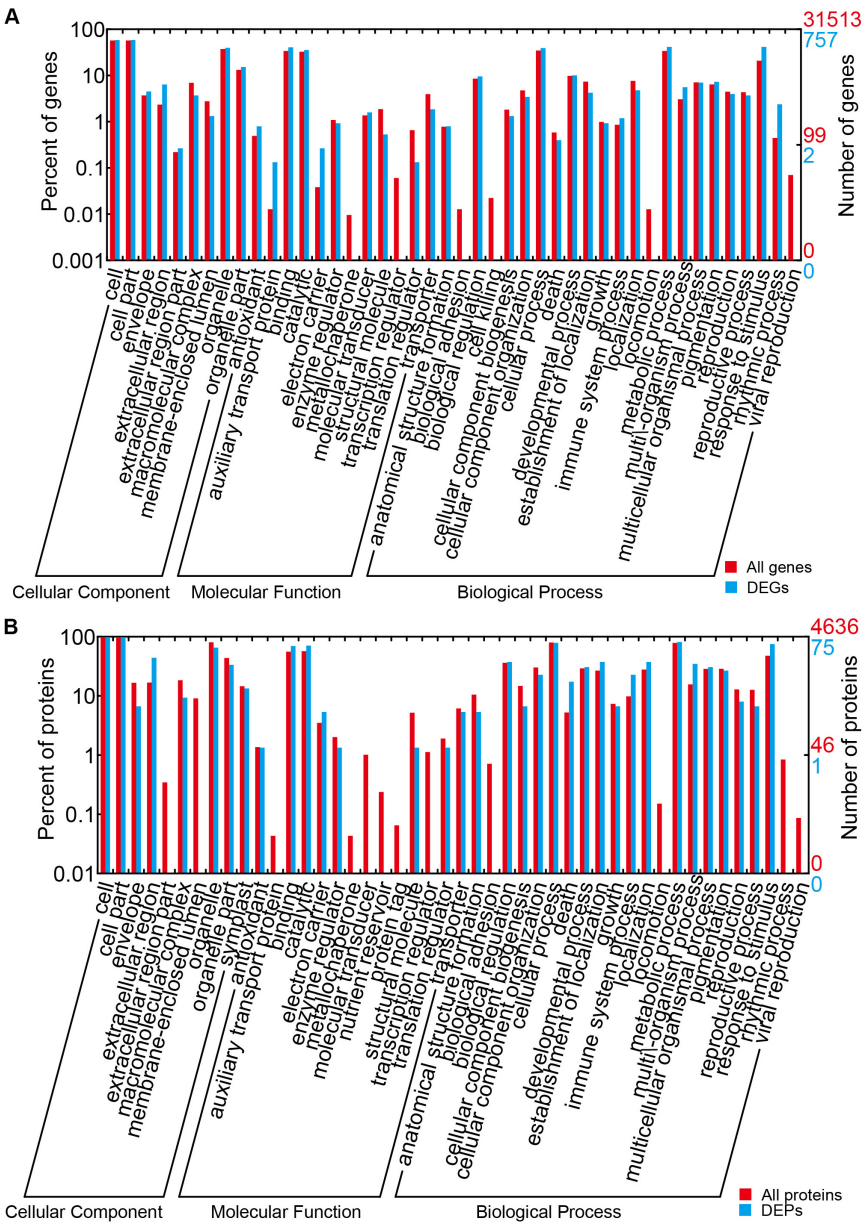


FIGURE 4 | GO term assignment of DEGs and DEPs in Chinese cabbage. **(A)** GO categories for DEGs in the transcriptome. **(B)** GO categories for DEPs in the proteome. The whole transcriptome or proteome was used as the background. Results were summarized for the three main GO categories: biological process (P), molecular function (F), and cellular component (C).

9.43%), pyruvate metabolism (5, 9.43%), arginine and proline metabolism (4, 7.55%), plant hormone signal transduction (4, 7.55%), and plant–pathogen interaction (4, 7.55%).

Correlation Analysis of Transcriptome and Proteome

To explore the congruence of fold change between the transcriptome and proteome data, we conducted correlation analysis. The quantitative relation of the proteins and genes are shown in **Table 2**. The filter criteria of proteins and genes used was fold change $\geq$ 1.2, p-value $\leq$ 0.05, fold change $\geq$ 2, and FDR $\leq$ 0.001.

After correlation analysis of the expression data at the proteomic and at the transcriptional level, we found that correlation between the two was low ($R = -0.0211$) (**Figure 6A** and **Supplementary Table S11**), and that the genes that can be detected in the proteome but not in the transcriptome and those that were detected in the transcriptome, were higher (**Supplementary Tables S12, S13**). This inferior correlation between the mRNA and protein abundance ratios showed the inconsistencies between transcription and translation in the plant

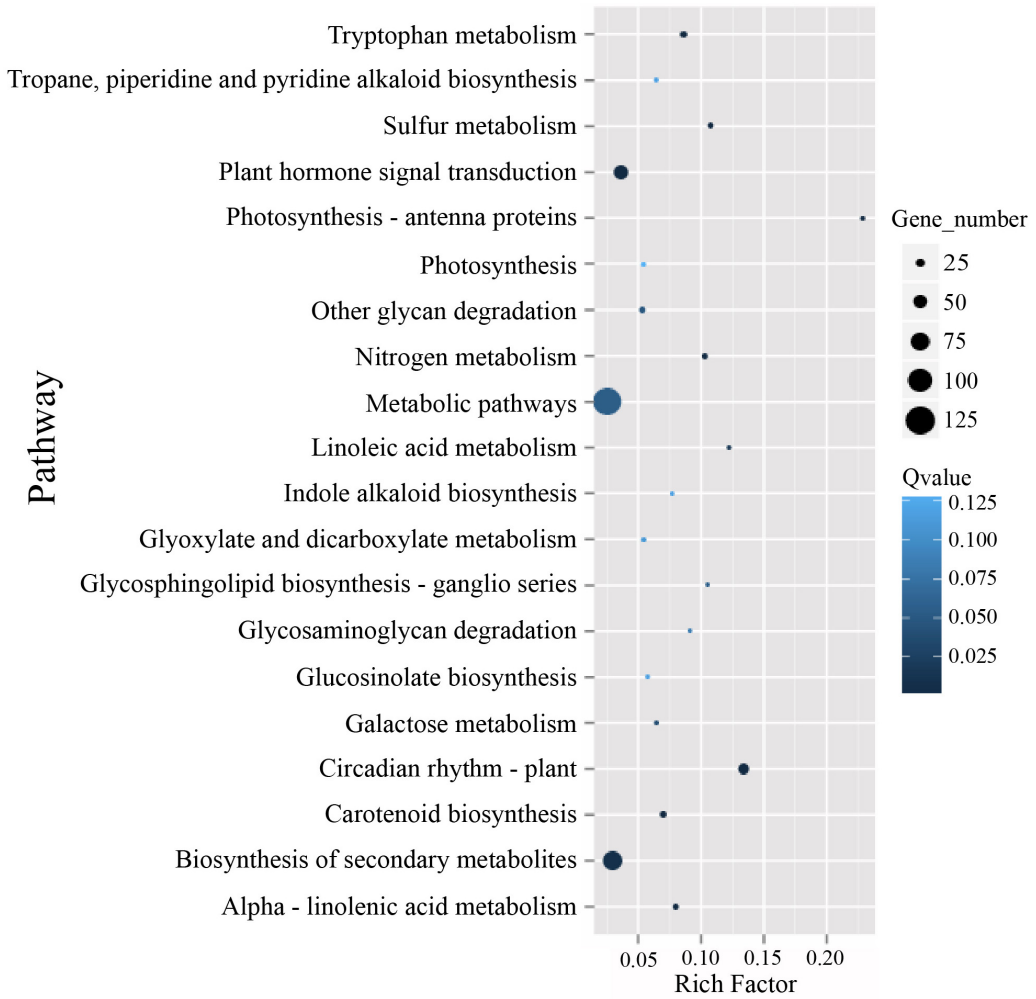


FIGURE 5 | The KEGG enrichment scatterplot of the top 20 statistically significant pathways plotted for the DEGs. The rich factor on the horizontal refers to the ratio between the gene numbers of DEGs which are located at the pathway entry as well as the total gene numbers of all annotated genes which are located at the pathway entry. *Q*-value was determined by multiple hypothesis testing corrections, and their values vary from 0 to 1.

system; moreover, temporal and spatial discrepancies in the function and the complexity of biological regulatory networks can also explain this result. Although the overall correlation degree was low, there were also a few genes that had a high correlation with the proteins. *Bra016926*, *Bra020322*, *Bra022468*, and *Bra006721* have the same trend in the changes of the *R* value = 0.8000 (**Figures 6B,D** and **Table 2**). Additionally, there were eight genes (*Bra028091*, *Bra000876*, *Bra025730*, *Bra003273*, *Bra000315*, *Bra036259*, *Bra004771*, and *Bra015656*) that showed an opposite change trend, with *R* value = −0.4762 (**Figures 6C,D** and **Table 2**).

Transcriptome and Proteome Data Mining

For further analysis, DEGs were then mapped to their predicted pathways (**Supplementary Table S7**). A total of 56 genes were mapped to the plant hormone signal transduction pathway. In this group, genes related to auxin were the majority, including Gretchen Hagen 3 (GH3) family genes and small auxin upregulated RNAs (SAUR) family genes. A clear majority of genes in this group were significantly downregulated. The cytokinin signal pathway-related genes ranked second, containing three cytokinin receptors and seven *Arabidopsis* response regulator (ARR) family genes. Most genes in this group were upregulated. The last major group was related to jasmonate (jasmonate ZIM domain-containing protein), and all the genes in this group were downregulated. Moreover, 34 genes were mapped to the plant–pathogen interaction pathway. The group of jasmonate related genes were also present in the hormone signal transduction pathway. There were five LRR receptor-like serine/threonine-protein kinase-related genes, and three of these were downregulated and two were upregulated. There were five WRKY transcription factors and only one was downregulated. As protein processing in the endoplasmic reticulum is important for virus reproduction and assembly, we also chose to focus on

TABLE 2 | Correlation analysis between the transcriptome and proteome.

Protein ID	Quantitation (TuMV/CK)	Sig	Diff Protein	Gene ID	log2(TuMV/CK)	Probability	Diff gene
Same trend							
Bra016926	1.28	*	+	Bra016926	1.57	0.84	+
Bra020322	1.78	*	+	Bra020322	1.63	0.86	+
Bra022468	0.75	*	−	Bra022468	−1.29	0.80	−
Bra006721	1.54	*	+	Bra006721	2.39	0.90	+
Different trend							
Bra028091	0.64	*	−	Bra028091	2.05	0.88	+
Bra000876	0.60	*	−	Bra000876	2.32	0.90	+
Bra025730	0.57	*	−	Bra025730	2.25	0.90	+
Bra003273	0.69	*	−	Bra003273	1.58	0.85	+
Bra000315	0.71	*	−	Bra000315	1.29	0.82	+
Bra036259	0.78	*	−	Bra036259	3.38	0.83	+
Bra004771	0.53	*	−	Bra004771	2.79	0.90	+
Bra015656	1.24	*	+	Bra015656	−1.22	0.81	−

**: Significance; +: upregulated; −: downregulated.*

this pathway. In this group, heat shock proteins (HSPs) were the majority, and most of the HSPs belonged to HSP20. In addition, we also mapped six genes to the regulation of autophagy pathway, and they were all downregulated. Since virus replication mainly occurs in the intracellular membrane structure, the genes in the protein processing in endoplasmic reticulum pathway may be related to virus replication. Eight out of ten genes in this group belonged to the HSPs. As previously reported, the autophagy pathways are related to *Potyvirus* infection (Cheng and Wang, 2017), thus, we also paid attention to this pathway. We found that it consisted of six genes and that all of them were downregulated.

In addition, there were also many disease-related genes found in the DEPs. In the upregulated proteins, there were four non-specific lipid transfer proteins (nsLTPs) (Bra020323, Bra020322, Bra006721, and Bra000377) that were predicted to encode pathogenesis-related (PR) proteins and EF1B (Bra008915), which is a translation elongation factor. Among the downregulated genes, there were three beta-1,3-glucanase genes, which have been reported to be induced by plant viruses and can accumulate in the vesicles and the cell wall to hydrolyze callose, resulting in the opening of the plasmodesmata (Epel, 2009). Although the correlation between proteome and transcriptome was low, there were still some valuable candidate genes. Bra006721 (nsLTP3) and Bra020322 (nsLTP4) present the same expression trend (upregulated) in the proteome and transcriptome. Bra025730 (pathogenesis-related thaumatin superfamily protein), Bra003273 (BETA-1, 3-GLUCANASE 3), Bra000315 (chitinase), Bra036259 (chitinase), and Bra004771 all showed an opposite expression trend (downregulated in the proteome, upregulated in the transcriptome) (**Table 2**). Based on the transcriptome and proteome data, we summarized the candidate genes and pathways related to TuMV infection on a preliminary network (**Figure 7**) that mainly focused on calcium signaling pathways, HSPs, WRKY transcription factors, and non-specific lipid transfer proteins.

DISCUSSION

Plant Hormones Are Correlated to TuMV Resistance

Many viral diseases can lead to developmental abnormalities and aberrant phenotypes. This phenomenon may be induced by disruptions in hormone metabolism. The past decades witnessed an increase in the studies that demonstrated that viral infections can affect the production and distribution of many plant hormones such as auxin, cytokines, and gibberellins (Alazem and Lin, 2015). An interaction has been found between *Tobacco mosaic virus* (TMV) infection and the auxin-responsive pathway, which can influence the development of disease symptoms (Padmanabhan et al., 2005). Herein, we were able to obtain several auxin-responsive genes that may be related to this pathway. Moreover, since the SAUR family proteins were previously reported to be related to elongation (Gil et al., 1994), their downregulated expression may explain the growth stunting. Combined, during viral infection, a set of Aux/IAA proteins can interact with TuMV and disrupt their normal function. Since there was an antagonistic effector of SA-mediated defense signaling, the ABA levels were originally considered to be negatively correlated with disease resistance (Thaler and Bostock, 2004). However, a positive role played by ABA in increasing pathogen resistance was later revealed, and this was conferred by the induction of callose deposition at plasmodesmata, which can inhibit viral cell-to-cell movement (Mauch-Mani and Mauch, 2005). Here, we found five abscisic acid receptor PYR/PYL family genes which function at the apex of a passive regulator pathway controlling ABA signaling (Park et al., 2009).

Several Resistant Genes Are Related to TuMV Infection

In the plant-pathogen interaction category, we found many resistance genes such as LRR related genes, WRKY genes,

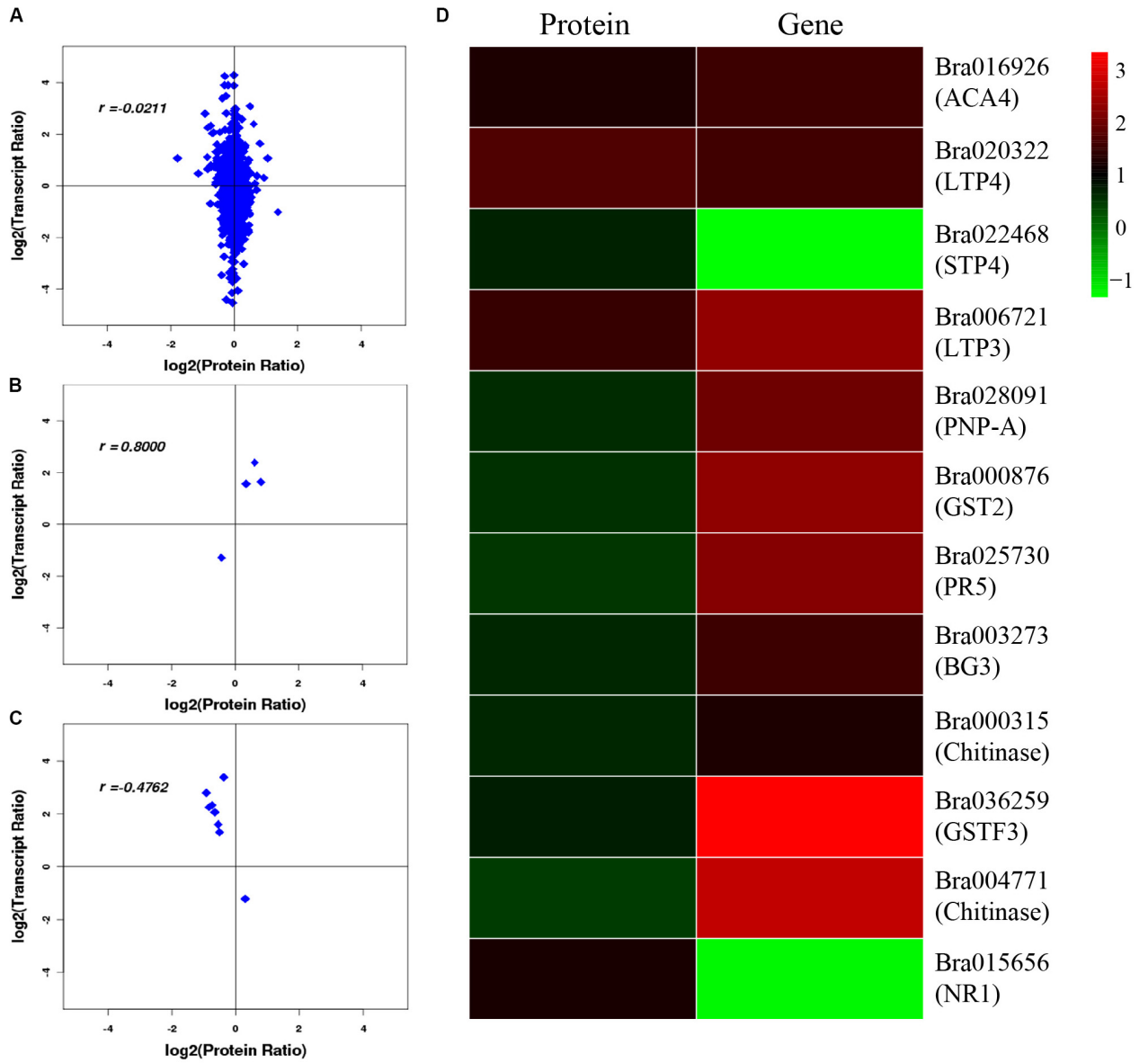


FIGURE 6 | Correlation analysis between transcriptome and proteome. **(A)** Correlation analysis of all DEGs and DEPs. **(B)** Correlation analysis of DEGs and DEPs with similar changing trends. **(C)** Correlation analysis of DEGs and DEPs with opposite changing trends. **(D)** Heatmap for 12 pairs of DEGs and DEPs. Red represents upregulated genes and green represents downregulated genes. Gene ID and gene names were attached on the right. ACA4: Autoinhibited Ca(2+)-atpase, isoform 4, LTP4: Lipid transfer protein 4, STP4: Sugar transporter 4, LTP3: Lipid transfer protein 3, PNP-A: Plant natriuretic peptide A, GST2: Glutathione s-transferase phi 2, PR5: Pathogenesis-related protein 5, PR2/BG2: Pathogenesis-related protein 2/beta-1,3-glucanase 2, Chitinase: Chitinase family protein, GSTF3: Glutathione s-transferase f3, Chitinase: Chitinase, NR1: Nitrate reductase 1. For the correlation analysis diagram **(B–D)**, the levels of the DEPs were labeled on the horizontal axis, and the levels of the DEGs were labeled on the vertical.

and transcription factor MYC genes. Leucine-rich repeat (LRR) proteins have been reported to be related to plant defense and in the resistance against various pathogens including fungi, bacteria, and viruses (Jones, 1997). Peng et al. (2013) cloned six members of the LRR II subfamily from non-heading Chinese cabbage and found that these genes were induced by TuMV infection. All the LRR genes identified here were flagellin-sensitive 2 (FLS2), which had structural and functional homologies with the other resistance genes and were also involved in the innate immune system (Gómez-Gómez and Boller, 2000). WRKY22 found in this study has been reported to be related to the MAPK pathway, which can confer resistance to both bacteria and fungi (Asai et al., 2002). Furthermore, WRKY33 was reported to be required for SA-dependent defense gene expression (Loake and Grant, 2007), thus, they may also be associated with TuMV resistance. MYC2 is an important bHLH transcription factor that can regulate the expression of JA-responsive genes (Zhai et al.,

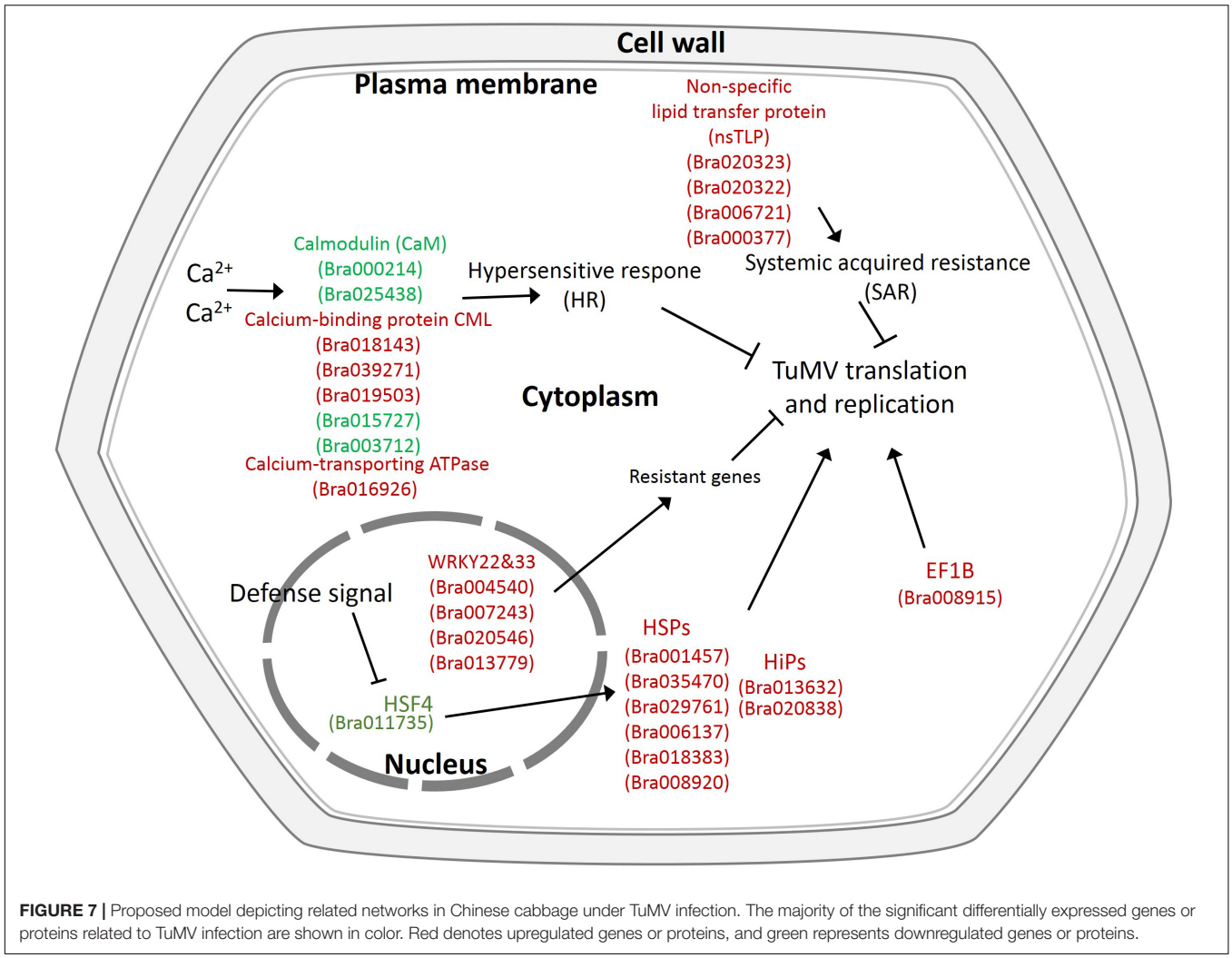


FIGURE 7 | Proposed model depicting related networks in Chinese cabbage under TuMV infection. The majority of the significant differentially expressed genes or proteins related to TuMV infection are shown in color. Red denotes upregulated genes or proteins, and green represents downregulated genes or proteins.

2013). Two MYC2 genes identified were both downregulated, consistent with the jasmonate ZIM domain-containing genes mentioned above. This may indicate that the JA pathway is involved in the TuMV infection process. Furthermore, two Calmodulin genes (CaMs; *Bra000214*, *Bra025438*) and five Calcium-like genes (CMLs; *Bra018143*, *Bra039271*, *Bra019503*, *Bra015727*, and *Bra003712*) were identified in this study. CaMs and CMLs are primary calcium sensors in all eukaryotes, and these have been reported to boost antiviral defense by causing rapid hypersensitive cell death either by forming necrotic lesions or by mediating RNA silencing (Yamakawa et al., 2001; McCormack and Braam, 2003; Perochon et al., 2011; Nakahara et al., 2012).

Several Host Factors Are Associated With TuMV Multiplication

As obligate intracellular parasites, viruses need to recruit host factors for their infection cycle, and the lack of certain host factors can confer resistance to the host (Truniger and Aranda, 2009; Nagy and Pogany, 2012). In terms of the DEGs, we found several HSP70s. Aside from plant closteroviruses, viruses do not have their own HSP70 genes, although many have been reported to recruit host HSP70s to assist in virion assembly, replication, and cell-to-cell movement (Mine et al., 2012). All the HSP70s (*Bra001457*, *Bra035470*, and *Bra029761*) were upregulated, which means they may take part in certain fundamental phases in the plant virus life cycles. A DnaJ homolog gene (*Bra008607*), functioning as part of the Hsp70 chaperone and can inhibit the *Brome mosaic virus* (BMV) negative-strand RNA synthesis (Wang et al., 2004), was identified and was also upregulated. Additionally, HSP20s, which are small heat shock proteins (sHSPs) that have the same function as HSP70, were also identified. Both *OsHSP20* and *NbHSP20* have been demonstrated to interact with RNA-dependent RNA polymerase (RdRp) of *Rice stripe virus* (RSV) and their sub-location and expression were significantly influenced by RSV (Li et al., 2015). As the HSP family is an abundant and complex group (Wang et al., 2004), they may exercise different functions, which can explain their inconsistent expression levels. As for the DEPs, the translation elongation factor EF1B (Bra008915) had also been identified in

Nicotiana benthamiana as a protein required for TMV infection (Hwang et al., 2013).

Some nsLTPs May Improve Plant Virus Resistance

We discovered many candidate genes and proteins related to TuMV infection in the correlation analysis of the transcriptome and proteome. nsLTPs are ubiquitous in animals, plants, fungi, and bacteria (Bureau et al., 1996). Although the lipid transport protein should promote virus replication, because of its participation in the assembly of viral replication (de Castro et al., 2016), nsLTPs have also been documented to assist in intercellular lipid transport and in the formation and accumulation of wax, thereby protecting against pathogens (Kader, 1996; Cameron et al., 2006). In this study, we found that nsLTP3 (Bra020323 and Bra006721) and nsLTP4 (Bra020322) were also significantly upregulated at the proteome level. Currently, there have been many studies that have demonstrated the disease resistance function of nsLTPs, such as in the resistance to fungi and bacteria in barley, maize, spinach, radish, and *Arabidopsis* (Terras et al., 1992; Molina and García-Olmedo, 1993; Segura et al., 1993). Consistent with this study, Park et al. found that *Tobacco mosaic virus* (TMV-P0) can induce *CaLTP1* expression (Park et al., 2002). Sujuon et al. also found that the silencing of *CALTPI* and *CALTPII* by VIGS can increase susceptibility to *Xanthomonas campestris* pv. *vescatoria* as well as to *Pepper mosaic mottle virus* (PMMV), and that the overexpression of *CALTPI* and *CALTPII* can enhance resistance (Sarowar et al., 2009). Moreover, nsLTPs were suggested to be involved in long-distance signaling by functioning as the systemic mobile signal for systemic acquired resistance (SAR) in plants (Sarowar et al., 2009). Despite knowledge of several biological functions of nsLTPs, their precise function(s) remains elusive; however, based on recent studies, nsLTPs have been verified to be irrelevant to intracellular lipid transport, due to their intercellular localization and intercellular secretion. For example, Ace-AMP1, which is homologous to plant nsLTPs and is isolated from onion (*Allium cepa* L.) seed, cannot transport phospholipids but have antifungal activity (Cammue et al., 1995; Tassin et al., 1998). There were several nsLTPs whose lipid-binding activities were not correlated with their antifungal activities (Sun et al., 2008). Although direct evidence is lacking, based on the enhanced susceptibility observed in this study and those in the pepper CALTPI and CALTPII study and in the *Arabidopsis* DIR1 study, it is logical to speculate that nsLTPs might serve as transporters for mobile signals during TuMV infection.

The Correlation Between Transcriptome and Proteome in TuMV Infection

To better elucidate the infection mechanism, we analyzed the correlation between the transcriptome and proteome by obtaining the PCC (**Figure 6**). Overall, the correlation between the transcriptome and proteome was rather low ($R = -0.0211$), and for those selected due to their significantly changed transcriptome and proteome, the same change trend was calculated to be $R = 0.8000$ and the opposite change trend was $R = -0.4762$. As previously reported, there was no necessary correlation between transcriptome and proteome, and normally the correlation was very low (Gygi et al., 1999; Foss et al., 2007; Ghazalpour et al., 2011). But there were exceptions, for example, an overall moderate correlation was obtained between transcriptome and proteome patterns of *Pseudomonas syringae* pv. *tomato* DC3000 *in planta* bacterial multi-omics (Nobori et al., 2020). Based on the fact that the transcriptome is the intermediate state of gene expression and that the proteome is the final state of gene expression, the two types of expression levels generally correspond to each other. On the other hand, organisms can also make full use of this network to regulate protein abundance through methods such as DNA methylation, RNA silencing, and protein degradation. Therefore, most data that lie between the transcriptome and proteome are relevant; furthermore, there exist a few genes that are not relevant due to these regulatory mechanisms. There are other abiotic factors that influence data correlation such as differences in the technical systems, differences between data types and methods, and the effects of proteins with high abundance.

DATA AVAILABILITY STATEMENT

Our RNA-sequencing data have been submitted to the GEO repository. The GEO accession number is GSE151932.

AUTHOR CONTRIBUTIONS

SL, CZ, LG, and XH conceived the study. SL, LG, and RZ completed the experiments. SL, CZ, and LG contributed to data analysis and manuscript preparation. CZ and XH participated in the planning of experiments and revising the manuscript. All the authors read and approved the final version of the manuscript.

FUNDING

This work was supported by the grants from the Jiangsu Agricultural Science and Technology Innovation Fund [CX (18) 3068]; the State Key Special Program 'Seven Main Crops Breeding' (2016YFD0101701); and the Nature Science Foundation of Jiangsu Province (BK20191308). SL was supported by a graduate fellowship from the China Scholarship Council.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00963/full#supplementary-material

REFERENCES

Alazem, M., and Lin, N. S. (2015). Roles of plant hormones in the regulation of host–virus interactions. *Mol.* Plant Pathol. 16, 529–540. doi: 10.1111/mpp.12204

Asai, T., Tena, G., Plotnikova, J., Willmann, M. R., Chiu, W.-L., Gomez-Gomez, L., et al. (2002). MAP kinase signalling cascade in *Arabidopsis* innate immunity. *Nature* 415, 977–983.

Bureau, T. E., Ronald, P. C., and Wessler, S. R. (1996). A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes. *Proc. Natl. Acad. Sci. U.S.A.* 93, 8524–8529.

Cameron, K. D., Teece, M. A., and Smart, L. B. (2006). Increased accumulation of cuticular wax and expression of lipid transfer protein in response to periodic drying events in leaves of tree tobacco. *Plant Physiol.* 140, 176–183.

Cammue, B. P. A., Thevissen, K., Hendriks, M., Eggermont, K., Goderis, I. J., Proost, P., et al. (1995). A potent antimicrobial protein from onion seeds showing sequence homology to plant lipid transfer proteins. *Plant Physiol.* 109, 445–455.

Cao, J., Yu, X., Ye, W., Lu, G., and Xiang, X. (2006). Functional analysis of a novel male fertility CYP86MF gene in Chinese cabbage (*Brassica campestris* L. ssp. chinensis Makino). *Plant Cell Rep.* 24, 715–723.

Cheng, F., Liu, S., Wu, J., Fang, L., Sun, S., Liu, B., et al. (2011). BRAD, the genetics and genomics database for Brassica plants. *BMC Plant Biol.* 11:136. doi: 10.1186/1471-2229-11-136

Cheng, X., and Wang, A. (2017). The potyvirus silencing suppressor protein VPg mediates degradation of SGS3 via ubiquitination and autophagy pathways. *J. Virol.* 91:e01478-16.

Chung, H., Jeong, Y.-M., Mun, J.-H., Lee, S.-S., Chung, W.-H., and Yu, H.-J. (2014). Construction of a genetic map based on high-throughput SNP genotyping and genetic mapping of a TuMV resistance locus in *Brassica rapa*. *Mol. Genet. Genomics* 289, 149–160.

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21, 3674–3676.

de Castro, I. F., Tenorio, R., and Risco, C. (2016). Virus assembly factories in a lipid world. *Curr. Opin. Virol.* 18, 20–26.

Epel, B. L. (2009). Plant viruses spread by diffusion on ER-associated movement-protein-rafts through plasmodesmata gated by viral induced host β-1, 3-glucanases. *Semin. Cell Dev. Biol.* 20, 1074–1081.

Foss, E. J., Radulovic, D., Shaffer, S. A., Ruderfer, D. M., Bedalov, A., Goodlett, D. R., et al. (2007). Genetic basis of proteome variation in yeast. *Nat. Genet.* 39, 1369–1375.

Fraser, R. (1990). The genetics of resistance to plant viruses. *Annu. Rev. Phytopathol.* 28, 179–200.

Fujiwara, A., Inukai, T., Kim, B. M., and Masuta, C. (2011). Combinations of a host resistance gene and the CI gene of *Turnip mosaic virus* differentially regulate symptom expression in *Brassica rapa* cultivars. *Arch. Virol.* 156, 1575–1581.

Ghazalpour, A., Bennett, B., Petyuk, V. A., Orozco, L., Hagopian, R., Mungrue, I. N., et al. (2011). Comparative analysis of proteome and transcriptome variation in mouse. *PLoS Genet.* 7:e1001393. doi: 10.1371/journal.pgen.1001393

Gil, P., Liu, Y., Orbovic, V., Verkamp, E., Poff, K. L., and Green, P. J. (1994). Characterization of the auxin-inducible SAUR-AC1 gene for use as a molecular genetic tool in *Arabidopsis*. *Plant Physiol.* 104, 777–784.

Gómez-Gómez, L., and Boller, T. (2000). FLS2: an LRR receptor–like kinase involved in the perception of the bacterial elicitor flagellin in *Arabidopsis*. *Mol. Cell* 5, 1003–1011.

Gygi, S. P., Rochon, Y., Franza, B. R., and Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* 19, 1720–1730.

Hughes, S., Green, S., Lydiate, D., and Walsh, J. (2002). Resistance to *Turnip mosaic virus* in *Brassica rapa* and *B. napus* and the analysis of genetic inheritance in selected lines. *Plant Pathol.* 51, 567–573.

Hughes, S., Hunter, P., Sharpe, A., Kearsey, M., Lydiate, D., and Walsh, J. (2003). Genetic mapping of the novel *Turnip mosaic virus* resistance gene TuRB03 in *Brassica napus*. *Theor. Appl. Genet.* 107, 1169–1173.

Hwang, J., Oh, C.-S., and Kang, B.-C. (2013). Translation elongation factor 1B (eEF1B) is an essential host factor for *Tobacco mosaic virus* infection in plants. *Virology* 439, 105–114.

Jenner, C. E., Tomimura, K., Ohshima, K., Hughes, S. L., and Walsh, J. A. (2002). Mutations in *Turnip mosaic virus* P3 and cylindrical inclusion proteins are separately required to overcome two *Brassica napus* resistance genes. *Virology* 300, 50–59.

Jenner, C. E., Wang, X., Tomimura, K., Ohshima, K., Ponz, F., and Walsh, J. A. (2003). The dual role of the potyvirus P3 protein of *Turnip mosaic virus* as a symptom and avirulence determinant in brassicas. *Mol. Plant Microbe Interact.* 16, 777–784.

Jones, J. (1997). The role of leucine-rich repeat proteins in plant defences. *Adv. Bot. Res.* 24, 89–167.

Jones, R., and Barbetti, M. (2012). Influence of climate change on plant disease infections and epidemics caused by viruses and bacteria. *CAB Rev.* 7, 1–32.

Kader, J.-C. (1996). Lipid-transfer proteins in plants. *Annu. Rev. Plant Biol.* 47, 627–654.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., et al. (2007). KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* 36, D480–D484.

Kang, B.-C., Yeam, I., and Jahn, M. M. (2005). Genetics of plant virus resistance. *Annu. Rev. Phytopathol.* 43, 581–621.

Kim, T., Seo, H. D., Hennighausen, L., Lee, D., and Kang, K. (2018). Octopus-toolkit: a workflow to automate mining of public epigenomic and transcriptomic next-generation sequencing data. *Nucleic Acids Res.* 46:e53.

Laliberte, J. F., and Sanfacon, H. (2010). Cellular remodeling during plant virus infection. *Annu. Rev. Phytopathol.* 48, 69–91.

Li, G., Lv, H., Zhang, S., Zhang, S., Li, F., Zhang, H., et al. (2019). TuMV management for brassica crops through host resistance: retrospect and prospects. *Plant Pathol.* 68, 1035–1044.

Li, G., Qian, W., Zhang, S., Zhang, S., Li, F., Zhang, H., et al. (2018). Variability in eukaryotic initiation factor iso4E in *Brassica rapa* influences interactions with the viral protein linked to the genome of *Turnip mosaic virus*. *Sci. Rep.* 8:13588.

Li, J., Xiang, C.-Y., Yang, J., Chen, J.-P., and Zhang, H.-M. (2015). Interaction of HSP20 with a viral RdRp changes its sub-cellular localization and distribution pattern in plants. *Sci. Rep.* 5:14016.

Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K., et al. (2009). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25, 1966–1967.

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

Loake, G., and Grant, M. (2007). Salicylic acid in plant defence—the players and protagonists. *Curr. Opin. Plant Biol.* 10, 466–472.

Lv, S., Changwei, Z., Tang, J., Li, Y., Wang, Z., Jiang, D., et al. (2015). Genome-wide analysis and identification of TIR-NBS-LRR genes in Chinese cabbage (*Brassica rapa* ssp. pekinensis) reveal expression patterns to TuMV infection. *Physiol. Mol. Plant Pathol.* 90, 89–97.

Lydiate, D. J., Pilcher, R. L. R., Higgins, E. E., and Walsh, J. A. (2014). Genetic control of immunity to *Turnip mosaic virus* (TuMV) pathotype 1 in *Brassica rapa* (Chinese cabbage). *Genome* 57, 419–425.

Ma, J., Hou, X., Xiao, D., Qi, L., Wang, F., Sun, F., et al. (2010). Cloning and characterization of the BcTuR3 gene related to resistance to *Turnip mosaic virus* (TuMV) from non-heading Chinese cabbage. *Plant Mol. Biol. Rep.* 28, 588–596.

Mauch-Mani, B., and Mauch, F. (2005). The role of abscisic acid in plant–pathogen interactions. *Curr. Opin. Plant Biol.* 8, 409–414.

Mazier, M., Flamain, F., Nicolaï, M., Sarnette, V., and Caranta, C. (2011). Knock-down of both eIF4E1 and eIF4E2 genes confers broad-spectrum resistance against potyviruses in tomato. *PLoS One* 12:e29595. doi: 10.1371/journal.pone.0029595

McCormack, E., and Braam, J. (2003). Calmodulins and related potential calcium sensors of *Arabidopsis*. *New Phytol.* 159, 585–598.

Mine, A., Hyodo, K., Tajima, Y., Kusumanegara, K., Taniguchi, T., Kaido, M., et al. (2012). Differential roles of Hsp70 and Hsp90 in the assembly of the replicase complex of a positive-strand RNA plant virus. *J. Virol.* 86, 12091–12104.

Molina, A., and García-Olmedo, F. (1993). Developmental and pathogen—induced expression of three barley genes encoding lipid transfer proteins. *Plant J.* 4, 983–991.

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

Nagy, P. D., and Pogany, J. (2012). The dependence of viral RNA replication on co-opted host factors. *Nat. Rev. Microbiol.* 10, 137–149.

Nakahara, K. S., Masuta, C., Yamada, S., Shimura, H., Kashihara, Y., Wada, T. S., et al. (2012). Tobacco calmodulin-like protein provides secondary defense by binding to and directing degradation of virus RNA silencing suppressors. *Proc. Natl. Acad. Sci. U.S.A.* 109, 10113–10118.

Nobori, T., Wang, Y., Wu, J., Stolze, S. C., Tsuda, Y., Finkemeier, I., et al. (2020). Multidimensional gene regulatory landscape of a bacterial pathogen in plants. *Nat. Plants* 6, 883–896.

Padmanabhan, M. S., Goregaoker, S. P., Golem, S., Shiferaw, H., and Culver, J. N. (2005). Interaction of the tobacco mosaic virus replicase protein with the Aux/IAA protein PAP1/IAA26 is associated with disease development. *J. Virol.* 79, 2549–2558.

Park, C.-J., Shin, R., Park, J. M., Lee, G.-J., You, J.-S., and Paek, K.-H. (2002). Induction of pepper cDNA encoding a lipid transfer protein during the resistance response to tobacco mosaic virus. *Plant Mol. Biol.* 48, 243–254.

Park, S.-Y., Fung, P., Nishimura, N., Jensen, D. R., Fujii, H., Zhao, Y., et al. (2009). Abscisic acid inhibits type 2C protein phosphatases via the PYR/PYL family of START proteins. *Science* 324, 1068–1071.

Peng, H., Li, S., Wang, L., Li, Y., Li, Y., Zhang, C., et al. (2013). *Turnip mosaic virus* induces expression of the LRR II subfamily genes and regulates the salicylic acid signaling pathway in non-heading Chinese cabbage. *Physiol. Mol. Plant Pathol.* 82, 64–72.

Perochon, A., Aldon, D., Galaud, J.-P., and Ranty, B. (2011). Calmodulin and calmodulin-like proteins in plant calcium signaling. *Biochimie* 93, 2048–2053.

Qian, W., Zhang, S., Zhang, S., Li, F., Zhang, H., Wu, J., et al. (2013). Mapping and candidate-gene screening of the novel *Turnip mosaic virus* resistance gene retr02 in Chinese cabbage (*Brassica rapa* L.). *Theor. Appl. Genet.* 126, 179–188.

Rodríguez-Hernández, A. M., Gosalvez, B., Sempere, R. N., Burgos, L., Aranda, M. A., and Truniger, V. (2012). Melon RNA interference (RNAi) lines silenced for Cm-eIF4E show broad virus resistance. *Mol. Plant Pathol.* 13, 755–763.

Rushholme, R. L. (2000). *The Genetic Control of Resistance to Turnip mosaic virus (TuMV) in Brassica*. Norwich: University of East Anglia.

Rusholme, R. L., Higgins, E. E., Walsh, J. A., and Lydiate, D. J. (2007). Genetic control of broad-spectrum resistance to *Turnip mosaic virus* in *Brassica rapa* (Chinese cabbage). *J. Gen. Virol.* 88, 3177–3186.

Sanfaçon, H. (2015). Plant translation factors and virus resistance. *Viruses* 7, 3392–3419.

Sarowar, S., Kim, Y. J., Kim, K. D., Hwang, B. K., Ok, S. H., and Shin, J. S. (2009). Overexpression of lipid transfer protein (LTP) genes enhances resistance to plant pathogens and LTP functions in long-distance systemic signaling in tobacco. *Plant Cell Rep.* 28, 419–427.

Segura, A., Moreno, M., and García-Olmedo, F. (1993). Purification and antipathogenic activity of lipid transfer proteins (LTPs) from the leaves of *Arabidopsis* and spinach. *FEBS Lett.* 332, 243–246.

Shattuck, V. (2010). The biology, epidemiology, and control of *Turnip mosaic virus*. *Hortic. Rev.* 14, 199–238.

Shopan, J., Mou, H., Zhang, L., Zhang, C., Ma, W., Walsh, J. A., et al. (2017). Eukaryotic translation initiation factor 2B—beta (eIF 2Bβ), a new class of plant virus resistance gene. *Plant J.* 90, 929–940.

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.

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., and Conesa, A. (2011). Differential expression in RNA-seq: a matter of depth. *Genome Res.* 21, 2213–2223.

Tassin, S., Broekaert, W. F., Marion, D., Acland, D. P., Ptak, M., Vovelle, F., et al. (1998). Solution structure of Ace-AMP1, a potent antimicrobial protein extracted from onion seeds. Structural analogies with plant nonspecific lipid transfer proteins. *Biochemistry* 37, 3623–3637.

Terras, F. R., Goderis, I. J., Van Leuven, F., Vanderleyden, J., Cammue, B. P., and Broekaert, W. F. (1992). In vitro antifungal activity of a radish (*Raphanus sativus* L.) seed protein homologous to nonspecific lipid transfer proteins. *Plant Physiol.* 100, 1055–1058.

Thaler, J. S., and Bostock, R. M. (2004). Interactions between abscisic—acid—mediated responses and plant resistance to pathogens and insects. *Ecology* 85, 48–58.

Tomlinson, J. (1987). Epidemiology and control of virus diseases of vegetables. *Ann. Appl. Biol.* 110, 661–681.

Truniger, V., and Aranda, M. A. (2009). Recessive resistance to plant viruses. *Adv. Virus Res.* 75, 119–231.

Walsh, J., Sharpe, A., Jenner, C., and Lydiate, D. (1999). Characterisation of resistance to *Turnip mosaic virus* in oilseed rape (*Brassica napus*) and genetic mapping of TuRB01. *Theor. Appl. Genet.* 99, 1149–1154.

Walsh, J. A., and Jenner, C. E. (2002). *Turnip mosaic virus* and the quest for durable resistance. *Mol. Plant Pathol.* 3, 289–300.

Wang, W., Vinocur, B., Shoseyov, O., and Altman, A. (2004). Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response. *Trends Plant Sci.* 9, 244–252.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10:57. doi: 10.1016/b978-0-12-411584-2.00003-2

Wang, Z., Jiang, D., Zhang, C., Tan, H., Li, Y., Lv, S., et al. (2015). Genome-wide identification of *Turnip mosaic virus*-responsive microRNAs in non-heading Chinese cabbage by high-throughput sequencing. *Gene* 571, 178–187.

Xinhua, W., Yang, L., and Huoying, C. (2011). A linkage map of pak—choi (*Brassica rapa* ssp. chinensis) based on AFLP and SSR markers and identification of AFLP markers for resistance to TuMV. *Plant Breed.* 130, 275–277.

Yamakawa, H., Mitsuhara, I., Ito, N., Seo, S., Kamada, H., and Ohashi, Y. (2001). Transcriptionally and post—transcriptionally regulated response of 13 calmodulin genes to tobacco mosaic virus—induced cell death and wounding in tobacco plant. *Eur. J. Biochem.* 268, 3916–3929.

Ye, J., Fang, L., Zheng, H., Zhang, Y., Chen, J., Zhang, Z., et al. (2006). WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res.* 34, W293–W297.

Yeam, I., Cavatorta, J. R., Ripoll, D. R., Kang, B.-C., and Jahn, M. M. (2007). Functional dissection of naturally occurring amino acid substitutions in eIF4E that confers recessive potyvirus resistance in plants. *Plant Cell* 19, 2913–2928.

Zhai, Q., Yan, L., Tan, D., Chen, R., Sun, J., Gao, L., et al. (2013). Phosphorylation-coupled proteolysis of the transcription factor MYC2 is important for jasmonate-signaled plant immunity. *PLoS Genet.* 9:e1003422. doi: 10.1371/journal.pgen.1003422

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.

Copyright © 2020 Lyu, Gao, Zhang, Zhang and Hou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 08 September 2020
doi: 10.3389/fgene.2020.565006

Similarities and Differences of Photosynthesis Establishment Related mRNAs and Novel lncRNAs in Early Seedlings (Coleoptile/Cotyledon vs. True Leaf) of Rice and *Arabidopsis*

*Yafei Shi, Jian Chen and Xin Hou**

State Key Laboratory of Hybrid Rice, College of Life Sciences, Wuhan University, Wuhan, China

Photosynthesis uses sunlight and carbon dioxide to produce biomass that is vital to all life on earth. In seed plants, leaf is the main organ for photosynthesis and production of organic nutrients. The seeds are mobilized to fuel post-germination seedling growth until seedling photosynthesis can be efficiently established. However, the photosynthesis and metabolism in the early growth and development have not been studied systematically and are still largely unknown. In this study, we used two model plants, rice (*Oryza sativa* L.; monocotyledonous) and *Arabidopsis* (*Arabidopsis thaliana*; dicotyledonous) to determine the similarities and differences in photosynthesis in cotyledons and true leaves during the early developmental stages. The photosynthesis-related genes and proteins, and chloroplast functions were determined through RNA-seq, real-time PCR, western blotting and chlorophyll fluorescence analysis. We found that in rice, the photosynthesis established gradually from coleoptile (cpt), incomplete leaf (icl) to first complete leaf (fcl); whereas, in *Arabidopsis*, photosynthesis well-developed in cotyledon, and the photosynthesis-related genes and proteins expressed comparably in cotyledon (cot), first true leaf (ftl) and second true leaf (stl). Additionally, we attempted to establish an mRNA-lncRNA signature to explore the similarities and differences in photosynthesis establishment between the two species, and found that DEGs, including encoding mRNAs and novel lncRNAs, related to photosynthesis in three stages have considerable differences between rice and *Arabidopsis*. Further GO and KEGG analysis systematically revealed the similarities and differences of expression styles of photosystem subunits and assembly factors, and starch and sucrose metabolisms between cotyledons and true leaves in the two species. Our results help to elucidate the gene functions of mRNA-lncRNA signatures.

Keywords: rice, *Arabidopsis*, messenger RNA, long non-coding RNA, cotyledon, true leaf, photosynthesis

Abbreviations: CNCI, Coding-Non-Coding Index; cot, cotyledon; CPAT, Coding Potential Assessment Tool; CPC, Coding Potential Calculator; cpt, coleoptile; DEGs, differentially expressed genes; fcl, first complete leaf; ftl, first true leaf; GO, Gene Ontology; icl, incomplete leaf; KEGG, Kyoto Encyclopedia of Genes and Genomes; KO, KEGG Orthology; lncRNA, long non-coding RNA; mRNA, messenger RNA; PS I, photosystem I; PS II, photosystem II; stl, second true leaf.

OPEN ACCESS

Edited by:
Weike Duan,
Huaiyin Institute of Technology, China

Reviewed by:
Himanshu Sharma,
National Agri-Food Biotechnology Institute, India
Vijay Gahlaut,
University of Delhi, India

***Correspondence:**
Xin Hou
xinhou@whu.edu.cn

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

Received: *23 May 2020*
Accepted: *17 August 2020*
Published: *08 September 2020*

Citation:
Shi Y, Chen J and Hou X (2020) Similarities and Differences of Photosynthesis Establishment Related mRNAs and Novel lncRNAs in Early Seedlings (Coleoptile/Cotyledon vs. True Leaf) of Rice and Arabidopsis. Front. Genet. 11:565006. doi: 10.3389/fgene.2020.565006

INTRODUCTION

Seeds provide energy and materials for the germination, early growth and development of plants. The size and storage of seeds vary considerably among species. In monocots, the endosperm constitutes the majority of the mature seed (Berger et al., 2006). Rice (*Oryza sativa*) endosperm consists of an outer aleurone layer and an inner starchy endosperm (Becraft and Asuncion-Crabb, 2000; Krishnan and Dayanandan, 2003). In maize (*Zea mays*), the endosperm typically consists of the vitreous endosperm and the starchy endosperm (Wang et al., 2012). In dicotyledons, such as the oilseed crop *Brassica napus* (rapeseed) and the model plant *Arabidopsis*, storage products consist of proteins and triacylglycerols (TAGs) synthesized mainly in the embryo (Baud et al., 2008). The endosperm and embryo are mobilized to fuel post-germination seedling growth until seedling photosynthesis can be efficiently established (Sela et al., 2020).

Rice, as a monocotyledonous model plant, is one of the most important crops in the world. Rice grain, as a cereal crop, is a food reserve to nourish growing seedlings during germination (Zheng and Wang, 2015; Wu et al., 2016). The starchy endosperm is the major energy source of rice and contains mostly starch and a small amount (8%) of proteins (Becraft and Yi, 2011). During rice seed germination, starch is decomposed into glucose, which is directly used as raw material for respiration, and cellulose is further synthesized to ensure the formation of cell walls in new cells (Leonova et al., 2010). Proteins stored in seeds are decomposed into soluble amino acids under the action of protease and transported to the growth site of embryos. Amino acids under the action of corresponding enzymes form structural proteins, which become cellular components in buds and roots. Rice seeds contain less fat, that is, approximately 1–2%. During seed germination, storage lipids are decomposed into fatty acids and glycerol by lipase and then converted into sugars, thereby providing raw materials for respiration. *Arabidopsis* seeds are considerably smaller than rice. Rice (Nipponbare) seed usually weighs 20 mg. By contrast, *Arabidopsis* (Col-0) seed weighs approximately 20 μg. *Arabidopsis* seeds contain 30–40% proteins (Baud et al., 2002), 38% model oil (O'Neill et al., 2003), and 2% oligosaccharides (Bentsink et al., 2000). The seed stored oil, TAGs, is metabolized to provide carbon skeleton and energy resources for germination and growth of young seedlings (Baud et al., 2002). At the stage of seeds maturation, the most abundant mRNAs in embryos are those encoding seed storage proteins, which are considered to be the source of seedling nitrogen for young seedlings (Pang et al., 1988).

The photosynthesis and carbon cycle of plants are the basis of all life activities. Photosynthesis plays a crucial role in biology and biogeochemistry, transforming sunlight energy and CO_2 into metabolic products of life (Liu and Last, 2017). After germination, leaves are the main organs of plant photosynthesis. In dicotyledonous plants, such as *Arabidopsis*, cotyledons are components of seeds and the first pair of leaves; the first true leaves appeared several days later. Both cotyledons and true leaves contain chloroplasts and have photosynthetic activity; however, the differentiation of chloroplasts follows different pathways in these two organs (Deng and Gruissem, 1987; Mansfield and Briarty, 1996). Cotyledons are capable of storing nutrients and performing photosynthesis and provide a major proportion of matter needed for seedling growth and development until the first true leaf becomes a significant exporter of photosynthates (Zhang et al., 2008; Zheng et al., 2011). Cotyledon deficiency experiments result in growth arrest or even death of young seedlings (Garcia-Cebrian et al., 2003; Hanley and May, 2006). The cotyledon chloroplast biogenesis factor CYO1 (shi-yo-u means cotyledon in Japanese) associated with PSI/LHCI and PSII/LHCII complexes is crucial for the biogenesis of cotyledon chloroplasts in *Arabidopsis* (Shimada et al., 2007; Muranaka et al., 2012). In rice, the coleoptile is the first leaf-like organ formed around the embryo during embryogenesis (Hirano, 2008). It is the basic structure of the seedling and plays a protective role when the embryo passes through the soil layers (Xiong et al., 2017). The incomplete leaf is the first green organ with only a sheath but no blade. The complete leaves with blades and sheaths are from the second green leaf (Du et al., 2018). OsCYO1 is expressed in leaf but not in coleoptile, and functions in the accumulation and/or assembly of PSI (Tominaga et al., 2016). There are structural and functional differences between monocotyledonous and dicotyledonous plants in early growth and development. In rice, photosynthetic activity may differ from that in *Arabidopsis*. However, the similarities and differences in photosynthesis and metabolism in the early growth and development of the two model plants have not been studied systematically and are still largely unknown.

In recent years, deep sequencing of transcripts has been increasingly used in the identification of differential expression (Lu et al., 2010; Bhargava et al., 2013; Gao et al., 2013; Yuan et al., 2016). Long non-coding RNAs (lncRNAs) are RNAs with lengths greater than 200 bp that lack protein-encoding functions and play important roles in essential biological processes. lncRNA mostly functions in the nucleus, often in association with chromatin processes (Chekanova, 2015). lncRNAs can be divided into three categories: long intergenic ncRNAs (lincRNAs), intronic ncRNAs (incRNAs), and natural antisense transcripts (NATs) transcribed from the complementary DNA strand of the associated genes (Chekanova, 2015). Although lncRNAs do not have the ability to encode proteins, it is widely believed that lncRNAs can be transcribed and regulate gene expression at the transcriptional and post-transcriptional levels (Kim and Sung, 2012). Many plant lncRNAs are developmentally and environmentally regulated and likely represent functional components of the transcriptome (Liu et al., 2012). The regulation of coding genes mediated by lncRNA at the transcriptional level can be divided into *cis*-regulation and trans-regulation. Epigenetic regulation of lncRNAs plays a key role in flowering regulation by controlling the expression of FLC (flowering site C) in *Arabidopsis*. FLC plays a key role in regulating flowering process and regulates flowering independently of environmental signals (Amasino and Michaels, 2010). COLDAIR and COOLAIR are two different classes of lncRNAs transcribed from FLC that participate in epigenetic silencing of FLC (Liu et al., 2010; Heo and Sung, 2011; Tian et al., 2019). A lncRNA HID1 with a length of 236 nt length was identified by chain-specific transcriptome sequencing. This lncRNA participates in seed photomorphogenesis by mediating

PIF3 expression in *Arabidopsis* (Wang et al., 2014b). In addition, some studies have found that lncRNAs can be transcribed from the antisense strands of genes. These lncRNAs are usually involved in post-translational splicing, editing, transport, translation, and degradation of mRNA (Liu et al., 2019).

In this study, high-throughput sequencing was used to analyze mRNA, novel lncRNAs and their target genes during the early growth and development of rice and *Arabidopsis*. We show that photosynthetic protein complexes gradually mature from cpt, icl to fcl in rice. However, the functions are fully developed during the cot stage of *Arabidopsis*. Photosynthesis and photosynthesis-related processes were clustered in the three stages of rice but not in *Arabidopsis* through Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses of differentially expressed genes (DEGs). Further analysis of DEGs, including mRNAs and lncRNAs, showed that the upregulated genes from icl to fcl were distributed in photosystem II (PSII), photosystem I (PSI), cytochrome b6/f complex, photosynthetic electron transport and F-type ATPase in rice but not in *Arabidopsis*. Additionally, assembling the individual subunits to form PSII and PSI are also different in the two plants and may not function at the transcriptional level.

MATERIALS AND METHODS

Plant Materials and Growth Conditions

Rice (Nipponbare) were grown under 16 h of light at 30°C and 8 h of darkness at 28°C. Seedlings were grown on plates containing only water and 0.8% agar. Leaves were harvested at 68 h for coleoptile (cpt), 82 h for incomplete leaf (icl), and 127 h for first complete leaf (fcl). *Arabidopsis* (Columbia-0) was grown under 16 h of light at 22°C and 8 h of darkness at 20°C. Seedlings were grown on 1/2 MS medium containing 1.0% sucrose (pH 5.7) or only water and 0.8% agar. Leaves were harvested at 111 h for cotyledon (cot), 183 h for first true leaf (ftl), and 236 h for second true leaf (stl).

RNA Extraction and RNA-Seq Preparation

Approximately 60 rice seedlings and 600 *Arabidopsis* seedlings were collected from each stages and frozen in liquid nitrogen immediately. Three biological replicates were performed. Total RNA was extracted using a RNeasy Plant Mini Kit followed by DNase I treatment according to the manufacturer's manual (Qiagen). The extracted RNA was analyzed by 1% (w/v) agarose gel for quality and completeness, and the concentration was measured on a NanoDrop 2000c spectrophotometer (NanoDrop Technologies, Wilmington, DE, United States). The RNA samples were sent to Novogene (Novogene, Co. Ltd.) for mRNA purification, cDNA library construction and sequencing using the platform Illumina. The original raw data from HiSeq X Ten was transformed to sequenced reads by base calling. Sequencing data analysis was completed in UniqueGene (Wuhan, China).

Sequence Analysis

Reads qualities of the RNA sequencing (RNA-Seq) were evaluated using FastQC (Version 0.11.5). The clean data was evaluated by removing the adapter and low-quality reads of the raw data obtained from Illumina sequencing. The clean reads were searched against the rice and *Arabidopsis* genome Using HISAT2 (Version 2.0.4). The obtained transcripts were assembled using StringTie (Version 1.0.4). All downstream analyses were based on the high quality clean data.

Identification of Novel lncRNAs

Based on the structural characteristics and non-coding function of lncRNA, we established a five-step screening method to obtain high quality lncRNA: (1) Transcripts with one exon were removed while the transcripts with two or more exons were selected; (2) the transcripts with length < 200 bp were removed; (3) Filtered out transcripts with reads coverage < 5 in all samples; (4) Filtered out the mRNA and other non-coding RNA (rRNA, tRNA, snoRNA, snRNA, etc.) by comparing gffcompare (Version 0.10.4) with the annotation files; (5) The potential lincRNA, intronic lncRNA and anti-sense lncRNA were screened according to the class_code information ("u," "i," "x") in the comparison results. Based on the length of the open reading frame, homology with known proteins and their coding potential, the CNCI (Coding-Non-Coding Index) (Sun et al., 2013), CPC (Coding Potential Calculator) (Kong et al., 2007), and CPAT (Coding Potential Assessment Tool), which rapidly recognizes coding and non-coding transcripts from a large pool of candidates (Wang et al., 2013) were combined to screen the lncRNAs. This analysis was combined with information from the Pfam protein database to ensure the predicted lncRNA transcripts did not contain protein-coding domains (Mistry et al., 2007). The transcripts that meet these four criteria were determined to be lncRNAs.

Differential Expression Analysis of mRNAs and lncRNAs

The clean reading was aligned with the reference genomes by HiSAT2 software, and then gene and transcript expression levels were calculated by performing RSEM. In order to make the data of gene expression comparable among samples, the level of gene expression was standardized. FPKM (fragments per kilobase of exon per million fragments mapped) values were used to analyze gene expression. Differential expression analysis of different stages was performed using the DESeq R package (1.10.1). DESeq provides a statistical program based on negative binomial distribution to determine differential expression in digital gene expression data. The resulting *P*-values were adjusted by Benjamini and Hochberg's approach for controlling the false discovery rate (Benjamini and Hochberg, 1995). Genes with an adjusted *P*-value < 0.05 and Fold change ≥ 2.00 found by DESeq were assigned as differentially expressed.

Target Genes of Differentially Expressed lncRNAs

The differentially expressed lncRNAs were selected as targets by *cis*- or *trans*- regulation. In the prediction of *cis* pathway target genes, the genes transcribed in 10 kb upstream or downstream of coding genes. Trans-acting target genes were selected by RNAplex software (Tafer and Hofacker, 2008).

GO and KEGG Pathway Analyses of Differentially Expressed mRNAs and lncRNAs

The GO enrichment analysis of DEGs was implemented by Wallenius non-central hypergeometric distribution based on GOseq R package (Young et al., 2010), which can be adjusted according to the gene length bias in DEGs. GO terms P-value < 0.05 were considered significantly enriched by differentially expressed genes. KEGG (Kanehisa and Goto, 2000) is a database resource for understanding the high-level functions and utilities of biological systems, such as cells, organisms and ecosystems, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies[1]. The statistical enrichment of DEGs in KEGG pathway were tested using KOBAS (Mao et al., 2005) software. The significantly enriched KEGG pathways were determined using the corrected P-value < 0.05 as a threshold.

Expression Analysis by Quantitative Real-Time PCR (q-RT PCR) of DEGs

The DEG results were confirmed by q-RT PCR using the same RNA samples that were used for RNA library construction. The isolated RNA was reverse transcribed to cDNA using a PrimeScript RT reagent Kit (Takara, Dalian China) according to the manufacturer's instructions. Real-time PCR analyses were performed on an Applied Biosystems 7300 Plus Real-Time PCR System with SYBR Green Mix (Takara, Dalian, China). The amplified PCR products were quantified and normalized using *actin1* for rice and *actin2* for *Arabidopsis* as a control (**Supplementary Table S1**). The correlation between RNA Seq and q-RT PCR results were analyzed through pearson's correlation coefficient (r).

Immunoblot Analysis

Protein was extracted from each tissue using extraction buffer [50 mM Tris-MES pH 7.5, 50 mM NaCl, 0.5% SDS (v/v)], and protease inhibitor cocktail P9599 (Sigma-Aldrich, St. Louise, MO, United States). The extract was centrifuged at 12000 rpm at 4°C for 30 min. For immunoblotting, protein samples (18 μg total protein) were separated on 12% SDS PAGE gels and transferred to nitrocellulose membranes (BioTrace^{TM} NT nitrocellulose, Mexico) followed by western blot analysis. After blocking non-specific binding with 5% milk, the blot was subsequently incubated with antibodies generated against the indicated proteins and detected using the Super Signal^{TM} West Pico PLUS Chemiluminescent Substrate kit (Thermo Scientific, United States).

Chlorophyll Fluorescence Measurements

Chlorophyll fluorescence was measured using FluorCam 800 MF (Photon Systems Instruments, spol. s r.o.). Plants were maintained in the dark for 30 min before the measurements. The intensities of the actinic light and saturating light were 300 and 2,100 μmol m^{-2} s^{-1} photosynthetic photon flux density for rice, and 80 and 1,600 μmol m^{-2} s^{-1} photosynthetic photon flux density for *Arabidopsis*, respectively. Three biological replicates for each treatment were performed with 10 plants.

Statistical Analysis

All data were expressed as mean ± SD and analyzed using GraphPad Prism software (version 7.0, La Jolla, CA, United States). Student's t-test was used to compare the different groups. And the significant differences between the three stages were calculated using Tukey's test. The Pearson's correlation coefficient (r) was used to analyze the correlation between qRT-PCR and RNA-seq results. Differences with $|\log_2 \text{ratio}| \geq 1$ and $P < 0.05$ were selected for the DEGs. Cytoscape software was used to visualize the inferred gene networks (Shannon et al., 2003). Modules were visualized using Cytoscape version 3.6.1.

Accession Numbers

The sequence data generated in the current study has been uploaded to the NCBI Sequence Read Archive. The accession numbers are SRR10907568–SRR10907573 (coleoptile sequencing of rice), SRR10913120-SRR10913123 (incomplete leaf sequencing of rice), SRR10913307-SRR10913312 (first complete leaf sequencing of rice), SRR10914245-SRR10914250 (cotyledon sequencing of *Arabidopsis*), SRR10914757-SRR10914762 (first true leaf sequencing of *Arabidopsis*), SRR10915176-SRR10915181 (second true leaf sequencing of *Arabidopsis*)[2].

RESULTS

Photosynthesis and Chloroplast Development of Early Growth and Development in Rice and *Arabidopsis*

To find the similarities and differences in photosynthesis and chloroplast development between monocotyledons and dicotyledons during early growth and development, photosynthesis was determined by measuring chlorophyll fluorescence in rice and *Arabidopsis*. We collected coleoptile (cpt) before the incomplete leaf (icl) appeared in rice and cotyledon (cot) before the first true leaf (ftl) appeared in *Arabidopsis* as stage 1 (S1), icl before the first complete leaf (fcl) appeared in rice and ftl before the second true leaf (stl) appeared in *Arabidopsis* as stage 2 (S2), fcl in rice and stl in *Arabidopsis* appeared fully as stage 3 (S3). The ratio of variable fluorescence to maximum fluorescence (Fv/Fm), which represents the maximum photochemical efficiency of PSII (Maxwell and Johnson, 2000; Hou et al., 2015), was significantly increased in the three stages of rice (**Figure 1A**). In contrast, unlike rice, *Arabidopsis* could germinate on 0.8% agar, but was arrested at the cotyledon stage without supplied nutrients (**Figure 1B**). In addition, the Fv/Fm was decreased during the growth (**Figure 1B**), suggesting that the nutrients stored in *Arabidopsis* seeds are insufficient to satisfy their normal growth and development. To ensure the

[1] http://www.genome.jp/kegg/

[2] https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=run_browser

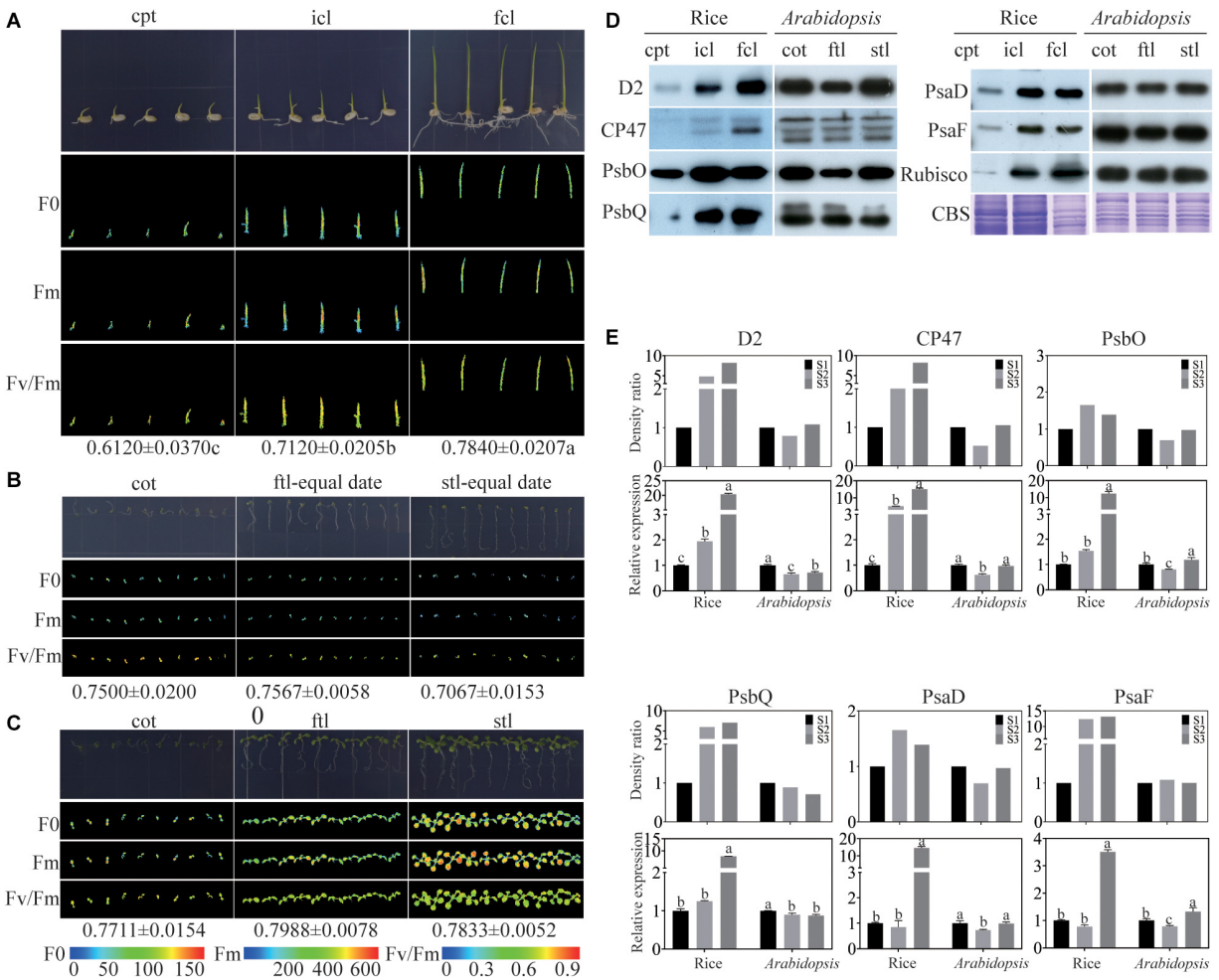


FIGURE 1 | Comparison of rice and *Arabidopsis* at the early developmental stages. **(A)** rice grown on 0.8% agar under 16 h of light at 30°C and 8 h of darkness at 28°C. **(B)** *Arabidopsis* plants grown on 0.8% agar under 16 h of light at 22°C and 8 h of darkness at 20°C conditions for different stages **(C)** *Arabidopsis* plants grown on 1/2MS under 16 h of light at 22°C and 8 h of darkness at 20°C at different stages. **(A–C)** Chlorophyll fluorescence analysis. F_0, the minimum fluorescence; Fm, the maximum fluorescence; Fv/Fm, the maximum efficiency of PSII photochemistry. cpt, rice coleoptile; icl, rice incomplete leaf; fcl, rice first complete leaf; cot, *Arabidopsis* cotyledons; ftl, *Arabidopsis* first true leaf; stl, *Arabidopsis* second true leaf. **(D)** Analysis of thylakoid membrane proteins in rice and *Arabidopsis* at different stages. 16 μg total proteins were separated by 12% SDS PAGE. And immunoblots were performed using antibodies as indicated. CBS, Coomassie Blue Staining. **(E)** The relative expression levels of chloroplast proteins in **(D)** and the relative expression levels of relavent genes of rice and *Arabidopsis*. Data was presented as the means ± SD of three biological replicates. S1, cpt and cot; S2, icl and ftl; S3, fcl and stl. a, b, and c indicate significant differences between the three stages according to Tukey's test.

normal growth of *Arabidopsis*, seedlings were grown on 1/2 MS medium containing 1% sucrose. No significant difference in Fv/Fm was found in the three stages of *Arabidopsis* (**Figure 1C**). Taken together, the results of the chlorophyll fluorescence assays suggested that differences exist in early growth and development between rice and *Arabidopsis* in PSII function.

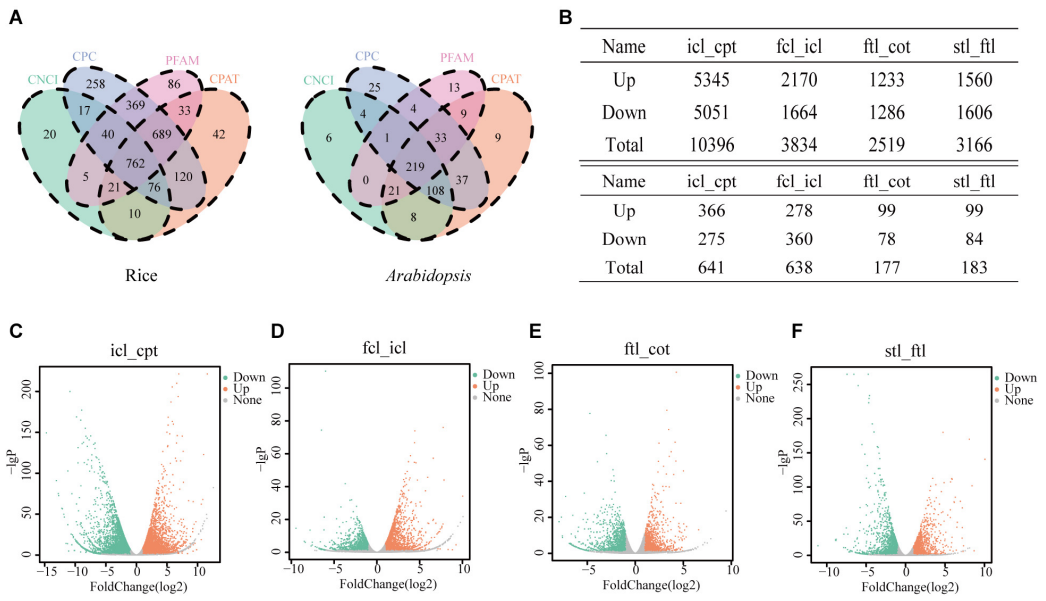
To further visualize differences in the photosynthetic complexes in two model plants, we used qRT-PCR and immunoblotting to examine the abundance of subunits of the photosynthetic complexes including PSII: D2, CP47, PsbO, and PsbQ; and PSI: PsaD and PsaF. The Rubisco protein and Coomassie Blue staining were used as a loading control. The results showed that the protein expression of PSII subunits increased from cpt to fcl in rice. Conversely, there was no significant difference in the abundance of PSII and PSI subunits between the three stages in *Arabidopsis* (**Figure 1D**). The mRNA levels of these six genes were consisting with the protein level in both rice and *Arabidopsis* (**Figure 1E**). These findings suggest that chloroplast functional protein complexes gradually mature from cpt, icl to fcl during the early growth and development of rice. Photosynthesis of the chloroplast is also gradually improved in the above three stages. Conversely, chloroplast proteins and photosynthesis are fully developed at the cot stage of *Arabidopsis*. More importantly, we can use RNA expression levels to represent the protein expression levels of photosynthetic genes because their expression trend was consistent.

Sequencing Data Summary

To systematically study the differences between rice and *Arabidopsis* during early growth and development, gene expression analyses were conducted by high-throughput RNA sequencing. A total of 54.9 Gb raw data for rice and 67.9 Gb raw data for *Arabidopsis* were generated. RNA sequencing was performed on the HiSeq X Ten platform, and we detected the expression of mRNA in all tissue samples. In detail, 91,035,666, 90,874,512, and 95,792,352 raw reads were obtained for rice coleoptile (cpt-A, B, and C, respectively); 86,766,618 and 95,714,768 raw reads were obtained for rice incomplete leaf (icl-A and B, respectively); and 88,510,902, 115,795,104, and 85,884,202 raw reads were obtained for rice first complete leaf (fcl-A, B, and C, respectively). A total of 99,311,856, 100,974.596, and 100,566,922 raw reads were obtained for *Arabidopsis* cotyledon (cot-A, B, and C, respectively); 85,926,770, 92,493,690, and 98,031,334 raw reads were obtained for *Arabidopsis* first true leaf (ftl-A, B, and C, respectively); 89,000,366, 125,026,956, and 98,057,458 raw reads were obtained for *Arabidopsis* second true leaf (stl-A, B, and C, respectively). The results of the RNA-Seq reads mapped on the reference are shown in **Supplementary Table S2**. Of the 17 library alignments, the unique mapping rate between clean reads and the reference genome was 93.96–98.83% (**Supplementary Table S3**). The three functional elements of gene proportions were 55.00–75.31% exon, 5.83–8.19% intron, 18.86–37.50% intergenic in rice and 83.98–88.71% exon, 4.76–8.39% intron, and 6.52–7.43% intergenic in *Arabidopsis* (**Supplementary Table S4**), and the normalized FPKM values for all mRNAs and lncRNAs in all samples are shown in **Supplementary Table S5**. The above results show that our sequencing results were reliable.

Identification and Characterization of mRNA and Novel lncRNA in Rice and *Arabidopsis*

In the present study, we collected three typical growth periods of rice and *Arabidopsis*. Using high-throughput sequencing, DEGs were detected in the leaves at different developmental stages. To study the basic features of lncRNAs in rice and *Arabidopsis*, lncRNAs were identified and compared with mRNAs. As a result, 762 novel transcripts in rice and 219 novel transcripts in *Arabidopsis* were identified as novel lncRNAs (**Figure 2A**). In addition, 43,148 known transcripts in rice and 25,481 in *Arabidopsis* were identified as mRNAs. Most identified lncRNAs had only 2–3 exons in their transcripts in both rice (**Supplementary Figures S1A,E**) and *Arabidopsis* (**Supplementary Figures S1B,F**) compared with mRNAs. The number of lncRNAs was notably small compared to the number of coding RNAs with 200–2000 nt transcripts in both rice (**Supplementary Figures S1C,G**) and *Arabidopsis* (**Supplementary Figures S1D,H**) compared with mRNAs. According to RNA-seq, the significantly dysregulated mRNAs and lncRNAs ($|\log_2 \text{ratio}| \geq 1$, $q < 0.05$), including upregulated and downregulated mRNAs and lncRNAs, were



Name	icl_cpt	fcl_icl	ftl_cot	stl_ftl
Up	5345	2170	1233	1560
Down	5051	1664	1286	1606
Total	10396	3834	2519	3166
Name	icl_cpt	fcl_icl	ftl_cot	stl_ftl
Up	366	278	99	99
Down	275	360	78	84
Total	641	638	177	183

FIGURE 2 | Differentially expressed unigenes and corresponding genes in each early growth developmental stage. **(A)** Three software CPC, CNCI, CPAT and a protein database pfam were used to predict lncRNAs, and the transcript was determined when the four methods were consistent in rice and *Arabidopsis*. **(B)** Number of differentially expressed genes obtained by comparisons between different stages and different lncRNAs in different stages. The genes were selected with "$|\log_2 \text{ratio}| \geq 1$" and "$q < 0.05$." **(C–F)** Volcanic map of differentially expressed genes in rice icl-vs.-cpt **(C)**, fcl-vs.-icl **(D)** and in *Arabidopsis* ftl-vs.-cot **(E)** and stl-vs.-ftl **(F)**. Volcano plot of the $\log_2$ fold change of gene expression levels (X-axis) against the $-\log_{10}$ *P*-value (Y-axis) comparing the two stages in rice and *Arabidopsis*. Splashes represent different genes. Gray splashes means genes without significant different expression. Orange splashes mean significantly upregulated genes. Green splashes mean significantly downregulated genes. The genes were selected with "$|\log_2 \text{ratio}| \geq 1$" and "$q < 0.05$."

found to be differentially expressed between icl-vs.-cpt and fcl-vs.-icl in rice; ftl-vs.-cot, stl-vs.-ftl in *Arabidopsis*. In rice, we identified 10396 and 3834 DEGs in the icl-vs.-cpt and fcl-vs.-icl comparisons, respectively (**Figure 2B**). Similarly, 2519 and 3166 DEGs were identified in the ftl-vs.-cot and stl-vs.-ftl comparisons, respectively, in *Arabidopsis* (**Figure 2B**). Similarly, according to RNA-seq, the significantly upregulated and downregulated lncRNAs ($|\log_2 \text{ratio}| \geq 1$, $q < 0.05$) were found to be differentially expressed between the different stages (**Supplementary Figure S2**). In rice, we identified 641 and 638 DEGs in the icl-vs.-cpt and fcl-vs.-icl comparisons, respectively (**Figures 2C,D**). Similarly, 177 and 183 DEGs were identified in the ftl-vs.-cot and stl-vs.-ftl comparisons (**Figures 2E,F**), respectively, in *Arabidopsis*.

GO Analysis of Differential Expressed mRNAs and Novel lncRNAs in Rice and *Arabidopsis*

To better understand the developmental pattern of the photosynthetic characteristics of early growth and development in rice and *Arabidopsis*, GO cluster analysis was conducted to further explore the biological processes enriched by the candidate mRNAs and lncRNAs. In rice, GO analysis for mRNAs revealed up- and downregulated genes in biological pathways, and "photosynthesis," "photosynthesis, light reaction," "photosynthetic electron transport in PSI," and "photosynthesis, light harvesting in PSI" were clearly enriched in icl-vs.-cpt. "Photosynthesis," "photosynthesis, dark reaction," "photosynthesis, light reaction," and "photosynthesis, light harvesting" were also clearly enriched in fcl-vs.-icl. In contrast, in *Arabidopsis*, no photosynthesis-related biological pathways were enriched in the ftl-vs.-cot and stl-vs.-ftl comparisons. Similarly, GO analysis for lncRNAs revealed that up- and downregulated genes in biological pathways in rice were "photosynthesis," "photosynthesis, light reaction," "photosynthesis, light harvesting," "phototransduction," and "red, far-red light phototransduction," which were clearly enriched in icl-vs.-cpt. "Photosynthesis," "photosynthesis, light reaction," "photorespiration," "photosynthesis, light harvesting," and "chlorophyll metabolic process" were also clearly enriched in fcl-vs.-icl. In *Arabidopsis*, similar to the mRNAs, no photosynthesis-related biological pathways were enriched in the ftl-vs.-cot and stl-vs.-ftl comparisons (**Supplementary Table S6**). This result suggests that chloroplast develop and mature gradually in the three stages of early growth and development in rice, unlike the fully developed chloroplast in cotyledon in *Arabidopsis*.

KEGG Analysis of Differential Expressed mRNAs and Novel lncRNAs in Rice and *Arabidopsis*

At the same time, KEGG pathway analysis showed that mRNAs in the network of photosynthesis were enriched in "Photosynthesis antenna proteins," "Photosynthesis," and "Starch and sucrose metabolism" in icl-vs.-cpt (**Figure 3A**). "Photosynthesis," "starch and sucrose metabolism," and "photosynthesis – antenna proteins" were also significantly enriched in fcl-vs.-icl (**Figure 3B**). Similar to the GO analysis, in *Arabidopsis*, no photosynthesis-related biological pathways were enriched in the ftl-vs.-cot and stl-vs.-ftl comparisons (**Figures 3C,D**). Interestingly, the "starch and sucrose metabolism" pathway was clearly enriched in the ftl-vs.-cot and stl-vs.-ftl comparisons (**Figures 3C,D**). Similarly, KEGG analysis for lncRNAs revealed that up- and downregulated photosynthesis-related genes in rice were "Photosynthesis – antenna proteins," "Carbon fixation in photosynthetic organisms" in icl-vs.-cpt (**Figure 4A**). "Carbon fixation in photosynthetic organisms" and "porphyrin and chlorophyll metabolism" in fcl-vs.-icl (**Figure 4B**). Additionally, "Carbon metabolism" and "Citrate cycle (TCA cycle)" were clearly enriched in fcl-vs.-icl (**Figure 4B**). This finding suggests that the accumulation of photosynthesis products increased gradually from cpt to fcl. Similar to the GO analysis in *Arabidopsis*, no photosynthesis-related biological pathways were enriched in the ftl-vs.-cot and stl-vs.-ftl comparisons (**Figures 4C,D**).

The KEGG analysis combined with GO analysis of DEGs indicated that there are great differences in the expression pattern of chloroplast-related genes and the function of chloroplast form integrity during early development between rice and *Arabidopsis*. Photosynthetic genes and functions develop and mature gradually in rice and fully develop when exposed to light in *Arabidopsis* in the three stages of early growth and development. More importantly, lncRNAs might play important roles in chloroplast development by regulating these coherent target genes. The difference in chloroplast development between rice and *Arabidopsis* is largely consistent with the mRNA level.

Different Patterns of Photosynthesis Metabolism in Early Development

To better describe the differences in gene expression at different stages between rice (from cpt to fcl) and *Arabidopsis* (from cot to stl). We studied DEGs related to photosynthesis in rice and *Arabidopsis*. The up-regulated genes in fcl compared to icl were distributed in PSII, PSI, cytochrome b6/f complex, photosynthetic electron transport and F-type ATPase (**Figure 5**). The functions of lncRNAs are executed on coding genes via *cis*- or *trans*-regulation. Different expression novel lncRNAs targeted photosynthetic genes in rice and *Arabidopsis* were showed in **Figure 6** and **Supplementary Table S7**. As showed in **Figure 6**, one novel lncRNA was correlated with several mRNAs and one mRNA was co-expressed with more than one novel lncRNA, further supporting the notion of inter-regulatory relationships. The novel lncRNAs targeted photosynthetic genes also showed upregulated chloroplast-associated genes, including *PsbA*, *PsbW*, *Psb27*, *PsaH*, *PsaO*, *PetC*, *PetF*, *PetH*, *ATPF1B,* and *ATPF1G* in icl-vs.-cpt, and *PsbA*, *Psb27*, *Psb28*, *PsaO*, *PetC*, *PetH*, *ATPF1B*, *ATPF1G,* and *ATPF1D* in fcl-vs.-icl. In contrast, for *Arabidopsis,* KEGG pathway analysis of novel lncRNAs targeted photosynthetic genes showed no significant difference in chloroplast-associated genes in the three stages (**Figure 7**). To validate the RNA-Seq data, DEGs and differentially expressed mRNAs and lncRNAs related to photosynthesis were selected

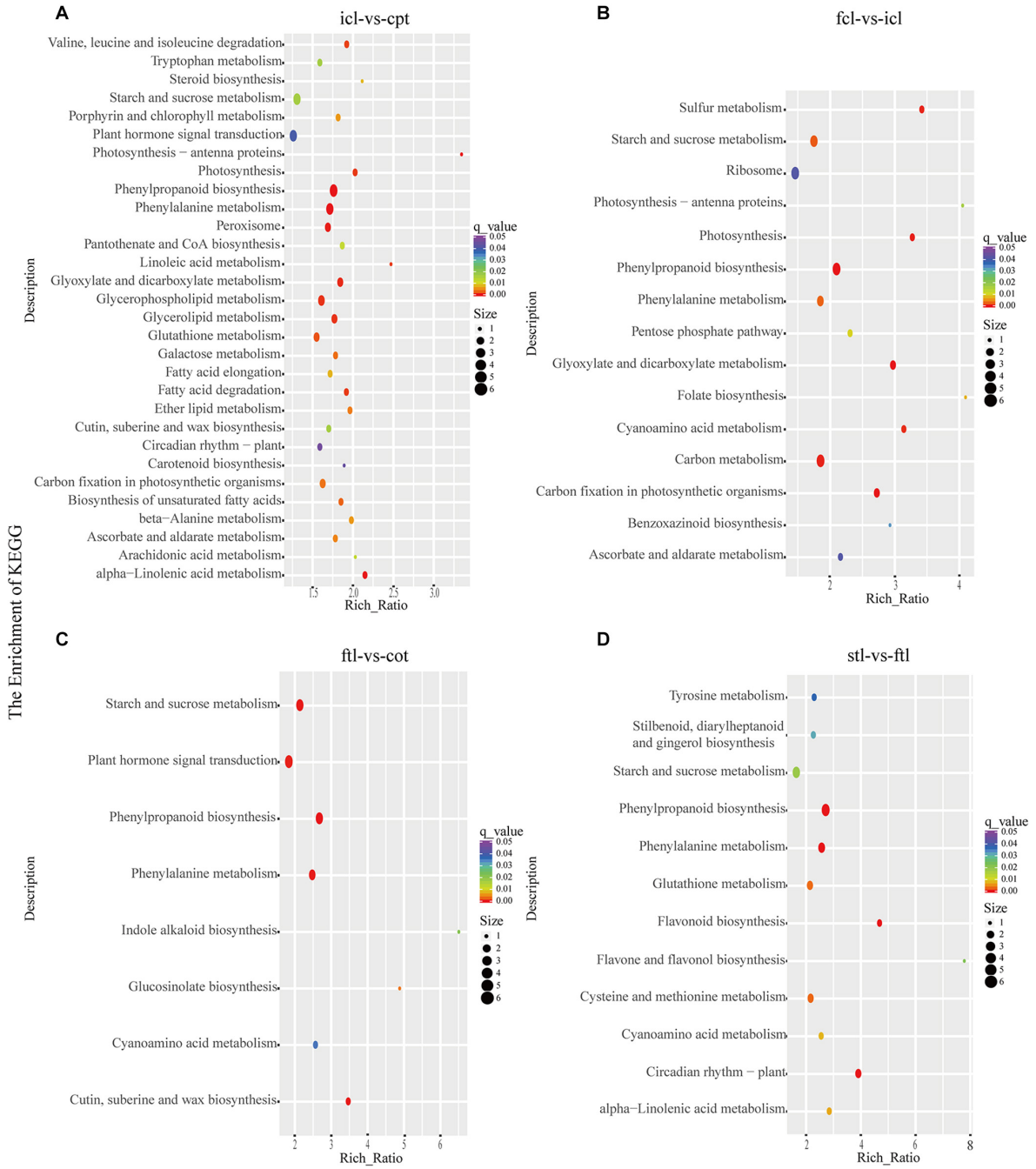


FIGURE 3 | The mRNAs KEGG classifications of predicted target genes in different comparisons. KEGG classification of the predicted target genes in different comparisons of icl-vs.-cpt **(A)** and fcl-vs.-icl **(B)** in rice. KEGG classification of the predicted target genes in different comparisons of ftl-vs.-cot **(C)** and stl-vs.-ftl **(D)** in *Arabidopsis*. In both **(A,B),** the y-axis shows the pathway category, and the x-axis shows the richness factor. A larger richness factor indicates greater enrichment. The size of the point represents the number of related DEGs., and the color of the bubble represents the range of *q*-values.

in rice and *Arabidopsis* at the three stages (**Supplementary Table S8**). The qRT-PCR expression patterns of mRNAs and lncRNAs genes were consistent with the sequencing results (**Supplementary Figure S3**). In addition, we performed qRT-PCR analysis of the genes mentioned in this paper, and further analysis of the correlation between RNA Seq and q-RT PCR results. As shown in **Supplementary Figure S4** and **Supplementary Table S9**, the expression trends of all 17 genes from qRT-PCR and RNA-seq analyses including mRNAs and lncRNAs were largely consistent (Pearson's correlation coefficient R^2 = 0.83 in rice and 0.91 in *Arabidopsis*), indicating that our sequencing results were reliable.

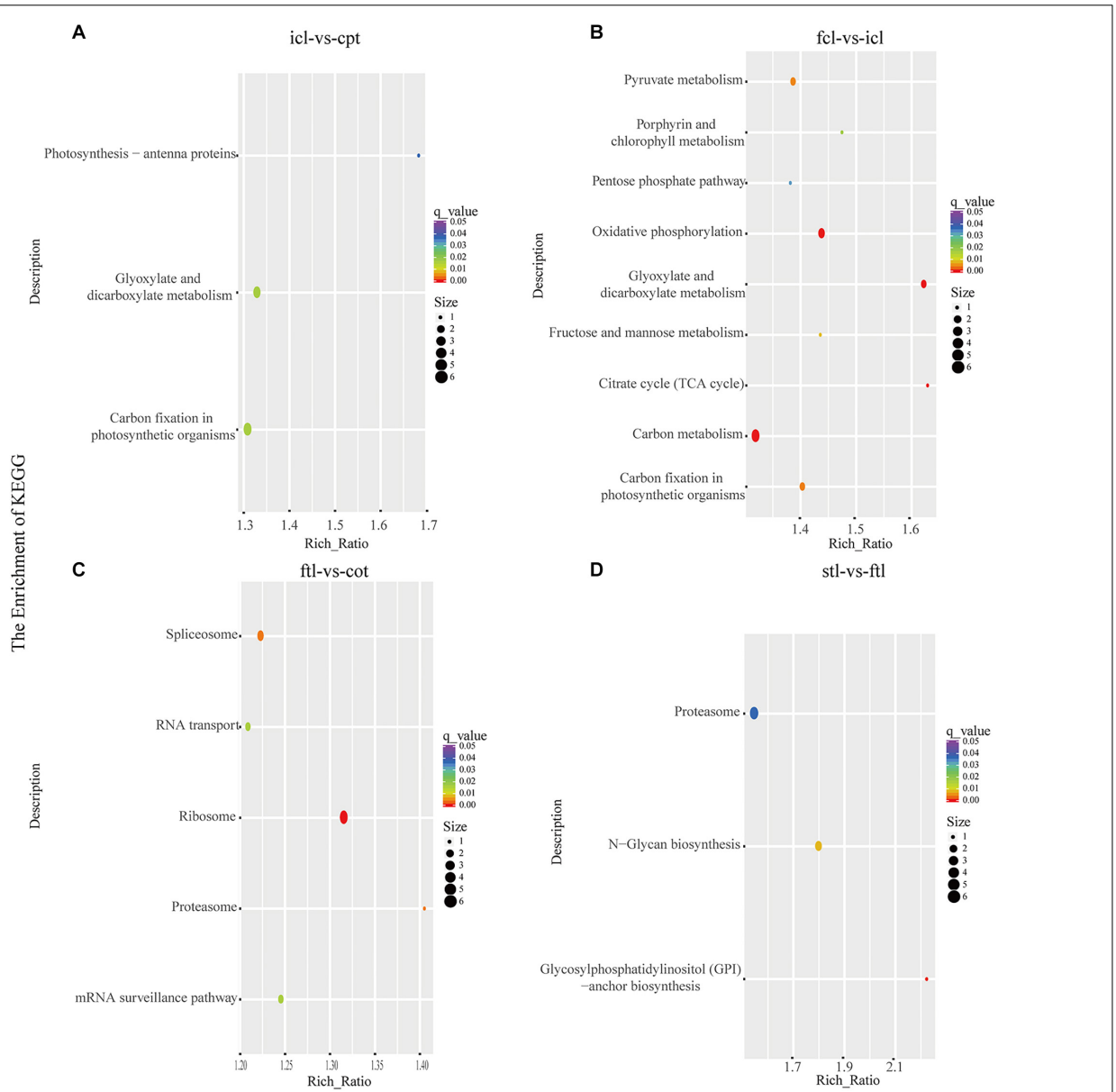


FIGURE 4 | The lncRNAs KEGG classifications of predicted target genes in different comparisons. KEGG classification of the predicted target genes in different comparisons of icl-vs.-cpt **(A)** and fcl-vs.-icl **(B)** in rice. KEGG classification of the predicted target genes in different comparisons of ftl-vs.-cot **(C)** and stl-vs.-ftl **(D)** in *Arabidopsis*. In both **(A,B)**, the y-axis shows the pathway category, and the x-axis shows the richness factor. A larger richness factor indicates greater enrichment. The size of the point represents the number of related DEGs, and the color of the bubble represents the range of *q*-values.

The light-harvesting chlorophyll protein complex (LHC) is also enriched in mRNA and lncRNA DEGs in rice. For icl-vs.-cpt, *LHCA2*, *LHCA5*, *LHCB1*, *LHCB2*, *LHCB4*, *LHCB5,* and *LHCB6* were upregulated by KEGG in both mRNA and lncRNA DEGs (**Figures 5**, **7**). For fcl-vs.-icl, *LHCA2*, *LHCA5*, *LHCB1*, *LHCB2*, *LHCB4,* and *LHCB5* were enriched in mRNA; *LHCA5*, *LHCB2*, *LHCB4,* and *LHCB5* were upregulated in lncRNA (**Figures 5**, **7**). In contrast, there was no significant difference in LHC genes in ftl-vs.-cot in either mRNA or lncRNA DEGs (**Figures 5**, **7**); *LHCB1* was upregulated and *LHCB4* was downregulated in mRNA DEGs for stl-vs.-ftl, and there was no significant difference in LHC genes in lncRNAs for stl-vs.-ftl (**Figures 5**, **7**).

Different Regulatory Styles of Assembly Factors for PSI and PSII in Early Development

The assembly of core subunits of PSII is regulated by the expression of nuclear genes and transported to plastids, and PSI assembly factors are encoded by both plastid and nuclear genes (Yang et al., 2015). We found that the expression

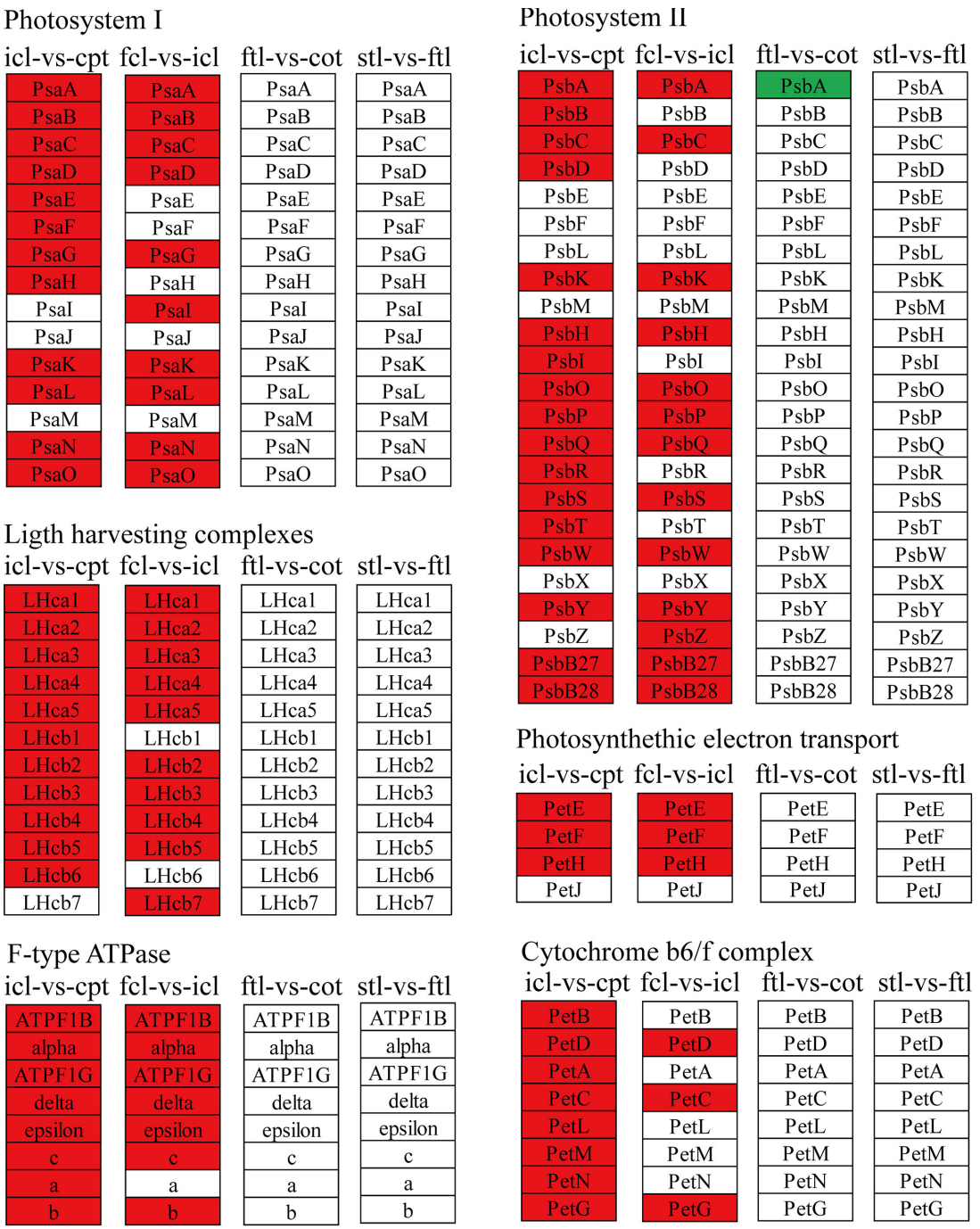


FIGURE 5 | The mRNAs KEGG classifications of photosynthetic genes in different comparisons. KEGG classification of the photosynthetic genes in different comparisons of icl-vs.-cpt and fcl-vs.-icl in rice, ftl-vs.-cot and stl-vs.-ftl in *Arabidopsis*. The change of expression amount was expressed by color change, the red color represented upregulated expression annotated to the KEGG Orthology (KO) system, the green color represented downregulated expression annotated to the KO system. The genes were selected with "$|\log_2 \text{ratio}| \geq 1$" and "$q < 0.05$."

values of the assembly factors of PSI and PSII were different between rice and *Arabidopsis* (**Figures 8A,B** and **Supplementary Table S10**). We used four genes as examples to explore the similarities and differences in PSI and PSII assembly during early development in rice and *Arabidopsis* (**Figures 8C,D**). The mRNA level of *FKBP20-2* was upregulated in icl-vs.-cpt and downregulated in fcl-vs.-icl in rice based on the sequencing results. In contrast, the expression levels were upregulated in cot-vs.-ftl and stl-vs.-ftl in *Arabidopsis*. To further visualize changes in the protein expression level, we used immunoblotting to

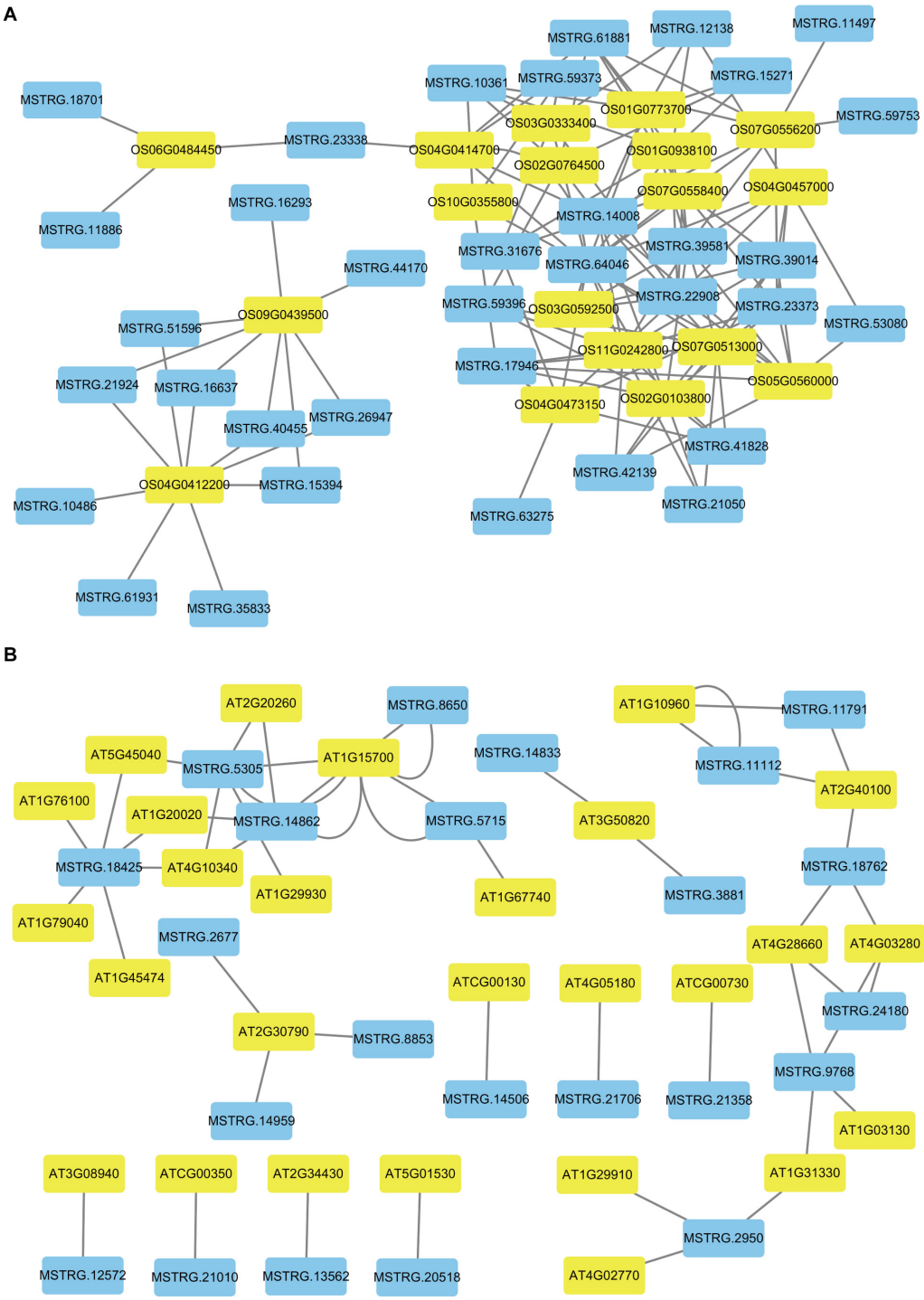


FIGURE 6 | The networks of novel lncRNAs and target mRNAs of DEGs related to photosynthesis. **(A)** Co-expression network analysis of novel lncRNAs and target mRNAs in rice. **(B)** Co-expression network analysis of novel lncRNAs and target mRNAs in *Arabidopsis*. The blue color indicated novel lncRNAs and the yellow color indicated target genes for novel lncRNAs.

examine the abundance of the assembly factors of three stages in rice and *Arabidopsis*. The results showed upregulation in icl-vs.-cpt and downregulation in fcl-vs.-icl in rice with no significant difference in cot-vs.-ftl and stl-vs.-ftl in *Arabidopsis*. For FKBP20-2, the RNA expression levels were consistent with the protein expression levels in rice for three stages but did

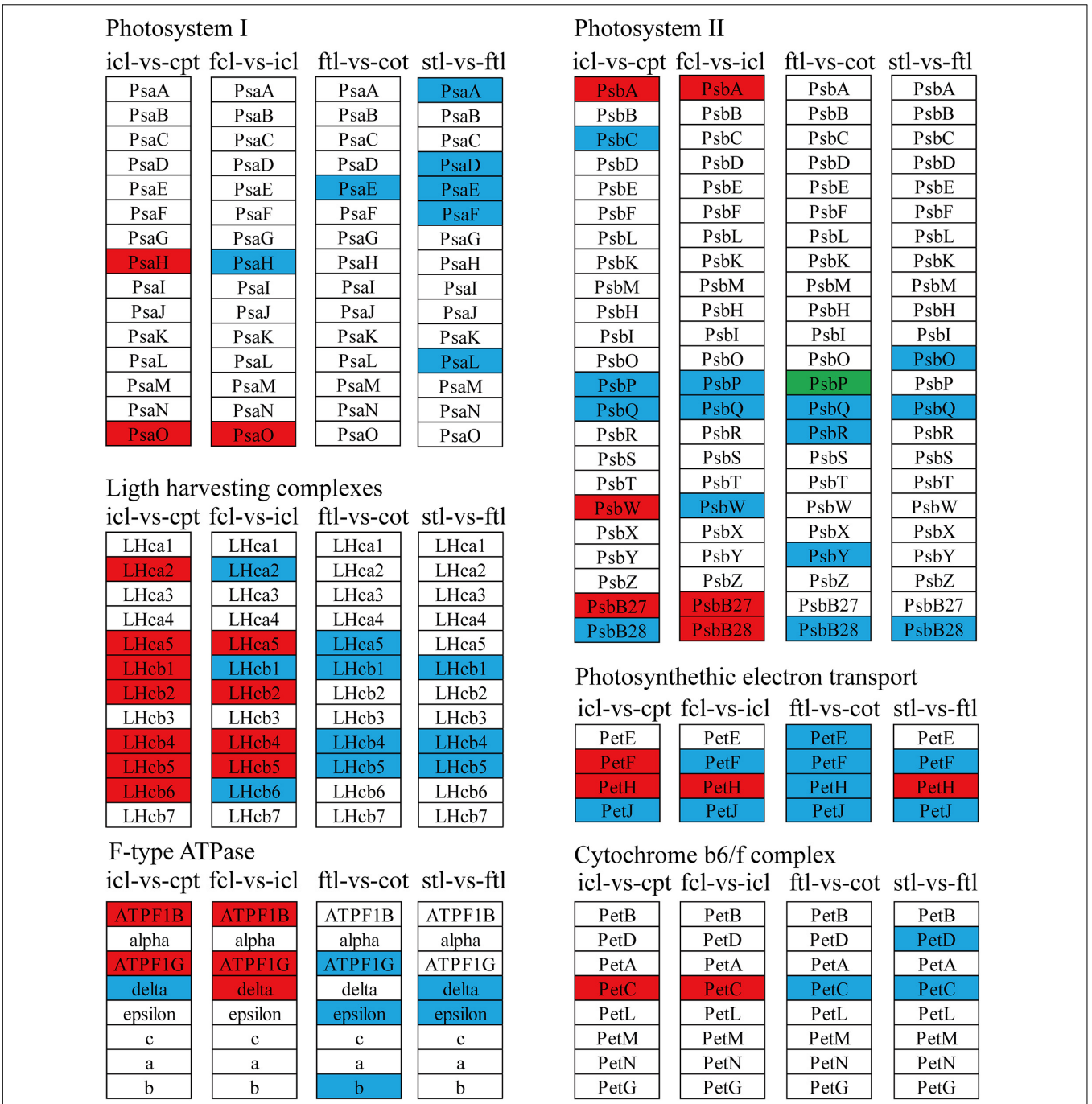


FIGURE 7 | The lncRNAs KEGG classifications of photosynthetic genes in different comparisons. KEGG classification of the photosynthetic genes in different comparisons of icl-vs.-cpt and fcl-vs.-icl in rice, ftl-vs.-cot, and stl-vs.-ftl in *Arabidopsis*. The change of expression amount was expressed by color change, the red color represented upregulated expression annotated to the KEGG Orthology (KO) system, the green color represented downregulated expression annotated to the KO system, blue color indicates that the target gene annotated to the KO system is not significantly up- and downregulated. The genes were selected with "$|\log_2 \text{ratio}| \geq 1$" and "$q < 0.05$."

not appear in *Arabidopsis*. The mRNA levels of *LQY1* were upregulated in icl-vs.-cpt and fcl-vs.-icl in rice, upregulated in cot-vs.-ftl and downregulated in stl-vs.-ftl in *Arabidopsis*. The immunoblots showed upregulation in icl-vs.-cpt and fcl-vs.-icl in rice, downregulation in cot-vs.-ftl and upregulation in stl-vs.-ftl in *Arabidopsis*. For LQY1, the RNA expression levels were consistent with the protein expression levels in rice and for three stages but did not appear in *Arabidopsis*. The mRNA levels of *TLP18.3* were upregulated in icl-vs.-cpt and fcl-vs.-icl in rice, downregulated in cot-vs.-ftl and upregulated in stl-vs.-ftl

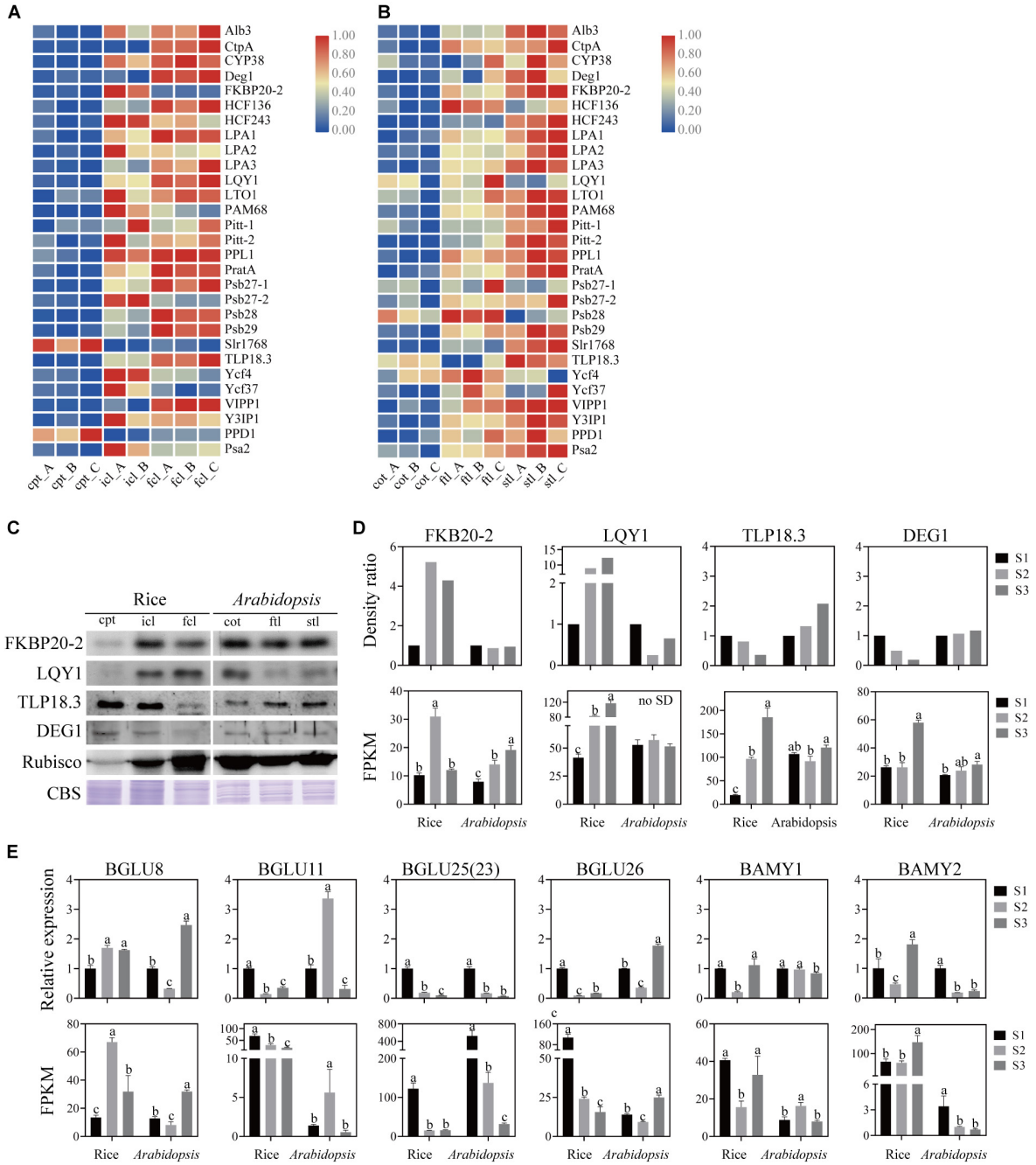


FIGURE 8 | Analysis of PSII and PSI assembly factors proteins and genes, starch and sucrose metabolism pathways marker genes of rice and *Arabidopsis* in three stages. Heatmap of the expression of photosystems assembly factors in rice **(A)** and *Arabidopsis* **(B)** at different stages. The gene expression levels are shown in different colors indicated by the scale bar. **(C)** Analysis of photosystems assembly factor proteins in rice and *Arabidopsis* at different stages. 16 μg total proteins were separated by 12% SDS PAGE. And immunoblots were performed using antibodies as indicated. CBS, Coomassie Blue Staining. **(D)** The density ratio levels of photosystems assembly factor proteins in **(C)** and RNA-seq-based gene expression values (FPKM) of relavent genes in rice and *Arabidopsis*. Data was presented as the means ± SD of three biological replicates. a, b, and c indicate significant differences between the three stages according to Tukey's test. **(E)** The relative expression levels of starch and sucrose metabolism pathways marker genes measured by real-time RT-PCR and FPKM of relavent genes in rice and *Arabidopsis*. Data was presented as means ± SD of three biological replicates. a, b, and c indicate significant differences between the three stages according to Tukey's test.

in *Arabidopsis*. The immunoblots showed downregulation in icl-vs.-cpt and fcl-vs.-icl in rice and upregulation in cot-vs.-ftl and stl-vs.-ftl in *Arabidopsis*. For TLP18.3, the RNA expression levels were not consistent with protein expression levels in rice and for three stages, but they appeared in *Arabidopsis*. The mRNA level of *Deg1* was not significantly different in icl-vs.-cpt and was

upregulated in fcl-vs.-icl in rice with no significant difference in cot-vs.-ftl and upregulated in stl-vs.-ftl in *Arabidopsis*. The immunoblots showed downregulated icl-vs.-cpt and fcl-vs.-icl in rice, with no significant difference in cot-vs.-ftl and stl-vs.-ftl in *Arabidopsis*. For Deg1, the RNA expression levels were not consistent with the protein expression levels in rice and *Arabidopsis* for three stages. Taken together, these results indicate that assembling the individual subunits to form PSII and PSI may not occur at the transcriptional level.

Different Patterns of Starch and Sucrose Metabolism in Early Development

Starch is a storage carbohydrate widely synthesized in plants. Starch serves as an important store of carbon that fuels plant metabolism and growth when they are unable to photosynthesize. The "starch and sucrose metabolism" terms were all clearly enriched among the DEGs in rice and *Arabidopsis*. To test the similarities and differences in expression patterns, four β-glucosidase (EC:3.2.1.21) genes, α-amylase (EC:3.2.1.1) gene and β-amylase (EC:3.2.1.2) gene were analyzed. Similarly, the qRT-PCR expression patterns of mRNA and lncRNA genes were consistent with the sequencing results (**Figure 8E**). These results indicate that although DEGs of starch and sucrose metabolism pathways are expressed both in rice and *Arabidopsis*, there are significant differences in the expression trend of key genes between the two species.

DISCUSSION

In this study, significant advances have been made in understanding the photosynthetic gene functions from coleoptile, incomplete leaf to first complete leaf in rice, from cotyledon, first true leaf to second true leaf in *Arabidopsis*. Many protein coding genes have been identified by traditional genetic and molecular analysis combined with recent genomic studies, which can act as positive or negative regulators of photomorphogenesis in seedlings under different light conditions (Jiao et al., 2007; Chen and Chory, 2011; Li et al., 2012). lncRNAs are an important class of non-coding RNAs (ncRNAs) which have been proved to have multiple functions in eukaryotes recently (Wang and Chang, 2011; Guttman and Rinn, 2012; Geisler and Coller, 2013). lncRNAs have been systematically identified or predicted in *Arabidopsis*, rice, wheat, and maize (De Lucia and Dean, 2011; Kim and Sung, 2012; Zhang and Chen, 2013). The non-coding RNA HIDDEN TREASURE 1 (HID1) is a positive regulator of photomorphogenesis (Wang et al., 2014a). lncRNAs can regulate photoperiod-sensitive male sterility (PSMS) in rice (Ding et al., 2012). Chloroplasts are green plastids found in land plants, algae, and some protists. Chloroplasts are the only site of photosynthesis in these cells. Therefore, chloroplasts are responsible for most of the world's primary productivity (Jarvis, 2011). In our study, chlorophyll fluorescence (Fv/Fm) from icl, cpt to fcl increased progressively in rice, indicating that the development of rice chloroplasts matured, in turn, from coleoptile to true leaf. In addition, western blotting also indicated that the development and maturation process of rice chloroplasts exhibits gradual maturation. In contrast, in *Arabidopsis*, the Fv/Fm ratio from cot, first true leaf to second true leaf was almost unchanged, and proteins analysis showed the same result. Therefore, there were significant differences in the early development of chloroplasts between rice and *Arabidopsis*.

In this study, we investigated mRNA-dependent mRNA-lncRNA interactions in different stages for early growth and development using bioinformatics approaches. Chloroplast is an essential semiautonomous organelle, which plays an important role in photosynthesis, carbon fixation, oxygen production, energy conversion, starch production, amino acid synthesis, fatty acid synthesis, pigment synthesis, hormone synthesis and sulfur and nitrogen metabolism (Neuhaus and Emes, 2000; Jarvis and Lopez-Juez, 2013; Xu et al., 2016; Wang et al., 2018). For rice, it is essential to know the development of photosynthetic genes. The gene expression of rice coleoptile and incomplete chloroplasts was significantly lower than that of the first complete leaf. This finding indicates that the photosynthetic genes are fully developed during the first complete development period. In contrast, in *Arabidopsis*, the expression of photosynthetic genes in cotyledons was similar to that in true leaves. This finding indicates that the photosynthetic genes have been developed in the cotyledon stage. In our study, we found that the function of the first complete leaf of rice is similar to that of the cotyledon of *Arabidopsis* in chloroplast development.

GO cluster analysis of rice showed that in the three stages of early development growth and development, the DEGs in biological pathways were mainly related to photosynthesis both in mRNA and lncRNA. Additionally, KEGG pathway analysis showed that mRNAs in the photosynthesis network were enriched in "Photosynthesis antenna proteins," "Photosynthesis," and "Starch and sucrose metabolism" in icl-vs.-cpt. "Photosynthesis," "Starch and sucrose metabolism," and "Photosynthesis antenna proteins" in fcl-vs.-icl. In *Arabidopsis*, no photosynthesis-related biological pathways were enriched in the ftl-vs.-cot and stl-vs.-ftl comparisons. Interestingly, the "starch and sucrose metabolism" pathway was clearly enriched in the ftl-vs.-cot and stl-vs.-ftl systems. KEGG analysis of lncRNAs revealed that up- and downregulated photosynthesis-related genes in rice were "Photosynthesis – antenna proteins," "Carbon fixation in photosynthetic organisms" in icl-vs.-cpt. "Carbon fixation in photosynthetic organisms" and "porphyrin and chlorophyll metabolism" in fcl-vs.- icl (**Figure 4B**). Additionally, "Carbon metabolism" and "Citrate cycle (TCA cycle)" were clearly enriched in fcl-vs.-icl. In *Arabidopsis*, no photosynthesis-related biological pathways were enriched in the ftl-vs.-cot and stl-vs.-ftl comparisons.

The main DEGs were chloroplast development and the carbon cycle pathway in early rice growth and development. Sucrose and starch contents due to alterations of the allocation pattern of photosynthetic fixed carbon (Higuchi et al., 2015; Pilkington et al., 2015). Temporary starch exists in cells with photosynthetic capacity, synthesized in the chloroplast during the day, and degraded at night to provide carbon for non-photosynthetic metabolism (Lunn, 2007). Leaf starch and sucrose allocation is a well-regulated process in carbohydrate metabolism. The amount of starch synthesis seems to be in harmony with the

length of the dark period (Cordenunsi-Lysenko et al., 2019). Starch-sucrose distribution is affected by plant development and sink activity, as well as environmental factors such as light intensity, temperature and photoperiod length (Pilkington et al., 2015). This finding indicates the involvement of lncRNAs in the regulation of chloroplast-related biological processes in rice. In *Arabidopsis*, the KEGG analysis of differentially expressed mRNAs revealed that the main enrichments in ftl-vs.-cot were phenylpropanoid biosynthesis, starch and sucrose metabolism, phenylalanine metabolism and cutin, suberine and wax biosynthesis; phenylpropanoid biosynthesis, flavonoid biosynthesis, circadian rhythm-plant and phenylalanine metabolism in stl-vs.-ftl. The KEGG analysis of differentially expressed lncRNAs showed that the main enrichments were ribosome in ftl-vs.-cot and glycosylphosphatidylinositol (GPI)-anchor biosynthesis in stl-vs.-ftl. This finding suggests that lncRNAs may not participate in chloroplast development and related biological processes in *Arabidopsis*. These results indicate that there are important differences between monocotyledon rice and dicotyledon *Arabidopsis* in early growth and development. In addition, we found that the plant could not grow without supplied nutrients for *Arabidopsis* (**Figure 1**), but rice grew normally under these conditions. It has been proven that there are differences between rice and *Arabidopsis* in early growth and development.

The chloroplast of eukaryotic cells is developed from an ancient endosymbiont related to cyanobacteria, which gives plants the ability of oxygenic photosynthesis (Gould et al., 2008; Nickelsen and Rengstl, 2013). The chloroplast is a semiautonomous organelle. The chloroplast has retained some genes, and different subunits of organellar protein complexes are derived from different genetic systems. PSII assembly factors are all encoded by the nucleus and transported to chloroplasts, and PSI assembly factors are encoded by nuclear genes and photosynthetic genes. From the mRNA sequencing data, we concluded that the assembly processes of PSI and PSII are not exactly the same. The expression trends of the four assembly genes for PSI and PSII differed from those of rice and *Arabidopsis*. FKBP20-2 participates in the assembly of the PSII supercomplexes in the chloroplast thylakoid lumen (Lima et al., 2006). LQY1 protein is involved in PSII photo-repair and reassembly of PSII complexes (Lu et al., 2011). TLP18.3 protein was involved in PSII regulating repair, turnover of damaged D1 and reassembly of PSII (Sirpio et al., 2007). Deg1 assists the assembly of the PSII complex through interaction with the PSII reaction center D2 protein (Sun et al., 2010). Together with the western blot results, FKB20-2 and LQY1 had the same expression pattern, but TLP18.3 and DEG1 had the opposite expression pattern of the three stages in rice. Similarly, in *Arabidopsis*, except for DEG1, FKB20-2, LQY1, and TLP18.3 had the opposite expression pattern of the three stages. The qRT–PCR and western blot results indicated that the PSI and PSII assembly were not correlated with its transcription level in the two species. Moreover, different expression levels of assembly factors in rice and *Arabidopsis* lead to different functions of PSI and PSII.

We take starch and sucrose metabolism as an example. The four β-glucosidase (EC:3.2.1.21) genes, α-amylase (EC:3.2.1.1) gene and β-amylase (EC:3.2.1.2) gene were analyzed. *BGLU25*/*BGLU23* represents an independent class of six thioglucoside glucohydrolases (Nakano et al., 2017), *BGLU8* encodes a putative beta glucosidase (Wang et al., 2017), *BGLU26* encodes an atypical myrosinase (Matern et al., 2019), *BGLU11* encodes beta glucosidase 11. *BAMY1* is involved in diurnal starch degradation in guard cells (Zanella et al., 2016), *AMY2/BAMY2* was a mixed-function OSC capable of synthesizing both β-amyrin and lupeol (Iturbe-Ormaetxe et al., 2003). In rice, β-glucosidase (EC:3.2.1.21), as a carbohydrate with biological functions, participates in energy acquisition, cell wall extension and signal transduction through selective hydrolysis of glycoside bonds in plants and plays an important role in the defense of pathogens (Davies and Henrissat, 1995; Mizutani et al., 2002; Barleben et al., 2005; Lee et al., 2006; Morant et al., 2008); alpha-amylase (EC:3.2.1.1) acts on the alpha-1,4-glycoside bond of starch but cannot cut the alpha-1,6-glycoside bond at the branch of amylopectin, thereby hydrolyzing the substrate (Wang et al., 2006); glucan 1,3-β-glucosidase (EC:3.2.1.58) is a growth-specific cell wall polysaccharide that exists in the endosperm of members of the Gramineae family (grasses) and commercially important grains (such as barley, rice, and wheat) among the main components of the cell wall (Hwang du et al., 2007). Those biological processes were significantly downregulated; endoglucanase (EC:3.2.1.4) hydrolyzed cellulose randomly in amorphous regions, producing oligosaccharides of different lengths, cellobiose, and glucose (Klippel et al., 2019), and showed significant upregulation in icl-vs.-cpt; in addition, no significant difference in fcl-vs.-icl. In *Arabidopsis,* the main upregulated gene is β-glucosidase (EC:3.2.1.21), and the downregulated gene is α-amylase (EC:3.2.1.1) in ftl-vs.-cot. The upregulated gene is β-glucosidase (EC:3.2.1.21), and the downregulated gene is β-amylase (EC:3.2.1.2), which hydrolyzes every other α-1,4-glycoside bond from the non-reductive end of starch substrate once, and the hydrolysate is maltose unit (Wang et al., 2006) in stl-vs.-ftl. In addition, no lncRNA DEGs showed significant differences in the three stages of the starch and sucrose metabolism pathways.

CONCLUSION

mRNA and lncRNAs sequencing revealed that photosynthesis was gradually established from coleoptile, incomplete leaf to complete leaf in rice and was fully functional in cotyledon of *Arabidopsis*. In summary, we examined the gene expression patterns of mRNAs and lncRNAs at different developmental stages of early growth and development in rice and *Arabidopsis*. In our study, functional modes were revealed for mRNAs and lncRNAs, the main DEGs enriched in multiple biological processes, such as photosynthetic genes and photosynthetic carbon assimilation genes. We found that the chloroplast protein from the coleoptile, incomplete leaf and first complete leaf gradually matured in rice. In contrast, in *Arabidopsis,* the maturation of chloroplast protein was immediately complete at the cotyledon stage. The results showed that monocotyledons and dicotyledons (at least in rice and *Arabidopsis*) differed greatly in

photosynthetic mechanisms. In conclusion, our studies revealed whole genome expression profiles of mRNAs and lncRNAs related to early growth and development in rice and *Arabidopsis,* thereby providing important evidence for further research on the molecular mechanisms of early growth and development.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/, SRR10907568–SRR10907573; https://www.ncbi.nlm.nih.gov/, SRR10913120-SRR10913123; https://www.ncbi.nlm.nih.gov/, SRR10913307–SRR10913312; https://www.ncbi.nlm.nih.gov/, SRR10914245–SRR10914250; https://www.ncbi.nlm.nih.gov/, SRR10914757–SRR10914762; https://www.ncbi.nlm.nih.gov/, SRR10915176–SRR10915181.

AUTHOR CONTRIBUTIONS

YS and XH designed the research and wrote the manuscript. YS and JC performed the research. All authors analyzed the data and read and approved the manuscript.

FUNDING

This work was supported by National Key Research and Development Program (2016YFD0100604).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.565006/full#supplementary-material

FIGURE S1 | The qualitative analysis of novel lncRNAs and the relationship about lncRNAs and mRNAs. The qualitative analysis of novel lncRNAs and the relationship about lncRNAs and mRNAs. Novel lncRNA exon distribution in rice **(A)** and *Arabidopsis* **(B)**; Novel lncRNA length distribution in rice **(C)** and *Arabidopsis* **(D)**; Exon number contrast of lncRNA vs. mRNA in rice **(E)** and *Arabidopsis* **(F)**; The length contrast of lncRNA vs. mRNA in rice **(G)** and *Arabidopsis* **(H)**.

FIGURE S2 | Cluster map of differential genes rice and *Arabidopsis*. Cluster map of differential genes rice mRNAs **(A)**, rice lncRNAs **(B)**, *Arabidopsis* mRNAs **(C),** and *Arabidopsis* lncRNAs **(D)** The change of expression amount was expressed by color change, the yellow color represented higher expression, and the blue color represented lower expression.

FIGURE S3 | The RNA-seq data and qRT-PCR validation of lncRNAs and their target protein-coding genes. The relative expression levels and RNA-seq-based gene expression values (FPKM) of six lncRNAs and their target protein-coding genes were shown. A lncRNA can have one or more target protein-coding genes. The bars denote the standard deviation.

FIGURE S4 | Verification of the transcriptome results through the experiments of qRT-PCR. The relationships between qRT-PCR and RNA-seq. Values were the $\log_2$ ratio (FPKM ratio by RNA-Seq) for genes in rice **(A)** and *Arabidopsis* **(B)**. The determine coefficient (R^2) was indicated in the figure. All qRT-PCR reactions were performed in three biological replicates.

TABLE S1 | The primers used for quantitative real-time PCR.

TABLE S2 | Summary of reads mapping to rice and *Arabidopsis* genome in three developmental stages.

TABLE S3 | Summary of reads quality of rice and *Arabidopsis* in three developmental stages.

TABLE S4 | Number and proportion of unique alignment sequences on three functional elements of gene (exon, intron and intergenic) in rice and *Arabidopsis*.

TABLE S5 | The normalized FPKM values for all lncRNAs and mRNAs in all samples in rice and *Arabidopsis*.

TABLE S6 | Gene ontology of biological pathways enriched in the transgenic lines based on the up- or down-regulated genes.

TABLE S7 | Trans and *cis* target photosynthetic genes of novel lncRNAs in rice and *Arabidopsis*.

TABLE S8 | The coherent target genes of differentially expressed lncRNAs in all comparisons.

TABLE S9 | The relationships between qRT-PCR analysis for validation of RNA seq data analysis in rice and *Arabidopsis*.

TABLE S10 | Summary of described PSII assembly factors in all samples in rice and *Arabidopsis*.

REFERENCES

Amasino, R. M., and Michaels, S. D. (2010). The timing of flowering. *Plant Physiol.* 154, 516–520. doi: 10.1104/pp.110.161653

Barleben, L., Ma, X., Koepke, J., Peng, G., Michel, H., and Stockigt, J. (2005). Expression, purification, crystallization and preliminary X-ray analysis of strictosidine glucosidase, an enzyme initiating biosynthetic pathways to a unique diversity of indole alkaloid skeletons. *Biochim. Biophys. Acta* 1747, 89–92. doi: 10.1016/j.bbapap.2004.09.026

Baud, S., Boutin, J. P., Miquel, M., Lepiniec, L., and Rochat, C. (2002). An integrated overview of seed development in *Arabidopsis thaliana* ecotype WS. *Plant Physiol. Biochem.* 40, 151–160.

Baud, S., Dubreucq, B., Miquel, M., Rochat, C., and Lepiniec, L. (2008). Storage reserve accumulation in *Arabidopsis*: metabolic and developmental control of seed filling. *Arabidopsis Book* 6:e0113. doi: 10.1199/tab.0113

Becraft, P. W., and Asuncion-Crabb, Y. (2000). Positional cues specify and maintain aleurone cell fate in maize endosperm development. *Development* 127, 4039–4048.

Becraft, P. W., and Yi, G. (2011). Regulation of aleurone development in cereal grains. *J. Exp. Bot.* 62, 1669–1675. doi: 10.1093/jxb/erq372

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate – A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300.

Bentsink, L., Alonso-Blanco, C., Vreugdenhil, D., Tesnier, K., Groot, S. P., and Koornneef, M. (2000). Genetic analysis of seed-soluble oligosaccharides in relation to seed storability of *Arabidopsis*. *Plant Physiol.* 124, 1595–1604.

Berger, F., Grini, P. E., and Schnittger, A. (2006). Endosperm: an integrator of seed growth and development. *Curr. Opin. Plant Biol.* 9, 664–670. doi: 10.1016/j.pbi.2006.09.015

Bhargava, A., Clabaugh, I., To, J. P., Maxwell, B. B., Chiang, Y. H., Schaller, G. E., et al. (2013). Identification of cytokinin-responsive genes using microarray meta-analysis and RNA-Seq in *Arabidopsis*. *Plant Physiol.* 162, 272–294. doi: 10.1104/pp.113.217026

Chekanova, J. A. (2015). Long non-coding RNAs and their functions in plants. *Curr. Opin. Plant Biol.* 27, 207–216. doi: 10.1016/j.pbi.2015.08.003

Chen, M., and Chory, J. (2011). Phytochrome signaling mechanisms and the control of plant development. *Trends Cell Biol.* 21, 664–671. doi: 10.1016/j.tcb.2011.07.002

Cordenunsi-Lysenko, B. R., Nascimento, J. R. O., Castro-Alves, V. C., Purgatto, E., Fabi, J. P., and Peroni-Okyta, F. H. G. (2019). The starch is (not) just another brick in the wall: the primary metabolism of sugars during banana ripening. *Front. Plant Sci.* 10:391. doi: 10.3389/fpls.2019.00391

Davies, G., and Henrissat, B. (1995). Structures and mechanisms of glycosyl hydrolases. *Structure* 3, 853–859. doi: 10.1016/S0969-2126(01)00220-9

De Lucia, F., and Dean, C. (2011). Long non-coding RNAs and chromatin regulation. *Curr. Opin. Plant Biol.* 14, 168–173. doi: 10.1016/j.pbi.2010.11.006

Deng, X. W., and Gruissem, W. (1987). Control of plastid gene expression during development: the limited role of transcriptional regulation. *Cell* 49, 379–387.

Ding, J. H., Lu, Q., Ouyang, Y. D., Mao, H. L., Zhang, P. B., Yao, J. L., et al. (2012). A long noncoding RNA regulates photoperiod-sensitive male sterility, an essential component of hybrid rice. *Proc. Natl. Acad. Sci. U.S.A.* 109, 2654–2659. doi: 10.1073/pnas.1121374109

Du, H., Wang, L., and Xiong, L. Z. (2018). Phenotypic observations of rice critical development stages. *Bio* 101:e1010178. doi: 10.21769/BioProtoc.1010178

Gao, Y., Xu, H., Shen, Y., and Wang, J. (2013). Transcriptomic analysis of rice (*Oryza sativa*) endosperm using the RNA-Seq technique. *Plant Mol. Biol.* 81, 363–378. doi: 10.1007/s11103-013-0009-4

Garcia-Cebrian, F., Esteso-Martinez, J., and Gil-Pelegrin, E. (2003). Influence of cotyledon removal on early seedling growth in *Quercus robur* L. *Ann. For. Sci.* 60, 69–73. doi: 10.1051/forest:2002075

Geisler, S., and Coller, J. (2013). RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat. Rev. Mol. Cell Biol.* 14, 699–712. doi: 10.1038/nrm3679

Gould, S. B., Waller, R. F., and McFadden, G. I. (2008). Plastid evolution. *Annu. Rev. Plant Biol.* 59, 491–517. doi: 10.1146/annurev.arplant.59.032607.092915

Guttman, M., and Rinn, J. L. (2012). Modular regulatory principles of large non-coding RNAs. *Nature* 482, 339–346. doi: 10.1038/nature10887

Hanley, M. E., and May, O. C. (2006). Cotyledon damage at the seedling stage affects growth and flowering potential in mature plants. *New Phytol.* 169, 243–250. doi: 10.1111/j.1469-8137.2005.01578.x

Heo, J. B., and Sung, S. (2011). Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. *Science* 331, 76–79. doi: 10.1126/science.1197349

Higuchi, K., Kanai, M., Tsuchiya, M., Ishii, H., Shibuya, N., Fujita, N., et al. (2015). Common reed accumulates starch in its stem by metabolic adaptation under Cd stress conditions. *Front. Plant Sci.* 6:138. doi: 10.3389/fpls.2015.00138

Hirano, H. Y. (2008). *Rice Biology in the Genomics Era*. Berlin: Springer.

Hou, X., Fu, A., Garcia, V. J., Buchanan, B. B., and Luan, S. (2015). PSB27: a thylakoid protein enabling *Arabidopsis* to adapt to changing light intensity. *Proc. Natl. Acad. Sci. U.S.A.* 112, 1613–1618. doi: 10.1073/pnas.1424040112

Hwang, du, H., Kim, S. T., Kim, S. G., and Kang, K. Y. (2007). Comprehensive analysis of the expression of twenty-seven beta-1, 3-glucanase genes in rice (*Oryza sativa* L.). *Mol. Cells* 23, 207–214.

Iturbe-Ormaetxe, I., Haralampidis, K., Papadopoulou, K., and Osbourn, A. E. (2003). Molecular cloning and characterization of triterpene synthases from *Medicago truncatula* and *Lotus japonicus*. *Plant Mol. Biol.* 51, 731–743. doi: 10.1023/a:1022519709298

Jarvis, P., and Lopez-Juez, E. (2013). Biogenesis and homeostasis of chloroplasts and other plastids. *Nat. Rev. Mol. Cell Biol.* 14, 787–802. doi: 10.1038/nrm3702

Jarvis, R. P. (2011). *Chloroplast Research in Arabidopsis: Methods and Protocols*. New York, NY: Humana Press.

Jiao, Y., Lau, O. S., and Deng, X. W. (2007). Light-regulated transcriptional networks in higher plants. *Nat. Rev. Genet.* 8, 217–230. doi: 10.1038/nrg2049

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kim, E. D., and Sung, S. (2012). Long noncoding RNA: unveiling hidden layer of gene regulatory networks. *Trends Plant Sci.* 17, 16–21. doi: 10.1016/j.tplants.2011.10.008

Klippel, B., Blank, S., Janzer, V. A., Piascheck, H., Moccand, C., Bel-Rhlid, R., et al. (2019). Characterization of a thermoactive endoglucanase isolated from a biogas plant metagenome. *Extremophiles* 23, 479–486. doi: 10.1007/s00792-019-01099-3

Kong, L., Zhang, Y., Ye, Z. Q., Liu, X. Q., Zhao, S. Q., Wei, L., et al. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* 35, W345–W349. doi: 10.1093/nar/gkm391

Krishnan, S., and Dayanandan, P. (2003). Structural and histochemical studies on grain-filling in the caryopsis of rice (*Oryza sativa* L.). *J. Biosci.* 28, 455–469. doi: 10.1007/Bf02705120

Lee, K. H., Piao, H. L., Kim, H. Y., Choi, S. M., Jiang, F., Hartung, W., et al. (2006). Activation of glucosidase via stress-induced polymerization rapidly increases active pools of abscisic acid. *Cell* 126, 1109–1120. doi: 10.1016/j.cell.2006.07.034

Leonova, S., Grimberg, A., Marttila, S., Stymne, S., and Carlsson, A. S. (2010). Mobilization of lipid reserves during germination of oat (*Avena sativa* L.), a cereal rich in endosperm oil. *J. Exp. Bot.* 61, 3089–3099. doi: 10.1093/jxb/erq141

Li, J., Terzaghi, W., and Deng, X. W. (2012). Genomic basis for light control of plant development. *Protein Cell* 3, 106–116. doi: 10.1007/s13238-012-2016-7

Lima, A., Lima, S., Wong, J. H., Phillips, R. S., Buchanan, B. B., and Luan, S. (2006). A redox-active FKBP-type immunophilin functions in accumulation of the photosystem II supercomplex in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U.S.A.* 103, 12631–12636. doi: 10.1073/pnas.0605452103

Liu, F., Marquardt, S., Lister, C., Swiezewski, S., and Dean, C. (2010). Targeted 3' processing of antisense transcripts triggers *Arabidopsis* FLC chromatin silencing. *Science* 327, 94–97. doi: 10.1126/science.1180278

Liu, H., Wang, R., Mao, B., Zhao, B., and Wang, J. (2019). Identification of lncRNAs involved in rice ovule development and female gametophyte abortion by genome-wide screening and functional analysis. *BMC Genomics* 20:90. doi: 10.1186/s12864-019-5442-6

Liu, J., Jung, C., Xu, J., Wang, H., Deng, S., Bernad, L., et al. (2012). Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in *Arabidopsis*. *Plant Cell* 24, 4333–4345. doi: 10.1105/tpc.112.102855

Liu, J., and Last, R. L. (2017). A chloroplast thylakoid lumen protein is required for proper photosynthetic acclimation of plants under fluctuating light environments. *Proc. Natl. Acad. Sci. U.S.A.* 114, E8110–E8117. doi: 10.1073/pnas.1712206114

Lu, T., Lu, G., Fan, D., Zhu, C., Li, W., Zhao, Q., et al. (2010). Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome Res.* 20, 1238–1249. doi: 10.1101/gr.106120.110

Lu, Y., Hall, D. A., and Last, R. L. (2011). A small zinc finger thylakoid protein plays a role in maintenance of photosystem II in *Arabidopsis thaliana*. *Plant Cell* 23, 1861–1875. doi: 10.1105/tpc.111.085456

Lunn, J. E. (2007). Compartmentation in plant metabolism. *J. Exp. Bot.* 58, 35–47. doi: 10.1093/jxb/erl134

Mansfield, S. G., and Briarty, L. G. (1996). The dynamics of seedling and cotyledon cell development in *Arabidopsis thaliana* during reserve mobilization. *Int. J. Plant Sci.* 157, 280–295. doi: 10.1086/297347

Mao, X. Z., Cai, T., Olyarchuk, J. G., and Wei, L. P. (2005). Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics* 21, 3787–3793. doi: 10.1093/bioinformatics/bti430

Matern, A., Bottcher, C., Eschen-Lippold, L., Westermann, B., Smolka, U., Doll, S., et al. (2019). A substrate of the ABC transporter PEN3 stimulates bacterial flagellin (flg22)-induced callose deposition in *Arabidopsis thaliana*. *J. Biol. Chem.* 294, 6857–6870. doi: 10.1074/jbc.RA119.007676

Maxwell, K., and Johnson, G. N. (2000). Chlorophyll fluorescence–a practical guide. *J. Exp. Bot.* 51, 659–668. doi: 10.1093/jxb/51.345.659

Mistry, J., Bateman, A., and Finn, R. D. (2007). Predicting active site residue annotations in the Pfam database. *BMC Bioinformatics* 8:298. doi: 10.1186/1471-2105-8-298

Mizutani, M., Nakanishi, H., Ema, J., Ma, S. J., Noguchi, E., Inohara-Ochiai, M., et al. (2002). Cloning of beta-primeverosidase from tea leaves, a key enzyme in tea aroma formation. *Plant Physiol.* 130, 2164–2176. doi: 10.1104/pp.102.011023

Morant, A. V., Jorgensen, K., Jorgensen, C., Paquette, S. M., Sanchez-Perez, R., Moller, B. L., et al. (2008). beta-glucosidases as detonators of plant chemical defense. *Phytochemistry* 69, 1795–1813. doi: 10.1016/j.phytochem.2008.03.006

Muranaka, A., Watanabe, S., Sakamoto, A., and Shimada, H. (2012). *Arabidopsis* cotyledon chloroplast biogenesis factor CYO1 uses glutathione as an electron donor and interacts with PSI (A1 and A2) and PSII (CP43 and CP47) subunits. *J. Plant Physiol.* 169, 1212–1215. doi: 10.1016/j.jplph.2012.04.001

Nakano, R. T., Pislewska-Bednarek, M., Yamada, K., Edger, P. P., Miyahara, M., Kondo, M., et al. (2017). PYK10 myrosinase reveals a functional coordination between endoplasmic reticulum bodies and glucosinolates in *Arabidopsis thaliana*. *Plant J.* 89, 204–220. doi: 10.1111/tpj.13377

Neuhaus, H. E., and Emes, M. J. (2000). Nonphotosynthetic metabolism in plastids. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 51, 111–140. doi: 10.1146/annurev.arplant.51.1.111

Nickelsen, J., and Rengstl, B. (2013). Photosystem II assembly: from cyanobacteria to plants. *Annu. Rev. Plant Biol.* 64, 609–635. doi: 10.1146/annurev-arplant-050312-120124

O'Neill, C. M., Gill, S., Hobbs, D., Morgan, C., and Bancroft, I. (2003). Natural variation for seed oil composition in *Arabidopsis thaliana*. *Phytochemistry* 64, 1077–1090.

Pang, P. P., Pruitt, R. E., and Meyerowitz, E. M. (1988). Molecular-cloning, genomic organization, expression and evolution of 12s-seed storage protein genes of *Arabidopsis thaliana*. *Plant Mol. Biol.* 11, 805–820.

Pilkington, S. M., Encke, B., Krohn, N., Hohne, M., Stitt, M., and Pyl, E. T. (2015). Relationship between starch degradation and carbon demand for maintenance and growth in *Arabidopsis thaliana* in different irradiance and temperature regimes. *Plant Cell Environ.* 38, 157–171. doi: 10.1111/pce.12381

Sela, A., Piskurewicz, U., Megies, C., Mene-Saffrane, L., Finazzi, G., and Lopez-Molina, L. (2020). Embryonic photosynthesis affects post-germination plant growth. *Plant Physiol.* 182, 2166–2181. doi: 10.1104/pp.20.00043

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

Shimada, H., Mochizuki, M., Ogura, K., Froehlich, J. E., Osteryoung, K. W., Shirano, Y., et al. (2007). *Arabidopsis* cotyledon-specific chloroplast biogenesis factor CYO1 is a protein disulfide isomerase. *Plant Cell* 19, 3157–3169. doi: 10.1105/tpc.107.051714

Sirpio, S., Allahverdiyeva, Y., Suorsa, M., Paakkarinen, V., Vainonen, J., Battchikova, N., et al. (2007). TLP18.3, a novel thylakoid lumen protein regulating photosystem II repair cycle. *Biochem. J.* 406, 415–425. doi: 10.1042/BJ20070460

Sun, L., Luo, H., Bu, D., Zhao, G., Yu, K., Zhang, C., et al. (2013). Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res.* 41, e166. doi: 10.1093/nar/gkt646

Sun, X., Ouyang, M., Guo, J., Ma, J., Lu, C., Adam, Z., et al. (2010). The thylakoid protease Deg1 is involved in photosystem-II assembly in *Arabidopsis thaliana*. *Plant J.* 62, 240–249. doi: 10.1111/j.1365-313X.2010.04140.x

Tafer, H., and Hofacker, I. L. (2008). RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics* 24, 2657–2663. doi: 10.1093/bioinformatics/btn193

Tian, Y., Zheng, H., Zhang, F., Wang, S., Ji, X., Xu, C., et al. (2019). PRC2 recruitment and H3K27me3 deposition at FLC require FCA binding of COOLAIR. *Sci. Adv.* 5:eaau7246. doi: 10.1126/sciadv.aau7246

Tominaga, J., Mizutani, H., Horikawa, D., Nakahara, Y., Takami, T., Sakamoto, W., et al. (2016). Rice CYO1, an ortholog of *Arabidopsis thaliana* cotyledon chloroplast biogenesis factor AtCYO1, is expressed in leaves and involved in photosynthetic performance. *J. Plant Physiol.* 207, 78–83. doi: 10.1016/j.jplph.2016.10.005

Wang, G. F., Wang, F., Wang, G., Wang, F., Zhang, X. W., Zhong, M. Y., et al. (2012). Opaque1 encodes a myosin XI motor protein that is required for endoplasmic reticulum motility and protein body formation in maize endosperm. *Plant Cell* 24, 3447–3462. doi: 10.1105/tpc.112.101360

Wang, J., Wang, Y., Gao, C., Jiang, L., and Guo, D. (2017). PPero, a computational model for plant PTS1 type peroxisomal protein prediction. *PLoS ONE* 12:e0168912. doi: 10.1371/journal.pone.0168912

Wang, K. C., and Chang, H. Y. (2011). Molecular mechanisms of long noncoding RNAs. *Mol. Cell* 43, 904–914. doi: 10.1016/j.molcel.2011.08.018

Wang, L., Park, H. J., Dasari, S., Wang, S. Q., Kocher, J. P., and Li, W. (2013). CPAT: coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Res.* 41:e74.

Wang, N., Zhang, Y., Wang, Q., Liu, J., Wang, H., Xue, Y., et al. (2006). Gene cloning and characterization of a novel alpha-amylase from alkaliphilic *Alkalimonas amylolytica*. *Biotechnol. J.* 1, 1258–1265. doi: 10.1002/biot.200600098

Wang, Y., Fan, X., Lin, F., He, G., Terzaghi, W., Zhu, D., et al. (2014a). *Arabidopsis* noncoding RNA mediates control of photomorphogenesis by red light. *Proc. Natl. Acad. Sci. U.S.A.* 111, 10359–10364. doi: 10.1073/pnas.1409457111

Wang, Y., Wang, X., Deng, W., Fan, X., Liu, T. T., He, G., et al. (2014b). Genomic features and regulatory roles of intermediate-sized non-coding RNAs in Arabidopsis. *Mol Plant* 7, 514–527. doi: 10.1093/mp/sst177

Wang, Z. W., Lv, J., Xie, S. Z., Zhang, Y., Qiu, Z. N., Chen, P., et al. (2018). OsSLA4 encodes a pentatricopeptide repeat protein essential for early chloroplast development and seedling growth in rice. *Plant Growth Regul.* 84, 249–260. doi: 10.1007/s10725-017-0336-6

Wu, X., Liu, J., Li, D., and Liu, C. M. (2016). Rice caryopsis development II: dynamic changes in the endosperm. *J. Integr. Plant Biol.* 58, 786–798. doi: 10.1111/jipb.12488

Xiong, Q., Ma, B., Lu, X., Huang, Y. H., He, S. J., Yang, C., et al. (2017). Ethylene-inhibited jasmonic acid biosynthesis promotes mesocotyl/coleoptile elongation of etiolated rice seedlings. *Plant Cell* 29, 1053–1072. doi: 10.1105/tpc.16.00981

Xu, X., Zhang, X. B., Shi, Y. F., Wang, H. M., Feng, B. H., Li, X. H., et al. (2016). A point mutation in an F-Box domain-containing protein is responsible for brown hull phenotype in rice. *Rice Sci.* 23, 1–8. doi: 10.1016/j.rsci.2016.01.001

Yang, H., Liu, J., Wen, X., and Lu, C. (2015). Molecular mechanism of photosystem I assembly in oxygenic organisms. *Biochim. Biophys. Acta* 1847, 838–848. doi: 10.1016/j.bbabio.2014.12.011

Young, M. D., Wakefield, M. J., Smyth, G. K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* 11:R14.

Yuan, J., Zhang, Y., Dong, J., Sun, Y., Lim, B. L., Liu, D., et al. (2016). Systematic characterization of novel lncRNAs responding to phosphate starvation in *Arabidopsis thaliana*. *BMC Genomics* 17:655. doi: 10.1186/s12864-016-2929-2

Zanella, M., Borghi, G. L., Pirone, C., Thalmann, M., Pazmino, D., Costa, A., et al. (2016). beta-amylase 1 (BAM1) degrades transitory starch to sustain proline biosynthesis during drought stress. *J. Exp. Bot.* 67, 1819–1826. doi: 10.1093/jxb/erv572

Zhang, H. X., Zhou, D. W., Matthew, C., Wang, P., and Zheng, W. (2008). Photosynthetic contribution of cotyledons to early seedling development in *Cynoglossum divaricatum* and *Amaranthus retroflexus*. *New. Z. J. Bot.* 46, 39–48.

Zhang, Y. C., and Chen, Y. Q. (2013). Long noncoding RNAs: new regulators in plant development. *Biochem. Biophys. Res. Commun.* 436, 111–114. doi: 10.1016/j.bbrc.2013.05.086

Zheng, W., Yang, J. Y., Luan, Z. H., Wang, P., Zhang, H. X., and Zhou, D. W. (2011). Compensatory growth and photosynthetic responses of *Pharbitis purpurea* seedlings to clipped cotyledon and second leaf. *Photosynthetica* 49, 21–28. doi: 10.1007/s11099-011-0004-4

Zheng, Y., and Wang, Z. (2015). The cereal starch endosperm development and its relationship with other endosperm tissues and embryo. *Protoplasma* 252, 33–40. doi: 10.1007/s00709-014-0687-z

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

Copyright © 2020 Shi, Chen and Hou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 11 September 2020
doi: 10.3389/fgene.2020.00788

Comprehensive Analysis of Respiratory Burst Oxidase Homologs (Rboh) Gene Family and Function of *GbRboh5/18* on Verticillium Wilt Resistance in *Gossypium barbadense*

*Ying Chang†, Bo Li†, Qian Shi, Rui Geng, Shuaipeng Geng, Jinlei Liu, Yuanyuan Zhang and Yingfan Cai**

State Key Laboratory of Cotton Biology, Key Laboratory of Plant Stress Biology, School of Life Sciences, Henan University, Kaifeng, China

OPEN ACCESS

Edited by:
Hongjian Wan, ZheJiang Academy of Agricultural Sciences, China

Reviewed by:
Huiming Guo, Chinese Academy of Agricultural Sciences (CAAS), China
Zhikun Li, Agricultural University of Hebei, China

***Correspondence:**
Yingfan Cai
yingfancai@outlook.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *19 April 2020*
Accepted: *02 July 2020*
Published: *11 September 2020*

Citation:
Chang Y, Li B, Shi Q, Geng R, Geng S, Liu J, Zhang Y and Cai Y (2020) Comprehensive Analysis of Respiratory Burst Oxidase Homologs (Rboh) Gene Family and Function of GbRboh5/18 on Verticillium Wilt Resistance in Gossypium barbadense. Front. Genet. 11:788. doi: 10.3389/fgene.2020.00788

Respiratory burst oxidase homologs (Rbohs) play a predominant role in reactive oxygen species (ROS) production, which is crucial in plant growth, differentiation, as well as their responses to biotic and abiotic stresses. To date, however, there is little knowledge about the function of cotton *Rboh* genes. Here, we identified a total of 87 *Rbohs* from five sequenced *Gossypium* species (the diploids *Gossypium arboreum*, *Gossypium raimondii*, and *Gossypium australe*, and the allotetraploids *Gossypium hirsutum* and *Gossypium barbadense*) via BLAST searching their genomes. Phylogenetic analysis of the putative 87 cotton *Rbohs* revealed that they were divided into seven clades. All members within the same clade are generally similar to each other in terms of gene structure and conserved domain arrangement. In *G. barbadense*, the expression levels of *GbRbohs* in the CladeD were induced in response to a fungal pathogen and to hormones (i.e., jasmonic acid and abscisic acid), based upon which the main functional member in CladeD was discerned to be *GbRboh5/18.* Further functional and physiological analyses showed that the knock-down of *GbRboh5/18* expression attenuates plant resistance to *Verticillium dahliae* infection. Combined with the molecular and biochemical analyses, we found less ROS accumulation in *GbRboh5/18-VIGS* plants than in control plants after *V. dahliae* infection. Overexpression of *GbRboh5/18* in *G. barbadense* resulted in more ROS accumulation than in control plants. These results suggest that *GbRboh5/18* enhances the cotton plants' resistance against *V. dahliae* by elevating the levels of ROS accumulation. By integrating phylogenetic, molecular, and biochemical approaches, this comprehensive study provides a detailed overview of the number, phylogeny, and evolution of the *Rboh* gene family from five sequenced

Gossypium species, as well as elucidating the function of *GbRboh5/18* for plant resistance against *V. dahliae*. This study sheds fresh light on the molecular evolutionary properties and function of *Rboh* genes in cotton, and provides a reference for improving cotton's responses to the pathogen *V. dahliae*.

Keywords: respiratory burst oxidase homologs (Rbohs), reactive oxygen species (ROS), *Gossypium*, *V. dahliae*, *GbRboh5/18*

INTRODUCTION

Reactive oxygen species (ROS) play dual roles in plant growth, and development, and in their responses to biotic and abiotic stresses (Torres and Dangl, 2005; Suzuki et al., 2011). Hydrogen peroxide (H_2O_2), the most stable species of ROS, can serve as a secondary signaling molecule for transduction in the cytoplasmic at a low basal level. However, because of the damage they cause to plant cell's DNA, proteins, and lipids, ROS species are toxic when the accumulation exceeds a certain threshold, which can occur for the superoxide anion radical (O_2^-), hydroxyl radical (·OH), and H_2O_2 (Baxter et al., 2014).

Respiratory burst oxidase homologs (Rbohs), which are plant-specific NADPH oxidases, are key enzymes that move electrons from cytoplasmic NADPH to O_2 and thus foster the generation of O_2^-, followed by the dismutation of O_2^- into H_2O_2 (Torres et al., 1998; Sagi and Fluhr, 2006). The first plant *Rboh* gene was found in rice, named *OsRbohA*, for which the mammalian homolog is $gp91^{phox}$/*Nox2* (Keller et al., 1998). Then more Rbohs from plants were successively identified from model plants (including Arabidopsis and tobacco) and crops, such as tomato, potato, and cucumber (Torres et al., 1998; Yoshioka et al., 2003; Sagi et al., 2004; Kobayashi et al., 2006; Xia et al., 2009; Yu et al., 2019). Generally, plant Rboh proteins are best described as integral plasma membrane proteins containing six transmembrane (TM) domains. The heme groups usually bind two pairs of histidine residues, respectively located in third and fifth TMs, to facilitate electron transport. Two Ca^{2+}-binding EF-hand motifs and phosphorylation sites in the cytological N-terminal region participate in regulating enzymatic activity. Similar to the mammalian $gp91^{phox}$, there are flavin adenine dinucleotide (FAD) and NADPH-binding domains on the C-terminus of Rboh proteins in plants (Wong et al., 2007).

Ten and nine Rbohs have been identified from Arabidopsis and rice, referred to as AtRbohA–AtRbohJ and OsRbohA–OsRbohI, respectively (Torres et al., 1998). Previous studies suggest they have distinct spatio-temporal expression, enzyme activity, and functional diversification (Lin et al., 2009; Kaur and Pati, 2018; Kaur et al., 2018; Kaya et al., 2019). *AtrbohB* fulfills important roles in the germination of Arabidopsis seeds through variant splicing transcripts (Muller et al., 2009). *AtRbohD* and *AtRbohF* present a different spatial expression pattern either in shoot or in root during Arabidopsis development (Jiao et al., 2013; Morales et al., 2016). Finally, AtRbohC, -D, -F, -H, and -J are synergistically activated by Ca^{2+}-binding and then phosphorylated to produce ROS (Ogasawara et al., 2008; Han et al., 2019; Kaya et al., 2019).

Previous studies have proven that Rbohs figure predominantly in ROS production during pathogen-associated molecular pattern-triggered immunity (PTI) and effector-triggered immunity (ETI) responses in Arabidopsis (Torres and Dangl, 2005; Nuhse et al., 2007; Kadota et al., 2014; Li et al., 2014b). Perhaps the best example is RbohD (and RbohF in some cases), a key player not only in intracellular oxidative stresses but also in plant responses and resistance to pathogens (Yun et al., 2011; Chaouch et al., 2012; Ma et al., 2016; Angelos and Brandizzi, 2018). AtRbohD showed the highest expression of the 10 *AtRbohs*: it functions via ROS production in abscisic acid-induced stomatal closure, flagellin-induced immune responses (Pogany et al., 2009), resistance to *Botrytis cinereal* (Galletti et al., 2008), as well as phosphorylation mediated by CRK2 in plant defense against the bacterial pathogen *Pseudomonas syringae* pv. tomato DC3000 (Kimura et al., 2020). AtRbohD-dependent H_2O_2 production was positively regulated by H2Bub1 in the responses against Vd-toxins in Arabidopsis (Zhao et al., 2020). DORN-mediated phosphorylation on the N-terminus of RbohD led to ROS accumulation and stomatal closure of Arabidopsis in biotic stresses (Chen et al., 2017). NbrbohB participates in H_2O_2 accumulation, resulting in elicitor-induced stomatal closure, but it does not function in elicitor-induced HR in *Nicotiana benthamiana* (Yoshioka et al., 2003; Kobayashi et al., 2007; Zhang et al., 2009). Ethylene signaling can regulate ROS production and enhance disease resistance, in which both are mediated by OsEIL2 binding to the promoter of *OsrbohA* and *OsrbohB* in rice (Yang et al., 2017).

Cotton (*Gossypium* spp.) is a globally important fiber crop that is the largest source of natural renewable textile fibers. The allotetraploids *Gossypium hirsutum* and *Gossypium barbadense* are cultivated worldwide because of their high yield and superior fiber quality, respectively (Wendel, 1989; Paterson et al., 2012; Wang et al., 2019). Both *G. hirsutum* and *G. barbadense* originated from two diploid progenitor species: *Gossypium arboreum* (A-genome donor) and *Gossypium raimondii* (D-genome donor) ca. 1–2 million years ago (Wendel, 1989). Because of their tetraploidy and repeat sequences, *G. hirsutum* and *G. barbadense* have large genomes; hence, their study from a genetics and genomics perspective is quite difficult (Paterson et al., 2012; Li et al., 2014a, 2015; Zhang et al., 2015). Moreover, the wild diploid cotton species *Gossypium australe* possesses excellent traits for plant resistance to disease, and its genome was fully sequenced and assembled last year (Cai et al., 2019). Nevertheless, with recent advances in sequencing technology and high-quality genome assemblies and annotations, it is much easier now to pursue detailed analyses of plant genomics (Wang et al., 2019).

Verticillium wilt, caused by the soil-born fungus *Verticillium dahliae* Kleb., is a serious widespread disease that can severely affect plant growth and fiber quality in cotton production fields

(Shaban et al., 2018; Wang et al., 2018). The disease symptoms of *V. dahliae* infection consists of necrotic areas on leaves, yellowing of leaves, wilting and defoliation, and discoloration of vascular tissues (Li et al., 2017). Currently, verticillium wilt has yet to be controlled effectively by either genetic means or fungicide applications, so analyzing the resistance mechanisms of cotton against verticillium wilt should help to resolve it. Although some studies have investigated the function of Rbohs in biotic stress defense responses of plants, detailed genome-wide phylogenetic and functional analyses of the *Rboh* genes family in *G. barbadense* are still lacking.

To better understand the dynamics of *Rboh* genes in *Gossypium* plants and to foster future research on this important enzyme family, here we aimed to provide a detailed overview of their number, phylogeny, and function in defense against *V. dahliae*, by investigating five sequenced *Gossypium* species. From these, we identified 87 *Rbohs*. We determined that not only did *GbRboh5/18* in the CladeD respond to *V. dahliae* infection and hormones, but it was also able to enhance plant resistance against verticillium wilt through elevated ROS accumulation in *G. barbadense*.

MATERIALS AND METHODS

Sequence Retrieval and Identification of Cotton *Rboh* Genes

The motif FAD_binding_8 (PF08022) and EF-hand 2 (PF09068) were each downloaded from the Pfam database[1]. The whole genome sequence from the ancestral diploids *G. arboreum* (A_2) (Li et al., 2014a; Du et al., 2018) and *G. raimondii* (D_5) (Paterson et al., 2012), as well as those of their descendant tetraploids, *G. barbadense* (AD_2) (Wang et al., 2019), and *G. hirsutum* (AD_1) (Li et al., 2015; Zhang et al., 2015; Wang et al., 2019) were downloaded from the CottonGen database[2], while that of *G. australe* (G) (Cai et al., 2019) was downloaded from NCBI[3] (accession number: SMMG00000000), to carry out a genome-wide search for Rboh genes in the *Gossypium* genus. Splice variants were excluded and only the first variant was retained for further analysis. The molecular weights (kDa) and isoelectric points (*p*I) of Rboh proteins were determined using ProtParam[4], and the subcellular localization of each protein was predicted by SherLoc2[5].

Phylogenetic, Gene Structure and Conserved Domain Analysis

A total of 10 Arabidopsis *Rboh* gene sequences were downloaded from TAIR 10[6]. Multiple sequence alignments of all identified Rbohs from cotton and Arabidopsis were performed in ClustalX 2.0. The phylogenetic tree of deduced amino acid sequences was constructed by applying the neighbor-joining (NJ) method, with default parameters and 1000 bootstrap replicates, in MEGA7.0[7]. To analyze the exon-intron distributions of the *Rboh* gene, the gene structure display server (GSDS)[8] was utilized. Conserved motifs were then predicted using the MEME[9] tool, under these parameters: number of repetitions, "any"; maximum number of motifs, "20"; and 60≥ "widths" ≥5, and between 6 and 300 residues. Only those motifs associated with an *E*-value $< e^{-5}$ were retained.

Chromosomal Localization

The positional information of a given cotton *Rboh* gene was obtained from parsed general feature format (GFF) files, downloaded from the CottonGen website. The *Rbohs* in the five *Gossypium* species genomes were all mapped onto the chromosome. A shared identity of genes between different *Gossypium* spp. above 91%, or an identity of genes between the A and D subgenomes on tetraploid *Gossypium* spp. higher than 98% were connected by different-colored lines and depicted with the Circos genome visualization tool[10].

Plant Materials, Growth Condition, and Treatments

Unless otherwise indicated, seeds of *G. barbadense* cultivar ("Xinhai15") were inoculated in ddH_2O for 24 h and then germinated for another 24 h under 25°C. After they had absorbed water, two seeds were sown per pot containing vermiculite. The pots were kept on a tray covered with a plastic dome, in a growth chamber at 25°C, with 120 $\mu E\ m^{-2}\ s^{-1}$ light illumination under a 12-h-light/12-h-dark photoperiod. The domes were removed when two cotyledons emerged on all seedlings. When the first two true leaves fully expanded and the third true leaf emerged, the cotton plants could be treated with *V. dahliae* or hormones. For the former, the *V. dahliae* suspension was injected into the vermiculite nearby root; for the latter, plants were sprayed with jasmonic acid (JA; 100 μM concentration) or with abscisic acid (ABA; 50 μM concentration).

RNA Extraction, Reverse Transcription and Quantitative Real-Time PCR Analyses

Total RNA was extracted from cotton leaves using the RNAprep Pure Plant Plus Kit (DP441, Tiangen). Reverse transcription was performed with the HiScript II Q Select RT SuperMix (R233, Vazyme Biotech, Nanjing, China) following the manufacturer's instructions. The quantitative PCR reactions employed the ChamQ SYBR Color qPCR Master Mix (Q411-02, Vazyme Biotech) and were run on the LightCycler480II Real-time PCR system (Roche) (every Q-PCR reaction was carried in triplicate for each sample). The cotton *UBQ7* gene served as an internal reference gene for the normalization of the data. The sequences

[1]http://pfam.xfam.org/
[2]https://www.cottongen.org/
[3]https://www.ncbi.nlm.nih.gov/
[4]http://web.expasy.org/protparam
[5]https://abi-services.informatik.uni-tuebingen.de/sherloc2/webloc.cgi
[6]http://www.arabidopsis.org
[7]https://www.megasoftware.net
[8]http://gsds.cbi.pku.edu.cn/
[9]http://meme-suite.org/tools/meme
[10]http://www.circos.ca/

of all primer used can be found in **Supplementary Table S2**. For a given sample, the gene's Ct values were standardized and its relative changes analyzed using the $2^{-\Delta\Delta CT}$ method.

Vector Construction

The *TRV-VIGS* vectors were constructed as described previously (Liu et al., 2002). The CDS fragment of *GbRboh5/18* (Gbar_A05G025370/Gbar_D05G026620) and the fragment of 297 bp for virus-induced gene silencing (VIGS) with *Bam*HI/*Kpn*I at the 5′ and 3′ ends, were amplified from *G. barbadense* L. "Xinhai15" cDNA by PCR with gene-specific primers (listed in **Supplementary Table S2**) and cloned into the *pMD18*-T vector. The *GbRboh5/18-VIGS* fragment was inserted into the *pTRV2* vector by *Bam*HI/*Kpn*I. All plasmids (including empty vector *pTRV1*, *pTRV2*, and recombinant vectors *pTRV2- GbRboh5/18*, *pTRV2-CLA* as the positive control) were respectively electroporated into the *Agrobacterium tumefaciens* strain GV3101, which was used to inject cotton cotyledons for VIGS.

The full length CDS fragment of *GbRboh5/18* was amplified from *G. barbadense* L. "Xinhai15" cDNA by PCR with specific primers (listed in **Supplementary Table S2**). The overexpression vector was constructed using the Gateway Technology (Invitrogen). The PCR product was first conducted in BP reaction with *pDONOR221* vector. Then, the positive plasmid *pDONOR221-GbRboh5/18* underwent an LR reaction with the *pK7WG2.0* vector to generate the *p35S::GbRboh5/18* construct. The positive plasmid *pK7WG2.0-GbRboh5/18* was electroporated into *A. tumefaciens* strain LBA4404, which was used to transform cotton plants for transient overexpression.

Virus-Induced Gene Silencing of *GbRboh5/18* in Cotton

Virus-induced gene silencing assays were performed as described previously (Li et al., 2016) with minor modifications. The cotton plants with two fully expanded cotyledons were selected for infiltration for the VIGS assay. The *A. tumefaciens* strain GV3101 cells carrying the corresponding plasmid were inoculated into an LB liquid medium (containing 50 μg mL^{-1} kanamycin, 50 μg mL^{-1} rifampicin, and 25 μg mL^{-1} gentamicin). Grown overnight at 28°C under continuous agitation (200 rpm), *A. tumefaciens* cells were collected until OD_{600} = 0.8–1.2 by centrifugation and resuspended in the infiltration buffer (10 mM 2-(N-morpholino) ethanesulfonic acid [MES], 10 mM $MgCl_2$, 200 μM AS). The density of *A. tumefaciens* cells suspension was set to OD_{600} = 0.6–0.8. Next, the *A. tumefaciens* cells containing *pTRV1*, *pTRV2,* or its derivatives, were mixed in a 1:1 ratio and these *A. tumefaciens* suspensions were incubated at room temperature for 3 h. Finally, the *A. tumefaciens* suspension was infiltrated fully on the underside of cotton cotyledons.

V. dahliae Culturing and Infection

Verticillium dahliae culturing and infection was performed as described (Zhao et al., 2017) with minor modifications. *V. dahliae* strain Vd991 was streak cultivated onto fresh PDA plates from glycerol stocks and incubated for 1 week at 25°C. The mycelia of one plate (90 mm) was transferred by a sterilized blade into a 200 mL Czapek-Dox medium and cultivated for 1 week with shaking at 200 rpm at 25°C in the dark to collect conidia. After filtrating the fungal cultures with 4 layers of sterilized gauze to remove mycelia, the conidial suspension was collected. The concentration of conidia in suspension was measured using a hemocytometer and was adjusted to approximately 1×10^7 conidia mL^{-1} for infection.

After two weeks of *A. tumefaciens* infiltration, cotton seedlings were inoculated with root-dipping into *V. dahliae* suspension, and then replanted into pots. After growing for another two weeks, the third true leaves of each plant were removed for staining experiments. Pretreated leaves were treated in 10 μM diphenyleneiodonium chloride (DPI, NADPH oxidase inhibitor) for 3 h before ROS staining or trypan blue staining.

Evaluation of Resistance to Verticillium Wilt

Verticillium wilt disease severity was performed as described (Meng et al., 2018) with minor modifications. The disease index of cotton plants was assessed 21 days and 25 days post-inoculation (dpi) with Vd991. The grade of disease severity was recorded and the disease index (DI) was calculated. The verticillium wilt disease severity level was recorded as five grades: Grade 0 (normal); Grade 1 (disease leaves less than 25%); Grade 2 (disease leaves between 25 and 50%); Grade 3 (disease leaves between 50 and 75%); and Grade 4 (disease leaves over 75%, almost all the dead plants). Disease index was scored with at least 30 plants per treatment and each treatment was repeated at least three times. Data were analyzed by Student's *t*-test: *, $p < 0.05$; **, $p < 0.01$.

The lesion area of the transverse section of the cotton stem at 25 dpi with Vd991 was photographed and the browning degree was calculated using imageJ as described (Long et al., 2020). Fungal recovery was performed as described (Zhao et al., 2017).

Transient Overexpression of *GbRboh5/18* in *G. barbadense*

Transient overexpression was performed as described (Li et al., 2018) with minor modifications. Overnight culturing of *A. tumefaciens* LBA4404 cells containing the *pK7WG2.0-GbRboh5/18* plasmid at 28°C until the OD_{600} = 1.2–1.5 *A. tumefaciens* cells were collected and suspended in the infection solution [1/2 MS liquid medium,165 μM acetosyringone (AS), 0.01% Tween-20], with this mixture adjusted to OD_{600} = 0.8–1.0 to infect the cotton seedlings. After seeds of *G. barbadense* L. "Xinhai15" germinated, they were cultured in Hoagland's solution until their cotyledons had fully expanded and the first true leaf emerged. These seedlings were washed with sterile water and 70% alcohol, and then sterilized in 0.01% $HgCl_2$ for 2 min. Seedlings were washed for 2 min with sterile water, 5 times, then placed in a sterile jar to which the infection buffer was added, without Tween-20, for 2 h. After vacuuming for 15 min, the buffer

was discarded. The seedlings were incubated in the *A. tumefaciens* suspensions at 28°C for 5 h (at 150 rpm) after seedlings were washed several times with sterile water to reduce the risk of subsequent bacterial contamination. Finally, the transformed seedlings were cultured in 1/2 MS solid medium and grown for 5 days; these seedlings were used to detect the expression levels or for H_2DCFDA staining.

ROS Staining, H_2DCFDA Staining and Trypan Blue Staining

The H_2O_2 content and cell death were detected by DAB staining and trypan blue staining, respectively. DAB staining and trypan blue staining were performed as described (Kumar et al., 2014) with minor modifications. Treated or control cotton true leaves were inoculated in 1 mg mL^{-1} of DAB staining solution in 10 mM Tris-acetic acid (pH 5.0) or in 0.4% (W/V) trypan blue staining buffer, and then vacuumed for 10 min. Tissue staining proceeded overnight (DAB staining) or for 3 h (trypan blue staining) at room temperature. Then the staining solution was drained off and absolute ethanol was added before tissue immersion in a boiling water-bath for 10 min to remove chlorophyll. The H_2O_2 was visualized as a reddish-brown color, while cell death appeared in blue. The H_2DCFDA staining was used to detect ROS accumulation in cells. Overexpression cotton true leaves (vector as the negative control) were inoculated in 10 μM H_2DCFDA solution in PBS buffer (pH 7.5), for 30 min at 37°C, after which the staining solution was drained off and the leaves washed four times with PBS buffer (pH 7.5). Fluorescence was visualized under a Zeiss LSM710 microscope with an excitation light of 488nm.

RESULTS

Identification of *Rbohs* in Five *Gossypium* Species

To identify all the *Rboh* genes in the five sequenced *Gossypium* species, we performed a BLASTP search in the diploids *G. arboreum* (A_2), *G. raimondii* (D_5), and *G. australe* (G), as well as the tetraploids *G. hirsutum* (AD_1) and *G. barbadense* (AD_2). The protein database was queried by PF08022 and PF09068, which are conserved in the AtRbohs. Overall, we found 87 *Rbohs* on the cotton genomes: 13 *Rbohs* in A_2, 13 *Rbohs* in D_5, 10 *Rbohs* in G, 24 *Rbohs* in AD_1, and 27 *Rbohs* in AD_2 (**Table 1**). The number of *Rboh* genes in the A_t sub-genome of tetraploid *Gossypium* species was the same as that in the diploid *G. arboretum* (A_2), while the *Rboh* genes in the D_t sub-genome of tetraploid *Gossypium* species was less than that in the diploid *G. raimondii* (D5), leaving the remainder 2 *GbRbohs* anchored to unmapped scaffolds. The genome harboring the fewest *Rboh* genes was that of *G. australe* (G).

According to the chromosomal locations of the corresponding genes, we renamed all *Rbohs* consecutively, as *GaRboh1–GaRboh13*, *GrRboh1–GrRboh13*, *GauRboh1–GauRboh10*, *GhRboh1–GhRboh24*, and *GbRboh1–GbRboh27*. For the 87 *Rboh* genes found in cotton, their encoded proteins consisted of 743–1015 amino acids. The predicted molecular weight of the Rboh proteins in the five *Gossypium* species ranged from 84.2 to 115.47 kDa, with a theoretical *p*I between 8.65 and 9.35. Their predicted subcellular localization showed that 71 of the 87 Rbohs were located in the cytoplasm, except for GaRboh9, GrRboh5, GhRboh2/15, and GbRboh2/15 that were restricted to mitochondria (**Supplementary Table S1**).

Phylogenetic Analysis of Rbohs in Five *Gossypium* Species

To understand their evolutionary relationships, a phylogenetic analysis of the putative 87 Rbohs from the five *Gossypium* species and 10 AtRbohs was carried out. Based on the homologous relationship in Arabidopsis, the 87 Rbohs in cotton were divided into seven clades (named CladeA, -B, -D, -E, -F, -H, and -I according to their homologs in Arabidopsis) (**Figures 1**, **2A**). There were 20 members in CladeA, -D, and -F, 7 members in CladeB, -E, and -H, 6 members in CladeI, respectively. The cotton Rboh members within each clade displayed a high similarity to their orthologs in Arabidopsis. These results suggested the Rbohs in the five *Gossypium* species are highly conserved.

The conserved motif analysis revealed that all these Rboh proteins contained at least 10 conserved motifs: two EF-hands in the N-terminus, six TM domains, and a NADPH-binding motif and a FAD-binding motif in the C-terminal region (**Figure 2B**). The number and arrangement of conserved domains within the same clade were similar. Comparing their gene structures, the exon-intron organizations revealed that *Rboh* genes in both CladeD and CladeH possessed densely arranged exons, with shorter introns, whereas longer introns characterized the rest of the cotton *Rboh* genes (**Figure 2C**). These results suggested a similar gene structure and conserved motif arrangement within the same clade.

Localization of *Rboh* Genes on Five *Gossypium* Chromosomes

To better understand their evolutionary relationships, we further analyzed the orthologous and paralogous *Rboh* genes in the five *Gossypium* species. Many gene blocks were found among the chromosomes of two sub-genomes of *G. hirsutum* and *G. barbadense*, marked by similarities to those in the sequenced diploid cotton genomes (A_2, D_5, and G). Compared to *G. raimondii*, *G. arboreum*, and *G. australe*, there, respectively were 22, 26, and 18 pairs of putative outparalogous genes in *G. barbadense* (**Figure 3A**). In *G. hirsutum,* there were 18, 20, and 14 pairs of putative orthologous *Rbohs* genes when compared to *G. raimondii*, *G. arboretum*, and *G. australe*, respectively (**Figure 3B**).

Further analysis of the paralogous Rboh genes, between the A and D subgenomes in the two tetraploid cotton species, showed that 24 of 27 could be one-to-one paired in *G. barbadense* between the A and D sub-genomes, in addition to *GbRboh12*, *GbRboh26*, and *GbRboh27* (**Figure 3C**). There were 11 syntenic gene blocks in *G. hirsutum*, with the exception of *GhRboh11* and *GhRboh12* (**Figure 3D**); these paired genes may exist as alleles on chromosomes between the A and D sub-genomes. These results

TABLE 1 | The number of *Rboh* genes in the five *Gossypium* species.

	G. arboreum	*G. raimondii*	*G. australe*	*G. hirsutum* (AD_1)			*G. barbadense* (AD_2)			Total
	(A_2)	(D_5)	(G)	At	Dt	Sca	At	Dt	Sca	
Rbohs	13	13	10	13	11	0	13	12	2	87

"Sca" indicates assembled scaffold. The "At," "Dt" denote, respectively, the A sub-genomes and D sub-genomes in the tetraploid; "t" indicates tetraploid.

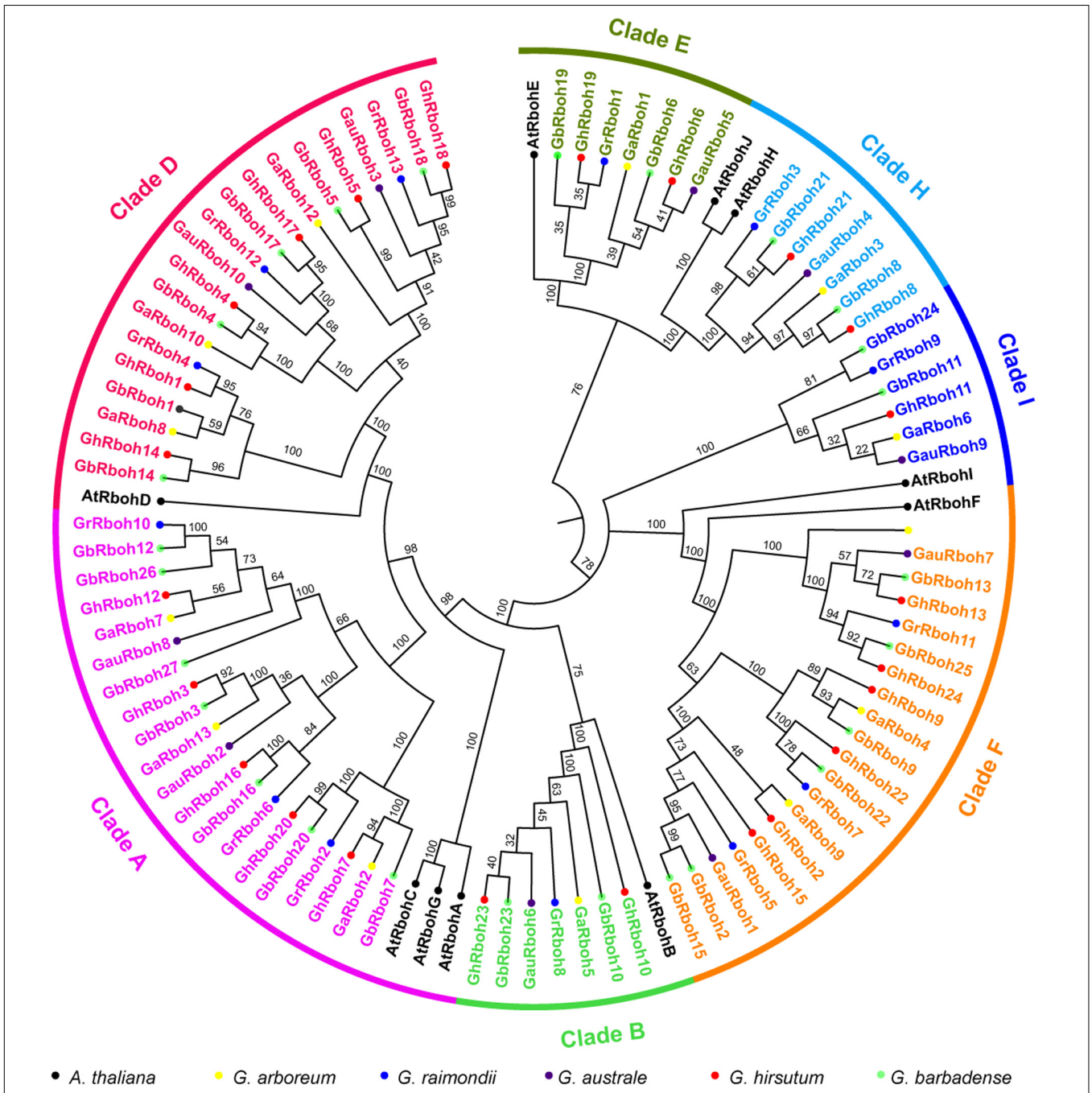


FIGURE 1 | Phylogenetic analysis of Rbohs in cotton and Arabidopsis. A phylogenetic tree of Rboh proteins from *G. arboreum*, *G. raimondii*, *G. australe*, *G. hirsutum*, *G. barbadense,* and *A. thaliana*. The full-length amino acid sequences of the Rboh proteins were aligned using ClustalX in MEGA7.0. The unrooted tree was generated by the neighbor-joining (NJ) method (n = 1000 bootstraps). Clades are distinguished by different colors. Cotton Rbohs are colored clade specifically and the Arabidopsis Rbohs are in black. Different-colored solid circles indicate genes in different *Gossypium* species or Arabidopsis.

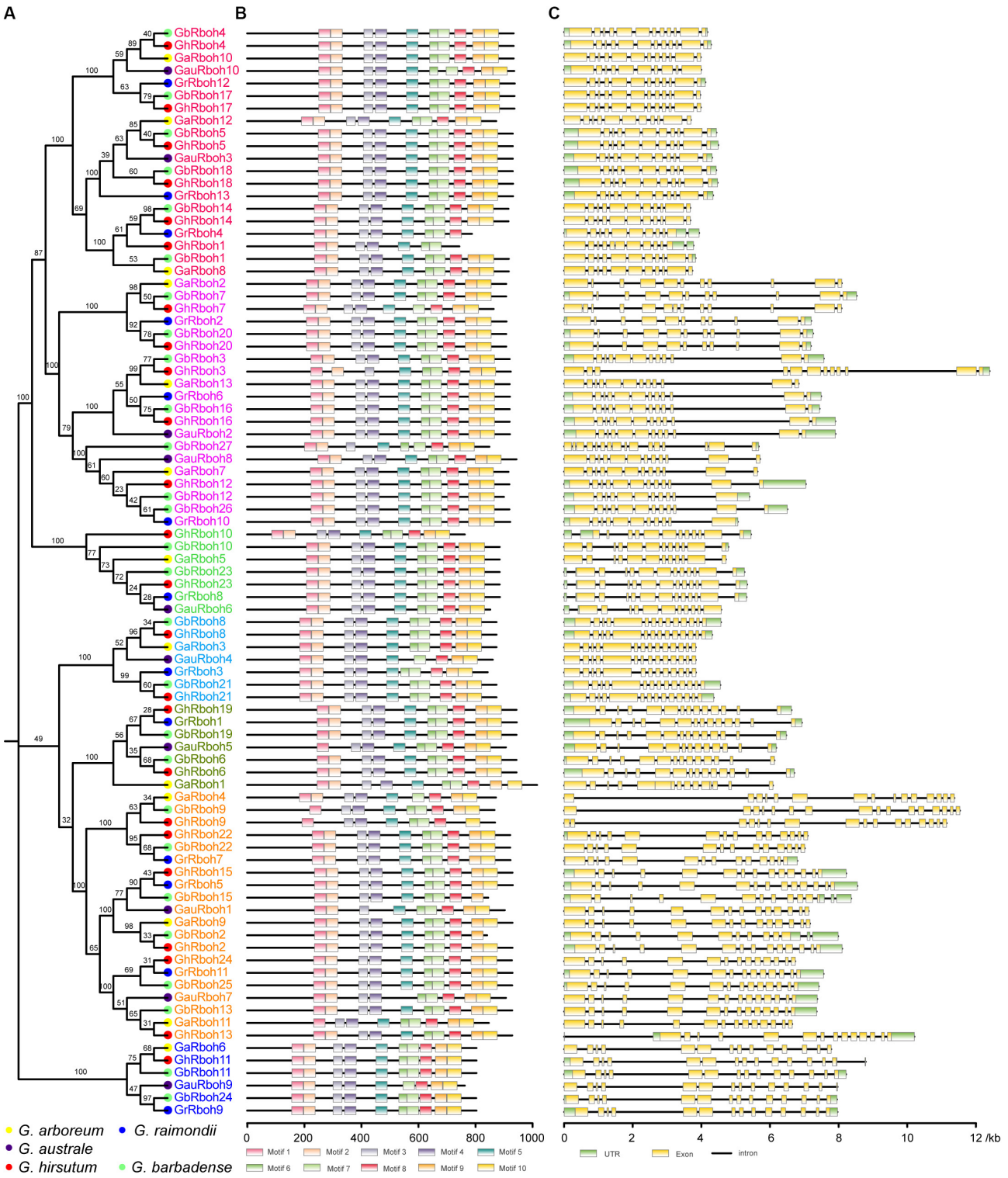


FIGURE 2 | Characterization of the 87 Rbohs in five *Gossypium* species. The characteristics include the homologous relationship, conserved domain location, and exon-intron structure. **(A)** Phylogenetic relationships among the Rbohs in five *Gossypium* species. Rbohs in same clade are indicated by the same color. Different-colored solid circles indicate genes in different *Gossypium* species. **(B)** Conserved domain architecture of cotton Rboh proteins. Motifs were identified through a MEME analysis (http://meme-suite.org/) and protein length was estimated via the scale at the bottom. **(C)** Exon-intron organization analyses of cotton *Rboh* genes. The gene length was estimated by the scale at the bottom through gene structure analyses (GSDS: http://gsds.cbi.pku.edu.cn/). Green boxes indicate 5′- or 3′-untranslated regions. Exons and introns are distinguished in yellow boxes and in black lines, respectively.

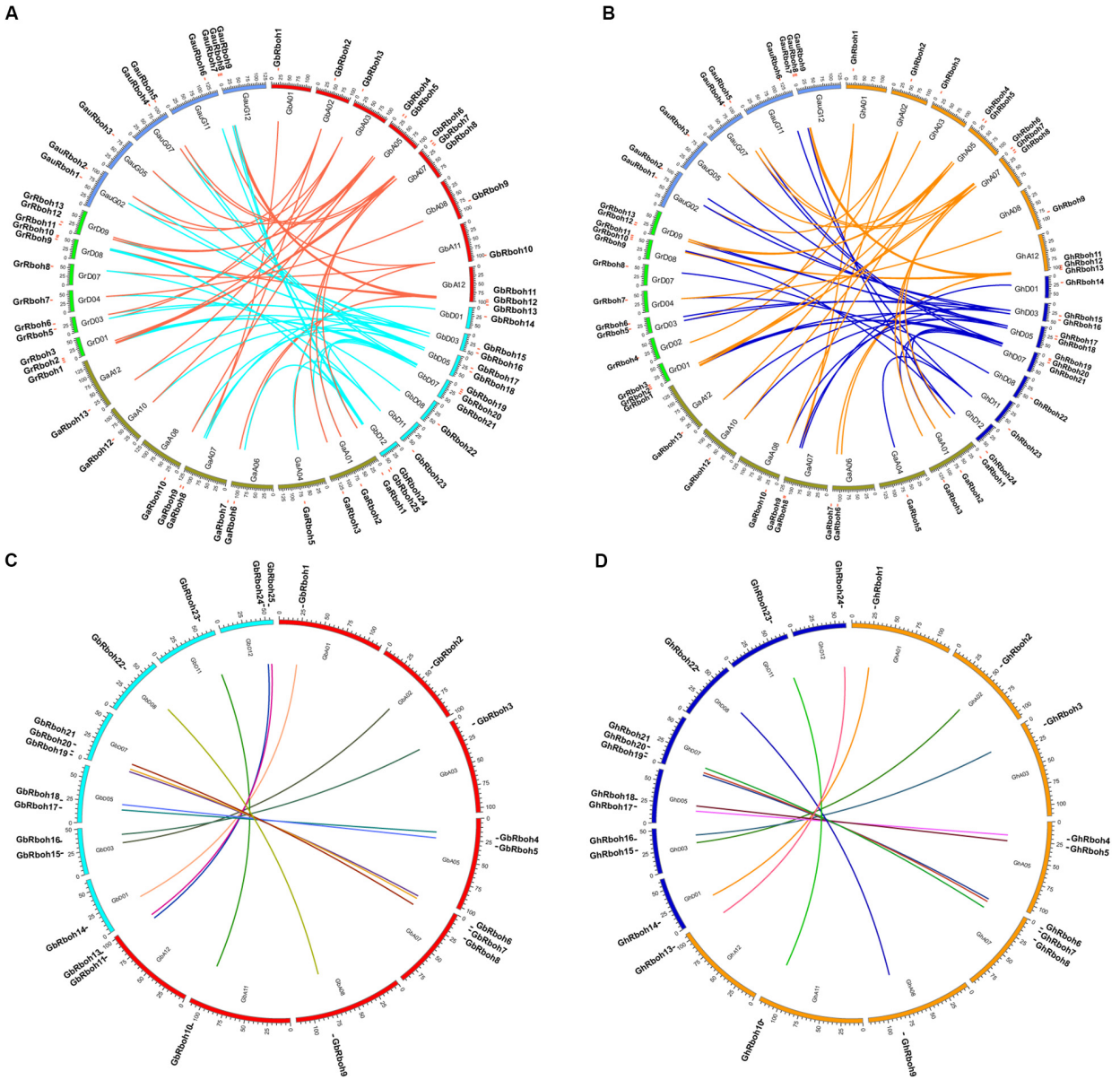


FIGURE 3 | Localization of cotton *Rbohs* on chromosomes. The genomic (At and Dt) analysis of Rboh genes in panels **(A)** *G. barbadense* and **(B)** *G. hirsutum* vis-à-vis the diploid *Gossypium* spp. Different-colored links indicate translocation from the A or D sub-genome of an allotetraploid cotton to the diploid cotton. Lined gene coupling represents a shared identity between them that exceeded 91%. The inter-genomic (At and Dt) analyses of *Rboh* genes in panels **(C)** *G. barbadense* and **(D)** *G. hirsutum* between the A and D sub-genomes. Different-colored lines indicate gene couples whose identity is over 98%, depicted using the Circos genome visualization tool (http://www.circos.ca/).

implied that Rbohs are generally highly conserved between A and D sub-genomes during the evolution of tetraploid cotton plants.

CladeD (mainly *GbRboh5/18*) Response to Pathogen and Hormones in *G. barbadense*

To detect which Rbohs may contribute to plant resistance against verticillium wilt, or which can respond to hormones in *G. barbadense*, we designed specific primers for the conserved regions to detect the post-treatment level of transcripts for each clade's genes. Within 24 h of treatment with *V. dahliae*, JA, or ABA, the transcripts of CladeA, -B, -D, and -F showed dramatic or slight changes, whereas the transcripts of *GbRbohs* in the other remaining three clades went undetected (**Figure 4**). Among the four responsive clades, CladeD underwent the starkest fold-change during the treatment with *V. dahliae* (6-fold at 24 h), JA (3.2-fold at 4 h), and ABA (9.6-fold at 4 h).

In *G. barbadense*, CladeD included three gene couples: *GbRboh1/14*, *GbRboh4/17*, and *GbRboh5/18*. We further detected

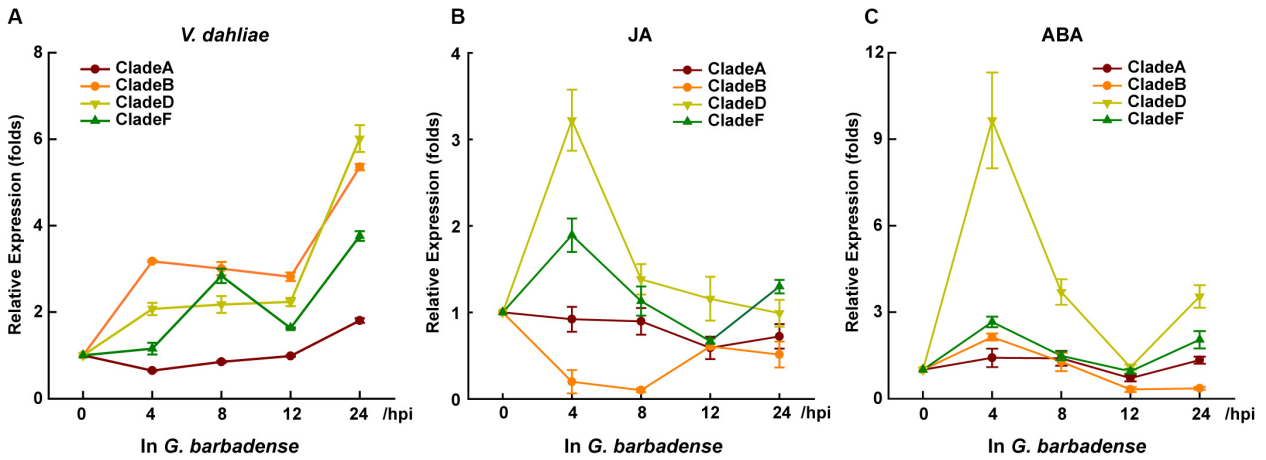


FIGURE 4 | *GbRbohs* in the CladeD respond to *V. dahliae* infection and hormones in G. *barbadense*. Quantitative analyses of each clade *GbRbohs*' expression levels in "Xinhai15" cotton plant leaves treated with **(A)** *V. dahliae*, **(B)** jasmonic acid, and **(C)** abscisic acid. Total RNAs were extracted from leaves of 3-week-old seedlings at 0, 4, 8, 12, and 24 h post-inoculation. The values are the mean ± SD for three technical replicates. Relative expression levels of all clade genes were determined after normalizing to the expression level at 0 h, which was set to 1.0. hpi, hours post-inoculation. The experiments were repeated at least three times, with similar results.

their respective transcript levels by Q-PCR to identify which members contributed chiefly. The transcript of *GbRboh5/18* was clearly induced by *V. dahliae,* or the JA, or ABA treatments (**Figure 5**), whereas the transcripts of *GbRboh1/14* and *GbRboh4/17* were not detected. Furthermore, the trend in the variation of *GbRboh5/18*'s transcript level was consistent with CladeD; hence we deduced that *GbRboh5/18* is the cardinal member of CladeD in *G. barbadense*.

Knock-Down of *GbRboh5/18* Attenuates Resistance to *V. dahliae* in *G. barbadense*

To verify the function of GbRbohD in cotton plant resistance to *V. dahliae*, we utilized VIGS, a powerful technique, to achieve knock-down in the transcript activity of *GbRboh5/18*, the gene mainly responsible in CladeD. When "Xinhai15" (cultivar of *G. barbadense*) seedlings had grown for 10 days, we infiltrated cotyledons with TRV: 00 (as negative control) or TRV: GbRboh5/18. After 2 weeks of treatment with *V. dahliae*, the *GbRboh5/18-VIGS* plants displayed more severe disease symptoms than the control plants, such as wilted leaves and dark vascular bundles (**Figure 6A**). The successful knock-down of the *GbRboh5/18* transcript was confirmed by the Q-PCR analysis (**Figure 6B**), and the disease index was consistent with the plant phenotype (**Figure 6C**). Further, a lesion area analysis and fungal recovery assay closely resembled the phenotype and disease index (**Figures 6D,E**). These results suggested that the knock-down of *GbRboh5/18* transcription is able to attenuate plant resistance to *V. dahliae* infection.

GbRboh5/18 Promotes ROS Accumulation in *G. barbadense*

To clarify the function of GbRboh5/18, we detected cell death in the *GbRboh5/18-VIGS* plants and control plants, via trypan blue staining. There was no significant difference between *GbRboh5/18-VIGS* and control leaves, but more of the blue stain appeared in *GbRboh5/18-VIGS* leaves than control leaves after infection with *V. dahliae*. Pretreatment with DPI, an inhibitor of the NADPH oxidase, before *V. dahliae* infection, resulted in almost no staining in all leaves (**Figure 7A**). Further, the DAB staining showed less ROS accumulation in *GbRboh5/18-VIGS* leaves than in control leaves after *V. dahliae* infection, yet this did not occur in both groups leaves when they were pretreated with DPI (**Figure 7B**). These results indicated that the knock-down of *GbRboh5/18* caused more cell death mediated by less ROS accumulation when plants were infected with the *V. dahliae* pathogen.

Nonetheless, we constructed the *GbRboh5/18*-OE vector to verify the functional mechanism of *GbRboh5/18* (**Figure 7C**). We utilized a transient overexpression strategy to detect ROS accumulation in cotton seedlings. Overexpression of *GbRboh5/18* was confirmed by the Q-PCR analysis (**Figure 7D**). The H_2DCF-DA fluorescent assay was used to observe ROS accumulation in *GbRboh5/18-OE* leaves, which revealed stronger fluorescent signals in *GbRboh5/18*-OE than in the control (**Figure 7E**). These results showed that GbRboh5/18 can increase the ROS accumulation.

DISCUSSION

Rbohs, a group of plant-specific NADPH oxidases, have been phylogenetically and functionally characterized for a variety of plant species (Torres et al., 1998; Xia et al., 2009; Kaur and Pati, 2018), but not yet clearly for cotton (*Gossypium* spp.). Five *Gossypium* genomes were sequenced and assembled over the past few years (Li et al., 2014a, 2015; Zhang et al., 2015; Du et al., 2018; Ma et al., 2018, 2019; Cai et al., 2019; Hu et al., 2019; Wang et al., 2019). With the development of

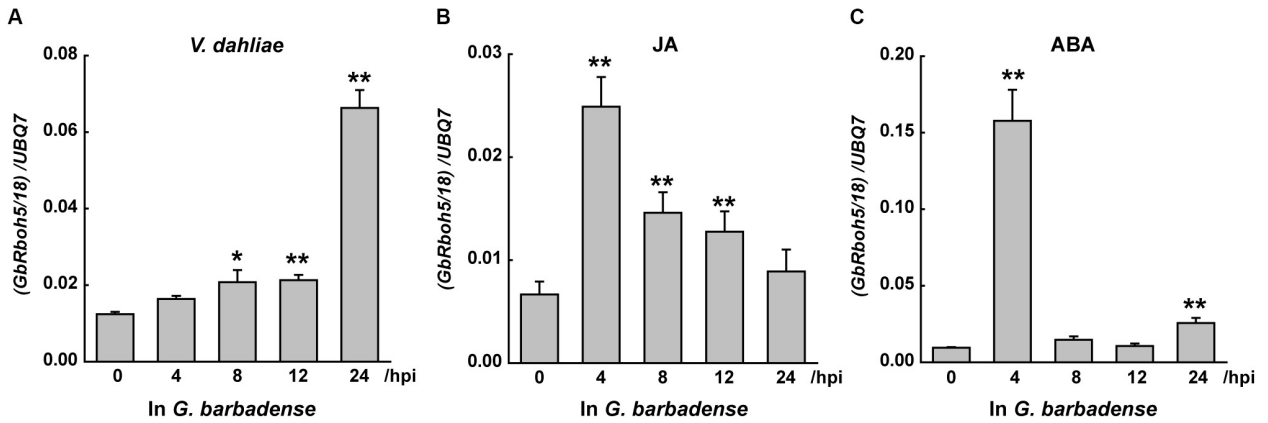


FIGURE 5 | *GbRboh5/18* are the main functional genes of CladeD in *G. barbadense*. Quantitative analyses of *GbRboh5/18* expression levels in "Xinhai15" cotton plant leaves inoculated with **(A)** *V. dahliae*, **(B)** jasmonic acid, and **(C)** abscisic acid. Total RNAs were extracted from leaves of 3-week-old seedlings at 0, 4, 8, 12, and 24 h post-inoculation. The experiments were repeated three times, with similar results. The values are the means ± SD for three technical replicates. Transcript levels of each gene couple were first normalized to *UBQ7*. Asterisks indicate significant differences compared with 0 h under same treatment (*$p < 0.05$; **$p < 0.01$, Student's *t*-test).

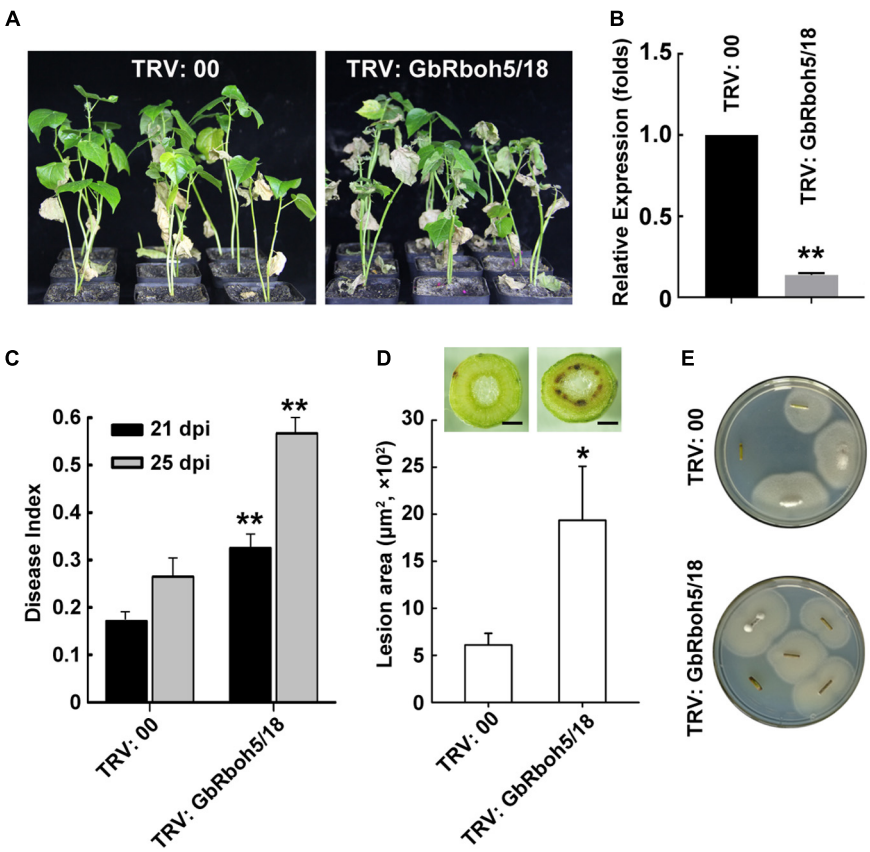


FIGURE 6 | Knock-down of *GbRboh5/18* attenuates plant resistance to *V. dahliae*. **(A)** Disease phenotype of cotton plants ("Xinhai15") at 25 days post-infection with *V. dahliae* strain Vd991. Seedlings were inoculated with *V. dahliae* two weeks after VIGS and photographed 25 days later. **(B)** Expression of *GbRboh5/18* in *GbRboh5/18-VIGS* and control plants. *UBQ7* was the internal reference gene (**$p < 0.01$; Student's *t*-tests). **(C)** Disease index of *GbRboh5/18-VIGS* and control plants at 21 and 25 dpi with *V. dahliae*. Asterisks indicate significant differences compared with TRV: 00 at same time (*$p < 0.05$; **$p < 0.01$, Student's *t*-test). **(D)** Lesion area of *GbRboh5/18-VIGS* and control plants' stems at 25 dpi with *V. dahliae*. Transverse section of stems were photographed with an OLYMPUS CX31 microscope. The values are the mean ± SD, $n \geq 30$. Bar = 0.5 mm. **(E)** Fungal recovery assay. Sterile stems approximately 1 cm from same position of plants was plated on PDA plates at 25°C to allow *V. dahliae* recovery; photographs were taken 3 days post-inoculation. All experiments were repeated at least three times, with similar results.

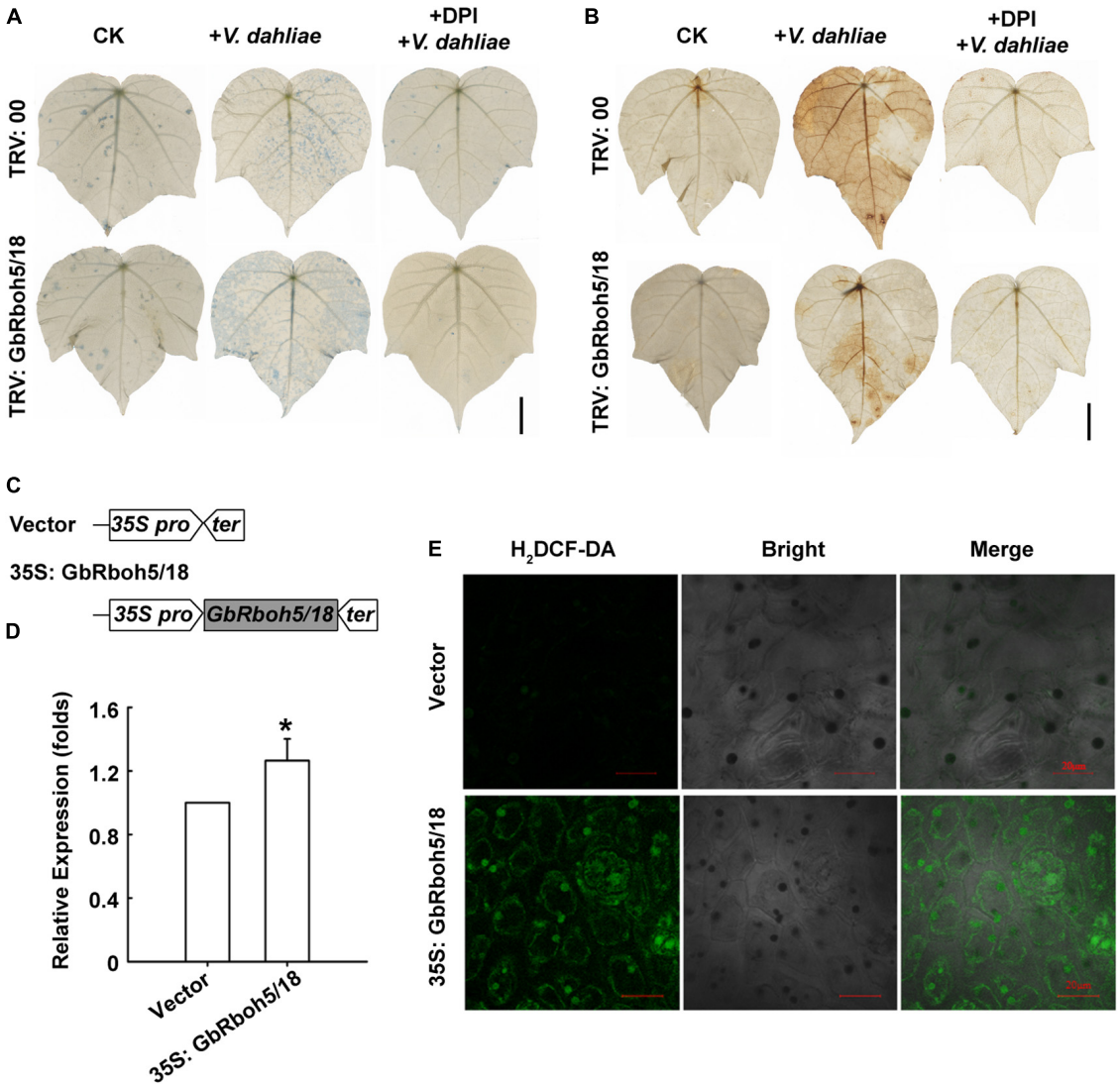


FIGURE 7 | GbRboh5/18 promotes ROS accumulation in *G. barbadense*. **(A)** Trypan blue staining of cell death and **(B)** DAB staining of ROS in *GbRboh5/18-VIGS* and control leaves. Leaves were stained at 48 h post-inoculation with or without *V. dahliae*. Leaves were pretreated with or without 10 μM DPI before trypan blue staining or DAB staining. Bar = 1 cm. **(C)** Schematic diagram showing the structure of overexpression vector. Gray-filled squares indicate ORF of *GbRboh5/18*; the left pentagon and right pentagon are the 35S promoter and 35S terminator, respectively. **(D)** Expression of *GbRboh5/18* in *GbRboh5/18-OE* and vector plant leaves. *UBQ7* was the internal reference gene. (*$p < 0.05$; Student's *t*-tests). **(E)** H_2DCF-DA fluorescence probe of ROS in *GbRboh5/18-OE* and vector leaves cells. Bar = 20 μm. After 5 days of growth, the first true leaves were used to Q-PCR or H_2DCF-DA staining. All experiments were repeated at least three times with similar results.

sequencing and assembly technology, information of genomes is becoming more and more complete. Utilizing bioinformatics, Q-PCR, VIGS, and plant-pathogen biology, in this study we systematically reported on the *Rboh* gene family in five sequenced *Gossypium* species. This integrated approach included building a phylogenetic tree of the *Rboh* genes in cotton and Arabidopsis; analyses of gene structure and conserved domains, and transcript levels' response to *V. dahliae*; and verifying the function of *GbRboh5/18* in conferring resistance to *V. dahliae*. Collectively, this provides a basic understanding of cotton *Rboh* genes and highlights the key role of *GbRboh5/18* in responses to *V. dahliae*.

We systematically identified 87 *Rboh* genes in these representative five sequenced *Gossypium* species: *G. arboretum*, *G. raimondii*, *G. australe*, *G. hirsutum*, and *G. barbadense*. Both *G. arboretum* and *G. raimondii* harbored 13 Rboh genes, a result that agrees with the findings of Zhang et al. (2020). The main difference with that work is that we found 24 and 27 *Rboh* genes in *G. hirsutum* and *G. barbadense*, while they found 26 and 19, respectively (Wang et al., 2020). For *G. hirsutum*, 4 of 26 *GhRbohs* were anchored to unmapped scaffolds in the NAU-NBI v1.1 assembly genome (assembled in 2015), whereas we found that all 24 *GhRbohs* in the HAU v1.0 assembly genome (assembled in 2018) mapped onto the A_t or D_t sub-genome chromosomes.

For *G. barbadense*, we found more *GbRbohs* in the HAU v2.0 assembly genome (assembled in 2018) than in the NAU assembly genome (**Table 1**, **Supplementary Table S1**). This disparity may be due to recent advances in assembly technology that have improved the quantity and quality of genomic information, resulted in a more accurate genome. In addition, this study uncovered 10 *GauRbohs* in the *G. austral* genome, making it now possible to further explore the function of the *Rboh* gene family in this species.

The number of *Rboh* genes in *G. hirsutum* was less than its sum in the two diploids, suggesting that gene loss events happened in the *Rboh* gene family during the polyploidization of *G. hirsutum*. By contrast, the number of *Rboh* genes in *G. barbadense* was more than the sum of the *Rboh* genes' number in the A_2 and D_5 genomes, suggesting gene expansion occurred in the *Rboh* gene family during the polyploidization of *G. barbadense*. These results are consistent with previous work showing that many gene loss and duplication events have occurred in allotetraploid cotton plants (Li et al., 2015; Zhang et al., 2015; Wang et al., 2019; Yang et al., 2019). According to their phylogenetic relationships with Arabidopsis orthologs, the Rbohs were assigned to seven clades. In addition to CladeA and -I, there were twice as many genes of other clades in the tetraploids as in the diploid ancestors (*G. arboretum* and *G. raimondii*). This suggests that the *Rboh* gene family was conserved during the evolution of cotton.

Rbohs play pivotal roles in ROS production during the PTI and ETI responses in Arabidopsis and rice (Pogany et al., 2009; Chaouch et al., 2012; Dubiella et al., 2013). Verticillium wilt, a "cancer" of cotton production, is a vascular bundle disease caused by the fungus *V. dahliae* (Fradin et al., 2009). Overexpression of polyamine oxidase GhPAO contributes to the resistance of plants against *V. dahliae* by elevating the level of H_2O_2 accumulation (Mo et al., 2015). These previous findings contended that ROS content influences plant resistance to *V. dahliae*. Yet in our study we did not detect the transcripts of *CladeE* (*GbRboh6/19*), *CladeH* (*GbRboh8/21*), or *CladeI* (*GbRboh11/24*) when plants were treated with either the pathogen or hormones. Accordingly, these results lead us to speculate that *CladeE*, *CladeH,* and *CladeI* may participate in certain physiological developmental processes, but not in the cotton plant defense response against this pathogen.

AtRbohD maintains hydrogen peroxide (H_2O_2) homeostasis in Arabidopsis plants under different stresses (Kobayashi et al., 2012; Morales et al., 2016; Yao et al., 2017). The latest findings suggest that RbohD-dependent H_2O_2 production is regulated by the PTP-MPK3/6-WRKY pathway in the response against Vd-toxins in Arabidopsis (Zhao et al., 2020). Since *GbRboh5/18* mainly contributes to *CladeD*, we also determined the function of *GbRboh5/18*, via VIGS and transient overexpression experiments. Knock-down of *GbRboh5/18* resulted in a lowered ROS content and reduced resistance to *V. dahliae*, whereas overexpressing *GbRboh5/18* resulted in higher ROS accumulations (**Figures 6**, **7**). So we may conclude that *GbRboh5/18* enhances cotton plant resistance by increasing its ROS accumulation.

In conclusion, we showed that the *Rboh* gene family and *GbRboh5/18*, in particular, contributed to resistance against verticillium wilt in cotton. The molecular functional mechanism of *GbRboh5/18* in the regulation of H_2O_2 production now awaits investigation.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.cottongen.org/ [BGI-CGB v2.0 assembly genome for *G. arboreum* (A2)], [JGI v2.0 assembly data for *G. raimondii* (D5)], [HAU v1.0 assembly genome for *G. hirsutum* (AD1)], [HAU v2.0 assembly genome for *G. barbadense* (AD2)]; https://www.ncbi.nlm.nih.gov/ [assembly CCRI_HENU_Gaus_1.1 for *G. australe* (G)].

AUTHOR CONTRIBUTIONS

YC and BL designed the experiments. BL and SG analyzed the genes' data. QS and RG performed the Q-PCR, VIGS, and the ROS detection. BL and YZ carried out the overexpression of cotton seedlings. JL constructed the phylogenetic tree. YC and YFC wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This study was funded by the National Key Research and Development Program of China (grant numbers 2016YFD0101902 and 2018YFD0100304), the Natural Science Foundation of China (grant numbers U1704104 and 31571724), and the Innovation and Talent Introduction Program on Crop Stress Biology (111 Project).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00788/full#supplementary-material

REFERENCES

Angelos, E., and Brandizzi, F. (2018). NADPH oxidase activity is required for ER stress survival in plants. *Plant J.* 96, 1106–1120. doi: 10.1111/tpj.14091

Baxter, A., Mittler, R., and Suzuki, N. (2014). ROS as key players in plant stress signalling. *J. Exp. Bot.* 65, 1229–1240. doi: 10.1093/jxb/ert375

Cai, Y., Cai, X., Wang, Q., Wang, P., Zhang, Y., Cai, C., et al. (2019). Genome sequencing of the Australian wild diploid species *Gossypium austral* highlights disease resistance and delayed gland morphogenesis. *Plant Biotechnol. J.* 18, 814–828 doi: 10.1111/pbi.13249

Chaouch, S., Queval, G., and Noctor, G. (2012). AtRbohF is a crucial modulator of defence-associated metabolism and a key actor in the interplay between intracellular oxidative stress and pathogenesis responses in Arabidopsis. *Plant J.* 69, 613–627. doi: 10.1111/j.1365-313X.2011.04816.x

Chen, D., Cao, Y., Li, H., Kim, D., Ahsan, N., Thelen, J., et al. (2017). Extracellular ATP elicits DORN1-mediated RBOHD phosphorylation to regulate stomatal aperture. *Nat. Commun.* 8, 2265–2278. doi: 10.1038/s41467-017-023402343

Du, X., Huang, G., He, S., Yang, Z., Sun, G., Ma, X., et al. (2018). Resequencing of 243 diploid cotton accessions based on an updated A genome identifies the genetic basis of key agronomic traits. *Nat. Genet.* 50, 796–802. doi: 10.1038/s41588-018-0116-x

Dubiella, U., Seybold, H., Durian, G., Komander, E., Lassig, R., Witte, C. P., et al. (2013). Calcium-dependent protein kinase/NADPH oxidase activation circuit is required for rapid defense signal propagation. *Proc. Natl. Acad. Sci. U S A* 110, 8744–8749. doi: 10.1073/pnas.1221294110

Fradin, E. F., Zhang, Z., Juarez Ayala, J. C., Castroverde, C. D., Nazar, R. N., Robb, J., et al. (2009). Genetic dissection of Verticillium wilt resistance mediated by tomato Ve1. *Plant Physiol.* 150, 320–332. doi: 10.1104/pp.109.136762

Galletti, R., Denoux, C., Gambetta, S., Dewdney, J., Ausubel, F. M., De Lorenzo, G., et al. (2008). The AtrbohD-mediated oxidative burst elicited by oligogalacturonides in Arabidopsis is dispensable for the activation of defense responses effective against *Botrytis cinerea*. *Plant Physiol.* 148, 1695–1706. doi: 10.1104/pp.108.127845

Han, J. P., Koster, P., Drerup, M. M., Scholz, M., Li, S., Edel, K. H., et al. (2019). Fine-tuning of RBOHF activity is achieved by differential phosphorylation and Ca^{2+} binding. *New. Phytol.* 221, 1935–1949. doi: 10.1111/nph.15543

Hu, Y., Chen, J., Fang, L., Zhang, Z., Ma, W., Niu, Y., et al. (2019). *Gossypium barbadense* and *Gossypium hirsutum* genomes provide insights into the origin and evolution of allotetraploid cotton. *Nat. Genet.* 51, 739–748. doi: 10.1038/s41588-019-0371375

Jiao, Y., Sun, L., Song, Y., Wang, L., Liu, L., Zhang, L., et al. (2013). AtrbohD and AtrbohF positively regulate abscisic acid-inhibited primary root growth by affecting Ca^{2+} signalling and auxin response of roots in Arabidopsis. *J. Exp. Bot.* 64, 4183–4192. doi: 10.1093/jxb/ert228

Kadota, Y., Sklenar, J., Derbyshire, P., Stransfeld, L., Asai, S., Ntoukakis, V., et al. (2014). Direct regulation of the NADPH oxidase RBOHD by the PRR-associated kinase BIK1 during plant immunity. *Mol. Cell.* 54, 43–55. doi: 10.1016/j.molcel.2014.02.021

Kaur, G., Guruprasad, K., Temple, B. R. S., Shirvanyants, D. G., Dokholyan, N. V., and Pati, P. K. (2018). Structural complexity and functional diversity of plant NADPH oxidases. *Amino. Acids* 50, 79–94. doi: 10.1007/s00726-017-24912495

Kaur, G., and Pati, P. K. (2018). In silico insights on diverse interacting partners and phosphorylation sites of respiratory burst oxidase homolog (Rbohs) gene families from Arabidopsis and rice. *BMC Plant Biol.* 18:161. doi: 10.1186/s12870-018-1378-2

Kaya, H., Takeda, S., Kobayashi, M. J., Kimura, S., Iizuka, A., Imai, A., et al. (2019). Comparative analysis of the reactive oxygen species-producing enzymatic activity of Arabidopsis NADPH oxidases. *Plant J.* 98, 291–300. doi: 10.1111/tpj.14212

Keller, T., Damude, H. G., Werner, D., Doerner, P., Dixon, R. A., and Lamb, C. (1998). A Plant Homolog of the Neutrophil NADPH Oxidase *gp91^{phox}* Subunit Gene Encodes a Plasma Membrane Protein with Ca^{2+} Binding Motifs. *Plant Cell* 10, 255–266. doi: 10.1105/tpc.10.2.255

Kimura, S., Hunter, K., Vaahtera, L., Tran, H. C., Citterico, M., Vaattovaara, A., et al. (2020). CRK2 and C-terminal Phosphorylation of NADPH Oxidase RBOHD Regulate Reactive Oxygen Species Production in Arabidopsis. *Plant Cell.* 32, 1063–1080. doi: 10.1105/tpc.19.00525

Kobayashi, M., Kawakita, K., Maeshima, M., Doke, N., and Yoshioka, H. (2006). Subcellular localization of Strboh proteins and NADPH-dependent O_2^--generating activity in potato tuber tissues. *J. Exp. Bot.* 57, 1373–1379. doi: 10.1093/jxb/erj113

Kobayashi, M., Ohura, I., Kawakita, K., Yokota, N., Fujiwara, M., Shimamoto, K., et al. (2007). Calcium-dependent protein kinases regulate the production of reactive oxygen species by potato NADPH oxidase. *Plant Cell* 19, 1065–1080. doi: 10.1105/tpc.106.048884

Kobayashi, M., Yoshioka, M., Asai, S., Nomura, H., Kuchimura, K., Mori, H., et al. (2012). StCDPK5 confers resistance to late blight pathogen but increases susceptibility to early blight pathogen in potato via reactive oxygen species burst. *New. Phytol.* 196, 223–237. doi: 10.1111/j.1469-8137.2012.04226.x

Kumar, D., Yusuf, M. A., Singh, P., Sardar, M., and Sarin, N. B. (2014). Histochemical Detection of Superoxide and H_2O_2 Accumulation in *Brassica juncea* Seedlings. *Bio. Protocol.* 4:e1108.

Li, F., Fan, G., Lu, C., Xiao, G., Zou, C., Kohel, R. J., et al. (2015). Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. *Nat. Biotechnol.* 33, 524–530. doi: 10.1038/nbt.3208

Li, F., Fan, G., Wang, K., Sun, F., Yuan, Y., Song, G., et al. (2014a). Genome sequence of the cultivated cotton *Gossypium arboreum*. *Nat. Genet.* 46, 567–572. doi: 10.1038/ng.2987

Li, L., Li, M., Yu, L., Zhou, Z., Liang, X., Liu, Z., et al. (2014b). The FLS2-associated kinase BIK1 directly phosphorylates the NADPH oxidase RbohD to control plant immunity. *Cell Host. Microbe.* 15, 329–338. doi: 10.1016/j.chom.2014.02.009

Li, H., Li, K., Guo, Y., Guo, J., Miao, K., Botella, J. R., et al. (2018). A transient transformation system for gene characterization in upland cotton (*Gossypium hirsutum*). *Plant Methods* 14:50. doi: 10.1186/s13007-018-0319312

Li, T., Ma, X., Li, N., Zhou, L., Liu, Z., Han, H., et al. (2017). Genome-wide association study discovered candidate genes of Verticillium wilt resistance in upland cotton (*Gossypium hirsutum* L.). *Plant Biotechnol. J.* 15, 1520–1532. doi: 10.1111/pbi.12734

Li, Y. B., Han, L. B., Wang, H. Y., Zhang, J., Sun, S. T., Feng, D. Q., et al. (2016). The thioredoxin GbNRX1 plays a crucial role in homeostasis of apoplastic reactive oxygen species in response to *Verticillium dahliae* infection in cotton. *Plant Physiol.* 170, 2392–2406. doi: 10.1104/pp.15.01930

Lin, F., Zhang, Y., and Jiang, M. Y. (2009). Alternative Splicing and Differential Expression of Two Transcripts of Nicotine Adenine Dinucleotide Phosphate Oxidase B Gene from *Zea mays*. *J. Integr. Plant Biol.* 51, 287–298. doi: 10.1111/j.1744-7909.2008.00808.x

Liu, Y., Schiff, M., Marathe, R., and Dinesh-Kumar, S. P. (2002). Tobacco *Rar1*, *EDS1* and *NPR1/NIM1* like genes are required for N-mediated resistance to tobacco mosaic virus. *Plant J.* 30, 415–429. doi: 10.1046/j.1365-313x.2002.01297.x

Long, L., Xu, F.-C., Zhao, J., Li, B., Xu, L., and Gao, W. (2020). GbMPK3 overexpression increases cotton sensitivity to *Verticillium dahliae* by regulating salicylic acid signaling. *Plant Sci.* 292:110374. doi: 10.1016/j.plantsci.2019.110374

Ma, X., Wang, W., Bittner, F., Schmidt, N., Berkey, R., Zhang, L., et al. (2016). Dual and Opposing Roles of Xanthine Dehydrogenase in Defense-Associated Reactive Oxygen Species Metabolism in Arabidopsis. *Plant Cell* 28, 1108–1126. doi: 10.1105/tpc.15.00880

Ma, X., Wang, Z., Li, W., Zhang, Y., Zhou, X., Liu, Y., et al. (2019). Resequencing core accessions of a pedigree identifies derivation of genomic segments and key agronomic trait loci during cotton improvement. *Plant Biotechnol. J.* 17, 762–775. doi: 10.1111/pbi.13013

Ma, Z., He, S., Wang, X., Sun, J., Zhang, Y., Zhang, G., et al. (2018). Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. *Nat. Genet.* 50, 803–813. doi: 10.1038/s41588-018-0119117

Meng, J., Gao, H., Zhai, W., Shi, J., Zhang, M., Zhang, W., et al. (2018). Subtle regulation of cotton resistance to Verticillium wilt mediated by MAPKK family members. *Plant Sci.* 272, 235–242. doi: 10.1016/j.plantsci.2018.05.003

Mo, H., Wang, X., Zhang, Y., Zhang, G., Zhang, J., and Ma, Z. (2015). Cotton polyamine oxidase is required for spermine and camalexin signalling in the defence response to *Verticillium dahliae*. *Plant J.* 83, 962–975. doi: 10.1111/tpj.12941

Morales, J., Kadota, Y., Zipfel, C., Molina, A., and Torres, M. A. (2016). The Arabidopsis NADPH oxidases RbohD and RbohF display differential expression patterns and contributions during plant immunity. *J. Exp. Bot.* 67, 1663–1676. doi: 10.1093/jxb/erv558

Muller, K., Carstens, A. C., Linkies, A., Torres, M. A., and Leubner-Metzger, G. (2009). The NADPH-oxidase AtrbohB plays a role in Arabidopsis seed after-ripening. *New Phytol.* 184, 885–897. doi: 10.1111/j.1469-8137.2009.03005.x

Nuhse, T. S., Bottrill, A. R., Jones, A. M., and Peck, S. C. (2007). Quantitative phosphoproteomic analysis of plasma membrane proteins reveals regulatory mechanisms of plant innate immune responses. *Plant J.* 51, 931–940. doi: 10.1111/j.1365-313X.2007.03192.x

Ogasawara, Y., Kaya, H., Hiraoka, G., Yumoto, F., Kimura, S., Kadota, Y., et al. (2008). Synergistic activation of the Arabidopsis NADPH oxidase AtrbohD by Ca^{2+} and phosphorylation. *J. Biol. Chem.* 283, 8885–8892. doi: 10.1074/jbc.M708106200

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

Pogany, M., von Rad, U., Grun, S., Dongo, A., Pintye, A., Simoneau, P., et al. (2009). Dual roles of reactive oxygen species and NADPH oxidase RBOHD in an Arabidopsis-Alternaria pathosystem. *Plant Physiol.* 151, 1459–1475. doi: 10.1104/pp.109.141994

Sagi, M., Davydov, O., Orazova, S., Yesbergenova, Z., Ophir, R., Stratmann, J. W., et al. (2004). Plant respiratory burst oxidase homologs impinge on wound responsiveness and development in *Lycopersicon esculentum*. *Plant Cell* 16, 616–628. doi: 10.1105/tpc.019398

Sagi, M., and Fluhr, R. (2006). Production of reactive oxygen species by plant NADPH oxidases. *Plant Physiol.* 141, 336–340. doi: 10.1104/pp.106.078089

Shaban, M., Miao, Y., Ullah, A., Khan, A. Q., Menghwar, H., Khan, A. H., et al. (2018). Physiological and molecular mechanism of defense in cotton against *Verticillium dahliae*. *Plant Physiol. Biochem.* 125, 193–204. doi: 10.1016/j.plaphy.2018.02.011

Suzuki, N., Miller, G., Morales, J., Shulaev, V., Torres, M. A., and Mittler, R. (2011). Respiratory burst oxidases: the engines of ROS signaling. *Curr. Opin. Plant Biol.* 14, 691–699. doi: 10.1016/j.pbi.2011.07.014

Torres, M. A., and Dangl, J. L. (2005). Functions of the respiratory burst oxidase in biotic interactions, abiotic stress and development. *Curr Opin. Plant Biol.* 8, 397–403. doi: 10.1016/j.pbi.2005.05.014

Torres, M. A., Onouchi, H., Hamada, S., Machida, C., Hammond-Kosack, K. E., and Jones, J. D. G. (1998). Six Arabidopsis thaliana homologues of the human respiratory burst oxidase (*gp91^{phox}*). *Plant J.* 14, 365–370. doi: 10.1046/j.1365-313x.1998.00136.x

Wang, J., Tian, L., Zhang, D. D., Short, D. P. G., Zhou, L., Song, S. S., et al. (2018). SNARE-encoding genes *VdSec22* and *VdSso1* mediate protein secretion required for full virulence in *Verticillium dahliae*. *Mol. Plant Microbe. Interact.* 31, 651–664. doi: 10.1094/MPMI-12-17-0289-R

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

Wang, W., Chen, D., Liu, D., Cheng, Y., Zhang, X., Song, L., et al. (2020). Comprehensive analysis of the *Gossypium hirsutum* L. respiratory burst oxidase homolog (*Ghrboh*) gene family. *BMC Genomics* 21:91. doi: 10.1186/s12864-020-65036506

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

Wong, H. L., Pinontoan, R., Hayashi, K., Tabata, R., Yaeno, T., Hasegawa, K., et al. (2007). Regulation of rice NADPH oxidase by binding of Rac GTPase to its N-terminal extension. *Plant Cell* 19, 4022–4034. doi: 10.1105/tpc.107.055624

Xia, X. J., Wang, Y. J., Zhou, Y. H., Tao, Y., Mao, W. H., Shi, K., et al. (2009). Reactive oxygen species are involved in brassinosteroid-induced stress tolerance in cucumber. *Plant Physiol.* 150, 801–814. doi: 10.1104/pp.109.138230

Yang, C., Li, W., Cao, J., Meng, F., Yu, Y., Huang, J., et al. (2017). Activation of ethylene signaling pathways enhances disease resistance by regulating ROS and phytoalexin production in rice. *Plant J.* 89, 338–353. doi: 10.1111/tpj.13388

Yang, Z., Ge, X., Yang, Z., Qin, W., Sun, G., Wang, Z., et al. (2019). Extensive intraspecific gene order and gene structural variations in upland cotton cultivars. *Nat. Commun.* 10:2989. doi: 10.1038/s41467-019-10820-x

Yao, Y., He, R. J., Xie, Q. L., Zhao, X. H., Deng, X. M., He, J. B., et al. (2017). ETHYLENE RESPONSE FACTOR 74 (ERF74) plays an essential role in controlling a respiratory burst oxidase homolog D (RbohD)-dependent mechanism in response to different stresses in Arabidopsis. *New Phytol.* 213, 1667–1681. doi: 10.1111/nph.14278

Yoshioka, H., Numata, N., Nakajima, K., Katou, S., Kawakita, K., Rowland, O., et al. (2003). *Nicotiana benthamiana gp91^{phox}* homologs *NbrbohA* and *NbrbohB* participate in H_2O_2 accumulation and resistance to *Phytophthora infestans*. *Plant Cell* 15, 706–718. doi: 10.1105/tpc.008680

Yu, S., Kakar, K. U., Yang, Z., Nawaz, Z., Lin, S., Guo, Y., et al. (2019). Systematic study of the stress-responsive Rboh gene family in *Nicotiana tabacum*: Genome-wide identification, evolution and role in disease resistance. *Genomics* 19, S0888–S7543. doi: 10.1016/j.ygeno.2019.08.010

Yun, B.-W., Feechan, A., Yin, M., Saidi, N. B. B., Le Bihan, T., Yu, M., et al. (2011). S-nitrosylation of NADPH oxidase regulates cell death in plant immunity. *Nature* 478, 264–268. doi: 10.1038/nature10427

Zhang, G., Yue, C., Lu, T., Sun, L., and Hao, F. (2020). Genome-wide identification and expression analysis of NADPH oxidase genes in response to ABA and abiotic stresses, and in fibre formation in *Gossypium*. *PeerJ.* 8:e8404. doi: 10.7717/peerj.8404

Zhang, H., Fang, Q., Zhang, Z., Wang, Y., and Zheng, X. (2009). The role of respiratory burst oxidase homologues in elicitor-induced stomatal closure and hypersensitive response in *Nicotiana benthamiana*. *J. Exp. Bot.* 60, 3109–3122. doi: 10.1093/jxb/erp146

Zhang, T., Hu, Y., Jiang, W., Fang, L., Guan, X., Chen, J., et al. (2015). Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. *Nat. Biotechnol.* 33, 531–537. doi: 10.1038/nbt.3207

Zhao, J., Chen, Q., Zhou, S., Sun, Y., Li, X., and Li, Y. (2020). H2Bub1 Regulates RbohD-Dependent Hydrogen Peroxide Signal Pathway in the Defense Responses to *Verticillium dahliae* Toxins. *Plant Physiol.* 182, 640–657. doi: 10.1104/pp.19.00913

Zhao, Y., Zhang, T., and Guo, H. (2017). Penetration assays, fungal recovery and pathogenicity assays for *Verticillium dahliae*. *Bio. Protocol.* 7:e213. doi: 10.21769/BioProtoc.2133

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.

Copyright © 2020 Chang, Li, Shi, Geng, Geng, Liu, Zhang and Cai. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 15 September 2020
doi: 10.3389/fgene.2020.562937

Transcriptomic Analysis Reveals Important Roles of Lignin and Flavonoid Biosynthetic Pathways in Rice Thermotolerance During Reproductive Stage

Zhenzhen Cai[1], Fengyu He[1], Xin Feng[1], Tong Liang[1], Hongwei Wang[1,2,3], Shuangcheng Ding[1,2,3] and Xiaohai Tian[1,2,3]**

[1] *Engineering Research Center of Ecology and Agricultural Use of Wetland, Ministry of Education, Agricultural College, Yangtze University, Jingzhou, China,* [2] *Hubei Key Laboratory of Waterlogging Disaster and Agricultural Use of Wetland, Agricultural College, Yangtze University, Jingzhou, China,* [3] *Hubei Collaborative Innovation Center for Grain Industry, Agricultural College, Yangtze University, Jingzhou, China*

OPEN ACCESS

Edited by:
Rong Zhou,
Aarhus University, Denmark

Reviewed by:
Kashmir Singh,
Panjab University, India
Qiu-Ning Liu,
Yancheng Teachers University, China
Mehar Hasan Asif,
National Botanical Research Institute (CSIR), India

***Correspondence:**
Shuangcheng Ding
shchding@yangtzeu.edu.cn
Xiaohai Tian
xiaohait@sina.com

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

Received: *17 May 2020*
Accepted: *27 August 2020*
Published: *15 September 2020*

Citation:
Cai Z, He F, Feng X, Liang T, Wang H, Ding S and Tian X (2020) Transcriptomic Analysis Reveals Important Roles of Lignin and Flavonoid Biosynthetic Pathways in Rice Thermotolerance During Reproductive Stage. Front. Genet. 11:562937. doi: 10.3389/fgene.2020.562937

Rice is one of the major staple cereals in the world, but heat stress is increasingly threatening its yield. Analyzing the thermotolerance mechanism from new thermotolerant germplasms is very important for rice improvement. Here, physiological and transcriptome analyses were used to characterize the difference between two germplasms, heat-sensitive MH101 and heat-tolerant SDWG005. Two genotypes exhibited diverse heat responses in pollen viability, pollination characteristics, and antioxidant enzymatic activity in leaves and spikelets. Through cluster analysis, the global transcriptomic changes indicated that the ability of SDWG005 to maintain a steady-state balance of metabolic processes played an important role in thermotolerance. After analyses of gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment, we found that the thermotolerance mechanism in SDWG00 was associated with reprogramming the cellular activities, such as response to abiotic stress and metabolic reorganization. In contrast, the down-regulated genes in MH101 that appeared to be involved in DNA replication and DNA repair proofreading, could cause serious injury to reproductive development when exposed to high temperature during meiosis. Furthermore, we identified 77 and 11 differentially expressed genes (DEGs) involved in lignin and flavonoids biosynthetic pathways, respectively. Moreover, we found that more lignin deposition and flavonoids accumulation happened in SDWG005 than in MH101 under heat stress. The results indicated that lignin and flavonoid biosynthetic pathways might play important roles in rice heat resistance during meiosis.

Keywords: rice, heat stress, transcriptome, spikelet, meiosis, lignin and flavonoid biosynthetic pathways

INTRODUCTION

With an increase in the frequency and intensity of extreme heat waves, high-temperature stress has become one of the severe limiting factors affecting crop production and distribution worldwide (Lobell and Gourdji, 2012; Wang et al., 2016; Yang et al., 2017). The mean global temperature has been predicted to rise 2.5–5.8°C by the end of the 21st century (IPCC 2014). The yield of crops significantly negatively correlated with temperature rise at least in the three major crops, rice, maize, and wheat (Yu et al., 2012; Kawahara et al., 2013; Li et al., 2016). The data showed that rice yields would decline by 10% with each 1°C increase in average minimum temperature (Peng et al., 2004).

The reproductive stage of rice was the most sensitive phase to heat stress. Of all the reproductive processes, the most susceptible stages to heat stress were flowering and microspore meiosis (Satake and Yoshida, 1978; Jagadish et al., 2007). Jagadish et al. (2007) reported that cumulative temperature over 33.7°C even as short as 1 hour at anthesis could lead to infertility in *indica* rice. Tian et al. (2009) observed that a mean temperature of 30°C, with approximately 70% relative humidity and low wind speeds, during the flowering period for longer than three days could lead to significantly lower seed setting in rice plants (Tian et al., 2009). Several previous studies have also revealed that pollen abortion and floret sterility were induced by exposure to a high temperature of about 33°C at meiosis of microspore mother cell stage (Prasad et al., 2006; Krishnan et al., 2011). In comparison to flowering, rice pollen abortion and floret sterility were relatively less susceptible to high temperature at the meiosis stage (Matsui et al., 2000; Jagadish et al., 2010, 2014; Shah et al., 2011). During meiosis stage, male sterility induced by high temperature appears to be associated with the poor pollen development (Matsui et al., 2000; Ku et al., 2003). Overall, the spikelet fertility was primarily determined by successful pollination and fertilization, thus the heat-induced decrease in spikelet fertility at reproductive stages was mainly attributed to lack of full development of pollens, abnormal anther dehiscence, poor pollen viability, and failed germination and/or elongation on the stigma (Satake and Yoshida, 1978; Matsui et al., 2000; Jagadish et al., 2010; Yu et al., 2016). Although many physiological studies have contributed to our understanding of heat-induced pollen fertility injury, molecular analyses are still largely lacking.

In recent decades, some heat stress response pathways and key regulators that function during anther development have been clarified (Frank et al., 2009; Jagadish et al., 2010). Several studies showed that during heat stress the imbalance between ROS and ROS scavenge system existed, and then induced the initiation of tapetum programmed cell death (PCD) (Suzuki et al., 2001; Yu et al., 2016). The genes participating in starch and sucrose metabolism in anthers were reported as crucial metabolic and transcriptional components in the acquisition of thermotolerance (Scharte et al., 2009; Li et al., 2015). Additionally, some researchers have also reported that plant hormones were key regulators of pollen development and thermotolerance (Tang et al., 2007; Mittler et al., 2012; Bokszczanin et al., 2013). Thus, different signal transduction events were triggered to establish homeostasis of metabolic processes by altering the composition of certain transcripts, proteins, metabolites, and lipids (McClung and Davis, 2010). Furthermore, a lot of heat-responsive genes and proteins, including heat shock proteins, antioxidant enzymes, and various transcription factors, have been identified and their functions well elucidated (Guo et al., 2008; Ramamoorthy et al., 2008; Scafaro et al., 2010).

However, currently, our understanding remains limited on how plants respond to heat stress at the whole plant level, especially on the core pathways leading to increased thermotolerance in the crucial floral organs that determine reproductive success. Therefore, it is necessary to discover the genes regulating thermotolerance and investigate the intrinsic molecular mechanism in new thermotolerant germplasms. Here, physiological and transcriptome analyses were used to characterize the difference between two germplasms, heat-sensitive MH101 and heat-tolerant SDWG005. SDWG005, belonging to *Oryza sativa* spp. *indica,* is an African landrace that exhibits heat-tolerant, while MH101, a conventional *indica* variety rice, is very sensitive to high temperature. We firstly compared the different responses to heat stress between SDWG005 and MH101 by observing their phenotypic characteristics, including pollen viability, pollination properties, and antioxidant enzymatic activities. Then we analyzed the global transcriptomic changes in response to heat stress in the two genotypes, and identified the important roles of lignin and flavonoids biosynthetic pathways in rice thermotolerance during meiosis. The results will lay an important foundation for understanding thermotolerance mechanisms in rice spikelet, and provide insights into the underlying molecular basis of heat responses.

MATERIALS AND METHODS

Plant Materials and Treatments

Two rice (*Oryza sativa* L.) genotypes, SDWG005 and MH101, were used in this study. SDWG005 belonging to *Oryza sativa* spp. *indica,* is an African landrace that exhibits heat-tolerant, while MH101, a conventional *indica* variety rice, is very sensitive to high temperature. Twenty-day-old seedlings were transplanted into a plastic pot (inner diameter 30 cm, height 30 cm) containing 12.5 kg soil and 8 g N: P: K compound fertilizer (26:10:15). The tillers were cut off leaving only the main stem during the plant culture. The plants were grown until the meiosis stage, when the distance between the pulvinus of the first leaf and second leaf from the top was about 2 cm. One day later, the rice plants were transferred to a growth chamber (AGC-MR, ZhejiangQiushi Environment Co., Ltd, Zhejiang, China) and started the heat treatment with a 14-h-day/10-h-night cycle and a 2-h change on the temperature simulating the local typical heat stress weather. The actual temperature regimes were 26, 30, and 34°C in daily mean temperature, respectively, and their corresponding daily maximum temperature were 28, 36 and 38°C. After heat treatment for three days, spikelet and flag leaves were collected from both heat stress and control plants (six independent plants of each) for each one of MH101 and SDWG005 for the extraction of total RNA or detection of antioxidant enzymatic activities. All

collected samples were immediately frozen in liquid nitrogen and stored at −80°C. Before anthesis (after heat treatment for eight days), the rice plants were moved out from growth chamber, and analyzed for the pollen viability and pollination properties.

Determination of Pollen Viability

The spikelets in the panicle containing anthers were collected on day 3 of the heat stress treatment. The pollen grains removed from the anthers were stained with 1% KI/I_2 on a glass slide, and subsequently observed and photographed with a magnification of 100× under a light microscope (DM4000B; Leica, Wetzlar, Germany). The pollen grains were divided into fertile and sterile groups, based on their colors. In each photo, 30–150 pollen grains were analyzed (10 repetitions). Percentage of pollen viability = (the number of fertile pollen grains/the total number of pollen grains) × 100%.

Determination of SOD, POD, APX, and MDA

Frozen spikelets or flag leaves (0.5 g) were ground into a fine powder in liquid nitrogen and then homogenized in a 50mM sodium phosphate buffer (pH 7.0). The homogenate was centrifuged at 13,000 g for 15min at 4°C, and the supernatant was stored in aliquots at −20°C until further use. The superoxide dismutase (SOD) activity was determined using the method of Giannopolitis and Ries (1977). The method of Maehly and Chance (1954), in which guaiacol is converted to tetraguaiacol and monitored at 470 nm, was used to determine the peroxidase (POD) activity. Ascorbate peroxidase (APX) activity was measured by the method of Bonnecarrère et al. (2011). The Malondialdehyde (MDA) content was measured by a modified TBA (Thiobarbituric acid) method as described (Peever and Higgins, 1989).

Determination of Lignin, Flavonoids, and Phenolics Content

Frozen spikelets were collected, oven-dried at 65°C to constant weight, and mortar ground over an 80-mesh sieve. Lignin content was determined from the cell wall fraction (removing of other compounds by phosphate buffer, 1% Triton X-100, 1 M NaCl and acetone) by thioglycolic acid reaction (Kovacik et al., 2007). Total phenolics were extracted by 80% methanol and calculated from the calibration curve prepared using gallic acid by the Folin-Ciocalteu method (Kovacik et al., 2007). The aluminum chloride colorimetric assay was used to determine the flavonoid content of both free phenolics and bound phenolics of plant extracts, of which, the total flavonoid content was established using a calibration curve and expressed as mg/g FW (Djeridane et al., 2006).

RNA Isolation and cDNA Library Construction

The spikelet sampled from control (28°C) and heat (36°C, 38°C) treated for MH101 and SDWG005 were collected for isolating RNA and cDNA library construction. Total RNA was prepared from spikelet with the Promega RNA extraction kit (Promega, China, LS1040). RNA integrity and quantity were measured with a Nanodrop N60 spectrophotometer (Implen) and an Agilent's Bioanalyzer 2100. The triplications were prepared for each sample, resulting in a total of 18 libraries sequenced by the Illumina HiSeq X Ten system to produce 150 bp paired-end reads per library. Sequencing was performed at Beijing Genomics Institute (Wuhan, China).

Detection of Differentially Expressed Genes

The rice genome version Rice Genome Annotation Project (RGAP7) was downloaded from http://rice.plantbiology.msu.edu (Kawahara et al., 2013). RNA-seq reads were aligned to the rice genome using the alignReads mode of STAR 2.4.1c (Dobin et al., 2013) with the default settings. FeatureCounts 1.4.6 (Liao et al., 2014) was used to summarize uniquely mapped paired-end reads and count fragments with "-t exon" and "-g gene_id". For MH101 or SDWG005, differentially expressed genes (DEGs) were derived from the comparison between control (28°C) and heat treatment under different temperatures (36°C, 38°C). DEGs were detected using the edgeR package (version 3.24.3) (Robinson et al., 2010) with a false discovery rate (FDR) adjusted *P*-value cutoff at 0.05 and an absolute value of $\log_2$ transformed fold change relative to control ($\log_2$FC) cutoff at 1.

Principal Component and Cluster Analyses

PCA was performed with the prcomp function of the base "stat" package (version 3.5.1) in R, using the $\log_2$ transformed raw reads counts as input. Differentially expressed genes in at least MH101 or SDWG005 were subjected to hierarchical clustering using the hclust function in R (distance: euclidean; method: ward.d) and the $\log_2$FC was used as the input. The hierarchical clustering was manually renumbered for simplicity of comparison.

GO Categories and KEGG Pathways Analysis

To identify the biological functions over-represented within the DEGs between up- and down-related DEGs or each numbered cluster, functional enrichment analyses were performed using the AgriGO program (Du et al., 2010). Singular enrichment analysis was used to identify enriched GO terms with the FDR adjusted *P*-value cutoff of 0.05. The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Ogata et al., 1999) enriched within DEGs were analyzed with the Fisher's exact test and the Benjamini-Hochberg multiple testing correction method.

Real-Time PCR Analysis

The relative expression of the selected genes was measured using quantitative real time-PCR (RT-qPCR) in 96-well plates using the ABI7500 Real-Time PCR Systems (Applied Biosystems, United States). The PCR reaction system consists of 1 μg cDNA, 200 nM primers, and 5 μl SYBR Premix Ex Taq II (Takara, China), and the reaction volume is 10 μl. The PCR reaction was conducted with the following conditions: 10 min at 94°C, 40 cycles of 15 s at 94°C, and 30 s at 60°C. The expression

level of the *O. sativa* β-*tubline* gene was used as internal control. For quantification, the $2^{-\Delta\Delta Cq}$ method was used, and variations (standard errors) were calculated from three biological replicates. Primers for qRT-PCR are listed in **Supplementary Table S1**.

RESULTS

Pollen Viability and Pollination Properties of SDWG005 and MH101 Display Differences Under Heat Stress

Previously, six rice varieties from various regions of Africa were treated with high temperature in flowering period to evaluate the thermotolerance based on the fertilization rate under different high-temperature treatments (Liu et al., 2019). The results showed that SDWG005, an extremely strong heat-resistant African landrace, was weakly affected by heat stress. In contrast, MH101 was more sensitive to heat stress. To further confirm their differential responses to high temperature, both varieties were treated under varied heat regimes for about eight days at meiosis stage, and the pollen viability and pollination properties were analyzed. The results showed that heat stress at the meiosis stage significantly affected pollen viability and pollination properties in rice plants, and the extent of the viability reduction and deficient pollination were dependent on the genotypes and temperature (**Figure 1**). Compared with the normal growth control (CK), treatments of T36 (T_{max} = 36°C) and T38 (T_{max} = 36°C) resulted in 76.83% and 79.73% reduction, respectively, in pollen viability in MH101 plants, whereas in only 13.74% and 37.82% reduction in SDWG005 plants (**Figures 1A,B**). The percentage of basal dehiscent anthers was measured in the flowers of SDWG005 and MH101 plants at anthesis, and the results showed that heat stress significantly reduced the rate of basal anther dehiscence in the two genotypes (**Figure 1C**). Compared to the normal growth control, treatments of T36 and T38 led to 13.98% and 30.41% decrease, respectively, in SDWG005. In MH101 plants, treatments of T36 and T38 caused 36.55% and 69.6% decrease, respectively, significantly greater than in SDWG005 (**Figure 1C**). We further compared the number of pollens on the stigma in SDWG005 and MH101 under different high-temperature treatments, and analyzed the percentage of stigma with ≥ 20 pollen grains. As shown in **Figure 1C**, MH101 was more significantly affected than SDWG005 by temperature with regard to the pollen number on stigma (**Figure 1C**). These results indicated that the pollen viability and pollination properties of SDWG005 decreased much slower than those of MH101 when temperature increased, confirming that SDWG005 (heat-tolerant) is much more tolerant to heat stress than MH101.

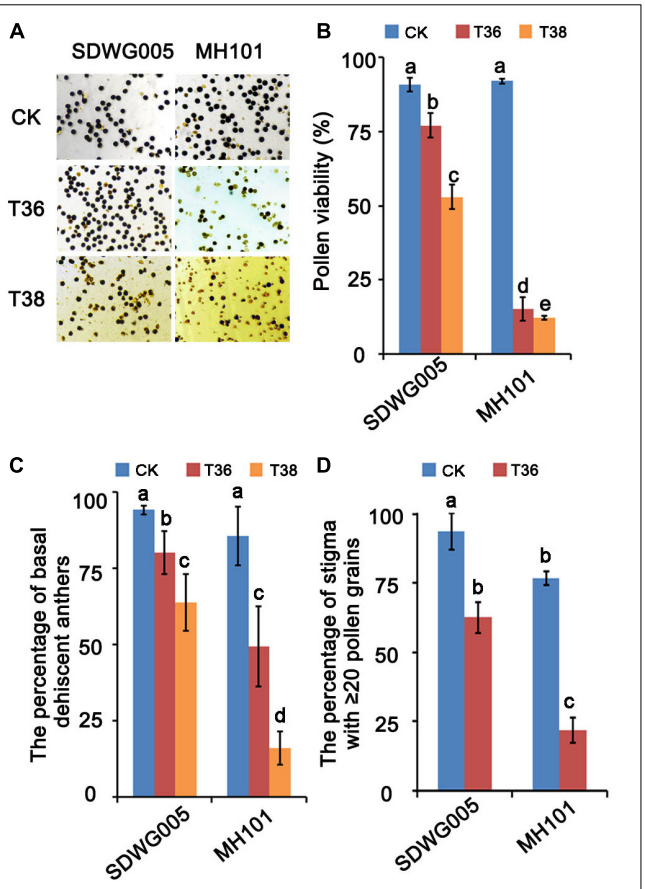


FIGURE 1 | Effect of heat stress on pollen viability and pollination properties of SDWG005 and MH101. **(A,B)** Differences in pollen viability of SDWG005 (heat-tolerant) and MH101 (heat-susceptible) exposed eight days to different temperature treatments. **(C)** The percentage of basal dehiscent anthers under high-temperature treatments. **(D)** The percentage of stigma with ≥ 20 pollen grains at anthesis. CK, T36 and T38 indicate normal temperature (T_{max} = 28°C) and high temperatures T36 (T_{max} = 36°C) and T38 (T_{max} = 38°C). Each data point represents the mean ± SD (n = 3) of three biological replicates. Significant differences are calculated by One-way ANOVA with Duncan's multiple test (SAS Institute, Inc., Cary, NC, United States). Different letters indicate statistical differences between samples at P = 0.05.

SDWG005 Carries Higher SOD, POD, and APX Activities Than MH101, and Is Less Affected by Heat

Enzymatic antioxidants, such as superoxide dismutase (SOD), peroxidase (POD), and ascorbate peroxidase (APX), are important reactive oxygen species (ROS) scavenging enzymes in reducing heat-induced ROS levels. We then examined the antioxidant enzymatic activities of SOD, POD, and APX in the flag leaves and spikelets of SDWG005 and MH101, respectively, after T36 and T38 treatments for three days. Whether upon heat stress or not, SDWG005 showed higher SOD and POD activities than MH101 in flag leaves (**Figure 2A**). No significant differences among SOD activity were observed in the flag leaves of two varieties under high-temperature stress. However, MH101 showed a significant decreases in POD and APX activities in flag leaves under T38 treatment; while SDWG005 was little affected, or even displayed a slight increase in POD and APX activities at T36 treatment (**Figure 2A**). Under normal conditions, SDWG005 and MH101 showed no differences among SOD and POD activities in spikelets, expect for APX activity (**Figure 2B**). Whereas MH101 had clear decreases in SOD, POD, and APX activities in spikelets under heat stress, SDWG005 only had a

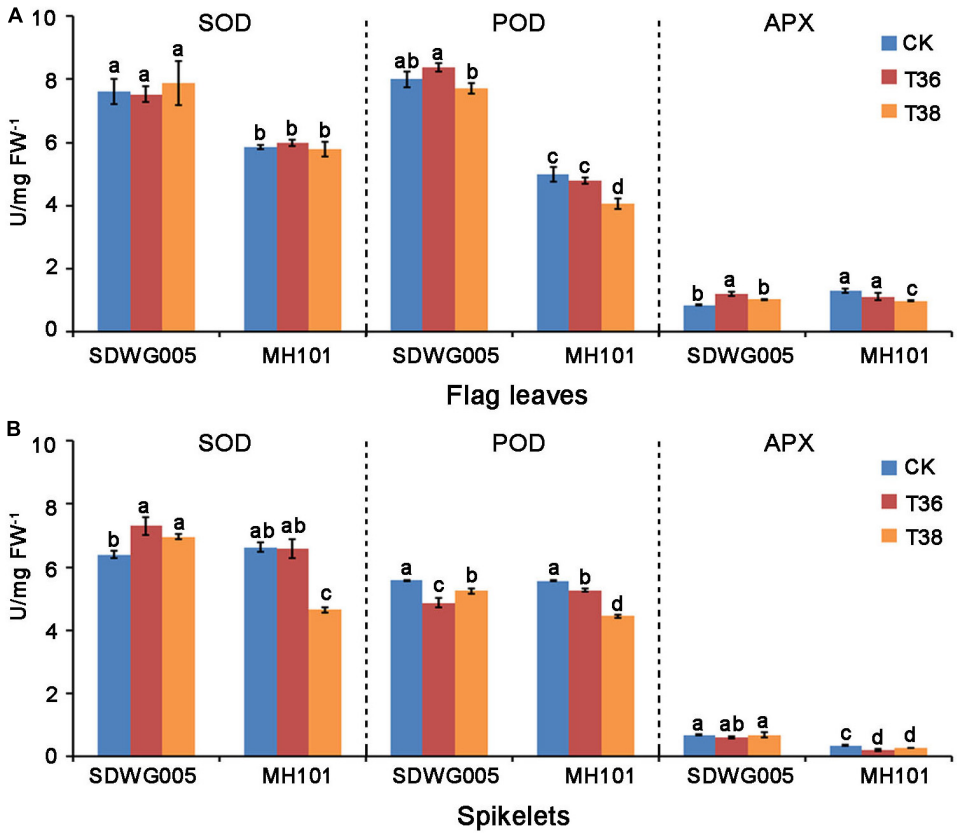


FIGURE 2 | SOD, POD, and APX activities in rice flag leaves and spikelets under different high-temperature stresses. **(A)** SOD, POD, and APX activities in rice flag leaves when exposed three days to different temperature treatments. **(B)** SOD, POD, and APX activities in rice spikelets when exposed three days to different temperature treatments. CK, T36 and T38 indicate normal temperature (about 28°C) and high temperatures T36 (T_{max} = 36°C) and T38 (T_{max} = 38°C). Each data point represents the mean ± SD of three biological replicates. Significant differences were calculated by One-way ANOVA with Duncan's multiple test (SAS Institute, Inc., Cary, NC, United States). Different letters indicate the statistical difference between samples at P = 0.05.

significant drop in POD activity at T36 (**Figure 2B**). These results suggested that SDWG005 carries higher SOD, POD, and APX activities than MH101 in flag leaves and spikelets, and is less affected by heat.

Transcriptome Sequencing, Detection of DEGs and Validation Analysis by RT-qPCR

To characterize the differential transcriptomic responses to heat stress between SDWG005 and MH101 during meiosis, we subjected spikelets of SDWG005 and MH101 to 36°C (T36) and 38°C (T38) treatments for three days or grown at 28°C (T28, control). Thus, for each one of SDWG005 and MH101, there were three conditions, including two treatments (T36 and T38) and one control (T28). With three biological replicates for each condition, a total of 18 cDNA libraries were constructed, and sequenced using the Illumina deep sequencing platform. After filtering out the low-quality reads, a total of ∼369 million paired-end clean reads were obtained with an average of 86.34% that were uniquely mapped to the rice reference genome (RGAP7[1]) (**Supplementary Table S2**). Principal component analysis (PCA) was performed to capture the overall variance among treatments. Principal components 1 (PC1) and 2 (PC2) captured 43.54% and 26.85% of the variation in the gene expression data, respectively. A clustering effect based on temperatures was observed along the PC1 dimension, while the PC2 dimension separated the effect of materials (**Figure 3A**). Moreover, replicates of each one of the heat-stressed conditions for MH101 or SDWG005 were clustered nearby but were separated from the corresponding controls.

To identify DEGs for each treatment relative to the control of MH101 or SDWG005, edgeR (Robinson et al., 2010) was used and the outputs were further filtered using the criteria of the absolute value of log2 transformed fold change ($|\log_2 FC| \geq 1$ and false discovery rate (FDR) adjusted P-value ≤ 0.05. As expected, numerous DEGs from the four treatments (T36-SDWG005, T38-SDWG005, T36-MH101, and T38-MH101) relative to the corresponding control were detected, with significantly higher/lower expression in SDWG005 (green and red) or MH101 (pink and blue) color-coded in **Figure 3B**. Compared to the control, a total of 3408 DEGs (1869 up- and 1539 down-regulated) in T36-MH101 and 6242 (3724 up- and 2518 down-regulated) in T38-MH101

[1]http://rice.plantbiology.msu.edu

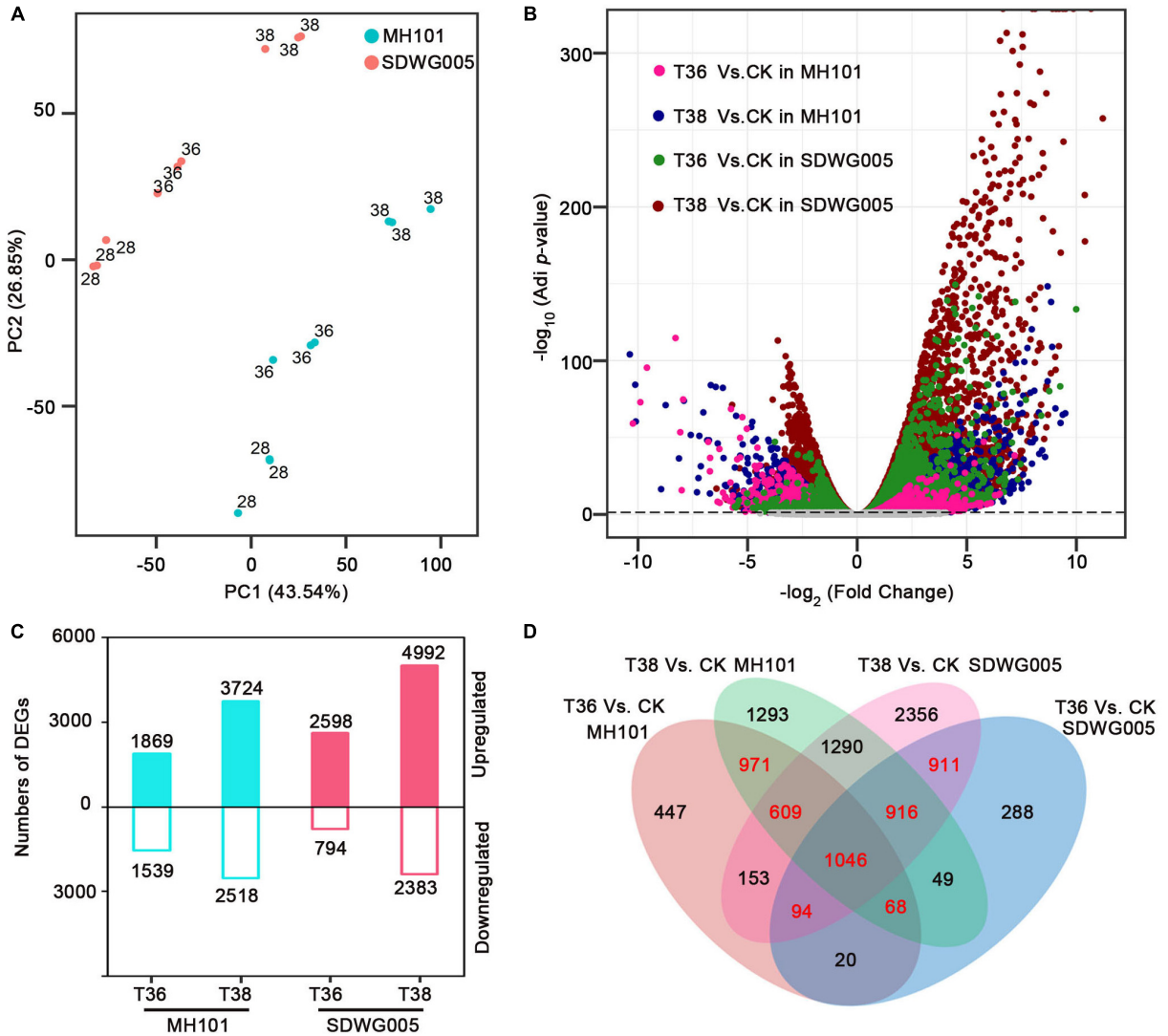


FIGURE 3 | Analysis of DEGs in rice varieties in response to heat stress. **(A)** PCAs of $\log_2$ transformed raw read counts for CK, T36, and T38 for MH101 and SDWG005, respectively. **(B)** Volcano plot of expressed genes under different heat stress. Log_2 transformed FC was plotted against the $-\log_{10}$ transformed adjusted *P*-value. Pink, blue, green and red dots represent DEGs of T36 relative to CK in MH101, T38 relative to CK in MH101, T36 relative to CK in SDWG005, T38 relative to CK in SDWG005, respectively. The gray line represents the $-\log_{10}$ transformed adjusted *P*-value cutoff. **(C)** Number of up- ($\log_2\text{FC} \geq 1$ and FDR adjusted *P*-value ≤ 0.05) and down-regulated ($\log_2\text{FC} \leq -1$ and FDR adjusted *P*-value ≤ 0.05) DEGs detected under different heat stress treatments relative to control for MH101 and SDWG005, respectively. **(D)** Venn's diagram of the comparison of DEGs under different heat stress treatments in MH101 and SDWG005.

were identified (**Figure 3C**); while a total of 3392 (2598 up- and 794 down-regulated) in T36-SDWG005 and 7375 (4992 up- and 2383 down-regulated) DEGs in T38-SDWG005 were identified. Additionally, as illustrated in the diagrams (**Figure 3D**), the partial overlap of DEGs among treatments were examined, and the results showed that in MH101(including 2694 DEGs) or SDWG005(including 2967 DEGs), a total of 4615 DEGs (in red) responded to T36 and T38 heat treatments (**Supplementary Table S3**).

To further examine the response to heat stress, DEGs with significant changes ($\log_2$FC values) in T36 treatment were plotted against those of T38 treatment for MH101 and SDWG005, respectively (**Supplementary Figure S1**). Consistent with the PCAs analysis, transcriptomic response to different heat stress treatments were highly correlated in each rice variety (R^2-values was 0.73 and 0.68, all *P*-values < 0.01; **Supplementary Figure S1**). The results show that under heat stress conditions, the transcriptomic changes are relatively consistent within treatments (T36 vs T38). In order to validate the RNA-seq data and confirm the differential expression genes, we performed RT-qPCR on ten random candidate DEGs. The RT-qPCR results indicated that these candidate DEGs displayed the same expression patterns compared with the sequencing data, thereby confirming the reliability of the RNA-seq data (**Supplementary Figure S2**).

The Importance of Maintaining the Homeostasis Balance of Metabolic Processes in Rice Resistance to Heat Damage

To further characterize gene expression changes upon heat stress treatments and among rice varieties, we performed a hierarchical clustering by using the expression changes (log_2FC value) for the identified 4615 DEGs shown as red number in **Figure 3D** (**Supplementary Table S3**). As shown in **Figure 4A**, analogous clusters with similar patterns of expression changes were readily observed. Therefore, we identified four clusters of DEGs, along with a red line reflecting the mean changes of expression describing each transcript accumulation pattern (**Figure 4B**). Cluster 1 included 712 DEGs that were progressively down-regulated under the two treatments (T36 and T38) in both genotypes, and the expression changes of these DEGs in SDWG005 were relatively more repressed than in MH101. Clusters 3 and 4 containing 873 and 2,037 DEGs, respectively, possessed most of the DEGs, and most of the DEGs were up-regulated by heat stress. The difference was that the genes in cluster 3 showed relatively higher induced expression levels

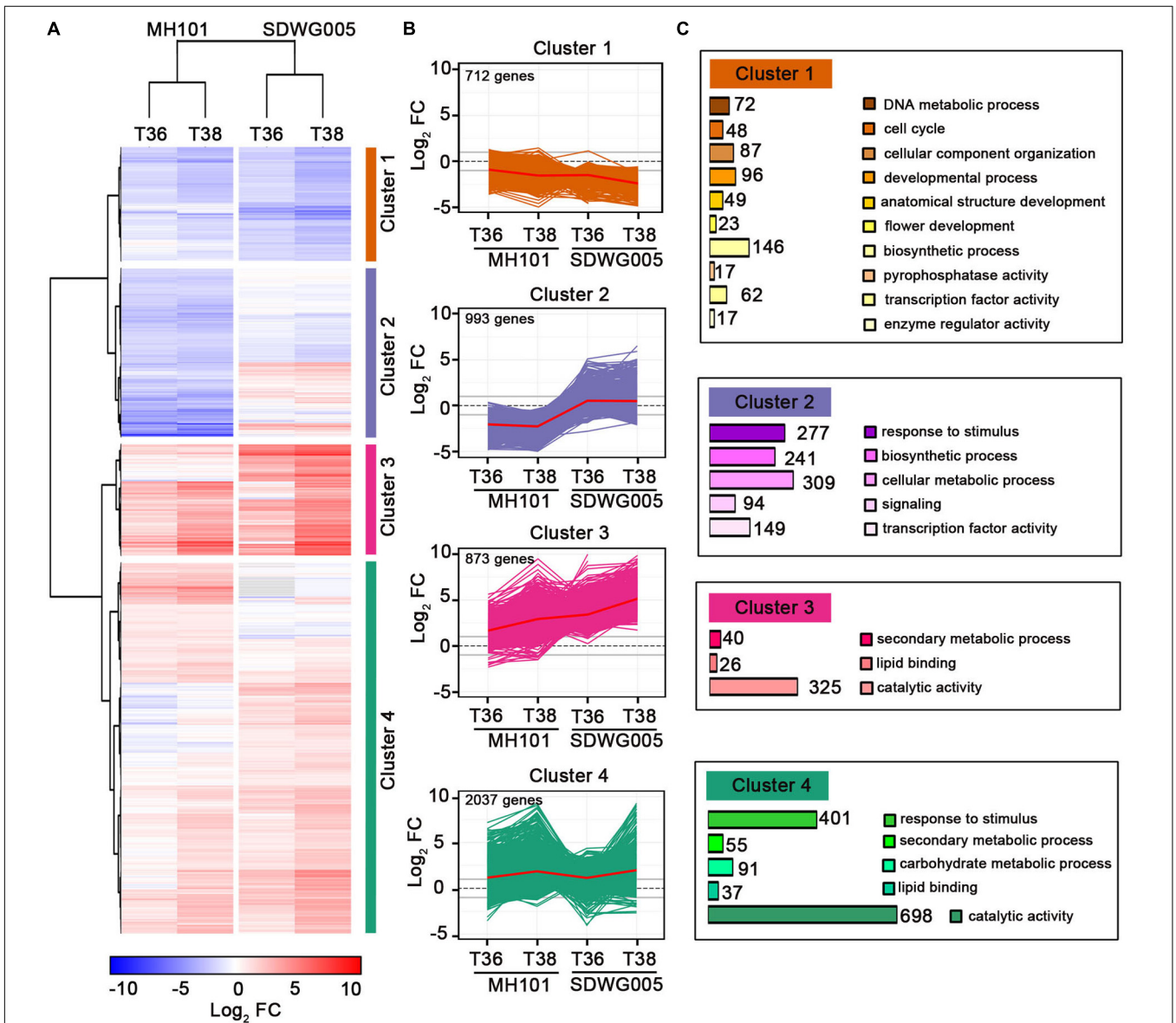


FIGURE 4 | Global patterns of gene expression during heat stress. **(A)** Hierarchical clustering analysis to reveal the transcript expression patterns of DEGs in MH101 and SDWG005 spikelets subjected to different high temperature treatments (T36 or T38). Colored bars show the fold changes of gene expression under T36 and T38 treatments compared to their corresponding controls. **(B)** Gene expression profiles in clusters 1, 2, 3, and 4 as indicated. Lines reflected expression changes for each DEGs across different treatments. The mean changes of expression in each cluster are shown with a red line. **(C)** GO categories of DEGs enriched in each cluster. The number show the DEGs enriched in each GO categories.

in SDWG005 than in MH101 in response to heat stress. Interestingly, cluster 2 contained 993 DEGs, whose expression patterns under heat stress were contrary to that of SDWG005 and MH101, and the expression levels of SDWG005 were relatively higher than that of MH101 (**Figure 4B**).

To relate DEGs in clusters to biological functions, we performed a Gene Ontology (GO) enrichment analysis for biological process by agriGO (Du et al., 2010). The relative expression of DEGs in clusters 3 and 4, which mainly participated in secondary metabolic processes, lipid binding, and catalytic activity, were up-regulated in response to heat stress (**Figure 4C**). The DEGs belonging to stimuli responses and carbohydrate metabolic processes were also greatly enriched in cluster 4. The enriched GO terms of cluster 1 indicated that some genes involved in metabolic processes (catabolism and biosynthesis), transcription factor activity, cellular component organization, and developmental growth were repressed upon heat stress in both genotypes. By comparison, in cluster 2, enriched GO terms involved in processes included response to stimulus, biosynthetic process, cellular metabolic process, signaling, and transcription activity (**Figure 4C**). Interestingly, the relative expression value were higher in heat-tolerant SDWG005 than in heat-susceptible MH101 The results indicated that compared with MH101, heat-tolerant rice SDWG005 might still be able to establish a new steady-state balance of metabolic processes under heat stress condition.

Identification of the Common and Genotype-Specific Mechanisms That Associated With Resistance to Heat

To further investigate the biological pathways of the up- or down-regulated DEGs in heat-tolerant and heat-sensitive varieties (**Supplementary Table S4**), we performed the GO and KEGG enrichment analyses. The representative GO categories enriched in SDWG005 and MH101 were selected respectively, and the heatmaps representing the $\log_{10}$-transformed *P*-values were constructed (**Figure 5**). We found that among up-regulated DEGs, GO terms associated with response to abiotic stress, oxidative stress, multiple metabolic processes, signal transduction, electron transport, and reproductive related development processes were more significantly enriched in tolerant variety SDWG005 (**Figure 5A**). While the GO terms involved in cell death, flower development, and reproductive process were specifically enriched in susceptible rice MH101 in response to high temperature (**Figure 5A**). On the other hand, GO terms for the down-regulated DEGs correlated with response to abiotic stress, signal transduction and reproductive

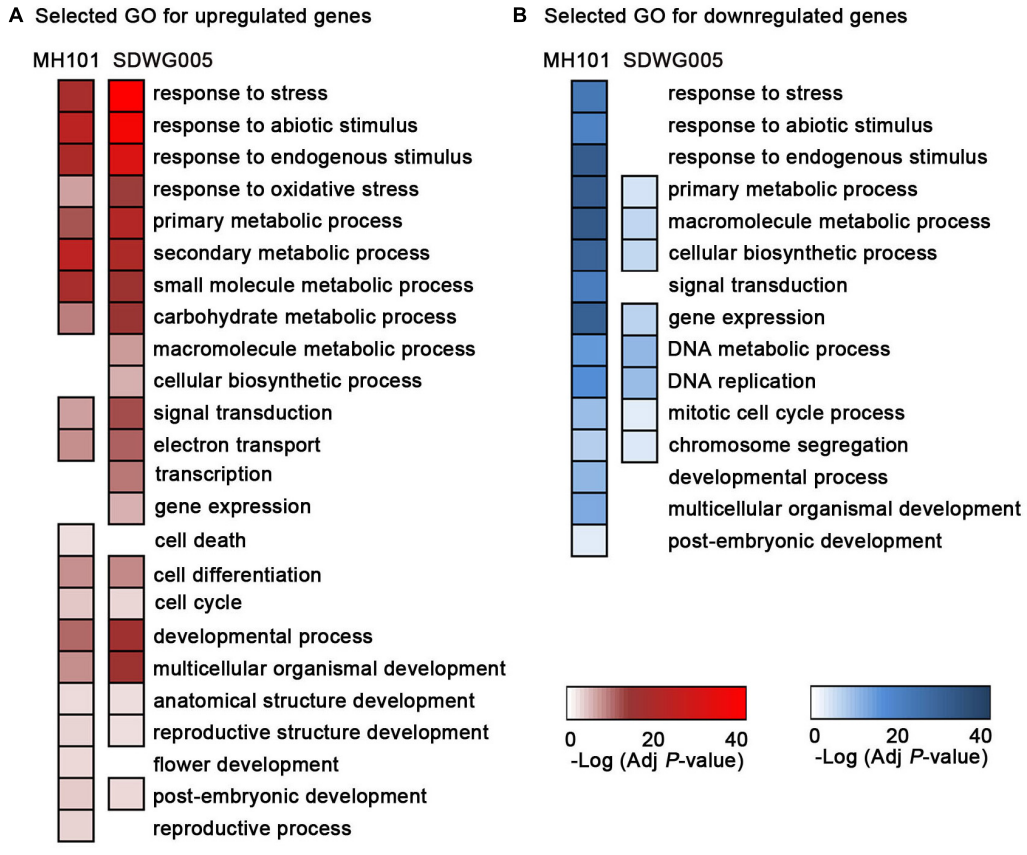


FIGURE 5 | GO enrichment analysis of DEGs for the biological process using AgriGO. Most representative enriched GO categories from the up-regulated DEGs **(A)** or down-regulated DEGs in rice spikelets upon heat stress **(B)**, respectively. Color bars: -$\log_{10}$ transformed *P*-value (adjusted by Benjamini-Hochberg).

development processes were also particularly observed in MH101 (**Figure 5B**). And GO terms for the down-regulated DEGs, including multiple cellular metabolic processes, DNA metabolic process, DNA replication, mitotic cell cycle process, and chromosome segregation, were more significantly enriched in MH101 than in SDWG005 (**Figure 5B**).

We further determined whether the heat-responsive DEGs were involved in specific pathways by searching against the KEGG pathway database. The results showed that among the up-regulated DEGs, six significant KEGG pathways (FDR adjusted *P*-value $\leq$ 5%), including phenylpropanoid biosynthesis, phenylalanine metabolism, flavonoid biosynthesis, metabolic pathways, biosynthesis of secondary metabolites, and cutin, suberin and wax biosynthesis, were collectively enriched in both genotypes, indicating the common mechanisms exist in both tolerant and sensitive varieties that were associated with resistance to heat stress (**Figure 6A** and **Supplementary Table S5**). While only one pathway of DNA replication was collectively enriched in both two genotypes among down-regulated DEGs (**Figure 6B** and **Supplementary Table S5**). Notably, the pathways involved in DNA repair processes, including mismatch repair, base excision repair, nucleotide excision repair, and pyrimidine and purine metabolism, were exclusively enriched in sensitive cultivar MH101, indicating that heat-induced abnormal anther development and deformed pollen grains formation, possibly act through suppressing the DNA replication process and even disturbing DNA repair proofreading to harm genome integrity (**Figure 6B** and **Supplementary Table S5**).

Taken together, these findings suggested both the common and genotype-specific mechanisms existed in both tolerant and sensitive varieties that were associated with resistance to heat stress. In tolerant cultivar SDWG005, the resistance mechanisms were associated with reprogramming the cellular activities, such as response to abiotic stress, metabolic reorganization, and secondary metabolites biosynthesis. In contrast, heat responses in susceptible cultivar MH101 were mainly associated with reproductive development injury, which appeared to act through inhibiting the transcriptional levels of the genes involved in DNA replication and DNA repair proofreading, to impair the processes of post-embryonic and reproductive structure development.

Analysis of DEGs Involved in Lignin and Flavonoids Biosynthetic Pathways

To further investigate the key genes involved in thermotolerance and their regulation pathway, the genes involved in the most significant enriched phenylpropanoid biosynthesis pathway were analyzed. The DEGs involved in phenylpropanoid biosynthesis were screened, and identified 77 DEGs that included almost all the key enzymes related to lignin biosynthesis (**Supplementary Table S6**). In response to heat stress, a dramatic up-regulation of most of these unigenes was observed in both genotypes, especially under T38 treatment (**Figure 7A**). Almost all of the laccase (LAC, 12 unigene) and peroxidase

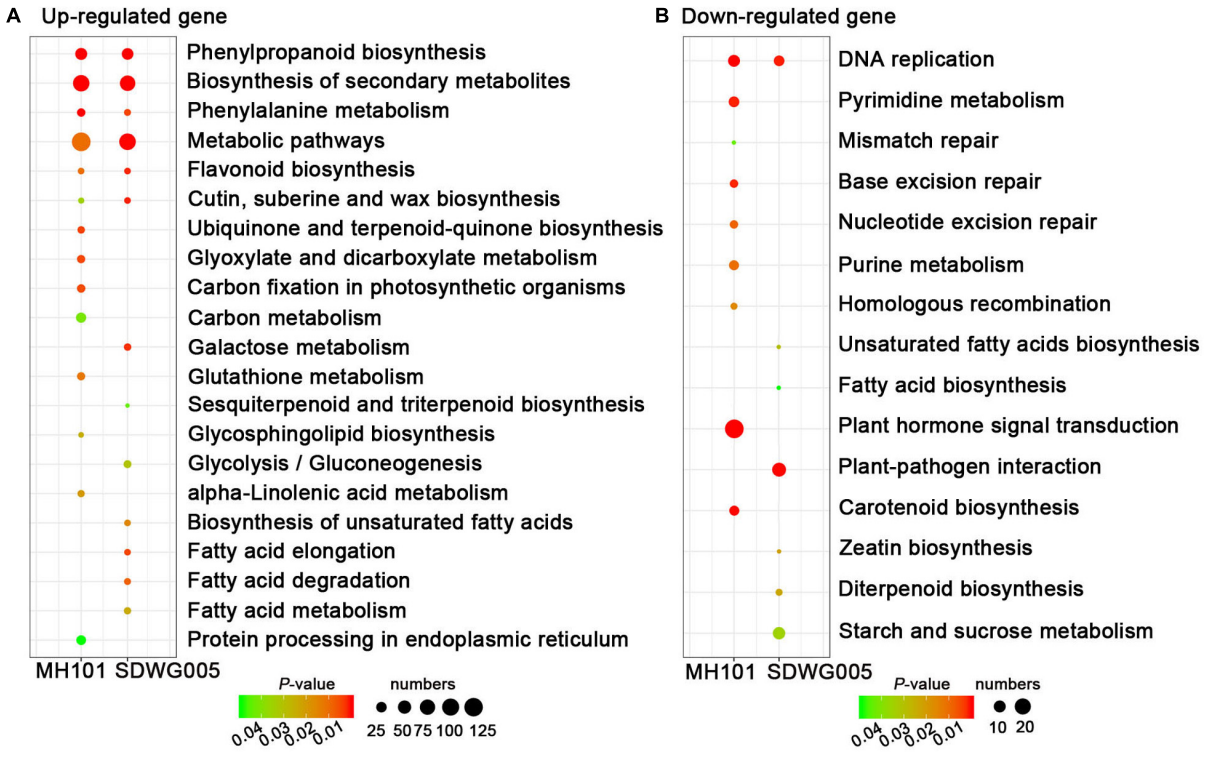


FIGURE 6 | The KEGG pathway enrichment scatter plots of DEGs. KEGG pathway enrichment analyses for up-regulated **(A)** or down-regulated genes **(B)**, respectively. Significance is indicated by color, and the number of DEGs is indicated by the size of a circle.

(PRX, 21 unigenes) family members exhibited a significantly heat-inducible response. Moreover, 11 DEGs were identified involving in flavonoid biosynthesis pathway, another branch of phenylpropane biosynthesis pathway, which was also enriched in KEGG pathway analysis (**Figure 7A** and **Supplementary Table S6**). Except for two chalcone isomerase (CHI) unigenes, 9 of 11 the unigenes were up-regulated upon heat stress, especially for the CHS and F3H unigenes, which exhibited a higher heat-inducible expression under T38 treatment (**Figure 7A**). Whether in lignin biosynthesis or flavonoids biosynthetic, more up-regulated genes were observed in SDWG005 than MH101 upon heat stress.

The DEGs related to the six main heat-responsive TFs were analyzed, such as MYBs, NACs, WRKYs, and bHLHs families, which might be involved in the regulation of flavonoid and lignin biosynthesis pathways (Xu et al., 2015; Yoon et al., 2015; Wessels et al., 2019). A total of 52 MYBs, 49 NACs, 21 WRKYs, 51 ERFs, 12 HSFs, and 39 bHLHs TFs were identified (**Supplementary Table S7**). Results indicated that except for partial bHLH TFs gene family members, the majority gene family members of MYBs, NACs, WRKYs, and ERFs showed significantly regulated under heat stresses (**Supplementary Figure S3**). HSFs encoding heat shock proteins, as direct transcriptional activators of genes regulated by thermal stress, whose expression levels showed higher up-regulated in SDWG005 than in MH101 (**Supplementary Figure S3**). The different expression levels of these TFs genes, especially participated in the regulation of flavonoid and lignin biosynthesis pathways, may be one factor that led to their different thermotolerance in the two genotypes during pollen development upon heat stress.

We then further analyzed the content of oxidative stress marker MDA, anti-oxidants (total phenolics and total flavonoids) and lignin in spikelets under heat stress. We found that lipid peroxidation (MDA concentration) of SDWG005 was always higher than that of MH101 whether under heat stress treatment or not (**Figure 7B**). Although the concentration of total phenolics was significantly accumulated under T38 treatment, no significant difference was observed between SDWG005 and MH101 (**Figure 7C**). While a higher accumulation of total flavonoids were observed in SDWG005 under normal growth conditions and heat stress (**Figure 7D**). Moreover, a more lignin deposition happened in SDWG005 than MH101 under T38 treatment (**Figure 7E**). The transcriptional profile of nine genes (*OsPAL*, *Os4CL*, *OsCCR*, *OsF5H*, *OsCHS*, *OsCHI*, *OsF3H*, and *OsDFR*) involved in lignin and flavonoid biosynthesis pathway were further confirmed by using RT-qPCR. As illustrated in **Figure 7F**, a dramatic upregulation of all these genes in two genotypes was observed in response to high temperature, especially in T38 treatment. Except for *Os4CL* (*Os08g14760*), the induction response of the remaining 8 genes in SDWG005 was higher than in MH101. Collectively, the genotype-dependent difference in biosynthesis and accumulation of lignin and flavonoids might be vital factors that led to their different thermotolerance in the two genotypes, suggesting the lignin and flavonoid biosynthetic pathways might play important roles in rice heat resistance during reproductive stage.

DISCUSSION

With global warming growing more severe, heat stress has profound effects on crop growth and development, even leads to significant yield loss worldwide (Fahad et al., 2017; Wang and Qin, 2017). Despite great advances on physiological and biochemical aspects of the high temperature induced fertility injury in rice, there is limited information regarding the molecular mechanism responding to heat stress (Satake and Yoshida, 1978; Matsui et al., 2000; Jagadish et al., 2010; Yu et al., 2016; Yan et al., 2017). Based on previous reports, SDWG005 and MH101 showed contrasting thermotolerance differences during the flowering stage (Zha et al., 2016). Recently, a transcriptional profile analysis under different time courses of heat treatment on the anthers of SDWG005 at anthesis were conducted and identified agmatine-coumarin-acyltransferase gene *OsACT* may involve in SDWG005 thermotolerance (Liu et al., 2020). Here, we further characterized the physiological observation and transcriptomic expression patterns analysis on spikelets of SDWG005 during meiosis under different high-temperature treatment. We observed higher pollen viability in SDWG005 after heat stress at meiosis stage than in sensitive cultivar MH101 (**Figure 1**). The pollination properties analysis also showed that SDWG005 was much more tolerant to heat stress than MH101. Through comparison among treatments, the number of DEGs was higher in SDWG005 than in MH101 (**Figure 3**), which might indicate that heat-tolerant rice SDWG005 had a more complex response to heat stress than the susceptible rice MH101. Unlike the *aus indica* type rice variety "Nagina22" ("N22") (Bahuguna et al., 2015), SDWG005 was highly heat tolerant but did not come with obvious poor agronomic performances. Hence, identification of key mechanisms of heat stress response in SDWG005 and understanding the molecular basis in such a novel extremely strong heat-resistant rice germplasm will help us to bring its thermotolerance into existing rice breeding programs.

It was reported that reproductive organs undergone both general and specific transcriptional alterations in response to high-temperature changes (Alba et al., 2004; Oshino et al., 2007; Endo et al., 2009; Bita et al., 2011). The general transcriptional changes refer to reprogramming of cellular activities by the activation of a number of genes related to stress perception, downstream signaling transductional and transcriptional controls, while the specific transcriptional changes involve silencing of cell proliferation- and DNA replication-regulatory genes (Abiko et al., 2005; Oshino et al., 2007; Zhang et al., 2012). Our results indicated that both common and genotype-specific mechanisms associated with resistance to heat stress existed in both tolerant and sensitive varieties. In the hierarchical clustering analysis, we observed that many DEGs, related to response to stimulus, biosynthetic process, cellular metabolic process, signaling, and transcription activity, showed the opposite pattern of expression (**Figure 4**), which indicated that heat-tolerant rice SDWG005 possibly established a new steady-state balance of metabolic processes to protect and repair proteins and membranes.

The GO and KEGG analyses further indicated that the up-regulated DEGs involved in multiple metabolic processes,

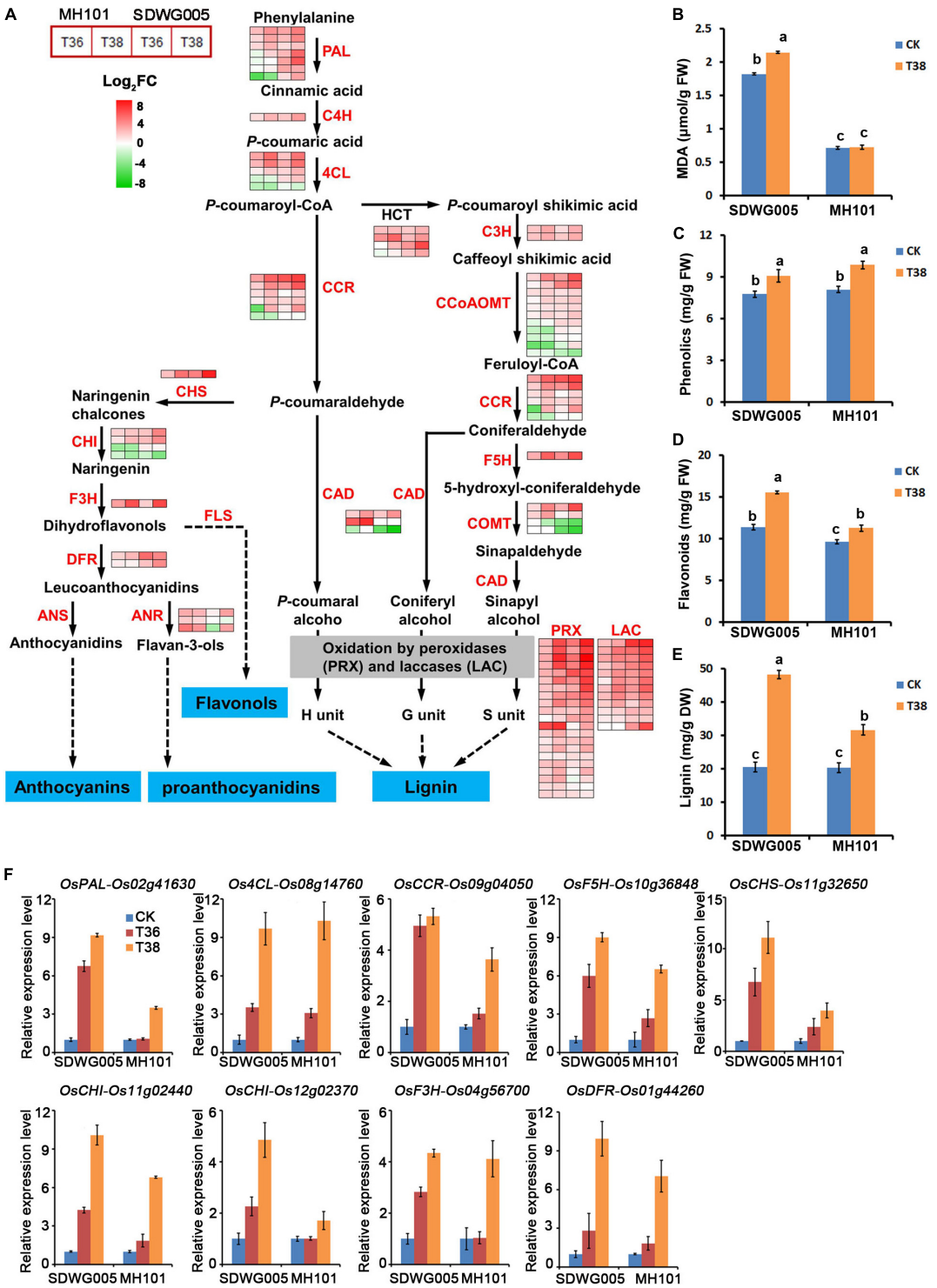


FIGURE 7 | Transcriptional changes of DEGs involved lignin and flavonoids biosynthetic pathways. **(A)** Expression heatmap of representative genes related to the lignin and flavonoids biosynthetic pathways. Data were $\log_2$ transformed and normalized to those of the corresponding normal growth controls to determine fold changes, as indicated by the colored bar. The detailed expression patterns and identities of the genes in each of these biosynthetic pathways are shown in **Supplementary Table S6**. Total content of MDA **(B)**, phenolics **(C)**, flavonoids **(D)**, and lignin **(E)** in spikelets of SDWG005 and MH101 varieties under heat treatment for three days. **(F)** The relative expression levels of *OsPAL*, *Os4CL*, *OsCCR*, *OsF5H*, *OsCHS*, *OsCHI*, *OsF3H*, and *OsDFR* in spikelets of SDWG005 and MH101 were determined by qRT-PCR under T36 or T38 conditions for three days. The *O. sativa* β-tubline gene was used for normalization. Each value represents the mean of three independent observations and SD determined. Significant differences were calculated by One-way ANOVA with Duncan's multiple test (SAS Institute, Inc., Cary, NC, United States). Different letters indicate the statistical difference between samples at $P = 0.05$.

transcription, and gene expression had higher expression in tolerant variety, which suggested that these events of rapid reprogramming of cellular activities possibly performed important roles in conferring thermotolerance (**Figures 5**, **6**). Whereas the enriched down-regulated DEGs involved in DNA replication, mitotic cell cycle process, chromosome segregation, and DNA repair protection mechanism were significantly inhibited in sensitive rice plants (**Figures 5**, **6**). It has reported that the genes that were down-regulated in response to high temperatures were predominantly expressed in tapetal cells, and the tapetum function might be impaired during microsporogenesis under high temperature (Endo et al., 2009). To determine whether the genes in the tapetum were affected under our high temperatures conditions, three known tapetum-specific genes, such as *TIP2*, *TDR* and *MTR*, were chose to examine the expression after exposure to high temperatures by RT-qPCR analysis. *TDR* and *TIP2* play a central role in differentiation, morphogenesis, and degradation of anther tapetum degradation (Li et al., 2006; Fu et al., 2014), and *MTR1* is thought to be involved in controlling the development of sporophytic and reproductive cells in rice (Tan et al., 2012). Compared with the normal growth control, the expression levels of *TDR* and *MTR1* showed slightly induction under T36 heat stress and recovered back to normal under T38 heat stress in SDWG005, while displayed abnormal high temperature-repressed expression patterns in MH101 (**Supplementary Figure S4**). A clear inhibition of *TIP2* was observed in both varieties under T38 heat stress, even though this gene was shown to be slightly induced in SDWG005 under T36 (**Supplementary Figure S4**). This was consistent with what we found previously that the flower development and reproductive process were impaired in susceptible rice MH101 in response to high temperature through GO and KEGG analyses, which indicated that the genotype-specific mechanisms might be responsible for the heat stress-induced pollen abortion and spikelets sterility.

The phenylpropanoid pathway is indispensable to plants, because of its role in the biosynthesis of lignin and the production of many other important compounds, such as the flavonoids, coumarins, and lignans (Vogt, 2010). The pathway begins with three reactions leading to the synthesis of *p*-coumaroyl CoA, the direct precursor for flavonoid or H-lignin biosynthesis, which represents the most important branch point within the central phenylpropanoid biosynthesis in plant. In this study, the phenylpropanoid biosynthesis pathway was the most significantly enriched pathway (**Figure 6A**). We further identified 77 DEGs incorporated all the key enzymes of lignin biosynthesis, and 11 DEGs related to flavonoids biosynthesis genes (**Figure 7A**). Some studies have demonstrated that cell wall remodeling proteins with enzymatic activity were involved in the modification of cell wall components in response to heat stress, which induced heat shock proteins expression and cell-wall remodeling to retain plasma membrane integrity, thus preventing cellular content leakage and conferring thermoprotection (Wu and Jinn, 2010; Wu et al., 2010; Huang et al., 2017). Here, we found that the expression levels of these genes related to lignin biosynthesis were highly induced in SDWG005, and more lignin were deposited in SDWG005 under heat stress (**Figures 7A,E,F**). The results were consistent with the notion that cell wall remodeling had a pronounced effect on thermotolerance, and will provide a foundation for further researches on fine-tuning the regulatory mechanisms to adjust the accumulation of lignin in response to heat stress.

Flavonoids are a class of secondary metabolites that are common in the plant kingdom, which as antioxidants can regulate the ROS levels and early anther development (Thompson et al., 2010; Ning et al., 2018). ROS are described as both signaling molecules and molecules that cause damage to cellular components, and it is generally well-accepted that correctly timed accumulation of ROS is beneficial for tapetum degradation and pollen viability (Satake and Yoshida, 1978; Saini et al., 1984; Zhao et al., 2018). In comparison to MH101, the decline of heat-induced injury of ROS scavenging enzymatic activity was not clearly observed in SDWG005 (**Figure 2**), indicating the ROS scavenging capacity in SDWG005 was better maintained for protecting the sensitive reproductive organs from heat-induced damage under heat stress. Moreover, more flavonoids were accumulated in SDWG005 under heat stress, indicating that the increased level of flavonoids related to the heat thermotolerance in rice (**Figure 7D**). Hence, we deduced that SDWG005 has a stronger defense ability to scavenge ROS molecules efficiently than MH101. The possible reason for the elevated heat tolerance of SDWG005 could be explained by the combination of two factors: Firstly, the increase of flavonoids biosynthesis and its accumulation in SDWG005 might contribute greatly to the enhancements of heat tolerance under heat stress, because of flavonoids being widely considered as non-enzymatic antioxidant molecules in plant cells; Secondly, the increase of flavonoids biosynthesis in SDWG005 may also play a regulatory role in the activation of the antioxidant enzyme system, because the antioxidant enzymes (SOD, CAT and APX) had been widely considered as an important mechanism actively used by plants to detoxify ROS damage. Several previous studies had also reported that the accumulation of flavonoids was closely associated with the activities of SOD and CAT for better antioxidant defense system under stressful environment (Winkel-Shirley, 2002; Dixon et al., 2005; Agati et al., 2011; Zaidi et al., 2019). Thus, the more accumulation of flavonoids in spikelets of SDWG005 may also stimulate the activation of the antioxidant enzyme system, and this occurrence was another important reason for the elevated heat tolerance of SDWG005 under the high temperature exposure at reproductive stage.

CONCLUSION

Overall, heat-responsive transcriptome analysis of the spikelets of SDWG005, a novel extremely strong heat-resistant rice germplasm, illustrated the molecular mechanism of its thermotolerance. Genes involved in regulation of reprogramming the cellular activities, such as response to abiotic stress, metabolic reorganization, and secondary metabolites biosynthesis, were significantly up-regulated in thermotolerant variety with heat treatment. In contrast, in sensitive variety, the down-regulation genes associated with

DNA replication and DNA repair proofreading could cause serious injury to reproductive development when exposed to high temperature during meiosis. Additionally, we found that the lignin and flavonoids levels played an important role in heat-induced defense responses. The results will provide useful information for the research on the mechanism of heat resistance and the application of heat resistance genes in rice breeding.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/, PRJNA633211.

AUTHOR CONTRIBUTIONS

SD and ZC designed and performed the experiments, analyzed the data, and prepared figures and tables. ZC performed the phenotyping experiments. TL, XF, and FH assisted in the phenotyping experiments. SD, XT, and HW wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Engineering Research Center of Ecology and Agricultural Use of Wetland, Ministry of Education (KFT201901). The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.562937/full#supplementary-material

REFERENCES

Abiko, M., Akibayashi, K., Sakata, T., Kimura, M., Kihara, M., Itoh, K., et al. (2005). High-temperature induction of male sterility during barley (*Hordeum vulgare* L.) anther development is mediated by transcriptional inhibition. *Sex. Plant Reprod.* 18, 91–100. doi: 10.1007/s00497-005-0004-2

Agati, G., Biricolti, S., Guidi, L., Ferrini, F., Fini, A., Tattini, M., et al. (2011). The biosynthesis of flavonoids is enhanced similarly by UV radiation and root zone salinity in L. vulgare leaves. *J Plant Physiol.* 168, 204–212. doi: 10.1016/j.jplph.2010.07.016

Alba, R., Fei, Z., Payton, P., Liu, Y., Moore, S. L., Debbie, P., et al. (2004). ESTs, cDNA microarrays, and gene expression profiling: tools for dissecting plant physiology and development. *Plant J.* 39, 697–714. doi: 10.1111/j.1365-313X.2004.02178.x

Bahuguna, R. N., Jha, J., Pal, M., Shah, D., Lawas, L. M., Khetarpal, S., et al. (2015). Physiological and biochemical characterization of NERICA-L-44: a novel source of heat tolerance at the vegetative and reproductive stages in rice. *Physiol. Plant.* 154, 543–559. doi: 10.1111/ppl.12299

Bita, C. E., Zenoni, S., Vriezen, W. H., Mariani, C., Pezzotti, M., and Gerats, T. (2011). Temperature stress differentially modulates transcription in meiotic anthers of heat-tolerant and heat-sensitive tomato plants. *BMC Genomics* 12:384. doi: 10.1186/1471-2164-12-384

Bokszczanin, K. L., Solanaceae Pollen Thermotolerance Initial Training Network (Spot-Itn) Consortium, and Sotirios Fragkostefanakis. (2013). Perspectives on deciphering mechanisms underlying plant heat stress response and thermotolerance. *Front. Plant Sci.* 4:315. doi: 10.3389/fpls.2013.00315

Bonnecarrère, V., Borsani, O., Diaz, P., Capdevielle, F., Blanco, P., and Monza, J. (2011). Response to photoxidative stress induced by cold in japonica rice is genotype dependent. *Plant Sci.* 180, 726–732. doi: 10.1016/j.plantsci.2011.01.023

Dixon, R. A., Xie, D. Y., and Sharma, S. B. (2005). Proanthocyanidins–a final frontier in flavonoid research? *New Phytol.* 165, 9–28. doi: 10.1111/j.1469-8137.2004.01217.x

Djeridane, A., Yousfi, M., Nadjemi, B., Boutassouna, D., Stocker, P., and Vidal, N. (2006). Antioxidant activity of some algerian medicinal plants extracts containing phenolic compounds. *Food Chem.* 97, 654–660. doi: 10.1016/j.foodchem.2005.04.028

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

Du, Z., Zhou, X., Ling, Y., Zhang, Z., and Su, Z. (2010). agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Res.* 38, W64–W70. doi: 10.1093/nar/gkq310

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

Fahad, S., Bajwa, A. A., Nazir, U., Anjum, S. A., Farooq, A., Zohaib, A., et al. (2017). Crop production under drought and heat stress: plant responses and management options. *Front. Plant Sci.* 8:1147. doi: 10.3389/fpls.2017.01147

Frank, G., Pressman, E., Ophir, R., Althan, L., Shaked, R., Freedman, M., et al. (2009). Transcriptional profiling of maturing tomato (*Solanum lycopersicum* L.) microspores reveals the involvement of heat shock proteins, ROS scavengers, hormones, and sugars in the heat stress response. *J. Exp. Bot.* 60, 3891–3908. doi: 10.1093/jxb/erp234

Fu, Z., Yu, J., Cheng, X., Zong, X., Xu, J., Chen, M., et al. (2014). The rice basic helix-loop-helix transcription factor TDR INTERACTING PROTEIN2 is a central switch in early anther development. *Plant Cell* 26, 1512–1524. doi: 10.1105/tpc.114.123745

Giannopolitis, C. N., and Ries, S. K. (1977). Superoxide dismutases: I. Occurrence in higher plants. *Plant Physiol.* 59, 309–314. doi: 10.1104/pp.59.2.309

Guo, J., Wu, J., Ji, Q., Wang, C., Luo, L., Yuan, Y., et al. (2008). Genome-wide analysis of heat shock transcription factor families in rice and *Arabidopsis*. *J. Genet. Genomics* 35, 105–118. doi: 10.1016/S1673-8527(08)60016-8

Huang, Y. C., Wu, H. C., Wang, Y. D., Liu, C. H., Lin, C. C., Luo, D. L., et al. (2017). PECTIN METHYLESTERASE34 contributes to heat tolerance through its role in promoting stomatal movement. *Plant Physiol.* 174, 748–763. doi: 10.1104/pp.17.00335

Jagadish, K. S. V., Craufurd, P., Shi, W. J., and Oane, R. (2014). A phenotypic marker for quantifying heat stress impact during microsporogenesis in rice (*Oryza sativa* L.). *Funct. Plant Biol.* 41, 48–55. doi: 10.1071/Fp13086

Jagadish, S. V., Craufurd, P. Q., and Wheeler, T. R. (2007). High temperature stress and spikelet fertility in rice (*Oryza sativa* L.). *J. Exp. Bot.* 58, 1627–1635. doi: 10.1093/jxb/erm003

Jagadish, S. V., Muthurajan, R., Oane, R., Wheeler, T. R., Heuer, S., Bennett, J., et al. (2010). Physiological and proteomic approaches to address heat tolerance during anthesis in rice (*Oryza sativa* L.). *J. Exp. Bot.* 61, 143–156. doi: 10.1093/jxb/erp289

Kawahara, Y., de la Bastide, M., Hamilton, J. P., Kanamori, H., McCombie, W. R., Ouyang, S., et al. (2013). Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* 6:4. doi: 10.1186/1939-8433-6-4

Kovacik, J., Klejdus, B., Backor, M., and Repcak, M. (2007). Phenylalanine ammonia-lyase activity and phenolic compounds accumulation in nitrogen-deficient *Matricaria chamomilla* leaf rosettes. *Plant Sci.* 172, 393–399. doi: 10.1016/j.plantsci.2006.10.001

Krishnan, R., Ramakrishnan, B., Reddy, K. R., and Reddy, V. R. (2011). High-temperature effects on rice growth, yield, and grain quality. *Adv. Agron.* 111, 87–206. doi: 10.1016/B978-0-12-387689-8.00004-7

Ku, S., Yoon, H., Suh, H. S., and Chung, Y. Y. (2003). Male-sterility of thermosensitive genic male-sterile rice is associated with premature programmed cell death of the tapetum. *Planta* 217, 559–565. doi: 10.1007/s00425-003-1030-7

Li, C., Wang, R. H., Ning, H. S., and Luo, Q. H. (2016). Changes in climate extremes and their impact on wheat yield in Tianshan Mountains region, northwest China. *Environ. Earth Sci.* 75, 1228. doi: 10.1007/s12665-016-6030-6

Li, N., Zhang, D. S., Liu, H. S., Yin, C. S., Li, X. X., Liang, W. Q., et al. (2006). The rice tapetum degeneration retardation gene is required for tapetum degradation and anther development. *Plant Cell* 18, 2999–3014. doi: 10.1105/tpc.106.044107

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

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

Liu, G., Zha, Z., Cai, H., Qin, D., Jia, H., Liu, C., et al. (2020). Dynamic transcriptome analysis of anther response to heat stress during anthesis in thermotolerant rice (*Oryza sativa* L.). *Int. J. Mol. Sci.* 21:1155. doi: 10.3390/ijms21031155

Liu, Y., Mu, Q., Wang, Y., Gao, Y., and Tian, X. H. (2019). Selecting high temperature tolerant germplasms from Africa Rice. *Chin. Agric. Sci. Bull.* 35, 8–12. doi: 10.11924/j.issn.1000-6850.casb18120035

Lobell, D. B., and Gourdji, S. M. (2012). The influence of climate change on global crop productivity. *Plant Physiol.* 160, 1686–1697. doi: 10.1104/pp.112.208298

Maehly, A. C., and Chance, B. (1954). The assay of catalases and peroxidases. *Methods Biochem. Anal.* 1:357. doi: 10.1002/9780470110171.ch14

Matsui, T., Omasa, K., and Horie, T. (2000). High temperature at flowering inhibits swelling of pollen grains, a driving force for thecae dehiscence in rice (*Oryza sativa* L.). *Plant Prod. Sci.* 3, 430–434. doi: 10.1626/pps.3.430

McClung, C. R., and Davis, S. J. (2010). Ambient thermometers in plants: from physiological outputs towards mechanisms of thermal sensing. *Curr. Biol.* 20, R1086–R1092. doi: 10.1016/j.cub.2010.10.035

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

Ning, L., Lin, Z., Gu, J., Gan, L., Li, Y., Wang, H., et al. (2018). The initial deficiency of protein processing and flavonoids biosynthesis were the main mechanisms for the male sterility induced by SX-1 in *Brassica napus*. *BMC Genomics* 19:806. doi: 10.1186/s12864-018-5203-y

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 27, 29–34. doi: 10.1093/nar/27.1.29

Oshino, T., Abiko, M., Saito, R., Ichiishi, E., Endo, M., Kawagishi-Kobayashi, M., et al. (2007). Premature progression of anther early developmental programs accompanied by comprehensive alterations in transcription during high-temperature injury in barley plants. *Mol. Genet. Genomics* 278, 31–42. doi: 10.1007/s00438-007-0229-x

Peever, T. L., and Higgins, V. J. (1989). Electrolyte leakage, lipoxygenase, and lipid peroxidation induced in tomato leaf tissue by specific and nonspecific elicitors from *Cladosporium fulvum*. *Plant Physiol.* 90, 867–875. doi: 10.1104/pp.90.3.867

Peng, S., Huang, J., Sheehy, J. E., Laza, R. C., Visperas, R. M., Zhong, X., et al. (2004). Rice yields decline with higher night temperature from global warming. *Proc. Natl. Acad. Sci. U.S.A.* 101, 9971–9975. doi: 10.1073/pnas.0403720101

Prasad, P. V. V., Boote, K. J., Allen, L. H., Sheehy, J. E., and Thomas, J. M. G. (2006). Species, ecotype and cultivar differences in spikelet fertility and harvest index of rice in response to high temperature stress. *Field Crops Res.* 95, 398–411. doi: 10.1016/j.fcr.2005.04.008

Ramamoorthy, R., Jiang, S. Y., Kumar, N., Venkatesh, P. N., and Ramachandran, S. (2008). A comprehensive transcriptional profiling of the WRKY gene family in rice under various abiotic and phytohormone treatments. *Plant Cell Physiol.* 49, 865–879. doi: 10.1093/pcp/pcn061

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Saini, H. S., Sedgley, M., and Aspinall, D. (1984). Developmental anatomy in wheat of male-sterility induced by heat-stress, water deficit or abscisic-acid. *Aust. J. Plant Physiol.* 11, 243–253. doi: 10.1071/Pp9840243

Satake, T., and Yoshida, S. (1978). High temperature-induced sterility in indica rices at flowering. *Jpn. J. Crop Sci.* 47, 6–17. doi: 10.1626/jcs.47.6

Scafaro, A. P., Haynes, P. A., and Atwell, B. J. (2010). Physiological and molecular changes in *Oryza meridionalis* Ng., a heat-tolerant species of wild rice. *J. Exp. Bot.* 61, 191–202. doi: 10.1093/jxb/erp294

Scharte, J., Schon, H., Tjaden, Z., Weis, E., and von Schaewen, A. (2009). Isoenzyme replacement of glucose-6-phosphate dehydrogenase in the cytosol improves stress tolerance in plants. *Proc. Natl. Acad. Sci. U.S.A.* 106, 8061–8066. doi: 10.1073/pnas.0812902106

Shah, F., Huang, J., Cui, K., Nie, L., Shah, T., Chen, C., et al. (2011). Impact of high-temperature stress on rice plant and its traits related to tolerance. *J. Agric. Sci.* 149, 545–556. doi: 10.1017/S0021859611000360

Suzuki, K., Takeda, H., Tsukaguchi, T., and Egawa, Y. (2001). Ultrastructural study on degeneration of tapetum in anther of snap bean (*Phaseolus vulgaris* L.) under heat stress. *Sex. Plant Reprod.* 13, 293–299. doi: 10.1007/s004970100071

Tan, H., Liang, W., Hu, J., and Zhang, D. (2012). MTR1 encodes a secretory fasciclin glycoprotein required for male reproductive development in rice. *Dev. Cell* 22, 1127–1137. doi: 10.1016/j.devcel.2012.04.011

Tang, R., Zheng, J., Jin, Z., Zhang, D., Huang, Y., and Chen, L. (2007). Possible correlation between high temperature-induced floret sterility and endogenous levels of IAA, GAs and ABA in rice (*Oryza sativa* L.). *Plant Growth Regul.* 54, 37–43. doi: 10.1007/s10725-007-9225-8

Thompson, E. P., Wilkins, C., Demidchik, V., Davies, J. M., and Glover, B. J. (2010). An Arabidopsis flavonoid transporter is required for anther dehiscence and pollen development. *J. Exp. Bot.* 61, 439–451. doi: 10.1093/jxb/erp312

Tian, X., Luo, H., Zhou, H., and Wu, C. (2009). Research on heat stress of rice in China: progress and prospect. *Chin. Agric. Sci. Bull.* 25, 166–168.

Vogt, T. (2010). Phenylpropanoid biosynthesis. *Mol. Plant* 3, 2–20. doi: 10.1093/mp/ssp106

Wang, D. P., Laza, M. R. C., Cassman, K. G., Huang, J. L., Nie, L. X., Ling, X. X., et al. (2016). Temperature explains the yield difference of double-season rice between tropical and subtropical environments. *Field Crops Res.* 198, 303–311. doi: 10.1016/j.fcr.2016.05.008

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

Wessels, B., Seyfferth, C., Escamez, S., Vain, T., Antos, K., Vahala, J., et al. (2019). An AP2/ERF transcription factor ERF139 coordinates xylem cell expansion and secondary cell wall deposition. *New Phytol.* 224, 1585–1599. doi: 10.1111/nph.15960

Winkel-Shirley, B. (2002). Biosynthesis of flavonoids and effects of stress. *Curr. Opin. Plant Biol.* 5, 218–223. doi: 10.1016/s1369-5266(02)00256-x

Wu, H. C., Hsu, S. F., Luo, D. L., Chen, S. J., Huang, W. D., Lur, H. S., et al. (2010). Recovery of heat shock-triggered released apoplastic Ca^{2+} accompanied by pectin methylesterase activity is required for thermotolerance in soybean seedlings. *J. Exp. Bot.* 61, 2843–2852. doi: 10.1093/jxb/erq121

Wu, H. C., and Jinn, T. L. (2010). Heat shock-triggered Ca^{2+} mobilization accompanied by pectin methylesterase activity and cytosolic Ca^{2+} oscillation are crucial for plant thermotolerance. *Plant Signal. Behav.* 5, 1252–1256. doi: 10.4161/psb.5.10.12607

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

Yan, H., Zhang, B., Zhang, Y., Chen, X., Xiong, H., Matsui, T., et al. (2017). High temperature induced glume closure resulted in lower fertility in hybrid rice seed production. *Front. Plant Sci.* 7:1960. doi: 10.3389/fpls.2016.01960

Yang, Z., Zhang, Z., Zhang, T., Fahad, S., Cui, K., Nie, L., et al. (2017). The effect of season-long temperature increases on rice cultivars grown in the central and southern regions of China. *Front. Plant Sci.* 8:1908. doi: 10.3389/fpls.2017.01908

Yoon, J., Choi, H., and An, G. (2015). Roles of lignin biosynthesis and regulatory genes in plant development. *J. Integr. Plant Biol.* 57, 902–912. doi: 10.1111/jipb.12422

Yu, C. Y., Xu, X. F., Ge, J., Guo, Y. F., Dong, J. G., and Dong, Z. S. (2016). Premature breakdown of tapetum associated with reverse thermo-sensitive genic male-sterile line Huiyou50S in rapeseed (*Brassica napus*). *Acta Physiol. Plant.* 38:54. doi: 10.1007/s11738-015-2039-9

Yu, Y. Q., Huang, Y., and Zhang, W. (2012). Changes in rice yields in China since 1980 associated with cultivar improvement, climate and crop management. *Field Crops Res.* 136, 65–75. doi: 10.1016/j.fcr.2012.07.021

Zaidi, S. H. R., Zakari, S. A., Zhao, Q., Khan, A. R., Shah, J. M., and Cheng, F. (2019). Anthocyanin accumulation in black kernel mutant rice and its contribution to ROS detoxification in response to high temperature at the filling stage. *Antioxidants* 8:510. doi: 10.3390/antiox8110510

Zha, Z., Yin, D., Wan, B., and Jiao, C. (2016). Analyzing of rice germplasm heat resistance during flowering stage. *Hubei Agric. Sci.* 55, 17–19.

Zhang, X., Li, J., Liu, A., Zou, J., Zhou, X., Xiang, J., et al. (2012). Expression profile in rice panicle: insights into heat response mechanism at reproductive stage. *PLoS One* 7:e49652. doi: 10.1371/journal.pone.0049652

Zhao, Q., Zhou, L., Liu, J., Du, X., Asad, M. A., Huang, F., et al. (2018). Relationship of ROS accumulation and superoxide dismutase isozymes in developing anther with floret fertility of rice under heat stress. *Plant Physiol. Biochem.* 122, 90–101. doi: 10.1016/j.plaphy.2017.11.009

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.

Copyright © 2020 Cai, He, Feng, Liang, Wang, Ding and Tian. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 25 September 2020
doi: 10.3389/fgene.2020.570156

Genome-Wide Identification and Characterization of the bHLH Transcription Factor Family in Pepper (*Capsicum annuum* L.)

Zhishuo Zhang[1,2], Juan Chen[2], Chengliang Liang[2], Feng Liu[2], Xilin Hou[1]* and Xuexiao Zou[1,2,3]**

[1] State Key Laboratory of Crop Genetics and Germplasm Enhancement, Key Laboratory of Biology and Germplasm Enhancement of Horticultural Crops in East China, Ministry of Agriculture, Nanjing Agricultural University, Nanjing, China, [2] Hunan Vegetable Research Institute, Changsha, China, [3] College of Horticulture, Hunan Agricultural University, Changsha, China

OPEN ACCESS

Edited by:
Wei Chen,
North China University of Science and Technology, China

Reviewed by:
Xiyin Wang,
North China University of Science and Technology, China
Peng Wu,
Yangzhou University, China

***Correspondence:**
Feng Liu
liufengrich@126.com
Xilin Hou
hxl@njau.edu.cn
Xuexiao Zou
zouxuexiao428@163.com

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

Received: 06 June 2020
Accepted: 03 September 2020
Published: 25 September 2020

Citation:
Zhang Z, Chen J, Liang C, Liu F, Hou X and Zou X (2020) Genome-Wide Identification and Characterization of the bHLH Transcription Factor Family in Pepper (Capsicum annuum L.). Front. Genet. 11:570156. doi: 10.3389/fgene.2020.570156

Plant basic helix–loop–helix (bHLH) transcription factors are involved in the regulation of various biological processes in plant growth, development, and stress response. However, members of this important transcription factor family have not been systematically identified and analyzed in pepper (*Capsicum annuum* L.). In this study, we identified 122 *CabHLH* genes in the pepper genome and renamed them based on their chromosomal locations. *CabHLHs* were divided into 21 subfamilies according to their phylogenetic relationships, and genes from the same subfamily had similar motif compositions and gene structures. Sixteen pairs of tandem and segmental duplicated genes were detected in the CabHLH family. *Cis*-elements identification and expression analysis of the *CabHLHs* revealed that they may be involved in plant development and stress responses. This study is the first comprehensive analysis of the *CabHLH* genes and will serve as a reference for further characterization of their molecular functions.

Keywords: genome-wide, pepper, bHLH transcription factor family, phylogenetic relationships, expression analysis

INTRODUCTION

The bHLH transcription factor (TF) family, named for its basic helix–loop–helix (bHLH) structure, is the second largest class of TFs and is widely distributed in animals, plants, and microorganisms (Guo et al., 2008). The bHLH domain consists of approximately 60 amino acids and is divided into a basic amino acid region and a helix–loop–helix region (Toledo-Ortiz et al., 2003). The basic region is located on the *N*-terminal side of the bHLH domain and is approximately 15 amino acids in length. These amino acids are mainly responsible for binding to *cis*-elements in DNA. The HLH region is located on the *C*-terminal side of the domain, consists of approximately 40 amino acids, and promotes the formation of homo- and heterodimer complexes (Murre et al., 1989; Ferre-D'Amare et al., 1994).

According to their evolutionary relationships, DNA binding abilities, and functional characteristics, bHLH proteins in animals have been divided into six groups, A–F (Atchley and Fitch, 1997). Many of the plant bHLH proteins that have been identified belong to Group B (Sailsbery and Dean, 2012). According to classification criteria developed in animals, the 133 *bHLH*

genes found in *Arabidopsis thaliana* have been divided into 12 subfamilies based on conserved amino acids at specific positions and on the presence or absence of additional conserved domains (Heim et al., 2003). Fourteen new bHLH TFs were subsequently discovered and further divided among 21 subfamilies, but this classification was limited to higher terrestrial plants (Toledo-Ortiz et al., 2003). As more family members were identified in species such as moss and seaweed, bHLH TFs were subdivided into 32 subfamilies (Carretero-Paulet et al., 2010). At present, the classification of the plant bHLH TF family is not clearly defined, and there are no corresponding names for each subfamily across species.

In plants, bHLH TFs are involved in various signal transduction and anabolic pathways, such as light signal transduction (Duek and Fankhauser, 2003; Castillon et al., 2007), anthocyanin synthesis (Ludwig et al., 1989; Goodrich et al., 1992; Quattrocchio et al., 1993; Nesi et al., 2000; Sakamoto et al., 2001), tryptophan synthesis (Smolen et al., 2002) and gibberellin synthesis (Arnaud et al., 2010). They also modulate stress responses, including responses to low temperature (LTR) (Feng et al., 2012; Lin et al., 2014; Xu et al., 2014), heat (Ko et al., 2009), drought (Li et al., 2007; Seo et al., 2011), and salt (Murre et al., 1989; Zhou et al., 2009; Liu et al., 2014). For example, *Arabidopsis* TT8 (*AtbHLH42*) typically regulates the synthesis of anthocyanins and procyanidins of vegetative organs through the formation of MYB-bHLH-WD (MBW) complexes, specifically the TT2-TT8-TTG1 complex (Baudry et al., 2004). *AtNIG1* was the first bHLH TF shown to be involved in the plant salt stress signaling pathway, and *Arabidopsis* atnig1-1 knockout mutants are significantly more sensitive to salt stress than wild-type plants (Kim and Kim, 2006). *AtbHLH122* overexpressing plants have stronger salt and anti-osmotic capacity than wild-type plants (Liu et al., 2014). The expression level of *AtbHLH92* is upregulated under NaCl, drought and cold stress (Jiang et al., 2009). In rice, *OsbHLH148* regulates the expression of genes associated with jasmonic acid signaling and serves as an initial response factor during drought stress, thereby participating in both drought and trauma responses (Kiribuchi et al., 2004). *OsbHLH1* can be specifically induced by cold stress, but is not induced by salt, PEG and ABA (Wang et al., 2003). *OsHLH006* participates in drought and wounding responses through the jasmonic acid signaling pathway (Kiribuchi et al., 2005).

Pepper (*Capsicum annuum* L.) is an economically important vegetable and the most widely grown cooking ingredient in the world. With the completion of the pepper genome sequence (Qin et al., 2014), genome-wide identification and classification of gene families can be performed to study genes that are critical for pepper growth and development. To date, a number of TF families have been characterized in pepper, such as the Dof (Wu et al., 2016) and Hsp70 (Guo et al., 2016) families. However, the pepper bHLH family has not been characterized previously. Here, we use a bioinformatics approach to identify and characterize members of the bHLH family in pepper. We report basic information about each gene, including its conserved domains, evolutionary relationships, chromosomal location, expression in various pepper tissues, and response to abiotic stress. These data provide a reference for further exploration of the molecular functions of *bHLH* genes in regulating pepper growth and stress responses.

MATERIALS AND METHODS

Identification of the bHLH Gene Family

Annotated sequences of pepper and tomato genes were downloaded from the Solanaceae Genomics Network[1], and annotated sequences of *Arabidopsis bHLHs* were obtained from TAIR[2]. We used HMMER 3.0 (Eddy, 1998) to identify *Arabidopsis*, tomato and pepper sequences that contained the complete bHLH domain (PF00010), using an E-value $< 1e^{-5}$ threshold. Candidate sequences were verified using the SMART[3] and NCBI databases[5]. Sequences with confirmed bHLH domains were retained for further analysis.

Phylogenetic Analysis and Classification of the *CabHLH* Gene Family

The sequences of the CabHLH and AtbHLH proteins were extracted, and a multiple alignment of the sequences was performed using ClustalW 2.0 (Larkin et al., 2007). A phylogenetic tree was constructed in MEGA 7.0 using the neighbor joining (NJ) method (Tamura et al., 2007) with the following parameters: 1,000 bootstrap replicates, Poisson model, and pairwise deletion. *CabHLHs* were placed into subfamilies based on the classification of closely related *AtbHLHs* and the bootstrap support values at relevant nodes.

Protein Properties, Conserved Motifs and Gene Structures

CabHLH protein sequences were uploaded to the ExPASy website[6] to calculate their molecular weights (MW) and isoelectric points (pI). MEME tools[7] v5.1.1 (Bailey et al., 2009) were used to identify up to ten conserved motifs in each CabHLH protein with an optimal motif width of 10–200 residues and all other parameters set to their default values. Intron locations were determined based on the GFF3 files of *Arabidopsis*, pepper and tomato sequences. Gene structures were drawn using TBtools v0.66833 (Chen et al., 2020).

Chromosomal Mapping and Gene Duplication Analysis

The chromosomal positions of the *CabHLH* genes were obtained from the gene annotation file and visualized using MapGene2Chromosome[8] v2. Within a genome, homologous gene pairs located within 100 kb on the same chromosome were considered to be tandem duplicates, whereas blocks of

[1]https://solgenomics.net/
[2]https://www.Arabidopsis.org/
[3,4]http://smart.embl-heidelberg.de/
[5]http://www.ncbi.nlm.nih.gov/
[6]http://web.expasy.org/
[7]http://meme-suite.org
[8]http://mg2c.iask.in/mg2c_v2.0/

genes copied from one region to another were considered to be segmental duplications (Tang et al., 2008; Liu et al., 2011). Segmental and tandem duplicated gene pairs within the pepper genome and collinear gene pairs among the pepper, tomato and *Arabidopsis* genomes were identified using MCScanX with a match score of 50, a match size of 5, a gap score of -3, and an E-value of $1e^{-05}$ (Wang et al., 2012). The non-synonymous substitution rate (Ka) and synonymous substitution rate (Ks) were calculated using KaKs_Calculator 2.0 (Wang et al., 2010), and a collinearity map was drawn with Circos software (Krzywinski et al., 2009).

Analysis of *cis*-Regulatory Elements

SeqKit v0.13.0 (Shen et al., 2016) was used to extract the promoter sequences of each *CabHLH* gene from the pepper genome file, 2000 bp upstream of the ATG start codon. Promoters were uploaded to the PlantCARE website[9] (Lescot et al., 2002) to predict their *cis*-elements.

Expression Analysis of the *CabHLH* Genes

RNA-Seq data were used to examine the expression of *CabHLH* genes in multiple tissues and in response to various abiotic stress treatments (Liu et al., 2017). The expression level of each gene was calculated as FPKM (fragments per kilobase of transcript per million mapped reads), transformed as $\log_2$ (FPKM + 1). Finally, expression heatmaps were generated in R v3.6.1.

Seeds of the pepper cultivar "6421," which exhibits good heat, drought, and disease tolerance, were obtained from the Vegetable Institute of the Hunan Academy of Agricultural Sciences. Plants were grown using the substrate floating seedling method at 24/16°C with a 16 h light/8 h dark photoperiod. Following a previously published treatment protocol (Liu et al., 2017), 40-day-old replicate pepper seedlings were exposed to 200 mM NaCl (salt stress), 400 mM mannitol (drought), 10°C (cold stress), or 42°C (heat stress). Salt stress was imposed by adding NaCl to a final concentration of 200 mM in the nutrient solution, and drought stress was applied by adding D-mannitol to a final concentration of 400 mM. For heat and cold stress treatments, the seedlings were transferred to a growth chamber at 42 or 10°C, and the illumination, photoperiod, and relative humidity were identical to those in the control treatment. Leaf tissue of treated and control plants was sampled at four time points 1, 6, 12, and 24 h after treatment initiation. Samples of treated and control plants were harvested at 7:00, 12:00, and 18:00 h on the first day and at 6:00 h on the following day. Three seedlings were randomly selected and combined to create one biological replicate, and three biological replicates were collected for each treatment and time point. Samples were frozen in liquid nitrogen and stored at -80°C until further use.

Total RNA was extracted from frozen leaf samples using an RNA kit (TaKaRa, Dalian, China) and reverse transcribed into cDNA with a PrimeScript RT reagent kit (TaKaRa). The SYBR Premix Ex Taq kit (TaKaRa) was used to measure relative gene expression levels following the manufacturer's instructions of the One-step Real-Time PCR System Time PCR Detection System (Applied Biosystems, Foster City, CA, United States). The cycling steps were 94°C for 30 s, 94°C for 10 s for 40 cycles, and 58°C for 30 s, followed by melting curve analysis at 65°C for 10 s for 61 cycles. The relative expression levels of selected genes were calculated using the $2^{-\Delta\Delta Ct}$ method (Schmittgen and Livak, 2008).

[9]http://bioinformatics.psb.ugent.be/webtools/plantcare/html/

RESULTS

Identification, Phylogenetic Analysis, Classification and Protein Properties of the *CabHLH* Genes

We used HMMER 3.0 to search for bHLH domains (PF00010) in the pepper and tomato proteins using an E-value threshold of $<1e^{-5}$. All candidate sequences were filtered with NCBI and SMART to further confirm that they contained complete bHLH domains. A total of 122 *CabHLH*s and 140 *SlbHLHs* were identified (**Supplementary Table S1**); the pepper *bHLH* genes were named *CabHLH1* to *CabHLH122* based on their arrangement on the pepper chromosomes. We constructed a phylogenetic tree of CabHLH and AtbHLH proteins in order to investigate their evolutionary relationships and to classify the *CabHLH*s into 21 established subfamilies according to the classifications of their *Arabidopsis* homologs (Li et al., 2006). Subfamily VII had the largest number of members in pepper (14 genes), whereas subfamilies IIIf and VIIIa had the fewest (one gene each) (**Figure 1**). Compared with *Arabidopsis*, pepper had no members of the XV subfamily but contained a unique X subfamily. In many cases, *Arabidopsis* and pepper had different numbers of genes in a given subfamily.

Comprehensive information on the *CabHLH* genes, including locus names, gene positions, protein lengths, exon numbers, molecular weights (MW), and isoelectric points (pI), is provided in **Supplementary Table S2**. The CabHLH proteins range in size from 117 (*CabHLH21*) to 633 (*CabHLH71*) amino acids, with an average length of approximately 344 amino acids. The MWs of the CabHLH proteins range from 12.9 kDa (*CabHLH31*) to 69.3 kDa (*CabHLH2*), and their pIs range from 4.6 (*CabHLH48*) to 10.32 (*CabHLH108*). The *CabHLH* genes contain 1 to 10 exons, highlighting the diversity of their structures.

Chromosome Locations and Duplication Analysis of the *CabHLH* Genes

A total of 111 *CabHLH* genes were located on 12 chromosomes (91%) (**Figure 2**), and the other 11 genes were mapped to scaffolds (*CabHLH112-122*). *CabHLH* genes were unevenly distributed on the 12 chromosomes, with the largest number of *CabHLH*s on chromosome 1 (17 genes) and the smallest number on chromosomes 5 and 7 (five genes each). Chromosomes 3, 8, 2, 10, 11, 6, 4, 12, and 9 contained 16, 14, 10, 10, 8, 7, 7, 6, and 6 *CabHLH* genes, respectively. *CabHLHs* from each subfamily were also unevenly distributed among the chromosomes, and most *CabHLH*s were clustered at the ends of the chromosomes.

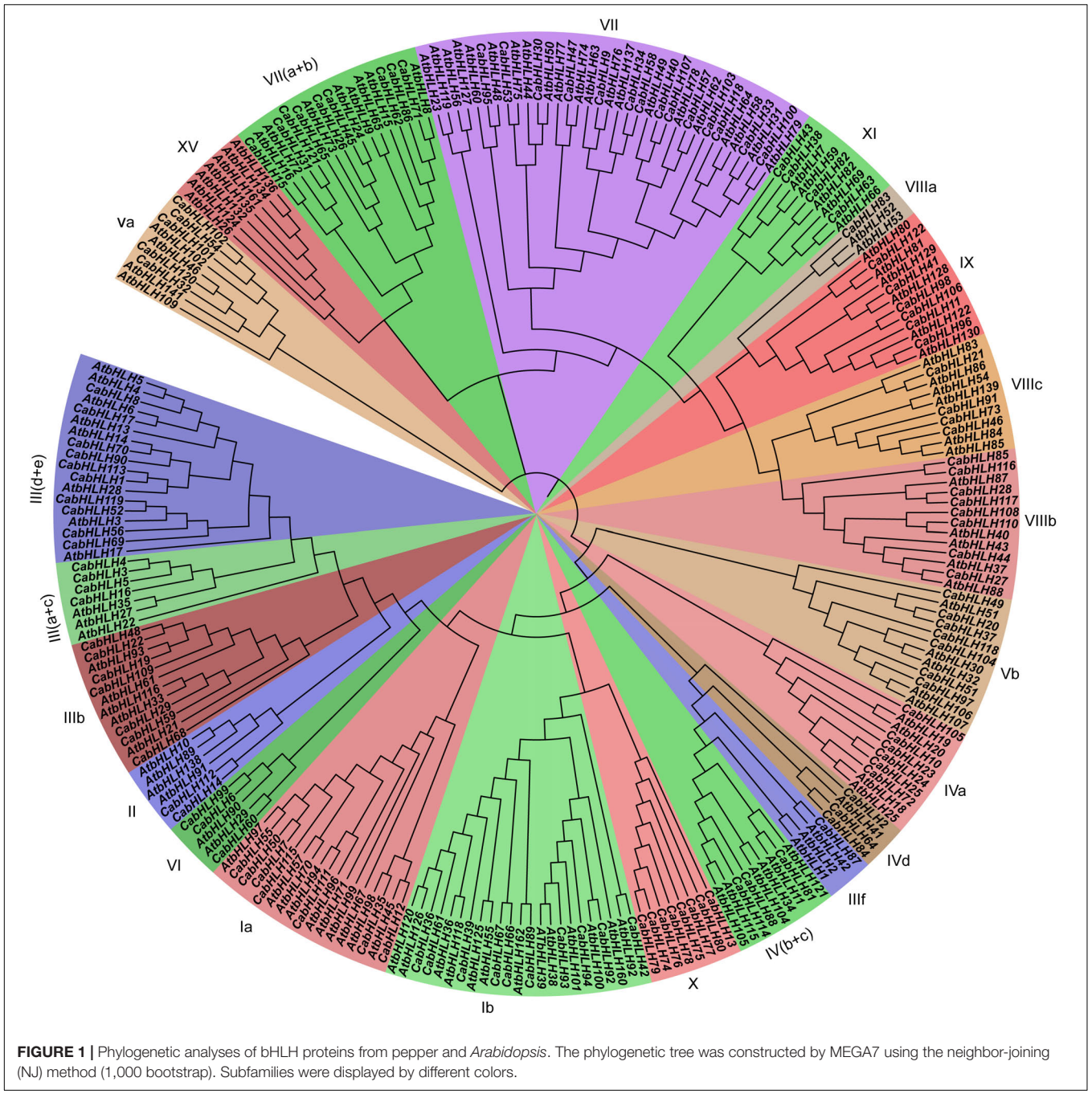


FIGURE 1 | Phylogenetic analyses of bHLH proteins from pepper and *Arabidopsis*. The phylogenetic tree was constructed by MEGA7 using the neighbor-joining (NJ) method (1,000 bootstrap). Subfamilies were displayed by different colors.

We analyzed the locations of *CabHLH* duplicates in the *C. annuum* genome, as tandem duplicates and segmental duplicates play an important role in the expansion of gene families and the generation of new gene functions. As shown in **Figure 2**, we identified six pairs of tandem duplicates on chromosomes 2, 8, 10, and 11: *CabHLH24/25*, *CabHLH75/76*, *CabHLH76/77*, *CabHLH79/80*, *CabHLH93/94*, and *CabHLH101/102*. In addition, ten gene pairs (*CabHLH8/17*, *CabHLH19/22*, *CabHLH22/48*, *CabHLH34/58*, *CabHLH29/59*, *CabHLH46/73*, *CabHLH92/93*, *CabHLH27/108*, *CabHLH44/108*, *CabHLH48/109*) appeared to have arisen through segmental duplication (**Figure 3**). To determine the order of these duplication events, we used the synonymous substitution rate (Ks) to estimate the duplicate divergence times. All segmental duplicates had larger Ks values (0.83–2.32) than did tandem duplicates (0.11–0.77), indicating that they had a relatively earlier origin (**Table 1**). In addition, the Ks values of *CabHLH46/73* and *CabHLH27/108* were significantly greater than those of other segmental duplicates, showing that they derived from more ancient duplication events.

To further explore the evolutionary relationships among bHLH TFs from different species, we constructed a collinearity

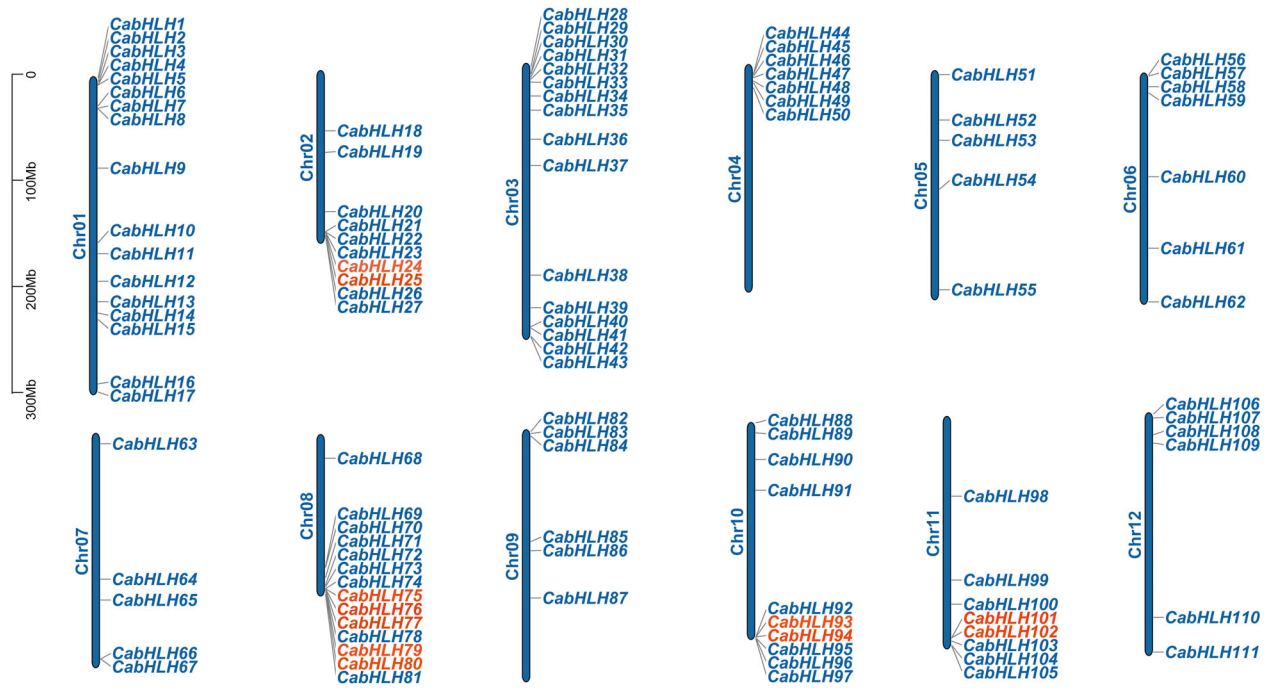


FIGURE 2 | Chromosomal locations of *CabHLH* genes. Tandem duplicated genes are marked in red.

plot of the pepper, tomato, and *Arabidopsis* bHLH gene families (**Figure 3**). A total of 117, 64, and 105 collinear gene pairs were identified between pepper and tomato, pepper and *Arabidopsis*, and tomato and *Arabidopsis*, respectively, indicating that significant expansion of the gene family had occurred before divergence of the three species (**Supplementary Table S3**). For example, 44 *CabHLH*s and 54 *AtbHLH*s had a collinear relationship, and most such relationships were one-to-one matches such as *CabHLH2*/*AtbHLH2* and *CabHLH12*/*AtbHLH45*. There were also one-to-many matches, such as *CabHLH17*/(*AtbHLH4, AtbHLH5, AtbHLH6*) and *CabHLH23*/(*AtbHLH18, AtbHLH25*). Many-to-one cases also existed, such as (*CabHLH6, CabHLH8, CabHLH17*)/*AtbHLH4* and (*CabHLH27, CaHLH108, CabHLH44*)/*AtbHLH88*. These results indicate that *bHLH*s are relatively conserved and that collinear *bHLH*s between species may originate from the same ancestor.

Gene Structure and Motif Analysis of CabHLH Family

Conserved motifs of the CabHLH proteins were analyzed using MEME tools, and ten conserved motifs from 26 to 154 amino acids in length were identified (**Supplementary Figure S1**). The number of conserved motifs in each CabHLH protein varied from one to five (**Figure 4**). Each subfamily contained several common motifs, while few subfamilies possessed unique motifs. For example, motifs 1 and 2 were present in almost all CabHLH proteins and represented the position of the bHLH domain, whereas motifs 9 and 10 were only found in subfamilies III (a + c) and VII, respectively, and may be related to unique functions of individual subfamilies. CabHLH proteins from the same subfamily exhibited similar motifs, suggesting that they may also share a degree of functional similarity. The diversity of motifs in different subfamilies suggests that *CabHLH* functions have tended to diversify during evolution.

We used TBtools to map the structures of the pepper, tomato and *Arabidopsis bHLH* genes (**Figure 4**, **Supplementary Figure S2**) and found that most *bHLHs* from the same subfamily shared similar gene structures. For example, subfamily III (d + e) contains 0–2 introns, subfamily IX has 4–6 introns. Intron gain and loss is a frequent phenomenon during evolution and can increase the complexity of gene structures (Roy and Gilbert, 2005). In the *CabHLHs*, most tandem duplicates (5/6) had different numbers of introns, whereas most segmental duplicates (7/10) had the same number of introns, suggesting that tandem duplicates may have undergone greater divergence in gene function over the course of evolution. In addition, we also analyzed the introns of collinear *bHLH* pairs. There were 53, 31, and 46 pairs of collinear *bHLH* pairs with different numbers of introns between pepper and tomato, pepper and *Arabidopsis*, and tomato and *Arabidopsis*, respectively, indicating that the functions of these collinear genes may have undergone a degree of differentiation (**Supplementary Table S3**).

Cis-Element Analysis of the *CabHLH* Genes

We extracted the 2,000 bp upstream promoter sequences of the *CabHLH* genes for *cis*-element analysis using the PlantCARE database. Ten common *cis*-elements were identified (**Supplementary Table S4**), and 119 *CabHLHs* contained at least

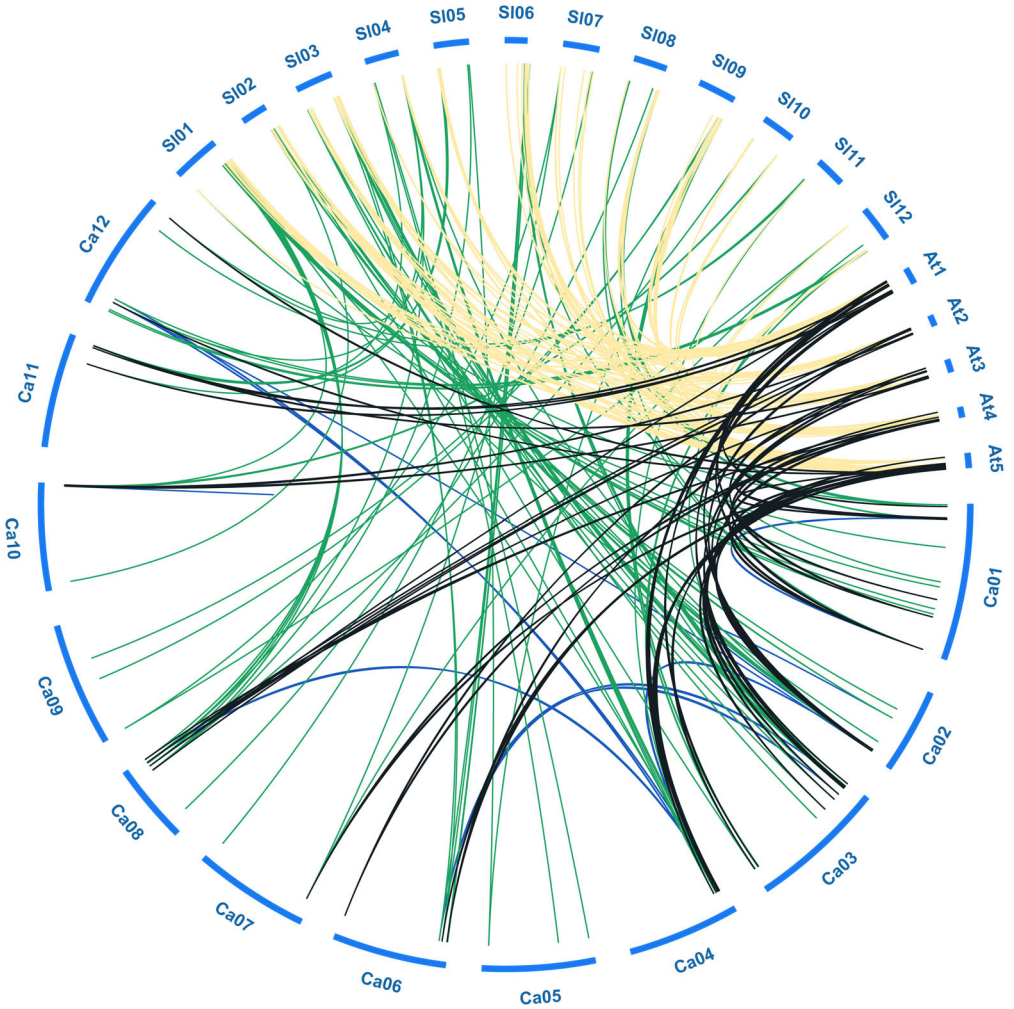


FIGURE 3 | Collinear analysis of *bHLH* genes among pepper (Ca), tomato (Sl), and *Arabidopsis* (At). Green, black and yellow lines represent the collinear gene pairs between pepper and tomato, pepper and *Arabidopsis*, tomato and *Arabidopsis* chromosomes, respectively. Blue lines indicate the segmental duplicated *bHLH* genes in pepper.

one *cis*-element. The ABRE, CGTCA-motif, and GARE-motif elements respond to ABA, JA, and GA stimulation, These motifs were present in the promoters of 86, 77, and 27 *CabHLHs*, suggesting that the expression of these genes responds to levels of the corresponding hormones. Two light-responsive elements (G-box and Sp1) that are ubiquitous in plants were identified in 90 and 12 *CabHLHs*, respectively. Stress-responsive cis-elements, including those associated with LTR, defense and stress (TC-rich repeats), drought (MBS), and anaerobic induction (ARE), were identified in the promoters of 33, 42, 52, and 27 *CabHLHs*, respectively. Diverse response elements indicate the importance of *CabHLHs* in stress responses.

Expression Patterns of *CabHLH* Genes in Various Tissues

We obtained the expression data of *CabHLH* genes from previous research (Liu et al., 2017) and removed 22 *CabHLH* genes with FPKM values of less than one in all tissues (Zhuo et al., 2018). An expression heatmap was created using the remaining 100 genes (**Figure 5**, **Supplementary Table S5**). Most *CabHLH*s differed in their expression patterns, although a few showed similar expression patterns. Some *CabHLH*s (such as *CabHLH100, CabHLH11, CabHLH8,* and *CabHLH43*) showed high expression levels (FPKM > 10) in most tissues analyzed, whereas other *CabHLH*s (such as *CabHLH23, CabHLH85, CabHLH39, CabHLH105*, and *CabHLH108*) were not expressed in any tissues. In addition, several *CabHLHs* showed extremely high expression in specific tissues, such as *CabHLH33*/*CabHLH100* in flower buds, *CabHLH33* in petals, and *CabHLH42* in the placenta. We obtained transcriptome data for *CabHLH33* and *CabHLH100* from another study and found that their expression was also significantly higher in flowers or flower buds than in any other tissues (**Figure 6**) (Qin et al., 2014). These genes may therefore have specific roles in flower development. We also analyzed the expression of duplicated genes in various tissues and found

TABLE 1 | Calculation of Ka and Ks ratios of 16 duplicated *CabHLH* gene pairs.

Genes	Ka	Ks	Duplication Type
CabHLH24/CabHLH25	0.2381	0.7698	Tandem
CabHLH75/CabHLH76	0.1017	0.1502	Tandem
CabHLH76/CabHLH77	0.0814	0.1058	Tandem
CabHLH79/CabHLH80	0.1069	0.2141	Tandem
CabHLH93/CabHLH94	0.1550	0.2800	Tandem
CabHLH101/CabHLH102	0.1081	0.1526	Tandem
CabHLH8/CabHLH17	0.1546	1.0551	Segmental
CabHLH19/CabHLH22	0.2626	1.0137	Segmental
CabHLH22/CabHLH48	0.4339	1.3220	Segmental
CabHLH34/CabHLH58	0.2341	0.8915	Segmental
CabHLH29/CabHLH59	0.2142	0.9823	Segmental
CabHLH46/CabHLH73	0.4674	2.3151	Segmental
CabHLH92/CabHLH93	0.4884	0.9128	Segmental
CabHLH27/CabHLH108	0.4385	2.1414	Segmental
CabHLH44/CabHLH108	0.3526	1.1955	Segmental
CabHLH48/CabHLH109	0.2438	0.8300	Segmental

that most duplicated gene pairs had similar expression patterns, such as *CabHLH48*/*CabHLH109*, *CabHLH75*/*CabHLH76*, and *CabHLH46*/*CabHLH73*, which were expressed at low levels in most tissues. By contrast, the expression of *CabHLH34* in flowers was much higher than that of its duplicate *CabHLH58*, and the expression of *CabHLH8* was higher than that of its duplicate *CabHLH17* in all tissues analyzed.

Expression Analysis of *CabHLH* Genes Under Abiotic Stresses

To analyze the response of the *CabHLH* genes to abiotic stress, we extracted transcriptome data for *CabHLH* gene expression after 6 h of cold, heat, salt, and drought stress. We used genes with FPKM values of bigger than one in at least one group to create a clustered heatmap and found many *CabHLHs* responded to abiotic stress (**Figure 7**, **Supplementary Table S6**). We also analyzed the relationship between transcriptome data and *cis*-elements and found that gene expression results were not clearly correlated with the presence/absence of specific *cis*-elements. For example, some *CabHLHs* with LTR promoter elements, such as *CabHLH5/17/32/65/90*, were upregulated under LTR treatment. However, some *CabHLHs* with LTR elements, such as *CabHLH16/36/48/114*, were downregulated or remained unchanged (**Figure 7**). This result indicates that the expression of these *CabHLHs* might induced by several *cis*-elements, and unidentified *cis*-elements might contribute to regulating the expression of these *CabHLHs* under abiotic stress.

To further validate the effects of abiotic stress on the expression of *CabHLH* genes, we selected eight genes that responded to abiotic stress (**Supplementary Table S6**) and verified the expression patterns of these genes using qRT-PCR (**Figure 8**). The specific primers used are listed in **Supplementary Table S7**. After cold stress treatment, the expression levels of *CabHLH30*, *CabHLH37*, *CabHLH42*, *CabHLH71*, and *CabHLH111* were upregulated, *CabHLH11* was downregulated, *CabHLH28* was first upregulated and then downregulated, and *CabHLH41* remained unchanged. After high temperature treatment, the expression levels of *CabHLH37* and *CabHLH42* were downregulated, and the expression levels of the remaining genes were upregulated. After drought treatment, *CabHLH30*, *CabHLH71*, and *CabHLH111* were upregulated, *CabHLH41* and *CabHLH37* were downregulated, and *CabHLH42* was first upregulated and then downregulated. Under salt stress, the expression levels of *CabHLH30*, *CabHLH37*, *CabHLH71*, and *CabHLH111* were upregulated, *CabHLH11* and *CabHLH28* were first upregulated and then downregulated, *CabHLH41* and *CabHLH42* were downregulated. In general, there was good correspondence between the RNA-seq data and the qRT-PCR results. However, few exceptions existed. For example, in the qRT-PCR experiment, the expression level of *CabHLH11* decreased after 6 h of cold treatment, but its expression was unchanged in the RNA-seq analysis, perhaps due to different sampling time points (qRT-PCR at 12:00, RNA-seq at 14:00).

DISCUSSION

A growing body of evidence suggests that plant *bHLH* genes are involved in physiological and biochemical processes such as stress resistance, growth and development, biosynthesis, and signaling (Duek and Fankhauser, 2003; Hernandez et al., 2004; Castillon et al., 2007). Members of the bHLH TF family have been identified in *Arabidopsis* (Toledo-Ortiz et al., 2003), rice (Li et al., 2006), apple (Yang et al., 2017), cabbage (Song et al., 2014), tomato (Sun et al., 2015), ginseng (Chu et al., 2018), and other species by comparative genomics. Until now, this family had not been characterized in pepper. In this study, we systematically analyzed the pepper bHLH TF family and provided a reference for further exploration of the roles of bHLH genes in regulation of pepper growth and stress responses.

A total of 122 *CabHLH* genes were identified and classified into 21 subfamilies according to their phylogenetic relationships with known *bHLH* genes from *Arabidopsis* (Li et al., 2006). Compared with *Arabidopsis*, pepper lacks members of the XV subfamily but contains a unique X subfamily. The acquired genes may counter gene losses, or even evolve novel functions (Qian et al., 2010). The functions of some *AtbHLH*s have been identified in previous studies. For example, *AtbHLH15* and *AtbHLH8* from subfamily VII (a + b) can combine with active phytochromes and mediate light signaling responses (Castillon et al., 2007). *AtbHLH44*, *AtbHLH58*, and *AtbHLH50* in subfamily VII are early response BR signaling components required for full BR response (Friedrichsen et al., 2002). Overexpression of *AtbHLH116* from subfamily IIIb in wild-type plants improves the expression of the *CBF* regulon in the cold and enhances freezing tolerance of transgenic plants (Chinnusamy et al., 2003). *AtbHLH1* from subfamily IIIf encodes a bHLH protein that regulates trichome development in *Arabidopsis* through interaction with GLABRA3 and TESTA GLABRA1 (Payne et al., 2000). *CabHLH*s and *AtbHLH*s from the same subfamilies may have similar functions, although this will require further experimental verification.

Gene duplication, including tandem duplication and segmental duplication, is the most important pathway for

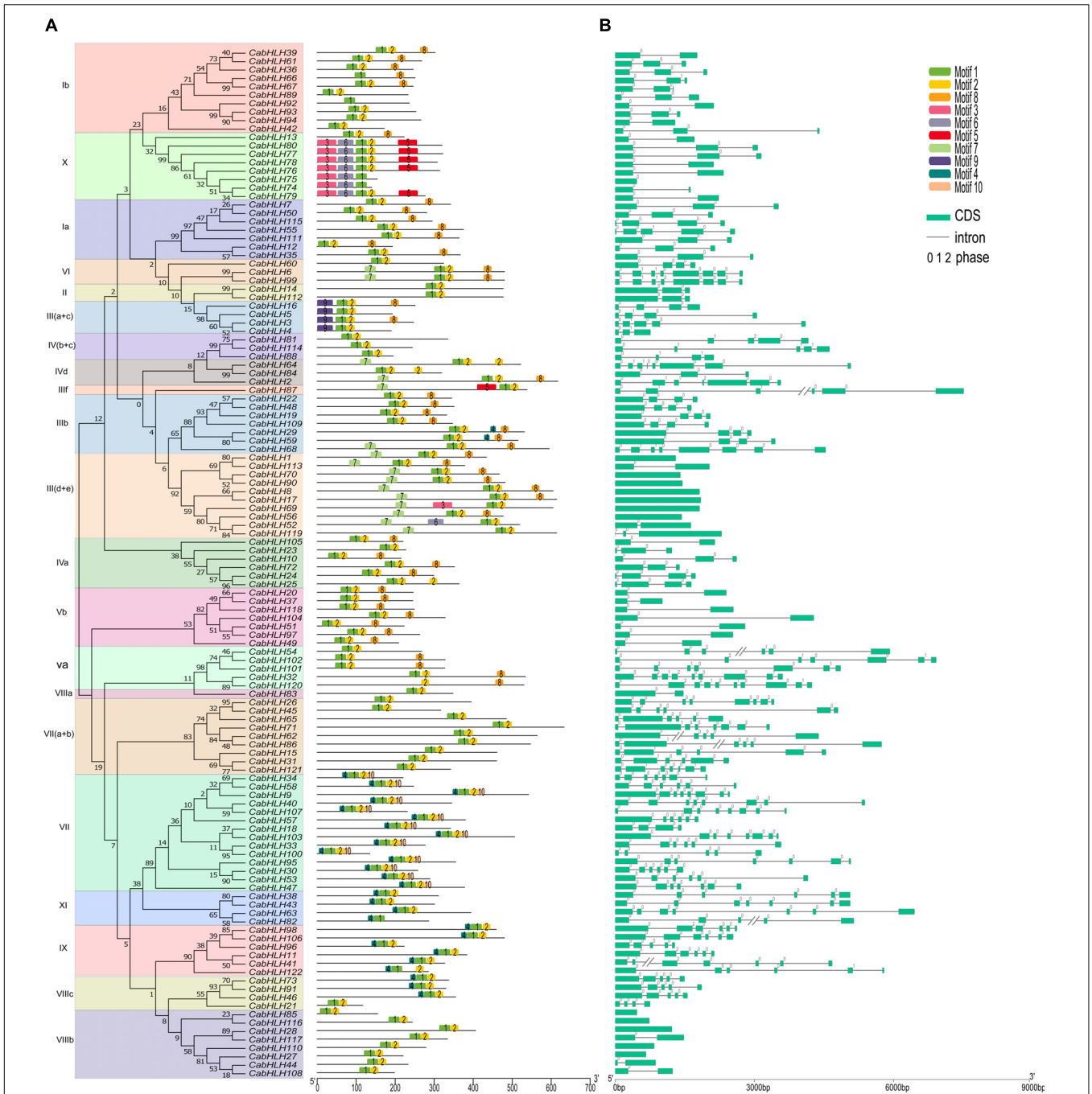


FIGURE 4 | Exon–intron structures of *CabHLH* genes and conserved motifs of CabHLH proteins. **(A)** Exon–intron organization of *CabHLH* genes. Green boxes represent exons and black lines indicate introns. The numbers 0, 1, and 2 denote the intron phases. **(B)** Conserved motifs in the CabHLH proteins. The conserved motifs were identified using MEME with complete protein sequences. Different motifs are displayed by various colors.

the evolution and expansion of gene families (Vision et al., 2000). We identified six tandem duplicated *CabHLH*s and ten segmental duplicated *CabHLH*s in the pepper genome. Collinear genes derive from a common ancestor and are present in the same relative positions in the genomes of two or more species. We identified 117, 64 and 105 collinear *bHLH* pairs between pepper and tomato, *Arabidopsis* and pepper, and *Arabidopsis* and tomato, respectively. In the process of evolution, collinear blocks may be disrupted by various factors. The greater the evolutionary distance, the fewer collinear gene pairs will be identified between species, and collinearity can therefore be used as a measure of the evolutionary distance between species (Wicker et al., 2010). There were more collinear gene pairs between tomato and pepper, consistent with the fact that both are members

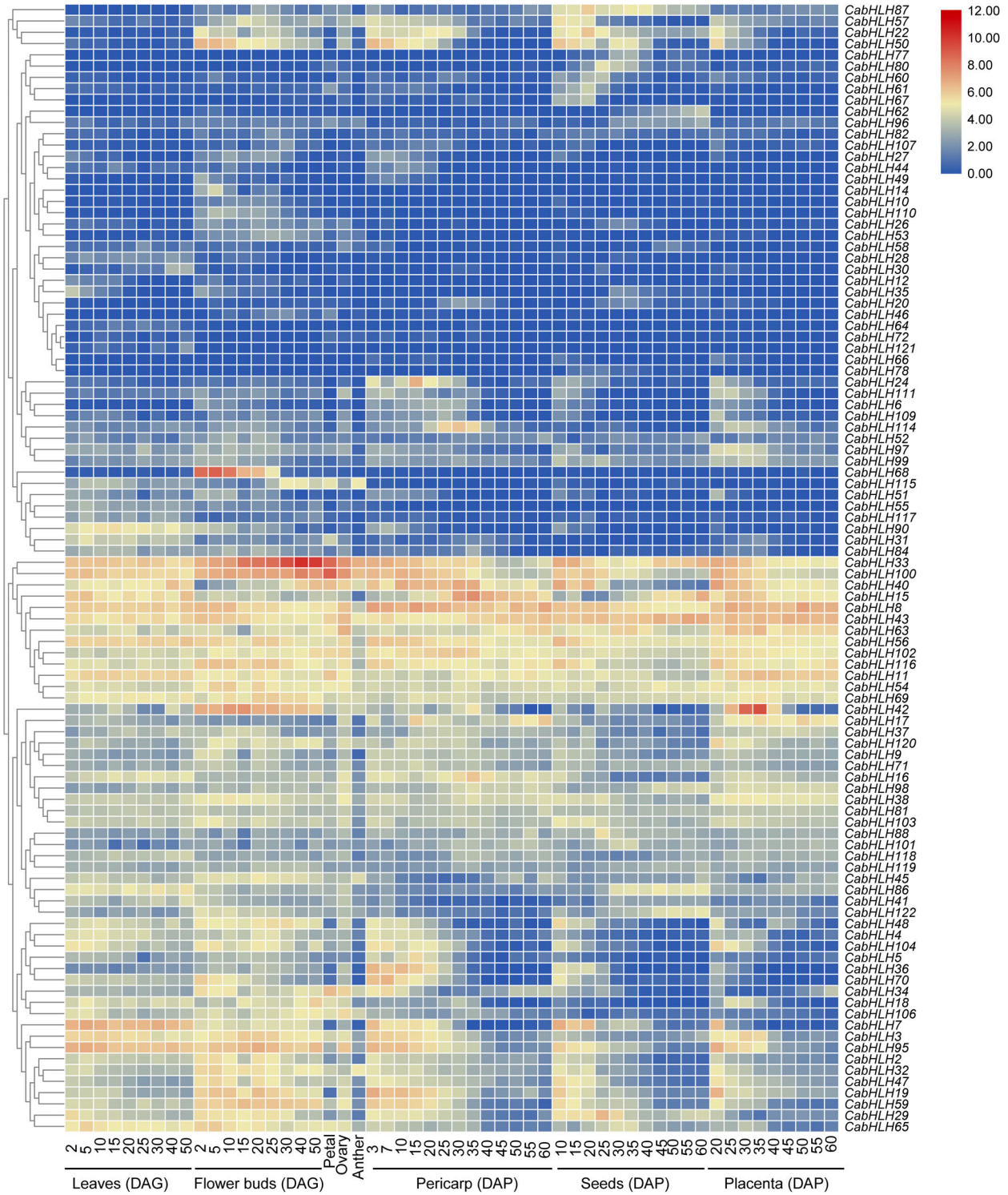


FIGURE 5 | Expression patterns of *CabHLH* genes in different tissues and organs. Heatmap of expression profiles [in $\log_2$ (FPKM + 1)] for *CabHLH* genes in different tissues and organs. The expression levels are displayed by the color bar. DAG, days after germination; DAP, days after pollination.

of the Solanaceae family (Qin et al., 2014). Previous studies have shown that the amplification of transposable elements has eroded collinearity in the pepper genome (Wicker et al., 2010; Qin et al., 2014), which may explain why the number of collinear gene pairs between pepper and *Arabidopsis* is much lower than that between *Arabidopsis* and tomato.

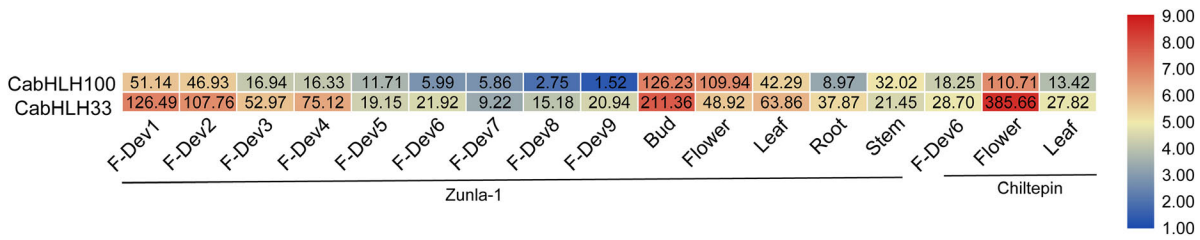


FIGURE 6 | Heatmap of expression profiles [in log 2 (RPKM + 1)] of *CabHLH33* and *CabHLH100* in two pepper cultivars "Zunla-1" (*Capsicum annuum* L.) and "Chiltepin" (*C. annuum var. glabriusculum*). The expression levels are displayed by the color bar. F-Dev-1, F-Dev-2, F-Dev-3, F-Dev-4, and F-Dev-5 (0–1 cm, 1–3 cm, 3–4 cm, 4–5 cm, and mature green fruit), F-Dev-6 (fruit turning red), F-Dev-7, F-Dev-8, and F-Dev-9 (3, 5, and 7 days after turning red). RPKM, reads per kilobase million.

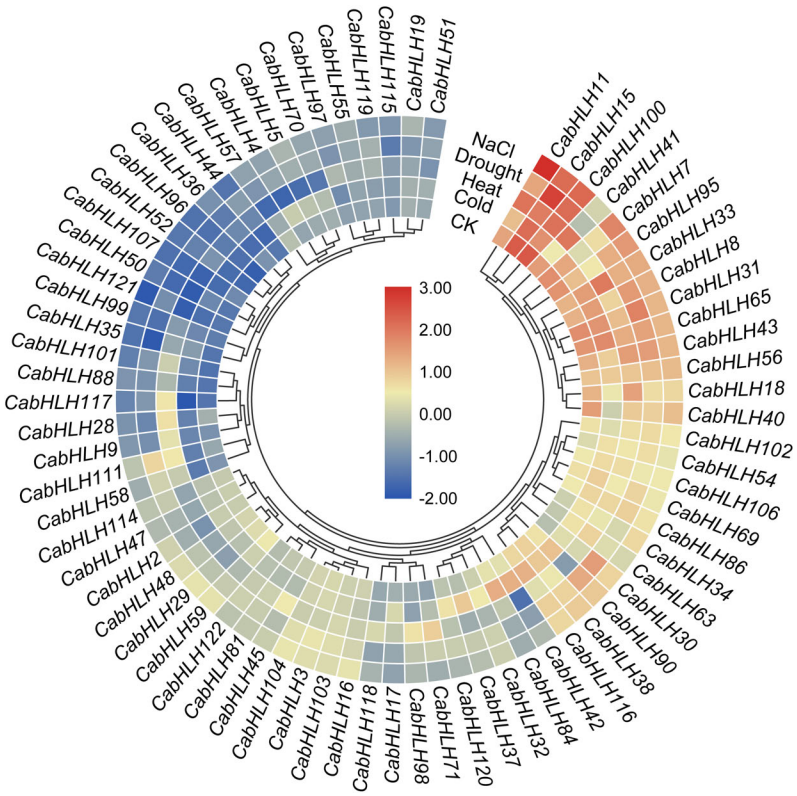


FIGURE 7 | Expression heatmap of *CabHLH* genes under multiple abiotic treatments. The color scale represents $\log_2$ (FPKM + 1) values.

We identified ten highly conserved motifs in the CabHLH proteins. Similar to the bHLHs of potato, lotus and *Arabidopsis* (Wang et al., 2018; Mao et al., 2019), motif 1 and motif 2 were present in almost all CabHLH proteins and represented the position of the bHLH domain, which is highly conserved among species. However, motif 9 and 10 were only present in subfamilies III (a + c) and VII, respectively. Variation in conserved motifs permits the classification of proteins into subfamilies and reflects each subfamily's specific functions (Jiang et al., 2019). Gene structure can also provide information for the study of gene family evolution (Guo et al., 2013). The number of introns varied from 0 to 9, indicating that the gain and loss of introns had occurred, which may be another reason for the differences among CabHLH subfamilies (Paquette et al., 2000).

We analyzed the expression profiles of *CabHLHs* in different tissues and found a large variety of expression patterns. Some *CabHLHs* (such as *CabHLH100, CabHLH11, CabHLH8*, and *CabHLH43*) were highly expressed (FPKM > 10) in most tissues analyzed and may participating in various development processes of pepper. Several *CabHLHs* were highly expressed in specific tissues, suggesting that they may have a role in those tissues' development. For example, *CabHLH33*, a homolog of *AtbHLH31*, was highly expressed in flower buds and petals. Previous studies suggest that *AtbHLH31* regulates petal growth by controlling cell expansion (Varaud et al., 2011), and *CabHLH33* may have

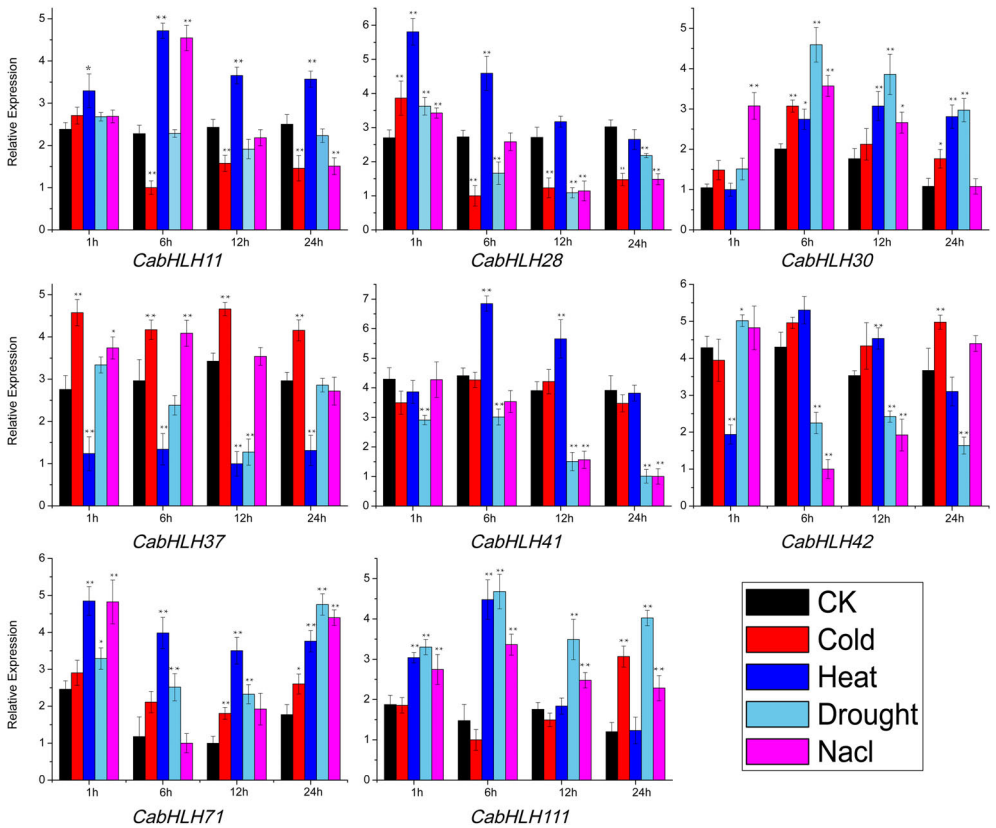


FIGURE 8 | qRT-PCR analysis of *CabHLH* genes under cold, heat, salt and drought treatments following a 24 h time course. *y*-axis: relative expression levels; *x*-axis: the time (hours) course of stress treatments. *t*-test: one asterisk denotes significant differences ($P < 0.05$) between treatment group and control group (CK); two asterisks denote extremely significant differences ($P < 0.01$).

a similar function in pepper. However, *CabHLH*s that were not expressed in any tissues (such as *CabHLH23*, *CabHLH85*, *CabHLH39*, *CabHLH105*, and *CabHLH108*) may have lost their functions during evolution and become pseudogenes, as has been demonstrated in the evolution of other plant genomes (Innan and Kondrashov, 2010; Xie et al., 2019). In addition, several duplicated pairs (such as *CabHLH34/CabHLH58* and *CabHLH8/CabHLH17*) had significantly different expression patterns, indicating that functional diversification of duplicated *CabHLH* pairs had occurred during the course of evolution (Blanc and Wolfe, 2004).

Plant *bHLH*s modulate stress responses, including responses to LTR (Feng et al., 2012; Lin et al., 2014; Xu et al., 2014), drought (Li et al., 2007; Seo et al., 2011), heat (Ko et al., 2009), and salt (Murre et al., 1989; Zhou et al., 2009; Liu et al., 2014). In this work, *cis*-element analysis indicated that *CabHLH*s contained elements (such as LTR, ABRE, and TC-rich) that are responsive to various stresses, which was consistent with previous research on potato and lotus *bHLH*s (Wang et al., 2018; Mao et al., 2019). In addition, we extracted the transcriptome data of *CabHLH*s and performed a qRT-PCR experiment to validate the response of the *CabHLH* genes to abiotic stress. These upregulated *CabHLH*s (such as *CabHLH30/37/42/71/111* under cold, *CabHLH11/28/30/41/71/111* under heat, *CabHLH30/71/111* under drought, and *CabHLH30/37/71/111* under NaCl) may regulate downstream bHLH-related genes, thus enhancing the stress tolerance of pepper. Our research provides a framework for further functional characterization of *CabHLH* genes.

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/**Supplementary Material.**

AUTHOR CONTRIBUTIONS

ZZ, FL, XH, and XZ conceived and designed the experiments. ZZ, JC, and CL performed the experiments. ZZ, CL, FL, and XH analyzed data. ZZ and XZ wrote the manuscript. All authors read and approved the manuscript.

FUNDING

This research was funded by the National Key Research and Development Program of China (2019YFD1000300).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.570156/full#supplementary-material

FIGURE S1 | Sequence logos of conserved motifs of CabHLHs.

FIGURE S2 | Exon-intron structures of *AtbHLHs* and *SlbHLHs*. The phylogenetic tree of *AtbHLHs* and *SlbHLH* proteins was constructed by MEGA7 using the neighbor-joining (NJ) method (1,000 bootstrap). Green boxes represent exons and black lines indicate introns.

TABLE S1 | Genes with complete bHLH domain in pepper, tomato, and *Arabidopsis*.

TABLE S2 | Basic information of the CabHLH transcription factor family.

TABLE S3 | Collinear *bHLHs* among tomato, pepper and *Arabidopsis*.

TABLE S4 | *Cis*-elements analysis of the *CabHLH* genes.

TABLE S5 | The FPKM values of *CabHLH* genes in different tissues and organs.

TABLE S6 | FPKM values of *CabHLH* genes under salt, drought, cold, and heat stress.

TABLE S7 | Primer sequences used in qRT-PCR.

REFERENCES

Arnaud, N., Girin, T., Sorefan, K., Fuentes, S., Wood, T. A., Lawrenson, T., et al. (2010). Gibberellins control fruit patterning in *Arabidopsis thaliana*. *Genes Dev.* 24, 2127–2132. doi: 10.1101/gad.593410

Atchley, W. R., and Fitch, W. M. (1997). A natural classification of the basic helix-loop-helix class of transcription factors. *Proc. Natl. Acad. Sci U.S.A.* 94, 5172–5176. doi: 10.1073/pnas.94.10.5172

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–W208. doi: 10.1093/nar/gkp335

Baudry, A., Heim, M. A., Dubreucq, B., Caboche, M., Weisshaar, B., and Lepiniec, L. (2004). *TT2*, *TT8*, and *TTG1* synergistically specify the expression of BANYULS and proanthocyanidin biosynthesis in *Arabidopsis thaliana*. *Plant J.* 39, 366–380. doi: 10.1111/j.1365-313X.2004.02138.x

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

Carretero-Paulet, L., Galstyan, A., Roig-Villanova, I., Martinez-Garcia, J. F., Bilbao-Castro, J. R., and Robertson, D. L. (2010). Genome-wide classification and evolutionary analysis of the bHLH family of transcription factors in Arabidopsis, poplar, rice, moss, and algae. *Plant Physiol.* 153, 1398–1412. doi: 10.1104/pp.110.153593

Castillon, A., Shen, H., and Huq, E. (2007). Phytochrome Interacting Factors: central players in phytochrome-mediated light signaling networks. *Trends Plant Sci.* 12, 514–521. doi: 10.1016/j.tplants.2007.10.001

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

Chinnusamy, V., Ohta, M., Kanrar, S., Lee, B. H., Hong, X., Agarwal, M., et al. (2003). ICE1: a regulator of cold-induced transcriptome and freezing tolerance in *Arabidopsis*. *Genes Dev.* 17, 1043–1054. doi: 10.1101/gad.1077503

Chu, Y., Xiao, S., Su, H., Liao, B., Zhang, J., Xu, J., et al. (2018). Genome-wide characterization and analysis of bHLH transcription factors in *Panax ginseng*. *Acta Pharm. Sin. B* 8, 666–677. doi: 10.1016/j.apsb.2018.04.004

Duek, P. D., and Fankhauser, C. (2003). HFR1, a putative bHLH transcription factor, mediates both phytochrome A and cryptochrome signalling. *Plant J.* 34, 827–836. doi: 10.1046/j.1365-313x.2003.01770.x

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics* 14, 755–763. doi: 10.1093/bioinformatics/14.9.755

Feng, X. M., Zhao, Q., Zhao, L. L., Qiao, Y., Xie, X. B., Li, H. F., et al. (2012). The cold-induced basic helix-loop-helix transcription factor gene MdCIbHLH1 encodes an ICE-like protein in apple. *BMC Plant Biol.* 12:22. doi: 10.1186/1471-2229-12-22

Ferre-D'Amare, A. R., Pognonec, P., Roeder, R. G., and Burley, S. K. (1994). Structure and function of the b/HLH/Z domain of USF. *EMBO J.* 13, 180–189. doi: 10.1002/j.1460-2075.1994.tb06247.x

Friedrichsen, D. M., Nemhauser, J., Muramitsu, T., Maloof, J. N., Alonso, J., Ecker, J. R., et al. (2002). Three redundant brassinosteroid early response genes encode putative bHLH transcription factors required for normal growth. *Genetics* 162, 1445–1456.

Goodrich, J., Carpenter, R., and Coen, E. S. (1992). A common gene regulates pigmentation pattern in diverse plant species. *Cell* 68, 955–964. doi: 10.1016/0092-8674(92)90038-e

Guo, A. Y., Chen, X., Gao, G., Zhang, H., Zhu, Q. H., Liu, X. C., et al. (2008). PlantTFDB: a comprehensive plant transcription factor database. *Nucleic Acids Res.* 36, D966–D969. doi: 10.1093/nar/gkm841

Guo, M., Liu, J. H., Ma, X., Zhai, Y. F., Gong, Z. H., and Lu, M. H. (2016). Genome-wide analysis of the Hsp70 family genes in pepper (*Capsicum annuum* L.) and functional identification of CaHsp70-2 involvement in heat stress. *Plant Sci.* 252, 246–256. doi: 10.1016/j.plantsci.2016.07.001

Guo, R., Xu, X., Carole, B., Li, X., Gao, M., Zheng, Y., et al. (2013). Genome-wide identification, evolutionary and expression analysis of the aspartic protease gene superfamily in grape. *BMC Genomics* 14:554. doi: 10.1186/1471-2164-14-554

Heim, M. A., Jakoby, M., Werber, M., Martin, C., Weisshaar, B., and Bailey, P. C. (2003). The basic helix-loop-helix transcription factor family in plants: a genome-wide study of protein structure and functional diversity. *Mol. Biol. Evol.* 20, 735–747. doi: 10.1093/molbev/msg088

Hernandez, J. M., Heine, G. F., Irani, N. G., Feller, A., Kim, M. G., Matulnik, T., et al. (2004). Different mechanisms participate in the R-dependent activity of the R2R3 MYB transcription factor C1. *J. Biol. Chem.* 279, 48205–48213. doi: 10.1074/jbc.M407845200

Innan, H., and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* 11, 97–108. doi: 10.1038/nrg2689

Jiang, S. Y., Jin, J., Sarojam, R., and Ramachandran, S. (2019). A comprehensive survey on the terpene synthase gene family provides new insight into its evolutionary patterns. *Genome Biol. Evol.* 11, 2078–2098. doi: 10.1093/gbe/evz142

Jiang, Y., Yang, B., and Deyholos, M. K. (2009). Functional characterization of the *Arabidopsis bHLH92* transcription factor in abiotic stress. *Mol. Genet. Genomics* 282, 503–516. doi: 10.1007/s00438-009-0481-3

Kim, J., and Kim, H. Y. (2006). Functional analysis of a calcium-binding transcription factor involved in plant salt stress signaling. *FEBS Lett.* 580, 5251–5256. doi: 10.1016/j.febslet.2006.08.050

Kiribuchi, K., Jikumaru, Y., Kaku, H., Minami, E., Hasegawa, M., Kodama, O., et al. (2005). Involvement of the basic helix-loop-helix transcription factor *RERJ1* in wounding and drought stress responses in rice plants. *Biosci. Biotechnol. Biochem.* 69, 1042–1044. doi: 10.1271/bbb.69.1042

Kiribuchi, K., Sugimori, M., Takeda, M., Otani, T., Okada, K., Onodera, H., et al. (2004). *RERJ1*, a jasmonic acid-responsive gene from rice, encodes a basic helix-loop-helix protein. *Biochem. Biophys. Res. Commun.* 325, 857–863. doi: 10.1016/j.bbrc.2004.10.126

Ko, D. K., Lee, M. O., Hahn, J. S., Kim, B. G., and Hong, C. B. (2009). Submergence-inducible and circadian rhythmic basic helix-loop-helix protein gene in *Nicotiana tabacum*. *J. Plant Physiol.* 166, 1090–1100. doi: 10.1016/j.jplph.2008.12.008

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

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

Lescot, M., Dehais, 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

Li, H., Sun, J., Xu, Y., Jiang, H., Wu, X., and Li, C. (2007). The bHLH-type transcription factor AtAIB positively regulates ABA response in *Arabidopsis*. *Plant Mol. Biol.* 65, 655–665. doi: 10.1007/s11103-007-9230-3

Li, X., Duan, X., Jiang, H., Sun, Y., Tang, Y., Yuan, Z., et al. (2006). Genome-wide analysis of basic/helix-loop-helix transcription factor family in rice and *Arabidopsis*. *Plant Physiol.* 141, 1167–1184. doi: 10.1104/pp.106.080580

Lin, Y., Zheng, H., Zhang, Q., Liu, C., and Zhang, Z. (2014). Functional profiling of EcaICE1 transcription factor gene from *Eucalyptus camaldulensis* involved in cold response in tobacco plants. *J. Plant Biochem. Biotechnol.* 23, 141–150. doi: 10.1007/s13562-013-0192-z

Liu, F., Yu, H., Deng, Y., Zheng, J., Liu, M., Ou, L., et al. (2017). PepperHub, an informatics hub for the chili pepper research community. *Mol. Plant* 10, 1129–1132. doi: 10.1016/j.molp.2017.03.005

Liu, W., Tai, H., Li, S., Gao, W., Zhao, M., Xie, C., et al. (2014). bHLH122 is important for drought and osmotic stress resistance in Arabidopsis and in the repression of ABA catabolism. *New Phytol.* 201, 1192–1204. doi: 10.1111/nph.12607

Liu, Y., Jiang, H., Chen, W., Qian, Y., Ma, Q., Cheng, B., et al. (2011). Genome-wide analysis of the auxin response factor (ARF) gene family in maize (*Zea mays*). *Plant Growth Regul.* 63, 225–234. doi: 10.1007/s10725-010-9519-0

Ludwig, S. R., Habera, L. F., Dellaporta, S. L., and Wessler, S. R. (1989). Lc, a member of the maize R gene family responsible for tissue-specific anthocyanin production, encodes a protein similar to transcriptional activators and contains the myc-homology region. *Proc. Natl. Acad. Sci U.S.A.* 86, 7092–7096. doi: 10.1073/pnas.86.18.7092

Mao, T. Y., Liu, Y. Y., Zhu, H. H., Zhang, J., Yang, J. X., Fu, Q., et al. (2019). Genome-wide analyses of the bHLH gene family reveals structural and functional characteristics in the aquatic plant Nelumbo nucifera. *PeerJ* 7:e7153. doi: 10.7717/peerj.7153

Murre, C., McCaw, P. S., and Baltimore, D. (1989). A new DNA binding and dimerization motif in immunoglobulin enhancer binding, daughterless, MyoD, and myc proteins. *Cell* 56, 777–783. doi: 10.1016/0092-8674(89)90682-x

Nesi, N., Debeaujon, I., Jond, C., Pelletier, G., Caboche, M., and Lepiniec, L. (2000). The TT8 gene encodes a basic helix-loop-helix domain protein required for expression of DFR and BAN genes in *Arabidopsis* siliques. *Plant Cell* 12, 1863–1878. doi: 10.1105/tpc.12.10.1863

Paquette, S. M., Bak, S., and Feyereisen, R. (2000). Intron-exon organization and phylogeny in a large superfamily, the paralogous cytochrome P450 genes of *Arabidopsis thaliana*. *DNA Cell Biol.* 19, 307–317. doi: 10.1089/10445490050021221

Payne, C. T., Zhang, F., and Lloyd, A. M. (2000). GL3 Encodes a bHLH Protein That Regulates Trichome Development in *Arabidopsis* Through Interaction With GL1 and TTG1. *Genetics* 156, 1349–1362.

Qian, W., Liao, B. Y., Chang, A. Y., and Zhang, J. (2010). Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends Genet.* 26, 425–430. doi: 10.1016/j.tig.2010.07.002

Qin, C., Yu, C., Shen, Y., Fang, X., Chen, L., Min, J., et al. (2014). Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. *Proc. Natl. Acad. Sci U.S.A.* 111, 5135–5140. doi: 10.1073/pnas.1400975111

Quattrocchio, F., Wing, J. F., Leppen, H., Mol, J., and Koes, R. E. (1993). Regulatory Genes Controlling Anthocyanin Pigmentation Are Functionally Conserved among Plant Species and Have Distinct Sets of Target Genes. *Plant Cell* 5, 1497–1512. doi: 10.1105/tpc.5.11.1497

Roy, S. W., and Gilbert, W. (2005). Rates of intron loss and gain: implications for early eukaryotic evolution. *Proc. Natl. Acad. Sci U.S.A.* 102, 5773–5778. doi: 10.1073/pnas.0500383102

Sailsbery, J. K., and Dean, R. A. (2012). Accurate discrimination of bHLH domains in plants, animals, and fungi using biologically meaningful sites. *BMC Evol. Biol.* 12:154. doi: 10.1186/1471-2148-12-154

Sakamoto, W., Ohmori, T., Kageyama, K., Miyazaki, C., Saito, A., Murata, M., et al. (2001). The Purple leaf (Pl) locus of rice: the Pl(w) allele has a complex organization and includes two genes encoding basic helix-loop-helix proteins involved in anthocyanin biosynthesis. *Plant Cell Physiol.* 42, 982–991. doi: 10.1093/pcp/pce128

Schmittgen, T. D., and Livak, K. J. (2008). Analyzing real-time PCR data by the comparative C(T) method. *Nat. Protoc.* 3, 1101–1108. doi: 10.1038/nprot.2008.73

Seo, J. S., Joo, J., Kim, M. J., Kim, Y. K., Nahm, B. H., Song, S. I., et al. (2011). *OsbHLH148*, a basic helix-loop-helix protein, interacts with OsJAZ proteins in a jasmonate signaling pathway leading to drought tolerance in rice. *Plant J.* 65, 907–921. doi: 10.1111/j.1365-313X.2010.04477.x

Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* 11:e0163962. doi: 10.1371/journal.pone.0163962

Smolen, G. A., Pawlowski, L., Wilensky, S. E., and Bender, J. (2002). Dominant alleles of the basic helix-loop-helix transcription factor ATR2 activate stress-responsive genes in *Arabidopsis*. *Genetics* 161, 1235–1246.

Song, X. M., Huang, Z. N., Duan, W. K., Ren, J., Liu, T. K., Li, Y., et al. (2014). Genome-wide analysis of the bHLH transcription factor family in Chinese cabbage (*Brassica rapa* ssp. *pekinensis*). *Mol. Genet. Genomics* 289, 77–91. doi: 10.1007/s00438-013-0791-3

Sun, H., Fan, H. J., and Ling, H. Q. (2015). Genome-wide identification and characterization of the bHLH gene family in tomato. *BMC Genomics* 16:9. doi: 10.1186/s12864-014-1209-2

Tamura, K., Dudley, J., Nei, M., and Kumar, S. (2007). MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* 24, 1596–1599. doi: 10.1093/molbev/msm092

Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., and Paterson, A. H. (2008). Synteny and collinearity in plant genomes. *Science* 320, 486–488. doi: 10.1126/science.1153917

Toledo-Ortiz, G., Huq, E., and Quail, P. H. (2003). The *Arabidopsis* basic/helix-loop-helix transcription factor family. *Plant Cell* 15, 1749–1770. doi: 10.1105/tpc.013839

Varaud, E., Brioudes, F., Szecsi, J., Leroux, J., Brown, S., Perrot-Rechenmann, C., et al. (2011). AUXIN RESPONSE FACTOR8 regulates Arabidopsis petal growth by interacting with the bHLH transcription factor BIGPETALp. *Plant Cell* 23, 973–983. doi: 10.1105/tpc.110.081653

Vision, T. J., Brown, D. G., and Tanksley, S. D. (2000). The origins of genomic duplications in Arabidopsis. *Science* 290, 2114–2117. doi: 10.1126/science.290.5499.2114

Wang, D., Zhang, Y., Zhang, Z., Zhu, J., and Yu, J. (2010). KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinformatics* 8, 77–80. doi: 10.1016/S1672-0229(10)60008-3

Wang, R., Zhao, P., Kong, N., Lu, R., Pei, Y., Huang, C., et al. (2018). Genome-wide identification and characterization of the potato bHLH transcription factor family. *Genes* 9:54. doi: 10.3390/genes9010054

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40:e49. doi: 10.1093/nar/gkr1293

Wang, Y. J., Zhang, Z. G., He, X. J., Zhou, H. L., Wen, Y. X., Dai, J. X., et al. (2003). A rice transcription factor *OsbHLH1* is involved in cold stress response. *Theor. Appl. Genet.* 107, 1402–1409. doi: 10.1007/s00122-003-1378-x

Wicker, T., Buchmann, J. P., and Keller, B. (2010). Patching gaps in plant genomes results in gene movement and erosion of colinearity. *Genome Res.* 20, 1229–1237. doi: 10.1101/gr.107284.110

Wu, Z., Cheng, J., Cui, J., Xu, X., Liang, G., Luo, X., et al. (2016). Genome-wide identification and expression profile of Dof transcription factor gene family in pepper (*Capsicum annuum* L.). *Front. Plant Sci.* 7:574. doi: 10.3389/fpls.2016.00574

Xie, J., Li, Y., Liu, X., Zhao, Y., Li, B., Ingvarsson, P. K., et al. (2019). Evolutionary origins of Pseudogenes and their association with regulatory sequences in plants. *Plant Cell* 31, 563–578. doi: 10.1105/tpc.18.00601

Xu, W., Jiao, Y., Li, R., Zhang, N., Xiao, D., Ding, X., et al. (2014). Chinese wild-growing Vitis amurensis ICE1 and ICE2 encode MYC-type bHLH transcription activators that regulate cold tolerance in *Arabidopsis*. *PLoS One* 9:e102303. doi: 10.1371/journal.pone.0102303

Yang, J., Gao, M., Huang, L., Wang, Y., van Nocker, S., Wan, R., et al. (2017). Identification and expression analysis of the apple (*Malus x domestica*) basic

helix-loop-helix transcription factor family. *Sci. Rep.* 7:28. doi: 10.1038/s41598-017-00040-y

Zhou, J., Li, F., Wang, J. L., Ma, Y., Chong, K., and Xu, Y. Y. (2009). Basic helix-loop-helix transcription factor from wild rice (OrbHLH2) improves tolerance to salt- and osmotic stress in *Arabidopsis*. *J. Plant Physiol.* 166, 1296–1306. doi: 10.1016/j.jplph.2009.02.007

Zhuo, X., Zheng, T., Zhang, Z., Zhang, Y., Jiang, L., Ahmad, S., et al. (2018). Genome-wide analysis of the NAC transcription factor gene family reveals differential expression patterns and cold-stress responses in the woody plant *Prunus mume*. *Genes* 9:494. doi: 10.3390/genes9100494

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.

Copyright © 2020 Zhang, Chen, Liang, Liu, Hou and Zou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 06 October 2020
doi: 10.3389/fgene.2020.583996

Genome-Wide Identification, Structural, and Gene Expression Analysis of BRI1-EMS-Suppressor 1 Transcription Factor Family in *Cucumis sativus*

Si Ma†, Tingting Ji†, Meiting Liang, Shihui Li, Yongqiang Tian and Lihong Gao**

Beijing Key Laboratory of Growth and Developmental Regulation for Protected Vegetable Crops, College of Horticulture, China Agricultural University, Beijing, China

OPEN ACCESS

Edited by:
Hongjian Wan, ZheJiang Academy of Agricultural Sciences, China

Reviewed by:
Yuzhou Wu, The University of Sydney, Australia
Vijay Gahlaut, University of Delhi, India
Wenjiao Wang, Shanxi Agricultural University, China

***Correspondence:**
Yongqiang Tian
tianyq1984@cau.edu.cn
Lihong Gao
gaolh@cau.edu.cn

†These authors have contributed equally to this work

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

Received: *16 July 2020*
Accepted: *07 September 2020*
Published: *06 October 2020*

Citation:
Ma S, Ji T, Liang M, Li S, Tian Y and Gao L (2020) Genome-Wide Identification, Structural, and Gene Expression Analysis of BRI1-EMS-Suppressor 1 Transcription Factor Family in Cucumis sativus. Front. Genet. 11:583996. doi: 10.3389/fgene.2020.583996

BRASSINOSTEROID INSENSITIVE1-EMS-suppressor 1 (BES1) is an essential regulator downstream of brassinosteroid signaling and plays important roles in plant stress response, growth, and development. To date, the regulation mechanisms of BES1 transcription factors have been identified and elucidated in model plants Arabidopsis and rice. However, little information is available regarding the BES1 family in *Cucumis sativus*. Therefore, this study conducted a genome-wide analysis of BES1 genes in cucumber. In cucumber, a total of six CsBES1 genes were identified, and their phylogenetic relationships, gene structures, and *cis*-elements in promoters were studied. *CsBES1* genes were distributed on four of seven chromosomes. Gene structure analysis showed that the intron–exon model of *CsBES1* genes was conserved and the CsBES1 protein contained a DUF822-conserved motif. Promoter *cis*-element prediction showed that plenty of developmental and stress- and hormone-related elements have been found in promoter regions of *CsBES1* genes. Meanwhile, BES1 was divided into three groups (I, II, and III) on the basis of phylogenetic relationship analysis in six plant species. In addition, *CsBES1* gene expression patterns were confirmed by transcription database and qRT-PCR analysis; the results showed that the expression of *CsBES1* genes had not only tissue-specific expression but also different types of *CsBES1* isoform which might respond to specific plant stresses. In summary, genome-wide identification, phylogeny, gene structure, and expression profile analysis of *CsBES1* genes in cucumber provided a referable theoretical information for further functional study of *CsBES1* genes and further facilitated the molecular breeding of cucumber.

Keywords: *Cucumis sativus*, BRI1-EMS-suppressor 1, gene structure, phylogeny, *cis*-element, abiotic stress

INTRODUCTION

Brassinosteroids (BRs), as a class of steroid hormone, are distributed ubiquitously in the plant kingdom. BRs can be used as stress alleviators to protect plants from a variety of abiotic and biotic stresses, such as drought, osmotic, cold, heat, wounding, and pathogen attack (Marcos et al., 2015; Krishna et al., 2017; Yan et al., 2018). In addition, BRs have been shown to act as key components of

multiple physiological and developmental processes such as vascular development, cell elongation, cell division, and leaf senescence (Guo et al., 2013; Tong et al., 2014; Unterholzner et al., 2015). Given this, extensive researches have been done to reveal the molecular mechanisms and potential application of BRs in plants. To date, great progress has been achieved in BR signal transduction by proteomic surveys and molecular genetics in Arabidopsis.

In the BR signal transduction process, BRs are perceived through plasma membrane receptor kinase BRASSINOSTEROID INSENSITIVE1 (BRI1) and the co-receptor BRI1-Associated Receptor Kinase 1 (BAK1; Kim and Wang, 2010; Kim et al., 2011). The activation of BRI1 leads to sequential phosphorylation or dephosphorylation reaction, including positive regulators BRI1-SUPPRESSOR 1 (BSU1; Mora-García et al., 2004), BR SIGNALING KINASE 1 (BSK1; Tang et al., 2008), and protein phosphatase 2A (PP2A; Tang et al., 2011), and negative regulators BRI1-KINASE INHIBITOR 1 (BKI1; Wang and Chory, 2006), BRASSINOSTEROID INSENSITIVE 2 (BIN2; Li et al., 2002), and the 14-3-3 proteins (Gampala et al., 2007). Finally, dephosphorylation brassinazole-resistant 1 (BZR1) or BRI1-EMS-suppressor 1 (BES1) are translocated into the nucleus to regulate thousands of BR-target genes (Wang et al., 2002; Yin et al., 2005; Sun et al., 2010).

Brassinazole-resistant 1 and BES1, as the essential regulators in the downstream of BR signaling, have been intensively investigated to reveal their roles in plant stress response, growth, and development (Yin et al., 2005; Ryu et al., 2010; Guo et al., 2013). BES1 and phytochrome-interacting factor 4 (PIF4) interaction promotes cell elongation and plant growth by binding to nearly 2000 common target genes and controlling a core transcription network (Oh et al., 2012). BZR1 acts as an important regulator mediating the trade-off between growth and immunity upon inducing the expression of several WRKY transcription factors (Lozano-Durán et al., 2013). bil1-1D/bzr1-1D is a gain-of-function mutant in which a point mutation causes the stabilization and accumulation of the BIL1/BZR1 protein in the nucleus. When overexpressed in *Lotus japonicus*, bil1-1D/bzr1-1D increases resistance against thrip feeding compared with the wild type (Miyaji et al., 2014). BES1 and BZR1 could involve in the regulation of glucosinolate biosynthesis through binding to glucosinolate biosynthetic genes by the conserved N-terminal DNA-binding domain or indirectly through MYB factors (Guo et al., 2013). Thus, BZR1 and BES1 likely play a critical role in determining the primary BR signaling outputs. Besides, BES1 also is considered to be involved in the strigolactone (SL) signaling pathway. BES1 could interact with more axillary growth locus 2 (MAX2), which is a key component in the SL signaling pathway, and also can be degraded as the direct substrate of MAX2 to regulate shoot branching (Wang et al., 2013). BZR1 interacts *in vitro* and *in vivo* with REPRESSOR OF ga1-3 (RGA), which belongs to a member of the DELLA family of transcriptional regulators, to control cell elongation and plant growth (Li et al., 2012).

Plenty of researches on the characterization and functions of BES1/BZR1 have been conducted in Arabidopsis and rice, and some genome-wide analysis of BES1/BZR1 in tomato, maize, and Chinese cabbage have been performed to explore its gene structural and functional diversity during evolution. However, there is still lack of research on the phylogenetic relationship and function of the *BES1/BZR1* gene family in cucumber. Cucumber (*Cucumis sativus* L.) is an economically significant vegetable crop around the world, as well as a model system for studying fresh fruit development not only due to its vascular bundle structure but also due to its diverse floral sex types (Tanurdzic and Banks, 2004). In this study, we comprehensively described the function of BES1 and systematically analyzed the relative complete profile of the BES1 gene family by using the bioinformatic method in cucumber plants. We identified the members and phylogenetic relationship of the *CsBES1* gene family; analyzed the gene structure and promoter elements; and carried out the expression pattern of *CsBES1* genes in different organs, under various abiotic, and biotic stresses by quantitative real-time PCR. In addition, these findings will provide an insight for understanding the biological function of CsBES1 and help for molecular breeding of cucumber in the future.

MATERIALS AND METHODS

Plant Material, Growth Conditions, and Stress Treatments

Cucumber (*Cucumis sativus* L.) cultivar "*Xintaimici*" was used for all the experiments. Cucumber seedlings were grown in a controlled-environment growth chamber under normal management (23/28°C, 10/14 h, dark/light, and 70–80% humidity). Seedlings with four mature leaves were treated with 150 mM NaCl or 10% PEG 6000 or transferred to a growth chamber set at 6°C as cold treatment. Leaves and roots were sampled on days 0, 6, 12, 24 h, 2 day, 3 day, 6 day, and 9 day after cold treatment and at 0, 1, 3, 6, 9, 12, and 24 h after NaCl and PEG treatments, respectively. All samples included at least three biological replicates. All samples were harvested, immediately frozen in liquid nitrogen, and stored at -80°C for further study.

Identification of the CsBES1 Gene Family in Cucumber

The BES1 genes and amino acid sequences of cucumber and melon were obtained in the Cucurbit Genomics Network[1] (Huang et al., 2009). The hidden Markov model (HMM) profile of the BES1_N domain (PF05687) was downloaded from the Pfam database (Pfam 29.0)[2] to identify BES1 genes in the cucumber genome using HMMER 3.0 software with an *E*-value cutoff of 1. Arabidopsis BES1 protein sequences[3] were used as a query against the predicted cucumber proteome sequences to identify all cucumber BES1 proteins. The BES1 protein sequences of Arabidopsis, rice, maize, and tomato were searched from the Arabidopsis Information Resource (Huala et al., 2001), the Rice Genome Annotation Project[4] (Kawahara et al., 2013), the

[1]http://cucurbitgenomics.org/organism/2

[2]http://pfam.sanger.ac.uk/

[3]http://www.arabidopsis.org

[4]http://rice.plantbiology.msu.edu/index.shtml

Maize Genetics and Genomics Database[5], and the Sol Genomics Network[6], respectively, for sequence analysis.

Phylogenetic Relationship, Sequence Alignment, and Gene Structure Analyses

To investigate the phylogenetic relationships of *CsBES1* genes, a total of 35 BES1 protein sequences were identified from six plant species including cucumber, Arabidopsis, melon, tomato, rice, and maize. Multiple sequence alignments of the 35 BES1 amino acid sequences were performed with ClustalW (Higgins et al., 1996) using default parameters. A phylogenetic tree, in which the degree of support for a particular grouping pattern was evaluated using bootstrap (1000 replicates) values, was constructed using the software MEGA 5.0 via the maximum-likelihood method (Tamura et al., 2011). The MEME server[7] and Tbtools were employed to analyze the motif composition of BZR1/BES1 in cucumber and Arabidopsis. The protparam tool[8] was use to analyze the physicochemical characteristics of deduced CsBES1 amino acid sequences, including theoretical molecular weight (MW), isoelectric point (pI), and grand average of hydropathicity (GRAVY). The Cell-PLoc 2.0[9] was used to analyze the subcellular localization of 6 CsBES1 amino acid sequences.

Gene Expression Analysis Upon the RNA-Seq Data

The transcription data of *BES1* genes from different tissues (BioProject Name: PRJNA80169; PRJNA271595; and PRJNA258122) and under biotic stresses (BioProject Name: PRJNA376073; PRJNA388584; and PRJNA292785) in cucumber were collected from the Cucurbit genomics expression database[10]. Hi-seq data of 10 cucumber tissues containing root, stem, leaf, male flower, female flower, ovary, expanded fertilized ovary (7 days after flowering), expanded unfertilized ovary (7 days after flowering), base of the tendril, and tendril were used for expression of *BES1*. We used the Reads Per Kilobase per Million mapped reads (RPKM) values of *BES1* genes and then plotted it using the Pheatmap package in R software. Three heat maps of *BES1* genes in different cucumber organs were obtained and three heat maps of *BES1* genes in response to gibberellic acid (GA), powdery mildew, and downy mildew.

RNA Extraction and qRT-PCR

Total RNA was extracted from frozen leaves and roots using RNeasy Plant kit (Huayueyang, Beijing, China) according to the manufacturer's protocol. cDNA was synthesized from 1000 ng RNA using FastKing cDNA Dispelling RT SuperMix Kit (TIANGEN, Beijing, China). cDNA was diluted 1:10 for further quantitative reverse transcription polymerase chain reaction (qRT-PCR). The qRT-PCR reactions for detecting gene expression by using SYBR Premix Ex Taq^{TM} (Tli RNase H Plus; TaKaRa Shuzo, Shiga, Japan) according to the manufacturer's protocol. Relative gene expression was calculated using $\Delta\Delta$CT values obtained from the formulas ΔCT = target CT – reference CT and $\Delta\Delta$CT = treated sample ΔCT – untreated sample ΔCT/$\Delta\Delta$CT = other organ ΔCT – root ΔCT, and the calculation of the gene expression levels followed the $2^{-\Delta\Delta CT}$ method described by Livak and Schmittgen (2001). All qRT-PCR reactions were conducted at at least three biological replications and three technical repeats. The cucumber Tublin gene (Csa4G000580) was used as the reference gene. The primers used for qRT-PCR are listed in **Supplementary Table 1**.

RESULTS

Identification of the BES1 Gene Family in Cucumber and Phylogenetic Relationship Analysis

A total of six *BES1* genes in cucumber were identified by comprehensive analysis in CuGenDB (available online: http://cucurbitgenomics.org/) and PlantTFDB (available online: http://planttfdb.cbi.pku.edu.cn/index.php?sp=Csg). The annotation IDs of *BES1/BZR1* genes in cucumber were Csa1G467200, Csa2G361450, Csa4G083490, Csa4G056530, Csa6G003450, and Csa6G501930. *CsBES1s* were located on chromosomes 1, 2, 4, and 6. To better understand characteristics of *CsBES1/CsBZR1*, we carried out systematically the statistical analysis on the protein physicochemical indices (**Table 1**). The six *CsBES1/CsBZR1* genes varied greatly in ORF length from 936 bp (Csa2G361450) to 2097 bp (Csa6G003450), and their encoded proteins ranged from 311 (Csa2G361450) to 698 (Csa6G003450) amino acids. The molecular weight ranged from 34.19 kDa (Csa2G361450) to 78.35 kDa (Csa6G003450), and the isoelectric point varied from 5.78 (Csa6G003450) to 9.17 (Csa2G361450). In addition, the GRAVY ranged from −0.689 (Csa6G501930) to 0.715 (Csa2G361450), and the subcellular localization predicted that all of the six CsBES1/CsBZR1 were localized in nucleus by Plant-mPLoc 2.0[11].

To establish the functional and evolutionary relationship of the *CsBES1/CsBZR1* gene family, we constructed a phylogenetic tree using the amino acid sequences of six putative cucumber *BES1*, eight Arabidopsis *BES1*, six melon *BES1*, nine tomato *BES1*, six rice *BES1*, and eleven maize *BES1* (**Table 2**). Based on phylogenetic tree classification, all selected BES1 proteins were classified into three groups, in which Csa2G361450 and Csa6G501930, Csa1G467200 and Csa4G083490, and Csa4G056530 and Csa6G003450 belonged to group I, group II, and group III, respectively (**Figure 1**).

Gene Structure and Conserved Motif Analysis in Cucumber

The distribution of exons and introns was the important information to understand gene structure. Thus, we analyzed the structural features of six *CsBES1* genes and eight *AtBES1*

[5]https://www.maizegdb.org/

[6]https://solgenomics.net/organism/

[7]http://meme-suite.org/tools/meme

[8]https://web.expasy.org/protparam/

[9]http://www.csbio.sjtu.edu.cn/bioinf/Cell-PLoc-2/

[10]http://cucurbitgenomics.org/rnaseq/home

[11]http://www.csbio.sjtu.edu.cn/bioinf/Cell-PLoc-2/

TABLE 1 | The characteristics of BZR1/BES1 genes from cucumber.

Genes	Locus name	Chr.	Location	Subcellular localization	ORF length	Protein physicochemical characteristics			
						Length (aa)	MW (KDa)	pI	GRAVY
CsBES1-1	Csa1G467200.1	1	16768701∼16771593 (-)	N	984	327	35.35	8.97	−0.644
CsBES1-2	Csa2G361450.1	2	17199641∼17201363 (+)	N	936	311	34.19	9.17	0.715
CsBES1-3	Csa4G056530.1	4	4704832∼4711322 (+)	N	2010	669	75.26	5.91	−0.404
CsBES1-4	Csa4G083490.1	4	5606824∼5610735 (+)	N	978	325	34.68	8.50	−0.572
CsBES1-5	Csa6G003450.1	6	340459∼346769 (-)	N	2097	698	78.35	5.78	−0.441
CsBES1-6	Csa6G501930.1	6	25188710∼25191519 (+)	N	960	319	34.64	8.96	−0.689

Abbreviation: Chr., chromosome; N, nucleus; Length (aa), length of amino acid; MW, molecular weight; pI, isoelectric point; and GRAVY, grand average of hydropathicity.

genes, including the number of exons and introns, and the length of exons, introns, and untranslated regions (**Figure 2**). The gene size varied from ∼1.5 kb in At3G50750 to 7.0 kb in At2G45880 and Csa4G056530. The structure analysis of *BES1/BZR1* genes from cucumber and Arabidopsis showed that there were three different types of structure distribution: the majority of *BES1/BZR1* genome contained two exons and one intron. However, only At1G19350 contained three exons and two introns, while At2G45880, At5G45300, Csa4G056530, and Csa6G003450 included ten exons and nine introns. Structural variation in *BES1* genes might be generated due to the gene duplication or incorporation after transcription during evolutionary history. These results suggested that different *BES1* genes may be functionally diversified.

To better understand the diversity and similarity of CsBES1 proteins, the conserved motifs were analyzed. As shown in **Figure 3**, a schematic overview of these conserved motifs was provided and demonstrated that all BES1 from cucumber and Arabidopsis included at least four conserved motifs located in N- and C-terminals. DUF822, also known as BES1_N (turquoise and red in **Figure 3** and **Supplementary Figure 1A**), was found in all CsBES1 and AtBES1 protein sequences. Besides, two glycosylation-conserved sites (yellow and purple colors in **Figure 3** and **Supplementary Figure 1B**) were found in At2G45880, At5G45300, Csa4G056530, and Csa6G003450.

Analysis of *Cis*-Elements in the CsBES1 Promoters

To better understand the potential regulatory mechanisms of *CsBES1* genes during growth and development stages, under abiotic or biotic stresses and in response to different hormones, we identified the presence of *cis*-elements in the promoter regions of *CsBES1* genes. A 2.0-kb promoter region of all the *CsBES1* genes was scanned, and the *cis*-elements, which could possibly control the expression of *CsBES1*, were predicted using PLANTCARE tools. The results are shown in **Figure 4**.

TABLE 2 | Identified members of the BES1/BZR1 gene family in different plants.

Species	Gene family (number)
Arabidopsis thaliana	*AtBES1*, *BZR1* and *BEH1* (8)
Cucumis sativus	*CsBZR1/BES1* (6)
Cucumis melo	*CmBZR1/BES1* (6)
Solanum lycopersicum	*SlBZR1/BES1/LAT61* (9)
Oryza sativa	*OsBZR1* (6)
Zea mays	*ZmBES1/BZR1* (11)

For stress-related elements, **(1)** the ARE element was found in four selected promoter regions of Csa1G467200, Csa2G361450, Csa6G003450, and Csa6G501930, among which two ARE elements were identified in Csa1G467200 and Csa6G003450. **(2)** The TC-rich repeat element was found in five selected promoter regions of Csa1G467200, Csa2G361450, Csa4G056530, Csa6G003450, and Csa6G501930, and at least two TC-rich repeats elements were predicted in Csa6G501930. **(3)** The LTR element was only found in two promoter regions of Csa6G003450 and Csa6G501930. **(4)** The GC-motif and MBS elements were only found in Csa4G083490 and Csa4G056530, respectively. This suggested that *CsBES1* family genes played different roles in response to various stresses. For example, Csa6G003450 and Csa6G501930 with LTR elements might respond to low-temperature stress, while Csa4G056530 with MBS elements might respond to drought stress.

For hormone-related elements, **(1)** the ABRE element was found in five promoter regions of Csa1G467200, Csa2G361450, Csa4G083490, Csa6G003450, and Csa6G501930, of which at least four ABRE elements were found in Csa1G467200, Csa4G083490, and Csa6G501930. Since ABRE is an important ABA response element, *CsBES1* is probably involved in the ABA signal pathway. **(2)** The CGTCA-motif and TGACG-motif elements were found in five promoter regions of Csa1G467200, Csa2G361450, Csa4G056530, Csa6G003450, and Csa6G501930, implying that *CsBES1* may respond to jasmonic acid. **(3)** The TATC box and TGA element were found only in Csa6G501930, indicating that Csa6G501930 might be involved in the GA or auxin pathway.

For growth and development-related elements, **(1)** there were at least three BOX4 elements predicted in all six CsBES1 promoter regions. **(2)** The GT1 motif was found in promoter regions of Csa1G467200, Csa4G056530, Csa4G083490, and Csa6G003450. **(3)** The HD-Zip and

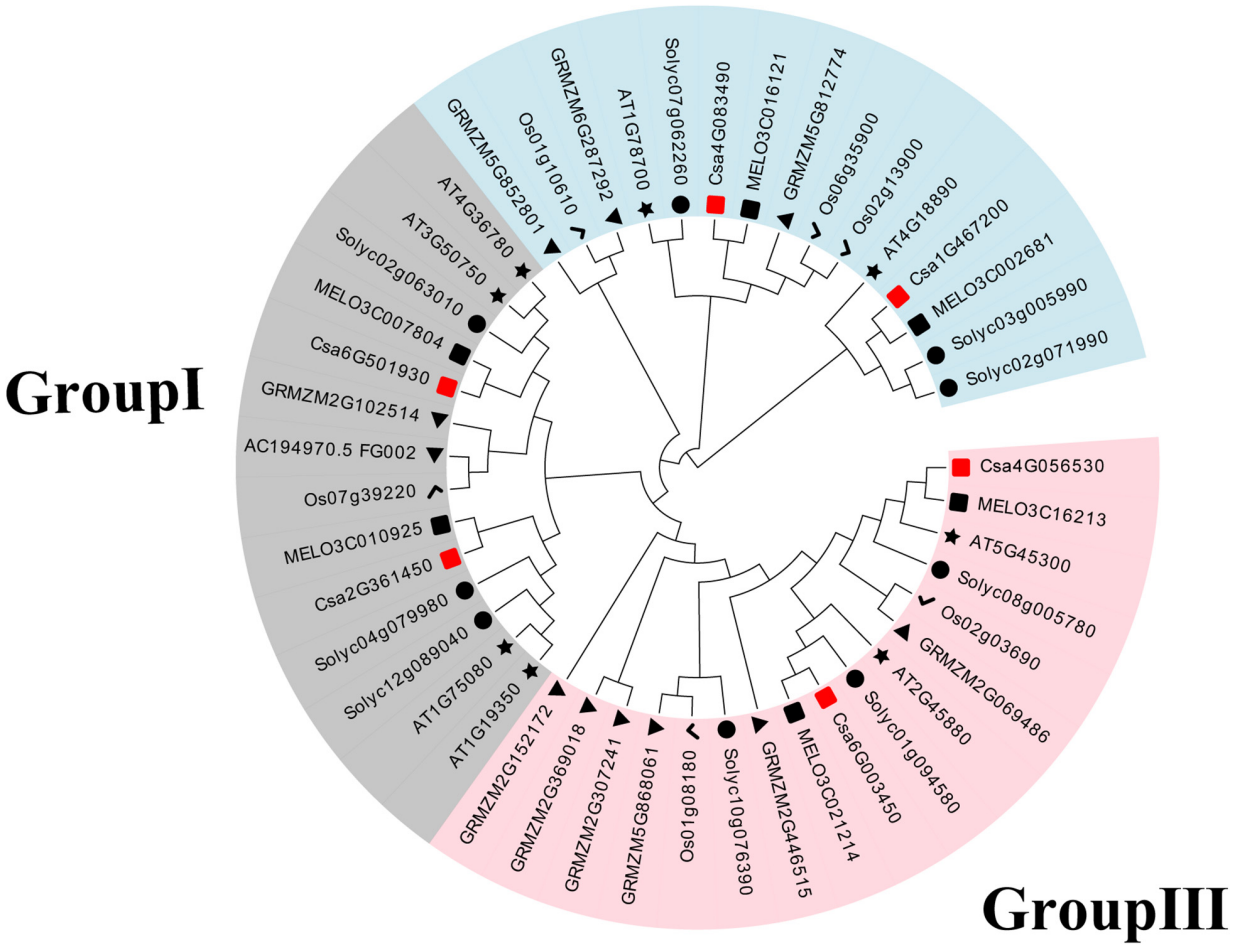


FIGURE 1 | Phylogenetic relationships in BES1 proteins from cucumber (Cs), Arabidopsis (AT), melon (MELO), tomato (Soly), maize (GRMZM), and rice (Os). The neighbor-joining tree was generated including 6 BES1 proteins from cucumber, 8 from Arabidopsis, 6 from melon, 9 from tomato, 11 from maize, and 6 from rice.

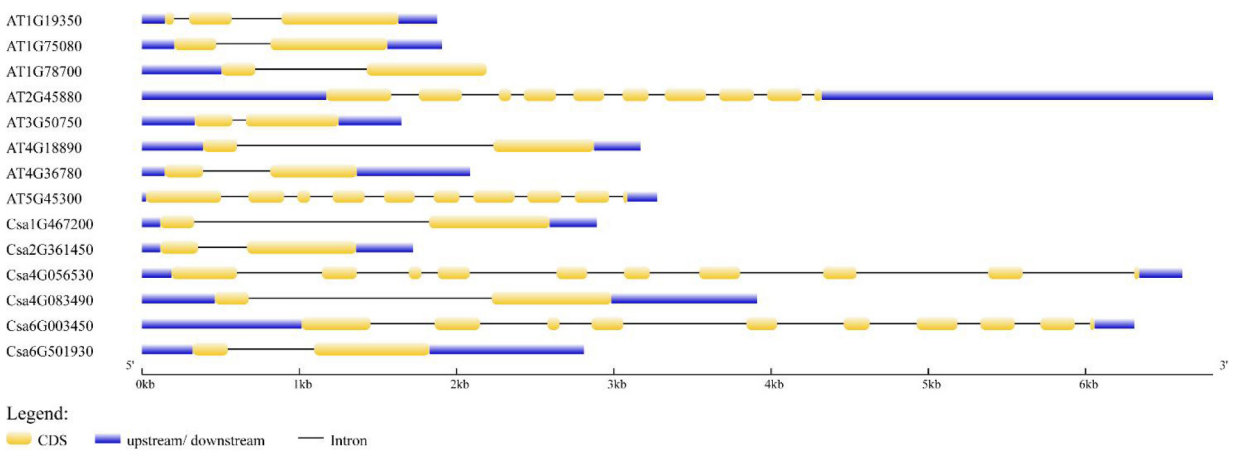


FIGURE 2 | Exon and intron structures of BEs1 in *Cucumis sativus* and Arabidopsis thaliana. Exons are shown as yellow boxes; introns were shown as black lines. Upstream and downstream untranslated regions are shown as blue boxes. Scales are shown at the bottom of the diagram.

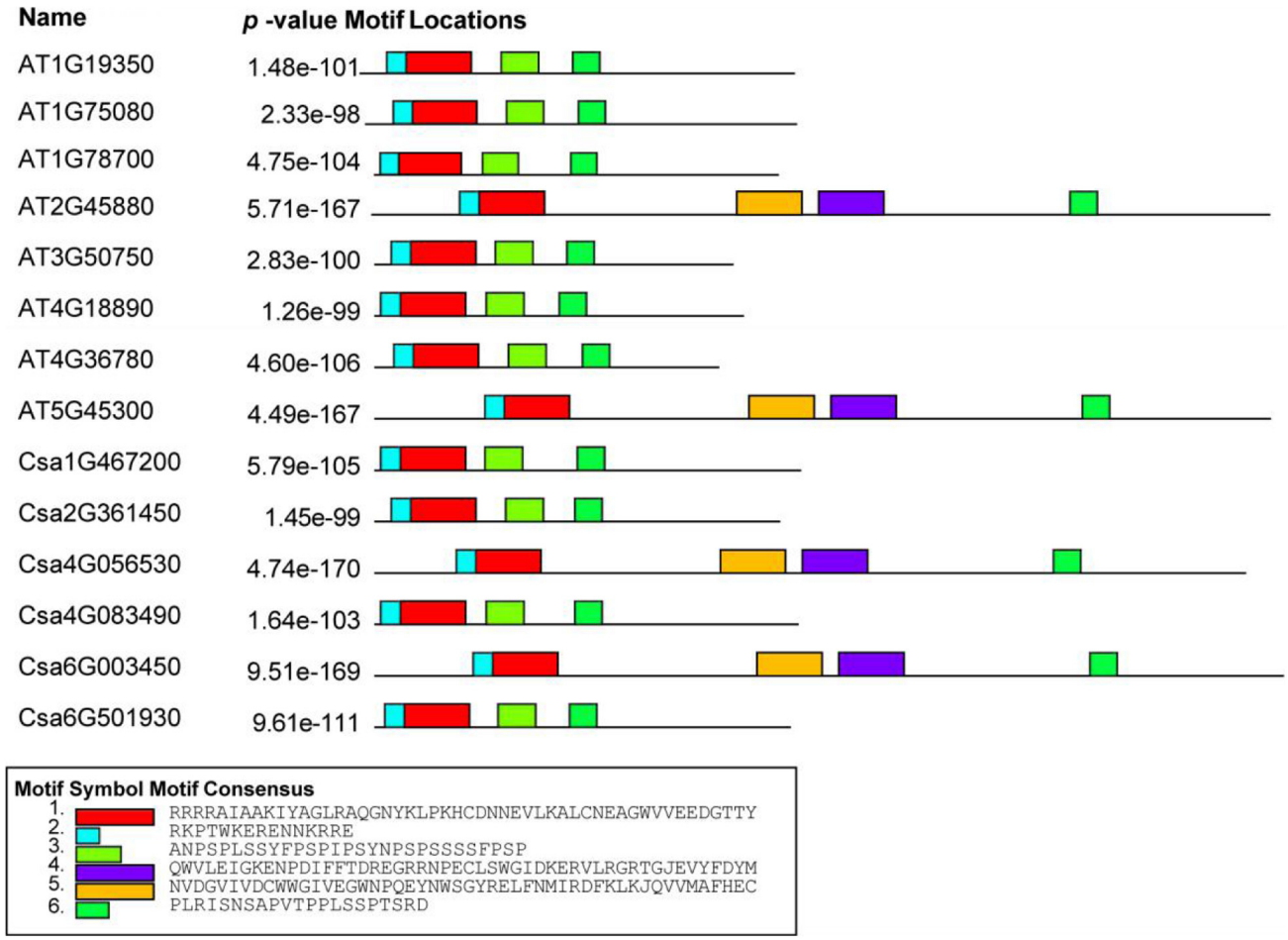


FIGURE 3 | Conserved motif analysis of BES1 amino acid from Arabidopsis and *Cucumis sativus*. Motif analyses were conducted using MEME online software and Tbtools as described in the method. Different color boxes represent various types of conserved motifs. The gene IDs were listed on the left part of this figure. The amino acid sequences of key motifs are listed at the bottom of this figure.

O2-site elements were only found in the promoter region of Csa1G467200.

Expression Patterns of BES1 in Cucumber Various Organs

To explore the possible functional roles of *CsBES1s* in developmental processes, the expression patterns of *CsBES1s* in root (R), stem (S), leaf (L), female flower (FF), male flower (MF), and fruit (F) were performed by qRT-PCR. As shown in **Figure 5**, Csa1G467200 and Csa2G361450 had a relatively low expression level in L and F, while Csa2G361450 showed the highest expression in flowers, especially in female flowers. Csa4G083490 was expressed highly in roots and fruits compared to its low expression in stems. Csa4G056530 and Csa6G003450 exhibited a similar expression pattern and had relatively high expression levels in stems and fruits compared with other tissues, and they did also respond similarly to some, although not all, of the abiotic stresses as analyzed in **Figures 6**, **7**. Interestingly, it is notable that Csa4G056530 and Csa6G003450 were from Group III of the phylogenetic tree and shared highly similar and distinct gene structures (nine exons compared with two exons for all other BES1 genes) and the particular long glycosylation-conserved site (**Figures 2**, **3**).

The RNA-sequencing database from the Cucurbit genomics database showed that the expression profiles of most *CsBES1*s were consistent with the results obtained by qRT-PCR except Csa2G361450. The Csa2G361450 transcription level showed the highest expression in flowers upon qRT-PCR data (**Figure 5**), but its expression was much higher in stems from the RNA-Seq database (**Supplementary Figure 2**). It is reasonable due to samples harvested in two different cucumber species and at different sampling times. Moreover, the expression profile of *CsBES1*s in different root development zones (differentiation, elongation, and meristematic zones) and fruit types (long and short fruits) were determined. The database indicated that Csa6G501930 was highly expressed in all tissues compared to other family genes, and in elongation zones and long fruits compared to other tissues (**Supplementary Figures 2B,C**). These

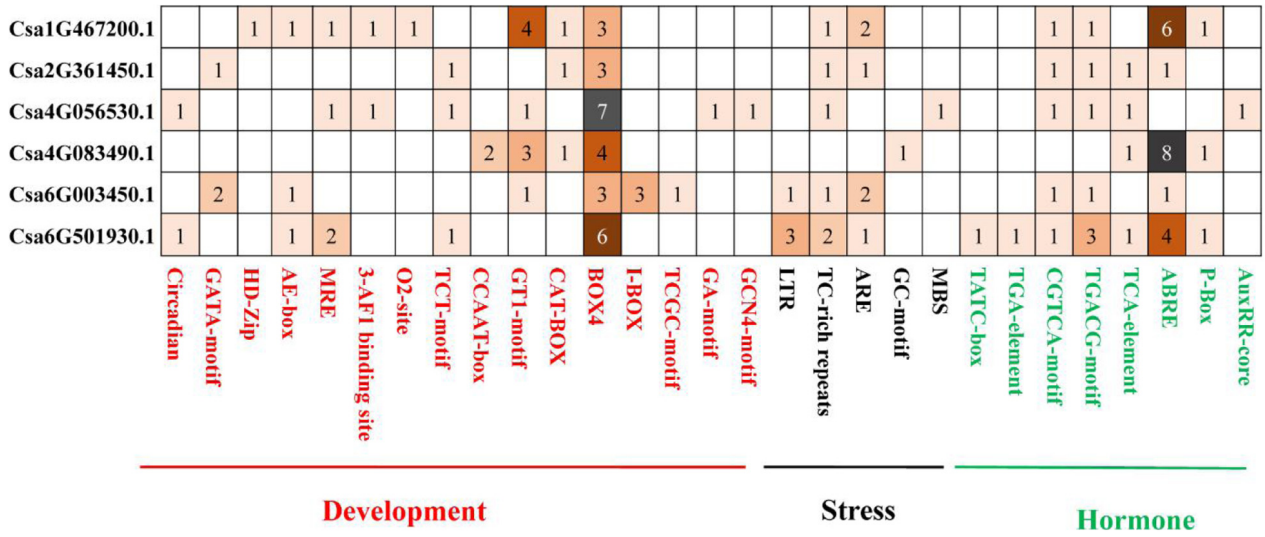


FIGURE 4 | Distribution of development and stress- and hormone-related cis elements in cucumber BES1 promoters. ~2.0-kb promoter sequences of six CsBES1 genes were predicted. The gene IDs were shown in the left part of this figure. The elements were shown at the bottom of this figure. The number in squares showed the number of elements in promoter regions.

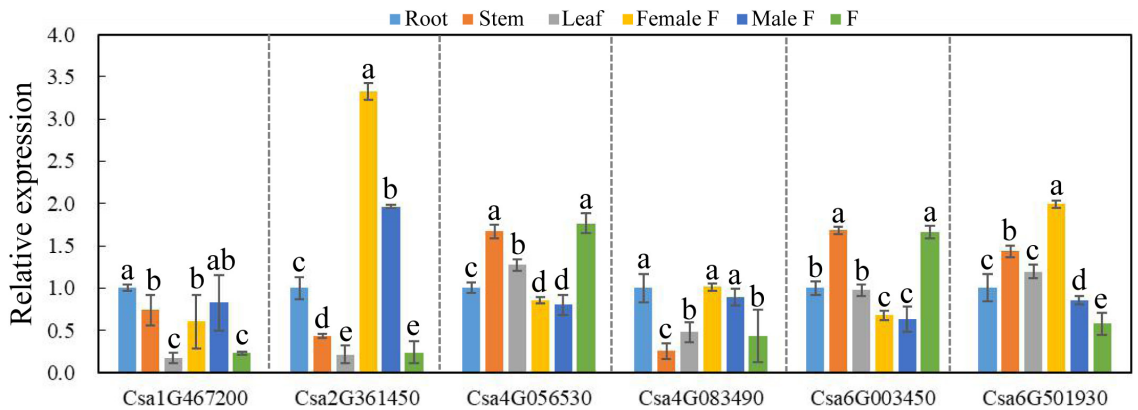


FIGURE 5 | The expression pattern of CsBES1 genes in different cucumber organs. The Cucumber Tublin gene was the reference gene used to calibrate the relative gene expression. Error bars represent ± SE (n = 3). One-way analysis of variance was conducted by Duncan's new multiple-range test, n = 3; different letters above the bars indicate significant differences ($P < 0.05$). Abbreviation: Female F, female flower; Male F, male flower; and F, fruit.

results suggested the various expression patterns of each BES1 in different cucumber organs and indicated that they probably play different roles in cucumber different organs.

Expression Pattern of BES1 Under Different Abiotic Stress Conditions

To elucidate the functions of *CsBES1*s in response to hormones and under different stresses (abiotic and biotic stresses), the expression profiles of *CsBES1*s in leaves and roots were conducted by qRT-PCR (**Figures 6**, **7**) and the RNA-sequence database (**Supplementary Figure 3**). We tested the expression level of *CsBES1*s in leaves and roots at 0, 1, 3, 6, 9, 12, and 24 h after NaCl (150 mM) and drought (20% PEG 6000) treatments, and at 0, 6, 12, 24, 2 day, 3 day, 6 day, and 9 day after cold (6°C) treatment.

After drought treatment, the expression of Csa1G467200, Csa2G361450, and Csa4G083490 in leaves significantly increased at 6 h and then decreased compared to the control, while the expression of Csa6G003450 and Csa6G501930 kept a relatively stable level until 6 h and then decreased about 4 folds after 9 h of treatment. The expression of Csa4G056530 had no obvious change after treatment in leaves (**Figure 6A**). In roots, all the *CsBES1* genes had a sharp increase at 9 h to 24 h after treatment. In particular, Csa1G467200, Csa6G501930, and Csa6G003450 attached the peak point at 24 h, while Csa4G056530 and Csa4G083490 showed the highest point at 12 h after treatment (**Figure 7A**).

After NaCl treatment, Csa1G467200 showed an up-regulation between 8 and 10 folds at 3 h and 6 h after treatment compared to control in leaves; Csa2G361450 had a higher expression at 12 h

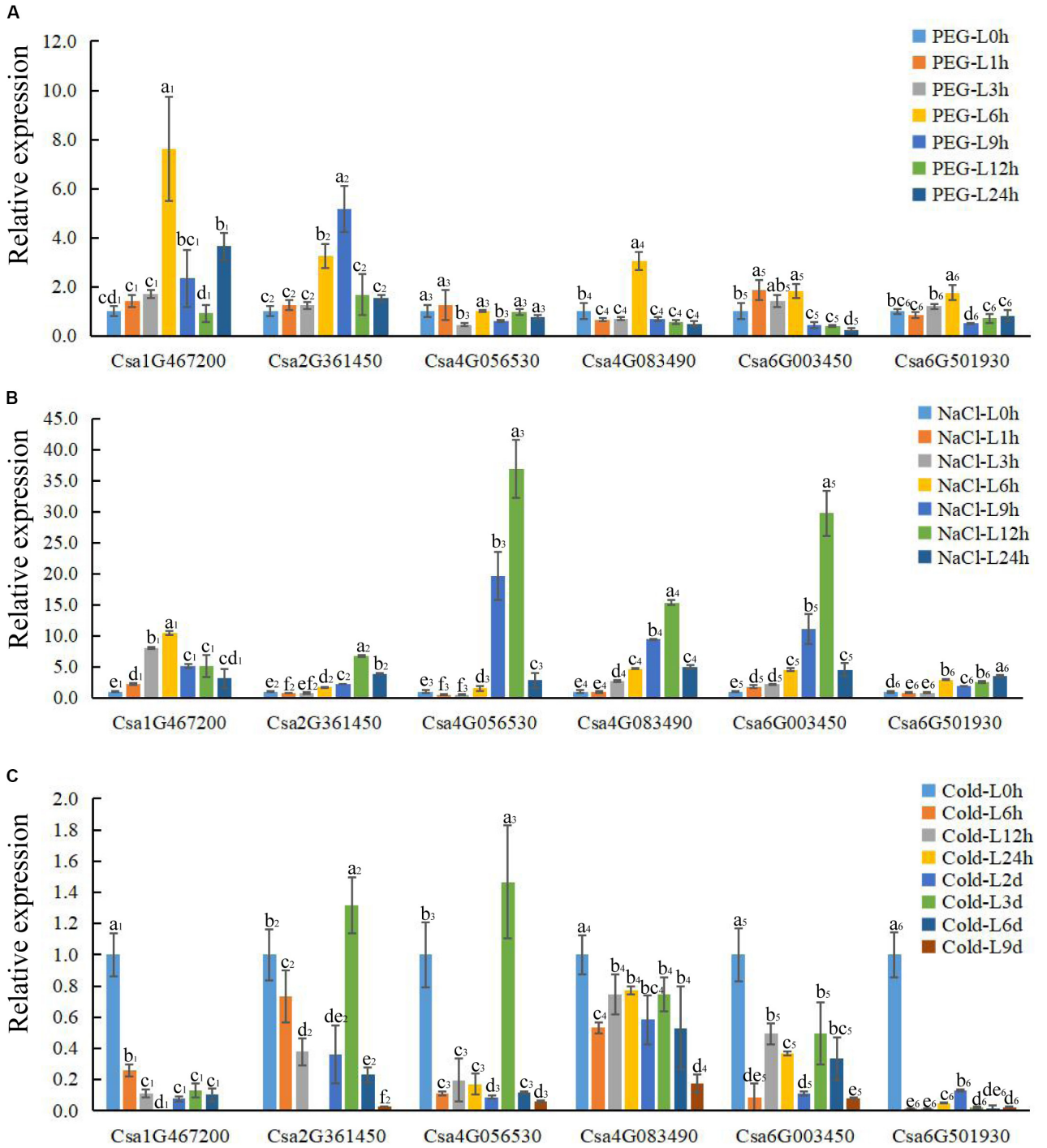


FIGURE 6 | The expression profile of CsBES1 genes in cucumber leaves in response to different abiotic stresses. **(A)** Drought treatment (20% PEG6000); **(B)** salt treatment (150 mM NaCl); and **(C)** cold treatment (6°C). Error bars represent ± SE (n = 3). One-way analysis of variance was conducted by Duncan's new multiple-range test, n = 3; different letters above the bars indicate significant differences ($P < 0.05$). Leaves were harvested at 0, 1, 3, 6, 9, 12, and 24 h after drought and salt treatments; the times of cold treatment were 0, 6, 12, 24 h, 2 day, 3 day, 6 day, and 9 day.

and 24 h compared with control; Csa4G056530, Csa4G083490, and Csa6G003450 increased sharply at 9 h and 12 h after treatment; and Csa6G501930 exhibited a slight upregulation from 6 h to 24 h after treatment (**Figure 6B**). In roots, the expression of Csa1G467200, Csa6G501930, and Csa6G003450 increased steeply at 1 h after treatment; Csa2G361450 had a high expression at 6 h, and the Csa4G056530 expression attached the peak point at 24 h after treatment compared to

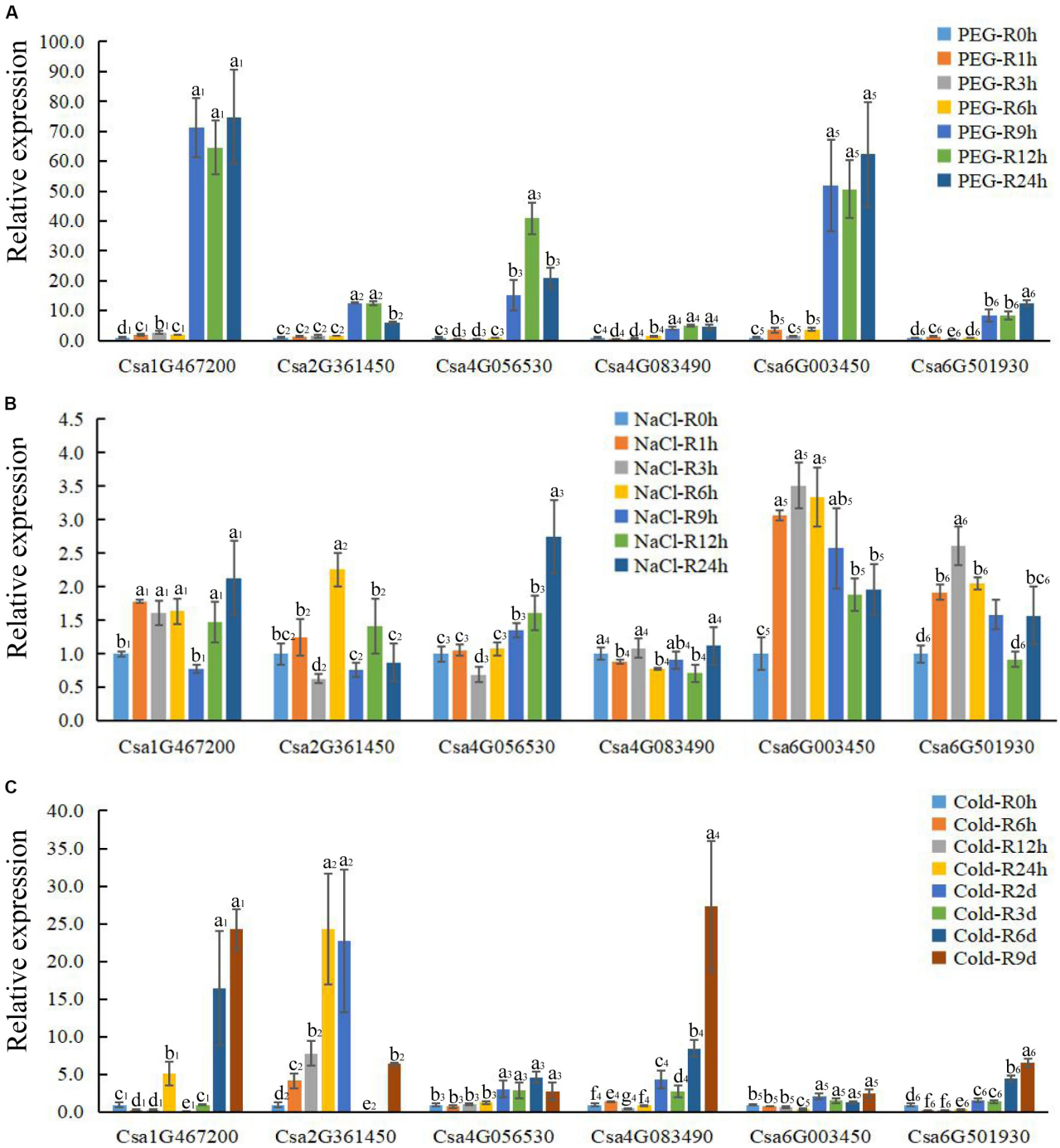


FIGURE 7 | The expression profile of CsBES1 genes in cucumber roots in response to different abiotic stresses. **(A)** Drought treatment (20% PEG6000); **(B)** salt treatment (150 mM NaCl); and **(C)** cold treatment (6°C). Error bars represent ± SE (n = 3). One-way analysis of variance was conducted by Duncan's new multiple-range test, n = 3; different letters above the bars indicate significant differences ($P < 0.05$). Roots were harvested at 0, 1, 3, 6, 9, 12, and 24 h after drought and salt treatments; times of cold treatment at 0, 6, 12, 24 h, 2 day, 3 day, 6 day, and 9 day.

the control; and Csa4G083490 showed no difference at all time points (**Figure 7B**).

After cold treatment, in leaves, Csa1G467200, Csa4G083490, Csa6G003450, and Csa6G501930 showed a decrease at all time points compared to the control; Csa2G361450 and Csa4G056530 showed the highest point at 3 day and then sharply decreased compared with control (**Figure 6C**). In roots, the expression of Csa4G056530, Csa4G083490, Csa6G003450, and Csa6G501930 was up-regulated at 2 day until to 9 day after treatment. The expression of Csa1G467200 and Csa2G361450

had similar changing patterns: increase–decrease–increase. Specifically, Csa1G467200 expressed an increase at 24 h, 6 day, and 9 day, while Csa2G361450 showed an increase at 6 h to 2 day and 9 day after treatment (**Figure 7C**).

In addition, we explored the expression pattern of *CsBES1* genes in leaves after GA treatment, downy mildew treatment, and powdery mildew treatment based on the RNA-sequence database. The results showed that Csa6G501930 had the highest expression at 6 h after GA treatment, and Csa2G361450 and Csa4G083490 had an increasing and decreasing trend after GA treatment, respectively, (**Supplementary Figure 3A**). The downy mildew and powdery mildew treatments lead to an overall decrease in Csa6G501930. Csa2G361450 was up-regulated at 6 h after downy mildew treatment, and Csa6G003450 showed an increase at 48 h after powdery mildew treatment (**Supplementary Figure 3B**).

DISCUSSION

The BES1 gene family has been previously reported in several plants, including Arabidopsis, tomato, Chinese cabbage, maize, and *Brassica napus* (Yin et al., 2005; Wu et al., 2016; Gao et al., 2018; Song et al., 2018; Yu et al., 2018; Cao et al., 2020). To date, the regulation mechanisms of BES1 transcription factors have been identified and elucidated in Arabidopsis and rice (Yin et al., 2005; Bai et al., 2012). However, little information is available regarding the BES1 family in cucumber. Therefore, this study conducted a genome-wide analysis of BES1 genes and proteins in cucumber. Based on phylogenetic relationship, conserved motifs, and gene structures analysis, it was found that there were six *CsBES1* genes distributed in three groups. Csa4G056530 and Csa6G003450, which belonged to group III, included ten exons and nine introns and possessed a particular long glycosylation conserved site. Accordingly, they had the longest amino acids and the largest molecular weight. By contrast, other four CsBES1s only included two exons and one intron. Specifically, Csa1G467200 and Csa4G083490 had a long segment of intron and were assigned to group II, while Csa2G361450 and Csa6G501930 contained a short segment of intron and were assigned to group I. Meanwhile, all of the CsBES1 proteins had one conserved DUF822 motif. These findings indicated that some intron (or motif) gain or loss events might be occurred during the evolution process in *CsBES1* genes and that various structures probably endowed these *CsBES1* genes with different functions.

The plant *cis*-element is an important molecular switch to regulate gene transcription during plant growth and development as well as in their responses to stresses and hormones. According to bioinformatics analysis, various *cis*-elements were found in *CsBES1* promoters, in which some gene specific *cis*-element was predicted. Notably, TATC-box (GA response element), and TGA-element (auxin response element) were only found in the Csa6G501930 promoter. This result was consistent with the transcription database, which showed that Csa6G501930 expression could be induced by GA (**Supplementary Figure 3A**). In the GA signaling pathway, RGA is a transcription factor belonging to the DELLA family protein. In Arabidopsis, the RGA protein could decrease the abundance and transcription level of AtBZR1. AtBZR1 (At1G75080) could interact with RGA *in vitro* and *in vivo* to participate the BR and GA signaling pathway (Li et al., 2012). In addition, AtBZR1 (At1G75080) also belonged to Group I similar to Csa6G501930 (**Figure 1**), and they had a similar gene structure (**Figure 2**) and conserved motif (**Figure 3**), which indicates that Csa6G501930 might participate in the brand GA signaling pathway as well as AtBZR1 (At1G75080). Auxin response factors (ARFs) are key transcription factors that regulate auxin-responsive gene expression. ARF6 and ARF7 have been identified to interact with BZR1 to coordinate in the growth and development of plants through the auxin and BR signaling pathway (Youn et al., 2016). Besides, many stress-related *cis*-elements were also predicted in *CsBES1* promoters, such as cold response element (LTR), and drought response element (MBS). In Arabidopsis, AtBES1 (At1G19350) was targeted for autophagy-induced degradation through directly interacting with DOMINANT SUPPRESSOR OF KAR 2 (DSK2) to respond abiotic stresses (Nolan et al., 2017). Together, these results provide referable evidence that CsBES1 may play important roles in the hormone regulation pathway and stress response in cucumber.

In this study, several experiments have been performed to characterize the expression profiles of CsBES1 genes in different cucumber organs in response to various abiotic stresses and hormones. It was noted that Csa2G361450 was highly expressed in flowers and showed a closer relationship with either At1G19350 or At1G75080 according to the phylogenetic tree (**Figure 5**). At1G75080 (BZR1) has been proven to be involved in photoperiodic flowering by regulating its target gene, BR ENHANCED EXPRESSION 1 (BEE), which directly binds to the FLOWERING LOCUTS T (FT) promoter to activate the transcription of FT and promote flowering initiation (Wang et al., 2019). Another study reported that AtBES1 (At1G19350) could directly bind to the promoter regions of *SPL/NZZ*, *TDF1*, *AMS*, *MS1*, and *MS2* genes, which are important transcription factors controlling anther and pollen development (Ye et al., 2010). These results suggested that Csa2G361450 probably participates in flower development. The transcription database showed that GA treatment induced an increase in the expression of Csa6G501930 and Csa2G361450 (**Supplementary Figure 3A**). BES1/BZR1, as a class of master transcription factor in the BR signaling pathway, mediated a direct cross talk between GA and other plant hormones (Li et al., 2018). In rice, OsBZR1 (Os07g39220) directly binds to the promoters of GA biosynthesis-related genes to regulate GA content for promoting cell elongation (Tong et al., 2014). Moreover, Csa6G501930 and Csa2G361450 had a close relationship in Group I based on phylogenetic tree analysis (**Figure 1**). This result suggested that Csa6G501930 and Csa2G361450 are probably involved in the GA signaling pathway.

In plants, BR has been universally recognized as an important phytohormone involved in a variety of stress responses (Krishna, 2003). Accordingly, BES1/BZR1, which control BR-regulated gene expression, has also been demonstrated to be critical in mediating plant responses to abiotic stresses including salt, drought, cold, heat, and freezing (Nolan et al., 2020). In this study, after drought treatment, Csa1G467200 expression

was induced in both leaves and roots, while Csa4G056530 and Csa6G003450 were induced only in roots. Csa6G003450 expression showed a significant increase in leaves but not in roots after salt treatment. However, the expression of Csa4G083490 was decreased in leaves but increased in root after cold treatment. It was interesting that Csa1G467200, Csa6G003450, and Csa6G501930 were reduced in leaves but induced in roots after cold treatment (**Figures 6**, **7**). These results revealed the different expression patterns of *CsBES1* genes under various abiotic stresses, thereby indicating that these genes may play different but essential roles in response to abiotic stresses.

CONCLUSION

In this study, we systematically analyzed the structures, conserved motifs, and phylogenetic relationships of *CsBES1* genes and their expression profiles in different cucumber organs under various abiotic stresses. A total of six putative *CsBES1* genes were identified in cucumber. These six genes exhibited different expression patterns in cucumber organs and response to variously abiotic stresses. The comprehensive analysis on *CsBES1* was expected to provide basic information for further studying the functions of *CsBES1* and for the molecular breeding of cucumber.

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/**Supplementary Material.**

AUTHOR CONTRIBUTIONS

LG, YT, and SM conceived and designed the experiments. SM, TJ, ML, and SL performed the experiments and collected the data. SM and TJ executed the data analyses. All authors contributed to the interpretation of the results. SM, LG, and YT wrote the drafts.

FUNDING

This work was financially supported by the National Key Research and Development Program of China (2019YFD1001903) and the China Agriculture Research System (CARS-23).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.583996/full#supplementary-material

Supplementary Figure 1 | Multiple alignments of BES1 domains from Arabidopsis and cucumber amino acid sequence.

Supplementary Figure 2 | The expression profile of CsBES1 genes in different cucumber organs based on RNA-sequence database.

Supplementary Figure 3 | Transcription data were used to analyze the expression profile of CsBES1 genes in cucumber leaves in response to different abiotic, biotic stresses and hormone.

Supplementary Table 1 | Primers used in the study.

REFERENCES

Bai, M. Y., Shang, J. X., Oh, E., Fan, M., Bai, Y., Zentella, R., et al. (2012). Brassinosteroid, gibberellin and phytochrome impinge on a common transcription module in *Arabidopsis*. *Nat. Cell Biol.* 14, 810–817. doi: 10.1038/ncb2546

Cao, X., Khaliq, A., Lu, S., Xie, M., Ma, Z., Mao, J., et al. (2020). Genome-wide identification and characterization of the BES1 gene family in apple (*Malus domestica*). *Plant Biol.* 22, 723–733. doi: 10.1111/plb.13109

Gampala, S. S., Kim, T. W., He, J. X., Tang, W., Deng, Z., and Wang, Z. Y. (2007). An essential role for 14-3-3 proteins in brassinosteroid signal transduction in *Arabidopsis*. *Dev. Cell* 13, 177–189. doi: 10.1016/j.devcel.2007.06.009

Gao, Y., Hu, J., Zhao, T., Xu, X., Jiang, J., and Li, J. (2018). Genome-wide identification and expression pattern analysis of BRI1-EMS-suppressor transcription factors in tomato under abiotic stresses. *J. Am. Soc. Hortic. Sci.* 143, 84–90. doi: 10.21273/jashs04312-17

Guo, R., Qian, H., Shen, W., Liu, L., Zhang, M., Cai, C., et al. (2013). BZR1 and BES1 participate in regulation of glucosinolate biosynthesis by brassinosteroids in *Arabidopsis*. *J. Exp. Bot.* 64, 2401–2412. doi: 10.1093/jxb/ert094

Higgins, D. G., Thompson, J. D., and Gibson, T. J. (1996). Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.* 266, 383–402. doi: 10.1016/s0076-6879(96)66024-8

Huala, E., Dickerman, A. W., Garcia-Hernandez, M., Weems, D., Reiser, L., and Mueller, L. A. (2001). The *Arabidopsis* information resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res.* 29, 102–105. doi: 10.1093/nar/29.1.102

Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., and Li, S. (2009). The genome of the cucumber, *Cucumis sativus* L. *Nat. Genet.* 41, 1275–1281.

Kawahara, Y., De La Bastide, M., Hamilton, J. P., and Kanamori, H. (2013). Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* 6:4. doi: 10.1186/1939-8433-6-4

Kim, T. W., Guan, S., Burlingame, A. L., and Wang, Z. Y. (2011). The CDG1 kinase mediates brassinosteroid signal transduction from BRI1 receptor kinase to BSU1 phosphatase and GSK3-like kinase BIN2. *Mol. Cell.* 43, 561–571. doi: 10.1016/j.molcel.2011.05.037

Kim, T. W., and Wang, Z. Y. (2010). Brassinosteroid signal transduction from receptor kinases to transcription factors. *Annu. Rev. Plant Biol.* 61, 681–704. doi: 10.1146/annurev.arplant.043008.092057

Krishna, P. (2003). Brassinosteroid-mediated stress responses. *J. Plant Growth Regul.* 22, 289–297. doi: 10.1007/s00344-003-0058-z

Krishna, P., Prasad, B. D., and Rahman, T. (2017). Brassinosteroid action in plant abiotic stress tolerance. *Methods Mol. Biol.* 1564, 193–202. doi: 10.1007/978-1-4939-6813-8_16

Li, J., Wen, J., Lease, K. A., Doke, J. T., Tax, F. E., and Walker, J. C. (2002). BAK1, an *Arabidopsis* LRR receptor-like protein kinase, interacts with BRI1 and modulates brassinosteroid signaling. *Cell* 110, 213–222. doi: 10.1016/s0092-8674(02)00812-7

Li, Q. F., Lu, J., Yu, J. W., Zhang, C. Q., He, J. X., and Liu, Q. Q. (2018). The brassinosteroid-regulated transcription factors BZR1/BES1 function as a coordinator in multisignal-regulated plant growth. *Biochim. Biophys. Acta Gene Regul. Mech.* 1861, 561–571. doi: 10.1016/j.bbagrm.2018.04.003

Li, Q. F., Wang, C., Jiang, L., Li, S., Sun, S. S., and He, J. X. (2012). An interaction between BZR1 and DELLAs mediates direct signaling crosstalk between brassinosteroids and gibberellins in *Arabidopsis*. *Sci. Signal.* 5:ra72. doi: 10.1126/scisignal.2002908

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

Lozano-Durán, R., Macho, A. P., Boutrot, F., Segonzac, C., Somssich, I. E., and Zipfel, C. (2013). The transcriptional regulator BZR1 mediates trade-off between plant innate immunity and growth. *eLife* 2:e00983.

Marcos, R., Izquierdo, Y., Vellosillo, T., Kulasekaran, S., Cascón, T., Hamberg, M., et al. (2015). 9-lipoxygenase-derived oxylipins activate brassinosteroid signaling to promote cell wall-based defense and limit pathogen infection. *Plant Physiol.* 169:2324.

Miyaji, T., Yamagami, A., Kume, N., Sakuta, M., Osada, H., Asami, T., et al. (2014). Brassinosteroid-related transcription factor BIL1/BZR1 increases plant resistance to insect feeding. *Biosci. Biotechnol. Biochemi.* 78, 960–968. doi: 10.1080/09168451.2014.910093

Mora-García, S., Vert, G., Yin, Y., Caño-Delgado, A., Cheong, H., and Chory, J. (2004). Nuclear protein phosphatases with kelch-repeat domains modulate the response to brassinosteroids in *Arabidopsis*. *Genes Dev.* 18, 448–460. doi: 10.1101/gad.1174204

Nolan, T. M., Brennan, B., Yang, M., Chen, J., Zhang, M., Li, Z., et al. (2017). Selective autophagy of BES1 mediated by DSK2 balances plant growth and survival. *Dev. Cell* 41, 33–46. doi: 10.1016/j.devcel.2017.03.013

Nolan, T. M., Vukašinović, N., Liu, D., Russinova, E., and Yin, Y. (2020). Brassinosteroids: multidimensional regulators of plant growth, development, and stress responses. *Plant Cell* 32:295. doi: 10.1105/tpc.19.00335

Oh, E., Zhu, J. Y., and Wang, Z. Y. (2012). Interaction between BZR1 and PIF4 integrates brassinosteroid and environmental responses. *Nat. Cell Biol.* 14, 802–809. doi: 10.1038/ncb2545

Ryu, H., Kim, K., Cho, H., and Hwang, I. (2010). Predominant actions of cytosolic BSU1 and nuclear BIN2 regulate subcellular localization of BES1 in brassinosteroid signaling. *Mol. Cells* 29, 291–296. doi: 10.1007/s10059-010-0034-y

Song, X., Ma, X., Li, C., Hu, J., Yang, Q., and Wang, X. (2018). Comprehensive analyses of the BES1 gene family in *Brassica napus* and examination of their evolutionary pattern in representative species. *BMC Genom.* 19:346. doi: 10.1186/s12864-018-4744-4

Sun, Y., Fan, X. Y., Cao, D. M., Tang, W., He, K., and Wang, Z. Y. (2010). Integration of brassinosteroid signal transduction with the transcription network for plant growth regulation in *Arabidopsis*. *Dev. Cell* 19, 765–777. doi: 10.1016/j.devcel.2010.10.010

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28, 2731–2739. doi: 10.1093/molbev/msr121

Tang, W., Kim, T., Oses-Prieto, J. A., Sun, Y., Deng, Z., Zhu, S., et al. (2008). BSKs mediate signal transduction from the receptor kinase BRI1 in *Arabidopsis*. *Science* 321, 557–560. doi: 10.1126/science.1156973

Tang, W., Yuan, M., Wang, R., Yang, Y., Wang, C., and Wang, Z. Y. (2011). PP2A activates brassinosteroid-responsive gene expression and plant growth by dephosphorylating BZR1. *Nat. Cell Biol.* 13, 124–131. doi: 10.1038/ncb2151

Tanurdzic, M. B., and Banks, J. A. (2004). Sex-determining mechanisms in land plants. *Plant Cell* 16, S61–S71.

Tong, H., Xiao, Y., Liu, D., Gao, S., Liu, L., and Chu, C. (2014). Brassinosteroid regulates cell elongation by modulating gibberellin metabolism in rice. *Plant Cell* 26:4376. doi: 10.1105/tpc.114.132092

Unterholzner, S. J., Rozhon, W., Papacek, M., Ciomas, J., Lange, T., Kugler, K. G., et al. (2015). Brassinosteroids are master regulators of gibberellin biosynthesis in *Arabidopsis*. *Plant Cell* 27, 2261–2272. doi: 10.1105/tpc.15.00433

Wang, F., Gao, Y., Liu, Y., Zhang, X., Gu, X., and Liu, H. (2019). BES1-regulated BEE1 controls photoperiodic flowering downstream of blue light signaling pathway in *Arabidopsis*. *New Phytol.* 223, 1407–1419.

Wang, X., and Chory, J. (2006). Brassinosteroids regulate dissociation of BKI1, a negative regulator of BRI1 signaling, from the plasma membrane. *Science* 313, 1118–1122. doi: 10.1126/science.1127593

Wang, Y., Sun, S., Zhu, W., Jia, K., Yang, H., and Wang, X. (2013). Strigolactone/MAX2-induced degradation of brassinosteroid transcriptional effector BES1 regulates shoot branching. *Dev. Cell* 27, 681–688. doi: 10.1016/j.devcel.2013.11.010

Wang, Z. Y., Nakano, T., Gendron, J., He, J., Chen, M., and Chory, J. (2002). Nuclear-localized BZR1 mediates brassinosteroid-induced growth and feedback suppression of brassinosteroid biosynthesis. *Dev. Cell* 2, 505–513. doi: 10.1016/s1534-5807(02)00153-3

Wu, P., Song, X., Wang, Z., Duan, W., Hu, R., and Hou, X. (2016). Genome-wide analysis of the BES1 transcription factor family in Chinese cabbage (*Brassica rapa* ssp. pekinensis). *Plant Growth Regul.* 80, 291–301. doi: 10.1007/s10725-016-0166-y

Yan, H., Zhao, Y., Shi, H., Li, J., Wang, Y., and Tang, D. (2018). BRASSINOSTEROID-SIGNALING KINASE1 phosphorylates MAPKKK5 to regulate immunity in *Arabidopsis*. *Plant Physiol.* 176, 2991. doi: 10.1104/pp.17.01757

Ye, Q., Zhu, W., Li, L., Zhang, S., Yin, Y., Ma, H., et al. (2010). Brassinosteroids control male fertility by regulating the expression of key genes involved in *Arabidopsis* anther and pollen development. *Proc. Natl. Acad. Sci. U.S.A.* 107, 6100–6105. doi: 10.1073/pnas.0912333107

Yin, Y., Vafeados, D., Tao, Y., Yoshida, S., Asami, T., and Chory, J. (2005). A new class of transcription factors mediates brassinosteroid-regulated gene expression in *Arabidopsis*. *Cell* 120, 249–259. doi: 10.1016/j.cell.2004.11.044

Youn, J. H., Kim, M. K., Kim, E. J., Son, S.-H., Lee, J. E., Jang, M. S., et al. (2016). ARF7 increases the endogenous contents of castasterone through suppression of BAS1 expression in *Arabidopsis* thaliana. *Phytochemistry* 122, 34–44. doi: 10.1016/j.phytochem.2015.11.006

Yu, H., Feng, W., Sun, F., Zhang, Y., Qu, J., Liu, B., et al. (2018). Cloning and characterization of BES1/BZR1 transcription factor genes in maize. *Plant Growth Regul.* 86, 235–249. doi: 10.1007/s10725-018-0424-2

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

Copyright © 2020 Ma, Ji, Liang, Li, Tian and Gao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 08 October 2020
doi: 10.3389/fgene.2020.590947

Genome-Wide Identification, Expression Profiling, and Evolution of Phosphate Transporter Gene Family in Green Algae

Long Wang[1,2†], Liang Xiao[1†], Haiyan Yang[1], Guanglei Chen[1], Houqing Zeng[3], Hongyu Zhao[2] and Yiyong Zhu[1*]*

[1] Agricultural Resource and Environment Experiment Teaching Center, College of Resource and Environment Science, Nanjing Agricultural University, Nanjing, China, [2] Key Laboratory of Plant Nutrition and Fertilizers, Ministry of Agriculture and Rural Affairs, Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences, Beijing, China, [3] College of Life and Environmental Sciences, Hangzhou Normal University, Hangzhou, China

OPEN ACCESS

Edited by:
Hongjian Wan,
ZheJiang Academy of Agricultural Sciences, China

Reviewed by:
Jian Yang,
Sichuan University, China
Zhong-Chang Wu,
Zhejiang University, China
Fuquan Liu,
Queen's University Belfast, United Kingdom

***Correspondence:**
Yiyong Zhu
Yiyong1973@njau.edu.cn
Hongyu Zhao
zhy200869@163.com

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: 03 August 2020
Accepted: 07 September 2020
Published: 08 October 2020

Citation:
Wang L, Xiao L, Yang H, Chen G, Zeng H, Zhao H and Zhu Y (2020) Genome-Wide Identification, Expression Profiling, and Evolution of Phosphate Transporter Gene Family in Green Algae. Front. Genet. 11:590947. doi: 10.3389/fgene.2020.590947

Phosphorus (P) is an essential nutrient for plant growth and development. Phosphate transporters (PHTs) are *trans*-membrane proteins that mediate the uptake and translocation of phosphate (Pi) in green plants. The PHT family including PHT1, PHT2, PHT3 and PHT4 subfamilies are well-studied in land plants; however, PHT genes in green algae are poorly documented and not comprehensively identified. Here, we analyzed the PHTs in a model green alga *Chlamydomonas reinhardtii* and found 25 putative PHT genes, which can be divided into four subfamilies. The subfamilies of CrPTA, CrPTB, CrPHT3, and CrPHT4 contain four, eleven, one, and nine genes, respectively. The structure, chromosomal distribution, subcellular localization, duplication, phylogenies, and motifs of these genes were systematically analyzed *in silico*. Expression profile analysis showed that CrPHT genes displayed differential expression patterns under P starvation condition. The expression levels of *CrPTA1* and *CrPTA3* were down-regulated, while the expression of most *CrPTB* genes was up-regulated under P starvation, which may be controlled by CrPSR1. The transcript abundance of most *CrPHT3* and *CrPHT4* genes was not significantly affected by P starvation except *CrPHT4-3*, *CrPHT4-4*, and *CrPHT4-6*. Our results provided basic information for understanding the evolution and features of the PHT family in green algae.

Keywords: phosphate transporters, phylogenetic analysis, expression pattern, Chlamydomonas, green algae

INTRODUCTION

Phosphorus (P) is one of the most important macronutrients for algae and land plants. It is a critical component of nucleic acids, phospholipids, and ATP and participates in numerous biochemical pathways, including gene expression, and signal transduction (Lorenzo-Orts et al., 2020). P acquisition is activated by transcriptional induction of genes encoding high- and low-affinity phosphate transporters (PHTs), phosphodiesterases, and phosphatases (Wykoff et al., 1999). In land plants, PHT1 (phosphate transporter 1), PHT2, PHT3, and PHT4 gene families are revealed

to encode proteins for uptake and transport of inorganic phosphate (Pi) between the cytoplasm and different subcellular compartments (Wang et al., 2017).

Phosphorus is mainly absorbed as Pi via the proton-coupled $H_2PO_4^-$ symporters PHT1 gene family, which are located in the plasma membrane of cells at the interface with the external environment in land plants (Nussaume et al., 2011; Wang et al., 2017). The PHT2, PHT3, and PHT4 gene families encode proteins that transport Pi between the cytoplasm and different subcellular compartments in land plants (Guo et al., 2008; Zhu et al., 2012; Liu et al., 2019). PHT2 proteins and some PHT4 members, which locate on chloroplast thylakoid membrane or inner envelope, are responsible for Pi homeostasis in chloroplasts (Guo et al., 2008; Liu et al., 2019). The mitochondrial phosphate transporters (PHT3) play crucial roles in respiration by absorbing Pi into the mitochondrial matrix, where Pi is utilized for the conversion of ADP to ATP (Zhu et al., 2012). A subgroup of conserved MYB transcription factors designated as phosphate starvation response (PHR) and PHR1-like (PHL) have been defined as central regulators of Pi starvation signaling in diverse plants (Sun et al., 2016). Through the physical interactions with a *cis* element, namely, PHR1 binding site (P1BS; GNATATNC), PHR transcription factors are responsible for the transcriptional activation of a considerable proportion of Pi starvation-induced genes, including PHT1 members (Zhou et al., 2008; Bustos et al., 2010; Sun et al., 2016). It has been convinced in paddy rice that OsPHR2 activates *OsPHT1;2* expressions by binding to the P1BS motif present in its promoter region (Wang et al., 2014).

In contrast, in diatoms and green algae, P starvation responses are regulated by an MYB family transcription factor phosphorus starvation response 1 (PSR1), which regulates Pi acquisition through the up-regulation of phosphatases and Pi transporters (Wykoff et al., 1999; Kumar Sharma et al., 2020). Green algae resemble the yeast in containing genes encoding both H^+/Pi (plant-like, PTA) and the Na^+/Pi (animal-like, PTB) transporters. Physiological evidence suggests that Na^+ uptake is necessary for Pi influx in *Chara corallina* under P-deficient conditions (Reid et al., 2000; Mimura et al., 2002). This is related to the sequence conservation and structural similarity of PTB proteins to the metazoan PHT and fungal PHT, which are known to be functional plasma membrane Na^+/Pi symporters (Kobayashi et al., 2003; Moseley et al., 2006). PHOSPHATE TRANSPORTER A (PTA) genes were homologous to PHO84 in yeast and encode the high-affinity H^+/Pi cotransporter (Moseley et al., 2009). The function of the PHT gene family has been extensively studied in land plants; however, the function of the PHT gene family in green algae is still largely unknown.

In this study, we identified 25 CrPHT genes in a model green alga, *Chlamydomonas reinhardtii*. We systematically analyzed the gene structures, chromosome localizations, phylogeny, and expression patterns in response to P starvation in both wild-type and *psr1* mutant backgrounds. Our work will provide a foundation for the further analysis of PHT genes function in green algae and the different mechanisms of Pi transport between green algae and land plants.

RESULTS

Phylogenetic Relationship of PHTs in Plants

To investigate the phylogenetic relationships among PHTs in green plants, a phylogenetic tree was reconstructed by maximum likelihood method using the PHT protein sequences from green plants, bacteria, and Cyanobacteria (**Figure 1**). It showed the PHTs in plants form six monophyletic groups. Cluster I (PHT2 subfamily) is sister to the cyanobacteria and bacterial PHT proteins and constitutes a sister group to the Cluster II (PTB subfamily). Cluster III contained the grouping of the PHT4 family. Cluster IV contained the grouping of the PHT3 subfamily. The PHT1 subfamily only includes PHT genes in land plants and constitutes a sister group to the PTA subfamily. PHTs in land plants consisted of PHT1, PHT2, PHT3, and PHT4 subfamilies, while PHTs in algae contained PTA, PTB, PHT3, and PHT4 subfamilies. PHT1 proteins are the H^+/Pi symporters that are responsible for Pi uptake in land plants, whereas PTB proteins are predicted to be the Na^+/Pi symporters catalyzing Pi uptake in algae. Endosymbiotic theory posits that chloroplasts and mitochondria arose from bacteria. Both organelles have retained their PHTs in membranes. The PHT2, PHT3, and PHT4 gene families encode proteins that transport Pi between the cytoplasm and different subcellular compartments (chloroplasts and mitochondria) in land plants, while how Pi is transported in subcellular compartments in algae was still unclear. To further understand the PHT in algae, we chose a model alga Chlamydomonas to analyze the PHT gene family.

Identification of *PHT* Gene Family in *Chlamydomonas reinhardtii*

To investigate the PHT genes in Chlamydomonas, we employed PHT protein sequences of Arabidopsis and rice in a global search against Chlamydomonas genomic databases. In all, 25 putative PHT genes were found in Chlamydomonas (**Table 1**). Details on gene name, locus name, open reading frame (ORF) length, protein length, molecular weight, isoelectric point (p*I*), predicted transmembrane domain (TM) number, predicted long hydrophilic loop (loop after), and P1BS element are listed in **Table 1**. The PTA subfamily contains 4 members, PTB contains 11 members, PHT3 contains 1 member, and PHT4 contains 9 members. We conducted blast analysis of the genes against the Pfam database and found that all proteins contain the MFS domain. Values for p*I* ranged from 6.21 to 9.63. The protein lengths of 25 putative PHTs range from 351 to 1666 aa, and their molecular weights are 37.44 to 171.66 kDa; most of them are around 60 kDa. PHT proteins contain 5 to 15 TM, most of them range from 10 to 12 TM, and most of them contain a large hydrophilic loop between the seventh and eighth TMDs in the PTB family except CrPTBx. The P1BS *cis* element sequence (GNATATNC) was determined previously as the binding motif for PHR1, and this element is present in the promoter regions of P starvation-induced genes in various higher plants (Zhou et al., 2008; Wang et al., 2014). To identify putative PSR1-binding sites, we searched the GNATATNC sequence in genome sequences of

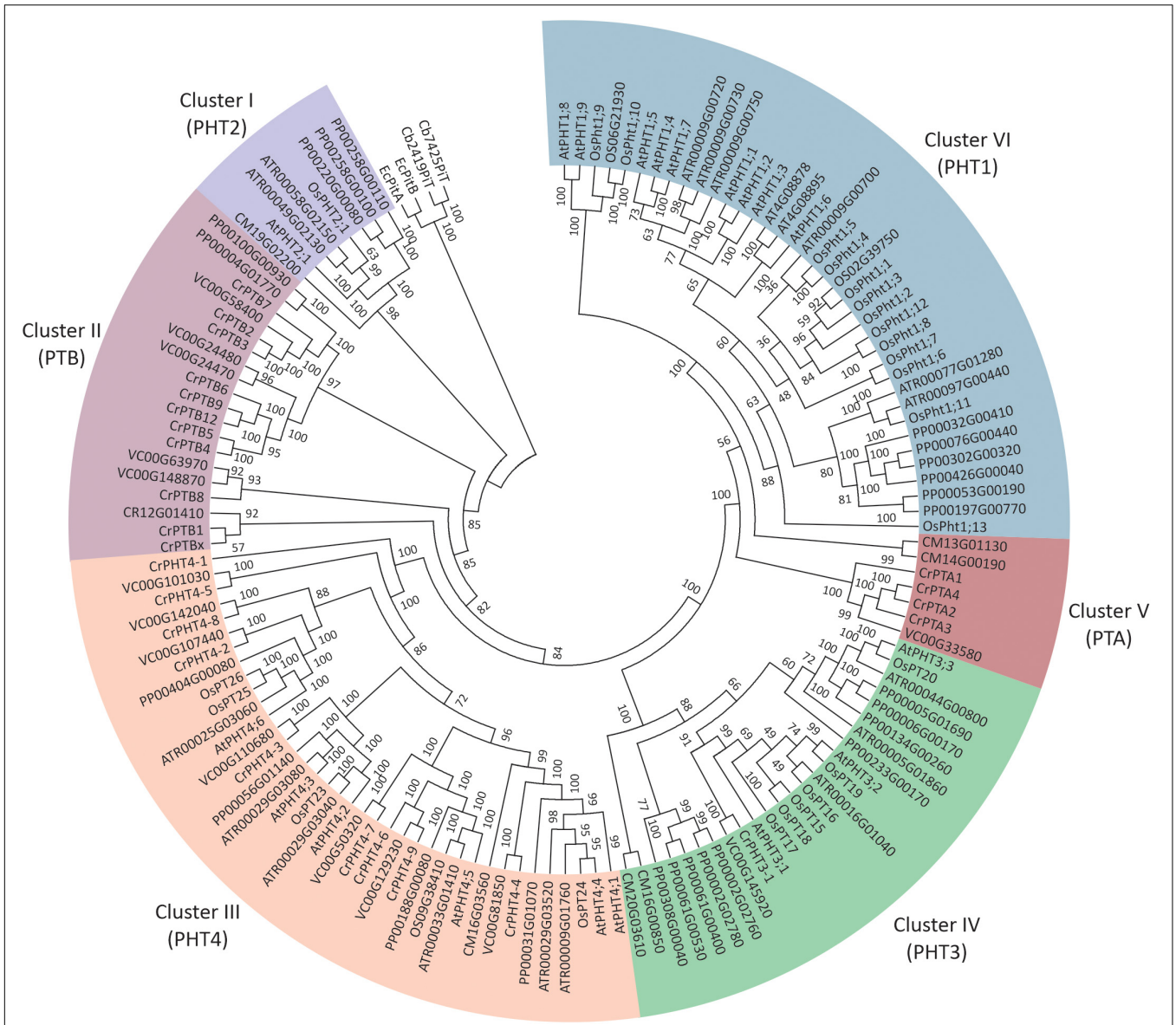


FIGURE 1 | Phylogenetic analyses of phosphate transporter genes in *Arabidopsis thaliana* (At), *Oryza sativa* (Os), *Amborella trichopoda* (Atr), and *Volvox carteri* (Vc). *Physcomitrella patens* (Pp) and *Chlamydomonas reinhardtii* (Cr). Four PHT genes from *Escherichia coli* and Cyanobacteria were selected as outgroup. Multi-sequence alignments were performed by Clustal Omega, and a phylogenetic tree was reconstructed by IQTREE using the maximum likelihood method. WAG + F + R6 was the best-fitting evolutionary model selected by ModelFinder under Bayesian Information Criterion.

all CrPHT genes (**Table 1**). P1BS elements were identified in most of the CrPHT genes. P1BS elements were present in the promoter regions of CrPTA1 and CrPTA3. All PTB subfamilies except CrPTB2, CrPTB9, and CrPTB12 have P1BS, while there is no P1BS in the PHT3 family. In the PTB4 family, only CrPHT4-2 and CrPHT4-4 have P1BS elements.

Prediction of Subcellular Targeting of PHT Proteins in Chlamydomonas

An N-terminal signal peptide targets the protein to the secretory pathway; proteins are inserted into the endoplasmic reticulum (ER), and either remain in the ER or move to the plasma membrane, vacuole, or endomembrane compartments (Chrispeels and Raikhel, 1992). Using four online tools (details in section "Materials and Methods"), which could predict subcellular localization by detecting subcellular targeting motifs, we discovered the signal peptide at the N-terminus of all PHT proteins (**Table 2**). Among them, the entire PTA family was predicted to locate in the plasma membrane with a very high reliability class. PTB proteins were predicted to target to the secretory pathway. However, based on Bonnot's report (Bonnot et al., 2017), PTB proteins are plasma membrane proteins, which suggests that they finally move to the plasma

TABLE 1 | Key features of all 25 PHTs in *Chlamydomonas reinhardtii*.

Locus	Gene name	Gene length (bp)	ORF length (bp)	Protein length (aa)	Mol wt (k Da)	pI	TM number	Hydrophilic loop after	P1BS locations on PHT gene
Cre02.g075050	PTA1	5518	1683	560	62.57	8.72	11	–	55, 1760, 1771, 2488
Cre16.g686800	PTA2	5067	1719	572	63.63	9.2	11	–	–
Cre16.g686750	PTA3	5450	1722	573	63.39	8.91	11	–	329
Cre16.g686850	PTA4	6985	1671	556	61.93	8.97	11	–	–
Cre12.g491600	PTB1	11369	5001	1666	171.66	6.77	15	7	6097, 10563
Cre07.g325741	PTB2	5012	1959	652	70.11	8.51	12	7	–
Cre07.g325740	PTB3	5542	1959	652	70.19	8.65	12	7	190, 2344
Cre02.g144750	PTB4	4957	1929	642	69.13	8.22	11	7	1026, 2506
Cre02.g144700	PTB5	4578	1878	625	67.05	9.18	11	7	2169
Cre16.g655200	PTB6	6690	1839	612	65.12	8.71	11	7	287, 3502
Cre12.g489400	PTB7	9082	2508	835	85.43	8.02	13	7	512, 7432, 7445, 8258
Cre16.g676757	PTB8	9198	4020	1339	139.34	6.21	12	7	2439, 8766
Cre02.g144600	PTB9	4542	1887	628	67.04	8.63	11	7	–
Cre02.g144650	PTB12	6218	1875	624	66.76	8.63	11	7	–
Cre12.g493404	PTBx	14786	3879	1292	131.42	6.6	8	3	5218
Cre03.g172300	PHT3-1	2782	1056	351	37.44	9.17	5	–	–
Cre01.g023000	PHT4-1	6591	1644	547	54.17	9.63	12	–	–
Cre03.g187050	PHT4-3	4012	1857	618	61.83	9.83	13	–	–
Cre08.g379550	PHT4-4	5842	1725	574	60.59	9.27	11	–	25, 4928
Cre09.g396950	PHT4-9	6056	2193	730	76.08	9.52	12	–	–
Cre09.g399141	PHT4-2	6414	1305	434	46.64	9.75	11	–	1353
Cre12.g515750	PHT4-5	4476	1500	499	52.81	9.13	10	–	–
Cre16.g648300	PHT4-6	6153	1926	641	65.25	7.59	11	–	–
Cre16.g663600	PHT4-7	5604	1413	470	49.53	8.66	11	–	–
Cre16.g675300	PHT4-8	6254	1740	579	60.17	9.06	11	–	–

The total number of predicted transmembrane domains (TM number) and the behind long hydrophilic loops (loop after) using a complete protein sequence. P1BS number in the genome regions of PT genes was identified by the NewPLACE database (http://sogo.dna.affrc.go.jp/cgi-bin).

membrane through the secretory pathway. PHT3 was predicted to target to the mitochondrion, with a very low reliability class. Most of the PHT4 family proteins were localized in chloroplast with a reliability class between 4 and 5. CrPHT4-2 and CrPHT4-7 were predicted to target to the secretory pathway. The CrPHT4-5 has no signal peptide and did not display predicted localization.

Conserved Motif Analysis of Chlamydomonas PHT Proteins

To analyze the conserved motifs and reveal the distinction between green algae and other ancient species, conserved motifs of PHT proteins from Chlamydomonas, bacteria, fungus, and metazoa were identified by MEME (Bailey and Elkan, 1994). The sequences of 4 Chlamydomonas PTA proteins; 11 PTB proteins; 8 PHT4 proteins; 5 PHT3 proteins; fungal PHT proteins ScPho89 and NcPho4; metazoan PHT proteins DmPit, HsPit1, and HsPit2; bacterial PHT proteins TePit, EcPitA, and EcPitB (from *Escherichia coli*); and 2 cyanobacteria proteins were selected for further analysis (**Table 3**). We identified a total of 35 conserved motifs among the selected PHT proteins and found that the conserved motifs could distinguish the Chlamydomonas PHT proteins from other members of the PHT proteins. Five motifs (motifs 1–5) were present in bacterial PHT, metazoan PHT, fungal PHT, and PTB family proteins. They contain the PHT signature GANDVAN domains (motif 2). Four motifs (motif 6–9) were common to the metazoan, fungal PHT, and PTB proteins. Twelve motifs (motifs 1–12) were present in PTB proteins. Eight motifs (motifs 16–23) were specific to the PTA proteins, and five motifs (motifs 24–28) were specific to the PHT3 proteins. Motifs 2, 4, and 29–35 were present in all PHT4 proteins. All motifs identified in the bacterial, metazoan, and fungal PHT proteins were present in the PTB proteins (**Table 3**). Meanwhile, motifs specific to the PTA and PHT3 family were found. Together, the conserved and specific motifs found in the different subfamilies of CrPHT proteins suggest their conservation in evolution and their divergence of functions.

Chromosomal Localization and Gene Duplication of *CrPHT* Genes

The chromosomal distributions of *CrPHT* genes were determined. As shown in **Figure 2**,25 PHT genes are distributed on 8 Chlamydomonas chromosomes. Out of the 25 predicted

TABLE 2 | Prediction of subcellular targeting of PHT proteins in *Chlamydomonas reinhardtii*.

Locus	Name	PM	cTP	mTP	SP	Other	loc	RC
Cre02.g075050	PTA1	0.63	0.08	0.01	0.22	0.06	P	1
Cre16.g686800	PTA2	0.43	0.2	0.03	0.27	0.07	P	1
Cre16.g686750	PTA3	0.48	0.09	0.02	0.35	0.07	P	1
Cre16.g686850	PTA4	0.55	0.06	0.01	0.31	0.07	P	1
Cre12.g491600	PTB1	–	0.008	0.019	0.953	0.06	S	1
Cre07.g325741	PTB2	–	0.004	0.008	0.913	0.395	S	3
Cre07.g325740	PTB3	–	0.004	0.007	0.894	0.475	S	3
Cre02.g144750	PTB4	–	0.002	0.012	0.979	0.309	S	3
Cre02.g144700	PTB5	–	0.002	0.013	0.982	0.28	S	2
Cre16.g655200	PTB6	–	0.011	0.015	0.943	0.339	S	2
Cre12.g489400	PTB7	–	0.139	0.22	0.075	0.513	–	–
Cre16.g676757	PTB8	–	0.011	0.029	0.688	0.108	S	3
Cre02.g144600	PTB9	–	0.008	0.013	0.958	0.264	S	2
Cre02.g144650	PTB12	–	0.008	0.014	0.957	0.243	S	2
Cre12.g493404	PTBx	–	0.003	0.043	0.914	0.076	S	1
Cre03.g172300	PHT3-1	0.12	0.23	0.54	0.02	0.08	M	4
Cre01.g023000	PHT4-1	0.04	0.88	0.05	0.02	0.01	C	4
Cre03.g187050	PHT4-3	0.11	0.76	0.07	0.05	0.01	C	4
Cre08.g379550	PHT4-4	0.06	0.84	0.08	0.01	0.01	C	5
Cre09.g396950	PHT4-9	0.04	0.9	0.05	0.01	0	C	4
Cre09.g399141	PHT4-2	0.13	0.01	0.01	0.85	0.01	S	5
Cre12.g515750	PHT4-5	0.17	0.35	0.1	0.34	0.03	–	–
Cre16.g648300	PHT4-6	0.05	0.89	0.03	0.02	0.01	C	5
Cre16.g663600	PHT4-7	0.03	0	0	0.97	0	S	1
Cre16.g675300	PHT4-8	0.11	0.69	0.11	0.07	0.02	C	4

PM, plasma membrane; cTP, chloroplast; mTP, mitochondrion; and SP, secretory pathway. The prediction was given with a reliability class (RC) indicated from 1 to 5. 1: 1.000 > diff > 0.800; 2: 0.800 > diff > 0.600; 3: 0.600 > diff > 0.400; 4: 0.400 > diff > 0.200; and 5: 0.200 > diff > 0.000. "–" means no data.

PHT genes, 18 are distributed across the 3 Chlamydomonas chromosomes (chromosome 2, 12 and 16) (**Figure 2**). Eight *PHT* genes scattered in clusters in chromosome 16. Chromosomes 1 and 8 each contain one gene. Chromosome 3, 7, and 9 each contain two genes. We further performed chromosome mapping to determine the gene duplication of *PHT* genes on the Chlamydomonas chromosomes. As shown in **Supplementary Table 1**, gene duplication of *CrPHT* genes contained four types, including tandem duplication (TD), proximal duplication (PD), transposed duplication (TRD), and dispersed duplication (DSD). Almost all the *PHT* genes were identified as duplicated genes (**Supplementary Table 1**). As shown in **Figure 2**, three groups of *PHT* genes can be identified as TD genes including group I (CrPTA2, CrPTA3, and CrPTA4) located on chromosome 16, group II (CrPTB4, CrPTB5, CrPTB9, and CrPTB12) located on chromosome 2, and group III (CrPTB2 and CrPTB3) located on chromosome 7, implying that the high density of PHT genes on this chromosome was mainly due to the TD events.

Expression Profiles of *CrPHT* Genes During Pi Starvation

To uncover the expression profiles of *CrPHT* genes under deprivation starvation and whether they are controlled by CrPSR1, we reanalyzed the RNA-seq data from the database[1] (Bajhaiya et al., 2016). Phylogenetic trees constructed by the NJ method showed that these CrPHT proteins could be divided into four clades (**Figure 3**). Transcription levels of *CrPHT* genes varied under Pi starvation (**Figure 3** and **Supplementary Table 2**). *CrPTA1* and *CrPTA3* transcripts declined following exposure of nutrient-replete cells to P starvation, and this decline is less significant in the *psr1* mutant. This characteristic makes CrPTA1 and CrPTA3 candidates for the low-affinity Pi uptake system that operates in cells that are P-replete. The abundance of CrPTA2 and CrPTA4 transcripts is not significantly affected by P starvation. Upon Pi starvation, expressions of *CrPTB2*, *CrPTB3*, *CrPTB4*, *CrPTB5*, *CrPTB7*, *CrPTB8*, and *CrPTB12* were induced, and this induction does not occur in the *psr1* mutant, indicating that these genes were regulated by PSR1. The abundance of *CrPTB1*, *CrPTB9*, *CrPTB12*, and *CrPTBx* transcripts were not significantly affected by P starvation. The expression levels of genes in the *CrPHT3* and *CrPHT4* subfamilies were not significantly affected by P starvation, except *CrPHT4-3*, *CrPHT4-4*, and *CrPHT4-6*.

Noticeably, the expression profiles of genes in each TD group (**Figure 2**) are variable. In Group I, *CrPTA2* and *CrPTA4* are poorly expressed, while *CrPTA3* is highly expressed in

[1]http://www.ebi.ac.uk/arrayexpress

TABLE 3 | Overall conserved motifs of PHT families in Chlamydomonas, bacteria, metazoans, and fungi.

Motif	AA	Sequence	Bacterial PiT	Metazoan PiT	Fungal PiT	Chlamydomonas			
						PTA	PTB	PHT3	PHT4
Motif1	16	GLPvStTHcivGavvG	Y	Y	Y		Y		
Motif2	10	GANDVANafG	Y	Y	Y		Y		Y
Motif3	10	TPsRGFcaEL	Y	Y	Y		Y		
Motif4	6	sWfysp	Y	Y	Y		Y		Y
Motif5	16	TLrQAvljAavfEFsG	Y	Y	Y		Y		
Motif6	21	HaaAEVFdPetEyvykYLQVF		Y	Y		Y		
Motif7	16	PiWvLayGGaGlvyGL		Y	Y		Y		
Motif8	29	LRRebsttlaiwvlPllvlTvfLNvFFV		Y	Y		Y		
Motif9	10	VtrTiaggla		Y	Y		Y		
Motif10	13	KkAAtrGlnvDiH					Y		
Motif11	10	AWVaAciaaG					Y		
Motif12	13	aPSjtmeKdiakY					Y		
Motif13	9	GsGAVVWnD					Y		
Motif14	10	NtKSiGYvdP					Y		
Motif15	12	ADaEanaGKpKE					Y		
Motif16	18	DKPWMGRMRMQvMGFaWM				Y			
Motif17	26	WRKFGLLMyYYWHRNFGTAMSWFVWD				Y			
Motif18	22	DVTGLDLREGDKRWLAiLDGHH				Y			
Motif19	24	FQFLYYFSSFWGQFGPNATTWLLP				Y			
Motif20	23	PGLGMFCEAYFVFAVGNLSAIWK				Y			
Motif21	17	FGQLFLGFFADRiGRKW				Y			
Motif22	18	KRRGETVVLVFSMQGWGN				Y			
Motif23	30	GNKLFQGTFiKiiNPsASLIQVLEWTLLNS				Y			
Motif24	19	VKvnvQtnPgfykgjsdGf						Y	
Motif25	25	CKFGFYEyFKKtYsDmAGpEyakKY						Y	
Motif26	30	WGRQIPYTMMKFasFEttVEmiYKyavPkP						Y	
Motif27	30	SFagGYiAGVFCAIVSHPADNLVSFLNnaK						Y	
Motif28	28	MIGTLTGaQWGiYDAFKVfvGLPTTGGV						Y	
Motif29	22	DjapryAgallGltNTaGalaG							Y
Motif30	21	CNmDrvnmSVAijPMaaeFGW							Y
Motif31	10	GLVqSAFywG							Y
Motif32	14	WgyytlLsWLPTYF							Y
Motif33	12	AGvLLGjsNTAG							Y
Motif34	11	RvlvGlGzGVA							Y
Motif35	7	tVRKilQ							Y

For each motif, the amino acid length (AA), the sequence, and the presence/absence (Y/–) in proteins of the PHT families (bacterial PiT, metazoan PiT, fungal PiT, PTA, PTB, PHT3, and PHT4) are given. Uppercase and lowercase indicate residues of the motifs conserved in >50% or <50%, respectively, of the concerned proteins.

nutrient-replete cells and its expression responded to phosphate starvation. This indicates that *CrPTA2* and *CrPTA4* are likely degenerated or redundant in function after duplication, compared to *CrPTA3*. In another duplicated group (Group III), the expression level of *CrPTB2* is much higher than *CrPTB3* under all the conditions, suggesting that the regulation of *CrPTB2* and *CrPTB3* are different. In Group III, *CrPTB4*, *CrPTB5*, and *CrPTB12* were sharply up-regulated under phosphate starvation, while CrPTB9 is barely expressed under phosphate starvation in both WT and *psr1* mutant. Different from other genes in Group III, *CrPTB4* still showed some expression in the *psr1* mutant. Together, functions of duplicated genes in each group are divergent and the gene duplication and subsequent divergence of *CrPHT* genes make up the complex network for P transport to adapt to a variable Pi environment.

qRT-PCR Analyses of *CrPHT* Genes Under P Stress Conditions

To explore the expression profiles of CrPHT genes under deprivation starvation, qRT-PCR analysis was performed (**Figure 4**). We firstly investigated the growth of WT and *psr1* mutant under P-replete or depleted conditions. In P-replete conditions, the growth of the *psr1* mutant is similar to that of WT. However, the *psr1* mutant was unable to acclimate properly to P starvation and exhibited much less growth than wild-type strains on the TA medium (Shimogawara et al., 1999). The *psr1*

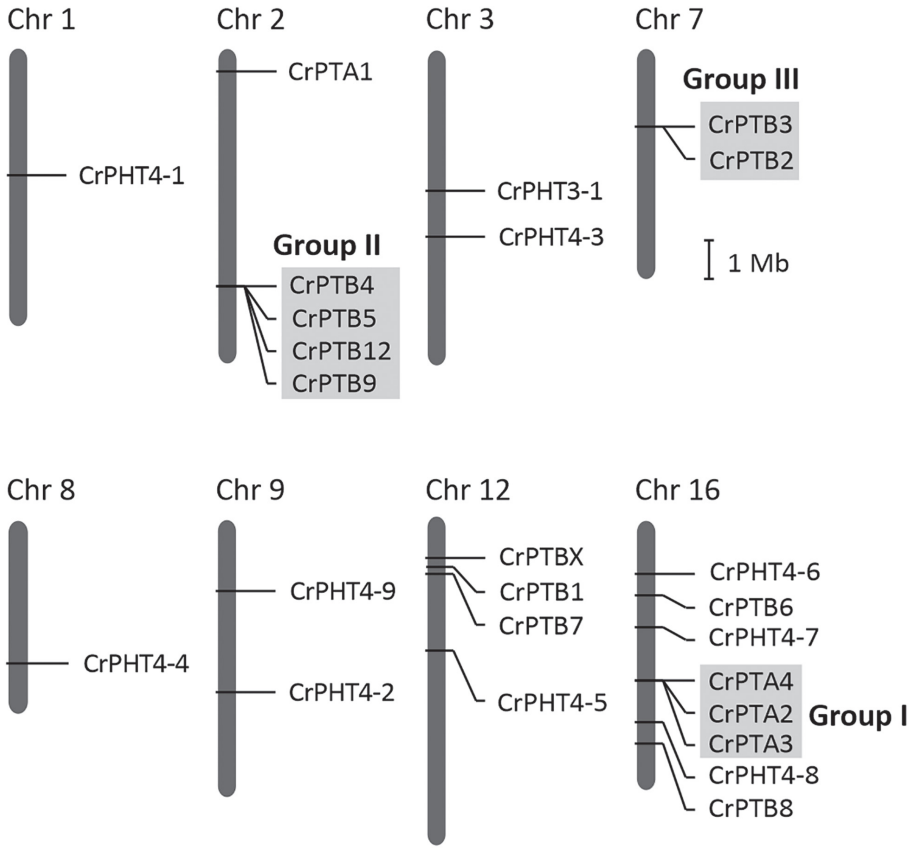


FIGURE 2 | Chromosomal location and gene duplication of PHT proteins in Chlamydomonas. Chromosome numbers are shown at the top of each bar. Recent tandem duplicated genes were marked with a gray background. The scale presented on the left indicates chromosome sizes in megabases (Mb).

mutant showed growth defect compared to the WT in P-depleted conditions (**Figure 4A**). We further analyzed the steady-state transcript levels of genes encoding the phosphatase CrPHOX; the PHTs CrPTB2, CrPTB4, CrPTB5 CrPTB12, and CrPTA1; and the transcription factor CrPSR1 in the WT and *psr1* mutant (**Figure 4**). Consistent with previous report (Shimogawara et al., 1999), the expression of *CrPSR1* was significantly induced by P starvation (**Figure 4B**). *CrPHOX* was induced by P starvation in WT and not in the *psr1* mutant (**Figure 4C**). Under Pi starvation conditions, the expressions of *CrPTB2*, *CrPTB4*, *CrPTB5*, and *CrPTB12* were induced in the WT while they showed no effect in the *psr1* mutant. The *CrPTA1* transcript reduced following exposure of nutrient-replete cells to P starvation, and this decline does not occur in the *psr1* mutant. These results suggest that different *CrPHT* genes have different expression patterns and CrPSR1 displays an important role in regulating Pi starvation-dependent signaling in Chlamydomonas.

DISCUSSION

Phosphate is one of the most important nutrients that significantly affect plant growth and metabolism (Raghothama, 1999). Although the evolution and phylogenetic relationship of PHTs in different plant species have been reported, none of these studies addressed the possible origin and early evolution of plant PHTs. In this study, we identified members of the PHT family in Chlamydomonas and characterized their evolutionary and expression profiles under P starvation conditions. Interestingly, based on phylogenetic analyses of PHT family members among different plant species, we found that two Pi uptake systems exist in chlorophytes, including H^+/Pi symport and Na^+/Pi symport (**Figure 1**), while higher plants only contain a H^+/Pi symport (Nussaume et al., 2011). PTB proteins are hypothesized to be Na^+/Pi symporters, which facilitate Pi uptake in chlorophytes, whereas PHT1 proteins are H^+/Pi symporters that facilitate Pi uptake in angiosperms and all streptophyte lineages (Ayadi et al., 2015; Bonnot et al., 2017). We suggest that the PTA subfamily in algae may also facilitate Pi uptake through H^+/Pi symport.

In this study, 25 *CrPHT* genes were identified in the Chlamydomonas genome, comprising all four major types of plant PHT genes. Chromosome location analysis revealed that almost all the PHT genes were derived from multiple gene duplication events in Chlamydomonas. We found 4 *CrPTA* genes in Chlamydomonas, in which CrPTA1 and CrPTA3 are through gene duplication. All PTA proteins were predicted to be located in the plasma membrane, similar to the PHT1 in land plants. PHT1 proteins are the well-documented plant Pi transporters.

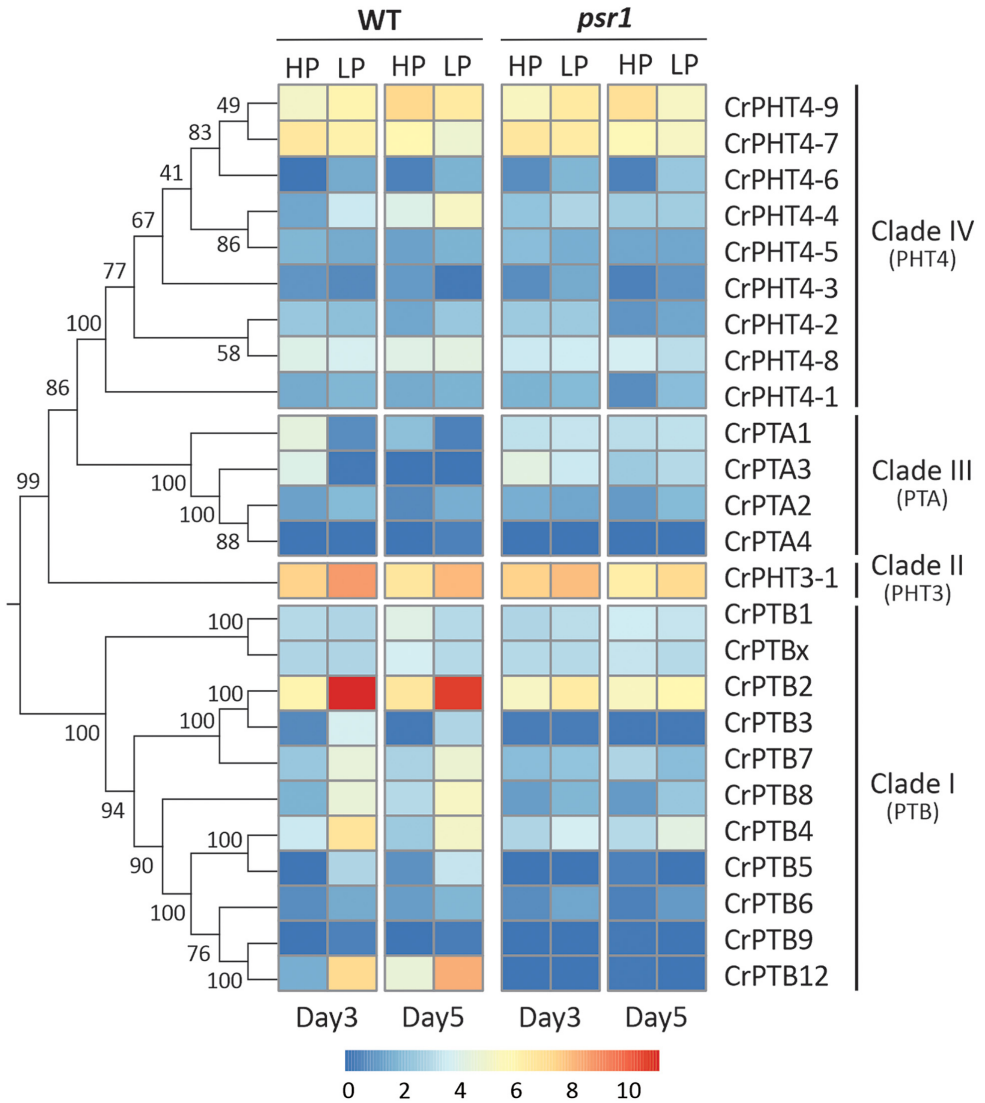


FIGURE 3 | Expression profiles of Chlamydomonas PHT genes under phosphate-replete and starvation conditions. The phylogenetic tree of PHT proteins, showing in the left panel, was constructed by the neighbor-joining method with 1000 bootstrap replications using MEGA 7.0 software. Bootstrap values were shown on the branches. Heat map showing the expression of phosphate transporter genes under high-P (HP) and low-P (LP) conditions in day 3 and day 5 which display change in expression relative to WT HP conditions or *psr1* HP. Each value corresponds to log2 of normalized counts. Day 3 and day 5 mean equivalence to the onset of P starvation and 48 h after the onset of P starvation, respectively.

These proteins have a conserved structure containing 12 transmembrane (TM) domains with a large hydrophilic loop between TM6 and TM7 and both hydrophilic N and C termini located in the cytoplasm (Raghothama, 1999; Nussaume et al., 2011). All plant PHT1 proteins have the conserved PHT1 signature GGDYPLSATIxSE and in fungi it is GGDYPLSxxIxSE (Karandashov and Bucher, 2005). In angiosperms, proteins in the PHT1 family catalyze the uptake of Pi from the environment through H^+/Pi symport (Ayadi et al., 2015). We propose that the origin of PHT1 genes in land plants is worth more studies, to uncover the mechanism of PHT type bias. All PTB genes were located in chromosomes 2, 7, 12, and 16 and formed 4 clusters. We found a signal peptide in PTB proteins, and all PTB were predicted to the secretory pathway. For example, MpPTB proteins traffic to the plasma membrane through the secretory pathway and function as the plasma membrane PHT for Pi uptake (Bonnot et al., 2017). It indicated that CrPTB may also locate on the plasma membrane and control the Pi uptake in Chlamydomonas.

The topology of the phylogenetic tree showed that the PHT2 subfamily is sister to the bacterial PHT proteins, together with a sister group to the PTB subfamily. We inferred that the PTB protein may move and locate to the chloroplast during evolution because we did not find PHT2 in chlorophyte algae, while in land plants, such as rice, PHT2;1 localized at the chloroplast envelope and functioned as a low-affinity

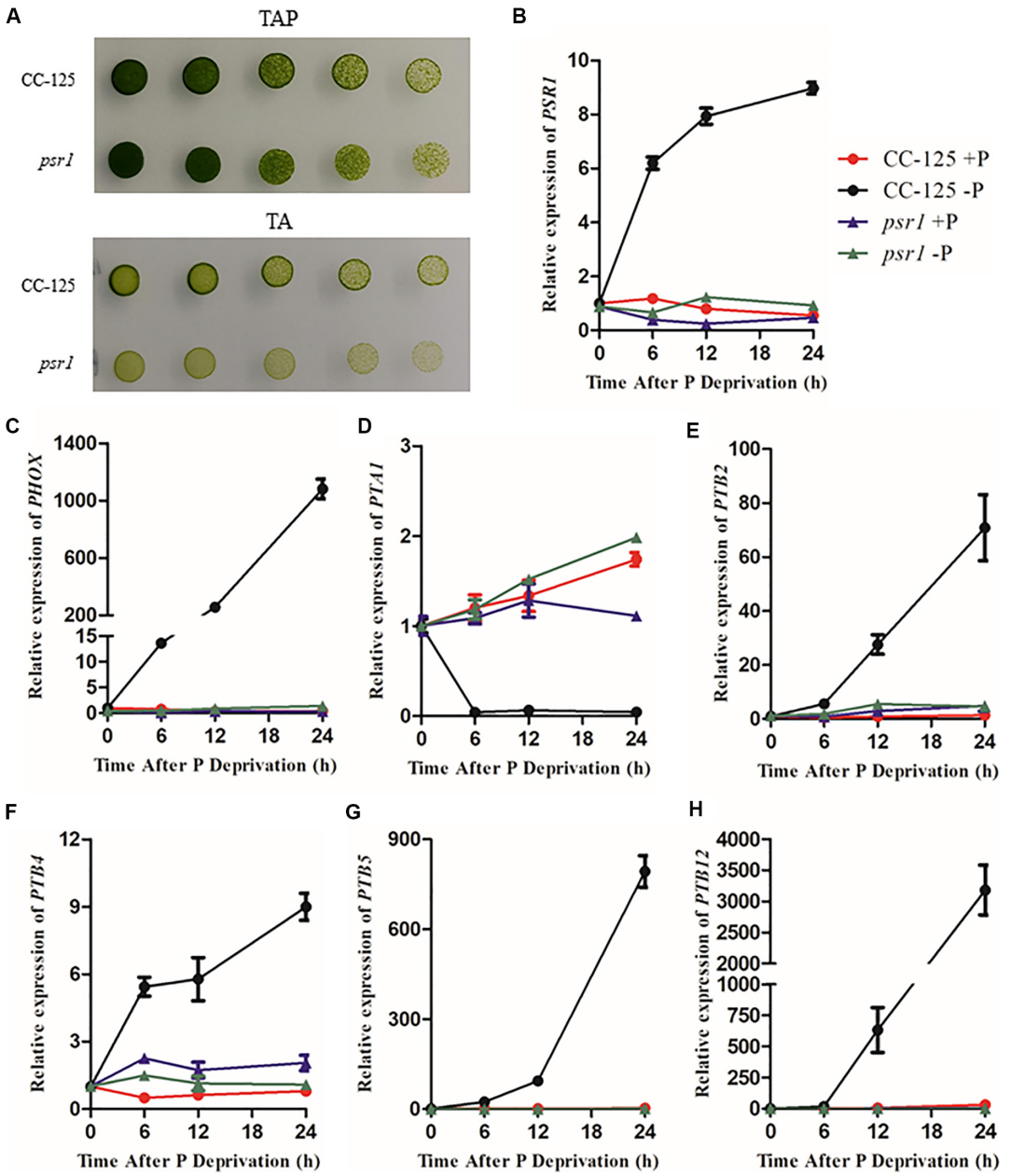


FIGURE 4 | The relative expression of Chlamydomonas PHT genes under phosphate replete and starvation conditions. **(A)** Growth of CC-125 and *psr1* in the TAP or TA plate. **(B)** to **(H)**, The *CrPSR1* **(B)**, *CrPHOX* **(C)**, *CrPTA1* **(D)**, *CrPTB2* **(E)**, *CrPTB4* **(F)**, *CrPTB5* **(G)**, and *CrPTB12* **(H)** mRNA levels were quantified by qRT-PCR over a period of 0, 6, 12, and 24 h of phosphate deprivation. *CBLP* was used as the internal control. For **B–H**, data are the means ± s.d. All experiments were repeated three times, and similar results were obtained.

Pi transporter (Liu et al., 2019). PHT3 proteins have two TM α-helices separated by a hydrophilic extramembrane loop, which are conserved in all analyzed mitochondrial transporter proteins, being essential for mitochondrial targeting (Zhu et al., 2012). There are three members in Arabidopsis, as well as in a basic angiosperm, *Amborella trichopoda*, while six transporters in rice and nine in *Physcomitrella patens* were clustered with AtPHT3 proteins in the phylogenetic tree (Wang et al., 2017). We only found one PHT3 gene in Chlamydomonas. Therefore, PHT3 family rapidly expanded in land plants, especially in mosses and monocots, through gene duplication during the evolution of plant lineages. In contrast to the rapid expansion of PHT3 in land plants, the members of the PHT4 family are ancient and relatively consistent. The PHT4 family was first characterized and designated by Guo et al. (2008) in Arabidopsis (Guo et al., 2008). There are six members in the PHT4 family in Arabidopsis and seven putative PHT4 members in rice (Wang et al., 2017). PHT4 proteins share a similarity with

SLC17/type I Pi transporters. We found nine PHT4 genes in Chlamydomonas which contained 11–13 TM and mainly located in the chloroplast.

Transcription of *CrPHT* genes varies in response to changes of phosphate levels. We found that, upon Pi starvation, expression of *CrPTA1* and *CrPTA3* is reduced in the cells. This characteristic makes CrPTA1 and CrPTA3 candidates for the low-affinity Pi uptake system that operates in cells under P-replete conditions. Most Arabidopsis PHT1 family genes, which encode H^+/Pi symporters, are up-regulated in phosphate-deprived conditions (Mudge et al., 2002). We also found that P starvation led to enhanced expression levels of *CrPTB2*, *CrPTB3*, *CrPTB4*, *CrPTB5*, *CrPTB6*, *CrPTB7*, and *CrPTB8*, indicating that they may be high-affinity PHTs operating at low Pi conditions. We also found that the reduction of *CrPTA* and induction of *CrPTB* transcription does not occur in *psr1* mutant (**Figure 4**), consistent with that the P1BS element is present in these genes (**Table 1**). Our results suggest that they may be regulated by CrPSR1.

MATERIALS AND METHODS

Sequence Collection and Phylogenetic Analyses

To illuminate the evolutionary relationship of PHT gene families in plant lineages, the representative PHT genes from *Arabidopsis thaliana* (At), *Oryza sativa* (Os), *Amborella trichopoda* (Atr), and *Volvox carteri* (Vc). *Physcomitrella patens* (Pp), *Chlamydomonas reinhardtii* (Cr), and *Cyanidioschyzon merolae* (Cm) were selected for constructing a phylogenetic tree with PHT genes from *Escherichia coli* (Ec) and Cyanobacteria as outgroup. Multi-sequence alignments were performed by Clustal Omega (Sievers et al., 2011), and a phylogenetic tree was reconstructed by IQTREE using the maximum likelihood method (Nguyen et al., 2015). WAG + F + R6 was the best-fitting evolutionary model selected by ModelFinder under Bayesian Information Criterion (Kalyaanamoorthy et al., 2017). The gene name referred to Liu's report (Liu et al., 2011).

Sequence Analysis and Structural Characterization

Information of candidate PHT genes in Chlamydomonas was obtained via searching the Phytozome database[2] (Goodstein et al., 2012), including chromosome locations, genomic sequences, coding sequences (CDS), and amino acid sequences. Conserved motifs were identified by MEME (Bailey and Elkan, 1994). Molecular weight, theoretical isoelectric point (theoretical p*I*), and instability index (II; with the value > 40 classified as unstable) of CrPHT proteins were analyzed by the ProtParam tool[3]. Hydropathy profiles were predicted using TMPRED[4]. P1BS was identified by the NewPLACE database[5] (Higo et al., 1999).

Subcellular Localization and Conserved Motif Analysis of CrPHT

Four Online tools were employed to predict subcellular localization, including Wolf TargetP[6](Emanuelsson et al., 2007), Psort[7], MultiLoc[8], and YLoc[9]. The prediction of the signal peptide and TM was performed with the SMART program[10]. The putative protein sequences were subjected to MEME program[11] to investigate conserved motifs with the following parameters: site distribution—any number of repetitions, number of motifs—35, the motif width between 5 and 30 (Bailey and Elkan, 1994).

Chromosome Localization and Gene Duplication

Chromosome mapping of the candidate CrPHT genes was viewed using the software MapDraw V2.1 (Yu et al., 2016). Gene duplication was identified based on plantDGD[12] (Qiao et al., 2019).

Expression Profile of CrPHT Genes Under P Stress Conditions

To get an insight into the expression profiles of the CrPHT gene family under P deprivation, we reanalyzed the RNA-seq sequence data in the ArrayExpress database[13] under the accession number E-MTAB-2556 (Bajhaiya et al., 2016). In detail, wild-type (CC125) and *psr1* mutant cells were cultured until day 3 and day 5 in high-P and low-P TAP media. For the low-P wild-type cells, day 3 and day 5 were equivalent to onset of P starvation and 48 h post-P starvation, respectively. Normalized log2 values of expression levels are shown in **Supplementary Table 2**.

Algae Growth Conditions, RNA Extraction, cDNA Synthesis, and Expression Analysis

The *C. reinhardtii* WT strain CC-125 and CC-4350 (*Crpsr1-1*) mutants were obtained from the Chlamydomonas Resource Center (Li et al., 2016). Cells were cultured in standard Tris–acetate-phosphate TAP medium at pH 7.0 under continuous illumination (50 mmol photons m^{-2} s^{-1}) on a rotating platform (150 rpm) at 24°C. For Pi deprivation, cells in the mid-logarithmic phase (5–8 × 10^6 cells mL^{-1}) were pelleted by

[2]https://phytozome.jgi.doe.gov/pz/portal.html
[3]http://web.expasy.org/protparam/
[4]http://www.ch.embnet.org/software/TMPRED_form.html
[5]http://sogo.dna.affrc.go.jp/cgi-bin/
[6]http://www.cbs.dtu.dk/services/TargetP/
[7]http://www.genscript.com/psort/wolf_psort.html
[8]http://abi.inf.uniuebingen.de/Services/MultiLoc2
[9]https://abi-services.informatik.uni-tuebingen.de/yloc/webloc.cgi
[10]http://smart.embl-heidelberg.de/smart/batch.pl
[11]http://meme-suite.org/tools/meme
[12]http://pdgd.njau.edu.cn:8080
[13]http://www.ebi.ac.uk/arrayexpress

centrifugation (2,000 *g*, 5 min), washed twice with TA in which 1.5 mM potassium chloride which was substituted for 1 mM potassium phosphate (Quisel et al., 1996), and then resuspended in TA medium.

Total RNA was extracted from frozen cell pellets using the RNeasy Mini Kit (Qiagen) and reverse transcribed to complementary DNA after DNase I treatment following the manufacturer's instructions (NEB). Quantitative real-time PCR was performed using a SYBR Premix kit (Roche) on a QuantStudio 6 Flex machine (Life Technologies). The *CBLP* gene was used as an internal control (Chang et al., 2005). The primer pairs used for RT-qPCR are given in the **Supplementary Table 3**.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

YZ and HYZ conceived and supervised the project. LW, LX, YZ, and HYZ designed the research. LW and LX performed the experiments. LW, HY, LX, GC, and HZ analyzed the data. LW, HZ, and HYZ wrote the manuscript. All authors discussed the results and commented on the manuscript.

FUNDING

This work was supported by the National Key R&D Program of China (2017YFD0200200/0200204/0200206) and National Natural Science Foundation (31801924).

ACKNOWLEDGMENTS

We are grateful to Dr. Lei Xu for helpful discussion.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.590947/full#supplementary-material

Supplementary Table 1 | List of gene duplication pairs identified in CrPHT genes.

Supplementary Table 2 | Expression level of CrPHT genes under phosphate starvation in the WT and *psr1* mutant.

Supplementary Table 3 | List of all primers used in this study.

REFERENCES

Ayadi, A., David, P., Arrighi, J. F., Chiarenza, S., Thibaud, M. C., Nussaume, L., et al. (2015). Reducing the genetic redundancy of *Arabidopsis* PHOSPHATE TRANSPORTER1 transporters to study phosphate uptake and signaling. *Plant Physiol.* 167, 1511–1526. doi: 10.1104/pp.114.252338

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

Bajhaiya, A. K., Dean, A. P., Zeef, L. A., Webster, R. E., and Pittman, J. K. (2016). PSR1 is a global transcriptional regulator of phosphorus deficiency responses and carbon storage metabolism in *Chlamydomonas reinhardtii*. *Plant Physiol.* 170, 1216–1234. doi: 10.1104/pp.15.01907

Bonnot, C., Proust, H., Pinson, B., Colbalchini, F. P. L., Lesly-Veillard, A., Breuninger, H., et al. (2017). Functional PTB phosphate transporters are present in streptophyte algae and early diverging land plants. *New Phytol.* 214, 1158–1171. doi: 10.1111/nph.14431

Bustos, R., Castrillo, G., Linhares, F., Puga, M. I., Rubio, V., Perez-Perez, J., et al. (2010). A central regulatory system largely controls transcriptional activation and repression responses to phosphate starvation in *Arabidopsis*. *PLoS Genet.* 6:e001102. doi: 10.1371/journal.pgen.1001102

Chang, C. W., Moseley, J. L., Wykoff, D., and Grossman, A. R. (2005). The LPB1 gene is important for acclimation of *Chlamydomonas reinhardtii* to phosphorus and sulfur deprivation. *Plant Physiol.* 138, 319–329. doi: 10.1104/pp.105.059550

Chrispeels, M. J., and Raikhel, N. V. (1992). Short peptide domains target proteins to plant vacuoles. *Cell* 68, 613–616. doi: 10.1016/0092-8674(92)90134-x

Emanuelsson, O., Brunak, S., Von Heijne, G., and Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.* 2, 953–971. doi: 10.1038/nprot.2007.131

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40, D1178–D1186.

Guo, B., Jin, Y., Wussler, C., Blancaflor, E. B., Motes, C. M., and Versaw, W. K. (2008). Functional analysis of the *Arabidopsis* PHT4 family of intracellular phosphate transporters. *New Phytol.* 177, 889–898. doi: 10.1111/j.1469-8137.2007.02331.x

Higo, K., Ugawa, Y., Iwamoto, M., and Korenaga, T. (1999). Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res.* 27, 297–300. doi: 10.1093/nar/27.1.297

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth.4285

Karandashov, V., and Bucher, M. (2005). Symbiotic phosphate transport in arbuscular mycorrhizas. *Trends Plant Sci.* 10, 22–29. doi: 10.1016/j.tplants.2004.12.003

Kobayashi, I., Fujiwara, S., Shimogawara, K., Kaise, T., Usuda, H., and Tsuzuki, M. (2003). Insertional mutagenesis in a homologue of a Pi transporter gene confers arsenate resistance on *Chlamydomonas*. *Plant Cell Physiol.* 44, 597–606. doi: 10.1093/pcp/pcg081

Kumar Sharma, A., Muhlroth, A., Jouhet, J., Marechal, E., Alipanah, L., Kissen, R., et al. (2020). The Myb-like transcription factor phosphorus starvation response (PtPSR) controls conditional P acquisition and remodelling in marine microalgae. *New Phytol.* 225, 2380–2395. doi: 10.1111/nph.16248

Li, X., Zhang, R., Patena, W., Gang, S. S., Blum, S. R., Ivanova, N., et al. (2016). An indexed, mapped mutant library enables reverse genetics studies of biological processes in *Chlamydomonas reinhardtii*. *Plant Cell* 28, 367–387. doi: 10.1105/tpc.15.00465

Liu, F., Chang, X.-J., Ye, Y., Xie, W.-B., Wu, P., Lian, X.-M. et al. (2011). Comprehensive sequence and whole-life-cycle expression profile analysis of the phosphate transporter gene family in rice. *Mol. Plant.* 4, 1105–1122. doi: 10.1093/mp/ssr058

Liu, X. L., Wang, L., Wang, X. W., Yan, Y., Yang, X. L., Xie, M. Y., et al. (2019). Mutation of the chloroplast-localized phosphate transporter OsPHT2;1 reduces flavonoids accumulation and UV tolerance in rice. *Plant J.* 102, 53–67. doi: 10.1111/tpj.14611

Lorenzo-Orts, L., Couto, D., and Hothorn, M. (2020). Identity and functions of inorganic and inositol polyphosphates in plants. *New Phytol.* 225, 637–652. doi: 10.1111/nph.16129

Mimura, T., Reid, R. J., Ohsumi, Y., and Smith, F. A. (2002). Induction of the Na+/Pi cotransport system in the plasma membrane of *Chara corallina* requires external Na+ and low levels of Pi. *Plant Cell Environ.* 25, 1475–1481. doi: 10.1046/j.1365-3040.2002.00921.x

Moseley, J. L., Chang, C. W., and Grossman, A. R. (2006). Genome-based approaches to understanding phosphorus deprivation responses and PSR1 control in *Chlamydomonas reinhardtii*. *Eukaryot. Cell* 5, 26–44. doi: 10.1128/ec.5.1.26-44.2006

Moseley, J. L., Gonzalez-Ballester, D., Pootakham, W., Bailey, S., and Grossman, A. R. (2009). Genetic interactions between regulators of *Chlamydomonas* phosphorus and sulfur deprivation responses. *Genetics* 181, 889–905. doi: 10.1534/genetics.108.099382

Mudge, S. R., Rae, A. L., Diatloff, E., and Smith, F. W. (2002). Expression analysis suggests novel roles for members of the Pht1 family of phosphate transporters in *Arabidopsis*. *Plant J.* 31, 341–353. doi: 10.1046/j.1365-313x.2002.01356.x

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Nussaume, L., Kanno, S., Javot, H., Marin, E., Pochon, N., Ayadi, A., et al. (2011). Phosphate import in plants: focus on the PHT1 transporters. *Front. Plant Sci.* 2:83. doi: 10.3389/fpls.2011.00083

Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., et al. (2019). Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. *Genome Biol.* 20:38.

Quisel, J. D., Wykoff, D. D., and Grossman, A. R. (1996). Biochemical characterization of the extracellular phosphatases produced by phosphorus-deprived *Chlamydomonas reinhardtii*. *Plant Physiol.* 111, 839–848. doi: 10.1104/pp.111.3.839

Raghothama, K. G. (1999). Phosphate acquisition. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 50, 665–693.

Reid, R. J., Mimura, T., Ohsumi, Y., Walker, N. A., and Smith, F. A. (2000). Phosphate uptake in chara: membrane transport via Na/Pi cotransport. *Plant Cell Environ.* 23, 223–228. doi: 10.1046/j.1365-3040.2000.00524.x

Shimogawara, K., Wykoff, D. D., Usuda, H., and Grossman, A. R. (1999). *Chlamydomonas reinhardtii* mutants abnormal in their responses to phosphorus deprivation. *Plant Physiol.* 120, 685–694. doi: 10.1104/pp.120.3.685

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W. Z., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7:539. doi: 10.1038/msb.2011.75

Sun, L., Song, L., Zhang, Y., Zheng, Z., and Liu, D. (2016). *Arabidopsis* PHL2 and PHR1 act redundantly as the key components of the central regulatory system controlling transcriptional responses to phosphate starvation. *Plant Physiol.* 170, 499–514. doi: 10.1104/pp.15.01336

Wang, D., Lv, S., Jiang, P., and Li, Y. (2017). Roles, regulation, and agricultural application of plant phosphate transporters. *Front. Plant Sci.* 8:817. doi: 10.3389/fpls.2011.00817

Wang, Z. Y., Ruan, W. Y., Shi, J., Zhang, L., Xiang, D., Yang, C., et al. (2014). Rice SPX1 and SPX2 inhibit phosphate starvation responses through interacting with PHR2 in a phosphate-dependent manner. *Proc. Natl. Acad. Sci. U.S.A.* 111, 14953–14958. doi: 10.1073/pnas.1404680111

Wykoff, D. D., Grossman, A. R., Weeks, D. P., Usuda, H., and Shimogawara, K. (1999). Psr1, a nuclear localized protein that regulates phosphorus metabolism in *Chlamydomonas*. *Proc. Natl. Acad. Sci. U.S.A.* 96, 15336–15341. doi: 10.1073/pnas.96.26.15336

Yu, J., Cheng, Y., Feng, K., Ruan, M., Ye, Q., Wang, R., et al. (2016). Genome-wide identification and expression profiling of tomato Hsp20 gene family in response to biotic and abiotic stresses. *Front. Plant Sci.* 7:1215. doi: 10.3389/fpls.2011.01215

Zhou, J., Jiao, F., Wu, Z., Li, Y., Wang, X., He, X., et al. (2008). OsPHR2 is involved in phosphate-starvation signaling and excessive phosphate accumulation in shoots of plants. *Plant Physiol.* 146, 1673–1686. doi: 10.1104/pp.107.111443

Zhu, W., Miao, Q., Sun, D., Yang, G. D., Wu, C. G., Huang, J. G., et al. (2012). The mitochondrial phosphate transporters modulate plant responses to salt stress via affecting ATP and gibberellin metabolism in *Arabidopsis thaliana*. *PLoS One* 7:e043530. doi: 10.1371/journal.pone.0043530

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.

Copyright © 2020 Wang, Xiao, Yang, Chen, Zeng, Zhao and Zhu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

BRIEF RESEARCH REPORT
published: 09 October 2020
doi: 10.3389/fgene.2020.576086

Recurrent Drought Conditions Enhance the Induction of Drought Stress Memory Genes in *Glycine max* L.

*Yeon-Ki Kim[1], Songhwa Chae[1], Nam-Iee Oh[2], Nguyen Hoai Nguyen[2] and Jong-Joo Cheong[2]**

[1] Department of Biosciences and Bioinformatics, Myongji University, Yongin, South Korea, [2] Center for Food and Bioconvergence, Seoul National University, Seoul, South Korea

Plants remember what they have experienced and are thereby able to confront repeated stresses more promptly and strongly. A subset of the drought responsive genes, called stress memory genes, displayed greatly elevated levels under recurrent drought conditions. To screen for a set of drought stress memory genes in soybean (*Glycine max* L.), we designed a 180K DNA chip comprising 60-bp probes synthesized *in situ* to examine 55,589 loci. Through microarray analysis using the DNA chip, we identified 2,162 and 2,385 genes with more than fourfold increases or decreases in transcript levels, respectively, under initial (first) drought stress conditions, when compared with the non-treated control. The transcript levels of the drought-responsive genes returned to basal levels during recovery (watered) states, and 392 and 613 genes displayed more than fourfold elevated or reduced levels, respectively, under subsequent (second) drought conditions, when compared to those observed under the first drought stress conditions. Gene Ontology and MapMan analyses classified the drought-induced memory genes exhibiting elevated levels of transcripts into several functional categories, including those involved in tolerance responses to abiotic stresses, which encode transcription factors, protein phosphatase 2Cs, and late embryogenesis abundant proteins. The drought-repressed memory genes exhibiting reduced levels of transcripts were classified into categories including photosynthesis and primary metabolism. Co-expression network analysis revealed that the soybean drought-induced and -repressed memory genes were equivalent to 172 and 311 Arabidopsis genes, respectively. The soybean drought stress memory genes include genes involved in the dehydration memory responses of Arabidopsis.

Keywords: DNA chip, drought, stress memory, gene ontology, transcript, microarray, soybean

OPEN ACCESS

Edited by:
Hongjian Wan,
Zhejiang Academy of Agricultural Sciences, China

Reviewed by:
Venkategowda Ramegowda, University of Arkansas, United States
Ashish Kumar Srivastava, Bhabha Atomic Research Centre (BARC), India
Yanwei Li, Zhejiang Academy of Agricultural Sciences, China

***Correspondence:**
Jong-Joo Cheong
cheongjj@snu.ac.kr

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

Received: *25 June 2020*
Accepted: *21 September 2020*
Published: *09 October 2020*

Citation:
Kim Y-K, Chae S, Oh N-I, Nguyen NH and Cheong J-J (2020) Recurrent Drought Conditions Enhance the Induction of Drought Stress Memory Genes in Glycine max L. Front. Genet. 11:576086. doi: 10.3389/fgene.2020.576086

INTRODUCTION

Prolonged or habitual drought may be one of the most serious detrimental stresses for crops during their lifetime. Plants have evolved various strategies to cope with such drought conditions by exhibiting many physiological and developmental changes through regulation of gene expression (Shinozaki and Yamaguchi-Shinozaki, 1996; Shanker et al., 2014). Under drought conditions,

many drought-responsive genes are expressed, and their transcript levels are returned to basal levels during recovery (watered) states. Thus, during recurring cycles of dehydration stresses, their transcript levels rise and fall repeatedly.

Plants can control their responses to repeated stress by altering the expression patterns of the responsive genes. A subset of genes called 'memory genes' are expressed at highly elevated or reduced levels during subsequent dehydration events (Ding et al., 2012; Kinoshita and Seki, 2014; Godwin and Farrona, 2020), thereby enabling plants to respond more promptly and strongly to the repeated drought stress. This response has been referred to as memory, imprinting, priming, training, and acclimation to stress (Bruce et al., 2007; Conrath, 2011; Walter et al., 2011). For example, Ding et al. (2012) showed that Arabidopsis plants trained with previous dehydration events wilted much slower than non-trained plants.

The potential of stress memory for enhancing crop productivity under drought conditions has been explored. In particular, Ramírez et al. (2015) improved potato drought tolerance through the induction of long-term water stress memory. Wang et al. (2014) improved tolerance to drought stress by priming before anthesis in wheat. Abdallah et al. (2017) performed drought priming to improve the productivity of olive under severe drought conditions. In addition, Tabassum et al. (2018) showed that terminal drought and seed priming improved drought tolerance in bread wheat.

The molecular mechanisms underlying stress memory are not yet fully understood. The most plausible mechanism involves epigenetic changes in the chromatin architecture of memory genes, such as promoter DNA methylation, histone modifications, or small RNA generation (Luo et al., 2012; Avramova, 2015; Kim et al., 2015; Lämke and Bäurle, 2017; Chang et al., 2019). Such an epigenetically modified status is transmitted mitotically to newly developed cells during the cell division process.

Furthermore, the epigenetic status of memory could be transmitted meiotically to the next generation of plants, exhibiting transgenerational epigenetic inheritance (Molinier et al., 2006; Boyko and Kovalchuk, 2011; Quadrana and Colot, 2016). The potential of transgenerational inheritance of acquired traits has been applied to improve crop productivity (Springer, 2013; Mickelbart et al., 2015; Bilichak and Kovalchuk, 2016). For instance, Verkest et al. (2015) improved drought tolerance in canola epi-lines by repeatedly selecting for increased drought tolerance for three generations. Tabassum et al. (2017) reported that seed priming and transgenerational drought memory improved tolerance against drought and salt stress in bread wheat. In addition, Raju et al. (2018) developed an epigenetic breeding system in soybean for increased yield and stability, by modulation of development, defense, phytohormone and abiotic stress response pathways.

The first step toward understanding the molecular mechanism of transcriptional memory and its wide application in crop improvement is the identification of stress memory genes. Many dehydration stress memory genes have been identified in Arabidopsis (Ding et al., 2013), *Zea mays* (Ding et al., 2014; Virlouvet et al., 2018), rice (Li et al., 2019), and potato (Chen et al., 2020). In this study, we identified drought stress memory genes in soybean by microarray analysis using a newly designed 180K DNA chip.

MATERIALS AND METHODS

Plant Materials and Treatments

Certified quality soybean seed (*Glycine max* L. cv. Daepoong) was obtained from the Korea Seed and Variety Service. The seeds were sown in potting soil at 15 mm depth in 50-well plates, with one seed per well (35 mm width × 35 mm length × 45 mm depth), and incubated in a growth chamber at 28°C and 50–60% relative humidity under a 16-h light (8,000 lux)/8-h dark photoperiod (**Supplementary Figure 1**). Water was supplied from a watering tray located beneath the plate. After growth for 7 days, when the primary leaves of soybean plantlets were fully open, one of the two plates was exposed to drought stress by withdrawing the watering tray for 4 days. Eight to twelve primary leaves were collected from the watered control (WT1) and drought-treated (DR1) plates, and frozen in liquid nitrogen. Then, the (first) drought-treated plate was re-watered for 1 day (WT2), and the second drought treatment was conducted for 4 days (DR2). This experiment was performed twice independently.

DNA Chip Design

Based on the whole soybean (*G. max* var. Williams 82) genome database available from Phytozome[1] (Goodstein et al., 2012), three 60-nt-long feature probes were designed based on a representative transcript of each gene, beginning 60 bp upstream from the end of the stop codon and shifting downstream at 30-bp intervals; the three resulting probes covered 120 bp: 60 bp of coding sequence (CDS) and 60 bp in the 3′ untranslated region (3′-UTR). Additional probes for alternatively spliced transcripts, chloroplast genes, mitochondrial genes, and selection markers, such as *gfp*, *gus*, *hyg*, *bar*, and *kan*, were included. A total of 175,391 probes were designed and synthesized *in situ* on the 180K DNA chip, which was manufactured by Agilent Technologies (Santa Clara, CA, United States).

Microarray Analysis

Two independent biological replicates of microarray experiments were performed using soybean plants treated with or without water deprivation. Total RNA was isolated from soybean leaves as described previously (Chae et al., 2017), and probes were labeled using the Low RNA Input Linear Amplification Kit PLUS (Agilent Technologies), following the manufacturer's protocol[2]. Total RNA (500 ng) was used to generate labeled cRNA by incorporating cyanine CTP. The microarray was scanned using a DNA microarray scanner (G2505C; Agilent Technologies) and raw intensity data were extracted using the feature extraction software. Background adjustment and normalization of microarray the data were achieved using

[1]http://www.phytozome.net

[2]http://www.agilent.com

the Limma package in the R computing environment[3]. The distribution of probe intensities from eight microarrays was analyzed using the plotDensities function. Frequency densities before and after normalization were plotted against log2-based intensities. Genes exhibiting > 4-fold enhanced transcription levels in two independent experiments were considered to show significantly increased expression. Expression profiling was conducted with the *Glycine max* Alternatively Spliced Transcript Detecting Microarray (GmASDM).

Bioinformatics Analysis

Multiple analyses were performed using Limma (Chae et al., 2016), which adopts the linear modeling approach implemented by lmFit and the empirical Bayesian statistics implemented by eBayes. Genes with an adjusted *P*-value < 0.05 were collected and further assessed in terns of whether they had expression levels > 1 or < -1 for at least one stage compared to the control.

Biological term enrichment among genes was assessed using GoMiner (Ashburner et al., 2000; Zeeberg et al., 2003). To identify a tentative ortholog of a *G. max* gene in the Arabidopsis genome, BLASTP analysis was performed for the two species; genes with a score ≥ 70 were tentatively considered as counterparts. The microarray contained 68,234 transcripts/54,076 genes, 19,750 of which matched Arabidopsis genes with scores ≥ 70; these genes were used as the total gene set in GoMiner. Terms with significant changes in the biological processes category were collected and subjected to hierarchical clustering using the Limma language of the Bioconductor project[4]. False discovery rate (FDR) values were obtained from 100 randomizations. Gene Ontology (GO) terms with FDR values < 0.05 in at least one group were collected. To present GO terms in hierarchical clusters, unique GO terms were first selected. FDR values < 0.05 were scaled from 0.5 to 5, such that the more enriched terms had values closer to 5 and a darker color on the heatmap(see footnote 4). Other non-significant FDR values were transformed to 0 using the Perl scripting language. To reduce redundancy, the top terms in each clade were selected and hierarchical clustering was performed with the Bioconductor R package.

To determine the enriched pathways, MapMan (ver. 3.6) analysis was conducted using AGI_TAIR9 as the mapping database. Pathways were generated with the highest homolog of Arabidopsis genes extracted from soybean gene transcripts, as described above.

Co-expression networks were established for the soybean transcripts using ArabidopsisNet[5]. The database was constructed using 994 microarrays from the Arabidopsis ATH1 Genome Array (GPL198), downloaded from the Gene Expression Omnibus (GEO[6]). The co-expression database was constructed based on Pearson coefficients, as described for RiceArrayNet (Lee et al., 2009) and RapaNet (Kim et al., 2017). To test the clustered genes, correlation information was generated in the Newick format and transformed into network and tree diagrams using the Molecular Evolutionary Genetics Analysis program (MEGA X; Kumar et al., 2018).

Quantitative Reverse-Transcription Polymerase Chain Reaction (qRT-PCR)

Total RNA was extracted from soybean leaves using the Spectrum Plant Total RNA kit (Sigma-Aldrich, St. Louis, MO, United States). First-strand cDNA was synthesized using the SuperScript III First-Strand Synthesis SuperMix for qRT-PCR (Invitrogen, Carlsbad, CA, United States). Quantitative PCR (qPCR) was carried out using SolGent 2 × Real-Time Smart Mix (SolGent, Daejeon, South Korea) with specific primers. Thermocycling and fluorescence detection were performed using the Mx3005P qPCR system (Agilent Technologies). The PCR program was as follows: 95°C for 15 min, followed by 40 cycles of 95°C for 30 s, 60°C for 30 s, and 72°C for 30 s. Soybean 60S rDNA was used as an internal control. The qRT-PCR experiments were repeated independently three times. Statistical analysis was performed using Duncan's test (Duncan, 1955) at a 95% confidence level.

RESULTS

DNA Chip Design

The whole soybean (*G. max* var. Williams 82) genome has been sequenced (Schmutz et al., 2010), and the database is available through Phytozome[7] (Goodstein et al., 2012). In total, 87,977 transcripts from 55,589 loci have been annotated at *Glycine max* Wm82.a2.v1[8]. Among these, single transcripts are expressed from 40,372 loci, whereas 47,605 transcripts are from 15,217 loci with alternative splicing (**Supplementary Table 1A**). We designed a microarray to contain 40,261 single transcripts and 27,873 alternative transcripts. In total, 68,234 transcripts were discerned among the 87,977 annotated transcripts.

In most cases, three 60-nt-long probes were designed from a representative transcript of each gene, starting 60 bp upstream of the stop codon and shifting downstream at 30-bp intervals (**Supplementary Figure 2A**). Thus, the three probes covered 120 bp: 60 bp of the CDS and 60 bp in the 3′-UTR of each transcript. Due to overlapping sequences, only two probes were designed for certain genes to ensure specificity. For some genes, a representative transcript is distinguished by a unique exon (UE) among the alternative spliced transcripts, and an additional probe was designed. In addition, 83 chloroplast and 88 mitochondrial genes (V1_YP) were included in the DNA chip. Several election markers such as *gfp*, *gus*, *hyg*, *bar*, and *kan* were also included as negative controls, to cover 68,954 transcripts. In total, 175,391 probes were designed and synthesized *in situ* on the 180K DNA chip.

[3]http://www.bioconductor.org/

[4]http://master.bioconductor.org/

[5]http://bioinfo.mju.ac.kr/arraynet/

[6]http://www.ncbi.nlm.nih.gov/geo/

[7]www.phytozome.net

[8]http://phytozome.jgi.doe.gov/

Drought Stress Memory Genes in Soybean

To test for genes involved in the recurring water stress, plants were exposed to cycles of watering and water deprivation. The seedlings were grown for 7 days, then exposed to drought stress for 4 days by removal of the water tray (**Supplementary Figure 1**). The first leaves of well-watered (WT1) and drought-treated (DR1) soybeans were collected and used for RNA extraction (**Supplementary Figure 2B**), and transcript levels of drought-responsive genes were analyzed with the 180K soybean DNA chip.

Microarray data were collected from biological replicates and the consistency between samples was tested (**Supplementary Figure 3A**). Log_2-based intensities of two WT1 samples were compared, and the linear model $y = 0.98x + 0.018$ was obtained, with a Pearson correlation coefficient of 0.97. The data were corrected for background and normalized using the Limma package (**Supplementary Figures 3B,C**), as described in the Methods section. Signal intensities after normalization are presented in **Supplementary Table 2A**.

The median intensity of sample WT1 was 231 and its maximum value was 535,859.7. The intensities of photosynthesis genes such as Chloroa_b-bind, RuBisCO_small subunit, and RbcS protein ranged from 366,252 to 393,034. In the first drought stress treatment (DR1), the median intensity of sample DR1 was 232.7 and its maximum value was 538,439.3. The intensities of genes such as Metallothio_2 family, PsbR family, ABA_WDS family, and RuBisCO_small domain proteins ranged from 296,733.2 to 369,185.6. Through the repeated watering treatments, photosynthesis genes such as RuBisCO_small subunit, Chloroa_b-bind, and RbcS proteins returned to their maximum induction levels. Meanwhile, through the recurring drought stress treatments, Dehydrin family, Metallothio_2 family, Tryp_alpha_amyl family, and ABA_WDS family reached the maximum expression levels among the highly induced genes.

From these normalized data, we classified the primary drought-responsive transcripts as those that were induced or repressed at the first drought treatment (DR1). Approximately 88% (60,386) of the 68,954 examined transcripts were detected in the microarray analysis (**Table 1** and **Supplementary Table 2A**). Compared to the well-watered control (WT1), 2,162 transcripts showed 4-fold increases (**Table 1** and **Supplementary Tables 1B, 2B**), whereas 2,385 transcripts showed 4-fold decreases (**Supplementary Table 1B, 2C**). These were categorized as drought-repressed transcripts (DRTs) and drought-induced transcripts (DITs), respectively.

When the DR1 plants were re-watered for 1 day (WT2), transcript levels recovered to those of non-stressed (WT1) plants. As the second drought stress proceeded for 4 days (DR2), transcript levels of the drought-responsive genes increased or decreased to the levels detected in the DR1 plants (**Supplementary Table 2B,C**). However, among them, transcript levels of 613 genes decreased 4-fold, while 392 genes increased 4-fold (**Supplementary Tables 1B, 2D,E**) under the second drought conditions (DR2), when compared to those observed under the first drought stress (DR1) conditions. We hypothesized that the drought-induced expression of these genes was influenced by the previous drought stress (DR1), and thus these genes were designated drought-induced memory transcripts (DIMTs) or drought-repressed memory transcripts (DRMTs), respectively. It is notable that no memory transcript was identified among the chloroplast or mitochondrial genes (**Supplementary Table 1B**).

The microarray datasets have been deposited at the National Center for Biotechnology Information (NCBI) GEO database (see footnote 6) under GEO accession number GSE153660.

GO and MapMan Analyses

To test for enriched GO terms, the DITs, DRTs, DIMTs, and DRMTs were subjected to GO analysis (**Supplementary Figure 4**). The biological function of each gene was inferred based on the conserved sequence motifs (domain) contained in the encoded protein and the most homologous Arabidopsis gene. The terms enriched among the DITs and DIMTs included the trehalose biosynthetic process, response to cold, response to wounding, negative regulation of abscisic acid (ABA)-activated signaling, ethylene-activated signaling pathway response to chitin, heat acclimation, and hyperosmotic salinity response were highly enriched (**Table 1** and **Supplementary Table 3A**). In addition, genes encoding protein phosphatase 2C (PP2C) family proteins and late embryogenesis abundant (LEA) proteins were also included in this group. By contrast, DRTs and DRMTs involved in the chlorophyll biosynthetic process, photosynthesis, light harvesting, and response to red or blue light was significantly repressed (**Table 1** and **Supplementary Table 3A**).

The MapMan analysis revealed that 190 gene transcripts out of 392 DIMTs have Arabidopsis equivalents functioning in the hormonal and transcriptional regulation in abiotic/biotic stress responses (**Supplementary Table 3B**). A number of ABA- and ethylene-responsive genes encoding a putative ABA deficient 2 (Glyma.11G180800-1), ABA 8′-hydroxylase (Glyma.16G109300-1), ABA-responsive protein-related (Glyma.19G100000-1), osmotin 34 (Glyma.11G025600-1), and ethylene responsive factor 1 (Glyma.10G007000-1) were highly upregulated under the second drought conditions (**Figure 1A**). Genes involved in ABA and ethylene synthesis were also induced, as shown by Glyma.08G176300-1 (9-*cis*-epoxycarotenoid dioxygenase), Glyma.08G255400-1 (9-*cis*-epoxycarotenoid dioxygenase), and Glyma.02G084400-1 (2-oxoglutarate-dependent dioxygenase). Among the biotic stress genes, the expression levels of several genes encoding disease resistance proteins, such as TIR-NBS-LRR class (Glyma.11G153000-1 and Glyma.16G214200-1), CC-NBS-LRR class (Glyma.05G082200-2), and downy mildew resistant 6 (Glyma.18G273200-1), were also increased. Genes encoding proteins involved in secondary metabolism, such as phenylpropanoid transferase family protein (Glyma.04G040000-1) and mannitol dehydrogenase (Glyma.07G175200-1, Glyma.16G096300-1, Glyma.01G021000-1, and Glyma.08G280500-1), were also included in the DIMTs.

Among 613 DRMTs, 349 genes have equivalents in Arabidopsis (**Supplementary Table 3C**). The metabolism overview showed that photosynthesis was severely affected, with significantly reduced gene transcript levels (**Figure 1B**).

TABLE 1 | Gene Ontology terms enriched with DRTs, DITs, DRMTs, and DIMTs.[a]

Gene Ontology terms (GO_id)	DRT	DIT	DRMT	DIMT
GO:0005992_trehalose biosynthetic process	0	0	0	1.58
GO:0009409_response to cold	0	5	0	5
GO:0009611_response to wounding	0	5	0	1.58
GO:0009788_negative regulation of ABA-activated signaling	0	1.04	0	1.31
GO:0009873_ethylene-activated signaling pathway	0	5	0	5
GO:0010200_response to chitin	0	1.85	0	5
GO:0010286_heat acclimation	0	5	0	5
GO:0042538_hyperosmotic salinity response	0	5	0	5
GO:0006364_rRNA processing	−1.37	0	−5	0
GO:0006636_unsaturated fatty acid biosynthetic process	−5	0	−5	0
GO:0009637_response to blue light	−5	0	−5	0
GO:0009765_photosynthesis, light harvesting	−5	0	−5	0
GO:0009773_photosynthetic electron transport in photosystem I	−5	0	−5	0
GO:0010075_regulation of meristem growth	−5	0	−5	0
GO:0010103_stomatal complex morphogenesis	−5	0	−5	0
GO:0010114_response to red light	−5	0	−5	0
GO:0010155_regulation of proton transport	−5	0	−5	0
GO:0010207_photosystem II assembly	−1.65	0	−5	0
GO:0010218_response to far red light	−5	0	−5	0
GO:0015995_chlorophyll biosynthetic process	−5	0	−5	0
GO:0016117_carotenoid biosynthetic process	−1.04	0	−5	0
GO:0016556_mRNA modification	−1.05	0	−5	0
GO:0018298_protein-chromophore linkage	−1.55	0	−5	0
GO:0019288_methylerythritol 4-phosphate pathway	−5	0	−5	0
GO:0019344_cysteine biosynthetic process	−5	0	−5	0
GO:0034660_ncRNA metabolic process	0	0	−5	0
GO:0042254_ribosome biogenesis	0	0	−5	0

[a]*Selected from* ***Supplementary Table 3A****.*

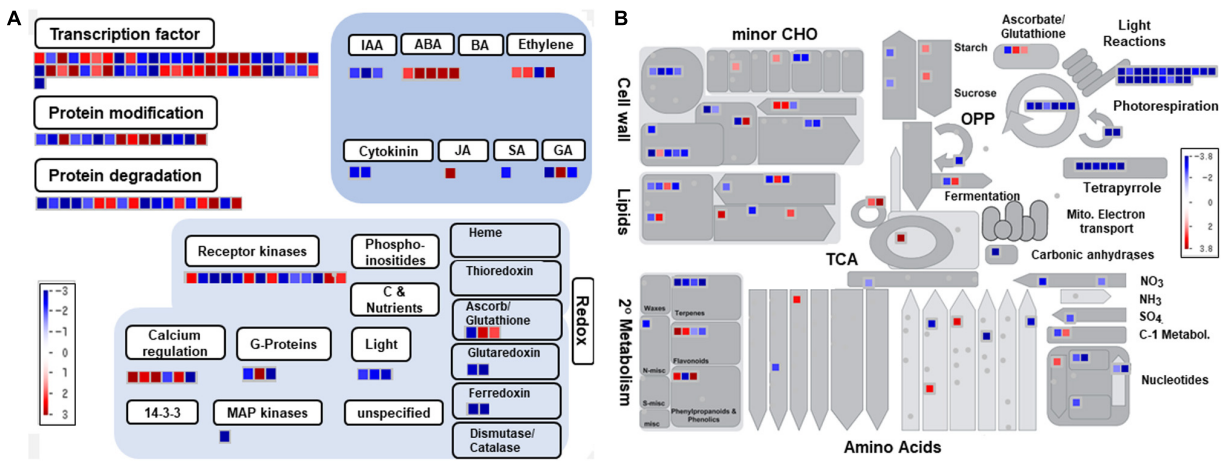


FIGURE 1 | Functional overview of soybean drought-induced and drought-repressed memory gene transcripts. **(A)** Regulation overview. **(B)** Metabolism overview. The MapMan program (ver. 3.6) was used along with AGI_TAIR9 as the mapping database. Red and blue spots represent drought-induced memory and drought-repressed memory signals, respectively.

Those genes encode a putative chloroplast thylakoid membrane protein (Glyma.07G019400-1), light harvesting complex PSII subunit 6 (Glyma.05G119000-1 and Glyma.08G074000-1), cytochrome b6f complex subunit (Glyma.07G163600-1, Glyma.20G008600-1, and Glyma.07G163600-1), chlorophyll binding (Glyma.16G162600-1, Glyma.16G165200-1, Glyma.16G165800-1, and Glyma.16G165500-1), photosystem II light harvesting complex protein 2.1 (Glyma.02G305400-1),

and ATPC1 (Glyma.13G204800-2). In addition, genes encoding proteins involved in primary metabolic pathways, such as oxidoreductase (Glyma.12G222200-1, Glyma.12G150400-1, Glyma.12G222200-1, and Glyma.06G247100-1), tetrapyrrole binding (Glyma.12G052500-1 and Glyma.11G127800-1), and glutamyl-tRNA reductase (Glyma.02G218300-1 and Glyma.14G185700-1), were repressed. Hormone metabolism was also affected by cytochrome b561/ferric reductase (Glyma.07G055800-1), and auxin-responsive family protein (Glyma.02G049200-1), GAST1 protein homolog 4 (Glyma.09G238300-1), and gibberellin 2-beta-dioxygenase (Glyma.11G003200-1) expression levels were decreased. Protein degradation processes mediated by serine-type endopeptidase (Glyma.05G152200-1), subtilase family protein (Glyma.16G018900-1 and Glyma.07G050200-1), and serine-type endopeptidase (Glyma.04G044600-1, Glyma.06G045100-1, Glyma.14G087500-1, and Glyma.17G236800-1) were repressed.

Co-expression Networks

The 392 transcripts in the DIMT group were equivalent to 172 Arabidopsis genes that constitute 11 groups including 6 major clusters, as shown in the network diagram (**Figure 2A**). The largest group, cluster I, comprises 68 soybean gene transcripts equivalent to 34 Arabidopsis genes (**Supplementary Table 3D**). This group of genes includes Glyma.19G147200-1 (LEA group 1 domain-containing protein), Glyma.10G277800-1 (Myb family transcription factor), Glyma.19G009100-1 (formate dehydrogenase), Glyma.01G195500-2 (ACT domain repeat 1), Glyma.06G177800-1 (C2 domain-containing protein/GRAM domain-containing protein), and Glyma.12G088300-1 (NAD^+ ADP-ribosyltransferase). The second largest group (cluster II) was composed of 31 soybean gene transcripts equivalent to 17 Arabidopsis members, including Glyma.08G178100-1 (2-nitropropane dioxygenase family/NPD family), Glyma.16G203500-1 (C_3HC_4-type ring finger protein), Glyma.19G137400-1 (transferase, transferring glycosyl groups), Glyma.11G242900-1 (epsin N-terminal homology domain-containing protein/clathrin assembly protein-related), and Glyma.01G146500-1 (protein binding/ubiquitin-protein ligase/zinc ion binding).

Using the 613 DRMTs, a tree diagram consisting of 102 clusters was produced with 311 Arabidopsis equivalents (**Figure 2B**). The first major group was composed of 341 transcripts that equivalent to 203 Arabidopsis genes (**Supplementary Table 3E**), including those encoding cytochrome b6f complex subunit (Glyma.07G163600-1, Glyma.20G008600-1, and Glyma.07G163600-1), chlorophyllide a oxygenase (Glyma.14G150600-1), and photosystem II reaction center PsbP family protein (Glyma.03G230300-1 and Glyma.19G227400-1).

qRT-PCR Analysis

We randomly selected eight DIMT genes that appeared to be involved in drought memory response and performed qRT-PCR analysis to confirm the microarray data. For each gene, a set of specific PCR primers was designed (**Supplementary Figure 5**), and their sequence specificity was examined by PCR with soybean genomic DNA.

The DIMT genes we tested included those containing domains encoding thaumatin (osmotin), WRKY, SMP (responsive ABA), MYB, NAM (NAC), PP2C, AP2 (DREB-responsive element), and LEA. These genes were also highly induced in the secondary drought stress conditions (**Figure 3**), confirming that the overall qRT-PCR data were consistent with the microarray data. In this experiment, the expression pattern of a PP2C gene (Glyma.14G162100-1) homologous to Arabidopsis *HAI1* was tested as a negative control. Transcript levels of this gene in the first and second stage of drought stress were comparable to each other. Thus, this gene is not a stress memory gene, but a simple drought-responsive gene.

DISCUSSION

We identified 392 DIMTs and 613 DRMTs in soybean through microarray analysis (**Supplementary Table 1B**). The DIMTs exhibiting significantly elevated transcript levels were involved in tolerance responses to abiotic stresses. The DIMTs include those encoding transcription factors, a trehalose biosynthesis enzyme, LEA proteins, and PP2C family proteins. Trehalose is a non-reducing disaccharide that has high water-holding activity and thereby maintains membrane fluidity under low-water conditions (Iordachescu and Imai, 2008; Delorge et al., 2014). LEA proteins are extremely hydrophilic; thus, the accumulation of these proteins gives plants the desiccation tolerance necessary for dealing with water-deficit stress (Hundertmark and Hincha, 2008; Hand et al., 2011). PP2Cs negatively regulate ABA-activated signaling by counteracting protein kinases, allowing plant cells to maintain the phosphorylation balance needed to control ABA signaling processes (Ma et al., 2009; Umezawa et al., 2009). By contrast, DRMTs exhibiting highly reduced levels of transcripts under second drought conditions include those involved in photosynthesis and primary metabolism.

However, it appears that gene sequence homology or protein structure does not determine whether the gene displays a memory function or is simply a drought-responsive gene. For example, among the 31 drought-induced transcripts encoding PP2C family proteins, only 10 exhibited expression patterns indicative of memory function (**Table 1**). As confirmed in a qPCR experiment (**Figure 3**), transcript levels of the PP2C gene (Glyma.14G162100-1) homologous to Arabidopsis *HAI1* in the first and second stage of drought stress were comparable to each other. Thus, this gene is not a stress memory gene; rather, it is a simple drought-responsive gene. Supporting this observation, the Arabidopsis gene *RD29B* displayed a memory function, whereas its homologous copy *RD29A* was non-trainable during repeated dry stress (Ding et al., 2012). Revealing the molecular mechanisms of transcriptional memory responses is critical for defining the characteristics of stress memory genes.

The soybean drought-induced stress memory genes identified in this study included those involved in the dehydration memory responses of Arabidopsis (Ding et al., 2013). In Arabidopsis, genes implicated in responses to salt, salinity,

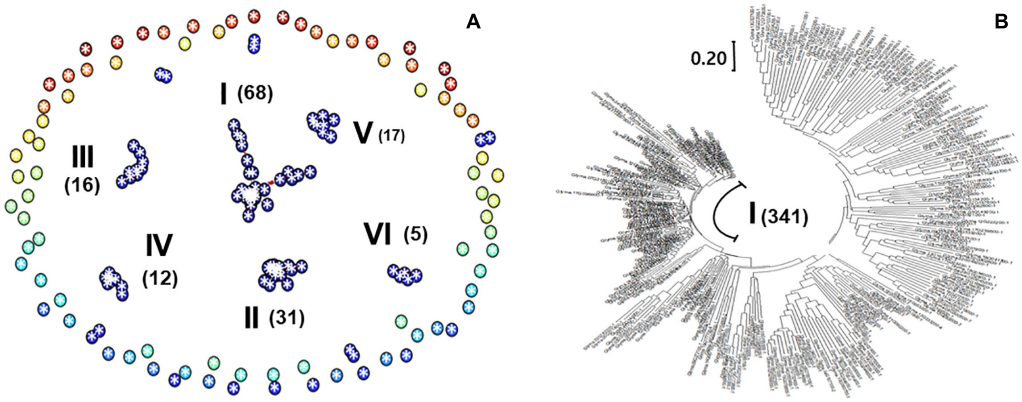


FIGURE 2 | Co-expression networks. The criteria for a depth search of 0 with an absolute correlation value of 0.65 were applied in ArabidopsisNet (http://bioinfo.mju.ac.kr/arraynet/Arabidopsis). **(A)** Network diagram for soybean drought-induced memory genes. Circles represent soybean genes equivalent to Arabidopsis genes. The 392 soybean transcripts equivalent to 172 Arabidopsis genes form 11 groups, including six major clusters (I–VI). Numbers of associated members are presented in the parentheses. **(B)** A tree diagram of soybean drought-repressed gene transcripts. A tree consisting of 102 clusters was produced for 613 soybean drought-repressed gene transcripts. The Newick format was first employed with 349 Arabidopsis equivalent genes; then, the genes were replaced with representative soybean transcripts. The tree was drawn with the MEGA X program (Kumar et al., 2018).

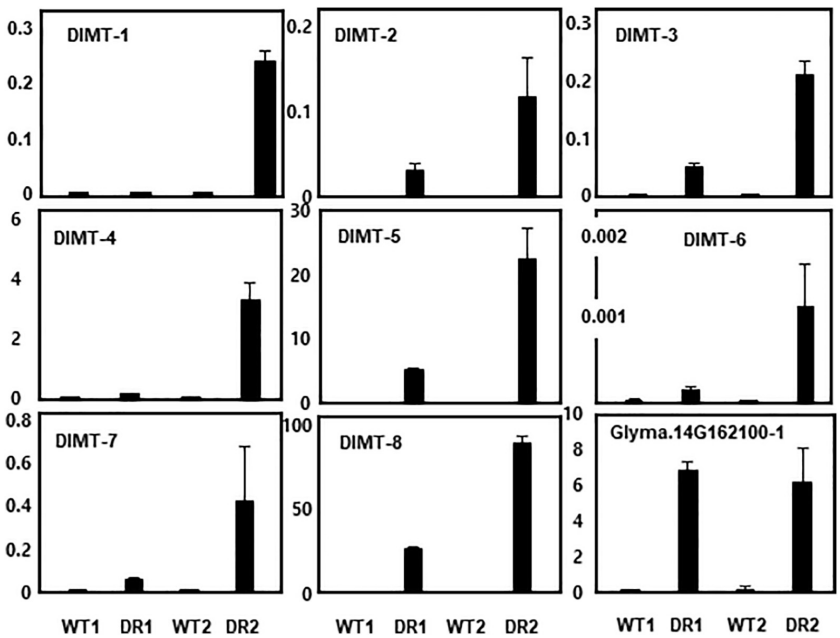


FIGURE 3 | qRT-PCR analysis of selected drought stress memory genes (DIMTs). Eight DIMTs were selected and their expression levels were determined by qRT-PCR. DIMT-1, Glyma.11G025600-1; DIMT-2, Glyma.06G061900-1; DIMT-3, Glyma.U018200-1; DIMT-4, Glyma.04G042300-1; DIMT-5, Glyma.12G149100-1; DIMT-6, Glyma.12G116800-1, DIMT-7, Glyma.20G155100-1; DIMT-8, Glyma.03G144400-1. Transcript levels of a drought inducible gene (Glyma.14G162100-1) were used as a negative control (non-stress memory gene).

cold/heat acclimation, and ABA constitute approximately one quarter of the drought-induced memory genes, in addition to LEA genes (Ding et al., 2013). Genes of chloroplast and thylakoid membrane-associated functions comprised the dehydration-repressed memory genes. Co-expression networks revealed that the soybean drought memory genes were equivalent to Arabidopsis genes having similar functions. Thus, dehydration stress memory in both dicot plants implies reinforcement of the cellular metabolic processes needed for adaptive and protective functions and repression of photosynthesis.

However, studies of other crop plants identified drought memory genes that function in species-specific ways, as well as genes that function similarly. *Z. mays* exhibited a conserved pattern of memory responses to dehydration stress, as also observed in Arabidopsis (Ding et al., 2014).

However, comparison of the cellular functions encoded by the memory genes in the two species revealed remarkable differences. *Z. mays* proteins associated with only four chloroplast and two thylakoid membrane functions were encoded by drought-repressed memory genes, in contrast with the 128 Arabidopsis drought-repressed memory genes implicated in these functions. *Z. mays* genes implicated in abiotic stress responses were also represented by much lower numbers of drought-repressed memory genes than in Arabidopsis.

In the case of potato, drought was memorized by genes involved in photosynthesis, signal transduction, sugar metabolism, flavonoid metabolism, and the biosynthesis of protease and protease inhibitors, transporters, and transcription factors (Chen et al., 2020). The expression levels of most photosynthesis-related genes during the second drought were higher than during the first drought. In addition, Li et al. (2019) identified clusters of rice stress memory gene transcripts involved in ABA signaling, the production of protective substances, and processes such as photosynthesis and DNA methylation. Interestingly, most memory transcripts associated with photosynthesis were sharply reduced by the first drought treatment, and then maintained at a stable level with the subsequent drought treatments. Thus, it appears that photosynthesis efficiency in rice plants was decreased by the initial drought stress, but recovered and improved during the subsequent treatments.

Wide exploitation of drought stress memory to improve crop yield has been hampered by the fact that stress-induced memories are relatively short term (Kinoshita and Seki, 2014; Avramova, 2015). In Arabidopsis, dehydration stress-induced transcriptional memory persists for 5 days under watered conditions (Ding et al., 2012). In most cases, the duration of a somatic memory is limited to one generation (Lämke and Bäurle, 2017). The acquired stress memory may be reset during recovery, subsequent growth, and meiosis, as observed in mammalians (Crisp et al., 2016). How plants overcome such resetting processes during meiosis, to transmit the stress memory to the next generation, remains to be intensively investigated.

The drought stress memory genes identified in this study will facilitate research on the epigenetic mechanisms and biotechnological applications of drought stress memory. Studies of the chromatin architecture of the gene loci should make it possible to understand the nature of the memory response. In addition, the memory genes could be used as markers to assess the transmission of an acquired memory across generations in soybean, and to develop proper priming techniques (method, strength, frequency, interval, etc.) to prolong the memory in the next generation or even farther.

DATA AVAILABILITY STATEMENT

The data can be accessed using the accession number GSE153660 on NCBI.

AUTHOR CONTRIBUTIONS

Y-KK designed the DNA chip and conducted the bioinformatics analyses. SC, N-IO, and NN performed the experiments. J-JC designed the whole research scheme and interpreted the data. Y-KK and J-JC wrote the manuscript. J-JC edited the manuscript. All the authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry (grant number 115080-2 to J-JC), and the National Research Foundation of Korea (grant numbers 2018R1D1A1B07049348 to Y-KK and 2020R111A1A01070089 to J-JC).

ACKNOWLEDGMENTS

The authors thank the research staff of GreenGene Biotech (Yongin, South Korea) for technical assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.576086/full#supplementary-material

REFERENCES

Abdallah, M. B., Methenni, K., Nouairi, I., Zarrouk, M., and Youssef, N. B. (2017). Drought priming improves subsequent more severe drought in a drought-sensitive cultivar of olive cv. *Chétoui*. *Sci. Hortic.* 221, 43–52. doi: 10.1016/j.scienta.2017.04.021

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. doi: 10.1038/75556

Avramova, Z. (2015). Transcriptional 'memory' of a stress: transient chromatin and memory (epigenetic) marks at stress-response genes. *Plant J.* 83, 149–159. doi: 10.1111/tpj.12832

Bilichak, A., and Kovalchuk, I. (2016). Transgenerational response to stress in plants and its application for breeding. *J. Exp. Bot*. 67, 2081–2092. doi: 10.1093/jxb/erw066

Boyko, A., and Kovalchuk, I. (2011). Genome instability and epigenetic modification-heritable responses to environmental stress? *Curr. Opin. Plant Biol.* 14, 260–266. doi: 10.1016/j.pbi.2011.03.003

Bruce, T. J. A., Matthes, M. C., Napier, J. A., and Pickett, J. A. (2007). Stressful "memories" of plants: evidence and possible mechanisms. *Plant Sci.* 173, 603–608. doi: 10.1016/j.plantsci.2007.09.002

Chae, S., Kim, J. S., Jun, K. M., Lee, S.-B., Kim, M. S., Nahm, B. H., et al. (2017). Analysis of genes with alternatively spliced transcripts in the leaf, root, panicle and seed of rice using a long oligomer microarray and RNA-Seq. *Mol. Cell.* 40, 714–730. doi: 10.14348/molcells.2017.2297

Chae, S., Kim, J. S., Jun, K. M., Pahk, Y.-M., Kim, M.-J., Lee, S.-B., et al. (2016). Analysis of representative organ-specific genes and promoters of rice using a 3' ORF-oriented long oligomer microarray. *J. Plant Biol.* 59, 579–593. doi: 10.1007/s12374-016-0276-0

Chang, Y.-N., Zhu, C., Jiang, J., Zhang, H., Zhu, J.-K., and Duan, C.-G. (2019). Epigenetic regulation in plant abiotic stress responses. *J. Integr. Plant Biol.* 62, 563–580. doi: 10.1111/jipb.12901

Chen, Y., Li, C., Yi, J., Yang, Y., Lei, C., and Gong, M. (2020). Transcriptome response to drought, rehydration and re-dehydration in potato. *Int. J. Mol. Sci.* 21:159. doi: 10.3390/ijms21010159

Conrath, U. (2011). Molecular aspects of defence priming. *Trends Plant Sci.* 16, 524–531. doi: 10.1016/j.tplants.2011.06.004

Crisp, P., Ganguly, D., Eichten, S. R., Borevitz, J. O., and Pogson, B. J. (2016). Reconsidering plant memory: intersections between stress recovery, RNA turnover, and epigenetics. *Sci. Adv.* 2:e1501340. doi: 10.1126/sciadv.1501340

Delorge, I., Janiak, M., Carpentier, S., and Van Dijck, P. (2014). Fine tuning of trehalose biosynthesis and hydrolysis as novel tools for the generation of abiotic stress tolerant plants. *Front. Plant Sci.* 5:147. doi: 10.3389/fpls.2014.00147

Ding, Y., Fromm, M., and Avramova, Z. (2012). Multiple exposures to drought 'train' transcriptional responses in *Arabidopsis*. *Nat. Commun*. 3:740. doi: 10.1038/ncomms1732

Ding, Y., Liu, N., Virlouvet, L., Riethoven, J.-J., Fromm, M., and Avramova, Z. (2013). Four distinct types of dehydration stress memory genes in *Arabidopsis thaliana*. *BMC Plant Biol.* 13:229. doi: 10.1186/1471-2229-13-229

Ding, Y., Virlouvet, L., Liu, N., Riethoven, J.-J., Fromm, M., and Avramova, Z. (2014). Dehydration stress memory genes of *Zea mays*; comparison with *Arabidopsis thaliana*. *BMC Plant Biol.* 14:141. doi: 10.1186/1471-2229-14-141

Duncan, D. B. (1955). Multiple range and multiple F tests. *Biometrics* 11, 1–42.

Godwin, J., and Farrona, S. (2020). Plant epigenetic stress memory induced by drought: a physiological and molecular perspective. *Methods Mol. Biol.* 2093, 243–259. doi: 10.1007/978-1-0716-0179-2_17

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nuc. Acids Res.* 40, D1178–D1186. doi: 10.1093/nar/gkr944

Hand, S. C., Menze, M. A., Toner, M., Boswell, L., and Moore, D. (2011). LEA proteins during water stress: not just for plants anymore. *Annu. Rev. Physiol.* 73, 115–134. doi: 10.1146/annurev-physiol-012110-142203

Hundertmark, M., and Hincha, D. K. (2008). LEA (Late Embryogenesis Abundant) proteins and their encoding genes in *Arabidopsis thaliana*. *BMC Genomics* 9:118. doi: 10.1186/1471-2164-9-1/118

Iordachescu, M., and Imai, R. (2008). Trehalose biosynthesis in response to abiotic stresses. *J. Integr. Plant Biol.* 50, 1223–1229. doi: 10.1111/j.1744-7909.2008.00736.x

Kim, J., Jun, K. M., Kim, J. S., Chae, S., Pahk, Y. M., Lee, T.-H., et al. (2017). RapaNet: a Web tool for the co-expression analysis of *Brassica rapa* Genes. *Evol. Bioinform.* 13, 1–11. doi: 10.1177/1176934317715421

Kim, J.-M., Sasaki, T., Ueda, M., Sako, K., and Seki, M. (2015). Chromatin changes in response to drought, salinity, heat, and cold stresses in plants. *Front. Plant Sci.* 8:114. doi: 10.3389/fpls.2015.00114

Kinoshita, T., and Seki, M. (2014). Epigenetic memory for stress response and adaptation in plants. *Plant Cell Physiol.* 55, 1859–1863. doi: 10.1093/pcp/pcu125

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096

Lämke, J., and Bäurle, I. (2017). Epigenetic and chromatin-based mechanisms in environmental stress adaptation and stress memory in plants. *Genome Biol.* 18:124. doi: 10.1186/s13059-017-1263-6

Lee, T.-H., Kim, Y.-K., Pham, T. T. M., Song, S. I., Kim, J.-K., Kang, K. Y., et al. (2009). A database for correlating gene expression from transcriptome profiling, and its application to the analysis of coexpressed genes in rice. *Plant Physiol.* 151, 16–33. doi: 10.1104/pp.109.139030

Li, P., Yang, H., Wang, L., Liu, H., Huo, H., Zhang, C., et al. (2019). Physiological and transcriptome analyses reveal short-term responses and formation of memory under drought stress in rice. *Front. Genet.* 10:55. doi: 10.3389/fgene.2019.00055

Luo, M., Liu, X., Singh, P., Cui, Y., Zimmerli, L., and Wu, K. (2012). Chromatin modifications and remodeling in plant abiotic stress responses. *Biochim. Biophys. Acta* 1819, 129–136. doi: 10.1016/j.bbagrm.2011.06.008

Ma, Y., Szostkiewicz, I., Korte, A., Moes, D., Yang, Y., Christmann, A., et al. (2009). Regulators of PP2C phosphatase activity function as abscisic acid sensors. *Science* 324, 1064–1068. doi: 10.1126/science.1172408

Mickelbart, M. V., Hasegawa, P. M., and Bailey-Serres, J. (2015). Genetic mechanisms of abiotic stress tolerance that translate to crop yield stability. *Nat. Rev. Genet.* 16, 237–251. doi: 10.1038/nrg3901

Molinier, J., Ries, G., Zipfel, C., and Hohn, B. (2006). Transgeneration memory of stress in plants. *Nature* 442, 1046–1049. doi: 10.1038/nature05022

Quadrana, L., and Colot, V. (2016). Plant transgenerational epigenetics. *Annu. Rev. Genet.* 50, 467–491. doi: 10.1146/annurev-genet-120215-035254

Raju, S. K. K., Shao, M.-R., Sanchez, R., Xu, Y.-Z., Sandhu, A., Graef, G., et al. (2018). An epigenetic breeding system in soybean for increased yield and stability. *Plant Biotechnol. J.* 16, 1836–1847. doi: 10.1111/pbi.12919

Ramírez, D. A., Rolando, J. L., Yactayo, W., Monneveux, P., Mares, V., and Quiroz, R. (2015). Improving potato drought tolerance through the induction of long-term water stress memory. *Plant Sci.* 238, 26–32. doi: 10.1016/j.plantsci.2015.05.016

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature* 463, 178–183. doi: 10.1038/nature08670

Shanker, A. K., Maheswari, M., Yadav, S. K., Desai, S., Bhanu, D., Attal, N. B., et al. (2014). Drought stress responses in crops. *Funct. Integr. Genom.* 14, 11–22. doi: 10.1007/s10142-013-0356-x

Shinozaki, K., and Yamaguchi-Shinozaki, K. (1996). Molecular responses to drought and cold stress. *Curr. Opin. Biotechnol.* 7, 161–167. doi: 10.1016/s0958-1669(96)80007-3

Springer, N. M. (2013). Epigenetics and crop improvement. *Trends Genet.* 29, 241–247. doi: 10.1016/j.tig.2012.10.009

Tabassum, T., Farooq, M., Ahmad, R., Zohaib, A., and Wahid, A. (2017). Seed priming and transgenerational drought memory improves tolerance against salt stress in bread wheat. *Plant Physiol. Biochem.* 118, 362–369. doi: 10.1016/j.plaphy.2017.07.007

Tabassum, T., Farooq, M., Ahmad, R., Zohaib, A., Wahid, A., and Shahid, M. (2018). Terminal drought and seed priming improves drought tolerance in wheat. *Physiol. Mol. Biol. Plants* 24, 845–856. doi: 10.1007/s12298-018-0547-y

Umezawa, T., Sugiyama, N., Mizoguchi, M., Hayashi, S., Myouga, F., Yamaguchi-Shinozaki, K., et al. (2009). Type 2C protein phosphatases directly regulate abscisic acid-activated protein kinases in *Arabidopsis*. *Proc. Natl. Acad. Sci. U.S.A.* 106, 17588–17593. doi: 10.1073/pnas.0907095106

Verkest, A., Byzova, M., Martens, C., Willems, P., Verwulgen, T., Slabbinck, B., et al. (2015). Selection for improved energy use efficiency and drought tolerance in canola results in distinct transcriptome and epigenome changes. *Plant Physiol.* 168, 1338–1350. doi: 10.1104/pp.15.00155

Virlouvet, L., Avenson, T. J., Du, Q., Zhang, C., Liu, N., Fromm, M., et al. (2018). Dehydration stress memory: gene networks linked to physiological responses during repeated stresses of *Zea mays*. *Front. Plant Sci.* 9:1058. doi: 10.3389/fpls.2018.01058

Walter, J., Nagy, L., Hein, R., Rascher, U., Beierkuhnlein, C., Willner, E., et al. (2011). Do plants remember drought? hints towards a drought-memory in grasses. *Environ. Exp. Bot.* 71, 34–40. doi: 10.1016/j.envexpbot.2010.10.020

Wang, X., Vignjevic, M., Jiang, D., Jacobsen, S., and Wollenweber, B. (2014). Improved tolerance to drought stress after anthesis due to priming before anthesis in wheat (*Triticum aestivum* L.) var. Vinjett. *J. Exp. Bot.* 65, 6441–6456. doi: 10.1093/jxb/eru362

Zeeberg, B. R., Feng, W., Wang, G., Wang, M. D., Fojo, A. T., Sunshine, M., et al. (2003). GoMiner, a resource for biological interpretation of genomic and proteomic data. *Genome Biol.* 4:R28. doi: 10.1186/gb-2003-4-4-r28

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.

Copyright © 2020 Kim, Chae, Oh, Nguyen and Cheong. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

METHODS
published: 09 October 2020
doi: 10.3389/fgene.2020.569100

Identification of Proteins of Tobacco Mosaic Virus by Using a Method of Feature Extraction

*Yu-Miao Chen, Xin-Ping Zu and Dan Li**

Information and Computer Engineering College, Northeast Forestry University, Harbin, China

Tobacco mosaic virus, TMV for short, is widely distributed in the global tobacco industry and has a significant impact on tobacco production. It can reduce the amount of tobacco grown by 50–70%. In this research of study, we aimed to identify tobacco mosaic virus proteins and healthy tobacco leaf proteins by using machine learning approaches. The experiment's results showed that the support vector machine algorithm achieved high accuracy in different feature extraction methods. And 188-dimensions feature extraction method improved the classification accuracy. In that the support vector machine algorithm and 188-dimensions feature extraction method were finally selected as the final experimental methods. In the 10-fold cross-validation processes, the SVM combined with 188-dimensions achieved 93.5% accuracy on the training set and 92.7% accuracy on the independent validation set. Besides, the evaluation index of the results of experiments indicate that the method developed by us is valid and robust.

Keywords: feature extraction, physicochemical properties, identification, tobacco mosaic virus, machine learning

OPEN ACCESS

Edited by:
Xiaoming Song,
North China University of Science and Technology, China

Reviewed by:
Fenglong Wang,
Tobacco Research Institute (CAAS), China
Guang Song,
Johns Hopkins University, United States
Hao Lin,
University of Electronic Science and Technology of China, China

****Correspondence:***
Dan Li
ld@nefu.edu.cn

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

Received: *03 June 2020*
Accepted: *09 September 2020*
Published: *09 October 2020*

Citation:
Chen Y-M, Zu X-P and Li D (2020) Identification of Proteins of Tobacco Mosaic Virus by Using a Method of Feature Extraction. Front. Genet. 11:569100. doi: 10.3389/fgene.2020.569100

INTRODUCTION

Tobacco mosaic virus is worldwide distribution and is the furthest invasive virus which is most harmful to crops. Tobacco is one of the important economic crops in our country, however, the existence of tobacco mosaic disease has greatly reduced the yield and quality of tobacco. Since plants do not have a complete immune system, once infected, the leaves can show mosaic symptoms or even deformities and the growth can also be chronically diseased, which makes tobacco mosaic virus is very difficult to control (Hu and Lee, 2015).

The study of viruses has attracted many scholars, and with the development of computer machine learning algorithms, many scholars have applied machine learning algorithms to the study of viruses. Metzler and Kalinina (2014) used one-class SVM method to detect atypical genes in viral families based on their statistical features, without the need for explicit knowledge of the source species. The simplicity of the statistical features used allows the method to be applied to a variety of viruses. Salama et al. (2016) predicted new drug-resistant strains that facilitate the design of antiviral therapies. In this study, neural network techniques were used to predict new strains, and using a rough set theory based on algorithm to extract these points mutation patterns. For phage virion proteins (PVPs) prior to *in vitro*, Manavalan et al. (2018) developed a SVM-based predictor that exhibited good performance and avoided the expensive costs required for experiments.

Using biochemical experiments to study all tobacco mosaic virus is a challenge because it is expensive and a waste of researchers' time, and there is no specific predictor to predict tobacco mosaic virus. So, in this research we evaluated the predictive performance of different classifiers in combination with different feature extraction methods. We have chosen classical machine

learning algorithms and classical feature extraction methods. The feature extraction methods AAC (Chou, 2001), 188-dimensions (Dubchak et al., 1995) and CKSAAGP (Chen et al., 2018) and their combination were chosen for the reasons. AAC is the first proposed feature extraction method that is widely used to predict the function of proteins. It based on their amino acid composition. CKSAAGP describes the spatial distribution information of amino acids, 188 dimensions in addition to the physicochemical properties of amino acids. Three feature extraction methods from different aspects, so these three feature extraction methods were chosen. The combined feature extraction method was attempted considering the expectation of better results. We finally chose the combination of support vector machine (SVM) with 188-dimensions as the final predictor because it has the best prediction effect.

MATERIALS AND METHODS

Our method was developed based on three steps (**Figure 1**). Step 1: we collected the data and preprocessed the dataset to obtain a non-redundant benchmark dataset that does not contain non-standard characters. Step 2: We used Amino Acid Composition (AAC), feature extraction method based on the composition of amino acid sequence and physicochemical properties (188-dimensions), composition of k-spaced amino acid pairs (CKSAAGP), and the combined methods AAC_CKSAAGP and 188_CKSAAGP which are proposed in this paper to extract features from protein sequences. Step 3: Five algorithms and 10-fold cross-validation are used to build and estimate the models, which are Random Forest (RF), Bagging, K-Nearest Neighbor (KNN), Naive Bayes (NB), and Support Vector Machine (SVM). Then, we validated experimental results using an independent validation set.

BENCHMARK DATASET

High-quality baseline data sets contribute to the accuracy of model predictions (Yang et al., 2019b; Cheng et al., 2020; Zhu et al., 2020). The dataset obtained for this experiment was derived from the Swiss-Prot database in The Uniprot (2018). Firstly, we used the keyword search method to collect data from the UniProt database. By entering the keywords "Tobacco mosaic virus" and "Tobacco leaf not virus" to obtain the positive and negative data needed for the experiment. For the sake of improving the reliability of the data, the following operations were performed: (1), Deleted the protein sequences containing non-standard letters, i.e. "B," "X," "Z," etc.; thus, we obtained 5,309 protein sequences of tobacco mosaic virus and 45,827 protein sequences of non-tobacco mosaic virus. (2), If the sample contains multiple similarity sequences, this sample is not statistically representative. We used the CD-HIT program (Fu et al., 2012) to delete sequences with similarity surpass 40% in positive and negative data sets (Zou et al., 2018). After removing the redundant sequences, we eventually obtained a dataset of 715 protein sequences of TMV proteins and 17,983 protein sequences of tobacco leaf proteins.

There are 715 sequences in the positive datasets and 17,893 sequences in the negative datasets. Much more negative data than positive data. For the purpose of balancing the datasets, we took a downsampling approach. We split the negative data by the size of the positive data. And randomly selected 10 of these copies as the negative dataset, so we obtained a negative dataset containing 7,150 sequences. The resulting positive and negative datasets were divided proportionally. The final training dataset consists of 500 positive data and 5,000 negative data. The test dataset consists of 215 positive data and 2,150 negative data. These data are available in our software package.

FEATURE EXTRACTION

Feature selection will affect the performance of machine learning methods for bioinformatics problems (Zhao et al., 2015). In the research of this paper, five feature extraction methods are selected, including amino acid composition (AAC), composition of k-spaced amino acid pairs (CKSAAGP), 188-dimensions feature extraction method, and the combined methods AAC_CKSAAGP and 188_CKSAAGP which are proposed in this paper.

Amino Acid Composition (AAC)

The coded amino acid composition coding scheme (Bhasin and Raghava, 2004) calculates the probability of occurrence of 20 natural amino acids (i.e., "ACDEFGHIKLMNPQRSTVWY") in protein sequences or peptide chains (Zhong et al., 2020). The calculation formula for each amino acid is as follows:

$$v_i = \frac{c_i}{len\left(seq\right)}, i \in (A, C, \cdots, Y) \tag{1}$$

Where c_i and $len(seq)$ represent the number of occurrences of amino acid i in the sequence or peptide chain and the length of the sequence or peptide chain, respectively (Lin et al., 2005; Lv et al., 2020).

Composition of k-Spaced Amino Acid Pairs (CKSAAGP)

The composition of K-spaced amino acid pairs (Chen et al., 2018) can be regarded as a variant of CKSAAP, which calculates the frequency of amino acid pairs separated by any k residues (the default maximum for k is set to 5) (Chen et al., 2020). Taking K = 0 as an e.g., a feature vector is defined as:

$$\gamma_0 = \left(\frac{C_{g1g1}}{C_n}, \frac{C_{g1g2}}{C_n}, \frac{C_{g1g3}}{C_n}, \cdots, \frac{C_{g5g5}}{C_n}\right)_{25} \tag{2}$$

wherein, $(g1g1, g1g2, g1g3, \cdots g5g5)$ represents 0-spacing amino acid group pairs. There are 25 groups in total, and each descriptor represents the composition of the corresponding residue pair in the protein sequence (Zhu et al., 2020). C_{g1g1} (Zhang et al., 2014) represents the number of times the residue pair $g1g1$ appears in the sequence and Cn represents the total number of residue pairs with a gap of 0 in the sequence. In a protein sequence of length N, for different values of K, the value of n can be defined as:

$$n = P - K - 1, K \in (0, 1, 2, 3, 4, 5) \tag{3}$$

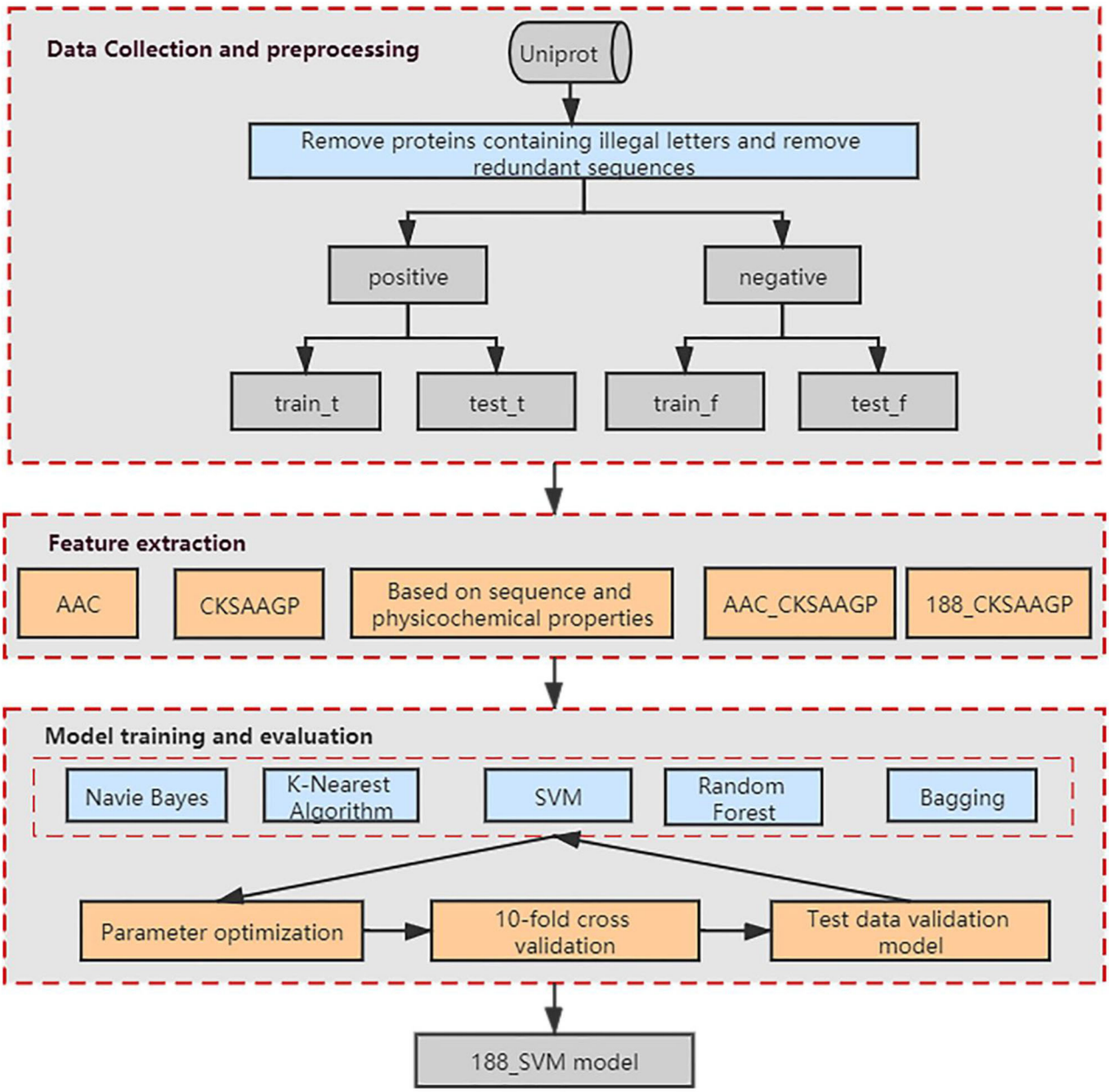


FIGURE 1 | Schematic showing the workflow used to develop our method. AAC, CKSAAGP, 188-dimensions, AAC_CKSAAGP, 188_ CKSAAGP are our five feature extraction methods. Naive Bayes, K-Nearest Algorithm, SVM, Random Forest and Bagging are our five classification algorithms.

When K takes each value, the CKSAAGP feature vector $(\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5)$ has a total size of 150 dimensions.

188-Dimensions

Each amino acid sequence has different physical and chemical properties including amino acid composition, hydrophobicity, normalized Van der Waals volume, polarity, polarizability, charge, surface tension, secondary structure, and solvent accessibility for each residue in the sequence (Cai et al., 2003). The feature extraction method for protein P is formulated as follows:

$$P = \{C, Tr, D\} \tag{4}$$

Where C represents the frequency of a kind of specific attribute (such as polarity) amino acid appearing in the global sequence. Tr represents the global percentage of transitions between a specific amino acid and another amino acid of a specific property. D is used to describe the first, 25%, 50%, 75% and last position of each specific amino acid in the peptide chain.

Assume that the protein sequence "VNVVVNVNNVVVVVNVNNNVVNNVNNNVVNVN" has 17 valines ($n_1 = 17$) and 15 asparagines ($n_2 = 15$). The components of these two amino acids are $\frac{n_1}{n_1+n_2} \times 100.00 = 53.13$ and $\frac{n_2}{n_1+n_2} \times 100.00 = 46.87$, respectively. The number of transitions from V to N or N to V is 17, so the percentage of these transitions is $(17/32) \times 100.00 = 53.13$. The first, 25, 50, 75% and last positions of valine in the peptide chain are 1, 5, 13, 21, 31, respectively, so the D-attribute of valine is D_V = (3.13, 15.63, 40.63, 65.23, 96.88). In the same way, the position of asparagine in the peptide chain can be found. In summary, the amino acid composition descriptors are C = (53.13, 64.87), Tr = (53.13), and D = (3.13, 15.63, 40.63, 65.23, 96.88, 6.25, 28.13, 68.75, 78.13, 100.00). Descriptors of other attributes can be described through a similar process, and then all the descriptors are combined to form a 188-dimensions feature vector.

Sequence	V	N	V	V	V	N	V	N	N	V	V	V	V	V	N	V	N	N	N	V	V	N	N	V	N	N	N	V	V	N	V	N
Sequence index	1				5					10					15					20					25					30		
Index for V	1		2	3	4		5			6	7	8	9	10		11				12	13			14				15	16		17	
Index for N		1				2		3	4						5		6	7	8			9	10		11	12	13			14		15
V/N transitions	|	|			|	|	|		|					|	|			|		|	|		|		|	|	|		

Combined Method

A combination of AAC, CKSAAGP, and feature extraction based on a combination of sequence and physicochemical properties constitute a new feature extraction method. The number of characterization dimensions for the AAC_CKSAAGP combination is 170, and the number of characterization dimensions for the 188_CKSAAGP combination is 338. Since the information of amino acid content is included in the 188-dimensions feature extraction method, the combination of AAC and 188-dimensions feature extraction method is not used in this paper.

Classifier

In this paper, five classifiers were used for the experiments. these classifiers were implemented through Waikato Environment for Knowledge Analysis software (Azuaje et al., 2006).

Random Forest

Random Forest (RF) is an integrated learning method first proposed by Leo Breiman and Adele Cutler (Azuaje et al., 2006; Goldstein et al., 2011; Cheng et al., 2018b,d), and it is a combination of multiple decision trees. Nowadays, many bioinformatics' problems use Random Forest (Tastan et al., 2012; Jamshid et al., 2018; Lyu et al., 2019; Ru et al., 2019; Lv et al., 2020). For processing large amounts of data, Random Forest is characterized by high accuracy, high speed and good robustness. In RF, we need to input the prediction samples into each tree for prediction, and finally use the voting algorithm to determine the result of prediction. The voting algorithm is shown speed and good robustness. The anti-noise capability of RF is strong, and it often shows good robustness when processing high-dimensional data. The voting algorithm is shown below:

$$\text{result} = \text{sgn}\left(\sum_{\text{i}=1}^{\text{n}} \text{pre}_{\text{label}}\right) \tag{5}$$

Where result is the final prediction, pre_{label} represents the predicted result for each decision tree, which equals to 1 or −1, and n represents the number of Decision Trees in the model.

Support Vector Machine

Support Vector Machine (SVM) is often applied to classification problems and is a supervised learning approach (Huang et al., 2012; Jiang et al., 2013; Xing et al., 2014; Kumar et al., 2015; Zhao et al., 2015; Liao et al., 2018; Wang et al., 2019). There are already a number of software packages that support the SVM algorithms. In this experiment, we used Libsvm (Chang and Lin, 2007) in Weka (version 3-8-2) (Hall et al., 2008) to implement the SVM, where we chose RBF, a radial basis function, to classify the proteins of tobacco mosaic virus. Then we determined the regularization parameter C and the kernel parameter g through grid search and 10-fold cross-validation (Wang et al., 2011).

K-Nearest Neighbor

The K-Nearest Neighbor (KNN) algorithm (Zhang and Zhou, 2007; Lan et al., 2013; Deng et al., 2016) which is one of the simplest, most convenient, and highly effective algorithms. Now it has been frequently used in the functional classification of proteins problems. The key step of KNN prediction is to find the K neighbors closest to the test data from the training set, and then use the category with the most K neighbors as the final category of the test data. In this experiment, we adopt the KNN algorithm based on the Harmanton distance and the Harmanton distance formula is summarized as follows:

$$L_{1(x_i,x_j)} = \sum_{k=1}^{n} |x_i^{(k)} - x_j^{(k)}| \tag{6}$$

where: x_i^k(k=1, 2, 3, ……n) is characteristic of the training set and x_j^k(k=1, 2, 3, ……n) is characteristic of the test dataset.

Naive Bayes

The Naive Bayes (NB) is a easily understand classification algorithm (Xue et al., 2006; Wang et al., 2008; Feng et al., 2013), which is based on the Bayesian classifier and assumes that the feature attributes of the data are simple and independent. In the

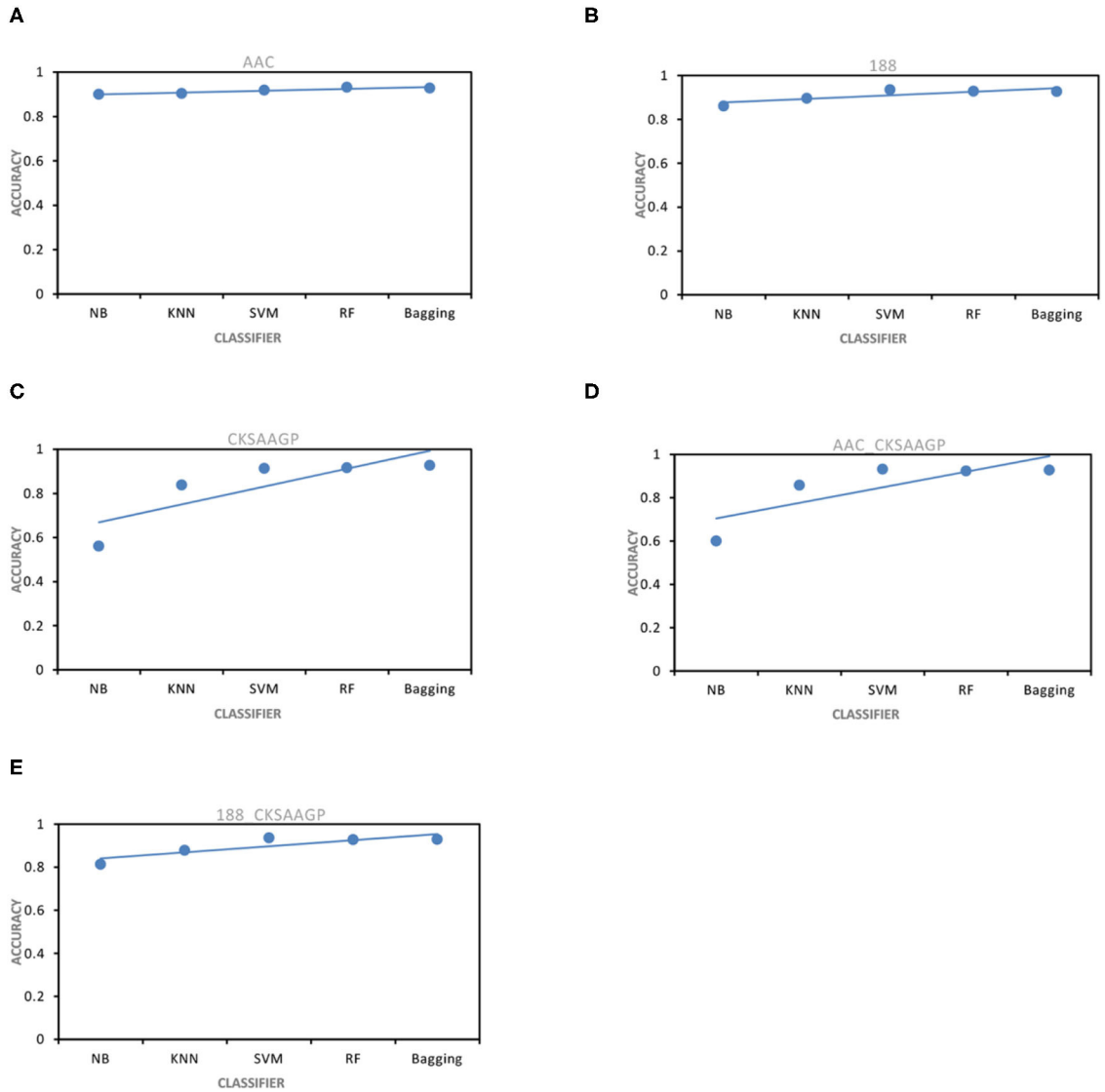


FIGURE 2 | ACC obtained by five feature extraction methods [**(A–E)** AAC, 188-dimensions, CKSAAGP, AAC_CKSAAGP, and 188-CKSAAGP] combined with five classifiers (NB, KNN, SVM, RF, Bagging) through 10-fold cross-validation.

classification scenario, it greatly reduces the complexity of the Bayesian classification algorithm. Suppose the sample data set is:

$$\left(x_1^{(1)}, x_2^{(1)}, \cdots, x_n^{(1)}, s_1\right), \left(x_1^{(2)}, x_2^{(2)}, \cdots, x_n^{(2)}, s_2\right), \cdots \left(x_1^{(m)}, x_2^{(m)}, \cdots, x_n^{(m)}, s_m\right) \quad (7)$$

there are m samples and each sample have n features. The data set has a total of class variable. Generally speaking, the m samples can be classified into s categories, where n features are independent of each other. The category of S is defined as follows:

$$S = \{s_1, s_2, s_3, \cdots, s_m\} \quad (8)$$

Among them, there are M class variables in the set S. Naive Bayes formula is defined as follows:

$$P\left(y_i|x_1, x_2, \cdots, x_n\right) = \frac{P_{y_i} \prod_{j=1}^{n} P\left(x_j|s_i\right)}{\prod_{j=1}^{n} P\left(x_j\right)} \quad (9)$$

Bagging

Bagging is a typical integrated learning algorithm (Abellán et al., 2017), which is directly based on autonomous sampling. For the input sample set $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$, a weak learner algorithm is used to classify each time, and a total of T iterations are made, and finally we will obtain a powerful classifier. Since it samples each training model, it has a strong generalization ability that can significantly reduce the variance of the training model.

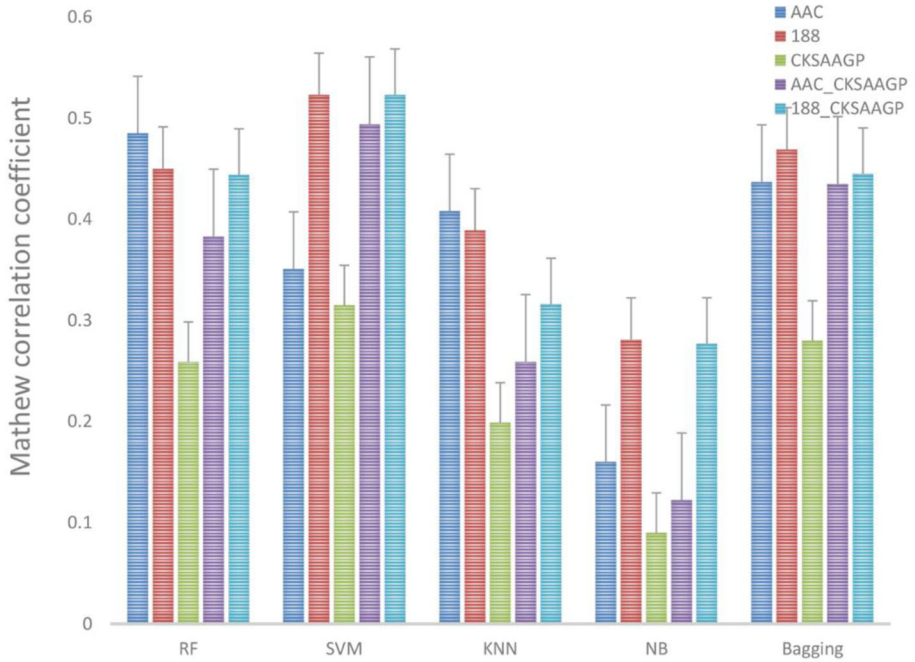


FIGURE 3 | Different feature extraction methods were used to predict the performance of different classification methods. The average MCC value of the classifier was tested by 10-fold cross-validation.

PERFORMANCE EVALUATION

There are five main parameters (Kou and Feng, 2015) to evaluate the predictive performance of this experiment, namely, sensitivity (Sn), specificity (Sp) (Ding et al., 2012; Tan et al., 2019), accuracy (ACC) (Thakur et al., 2016; Cheng et al., 2018a,b), Matthews correlation coefficient (MCC) (Yang et al., 2019a,b) and area of ROC curve (AUC) (Lobo et al., 2008; Li and Fine, 2010; Wang et al., 2010; Hajian-Tilaki, 2014; Baratloo et al., 2015). Defined as follows:

$$S_n = \frac{TP}{TP + FN} \tag{10}$$

$$S_P = \frac{TN}{TN + FP} \tag{11}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$MCC = \frac{(TP + TN) - (FP + FN)}{\sqrt{(TP + FP) \times (TN + FN) \times (TP + FN) \times (TN + FP)}} \tag{13}$$

$$AUC = \frac{\frac{\sum_{i \in positiveClAss} rank_i - M \times (M+1)}{2}}{M \times N} \tag{14}$$

Where TP represents the amount of tobacco mosaic virus correctly predicted by the model (Dong et al., 2015); TN indicates the amount of non-tobacco mosaic virus correctly predicted by the model (Niu et al., 2018); FN indicates the amount of non-tobacco mosaic virus incorrectly predicted by the model (Kim et al., 2016); FP indicates the amount of non-tobacco mosaic virus predicted by the model; M and N indicate the amount of positive and negative data, respectively; and $rank_i$ is the score of the i-th positive sample was calculated by classification. The higher the value of the five evaluation indicators above, the better the model prediction.

RESULTS AND DISCUSSION

Performance Evaluation of Different Classifiers

The ACC and MCC of SVM and RF were mostly higher than the predictors of NB, KNN and Bagging under different feature extraction methods (**Figures 2**, **3**). When the feature extraction method selects 188_CKSAAGP or 188-dimensions, SVM reach the highest ACC. When the feature extraction method uses AAC, RF achieves the highest ACC. Through 10-fold cross-validation, the MCC of SVM is higher than that of RF (**Figure 3**).

However, when comparing predictor superiority, it is possible to use not only the predicted ACC and MCC comparison, but also the trade-off between Sn and Sp. The sensitivity (Sn) and specificity (Sp) of SVM, RF, and Bagging predictor variables are greater than those of NB and KNN (**Table 1** and **Figure 5**). This result shows that SVM, RF, and Bagging predict tobacco mosaic virus are better than NB and KNN due to the difference in the ability of these five common classification algorithms to handle multidimensional datasets. NB is a naive algorithm based on the assumption that the individual properties are independent of each other, and NB

TABLE 1 | Predictive effect of NB, KNN, SVM, RF, and Bagging on different trait extraction methods for tobacco mosaic virus.

Method	Feature	Sp	Sn	ACC (%)	MCC	AUC
NB	AAC	0.866	0.899	89.95	0.16	0.719
KNN		0.902	0.905	90.45	0.408	0.699
SVM (c = 9, g = 3)		0.901	0.918	91.80	0.351	0.61
RF		**0.932**	**0.932**	**93.22**	**0.485**	**0.833**
Bagging		0.922	0.928	92.76	0.437	0.799
NB	188-dimensions	0.882	0.862	86.16	0.281	0.732
KNN		0.899	0.897	89.65	0.389	0.699
SVM		**0.936**	**0.936**	**93.58**	**0.523**	0.657
RF		0.929	0.929	92.91	0.45	**0.838**
Bagging		0.924	0.93	93.04	0.469	0.807
NB	CKSAAGP	0.859	0.561	56.05	0.09	0.64
KNN		0.868	0.839	83.87	0.199	0.618
SVM (c = 5, g = 3)		0.894	0.914	91.38	**0.315**	0.601
RF		0.917	0.916	91.58	0.259	**0.779**
Bagging		**0.919**	**0.917**	**91.69**	0.28	0.711
NB	AAC_CKSAAGP	0.866	0.601	60.07	0.122	0.669
KNN		0.878	0.858	85.82	0.259	0.648
SVM (c = 3, g = 3)		0.923	**0.931**	**93.15**	**0.494**	0.675
RF		**0.927**	0.924	92.36	0.383	**0.831**
Bagging		0.921	0.927	92.75	0.435	0.795
NB	188_ CKSAAGP	0.885	0.814	81.40	0.277	0.726
KNN		0.887	0.878	87.80	0.316	0.669
SVM (c = 3, g = −11)		**0.936**	**0.936**	**93.58**	**0.523**	0.657
RF		0.929	0.929	92.85	0.444	**0.823**
Bagging		0.925	0.93	93.00	0.463	0.803

Sp, specificity; Sn, sensitivity; ACC, accuracy; MCC, Matthews correlation coefficient; AUC, area of ROC curve. AAC, amino acid composition; 188-dimensions, composition based on sequence and physicochemical properties; CKSAAGP, composition of k-spaced amino acid pairs; AAC_CKSAAGP, combination of AAC and CKSAAGP; 188_CKSAAGP, combination of 188-dimensions and CKSAAGP. The bold values represent the highest score of current feature extraction method in different classifiers.

TABLE 2 | Through the use of test data, the evaluation results of the prediction model of NB, KNN, SVM, RF, and Bagging combined with different types of extraction methods.

Method	Feature	Sp	Sn	ACC(%)	MCC	AUC
NB	AAC	0.845	0.89	0.8905	0.051	0.677
KNN		0.896	0.901	0.9011	0.368	0.673
SVM		0.893	0.914	0.9142	0.289	0.508
RF		**0.912**	**0.923**	**0.9226**	**0.372**	**0.781**
Bagging		**0.912**	0.922	0.9222	**0.372**	0.734
NB	188-dimensions	0.877	0.859	0.8588	0.255	0.709
KNN		0.886	0.888	0.8875	0.311	0.653
SVM		**0.925**	**0.928**	**0.9277**	**0.434**	0.615
RF		0.92	0.924	0.9243	0.393	**0.799**
Bagging		0.909	0.922	0.9218	0.377	0.774
NB	CKSAAGP	0.858	0.558	0.5577	0.086	0.634
KNN		0.86	0.832	0.8321	0.15	0.587
SVM		0.886	0.91	0.9099	**0.268**	0.582
RF		**0.922**	**0.915**	0.9146	0.235	**0.736**
Bagging		0.917	**0.915**	**0.9150**	0.243	0.692
NB	AAC_CKSAAGP	0.861	0.589	0.5886	0.104	0.647
KNN		0.87	0.859	0.8592	0.213	0.615
SVM		0.915	**0.926**	**0.9256**	**0.423**	0.63
RF		**0.922**	0.919	0.9188	0.312	**0.777**
Bagging		0.915	0.923	0.9226	0.373	0.726
NB	188_CKSAAGP	0.879	0.811	0.8106	0.244	0.705
KNN		0.877	0.87	0.8702	0.257	0.636
SVM		**0.925**	**0.928**	**0.9277**	**0.434**	0.615
RF		0.92	0.925	0.9247	0.399	**0.776**
Bagging		0.915	0.924	0.9239	0.393	0.75

The bold values represent the highest score of current feature extraction method in different classifiers.

is very friendly to low dimensional features. However, for multidimensional datasets, there is often some correlation between attribute features. The low ACC of KNN may be because the small size of the training datasets. The SVM, RF and Bagging classification algorithms do not require much in terms of dataset dimensionality, and they can handle high-dimensional, noisy and missing datasets with strong correlation between attributes.

In addition, we also used the test datasets to verify the model. The results are shown in **Table 2**. The results show that the model constructed by SVM combined with 188-dimensions or 188_CKSAAGP achieves a high AAC, which shows that this model is reliable. Although the model constructed by SVM combined with 188-dimensions or 188_CKSAAGP is lower than other algorithms in terms of AUC, evaluation indicators such as Sp, Sn, and MCC have all achieved the best results. Therefore, the SVM algorithm is very promising in TMV classification. Due to the above reasons, this experiment chose SVM as the final classifier to predict TMV.

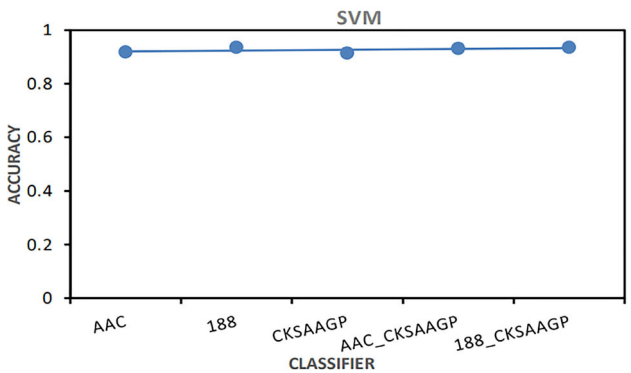


FIGURE 4 | The AAC value obtained by the predictor constructed by SVM combined with five feature selection methods (AAC, 188-dimensions, CKSAAGP, AAC_CKSAAGP, and 188-CKSAAGP).

Performance Evaluation of Different Feature Extraction Methods

Among different feature extraction, Bagging predictor combined with 188-dimensions feature extraction to obtain the best

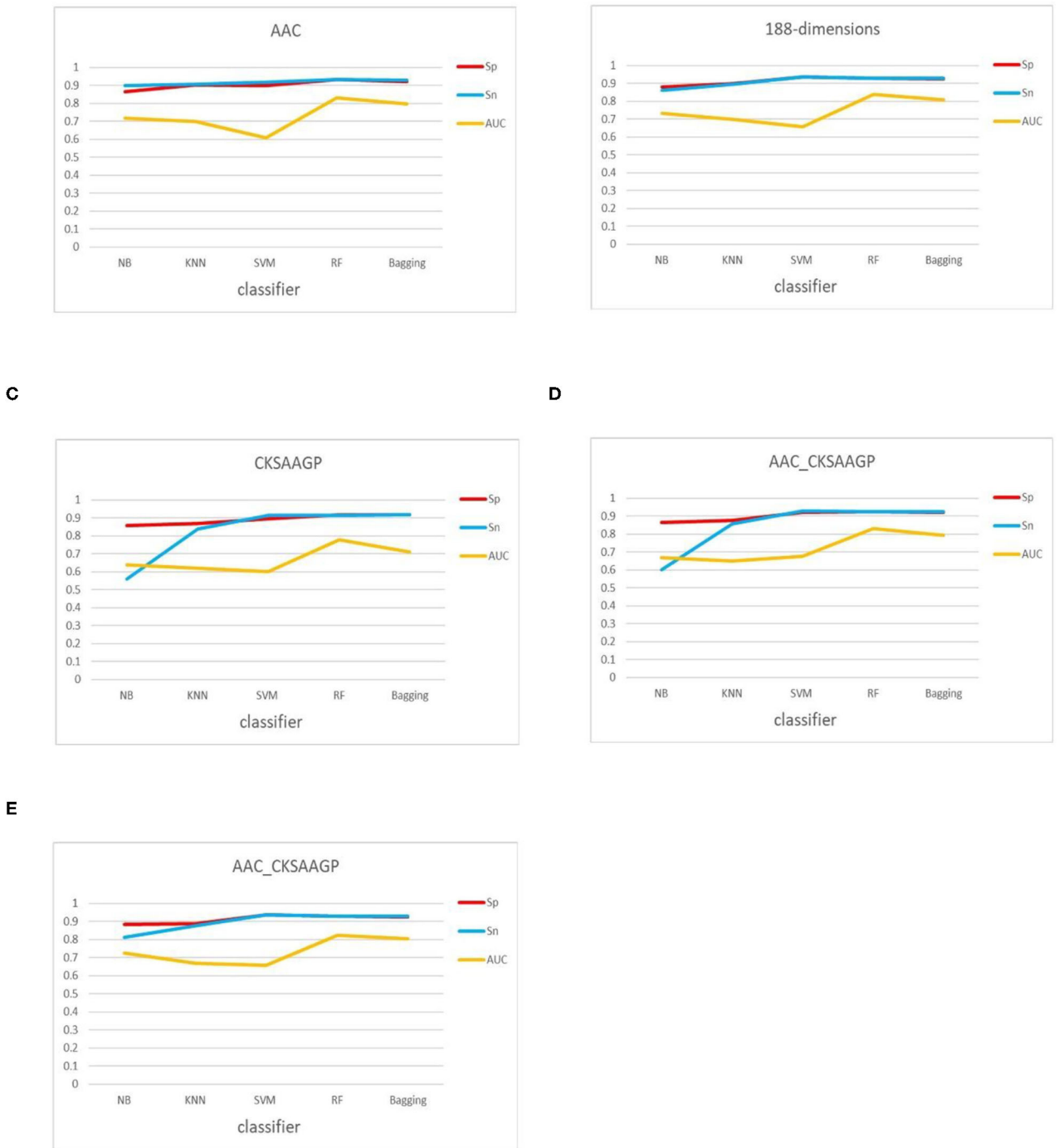


FIGURE 5 | Sn, Sp, and AUC values obtained by 10-fold cross-validation through five feature extraction methods [**(A–E)** AAC, 188-dimensions, CKSAAGP, AAC_CKSAAGP, and 188-CKSAAGP] combined with five classifiers (NB, KNN, SVM, RF, Bagging).

prediction performance. NB, KNN and RF predictors combined with AAC feature extraction to obtain the best prediction performance. The prediction models built by SVM combined with 188-dimensions or 188_CKSAAGP feature extraction have obtained the best prediction results (**Figure 4**).

In addition, the classification effect of the classifier constructed by 188-dimensions combined with different classification

algorithms in terms of Sn, Sp, MCC, AUC is higher than other feature extraction methods, which proves that the prediction model of the former is better than other models (**Figure 5**). In the test datasets, SVM combined with 188-dimensions obtained a prediction accuracy of 92.77%, which proves that the prediction model is reliable. Therefore, in this study, we use 188-dimensions as the final feature extraction method.

CONCLUSION

Rapid and accurate identification of tobacco mosaic virus is the key to successfully protecting tobacco from poison. Kumar and Prakash (2016) used direct antigen coating enzyme linked immunoassay (DAC-ELISA) technique to detect the TMV virus from pepper samples. However, this method is very complicated in sample preparation and detection processes, which is time-consuming and labor-intensive. Our goal was to distinguish between tobacco mosaic virus proteins and healthy tobacco leaf proteins in a large amount of data. The work in this paper provides an effective method to solve this problem.

In this experiment, first, we constructed a high-quality benchmark tobacco mosaic virus protein data set, which ensures the reliability of the classification tool. Secondly, we compared the performance of five feature extraction and five classifier constructs as predictors through 10-fold cross-validation, and then validated each model with the test datasets. The results show that SVM combined with 188-dimensions feature extraction method has the best prediction performance. It has obtained 93.58% accuracy on the train datasets and 92.77% accuracy on the test datasets, which proves that the prediction model has good robustness, so this paper chooses support vector machine as the prediction engine. We hope that these findings will help the development of identification of tobacco mosaic virus.

In future research, because feature selection technology has been successfully applied to some biological information experiments (Dong et al., 2015), feature selection on protein data can improve the prediction effect of the classifier. In addition, we will also try to use machine learning methods to solve analytical problems in genomics (Cheng et al., 2018c), epigenomics (Wang et al., 2017), and other proteomics fields.

DATA AVAILABILITY STATEMENT

Experimental data can be obtained from the corresponding author according to the reasonable request.

AUTHOR CONTRIBUTIONS

Y-MC collected the datasets. X-PZ processed the datasets. Y-MC and X-PZ designed the experiments. X-PZ did and analyzed the experiments' result. Y-MC contributed to the writing of this paper. All authors contributed to the article and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.569100/full#supplementary-material

REFERENCES

Abellán, J., Castellano, J. G., and Mantas, C. J. (2017). A new robust classifier on noise domains: bagging of credal C4.5 trees. *Complexity* 2017:9023970. doi: 10.1155/2017/9023970

Azuaje, F., Witten, I. H., Frank, E. (2006). Data mining: practical machine learning tools and techniques 2nd edition. *BioMed. Eng. Online* 5:51. doi: 10.1186/1475-925X-5-51

Baratloo, A., Hosseini, M., Negida, A., and El Ashal, G. (2015). Part 1: simple definition and calculation of accuracy, sensitivity and specificity. *Emergency* 3, 48–49. doi: 10.1111/j.1945-5100.2007.tb00551.x

Bhasin, M., and Raghava, G. P. S. (2004). Classification of nuclear receptors based on amino acid composition and dipeptide composition. *J. Biol. Chem.* 279, 23262–23266. doi: 10.1074/jbc.M401932200

Cai, C. Z., Han, L. Y., Ji, Z. L., Chen, X., and Chen, Y. Z. (2003). SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.* 31, 3692–3697. doi: 10.1093/nar/gkg600

Chang, C.-C., and Lin, C.-J. (2007). LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2:27. doi: 10.1145/1961189.1961199

Chen, Z., Zhao, P., Li, F., André, L., Marquez-Lago, T. T., Wang, Y., et al. (2018). iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* 34, 2499–2502. doi: 10.1093/bioinformatics/bty140

Chen, Z., Zhao, P., Li, F., Leier, A., Song, J., Marquez-Lago, T. T., et al. (2020). PROSPECT: A web server for predicting protein histidine phosphorylation sites. *J. Bioinform. Comput. Biol.* 18:2050018. doi: 10.1142/S0219720020500183

Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018a). DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. *Bioinformatics* 11, 1953–1956. doi: 10.1093/bioinformatics/bty002

Cheng, L., Jiang, Y., Ju, H., Sun, J., Peng, J., Zhou, M., et al. (2018b). InfAcrOnt: calculating cross-ontology term similarities using information flow by a random walk. *BMC Genomics* 19:919. doi: 10.1186/s12864-017-4338-6

Cheng, L., Qi, C., Zhuang, H., Fu, T., and Zhang, X. (2020). gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucleic Acids Res.* 48, D556–D560. doi: 10.1093/nar/gkaa511

Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., et al. (2018c). LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. *Nucleic Acids Res.* 47, D140–D144. doi: 10.1093/nar/gky1051

Cheng, L., Zhuang, H., Yang, S., Jiang, H., Wang, S., and Zhang, J. (2018d). Exposing the causal effect of c-reactive protein on the risk of type 2 diabetes mellitus: a mendelian randomization study. *Front. Genet.* 9:657. doi: 10.3389/fgene.2019.00085

Chou, K.-C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* 43, 246–255. doi: 10.1002/prot.1035

Deng, Z., Zhu, X., Cheng, D., Zong, M., and Zhang, S. (2016). Efficient kNN classification algorithm for big data. *Neurocomputing* 195, 143–148. doi: 10.1016/j.neucom.2015.08.112

Ding, C., Yuan, L. F., Guo, S. H., Lin, H., and Chen, W. (2012). Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions. *J. Proteomics* 77, 321–328. doi: 10.1016/j.jprot.2012.09.006

Dong, Q., Shanyi, W., Kai, W., Xuan, L., and Liu, B. (2015). "Identification of DNA-binding proteins by auto-cross covariance transformation," in *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (Washington, DC), 470–5. doi: 10.1109/BIBM.2015.7359730

Dubchak, I., Muchnik, I., Holbrook, S. R., and Kim, S. H. (1995). Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl. Acad. Sci. U.S.A.* 92:8700. doi: 10.1073/pnas.92.19.8700

Feng, P.-M., Ding, H., Chen, W., and Lin, H. (2013). Naïve bayes classifier with feature selection to identify phage virion proteins. *Comput. Math. Methods Med.* 2013:530696. doi: 10.1155/2013/530696

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Goldstein, B. A., Polley, E. C., and Briggs, F. B. S. (2011). Random forests for genetic association studies. *Stat. Appl. Genet. Mol. Biol.* 10, 32–32. doi: 10.2202/1544-6115.1691

Hajian-Tilaki, K. (2014). Sample size estimation in diagnostic test studies of biomedical informatics. *J. Biomed. Inform.* 48, 193–204. doi: 10.1016/j.jbi.2014.02.013

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. (2008). The WEKA data mining software: an update. *SIGKDD Explor. Newsl* 11, 10–18. doi: 10.1145/1656274.1656278

Hu, T.-W., and Lee, A. H. (2015). Commentary: Tobacco control and tobacco farming in African countries. *J. Public Health Policy* 36, 41–51. doi: 10.1057/jphp.2014.47

Huang, Y., Liu, N., Wang, J. P., Wang, Y. Q., Yu, X. L., Wang, Z. B., et al. (2012). Regulatory long non-coding RNA and its functions. *Biochemistry* 68, 611–618. doi: 10.1007/s13105-012-0166-y

Jamshid, P., Ali Reza, K., and Maryam, J. (2018). GENIRF: an algorithm for gene regulatory network inference using rotation forest. *Curr. Bioinform.* 13, 407–419. doi: 10.2174/1574893612666170731120830

Jiang, Q., Wang, G., Jin, S., Li, Y., and Wang, Y. J. (2013). Predicting human microRNA-disease associations based on support vector machine. *Int. J. Data Min. Bioinform.* 8, 282–293. doi: 10.1504/IJDMB.2013.056078

Kim, H.-J., Jo, N.-O., and Shin, K.-S. (2016). Optimization of cluster-based evolutionary undersampling for the artificial neural networks in corporate bankruptcy prediction. *Expert Syst. Appl.* 59, 226–234. doi: 10.1016/j.eswa.2016.04.027

Kou, G., and Feng, Y. J. (2015). Identify five kinds of simple super-secondary structures with quadratic discriminant algorithm based on the chemical shifts. *J. Theor. Biol.* 380, 392–398. doi: 10.1016/j.jtbi.2015.06.006

Kumar, M., Gromiha, M. M., and Raghava, G. P. S. (2015). SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. *J. Mol. Recognit.* 24, 303–313. doi: 10.1002/jmr.1061

Kumar, S., and Prakash, H. S. (2016). Detection of tobacco mosaic virus and tomato mosaic virus in pepper seeds by enzyme linked immunosorbent assay (ELISA). *Arch. Phytopathol. Plant Protect.* 49, 59–63. doi: 10.1080/03235408.2012.658991

Lan, L., Djuric, N., Guo, Y., and Vucetic, S. (2013). MS-kNN: protein function prediction by integrating multiple data sources. *BMC Bioinform.* 14(Suppl. 3):S8. doi: 10.1186/1471-2105-14-S3-S8

Li, J., and Fine, J. P. (2010). Weighted area under the receiver operating characteristic curve and its application to gene selection. *J. Royal Stat. Soc.* 59, 673–692. doi: 10.1111/j.1467-9876.2010.00713.x

Liao, Z., Li, D., Wang, X., Li, L., and Zhou, Q. (2018). Cancer diagnosis through isomiR expression with machine learning method. *Curr. Bioinform.* 13, 57–63. doi: 10.2174/1574893611666160609081155

Lin, H., Han, L. Y., Cai, C. Z., Ji, Z., and Chen, Y. Z. (2005). Prediction of transporter family from protein sequence by support vector machine approach. *Proteins* 62, 218–231. doi: 10.1002/prot.20605

Lobo, J. M., Jiménez-Valverde, A., and Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. *Glob. Ecol. Biogeogr.* 17, 145–151. doi: 10.1111/j.1466-8238.2007.00358.x

Lv, H., Dao, F.-Y., Zhang, D., Guan, Z.-X., Yang, H., Su, W., et al. (2020). iDNA-MS: an integrated computational tool for detecting DNA modification sites in multiple genomes. *iScience* 23:100991. doi: 10.1016/j.isci.2020.100991

Lyu, Z., Jin, S., Ding, H., and Zou, Q. (2019). A random forest sub-golgi protein classifier optimized via dipeptide and amino acid composition features. *Front. Bioeng. Biotechnol.* 7:215. doi: 10.3389/fbioe.2019.00215

Manavalan, B., Shin, T. H., and Lee, G. (2018). PVP-SVM: sequence-based prediction of phage virion proteins using a support vector machine. *Front Microbiol.* 9:476. doi: 10.3389/fmicb.2018.00476

Metzler, S., and Kalinina, O. V. (2014). Detection of atypical genes in virus families using a one-class SVM. *BMC Genomics* 15:913. doi: 10.1186/1471-2164-15-913

Niu, M., Li, Y., Wang, C., and Ke, H. (2018). RFAmyloid: a web server for predicting amyloid proteins. *Int. J. Mol. Sci.* 19:2071. doi: 10.3390/ijms19072071

Ru, X., Li, L., and Zou, Q. (2019). Incorporating distance-based top-n-gram and random forest to identify electron transport proteins. *J. Proteome. Res.* 18, 2931–2939. doi: 10.1021/acs.jproteome.9b00250

Salama, M. A., Hassanien, A. E., and Mostafa, A. (2016). The prediction of virus mutation using neural networks and rough set techniques. *EURASIP J. Bioinform. Syst. Biol.* 2016:10. doi: 10.1186/s13637-016-0042-0

Tan, J. X., Li, S. H., Zhang, Z. M., Chen, C. X., Chen, W., Tang, H., et al. (2019). Identification of hormone binding proteins based on machine learning methods. *Math. Biosci. Eng.* 16, 2466–2480. doi: 10.3934/mbe.2019123

Tastan, O., Qi, Y., Carbonell, J. G., and Kleinseetharaman, J. (2012). Prediction of interactions between HIV-1 and human proteins by information integration. *Pac. Symp. Biocomput.* 2009, 516–527. doi: 10.1142/9789812836939_0049

Thakur, A., Rajput, A., and Kumar, M. (2016). MSLVP: prediction of multiple subcellular localization of viral proteins using a support vector machine. *Mol. Syst.* 12, 2572–2586 doi: 10.1039/C6MB00241B

The Uniprot, C. (2018). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515. doi: 10.1093/nar/gky1049

Wang, G., Luo, X., Wang, J., Wan, J., Xia, S., Zhu, H., et al. (2017). MeDReaders: a database for transcription factors that bind to methylated DNA. *Nucleic Acids Res.* 46, D146–D151. doi: 10.1093/nar/gkx1096

Wang, G., Wang, Y., Feng, W., Wang, X., Yang, J. Y., Zhao, Y., et al. (2008). Transcription factor and microRNA regulation in androgen-dependent and -independent prostate cancer cells. *BMC Genomics* 9:S22. doi: 10.1186/1471-2164-9-S2-S22

Wang, G., Wang, Y., Teng, M., Zhang, D., Li, L., and Liu, Y. (2010). Signal transducers and activators of transcription-1 (STAT1) regulates microRNA transcription in interferon γ-stimulated HeLa cells. *PLoS ONE* 5:e11794. doi: 10.1371/journal.pone.0011794

Wang, Y., Shi, F., Cao, L., Dey, N., Wu, Q., Ashour, A. S., et al. (2019). Morphological segmentation analysis and texture-based support vector machines classification on mice liver fibrosis microscopic images. *Curr. Bioinform.* 14, 282–294. doi: 10.2174/1574893614666190304125221

Wang, Y., Wang, Y., Yang, Z., and Deng, N. (2011). Support vector machine prediction of enzyme function with conjoint triad feature and hierarchical *context*. *BMC Syst. Biol.* 5 (Suppl. 1):S6. doi: 10.1186/1752-0509-5-S1-S6

Xing, Y.-Q., Liu, G.-Q., Zhao, X.-J., Zhao, H.-Y., and Cai, L. (2014). Genome-wide characterization and prediction of Arabidopsis thaliana replication origins. *Biosystems* 124, 1–6. doi: 10.1016/j.biosystems.2014.07.001

Xue, Y., Chen, H., Jin, C., Sun, Z., and Yao, X. (2006). NBA-Palm: prediction of palmitoylation site implemented in Naïve Bayes algorithm. *BMC Bioinform.* 7:458. doi: 10.1186/1471-2105-7-458

Yang, H., Yang, W., Dao, F.-Y., Lv, H., Ding, H., Chen, W., et al. (2019a). A comparison and assessment of computational method for identifying recombination hotspots in Saccharomyces cerevisiae. *Brief. Bioinform.* 21, 1568–1580. doi: 10.1093/bib/bbz123

Yang, W., Zhu, X.-J., Huang, J., Ding, H., and Lin, H. (2019b). A brief survey of machine learning methods in protein sub-golgi localization. *Curr. Bioinform.* 14, 234–240. doi: 10.2174/1574893613666181113131415

Zhang, M., and Zhou, Z. (2007). ML-KNN: A lazy learning approach to multi-label learning. *Pattern. Recogn.* 40, 2038–2048. doi: 10.1016/j.patcog.2006.12.019

Zhang, S., Hao, L., and Zhang, T. (2014). Prediction of protein–protein interaction with pairwise kernel support vector machine. *Int. J. Mol. Sci.* 15, 3220–3233 doi: 10.3390/ijms15023220

Zhao, Y., Wang, F., and Juan, L. (2015). MicroRNA promoter identification in arabidopsis using multiple histone markers. *BioMed. Res. Int.* 2015:861402. doi: 10.1155/2015/861402

Zhong, W., Zhong, B., Zhang, H., Chen, Z., Chen, Y. (2020). Identification of anti-cancer peptides based on multi-classifier system. *Comb. Chem. High Throughput Screen.* 22, 694–704 doi: 10.2174/1386207322666191203141102

Zhu, Y., Jia, C., Li, F., and Song, J. (2020). Inspector: a lysine succinylation predictor based on edited nearest-neighbor undersampling and adaptive synthetic oversampling. *Anal. Biochem.* 593:113592. doi: 10.1016/j.ab.2020.113592

Zou, Q., Lin, G., Jiang, X., Liu, X., and Zeng, X. (2018). *Sequence clustering in* bioinformatics: an empirical study. *Brief Bioinform.* 21, 1–10. doi: 10.1093/bib/bby090

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.

Copyright © 2020 Chen, Zu and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 19 October 2020
doi: 10.3389/fgene.2020.592596

Integrating Sugar Metabolism With Transport: Elevation of Endogenous Cell Wall Invertase Activity Up-Regulates *SlHT2* and *SlSWEET12c* Expression for Early Fruit Development in Tomato

*Lei Ru[1,2], Yong He[1], Zhujun Zhu[1], John W. Patrick[2] and Yong-Ling Ruan[2]**

[1]*The Key Laboratory for Quality Improvement of Agricultural Products of Zhejiang Province, College of Agriculture and Food Science, Zhejiang Agriculture and Forestry University, Hangzhou, China,* [2]*School of Environmental and Life Sciences, Australia-China Research Centre for Crop Improvement, The University of Newcastle, Callaghan, NSW, Australia*

OPEN ACCESS

Edited by:
Wei Chen,
North China University of Science and Technology, China

Reviewed by:
Chuankui Song,
Anhui Agricultural University, China
Asif Ameen,
Gomal University, Pakistan

***Correspondence:**
Yong-Ling Ruan
yong-ling.ruan@newcastle.edu.au

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

Received: *07 August 2020*
Accepted: *22 September 2020*
Published: *19 October 2020*

Citation:
Ru L, He Y, Zhu Z, Patrick JW and Ruan Y-L (2020) Integrating Sugar Metabolism With Transport: Elevation of Endogenous Cell Wall Invertase Activity Up-Regulates SlHT2 and SlSWEET12c Expression for Early Fruit Development in Tomato. Front. Genet. 11:592596. doi: 10.3389/fgene.2020.592596

Early fruit development is critical for determining crop yield. Cell wall invertase (CWIN) and sugar transporters both play important roles in carbon allocation and plant development. However, there is little information about the relationship between CWIN and those functionally related sugar transporters during fruit development. By using transgenic tomato with an elevated CWIN activity, we investigated how an increase in CWIN activity may regulate the expression of sugar transporter genes during fruit development. Our analyses indicate that CWIN activity may be under tight regulation by multiple regulators, including two invertase inhibitors (INVINHs) and one defective CWIN (deCWIN) in tomato ovaries prior to anthesis. Among the sugar transporters, expression of *SlSWEET12c* for sucrose efflux and *SlHT2* for hexose uptake was enhanced by the elevated CWIN activity at 10 and 15 days after anthesis of tomato fruit development, respectively. The findings show that some specific sugars will eventually be exported transporters (SWEETs) and hexose transporters (HTs) respond to elevate CWIN activity probably to promote rapid fruit expansion when sucrose efflux from phloem and hexose uptake by parenchyma cell are in high demand. The analyses provide new leads for improving crop yield by manipulating CWIN-responsive sugar transporters, together with CWIN itself, to enhance fruit development and sugar accumulation.

Keywords: cell wall invertase, tomato fruit, SWEET, hexose transporter, sucrose transporter, sugar metabolism

INTRODUCTION

Tomato is one of the most important fleshy fruit crops globally (Schwarz et al., 2014). Following pollination and fertilization, tomato fruit undergoes rapid cell division and expansion prior to maturation. Early development of tomato fruit is a critical period for determining fruit traits, such as fruit size and sugar accumulation (Mounet et al., 2009), hence the final fruit yield and quality (Ruan et al., 2012; Azzi et al., 2015).

Ovary and young fruits are typical carbon sinks, which require sugars transported from source leaves. In these sinks, sucrose can be unloaded into sink cells through either symplasmic or apoplasmic unloading or both simultaneously (Patrick, 1997; Patrick et al., 2001). Phloem unloading is often temporally and spatially dynamic to match developmental and physiological demands. For example, in tomato fruit, unloading shifts from an apoplasmic route in the ovary wall to a symplasmic pathway in the pericarp of young fruit (Jin et al., 2009; Palmer et al., 2015) before switching to an apoplasmic route in the pericarp at the onset of the rapid phase of hexose accumulation (Ruan and Patrick, 1995). In contrast, unloading appears to occur apoplasmically from fruit placenta to ovules or developing seeds (Jin et al., 2009; Palmer et al., 2015). Symplasmic unloading requires intercellular interconnection by functional plasmodesmata (Lalonde et al., 2003), while apoplasmic unloading typically involves the expression of plasma-membrane sugar transporters and cell wall invertase (CWIN) hydrolyzing sucrose into glucose and fructose (Slewinski, 2011; Ludewig and Flugge, 2013; Li et al., 2017). The sugar transporters include sucrose transporters (SUTs), hexose transporters (HTs), as well as sugars will eventually be exported transporters (SWEETs).

Cell wall invertase (CWIN) and sugar transporters have been studied extensively but separately. SlLin5 is the major CWIN expressed during tomato fruit development (Fridman and Zamir, 2003; Ruan, 2014; Palmer et al., 2015). SlLin5 has also been shown to locate in QTL sites that determine fruit weight (Fridman et al., 2004) and sugar content (Tieman et al., 2017). Silencing *SlLin5* resulted in fruit abortion (Zanor et al., 2009), while increasing CWIN activity sustained fruit set under heat stress by suppressing programmed cell death (Liu et al., 2016). Identification and expression of SUTs and HTs have been investigated in tomato (e.g., Reuscher et al., 2014). SUTs, the smallest family among the three sugar transporters, have been extensively reviewed (e.g., Kuhn and Grof, 2010; Ayre, 2011; Reinders et al., 2012). In tomato, there are three SUTs, namely, SlSUT1, SlSUT2, and SlSUT4 (Hackel et al., 2006). For the HT family, three SlHTs (SlHT1–3) have been experimentally analyzed, which revealed that their preferred substrate is hexose (McCurdy et al., 2010). Other newly identified HTs are yet to be experimentally characterized (Reuscher et al., 2014). The tomato SWEET family has been examined on a genome-wide scale, including a total of 31 putative SlSWEET genes in the tomato genome (Feng et al., 2015; Shammai et al., 2018; Ho et al., 2019). The SWEETs may facilitate bidirectional transport of glucose and fructose (clades I and II), sucrose (clade III) across plasma membranes, and of sugars across tonoplasts of vacuoles (clade VI) (Chen et al., 2015a).

Both CWIN and sugar transporters can control sucrose allocation (Ludewig and Flugge, 2013; Schroeder et al., 2013). It is widely accepted that co-expression of CWIN and HTs exists to support the development of many sink organs, such as developing barley grains (Weschke et al., 2003), tomato (Shen et al., 2019), and apple fruits (Wang et al., 2020). Recently, heterologous expression of *MdHT2* increased CWIN activity (Wang et al., 2020). Conversely, Liao et al. (2020) reported that silencing AtCWIN2 and AtCWIN4 genes reduced the expression level of many clade II SWEETs for hexoses and H^+-coupled HTs during ovule formation in *Arabidopsis*. Similarly, Sosso et al. (2015) found that ZmSWEET4c, transferring CWIN-mediated hexoses for seed filling in maize, was downregulated in a CWIN mutant, while reciprocally the expression of CWIN was reduced in a SWEET mutant (Sosso et al., 2015).

Despite the progress, as outlined above, on understanding the expression relationship between CWIN and sugar transporter genes in wild type (WT) plants and mutants or transgenic lines, where one of the two players was suppressed or knocked out, there has been no report as how sugar transporters may respond to elevated endogenous CWIN activity during tomato fruit development (Ru et al., 2017). Answering this question is important as it will not only fill the knowledge gap in understanding sugar metabolism and transport but could also provide new avenues to improve crop yield directly (Ruan, 2014). Here, we address this question by using transgenic tomato in which an inhibitor gene against CWIN was silenced leading to elevated CWIN activity (Jin et al., 2009).

MATERIALS AND METHODS

Plant Materials

Transgenic plants (*SlINVINH1*-RNAi), in which CWIN activity is elevated by silencing its inhibitor gene, *SlINVINH1*, were used (Jin et al., 2009). Tomato plants (*Solanum lycopersicum* XF-2) were raised in a glasshouse under natural light and exposed to day/night temperatures of 25/18°C for 14 and 10 h, respectively. Plants were grown in 25 cm diameter (8 L) pots filled with potting mix (one part coarse sand, one part perlite, and one part coir-peat), with one plant per pot. Standard Osmocote™/Osmocote high (K) potassium™ 1:1 slow release fertilizer (Scotts) was applied at a rate of 20 g per pot, supplemented with a weekly liquid fertilizer regime of Jurox Wuxal Liquid Foliar Nutrient Fertilizer™ at a diluted concentration of 4 $ml.L^{-1}$. Pots were maintained at field capacity by being irrigated twice a day, each of 3-min duration, by an automated drip irrigation system.

Fruit age was determined by tagging flowers on the day of anthesis. Hand-pollination was applied on the day of anthesis to ensure every flower was pollinated. Ovaries at 2-days before anthesis (2 dba) and fruits at 2, 5, 10, and 15 days after anthesis (daa) were harvested as reproductive tissues. This sampling regime captured important tomato fruit development, such as fruit set (2 dba ovaries to 2 daa fruitlets), cell division (2 daa fruitlets to 10 daa fruits), and onset of cell expansion (15 daa fruits). Sink and source leaves were collected from expanding and expanded leaves of the same grown plants, while roots and shoots were harvested from 2-week old young seedlings. Samples were stored at −80°C immediately after harvest.

RNA Extraction and cDNA Synthesis

RNA was extracted with RNeasy Plant Mini kit (QIAGEN) with 20–50 mg fresh weight tissue depending on the samples used. Thirty microliter of RNase-free H_2O was added to dissolve

RNA in the final step. RNA quality and concentration were measured with NanoDrop1000™ spectrophotometer (Thermo Scientific). For all samples, OD260/280 was between 1.8 and 2.4, which is an indicator for good RNA quality. One microgram of total RNA was treated with RQ1 RNase-Free DNase (Promega) to degrade residual genomic DNA. The DNase-treated RNA samples were then converted into cDNA by using oligo$(dT)_{20}$ primer and SuperScript™III reverse transcriptase (Invitrogen life technology), according to the manufacturer's instructions. The cDNA was diluted 1:20 for semi-quantitative real-time PCR (semi-qRT-PCR) and quantitative real-time PCR (qRT-PCR).

Semi-qRT-PCR

Semi-qRT-PCR was performed with Taq DNA polymerase (New England Biolabs). A PCR reaction (10 μl) comprised 10 × ThermoPol reaction buffer (1 μl), 10 mM dNTP (0.2 μl), 10 mM forward/reverse primers (0.25 μl each), Taq DNA polymerase (0.05 μl), diluted cDNA (1 μl), and autoclaved water (7.5 μl). PCR program was as follows: initial denaturation at 95°C for 3 min, followed by 25 cycles of denaturation at 95°C for 30 s, annealing at 60°C for 30 s, extension at 68°C for 30 s, and a final extension at 68°C for 5 min, and products were held at 4–10°C. After that, 2 μl loading dye (Thermo Scientific) was added into PCR products, and then 10 μl products with loading dye were electrophoresed on 1.5% agarose gel with 1:1,000 GelRed (Biotium). Bands of cDNAs were visualized in GelDoc XR Imaging System (BIORAD), and band intensity was measured with Image J 1.45 s. The cDNA templates were adjusted to make sure that the band intensity of PCR products from different tissues were equalized before amplifying *SlEF1a*. After the equalization, the same amount of cDNA of different tissues was applied to amplify all the genes of interest. The gene specific primers are listed in **Supplementary Table S1**.

Selection of Reference Genes for qRT-PCR

There are no universal reference genes for all experimental conditions, thus screening stable reference genes is necessary for different tissues and stages selected in our study. Eight tomato tissues from different organs and fruits at different developmental stages were investigated including four vegetative tissues of roots and shoots from 2-week old seedlings, sink and source leaves, as well as four reproductive tissues of ovaries at 2 dba and fruits at 2, 10, and 15 daa. Seven candidate reference genes were assessed in this study (**Supplementary Table S2**).

Quantitative Real-Time-PCR

Quantitative real-time-PCR was conducted according to Ru et al. (2017). Briefly, the same cDNAs for semi-qRT-PCR were used in qRT-PCR analyses. qRT-PCR was performed using Platinum Taq DNA polymerase and SYBR Green fluorescent dye (Invitrogen) on Rotor-gene Q (QIAGEN). PCR program was as follows: initial denaturation at 95°C for 10 min, followed by 45 cycles of denaturation at 95°C for 20 s, annealing at 60°C for 20 s, and extension at 72°C for 20 s. A melt curve was performed in each run to verify the amplification of a specific product with 0.5°C increment from 60 to 90°C. Relative expression was calculated using the Pfaffl method (Pfaffl, 2001). See **Supplementary Table S3** for the gene specific primers used for qRT-PCR.

RESULTS

Multiple Potential CWIN Regulators Expressed at Ovary Stages, Including Two INVINHs and One Defective CWIN

Cell wall invertase activity may be regulated by invertase inhibitors (INVINHs) and defective CWIN (deCWIN) together at the posttranslational level (Palmer et al., 2015). However, how CWIN activity is tightly regulated during early tomato fruit development is still unclear. To address this issue, the expression of four known CWINs (Fridman and Zamir, 2003), INVINHs (Qin et al., 2016), and deCWIN (Palmer et al., 2015) were investigated using RT-PCR. Among the four CWINs, the expression of *SlCWIN1*, previously named *SlLin5*, was only detected in reproductive tissues, not in vegetative tissues. Moreover, *SlCWIN1* is the main CWIN expressed during tomato fruit development, compared to SlCWIN2–SlCWIN4. Expression level of *SlCWIN1* decreased sharply in 2 daa fruitlets as compared to 2 dba ovaries and then was maintained at a relative stable level from 2 to 10 daa fruits (**Figure 1A**). The expression of a previously characterized CWIN inhibitor, *SlINVINH1* (Jin et al., 2009), is constitutively expressed in both vegetative and reproductive tissues (**Figure 1A**). However, *SlINVINH2*, a new CWIN inhibitor which clustered closely together with functional CWIN INVINH in tomato and NtCIF in tobacco (**Figure 1B**), was only expressed in the 2 dba ovary stages but not in vegetative tissues and other fruit stages (**Figure 1A**). *SldeCWIN1* is another potential regulator for regulating CWIN activity at post-translational level (Palmer et al., 2015) with its expression was only detected in 2 dba ovaries and sink leaves (**Figure 1A**). The findings indicate that CWIN activity could be regulated by two inhibitors, SlINVINH1 and SlINVINH2, and one defective invertase, SldeCWIN1, at the posttranslational level in tomato ovaries prior to fruit development.

Elevated CWIN Activity Does Not Affect the Expression of SlSUTs

Firstly, Genorm was applied to obtain stable reference genes for qRT-PCR analyses (Vandesompele et al., 2002). The data were used to rank the reference genes expression in different tissues in Genorm (**Supplementary Table S4**). *SlCAC* and *SlSAND* were identified as the most stable reference genes for the qRT-PCR experiment (**Supplementary Table S5**).

Three sucrose transporter genes (*SlSUT1*, *SlSUT2*, and *SlSUT4*), the only three members in SUT family in tomato (Reuscher et al., 2014), were analyzed. Among them, *SlSUT1* was dominantly expressed in source leaves but extremely low in reproductive fruit (**Figure 2A**), while *SlSUT2* and *SlSUT4* were constitutively expressed in both vegetative and different fruit developmental stages (**Figures 2B,C**). Thus, only the expression of SlSUT2 and SlSUT4 was further investigated in response to elevated

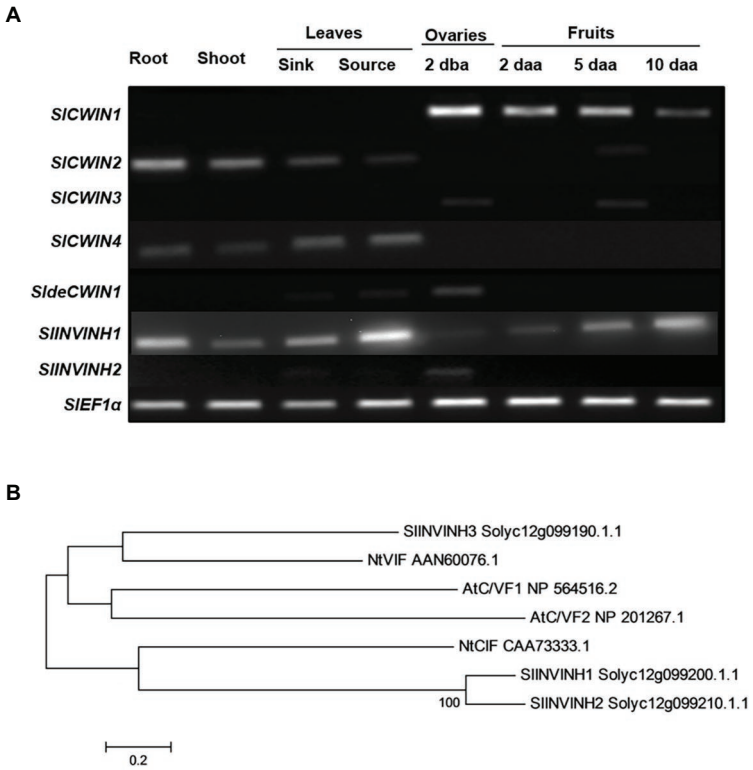


FIGURE 1 | Cell wall invertase (CWIN) and its potential regulators. **(A)** Expression profile of CWIN (*SlCWIN1–SlCWIN4*), a defective CWIN (deCWIN; *SldeCWIN1*), CWIN inhibitors (*SlINVINH1* and *SlINVINH2*) in vegetative and reproductive tissues. Vegetative tissues included roots and shoots from 2-week old seedlings, sink, and source leaves from similar raised plants. Reproductive tissues included 2 days before anthesis (2 dba) ovaries and 2, 5, and 10 days after anthesis (daa) fruits. SlEF1α was used as reference gene. Original gel images in **(A)** are shown in **Supplementary Figure S1**. **(B)** Phylogenetic tree of invertase inhibitors. The tree was generated with MEGA 5.10 using the neighbor-joining method, following multiple sequences alignments with ClustalW. Species abbreviations: At, *Arabidopsis thaliana*; Nt, *Nicotiana tabacum*; and Sl, *Solanum lycopersicum*.

CWIN activity in relation to fruit development. The analyses revealed no significant changes of *SlSUT2* and *SlSUT4* mRNA levels in response to elevated CWIN activity during tomato fruit development (**Figures 2D,E**).

Elevated CWIN Activity Enhanced the Expression of *SlHT2* and *SlSWEET12c* in Early Tomato Fruit Development

SlHT1–3 are the three main contributors for hexose accumulation in tomato fruit (McCurdy et al., 2010). While *SlHT1* was expressed in vegetative tissues, its transcript was only weakly expressed in 2 dba ovaries and developing fruits from 2 to 15 daa (**Figure 3A**). Expression of *SlHT2* was high in vegetative tissues and 2 dba ovaries, as well as 2, 10, and 15 daa fruits (**Figure 3B**). Transcript level of *SlHT3* was relatively higher in ovaries and fruits than in vegetative tissues (**Figure 3C**). The expression level of *SlHT2* significantly increased only in 15 daa fruit of *SlINVINH1*-RNAi (CWIN activity-elevated) when compared to WT plants (**Figure 3D**), with no response observed for *SlHT3* across all stages examined (**Figure 3E**).

SWEET transporter family has been previously examined in tomato. There are a total of 29 *SlSWEET* genes as reported by Feng et al. (2015), and 31 *SlSWEET* genes as reported by Ho et al. (2019). SWEET transporters contain seven transmembranne domains. However, there are two SlSWEET genes containing only six and five TM domains among the 31 SWEETs, respectively (**Supplementary Table S6**), which are deemed to be non-functional. Hence, on these grounds, two *SlSWWET* genes, *SlSWEET17* and *SlSWEET18*, were excluded from our analyses. Phylogenetic tree of 30 *SlSWEET* genes were clustered with 17 *Arabidopsis SWEET* genes to find closet orthologs in tomato (**Supplementary Figure S2**). The *SWEET* genes were divided into four clades. Members in clade I, encode transporters such as AtSWEET1a and AtSWEET1b that have been shown to facilitate glucose transport. Members from clade III, such as AtSWEET11 and AtSWEET12 facilitate sucrose efflux (Chen et al., 2012). Therefore, four putative glucose-transporting SWEETs belong to clade I (*SlSWEET1a*, *1b*, *1c*, and *1d*) and three sucrose-transporting SWEETs from clade III (*SlSWEET12a*, *12b*, and *12c*) were selected for further expression analyses (**Supplementary Figure S2**). Given *SlSWEET1c*, *SlSWEEET1d*, and *SlSWEEET12a* were not expressed during reproductive stages based on semi-RT-PCR analyses (**Supplementary Figure S3**), only expression of *SlSWEET1a*, *1b*, *12b*, and *12c* was further analyzed during tomato fruit development.

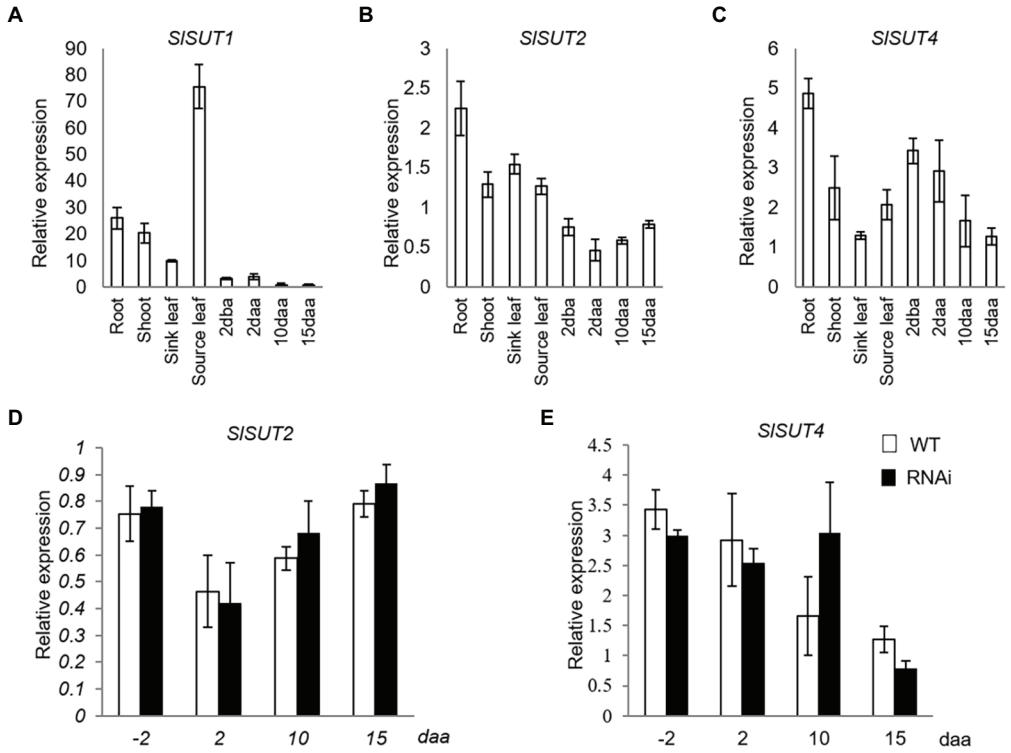


FIGURE 2 | Quantitative real-time-PCR (qRT-PCR) analysis of the expression profiles of sucrose transporter (SUT) genes. **(A–C)** are expression profiles of SUTs in different tissues of tomato plants. **(A)** *SlSUT1*. **(B)** *SlSUT2*. **(C)** *SlSUT4*; **(D)** and **(E)** are expression profiles of SUTs in response to elevated CWIN activity in *SlINVINH1*-RNAi as compared to WT plants **(D)** *SlSUT2*. **(E)** *SlSUT4*. Tissues examined were vegetative and reproductive tissues. Vegetative tissues included root and shoots from 2-week old seedlings, sink, and source leaves from the similar-raised plants. Reproductive tissues included 2 dba ovaries and 2, 10, and 15 daa fruits. *SlCAC* and *SlSAND* were used as reference genes. Each value is the mean ± SE of three biological replicates.

Until now, only SlSWEET1a has been experimentally researched among the *SlSWEET* genes in tomato. Here, SlSWEET1a has been shown to be expressed in young leaves, involved in glucose efflux during phloem unloading (Ho et al., 2019); an observation verified in our study (**Figure 4A**). We further found that the *SlSWEET1a* was abundantly expressed in 2, 10, and 15 daa fruits, with expression levels increased 16-times in 2 daa fruitlets as compared to 2 dba ovaries (**Figure 4A**). Interestingly, *SlSWEET1b* was only highly expressed in 10 and 15 daa fruits but weakly in 2 daa fruit (**Figure 4B**). For the three potential *SlSWEET* genes encoding sucrose-transporting from clade III, expression levels of *SlSWEET12b* was abundant in vegetative tissues, especially in source leaves, but low in reproductive tissues (**Figure 4C**). Similar to the expression pattern of *SlSWEET1b*, the transcripts level of *SlSWEET12c* was highly expressed in 10 and 15 daa fruits (**Figure 4D**).

Among the *SlSWEET* genes investigated, *SlSWEET1a*, *1b*, and *12b* genes did not show any response in their mRNA levels to elevation of CWIN activity in tomato ovaries and developing fruit across all stages examined in the transgenic plants in comparison to the WT (**Figures 4E–G**). However, the transcript level of *SlSWEET12c* was significantly increased by 2-fold in 10 daa fruit in response to elevated CWIN activity (**Figure 4H**).

DISCUSSION

CWIN Activity May Be Regulated by Two Inhibitors and One Defective CWIN at the Ovary Stage

Our previous studies have shown that CWIN activity is regulated by *SlINVINH1* at the posttranslational level (Jin et al., 2009). Comparing with fruits, tight regulation of CWIN activity is needed at the ovary stage, probably through interaction with SldeCWIN1 and SlINVINH1, to ensure that the ovaries are well-protected and ready for fruit set upon fertilization (Palmer et al., 2015). Indeed, our transcriptomic analyses revealed that ovaries are more sensitive to elevated CWIN activity as compared to fruitlets (Ru et al., 2017).

In this study, we identified one more inhibitor-*SlINVINH2*, which is specifically expressed at ovary stage. It has been reported that invertase activity could be modulated by INVINHs and defective invertases at the posttranslational level (Jin et al., 2009; Le Roy et al., 2013; Lin et al., 2013). The presence of two INVINHs, *SlINVINH1* and the newly-identified *SlINVINH2*, may act in repressing CWIN activity through the action of SldeCWIN1 at ovary stages to prevent pre-mature cell division in the ovaries prior to fertilization, whereas the near absence of *SldeCWIN1* and *SlINVINH2* expression in 2, 5, and 10 daa fruits could

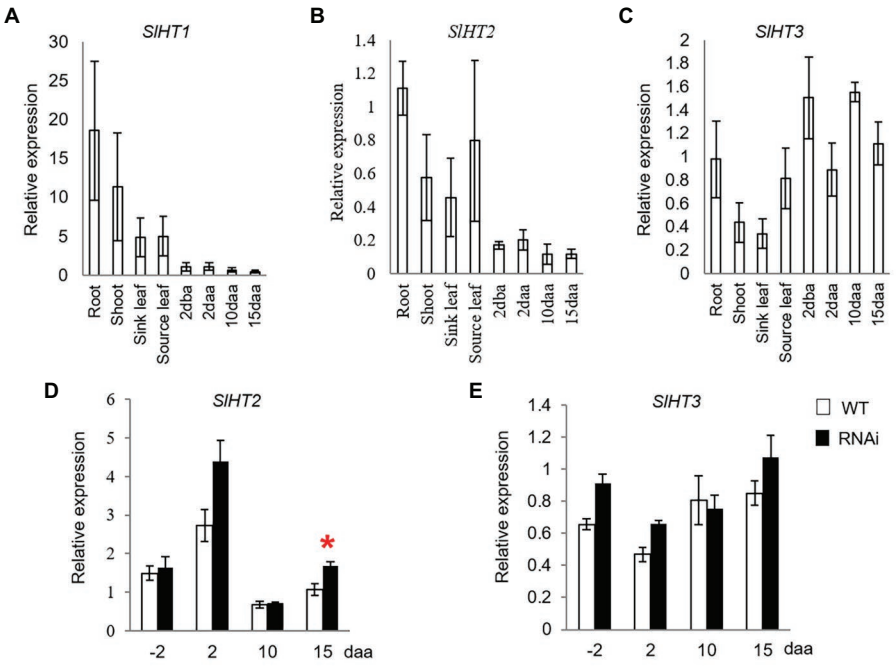


FIGURE 3 | Quantitative real-time-PCR analysis of the expression profiles of hexose transporter (HT) genes. **(A–C)** are expression profiles of HTs in different tissues of tomato plants. **(A)** *SlHT1*. **(B)** *SlHT2*. **(C)** *SlHT3*; **(D)** and **(E)** are expression profiles of HTs in response to elevated CWIN activity in *SlINVINH1*-RNAi as compared to wild type (WT) plants. **(D)** *SlHT2*. **(E)** *SlHT3*. Tissues examined were vegetative and reproductive tissues. Vegetative tissues included root and shoots from 2-week old seedlings, sink, and source leaves from similar raised plants. Reproductive tissues included 2 dba ovaries and 2, 10, and 15 daa fruits. *SlCAC* and *SlSAND* were used as reference genes. Each value is the mean ± SE of three biological replicates. Student *t*-test was applied in comparing *SlINVINH1*-RNAi and WT plants in each individual stage. *Indicates significance at $p < 0.05$.

alleviate the inhibition, leading to an increase in CWIN activity to stimulate cell division of the newly formed fruitlets following fertilization. Further experiments are required to determine whether and how SldeCWIN1 and two INHs interact with each other and with CWIN to keep the CWIN activity low prior to fertilization.

Enhanced Expression of *SlHT2* Could Facilitate Hexose Import Into Fruit Parenchyma Cells

Hexose produced by CWIN in the apoplasm has been shown to be transported into cells by SlHTs. SlHT3 is the main contributor to hexose accumulation in tomato fruit at the middle phase of development (Dibley et al., 2005; McCurdy et al., 2010). Interestingly, it is the expression level of *SlHT2*, but not *SlHT3*, that was increased in response to elevated CWIN activity in 15 daa fruits of the CWIN-elevated transgenic plants (**Figure 3D**). Fruits at this stage onward are at the onset of rapid hexose accumulation in tomato fruit with sucrose being unloaded apoplasmically from the phloem in the fruit pericarp (Ruan and Patrick, 1995). Extra hexoses released from elevated CWIN activity in 15 daa fruits of the transgenic plants may be taken up by SlHT2 but not SlHT3.

Phloem unloading of sucrose also follows an aoplasmic pathway in the ovary walls and placenta connecting to ovules in 2 dba ovaries (Palmer et al., 2015). However, in contrast to that in 15 daa fruit, no HTs was responsive to elevated CWIN activity in ovaries (**Figures 3D,E**), indicating that HTs are not sensitive to alteration of CWIN activity or that high expression of HTs is not required at this stage. To this end, it is worth noting that ovaries do not undergo rapid cell division and hexose accumulation, partly owing to low CWIN activity (Palmer et al., 2015) possible through the triple inhibition imposed by SldeCWIN1, SlINVINH1, and SlINVINH2 as hypothesized above. The low CWIN activity and weak expression of *HTs* in the ovaries are also consistent with the fact that only tiny amount of sugar is required to maintain ovary development (Palmer et al., 2015; Ru et al., 2017).

Enhanced Expression of *SlSWEET12c* Could Facilitated Sucrose Efflux From Phloem to Support Seed Filling

SWEETs are a new class sugar transporter identified over the last 5 years or so (Chen et al., 2015a,b). SWEET may facilitate sucrose or hexose efflux across membranes. Thus, apart from SlHTs, some SWEETs may also facilitate hexose transport. In this context, ZmSWEET4 has been reported to act as a hexose transporter downstream of a CWIN in maize since the expression levels of ZmSWEET4 were suppressed in the CWIN mutant (Sosso et al., 2015). In tomato, SlSWEET1a is localized to the plasma membranes of phloem parenchyma cells of immature leaves to take up glucose from the apoplasm of the phloem likely derived from the CWIN-mediated hydrolysis of unloaded sucrose (Ho et al., 2019). From our phytogenic analysis,

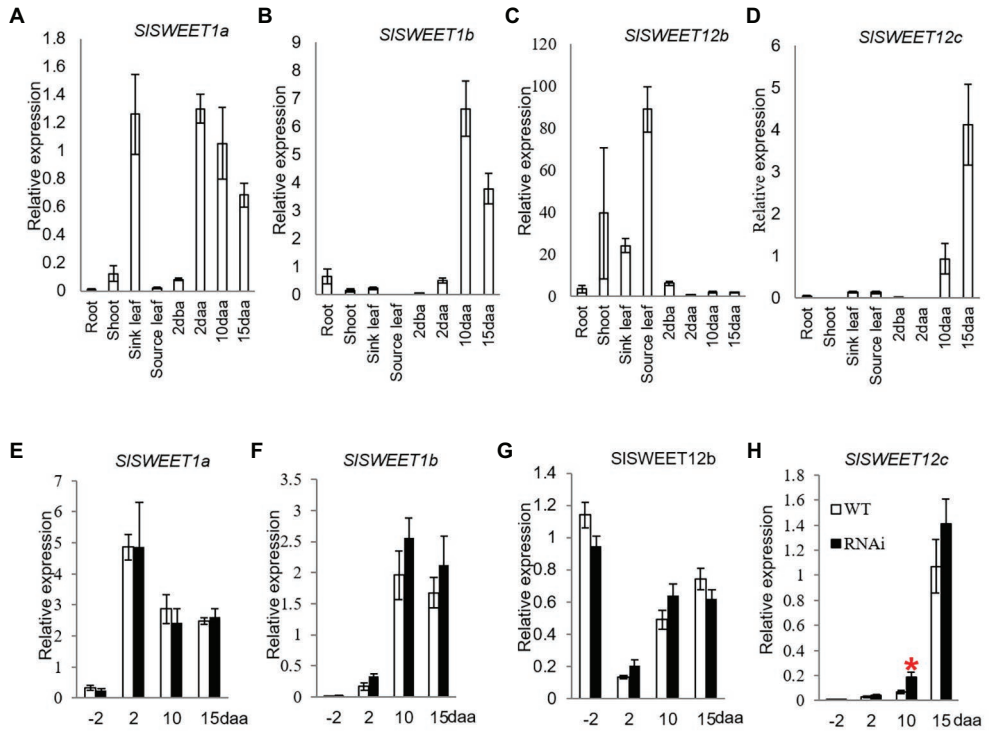


FIGURE 4 | Quantitative real-time-PCR analysis of the expression profiles of sugars will eventually be exported transporter (SWEET) genes. **(A–D)** are expression profiles of SWEETs in different tissues of tomato plants. **(A)** *SlSWEET1a*. **(B)** *SlSWEET1b*. **(C)** *SlSWEET12b*. **(D)** *SlSWEET12c*. **(D–G)** are expression profiles of SWEETs in response to elevated CWIN activity in *SlINVINH1*-RNAi as compared to WT plants. **(E)** *SlSWEET1a*. **(F)** *SlSWEET1b*. **(G)** *SlSWEET12b*. **(H)** *SlSWEET12c*. Tissues examined were vegetative and reproductive tissues. Tissues included 2 dba ovaries and 2, 10, and 15 daa fruits. *SlCAC* and *SlSAND* were used as reference genes. Each value is the mean ± SE of three biological replicates. Student *t*-test was applied in comparing *SlINVINH1*-RNAi and WT plants in each individual stage. *Indicates significance at $p < 0.05$.

SlSWEET1b is another putative hexose-transporting SWEET which might transport hexose into the cytoplasm of sink cells after phloem unloaded sucrose is degraded by CWIN. Expression levels of *SlSWEET1a* and *SlSWEET1b* were not changed in CWIN-elevated fruits (**Figures 4E,F**). Expression of *SlSWEET1a* showed an increased trend as fruit develops in tomato (Ho et al., 2019). However, this was not the case in our study. Indeed, its expression level decreased in fruit from 2 to 15 daa (**Figure 4A**). This result indicates that *SlSWEET1a* expression may be cultivar-dependent. Nevertheless, SlSWEET1a seems to act as a glucose effluxer in both vegetative and reproductive organs, such as developing leaves and fruits.

In developing tomato fruit, phloem unloading of sucrose occurs apoplasmically in the sites of placenta connecting to seeds (Jin et al., 2009; Palmer et al., 2015). Here, sucrose must move out of the phloem likely facilitated by clade III SWEET transporters. For instance, the AtSWEET12 facilitates sucrose efflux from phloem of seed coat to embryo in *Arabidopsis* seeds (Chen et al., 2015b). In tomato, the orthologs of AtSWEET12, SlSWEET12b, and SlSWEET12c might play a similar role in facilitating sucrose efflux down a sucrose concentration gradient out of the phloem into seeds of 10 daa fruit. High expression of *CWIN* in this region (Jin et al., 2009; Palmer et al., 2015) could create a steep sucrose gradient between

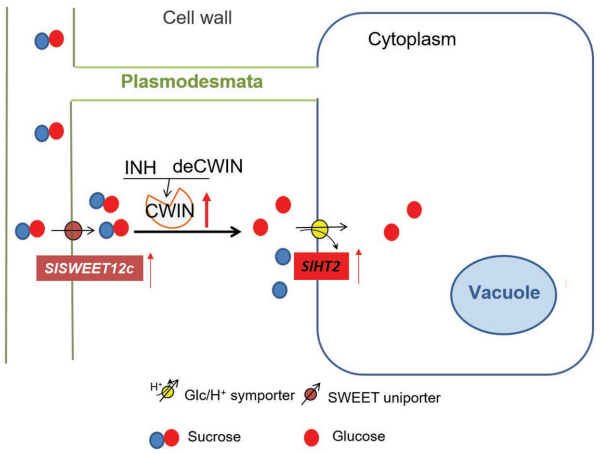


FIGURE 5 | A model of sugar transporters in response to elevated CWIN activity during tomato fruit development. CWIN degarades sucrose (Suc) into glucose (Glc) and fructose. CWIN activity could be tightly regulated by SlINVINHs and/or SldeCWIN1. Suc efflux may be facilitated by Suc-transporting SlSWEETs to the apoplasmic space. Elevated CWIN activity enhanced the expression of *SlSWEET12c*, which may unload extra Suc into cell wall space. Extra Glu produced through elevated CWIN activity is taken up by SlHT2 into the cytosol.

sieve element/companion cell complex and the adjacent apoplasm of the seeds by degrading sucrose into hexose (Patrick, 1997; Ruan, 2014). *SlSWEET12c* is highly expressed in tomato seeds (**Supplementary Figure S4**) and exhibited an increased expression in response to elevated CWIN activity at 10 daa fruit (**Figure 4H**). We thus speculate that SlSWEET12c may facilitate sucrose unloading from the phloem down a concentration gradient maintained by CWIN to support seed filling.

In summary, our results indicate that CWIN activity appears to be tightly controlled by multiple regulators at the ovary stage. Among the sugar transporters investigated, only SWEET12c and SlHT2 were selectively upregulated by elevated CWIN activity during tomato fruit development (**Figure 5**). The nature of this CWIN-mediated upregulation of these specific sugar transporters is yet to be determined. However, our recent study demonstrates that CWIN positively regulates ovule development through sugar signaling that is potentially perceived by a cohort of HTs and clades I and II SWEETs (Liao et al., 2020). Experiments are underway to test these possibilities along with the functional implications of this coupling between sucrose catabolism and sugar transport in fruit development.

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

Y-LR conceived the project. LR and Y-LR designed the experiments. LR and YH conducted the experiments. LR, Y-LR, and JP analyzed the data. LR, Y-LR, JP, YH, and ZZ wrote the manuscripts. All authors contributed to the article and approved the submitted version.

FUNDING

The work was partially supported by Australian Research Council (DP180103834) and National Science Foundation of China (31801915).

ACKNOWLEDGMENTS

LR gratefully acknowledges scholarship support from the Chinese Scholarship Council and the University of Newcastle.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.592596/full#supplementary-material

REFERENCES

Ayre, B. G. (2011). Membrane-transport systems for sucrose in relation to whole-plant carbon partitioning. *Mol. Plant* 4, 377–394. doi: 10.1093/mp/ssr014

Azzi, L., Deluche, C., Gevaudant, F., Frangne, N., Delmas, F., Hernould, M., et al. (2015). Fruit growth-related genes in tomato. *J. Exp. Bot.* 66, 1075–1086. doi: 10.1093/jxb/eru527

Chen, L. Q., Cheung, L. S., Feng, L., Tanner, W., and Frommer, W. B. (2015a). Transport of sugars. *Annu. Rev. Biochem.* 84, 865–894. doi: 10.1146/annurev-biochem-060614-033904

Chen, L. Q., Lin, I. W., Qu, X. Q., Sosso, D., McFarlane, H. E., Londono, A., et al. (2015b). A cascade of sequentially expressed sucrose transporters in the seed coat and endosperm provides nutrition for the *Arabidopsis* embryo. *Plant Cell* 27, 607–619. doi: 10.1105/tpc.114.134585

Chen, L. Q., Qu, X. Q., Hou, B. H., Sosso, D., Osorio, S., Fernie, A. R., et al. (2012). Sucrose efflux mediated by SWEET proteins as a key step for phloem transport. *Science* 335, 207–211. doi: 10.1126/science.1213351

Dibley, S. J., Gear, M. L., Yang, X., Rosche, E. G., Offler, C. E., McCurdy, D. W., et al. (2005). Temporal and spatial expression of hexose transporters in developing tomato (*Lycopersicon esculentum*) fruit. *Funct. Plant Biol.* 32, 777–785. doi: 10.1071/fp04224

Feng, C. Y., Han, J. X., Han, X. X., and Jiang, J. (2015). Genome-wide identification, phylogeny, and expression analysis of the SWEET gene family in tomato. *Gene* 573, 261–272. doi: 10.1016/j.gene.2015.07.055

Fridman, E., Carrari, F., Liu, Y. S., Fernie, A. R., and Zamir, D. (2004). Zooming in on a quantitative trait for tomato yield using interspecific introgressions. *Science* 305, 1786–1789. doi: 10.1126/science.1101666

Fridman, E., and Zamir, D. (2003). Functional divergence of a syntenic invertase gene family in tomato, potato, and *Arabidopsis*. *Plant Physiol.* 131, 603–609. doi: 10.1104/pp.014431

Hackel, A., Schauer, N., Carrari, F., Fernie, A. R., Grimm, B., and Kuhn, C. (2006). Sucrose transporter LeSUT1 and LeSUT2 inhibition affects tomato fruit development in different ways. *Plant J.* 45, 180–192. doi: 10.1111/j.1365-313X.2005.02572.x

Ho, L. H., Klemens, P. A. W., Neuhaus, H. E., Ko, H. Y., Hsieh, S. Y., and Guo, W. J. (2019). SlSWEET1a is involved in glucose import to young leaves in tomato plants. *J. Exp. Bot.* 70, 3241–3254. doi: 10.1093/jxb/erz154

Jin, Y., Ni, D. A., and Ruan, Y. L. (2009). Posttranslational elevation of cell wall invertase activity by silencing its inhibitor in tomato delays leaf senescence and increases seed weight and fruit hexose level. *Plant Cell* 21, 2072–2089. doi: 10.1105/tpc.108.063719

Kuhn, C., and Grof, C. P. (2010). Sucrose transporters of higher plants. *Curr. Opin. Plant Biol.* 13, 288–298. doi: 10.1016/j.pbi.2010.02.001

Lalonde, S., Tegeder, M., Throne-Holst, M., Frommer, W. B., and Patrick, J. W. (2003). Phloem loading and unloading of sugars and amino acids. *Plant Cell Environ.* 26, 37–56. doi: 10.1046/j.1365-3040.2003.00847.x

Li, J., Wu, L., Foster, R., and Ruan, Y. L. (2017). Molecular regulation of sucrose catabolism and sugar transport for development, defence and phloem function. *J. Integr. Plant Biol.* 59, 322–335. doi: 10.1111/jipb.12539

Liao, S., Wang, L., Li, J., and Ruan, Y. L. (2020). Cell wall invertase is essential for ovule development through sugar signaling rather than provision of carbon nutrients. *Plant Physiol.* 183, 1126–1144. doi: 10.1104/pp.20.00400

Le Roy, K., Vergauwen, R., Struyf, T., Yuan, S., Lammens, W., Mátrai, J., et al. (2013). Understanding the role of defective invertases in plants: tobacco Nin88 fails to degrade sucrose. *Plant Physiol.* 161, 1670–1681. doi: 10.1104/pp.112.209460

Lin, Y., Liu, J., Liu, X., Ou, Y., Li, M., Zhang, H., et al. (2013). Interaction proteins of invertase and invertase inhibitor in cold-stored potato tubers suggested a protein complex underlying post-translational regulation of invertase. *Plant Physiol. Biochem.* 73, 237–244. doi: 10.1016/j.plaphy.2013.09.012

Liu, Y. -H., Offler, C. E., and Ruan, Y. -L. (2016). Cell wall invertase promotes fruit set under heat stress by suppressing ROS-independent cell death. *Plant Physiol.* 172, 163–180. doi: 10.1104/pp.16.00959

Ludewig, F., and Flugge, U. I. (2013). Role of metabolite transporters in source-sink carbon allocation. *Front. Plant Sci.* 4:231. doi: 10.3389/fpls.2013.00231

McCurdy, D. W., Dibley, S., Cahyanegara, R., Martin, A., and Patrick, J. W. (2010). Functional characterization and RNAi-mediated suppression reveals roles for hexose transporters in sugar accumulation by tomato fruit. *Mol. Plant* 3, 1049–1063. doi: 10.1093/mp/ssq050

Mounet, F., Moing, A., Garcia, V., Petit, J., Maucourt, M., Deborde, C., et al. (2009). Gene and metabolite regulatory network analysis of early developing fruit tissues highlights new candidate genes for the control of tomato fruit composition and development. *Plant Physiol.* 149, 1505–1528. doi: 10.1104/pp.108.133967

Palmer, W. M., Ru, L., Jin, Y., Patrick, J. W., and Ruan, Y. L. (2015). Tomato ovary-to-fruit transition is characterized by a spatial shift of mRNAs for cell wall invertase and its inhibitor with the encoded proteins localized to sieve elements. *Mol. Plant* 8, 315–328. doi: 10.1016/j.molp.2014.12.019

Patrick, J. W. (1997). Phloem unloading: sieve element unloading and post-sieve element transport. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 48, 191–222. doi: 10.1146/annurev.arplant.48.1.191

Patrick, J. W., Zhang, W., Tyerman, S. D., Offler, C. E., and Walker, N. A. (2001). Role of membrane transport in phloem translocation of assimilates and water. *Aust. J. Plant Physiol.* 28, 695–707. doi: 10.1071/PP01023

Pfaffl, M. W. (2001). A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res.* 29, 2002–2007. doi: 10.1093/nar/29.9.e45

Qin, G., Zhu, Z., Wang, W., Cai, J., Chen, Y., Li, L., et al. (2016). A tomato vacuolar invertase inhibitor mediates sucrose metabolism and influences fruit ripening. *Plant Physiol.* 172, 1596–1611. doi: 10.1104/pp.16.01269

Reinders, A., Sivitz, A. B., and Ward, J. M. (2012). Evolution of plant sucrose uptake transporters. *Front. Plant Sci.* 3:22. doi: 10.3389/fpls.2012.00022

Reuscher, S., Akiyama, M., Yasuda, T., Makino, H., Aoki, K., Shibata, D., et al. (2014). The sugar transporter inventory of tomato: genome-wide identification and expression analysis. *Plant Cell Physiol.* 55, 1123–1141. doi: 10.1093/pcp/pcu052

Ru, L., Osorio, S., Wang, L., Fernie, A. R., Patrick, J. W., and Ruan, Y. L. (2017). Transcriptomic and metabolomics responses to elevated cell wall invertase activity during tomato fruit set. *J. Exp. Bot.* 68, 4263–4279. doi: 10.1093/jxb/erx219

Ruan, Y. L. (2014). Sucrose metabolism: gateway to diverse carbon use and sugar signaling. *Annu. Rev. Plant Biol.* 65, 33–67. doi: 10.1146/annurev-arplant-050213-040251

Ruan, Y. L., and Patrick, J. W. (1995). The cellular pathway of postphloem sugar-transport in developing tomato fruit. *Planta* 196, 434–444. doi: 10.1007/bf00203641

Ruan, Y. L., Patrick, J. W., Bouzayen, M., Osorio, S., and Fernie, A. R. (2012). Molecular regulation of seed and fruit set. *Trends Plant Sci.* 17, 656–665. doi: 10.1016/j.tplants.2012.06.005

Schroeder, J. I., Delhaize, E., Frommer, W. B., Guerinot, M. L., Harrison, M. J., Herrera-Estrella, L., et al. (2013). Using membrane transporters to improve crops for sustainable food production. *Nature* 497, 60–66. doi: 10.1038/nature11909

Schwarz, D., Thompson, A. J., and Kläring, H. -P. (2014). Guidelines to use tomato in experiments with a controlled environment. *Front. Plant Sci.* 5:625. doi: 10.3389/fpls.2014.00625

Shammai, A., Petreikov, M., Yeselson, Y., Faigenboim, A., Moy-Komemi, M., Cohen, S., et al. (2018). Natural genetic variation for expression of a SWEET transporter among wild species of *Solanum lycopersicum* (tomato) determines the hexose composition of ripening tomato fruit. *Plant J.* 96, 343–357. doi: 10.1111/tpj.14035

Shen, S., Ma, S., Liu, Y., Liao, S., Li, J., Wu, L., et al. (2019). Cell wall invertase and sugar transporters are differentially activated in tomato styles and ovaries during pollination and fertilization. *Front. Plant Sci.* 10:506. doi: 10.3389/fpls.2019.00506

Slewinski, T. L. (2011). Diverse functional roles of monosaccharide transporters and their homologs in vascular plants: a physiological perspective. *Mol. Plant* 4, 641–662. doi: 10.1093/mp/ssr051

Sosso, D., Luo, D., Li, Q. -B., Sasse, J., Yang, J., Gendrot, G., et al. (2015). Seed filling in domesticated maize and rice depends on SWEET-mediated hexose transport. *Nat. Genet.* 47, 1489–1493. doi: 10.1038/ng.3422

Tieman, D., Zhu, G., Resende, M. F. Jr., Lin, T., Nguyen, C., Bies, D., et al. (2017). A chemical genetic roadmap to improved tomato flavor. *Science* 355, 391–394. doi: 10.1126/science.aal1556

Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., et al. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* 3:Research0034. doi: 10.1186/gb-2002-3-7-research0034

Wang, Z., Wei, X., Yang, J., Li, H., Ma, B., Zhang, K., et al. (2020). Heterologous expression of the apple hexose transporter MdHT2.2 altered sugar concentration with increasing cell wall invertase activity in tomato fruit. *Plant Biotechnol. J.* 18, 540–552. doi: 10.1111/pbi.13222

Weschke, W., Panitz, R., Gubatz, S., Wang, Q., Radchuk, R., Weber, H., et al. (2003). The role of invertases and hexose transporters in controlling sugar ratios in maternal and filial tissues of barley caryopses during early development. *Plant J.* 33, 395–411. doi: 10.1046/j.1365-313X.2003.01633.x

Zanor, M. I., Osorio, S., Nunes-Nesi, A., Carrari, F., Lohse, M., Usadel, B., et al. (2009). RNA interference of LIN5 in tomato confirms its role in controlling Brix content, uncovers the influence of sugars on the levels of fruit hormones, and demonstrates the importance of sucrose cleavage for normal fruit development and fertility. *Plant Physiol.* 150, 1204–1218. doi: 10.1104/pp.109.136598

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.

Copyright © 2020 Ru, He, Zhu, Patrick and Ruan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 21 October 2020
doi: 10.3389/fgene.2020.570138

HpeNet: Co-expression Network Database for *de novo* Transcriptome Assembly of *Paeonia lactiflora* Pall

Minghao Sheng[1,2†], Jiajie She[2†], Wenying Xu[2], Yan Hong[1], Zhen Su[2] and Xiaodong Zhang[1]**

[1] *Beijing Agricultural Biotechnology Research Center, Beijing Engineering Research Center of Functional Floriculture, Beijing Academy of Agriculture and Forestry Science, Beijing, China,* [2] *State Key Laboratory of Plant Physiology and Biochemistry, College of Biological Sciences, China Agricultural University, Beijing, China*

The herbaceous peony (*Paeonia lactiflora* Pall.) is a well-known ornamental flowering and pharmaceutical plant found in China. Its high medicinal value has long been recognized by traditional Chinese medicine (as *Radix paeoniae Alba* and *Radix paeoniae Rubra*), and it has become economically valued for its oilseed in recent years; like other *Paeonia* species, it has been identified as a novel resource for the α-linolenic acid used in seed oil production. However, its genome has not yet been sequenced, and little transcriptome data on *Paeonia lactiflora* are available. To obtain a comprehensive transcriptome for *Paeonia lactiflora*, RNAs from 10 tissues of the *Paeonia lactiflora* Pall. cv Shaoyou17C were used for *de novo* assembly, and 416,062 unigenes were obtained. Using a homology search, it was found that 236,222 (approximately 57%) unigenes had at least one BLAST hit in one or more public data resources. The construction of co-expression networks is a feasible means for improving unigene annotation. Using in-house transcriptome data, we obtained a co-expression network covering 95.13% of the unigenes. Then we integrated co-expression network analyses and lipid-related pathway genes to study lipid metabolism in *Paeonia lactiflora* cultivars. Finally, we constructed the online database HpeNet (http://bioinformatics.cau.edu.cn/HpeNet) to integrate transcriptome data, gene information, the co-expression network, and so forth. The database can also be searched for gene details, gene functions, orthologous matches, and other data. Our online database may help the research community identify functional genes and perform research on *Paeonia lactiflora* more conveniently. We hope that *de novo* transcriptome assembly, combined with co-expression networks, can provide a feasible means to predict the gene function of species that do not have a reference genome.

Keywords: ***Paeonia lactiflora*** **Pall.,** ***de novo*** **assembly, functional annotation, co-expression network, database**

OPEN ACCESS

Edited by:
Weike Duan,
Huaiyin Institute of Technology, China

Reviewed by:
Ze Peng,
University of Florida, United States
Rajesh Kumar Gazara,
Indian Institute of Technology Roorkee, India

****Correspondence:***
Xiaodong Zhang
zhangxiaodong@baafs.net.cn
Zhen Su
zhensu@cau.edu.cn

† These authors have contributed equally to this work

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

Received: *06 June 2020*
Accepted: *18 August 2020*
Published: *21 October 2020*

Citation:
Sheng M, She J, Xu W, Hong Y, Su Z and Zhang X (2020) HpeNet: Co-expression Network Database for de novo Transcriptome Assembly of Paeonia lactiflora Pall. Front. Genet. 11:570138. doi: 10.3389/fgene.2020.570138

INTRODUCTION

Paeonia lactiflora Pall., commonly known as the herbaceous peony, is an ornamental flowering plant, well-known around the world. It is cultivated in more than 50 countries, including China, United States, New Zealand, and Turkey (Zhao et al., 2016). It has significant medicinal value (Yang et al., 2020) and can be used together with liquorice to treat a range of conditions, including

spasmolysis; it also provides pain relief and calm (Wang et al., 2009). In recent years, the oil of *Paeonia lactiflora* has attracted increasing attention because it contains a high proportion of polyunsaturated fatty acids (PUFAs). Using PUFAs in place of saturated fat can help reduce the risk for cardiovascular disease and coronary heart disease in humans (Mozaffarian et al., 2010; Rizos et al., 2012). *Paeonia lactiflora* has recently attracted greater research attention. However, unlike other model species, the genome of *Paeonia lactiflora* has not been published, and population genetic resources are lacking, inhibiting gene function investigation and molecular breeding at the whole-genome level.

Recent years have seen extraordinary progress in next-generation sequencing technologies and many discoveries in high-throughput technology. Eight flavonoid metabolic pathway genes have been isolated, and the expression patterns of these eight genes in the flowers have been investigated (Zhao et al., 2012), enabling the mechanism of *Paeonia lactiflora* flower diversity to be described. It has also been found via transcriptome data analyses that the inhibition of anthocyanin biosynthesis mediates yellow formation (the color of the inner-petal is yellow) in *Paeonia lactiflora* (Zhao et al., 2014). Zifengyu and Hongyanzhenghui have been identified as thermo-tolerant and moderately thermo-tolerant cultivars, respectively. Analyses of the transcriptome data of these two cultivars have led to the detection of 161 heat stress response genes (Hao et al., 2016). Some miRNAs that might be involved in heat stress response have also been identified by miRNA sequencing (Hao et al., 2017). In addition, a combination of transcriptome and isobaric tags for relative and absolute quantification proteome (iTRAQ) analyses have revealed that APETALA2 regulates the development of flower shape (Wu et al., 2018). However, unlike other well-studied species, little transcriptome data have been assembled for *Paeonia lactiflora*, and these data focus on specific tissues and different cultivars. It is difficult to study *Paeonia lactiflora* at the whole-genome level using only available transcriptome data. These data have been generated in response to scientific questions, so certain tissue-specific genes and regulated genes are unlikely to have been detected. Additionally, transcriptome data from different cultivars show large differences, which make it more difficult to compare transcriptome data.

Without a reference genome for *Paeonia lactiflora*, *de novo* transcriptome assembly, a proven technology, can allow us to identify complete and contiguous mRNA sequences from RNA-seq data (Guttman et al., 2010; Grabherr et al., 2011; Davidson and Oshlack, 2014). There are great demands for high-quality transcriptome data from different tissues of the same cultivar. In this way, it is possible to predict and annotate the genes of *Paeonia lactiflora*, and these results can support investigations of gene function and molecular breeding. Using sequence similarity, identified assembled unigenes have been annotated by known datasets. In addition, co-expression analyses are a helpful way to improve functional annotation (Eisen et al., 1998; Walker et al., 1999). Gene function is closely related to gene-expression profiles, which can be established through the acquisition of transcriptome data (Inoue and Horimoto, 2017) (Moretto et al., 2019). Through the integration of available transcriptome datasets, constructed gene co-expression networks can link genes to specialized metabolic pathways, which can identify important genes and infer their potential functions (Rhee and Mutwil, 2014; Wisecaver et al., 2017; Salleh et al., 2018; van Dam et al., 2018). Co-expression networks must be constructed with high-quality and more complete transcriptome data. Transcriptome data have already been accumulated through model plants, and some co-expression network databases have been constructed, such as for *Arabidopsis*, rice, maize and cotton (You et al., 2015; Lee and Lee, 2017; You et al., 2017; Ma et al., 2018; Obayashi et al., 2018; Tian et al., 2018). For instance, ATTED-II is a co-expression database for nine plant species, with microarray data from 16,033 samples and RNA-seq data from 2,120 samples only for *Arabidopsis*. The RiceFREND database includes a large collection of rice microarray data drawn from various tissues at different stages of growth. The creation of similar resources for *Paeonia lactiflora* would require high-quality transcriptome data from the same cultivar and from different tissues to construct co-expression networks. Once this is done, functional annotation on the whole-genome level can be performed.

To improve the annotation of *Paeonia lactiflora*, we employed high-throughput sequencing technology and sequenced the transcriptomes of 10 tissues drawn from different growth stages of *Paeonia lactiflora* Pall. cv Shaoyou17C. With *de novo* assembly and gene function prediction, we produced reference transcriptome datasets. In addition, to further annotate the unigenes of *Paeonia lactiflora* and study gene functions, we constructed a co-expression network and created an online database, HpeNet[1]. The database contains transcriptome data, gene information, the co-expression network, and so forth. We hope that the resulting database will be beneficial to researchers conducting future work on *Paeonia lactiflora* Pall.

MATERIALS AND METHODS

Plant Materials, Growth Conditions, and RNA Isolation

Paeonia lactiflora Pall. cv Shaoyou17C (a cultivar with high oilseed yield and high oil content bred by the Beijing Academy of Agriculture and Forestry Science), cv Hangbaishao (*Radix paeoniae Alba*), and cv Chishao (*Radix paeoniae Rubra*) were obtained from Yucheng County (Henan Province, China) and cultivated in open fields at Beijing Academy of Agriculture and Forestry Science in Beijing, China. All of the plants were 5 years old and were normally flowered and fruited, using standard agronomic practices.

To identify and annotate the transcripts more comprehensively, 10 tissues of *Paeonia lactiflora* Pall. cv Shaoyou17C were harvested from three biological replicates: the leaf (LY), flower (FY), kernel (KnY), ovary (OvY), stigma (StY), androecium (AnY), fibrous root (FRY), pod (CpY), seed coat (HsY) and four growth stages of seed (SY). In addition, the leaf (LMS) of *Paeonia lactiflora* Pall. cv Chishao and the flower (FW)

[1]http://bioinformatics.cau.edu.cn/HpeNet/

and the leaf (LW) of *Paeonia lactiflora* Pall. cv Hangbaishao were also harvested from three biological replicates to isolate the RNA. All fresh samples were homogenized in liquid nitrogen and stored at −80°C.

Total RNA was isolated from each sample using TRIZOL® reagent (Invitrogen, Carlsbad, CA, United States), according to the manufacturer's protocol. The RNA was quantified and qualified using the Qubit® RNA Assay Kit in a Qubit® 2.0 Fluorometer (Life Technologies, Carlsbad, CA, United States) to measure RNA concentrations. RNA integrity was assessed using the RNA Nano 6000 Assay Kit of the Agilent Bioanalyzer 2100 system (Agilent Technologies, Santa Clara, CA, United States). Finally, the total RNA of 40 samples from different tissues and cultivars were obtained.

Library Preparation and Sequencing

All library constructions and Illumina paired-end sequencing were done with Novogene. The sequencing libraries were generated using the NEBNext® Ultra™ RNA Library Prep Kit for Illumina® (NEB, United States), following the manufacturer's recommendations. Poly-T oligo-attached magnetic beads were used to purify the mRNA from total RNA. The quality of these libraries was assessed on the Agilent Bioanalyzer 2100 system. The libraries were prepared from three biological replicates of 10 tissues, sequenced on the IlluminaHiSeq™ 2000 platform, following standard protocols. The paired-end reads were generated in 150 bp format.

Data Processing and Transcriptome *de novo* Assembly

The experimental design and mRNA-seq procedures are shown in **Supplementary Figure 1**. The Q20, Q30, GC-content, and sequence duplication levels of the clean data were calculated with FastQC. We defined the number of bases using a mass value of $Q_{phred} \leq 20$, which accounted for more than 50% of the entire read as low-quality reads. To generate clean data, all reads containing adapter, reads containing Ploy-N of more than 10%, and low-quality reads were removed. All of the clean reads of *Paeonia lactiflora* Pall. cv Shaoyou17C were used for *de novo* transcriptome assembly using Trinity (r20140413p1) (Grabherr et al., 2011), and all of the parameters were set to the defaults (min_contig_length 200, min_glue 2, min_kmer_cov 2). The transcript sequence obtained by Trinity was used as a reference sequence for subsequent analyses. Then Corset (Davidson and Oshlack, 2014) was employed to create a hierarchical cluster from the transcripts according to the number of mapped reads and expression patterns. After clustering was done with the Corset hierarchy, the longest cluster sequence (unigene) was obtained to produce functional annotation and transcriptome expression analyses.

Functional Annotation of Unigenes

We predicted the annotations of the unigenes in *Paeonia lactiflora* Pall. cv Shaoyou17C using sequence alignment searches for homologous proteins against seven authoritative public databases (Altschul et al., 1997), including NCBI non-redundant protein sequences (NR[2]), NCBI non-redundant nucleotide sequences (NT[2]), Protein family (Pfam (Finn et al., 2008)[3]), the euKaryotic Ortholog Groups (KOG[4]), Swiss-Prot[5], Kyoto Encyclopedia of Genes and Genomes (KEGG[6]), and Gene Ontology (GO[7]). Blast2GO software (b2g4pipe_v2.5; Gotz et al., 2008) was used to predict the GO annotation of unigenes with an E-value cutoff of 1E−6. KEGG annotations were generated using the KEGG Automatic Annotation Server (KAAS, r140224; (Moriya et al., 2007), with an E-value cutoff of 1E−10. We obtained the domain architecture of the protein family from the Pfam database[8] (Finn et al., 2010) and the HMMER 3.0 package, with the cutoff for the E-value set to 0.01. The NT database annotation was based on NCBI BLAST (v2.2.28+), and an E-value of 1E−5 was chosen as the cutoff. The remaining three databases, NR, Swiss-Prot, and KOG, were searched using diamond software (v0.8.22), with E-values of 1E−5, 1E−5, and 1E−3, respectively. To identify the possible orthologous genes in *Arabidopsis*, the unigenes were compared with The *Arabidopsis* Information Resource 10 (TAIR10) for *Arabidopsis thaliana*[9] using the BLASTX method, with an E-value cutoff of 1E−5, and the top hits were selected as orthologous genes.

Transcriptome Expression Analyses

The unigene expression level for each sample was estimated using RSEM (Li and Dewey, 2011). RSEM does not require a reference genome and, it is capable of handling multiple mapped reads. First, clean data were mapped back onto the assembled transcriptome. Then, read counts for each unigene were obtained from the mapping results, and the read counts were converted into fragments per million mapped reads (FPKM) to analyze the expression level of the unigene. We set FPKM > 0.3 as the minimum FPKM value to call a unigene as an expressed unigene. This threshold is recommended by mainstream magazines and, can also reflect gene expression levels well. To analyze the differential gene expression, the DESeq package (Love et al., 2014) was employed, and the methods of Benjamini and Hochberg were used to control the false discovery rate and adjust the resulting P-value. We screened the threshold for Padj < 0.05 and $|\log2FoldChange| > 1$ as differentially expressed genes (DEGs).

Co-expression Network Construction Algorithm

Pearson's correlation coefficient (PCC) was used to calculate the correlations between the expression patterns of two unigenes among all of the samples, and mutual rank (MR) took a geometric average of the PCC rank (Obayashi and Kinoshita, 2009; Klinkenberg et al., 2014; Aoki et al., 2016). The mathematical

[2]https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/

[3]http://pfam.xfam.org/

[4]http://www.ncbi.nlm.nih.gov/COG/

[5]https://www.ebi.ac.uk/uniprot/

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

[7]http://geneontology.org/

[8]http://pfam.xfam.org/search#tabview=tab3

[9]https://www.arabidopsis.org/

formulas for PCC and MR are as follows:

$$PCC = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

$$MR(AB) = \sqrt{(Rank(A \rightarrow B) \times Rank(B \rightarrow A))} \quad (2)$$

where X and Y represent the FPKM values for two genes, and n represents the number of samples; i represents samples in the formula under different conditions. To calculate the MR values, we ranked all gene pairs from high to low by their PCC values. Using formula (2), MR takes the geometric average of the relative positions of genes A and B based on the ranked PCC values (You et al., 2015, 2017; Ma et al., 2018; Tian et al., 2018; She et al., 2019).

Statistical Test Method

Fisher's exact test was used to compute significance in gene-set enrichment analyses (Yi et al., 2013). The mathematical formula for this method is given as follows:

$$P = \frac{\binom{n}{k}\binom{N-n}{K-k}}{\binom{N}{K}} \quad (3)$$

where N represents the number of all protein-coding unigenes, n represents the gene number in the query gene list, K is the gene number for a certain background gene set, and k is the number of overlapping genes between the query gene list and the background gene set.

Search and Visualization Platform

The database was constructed in the LAMP (Red Hat Linux + Apache[10] + MySQL[11] + PHP[12]) environment. The network visualization was dynamically displayed in Cytoscape.js[13], an open-source java script package (Franz et al., 2016).

DATABASE CONSTRUCTION AND CONTENT

Transcriptome of 10 *Paeonia lactiflora* Pall. Tissues Determined via Sequencing and *de novo* Assembly

Paeonia lactiflora Pall. cv Shaoyou17C is a hybrid cultivar of *Paeonia lactiflora* Pall. cv Chishao and *Paeonia lactiflora* Pall. cv Hangbaishao. *Paeonia lactiflora* Pall. cv Shaoyou17C has high seed production and oil content bred. By sequencing and analyzing its transcriptome, we wish to discover the related genes regulating the high oil content and high proportion of PUFAs. There is no reference genome for *Paeonia lactiflora* Pall at present, so we chose transcriptome sequencing with *de novo* assembly to investigate transcripts in the 10 tissues of *Paeonia lactiflora* Pall. cv Shaoyou17C. The total RNA of each sample with three biological replicates (except for the four stages of the seed) was individually isolated, and the transcriptome profiles were generated on the Illumina HiSeq2000 platform (see detailed description in section "MATERIALS AND METHODS"). After data cleaning and quality control, a total of 2 billion 327 million high-quality 150 bp paired–end reads, at about 349.05 Gb, were generated (**Supplementary Table 1**).

To obtain a high-quality and comprehensive transcriptome, all of the high-quality reads from 10 tissues of Shaoyou17C were used for transcript assembly. The overlapping information of the high-quality reads provided a total of 519,560 transcripts for successful assembly using Trinity software. Then Corset software was used to cluster contigs, count reads, and remove redundancy from the assembly. At the end of this process, we obtained 416,062 unigenes. Next, we assessed the quality of the assembled transcriptome. The lengths of these unigenes ranged from 201 bp to 13,684 bp, with a mean length of 965 bp and an N50 of 1,462 bp (**Supplementary Table 2**). The length frequency of assembled unigenes was further analyzed (**Figure 1A**). We found that 167,108 unigenes (40.16%) were less than 500 bp, and 114,982 unigenes (27.64%) ranged from 500 to 1000 bp. There were 117,052 unigenes (20.90%) that ranged from 1000 bp to 3000 bp and 16,920 unigenes (11.3%) that were longer than 3000 bp (**Figure 1A**).

Functional Annotation of the Assembled Unigenes

All of the unigenes were annotated using the results of the BLAST comparison public databases, including the NR, NT, Swiss-Prot, Pfam, KOG, and the *Arabidopsis* protein (TAIR10) databases. The results showed that 197,488 (47.46%) and 129,014 (31%) unigenes were annotated with the NR (E-value $\leq$ 1E–5) and NT (E-value $\leq$ 1E–5), respectively. Next, we searched the predicted unigene sequences against two of the other databases. A homology search in the Pfam database (E-value $\leq$ 0.01) indicated that 139,121 transcripts were annotated at least once, while in Swiss-Prot database (E-value $\leq$ 1E–5), 138,800 unigenes were annotated. To provide further functional prediction and classification, all assembled unigenes were aligned using the KOG database (E-value $\leq$ 1E–3). A total of 45,348 unigenes were annotated into 26 clusters (**Figure 1B**). Putative functions were assigned for over 50% of the unigenes (**Table 1**).

KEGG annotation provides information on the molecular network. A total of 72,433 (17.4%) unigenes with E-values < 1E–10 were mapped to 130 pathways belonging to all five major categories of KEGG (**Figure 1C**, **Supplementary Table 3**). Most unigenes were annotated in the metabolism category, with 43,031 unigenes. Lipid metabolism is a metabolic pathway related to oil content featuring 4880 unigenes. The oil of *Paeonia lactiflora* contains a high proportion of linolenic acid. Of all of the pathways, ko00061 (Fatty acid biosynthesis) and ko00071 (Fatty acid degradation) were most likely to significantly contribute

[10]https://www.apache.org/
[11]https://www.mysql.com/
[12]https://php.net/
[13]http://js.cytoscape.org/

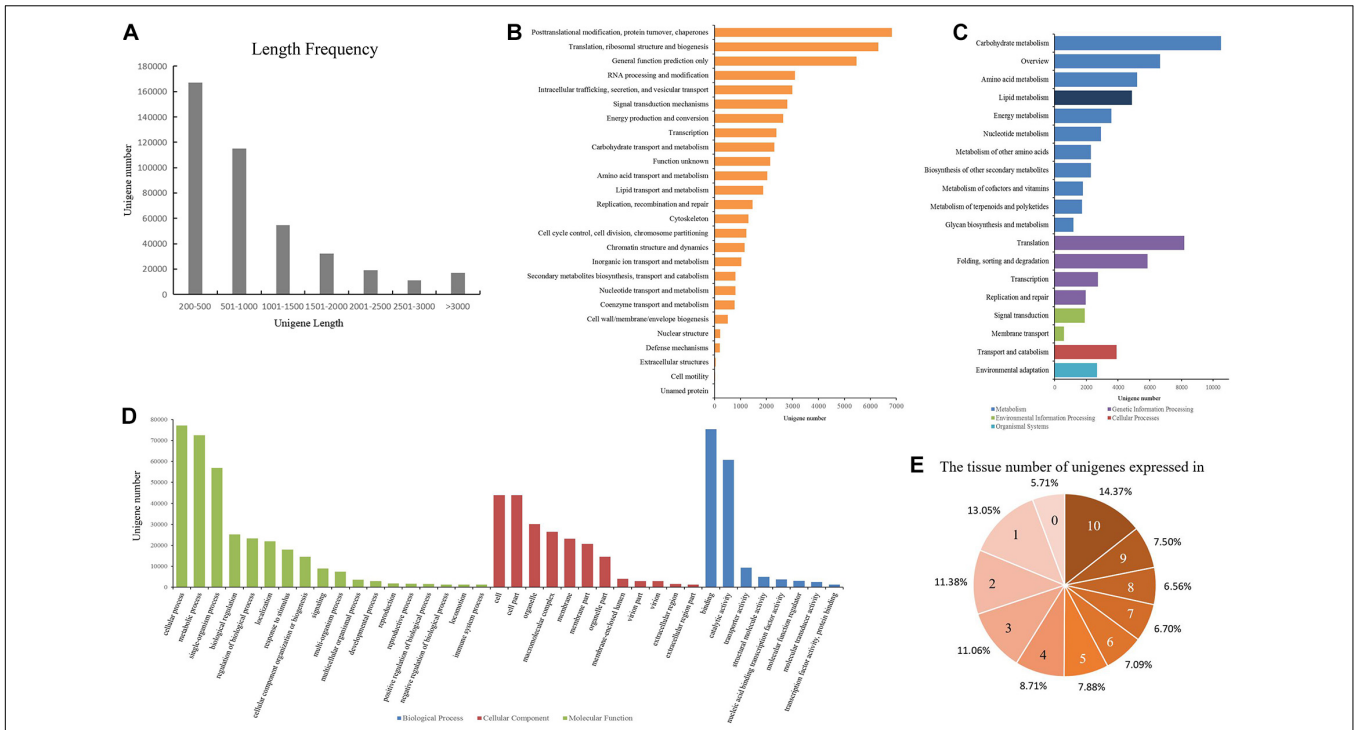


FIGURE 1 | Data quality and annotation of unigenes assembled with mRNA-seq. **(A)** Length distribution of unigenes. **(B)** Unigenes annotated using the Eukaryotic Orthologous Groups (KOG) tool. All unigenes were divided into 26 groups. **(C)** Functional annotation of the unigenes based on the KEGG classification. The numbers of unigenes mapped to each pathway group are shown in the bar chart. Colors indicate KEGG categories: blue for metabolism, purple for genetic information processing, green for environmental information processing, red for cellular processes, and cyan for organismal systems. **(D)** Gene ontology (GO) classification of the unigenes. Each bar represents the number of unigenes mapped to each GO category. Bar color indicates GO category: green for biological process (BP), red for cellular component (CC), and blue for molecular function (MF). **(E)** Number of the tissues in which each unigene is expressed.

to linolenic acid biosynthesis. These had 404 unigenes and 612 unigenes, respectively (**Table 2**).

GO is an international standardized functional gene classification that associates GO terms with other functional annotations of *de novo* assembly unigenes from homology searches. There were 142,163 unigenes found with at least one GO term, obtained using Blast2GO software (E-value ≤ 1E–6). The major categories in GO [biological process (BP), molecular function (MF), and cellular component (CC)] are represented in **Figure 1D**. Among the secondary entries, 292 unigenes, 25,177 unigenes and 72,426 unigenes were annotated as growth (GO:0040007), biological regulation (GO:0065007) and metabolic process (GO:0008152), respectively (**Supplementary Table 4**).

TABLE 1 | Annotation of unigenes based on multiple database resources.

Source	Number of unigenes	Percentage (%)	E-value
NR	197488	47.46%	1E–5
NT	129014	31%	1E–5
Pfam domains	139121	33.43%	0.01
Swiss-Prot	138800	33.36%	1E–5
KOG annotation	45348	10.89%	1E–3
GO annotation entries	142163	34.16%	1E–10
KEGG pathways	72433	17.40%	1E–6
Orthologous pairs in *Arabidopsis*	166486	40.01%	1E–5
Annotated in all databases	25291	6.07%	–
Annotated in at least one database	236222	56.77%	–
Total unigenes	416062	–	–

TABLE 2 | Unigenes related to lipid metabolism, based on KEGG classification.

KEGG Pathway	Gene Number
Biosynthesis of unsaturated fatty acids	481
Linoleic acid metabolism	161
Alpha-linolenic acid metabolism	652
Fatty acid biosynthesis	404
Fatty acid degradation	612
Fatty acid elongation	174
Glycerolipid metabolism	437
Glycerophospholipid metabolism	665
Sphingolipid metabolism	284
Steroid biosynthesis	333
Synthesis and degradation of ketone bodies	58
Arachidonic acid metabolism	115
Ether lipid metabolism	214
Cutin, suberine, and wax biosynthesis	290

Co-expression Network Construction

Approximately 56.77% of unigenes were annotated using public databases, (**Table 1**). To improve the functional annotation of the assembled unigenes and the functional genomics research for *Paeonia lactiflora* Pall. cv Shaoyou17C, we constructed a co-expression network. First, we checked the correlation of the replicates. RSEM was used to map the clean reads to the unigenes. In this mapping, about 69% to 85% of reads were mapped to the reference transcripts (**Supplementary Table 5**). From these mapping results, the FPKM value for each unigene in each sample was computed. Next, we calculated the correlation coefficient for each sample. Every three biological replicates in each sample had a high correlation (**Supplementary Figure 2**). The low correlation for the four seed stages could be due to the large differences between stages of seed growth. The expression patterns of the unigenes were checked, and then the distribution of expression for unigenes in different tissues was summarized. In all, 394,329 unigenes are expressed at least in one tissue and, 59,199 unigenes were expressed in all 10 tissues of Shaoyou17C (**Figure 1E**), and 28.43% were expressed in more than eight tissues. Half of the unigenes were detected in more than five tissues. These results indicate that most unigenes have good coverage and high quality.

Using transcriptome data, the PCC and MR (Umrania et al., 2017) algorithms were combined to construct a co-expression network. PCC and MR were computed to relate all pairs of unigenes based on a correlation algorithm (see detailed description in Methods). The PCC distributions (**Supplementary Figure 3A**) indicated that the co-expression network featured a similar normal distribution. To obtain the optimal co-expressed gene pairs, we set strict parameters. We found that the threshold of PCC had a cut-off of 0.7, and the MR top3 + MR $\leq$ 30, which were chosen to construct the co-expression network and ensured its coverage and creditability. There were 395,501 nodes in the constructed network, which covered 95.05% of unigenes in *Paeonia lactiflora* Pall. In all, there were 3,376,888 and 1,197,375 pairs in the positive co-expression network (PCC-value $>$ 0) and negative co-expression network (PCC-value $<$ 0), respectively. Whether it is in the positive or negative co-expression network, as the number of edges increases, the number of nodes decreases (**Supplementary Figures 3B,C**). The statistics for gene pairs suggest that positive and negative gene pairs were in accordance with the scale-free network.

Tissue-Preferential Analyses and Comparative Analyses Between Different Cultivars

We investigated unigene-expression patterns in different tissues. For Shaoyou17C, the 10 aforementioned tissues (LY, FY, KnY, OvY, StY, AnY, FRY, CpY, HsY and SY) were subjected to co-expression network analyses. The gene expression value was overlaid onto the co-expression network to estimate whether a given gene was expressed in the given tissue.

Chishao has a low seed setting rate but large seeds, Hangbaishao has small seeds but a high seed setting rate, and Shaoyou17C combines the advantages of both. Further, because differences between different cultivars could be caused by DEGs, the expression patterns were compared between different cultivars. We obtained transcriptome data from the leaf tissue of three cultivars, namely, Shaoyou17C, Chishao, and Hangbaishao, and from the flower tissues of Shaoyou17C and Chishao. Each tissue had three independent biological replicates. Following previous studies, Padj $<$ 0.05 and $|\log2FoldChange| > 1$ were chosen as thresholds. DEGs were identified between two different cultivars in leaf and flower. Among the leaf samples, there were 46,736 DEGs, 52,850 DEGs, and 56,414 DEGs (Padj $<$ 0.05 $|\log2FoldChange| > 1$) between Shaoyou17C and Chishao, Shaoyou17C and Hangbaishao, and Chishao and Hangbaishao, respectively. In the flower of Shaoyou17C and Chishao, 25,541 unigenes were DEGs (**Supplementary Table 6**). We also checked the number of DEGs under a stricter threshold (Padj $<$ 0.05 $|\log2FoldChange| > 2$). For a more intuitive presentation, red and blue were chosen to represent up-regulated and down-regulated genes, respectively. All of the DEGs were overlaid onto a co-expression network (**Supplementary Figure 4**). Users are able to check the DEGs between different cultivars. We also provided a user-friendly interface to show the expression patterns of co-expressed unigenes in different tissues and cultivars.

Construction of a Database With the Co-expression Network

Based on previous work, a database (HpeNet) was constructed for a co-expression network with functional analyses in *Paeonia lactiflora* Pall. This database was developed in the LAMP environment. It features three main parts: tissue-preferential expression view and co-expression network, including tissue-preferential and comparative analysis between different cultivars; functional annotation, including GO terms, the lipid pathway, KOG annotation, transcription factors, and Pfam domains, and so forth; and analysis tools for the co-expression network, including gene expression profiling analysis and gene set enrichment analysis (GSEA) (**Figure 2**). HpeNet can be accessed online at http://bioinformatics.cau.edu.cn/HpeNet/.

We took Cluster-55448.207014 as an example to illustrate the functions offered by HpeNet (**Figure 3A**). A gene detail search indicated that it was annotated as photosystem I light harvesting complex gene 1 (LHCA1). Then we explored possible functions of the unigene in the database. The annotation information found on the gene detail page shows that AT3G54890 (LHCA1) is the orthologous gene in *Arabidopsis thaliana*, and the sequence of Cluster-55448.207014 was presented in the database. GO annotation and the KEGG pathway were also available. This was annotated as "glucose catabolic process" (GO:0006007), "2,3-bisphosphoglycerate-independent phosphoglycerate mutase activity" (GO:0046537), "catalytic activity" (GO:0003824), and light-harvesting complex II chlorophyll a/b binding protein 1 (K08912). This suggests that Cluster-55448.207014 is involved in the glucose catabolic process and functions as a light-harvesting complex II a/b binding protein. The annotation information on the Pfam domain indicates that Cluster-55448.207014 has a PF10143 domain, which functions as 2,3-bisphosphoglycerate-independent phosphoglycerate mutase.

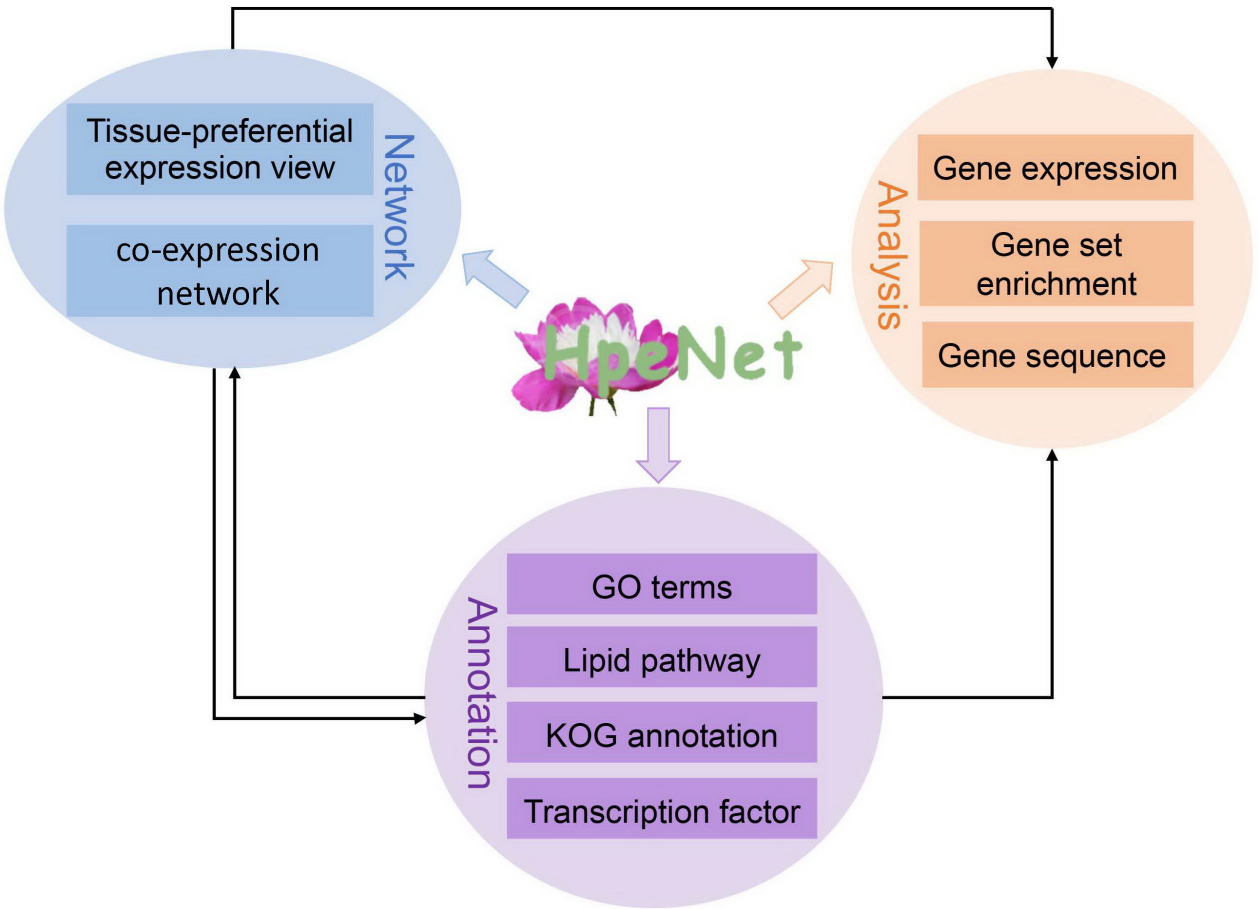


FIGURE 2 | Database architecture. The HpeNet database is divided into three main sections: network, annotations, and analytical tools. Lines with different colors indicate different pieces.

Through a network search, users can obtain co-expressed relationships and detailed information on network members to enable further analyses, including network analysis (tissue-preferential analysis and comparative analysis between different cultivars), gene expression profiling analysis and gene set enrichment analysis. Using the lipid-related pathway, users can search a pathway of interest and find detailed information on key enzymes and conduct network analyses. With the supported analytical tools, users can conduct BLAST alignment and gene set enrichment analysis for the gene list of interest. Furthermore, a database manual is provided.

CASE STUDY

Gene Network Analyses of Photosynthesis-Related Genes

Photosynthesis is the major source of energy for plant growth and development. In the higher plants, Photosystem I (PSI) and Photosystem II (PSII) harvest light in the thylakoid membrane. The peripheral light-harvesting complex of PSI (LHCI) is one of the two most important contents in PSI. It is composed of four LHCA complexes and is assembled into two heterodimers, LHCA1/4 and LHCA2/3. The study of the light-harvesting complex (LHC) genes of PSI and PSII in *Paeonia lactiflora* can help us understand its growth and development. Using orthologous genes, 175 LHC unigenes in *Paeonia lactiflora* were identified with 16 LHC unigenes in *Arabidopsis* (**Supplementary Table 7**). Cluster-55448.207014 and Cluster-55448.124893, both homologous to AT3G45890 (LHCA1), were selected, and their detailed information and co-expression networks were searched in the database (MR $\leq$ 50) (**Figures 3A,B**). In the KEGG pathway annotation, both of the unigenes were annotated with the light-harvesting complex KEGG term.

A co-expression relationship was found between Cluster-55448.207014 and Cluster-55448.124893 and the five other unigenes of the light harvesting complex, Cluster-55448.124892 (orthologous gene of LHCA3), Cluster-55448.145353 (orthologous gene of LHCA5), Cluster-55448.162485 (orthologous gene of LHCA6), Cluster-55448.152840 (orthologous gene of LHCB7), and Cluster-55448.200801 (orthologous gene of LHB1B1) were in the co-expression network (**Figure 3C**). To validate the networks and further investigate the function of the co-expressed genes, gene set enrichment

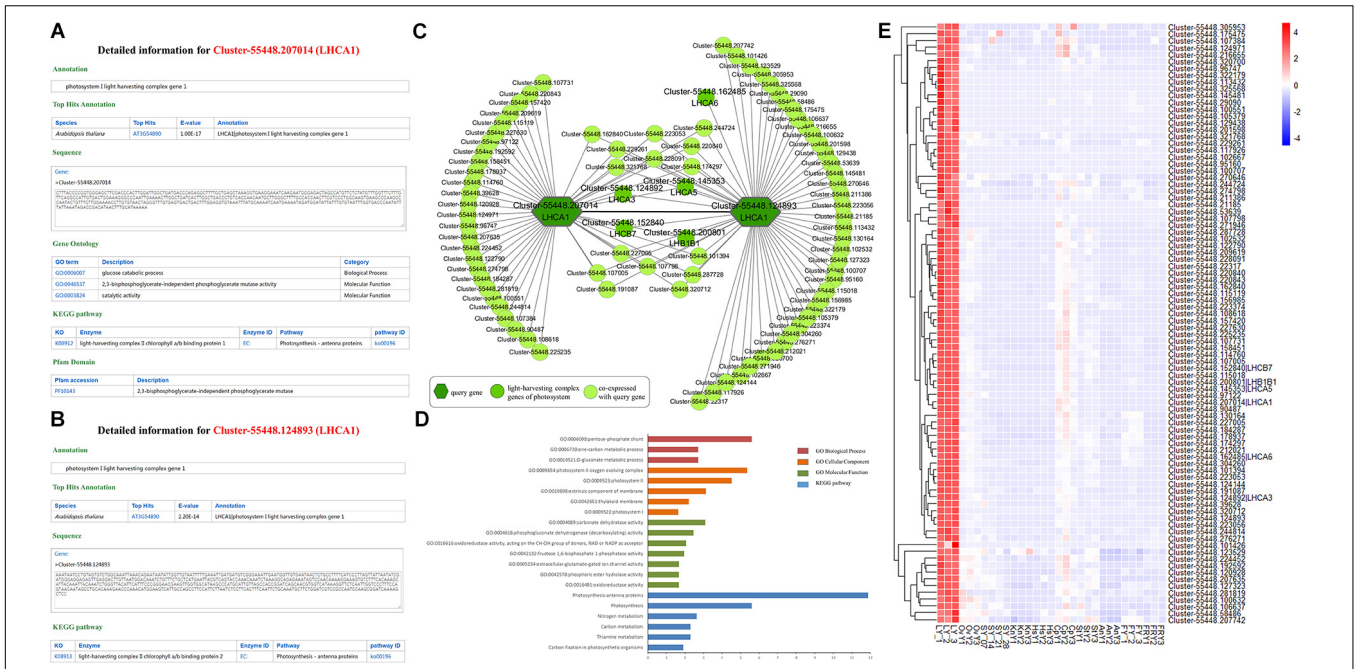


FIGURE 3 | Comprehensive explorations for the function of LHCA1 (Cluster-55448.207014 and Cluster-55448.124893). **(A)** and **(B)** Individual gene detail pages for Cluster-55448.207014 and Cluster-55448.124893, respectively. **(C)** Co-expression network for the LHCA1 genes (Cluster-55448.207014 and Cluster-55448.124893) in *Paeonia lactiflora*. The other light-harvesting complex genes of photosystem are highlighted with a dark green circle. **(D)** GSEA enrichment analyses of all genes from the co-expression network for LHCA1. **(E)** Heatmap of all genes from the co-expression network expressed in different samples.

analysis was conducted. The GO terms "photosystem I" (GO:0009522, FDR = 0.0237), "photosystem II oxygen evolving complex" (GO:0009654, FDR = 4.45E-06) and oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor" (GO:0016616, FDR = 8.75E-03) were significantly enriched. Additionally, the co-expression genes were also associated with the KEGG terms "Photosynthesis-antenna proteins" (FDR = 1.33E-12) and "Carbon fixation in photosynthetic organisms" (FDR = 0.0124) (**Figure 3D**). Consistent with the findings for *Arabidopsis* and other species, Cluster-55448.207014, Cluster-55448.124893, and their co-expressed genes were closely related to photosynthesis. Next, the expression patterns for Cluster-55448.207014, Cluster-55448.124893, and their co-expressed genes were checked (**Figure 3E**). As with LHCA1 in *Arabidopsis*, the unigenes were preferentially expressed in the leaf and hardly expressed at all in other tissues, particularly the fibrous root. This corresponds to the organization of photosynthesis. Some unannotated unigenes in the network may also be involved in photosynthesis.

Network Analyses for Genes Related to Oleate Desaturase

The seed oils were obtained through fatty acid biosynthesis, which includes synthesis, transport, and desaturation. The endoplasmic reticulum, plastids and cytosol were the main locations of reaction. (Bourgis et al., 2011; Li-Beisson et al., 2013). There are three types of enzymes in plant seeds that are primarily responsible for fatty acid composition: stearoyl-ACP desaturase, SAD (Klinkenberg et al., 2014), oleate desaturase, FAD2 (Okuley et al., 1994) and linoleate desaturase, FAD3 (Browse et al., 1993). They act in sequence to desaturate stearic acid (18:0) to oleic acid (18:1), to linoleic acid (18:2), and then to linolenic acid (18:3) (Miquel and Browse, 1992; Browse et al., 1993). In soybean, the largest oilseed crop, double mutants of the FAD2 and FAD3 genes result in high 18:1 and low 18:2 (Pham et al., 2012). Soybean contains 23% oleic acid, 54% linoleic acid, and only 8% linolenic acid (Wilson, 2004). Unlike that of oil-producing crops, the oil of *Paeonia lactiflora* Pall. cv Shaoyou17C contains a high proportion of PUFA. It has 35.28% oleic acid (18:1), 26.9% linoleic (18:2) acid, and 32.2% linolenic acid (18:3) content. This stimulated our great interest in investigating the differences between *Paeonia lactiflora* and other oilseed crops.

To investigate why Shaoyou17C seeds contain such a high proportion of linoleic acid and linolenic acid, the fatty acid synthesis pathway enzyme in Shaoyou17C was identified first through a prediction of the homologous genes in *Arabidopsis*. After genes with poor reproducibility were removed, we chose the median values of three repeats to represent the expression pattern for each tissue. A comparison of the expression pattern of fatty acid synthesis pathway genes in different tissues showed that an oleate desaturase, Cluster-55448.309693 (orthologous gene of FAD2), preferentially expressed in seed organ (**Supplementary Table 8**).

Detailed information on Cluster-55448.309693 was searched (MR $\leq$ 30) in the database. The tissue preferential network with expression of the unigenes in the 10 tissues is shown in **Figure 4**. The gray node was not expressed in the corresponding

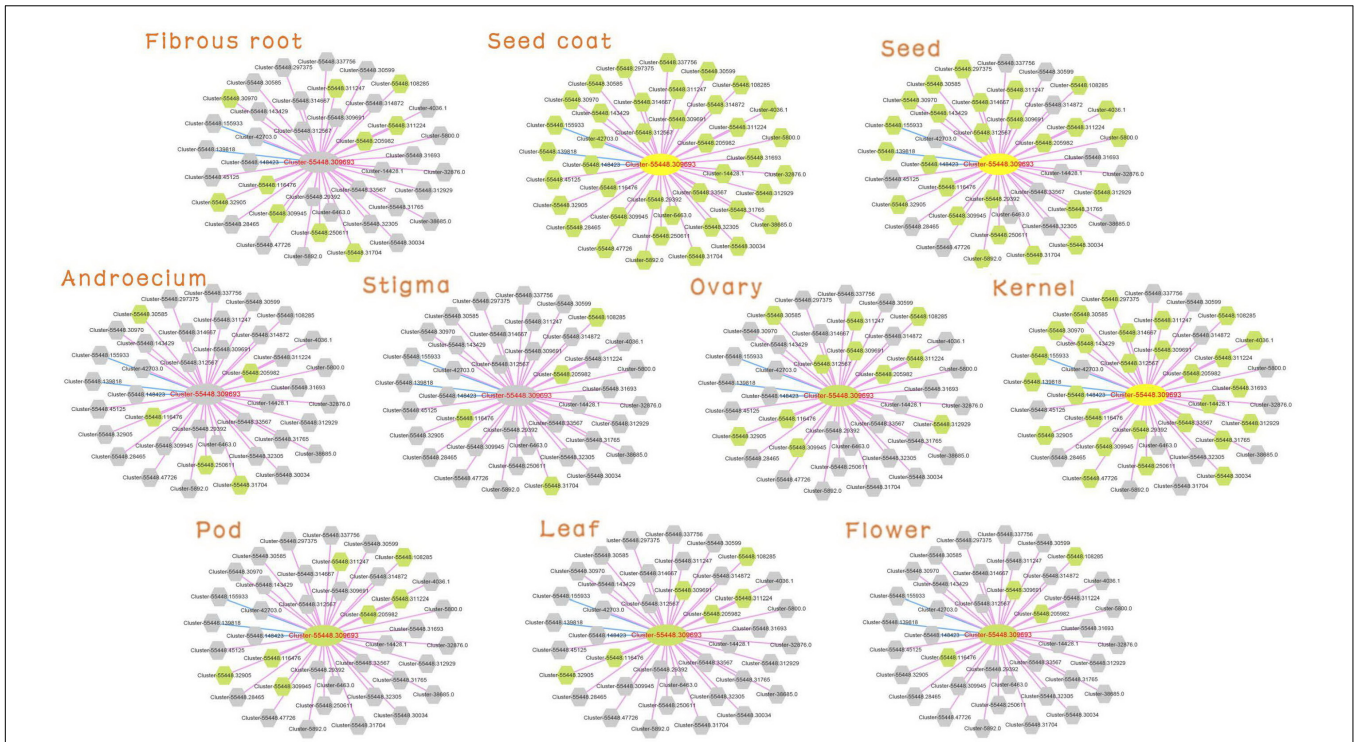


FIGURE 4 | Expression view of FAD2 network in different tissues. Query gene and co-expressed genes highlighted with yellow and green circles, respectively. Gray circles represent genes without an expression profile in that tissue.

tissue. Consistent with the expression characteristic of Cluster-55448.309693, the network in seed, seed coat and kernel had more expressed genes, and the network structure was relatively complete. Cluster-55448.309693 together with its co-expressed genes did not express in fibrous root, androecium, stigma.

The co-expressed genes of Cluster-55448.309693 contained two fatty acid desaturases, Cluster-55448.32905 and Cluster-55448.309691 (orthologous gene of FAD6), co-expressed with FAD2. Then the expression network of these three fatty acid desaturases was analyzed (**Supplementary Table 9**, **Figure 5**). In GSEA analyses, it was found that genes in the co-expression network were enriched with gene sets related to FA metabolism, such as the "lipid metabolic process" (GO:0006629, FDR = 0.0178), "oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor" (GO:0016655, FDR = 0.0142), "Fatty acid metabolism" (FDR = 6.82E-03) and "biosynthesis of unsaturated fatty acids" (FDR = 2.80E-03). These could help us study seed yield and the oil content of peony. Therefore, in addition to FAD2, FAD6 might play an important role in fatty acid desaturation in *Paeonia lactiflora* seeds.

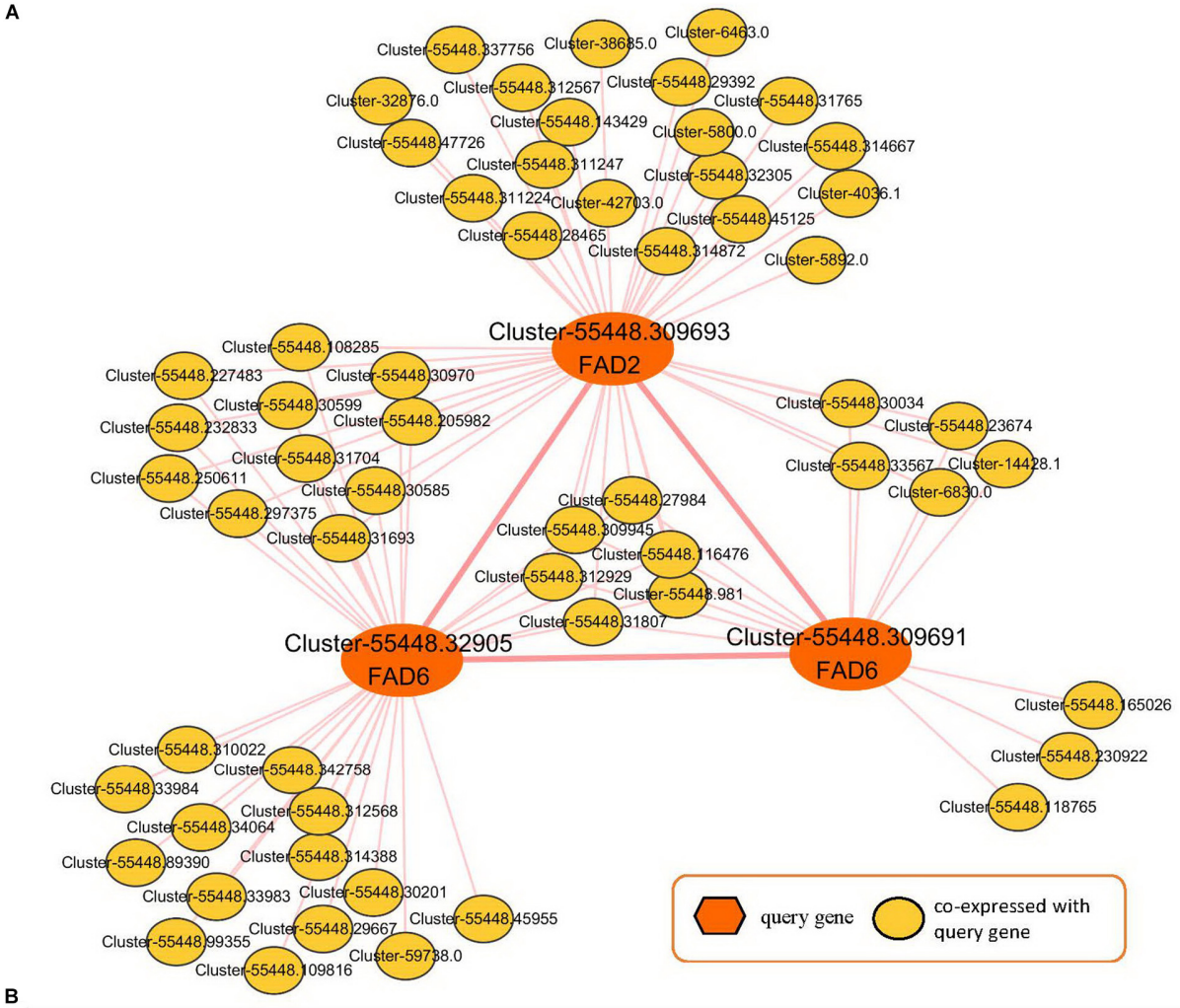
In summary, the transcriptome of peony was *de novo* assembled, with annotated and improved information on unigenes through the co-expression network. We developed an online database, HpeNet, that integrated transcriptome data, gene information, the co-expression network, and so forth. This made up for the lack of transcriptome data for *Paeonia lactiflora*. In addition, the database provided gene detail search, co-expressed network, lipid-related pathway gene inquiry, and several functional analytic tools that can help investigate gene function and molecular breeding.

DISCUSSION

We sequenced the transcriptomes of 10 tissues from different growth stages of *Paeonia lactiflora* Pall. cv Shaoyou17C. After the data were cleaned and quality control was implemented, a total of 2 billion 327 million reads with about 349.05 Gb transcriptome data were obtained. Then we *de novo* assembled the transcriptome of *Paeonia lactiflora* Pall. and as a result 416,062 unigenes were predicted. Furthermore, using public authoritative databases, 236,222 (approximately 57%) unigenes were annotated. In addition, a co-expression network was constructed to gather all available data into the online database HpeNet, which is easy for international users to access. The database also allows researcher to search by gene detail, gene function, and orthologous matches, as well as tools for co-expressed network analysis, lipid-related pathway gene inquiry, and several for functional analysis. We hope that our study will be beneficial to studies on *Paeonia lactiflora* Pall. and some related species in Ranales.

The genome of *Paeonia lactiflora* Pall. has not yet been decoded, and gene annotation is far from complete. The prohibitive costs associated with sequencing and assembling the genome make it infeasible to consider whole-genome sequencing in the near future as NGS technology evolves

A



B

Category	Description	FDR
GFam	Transcription_related, Transcription factor: MADS	3.45E-03
GO_BP	GO:0015743 malate transport	1.78E-02
GO_BP	GO:0006629 lipid metabolic process	1.78E-02
GO_MF	GO:0008508 bile acid:sodium symporter activity	3.15E-02
GO_MF	GO:0016655 oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor	1.42E-02
GO_MF	GO:0016717 oxidoreductase activity, acting on paired donors, with oxidation of a pair of donors resulting in the reduction of molecular oxygen to two molecules of water,	1.54E-03
KEGG	Fatty_acid_metabolism	6.82E-03
KEGG	Biosynthesis_of_unsaturated_fatty_acids	2.80E-03

FIGURE 5 | Co-expression network analyses of *Paeonia lactiflora* fatty acid desaturase genes. **(A)** Co-expression network of FAD2 (Cluster-55448.309693) and FAD6 (Cluster-55448.32905 and Cluster-55448.309691). **(B)** GSEA enrichment analyses for all genes from **(A)**. The results shown list the category, description, and FDR value.

(Goodwin et al., 2016). Transcriptome analysis is an important way in which gene discovery can go forward, and the transcribed sequences contain a high content of functional information. To build a comprehensive transcriptome, we chose 10 representative tissues with three biological replicates at different stages of *Paeonia lactiflora* Pall. cv Shaoyou17C. To the best of our knowledge, this is the most comprehensive measurement of the *Paeonia lactiflora* transcriptome, and it covers various tissues of *Paeonia lactiflora*. High-quality sequences of unigenes were obtained through *de novo* assembly. This approach

goes beyond supporting research on the molecular biology of *Paeonia lactiflora*, including gene cloning and homologous and phylogenetic analysis, and maybe beneficial to decoding its entire genome.

Co-expression network is an optimal instrument to use in exploring gene function. Many co-expression network databases have been constructed for model plants and important crops, such as ATTED-II (Obayashi et al., 2018), PlaNet (Proost and Mutwil, 2017), AraNet (Lee and Lee, 2017), RiceNet (Lee et al., 2015), RiceFREND (Sato et al., 2013), ccNET (You et al., 2017), MCENet (Tian et al., 2018) and WheatNet (Lee et al., 2017), but there are few co-expression databases for no-reference genome plants such as *Paeonia lactiflora* Pall. Annotation could help develop analyses of the function of unigenes and molecular breeding for *Paeonia lactiflora*.

Using the co-expression network, we found that LHCA1, Cluster-55448.207014, and Cluster-55448.124893 were co-expressed with five other genes of the light-harvesting complex, namely, LHCA3, LHCA5, LHCA6, LHCB7 and LHB1B1. In addition, Cluster-55448.309693 (orthologous gene of FAD2), an important fatty acid desaturase, was co-expressed with two other fatty acid desaturases, Cluster-55448.32905 and Cluster-55448.309691 (orthologous gene of FAD6). These results show that the network is reliable. Taking Cluster-55448.309693 as an example, comparative analyses between different cultivars indicated that Cluster-55448.309693 had a similar network structure in different cultivars, and fewer than three DEGs existed in the network. However, as we are limited by the data, only a few DEGs were noted in leaves and flowers. It is noteworthy that the network also contains a large number of unannotated unigenes. Using the database, it is possible to query the expression changes in co-expressed genes of different cultivars.

To facilitate the development of this information, we constructed HpeNet, a database with unigene information, a co-expression network, and analytical tools. We have not yet found any databases related to *Paeonia lactiflora* or even to any Ranales species. HpeNet may be the first online database for a Ranales species that offers gene information and co-expression networks. Using this database, users can query unigene information, including predicted annotation information, sequence information, and expression information for different tissues. We annotated 236,222 unigenes using seven public and authoritative databases. It is worth noting that about 57% of unigenes were annotated from public authoritative databases, a rate that is similar to that of many other *de novo* assembly plant transcriptomes, such as *Cuscuta pentagona* (Ranjan et al., 2014), *Lens culinaris* Medik. (Verma et al., 2013), and *Arceuthobium sichuanense* (Wang et al., 2016). This indicates that unigene annotation should be improved.

This study had limitations and it could be improved. Although 10 organs from the series growth stages of *Paeonia lactiflora* Pall. were included, some were excluded. This necessarily prevented us from being able to predict all genes. *De novo* assembly circumvents the need for a reference genome, but this approach has some deficiencies. Investigations have found that *de novo* assembly often generates fragments of a transcript or multiple contigs (Davidson and Oshlack, 2014). Furthermore, we obtained a large number of unigenes, although we employed Corset to merge the results and hierarchically cluster the contigs. Additionally, genes that share sequences, sequencing errors, variations and other issues can increase the number of unigenes (Vijay et al., 2013). Transcriptome data on seeds at different time points of development and on different tissues from flowers and seeds can be used for more detailed analyses. The reasons for the high proportion of unsaturated fatty acids, the key genes that regulate flowering, and the important genes that control the development of oil seed and other key functions. all require additional research.

In summary, through the transcriptomes of 10 tissues at different growth stages of *Paeonia lactiflora* Pall. cv Shaoyou17C, we obtained a comprehensive transcriptome with 416,062 unigenes. In addition, key insights were obtained from them through transcriptome profiling. Utilizing both sequence homology and co-expression networks, the functional annotation of the unigenes was refined and the functional genomic understanding of *Paeonia lactiflora* Pall. was developed. Further, HpeNet provides an online platform with transcriptome data, gene information, co-expression networks, and analytical tools. It will be convenient for researchers to obtain this information. This study improves our understanding of *Paeonia lactiflora* Pall. and provides a feasible way to study species without a reference genome.

DATA AVAILABILITY STATEMENT

The RNA-seq datasets generated for this study have been deposited in the Sequence Read Archive (www.ncbi.nlm.nih.gov/sra) under the accession number PRJNA656429.

AUTHOR CONTRIBUTIONS

XZ, WX, and ZS designed this project. XZ and HY generated the transcriptome data and annotations. MS calculated the co-expression network. MS and JS completed the database construction and draft. All authors read and approved the final manuscript.

FUNDING

This work was financially supported by the Project for Interdisciplinary Research for Science and Technology Star (Z171100001117123, Beijing Science and Technology Committee), grant of the National Natural Science Foundation of China (31371291), Tianjie Shanghe Agri & Forest Sci Co. Ltd., and Shanghe Mudan Agri Development Co. Ltd.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.570138/full#supplementary-material

REFERENCES

Altschul, S. F., Madden, T. L., Schaffer, 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

Aoki, Y., Okamura, Y., Tadaka, S., Kinoshita, K., and Obayashi, T. (2016). ATTED-II in 2016: a plant coexpression database towards lineage-specific coexpression. *Plant Cell Physiol.* 57:e5. doi: 10.1093/pcp/pcv165

Bourgis, F., Kilaru, A., Cao, X., Ngando-Ebongue, G. F., Drira, N., Ohlrogge, J. B., et al. (2011). Comparative transcriptome and metabolite analysis of oil palm and date palm mesocarp that differ dramatically in carbon partitioning. *Proc. Natl. Acad. Sci. U.S.A.* 108, 12527–12532. doi: 10.1073/pnas.1106502108

Browse, J., McConn, M., James, D. Jr., and Miquel, M. (1993). Mutants of Arabidopsis deficient in the synthesis of alpha-linolenate. biochemical and genetic characterization of the endoplasmic reticulum linoleoyl desaturase. *J. Biol. Chem.* 268, 16345–16351.

Davidson, N. M., and Oshlack, A. (2014). Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. *Genome Biol.* 15:410. doi: 10.1186/s13059-014-0410-416

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.* 95, 14863–14868. doi: 10.1073/pnas.95.25.14863

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H. R., et al. (2008). The Pfam protein families database. *Nucleic Acids Res.* 36, D281–D288. doi: 10.1093/nar/gkm960

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., et al. (2010). The Pfam protein families database. *Nucleic Acids Res.* 38, D211–D222. doi: 10.1093/nar/gkp985

Franz, M., Lopes, C. T., Huck, G., Dong, Y., Sumer, O., and Bader, G. D. (2016). Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics* 32, 309–311. doi: 10.1093/bioinformatics/btv557

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–351. doi: 10.1038/nrg.2016.49

Gotz, S., Garcia-Gomez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., et al. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* 36, 3420–3435. doi: 10.1093/nar/gkn176

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–U130. doi: 10.1038/nbt.1883

Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., et al. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* 28, 503–510. doi: 10.1038/nbt.1633

Hao, Z. J., Liu, D., Gong, S. J., Zhao, D. Q., and Tao, J. (2017). High throughput sequencing of herbaceous peony small RNAs to screen thermo-tolerance related microRNAs. *Genes Genom.* 39, 397–408. doi: 10.1007/s13258-016-0505-504

Hao, Z. J., Wei, M. R., Gong, S. J., Zhao, D. Q., and Tao, J. (2016). Transcriptome and digital gene expression analysis of herbaceous peony (Paeonia lactiflora Pall.) to screen thermo-tolerant related differently expressed genes. *Genes Genom.* 38, 1201–1215. doi: 10.1007/s13258-016-0465-468

Inoue, M., and Horimoto, K. (2017). Relationship between regulatory pattern of gene expression level and gene function. *PLoS One* 12:e0177430. doi: 10.1371/journal.pone.0177430

Klinkenberg, J., Faist, H., Saupe, S., Lambertz, S., Krischke, M., Stingl, N., et al. (2014). Two fatty acid desaturases, stearoyl-acyl carrier protein delta9-desaturase6 and fatty acid desaturase3, are involved in drought and hypoxia stress signaling in Arabidopsis crown galls. *Plant Physiol.* 164, 570–583. doi: 10.1104/pp.113.230326

Lee, T., Hwang, S., Kim, C. Y., Shim, H., Kim, H., Ronald, P. C., et al. (2017). WheatNet: a genome-scale functional network for hexaploid bread wheat. *Triticum aestivum*. *Mol. Plant* 10, 1133–1136. doi: 10.1016/j.molp.2017.04.006

Lee, T., and Lee, I. (2017). AraNet: a network biology server for arabidopsis thaliana and other non-model plant species. *Methods Mol. Biol.* 1629, 225–238. doi: 10.1007/978-1-4939-7125-1_15

Lee, T., Oh, T., Yang, S., Shin, J., Hwang, S., Kim, C. Y., et al. (2015). RiceNet v2: an improved network prioritization server for rice genes. *Nucleic Acids Res.* 43, W122–W127. doi: 10.1093/nar/gkv253

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

Li-Beisson, Y., Shorrosh, B., Beisson, F., Andersson, M. X., Arondel, V., Bates, P. D., et al. (2013). Acyl-lipid metabolism. *Arabidopsis Book* 11:e0161. doi: 10.1199/tab.0161

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/s13059-014-0550-558

Ma, X., Zhao, H., Xu, W., You, Q., Yan, H., Gao, Z., et al. (2018). Co-expression gene network analysis and functional module identification in bamboo growth and development. *Front. Genet.* 9:574. doi: 10.3389/fgene.2018.00574

Miquel, M., and Browse, J. (1992). Arabidopsis mutants deficient in polyunsaturated fatty acid synthesis. biochemical and genetic characterization of a plant oleoyl-phosphatidylcholine desaturase. *J. Biol. Chem.* 267, 1502–1509.

Moretto, M., Sonego, P., Villasenor-Altamirano, A. B., and Engelen, K. (2019). First step toward gene expression data integration: transcriptomic data acquisition with COMMAND>_. *BMC Bioinform* 20:54. doi: 10.1186/s12859-019-2643-2646

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* 35, W182–W185. doi: 10.1093/nar/gkm321

Mozaffarian, D., Micha, R., and Wallace, S. (2010). Effects on coronary heart disease of increasing polyunsaturated fat in place of saturated fat: a systematic review and meta-analysis of randomized controlled trials. *PLoS Med.* 7:e1000252. doi: 10.1371/journal.pmed.1000252

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.* 59:440. doi: 10.1093/pcp/pcx191

Obayashi, T., and Kinoshita, K. (2009). Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. *DNA Res.* 16, 249–260. doi: 10.1093/dnares/dsp016

Okuley, J., Lightner, J., Feldmann, K., Yadav, N., Lark, E., and Browse, J. (1994). Arabidopsis FAD2 gene encodes the enzyme that is essential for polyunsaturated lipid synthesis. *Plant Cell* 6, 147–158. doi: 10.1105/tpc.6.1.147

Pham, A. T., Shannon, J. G., and Bilyeu, K. D. (2012). Combinations of mutant FAD2 and FAD3 genes to produce high oleic acid and low linolenic acid soybean oil. *Theor. Appl. Genet.* 125, 503–515. doi: 10.1007/s00122-012-1849-z

Proost, S., and Mutwil, M. (2017). PlaNet: comparative co-expression network analyses for plants. *Methods Mol. Biol.* 1533, 213–227. doi: 10.1007/978-1-4939-6658-5_12

Ranjan, A., Ichihashi, Y., Farhi, M., Zumstein, K., Townsley, B., David-Schwartz, R., et al. (2014). De novo assembly and characterization of the transcriptome of the parasitic weed dodder identifies genes associated with plant parasitism. *Plant Physiol.* 166, 1186–1199. doi: 10.1104/pp.113.234864

Rhee, S. Y., and Mutwil, M. (2014). Towards revealing the functions of all genes in plants. *Trends Plant Sci.* 19, 212–221. doi: 10.1016/j.tplants.2013.10.006

Rizos, E. C., Ntzani, E. E., Bika, E., Kostapanos, M. S., and Elisaf, M. S. (2012). Association between omega-3 fatty acid supplementation and risk of major cardiovascular disease events a systematic review and meta-analysis. *JAMA-J. Am. Med. Assoc.* 308, 1024–1033. doi: 10.1001/2012.jama.11374

Salleh, S. M., Mazzoni, G., Lovendahl, P., and Kadarmideen, H. N. (2018). Gene co-expression networks from RNA sequencing of dairy cattle identifies genes and pathways affecting feed efficiency. *BMC Bioinform.* 19:513. doi: 10.1186/s12859-018-2553-z

Sato, Y., Namiki, N., Takehisa, H., Kamatsuki, K., Minami, H., Ikawa, H., et al. (2013). RiceFREND: a platform for retrieving coexpressed gene networks in rice. *Nucleic Acids Res.* 41, D1214–D1221. doi: 10.1093/nar/gks1122

She, J., Yan, H., Yang, J., Xu, W., and Su, Z. (2019). croFGD: catharanthus roseus functional genomics database. *Front. Genet.* 10:238. doi: 10.3389/fgene.2019.00238

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. Genom.* 45, 351–360. doi: 10.1016/j.jgg.2018.05.007

Umrania, V. V., Rao Deepika, P. C., and Kulkarni, M. (2017). Evaluation of dietary supplementation of omega-3 polyunsaturated fatty acids as an adjunct

to scaling and root planning on salivary interleukin-1beta levels in patients with chronic periodontitis: a clinico-immunological study. *J. Ind. Soc. Periodontol.* 21, 386–390. doi: 10.4103/jisp.jisp_16_16

van Dam, S., Vosa, U., van der Graaf, A., Franke, L., and de Magalhaes, J. P. (2018). Gene co-expression analysis for functional classification and gene-disease predictions. *Brief Bioinform.* 19, 575–592. doi: 10.1093/bib/bbw139

Verma, P., Shah, N., and Bhatia, S. (2013). Development of an expressed gene catalogue and molecular markers from the de novo assembly of short sequence reads of the lentil (Lens culinaris Medik.) *transcriptome*. *Plant Biotechnol. J.* 11, 894–905. doi: 10.1111/pbi.12082

Vijay, N., Poelstra, J. W., Kunstner, A., and Wolf, J. B. (2013). Challenges and strategies in transcriptome assembly and differential gene expression quantification. a comprehensive in silico assessment of RNA-seq experiments. *Mol. Ecol.* 22, 620–634. doi: 10.1111/mec.12014

Walker, M. G., Volkmuth, W., Sprinzak, E., Hodgson, D., and Klingler, T. (1999). Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. *Genome Res.* 9, 1198–1203. doi: 10.1101/gr.9.12.1198

Wang, W. P., Wang, C. J., Gu, S., Cao, Q., Lv, Y. P., Gao, J. J., et al. (2009). Pharmacokinetic studies of the significance of herbaceous compatibility of peony liquorice decoction. *Mode Tradit Chin. Med. Mater. Med.* 11, 382–387. doi: 10.1016/s1876-3553(10)60020-8

Wang, Y., Li, X., Zhou, W., Li, T., and Tian, C. (2016). De novo assembly and transcriptome characterization of spruce dwarf mistletoe *Arceuthobium sichuanense* uncovers gene expression profiling associated with plant development. *BMC Genom.* 17:771. doi: 10.1186/s12864-016-3127-y

Wilson, R. (2004). "Seed composition," in *Soybeans: Improvement, Production, and Uses*, eds H. R. Boerma and J. E. Specht (Madison, WI: ASA), 621–677.

Wisecaver, J. H., Borowsky, A. T., Tzin, V., Jander, G., Kliebenstein, D. J., and Rokas, A. (2017). A global coexpression network approach for connecting genes to specialized metabolic pathways in plants. *Plant Cell* 29, 944–959. doi: 10.1105/tpc.17.00009

Wu, Y., Tang, Y., Jiang, Y., Zhao, D., Shang, J., and Tao, J. (2018). Combination of transcriptome sequencing and iTRAQ proteome reveals the molecular mechanisms determining petal shape in herbaceous peony (Paeonia lactiflora Pall.). *Biosci. Rep.* 38:BSR20181485. doi: 10.1042/BSR20181485

Yang, Y., Li, S. S., Teixeira da Silva, J. A., Yu, X. N., and Wang, L. S. (2020). Characterization of phytochemicals in the roots of wild herbaceous peonies from China and screening for medicinal resources. *Phytochemistry* 174:112331. doi: 10.1016/j.phytochem.2020.112331

Yi, X., Du, Z., and Su, Z. (2013). PlantGSEA: a gene set enrichment analysis toolkit for plant community. *Nucleic Acids Res.* 41, W98–W103. doi: 10.1093/nar/gkt281

You, Q., Xu, W., Zhang, K., Zhang, L., Yi, X., Yao, D., et al. (2017). ccNET: database of co-expression networks with functional modules for diploid and polyploid *Gossypium*. *Nucleic Acids Res.* 45, D1090–D1099. doi: 10.1093/nar/gkw910

You, Q., Zhang, L., Yi, X., Zhang, Z., Xu, W., and Su, Z. (2015). SIFGD: setaria italica functional genomics database. *Mol. Plant* 8, 967–970. doi: 10.1016/j.molp.2015.02.001

Zhao, D., Jiang, Y., Ning, C., Meng, J., Lin, S., Ding, W., et al. (2014). Transcriptome sequencing of a chimaera reveals coordinated expression of anthocyanin biosynthetic genes mediating yellow formation in herbaceous peony (Paeonia lactiflora Pall.). *BMC Genom.* 15:689. doi: 10.1186/1471-2164-15-689

Zhao, D. Q., Tao, J., Han, C. X., and Ge, J. T. (2012). Flower color diversity revealed by differential expression of flavonoid biosynthetic genes and flavonoid accumulation in herbaceous peony (Paeonia lactiflora Pall.). *Mol. Biol. Rep.* 39, 11263–11275. doi: 10.1007/s11033-012-2036-2037

Zhao, D. Q., Wei, M. R., Liu, D., and Tao, J. (2016). Anatomical and biochemical analysis reveal the role of anthocyanins in flower coloration of herbaceous peony. *Plant Physiol. Biochem.* 102, 97–106. doi: 10.1016/j.plaphy.2016.02.023

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.

Copyright © 2020 Sheng, She, Xu, Hong, Su and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 22 October 2020
doi: 10.3389/fgene.2020.574962

Characterization of Chloroplast Genomes From Two *Salvia* Medicinal Plants and Gene Transfer Among Their Mitochondrial and Chloroplast Genomes

Chengwen Gao, Chuanhong Wu, Qian Zhang, Xia Zhao, Mingxuan Wu, Ruirui Chen, Yalin Zhao and Zhiqiang Li**

Laboratory of Medical Biology, Medical Research Center, The Affiliated Hospital of Qingdao University, Qingdao, China

Salvia species have been widely used as medicinal plants and have played an important role in the treatment and recovery of individuals with COVID-19. In this study, we reported two newly identified whole chloroplast genome sequences of *Salvia* medicinal plants (*Salvia yangii* and *Salvia miltiorrhiza* f. *alba*) and compared them with those of seven other reported *Salvia* chloroplast genomes. These were proven to be highly similar in terms of overall size, genome structure, gene content, and gene order. We identified 10 mutation hot spots (*trnK-rps16*, *atpH-atpI*, *psaA-ycf3*, *ndhC-trnV*, *ndhF*, *rpl32-trnL*, *ndhG-ndhI*, *rps15-ycf1*, *ycf1a*, and *ycf1b*) as candidate DNA barcodes for *Salvia*. Additionally, we observed the transfer of nine large-sized chloroplast genome fragments, with a total size of 49,895 bp (accounting for 32.97% of the chloroplast genome), into the mitochondrial genome as they shared >97% sequence similarity. Phylogenetic analyses of the whole chloroplast genome provided a high resolution of *Salvia*. This study will pave the way for the identification and breeding of *Salvia* medicinal plants and further phylogenetic evolutionary research on them as well.

Keywords: *Salvia*, chloroplast genome, DNA barcodes, gene transfer, phylogenetic

OPEN ACCESS

Edited by:
Rong Zhou,
Aarhus University, Denmark

Reviewed by:
Cheng Qin,
Zunyi Vocational and Technical College, China
Rajesh Kumar Gazara,
Indian Institute of Technology Roorkee, India

****Correspondence:***
Chengwen Gao
gaochengwen6@126.com
Zhiqiang Li
lizqsjtu@163.com

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

Received: *22 June 2020*
Accepted: *23 September 2020*
Published: *22 October 2020*

Citation:
Gao C, Wu C, Zhang Q, Zhao X, Wu M, Chen R, Zhao Y and Li Z (2020) Characterization of Chloroplast Genomes From Two Salvia Medicinal Plants and Gene Transfer Among Their Mitochondrial and Chloroplast Genomes. Front. Genet. 11:574962. doi: 10.3389/fgene.2020.574962

INTRODUCTION

Salvia L. species constitute the largest genus in the family Lamiaceae and comprise approximately 1,000 species worldwide. Of these, 84 *Salvia* species originated from southwestern China, especially from the Hengduan Mountains region (Li et al., 2013). The name *Salvia* comes from the Latin word "salvare," which means "healing." *Salvia* species have been widely used as medicinal plants, and as an important component of the Xuebijing injection, they have played an important role in the treatment and recovery of individuals with COVID-19 (Pan et al., 2020; Wen et al., 2020; Zhang Y.L. et al., 2020). For thousands of years in China, more than 40 *Salvia* species have been used for treating common cold, tuberculosis, bronchitis, hemorrhages, and menstrual disorders (Topcu, 2006). Among these, *Salvia miltiorrhiza* Bunge (Danshen), a characteristic traditional Chinese medicine (TCM) (Li et al., 2018), has been used for treating cerebrovascular and cardiovascular diseases effectively (Chen and Chen, 2017; Wang et al., 2017). However, compared with the purple-flowered *S. miltiorrhiza* variety, the white-flowered landrace of *Salvia miltiorrhiza* f. *alba* exerts

superior medicinal qualities (Zhang et al., 2016). *Salvia yangii* also possesses high medicinal value (Mazandarani and Ghaemi, 2010). However, the difficulty in identifying these plants leads to an uneven quality of *Salvia* medicinal materials (Hu G.X. et al., 2016; Xiang et al., 2016; Drew et al., 2017). With the development of sequencing technology, the sequencing of chloroplast genomes can be fulfilled in a short time at a relatively low cost currently. Whole chloroplast genome sequences can provide more genetic information and higher species resolution ability than other molecular data, and they provide a possible resolution to these issues. However, the chloroplast genomes of most *Salvia* plants remain unknown.

Chloroplasts are specialized organelles where photosynthesis occurs and are composed of stacked thylakoids interconnected by lamellae (Douglas, 1998; Brunkard et al., 2015). The highly conserved chloroplast genomes possess a quadripartite structure including small single-copy (SSC) and large single-copy (LSC) regions along with dual copies of inverted repeat (IR) regions (Wicke et al., 2011; Wang et al., 2015). Whole chloroplast genomes provide crucial information, which can be used for understanding evolutionary differences, exploring genetic diversity, and generating highly resolved phylogenies, especially in complex/low taxonomic levels (Hu H. et al., 2016; Xu et al., 2019; Huang et al., 2020).

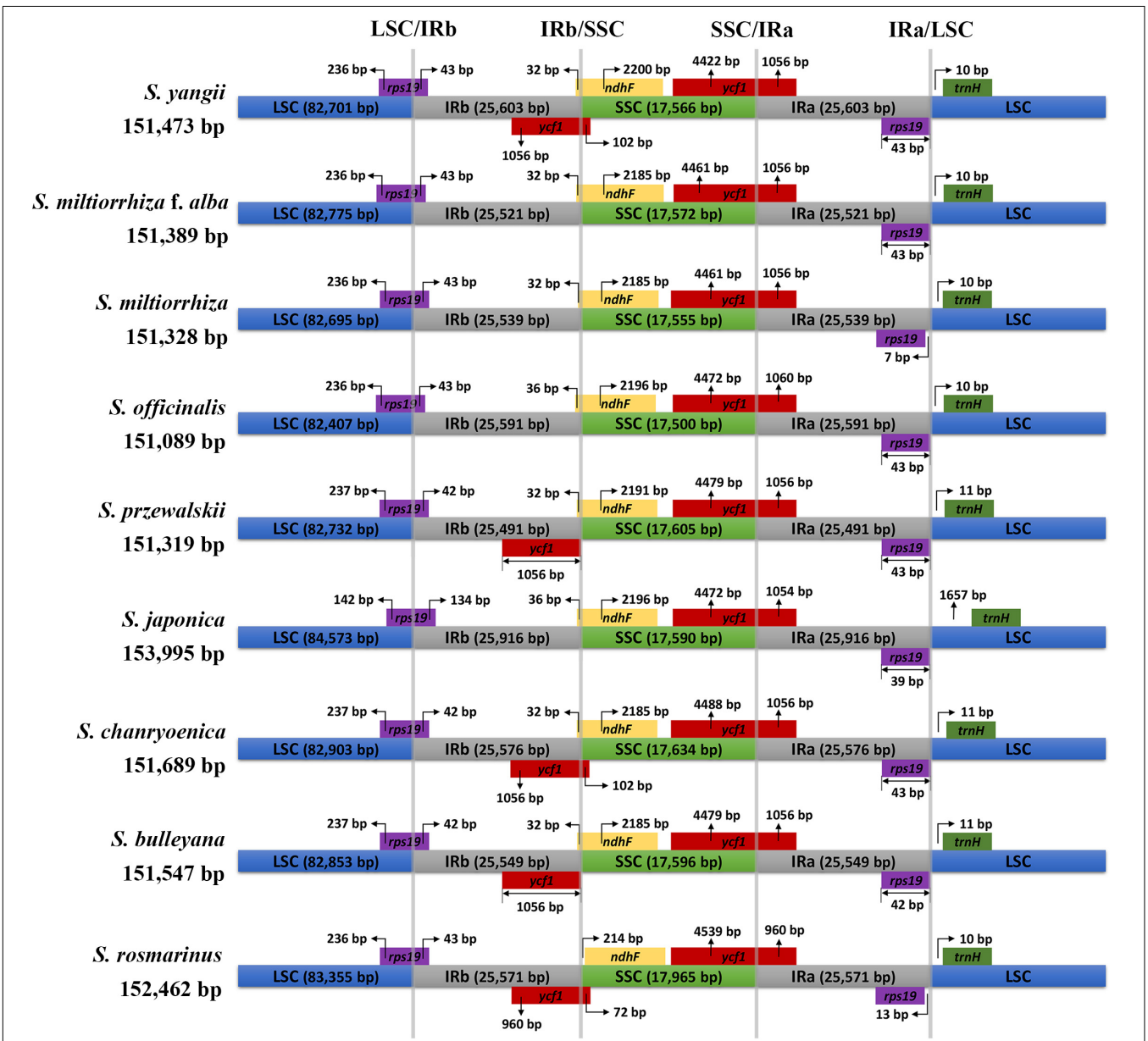


FIGURE 1 | Comparison of inverted repeat (IR)/small single-copy (SSC) expansions in *Salvia* species. *Colored boxes above the horizontal line* indicate the genes and gene segments across the IRa/b junctions. IR segments and genes are not to scale.

TABLE 1 | Summary of the features of the chloroplast genome of nine *Salvia* species.

Species	Genome size (bp)	LSC length (bp)	IR length (bp)	SSC length (bp)	GenBank number	GC%	Number of genes			
							Total	CDS	rRNA	tRNA
S. yangii	151,473	82,701	25,603	17,566	MT012421	38.1	116	82	4	30
S. miltiorrhiza f. *alba*	151,389	82,775	25,521	17,572	MT012420	38.0	115	81	4	30
S. miltiorrhiza	151,328	82,695	25,539	17,555	NC_020431	38.0	115	81	4	30
S. officinalis	151,089	82,407	25,591	17,500	NC_038165	38.0	115	81	4	30
S. przewalskii	151,319	82,732	25,491	17,605	NC_041091	38.0	115	81	4	30
S. japonica	153,995	84,573	25,916	17,590	NC_035233	38.0	115	81	4	30
S. chanryoenica	151,689	82,903	25,576	17,634	NC_040121	38.0	115	81	4	30
S. bulleyana	151,547	82,853	25,549	17,596	NC_041092	38.0	115	81	4	30
S. rosmarinus	152,462	83,355	25,571	17,965	NC_027259	38.0	115	81	4	30

Here, we assembled the chloroplast genomes of *S. miltiorrhiza* f. *alba* and *S. yangii*, followed by comparing them with seven previously reported *Salvia* chloroplast genomes from GenBank. This study aimed to (1) perform a comparative analysis of the chloroplast genomes of these nine *Salvia* species; (2) ascertain the highly divergent sequences in the *Salvia* chloroplast genomes; (3) identify chloroplast gene insertion in the mitochondria; and (4) explore the evolutionary differences and similarities in the genus *Salvia* and lamiids. Thus, the information generated in this study would expand our knowledge on the evolution of plastome as well as the phylogenies of *Salvia* species.

MATERIALS AND METHODS

Plant Material Sampling

The *Salvia* accessions in this study were from the *Salvia* nuclear genome sequencing project, and all plant materials were conserved in China Academy of Chinese Medical Sciences, Beijing, China (Xu et al., 2016; Boachon et al., 2018). The DNA was extracted from fresh leaves, followed by the use of Illumina sequencing technology to generate libraries. Post-sequencing, the paired-end sequence reads were obtained through next-generation sequencing.

Chloroplast Genome Assembly and Annotation

Trimmomatic v0.38 was used to filter the raw sequencing data (Bolger et al., 2014). N-containing sequences and adapter sequences were removed. Sequences with a *Q* value less than 20 were also removed. Then, the clean data were used to perform *de novo* assembly using SPAdes 3.61 with varying *K*-mer parameters (Bankevich et al., 2012). We ordered *de novo* scaffolds that were positively correlated to the chloroplasts on the reference chloroplast genome of *S. miltiorrhiza* (NC_020431). Next, the Geneious Prime software v2020.0.4 (Kearse et al., 2012) was used to remap the paired-end reads to fill gaps in the final consensus sequence with multiple iterations.

GeSeq was used to conduct chloroplast genome annotation to predict transfer RNA (tRNA), gene-encoding proteins, and ribosomal RNA (rRNA), with manual adjustments as required (Tillich et al., 2017). We manually examined the IR junctions of *Salvia* species. A circular map of *Salvia* chloroplast genomes was subsequently drawn using OGDraw (Greiner et al., 2019).

Genome Comparative Analysis and Identification of Hypervariable Regions

MAFFT 7.221 was used to align the chloroplast genome sequences of *S. miltiorrhiza* f. *alba* and *S. yangii* (Katoh and Standley, 2013). Next, DnaSP 6.12 was used to perform a sliding window analysis with a step size of 200 bp and window length of 600 bp in order to detect the rapidly evolving molecular markers for performing phylogenetic analysis (Librado and Rozas, 2009).

Tree-based methods were utilized to assess the hypervariable barcodes and to compare the chloroplast genes *matK* and *rbcL*. MEGA 7.0 software was used to build maximum likelihood (ML) trees with 1,000 bootstrap replicates for each hypervariable marker (Kumar et al., 2016).

Identification of Chloroplast Gene Insertion in the Mitochondria

The mitochondrial genome of *S. miltiorrhiza* was retrieved from GenBank (NC_023209), followed by a homology search between chloroplast and mitochondria genomes using BLAST (with default parameters) to identify the transferred genes between the mitochondrial and chloroplast genomes. Circos was used to draw the chloroplast and mitochondrial maps from *Salvia* species, along with the fragments of gene transfers (Krzywinski et al., 2009).

Phylogenetic Analysis

We used the following three methods to perform phylogenetic analyses of *Salvia* species: Bayesian inference (BI) with a GTR + I + G model using MrBayes 3.2 [the Markov chain Monte Carlo (MCMC) algorithm was run for 1 million generations and sampled every 100 generations] (Ronquist et al., 2012); maximum likelihood (ML) using MEGA 7.0 with 1,000 bootstrap replicates (Kumar et al., 2016); and maximum parsimony (MP) with a

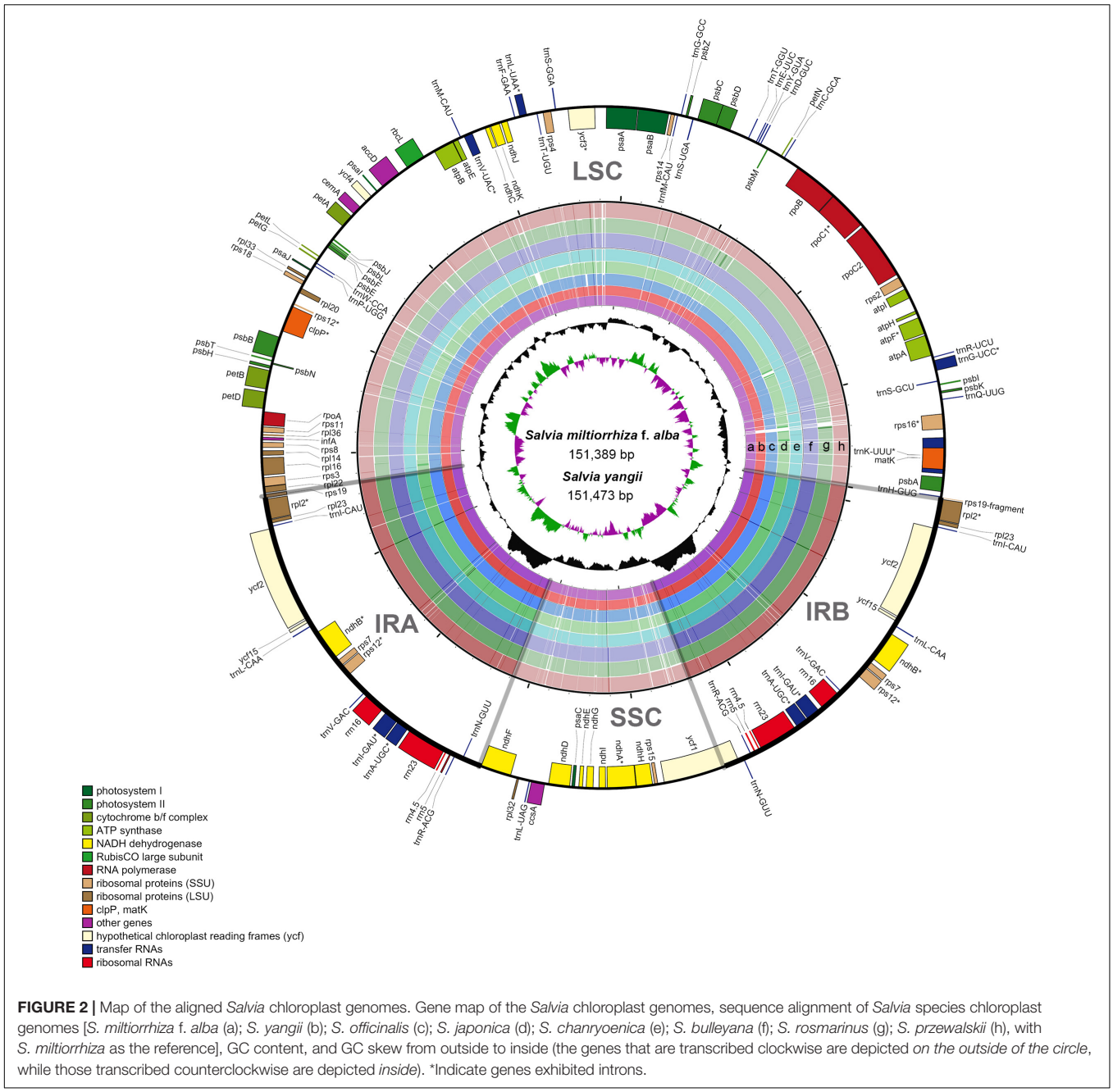


FIGURE 2 | Map of the aligned *Salvia* chloroplast genomes. Gene map of the *Salvia* chloroplast genomes, sequence alignment of *Salvia* species chloroplast genomes [*S. miltiorrhiza* f. *alba* (a); *S. yangii* (b); *S. officinalis* (c); *S. japonica* (d); *S. chanryoenica* (e); *S. bulleyana* (f); *S. rosmarinus* (g); *S. przewalskii* (h), with *S. miltiorrhiza* as the reference], GC content, and GC skew from outside to inside (the genes that are transcribed clockwise are depicted *on the outside of the circle*, while those transcribed counterclockwise are depicted *inside*). *Indicate genes exhibited introns.

heuristic search in PAUP 4.0 with 1,000 random taxon stepwise addition sequences (Swofford, 1993).

RESULTS

Chloroplast Genome Organization and Features

The chloroplast genomes of *S. yangii* and *S. miltiorrhiza* f. *alba* were 151,473 and 151,389 bp long, respectively, and exhibited a quadripartite structure with dual IR regions (25,603 and 25,521 bp), an SSC region (17,566 and 17,572 bp), and an LSC region (82,701 and 82,775 bp) (**Figure 1**). *S. yangii* and *S. miltiorrhiza* f. *alba* contained 116 and 115 unique genes in their chloroplast genomes, respectively, of which, 82 and 81 were protein-coding genes, four were rRNAs, and 30 were tRNAs (**Table 1**). Eighteen genes exhibited introns, of which six tRNAs (*trnV-UAC*, *trnG-UCC*, *trnK-UUU*, *trnL-UAA*, *trnI-GAU*, and *trnA-UGC*) and nine protein-coding genes (*accD*, *rps16*, *rpl16*, *ndhB*, *petB*, *rpl2*, *atpF*, *rpoC1*, and *ndhA*) had a single intron and the remaining three genes displayed two introns (*ycf3*, *rps12*, and *clpP*). For both *S. yangii* and *S. miltiorrhiza* f. *alba*, the *trnK-UUU* intron, comprising the *matK* gene, had the largest intron for both species (2,515 and 2,522 bp, respectively). Furthermore,

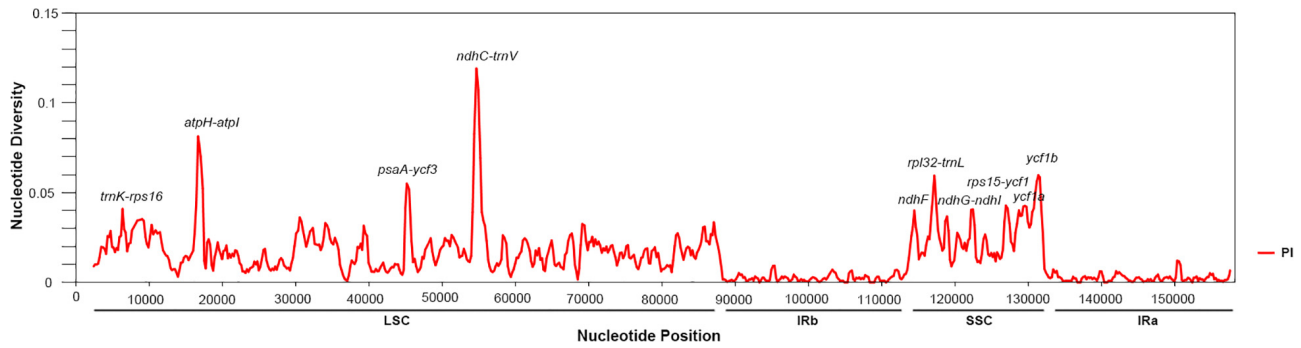


FIGURE 3 | Sliding window analysis of the *Salvia* chloroplast genomes (step size, 200 bp; window length, 600 bp). *X*-axis: midpoint position of the window; *Y*-axis: nucleotide diversity in each window.

TABLE 2 | Variability in the two universal chloroplast DNA barcodes and the 10 novel markers in *Salvia* species.

Markers	Length (bp)	Variable sites		PI sites[a]		Nucleotide diversity	Number of haplotypes
		Number	%	Number	%		
trnK-rps16	557	60	10.77	31	5.57	0.041	9
atpH-atpI	840	203	24.17	36	4.29	0.067	8
psaA-ycf3	625	117	18.72	18	2.88	0.052	8
ndhC-trnV	862	334	38.75	31	3.60	0.096	9
ndhF	2,217	129	5.82	68	3.07	0.022	8
rpl32-trnL	619	97	15.67	47	7.59	0.058	9
ndhG-ndhI	339	53	15.63	31	9.14	0.061	9
rps15-ycf1	369	51	13.82	29	7.86	0.053	8
ycf1a	974	58	5.95	27	2.77	0.020	9
ycf1b	1,024	115	11.23	47	4.59	0.039	8
rbcL	1,446	37	2.56	19	1.31	0.009	7
matK	1,535	90	5.86	49	3.19	0.023	9

[a]*Parsimony informative sites.*

the *Salvia* chloroplast genomes possessed similar GC contents (38.0–38.1%). IR expansion was considered the main cause for the size differences in the *Salvia* chloroplast genomes. Both *Salvia* species were found to be generally conserved in terms of gene order and genome structure (**Table 1**). The GenBank accession numbers for the complete chloroplast genome sequences of the two *Salvia* species were MT012420 and MT012421.

Comparisons of the fully annotated IR/SC junction regions were found to exhibit almost the same relative positions among the nine *Salvia* chloroplast genomes (**Figure 1**). The *rps19* gene contained all LSC/IRb junctions, resulting in the partial expansion (42–134 bp) of the IRb region toward the *rps19* gene. The IRb/SSC boundary positions were located on the *ycf1* and *ndhF* genes, and the SSC/IRa borders in the nine chloroplast genomes were located on the *ycf1* gene.

Comparative *Salvia* Chloroplast Genomes and Divergence Hot Spot Regions

The results of the comprehensive sequence divergence of the two newly assembled and the seven previously reported *Salvia* chloroplast genomes with *S. miltiorrhiza* as the control displayed high sequence similarity (**Figure 2**). As expected, the LSC and SSC regions exhibited comparatively higher sequence divergence than did the IR regions. A search for nucleotide substitutions identified 5,833 variable sites (3.69%), including 2,486 parsimony-informative sites (1.57%), across the nine *Salvia* chloroplast genomes.

Next, in the *Salvia* chloroplast genomes, we calculated the nucleotide diversity (π) values within 600-bp windows to detect the sequence divergence hot spots. The π values were in the range of 0–0.119, with extremely high values ($\pi > 0.002$) in the following 10 regions (*trnK-rps16*, *atpH-atpI*, *psaA-ycf3*, *ndhC-trnV*, *ndhF*, *rpl32-trnL*, *ndhG-ndhI*, *rps15-ycf1*, *ycf1a*, and *ycf1b*) (**Figure 3** and **Table 2**). Divergence hot spot regions could be the ideal molecular markers to distinguish *Salvia* species. Two conventional candidate DNA barcodes (*matK* and *rbcL*) were used to compare the marker divergence. The results revealed a lower variability in these barcodes compared with the newly identified markers (**Table 2**). The *ndhC-trnV* region exhibited the highest variability (38.75%). **Supplementary Figure S1** presents the graphical representation of these results using the ML method.

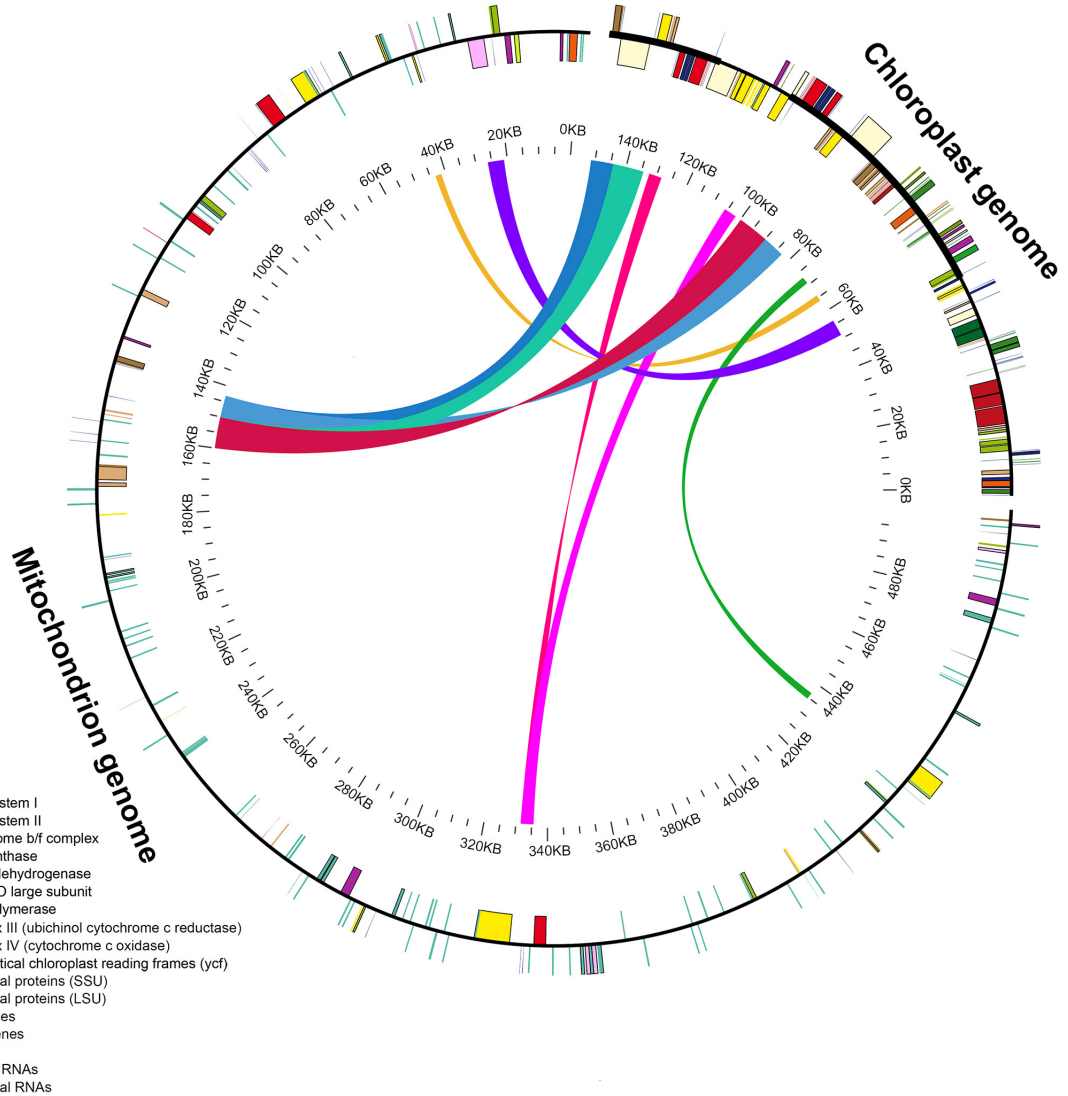


FIGURE 4 | Schematic for the chloroplast-to-mitochondrial gene transfer in *Salvia* species. A *colored line within the circle* represents the areas of the chloroplast genome that were transferred across the specified position in the mitochondrial genome. Genes shown *inside* and *outside the circle* have been transcribed in a clockwise and counterclockwise manner, respectively.

Characterization of *Salvia* Chloroplast Genome Transfer Into the Mitochondrial Genome

The length of the GenBank mitochondrial genome sequence of *S. miltiorrhiza* (499,236 bp) was found to be approximately 3.3 times longer than that of the chloroplast genome. We identified nine large chloroplast genome fragments in the mitochondrial genome, including both genes and intergenic regions. The fragments ranged from 1,737 to 9,734 bp and retained >97% of their sequence identity with their original chloroplast counterparts. These fragments had a total length of 49,895 bp, accounting for ∼32.97% of the chloroplast genome (**Figure 4** and **Supplementary Table S1**). Eleven intact chloroplast genes (*ndhB*, *rps7*, *rps12*, *ycf2*, *rpl23*, *atpE*, *atpB*, *rbcL*, *psbB*, *petL*, and *petG*), six tRNAs (*trnL-CAA*, *trnV-GAC*, *tRNA-Ile*, *trnM-CAU*, *trnW-CCA*, and *trnP-UGG*), one rRNA (*rrn23*), and numerous partial genes and intergenic spacer regions were identified.

Phylogenetic Analysis

In this study, the phylogenetic position of *Salvia* in the lamiids was standardized using chloroplast genomes (**Figure 5**). Three separate methods were used for conducting the phylogenetic analyses of the chloroplast genomes: ML, BI, and MP. Approximately identical topologies were generated by the ML and BI methods, with a high support for the majority of the branches (**Supplementary Figure S2**). However, marginally different positions of a few taxa were revealed by the MP trees compared with those by ML and BI (**Supplementary Figure S3**).

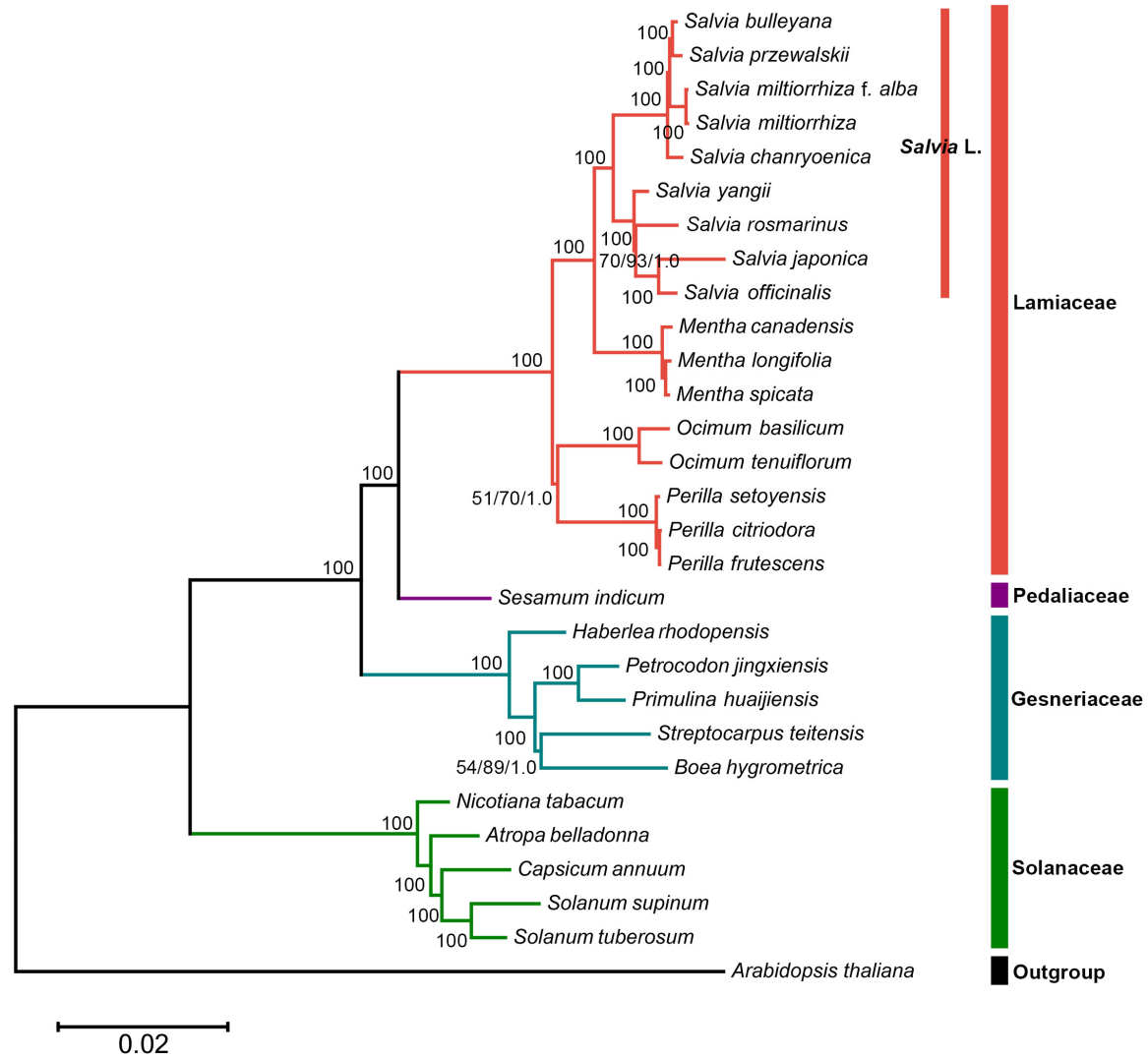


FIGURE 5 | The whole chloroplast genome sequence-based maximum likelihood (ML) phylogenetic tree of the nine *Salvia* species with 19 related species in the lamiids. *Numbers posted with branches* are the ML bootstrap values, maximum parsimony (MP) bootstrap values, and Bayesian posterior probabilities, respectively. *100* indicate 100% ML and MP bootstrap support and 1.0 Bayesian posterior probability.

Despite these variations, a majority of taxonomical relationships were highly supported and well resolved, which supported that the use of chloroplast genomic data promoted the resolution of phylogenetic analyses. Thus, the phylogenetic tree showed that *S. miltiorrhiza* f. *alba*, *S. miltiorrhiza*, *Salvia bulleyana*, and *Salvia przewalskii* were clustered on a single terminal branch.

DISCUSSION

Characterization of Complete Chloroplast Genome Structures

This is the first study to report the complete chloroplast genome of *S. yangii* and *S. miltiorrhiza* f. *alba*. We found the genomic structures along with gene type, number, and order to be fairly identical among the nine *Salvia* chloroplast genomes, with the exception of variations in the numbers of SNPs and small Indels (Qian et al., 2013; He et al., 2017; Du et al., 2019; Liang et al., 2019; Tao et al., 2019; Zhang X.J. et al., 2020). Thus, the study of the complete chloroplast genomes would afford useful genetic information to study the authentication, breeding, and evolutionary similarities/differences between *Salvia* species.

Comparative Analysis of *Salvia* Chloroplast Genomes and DNA Barcodes

The chloroplast genomes were found to be fairly conserved across the nine taxa, except for a few variable regions, based on the results of the comparative analysis. The coding regions contained most of the conserved sequences, whereas the non-coding

regions had most of the variable sequences. Consistent with the results of similar studies on other plants, the LSC and SSC regions were less conserved than the IR region (**Figure 1**; Song et al., 2019).

DNA barcoding was first proposed in 2003 (Hebert et al., 2003). DNA barcoding is a new technique that is widely used as a biological tool for breeding, species identification, and evolutionary research (Techen et al., 2014; Mishra et al., 2016). The herbal medicine industry relies on the identification of novel plant species. Due to their high medicinal value and cost-effective processing, it has become necessary to develop easy and safe methods for the identification and development of *Salvia* species.

The chloroplast genomes possess a smaller size and multiple copies in a cell compared with the nuclear genome. Furthermore, due to the presence of adequate interspecific divergence, in chloroplast genomes, the best species authentication methods are based on chloroplast genome-based DNA barcodes (Hollingsworth et al., 2011).

Here, we found an increase in the number of variable sites in the following 10 specific regions based on the results of pairwise chloroplast genomic alignment and SNP analysis: *trnK-rps16*, *atpH-atpI*, *psaA-ycf3*, *ndhC-trnV*, *ndhF*, *rpl32-trnL*, *ndhG-ndhI*, *rps15-ycf1*, *ycf1a*, and *ycf1b*. Thus, *Salvia* species may be detected using these regions as novel candidate fragments. However, further experiments are required to support this *Salvia* chloroplast sequence data.

Chloroplast Genome Fragments Were Found in the Mitochondrial Genome

Specific information pertaining to the intracellular gene transfer between different genomes (mitochondrial, nuclear, and chloroplast) has been disclosed through sequencing analysis (Timmis et al., 2004; Nguyen et al., 2020). Previous research has detected high amounts of transfer of nuclear DNA from the organelle in angiosperms (Hazkani-Covo et al., 2010; Smith, 2011; Park et al., 2014). Additionally, a characteristic feature of long-term evolution has been identified as chloroplast-to-mitochondrial gene transfer (Gui et al., 2016; Nguyen et al., 2020). Here, in *Salvia* species, we identified nine large chloroplast genome (32.97% of the chloroplast genome) fragments in the mitochondrial genomes.

Phylogenetic Analysis of *Salvia* Species

The whole chloroplast genome sequence-based phylogenetic tree was built to explore the evolutionary similarities/differences between *Salvia* species and between genera in the lamiids. We found that *S. przewalskii*, *S. miltiorrhiza*, *S. bulleyana*, and *S. miltiorrhiza* f. *alba* were clustered on a single terminal branch. Several studies have revealed similarities in appearance and characteristics between these species (Li et al., 2008; Wang et al., 2014). Regarding their compositions, common substituents of *S. miltiorrhiza* (*S. bulleyana*, *S. przewalskii*, and *S. miltiorrhiza* f. *alba*) shared the maximum chemical composition of *S. miltiorrhiza* (Danshen) (Wang et al., 2018).

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in GenBank of NCBI at https://www.ncbi.nlm.nih.gov, reference number MT012420 and MT012421. Chloroplast genome raw sequencing data have been deposited at the NCBI Sequence Read Archive (SRA) under accession PRJNA646330.

AUTHOR CONTRIBUTIONS

ZL and CG conceived the study and acquired the funding. CG performed the data analyses and drafted the earlier version of the manuscript. CW, QZ, XZ, MW, RC, and YZ revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This study was supported by the Clinical Medicine +X, The Affiliated Hospital of Qingdao University, and Qingdao University.

ACKNOWLEDGMENTS

We would like to thank Editage (www.editage.com) for English language editing. We finally thank the reviewers for reviewing the manuscript and providing valuable comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.574962/full#supplementary-material

REFERENCES

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Boachon, B., Buell, C. R., Crisovan, E., Dudareva, N., Garcia, N., Godden, G., et al. (2018). Phylogenomic mining of the mints reveals multiple mechanisms contributing to the evolution of chemical diversity in lamiaceae. *Mol. Plant* 11, 1084–1096. doi: 10.1016/j.molp.2018.06.002

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

Brunkard, J. O., Runkel, A. M., and Zambryski, P. C. (2015). Chloroplasts extend stromules independently and in response to internal redox signals. *Proc. Natl. Acad. Sci. U.S.A.* 112, 10044–10049. doi: 10.1073/pnas.1511570112

Chen, W., and Chen, G. X. (2017). Danshen (*Salvia miltiorrhiza* Bunge): a prospective healing sage for cardiovascular diseases. *Curr. Pharm. Des.* 23, 5125–5135. doi: 10.2174/1381612823666170822101112

Douglas, S. E. (1998). Plastid evolution: origins, diversity, trends. *Curr. Opin. Genet. Dev.* 8, 655–661. doi: 10.1016/s0959-437x(98)80033-6

Drew, B. T., Gonzalez-Gallegos, J. G., Xiang, C. L., Kriebel, R., Drummond, C. P., Walker, J. B., et al. (2017). Salvia united: the greatest good for the greatest number. *Taxon* 66, 133–145. doi: 10.12705/661.7

Du, Y., Wang, Y. Y., Xiang, C. L., and Yang, M. Q. (2019). Characterization of the complete chloroplast genome of salvia przewalskii Maxim (Lamiaceae), a substitute for *Dan-Shen Salvia miltiorrhiza* Bunge. *Mitochond. DNA Part B Resour.* 4, 981–982. doi: 10.1080/23802359.2019.1581107

Greiner, S., Lehwark, P., and Bock, R. (2019). OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* 47, W59–W64. doi: 10.1093/nar/gkz238

Gui, S. T., Wu, Z. H., Zhang, H. Y., Zheng, Y. Z., Zhu, Z. X., Liang, D. Q., et al. (2016). The mitochondrial genome map of *Nelumbo nucifera* reveals ancient evolutionary features. *Sci. Rep.* 6:11. doi: 10.1038/srep30158

Hazkani-Covo, E., Zeller, R. M., and Martin, W. (2010). Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet.* 6:834. doi: 10.1371/journal.pgen.1000834

He, Y., Han, L., Liu, Y., Tian, N., Su, X., and Wang, Z. (2017). Complete sequence analysis of chloroplast genome of salvia japonica. *Bull. Bot. Res.* 37, 572–578.

Hebert, P. D. N., Ratnasingham, S., and Dewaard, J. R. (2003). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc. R. Soc. B Biol. Sci.* 270, S96–S99. doi: 10.1098/rsbl.2003.0025

Hollingsworth, P. M., Graham, S. W., and Little, D. P. (2011). Choosing and using a plant DNA barcode. *PLoS One* 6:e019254. doi: 10.1371/journal.pone.0019254

Hu, G. X., Xiang, C. L., Liu, E. D., Dong, H. J., and Funamoto, T. (2016). Karyotypic study of eighteen taxa of Salvia (Lamiaceae) from China. *Caryologia* 69, 50–57. doi: 10.1080/00087114.2015.1109955

Hu, H., Hu, Q. J., Al-Shehbaz, I. A., Luo, X., Zeng, T. T., Guo, X. Y., et al. (2016). Species delimitation and interspecific relationships of the genus *Orychophragmus* (Brassicaceae) inferred from whole chloroplast genomes. *Front. Plant Sci.* 7:1826. doi: 10.3389/fpls.2016.01826

Huang, J., Yu, Y., Liu, Y. M., Xie, D. F., He, X. J., and Zhou, S. D. (2020). Comparative chloroplast genomics of *Fritillaria* (Liliaceae), inferences for phylogenetic relationships between *Fritillaria* and Lilium and plastome evolution. *Plants Basel* 9:133. doi: 10.3390/plants9020133

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649. doi: 10.1093/bioinformatics/bts199

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

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

Li, M. H., Li, Q. Q., Zhang, C. H., Zhang, N., Cui, Z. H., Huang, L. Q., et al. (2013). An ethnopharmacological investigation of medicinal *Salvia* plants (*Lamiaceae*) in China. *Acta Pharm. Sin. B* 3, 273–280. doi: 10.1016/j.apsb.2013.06.001

Li, Z., Zaijun, Y., Xia, H., Jin, L. I, and Deguang, W. A. N. (2008). Study on the leaf epidermal structural characters of *Salvia miltiorrhiza* and *Salvia* from Sichuan. *J. Sichuan Univer. Nat. Sci. Edn.* 45, 674–680.

Li, Z. M., Xu, S. W., and Liu, P. Q. (2018). *Salvia miltiorrhiza* Burge (Danshen): a golden herbal medicine in cardiovascular therapeutics. *Acta Pharmacol. Sin.* 39, 802–824. doi: 10.1038/aps.2017.193

Liang, C. L., Wang, L., Lei, J., Duan, B. Z., Ma, W. S., Xiao, S. M., et al. (2019). A Comparative analysis of the chloroplast genomes of four salvia medicinal plants. *Engineering* 5, 907–915. doi: 10.1016/j.eng.2019.01.017

Librado, P., and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25, 1451–1452. doi: 10.1093/bioinformatics/btp187

Mazandarani, M., and Ghaemi, E. (2010). Ethnopharmacological investigation of different parts of *Perovskia abrotanoides* Karel. *Planta Med.* 76, 1223–1223.

Mishra, P., Kumar, A., Nagireddy, A., Mani, D. N., Shukla, A. K., Tiwari, R., et al. (2016). DNA barcoding: an efficient tool to overcome authentication challenges in the herbal market. *Plant Biotechnol. J.* 14, 8–21. doi: 10.1111/pbi.12419

Nguyen, V. B., Giang, V. N. L., Waminal, N. E., Park, H. S., Kim, N. H., Jang, W., et al. (2020). Comprehensive comparative analysis of chloroplast genomes from seven Panax species and development of an authentication system based on species-unique single nucleotide polymorphism markers. *J. Ginseng Res.* 44, 135–144. doi: 10.1016/j.jgr.2018.06.003

Pan, X., Dong, L., Yang, L., Chen, D., and Peng, C. (2020). Potential drugs for the treatment of the novel coronavirus pneumonia (COVID-19) in China. *Virus Res.* 286:198057. doi: 10.1016/j.virusres.2020.198057

Park, S., Ruhlman, T. A., Sabir, J. S. M., Mutwakil, M. H. Z., Baeshen, M. N., Sabir, M. J., et al. (2014). Complete sequences of organelle genomes from the medicinal plant *Rhazya stricta* (Apocynaceae) and contrasting patterns of mitochondrial genome evolution across asterids. *BMC Genom.* 15:405. doi: 10.1186/1471-2164-15-405

Qian, J., Song, J. Y., Gao, H. H., Zhu, Y. J., Xu, J., Pang, X. H., et al. (2013). The complete chloroplast genome sequence of the medicinal plant *Salvia miltiorrhiza*. *PLoS One* 8:e057607. doi: 10.1371/journal.pone.0057607

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Hohna, S., et al. (2012). MrBayes 3.2: efficient Bayesian Phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. doi: 10.1093/sysbio/sys029

Smith, D. R. (2011). Extending the limited transfer window hypothesis to inter-organelle DNA migration. *Genome Biol. Evol.* 3, 743–748. doi: 10.1093/gbe/evr068

Song, Y., Zhang, Y. J., Xu, J., Li, W. M., and Li, M. F. (2019). Characterization of the complete chloroplast genome sequence of *Dalbergia* species and its phylogenetic implications. *Sci. Rep.* 9:20401. doi: 10.1038/s41598-019-56727-x

Swofford, D. L. (1993). PAUP – a Computer-Program for phylogenetic inference using maximum parsimony. *J. General Physiol.* 102, A9–A9.

Tao, A., Zhao, F. Y., and Qian, J. F. (2019). The complete chloroplast genome sequence of the medicinal plant *Salvia yunnanensis* C. H. Wright (Lamiaceae). *Mitochond. DNA Part B Resour.* 4, 3603–3605. doi: 10.1080/23802359.2019.1677523

Techen, N., Parveen, I., Pan, Z. Q., and Khan, I. A. (2014). DNA barcoding of medicinal plant material for identification. *Curr. Opin. Biotechnol.* 25, 103–110. doi: 10.1016/j.copbio.2013.09.010

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., et al. (2017). GeSeq - versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* 45, W6–W11. doi: 10.1093/nar/gkx391

Timmis, J. N., Ayliffe, M. A., Huang, C. Y., and Martin, W. (2004). Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat. Rev. Genet.* 5, 123–U116. doi: 10.1038/nrg1271

Topcu, G. (2006). Bioactive triterpenoids from *Salvia* species. *J. Nat. Prod.* 69, 482–487. doi: 10.1021/np0600402

Wang, L. L., Ma, R. F., Liu, C. Y., Liu, H. X., Zhu, R. Y., Guo, S. Z., et al. (2017). *Salvia miltiorrhiza*: a potential red light to the development of cardiovascular diseases. *Curr. Pharm. Des.* 23, 1077–1097. doi: 10.2174/1381612822666161010105242

Wang, M. X., Cui, L. C., Feng, K. W., Deng, P. C., Du, X. H., Wan, F. H., et al. (2015). Comparative analysis of *Asteraceae* chloroplast genomes: structural organization. RNA Editing and Evolution. *Plant Mol. Biol. Rep.* 33, 1526–1538. doi: 10.1007/s11105-015-0853-2

Wang, T., Liu, S.-Y., Wang, L., Wang, H.-Y., and Zhang, L. (2014). Anatomical characteristics of laminae and petioles of 11 species of *Salvia* and their taxonomic significance. *China J. Chin. Mater. Med.* 39, 2629–2634.

Wang, Y., Jiang, K., Wang, L. J., Han, D. Q., Yin, G., Wang, J., et al. (2018). Identification of Salvia species using high-performance liquid chromatography combined with chemical pattern recognition analysis. *J. Separ. Sci.* 41, 609–617. doi: 10.1002/jssc.201701066

Wen, L., Zhou, Z., Jiang, D., and Huang, K. (2020). Effect of Xuebijing injection on inflammatory markers and disease outcome of coronavirus disease 2019. *Zhonghua Wei Zhong Bing Ji Jiu Yi Xue* 32, 426–429. doi: 10.3760/cma.j.cn121430-20200406-00386

Wicke, S., Schneeweiss, G. M., Depamphilis, C. W., Muller, K. F., and Quandt, D. (2011). The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Mol. Biol.* 76, 273–297. doi: 10.1007/s11103-011-9762-4

Xiang, C. L., Hu, G. X., and Peng, H. (2016). *Salvia wuana* (Lamiaceae), a new name for *S-pauciflora* E. *Peter. Phytotaxa* 255, 99–100. doi: 10.11646/phytotaxa.255.1.10

Xu, H. B., Song, J. Y., Luo, H. M., Zhang, Y. J., Li, Q. S., Zhu, Y. J., et al. (2016). Analysis of the genome sequence of the medicinal plant *Salvia miltiorrhiza*. *Mol. Plant* 9, 949–952. doi: 10.1016/j.molp.2016.03.010

Xu, W. Q., Losh, J., Chen, C., Li, P., Wang, R. H., Zhao, Y. P., et al. (2019). Comparative genomics of figworts (Scrophularia, Scrophulariaceae), with implications for the evolution of *Scrophularia* and *Lamiales*. *J. Syst. Evol.* 57, 55–65. doi: 10.1111/jse.12421

Zhang, L. L., Liu, Y. Y., Liu, Z. L., Wang, C., Song, Z. Q., Liu, Y. X., et al. (2016). Comparison of the roots of *Salvia miltiorrhiza* Bunge (Danshen) and its variety *S. miltiorrhiza* Bge f. Alba (Baihua Danshen) based on multi-wavelength HPLC-fingerprinting and contents of nine active components. *Analyt. Methods* 8, 3171–3182. doi: 10.1039/c6ay00185h

Zhang, X. J., Chen, C., Wang, R., Yao, Y., Liu, L. X., and Zhang, L. Y. (2020). Characterization of the complete chloroplast genome of *Salvia hispanica* (Lamiaceae). *Mitochondr. DNA Part B Resour.* 5, 1748–1750. doi: 10.1080/23802359.2020.1749162

Zhang, Y. L., Cui, Q., Zhang, D., Ma, X., and Zhang, G. W. (2020). Efficacy of Xuebijing injection for the treatment of coronavirus disease 2019 via network pharmacology. *Trad. Med. Res.* 5, 201–215. doi: 10.12032/tmr20200507178

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.

Copyright © 2020 Gao, Wu, Zhang, Zhao, Wu, Chen, Zhao and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

DATA REPORT
published: 23 October 2020
doi: 10.3389/fgene.2020.590065

First Draft Genome Assembly of the Seaweed *Sargassum fusiforme*

Shengqin Wang[1,2†], *Lidong Lin*[2,3†], *Yijian Shi*[4], *Weiguo Qian*[2], *Nan Li*[1,2], *Xiufeng Yan*[1,2], *Huixi Zou*[1,2*] *and Mingjiang Wu*[1,2*]

[1] National and Local Joint Engineering Research Center of Ecological Treatment Technology for Urban Water Pollution, Wenzhou University, Wenzhou, China, [2] College of Life and Environmental Science, Wenzhou University, Wenzhou, China, [3] Dongtou Fisheries Science and Technology Research Institute, Wenzhou, China, [4] College of Electrical and Electronic Engineering, Wenzhou University, Wenzhou, China

Keywords: ***sargassum fusiforme*, genetic breeding, brown algae, phaeophyceae, sargassaceae**

OPEN ACCESS

Edited by:
Hongjian Wan,
ZheJiang Academy of Agricultural Sciences, China

Reviewed by:
Peng Xu,
Guangzhou University, China
Lijing Bu,
University of New Mexico, United States

***Correspondence:**
Huixi Zou
zjuzhx@wzu.edu.cn
Mingjiang Wu
wmj@wzu.edu.cn

†These authors have contributed equally to this work

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

Received: *31 July 2020*
Accepted: *17 September 2020*
Published: *23 October 2020*

Citation:
Wang S, Lin L, Shi Y, Qian W, Li N, Yan X, Zou H and Wu M (2020) First Draft Genome Assembly of the Seaweed Sargassum fusiforme. Front. Genet. 11:590065. doi: 10.3389/fgene.2020.590065

INTRODUCTION

Sargassum fusiforme is a major edible brown macroalga grown in lower intertidal zones along coastlines. The macroalga is cultivated in Japan, Korea, and China (Ma et al., 2017). *S. fusiforme* is a source of bioactive compounds, such as polysaccharides, which have various functions, including antioxidants, anti-aging, memory improvement and immune regulation (Chen et al., 2016; Bogie et al., 2019; Zhang et al., 2020). It is also an essential herb used to disperse phlegm in traditional Chinese medicine (Li and Wei, 2002). However, genomic and genetic resources for this economically important seaweed are limited. This limitation could be attributed to, in part, the high levels of polysaccharides in the brown algae, which complicate DNA extraction. Recent research has shown that *S. fusiforme* contains potentially toxic quantities of inorganic arsenic, although the levels differ significantly between different growing areas (Yokoi and Konomi, 2012). Notably, a study on the molecular genetics and genomics of *S. fusiforme* could give a clue on strategies for reducing the levels of inorganic arsenic.

The genus *Sargassum* (Fucales, Phaeophyceae) consists of over 300 species of multicellular organisms, which are widely distributed in marine habitats. Many of these species are essential in the ecosystem and are economically important marine crops. For example, *Sargassum horneri* is the dominant golden-tide seaweed in the ocean (Liu et al., 2018). Despite its wide distribution, there are no available complete or draft genomes for any of the *Sargassum* genus microalgae to date. Besides, the entire Phaeophyceae class only has six decoded complete or draft genome sequences of brown algae species. These species include *Ectocarpus siliculosus* (Cock et al., 2010), *Saccharina japonica* (Ye et al., 2015), *Macrocystis pyrifera* (https://www.ncbi.nlm.nih.gov/genome/?term=Macrocystis+pyrifera), *Nemacystus decipiens* (Nishitsuji et al., 2019), *Cladosiphon okamuranus* (Nishitsuji et al., 2016), and *Undaria pinnatifida* (Shan et al., 2020). The availability of genomic resources for these essential algae will assist future selective breeding programs and fundamental genomic and evolutionary studies, among other roles.

The complete chloroplast and mitochondrial genome sequences of *S. fusiforme* have been decoded and phylogenetically analyzed (Liu et al., 2016; Yonghua et al., 2019). Recently, genome assembly using a sequencing technology that combines PacBio and Illumina reads has shown high performance. In this study, *de novo* whole-genome sequencing in a strain of *S. fusiforme* was performed using PacBio and Illumina reads. The PacBio long reads were assembled using wtdbg2,

and the Illumina short reads were used to correct the primary assembly by Pillon (Walker et al., 2014; Ruan and Li, 2020). To our knowledge, this is the first reported genome assembly in a member of the Fucales order.

Value of the Data

The genomic sequence data generated in the present study provides essential information for further *S. fusiforme* studies, such as functional genomic analysis and genome-assisted breeding. Since these results offer the first reference genome of the order Fucales, they can be used in evolutionary studies in the Chromista kingdom. Further, the draft genome can be used as a genomic reference for the discovery of functional products and other genome mining applications.

MATERIALS AND METHODS

Materials and DNA Extraction

The *S. fusiforme* strain used for genome sequencing in this study was collected from the Wenzhou Dongtou District, Zhejiang Province of China. Dozens of picked vesicles were cleaned in ddH_2O, then frozen and stored in liquid nitrogen for genomic DNA extraction. The genomic DNA was extracted by the cesium chloride density gradient centrifugation method (Jagielski et al., 2017). DNA quantitation and quality checks were performed using a NanoDrop 2000 microspectrophotometer, a Qubit fluorometer, and 1% agarose gel electrophoresis (Novogene Corporation, Beijing, China).

Illumina Sequencing and Genome Size Estimation

The isolated genomic DNA was subjected to short-read library preparation with stranded Illumina paired-end protocols (Illumina Inc., San Diego, CA, USA). Subsequently, libraries with insert lengths of 500 bp were constructed and sequenced by the Illumina NovaSeq platform. The raw reads were processed with the fastp program with the following parameters: -g -q 5 -u 50 -n 15 -l 150 (Chen et al., 2018), to obtain more than 197.5 million paired-end clean reads. The kmer frequencies ($k = 21$) were calculated using Jellyfish v2.2.10 (Marçais and Kingsford, 2011). The kmer-based statistical method from Genomescope (Vurture et al., 2017), predicted the *S. fusiforme* haploid genome length to be ∼ 450 MB, with heterozygosity of ∼1.0%. Notably, this estimated size is larger than the genome size of the *Sargassum* genus (196–319 Mb) previously estimated in three other species by static microspectrophotometry (Phillips et al., 2011).

PacBio Sequencing and Genome Assembly

For long-read sequencing, the isolated genomic DNA was subjected to a large insert library preparation using the PacBio Sequel sequencing platform (PacBio, Menlo Park, CA, USA). A total of ∼16 GB of polymerase reads were obtained following the manufacturer's instructions. Long reads were assembled by wtdbg2 (Ruan and Li, 2020). Due to wtdbg2 constructed the consensus using a fuzzy Bruijn graph and assembled raw reads without error correction, the primary assembly was realigned against the long reads using Minimap2, then polished with Racon, which does three replications(Vaser et al., 2017; Li, 2018). Finally, the Illumina paired-end clean reads were mapped onto the assembled contigs using bowtie2 and further corrected by Pilon (Langmead and Salzberg, 2012; Walker et al., 2014). To remove potential contaminants, contigs with a biased GC content (>0.6 or <0.4) were aligned to the NCBI non-redundant nucleic acid database with an *E*-value 1e-5. Then, a total of 112 contigs with the top 10 no matches to eukaryotes were regarded as contaminant contigs and removed from the final assembly. At last, 6,750 contigs were retained as the final assembly, with a total length of ∼394.4 MB, which has an N50 value of ∼142.1 KB (**Table 1**). Therefore, the estimated coverage for PacBio data and Illumina data were ∼42 x and 180 x, respectively.

TABLE 1 | Comparison of genome assemblies of seven brown algae.

Species	*E. siliculosus*	*S. japonica*	*N. decipiens*	*C. okamuranus*	*U. pinnatifida*	*M. pyrifera*	*S. fusiforme*
Total genome size (Mbp)	197	537	154	130	502.8	409.1	394.4
Scaffold N50 length (Kbp)	6,528	252	1,863	418	16,510	–	–
Contig N50 length (Bp)	–	58,867	6,265	21,705	1,707,374	2,573	142,085
Number of scaffolds	30	13,327	685	541	114	–	–
Number of contigs	–	–	411,597	31,858	515	246,830	6,750
Length of longest contig (Kbp)	324	466	–	–	7,337	110	931
Length of shortest contig (Bp)	–	–	–	–	–	280	319
Length of longest scaffolds (Kbp)	18,946	1,469	–	–	32,303	–	–
Length of shortest scaffolds (Bp)	869,870	500	–	–	1,000	–	–
GC contents (%)	56	49.6	54	54	50.14	50	48.4
The identified complete BUSCOs (%)	–	–	–	–	82.9	–	90.4

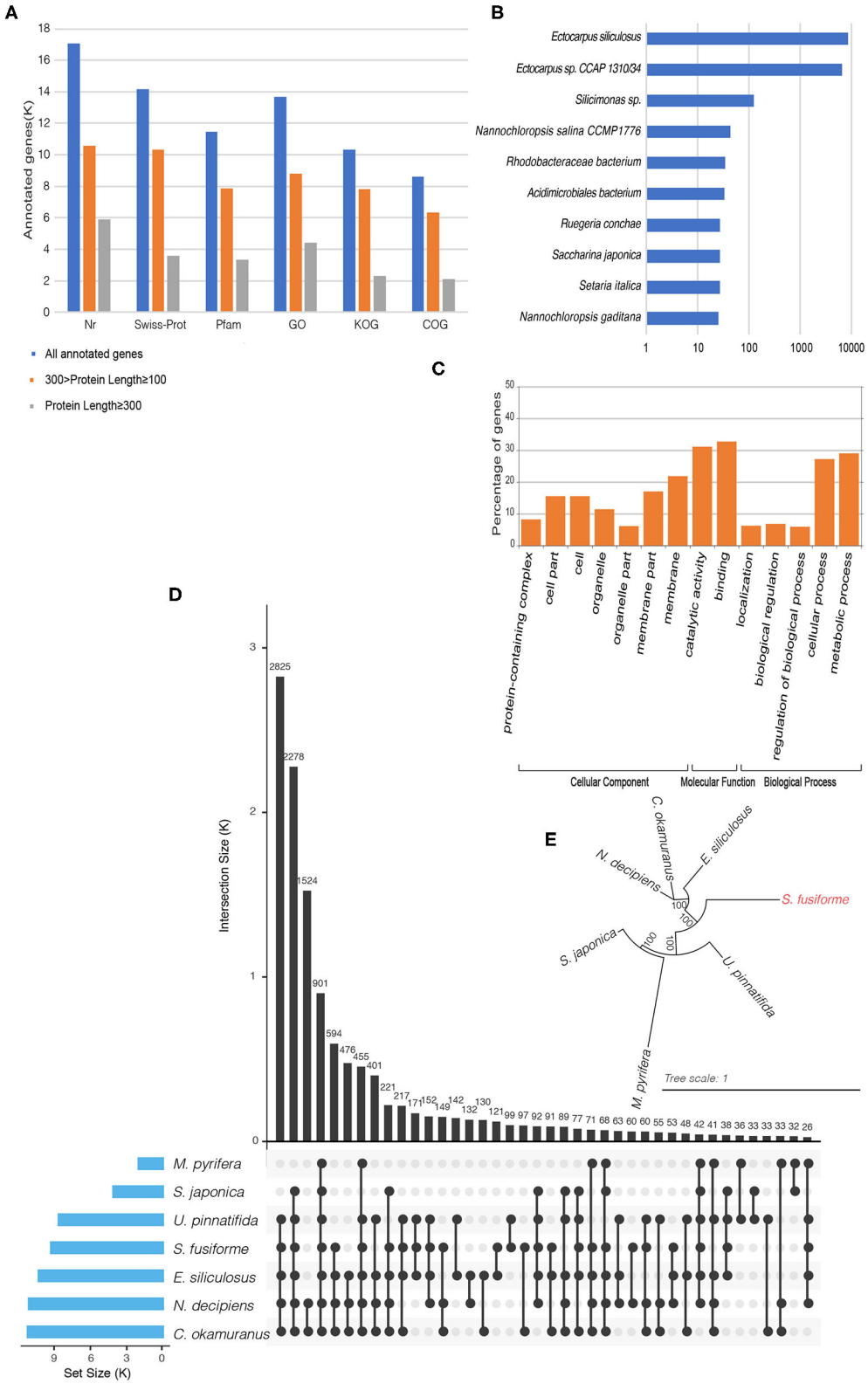


FIGURE 1 | (A) Statistics of gene functional annotations in the *S. fusiforme* genome. **(B)** Species distribution in Nr homolog. **(C)** Gene Ontology classification. Visualization of gene ontology terms (gene number > 1,000) for the putative genes. **(D)** UpSet plot showing ortholog intersections across the seven brown algae. **(E)** Phylogenetic tree of seven brown algae using the maximum likelihood method.

Repeat Region Prediction

A *de novo* repeat library for *S. fusiforme* was conducted using RepeatModeler (Flynn et al., 2020), which employs three repeat-finding methods; RECON (Bao and Eddy, 2002), RepeatScout (Price et al., 2005), and TRF (Benson, 1999). The repeat library was then subjected to RepeatMasker to find and mask homologous repeats in the assembled genome using rmblast as the default search engine. Finally, the total length of repetitive sequences was ∼239.4 Mb, accounting for ∼60.7% of the draft genome size.

Gene Prediction and Annotation

The BRAKER1 pipeline was used to perform gene prediction by integrating *ab initio* gene prediction and RNA-seq based prediction, which combined the advantages of both GeneMark-ET and AUGUSTUS (Hoff et al., 2016). Raw RNA-seq data from three independent transcriptomic projects were used for gene prediction: ERR2041176, SRR12206544, and SRR5357673 (One Thousand Plant Transcriptomes Initiative, 2019). Transcript alignment was performed on the repeat masked assembly based on Tophat2, with default parameters (Kim et al., 2013). The aligned results, combined with transcript data, were used to generate initial gene structures using the GeneMark-ET tool (Lomsadze et al., 2014). The initial gene structures were used for training by AUGUSTUS to produce the final gene predictions (Stanke and Waack, 2003). Finally, 20,222 putative genes were detected. Of these putative genes, 84.45% genes had Nr homologs, 69.93% had Swiss-Prot homologs, 56.57% had Pfam homologs, 67.65% had gene ontology (GO) homologs, 51.11% had KOG homologs, and 42.60% had COG homologs (**Figure 1A**). In Nr homolog analysis (with an *E*-value of 0.0001), the best hit species with 15,284 genes belong to the *Ectocarpus* genus (**Figure 1B**). Here, the GO term distribution of the putative genes into the molecular function, cellular component, and biological process categories (**Figure 1C**) was determined by the Blast2GO suite with an *E*-value of 0.0001 and visualized with the WEGO 2.0 web tool at the macro-level (Conesa and Götz, 2008). COG and KOG analyses were performed with an *E*-value of 0.001 by the WebMGA server (Wu et al., 2011). KAAS web services were used to map the putative *S. fusiforme* genes onto the KEGG metabolic pathways along with genes in other plant species, including *Amborella trichopoda*, *Chlamydomonas reinhardtii*, *Ostreococcus lucimarinus*, *Ostreococcus tauri*, *Micromonas commoda*, *Cyanidioschyzon merolae*, *Galdieria sulphuraria*, and *Arabidopsis thaliana*. A total of 3,796 genes had homologs in the KEGG database, of which 1,140 were mapped onto 119 enzymes listed in the pathway categorized "Metabolism."

Completeness and Accuracy of the Assembly

The predicted genes from the BRAKER1 pipeline were subject to BUSCO version 3.0.2 to evaluate the completeness of the assembled genome, based on the eukaryota_odb9 database. More than 90% complete BUSCOs were detected at the protein level, with the single-copy, duplicated, fragmented, and missing accounting for 82.8, 7.6, 6.3, and 3.3%, respectively. The accuracy of the genome assembly was evaluated by mapping the clean reads against the genome with bowtie2. Finally, more than 90.7% of reads were remapped.

Comparative Genomics

Ortholog analysis was conducted using OrthoMCL v2.0.9 based on protein datasets from the BRAKER1 pipeline and six other brown algae species: *E. siliculosus*, *S. japonica*, *M. pyrifera*, *N. decipiens*, *C. okamuranus*, and *U. pinnatifida*. Here, the protein sequences for *S. japonica* were determined by combining transcriptome data from the 1 KP (One Thousand Plant Transcriptomes Initiative, 2019), and the protein sequences for *M. pyrifera* were called using FRAGGENESCAN from pyrosequencing technology data (Konotchick et al., 2013). CD-HIT was used to remove redundant sequences (90% identity or more) in each organism. Then, non-redundant protein sequences were subjected to all-again-all Blastp with an *E*-value of 1e-5. Finally, a total of 1,4819 groups were identified by OrthoMCL and ortholog intersections across the seven brown algae, as shown in **Figure 1D**. Subsequently, 2,287 groups were identified as one to one orthologs from these species. These sequences were concatenated into a supergene and used for multiple sequence alignment by MAFFT. Then, a phylogenetic tree was built by PhyML (**Figure 1E**) using the maximum likelihood method.

DATA AVAILABILITY STATEMENT

The data used to support the findings of this study are available from Short Read Archive (SRA) database (http://www.ncbi.nlm.nih.gov/sra/) under the accession number: PRJNA597239. The raw SMRT sequencing data is available in NCBI SRA database with accession number: SRR12124361. Genome assembly and annotation data has been deposited at Figshare repository (https://figshare.com/s/517f180deb56a5b31b14).

AUTHOR CONTRIBUTIONS

LL, XY, HZ, and MW conceived the project and designed the objectives. LL, WQ, and NL collected the specimens. WQ and NL perform DNA/RNA extraction and sequencing. SW and YS implemented genome assembly and data analysis. SW drafted the manuscript. All authors edited and contributed to the article.

FUNDING

This work was supported by the Special Project on Blue Granary Science and Technology Innovation under the National Key R&D Program (2018YFD0901501), the Special Science and Technology Innovation Project for Seeds and Seedlings of Wenzhou City (N20160016), the National Natural Science Foundation of China (41876197 and 31670402), and the Key Fishing and Agricultural Science and Technology Project in Dongtou District, Wenzhou (N2018Y03A).

REFERENCES

Bao, Z., and Eddy, S. R. (2002). Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Res.* 12, 1269–1276. doi: 10.1101/gr.88502

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi: 10.1093/nar/27.2.573

Bogie, J., Hoeks, C., Schepers, M., Tiane, A., Cuypers, A., Leijten, F., et al. (2019). Dietary Sargassum fusiforme improves memory and reduces amyloid plaque load in an Alzheimer's disease mouse model. *Sci. Rep.* 9:4908. doi: 10.1038/s41598-019-41399-4

Chen, P., He, D., Zhang, Y., Yang, S., Chen, L., Wang, S., et al. (2016). Sargassum fusiforme polysaccharides activate antioxidant defense by promoting Nrf2-dependent cytoprotection and ameliorate stress insult during aging. *Food Funct.* 7, 4576–4588. doi: 10.1039/c6fo00628k

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

Cock, J. M., Sterck, L., Rouzé, P., Scornet, D., Allen, A. E., Amoutzias, G., et al. (2010). The Ectocarpus genome and the independent evolution of multicellularity in brown algae. *Nature* 465, 617–621. doi: 10.1038/nature09016

Conesa, A., and Götz, S. (2008). Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genomics* 2008:619832. doi: 10.1155/2008/619832

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., et al. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U.S.A.* 117, 9451–9457. doi: 10.1073/pnas.1921046117

Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., and Stanke, M. (2016). BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* 32, 767–769. doi: 10.1093/bioinformatics/btv661

Jagielski, T., Gawor, J., Bakuła, Z., Zuchniewicz, K., Zak, I., and Gromadka, R. (2017). An optimized method for high quality DNA extraction from microalga prototheca wickerhamii for genome sequencing. *Plant Methods* 13, 77–78. doi: 10.1186/s13007-017-0228-9

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36. doi: 10.1186/gb-2013-14-4-r36

Konotchick, T., Dupont, C. L., Valas, R. E., Badger, J. H., and Allen, A. E. (2013). Transcriptomic analysis of metabolic function in the giant kelp, Macrocystis pyrifera, across depth and season. *New Phytol.* 198, 398–407. doi: 10.1111/nph.12160

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Meth.* 9, 357–359. doi: 10.1038/nmeth.1923

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191

Li, X., and Wei, W. (2002). *Chinese Materia Medica: Combinations & Applications.* Donica Publishing Ltd.; Elsevier Health Sciences.

Liu, F., Liu, X., Wang, Y., Jin, Z., Moejes, F. W., and Sun, S. (2018). Insights on the *Sargassum horneri* golden tides in the Yellow Sea inferred from morphological and molecular data. *Limnol. Oceanograp.* 63, 1762–1773. doi: 10.1002/lno.10806

Liu, F., Pang, S., and Luo, M. (2016). Complete mitochondrial genome of the brown alga *Sargassum fusiforme* (*Sargassaceae, Phaeophyceae*): genome architecture and taxonomic consideration. *Mitochondrial DNA A DNA Mapp. Seq. Anal.* 27, 1158–1160. doi: 10.3109/19401736.2014.936417

Lomsadze, A., Burns, P. D., and Borodovsky, M. (2014). Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res.* 42, e119–e119. doi: 10.1093/nar/gku557

Ma, Z., Wu, M., Lin, L., Thring, R. W., Yu, H., Zhang, X., et al. (2017). Allelopathic interactions between the macroalga Hizikia fusiformis (Harvey) and the harmful blooms-forming dinoflagellate Karenia mikimotoi. *Harmful Algae* 65, 19–26. doi: 10.1016/j.hal.2017.04.003

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770. doi: 10.1093/bioinformatics/btr011

Nishitsuji, K., Arimoto, A., Higa, Y., Mekaru, M., Kawamitsu, M., Satoh, N., et al. (2019). Draft genome of the brown alga, Nemacystus decipiens, Onna-1 strain: fusion of genes involved in the sulfated fucan biosynthesis pathway. *Sci. Rep.* 9, 4607–4611. doi: 10.1038/s41598-019-40955-2

Nishitsuji, K., Arimoto, A., Iwai, K., Sudo, Y., Hisata, K., Fujie, M., et al. (2016). A draft genome of the brown alga, Cladosiphon okamuranus, S-strain: a platform for future studies of "mozuku" biology. *DNA Res.* 23, 561–570. doi: 10.1093/dnares/dsw039

One Thousand Plant Transcriptomes Initiative (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574, 679–685. doi: 10.1038/s41586-019-1693-2

Phillips, N., Kapraun, D. F., Gómez Garreta, A., Ribera Siguan, M. A., Rull, J. L., Salvador Soler, N., et al. (2011). Estimates of nuclear DNA content in 98 species of brown algae (Phaeophyta). *AoB Plants* 2011:plr001. doi: 10.1093/aobpla/plr001

Price, A. L., Jones, N. C., and Pevzner, P. A. (2005). *De novo* identification of repeat families in large genomes. *Bioinformatics* 21(Suppl. 1), i351–i358. doi: 10.1093/bioinformatics/bti1018

Ruan, J., and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nat. Meth.* 17, 155–158. doi: 10.1038/s41592-019-0669-3

Shan, T., Yuan, J., Su, L., Li, J., Leng, X., Zhang, Y., et al. (2020). First genome of the brown alga undaria pinnatifida: chromosome-level assembly using PacBio and Hi-C technologies. *Front. Genet.* 11:140. doi: 10.3389/fgene.2020.00140

Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19(Suppl. 2), ii215–ii225. doi: 10.1093/bioinformatics/btg1080

Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017). Fast and accurate *de novo* genome assembly from long uncorrected reads. *Genome Res.* 27, 737–746. doi: 10.1101/gr.214270.116

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., et al. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* 33, 2202–2204. doi: 10.1093/bioinformatics/btx153

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* 9:e112963. doi: 10.1371/journal.pone.0112963

Wu, S., Zhu, Z., Fu, L., Niu, B., and Li, W. (2011). WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genomics* 12, 444–449. doi: 10.1186/1471-2164-12-444

Ye, N., Zhang, X., Miao, M., Fan, X., Zheng, Y., Xu, D., et al. (2015). Saccharina genomes provide novel insight into kelp biology. *Nat. Commun.* 6:6986. doi: 10.1038/ncomms7986

Yokoi, K., and Konomi, A. (2012). Toxicity of so-called edible hijiki seaweed (*Sargassum fusiforme*) containing inorganic arsenic. *Regul. Toxicol. Pharmacol.* 63, 291–297. doi: 10.1016/j.yrtph.2012.04.006

Yonghua, Z., Shengqin, W., Weiguo, Q., Yijian, S., Nan, L., Xiufeng, Y., et al. (2019). Characterization of the complete chloroplast genome of *Sargassum fusiforme* and its phylogenomic position within phaeophyceae. *Mitochondrial DNA Part B* 4, 3258–3259. doi: 10.1080/23802359.2019.1671246

Zhang, R., Zhang, X., Tang, Y., and Mao, J. (2020). Composition, isolation, purification and biological activities of *Sargassum fusiforme* polysaccharides: A review. *Carbohyd. Poly.* 228:115381. doi: 10.1016/j.carbpol.2019.115381

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.

Copyright © 2020 Wang, Lin, Shi, Qian, Li, Yan, Zou and Wu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 27 October 2020
doi: 10.3389/fgene.2020.591984

Genome-Wide Identification and Expression Profile of the SNAT Gene Family in Tobacco (*Nicotiana tabacum*)

Jiemei Zhang[1,2], Zhengping Yao[1,2], Renjun Zhang[1,2], Zongmin Mou[1,3], Honghui Yin[4], Tianyang Xu[4], Dake Zhao[1,3] and Suiyun Chen[1,3]**

[1] Biocontrol Engineering Research Center of Plant Disease and Pest, Biocontrol Engineering Research Center of Crop Disease and Pest, Yunnan University, Kunming, China, [2] School of Life Sciences, Yunnan University, Kunming, China, [3] School of Ecology and Environmental Science, Yunnan University, Kunming, China, [4] Wenshan Branch of Yunnan Tobacco Company, Wenshan, China

Melatonin plays key roles in development and confers stress tolerance to plants. Serotonin *N*-acetyltransferase (SNAT) is either the enzyme involved in the last step or the penultimate enzyme of phytomelatonin biosynthesis. To date, *SNAT* genes have not been characterized in tobacco (*Nicotiana tabacum*), an economically important plant species. The sequence of the Acetyltransf_7 conserved domain was used as a query sequence, and 12 *NtSNAT* candidate genes were in turn identified in the genome of tobacco. These *NtSNAT*s could be divided into two groups based on the phylogenetic tree. *NtSNAT1* and *NtSNAT2* clustered together with the other typical *SNAT*s, but the other 10 *NtSNAT*s separately clustered outside of the typical *SNAT*s. These 10 *NtSNAT*s have only motif 1, whereas representative *SNAT*s, such as *NtSNAT1* and *NtSNAT2* or a *SNAT* from cyanobacteria, have five motifs. In addition, *NtSNAT1* and *NtSNAT2* are highly homologous to the characterized *OsSNAT1,* 62.95 and 71.36%, respectively; however, the homology between the other 10 *NtSNAT* genes and *OsSNAT1* is low. Concomitantly, it is hypothesized that *NtSNAT1* and *NtSNAT2* are the homolog of *SNATs*, whereas the other 10 candidates could be considered *NtSNAT*-like genes. Furthermore, both *Nicotiana tomentosiformis* and *Nicotiana sylvestris*, two diploid ancestor species of *N. tabacum*, have two *SNAT* candidates; therefore, it is speculated that gene rearrangement or deletion during the process of genomic stabilization after whole-genome duplication or polyploidization led to the preservation of *NtSNAT1* and *NtSNAT2* during the evolution of tobacco from the ancestral diploid to the allotetraploid. *NtSNAT* and *NtSNAT*-like genes were differentially expressed in all organs under different stress conditions, indicating that these genes potentially associated with plant growth and development and stress resistance. Under different stress conditions, the expression of *NtSNAT1* was significantly upregulated upon high-temperature and cadmium stresses, while the expression of *NtSNAT2* did not significantly increase under any of the tested stress treatments. These results provide valuable information for elucidating the evolutionary relationship of *SNAT* genes in tobacco and genetic resources for improving tobacco production in the future.

Keywords: *Nicotiana tabacum*, melatonin, serotonin *N*-acetyltransferase, tissue-specific expression, abiotic stress

OPEN ACCESS

Edited by:
Rong Zhou,
Aarhus University, Denmark

Reviewed by:
Tanweer Kumar,
Lanzhou University, China
Jianghua Chen,
Xishuangbanna Tropical Botanical Garden (CAS), China

****Correspondence:***
Dake Zhao
zhaodk2012@ynu.edu.cn
Suiyun Chen
chensuiyun@ynu.edu.cn

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

Received: *05 August 2020*
Accepted: *30 September 2020*
Published: *27 October 2020*

Citation:
Zhang J, Yao Z, Zhang R, Mou Z, Yin H, Xu T, Zhao D and Chen S (2020) Genome-Wide Identification and Expression Profile of the SNAT Gene Family in Tobacco (Nicotiana tabacum). Front. Genet. 11:591984. doi: 10.3389/fgene.2020.591984

INTRODUCTION

Melatonin (*N*-acetyl-5-methoxytryptamine) is an evolutionarily conserved pleiotropic molecule that exists ubiquitously in living organisms (Tan et al., 2011; Lee et al., 2014a; Kanwar et al., 2018). Melatonin is considered to be a growth regulator and is involved in regulating numerous plant biological processes, including seed germination, rooting, flowering, senescence, and photosynthesis (Chen et al., 2009; Kolář et al., 2010; Wang et al., 2012; Hernández et al., 2015; Wei et al., 2015; Zhang et al., 2017; Arnao and Hernández-Ruiz, 2019; Turk and Genisel, 2019). In addition to its roles in plant development, melatonin is also involved in the tolerance of plants to a range of biotic and abiotic stresses, such as pathogen infection, drought, high temperature, cold, and salinity (Tal et al., 2011; Byeon et al., 2014; Lee et al., 2014a; Shi et al., 2015a; Zhang et al., 2015; Arnao and Hernández-Ruiz, 2017; Zhao D. et al., 2019).

Melatonin is synthesized from tryptophan through four distinct reaction steps. A total of six enzymes, tryptophan decarboxylase (TDC), TPH, tryptamine 5-hydroxylase (T5H), serotonin *N*-acetyltransferase (SNAT), acetylserotonin-*O*-methyltransferase (ASMT), and caffeic acid *O*-methyltransferase (COMT), participate in the synthesis of melatonin in plants, among which SNAT is either the penultimate enzyme or the last step enzyme involved in melatonin biosynthesis (Back et al., 2016; Hardeland, 2016; Wang et al., 2017; Yu et al., 2018). Since the first plant *SNAT* gene was identified and cloned in rice, homologous *SNAT* genes in other species, such as cyanobacteria, alga laver, *Arabidopsis thaliana*, grapevine and apple, also have been identified, found to be present at even higher frequencies than the frequency in rice, and enzymatically characterized (Byeon et al., 2013, 2015, 2016b; Wang et al., 2017; Yu et al., 2019).

*SNAT*s have been reported to play a regulatory role in maintaining the steady-state level of melatonin, and there is a relationship between *SNAT*s and the response to abiotic stress (Park et al., 2014). A growing body of evidence indicates that different plant SNATs have different thermophilic properties: a SNAT protein of cyanobacteria presented increased catalytic activity at 70℃, a SNAT protein of loblolly pine presented increased catalytic activity at 55℃, and a SNAT protein of apple presented increased catalytic activity at 35℃. The heat resistance of SNATs shows that it functions in the heat stress response (Byeon et al., 2013, 2016b; Kang et al., 2013; Park et al., 2014; Yu et al., 2019). Additionally, transgenic *Arabidopsis* ectopically expressing *MzSNAT5* presented elevated melatonin levels, thus resulting in enhanced drought tolerance (Wang et al., 2017). Low melatonin production by suppression of *SNAT*s in rice causes stunted seedling growth together with yield penalty, increased abiotic stress susceptibility, and increased coleoptile growth under anoxic conditions (Byeon and Back, 2016). In addition to these abiotic defense responses, inhibition of the *GhSNAT1* melatonin biosynthesis-relate genes reduced the resistance of cotton inoculated with pathogenic bacteria (Li et al., 2019). Hence, *SNAT*s play an imperative role in the plant response to biotic and abiotic stresses.

Tobacco (*Nicotiana tabacum*) is an allotetraploid (2n = 48) that originated from chromosome doubling after an intraspecific hybridization event between *Nicotiana tomentosiformis* (2n = 24) and *Nicotiana sylvestris* (2n = 24) (Lim et al., 2004). Tobacco is an essential commercial crop species within the Solanaceae family. Tobacco is cultivated in more than 120 countries contributes substantially to the economic development of countries worldwide (Wang and Bennetzen, 2015; Tong et al., 2019). However, tobacco is often affected by various biotic and abiotic stresses during its growth and development, such as drought, pathogens, cold, high temperature, and heavy metals (Cho and Hong, 2006; Liu et al., 2011; Ma et al., 2014). These stresses cause stunted growth, senescence, reduce yields, and even death (Debnath et al., 2019). A large number of studies have confirmed that melatonin plays a pivotal role in plant growth and development and the response to biotic or abiotic stress (Hernández-Ruiz et al., 2004; Shi et al., 2015a,b; Yu et al., 2018; Zhao L. et al., 2019). However, numerous important biological pathways and gene families, including SNAT family members responsible for melatonin biosynthesis, remain unexplored in tobacco due to the lack of fully annotated reference genomes (Yu et al., 2020).

To further explore the roles of *SNAT* genes in tobacco, we used tobacco genomic data and performed a genome-wide investigation of the *SNAT* gene family. The phylogenetic relationships, sequence features, gene structures, and protein motifs of these *NtSNAT* genes were analyzed. Moreover, the functional diversity of *SNAT*s in tobacco was studied through the analysis of tissue-specific expression patterns and stress responses. Our results provide a reference for the identification of tobacco *SNAT* gene function and facilitate further work on improving the stress resistance of tobacco.

MATERIALS AND METHODS

Tobacco *SNAT* Gene Sequence Retrieval and Gene Identification

To identify the tobacco *SNAT* candidates, the hidden Markov model (HMM) profile of the *SNAT* conserved domain (Pfam13508) was used as a query to search the genomic databases of *N. tabacum* (tobacco), *N. tomentosiformis*, *N. sylvestris*, and *Rhodospirillum rubrum* by the Markov model-based HMMER program. *R. rubrum* is the first photosynthetic α-proteobacterium indicated to synthesize melatonin (Manchester et al., 1995). All the candidate sequences were analyzed via the PFAM[1] and SMART[2] databases (Zhao Y. et al., 2019), and proteins without a typical *SNAT* conserved domain were removed. Chloroplast transit signal peptides were identified via ChloroP[3] (Byeon et al., 2016a). The molecular weight (MW) and isoelectric point (pI) of each protein sequence were calculated using the online tool ExPASY[4] (Liu et al., 2017).

[1]http://pfam.xfam.org/

[2]http://smart.embl-heidelberg.de/

[3]http://www.cbs.dtu.dk/services/Chlorop/

[4]http://web.expas-y.org/compute_pi

Distribution of Conserved Domains and Analysis of *NtSNAT* Structure

Conserved domains within *NtSNAT* genes were identified via the NCBI database[5]. Exon-intron structures were determined according to the alignments of their transcribed sequences with corresponding genomic sequences, and a diagram was generated with the online Gene Structure Display Server[6]. All the conserved motifs of the SNATs and NtSNAT proteins were subsequently identified by the MEME program[7].

Sequence Alignment and Phylogenetic Tree Construction

The deduced protein sequences of the NtSNATs were aligned with the sequence of OsSNAT1 (AK059369) using ClustalX (2.0.9). We performed phylogenetic analyses based on the neighbor-joining method using MEGA 7.0 software. Except for these of the *SNAT*s identified in *Nicotiana*, the sequences of other putative *SNAT*s were retrieved from the NCBI database (**Supplementary Table S1**). Bootstrap analysis was performed using 1,000 resampling replications, and branch lengths were assigned through pairwise calculations of the genetic distances.

Plant Growth Conditions and Stress Treatments

Seeds of tobacco (K326 cultivar) were germinated in growth media consisting of a mixture of vermiculite and humus (v:v = 1:2) in polystyrene, square dishes. The seedlings were grown in a greenhouse at a day/night temperature of 25/18°C, an air humidity of 50–60%, and a photoperiod consisting of 16 h light/8 h dark; the light intensity was 100 μmol/m^2/s. For heat-stress treatment, a group of 6 week-old tobacco seedlings was treated at 45°C in a growth chamber for 3, 6, and 9 h. For cold stress treatment, the seedlings were placed in a 4°C incubator for durations of 3, 6, and 24 h. The drought stress treatment was stopped, after which the seedlings were watered for 1, 4, and 7 days. For cadmium treatment, the cadmium concentration was determined on the basis of 10 mg/kg mixed soil, and samples were taken at 1, 4, and 7 days. The photoperiod and humidity of all the stress treatments were the same as those of the growth conditions described above. The conditions of the control treatment were the same those during seedling growth. Each treatment involved three biological replicates, and all the samples were immediately frozen in liquid nitrogen and stored at −80°C until analysis.

Isolation of Total RNA and Reverse Transcription

Total RNA was extracted from tobacco leaves, stems, flowers, fruits, and roots using an RNAprep Pure Plant Kit (TIANGEN). First-strand cDNA was synthesized via a cDNA synthesis kit. Gene-specific primers were designed against the genome sequence of tobacco present in the NCBI database. qRT-PCRs assay were performed in a 10.0 μl reaction volume using Super Real PreMix Plus (TIANGEN), and the actin gene was used as an internal control. All the primers used for qRT-PCR-based analysis are presented in **Supplementary Table S2**. Expression data were calculated using the $2^{-\Delta\Delta CT}$ method, and the actin gene was used as a reference for the expression analysis of the *NtSNAT* genes in tobacco. All the results were generated via six samples: three biological replicates and three technical replicates. The $2^{-\Delta\Delta CT}$ values were subsequently used to draw heatmaps via MeV 4.9 and GraphPad Prism 5 software.

Statistical Analysis

All the data were statistically analyzed using SPSS 23.0 statistical software. To investigate the expression differences of *SNAT* genes in all the samples, the *T*-test was used. Single asterisk (*) indicate significant differences, at $p \leq 0.05$, and double asterisks (**) indicate extremely significant differences, at $p \leq 0.01$.

RESULTS

Primary Identification of *NtSNAT* Genes in the Tobacco Genome

Name searches and HMM analysis revealed a total of 12 candidate *SNAT* genes in tobacco. For convenience, we named the *NtSNAT* genes according to their MW. Details concerning the gene name, locus name, open reading frame (ORF) length, exon and intron numbers, protein length, MW, pI, and chloroplast transit signal peptide are listed in **Table 1** and **Supplementary Table S3**. The MWs of the predicted NtSNAT proteins ranged from 20.4 to 34.9 kDa. In addition, the pIs ranged from 5.25 to 9.28, the ORF lengths ranged from 525 to 906, and protein lengths ranged from 174 to 301 amino acids (aa). A chloroplast transit signal peptide was identified in five of the candidate genes: *NtSNAT1*, *NtSNAT2*, *NtSNAT3*, *NtSNAT8*, and *NtSNAT9*. To examine the structural features of the *NtSNAT* genes, the exon/intron configurations of *NtSNAT* genes in the tobacco plants were compared. Sequence analysis revealed that introns were present in the coding DNA sequences (CDSs) of these genes, except in *NtSNAT7* and *NtSNAT12*, and the number of introns varied from 2 to 8 (**Figure 1**). On the basis of the established *SNAT* identification standard, for the two diploid ancestors of tobacco, *N. tomentosiformis*, and *N. sylvestris*, each have two *SNAT* genes.

Multiple Sequence Alignment of *NtSNATs* and the Characterized Rice *OsSNAT1* Suggests *NtSNAT1* and *NtSNAT2* Are Authentic Tobacco *SNATs*

Sequence alignment of the NtSNAT proteins is shown in **Figure 2**. The results showed that *NtSNAT1* and *NtSNAT2* are highly homologous to the identified *OsSNAT1* (62.95 and 71.36%, respectively), while several other candidate genes have low homology with *OsSNAT1*. Therefore, we preliminarily infer that *NtSNAT1* and *NtSNAT2* are homologs of *SNATs* and that the other 10 members belong to another subgroup. It was found that aa 169–242 of *NtSNAT1* and *NtSNAT2* compose conserved domains according to NCBI conserved domain analysis. Each of these two

[5]https://www.ncbi.nlm.nih.gov/Strucure/cd-d/Wrpsb.cgi

[6]http://www.cbi.pku.edu.cn/resource/index.htm

[7]http://meme-suite.org/tools/meme

TABLE 1 | *SNAT* gene candidates in tobacco, *Nicotiana tomentosiformis*, and *Nicotiana sylvestris*.

Name	Gene locus	ORF length (bp)	Chloroplast transit peptide	No. exons	No. introns	Deduced polypeptide		
						Length (aa)	Mw (kDa)	pI
NtSNAT1	LOC107779554	771	Yes	8	7	256	28.6	5.25
NtSNAT2	LOC107817186	768	Yes	8	7	255	28.4	5.38
NtSNAT3	LOC107796590	834	Yes	3	2	277	30.8	9.44
NtSNAT4	LOC107771832	651	No	6	5	216	25	8.95
NtSNAT5	LOC107772525	525	No	6	5	174	20.4	6.91
NtSNAT6	LOC107774960	906	No	9	8	301	34.9	9.23
NtSNAT7	LOC107791297	708	No	1	0	235	26.1	5.86
NtSNAT8	LOC107802269	855	Yes	5	4	284	31.7	7.53
NtSNAT9	LOC107820131	849	Yes	5	4	282	31.6	6.94
NtSNAT10	LOC107823851	525	No	6	5	174	20.4	6.91
NtSNAT11	LOC107827352	906	No	9	8	301	34.9	9.28
NtSNAT12	LOC107827499	708	No	1	0	235	26.3	5.68
NsSNAT1	LOC104229486	768	Yes	8	7	255	28.42	5.38
NsSNAT2	LOC104224393	780	Yes	1	0	259	21.91	9.8
NtoSNAT1	LOC104107124	768	Yes	8	7	255	28.59	5.15
NtoSNAT2	LOC104104574	780	Yes	1	0	259	21.86	9.58

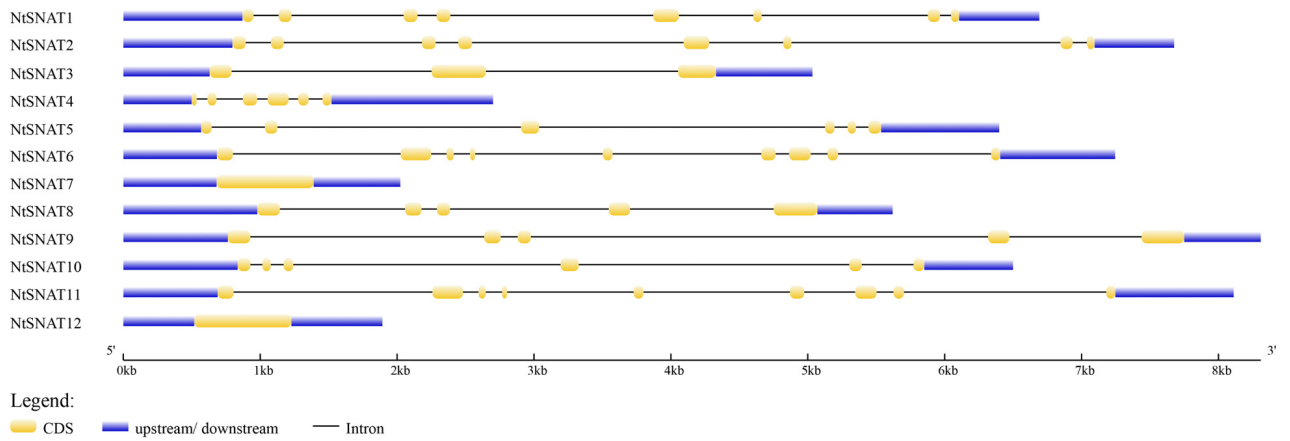


FIGURE 1 | Exon-intron configurations of *NtSNAT* genes in tobacco plants. The introns and exons are drawn to scale together with the complete coding regions of their respective genes. The boxes indicate the exons, and lines indicate the introns.

proteins has a GNAT functional domain with a coenzyme-A-binding site, and it is suggested that these two proteins could be tobacco *SNAT*s. In addition, the motif 1 conserved region, which may have important functions, was detected in these proteins.

Further Validation of the *NtSNAT1* and *NtSNAT2* Members of *NtSNAT* Family Based on Phylogenetic Tree and Motif Analysis

For the phylogenetic tree, a BLAST search was performed using the rice *OsSNAT1* amino acid sequence. *SNAT* homologs were found in various taxa, including bacteria, algae, mosses, ferns, gymnosperms, and angiosperms (**Figure 3A**). These homologous genes not only have typical *SNAT* conserved domains but also complete conserved motifs. The phylogenetic tree indicated that the *NtSNAT* and *SNAT* genes from the included species could be divided into two clades. *NtSNAT1* and *NtSNAT2* could be classified as typical *SNAT*s, while the remaining 10 *NtSNAT*s clustered together separately. Additionally, four *SNAT*s from two diploid ancestors of tobacco, *N. tomentosiformis* and *N. sylvestris*, clustered into clade I together with the typical *SNAT*s. Protein motifs are often used to predict protein function. Twelve *NtSNAT* and twenty-three *SNAT* genes were analyzed according to their conserved motifs (**Figure 3B** and **Supplementary Figure S1**). The results showed that motif 1 was the most widely distributed and was present in all the members. However, except in *R. rubrum*, motif 2, motif 3, motif 4, and motif 5 were present in all the members of clade I. The *SNAT* genes in clade II contained only motif 1. Therefore, it was further speculated that *NtSNAT1* and *NtSNAT2* were likely tobacco *SNAT*s, and the other 10 candidates were considered *NtSNAT*-like genes that belonged to another subgroup. Each of the two ancestor diploids has two possible *SNAT* homologs. During the evolution from diploids

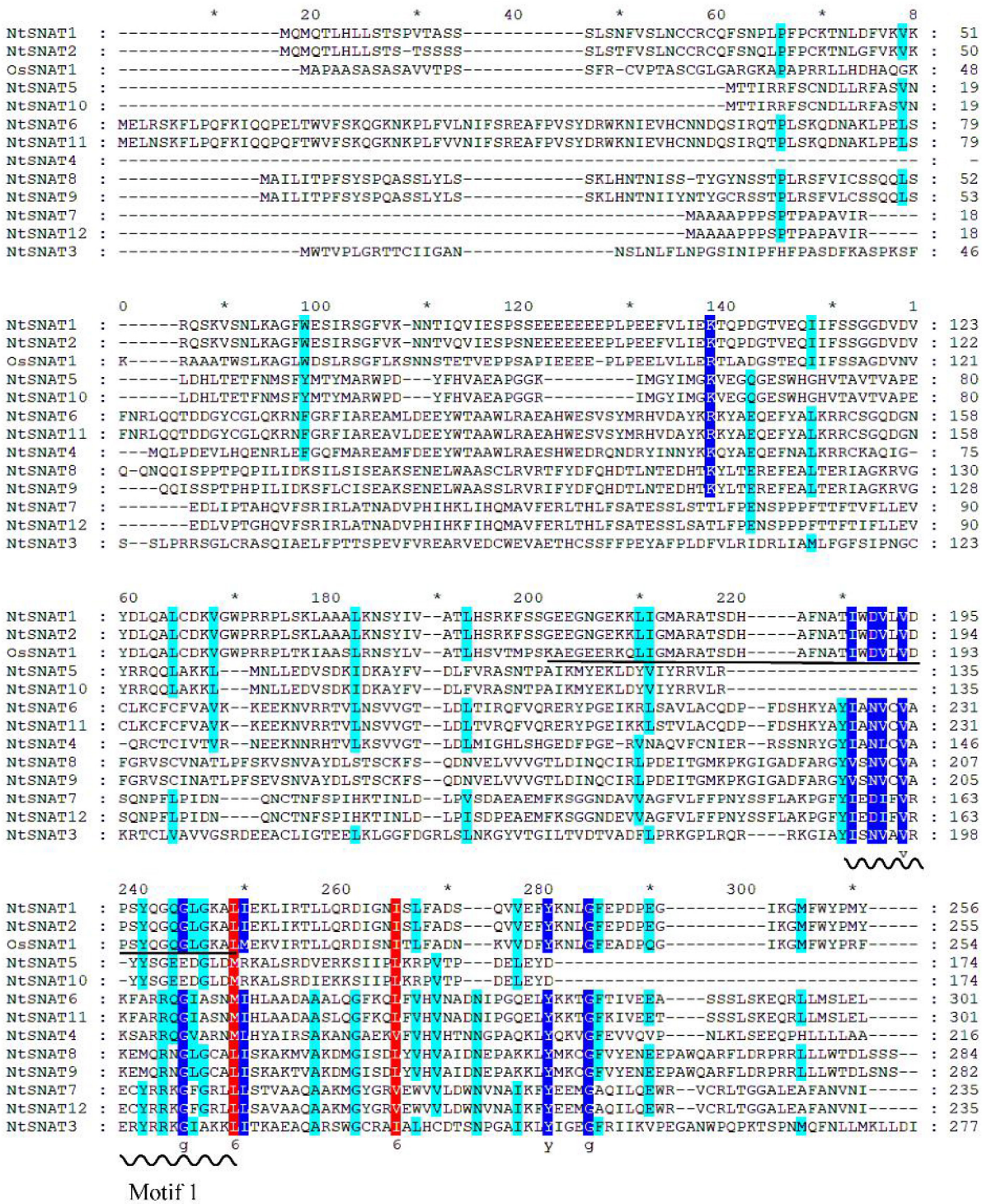


FIGURE 2 | Multiple sequence alignment of OsSNAT1 and NtSNAT proteins. The wavy line in the figure indicates conserved motif 1, and the conserved acetyl-CoA binding motif is underlined. The conserved amino acid residues are indicated by colored shading. The asterisk "*" represent the length of every twenty amino acids (including blanks inside).

to allotetraploids, there should be eight *NtSNAT* homologs in allotetraploid tobacco under normal circumstances, but only two *SNAT* homologs were identified in tobacco. These results suggest that the current typical *SNAT*s in tobacco may have arisen via gene loss during the process of genomic stabilization after the occurrence of polyploidization or whole-genome duplication.

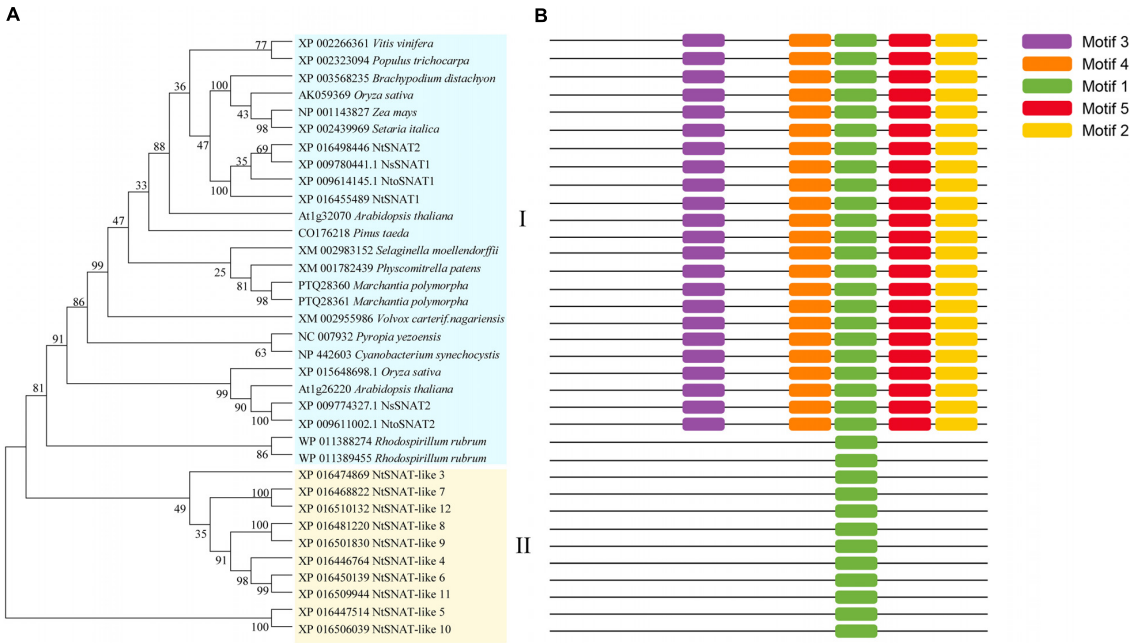


FIGURE 3 | Phylogenetic tree of *NtSNAT*s and other *SNAT*s **(A)** and corresponding conserved motifs for each gene **(B)**. Different motifs are highlighted with different colored boxes with numbers 1–5.

Moreover, motif 1, motif 2, motif 3, and motif 4 were present at the C-terminus, suggesting important biological functions of the C-terminus for members of the *SNAT* gene family.

Organ-Specific Expression Analysis of *NtSNAT* and *NtSNAT*-Like Genes

Expression profiling provides useful clues about gene function. To examine the expression patterns of the candidate *NtSNAT* genes, we evaluated their expression levels in the roots, stems, young leaves, flowers, and fruits (**Figure 4**). The results revealed that *NtSNAT* and *NtSNAT*-like genes were expressed in all of the tested organs. Among them, the expression of *NtSNAT1*, *NtSNAT2*, and *NtSNAT*-like 10 in the leaves was significantly higher than that in other organs. In addition, *NtSNAT*-like 3 and *NtSNAT*-like 5 were expressed at relatively high levels in the flowers. Similarly, *NtSNAT*-like 6, *NtSNAT*-like 8, *NtSNAT*-like 9, *NtSNAT*-like 11, and *NtSNAT*-like 12 were highly expressed in the fruits. The expression of *NtSNAT*-like 4 was extremely low in the stems and fruits, and this gene was expressed mainly in the roots. Last, the expression of *NtSNAT*-like 7 was significantly higher in the stems than in the other organs. And these genes related Gene Ontology terms were listed in **Supplementary Table S4**, further indicating the functions of 12 SNAT genes in different organs of tobacco.

Expression Profiles of *NtSNAT* and *NtSNAT*-Like Genes Under Different Stress Conditions

Under various stress conditions, it is evident that the expression of *NtSNAT* and *NtSNAT*-like genes was induced to a greater extent when the plants were subjected to abiotic stress, including heat, cold, cadmium, and drought than when they were under no stress (**Figure 5**). Furthermore, the expression of most of these genes was highly variable. The expression of six *NtSNAT* genes, *NtSNAT1*, *NtSNAT*-like 3, *NtSNAT*-like 5, *NtSNAT*-like 6, *NtSNAT*-like 10, and *NtSNAT*-like 11, dramatically increased in response to high-temperature conditions. Under cadmium stress, the expression of the *NtSNAT1*, *NtSNAT*-like 3, *NtSNAT*-like 10, *NtSNAT*-like 11, and *NtSNAT*-like 12 genes strongly increased. When the plants were under drought stress, the expression of only three genes, *NtSNAT*-like 3, *NtSNAT*-like 8, and *NtSNAT*-like 9, was significantly upregulated, whereas the *NtSNAT*-like 7, *NtSNAT*-like 8, *NtSNAT*-like 10, *NtSNAT*-like 11, and *NtSNAT*-like 12 genes responded strongly to cold stress. Furthermore, the expression of three of these genes (*NtSNAT2*, *NtSNAT*-like 4, and *NtSNAT*-like 9) did not significantly increase under all the tested stresses.

DISCUSSION

The SNAT family belongs to the GNAT superfamily, and the members of the SNAT family play a critical role in regulating the accumulation of melatonin (Dyda et al., 2000; Lee et al., 2015). *SNAT* genes have been identified in various plant species (Byeon et al., 2016b; Yu et al., 2019). Twelve candidate *SNAT* genes in tobacco were recognized. When aligned with the defined *OsSNAT1* amino acid sequence (Kang et al., 2013), *NtSNAT1* and *NtSNAT2* were highly homologous to *OsSNAT1* (62.95 and 71.36%, respectively), while several other *NtSNAT* genes were found to have low homology with *OsSNAT1*. Hence, we

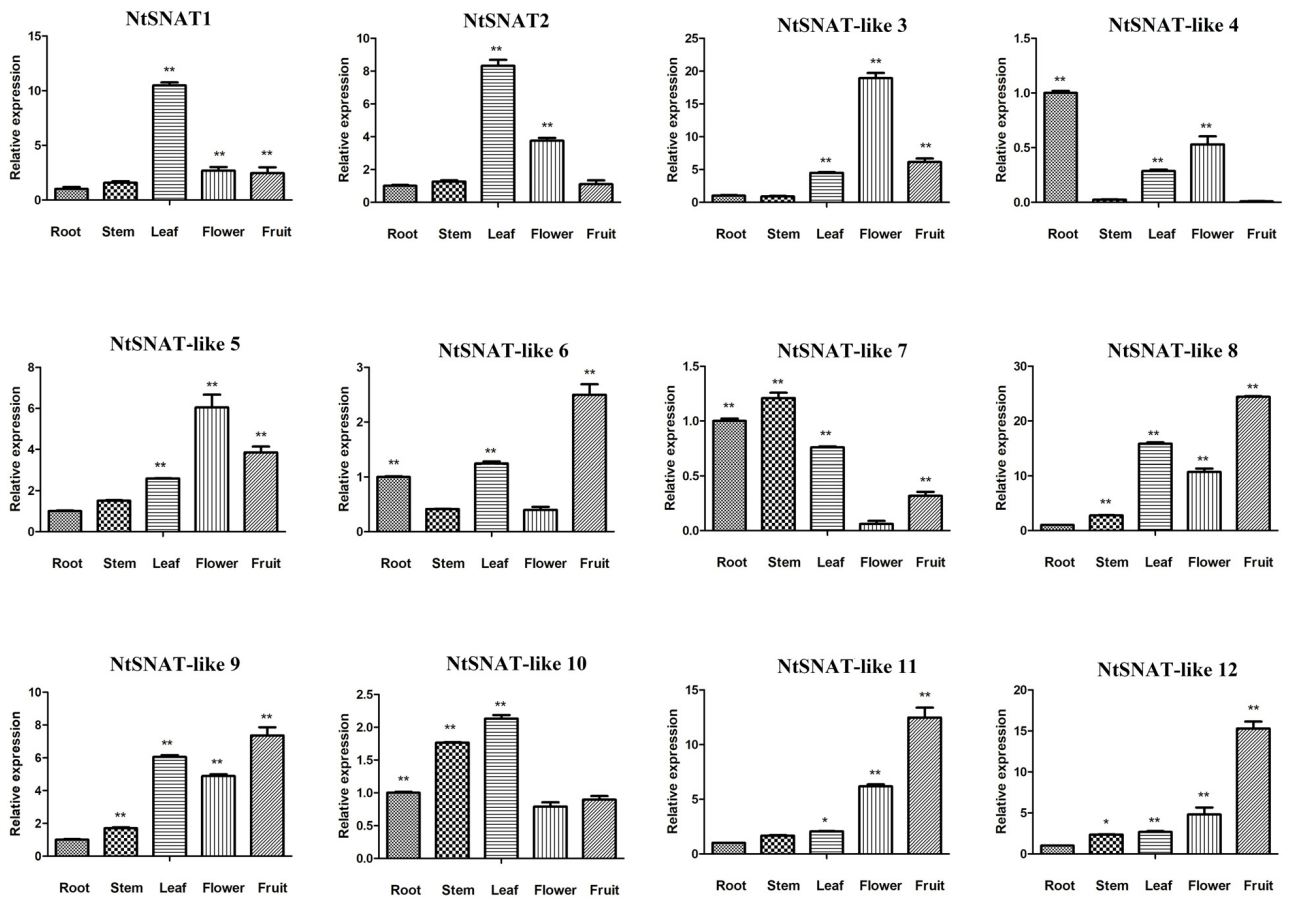


FIGURE 4 | Expression profile of *NtSNAT* candidates in different tobacco organs based on qRT-PCR results. Single asterisk (*) indicates significant differences at $p \leq 0.05$, and double asterisks (**) indicate extremely significant differences at $p \leq 0.01$.

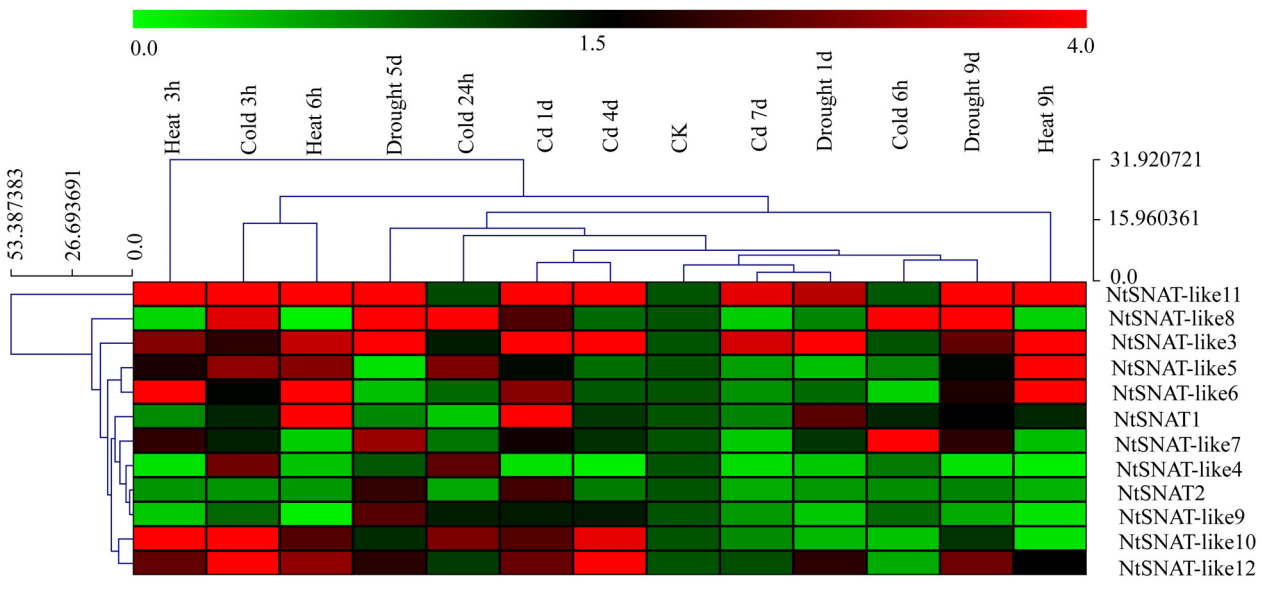


FIGURE 5 | Hierarchical cluster display of expression profiles of *NtSNAT* and *NtSNAT*-like genes in tobacco leaves in response to drought, heat, cold, and cadmium (Cd) treatments. The blocks with colors indicate low (green) or high (red) transcript accumulation.

preliminarily determined that *NtSNAT1* and *NtSNAT2* might be tobacco *SNAT*s. Additionally, among the 12 candidate *NtSNAT* genes, only *NtSNAT1* and *NtSNAT2* have five conserved motifs, which is consistent with the findings of typical *SNAT*s. Moreover, *NtSNAT1* and *NtSNAT2* clustered together with the classic *SNAT*s, while the rest of the *NtSNAT*s clustered separately from the identified *SNAT*s. Therefore, these findings further confirmed our speculation.

Previous studies have shown that plant *SNAT*s are vertically transferred to descendants through endosymbiosis (Kang et al., 2013; Lee et al., 2014b; Byeon et al., 2015; Wang et al., 2020). The evolution of *SNAT* genes from cyanobacteria to higher plants verified that *NtSNAT1* and *NtSNAT2* indeed evolved vertically through endosymbiosis. From bacteria to higher plants, SNAT proteins are ubiquitous, which means that SNAT proteins may have evolved in the early stages of biological history. Aside from cyanobacteria, we also identified *SNAT* homologous genes from the purple non-sulfur bacteria *R. rubrum*, which has been proven to be able to synthesize melatonin (Tilden et al., 1997; Tan et al., 2013). Interestingly, *R. rubrum* is one of the most ancestral species of living organisms and is the first photosynthetic α-proteobacterium shown to synthesize melatonin (Manchester et al., 1995), despite only motif 1 being present in the *SNAT*s of *R. rubrum*. These facts, when taken together with all the tested species of common *SNAT* genes containing conserved motif 1, suggest that motif 1 is probably the core factor of *SNAT* genes.

Whole-genome duplication or polyploidization is an important driver of adaptation and speciation in plants (Hovav et al., 2008; Liu et al., 2019). In the polyploid genome, there are a large number of duplicated genes and duplicated genes from different diploid ancestors. Due to rearrangement or deletion of these duplicated genes, there are approximately three different fates of a polyploid genome, namely, subfunctionalization, pseudogenization or functional diversification (Adams et al., 2003; Hovav et al., 2008; Sojli et al., 2020). Therefore, it is speculated that the evolution of tobacco from being diploid (the two ancestral species of which were *N. tomentosiformis* and *N. sylvestris*) to allotetraploid may have caused the rearrangement or deletion of different genes due to gene duplication. In the present investigation, each of the two ancestor diploids has two possible *SNAT*s; however, only two typical *SNAT*s, *NtSNAT1* and *NtSNAT2*, were identified in tobacco. These results suggest that the typical *SNAT*s currently in tobacco may have arisen by gene loss during the process of genomic stabilization following polyploidization or whole-genome duplication. Ten additional *NtSNAT*-like genes branching independently from the typical *SNAT*s were identified, and the proteins encoded by these genes had markedly fewer motifs than *NtSNAT1* and *NtSNAT2* did or the representative *SNAT* from cyanobacteria did. Therefore, motif 1 has been stable throughout the evolutionary process.

To explore the possible functional differences of *NtSNAT* and *NtSNAT*-like genes, their expression patterns in different tissues and under different stresses were determined. The results demonstrated different types of expression patterns among these genes. With respect to the expression in different organs, 2 *NtSNAT* and 10 *NtSNAT*-like genes were expressed in the roots, stems, leaves, flowers, and fruits, indicating that these genes may have potential effects in vegetative and reproductive growth. Moreover, the high expression of *NtSNAT1* and *NtSNAT2* in the leaves indicates that these two genes could be pivotal in leaf growth and development. In addition, studies have shown that *SNAT*s play an important role in the process of plant stress resistance (Wang et al., 2017). The SNAT enzyme is involved in the biosynthesis of melatonin, which is reported to regulate the thermotolerance of many plant species. For example, the cyanobacteria *SNAT* gene was proven to be involved in melatonin in response to high temperature (Byeon et al., 2013). Under heat-stress conditions, *SlSNAT* interacts with HSP40 to maintain melatonin levels, thereby increasing the heat resistance of tomato plants (Wang et al., 2020). In this study, *NtSNAT1*, *NtSNAT*-like 3, *NtSNAT*-like 5, *NtSNAT*-like 6, and *NtSNAT*-like 10 responded to heat stress at different heat treatment time points, indicating that these genes may have potential effects in increasing heat resistance. Similarly, Lee and Back (2017) showed that overexpression of *OsSNAT* in rice can significantly increase plant resistance to cadmium stress and senescence. In this paper, under cadmium-stress conditions, *NtSNAT1*, *NtSNAT*-like 3, *NtSNAT*-like 10, *NtSNAT*-like 11, and *NtSNAT*-like 12 were found to respond to cadmium stress at different time points, suggesting that these genes potentially associated with increasing plant tolerance to cadmium. Notably, the expression of *NtSNAT2*, a representative *SNAT* gene, was not significantly upregulated compared with that in the control group under all the stress treatments. Accordingly, it is speculated that *NtSNAT2* may not respond during the actual stress time and may be expressed at other time points. The exact role of *NtSNAT1* and *NtSNAT2* and the catalytic activities of their encoded proteins require further study.

CONCLUSION

Serotonin *N*-acetyltransferase is a key enzyme in the melatonin biosynthesis pathway. *NtSNAT1* and *NtSNAT2*, together with 10 additional *NtSNAT*-like genes, were identified as candidate genes for improving tobacco production. Among the five motifs typically present within *SNAT*s, motif 1 is indispensable for melatonin biosynthesis. During the evolutionary process through which tobacco changed from being diploid to allotetraploid, *NtSNAT1* and *NtSNAT2* were retained due to gene rearrangement or deletion during genome stabilization after whole-genome duplication and polyploidization. *NtSNAT1* potentially associated with regulating plant growth and development and increasing plant tolerance to stress.

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

DZ and SC designed the experiments. JZ designed and carried out the experiments and wrote the manuscript. ZY, RZ, and ZM participated and analyzed the data from the experiments. All the authors reviewed and approved the manuscript in its final form.

FUNDING

This work was supported financially by the National Natural Science Foundation of China (Nos. 31960082 and 81560622) and the Key Project of China National Tobacco Corporation's Yunnan Provincial Company Science and Technology Plan (201853010020082).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.591984/full#supplementary-material

Supplementary Figure 1 | Conserved amino acid sequences of 12 motifs.

Supplementary Table 1 | Accession number and source organism of 35 SNATs from bacteria, algae, mosses, ferns, gymnosperms, and angiosperms.

Supplementary Table 2 | Primers used for the qRT-PCR assay of 12 candidate *NtSNAT*s.

Supplementary Table 3 | Full amino acid sequences of 12 candidate *NtSNAT*s.

Supplementary Table 4 | *NtSNAT* and *NtSNAT*-like genes and their related Gene Ontology (GO) terms.

REFERENCES

Adams, K. L., Richard, C., Ryan, P., and Wendel, J. (2003). Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc. Natl. Acad. Sci. U.S.A.* 100, 4649–4654. doi: 10.1073/pnas.0630618100

Arnao, M. B., and Hernández-Ruiz, J. (2017). Melatonin and its relationship to plant hormones. *Ann. Bot.* 121, 195–207. doi: 10.1093/aob/mcx114

Arnao, M. B., and Hernández-Ruiz, J. (2019). Melatonin: a new plant hormone and/or a plant master regulator? *Trends Plant Sci.* 24, 38–48. doi: 10.1016/j.tplants.2018.10.010

Back, K., Tan, D. X., and Reiter, R. J. (2016). Melatonin biosynthesis in plants: multiple pathways catalyze tryptophan to melatonin in the cytoplasm or chloroplasts. *J. Pineal Res.* 61, 426–437. doi: 10.1111/jpi.12364

Byeon, Y., and Back, K. (2014). Melatonin synthesis in rice seedlings *in vivo* is enhanced at high temperatures and under dark conditions due to increased serotonin *N*-acetyltransferase and *N*-acetylserotonin methyltransferase activities. *J. Pineal Res.* 56, 189–195. doi: 10.1111/jpi.12111

Byeon, Y., and Back, K. (2016). Low melatonin production by suppression of either serotonin *N*-acetyltransferase or *N*-acetylserotonin methyltransferase in rice causes seedling growth retardation with yield penalty, abiotic stress susceptibility, and enhanced coleoptile growth under anoxi. *J. Pineal Res.* 60, 348–359. doi: 10.1111/jpi.12317

Byeon, Y., Lee, H., Choi, D. W., and Back, K. (2015). Chloroplast-encoded serotonin *N*-acetyltransferase in the red alga *Pyropia yezoensis*: gene transition to the nucleus from chloroplasts. *J. Exp. Bot.* 66, 709–717. doi: 10.1093/jxb/eru357

Byeon, Y., Lee, H. Y., and Back, K. (2016a). Cloning and characterization of the serotonin *N*–acetyltransferase–2 gene (*SNAT2*) in rice (*Oryza sativa*). *J. Pineal Res.* 61, 198–207. doi: 10.1111/jpi.12339

Byeon, Y., Lee, H. J., Lee, H. Y., and Back, K. (2016b). Cloning and functional characterization of the *Arabidopsis N*-acetylserotonin *O*-methyltransferase responsible for melatonin synthesis. *J. Pineal Res.* 60, 65–73. doi: 10.1111/jpi.12289

Byeon, Y., Lee, K., Park, Y. I., Park, S., and Back, K. (2013). Molecular cloning and functional analysis of serotonin *N*-acetyltransferase from the cyanobacterium *Synechocystis* sp. PCC 6803. *J. Pineal Res.* 55, 371–376. doi: 10.1111/jpi.12080

Chen, Q., Qi, W. B., Reiter, R. J., Wei, W., and Wang, B. M. (2009). Exogenously applied melatonin stimulates root growth and raises endogenous indoleacetic acid in roots of etiolated seedlings of *Brassica juncea*. *J. Plant Physiol.* 166, 324–328. doi: 10.1016/j.jplph.2008.06.002

Cho, E. K., and Hong, C. B. (2006). Over-expression of tobacco *NtHSP70-1* contributes to drought-stress tolerance in plants. *Plant Cell Rep.* 25, 349–358. doi: 10.1007/s00299-005-0093-2

Debnath, B., Islam, W., Li, M., Sun, Y., Lu, X., and Mitra, S. (2019). Melatonin mediates enhancement of stress tolerance in plants. *Int. J. Mol. Sci.* 20:1040. doi: 10.3390/ijms20051040

Dyda, F., Klein, D. C., and Hickman, A. B. (2000). GCN5-related *N*-acetyltransferases: a structural overview. *Annu. Rev. Bioph. Biomol.* 29, 81–103. doi: 10.1146/annurev.biophys.29.1.81

Hardeland, R. (2016). Melatonin in plants-diversity of levels and multiplicity of functions. *Front. Plant Sci.* 7:198. doi: 10.3389/fpls.2016.00198

Hernández, I. G., Gomez, F. J. V., Cerutti, S., Arana, M. V., and Silva, M. F. (2015). Melatonin in *Arabidopsis thaliana* acts as plant growth regulator at low concentrations and preserves seed viability at high concentrations. *Plant Physiol. Biochem.* 94, 191–196. doi: 10.1016/j.plaphy.2015.06.011

Hernández-Ruiz, J., Cano, A., and Arnao, M. B. (2004). Melatonin: a growth-stimulating compound present in lupin tissues. *Planta* 220, 140–144. doi: 10.1007/s00425-004-1317-3

Hovav, R., Udall, J. A., Chaudhary, B., Rapp, R., Flagel, L., and Wendel, J. F. (2008). Partitioned expression of duplicated genes during development and evolution of a single cell in a polyploid plant. *Proc. Natl. Acad. Sci. U.S.A.* 105, 6191–6195. doi: 10.1073/pnas.0711569105

Kang, K., Lee, K., Park, S., Byeon, Y., and Back, K. (2013). Molecular cloning of rice serotonin *N*-acetyltransferase, the penultimate gene in plant melatonin biosynthesis. *J. Pineal Res.* 55, 7–13. doi: 10.1111/jpi.12011

Kanwar, M. K., Yu, J., and Zhou, J. (2018). Phytomelatonin: recent advances and future prospects. *J. Pineal Res.* 65:e12526. doi: 10.1111/jpi.12526

Kolář, J., Johnson, C. H., and Macháčková, I. (2010). Exogenously applied melatonin (*N*-acetyl-5-methoxytryptamine) affects flowering of the short-day plant *Chenopodium rubrum*. *Physiol. Plant* 118, 605–612. doi: 10.1034/j.1399-3054.2003.00114.x

Lee, H. Y., Byeon, Y., and Back, K. (2014a). Melatonin as a signal molecule triggering defense responses against pathogen attack in *Arabidopsis* and tobacco. *J. Pineal Res.* 57, 262–268. doi: 10.1111/jpi.12165

Lee, H. Y., Byeon, Y., Lee, K., Lee, H., and Back, K. (2014b). Cloning of *Arabidopsis* SNAT and its role with caffeic acid *O*-methyltransferase in the biosynthesis of melatonin *in vitro* despite their different subcellular localizations. *J. Pineal Res.* 57, 418–426. doi: 10.1111/jpi.12181

Lee, H. Y., Byeon, Y., Tan, D. X., Reiter, R. J., and Back, K. (2015). *Arabidopsis* serotonin *N*-acetyltransferase knockout mutant plants exhibit decreased melatonin and salicylic acid levels resulting in susceptibility to an avirulent pathogen. *J. Pineal Res.* 58, 291–299. doi: 10.1111/jpi.12214

Lee, K., and Back, K. (2017). Overexpression of rice serotonin *N*-acetyltransferase 1 in transgenic rice plants confers resistance to cadmium and senescence and increases grain yield. *J. Pineal Res.* 62:e12392. doi: 10.1111/jpi.12392

Li, C., He, Q., Zhang, F., Yu, J. W., Li, C., and Zhao, T. L. (2019). Melatonin enhances cotton immunity to Verticillium wilt via manipulating lignin and gossypol biosynthesis. *Plant J.* 100, 784–800. doi: 10.1111/tpj.14477

Lim, K. Y., Matyasek, R., Kovarik, A., and Leitch, A. (2004). Genome evolution in allotetraploid *Nicotian*a. *Biol. J. Linn. Soc.* 82, 599–606. doi: 10.1111/j.1095-8312.2004.00344.x

Liu, H. M., Ekrt, L., Kouteck, P., Pellicer, J., Hidalgo, O., and Marquardt, J. (2019). Polyploidy does not control all: lineage-specific average chromosome

length constrains genome size evolution in ferns. *J. Syst. Evol.* 57, 418–430. doi: 10.1111/jse.12525

Liu, W., Zhao, D., Zheng, C., Chen, C., Xin, P., and Yuan, C. (2017). Genomic analysis of the ASMT gene Family in *Solanum lycopersicum*. *Molecules* 22:1984. doi: 10.3390/molecules22111984

Liu, H., Zhou, X., Dong, N., Liu, X., Zhang, H., and Zhang, Z. (2011). Expression of a wheat MBY gene in transgenic tobacco enhances resistance to *Ralstonia solanacearum*, and to drought and salt stresses. *Funct. Integ. Genom.* 11, 431–443. doi: 10.1007/s10142-011-0228-1

Ma, Y., Zhang, L., Zhang, J., Chen, J., Wu, T., Zhu, S., et al. (2014). Expressing a *Citrus* ortholog of *Arabidopsis* ERF1 enhanced cold-tolerance in tobacco. *Sci. Hortic.* 174, 65–76. doi: 10.1016/j.scienta.2014.05.009

Manchester, L. C., Poeggeler, B., Alvares, F. L., Ogden, G. B., and Reiter, R. J. (1995). Melatonin immunoreactivity in the photosynthetic prokaryote *Rhodospirillum rubrum*: implications for an ancient antioxidant system. *Cell. Mol. Biol. Res.* 41, 391–395. doi: 10.1016/0008-8749(95)80021-A

Park, S., Byeon, Y., Lee, H. Y., Kim, Y. S., Ahn, T., and Back, K. (2014). Cloning and characterization of a serotonin N-acetyltransferase from a gymnosperm, loblolly pine (*Pinus taeda*). *J. Pineal Res.* 57, 348–355. doi: 10.1111/jpi.12174

Shi, H., Chen, Y., Tan, D. X., Reiter, R. J., Chan, Z., and He, C. (2015a). Melatonin induces nitric oxide and the potential mechanisms relate to innate immunity against bacterial pathogen infection in *Arabidopsis*. *J. Pineal Res.* 59, 102–108. doi: 10.1111/jpi.12244

Shi, H., Qian, Y., Tan, D. X., Reiter, R. J., and He, C. (2015b). Melatonin induces the transcripts of CBF/DREB1s and their involvement in both abiotic and biotic stresses in *Arabidopsis*. *J. Pineal Res*. 59, 334–342. doi: 10.1111/jpi.12262

Sojli, L., Sehyun, C., Jeon, D., Donghyun, J., Yuna, K., and Chang, K. (2020). Evolutionary impact of whole genome duplication in Poaceae family. *J. Crop Sci. Biotechnol.* 57, doi: 10.1007/s12892-020-00049-2

Tal, O., Haim, A., Harel, O., and Gerchman, Y. (2011). Melatonin as an antioxidant and its semi-lunar rhythm in green macroalga *Ulva* sp. *J. Exp. Bot.* 62, 1903–1910. doi: 10.1093/jxb/erq378

Tan, D. X., Hardeland, R., Manchester, L. C., Korkmaz, A., Ma, S., Rosales-Corral, S., et al. (2011). Functional roles of melatonin in plants, and perspectives in nutritional and agricultural science. *J. Exp. Bot.* 63, 577–597. doi: 10.1093/jxb/err256

Tan, D. X., Manchester, L. C., Liu, X., Rosales-Corral, S. A., Acuna-Castroviejo, D., and Reiter, R. J. (2013). Mitochondria and chloroplasts as the original sites of melatonin synthesis: a hypothesis related to melatonin's primary function and evolution in eukaryotes. *J. Pineal Res.* 54, 127–138. doi: 10.1111/jpi.12026

Tilden, A. R., Becker, M. A., Amma, L. L., Arciniega, J., and McGaw, A. K. (1997). Melatonin production in an aerobic photosynthetic bacterium: an evolutionarily early association with darkness. *J. Pineal Res.* 22, 102–106. doi: 10.1111/j.1600-079X.1997.tb00310.x

Tong, Z., Zhou, J., Xiu, Z., Jiao, F., Hu, Y., Zheng, F., et al. (2019). Construction of a high-density genetic map with whole genome sequencing in *Nicotiana tabacum* L. *Genomics* 112, 2028–2033. doi: 10.1016/j.ygeno.2019.11.015

Turk, H., and Genisel, M. (2019). Melatonin-related mitochondrial respiration responses are associated with growth promotion and cold tolerance in plants. *Cryobiology* 92, 76–85. doi: 10.1016/j.cryobiol.2019.11.006

Wang, L., Feng, C., Zheng, X., Guo, Y., Zhou, F., Shan, D., et al. (2017). Plant mitochondria synthesize melatonin and enhance the tolerance of plants to drought stress. *J. Pineal Res*. 63:e12429. doi: 10.1111/jpi.12429

Wang, P., Yin, L., Liang, D., Li, C., Ma, F., and Yue, Z. (2012). Delayed senescence of apple leaves by exogenous melatonin treatment: toward regulating the ascorbate-glutathione cycle. *J. Pineal Res.* 53, 11–20. doi: 10.1111/j.1600-079x.2011.00966.x

Wang, X., and Bennetzen, J. L. (2015). Current status and prospects for the study of *Nicotiana* genomics, genetics, and nicotine biosynthesis genes. *Mol. Genet. Genomics* 290, 11–21. doi: 10.1007/s00438-015-0989-7

Wang, X., Zhang, H., Xie, Q., Liu, Y., Lv, H., Bai, R., et al. (2020). *SlSNAT* interacts with HSP40, a molecular chaperone, to regulate melatonin biosynthesis and promote thermotolerance in tomato. *Plant Cell Physiol.* 61, 909–921. doi: 10.1093/pcp/pcaa018

Wei, W., Li, Q. T., Chu, Y. N., Reiter, R. J., Yu, X. M., Zhu, D. H., et al. (2015). Melatonin enhances plant growth and abiotic stress tolerance in soybean plants. *J. Exp. Bot.* 66, 695–707. doi: 10.1093/jxb/eru392

Yu, S., Kakar, K., Yang, Z., Nawaz, Z., Lin, S., Guo, Y., et al. (2020). Systematic study of the stress-responsive Rboh gene family in *Nicotiana tabacum*: genome-wide identification, evolution and role in disease resistance. *Genomics* 112, 1404–1418. doi: 10.1016/j.ygeno.2019.08.010

Yu, Y., Bian, L., Jiao, Z., Yu, K., Wan, Y., Zhang, G., et al. (2019). Molecular cloning and characterization of a grapevine (*Vitis vinifera* L.) serotonin *N*-acetyltransferase (*VvSNAT2*) gene involved in plant defense. *BMC Genomics* 20:880. doi: 10.1186/s12864-019-6085-3

Yu, Y., Lv, Y., Shi, Y., Li, T., Chen, Y., Zhao, D., et al. (2018). The role of phyto-melatonin and related metabolites in response to stress. *Molecules* 23:1887. doi: 10.3390/molecules23081887

Zhang, H. J., Sun, Q. Q., Cao, Y. Y., Li, X., Zhao, B., Wu, P., et al. (2017). Proteomic analysis reveals a role of melatonin in promoting cucumber seed germination under high salinity by regulating energy production. *Sci. Rep.* 7:503. doi: 10.1038/s41598-017-00566-1

Zhang, N., Sun, Q., Zhang, H., Cao, Y., Weeda, S., Ren, S., et al. (2015). Roles of melatonin in abiotic stress resistance in plants. *J. Exp. Bot.* 66, 647–656. doi: 10.1093/jxb/eru336

Zhao, D., Yu, Y., Shen, Y., Liu, Q., Zhao, Z., Sharma, R., et al. (2019). Melatonin synthesis and function: evolutionary history in animals and plants. *Front. Endocrinol.* 10:249. doi: 10.3389/fendo.2019.00249

Zhao, L., Chen, L., Gu, P., Zhan, X., Zhang, Y., Hou, C., et al. (2019). Exogenous application of melatonin improves plant resistance to virus infection. *Plant Pathol.* 68, 1287–1295. doi: 10.1111/ppa.13057

Zhao, Y., Ma, R., Xu, D., Bi, H., Xia, Z., and Peng, H. (2019). Genome-wide identification and analysis of the AP2 transcription factor gene family in wheat (*Triticum aestivum* L.). *Front. Plant Sci.* 10:1286. doi: 10.3389/fpls.2019.01286

Conflict of Interest: HY and TX was employed by the Wenshan Branch of Yunnan Tobacco Company.

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

Copyright © 2020 Zhang, Yao, Zhang, Mou, Yin, Xu, Zhao and Chen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 05 November 2020
doi: 10.3389/fgene.2020.591806

Genome-Wide Identification of Circular RNAs in Response to Low-Temperature Stress in Tomato Leaves

Xuedong Yang[1†], Yahui Liu[1†], Hui Zhang[1†], Jinyu Wang[2], Gaurav Zinta[3], Shangbo Xie[4], Weimin Zhu[1] and Wen-Feng Nie[2,5*]*

[1] Shanghai Key Laboratory of Protected Horticulture Technology, The Protected Horticulture Institute, Shanghai Academy of Agricultural Sciences, Shanghai, China, [2] Department of Horticulture, College of Horticulture and Plant Protection, Yangzhou University, Yangzhou, China, [3] Biotechnology Division, CSIR-Institute of Himalayan Bioresource Technology, Palampur, India, [4] BGI-Shenzhen, Shenzhen, China, [5] Joint International Research Laboratory of Agriculture and Agri-Product Safety, The Ministry of Education of China, Yangzhou University, Yangzhou, China

OPEN ACCESS

Edited by:
Xiaoming Song,
North China University of Science and Technology, China

Reviewed by:
Kailiang Bo,
Chinese Academy of Agricultural Sciences, China
Yue Guo,
Jiangsu Academy of Agricultural Sciences, China

***Correspondence:**
Weimin Zhu
yy17@saas.sh.cn
Wen-Feng Nie
wfnie@yzu.edu.cn

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: 05 August 2020
Accepted: 06 October 2020
Published: 05 November 2020

Citation:
Yang X, Liu Y, Zhang H, Wang J, Zinta G, Xie S, Zhu W and Nie W-F (2020) Genome-Wide Identification of Circular RNAs in Response to Low-Temperature Stress in Tomato Leaves. Front. Genet. 11:591806. doi: 10.3389/fgene.2020.591806

Abiotic stress adversely inhibits the growth and development of plants, by changing the expression of multiple genes. Circular RNAs (circRNAs), as a class of non-coding RNAs, function in transcriptional and posttranscriptional regulation. Yet, the involvement of circRNAs in abiotic stress response is rarely reported. In this study, the participation and function of circRNAs in low-temperature (LT)-induced stress response were investigated in tomato leaves. We generated genome-wide profiles of circRNAs and mRNAs in tomato leaves grown at 25°C room temperature (RT) and 12°C LT. Our results show that 1,830 circRNAs were identified in tomato leaves in both RT and LT treatments, among which 1,759 were differentially induced by the LT treatment. We find that the identified circRNAs are mainly located at exons of genes, but less distributed at introns of genes or intergenic regions. Our results suggest that there are 383 differentially expressed circRNAs predicted to function as putative sponges of 266 miRNAs to target 4,476 mRNAs in total. Moreover, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis assays indicate that multiple pathways were enriched in both differentially expressed genes induced by LT and parental genes of differentially expressed circRNAs induced by LT, revealing the key functions of circRNAs and the corresponding targeted genes in response to LT stress. Our results suggest that circRNAs may be involved in regulating metabolism (i.e., carbohydrate, amino acid, lipid, and energy), signal transduction, and environmental adaptation-related pathways and that these circRNAs were predicted to regulate the expression of transcription factors, genes in signal transduction pathways, and genes related to the Ca^{2+} channel through targeting the corresponding proteins, such as WRKY, NAC, cytochrome P450, and calmodulin binding protein. Taken together, our study uncovers that multiple circRNAs are isolated and differently regulated in response to LT stress and provides the resource and potential networks of circRNA–miRNA–mRNA under LT stress for further investigations in tomato leaves.

Keywords: circRNAs, tomato, cold stress, miRNA sponge, post-transcriptional regulation

INTRODUCTION

Most of the genomes in eukaryotes generate RNA transcripts that do not code for proteins. Short non-coding RNAs (smaller than 200 nt) including small interfering RNAs (siRNAs), microRNAs (miRNAs), and PIWI-interacting RNAs (piRNAs) are well studied in the regulation of gene expression (Zhang et al., 2018). Long non-coding RNAs (lncRNAs) regulate multiple gene expressions that are enriched in various biological processes in eukaryotes (Zaynab et al., 2018). The lncRNAs are the key regulators of chromatin remodeling and genome pattern, RNA stabilization, and transcription regulation. The lncRNAs are well known to function as important regulators in response to abiotic stress response in plants. Circular RNAs (circRNAs) are recently discovered endogenous non-coding RNAs (Salzman et al., 2012; Memczak et al., 2013; Wang et al., 2014). The circRNAs are derived from linear precursor mRNAs, which can be generated from circularization of lariat-driven and intron-pairing-driven, alternative splicing, and protein factors associated circulation (Zhang et al., 2014; Starke et al., 2015; Chen, 2016; Liu J. et al., 2017). According to the original positions in pre-mRNAs, the circRNAs are classified into three groups: exonic, intronic, and intergenic circRNAs (Zhang et al., 2013; Jeck and Sharpless, 2014). Although circRNAs were firstly discovered in humans by electron microscopy nearly three decades ago, circRNAs had been previously considered as a rare existence in nature and as transcriptional noise or artifacts for a long time (Salzman, 2016). The circRNAs are generated in the nucleus and then transferred to the cytoplasm (Hansen et al., 2013; Memczak et al., 2013; Jeck and Sharpless, 2014). CircRNAs are highly abundant and stable, compared to their linear counterparts (Hansen et al., 2011; Salzman et al., 2012, 2013; Jeck et al., 2013; Memczak et al., 2013), suggesting their potential biological significance. Recent studies reveal that circRNAs are conserved among species and are mainly expressed in a cell-, tissue-, and developmental stage-dependent manner (Jeck et al., 2013; Memczak et al., 2013; Salzman et al., 2013; Szabo et al., 2015; Venø et al., 2015).

With the development of high-throughput sequencing methods, circRNAs have been isolated in bacteria (Danan et al., 2012), fungi (Wang et al., 2014), animals (Zhang et al., 2013; Ashwal-Fluss et al., 2014), and humans (Memczak et al., 2013). In plants, circRNAs were firstly discovered in *Arabidopsis thaliana* (Wang et al., 2014; Ye et al., 2015; Conn et al., 2017). After that, circRNAs have been gradually identified in rice (Lu et al., 2015; Ye et al., 2015), barley (Darbani et al., 2016), kiwi (Wang Z.P. et al., 2017), and wheat (Wang et al., 2017b), indicating that circRNAs commonly exist in monocots and dicots. Compared to the circRNAs identified in animals, plant circRNAs likely are shorter in the sequences of the flanking introns (Ye et al., 2015). Moreover, plant circRNAs display the feature of regulating multiple biological pathways. The fact that circRNAs and their linear forms might negatively regulate the posttranscriptional level of their parental gene was further confirmed by overexpression of Os08circ16564 in rice (Lu et al., 2015). Six differentially expressed circRNAs were identified in response to dehydration stress in wheat (Wang et al., 2017b). Recent studies suggest that circRNAs could function as miRNA sponges and thereby impair miRNA-directed gene silencing (Hansen et al., 2013; Memczak et al., 2013). These studies hint that circRNAs plausibly respond to abiotic stress through regulating the expression of stress-related genes.

Cold stress is a major environmental challenge that largely influences the growth and development of plants (Agarwal et al., 2006; Chinnusamy et al., 2006; Jiang et al., 2013). Cold stress is classed into chilling stress (0–20°C) and freezing stress (<0°C; Chinnusamy et al., 2010); low temperature (LT) greatly influences crop productivity, especially in chilling-sensitive crops (Zhou et al., 2011). Cold stress tolerance increases in plants after previous exposure to LT. This mechanism of adaptation is usually called cold acclimation, which results in numerous changes in gene expression (Steponkus, 1984; Guy, 1990; Thomashow, 1999). It has been shown that two calcium/calmodulin-regulated receptor-like cytoplasmic kinases (RLCKs), CRLK1, and CRLK2, positively regulate freezing tolerance (Yang et al., 2010a,b; Zhao C. et al., 2017). Recently, a plasma membrane-localized RLCK, cold-responsive protein kinase 1 (CRPK1), was reported to negatively regulate freezing tolerance (Liu Z. et al., 2017). Recent studies show that the two non-membrane proteins located in the nucleus and cytosol, including the histone variant H2A.Z and the photoreceptor phytochrome B (phyB), probably function as temperature sensing (Kumar and Wigge, 2010; Jung et al., 2016; Legris et al., 2016). The reactive oxygen species (ROS) induced by the cold may activate a mitogen-activated protein kinase cascade (AtMEKK1-AtMKK2-AtMPK4/6) that enhances the tolerance to freezing and other abiotic stresses (Kovtun et al., 2000; Teige et al., 2004). Although circRNAs have been identified and are reported to be involved in tomato fruit ripening, coloration, pigment accumulation, and ethylene pathway; chilling injury on fruits; and *Phytophthora infestans* infection in tomato (Zuo et al., 2016; Tan et al., 2017; Wang et al., 2017a; Yin et al., 2018; Zhou et al., 2018b; Hong et al., 2020), the roles of circRNAs in response to abiotic stresses (i.e., LT) in tomato plants are still unknown.

In this study, we used high-throughput sequencing to identify genome-wide mRNAs and circRNAs in response to LT stress in tomato leaves. We identified a total of 1,830 circRNAs, among which 1,759 were differentially regulated under LT stress. Of these circRNAs, 383 differentially expressed circRNAs were identified to function as putative sponges of 266 miRNAs to target 4,476 mRNAs. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) annotations for parental genes of circRNAs and mRNAs revealed that circRNAs and mRNAs were involved in regulating metabolism (amino acid, carbohydrate, lipid, and energy), signal transduction, and environmental adaptation. Furthermore, these circRNAs were predicted to regulate the genes related to transcription and the Ca^{2+} channel by targeting the associated proteins such as NAC, MYB, WRKY, and calmodulin-binding protein. Hence, our study improves the understanding on the function of circRNAs in cold response in tomato plants.

MATERIALS AND METHODS

Plant Materials and Cold Stress Treatment

The wild-type tomato (*Solanum lycopersicum*) "cv 1479" were used to perform LT stress assays. The seeds of tomato were germinated at trays with vermiculite substrate, and then the seedlings were transferred to pots. The well-grown and 2-week-old seedlings were divided into two groups (n = 12). The two groups were treated with room temperature (25°C/22°C, day/night, 14-h photoperiod, referred to as RT) and low temperature (15°C/12°C, day/night, 14-h photoperiod, referred to as LT), with 60–70% relative humidity for 3 weeks, respectively. After the treatments, the fifth leaf of each plant was harvested and frozen immediately in liquid nitrogen for further analysis. Three biological replicates were performed, and the fifth leaves from four individual plants were mixed and referred to as one biological replicate.

Library Preparation for circRNA-seq and mRNA-seq

All the sequencing was performed at BGI, China. The total RNAs were isolated from the leaves of tomato plants using the TRIzol reagent (Invitrogen, United States) according to the manufacturer's instruction. The concentration and purity of total RNAs were checked by the NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, United States). Ribosomal RNA (rRNA) was removed by Ribo-Zero rRNA Removal Kits (Illumina, United States). The RNA integrity was checked by the Agilent 2100 Bioanalyzer Lab-on-Chip system (Agilent Technologies, United States). The RNAs were further incubated at 37°C for 1 h for the digestion with RNase R (Epicentre, WI, United States). The purified RNAs were used as templates for the construction of cDNA libraries according to the protocol for the mRNA-seq sample preparation kit (Illumina, United States). The clustering of samples was performed on a cBot Cluster Generation System using TruSeq PE Cluster Kit v3-cBot-HS (Illumina, United States) following the manufacturer's procedure. The paired-end sequencing was detected with the Illumina HiSeq X Ten platform (BGI, Wuhan, China).

Read Alignment and Identification of circRNA

The clean reads of three biological replicates in RT and LT samples were assembled into two data libraries. The low-quality reads, including unknown bases greater than 5%, those containing more than 50% bases with $Q \leq 20$ ploy-N, and adaptor sequences, were removed. Q20, Q30, and GC contents were calculated. Reads from every sample were mapped to the *S. lycopersicum* reference genome[1] using TopHat version 2.1.1 (Kim et al., 2013). The SAM files were then treated by CIRI (v2.05; Gao et al., 2015) and find_circ (v1.2) with genomic annotations from International Tomato Annotation Group 2.4 (ITAG2.4) to identify circRNAs. The CIRI software scans SAM files twice and then collects the related information to isolate the circRNAs. CIRI exactly detects the junction reads with PCC signals that reflect the candidate of a circRNA during the first scanning of SAM alignment. Preliminary filtering is performed by the paired-end mapping (PEM) and GT-AG splicing signals for the junctions. When the clustering of junction reads and recording of each circRNA candidate are finished, CIRI scans the SAM alignment one more time to detect additional junction reads and simultaneously performs further filtering to remove the false-positive candidates generated from the reads incorrectly mapped to the homologous genes or repetitive sequences. The circRNA calling was done by using find-circ software, meeting the following conditions: GU/AG, on the sides of splice site; clear breakpoint; two mismatches; appearance of breakpoint in the position within 2 nucleotides (nt); more than two reads supporting the junction; a score of blasting to the right position of short sequence that is at least 35 higher than blasting to the other positions. Finally, identified circRNAs are outputted with annotation information.

Expression Level Analysis of circRNAs

The expression levels of circRNAs were calculated by the number of junction reads at both ends of the circRNAs identified by CIRI (v2.0.5) and find_circ (v1.2) tools. Since two software, CIRI and find_circ, were used for the prediction, the final junction reads were averaged. We used the DESeq2 algorithm to detect differentially expressed circRNAs. Genes with an absolute value of log2 (fold change) ≥ 1 and P value < 0.05 found by DESeq2 were defined as differentially expressed genes. The expression levels of each circRNA identified in three biological replicates was averaged to get the final expression value of the corresponding circRNA.

circRNA Annotation and GO Enrichment

The function of genes was annotated according to the following databases: Swiss-Prot (a manually annotated and reviewed protein sequence database), NCBI non-redundant protein sequences, KOG/COG (Clusters of Orthologous Groups of proteins), Pfam (protein family), KEGG, and GO. The corresponding parental genes of differentially expressed circRNAs were collected. The GO enrichment assays were performed using topGO R packages.

Prediction for circRNA–miRNA–mRNA Relationships

The miRNA binding sites of tomato circRNAs upon the alignment against miRBase21.0[2] were predicted by miRanda (3.3a) and TargetScan (V7.0; Kozomara and Griffiths-Jones, 2014). The FASTA-formatted sequences of predicted target miRNAs were detected by a website[3] to find the corresponding target coding genes (*S. lycopersicum*, transcript, cDNA library, version 2.4). All the results were visualized by Cytoscape (v3.5.0) to clearly display the relationships among the circRNA–miRNA–mRNA (Shannon et al., 2003).

[1] ftp://ftp.solgenomics.net/tomato_genome/assembly/build_3.00/, version SL3.0.

[2] http://www.mirbase.org/

[3] http://plantgrn.noble.org/psRNATarget/

RESULTS

Identification of circRNAs in Tomato Leaves

In order to isolate potential circRNAs in tomato leaves, polyA-depleted RNA libraries from 5-week-old tomato leaves grown at RT and LT were constructed and sequenced by the Illumina HiSeq X Ten platform. In total, about 200 million paired-end raw reads were generated from samples in RT and LT (**Supplementary Table S1**). Low-quality reads were firstly filtered, and clean reads were then mapped to reference genome *S. lycopersicum* (SL3.0), resulting in about 160 million unique mapped reads in RT and LT samples used for further analysis (**Supplementary Table S1**).

There were 1,830 circRNAs identified, 530 of which existed in both RT and LT samples (**Figure 1A** and **Supplementary Table S2**). Among the identified circRNAs, 1,709 (93.39%) were exonic circRNAs, 90 (4.92%) were intergenic regions (intergenic circRNAs), and 31 (1.69%) were intronic circRNAs (**Figure 1B**). The length-dependent count analysis showed that more than half of the circRNAs are shorter than 2 kb (**Figure 1C**). The circRNAs identified in RT treatment and LT treatment displayed similar expression patterns on tomato chromosomes (**Figure 1D**). These results indicated that circRNAs in tomato leaves were generated from different genomic regions and chromosomes, showing the characterization of various lengths.

LT Treatment Induces the Expression of circRNAs in Tomato

To test the induction of LT on the expression of circRNAs in tomato leaves, we compared the expression of circRNAs between the treatments of RT and LT. We found that 1,759 circRNAs were significantly expressed in the LT treatment, compared to RT, and that 1,115 and 644 circRNAs were upregulated and downregulated, respectively, (**Figures 2A,B** and **Supplementary Table S2**). The distribution of circRNAs, mRNA, and miRNA on chromosomes was displayed (**Figure 2A**), indicating a high association among the circRNAs, miRNA, and the targeted mRNA. In terms of the junction read number, the expression of circRNAs in samples at LT were higher than that at RT by boxplot representation (**Figure 2C**). These results suggest that LT treatment greatly induced the expression of circRNAs in tomato leaves.

Biological Function Analysis of circRNAs Induced by LT

To understand the function of the circRNAs, GO and KEGG analyses were performed in the parental genes of the circRNAs and in the differentially expressed genes induced by LT. By GO analysis, we found that multiple pathways were enriched in the parental genes of the circRNAs induced by LT treatment; among these pathways, nucleotide binding and purine binding were significantly enriched (**Figure 3A**). Pathways related to abiotic stress, DNA binding, regulation of biological process, and transcription regulator activity were enriched in the differentially expressed genes induced by LT treatment (**Figure 3B**).

Kyoto Encyclopedia of Genes and Genomes pathway analysis showed that the parental genes of circRNAs (**Figure 4A**) and LT-induced genes (**Figure 4B**) were mostly enriched in transport and catabolism, signal transduction, metabolism (carbohydrate, amino acid, lipid, energy, nucleotide, terpenoids, and polyketides), and environmental adaptation. The results suggested that circRNAs likely participated in cellular processes and metabolic processes. Notably, the specific and detailed function of circRNAs needs to be investigated in further studies.

circRNAs Function As Putative miRNA Sponges to Regulate Transcription of Genes

The circRNAs function to target miRNAs and inhibit the activity of the miRNAs. The miRNAs have been shown to take part in several physiological processes by controlling the expression of genes in plants. To reveal the role of circRNAs in response to LT in tomato plants, the miRNA targets of circRNAs were predicted, and the circRNA–miRNA–mRNA networks were generated. Among these identified circRNAs in RT and LT samples, 383 differentially expressed circRNAs were predicted to function as the putative sponges of 266 miRNA to further regulate 4,476 targeted mRNAs (**Figure 5** and **Supplementary Table S3**). We found that targeted mRNAs included various genes that were presented in transcription and signaling transduction pathways, such as several transcription factor families and proteins including calcium (Ca^{2+}) channels and receptor-like kinases (RLKs), cytochrome P450, and E3 ubiquitin-protein ligase (**Supplementary Table S4**). Many miRNAs targeted transcription factors, including NAC, MYB, WRKY, WD40, zinc finger transcription factor, bHLH, and Hsf (**Supplementary Table S4**). For example, miR-9479-3p targeted cytochrome P450, MYB, and bHLH; miR-477-3p targeted bHLH, calmodulin, WRKY, and MYB (**Supplementary Table S4**). These results further confirm that circRNAs are critical in regulating transcription by playing as miRNA sponges in response to LT treatment in tomato leaves.

DISCUSSION

The circRNAs are a novel class of non-coding RNAs that are commonly explored. Various circRNAs are highly expressed and evolutionarily conserved (Salzman, 2016). Recently, plant circRNAs have been identified and reported in *Arabidopsis* (Ye et al., 2015), barley (Darbani et al., 2016), rice (Ye et al., 2015), tomato (Zuo et al., 2016), soybean (Zhao W. et al., 2017), wheat (Wang et al., 2017b), cotton (Zhao T. et al., 2017), kiwi (Wang Z.P. et al., 2017), potato (Zhou et al., 2018b), maize (Chen et al., 2018), and orange (Zeng et al., 2018). The identification of circRNAs accelerates the understanding of the functions of circRNAs in different biological processes in plants. Cold stress is a great challenge to plants, inducing loss of yield and quality damage to crops. Moreover, cold response in plants is a complex process, including the processes of inducing expression changes of the calcium channel protein, metabolism-related proteins, and multiple transcription factors (An et al., 2018; Sun et al., 2018; Xie et al., 2018; Fürtauer et al., 2019; Liu et al., 2019;

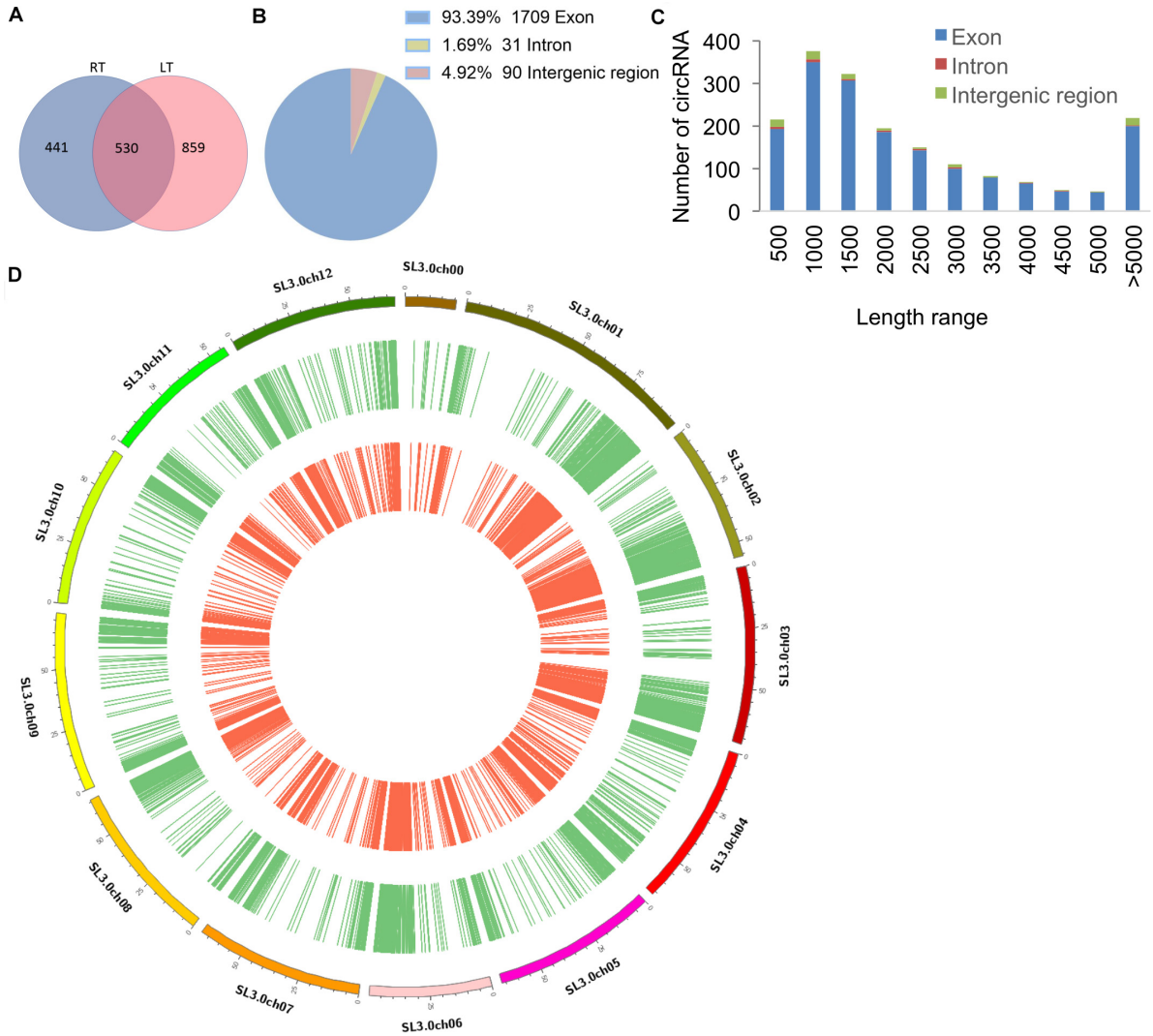


FIGURE 1 | Characterization of tomato circRNAs in different temperature treatments. **(A)** Venn diagram shows the number and distribution of detected circRNAs in tomato samples at room temperature (RT) and low temperature (LT). The chromosomes of *S. lycopersicum* (version SL3.0) are displayed. **(B)** The pie chart represents the classification of the circRNAs according to the location of the genomic region. **(C)** The histogram shows the number of circRNAs in different ranges of length. **(D)** The distribution of all the identified circRNAs in samples of RT and LT (n = 1,830). circRNAs identified in tomato leaves at LT treatment are displayed in green, and circRNAs identified at RT treatment are displayed in red.

Gong et al., 2020). Epigenetic regulations, including DNA methylation, histone modifications (variants), and non-coding RNAs are essential to gene expression (Zhang et al., 2018; Nie et al., 2019). Hence, understanding the epigenetic regulation in response to cold stress in plants is of interest. So far, the roles of circRNAs in the processes of cold tolerance are not well known, especially in thermophilic vegetable crops, such as in tomato plants (Camejo et al., 2005). In this study, we report the isolation of 1,830 circRNAs in the leaves of 5-week-old tomato seedlings and find 859 circRNAs specifically expressed in response to LT treatment (**Figure 1A**). We characterize the whole-genome-wide pattern and the distribution on chromosomes of tomato circRNAs (**Figures 1D**, **2A**). In addition, we clearly map the relationship among differentially expressed mRNAs, miRNAs, and circRNAs under LT treatment (**Figure 2A**). These sequencing data provide much information for further studies on the function of circRNAs in tomato plants, especially on the investigations of tomato plants responding to abiotic stress.

Circular RNAs are classified into three groups: exonic circRNAs, intronic circRNAs, and intergenic circRNAs (Jeck and Sharpless, 2014). In our study, we find that 1,709, 90, and 31 circRNAs were exonic, intergenic, and intronic circRNAs, respectively, (**Figure 1B**). These results are consistent with the previous reports in *Arabidopsis*, rice, and tomato (Lu et al., 2015; Ye et al., 2015; Zuo et al., 2016; Wang et al., 2017a; Pan et al., 2018; Zhou et al., 2018b), suggesting that the circRNA-calling method in our study is reliable. Meanwhile, circRNAs were largely distributed at chromosome 1, consistent with results

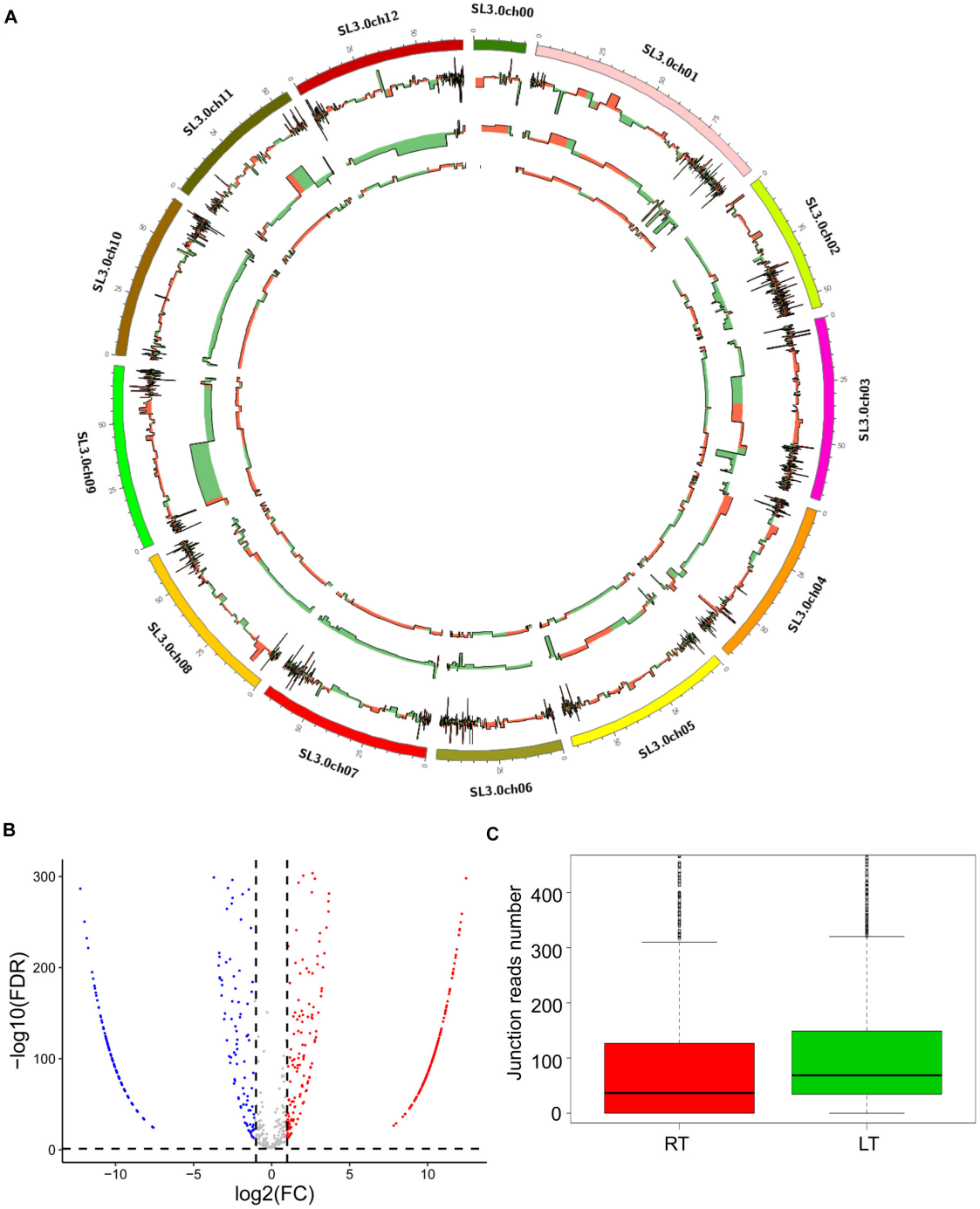


FIGURE 2 | LT treatment induces the expression of various circRNAs in tomato leaves. **(A)** Distribution of differentially expressed mRNA, miRNA, and circRNA on the tomato chromosomes. The first circle (from outside to inside) represents chromosomes of *S. lycopersicum* (version SL3.0), the second circle represents mRNA expression, the third circle represents miRNA expression, and the fourth circle represents circRNA expression. Red indicates upregulation, and green indicates downregulation. **(B)** The expression profile of the circRNAs in the leaves grown at RT and LT. Red dots represent the upregulated circRNAs, blue dots represent the downregulated circRNAs, and gray dots represent the unchanged circRNAs. **(C)** The box plot shows back-splice junction reads at RT and LT treatments.

of previous studies in tomato (Zuo et al., 2016; Wang et al., 2017a; Yin et al., 2018; Zhou et al., 2018a). These results, as well as our methods in identification of circRNAs, contribute to the progress of the development of the field of plant circRNAs.

In tomato, most studies on circRNAs are mainly performed in fruit (Zuo et al., 2016; Wang et al., 2017a; Yin et al., 2018; Zhou et al., 2018a). A previous study reported that circRNAs show ripening-dependent characterizations and that the parental genes

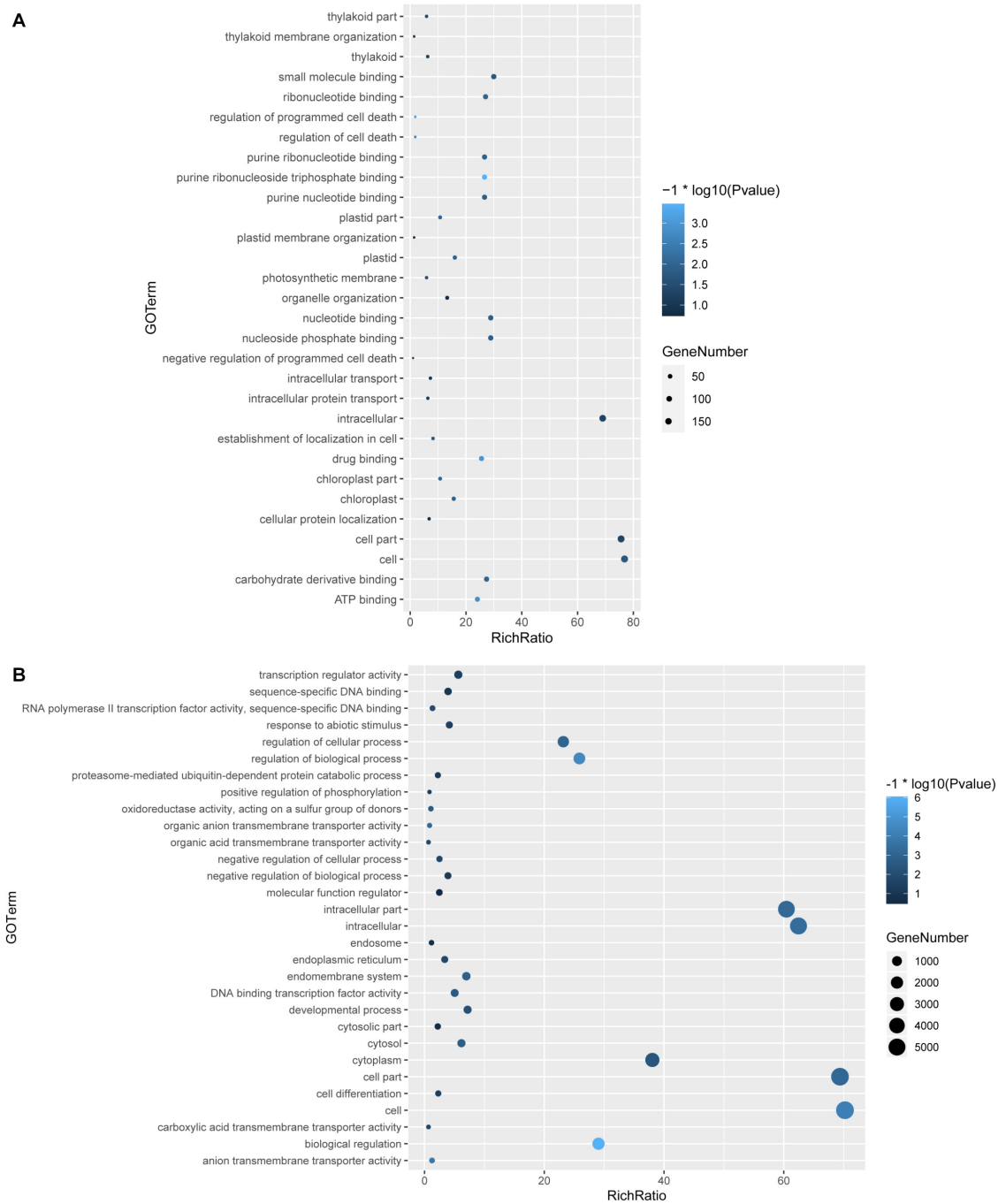


FIGURE 3 | GO analysis of genes and circRNAs induced by LT treatment. **(A)** GO analysis of parental genes of differentially expressed circRNAs under LT treatment. **(B)** GO analysis of differentially expressed genes under LT treatment.

of the circRNAs with significantly different expression levels were mainly involved in metabolic, cellular, and single-organism processes and catalytic and binding activities (Zhou et al., 2018a). It has been shown that several circRNAs were specific in tomato fruits at different developmental stages, which enriches the number of plant circRNAs involved in fruit coloration and ripening (Zhou et al., 2018a). Besides, circRNAs were predicted to be involved in transcription regulation through targeting transcription factors such as ethylene-responsive transcription factor (ERF), squamosa promoter binding-like protein (SBP), and myeloblastosis (MYB) proto-oncogene protein (Yin et al., 2018). These results uncover the role of circRNAs in fruit ripening and isolate the plausible circRNAs in tomato fruits. The circRNAs are likely to regulate fruit ripening by controlling

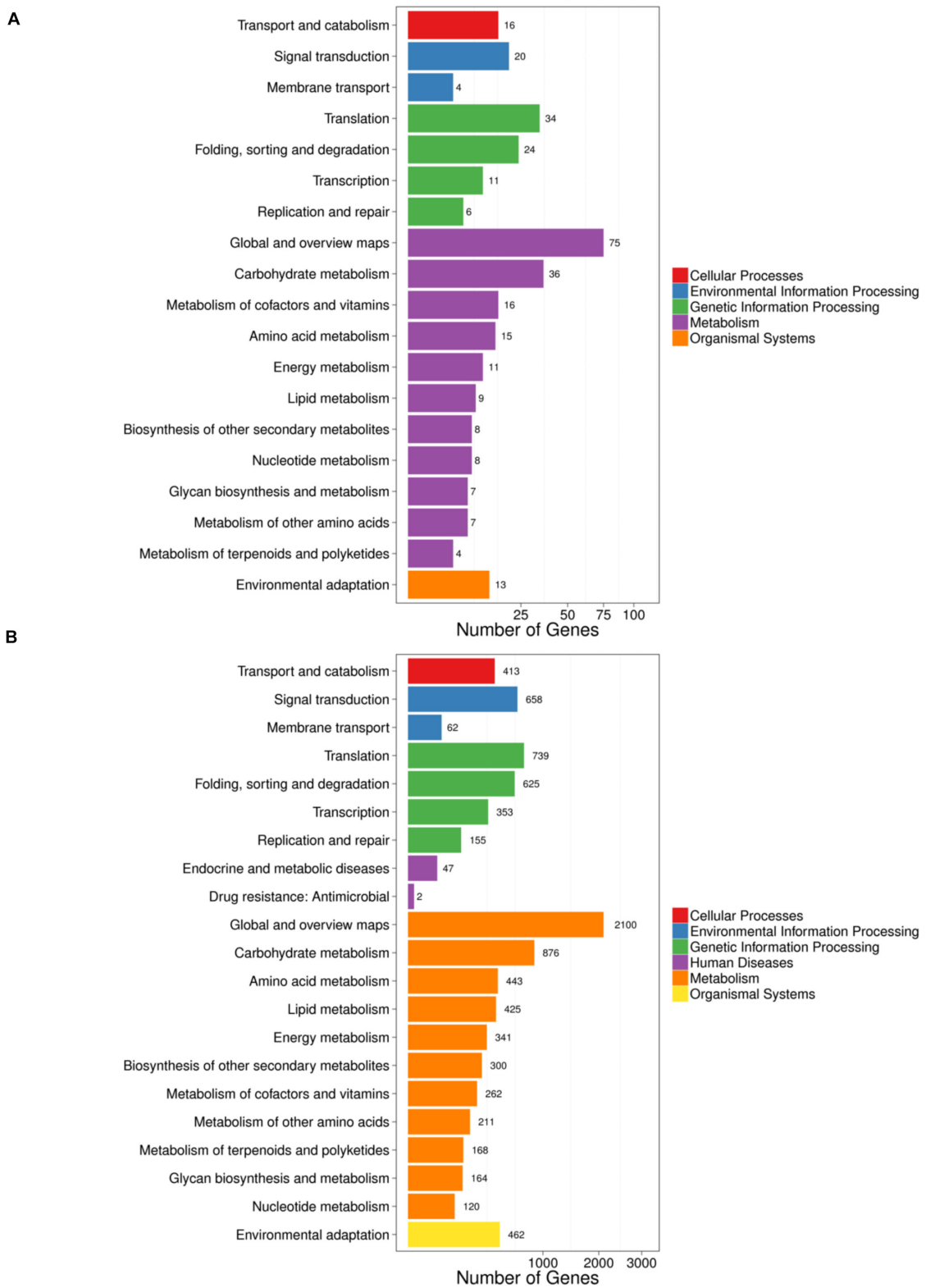


FIGURE 4 | KEGG analysis of genes and circRNAs induced by LT treatment. **(A)** KEGG analysis of parental genes of circRNAs differentially expressed under LT treatment. **(B)** KEGG analysis of differentially expressed genes under LT treatment. The rectangles indicate the mRNAs. The ellipses indicate the circRNAs. Red, upregulated; green, downregulated.

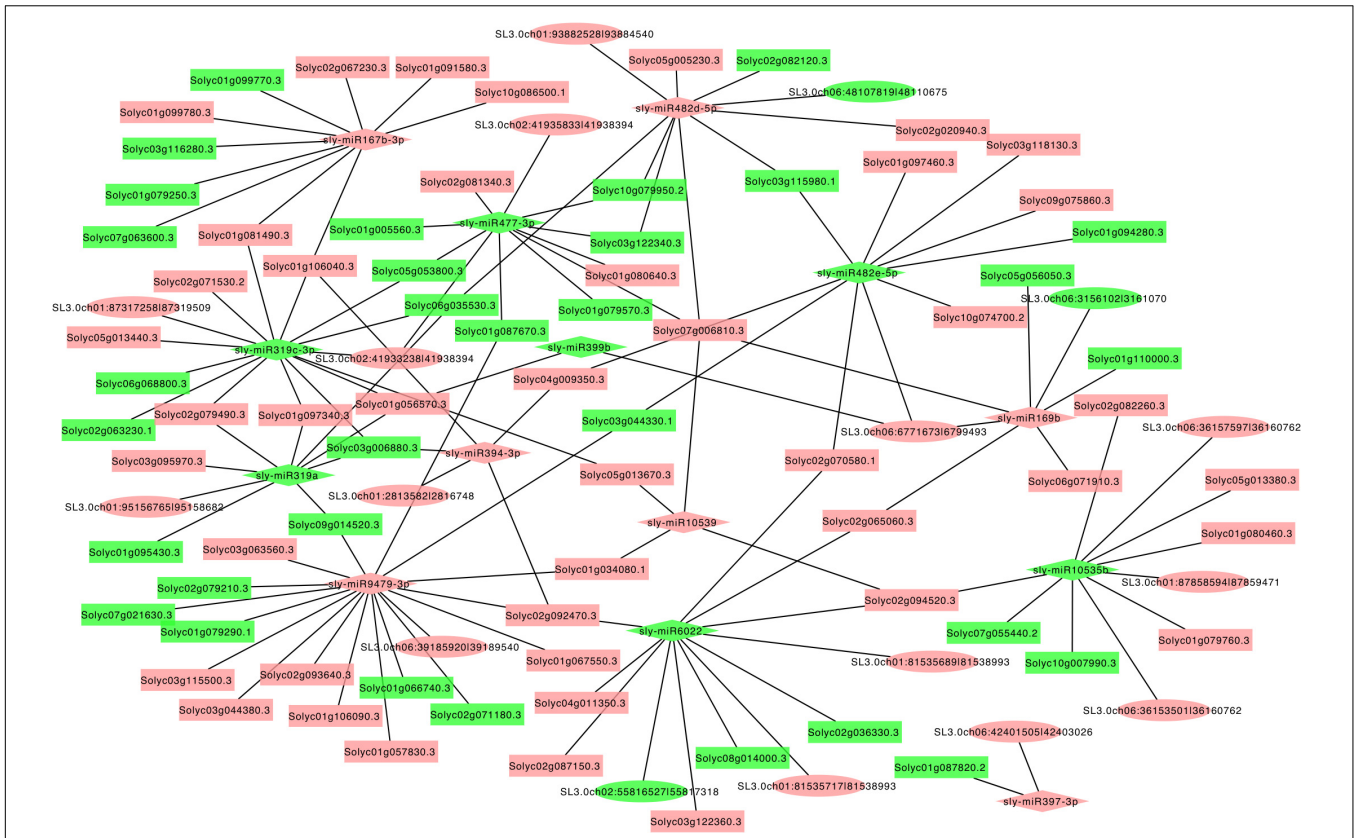


FIGURE 5 | The predicted regulatory networks regulated by circRNAs and the corresponding targeted genes during LT stress response in tomato leaves. The diamonds indicate miRNAs; the ellipses indicate circRNAs; the rectangles indicate targeted genes. Red indicates upregulation, and blue indicates downregulation.

metabolism adaptation, hormone content, and photosynthesis related pathways (Yin et al., 2018). Unlike the investigation of circRNAs in tomato fruits, here, we report the role of circRNAs in LT stress response in tomato leaves, generating genome-wide profiles of circRNAs and mRNAs in tomato leaves with LT treatment. Our results show that 1,759 circRNAs were LT dependent in tomato leaves. The numbers of circRNAs identified in our study is different to that identified in the fruits (Yin et al., 2018; Zhou et al., 2018a), indicating that the abundance and the kind of circRNAs in different tissues are various. Under LT stress treatment, the parental genes of the differentially expressed circRNAs in tomato leaves are enriched in the pathways of regulating metabolism (i.e., carbohydrate, amino acid, lipid, and energy), signal transduction, and environmental adaptation, and these circRNAs were predicted to regulate the expression of genes including transcription factors, signal transduction pathways, and Ca^{2+} channel through targeting the corresponding proteins, such as WRKY, NAC, cytochrome P450, and calmodulin-binding protein. Thus, our results connect the relationship between circRNAs and abiotic stress response in tomato leaves, providing a useful perspective for further understanding the mechanisms of abiotic stress response (i.e., LT) in tomato plants.

Studies have shown that photosynthetic light reactions and the central carbohydrate metabolism are immediately reprogrammed to prevent any imbalances that would cause the production of ROS, cell damage, and cell death in response to changes of temperature (Huner et al., 1998; Hu et al., 2008). Changing temperature affects photosynthesis immediately (Hu et al., 2006; Nägele et al., 2011; Nie et al., 2013). In this study, GO enrichment analysis of parental genes of LT-induced circRNAs indicates that various photosynthesis-related pathways were enriched in LT treatment, such as the thylakoid part, thylakoid membrane organization, regulation of programmed cell death, photosynthetic membrane, and chloroplast (**Figure 3A**). These results suggest that LT-induced circRNAs likely function on regulating photosynthesis progress in response to LT, consistent with the importance of photosynthesis in response to abiotic stress (Nie et al., 2013; Gururani et al., 2015). It is meaningful to test whether circRNAs could regulate the photosynthesis rate and thus further regulate the biomass of tomato plants in the future, which is conveniently confirmed by the genetic mutants by CRISPR. Unlike the GOs enriched in LT-treated tomato leaves, chilling injury treatment on tomato fruits could induce some circRNAs that were predicted to regulate chilling-responsive processes, such as redox reaction, arginine and polyamine metabolism, cell wall degradation, heat and cold shock protein, energy metabolism, and metabolism of jasmonic acid and abscisic acid (Zhou et al., 2018a). Hence, LT-induced circRNAs display tissue-dependent properties and functions.

In rice and *Arabidopsis*, the expression levels of most circRNAs and their corresponding parental genes were positively correlated (Ye et al., 2015). It has been shown that carbohydrates are essential for the metabolic reprogramming during cold acclimation (Fürtauer et al., 2019); freezing tolerance in plants is dependent on lipid remodeling at the outer chloroplast membrane (Moellering et al., 2010). In this study, by KEGG analysis, we found that the parental genes of circRNAs (**Figure 4A**) and differentially expressed genes (**Figure 4B**) share lots of pathways, such as transport and catabolism, signal transduction, metabolism (carbohydrate, amino acid, lipid, energy, nucleotide, terpenoids, and polyketides), and environmental adaptation, suggesting circRNAs and mRNA are coordinately regulated in plants when responding to cold stress. Given the high correlation between circRNAs and the corresponding parental genes, it is possible to investigate the detailed genetic function of these parental genes in the signaling pathway in response to LT stress in further studies.

Several studies reported that circRNAs play important roles in miRNA function and transcriptional control by competing with endogenous RNAs or positive regulators on their parental genes (Hansen et al., 2013; Memczak et al., 2013; Ye et al., 2015). Following the published methods of bioinformatics analysis (Wu et al., 2013; Ye et al., 2014), we found that 383 circRNAs were potential sponges of 266 miRNAs, which target 4,476 mRNAs (**Supplementary Table S3**). These targeted genes of miRNAs include genes participating in signaling transduction pathways, such as calcium (Ca^{2+}) channels and RLKs, cytochrome P450, and E3 ubiquitin-protein ligase (**Supplementary Table S4**). Moreover, multiple miRNAs target transcription factors such as NAC, MYB, WRKY, WD40, zinc finger transcription factor, bHLH, and HSF (**Supplementary Table S4**). For example, we find that miR-9479-3p is predicted to target cytochrome P450, MYB, and bHLH; miR-477-3p is predicted to target bHLH, calmodulin, WRKY, and MYB (**Supplementary Table S4**), indicating that circRNAs likely regulate the expression of transcriptional factors through targeting miRNAs. Although the putative miRNA-binding sites of circRNAs have been identified, genetic evidence for their biological function in tomato leaves needs further investigations. Our results provide a resource to uncover the functions of circRNAs in response to abiotic stress in tomato leaves. It should be noted that further studies on the networks of circRNA–miRNA–mRNA are essential to fully understand the mechanisms of cold stress tolerance mediated by circRNAs.

DATA AVAILABILITY STATEMENT

The sequencing data was uploaded to NCBI and the accession Nos. are SRR12489163–SRR12489168.

AUTHOR CONTRIBUTIONS

XY, WZ, and W-FN designed the research. XY, YL, HZ, and JW performed the research. XY, YL, SX, and W-FN analyzed and interpreted data. XY, YL, and W-FN wrote the manuscript. XY, W-FN, and GZ contributed to the manuscript editing. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Key R&D Program of China (2017YFD0101902 and 2016YFD0101703) and Shanghai Agriculture Applied Technology Development Program, China (Grant No. Z2017030102), and by Basic Research Program of Jiangsu Province (SBK2020041923).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.591806/full#supplementary-material

Supplementary Table 1 | The summary of sequencing quality.

Supplementary Table 2 | The circRNAs identified in tomato plants.

Supplementary Table 3 | The prediction of miRNA-circRNA-mRNA networks.

Supplementary Table 4 | The predicted targets of miRNA and the annotation of the corresponding genes.

REFERENCES

Agarwal, M., Hao, Y., Kapoor, A., Dong, C. H., Fujii, H., Zheng, X., et al. (2006). A R2R3 type MYB transcription factor is involved in the cold regulation of CBF genes and in acquired freezing tolerance. *J. Biol. Chem.* 281, 37636–37645. doi: 10.1074/jbc.m605895200

An, J. P., Li, R., Qu, F. J., You, C. X., Wang, X. F., and Hao, Y. J. (2018). An apple NAC transcription factor negatively regulates cold tolerance via CBF-dependent pathway. *J. Plant Physiol.* 221, 74–80. doi: 10.1016/j.jplph.2017.12.009

Ashwal-Fluss, R., Meyer, M., Pamudurti, N. R., Ivanov, A., Bartok, O., Hanan, M., et al. (2014). CircRNA biogenesis competes with Pre-mRNA splicing. *Mol. Cell* 56, 55–66. doi: 10.1016/j.molcel.2014.08.019

Camejo, D., Rodriguez, P., Morales, M. A., Dell'Amico, J. M., Torrecillas, A., and Alarcon, J. J. (2005). High temperature effects on photosynthetic activity of two tomato cultivars with different heat susceptibility. *J. Plant Physiol.* 162, 281–289. doi: 10.1016/j.jplph.2004.07.014

Chen, L., Zhang, P., Fan, Y., Lu, Q., Li, Q., Yan, J. B., et al. (2018). Circular RNAs mediated by transposons are associated with transcriptomic and phenotypic variation in maize. *New Phytol.* 217, 1292–1306. doi: 10.1111/nph.14901

Chen, L. L. (2016). The biogenesis and emerging roles of circular RNAs. *Nat. Rev. Mol. Cell Biol.* 17, 205–211. doi: 10.1038/nrm.2015.32

Chinnusamy, V., Zhu, J., and Zhu, J.-K. (2006). Gene regulation during cold acclimation in plants. *Physiol. Plant.* 126, 52–61. doi: 10.1111/j.1399-3054.2006.00596.x

Chinnusamy, V., Zhu, J.-K., and Sunkar, R. (2010). Gene regulation during cold stress acclimation in plants. *Methods Mol. Biol.* 639, 39–55. doi: 10.1007/978-1-60761-702-0_3

Conn, V. M., Hugouvieux, V., Nayak, A., Conos, S. A., Capovilla, G., Cildir, G., et al. (2017). A circRNA from SEPALLATA3 regulates splicing of its cognate mRNA through R-loop formation. *Nat. Plants* 3:17053.

Danan, M., Schwartz, S., Edelheit, S., and Sorek, R. (2012). Transcriptome-wide discovery of circular RNAs in Archaea. *Nucleic Acids Res.* 40, 3131–3142. doi: 10.1093/nar/gkr1009

Darbani, B., Noeparvar, S., and Borg, S. (2016). Identification of circular RNAs from the parental genes involved in multiple aspects of cellular metabolism in barley. *Front. Plant Sci.* 7:776. doi: 10.3389/fpls.2016.00776

Fürtauer, L., Weiszmann, J., Weckwerth, W., and Nägele, T. (2019). Dynamics of plant metabolism during cold acclimation. *Int. J. Mol. Sci.* 20:5411. doi: 10.3390/ijms20215411

Gao, Y., Wang, J., and Zhao, F. (2015). CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. *Genome Biol.* 16:4.

Gong, Z., Xiong, L., Shi, H., Yang, S., Herrera-Estrella, L. R., Xu, G., et al. (2020). Plant abiotic stress response and nutrient use efficiency. *Sci. China Life Sci.* 63, 635–674.

Gururani, M. A., Venkatesh, J., and Tran, L. S. (2015). Regulation of photosynthesis during abiotic stress-induced photoinhibition. *Mol. Plant* 8, 1304–1320. doi: 10.1016/j.molp.2015.05.005

Guy, C. L. (1990). Cold acclimation and freezing stress tolerance: role of protein metabolism. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 41, 187–223. doi: 10.1146/annurev.pp.41.060190.001155

Hansen, T. B., Jensen, T. I., Clausen, B. H., Bramsen, J. B., Finsen, B., Damgaard, C. K., et al. (2013). Natural RNA circles function as efficient microRNA sponges. *Nature* 495, 384–388. doi: 10.1038/nature11993

Hansen, T. B., Wiklund, E. D., Bramsen, J. B., Villadsen, S. B., Statham, A. L., Clark, S. J., et al. (2011). MiRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA. *EMBO J.* 30, 4414–4422. doi: 10.1038/emboj.2011.359

Hong, Y. H., Meng, J., Zhang, M., and Luan, Y. S. (2020). Identification of tomato circular RNAs responsive to *Phytophthora infestans*. *Gene* 746:144652. doi: 10.1016/j.gene.2020.144652

Hu, W. H., Song, X. S., Shi, K., Xia, X. J., Zhou, Y. H., and Yu, J. Q. (2008). Changes in electron transport, superoxide dismutase and ascorbate peroxidase isoenzymes in chloroplasts and mitochondria of cucumber leaves as influenced by chilling. *Photosynthetica* 46:581.

Hu, W. H., Zhou, Y. H., Du, Y. S., Xia, X. J., and Yu, J. Q. (2006). Differential response of photosynthesis in greenhouse- and field-ecotypes of tomato to long-term chilling under low light. *J. Plant Physiol.* 163, 1238–1246. doi: 10.1016/j.jplph.2005.10.006

Huner, N. P. A., Öquist, G., and Sarhan, F. (1998). Energy balance and acclimation to light and cold. *Trends Plant Sci.* 3, 224–230. doi: 10.1016/s1360-1385(98)01248-5

Jeck, W. R., and Sharpless, N. E. (2014). Detecting and characterizing circular RNAs. *Nat. Biotechnol.* 32, 453–461. doi: 10.1038/nbt.2890

Jeck, W. R., Sorrentino, J. A., Wang, K., Slevin, M. K., Burd, C. E., Liu, J., et al. (2013). Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* 19, 141–157. doi: 10.1261/rna.035667.112

Jiang, Y. P., Huang, L. F., Cheng, F., Zhou, Y. H., Xia, X. J., Mao, W. H., et al. (2013). Brassinosteroids accelerate recovery of photosynthetic apparatus from cold stress by balancing the electron partitioning, carboxylation and redox homeostasis in cucumber. *Physiol. Plant.* 148, 133–145. doi: 10.1111/j.1399-3054.2012.01696.x

Jung, J. H., Domijan, M., Klose, C., Biswas, S., Ezer, D., Gao, M., et al. (2016). Phytochromes function as thermosensors in *Arabidopsis*. *Science* 354, 886–889. doi: 10.1126/science.aaf6005

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36.

Kovtun, Y., Chiu, W.-L., Tena, G., and Sheen, J. (2000). Functional analysis of oxidative stress-activated mitogen-activated protein kinase cascade in plants. *Proc. Natl. Acad. Sci. U.S.A.* 97, 2940–2945. doi: 10.1073/pnas.97.6.2940

Kozomara, A., and Griffiths-Jones, S. (2014). MiRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* 42, D68–D73.

Kumar, S. V., and Wigge, P. A. (2010). H2A.Z-containing nucleosomes mediate the thermosensory response in *Arabidopsis*. *Cell* 140, 136–147. doi: 10.1016/j.cell.2009.11.006

Legris, M., Klose, C., Burgie, E. S., Rojas, C. C., Neme, M., Hiltbrunner, A., et al. (2016). Phytochrome B integrates light and temperature signals in *Arabidopsis*. *Science* 354, 897–900. doi: 10.1126/science.aaf5656

Liu, C., Schläppi, M. R., Mao, B., Wang, W., Wang, A., and Chu, C. (2019). The bZIP73 transcription factor controls rice cold tolerance at the reproductive stage. *Plant Biotechnol. J.* 17, 1834–1849.

Liu, J., Liu, T., Wang, X., and He, A. (2017). Circles reshaping the RNA world: from waste to treasure. *Mol. Cancer* 16:58.

Liu, Z., Jia, Y., Ding, Y., Shi, Y., Li, Z., Guo, Y., et al. (2017). Plasma membrane CRPK1-mediated phosphorylation of 14-3-3 proteins induces their nuclear import to fine-tune CBF signaling during cold response. *Mol. Cell* 66, 117–128.e5.

Lu, T., Cui, L., Zhou, Y., Zhu, C., Fan, D., Gong, H., et al. (2015). Transcriptome-wide investigation of circular RNAs in rice. *RNA* 21, 2076–2087. doi: 10.1261/rna.052282.115

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., et al. (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 495, 333–338. doi: 10.1038/nature11928

Moellering, E. R., Muthan, B., and Benning, C. (2010). Freezing tolerance in plants requires lipid remodeling at the outer chloroplast membrane. *Science* 330, 226–228. doi: 10.1126/science.1191803

Nägele, T., Kandel, B. A., Frana, S., Meissner, M., and Heyer, A. G. (2011). A systems biology approach for the analysis of carbohydrate dynamics during acclimation to low temperature in *Arabidopsis thaliana*. *FEBS J.* 278, 506–518. doi: 10.1111/j.1742-4658.2010.07971.x

Nie, W. F., Lei, M., Zhang, M., Tang, K., Huang, H., Zhang, C., et al. (2019). Histone acetylation recruits the SWR1 complex to regulate active DNA demethylation in *Arabidopsis*. *Proc. Natl. Acad. Sci. U.S.A.* 116, 16641–16650. doi: 10.1073/pnas.1906023116

Nie, W. F., Wang, M. M., Xia, X. J., Zhou, Y. H., Shi, K., Chen, Z. X., et al. (2013). Silencing of tomato RBOH1 and MPK2 abolishes brassinosteroid-induced H2O2 generation and stress tolerance. *Plant Cell Environ.* 36, 789–803. doi: 10.1111/pce.12014

Pan, T., Sun, X., Liu, Y., Li, H., Deng, G., Lin, H., et al. (2018). Heat stress alters genome-wide profiles of circular RNAs in *Arabidopsis*. *Plant Mol. Biol.* 96, 217–229. doi: 10.1007/s11103-017-0684-7

Salzman, J. (2016). Circular RNA expression: its potential regulation and function. *Trends Genet.* 32, 309–316. doi: 10.1016/j.tig.2016.03.002

Salzman, J., Chen, R. E., Olsen, M. N., Wang, P. L., and Brown, P. O. (2013). Cell-type specific features of circular RNA expression. *PLoS Genet.* 9:e1003777. doi: 10.1371/journal.pgen.1003777

Salzman, J., Gawad, C., Wang, P. L., Lacayo, N., and Brown, P. O. (2012). Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS One* 7:e30733. doi: 10.1371/journal.pone.0030733

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

Starke, S., Jost, I., Rossbach, O., Schneider, T., Schreiner, S., Hung, L. H., et al. (2015). Exon circularization requires canonical splice signals. *Cell Rep.* 10, 103–111. doi: 10.1016/j.celrep.2014.12.002

Steponkus, P. L. (1984). Role of the plasma membrane in freezing injury and cold acclimation. *Annu. Rev. Plant Physiol.* 35, 543–584. doi: 10.1146/annurev.pp.35.060184.002551

Sun, X., Wang, Y., and Sui, N. (2018). Transcriptional regulation of bHLH during plant response to stress. *Biochem. Biophys. Res. Commun.* 503, 397–401. doi: 10.1016/j.bbrc.2018.07.123

Szabo, L., Morey, R., Palpant, N. J., Wang, P. L., Afari, N., Jiang, C., et al. (2015). Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. *Genome Biol.* 16:126.

Tan, J., Zhou, Z., Niu, Y., Sun, X., and Deng, Z. (2017). Identification and functional characterization of tomato CircRNAs derived from genes involved in fruit pigment accumulation. *Sci. Rep.* 7:8594.

Teige, M., Scheikl, E., Eulgem, T., Dóczi, R., Ichimura, K., Shinozaki, K., et al. (2004). The MKK2 pathway mediates cold and salt stress signaling in *Arabidopsis*. *Mol. Cell* 15, 141–152. doi: 10.1016/j.molcel.2004.06.023

Thomashow, M. F. (1999). PLANT COLD ACCLIMATION: freezing tolerance genes and regulatory mechanisms. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 50, 571–599. doi: 10.1146/annurev.arplant.50.1.571

Venø, M. T., Hansen, T. B., Venø, S. T., Clausen, B. H., Grebing, M., Finsen, B., et al. (2015). Spatio-temporal regulation of circular RNA expression during porcine embryonic brain development. *Genome Biol.* 16:245.

Wang, P. L., Bao, Y., Yee, M. C., Barrett, S. P., Hogan, G. J., Olsen, M. N., et al. (2014). Circular RNA is expressed across the eukaryotic tree of life. *PLoS One* 9:e90859. doi: 10.1371/journal.pone.0090859

Wang, Y. X., Wang, Q., Gao, L. P., Zhu, B. Z., Luo, Y. B., Deng, Z. P., et al. (2017a). Integrative analysis of circRNAs acting as ceRNAs involved in ethylene pathway in tomato. *Physiol. Plant.* 161, 311–321. doi: 10.1111/ppl.12600

Wang, Y. X., Yang, M., Wei, S. M., Qin, F. J., Zhao, H. J., and Suo, B. (2017b). Identification of circular RNAs and their targets in leaves of *Triticum aestivum* L. under dehydration stress. *Front. Plant Sci.* 7:2024. doi: 10.3389/fpls.2016.02024

Wang, Z. P., Liu, Y. F., Li, D. W., Li, L., Zhang, Q., Wang, S. B., et al. (2017). Identification of circular RNAs in kiwifruit and their species-specific response to bacterial canker pathogen invasion. *Front. Plant Sci.* 8:413. doi: 10.3389/fpls.2017.00413

Wu, H.-J., Wang, Z.-M., Wang, M., and Wang, X.-J. (2013). Widespread long noncoding RNAs as endogenous target mimics for microRNAs in plants. *Plant Physiol.* 161, 1875–1884. doi: 10.1104/pp.113.215962

Xie, Y., Chen, P., Yan, Y., Bao, C., Li, X., Wang, L., et al. (2018). An atypical R2R3 MYB transcription factor increases cold hardiness by CBF-dependent and CBF-independent pathways in apple. *New Phytol.* 218, 201–218. doi: 10.1111/nph.14952

Yang, T., Chaudhuri, S., Yang, L., Du, L., and Poovaiah, B. W. (2010a). A calcium/calmodulin-regulated member of the receptor-like kinase family confers cold tolerance in plants. *J. Biol. Chem.* 285, 7119–7126. doi: 10.1074/jbc.m109.035659

Yang, T., Shad Ali, G., Yang, L., Du, L., Reddy, A. S., and Poovaiah, B. W. (2010b). Calcium/calmodulin-regulated receptor-like kinase CRLK1 interacts with MEKK1 in plants. *Plant Signal. Behav.* 5, 991–994. doi: 10.4161/psb.5.8.12225

Ye, C. Y., Chen, L., Liu, C., Zhu, Q. H., and Fan, L. J. (2015). Widespread noncoding circular RNAs in plants. *New Phytol.* 208, 88–95. doi: 10.1111/nph.13585

Ye, C.-Y., Xu, H., Shen, E., Liu, Y., Wang, Y., Shen, Y., et al. (2014). Genome-wide identification of non-coding RNAs interacted with microRNAs in soybean. *Front. Plant Sci.* 5:743. doi: 10.3389/fpls.2014.00743

Yin, J. L., Liu, M. Y., Ma, D. F., Wu, J. W., Li, S. L., Zhu, Y. X., et al. (2018). Identification of circular RNAs and their targets during tomato fruit ripening. *Postharvest Biol. Technol.* 136, 90–98. doi: 10.1016/j.postharvbio.2017.10.013

Zaynab, M., Fatima, M., Abbas, S., Umair, M., Sharif, Y., and Raza, M. A. (2018). Long non-coding RNAs as molecular players in plant defense against pathogens. *Microb. Pathog.* 121, 277–282. doi: 10.1016/j.micpath.2018.05.050

Zeng, R. F., Zhou, J. J., Hu, C. G., and Zhang, J. Z. (2018). Transcriptome-wide identification and functional prediction of novel and flowering-related circular RNAs from trifoliate orange (*Poncirus trifoliata* L. Raf.). *Planta* 247, 1191–1202. doi: 10.1007/s00425-018-2857-2

Zhang, H. M., Lang, Z. B., 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. O., Wang, H. B., Zhang, Y., Lu, X., Chen, L. L., and Yang, L. (2014). Complementary sequence-mediated exon circularization. *Cell* 159, 134–147. doi: 10.1016/j.cell.2014.09.001

Zhang, Y., Zhang, X. O., Chen, T., Xiang, J. F., Yin, Q. F., Xing, Y. H., et al. (2013). Circular intronic long noncoding RNAs. *Mol. Cell* 51, 792–806. doi: 10.1016/j.molcel.2013.08.017

Zhao, C., Wang, P., Si, T., Hsu, C. C., Wang, L., Zayed, O., et al. (2017). MAP kinase cascades regulate the cold response by modulating ICE1 protein stability. *Dev. Cell* 43, 618–629.e5.

Zhao, T., Wang, L. Y., Li, S., Xu, M., Guan, X. Y., and Zhou, B. L. (2017). Characterization of conserved circular RNA in polyploid *Gossypium* species and their ancestors. *FEBS Lett.* 591, 3660–3669. doi: 10.1002/1873-3468.12868

Zhao, W., Cheng, Y. H., Zhang, C., You, Q. B., Shen, X. J., Guo, W., et al. (2017). Genome-wide identification and characterization of circular RNAs by high throughput sequencing in soybean. *Sci. Rep.* 7:5636.

Zhou, M. Q., Shen, C., Wu, L. H., Tang, K. X., and Lin, J. (2011). CBF-dependent signaling pathway: a key responder to low temperature stress in plants. *Crit. Rev. Biotechnol.* 31, 186–192. doi: 10.3109/07388551.2010.505910

Zhou, R., Xu, L. P., Zhao, L. P., Wang, Y. L., and Zhao, T. M. (2018a). Genome-wide identification of circRNAs involved in tomato fruit coloration. *Biochem. Biophys. Res. Commun.* 499, 466–469. doi: 10.1016/j.bbrc.2018.03.167

Zhou, R., Zhu, Y. X., Zhao, J., Fang, Z. W., Wang, S. P., Yin, J. L., et al. (2018b). Transcriptome-wide identification and characterization of potato circular RNAs in response to *Pectobacterium carotovorum* subspecies *brasiliense* infection. *Int. J. Mol. Sci.* 19:71. doi: 10.3390/ijms19010071

Zuo, J., Wang, Q., Zhu, B., Luo, Y., and Gao, L. (2016). Deciphering the roles of circRNAs on chilling injury in tomato. *Biochem. Biophys. Res. Commun.* 479, 132–138. doi: 10.1016/j.bbrc.2016.07.032

Conflict of Interest: SX was employed by company BGI-Shenzhen.

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

Copyright © 2020 Yang, Liu, Zhang, Wang, Zinta, Xie, Zhu and Nie. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 12 November 2020
doi: 10.3389/fgene.2020.598714

Transcriptional Analysis of Metabolic Pathways and Regulatory Mechanisms of Essential Oil Biosynthesis in the Leaves of *Cinnamomum camphora* (L.) Presl

Jiexi Hou[1], Jie Zhang[1], Beihong Zhang[1,2], Xiaofang Jin[1], Haiyan Zhang[1]* and Zhinong Jin[1]**

[1] *Jiangxi Provincial Engineering Research Center for Seed-Breeding and Utilization of Camphor Trees, The School of Hydraulic and Ecological Engineering, Nanchang Institute of Technology, Nanchang, China,* [2] *Key Laboratory of Silviculture, Co-Innovation Center of Jiangxi Typical Trees Cultivation and Utilization, College of Forestry, Jiangxi Agricultural University, Nanchang, China*

OPEN ACCESS

Edited by:
Hongjian Wan, ZheJiang Academy of Agricultural Sciences, China

Reviewed by:
Changjun Ding, Chinese Academy of Forestry, China
Duc Quan Nguyen, The University of Newcastle, Australia

***Correspondence:**
Jiexi Hou
jxhou@cctnit.com
Haiyan Zhang
HYZhang@cctnit.com
Zhinong Jin
agroking@cctnit.com

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

Received: *25 August 2020*
Accepted: *14 October 2020*
Published: *12 November 2020*

Citation:
Hou J, Zhang J, Zhang B, Jin X, Zhang H and Jin Z (2020) Transcriptional Analysis of Metabolic Pathways and Regulatory Mechanisms of Essential Oil Biosynthesis in the Leaves of Cinnamomum camphora (L.) Presl. Front. Genet. 11:598714. doi: 10.3389/fgene.2020.598714

The roots, bark, and leaves of *Cinnamomum camphora* are rich in essential oils, which mainly comprised monoterpenes and sesquiterpenes. Although the essential oils obtained from *C. camphora* have been widely used in pharmaceutical, medicinal, perfume, and food industries, the molecular mechanisms underlying terpenoid biosynthesis are poorly understood. To address this lack of knowledge, we performed transcriptome analysis to investigate the key regulatory genes involved in terpenoid biosynthesis in *C. camphora*. High-oil-yield trees of linalool type and low-oil-yield trees were used to assemble a *de novo* transcriptome of *C. camphora*. A total of 121,285 unigenes were assembled, and the total length, average length, N50, and GC content of unigenes were 87,869,987, 724, 1,063, and 41.1%, respectively. Comparison of the transcriptome profiles of linalool-type *C. camphora* with trees of low oil yield resulted in a total of 3,689 differentially expressed unigenes, among them 31 candidate genes had annotations associated with metabolism of terpenoids and polyketides, including four in the monoterpenoid biosynthesis pathway and three in the terpenoid backbone biosynthesis pathway. Collectively, this genome-wide transcriptome provides a valuable tool for future identification of genes related to essential oil biosynthesis. Additionally, the identification of a cohort of genes in the biosynthetic pathways of terpenoids provides a theoretical basis for metabolic engineering of essential oils in *C. camphora*.

Keywords: transcriptional analysis, terpenoid biosynthesis, chemotypes, leaves, *Cinnamomum camphora*

INTRODUCTION

As a member of the Lauraceae family, *Cinnamomum camphora* is an evergreen tree widely distributed around the southern portions of the Yantze River in China, with the largest numbers in Jiangxi, Guangdong, Fujian, Zhejiang, Yunnan, and Sichuan. Many parts of the camphor tree, including roots, stems, leaves, flowers, fruits, and bark, are rich in essential oils that have a wide

range of uses in the pest control, spice, cosmetics, and pharmaceutical industries (Guo et al., 2016; Jiang et al., 2016). In order to achieve sustainable production, large-scale plantations have been established in China. Under this model, camphor trees are maintained as bushes, and the leaves and twigs are harvested every year for essential oil production. *C. camphora* plantations are characterized by fast growth and high afforestation density, and therefore, they can provide a large number of raw materials for the essential oil industry.

Essential oils in the camphor tree primarily consist of monoterpenes and sesquiterpenes. Depending on the principal chemical component of their essential oils, camphor trees can be divided into different chemical types, such as linalool, borneol, camphor, cineol, and iso-nerolidol (Guo et al., 2017). With the gradual improvements in *C. camphora* cultivation and essential oil extraction techniques, use of *C. camphora* as a source of natural linalool, which has long been preferred over its synthetic form for the perfume industry, is becoming more popular. However, the molecular mechanisms underlying the terpenoid biosynthesis in *C. camphora* are still poorly understood, which impedes the progress of identifying "superior" *C. camphora* clones that could increase the efficiency of using *C. camphora* for essential oil production.

The biosynthetic pathway of terpenoids in plants starts with the synthesis of the universal terpenoid precursors isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP) through the mevalonate (MVA) or the methylerythritol 4-phosphate (MEP) pathways. Isoprenyl diphosphate synthases (IDSs) catalyze reactions involving IPP and DMAPP, forming the intermediate phosphate precursors, geranyl diphosphate (GPP), geranylgeranyl diphosphate (GGPP), and farnesyl diphosphate, which are then catalyzed by terpene synthases (TPSs) to form diverse terpenoids (Chen et al., 2003; Cheng et al., 2007; Ma et al., 2012). The enzymes in the MVA pathway have been shown to be located in cytosol or peroxisomes, whereas all the enzymes involved in the MEP pathway are found in plastids (Reumann et al., 2007; Simkin et al., 2011). In contrast, the localizations of IDSs and TPSs are more diverse and are often at the site of terpenoid biosynthesis (Ma et al., 2012). Although characterizations of the genes involved in terpenoid biosynthesis have been reported in many plant species, including peppermint, *Arabidopsis*, conifers, *Hevea brasiliensis*, and *Salvia miltiorrhiza* (Battaile and Loomis, 1961; Lange and Ghassemian, 2003; Sando et al., 2008a,b; Zulak and Bohlmann, 2010; Ma et al., 2012), many of the enzyme-encoding genes have yet to be elucidated because of the complexity of the terpenoid biosynthesis pathway, especially in the species without a whole-genome sequence.

In the studies of *C. camphora*, transcriptome analyses for leaves of camphor trees of linalool-, borneol-, camphor-, cineol-, and iso-nerolidol types led to the identification of 424 unigenes in the biosynthetic pathways of terpenoids (Jiang et al., 2014) and transcriptome profiling of linalool and borneol types of *C. camphora* identified three monoterpene synthase genes possibly involved in the biosynthesis of borneol (Chen et al., 2018). Although these two transcriptome studies have provided valuable insights into the terpenoid biosynthesis pathway, comparisons between only high-oil-yielding individuals could exclude the identification of genes upstream of the biosynthetic pathway of terpenoids. In this study, we used RNA-Seq to identify key enzyme-encoding genes putatively associated with terpenoid biosynthesis in *C. camphora*, which may partially determine its essential oil composition. Differentially expressed genes (DEGs) between camphor trees of linalool-type and low-oil-yield camphor trees were identified. In addition to four unigenes in the "monoterpenoid biosynthesis" pathway, this analysis also identified three unigenes in the "terpenoid backbone biosynthesis" pathway. Overall, these results contribute to the general understanding of terpenoid biosynthesis in *C. camphora* and provide a theoretical basis for metabolic engineering of terpenoids in this species.

MATERIALS AND METHODS

Plant Materials

Mature leaves of linalool-type and low-oil-yield *C. camphora* were harvested from 3-year-old clones grown at the experimental garden of NanChang Institute of Technology, China, in October. The clones were propagated from mother trees through cutting propagation. For each biological replicate ($n = 3$), leaves from at least five tree clones, which were cloned from the same mother tree, were collected and mixed. The three replicates each for the low-oil-yield variety and linalool-type *C. camphora* were termed N- (1–3) and LI- (1–3) separately. All the samples were harvested and snap-frozen in liquid nitrogen and then stored at −80°C until RNA extraction. Meanwhile, leaves were also collected for oil extraction and chemical component determination.

Essential Oil Extraction and Determination of Essential Oil Yield

Fresh leaves were harvested and processed for oil extraction immediately after harvesting. Leaf samples were hydrodistilled in a modified Clevenger-type apparatus for 4 h. Extracted essential oils were weighed, and oil yields were calculated using the formula:

$$\text{Essential oil yield (\%)} = \text{W1/W2} \times 100$$

where W1 is the weight of extracted essential oil, and W2 is the weight of fresh leaves.

Gas Chromatography–Mass Spectrometry Analysis

Analyses of essential oils were performed on a gas chromatography system (Agilent 7890B, United States) equipped with a TG-5MS capillary column (30 m × 0.25-mm internal diameter × 0.25-μm film thickness). Helium with a flow rate of 1 mL/min was used as the carrier gas. The injector was maintained at 220°C. The initial temperature of the column was set at 40°C and increased to 280–300°C at a rate of 6.5°C/min. The mass spectrometer was set to scan in the range of 50–650 m/z with a scan rate of 0.5 scans/s. Compounds

were determined based on their relative retention time and mas spectra data. Details were described in our previous study (Zhang et al., 2019).

cDNA Library Preparation and RNA Sequencing

Total RNA was extracted from samples using the RNeasy Plant Mini Kit (Qiagen, United States) following the manufacturer's instructions. Next, gDNA was removed using DNase I digestion (Qiagen, United States). The integrity of extracted RNA was determined by the RNA Nano 6000 Assay Kit of the Agilent 2100 Bioanalyzer (Agilent Technologies, United States). Extracted RNA was sent to Gene *Denovo* Biotechnology Co. (Guangzhou, China) for cDNA library preparation and sequencing. Up to 1 μg of total RNA from each sample was used for library construction using the NEB #7530 kit for Illumina® (New England Biolabs, United States), as described by the manufacturer. mRNA was first enriched by oligo(dT) beads and fragmented into small fragments using fragmentation buffer and reverse-transcribed into cDNA with ProtoScript II Reverse Transcriptase and random primers. Second-strand cDNA was synthesized by DNA polymerase I along with RNase H, dNTP, and buffer. Next, cDNA fragments were purified with 1.8 × Agencourt AMPure XP Beads, end-repaired, and ligated to Illumina sequencing adapters after addition of poly(A) tails. Ligated fragments were subjected to size selection by gel electrophoresis and polymerase chain reaction (PCR) amplified. The resulting cDNA library was sequenced using Illumina HiSeq^{TM} 4000.

De novo Assembly and Functional Annotation

Raw reads from each sample were filtered by removing reads containing more than 10% of unknown nucleotides, low-quality reads that contained more than 40% of low-quality (Q value ≤ 20) bases, and adapter sequences according to the Illumina adapter list. Residual ribosome RNA (rRNA) reads were identified by mapping the filtered reads to rRNA, and the identified sequences were then removed. The resulting 87.9 M reads were assembled using short read assembler, Trinity (Grabherr et al., 2011), and further clustered to obtain non-redundant unigenes by Corset (Davidson and Oshlack, 2014). The completeness of the transcriptomic assembly was assessed using BUSCO (Simao et al., 2015) by comparing the 1,440 embryophyta-specific genes to our unigenes. Unigenes were annotated by BLASTx (with an E-value threshold of $1e^{-5}$) to the Swiss-Prot protein database (Bairoch and Apweiler, 2000) and NCBI non-redundant protein (Nr) database (O'Leary et al., 2016). Nr annotation results of unigenes were used for Gene Ontology (GO) annotation by Blast2GO software (Conesa et al., 2005), and functional classification of unigenes was performed using WEGO software (Ye et al., 2006).

Identification of Differentially Expressed Genes

Expression values of each gene were calculated as reads per kilobase (of exon) per million (RPKM) mapped reads. Biological samples with low and high oil contents were grouped, and comparison was undertaken to identify DEGs between the two experimental groups. Principal component analysis (PCA) was also performed as a quality control. DEGs with a fold change ≥ 2 and $p < 0.01$ across the two experimental groups were identified using the edgeR package (Robinson et al., 2010).

Quantitative Reverse Transcriptase–PCR

Leaves from five randomly selected 3-year-old tree clones of each chemotype were collected for quantitative reverse transcriptase (RT)–PCR analysis. Total RNA was extracted and reverse-transcribed into cDNA using the HiScript II Q RT SuperMix for quantitative PCR (qPCR) (Vazyme, China). The PCRs were performed using ChamQTM SYBR® qPCR Master Mix (Vazyme, China) according to the manufacturer's instructions on an ABI Step One Plus System with three replicates. Sequences of the gene-specific primers are listed in **Supplementary Table S1**. Relative gene expression was calculated using the $2^{-\Delta\Delta Ct}$ method, and *ACTIN* was chosen as the internal reference gene to normalize expression levels.

RESULTS

Chemical Composition of Leaf Essential Oils From *C. camphora* of Different Oil Contents

In general, there were no major differences in morphology between camphor trees of different chemotypes, and the trees were therefore initially classified by odor alone. Fresh leaves of the six samples were harvested, and the essential oils were obtained by water distillation. The average oil outputs of the samples from linalool- and low-oil-yield chemotypes were 1.3 and 0.3%, respectively. Chemical composition of the extracted essential oils was determined by gas chromatography–mass spectrometry (GC-MS). **Table 1** lists the major components detected, 1–15 are monoterpenes, and the remaining are sesquiterpenes. In the essential oils of the linalool type, β-linalool was the principal constituent, making up ∼90% in average. Among the various constituents of the essential oils extracted from the different chemotypes of camphor trees, the major difference was observed for the monoterpene content. Therefore, transcriptome analyses between trees of linalool type and trees with low oil yield were performed to investigate the molecular mechanisms underlying the biosynthesis of monoterpenes.

De novo Assembly of the *C. camphora* Transcriptome

To construct a genome-wide transcriptome of *C. camphora*, total RNA of the experimental samples was isolated. This pooled total RNA was sequenced and a total of 87.9 M 150 bp paired-end reads remained after raw reads were processed with sequence filtering and quality control (**Supplementary Table S2**). *De novo* assembly was carried out with short reads using Trinity (see section "Materials and Methods"). A total of 121,285 unigenes were assembled with a high N50 value (1063), as well as high

TABLE 1 | Chemical composition of the essential oils from leaf extracts of *C. camphora*.

No.	Compound	N-1 (%)	N-2 (%)	N-3 (%)	LI-1(%)	LI-2(%)	LI-3 (%)
1	α-Thujene	1.51	—	—	—	0.1	—
2	Pinene	8.28	4.63	—	—	0.36	0.19
3	Camphene	—	1.49	—	—	—	0.14
4	Phellandrene	26.19	0.88	—	0.19	0.12	0.47
5	β-Myrcene	2.26	—	—	0.16	0.16	—
6	D-Limonene	—	—	—	0.15	0.26	0.26
7	Eucalyptol	37.7	2.14	—	0.59	—	0.46
8	β-Ocimene	0.56	4.87	—	—	—	0.68
9	Terpinene	3.48		—	—	—	—
10	β-Linalool	—	2.06	—	90.38	86.8	89.16
11	Borneol acetate	—	1.95	—	—	—	—
12	Camphor	—	—	—	—	1.36	—
13	Terpineol	10.57	—	—	—	0.13	0.07
14	Citronellol	—	—	—	—	0.13	—
15	Citral	—	—	—	0.43	1.16	0.54
16	Caryophyllene	2.2	13.16	8.58	2.07	1.71	1.03
17	Alloaromadendrene	—	1.24	0.18	—	—	—
18	Germacrene D	1.37	1.28	3.11	1.26	0.24	0.59
19	Naphthalene	—	1.36		—	—	—
20	Humulene	1.76	3.41	5.49	—	—	0.59
21	ç-Elemene	1.25	15.01	4.43	0.86	3.36	1.18
22	Nerolidol	—	22.2	44.95	0.11	0.14	0.24
23	(−)-Spathulenol	—	1.76	3.72	—	0.94	0.41
24	Nerolidyl acetate	—	1.29	—	—	—	—
25	Guaiol	0.9	—	0.24	—	—	—
26	Caryophyllene oxide	—	1.17	1.1	0.28	0.34	0.58
27	Others and Unknown	1.97	20.1	28.2	3.52	2.69	3.41

N: C. camphora with low oil yield; LI: C. camphora of linalool type; "—" represents a compound not detected or in trace amounts (< 0.1%).

values for both the maximum unigene length (15,713) and percentage of mapped reads (>98%), indicating a good assembly quality (**Table 2**). In addition, BUSCO analysis was performed to assess the completeness of the final genome assembly. Based on the 1,440 embryophytic-specfic genes, 79.8% were identified as complete genes, including 76.3% complete and single-copy genes and 3.5% complete and duplicated genes, and 8.8% were identified as fragmented genes, indicating good completeness of the genome assembly (**Supplementary Figure S1**).

Functional Annotation of Unigenes

Functional annotation of 121,285 *C. camphora* unigenes was performed by comparing unigene sequences to the protein databases Nr and Swissprot with an *E*-value threshold of $1e^{-5}$ using BLASTx analysis. This analysis yielded a total of 50,545 annotated genes, with 34,789 unigenes in common between two databases (**Figure 1A**). A total of 49,295 unigenes had significant hits in the Nr database. Further GO annotation analysis assigned the 49,295 unigenes to three main categories, including biological process, cellular component, and molecular function (**Figure 1B**). Within the biological process, "metabolic process" was highly represented (29.5%) followed by "cellular process" (24.1%). Within the cellular component, "cell" (23.7%), "cell part" (23.7%), and "organelle" (16.0%) were most represented. Similarly, "catalytic activity" and "binding" were the most dominant terms under molecular function, accounting for 55.4 and 36.1%, respectively (**Supplementary Table S3**). To systematically analyze the metabolic pathways of unigene products in cells and predict the function of these unigene products, Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis was performed, and 11,111 unigenes were annotated and classified into 19 metabolic pathways. The category "Global and Overview" included the largest number of unigenes (**Figure 1C**), among which "metabolic pathways" and "biosynthesis of secondary metabolites" were dominant, accounting for 40.8 and 22.2%, respectively (**Supplementary Table S4**). Additionally, 359 unigenes were

TABLE 2 | Statistics for *de novo* assembly of the *C. camphora* transcriptome.

Total assembled bases	87.9M
Total no. of unigenes	121,285
N50 (bp)	1,063
Shortest unigene length (bp)	201
Longest unigene length (bp)	15,713
Average unigene length (bp)	724
Mapped reads (%)	>98%

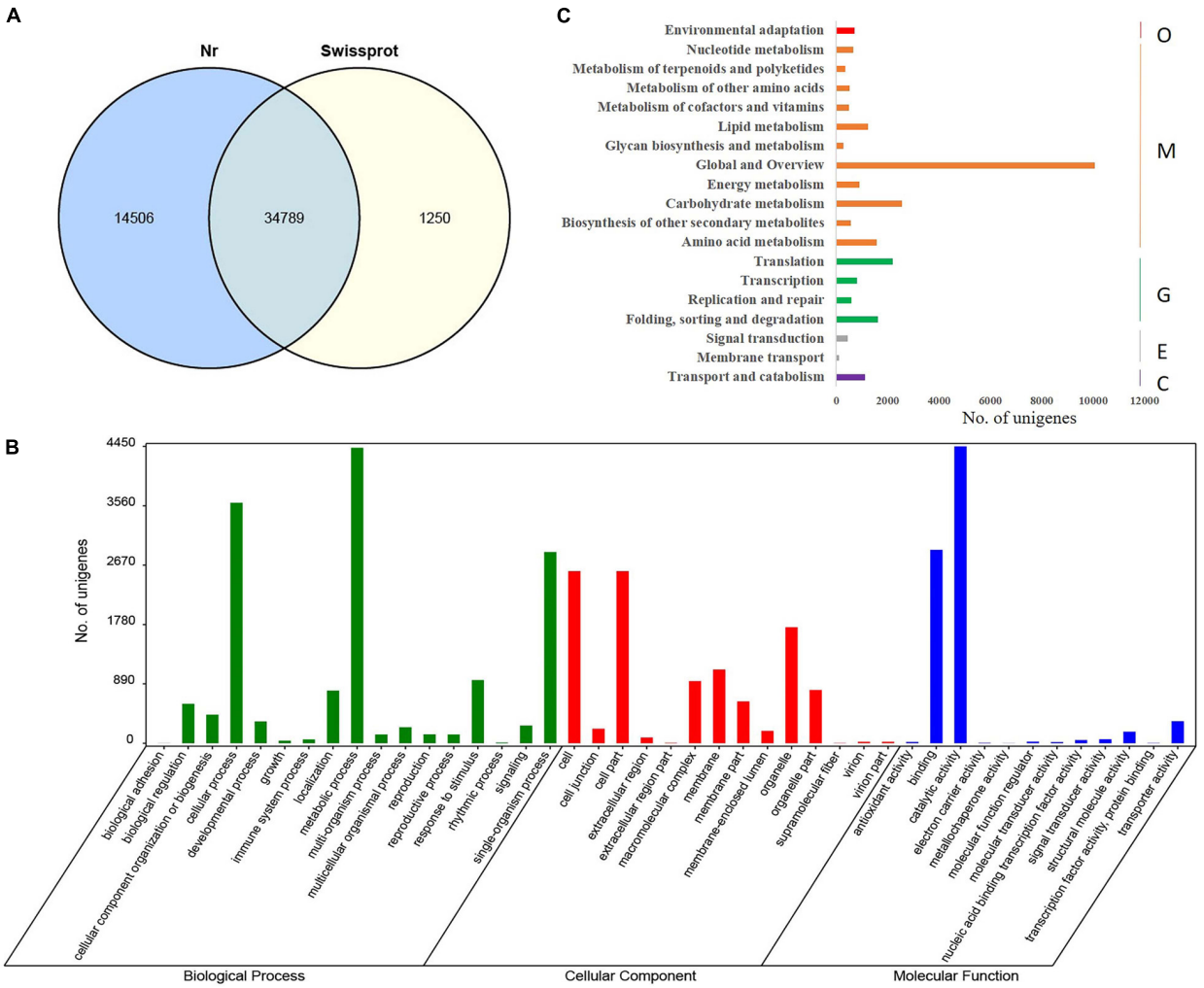


FIGURE 1 | Functional annotation and GO classification of assembled unigenes in *C. camphora*. **(A)** Functional annotation of assembled unigenes by BLASTx against Nr and Swissprot databases with an *E*-value threshold of $1e^{-5}$. A total of 34,789 unigenes were commonly annotated in the two databases. **(B)** GO terms of *C. camphora* unigenes determined by Blast2GO fell into the three categories of biological process, cellular component, and molecular function, where metabolic process, cell part, and catalytic activity were the three most prevalent subcategories, respectively. **(C)** KEGG classification of annotated unigenes. The *x*-axis shows the number of unigenes annotated, and the *y* axis shows the KEGG metabolic pathways the unigenes fell into, including organismal system (O), metabolism (M), genetic information processing (G), environmental information processing (E), and cellular processes (C).

placed into the category "metabolism of terpenoids and polyketides," which included "terpenoid backbone biosynthesis" (32.3%), "monoterpenoid biosynthesis" (5.0%), "diterpenoid biosynthesis" (13.4%), "sesquiterpenoid and triterpenoid biosynthesis" (5.0%), "carotenoid biosynthesis" (15.3%), "brassinosteroid biosynthesis" (8.9%), "zeatin biosynthesis" (8.4%), and "limonene and pinene degradation" (11.7%) (**Supplementary Table S4**).

RNA-Seq Analyses of DEGs Between *C. camphora* Leaves With High and Low Oil Contents

To get a better understanding of the dynamic processes of terpenoid biosynthesis in *C. camphora*, we performed transcriptome analyses of camphor leaves with high and low oil contents based on the oil yield and GC-MS results (**Table 1**). Trimmed reads were mapped to the *de novo* transcriptome assembly of *C. camphora*. PCA demonstrated a good similarity between samples in the high- and low-oil-content groups (**Supplementary Figure S2A**). Similarly, the range of Pearson correlation coefficient between two samples in each experimental group is 0.93–0.96 (**Supplementary Figure S2B**), also indicating that the samples in the same groups are highly correlated. Further transcriptome analysis identified 3,689 DEGs with $p < 0.01$ and $|\log2 \text{ fold_change}| > 2$, which comprised 2,120 up-regulated and 1,569 down-regulated unigenes. The GO terms "metabolic process" in the "biological process" category and "catalytic activity" in the "molecular function" category were the two most highly represented

within DEGs (**Supplementary Figure S3**). In accordance with the GO analysis, KEGG enrichment analysis revealed that most of the 3,689 DEGs were in "metabolism" class (**Figure 2A**), which included "biosynthesis of secondary metabolites," "phenylpropanoid biosynthesis," "monoterpenoid biosynthesis," "terpenoid backbone biosynthesis," and other pathways (**Supplementary Table S5**).

Putative Genes Involved in Terpenoid Backbone and Terpenoid Biosynthesis in *C. camphora*

Annotated protein-coding enzymes were mapped into the terpenoid backbone biosynthesis pathway, and gene expression data of each enzyme in the two experimental groups were indicated with logRPKM values (**Figure 2B** and **Supplementary Table S6**). Two unigenes encoding key enzymes, *AACT-like* in the MVA pathway and *DXS-like* in the MEP pathway, exhibited a higher expression level in the groups with high oil contents. Because both MVA and MEP pathways generate the precursors IPP and DMAPP for the final production of terpenoids, increased expression of these two genes could explain the higher amount of terpenoids we observed in this group (**Table 1**). Additionally, the *GGPPS-like* gene also had a higher expression level in the leaves of linalool type. Among the DEGs, 14 unigenes were annotated as terpenoid synthase–encoding genes, including four in monoterpenoid biosynthesis, four in sesquiterpenoid and triterpenoid biosynthesis, and six in diterpenoid biosynthesis (**Figure 2B** and **Supplementary Table S6**). One *LIS-like* and one *GerS-like* unigene were up-regulated in the monoterpenoid biosynthesis pathway, indicating that increased availability of terpenoid precursors, along with increased levels of monoterpene synthases, might be the reason for the higher oil yield and higher accumulation of linalool and other monoterpenes in the trees of linalool type. Besides monoterpenoid synthases, four sesquiterpenoid synthase unigenes encoding GerD and four diterpenoid synthase unigenes encoding CLS, CYP88D6, GA2OX1, and GA2OX8 were also up-regulated. In general, the upregulation of terpenoid synthase unigenes in the trees of linalool-type *C. camphor* is consistent with their increased accumulation of terpenoids, suggesting that the presence of various terpenoid synthases might be the reason for the occurrence of varied terpenoids in the essential oils of *C. camphora*. To validate the reliability of our RNA-Seq data, four DEGs were chosen for qRT-PCR examination. As shown in **Figure 2C**, the expression profiles of these unigenes observed by qRT-PCR are identical to their profiles identified in the RNA-Seq analysis.

DISCUSSION

C. camphora oils have a wide variety of uses, including ornamental, medicinal, and industrial, which makes understanding the regulation of essential oil biosynthesis in camphor tree highly important (Guo et al., 2016; Jiang et al., 2016; Chen et al., 2018). Because of earlier unregulated harvesting, many natural sources of camphor trees have been exhausted. In response to this, camphor tree plantations have been established in many countries including China and Japan. These camphor trees are maintained as bushes and are harvested for twigs and leaves before essential oil extraction. Camphor trees can be divided into different chemotypes based on their main terpenoid components. In this study, we identified camphor trees of linalool type and low oil contents without a dominant terpene component. A total of 26 main components were identified by GC-MS. We further demonstrated that the chemical components in the essential oils of *C. camphora* leaf extracts contained mainly monoterpenes and sesquiterpenes with very minor quantities of diterpenes (**Table 1**). Moreover, linalool took up ∼90 of the total essential oil contents in linalool type, indicating that monoterpene levels account for the major difference between the essential oils of different chemotypes.

Terpenoid biosynthesis and its regulation are extensively documented in many plant species. Despite having high structural diversity, all terpenes are derived from two isomeric basic backbone molecules, IPP and DMAPP, which are synthesized either through the MVA or MEP pathways (Sacchettini and Poulter, 1997). Although many members of the Lauraceae family are rich in terpenoids and are economically important tree species, limited studies are available analyzing genes involved in its terpenoid biosynthesis pathway (Yang et al., 2005; Lin et al., 2014; Chen et al., 2018; Chen et al., 2020). This study was focused on identifying genes involved in terpenoid biosynthesis and key enzymes involved in the biosynthesis of the backbone IPP and DMAPP molecules in *C. camphora*. A transcriptome of *C. camphora* was also generated to provide a reference for future RNA-Seq–based expression profiling of terpenoid biosynthesis. Functional annotation analysis showed significant enrichment in genes associated with "metabolic pathways" and "biosynthesis of secondary metabolites" (**Figure 1**). This is consistent with our observation that September–November is the time of the year when *C. camphora* harbors the highest oil contents (Qin et al., 2015), and it is also the harvesting time of our samples. Furthermore, 116 and 18 expressed unigenes were identified to be highly homologous with 49 and 11 known enzymes in the terpenoid backbone biosynthesis and monoterpenoid biosynthesis pathways, respectively.

By comparing the transcriptome profiles of leaves of linalool and leaves with low oil contents, 3,689 unigenes were identified as DEGs. Among them, expression levels of *AACT-like* and *DXS-like*, catalyzing the first steps toward IPP and DMAPP in the MVA and MEP pathways, were up-regulated in the leaves of linalool type. *DXS* genes are believed to be a significant control point in the MEP pathway (Wolfertz et al., 2004; Rohdich et al., 2006), and manipulation of gene expression levels of *DXS* in *Arabidopsis* altered the levels of tocopherols, carotenoids, chlorophylls, abscisic acid, and various other isoprenoids (Estévez et al., 2001). However, the contribution of each pathway to the generation of the common precursors for varied terpenoid biosynthesis is still unknown, because crosstalk between the two pathways has been documented in

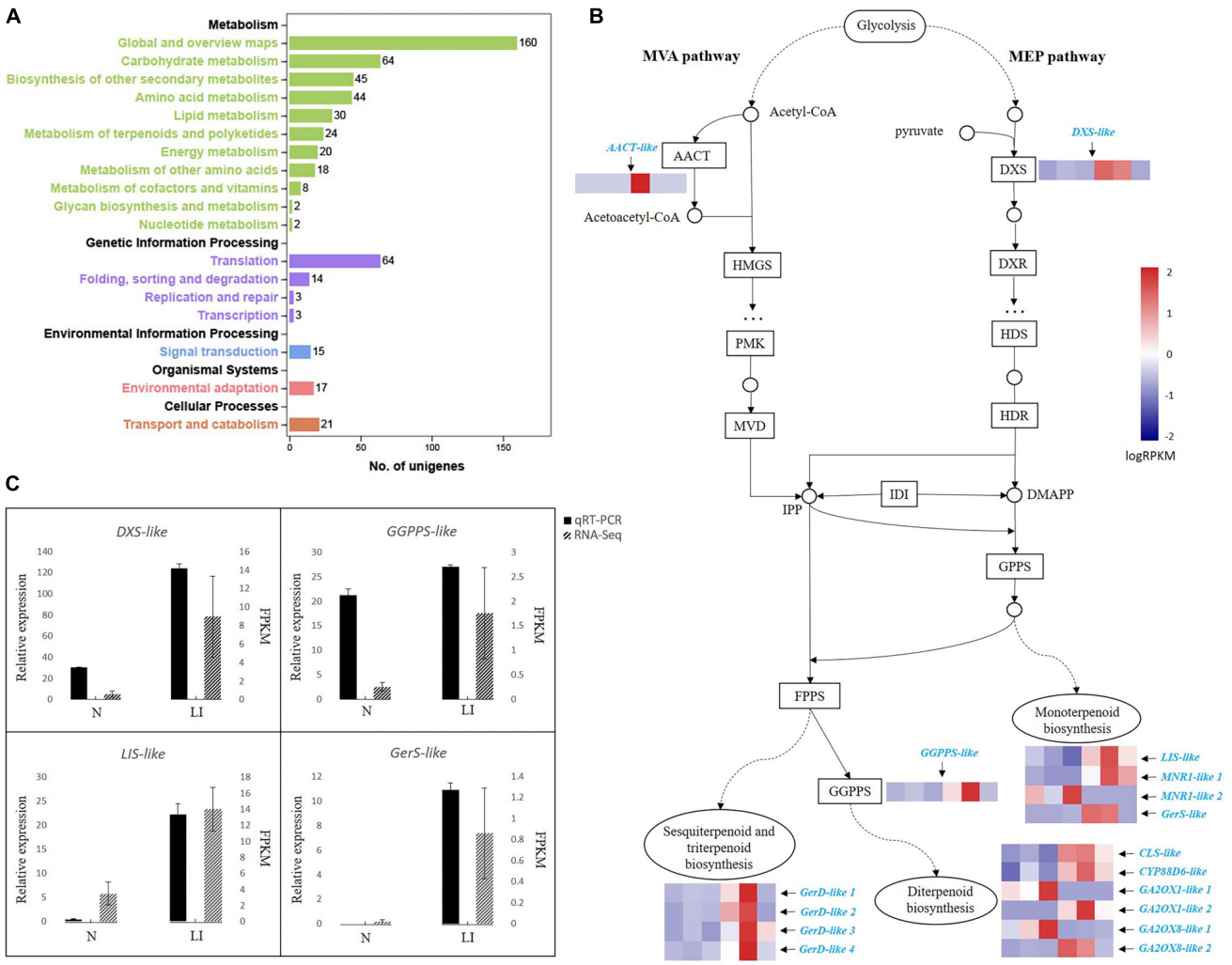


FIGURE 2 | Transcriptional cascade of DEGs in terpenoid backbone biosynthesis in linalool-type *C. camphora*. **(A)** KEGG pathway annotations of DEGs. The *x*-axis indicates the number of unigenes annotated in each corresponding pathway. The five categories of KEGG pathways are divided and color-coded. **(B)** DEGs in terpenoid backbone biosynthesis. The transcriptional cascade of terpenoid backbone biosynthesis is shown in the form of protein-coding enzymes or nodes annotated by KEGG pathway analysis. Enzyme expression patterns are indicated with logRPKM values, and DEGs are in azure. The left three and right three columns represent low oil content type (N) and linalool type (LI), respectively. AACT, acetyl-CoA C-acetyltransferase; DXS, 1-deoxy-D-xylulose-5-phosphate synthase; HMGS, hydroxymethylglutaryl-CoA synthase; PMK, phosphomevalonate kinase; MVD, mevalonate diphosphate decarboxylase; DXR, 1-deoxy-d-xylulose-5-phosphate reductoisomerase; HDS, hydroxymethylbutenyl 4-diphosphate synthase; HDR, (E)-4-hydroxy-3-methyl-but-2-enyl diphosphate reductase; IDI, isopentenyl-diphosphate delta-isomerase; GPPS, geranyl diphosphate synthase; FPPS, farnesyl pyrophosphate synthase; GGPPS, geranylgeranyl pyrophosphate synthase; LIS, linalool synthase; MNR1, (+)-neomenthol dehydrogenase-like isoform; GerS, geraniol synthase; GerD, (−)-germacrene D synthase; CLS, ent-copalyl diphosphate synthase; CYP88D6, cytochrome P450 88D6; GA2OX, Gibberellin 2-beta-dioxygenase. **(C)** qRT-PCR validation of RNA-Seq data. The black bars represent relative gene expression determined by qRT-PCR (left *y*-axis) and the dotted gray bars represent gene expression levels detected by RNA-Seq (right *y*-axis). Standard errors of the three technical replicates of each sample of the qRT-PCR analysis and three biological replicates in the two experimental groups of the RNA-Seq analysis are indicated by error bars.

many species including chamomile, *Arabidopsis*, tobacco, and snapdragon (Adam and Zapp, 1998; Bick and Lange, 2003; Hemmerlin et al., 2003; Laule et al., 2003; Dudareva et al., 2005). In addition, the expression level of *GGPPS-like* in the "terpenoid backbone biosynthesis" pathway was up-regulated in the comparison, indicating increased supply of precursors for diterpenes biosynthesis. Although the percentage of diterpenes in leaf extracts of linalool-type camphor trees is not high, their oil yield is much higher than trees of low oil yield. Therefore, higher *GGPPS* expression level may simply be correlated with a general increase in oil content. Interestingly, it has been documented that GPP, the precursor for monoterpene, is constitutively and ubiquitously expressed in *Arabidopsis* (Van Schie et al., 2013), and our results also showed that the level of *GPPS* was indifferent between the groups regardless of different accumulation of monoterpenes.

Noticeably, *LIS-like* shared a high sequence identity (>99%) with some of the identified linalool synthases in the Lauraceae family. The conserved DDxxD and NSE/DTE motifs for Mg^{2+} or Mn^{2+} ion binding as well as the predicted D562 active site cleft were all properly distributed in the LIS-like protein (**Supplementary Figure S4** and **Supplementary Table S7**). However, caution should be used when assigning protein functions to terpenoid synthase genes predicted based on sequence similarities as it has been reported that point mutations alone were sufficient to significantly change the kinetic activity of a linalool synthase gene in *Cinnamomum osmophloeum* (Lin et al., 2014). Additionally, alternative splicing has been shown to alter linalool biosynthesis, as reported in *Camellia sinensis* (Liu et al., 2018). Higher levels of monoterpene production could potentially lower production of other terpenes and isoprenoid derivatives due to the competition for common precursors. Indeed, four unigenes in the "carotenoid biosynthesis pathway" and two unigenes in the "steroid biosynthesis pathway" were down-regulated in the comparison (**Supplementary Table S6**). Similar examples can be seen in tobacco plants producing monoterpene where lower levels of β-caryophyllene were detected in the flowers of these plants (Lücker et al., 2004). Additionally, both transgenic tomato and *Arabidopsis* plants producing a phytoene synthase and the strawberry linalool/nerolidol synthase separately had significant growth retardation due to the severe inhibition of gibberellin biosynthesis (Fray et al., 1995; Aharoni et al., 2004).

Other genes of interest emerged from this study, which include several members of the WRKY family. *WRKY21-like*, *WRKY47-like*, and *WRKY65-like* were all up-regulated in the leaves with high oil content (**Supplementary Table S6**). WRKY transcription factors have been reported to regulate the expression of genes involved in terpene biosynthesis in many species, including *Arabidopsis*, cotton, *Artemisia annua*, *Catharanthus roseus*, and others (Xu et al., 2004; Ma et al., 2009; Mao et al., 2011; Suttipanta et al., 2011). Moreover, it was suggested that the linalool-type *C. osmophloeum* contains a W-box element (T)(T)TGAC(C/T) in the promoter region of the linalool synthase gene, enabling regulation by WRKY transcription factors (Lin et al., 2014). Based on these previous studies, we suspect that members of the WRKY family might be involved in the regulation of terpene synthase genes of *C. camphora* and therefore could potentially play critical roles in the quantitative difference of oil contents between different chemotypes.

All in all, this study identified a cohort of genes in the terpenoid backbone biosynthesis and monoterpenoid biosynthesis pathways that were commonly up-regulated in leaves with high linalool contents compared to leaves with low oil contents. The knowledge obtained from this study could facilitate manipulation of camphor tree essential oil production through metabolic engineering of essential oil biosynthesis. Additional studies should focus on the isolation of monoterpene synthases genes from different chemotypes of *C. camphora*. Additionally, functional analyses are needed to demonstrate the impact of these genes on terpenoid production.

DATA AVAILABILITY STATEMENT

Raw reads yielded from Illumina sequencing have been uploaded to the NCBI Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) and accession numbers for the six samples are as below: N-1 (SRR11362618); N-2 (SRR11362617); N-3 (SRR12576812); LI-1 (SRR11362614); LI-2 527 (SRR11362613); and LI-3 (SRR12576811).

AUTHOR CONTRIBUTIONS

HZ, JH, and ZJ designed the experiments. BZ, JZ, XJ, and JH performed the experiments. HZ and JH analyzed the data. JH wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (grant no. 31560242) and the Jiangxi Provincial Education Department Fund (grant no. GJJ190975).

ACKNOWLEDGMENTS

We thank HuiJuan Cheng and Yiqun Zhang for their assistance in sample harvesting and essential oil extraction. We would also like to thank Yanbo Wang and reviewers for their help in improving the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.598714/full#supplementary-material

Supplementary Figure 1 | BUSCO analysis of the final transcriptomic assembly.

Supplementary Figure 2 | Relationship between the samples.

Supplementary Figure 3 | GO enrichment histogram of DEGs.

Supplementary Figure 4 | Amino acid sequence alignment of identified LIS-like in *C. camphora* with linalool synthase CoLIS-LL in *C. osmophloeum* (accession number: JX028232).

Supplementary Table 1 | Sequences of the gene-specific primers used in quantitative RT-PCR.

Supplementary Table 2 | Raw transcriptomic data.

Supplementary Table 3 | GO enrichment analysis of *C. camphora* unigenes.

Supplementary Table 4 | KEGG enrichment of *C. camphora* unigenes.

Supplementary Table 5 | KEGG enrichment of DEGs.

Supplementary Table 6 | Putative unigenes associated with terpenoid biosynthesis.

Supplementary Table 7 | Sequences of putative unigenes associated with terpenoid biosynthesis.

REFERENCES

Adam, K.-P., and Zapp, J. (1998). Biosynthesis of the isoprene units of chamomile sesquiterpenes. *Phytochemistry* 48, 953–959. doi: 10.1016/S0031-9422(97)00992-8

Aharoni, A., Giri, A. P., Verstappen, F. W. A., Bertea, C. M., Sevenier, R., Sun, Z., et al. (2004). Gain and loss of fruit flavor compounds produced by wild and cultivated strawberry species. *Plant Cell* 16, 3110–3131. doi: 10.1105/tpc.104.023895

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28, 45–48. doi: 10.1093/nar/28.1.45

Battaile, J., and Loomis, W. D. (1961). Biosynthesis of terpenes: II. The site and sequence of terpene formation in peppermint. *BBA* 51, 545–552. doi: 10.1016/0006-3002(61)90612-6

Bick, J. A., and Lange, B. M. (2003). Metabolic cross talk between cytosolic and plastidial pathways of isoprenoid biosynthesis: unidirectional transport of intermediates across the chloroplast envelope membrane. *Arch. Biochem. Biophys.* 415, 146–154. doi: 10.1016/s0003-9861(03)00233-9

Chen, C., Zheng, Y., Zhong, Y., Wu, Y., Li, Z., Xu, L.-A., et al. (2018). Transcriptome analysis and identification of genes related to terpenoid biosynthesis in *Cinnamomum camphora*. *BMC Genomics* 19:550. doi: 10.1186/s12864-018-4941-1

Chen, F., Tholl, D., D'Auria, J. C., Farooq, A., Pichersky, E., and Gershenzon, J. (2003). Biosynthesis and emission of terpenoid volatiles from *Arabidopsis* flowers. *Plant Cell* 15, 481–494. doi: 10.1105/tpc.007989

Chen, Y., Li, Z., Zhao, Y., Gao, M., Wang, J., Liu, K., et al. (2020). The Litsea genome and the evolution of the laurel family. *Nat. Commun.* 11:1675. doi: 10.1038/s41467-020-15493-5

Cheng, A.-X., Lou, Y.-G., Mao, Y.-B., Lu, S., Wang, L.-J., and Chen, X.-Y. (2007). Plant terpenoids: biosynthesis and ecological functions. *J. Integr. Plant Biol.* 49, 179–186. doi: 10.1111/j.1744-7909.2007.00395.x

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21, 3674–3676. doi: 10.1093/bioinformatics/bti610

Davidson, N. M., and Oshlack, A. (2014). Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. *Genome Biol.* 15:410. doi: 10.1186/s13059-014-0410-6

Dudareva, N., Andersson, S., Orlova, I., Gatto, N., Reichelt, M., Rhodes, D., et al. (2005). The nonmevalonate pathway supports both monoterpene and sesquiterpene formation in snapdragon flowers. *Proc. Nat. Acad. Sci. U.S.A.* 102, 933–938. doi: 10.1073/pnas.0407360102

Estévez, J. M., Cantero, A., Reindl, A., Reichler, S., and León, P. (2001). 1-Deoxy-D-xylulose-5-phosphate synthase, a limiting enzyme for plastidic isoprenoid biosynthesis in plants. *J. Biol. Chem.* 276, 22901–22909. doi: 10.1074/jbc.M100854200

Fray, R. G., Wallace, A., Fraser, P. D., Valero, D., Hedden, P., Bramley, P. M., et al. (1995). Constitutive expression of a fruit phytoene synthase gene in transgenic tomatoes causes dwarfism by redirecting metabolites from the gibberellin pathway. *Plant J.* 8, 693–701. doi: 10.1046/j.1365-313X.1995.08050693.x

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Guo, S., Geng, Z., Zhang, W., Liang, J., Wang, C., Deng, Z., et al. (2016). The chemical composition of essential oils from *Cinnamomum camphora* and their insecticidal activity against the stored product pests. *Int. J. Mol. Sci.* 17:1836. doi: 10.3390/ijms17111836

Guo, X., Cui, M., Deng, M., Liu, X., Huang, X., Zhang, X., et al. (2017). Molecular differentiation of five *Cinnamomum camphora* chemotypes using desorption atmospheric pressure chemical ionization mass spectrometry of raw leaves. *Sci. Rep.* 7:46579. doi: 10.1038/srep46579

Hemmerlin, A., Hoeffler, J.-F., Meyer, O., Tritsch, D., Kagan, I. A., Grosdemange-Billiard, C., et al. (2003). Cross-talk between the cytosolic mevalonate and the plastidial methylerythritol phosphate pathways in tobacco bright yellow-2 cells. *J. Biol. Chem.* 278, 26666–26676. doi: 10.1074/jbc.M302526200

Jiang, H., Wang, J., Song, L., Cao, X., Yao, X., Tang, F., et al. (2016). GC×GC-TOFMS Analysis of essential oils composition from leaves, twigs and seeds of Cinnamomum camphora L. Presl and their insecticidal and repellent activities. *Molecules* 21, 423–423. doi: 10.3390/molecules21040423

Jiang, X., Wu, Y., Xiao, F., Xiong, Z., and Xu, H. (2014). Transcriptome analysis for leaves of five chemical types in *Cinnamomum camphora*. *Yi chuan* 36, 58–68. doi: 10.3724/sp.j.1005.2014.00058

Lange, B. M., and Ghassemian, M. (2003). Genome organization in Arabidopsis thaliana: a survey for genes involved in isoprenoid and chlorophyll metabolism. *Plant Mol. Biol.* 51, 925–948. doi: 10.1023/a:1023005504702

Laule, O., Fürholz, A., Chang, H.-S., Zhu, T., Wang, X., Heifetz, P. B., et al. (2003). Crosstalk between cytosolic and plastidial pathways of isoprenoid biosynthesis in *Arabidopsis thaliana*. *PNAS* 100, 6866–6871. doi: 10.1073/pnas.1031755100

Lin, Y.-L., Lee, Y.-R., Huang, W.-K., Chang, S.-T., and Chu, F.-H. (2014). Characterization of S-(+)-linalool synthase from several provenances of *Cinnamomum osmophloeum*. *Tree Genet Genomes* 10, 75–86. doi: 10.1007/s11295-013-0665-1

Liu, G.-F., Liu, J.-J., He, Z.-R., Wang, F.-M., Yang, H., Yan, Y.-F., et al. (2018). Implementation of CsLIS/NES in linalool biosynthesis involves transcript splicing regulation in *Camellia sinensis*. *Plant Cell Environ.* 41, 176–186. doi: 10.1111/pce.13080

Lücker, J., Schwab, W., van Hautum, B., Blaas, J., van der Plas, L. H. W., Bouwmeester, H. J., et al. (2004). Increased and altered fragrance of tobacco plants after metabolic engineering using three monoterpene synthases from lemon. *Plant Physiol.* 134, 510–519. doi: 10.1104/pp.103.030189

Ma, D., Pu, G., Lei, C., Ma, L., Wang, H., Guo, Y., et al. (2009). Isolation and characterization of AaWRKY1, an Artemisia annua transcription factor that regulates the amorpha-4,11-diene synthase gene, a key gene of artemisinin biosynthesis. *Plant Cell Physiol.* 50, 2146–2161. doi: 10.1093/pcp/pcp149

Ma, Y., Yuan, L., Wu, B., Li, X. E., Chen, S., and Lu, S. (2012). Genome-wide identification and characterization of novel genes involved in terpenoid biosynthesis in *Salvia miltiorrhiza*. *J. Exp. Bot.* 63, 2809–2823. doi: 10.1093/jxb/err466

Mao, G., Meng, X., Liu, Y., Zheng, Z., Chen, Z., and Zhang, S. (2011). Phosphorylation of a WRKY transcription factor by two pathogen-responsive MAPKs drives phytoalexin biosynthesis in *Arabidopsis*. *Plant Cell* 23, 1639–1653. doi: 10.1105/tpc.111.084996

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733–D745. doi: 10.1093/nar/gkv1189

Qin, Z.-H., Li, J.-F., Shi, Y., Lu, K.-C., Tang, F.-C., Liu, H.-L., et al. (2015). Content and main compositions analysis of essential oil in branches and leaves of *Cinnamomum camphora* in different months (in Chinese). *Guangxi For. Sci.* 44, 428–430. doi: 10.19692/j.cnki.gfs.2015.04.021

Reumann, S., Babujee, L., Ma, C., Wienkoop, S., Siemsen, T., Antonicelli, G. E., et al. (2007). Proteome analysis of *Arabidopsis* leaf peroxisomes reveals novel targeting peptides, metabolic pathways, and defense mechanisms. *Plant Cell* 19, 3170–3193. doi: 10.1105/tpc.107.050989

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Rohdich, F., Lauw, S., Kaiser, J., Feicht, R., Köhler, P., Bacher, A., et al. (2006). Isoprenoid biosynthesis in plants – 2C-methyl-d-erythritol-4-phosphate synthase (IspC protein) of *Arabidopsis thaliana*. *FEBS J.* 273, 4446–4458. doi: 10.1111/j.1742-4658.2006.05446.x

Sacchettini, J. C., and Poulter, C. D. (1997). Creating isoprenoid diversity. *Science* 277, 1788–1789. doi: 10.1126/science.277.5333.1788

Sando, T., Takaoka, C., Mukai, Y., Yamashita, A., Hattori, M., Ogasawara, N., et al. (2008a). Cloning and characterization of mevalonate pathway genes in a natural rubber producing plant, *Hevea brasiliensis*. *Biosci. Biotechnol. Biochem.* 72, 2049–2060. doi: 10.1271/bbb.80165

Sando, T., Takeno, S., Watanabe, N., Okumoto, H., Kuzuyama, T., Yamashita, A., et al. (2008b). Cloning and characterization of the 2-C-methyl-D-erythritol 4-phosphate (MEP) pathway genes of a natural-rubber producing plant, *Hevea brasiliensis*. *Biosci. Biotechnol. Biochem.* 72, 2903–2917. doi: 10.1271/bbb.80387

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness

with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Simkin, A. J., Guirimand, G., Papon, N., Courdavault, V., Thabet, I., Ginis, O., et al. (2011). Peroxisomal localisation of the final steps of the mevalonic acid pathway in planta. *Planta* 234:903. doi: 10.1007/s00425-011-1444-6

Suttipanta, N., Pattanaik, S., Kulshrestha, M., Patra, B., Singh, S. K., and Yuan, L. (2011). The transcription factor CrWRKY1 positively regulates the terpenoid indole alkaloid biosynthesis in *Catharanthus roseus*. *Plant Physiol.* 157, 2081–2093. doi: 10.1104/pp.111.181834

Van Schie, C. C. N., Haring, M. A., and Schuurink, R. C. (2013). "Prenyldiphosphate synthases and gibberellin biosynthesis," in *Isoprenoid Synthesis in Plants and Microorganisms: New Concepts and Experimental Approaches*, eds T. Bach and M. Rohmer (New York: Springer), 214–220.

Wolfertz, M., Sharkey, T. D., Boland, W., and Kühnemann, F. (2004). Rapid regulation of the methylerythritol 4-Phosphate pathway during isoprene synthesis. *Plant Physiol.* 135, 1939–1945. doi: 10.1104/pp.104.043737

Xu, Y.-H., Wang, J.-W., Wang, S., Wang, J.-Y., and Chen, X.-Y. (2004). Characterization of GaWRKY1, a cotton transcription factor that regulates the sesquiterpene synthase gene (+)-delta-cadinene synthase-A. *Plant Physiol.* 135, 507–515. doi: 10.1104/pp.104.038612

Yang, T., Li, J., Wang, H.-X., and Zeng, Y. (2005). A geraniol-synthase gene from *Cinnamomum tenuipilum*. *Phytochemistry* 66, 285–293. doi: 10.1016/j.phytochem.2004.12.004

Ye, J., Fang, L., Zheng, H., Zhang, Y., Chen, J., Zhang, Z., et al. (2006). WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res.* 34, W293–W297. doi: 10.1093/nar/gkl031

Zhang, B., Wu, C., Xiao, Z., Zhang, H., Cao, M., Liu, Y., et al. (2019). Chemical constituents and chemotypes of fresh leaf essential oil of wild species belonging to Sect. Camphor (Trew.) Meissn. in southeastern China. *J. Essent.* 22, 1115–1122. doi: 10.1080/0972060X.2019.1662331

Zulak, K. G., and Bohlmann, J. (2010). Terpenoid biosynthesis and specialized vascular cells of conifer defense. *J. Integr. Plant Biol.* 52, 86–97. doi: 10.1111/j.1744-7909.2010.00910.x

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.

Copyright © 2020 Hou, Zhang, Zhang, Jin, Zhang and Jin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 19 November 2020
doi: 10.3389/fgene.2020.580568

Improved *de novo* Assembly of the Achlorophyllous Orchid *Gastrodia elata*

Shanshan Chen[1,2†], Xiao Wang[3,4†], Yangzi Wang[5†], Guanghui Zhang[6], Wanling Song[5], Xiao Dong[5], Michael L. Arnold[7], Wen Wang[8,9,10], Jianhua Miao[11], Wei Chen[12,13,14]* and Yang Dong[5,11,12,14]**

[1] School of Life Sciences, Zhengzhou University, Zhengzhou, China, [2] BGI College, Zhengzhou University, Zhengzhou, China, [3] State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China, [4] Jiaxing Synbiolab Biotechnology Co., Ltd., Jiaxing, China, [5] College of Biological Big Data, Yunnan Agricultural University, Kunming, China, [6] The Key Laboratory of Medicinal Plant Biology of Yunnan Province, National and Local Joint Engineering Research Center on Germplasm Innovation and Utilization of Chinese Medicinal Materials in Southwest China, Yunnan Agricultural University, Kunming, China, [7] Department of Genetics, University of Georgia, Athens, GA, United States, [8] State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China, [9] Center for Ecological and Environmental Sciences, Northwestern Polytechnical University, Xi'an, China, [10] Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China, [11] Guangxi Key Laboratory of Medicinal Resources Protection and Genetic Improvement, Guangxi Botanical Garden of Medicinal Plants, Nanning, China, [12] State Key Laboratory of Conservation and Utilization of Bio-Resources in Yunnan, Yunnan Agricultural University, Kunming, China, [13] College of Agronomy and Biotechnology, Yunnan Agricultural University, Kunming, China, [14] Yunnan Research Institute for Local Plateau Agriculture and Industry, Kunming, China

Achlorophyllous plants are full mycoheterotrophic plants with no chlorophyll and they obtain their nutrients from soil fungi. *Gastrodia elata* is a perennial, achlorophyllous orchid that displays distinctive evolutionary strategy of adaptation to the non-photosynthetic lifestyle. Here in this study, the genome of *G. elata* was assembled to 1.12 Gb with a contig N50 size of 110 kb and a scaffold N50 size of 1.64 Mb so that it helped unveil the genetic basics of those adaptive changes. Based on the genomic data, key genes related to photosynthesis, leaf development, and plastid division pathways were found to be lost or under relaxed selection during the course of evolution. Thus, the genome sequence of *G. elata* provides a good resource for future investigations of the evolution of orchids and other achlorophyllous plants.

Keywords: ***Gastrodia elata*, genome, achlorophyllous, relaxed selection, non-photosynthetic**

OPEN ACCESS

Edited by:
Wei Chen,
North China University of Science and Technology, China

Reviewed by:
Liangsheng Zhang,
Zhejiang University, China
Xiaohua Jin,
Institute of Botany (CAS), China

***Correspondence:**
Jianhua Miao
mjh1962@vip.163.com
Wei Chen
wchenntr@gmail.com
Yang Dong
loyalyang@163.com

† These authors have contributed equally to this work

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

Received: *06 July 2020*
Accepted: *19 October 2020*
Published: *19 November 2020*

Citation:
Chen S, Wang X, Wang Y, Zhang G, Song W, Dong X, Arnold ML, Wang W, Miao J, Chen W and Dong Y (2020) Improved de novo Assembly of the Achlorophyllous Orchid Gastrodia elata. Front. Genet. 11:580568. doi: 10.3389/fgene.2020.580568

INTRODUCTION

In the autotroph-dominant plant world, the symbiotic relationship between plants and fungi plays an indispensable role in the maintenance of ecosystem (Read, 1991). In an extreme situation, some plants (at least 50 independent origins) have become solely dependent upon their fungal associates for energy source and other nutrients during the course of evolution (Merckx and Freudenstein, 2010). These fully mycoheterotrophic plants often lack a functional photosynthetic mechanism, so they are termed as "non-photosynthetic plants (Merckx et al., 2009)."

The loss of photosynthesis was independently evolved over 40 times in a diverse range of plant families and genera (Bidartondo, 2005). As an evolutionary adaptation to a low-light undergrowth environment, this extreme phenotype was also associated with some quite remarkable parallel evolutionary traits in plant, such as smaller biomass, specialized leaves without stomata, lack of root hair, reduced vascular tissues, and miniatured seeds (Leake, 1994). Even though multiple studies of plastomes from heterotrophic plants have showed reduction in plastome sizes and housekeeping gene numbers (Krause, 2008; Barrett and Davis, 2012; Wicke et al., 2013, 2014), the nuclear genomic basis of the other parallel traits in non-photosynthetic plants remains elusive.

All members of the Orchidaceae family (up to 26,567 species in 880 genera) rely on fungal associates in some or all stages of their life cycle (Leake, 1994; Cai et al., 2015). In particular, more than 200 orchid species from a myriad genera are fully mycoheterotrophic and non-photosynthetic (Leake, 1994; Merckx et al., 2009; Merckx and Freudenstein, 2010). Since *Gastrodia* is one of the largest genera consisting of fully mycoheterotrophic species from a wide geographic area (Cribb and Killmann, 2010), members in this genus are excellent models to investigate the reconfigured traits in non-photosynthetic plants.

In China, *Gastrodia elata* is cultivated for medicinal uses. It is a perennial, non-photosynthetic orchid with an enlarged rhizome and vestigial leaves on an upright flower-bearing stem (**Figure 1A**). The completion of its life cycle requires at least two types of fungi: *Armillaria* and *Mycena* (Lan et al., 1994; Ning et al., 2010) (**Supplementary Figure 1**). In this study, we presented a high quality *de novo* assembly of the *G. elata* nuclear genome, which had a higher coverage, contig N50 size, and Benchmarking Universal Single-Copy Ortholog (BUSCO) completeness than that of a previous report (Yuan et al., 2018). Compared with the genomes of photosynthetic orchid species, such as *Phalaenopsis equestris* (Cai et al., 2015), *Apostasia Shenzhenica* (Zhang et al., 2017), and *Dendrobium officinale* (Yan et al., 2015), the new *G. elata* nuclear genome and plastome assemblies showed key gene loss and relaxed selection related to its non-functional photosystem and retarded leaf development. These data demonstrated that the improved *G. elata* genome provides more insights into the genomic and evolutionary mechanisms underlying the morphological and physiological adaptations associated with a non-photosynthetic lifestyle.

MATERIALS AND METHODS

Plant Material, Genomic DNA Extraction, and Library Construction

A flowering *G. elata* Bl. f. glauca S. Chow plant was collected from Zhaotong, Yunnan Province, China (103°43′E, 27°20′N). Genomic DNA was extracted from fresh stems by the Qiagen DNeasy Plant Mini Kit (Cat. No. 69104). After genomic DNA purification and quality validation, 10 paired-end and mate-pair DNA libraries with insert sizes of 169 bp (×2), 300 bp (×2), 374 bp, 545 bp, 753 kb, 2 kb, 5 kb, and 10 kb were constructed using library construction kits (Illumina) as previously described (Liang et al., 2015; **Supplementary Table 1**).

Genome Sequencing and Assembly

A whole-genome shotgun strategy was adopted to sequence and assemble the genome of *G. elata*. All ten DNA libraries were sequenced on an Illumina HiSeq 2000 platform and 483.07 Gb of data was obtained. The genome size of *G. elata* was estimated according to the 17-mer frequency distribution with the formula: Genome size = *K*-mer_num/Peak_depth. Total of 84.67 Gb data was retained for 17-mer analysis. The 17-mer frequency distribution plot shows the peak at 66 (**Supplementary Figure 2**), and the total *K*-mer count is 90,917,019,718, then the genome size is estimated as 1.378 Gb.

The assembly of *G. elata* genome was carried out using the SOAPdenovo. A *de Bruijn* graph was constructed using the parameter "−K 83 −d 3", then the graph was simplified using "−M 3" parameters to clip tips, remove low coverage links, merge bubbles, and solve tiny repeats. Those repeats on the simplified graph were broken at the boundaries and unambiguous DNA fragments were produced as contigs. Filtered sequencing data was then realigned to contigs with the parameters "−k 83". Parameter "−F" was used to deal with the paired-end information, and unique contigs were joined to construct scaffolds. Gapcloser (v1.12) was used to fill the gaps between scaffolds with the default parameters.[1] Contigs short than 100 bp were excluded within the whole assembly process.

The completeness and accuracy of the gap closed assembly was assessed using short insert-sized sequencing reads, conserved genes, and RNA-seq data. Filtered reads from three small-insert libraries (169, 374, and 753 bp) were mapped to the assembled genome using the bowtie2 (version 2.2.9[2]) (Langmead and Salzberg, 2012) with default parameters to estimate the quality of the assembly. The overall mapping rates of these small-insert size reads is higher than 92%, suggesting that most of the *G. elata* genome has been assembled. BUSCOs approach (Waterhouse et al., 2017) was also used to evaluate the completeness of the *G. elata* genome. Three hundred and three (303) conserved single-copy orthologs has been used as reference, of which 248 (81.85%) were identified in our genome assembly (**Table 1**). Over 74.59 millions of cleaned RNA-seq reads from three different *G. elata* tissues (flower, stem, and tuber) were mapped to the assembled genome with the default settings using the TopHat2 (version 2.0.14[3]) (Trapnell et al., 2009; **Supplementary Table 2**), and we observed the average mapping ratio of 83%, further validate the high completeness of *G. elata* genome assembly.

RNA-Sequencing Data Analysis

Overall, 82.37 million transcriptome raw reads were obtained from *G. elata* leaf, stem, and tuber tissues (**Supplementary Table 2**). First, all RNA-sequencing (RNA-seq) data were filtered by the strict quality control process, then about 74.59 millions

[1]http://sourceforge.net/projects/soapdenovo2/files/GapCloser/

[2]https://sourceforge.net/projects/bowtie-bio/files/bowtie2/2.3.4.3

[3]http://ccb.jhu.edu/software/tophat/index.shtml

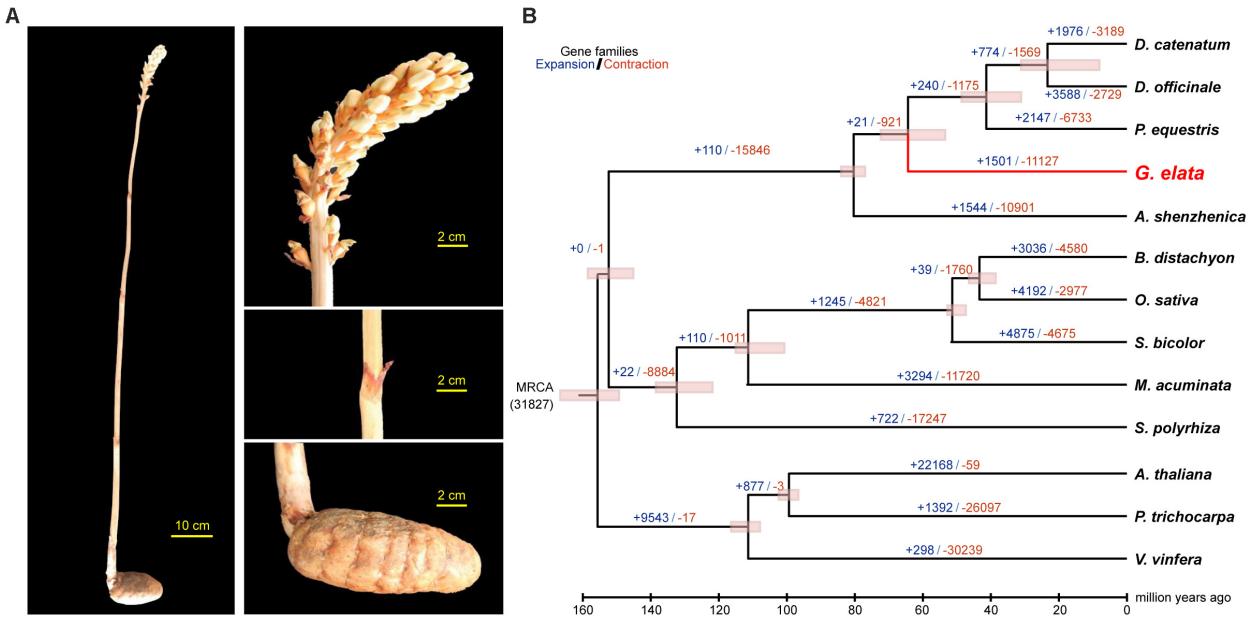


FIGURE 1 | Divergence time and morphology of *G. elata*. **(A)** Morphology of *G. elata*: photos showing a whole plant, flower, stem with degraded blade and tuber. **(B)** Divergence times are indicated by pink bars, expanded gene family numbers by dark blue letters and contracted gene families by orange letters.

TABLE 1 | Various assembly parameters of the *G. elata* genome.

	***G. elata* (This study)**	***G. elata* (Yuan et al., 2018)**
Sequencing platform	Illumina Hiseq 2000	Illumina Hiseq 2500
Sequenced data (Gb)	483.07	179.1
Genome coverage (×)	350.56	151.78
Estimated genome size (Gb)	1.37	1.18
Assembled genome (Gb)	1.12	1.06
Contig N50 (Kb)	110.03	68.97
Scaffold N50 (Mb)	1.64	4.91
TE proportion (%)	69.81%	66.18%
Total BUSCO groups searched	303	956
Complete BUSCOs	81.85%	67.15%
Duplicated BUSCOs	17.49%	9.21%
Fragmented BUSCOs	3.96%	4.39%
Missing BUSCOs	14.19%	28.45%

cleaned reads were *de novo* assembled using trinity[4] (Grabherr et al., 2011) with default settings to yield transcripts that prepared for the genome annotation. Next, all RNA-seq reads were mapped to the *G. elata* genome assembly using the TopHat2 (version 2.0.14; see text footnote 3) (Trapnell et al., 2009) with default settings. The fragments per kilobase of exon model per million reads mapped (FPKM) of each predicted protein-coding gene was calculated by the Cufflinks[5] using default parameters. FPKM > = 0.05 was set as the threshold to identify expressed genes.

[4]https://github.com/trinityrnaseq/trinityrnaseq/releases

[5]http://cole-trapnell-lab.github.io/cufflinks/cuffdiff/index.html

Repeat Sequence Annotation

RepeatMasker (version 3.2.6) with the Repbase TE library was used to identify known transposable elements (TEs) with the default parameters. The library was constructed by generating the consensus sequence of each TE family, which was used for the RepeatMasker to identify additional high and medium copy repeats in the *G. elata* genome assembly. TRF with parameters set to "Match = 2, Mismatch = 7, Delta = 7, PM = 80, PI = 10, Minscore = 50, and MaxPeriod = 2000" was used to predict tandem repeats.

Protein-Coding Gene Annotation

A combination of *de novo*-, transcriptome-based prediction, and homology aligning were used to process gene annotation. Gene sets from ten species (*Arabidopsis thaliana*, *Brachypodium distachyon*, *Oryza sativa*, *Sorghum bicolor*, *Solanum tuberosum*, *Triticum aestivum*, *Hordeum vulgare*, *P. equestris*, *Dendrobium catenatum,* and *A. shenzhenica*) were used for homology-based predictions, one species at a time. We used the TBLASTN to search the non-redundant protein sequences of each gene set with an E-value < 1e-2. Only regions with homologous blocks longer than 80% of the query protein were retained. The best hits were selected. Then, the EVM was used to construct the gene structures. After repeat sequences were masked using the homology-based approach, three softwares, AUGUSTUS (Stanke et al., 2006), SNAP (Korf, 2004), and GlimmerHMM (Majoros et al., 2004), were used for the genes *de novo* prediction. Information obtained from the homology-based predictions and *de novo* predictions were then integrated in the GLEAN to generate a consensus gene set. Finally, RNA-seq data from a single *G. elata* plant's flower, stem, and tuber tissues were

obtained and assembled for the facilitating of the protein-coding gene annotation.

Functional Annotation

Best hits were selected from alignments to the SwissProt and TrEMBL databases to assign gene function information. Gene motifs and domains were identified using the InterProScan5[6] (Jones et al., 2014) by alignment to databases including the ProDom, PRINTS, Pfam, SMART, PANTHER, and PROSITE. According to the corresponding SwissProt and TrEMBL entries information, Gene Ontology (GO) terms and ID for each gene was obtained. Kyoto Encyclopedia of Genes and Genomes (KEGG) protein database was used as reference for gene alignments to obtain KEGG IDs as well as the corresponding pathways information.

Gene Family Clustering

All proteins from selected 13 species *(Gastodia elata*, *P. equestris*, *D. officinale*, *D. catenatum*, *A. shenzhenica*, *B. distachyon*, *O. sativa*, *S. bicolor*, *Spirodela polyrhiza*, *Musa acuminata, A. thaliana*, *Populus trichocarpa,* and *Vitis vinifera*) were aligned using the BLASTP, then, the gene families were defined using the OrthoMCL (Li et al., 2003). CAFÉ (De Bie et al., 2006) was then used to identify gene family expansions and contractions in *G. elata*. To identify gene family clusters in these species and *G. elata*, we performed all-versus-all protein alignments using the BLASTP with the E-value threshold set to "1e-5." We used the OrthoMCL to process high scoring segment pairs. The MCL module from OrthoMCL was then used to define final paralogous and orthologous genes with the parameter set as "−abc –I = 1.5".

Phylogenetic Tree Construction and Divergence Time Estimation

Those single-copy orthologs identified from gene family cluster analysis of the aforementioned species were used to construct a phylogenetic tree. MUSCLE version 3.6[7] (Edgar, 2004) was used with default settings to perform multiple sequence alignments. Fourfold degenerate sites of genes were collected and concatenated into a "super sequence" for each species. We used the MrBayes[8] (Huelsenbeck and Ronquist, 2001) to reconstruct phylogenetic trees between species. The "MCMCTREE" module from the PAML package[9] (Yang, 2007) was used to estimate the divergence time among species.

Chloroplast Genome (Plastome) Assembly and Annotation

The clean reads were aligned to the plastomes of *P. equestris* and *D. officinale* (Jheng et al., 2012; Yang et al., 2016) using the bowtie2 with default settings, respectively. Reads extracted from the "paired-aligned" alignments were merged and assembled using the SOAPdenovo with default parameters. Contigs shorter than 100 bp were excluded. Filtered contigs were joined to scaffolds based on the paired-end information and gaps between scaffolds were closured. DOGMA[10] (Wyman et al., 2004) was used to annotate the protein-coding genes and tRNA genes with the cut-off set to 80%. The boundaries of protein-coding genes and plastome structures were manually checked by comparison to the plastomes of *P. equestris* and *D. officinale*. The linear plot of *G. elata* plastome was yielded by the OGDRAW[11] (Lohse et al., 2013), followed with some manual adjustments.

Relaxation Selection Analyses and Symbiotic Gene Analysis

These nuclear-encoded photosynthesis-related proteins were identified based on the National Center for Biotechnology Information (NCBI) database. Genome sequences of seven species (*G. elata, A. shenzhenica, P. equestris, D. catenatum, D. officinale, O. sativa,* and *A. thaliana*) were searched for all of the known plant nuclear-encoded photosynthesis-related genes. Only genes that were in a one-to-one orthologs for every pair of genomes of the seven species were used in our analyses. For genes that have more than one transcript, we aligned all of the possible transcript pairs to all seven species and retained those that provided the highest alignment scores. Alignments and consensus trees were used for posterior molecular evolutionary analysis. We used a gene-level approach based on the ratio of non-synonymous (K_a) to synonymous (K_s) substitutions rate ($\omega = K_a/K_s$) to identify potential relaxation of selective constraints, using the CODEML likelihood ratio tests (LRTs) algorithm from the PAML package. First, we tested branch models M0, the simplest model, which has a single ω ratio for the entire tree. Subsequently, we used two-ratio models that allow a background ω ratio and a different ω on the branch of interest. For null hypotheses, we used the one-ratio model, two-ratio model, and more models with a fixed $\omega = 1$ on the branch under analysis. The level of significance for these LRTs was calculated using a χ^2 approximation, where twice the difference of log likelihood between the models (2ΔlnL) would be asymptotic to a χ^2 distribution, with the number of degrees of freedom corresponding to the difference in the number of parameters between the nested models.

These Gastrodia antifungal protein (GAFP) were downloaded from the NCBI database. Genome sequences of seven species (*G. elata, A. shenzhenica, P. equestris, D. catenatum, D. officinale, O. sativa,* and *A. thaliana*) were searched for all of the known GAFP genes with blast software. Finally, the genes with E-value $\leq$ 1e-6 in the comparison result were selected as a candidate GAFP gene. And the protein domain of carotenoid cleavage dioxygenases (CCDs) was downloaded from the pfam website. The hmmersearch software were used for sequence alignment to identify CCDs genes in other species. The identified genes are calculated

[6] ftp://ftp.ebi.ac.uk/pub/software/unix/iprscan/5/5.32-71.0/interproscan-5.32-71.0-64-bit.tar.gz

[7] http://www.drive5.com/muscle/downloads.htm

[8] http://nbisweden.github.io/MrBayes/download.html

[9] http://evomics.org/resources/software/molecular-evolution-software/paml/

[10] http://dogma.ccbb.utexas.edu

[11] https://chlorobox.mpimp-golm.mpg.de/OGDraw-Downloads.html

by K_a/K_s to verify whether positive selection occurred in *G. elata*.

High Performance Liquid Chromatography Analysis of Photosynthetic Pigments

From each sample of *G. elata*'s vestigial scalelike leaves and *P. equestris*'s and *D. Officinale*'s normal leaves, about 1 g of leave tissue was collected and dried as powder. Then, the dried powder was dealt with 100% 40 mL of methanol for half an hour and sonicated for an hour, and then, methanol diluted to 50 mL. The methanol extract was then filtered by the 0.45 μm membrane filter. Ten microliters of filtrate was prepared for high performance liquid chromatography (HPLC). Quantitative analysis of the photosynthetic pigments of three species was performed using the chromatographic column Inertsil ODS-3 (250 mm × 4.6 mm, 5 μm) and the column temperature was maintained at 25°C with the flow rate set as 1 mL × min^{-1}. The mobile phase was consisted of A: acetonitrile and 0.05 mol/L of Tris–HC1 buffer (70:3); and B: methanol and n-hexane (5:1). Gradient elution was then used with the following system: 100% A at initiation, 100% A at 18 min, 100% B at 20.5 min, and 100% B at 46 min. 445 nm was used to detect photosynthetic pigments.

RESULTS

De novo Assembly of the *G. elata* Genome

The genome size of *G. elata* was estimated to be 1.378 Gb using the *K*-mer distribution analysis (**Supplementary Figure 2**). All 483.07 Gb clean data (350.56 × genome coverage) were *de novo* assembled into 69,353 contigs (1.11 Gb in total length) with a contig N50 size of 110.03 kb, and 45,884 scaffolds (1.12 Gb in total length) with a scaffold N50 size of 1.64 Mb (**Supplementary Table 3**). As a result of high genome coverage, our *G. elata* assembly had a much longer contig N50 size (110.03 kb) than that from a previous study (68.9 kb, **Table 1**; Yuan et al., 2018). Preliminary evaluation of the quality of our assembly showed that at least 92% of the clean reads from different insert-sizes of paired-end libraries could be mapped back to the assembled genome (**Supplementary Table 4**). In addition, over 82.37 million RNA sequencing reads from each of the flower, stem, and tuber tissues were generated to further verify the quality of the assembly. The overall mapping ratios of these reads to the genome assembly were 90.7% for the flower, 93.8% for the stem, and 67.1% for the tuber (**Supplementary Table 2**), indicating the high quality of the *G. elata* genome.

The completeness of the genic regions and other genomic elements in our *G. elata* genome was evaluated by the BUSCOs approach (Waterhouse et al., 2017). The result showed that 248 out of 303 (81.85%) near-universal single-copy orthologs were identified in our *G. elata* genome. This number was much higher than that reported in the previous study (67.15% BUSCO completeness, **Table 1**; Yuan et al., 2018). Additionally, 12 BUSCO genes (3.96%) had fragmented matches, and 43 BUSCO genes (14.19%) were missing in our assembly. Both parameters were lower than that reported in the previous study (4.39 and 28.45%, respectively, **Table 1**; Yuan et al., 2018). These data demonstrated that our *G. elata* genome had a much higher completeness for subsequent functional analysis.

Gastrodia elata Genome Repeat Analysis

Repetitive DNA elements constituted approximately 68.34% of the acquired genome (**Supplementary Table 5**). This proportion was higher than those of the other orchids genomes, including *P. equestris* (62%) (Cai et al., 2015), *D. officinale* (63.33%) (Liang et al., 2015), and *A. shenzhenica* (42.05%) (Zhang et al., 2017). Long interspersed nuclear elements (LINEs) and DNA transposons constituted 6.20 and 7.49% of the total genome assembly, respectively (**Supplementary Table 5**). The long terminal repeats (LTRs) were the most abundant TEs, which constituted 59.95% of the genome (**Supplementary Table 5**). Notably, Gypsy-type (37.41%) and Copia-type (8.07%) TEs accounted for most of the LTRs (**Supplementary Table 6**). Sequence divergence line plot of Copia and Gypsy types of TEs shows that both of the two types experienced two recently expansion events during the same periods (**Supplementary Figure 3a**). Compared to the conserved single copy genes, those duplicated genes ("accessory genes") in *G. elata* are located closer to the TEs (Fisher's exact test, $P < 0.005$), suggesting their contribution in gene copy number variation and the important role in *G. elata* evolution (**Supplementary Figure 3b**).

Orthologous Gene Analysis in the *G. elata* Genome

Overall, we identified 24,484 protein-coding genes (**Supplementary Table 7**) with about 10.05 kb average length in the *G. elata* genome that belonged to 11,065 unique gene families (**Supplementary Table 8**). Besides, we also identified a myriad of microRNAs (1,125), transfer RNAs (1,123), ribosomal RNAs (1,596), and small nuclear RNAs (868; **Supplementary Table 9**) in the *G. elata* genome.

Phylogenetic analysis with divergence time estimation showed that the *G. elata* diverged from other orchids somewhere between 72.3 and 55.8 million years ago (Mya) (**Figure 1B**). During the course of evolution, it is apparent that 19 gene families were significantly expanded ($P < 0.001$; **Supplementary Table 10**), whereas six gene families were significantly contracted in the *G. elata* genome ($P < 0.001$; **Supplementary Table 11**). KEGG enrichment analyses showed that gene families involved in flavonoid biosynthesis, plant-pathogen interaction, and circadian rhythm were significantly contracted in the *G. elata* genome (**Supplementary Table 12**). In comparison, the expanded gene families are mainly involved in the glycan, sphingolipid, and galactose metabolisms, and intriguingly, also plant-pathogen interaction (**Supplementary Table 13**). Since some members of the plant-pathogen interaction gene families were contracted and other members were expanded, it is possible that *G. elata* may

have rewired its pathogen resistance pathway when adapting to the low-light environment. Indeed, *G. elata* as a fully mycoheterotrophic plant not only needs to obtain nutrients from the fungi associates, but also have to prevent the fungi from invading into the inner most section of the tuber (Zhang, 1980).

Gastrodia elata Plastome Assembly and Relaxed Selection of Plastid Genes

The conserved plastome structure of most land plants generally is a quadripartite single circular molecule of 100–220 kb. It consists of a small and a large single-copy regions (SSC and LSC) which are separated by a pair of inverted repeats (IRs) (Sandelius and Aronsson, 2009). The plastome of the photosynthetic plant harbors around 100 essential genes that primarily encode RNAs and proteins involved in photosynthesis, transcription, and translation (Daniell et al., 2016). *P. equestris* and *D. officinale* are two photosynthetic orchids with fully functional plastomes as other land plants (Jheng et al., 2012; Yang et al., 2016). The plastomes of these two orchids (with sizes of 148,959 and 152,018 bp, respectively) both have the typical quadripartite single circular molecule structure with relatively complete photosynthesis-related gene sets (**Figure 2**). In comparison, the plastome of *G. elata* was assembled with the size of 40,037 bp, meaning that more than two thirds of the plastome sequence was lost after the transition to a low-light heterotrophic lifestyle (**Figure 2**). IRs are considered to play roles in plastome stabilization (Palmer and Thompson, 1982). In the *G. elata* plastome, one copy of IR was completely lost, and the remaining IR sequence was less than half the size of its counterparts in the other two orchids. The *G. elata* LSC retained less than a quarter of the sequence size compared to the two orchid plastomes, and the *G. elata* SSC also lost half of its size.

Only 50 genes were identified in the *G. elata* plastome, among which 10 genes were manually checked to be pseudogenes (with frameshifts or premature stop codons) (**Supplementary Table 14**). The photosynthesis-related genes in the *G. elata* plastome were carefully checked. The result showed that the paralogs of *psa* and *psb* (photosystem I and II genes), *pet* (cytochrome b/f complex genes), *rbc* (rubisco subunit genes), and *atp* (ATP synthase genes) were completely lost or were found to have incomplete gene structures due to a reduced LSC. Two of the six remaining *ndh* (NADH dehydrogenase) genes were found to be non-functional. All these changes concluded that the genomic basis of photosynthesis in the *G. elata* plastome was highly degenerated. These observations are in accordance with a previous report (Yuan et al., 2018), which includes small assembled plastome sizes (40,026 vs. 35,326 bp), extensively loss of DNA fragments, missing of one copy of IR structure, and loss or non-functional of photosynthesis-related genes.

Natural selection is essential in the maintenance of trait for the plant population. The weakening or disappear of selection strength that leads to trait variations is referred to as "relaxed selection" (Lahti et al., 2009). In its evolutionary history, *G. elata* experienced the transition to a low-light, undergrowth environment (Bidartondo, 2005) and a fully heterotrophic lifestyle. This transition made photosynthesis dispensable for *G. elata* to survive. We therefore hypothesized that the remaining photosynthesis-related plastome genes might be under relaxed selection. Indeed, the K_a/K_s values of these genes (includes *psa*, *psb*, *atp*, *pet* genes, and so on) in *G. elata* were higher than *D. officinale* and *P. equestris*, and non-orchid plant species (**Supplementary Table 15**). This indicates the reduction of functional constraints, and, thus, these genes were under a significantly different selective regime in *G. elata* than in those photosynthetically active species.

Relaxation of Selective Constraints of Nuclear-Coded Photosynthesis-Related Genes in *G. elata*

We define "photosynthesis-related genes" as all known genes that are involved in the structure, development, and normal function of the plastid and leaf in land plants. The plastome only contains a small part of these photosynthesis-related genes, and the rest are within the plant nuclear genome. Based on the aforementioned hypothesis, we also checked the molecular evolution of the photosynthesis-related nuclear genes in the *G. elata* genome. In brief, they consisted of 4,818 plastid-related genes, 203 leaf development-related genes, and 1,408 genes that were directly involved in the photosynthesis. The K_a/K_s values of these genes in *G. elata* were significantly higher than those in other orchids and non-orchid plant species, indicating relaxation of selective constraints on the leaf development and the photosynthesis process (*t*-test and two-way ANOVA test, $p < 0.05$, **Figure 3** and **Supplementary Figure 4**). We also found that the K_a/K_s values of these searched genes in orchids were higher than those in non-orchid species. It suggested that there might be a partial relaxation of selection on plastid and leave functions in the photosynthetic orchids due to their symbiotic relationships with fungi. These findings also suggested a positive correlation between the degree of heterotrophy in plants and the non-synonymous mutation rates in genes that were involved in the photosynthetic process, plastid and leave functions.

Analysis of Symbiotic Genes in *G. elata* Genome

Since *G. elata* cannot perform photosynthesis, it relies on symbiotic genes for nutrition, so symbiotic genes are also attracting attention in the *Analysis of Symbiotic Genes in Gastrodia Genome*. *G. elata* exists underground without leaves and chloroplasts, with the nutrients necessary for growth being supplied by the symbiotic fungus *Armillaria mellea* (Sa et al., 2003) fungus. Gastrodia antifungal protein (GAFP, also known as gastrodianin) was first purified from the cortex of the terminal corm of *G. elata* (Hu et al., 1988). It is mainly distributed throughout the epidermis and cortex layer of *G. elata* (Liu et al., 1993; Hu and Huang, 1994) and shows a strong fungistatic activity against a broad spectrum of fungi, including *A. mellea* (Hu and Huang, 1994), *Rhizoctonia solani*, *Valsa ambiens*, *Gibberella zeae*, *Ganoderma lucidum,* and *Botrytis*

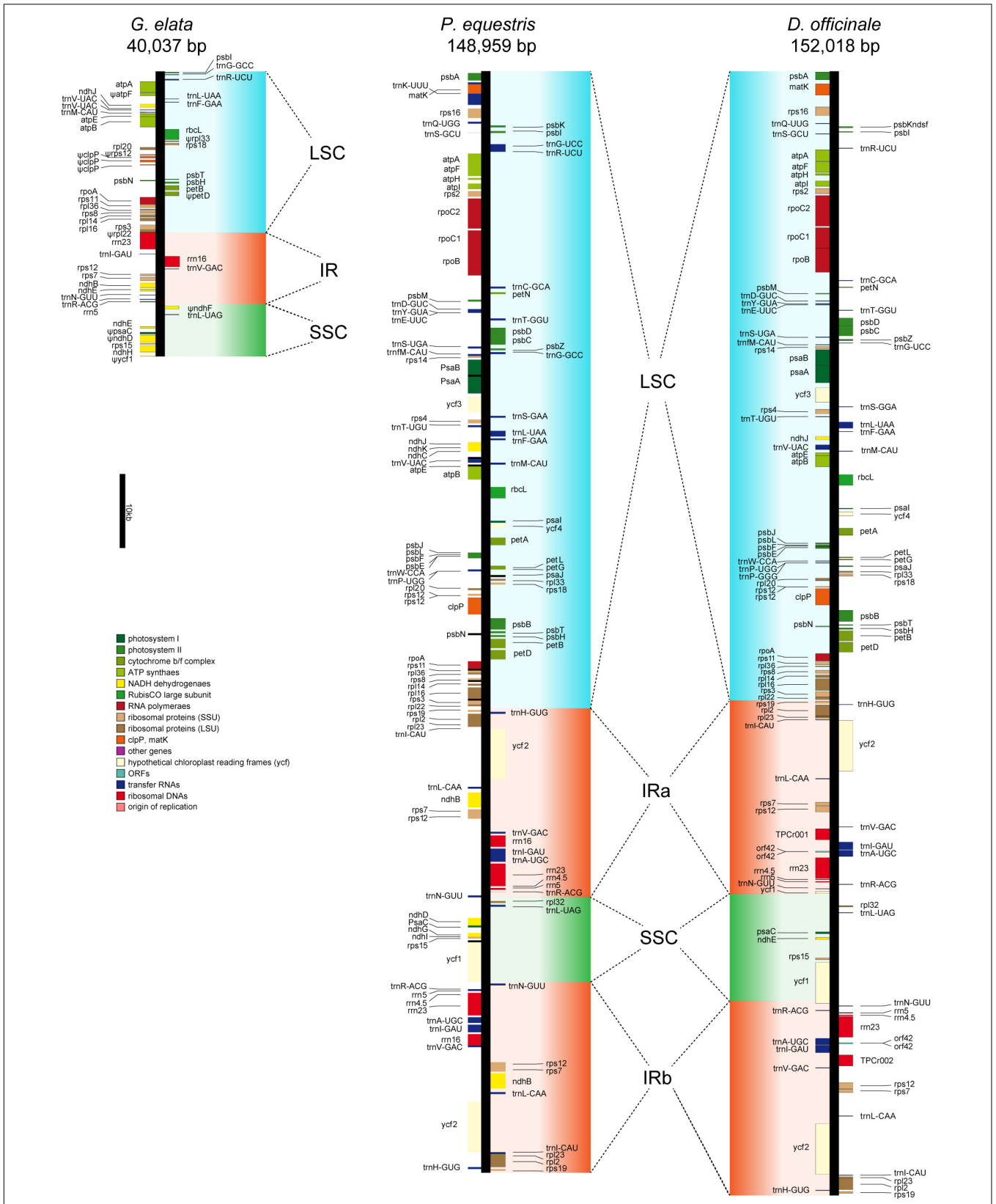


FIGURE 2 | Plastid map of three orchid plants. The plastid of *G. elata* shows significant degradation comparing with *P. equestris* and *D. officinale*. Ψ indicates the pseudogenes.

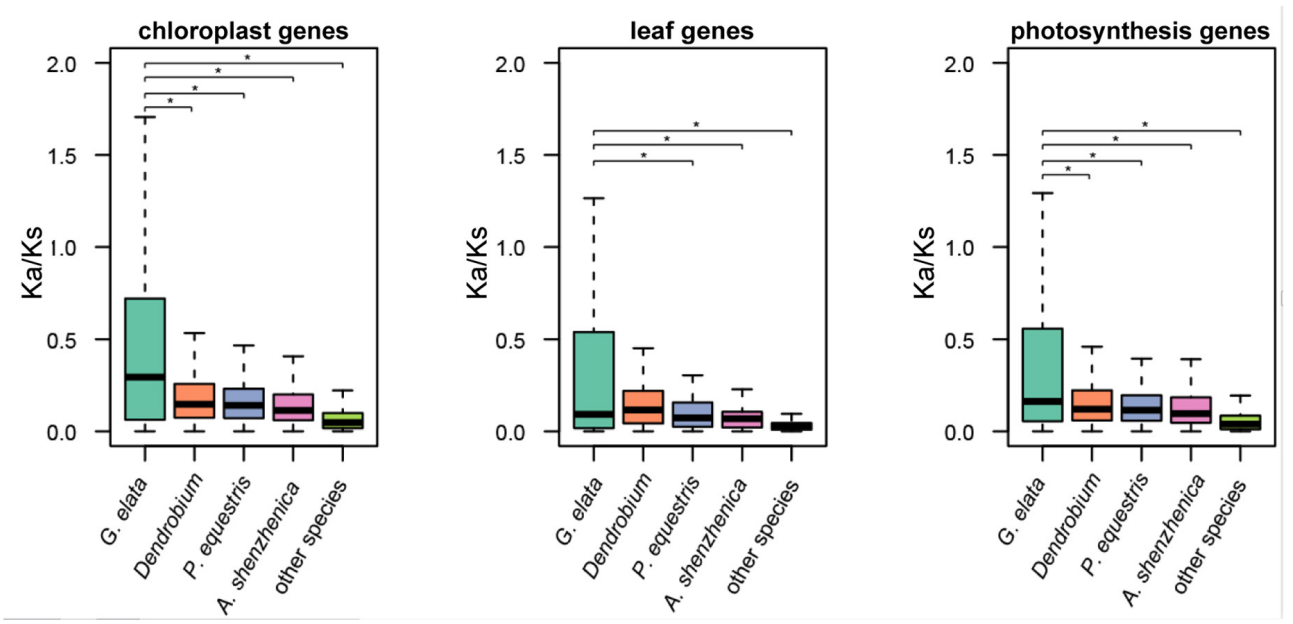


FIGURE 3 | Relaxation of selective constraints on nuclear-encoded genes. K_a/K_s ratios of nuclear-encoded plastid genes and genes involved in leaf development in *G. elata*, other orchids (including *D. officinale*, *P. equestris*, *D. catenatum,* and *A. shenzhenica*) and other green plants (including *A. thaliana* and *O. sativa*). *$P < 0.05$.

cinerea in vitro to reach plants symbiosis with microorganisms (Wang et al., 2001). In the *G. elata* genome, we identified 20 GAFP genes, 10 GAFP genes in *A. shenzhenica*, and 0 in *A. thaliana*, which indicates that GAFP has expanded in *G. elata* (**Supplementary Table 16**). It is known that strigolactone can stimulate hyphal branching and development of arbuscular mycorrhizal fungi, which increases the chances of an encounter with a host plant (Kretzschmar et al., 2012). Strigolactone is an important signal for the establishment of the symbiosis relationship between *G. elata* and *A. mellea*. Its mechanism of action is similar to that of promoting symbiosis between plants and mycorrhiza. In this article, we have identified the key genes for the biosynthesis of strigolactone CCDs (Delaux et al., 2012). In the *G. elata* genome, 12 CCDs genes were identified, whereas ten and nine CCDs genes were identified in *A. shenzhenica* and *A. thaliana*, respectively. In order to verify whether the CCDs gene has positive selection in *G. elata*, we further performed the K_a/K_s analysis. The K_a/K_s values of these genes in *G. elata* were significantly higher than those in *D. catenatum* and *P. equestris*. It showed that, compared with some orchid and non-orchid plant species, CCDs have expanded in the *G. elata* genome.

Genomic Basis for the Achlorophyllous Phenotype in *G. elata*

We next investigated nuclear-encoded genes involved in the plant pigment synthesis in the *G. elata* genome. We detected no copies of four core genes in the chlorophyll biosynthetic pathway (*UROS*, *CAO*, *LPOR*, and *VDE*; **Supplementary Figure 5a**). This is in line with the result that no signals of major photosynthetic pigments could be detected via the HPLC in the *G. elata* plant tissue (compared to the other two orchids *D. officinale* and *P. equestris*) (**Supplementary Figure 5b**). Moreover, three essential genes (*PDV1*, *PDV2,* and *ARC3*) involved in the chloroplast division were also not detected (**Supplementary Figure 6**). These results clearly showed the genomic basis for the achlorophyllous phenotype in *G. elata*.

CONCLUSION

The *G. elata* genome provides an illuminating model for probing evolutionary molecular changes associated with leaf loss, plastid degeneration, and plant-fungal symbioses. Both plastid-encoded and nuclear-encoded photosynthesis-related genes in *G. elata* showed evidence of the relaxation of selective constraints. Besides the profound changes in the photosynthetic system, it is possible that future research could identify the underlying mechanisms for the rewiring of basic metabolic pathways, the evolution of TEs, gene death, and generation of new genes, since an improved *G. elata* genome is now available. Thus, the genome sequence of *G. elata* would be a valuable resource for future investigations of the evolution of orchids and non-photosynthetic plants at both the genomic and the whole-organism levels.

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: NCBI database under BioProject PRJNA394702.

AUTHOR CONTRIBUTIONS

SSC, GHZ, and WLS collected the samples and performed the experiments. XW, YZW, and XD completed the data analysis. YD, WC, WW, MA, and JHM edited and modified the manuscript. All authors read and approved the manuscript.

FUNDING

This study was supported by "The Thousand Talents Program" (Class B) of the government of China, National Key R&D Program of China (2019YFC1711100), Yunnan Provincial Key Programs of Yunnan Eco-friendly Food International Cooperation Research Center Project (2019ZG00908), and the Guangxi Innovation-Driven Development Project (GuiKe AA18242040).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.580568/full#supplementary-material

REFERENCES

Barrett, C. F., and Davis, J. I. (2012). The plastid genome of the mycoheterotrophic *Corallorhiza striata* (Orchidaceae) is in the relatively early stages of degradation. *Am. J. Bot.* 99, 1513–1523. doi: 10.3732/ajb.1200256

Bidartondo, M. I. (2005). The evolutionary ecology of myco-heterotrophy. *New Phytol.* 167, 335–352. doi: 10.1111/j.1469-8137.2005.01429.x

Cai, J., Liu, X., Vanneste, K., Proost, S., Tsai, W. C., Liu, K. W., et al. (2015). The genome sequence of the orchid *Phalaenopsis equestris*. *Nat. Genet.* 47, 65–72. doi: 10.1038/ng.3149

Cribb, P., and Killmann, E. F. A. D. (2010). A revision of Gastrodia (Orchidaceae: Epidendroideae, Gastrodieae) in tropical Africa. *Kew Bull.* 65, 315–321. doi: 10.1007/s12225-010-9193-4

Daniell, H., Lin, C. S., Yu, M., and Chang, W. J. (2016). Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biol.* 17:134. doi: 10.1186/s13059-016-1004-2

De Bie, T., Cristianini, N., Demuth, J. P., and Hahn, M. W. (2006). CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22, 1269–1271. doi: 10.1093/bioinformatics/btl097

Delaux, P. M., Xie, X., Timme, R. E., Puech-Pages, V., Dunand, C., Lecompte, E., et al. (2012). Origin of strigolactones in the green lineage. *New Phytol.* 195, 857–871. doi: 10.1111/j.1469-8137.2012.04209.x

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Hu, Z., and Huang, Q. Z. (1994). Induction and accumulation of the antifungal protein in *Gastrodia elata*. *Acta Bot. Yunnan* 16, 169–177.

Hu, Z., Yang, Z., and Wang, J. (1988). Isolation and partial characterization of an antifungal protein from Gastrodia elata corm. *Acta Bot. Yunnanica* 10, 373–380.

Huelsenbeck, J. P., and Ronquist, F. (2001). MRBAYES: bayesian inference of phylogenetic trees. *Bioinformatics* 17, 754–755. doi: 10.1093/bioinformatics/17.8.754

Jheng, C. F., Chen, T. C., Lin, J. Y., Chen, T. C., Wu, W. L., and Chang, C. C. (2012). The comparative chloroplast genomic analysis of photosynthetic orchids and developing DNA markers to distinguish *Phalaenopsis orchids*. *Plant Sci.* 190, 62–73. doi: 10.1016/j.plantsci.2012.04.001

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, 1236–1240. doi: 10.1093/bioinformatics/btu031

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics* 5:59. doi: 10.1186/1471-2105-5-59

Krause, K. (2008). From chloroplasts to "cryptic" plastids: evolution of plastid genomes in parasitic plants. *Curr. Genet.* 54, 111–121. doi: 10.1007/s00294-008-0208-8

Kretzschmar, T., Kohlen, W., Sasse, J., Borghi, L., Schlegel, M., Bachelier, J. B., et al. (2012). A petunia ABC protein controls strigolactone-dependent symbiotic signalling and branching. *Nature* 483, 341–344. doi: 10.1038/nature10873

Lahti, D. C., Johnson, N. A., Ajie, B. C., Otto, S. P., Hendry, A. P., Blumstein, D. T., et al. (2009). Relaxed selection in the wild. *Trends Ecol. Evol.* 24, 487–496. doi: 10.1016/j.tree.2009.03.010

Lan, J., Xu, J. T., and Li, J. S. (1994). Study on symbiotic relation between Gastrodia elata and Armillariella mellea by autoradiography. *Mycosystema* 13, 219–222.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Leake, J. R. (1994). The biology of myco-heterotrophic ('saprophytic') plants. *New Phytol.* 127, 171–216. doi: 10.1111/j.1469-8137.1994.tb04272.x

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189. doi: 10.1101/gr.1224503

Liang, Y., Xiao, W., Hui, L., Yang, T., Lian, J., Yang, R., et al. (2015). The genome of dendrobium officinale illuminates the biology of the important traditional chinese orchid herb. *Mol. Plant* 008, 922–934. doi: 10.1016/j.molp.2014.12.011

Liu, J., Jin-tang, X., He, W., Hong-xiang, L., and Yong-ru, S. (1993). Detection and immunofiuorescence localization of antifungal protein of gastrodiaelata. *J. Integr. Plant Biol.* 35, 593–599.

Lohse, M., Drechsel, O., Kahlau, S., and Bock, R. (2013). OrganellarGenomeDRAW–a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Res.* 41, W575–W581. doi: 10.1093/nar/gkt289

Majoros, W. H., Pertea, M., and Salzberg, S. L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* 20, 2878–2879. doi: 10.1093/bioinformatics/bth315

Merckx, V., Bidartondo, M. I., and Hynson, N. A. (2009). Myco-heterotrophy: when fungi host plants. *Ann. Bot.* 104, 1255–1261. doi: 10.1093/aob/mcp235

Merckx, V., and Freudenstein, J. V. (2010). Evolution of mycoheterotrophy in plants: a phylogenetic perspective. *New Phytol.* 185, 605–609. doi: 10.1111/j.1469-8137.2009.03155.x

Ning, Z., Bai, X. F., Jian-Tao, L. V., Yang, J. F., and Gui-Ping, X. U. (2010). Study on symbiotic mechanism between gastrodia elata blume and armillaria mellea in tissue culture system. *Med. Plant* 34, 95–96.

Palmer, J. D., and Thompson, W. F. (1982). Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. *Cell* 29, 537–550. doi: 10.1016/0092-8674(82)90170-2

Read, D. J. (1991). Mycorrhizas in ecosystems. *Cell. Mol. Life Sci.* 47, 376–391. doi: 10.1007/BF01972080

Sa, Q., Wang, Y., Li, W., Zhang, L., and Sun, Y. (2003). The promoter of an antifungal protein gene from *Gastrodia elata* confers tissue -specific and fungus-inducible expression patterns and responds to both salicylic acid and jasmonic acid. *Plant Cell Rep.* 22, 79–84. doi: 10.1007/s00299-003-0664-z

Sandelius, A. S., and Aronsson, H. (2009). "The chloroplast: interactions with the environment," in *Plant Cell Monographs*, ed. P. Nick (Cham: Springer).

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Res.* 34, W435–W439. doi: 10.1093/nar/gkl200

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–1111. doi: 10.1093/bioinformatics/btp120

Wang, X., Bauw, G., Van Damme, E. J., Peumans, W. J., Chen, Z. L., Van Montagu, M., et al. (2001). Gastrodianin-like mannose-binding proteins: a novel class of plant proteins with antifungal properties. *Plant J.* 25, 651–661. doi: 10.1046/j.1365-313x.2001.00999.x

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., et al. (2017). BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* 35, 543–548. doi: 10.1093/molbev/msx319

Wicke, S., Muller, K. F., De Pamphilis, C. W., Quandt, D., Wickett, N. J., Zhang, Y., et al. (2013). Mechanisms of functional and physical genome reduction in photosynthetic and nonphotosynthetic parasitic plants of the broomrape family. *Plant Cell* 25, 3711–3725. doi: 10.1105/tpc.113.113373

Wicke, S., Schaferhoff, B., DePamphilis, C. W., and Muller, K. F. (2014). Disproportional plastome-wide increase of substitution rates and relaxed purifying selection in genes of carnivorous Lentibulariaceae. *Mol. Biol. Evol.* 31, 529–545. doi: 10.1093/molbev/mst261

Wyman, S. K., Jansen, R. K., and Boore, J. L. (2004). Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20, 3252–3255. doi: 10.1093/bioinformatics/bth352

Yan, L., Wang, X., Liu, H., Tian, Y., Lian, J., Yang, R., et al. (2015). The genome of dendrobium officinale illuminates the biology of the important traditional chinese orchid herb. *Mol. Plant* 8, 922–934. doi: 10.1016/j.molp.2014.12.011

Yang, P., Zhou, H., Qian, J., Xu, H., Shao, Q., Li, Y., et al. (2016). The complete chloroplast genome sequence of *Dendrobium officinale*. *Mitochondrial DNA DNA Mapp. Seq. Anal.* 27, 1262–1264. doi: 10.3109/19401736.2014.945547

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Yuan, Y., Jin, X., Liu, J., Zhao, X., Zhou, J., Wang, X., et al. (2018). The Gastrodia elata genome provides insights into plant adaptation to heterotrophy. *Nat. Commun.* 9:1615. doi: 10.1038/s41467-018-03423-5

Zhang, G. Q., Liu, K. W., Li, Z., Lohaus, R., Hsiao, Y. Y., Niu, S. C., et al. (2017). The Apostasia genome and the evolution of orchids. *Nature* 549, 379–383. doi: 10.1038/nature23897

Zhang, W. J. (1980). The biological relationship of gastrodia elata and armillaria mellea. *Bull. Bot.* 22, 57–62.

Conflict of Interest: XW was employed by the company Jiaxing Synbiolab Biotechnology Co., Ltd. and she declared that Jiaxing Synbiolab Biotechnology Co., Ltd. plays no role in the funding, design, analysis, and publication of this manuscript.

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

Copyright © 2020 Chen, Wang, Wang, Zhang, Song, Dong, Arnold, Wang, Miao, Chen and Dong. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

DATA REPORT
published: 30 November 2020
doi: 10.3389/fgene.2020.565146

Transcriptomes of Different Tissues of Flax (*Linum usitatissimum* L.) Cultivars With Diverse Characteristics

Alexey A. Dmitriev[1†], *Roman O. Novakovskiy*[1†], *Elena N. Pushkova*[1†], *Tatiana A. Rozhmina*[1,2], *Alexander A. Zhuchenko*[2,3], *Nadezhda L. Bolsheva*[1], *Artemy D. Beniaminov*[1], *Vladimir A. Mitkevich*[1], *Liubov V. Povkhova*[1,4], *Ekaterina M. Dvorianinova*[1,4], *Anastasiya V. Snezhkina*[1], *Anna V. Kudryavtseva*[1], *George S. Krasnov*[1] *and Nataliya V. Melnikova*[1*]

[1] *Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia,* [2] *Federal Research Center for Bast Fiber Crops, Torzhok, Russia,* [3] *All-Russian Horticultural Institute for Breeding, Agrotechnology and Nursery, Moscow, Russia,* [4] *Moscow Institute of Physics and Technology, Dolgoprudny, Russia*

Keywords: flax, *Linum usitatissimum* L., transcriptome sequencing, comparative genomics, genotype-phenotype, gene expression, DNA polymorphism

OPEN ACCESS

Edited by:
Weike Duan,
Huaiyin Institute of Technology, China

Reviewed by:
Himanshu Sharma,
National Agri-Food Biotechnology Institute, India
Catalina Iulia Pislariu,
Texas Woman's University, United States

***Correspondence:**
Nataliya V. Melnikova
mnv-4529264@yandex.ru

[†]*These authors have contributed equally to this work*

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

Received: *24 May 2020*
Accepted: *20 October 2020*
Published: *30 November 2020*

Citation:
Dmitriev AA, Novakovskiy RO, Pushkova EN, Rozhmina TA, Zhuchenko AA, Bolsheva NL, Beniaminov AD, Mitkevich VA, Povkhova LV, Dvorianinova EM, Snezhkina AV, Kudryavtseva AV, Krasnov GS and Melnikova NV (2020) Transcriptomes of Different Tissues of Flax (Linum usitatissimum L.) Cultivars With Diverse Characteristics. Front. Genet. 11:565146. doi: 10.3389/fgene.2020.565146

INTRODUCTION

Flax (*Linum usitatissimum* L.) is widely used to produce fiber and seed. Linseed is the richest source of omega-3 fatty acids, which reduce the risk of cancer and cardiovascular diseases, and lignans, which have antibacterial, antifungicide, antioxidant, and anticarcinogenic activities, and also contains easily digestible proteins, dietary fibers, vitamins, and minerals (Muir and Westcott, 2003; Goyal et al., 2014; Imran et al., 2015; Kezimana et al., 2018; Parikh et al., 2018; Cullis, 2019; Mali et al., 2019). Linseed is used in environment-friendly paints and varnishes and also in animal feed to obtain products with increased content of polyunsaturated fatty acids (Kouba and Mourot, 2011; Singh et al., 2011; Goyal et al., 2014). Flax fiber is valuable for the production of textile and composite materials (Costa et al., 2018; Baley et al., 2019). To obtain high and stable yields of organic flax products, it is necessary to cultivate varieties that are resistant to adverse environmental factors and possess a complex of economically valuable traits. The use of traditional methods of breeding requires up to 12–15 years to create a new cultivar. Biotechnologies, including genome editing and marker-assisted and genomic selection, can significantly increase the accuracy and efficiency of the breeding process (Dwivedi et al., 2018; Cobb et al., 2019; Gionfriddo et al., 2019; Mascher et al., 2019; Varshney et al., 2019). For the development and introduction of biotechnologies into practice and breeding of improved cultivars of *L. usitatissimum*, large-scale studies of genomes and transcriptomes on representative sets of flax samples with diverse characteristics are needed.

The flax genome was sequenced and a significant number of transcriptomic studies were performed (Wang et al., 2012; You et al., 2018; Cullis, 2019; Akhmetshina et al., 2020) that laid the foundation for identification of genes that are responsible for valuable traits. Using transcriptome sequencing, the search was performed for genes that are associated with the following flax features:

1) Fiber characteristics in varieties Baihua (Long et al., 2012), Zhongya 2 (Guo et al., 2017), and Mogilevsky (Gorshkov et al., 2017; Mokshina et al., 2017; Gorshkova et al., 2018);
2) Seed characteristics in varieties NEW and Shuangya 4 (Xie et al., 2019);

3) Response to *Fusarium oxysporum* infection in varieties CDC Bethune, Lutea (Galindo-Gonzalez and Deyholos, 2016), Dakota, #3896, AP5, TOST (Dmitriev et al., 2017), and Nike (Preisner et al., 2018);
4) Response to drought in variety T-397 (Dash et al., 2017);
5) Response to unfavorable soil pH and content of macro- and microelements in varieties Heiya No.19 (Yu et al., 2014), Hermes, TMP1919, Lira, Orshanskiy (Dmitriev et al., 2016a), CDC Bethune, Stormont Cirrus (Dmitriev et al., 2016b), Norlin, Mogilevsky (Dmitriev et al., 2019), and Agatha (Wu et al., 2019).

However, in most studies, only one or two cultivars/lines were used, but for the comparative analysis and search for a gene function, much more genotypes with diverse agronomically important traits should be investigated. In the present work, we performed transcriptome sequencing of five different tissues of six flax cultivars/lines with a diverse feature set.

MATERIALS AND METHODS

Plant Materials

Five flax cultivars (Alizee, Atlant, Diplomat, LM98, and Universal) and one line (#3896) were chosen for the present study based on their breeding value and differences in agronomically important parameters such as productivity, fiber and seed characteristics, and resistance to stresses. The characteristics of examined genotypes are represented in **Table 1** (Ryzhov et al., 2012; Rozhmina and Loshakova, 2016; Pavlova et al., 2018; Kolotov, 2020; Rozhmina et al., 2020). Seeds were obtained from the Institute for Flax (Torzhok, Russia).

Flax seeds were sterilized in 1% sodium hypochlorite for 2 min. Seedlings were grown in Petri dishes for 7 days, and then roots and shoots were collected from five plants for each cultivar/line and frozen in liquid nitrogen until further use. Also, plants were grown in the greenhouse for 6 weeks to the flowering stage, and after that, plant materials were collected from leaves, stems (the upper and middle part of the plant), and flowers of five plants for each cultivar/line and immediately frozen in liquid nitrogen until further use. Samples were stored at −75°C before RNA extraction.

RNA Extraction and Transcriptome Sequencing

RNA was extracted from pools of five plants for each combination of tissue-genotype. Plant materials were homogenized using MagNA Lyser (Roche, Switzerland) in 600 µl of RNA lysis buffer from a Quick-RNA Miniprep Kit (Zymo Research, United States) with solid-glass beads (Sigma-Aldrich, United States) and then RNA was extracted according to the Quick-RNA Miniprep Kit protocol with in-column DNase I treatment. Quality and concentration of RNA were evaluated using 2100 Bioanalyzer (Agilent Technologies, United States) and a Qubit 2.0 fluorometer (Thermo Fisher Scientific, United States). RIN (RNA Integrity Number) values were more than 8 and close between samples. NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs, United Kingdom) and NEBNext

TABLE 1 | Characteristics of six examined flax cultivars/lines.

Name	Type	Height, cm	Productivity, c/ha			Fiber content, %	Long fiber, %	Long fiber number	Oil content, %	Weight of 1,000 seeds	Vegetation period, days	Resistance to lodging, score	Resistance, %		
			Straw	Seed	Fiber								Fusarium wilt	Rust	pH 6.2
Alizee	Fiber flax	71	75.3	8.6	23.5	31.2	17.3	12.3	37.1	5.44	90	9.0	54.0	99.1	44.6
Atlant	Fiber flax	84	80.7	8.7	20.5	25.4	19.6	14.2	38.4	4.29	82	8.0	90.9	88.4	75.3
Diplomat	Fiber flax	69	63.2	10.7	21.1	33.4	13.4	12.9	37.5	4.65	77	8.5	80.2	99.7	31.3
Universal	Fiber flax	73	56.4	10.6	17.2	30.5	13.7	12.7	38.9	4.79	71	7.5	92.5	89.4	70.9
#3896	Linseed	66	24.3	16.0	4.7	19.4	6.8	-	44.0	7.05	91	8.5	94.4	87.9	95.2
LM98	Linseed	60	52.9	19.1	11.4	21.6	9.4	-	42.8	5.34	108	8.0	6.7	77.5	58.4

c/ha—centners per hectare.

Ultra II Directional RNA Library Prep Kit for Illumina (New England Biolabs) were used for isolation of mRNA from 1 μg of total RNA and cDNA library preparation according to the manufacturer's protocols. In total, 30 libraries were obtained—from roots and shoots of seedlings and leaves, flowers, and stems of adult plants for each of six flax cultivars/lines. The quality of cDNA libraries was evaluated using 2100 Bioanalyzer (Agilent Technologies)—they had an optimal length distribution and were free of adapter dimers. After that, the libraries were sequenced on NextSeq 500 (Illumina, United States) with a read length of 86 bp.

Preliminary Data Analysis

Transcriptome sequencing of 30 cDNA libraries from five different tissues (leaves, flowers, stems, seedling roots, and seedling shoots) of six flax cultivars/lines (#3896, Alizee, Atlant, Diplomat, LM98, and Universal) was performed, and from 6 to 16 million reads were obtained for each library. The raw data were deposited in the NCBI Sequence Read Archive (SRA) under the BioProject accession number PRJNA634481.

Reads for each library were trimmed with Trimmomatic (Bolger et al., 2014) and mapped to the NCBI representative *L. usitatissimum* genome (GenBank assembly: GCA_000224295.2) using STAR (Dobin et al., 2013), and plots were generated using MultiQC (Ewels et al., 2016). For the majority of samples, more than 87% of reads were uniquely mapped to the *L. usitatissimum* genome and about 8–10% of reads were mapped to several loci (**Supplementary Data 1**).

For evaluation of gene expression in examined flax tissues and genotypes, reads mapped to the *L. usitatissimum* genome were quantified using BEDTools (Quinlan and Hall, 2010). Expression levels were quantified as read counts per million (CPM) for 200-bp intervals (in case of absence of reads aligned to the particular region, intervals were increased). Obtained results are represented in **Supplementary Data 2**. This table is a valuable resource for differential expression analysis.

For visualization of differences between gene expression profiles of flax tissues and genotypes, multidimensional scaling (MDS) plots were generated using edgeR (Robinson et al., 2010). MDS for 30 flax samples (five tissues of six genotypes) from the current project are represented in **Figure 1**. As can be seen from the figure, samples were grouped according to the type of plant material: flowers and roots formed two distant groups, while leaves, stems, and seedling shoots were close to each other. Next, the present data were combined with the data from four NCBI BioProjects, in which transcriptome sequencing was performed for flax shoots (PRJNA229810), bast fiber and xylem (PRJNA251268), roots (PRJNA412801), and developing seeds (PRJNA539945). We used only forward reads and trimmed them to 70 nucleotides to unify data and reduce the batch effect. Grouping of expression data for five different research projects, including the current one, was again consistent with the type of plant material. Three groups were revealed: group 1—flowers and seeds, group 2—roots, and group 3—leaves, stems, and shoots (**Supplementary Data 3**). This points to the quality of the obtained data and the possibility of a joint analysis of expression data from several research projects that is important for the identification of common regularities in gene expression for particular flax organs and tissues.

Due to the sequencing of a representative set of tissues and genotypes (five tissues of six cultivars/lines), the obtained data are the basis for gene expression analysis in a particular tissue that is important for understanding the key molecular processes occurring in flax plants. Moreover, using these data, the search for genes with the most significant differences in expression between flax genotypes with diverse characteristics can be performed that is necessary for revealing associations between cultivar/line phenotype and gene expression profile. For example, we compared gene expression levels between groups of two linseed and four fiber flax genotypes under study. As annotation is currently absent for the NCBI representative flax genome or other flax genome assemblies (https://www.ncbi.nlm.nih.gov/genome/browse/#!/eukaryotes/6953/), the representative genome (GenBank assembly: GCA_000224295.2) was divided into 1,000-bp intervals, the expression level was quantified as CPM for each interval, and differential expression analysis was performed using edgeR (Robinson et al., 2010). For each tissue, genomic regions were sorted according to the score calculated as $-\log(p\text{-value})^{*}\text{abs}(\log FC)$, where p-value was estimated using quasi-likelihood methods (Lund et al., 2012) and *FC* (fold change) was equal to the ratio of average CPM in the fiber flax group to average CPM in the linseed group. The highest number of differentially expressed transcripts between linseed and fiber flax genotypes was revealed for seedling shoots and leaves, while the lowest was revealed for flowers (**Supplementary Data 4–8**).

Our data are also valuable for preliminary analysis of the expression of particular genes, gene families, or genes involved in the same pathway. This can be performed using **Supplementary Data 1**—knowing the coordinates of particular genes in the flax genome, one can find data on their expression in different genotypes and tissues. An example of such analysis is presented in our previous work on expression of cinnamyl-alcohol dehydrogenase (CAD) encoding genes in roots of flax cultivars/lines susceptible (TOST and AP5) and resistant (#3896 and Dakota) to *F. oxysporum* infection under control and the biotic stress conditions (Novakovskiy et al., 2019). Basing on the data of the present study, we performed a similar analysis of expression of 13 *CAD* genes in five tissues of six examined cultivars/lines. The results are represented in **Supplementary Data 9**. Tissue-specific expression was identified—*CAD2A*, *CAD4B*, *CAD5A*, *CAD5B*, and *CAD6* genes were expressed predominantly in seedling roots, while *CAD3B* was expressed in seedling shoots. Genotype-specific expression profiles were also observed, especially for *CAD2B*, *CAD3A*, *CAD4A*, *CAD7*, and *CAD8* genes. It is worth noting that the present data on expression profiles of *CAD* genes in seedling roots are highly concordant with the results of our aforementioned work (Novakovskiy et al., 2019). In both studies, *CAD6* had the highest expression within *CAD* genes in roots free from *F. oxysporum* infection, and *CAD3A* had the lowest one; expression levels of the other genes were also very similar between studies, indicating the reproducibility of our experiments.

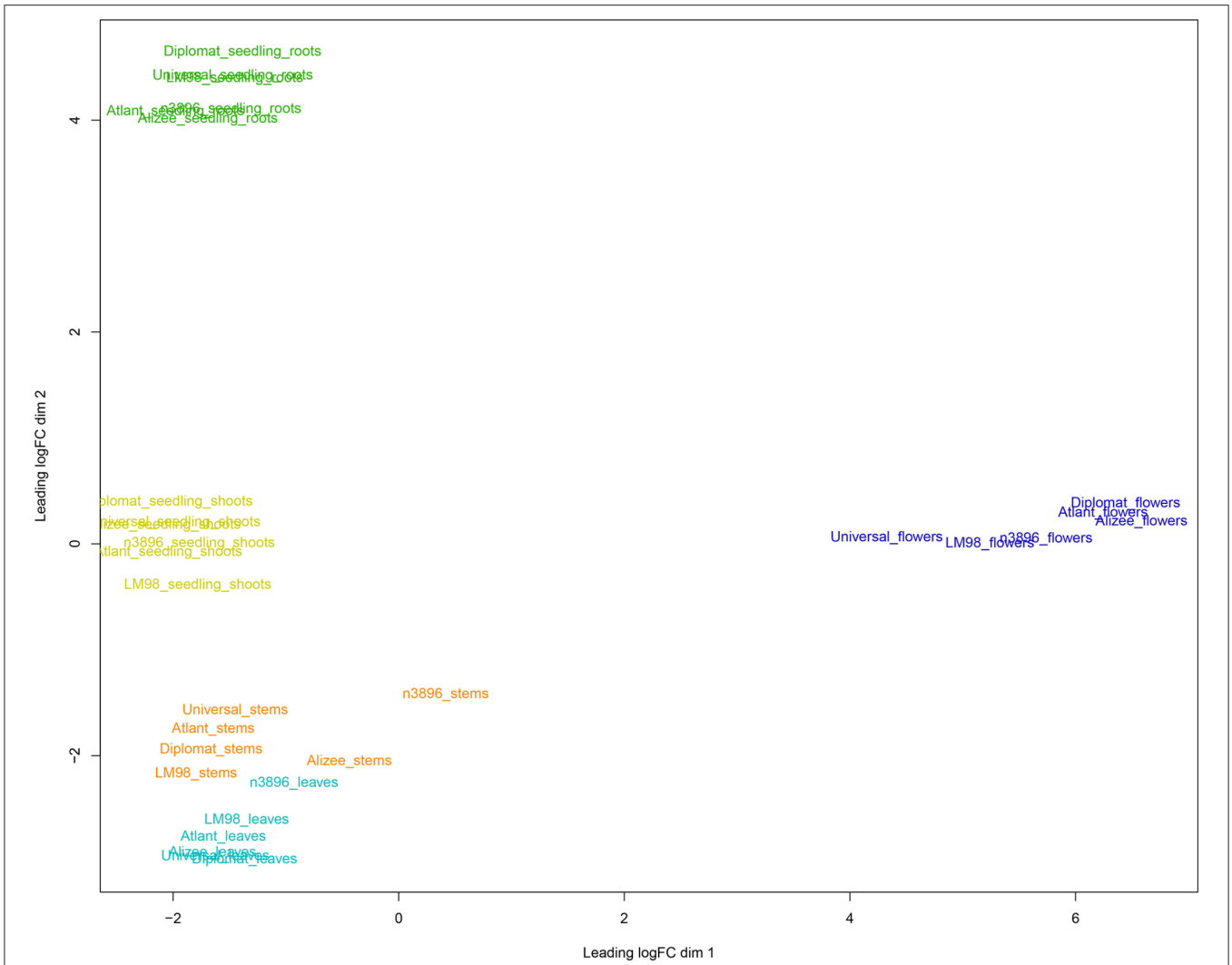


FIGURE 1 | Multidimensional scaling plot for gene expression profiles of five tissues (leaves, flowers, stems, seedling roots, and seedling shoots) of six flax cultivars/lines (#3896, Alizee, Atlant, Diplomat, LM98, and Universal). Different tissues are marked with different colors: leaves—cyan, flowers—blue, stems—orange, seedling roots—green, and seedling shoots—lime.

Our dataset can also be used to search for polymorphisms in expressed regions of the genome within the studied flax genotypes. As an example, variant calling was performed using VarScan (Koboldt et al., 2012) for 13 *CAD* genes, and the largest number (eight) of single-nucleotide polymorphisms (SNPs) was identified for *CAD6*—positions CP027622.1 2160315, CP027622.1 2161357, CP027622.1 2161475, CP027622.1 2161540, CP027622.1 2162146, CP027622.1 2162234, CP027622.1 2162246, and CP027622.1 2162410 according to the NCBI representative *L. usitatissimum* genome GCA_000224295.2. Therefore, this gene may be of interest for the DNA certification of flax cultivars.

The present dataset is especially valuable for revealing trends of interest at the level of gene expression or DNA polymorphisms in expressed genomic regions. However, the validation of the identified trends on extended sample sets is necessary, and for these purposes, other approaches, such as quantitative PCR and targeted sequencing, are more appropriate. Besides, our data are valuable for obtaining complete flax genome annotation, whose absence for the NCBI representative *L. usitatissimum* genome complicates molecular genetic studies of this crop.

CONCLUSIONS

The obtained data on 30 flax transcriptomes are the basis for the evaluation of expression of genes of interest in particular tissues and genotypes, search for genes with differential expression between genotypes with diverse characteristics, identification of polymorphisms in particular genes, and assessment of genetic diversity. Such information is necessary to establish associations between gene expression or DNA polymorphisms and valuable traits. Thus, the present dataset opens up novel opportunities for functional research, development of genome editing, and marker-assisted and genomic breeding. It creates

the necessary basis for the effective application of biotechnology approaches on flax that will allow the breeding of cultivars with desirable characteristics.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://www.ncbi.nlm.nih.gov/sra/~PRJNA634481.

AUTHOR CONTRIBUTIONS

AD, TR, and NM conceived and designed the work. RN, EP, TR, NB, AB, VM, LP, ED, and AS performed the experiments. AD, TR, AZ, AK, GK, and NM analyzed the data. AD, EP, GK, and NM wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was financially supported by the Russian Science Foundation, Grant No. 16-16-00114.

ACKNOWLEDGMENTS

We thank the Center for Precision Genome Editing and Genetic Technologies for Biomedicine, EIMB RAS for providing the computing power and sequencing techniques. This work was performed using the equipment of EIMB RAS Genome center (http://www.eimb.ru/ru1/ckp/ccu_genome_ce.php).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.565146/full#supplementary-material

REFERENCES

Akhmetshina, A. O., Strygina, K. V., Khlestkina, E. K., Porokhovinova, E. A., and Brutch, N. B. (2020). High-throughput sequencing techniques to flax genetics and breeding. *Ecol. Genet*. 18, 103–124. doi: 10.17816/ecogen16126

Baley, C., Gomina, M., Breard, J., Bourmaud, A., and Davies, P. (2019). Variability of mechanical properties of flax fibres for composite reinforcement. A review. *Ind. Crops Prod*. 145:111984. doi: 10.1016/j.indcrop.2019.111984

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

Cobb, J. N., Biswas, P. S., and Platten, J. D. (2019). Back to the future: revisiting MAS as a tool for modern plant breeding. *Theor. Appl. Genet*. 132, 647–667. doi: 10.1007/s00122-018-3266-4

Costa, S. M., Ferreira, D. P., Ferreira, A., Vaz, F., and Fangueiro, R. (2018). Multifunctional flax fibres based on the combined effect of silver and zinc oxide (Ag/ZnO) nanostructures. *Nanomaterials* 8:1069. doi: 10.3390/nano8121069

Cullis, C. A. (2019). *Genetics and Genomics of Linum.* Cham: Springer.

Dash, P. K., Rai, R., Mahato, A. K., Gaikwad, K., and Singh, N. K. (2017). Transcriptome landscape at different developmental stages of a drought tolerant cultivar of flax (*Linum usitatissimum*). *Front Chem*. 5:82. doi: 10.3389/fchem.2017.00082

Dmitriev, A. A., Krasnov, G. S., Rozhmina, T. A., Kishlyan, N. V., Zyablitsin, A. V., Sadritdinova, A. F., et al. (2016a). Glutathione S-transferases and UDP-glycosyltransferases Are Involved in Response to Aluminum Stress in Flax. *Front. Plant Sci* 7, 1920. doi: 10.3389/fpls.2016.01920

Dmitriev, A. A., Krasnov, G. S., Rozhmina, T. A., Novakovskiy, R. O., Snezhkina, A. V., Fedorova, M. S., et al. (2017). Differential gene expression in response to *Fusarium oxysporum* infection in resistant and susceptible genotypes of flax (*Linum usitatissimum* L.). *BMC Plant Biol.* 17 (Suppl. 2):253. doi: 10.1186/s12870-017-1192-2

Dmitriev, A. A., Krasnov, G. S., Rozhmina, T. A., Zyablitsin, A. V., Snezhkina, A. V., Fedorova, M. S., et al. (2019). Flax (*Linum usitatissimum* L.) response to non-optimal soil acidity and zinc deficiency. *BMC Plant Biol.* 19 (Suppl. 1):54. doi: 10.1186/s12870-019-1641-1

Dmitriev, A. A., Kudryavtseva, A. V., Krasnov, G. S., Koroban, N. V., Speranskaya, A. S., Krinitsina, A. A., et al. (2016b). Gene expression profiling of flax (*Linum usitatissimum* L.) under edaphic stress. *BMC Plant Biol.* 16 (Suppl. 3):237. doi: 10.1186/s12870-016-0927-9

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

Dwivedi, S. L., Siddique, K. H. M., Farooq, M., Thornton, P. K., and Ortiz, R. (2018). Using Biotechnology-led approaches to uplift cereal and food legume yields in dryland environments. *Front. Plant Sci.* 9:1249. doi: 10.3389/fpls.2018.01249

Ewels, P., Magnusson, M., Lundin, S., and Kaller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32, 3047–3048. doi: 10.1093/bioinformatics/btw354

Galindo-Gonzalez, L., and Deyholos, M. K. (2016). RNA-seq transcriptome response of flax (*Linum usitatissimum* L.) to the pathogenic fungus *Fusarium oxysporum* f. *sp. lini. Front Plant Sci*. 7:1766. doi: 10.3389/fpls.2016.01766

Gionfriddo, M., De Gara, L., and Loreto, F. (2019). Directed evolution of plant processes: towards a green (r)evolution? *Trends Plant Sci.* 24, 999–1007. doi: 10.1016/j.tplants.2019.08.004

Gorshkov, O., Mokshina, N., Gorshkov, V., Chemikosova, S., Gogolev, Y., and Gorshkova, T. (2017). Transcriptome portrait of cellulose-enriched flax fibres at advanced stage of specialization. *Plant Mol. Biol.* 93, 431–449. doi: 10.1007/s11103-016-0571-7

Gorshkova, T., Chernova, T., Mokshina, N., Gorshkov, V., Kozlova, L., and Gorshkov, O. (2018). Transcriptome analysis of intrusively growing flax fibers isolated by laser microdissection. *Sci. Rep*. 8:14570. doi: 10.1038/s41598-018-32869-2

Goyal, A., Sharma, V., Upadhyay, N., Gill, S., and Sihag, M. (2014). Flax and flaxseed oil: an ancient medicine & modern functional food. *J. Food Sci. Technol*. 51, 1633–1653. doi: 10.1007/s13197-013-1247-9

Guo, Y., Qiu, C., Long, S., Chen, P., Hao, D., Preisner, M., et al. (2017). Digital gene expression profiling of flax (*Linum usitatissimum* L.) stem peel identifies genes enriched in fiber-bearing phloem tissue. *Gene* 626, 32–40. doi: 10.1016/j.gene.2017.05.002

Imran, M., Ahmad, N., Anjum, F. M., Khan, M. K., Mushtaq, Z., Nadeem, M., et al. (2015). Potential protective properties of flax lignan secoisolariciresinol diglucoside. *Nutr. J*. 14:71. doi: 10.1186/s12937-015-0059-3

Kezimana, P., Dmitriev, A. A., Kudryavtseva, A. V., Romanova, E. V., and Melnikova, N. V. (2018). Secoisolariciresinol diglucoside of flaxseed and its metabolites: biosynthesis and potential for nutraceuticals. *Front. Genet*. 9:641. doi: 10.3389/fgene.2018.00641

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., et al. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 22, 568–576. doi: 10.1101/gr.129684.111

Kolotov, A. P. (2020). *Oil Flax in the Middle Urals.* Ekaterinburg: UrFASRC.

Kouba, M., and Mourot, J. (2011). A review of nutritional effects on fat composition of animal products with special emphasis on n-3 polyunsaturated fatty acids. *Biochimie* 93, 13–17. doi: 10.1016/j.biochi.2010.02.027

Long, S. H., Deng, X., Wang, Y. F., Li, X., Qiao, R. Q., Qiu, C. S., et al. (2012). Analysis of 2,297 expressed sequence tags (ESTs) from a cDNA library of flax (*Linum ustitatissimum* L.) bark tissue. *Mol. Biol. Rep*. 39, 6289–6296. doi: 10.1007/s11033-012-1450-1

Lund, S. P., Nettleton, D., McCarthy, D. J., and Smyth, G. K. (2012). Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Stat. Appl. Genet. Mol. Biol*. 11, 1–42. doi: 10.1515/1544-6115.1826

Mali, A. V., Padhye, S. B., Anant, S., Hegde, M. V., and Kadam, S. S. (2019). Anticancer and antimetastatic potential of enterolactone: clinical, preclinical and mechanistic perspectives. *Eur. J. Pharmacol*. 852, 107–124. doi: 10.1016/j.ejphar.2019.02.022

Mascher, M., Schreiber, M., Scholz, U., Graner, A., Reif, J. C., and Stein, N. (2019). Genebank genomics bridges the gap between the conservation of crop diversity and plant breeding. *Nat. Genet*. 51, 1076–1081. doi: 10.1038/s41588-019-0443-6

Mokshina, N., Gorshkov, O., Ibragimova, N., Chernova, T., and Gorshkova, T. (2017). Cellulosic fibres of flax recruit both primary and secondary cell wall cellulose synthases during deposition of thick tertiary cell walls and in the course of graviresponse. *Funct. Plant Biol*. 44, 820–831. doi: 10.1071/FP17105

Muir, A. D., and Westcott, N. D. (2003). *Flax: The Genus Linum*. London: CRC Press.

Novakovskiy, R. O., Povkhova, L. V., Krasnov, G. S., Rozhmina, T. A., Zhuchenko, A. A., Kudryavtseva, L. P., et al. (2019). The cinnamyl alcohol dehydrogenase gene family is involved in the response to *Fusarium oxysporum* in resistant and susceptible flax genotypes. *Vavilov J. Genet. Breed*. 23, 896–901. doi: 10.18699/VJ19.564

Parikh, M., Netticadan, T., and Pierce, G. N. (2018). Flaxseed: its bioactive components and their cardiovascular benefits. *Am. J. Physiol. Heart Circ. Physiol*. 314, H146–H159. doi: 10.1152/ajpheart.00400.2017

Pavlova, L. N., Rozhmina, T. A., Gerasimova, E. G., Rumyanceva, V. N., Kudryavceva, L. P., and Kiseleva, T. S. (2018). "Hozyajstvennaya cennost' novyh sortov l'na-dolgunca," in *Nauchnoe Obespechenie Proizvodstva Pryadil'nyh kul'tur: Sostoyanie, Problemy i Perspektivy*. T. A. Rozhmina, V. P. Ponazhev, L. N. Pavlova, and N. V. Proletova (Tver: Tver State University), 18–20.

Preisner, M., Wojtasik, W., Kostyn, K., Boba, A., Czuj, T., Szopa, J., et al. (2018). The cinnamyl alcohol dehydrogenase family in flax: differentiation during plant growth and under stress conditions. *J. Plant Physiol*. 221, 132–143. doi: 10.1016/j.jplph.2017.11.015

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Rozhmina, T. A., and Loshakova, N. I. (2016). New sources of effective resistance genes to fusarium wilt in flax (*Linum usitatissimum* L.) depending on temperature. *Agric. Biol*. 51, 310–317. doi: 10.15389/agrobiology.2016.3.310eng

Rozhmina, T. A., Zhuchenko, A. A., Melnikova, N. V., and Smirnova, A. D. (2020). Resistance of flax gene pool samples to edaphic stress caused by low acidity. *Agri. Sci. Euro North East* 21, 133–140. doi: 10.30766/2072-9081.2020.21.2.133-140

Ryzhov, A. I., Rozhmina, T. A., and Golubeva, L. M. (2012). The role of the flax gene pool in obtaining competitive fiber products. *Technologies of the 21st century in light industry*. 6, 1–11.

Singh, K. K., Mridula, D., Rehal, J., and Barnwal, P. (2011). Flaxseed: a potential source of food, feed and fiber. *Crit. Rev. Food Sci. Nutr*. 51, 210–222. doi: 10.1080/10408390903537241

Varshney, R. K., Pandey, M. K., Bohra, A., Singh, V. K., Thudi, M., and Saxena, R. K. (2019). Toward the sequence-based breeding in legumes in the post-genome sequencing era. *Theor. Appl. Genet*. 132, 797–816. doi: 10.1007/s00122-018-3252-x

Wang, Z., Hobson, N., Galindo, L., Zhu, S., Shi, D., McDill, J., et al. (2012). The genome of flax (*Linum usitatissimum*) assembled *de novo* from short shotgun sequence reads. *Plant J*. 72, 461–473. doi: 10.1111/j.1365-313X.2012.05093.x

Wu, J., Zhao, Q., Wu, G., Yuan, H., Ma, Y., Lin, H., et al. (2019). Comprehensive Analysis of Differentially expressed unigenes under NaCl stress in flax (*Linum usitatissimum* L.) using RNA-Seq. *Int. J. Mol. Sci*. 20:369. doi: 10.3390/ijms20020369

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

You, F. M., Xiao, J., Li, P., Yao, Z., Jia, G., He, L., et al. (2018). Chromosome-scale pseudomolecules refined by optical, physical and genetic maps in flax. *Plant J*. 95, 371–384. doi: 10.1111/tpj.13944

Yu, Y., Huang, W., Chen, H., Wu, G., Yuan, H., Song, X., et al. (2014). Identification of differentially expressed genes in flax (*Linum usitatissimum* L.) under saline-alkaline stress by digital gene expression. *Gene* 549, 113–122. doi: 10.1016/j.gene.2014.07.053

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.

Copyright © 2020 Dmitriev, Novakovskiy, Pushkova, Rozhmina, Zhuchenko, Bolsheva, Beniaminov, Mitkevich, Povkhova, Dvorianinova, Snezhkina, Kudryavtseva, Krasnov and Melnikova. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 03 December 2020
doi: 10.3389/fgene.2020.593577

Genome-Wide Identification and Expression Analysis of the Class III Peroxidase Gene Family in Potato (*Solanum tuberosum* L.)

Xuanshong Yang[1,2], Jiazheng Yuan[3], Wenbin Luo[4], Mingyue Qin[1,2], Jiahan Yang[1,2], Weiren Wu[2] and Xiaofang Xie[1,2]**

[1] College of Life Sciences, Fujian Agriculture & Forestry University, Fuzhou, China, [2] Fujian Key Laboratory of Crop Breeding by Design, Fujian Agriculture & Forestry University, Fuzhou, China, [3] Department of Biological and Forensic Sciences, Fayetteville State University, Fayetteville, NC, United States, [4] The Crop Institute, Fujian Academy of Agricultural Sciences, Fuzhou, China

Class III peroxidases (PRXs) are plant-specific enzymes and play important roles in plant growth, development and stress response. In this study, a total of 102 non-redundant *PRX* gene members (*StPRXs*) were identified in potato (*Solanum tuberosum* L.). They were divided into 9 subfamilies based on phylogenetic analysis. The members of each subfamily were found to contain similar organizations of the exon/intron structures and protein motifs. The *StPRX* genes were not equally distributed among chromosomes. There were 57 gene pairs of segmental duplication and 26 gene pairs of tandem duplication. Expression pattern analysis based on the RNA-seq data of potato from public databases indicated that *StPRX* genes were expressed differently in various tissues and responded specifically to heat, salt and drought stresses. Most of the *StPRX* genes were expressed at significantly higher levels in root than in other tissues. In addition, real-time quantitative PCR (qRT-PCR) analysis for 7 selected *StPRX* genes indicated that these genes displayed various expression levels under abiotic stresses. Our results provide valuable information for better understanding the evolution of *StPRX* gene family in potato and lay the vital foundation for further exploration of *PRX* gene function in plants.

Keywords: potato, class III peroxidase, phylogenetic analysis, expression pattern, abiotic stress

OPEN ACCESS

Edited by:
Weike Duan,
Huaiyin Institute of Technology, China

Reviewed by:
Robert A. Haney,
Ball State University, United States
Fernando Cardona,
Consejo Superior de Investigaciones Científicas (CSIC), Spain

****Correspondence:***
Xiaofang Xie
xxf317@fafu.edu.cn
Weiren Wu
wuwr@fafu.edu.cn

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *11 August 2020*
Accepted: *09 November 2020*
Published: *03 December 2020*

Citation:
Yang X, Yuan J, Luo W, Qin M, Yang J, Wu W and Xie X (2020) Genome-Wide Identification and Expression Analysis of the Class III Peroxidase Gene Family in Potato (Solanum tuberosum L.). Front. Genet. 11:593577. doi: 10.3389/fgene.2020.593577

INTRODUCTION

As a large family of isozymes, peroxidases (POD) play important roles in the growth, development and defense processes in plants (Hiraga et al., 2001). Peroxidases are divided into two major groups, hemoglobin peroxidases and non-hemoglobin peroxidases, according to their protein structures (Hiraga et al., 2000). Exception of animal peroxidases, the hemoglobin peroxidases have been further divided into three classes based on their sequences and catalytic properties, namely, I, II and III peroxidases. Class I peroxidases are widely distributed in most living organisms other than animals, and play an important role in removing excess H_2O_2 to prevent cell damage (Erman and Vitello, 2002; Shigeoka et al., 2002). Class II peroxidases only present in fungi and are mainly involved in lignin degradation (Piontek et al., 2001). Class III peroxidases (PRX, EC 1.11.1.7)

exist in various plants as a multi-genic family (Tognolli et al., 2002; Duroux and Welinder, 2003; Passardi et al., 2004a). The PRX protein contains highly conserved amino acid residues, including a single peptide chain and the protoporphyrin IX domain (Welinder, 1993). Most plant PRXs fuse with carbohydrate side chains to form glycosylated proteins. This glycosylation prevents the protein being degraded by protease and maintains the enzyme stability (Zheng and Van Huystee, 1991). In addition, two histidine residues interact with a heme group and eight cysteine residues, forming disulfide bridges; the distal histidine is essential for catalytic activity (Passardi et al., 2004a). The functions of PRXs have been illustrated in several studies as important proteins for a wide range of physiological processes of plants, such as growth hormone metabolism (Gazaryan et al., 1996), formation of lignin and liposites, crosslinking of cell walls (Barcelo and Pomar, 2001; Passardi et al., 2004b), cell growth and elongation, and various defense processes against biotic and abiotic stresses (Schopfer et al., 2002; Liszkay et al., 2004; Bindschedler et al., 2006). For example, *Arabidopsis* peroxidases *AtPrx33* and *AtPrx34* are associated with root elongation (Passardi et al., 2006), and *AtPrx72* has an important role in lignification (Herrero et al., 2013). Cotton *GhPOX1* is involved in cotton fiber elongation by means of maintaining high levels of reactive oxygen species production (Mei et al., 2009). The expression of several *ZmPRX* is altered significantly in response to abiotic stress based on microarray analysis in maize, suggesting that these genes play important role in resistance to abiotic stress (Wang et al., 2015). It was also reported that either present or lack of specific peroxidase isoforms appears to be correlated to a particular cellular process or participating a particular protein localization (Loukili et al., 1999; Allison and Schultz, 2004). Genome-wide analysis of this large multigenic family will be greatly helpful to understand its physiological roles and characteristics. In recent years, the PRX family has been widely studied in many species, such as *Arabidopsis thaliana* (Tognolli et al., 2002), rice (Passardi et al., 2004a), and maize (Wang et al., 2015).

Potato (*Solanum tuberosum*) is one of the most important food crops in the world. Potato tubers are rich in nutrients and are valuable processing raw materials for food industry (Pang, 2019). However, potato contains rich phenolic substances, which are the substrates of peroxidase reaction and lead to an enzymatic browning reaction frequently caused by POD (peroxidase) in the storage and potato processing (Zhu and Hu, 2013), affecting the quality of products (Zhou et al., 2010; Zhu, 2017). In addition, potato plants are often subjected to various types of abiotic and biotic stresses during growth and development. Although *PRX* gene family members play important roles in the plant growth and development, their functions in potato are poorly deciphered. Thoroughly analyzing the *PRX* gene family in potato is a primary step to understand its physiological roles and characteristics.

In this study, a systematic investigation of the *PRX* gene family in potato, including *PRX* gene structure, chromosomal localization, gene duplication and phylogenetic relationship, was performed using the sequences from the genome database. Moreover, the tissue-specific expression profile and expression patterns under abiotic stresses such as drought, heat and salt treatments were also investigated. The objectives of this study were to identify and assess the sequence structures of the potato *PRX* gene family and analyze the evolutionary relationship of *PRX* gene family in plants. The results of this study provide a solid foundation for the next phase functional investigation of *PRX* genes in potato.

MATERIALS AND METHODS

Screening and Domain Identification of Potato PRX Proteins

The protein sequences of the *Arabidopsis thaliana* PRX members (Tognolli et al., 2002) were downloaded from the *Arabidopsis* database[1]. These protein sequences of *Arabidopsis* were used as the queries to identify the PRX orthologs in potato using the BLASTP tool SpudDB[2] and Phytozome v12.1[3]. Proteins with more than 30% similarity to the query sequence and an $E < E^{-10}$ were selected. The domains for PRX proteins were further confirmed using the Conserved Domain Database of NCBI[4]. The sequences possessing PRX conserved domain were selected as the final candidates of *PRX* genes (*StPRXs*) and were renamed according to their physical position in the potato genome. The information of these genes was obtained from Phytozome[5], including gene IDs, physical position, gene sequence and protein sequence. The parameters for the predicted StPRX proteins were calculated using online ExPASy tools (Gasteiger et al., 2003)[6].

Gene Structure and Conserved Motif Analysis of Potato PRX Protein

The *StPRX* gene structures were identified using the Gene Structure Display Server (GSDS[7]; Guo et al., 2007). The conserved motifs were identified using the MEME software (version 5.0.3[8]; Bailey et al., 2009). Parameters were set as 20 motifs with the optimum motif width of 50–300 residues. The conserved motifs were then further annotated with the CDD program[9] (Marchler et al., 2017).

Phylogenetic Analysis of PRX Proteins

The protein sequences of StPRXs were aligned using the multiple sequence alignment tool ClustalX (Thompson et al., 1997). The phylogenetic tree of PRX family proteins was generated using the MEGA-X maximum-likelihood model (Kumar et al., 2016) with 1000 bootstrap replicates. Orthologs identification method is based on a report (Blanc, 2004). To identify putative orthologs between two different species, each sequence from potato was

[1]http://www.arabidopsis.org/
[2]http://solanaceae.plantbiology.msu.edu/
[3]https://phytozome.jgi.doe.gov
[4]https://www.ncbi.nlm.nih.gov/cdd/
[5]https://phytozome.jgi.doe.gov/pz/portal.html
[6]http://web.expasy.org/protparam/
[7]http://gsds.cbi.pku.edu.cn/
[8]http://alternate.meme-suite.org/tools/meme
[9]https://www.ncbi.nlm.nih.gov/

searched against all sequences from maize and *Arabidopsis* using the BLASTN tool (Altschul et al., 1997). For each query sequence, the best hit among those that met the criteria of alignment ≥ 300 bp and similarity ≥ 40% was considered as the ortholog.

Chromosomal Location and Gene Duplication

Information of the chromosomal location image of *StPRX* genes was retrieved by the MapInspect tool[10]. To assess gene duplication, the parameters for the proportion of overlap and the similarity between the two sequences were set to be > 70% (Gu et al., 2002; Yang et al., 2008). Two nearby duplicated genes were defined as tandem duplicated genes when the physical space between them was less than 100 kb and contained less than three intervening genes (Wang et al., 2010), while any other two duplicated genes that did not meet the condition of tandem duplicated genes, including those located on the same chromosome or different chromosomes, were all defined as segmental duplicated genes.

Expression Analysis

The FPKM (fragments per kilobase per million) values of *StRPX* in various tissues and treatments (salt, drought and heat) and their control (CK) generated by RNA-seq (DM_v4.03) were extracted from the Potato Genome Database (see footnote). The expression profile of *StRPX* genes was generated using the R package[11] of the heatmap function (Warnes et al., 2016).

Plant Treatments and Quantitative Real-Time PCR Analysis

T virus-free plantlets (*S. tuberosum* L. autotetraploid cultivar Zhongshu 3) were generated by *in vitro* nodal cutting method. Potato shots placed on full MS solid medium were cultured in a growth chamber under the condition of 22°C and 16 h light/8 h dark photoperiod for 1 month. The plantlets were transplanted into a tray with a half-strength modified Hoagland solution (Hammer et al., 1978) for 6 days, and then were exposed to the abiotic stress conditions, including heat (35°C), drought (260 mM mannitol) and salt (150 mM NaCl) treatments. Untreated plantlets were used as control (CK). The treated and control plantlets were collected 6 h after treatment and then stored at −80°C before RNA extraction.

The total RNA of the plantlets was extracted using TRIzol reagent (Invitrogen)[12] according to the manufacturer's instructions. The cDNA samples were then assessed by qRT-PCR using SYBR Premix Ex Taq (Takara). *Actin* was used as an internal control gene. Three biological replicates (each containing 6 plants) and three technical replicates were measured for each treatment. The relative expression level of a gene was calculated according to the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001). The primers used for qRT-PCR analysis are listed in **Supplementary Table S1**.

[10]http://www.plantbreeding.wur.nl/uk/software_mapinspect.html

[11]http://www.r-project.org/

[12]http://www.invitrogen.com

RESULTS

Identification and Characterization of *PRX* Genes

Using 73 *Arabidopsis* PRX sequences as queries (Tognolli et al., 2002) and validating the candidate sequences by conservative domain analysis based on CDD, a total of 102 PRX genes (*StPRXs*) were identified from potato genome and were renamed from *StPRX1* to *StPRX102* based on their physical position on chromosomes (**Table 1**). However, the location of *StPRX1* on the chromosome could not yet be defined because it was located in the unmapped scaffold. The protein lengths of the 102 *StPRX* genes varied from 152 (*StPRX29*) to 592 (*StPRX60*) amino acids, with an average of 310.8 amino acids. The molecular weights ranged from 16866.8 Da (*StPRX29*) to 63735.06 Da (*StPRX60*). The theoretical isoelectric points (pI) of these *StPRX* genes varied from 4.43 (*StPRX20*) to 10.02 (*StPRX81*). The detailed information for *StPRX* genes was listed in **Table 1**.

Phylogenetic Analysis

To reveal the evolutionary relationship of the *PRX* gene family, an unrooted phylogenetic tree (**Figure 1A**) was obtained basing the MEGA-X maximum-likelihood model. The 102 *StPRXs* were classified into 9 subfamily (I-IX) with the bootstrap values (≥50%) on the phylogenetic tree. However, 2 *StPRX* genes (*StPRX 8* and *StPRX22*) could not be assigned to any of the 9 subfamilies due to the low bootstrap values (<50%). Among these 9 groups, subfamily I possessed the largest clade, which contained 35 *StPRX* genes, followed by group VIII, which had 22 PRX members. These two subfamilies accounted for 55.88% of the total *StPRXs*. In contrast, groups III and IV only had two or three *StPRX* genes.

To further explore the evolutionary process of the PRX family in potato, the 102 PRX protein sequences of potato were aligned with 73 PRX proteins of *Arabidopsis thaliana* and 119 PRX proteins of maize (Tognolli et al., 2002; Wang et al., 2015). The phylogenetic tree was divided into 12 different groups (groups A–L; **Figure 2**). Among these groups, group G was the largest, which contained 92 PRX members, including 35 of potato, 30 of maize and 27 of *Arabidopsis*. Groups A, D and K also had a large number of genes, containing 40, 35, and 40 members, respectively. In contrast, group F was the smallest, which contained only 5 members, including 1 of potato, 1 of maize and 3 of *Arabidopsis*. Interestingly, although group D was large, it had no members from *Arabidopsis*. Moreover, group I and L only contained 5 and 8 members from maize, respectively. A few groups were supported by low bootstrap values, which might be due to the relative less informative character positions beside the conserved PRX domains. This scenario has been also found in the analysis of other gene families (Li et al., 2006; Wang et al., 2018). In addition, a total of 82 ortholog pairs were identified between potato and *Arabidopsis* (St-At; **Supplementary Table S4**). However, only four ortholog pairs were retrieved between potato and maize (St-Zm; **Supplementary Table S5**). The orthologs between potato and *Arabidopsis* were much greater than that between potato

TABLE 1 | The characters of 102 *PRX* gene family members in *Solanum tuberosum*.

Gene name	Gene ID	Location (bp)	Chr.	PL (aa)	MW (Da)	PI
StPRX1	PGSC0003DMG400011948	22096415..22097488	0	299	32921.88	5.23
StPRX2	PGSC0003DMG400032147	1754474..1757591	1	305	33739.61	8.10
StPRX3	PGSC0003DMG400032199	1761660..1764244	1	324	34984.22	4.60
StPRX4	PGSC0003DMG400016371	3596648..3598780	1	440	48820.52	6.32
StPRX5	PGSC0003DMG400014725	6806225..6809227	1	309	34613.11	8.58
StPRX6	PGSC0003DMG400014726	6812744..6815906	1	252	27532.73	5.63
StPRX7	PGSC0003DMG400014728	6864443..6867718	1	218	24433.11	5.65
StPRX8	PGSC0003DMG400024019	45907335..45908930	1	235	25675.13	5.86
StPRX9	PGSC0003DMG400011458	47297311..47300055	1	297	32037.04	4.83
StPRX10	PGSC0003DMG400022848	59916267..59918783	1	325	35564.60	8.56
StPRX11	PGSC0003DMG400022850	59968801..59970227	1	325	35441.54	7.95
StPRX12	PGSC0003DMG401018293	78690854..78691752	1	258	28158.11	7.65
StPRX13	PGSC0003DMG400012589	81814094..81815527	1	331	35958.90	8.38
StPRX14	PGSC0003DMG400025803	84743624..84745033	1	311	34505.20	5.33
StPRX15	PGSC0003DMG402015497	20759728..20761348	2	331	36729.25	9.34
StPRX16	PGSC0003DMG400025478	23460843..23462610	2	326	35405.94	5.99
StPRX17	PGSC0003DMG400022955	32712360..32714253	2	253	27718.73	8.74
StPRX18	PGSC0003DMG400010715	32796108..32797987	2	253	27718.73	8.74
StPRX19	PGSC0003DMG400022342	35169225..35172205	2	363	38682.57	4.69
StPRX20	PGSC0003DMG400022341	35183633..35186491	2	358	38031.64	4.43
StPRX21	PGSC0003DMG400016453	36329136..36330486	2	326	37243.99	8.90
StPRX22	PGSC0003DMG400013654	37854602..37856626	2	324	36500.79	5.77
StPRX23	PGSC0003DMG400003654	39070302..39074823	2	319	34770.84	9.35
StPRX24	PGSC0003DMG400012668	41145199..41150765	2	316	35141.91	4.95
StPRX25	PGSC0003DMG400024967	43038498..43040517	2	332	36390.55	9.22
StPRX26	PGSC0003DMG400010061	44288870..44290076	2	260	28291.97	6.88
StPRX27	PGSC0003DMG400010064	44310085..44311431	2	325	35728.53	6.42
StPRX28	PGSC0003DMG400000511	46773640..46774626	2	328	36508.42	8.57
StPRX29	PGSC0003DMG400020252	48001688..48003039	2	152	16866.80	5.05
StPRX30	PGSC0003DMG400005062	2282144..2284899	3	334	36318.98	6.32
StPRX31	PGSC0003DMG400022567	3855707..3861892	3	445	49360.67	8.37
StPRX32	PGSC0003DMG400022541	3883852..3886414	3	333	36913.24	8.47
StPRX33	PGSC0003DMG400014867	7132984..7135756	3	331	35845.02	8.53
StPRX34	PGSC0003DMG400024253	41651231..41652456	3	317	34448.21	6.98
StPRX35	PGSC0003DMG400015801	43497865..43500054	3	319	34444.9	8.87
StPRX36	PGSC0003DMG400000559	46475042..46476231	3	290	31249.31	5.83
StPRX37	PGSC0003DMG401002540	60230456..60232244	3	355	38956.21	5.28
StPRX38	PGSC0003DMG400024813	58199228..58201394	4	273	29849.00	5.59
StPRX39	PGSC0003DMG400025084	61089768..61091391	4	264	28737.56	7.61
StPRX40	PGSC0003DMG401025083	61081697..61083354	4	348	38241.36	5.88
StPRX41	PGSC0003DMG402025083	61099781..61102899	4	359	39717.90	7.58
StPRX42	PGSC0003DMG400006386	65556167..65558314	4	323	35728.04	8.77
StPRX43	PGSC0003DMG400005152	69721596..69723452	4	310	34424.07	5.71
StPRX44	PGSC0003DMG400003748	70184517..70186504	4	331	36595.03	8.29
StPRX45	PGSC0003DMG400009950	71149878..71152497	4	271	29647.46	8.64
StPRX46	PGSC0003DMG400015584	7054308..7055315	5	335	37292.16	8.03
StPRX47	PGSC0003DMG400018624	10018838..10020432	5	485	52645.89	5.18
StPRX48	PGSC0003DMG400020975	23721320..23722830	5	241	26599.03	8.44
StPRX49	PGSC0003DMG400005272	42500860..42503146	5	322	35270.83	8.77
StPRX50	PGSC0003DMG400005279	42524175..42525868	5	327	35795.19	8.26
StPRX51	PGSC0003DMG400005273	42534489..42535934	5	327	35736.75	7.54

(Continued)

TABLE 1 | Continued

Gene name	Gene ID	Location (bp)	Chr.	PL (aa)	MW (Da)	PI
StPRX52	PGSC0003DMG400015035	42586204..42587629	5	327	35795.79	7.57
StPRX53	PGSC0003DMG400006993	45422092..45429842	5	273	29118.81	5.92
StPRX54	PGSC0003DMG400023491	50994131..50995456	5	334	36697.30	9.09
StPRX55	PGSC0003DMG400014055	31519344..31521327	6	319	34526.11	9.41
StPRX56	PGSC0003DMG400030764	40825442..40828587	6	312	34193.31	8.62
StPRX57	PGSC0003DMG400030430	56939983..56941180	6	316	35275.13	5.73
StPRX58	PGSC0003DMG400030382	57556039..57558160	6	312	33614.90	6.06
StPRX59	PGSC0003DMG400004090	10927329..10929258	7	329	36314.51	9.33
StPRX60	PGSC0003DMG400000694	44979030..44982593	7	592	63735.06	8.90
StPRX61	PGSC0003DMG400000693	44995516..44996807	7	331	35726.00	9.03
StPRX62	PGSC0003DMG400025492	47159208..47160591	7	328	36021.11	7.51
StPRX63	PGSC0003DMG400025491	47171967..47173483	7	329	35706.95	8.97
StPRX64	PGSC0003DMG400025490	47178254..47179727	7	329	35804.15	8.96
StPRX65	PGSC0003DMG400020437	50315313..50316847	7	224	23723.53	8.30
StPRX66	PGSC0003DMG400020494	1070500..1071459	8	319	35247.52	6.35
StPRX67	PGSC0003DMG400005872	5883543..5889021	8	345	38898.58	8.71
StPRX68	PGSC0003DMG400029546	44803067..44804385	8	336	37119.67	9.41
StPRX69	PGSC0003DMG400001774	6294037..6297308	9	324	36832.21	7.12
StPRX70	PGSC0003DMG400026575	13125916..13128279	9	316	33238.39	8.63
StPRX71	PGSC0003DMG400037550	13143159..13144652	9	204	22597.61	8.70
StPRX72	PGSC0003DMG400024285	50464266..50465966	9	322	35077.57	8.88
StPRX73	PGSC0003DMG400035475	28321075..28323202	10	322	35248.38	6.39
StPRX74	PGSC0003DMG400006679	49022687..49023709	10	254	28268.72	5.24
StPRX75	PGSC0003DMG400020800	49084193..49085040	10	250	27069.43	5.68
StPRX76	PGSC0003DMG400020799	49151590..49152437	10	250	27197.52	5.68
StPRX77	PGSC0003DMG400020798	49172630..49173264	10	179	18985.45	4.83
StPRX78	PGSC0003DMG400020801	49258779..49259876	10	300	32593.1	9.03
StPRX79	PGSC0003DMG400010465	54852020..54854941	10	329	37248.41	6.63
StPRX80	PGSC0003DMG401010480	54880317..54881406	10	336	36523.78	9.27
StPRX81	PGSC0003DMG400010479	54899402..54900633	10	383	42139.81	10.02
StPRX82	PGSC0003DMG400034594	58853562..58854760	10	203	21774.81	6.07
StPRX83	PGSC0003DMG400016223	5090584..5091685	11	337	37722.07	5.46
StPRX84	PGSC0003DMG400019492	5101081..5102175	11	335	36560.76	8.51
StPRX85	PGSC0003DMG400019491	5104999..5106087	11	335	36652.79	8.32
StPRX86	PGSC0003DMG400019490	5107998..5108778	11	332	36196.38	8.51
StPRX87	PGSC0003DMG400019474	5116202..5117294	11	232	25085.99	8.56
StPRX88	PGSC0003DMG400019473	11796060..11796799	11	305	33236.00	8.67
StPRX89	PGSC0003DMG400018031	11937469..11938723	11	178	19158.80	8.26
StPRX90	PGSC0003DMG400015106	12404341..12405565	11	319	34436.96	6.52
StPRX91	PGSC0003DMG400027614	20684710..20686486	11	300	32441.69	6.94
StPRX92	PGSC0003DMG400019766	2724753..2727047	11	252	26855.16	4.88
StPRX93	PGSC0003DMG400015548	45091925..45093213	11	326	34682.53	4.84
StPRX94	PGSC0003DMG400039106	2392106..2393703	12	328	36283.93	9.22
StPRX95	PGSC0003DMG400024329	6654188..6656445	12	324	35318.86	9.00
StPRX96	PGSC0003DMG402024332	6656986..6658371	12	335	37699.31	8.05
StPRX97	PGSC0003DMG401029332	6660893..6662191	12	322	35245.52	8.65
StPRX98	PGSC0003DMG400024330	6664412..6668643	12	322	35330.71	8.77
StPRX99	PGSC0003DMG400021801	7978370..7980568	12	328	36224.44	8.66
StPRX100	PGSC0003DMG400020388	11649782..11651610	12	353	38921.57	8.79
StPRX101	PGSC0003DMT400075415	57162212..57162790	12	154	17167.35	9.47
StPRX102	PGSC0003DMG402029332	57165238..57166864	12	318	34969.76	8.81

Chr., chromosome; PL, protein length; MW, molecular weight; PI, theoretical isoelectric points.

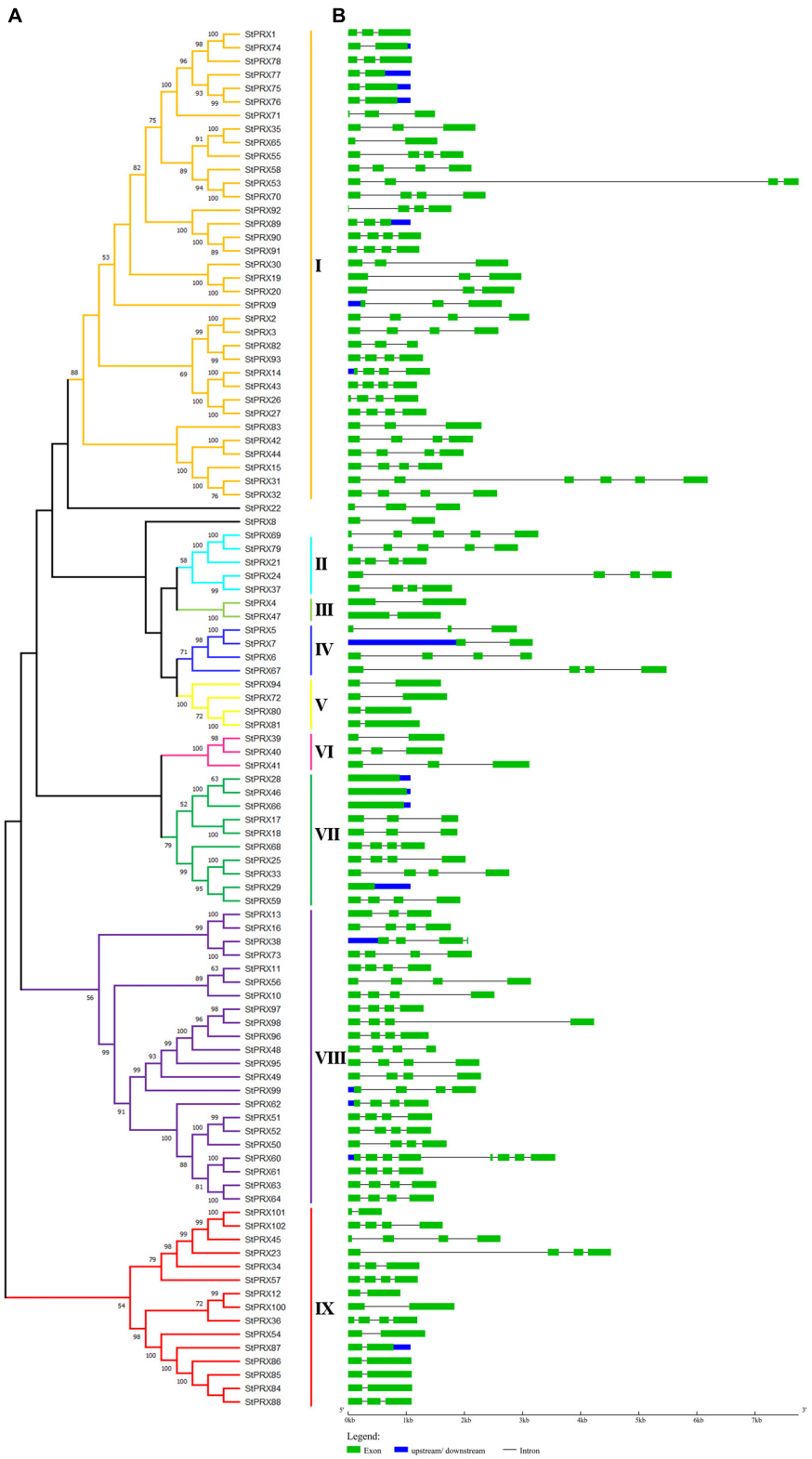


FIGURE 1 | Phylogenetic tree and gene structure of *StPRX* genes. **(A)** The maximum-likelihood phylogenetic tree generated by MEGA X with bootstrapping analysis (1,000 replicates). Each subfamily is distinguished by different colors. **(B)** The exon-intron structure of *StPRX* genes generated by the online software GSDS. The horizontal black lines and the green boxes represent introns and exons, respectively. The blue boxes represent upstream or downstream sequences. The sizes of exons and introns can be estimated according to the scale at bottom.

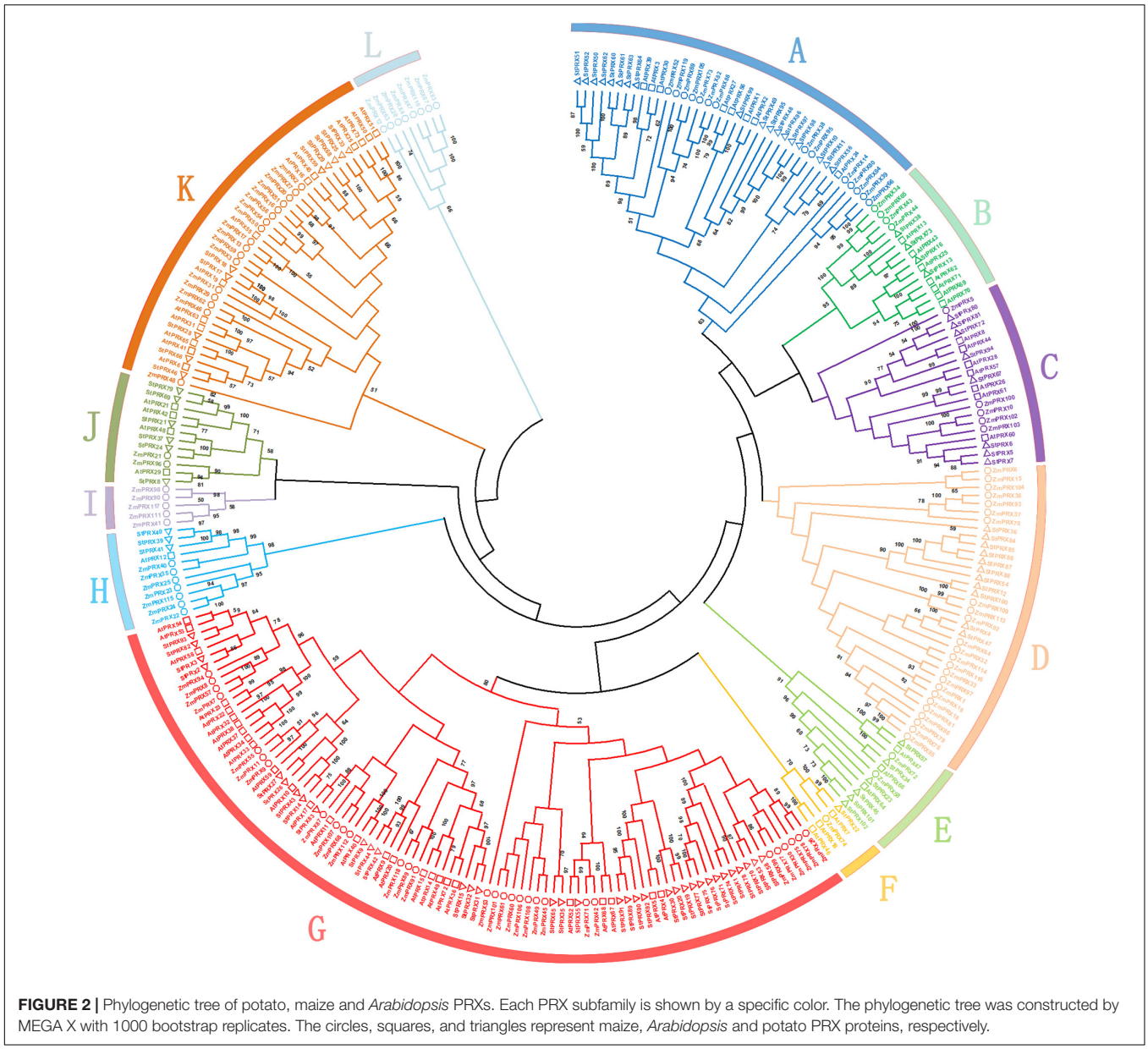


FIGURE 2 | Phylogenetic tree of potato, maize and *Arabidopsis* PRXs. Each PRX subfamily is shown by a specific color. The phylogenetic tree was constructed by MEGA X with 1000 bootstrap replicates. The circles, squares, and triangles represent maize, *Arabidopsis* and potato PRX proteins, respectively.

and maize, probably because of the closer evolutionary distance between these eudicot species.

Gene Structure and Protein Motif Analysis of *StPRX*

Structure analysis of *StPRX* genes showed that the number of introns varied from 0 to 7 (**Figure 1B**). Most of them had 1–3 introns, with *StPRX60* containing the maximum (7) while four genes lacked introns (*StPRX28*, *StPRX29*, *StPRX46,* and *StPRX66*) (**Figure 1B**). Genes with similar exon/intron structure were grouped together, but structural variation was also found among these *StPRX* genes (**Figure 1**).

To investigate the diversity of motif components among *StPRXs*, the motif distribution in 102 StPRX proteins was investigated using the online tool MEME program. A total of 18 conserved motifs were identified (**Figure 3** and **Supplementary Table S2**). The majority of StPRX proteins contained two to three conserved motifs. The StPRX proteins on the same branch had similar conserved motif composition and sorting order, suggesting that StPRX proteins in the same branch might share similar function. Using the CDD tool, a total of 11 motifs (motif 1/2/3/4/5/6/7/9/10/12/15) were functionally annotated for the components of the conserved PRX domain (**Figure 3** and **Supplementary Table S2**). All the members of the potato PRX family contained at least one motif belonging to the typical domains of PRX family. In addition, some motifs appeared to be unknown in function. For example, the functions of three motifs (motif 11/14/18) in subgroup III and of four motifs (motif

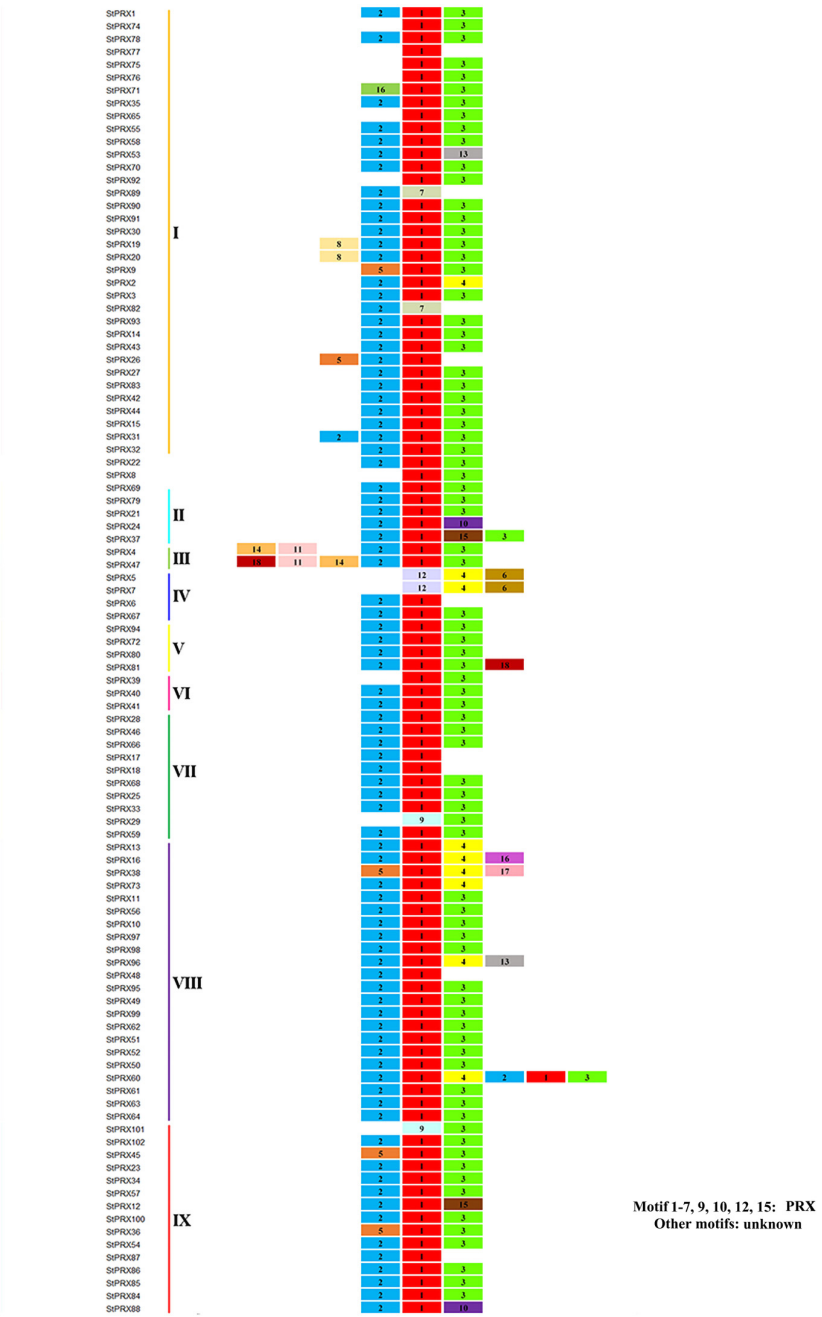


FIGURE 3 | Conserved motifs of PRX proteins in potato. The protein list on the left is the same as that in **Figure 1**. The genes of different subfamilies are highlighted with different colors. Different motifs are exhibited with different colored boxes and numbers (1–18). The detailed sequences of motifs are listed in **Supplementary Table S2**.

8/13/16/17) in subgroups I, III, V, and VIII were yet to be determined.

Chromosomal Locations and Duplications of *StPRX* Genes

To reveal the genome organization and distribution of *StPRX* on different chromosomes in potato, a graph of chromosomes was constructed using the MapInspect tool. A total of 101 out of the 102 *StPRX* genes were located on the 12 potato chromosomes (**Figure 4**). Among them, the largest number of *StPRX* genes (15) was located on chromosome 2, followed by chromosomes 1 (13), and chromosomes 11 (11) and 10 (10). In contrast, only a few *StPRX* genes were located on chromosomes 8 (3), 6 (4), and 9 (4). In addition, some chromosomes showed a dense cluster of *StPRXs,* such as near the telomeric region of chromosomes

2, 4, 5, 7, 10, 11, and 12. Gene duplication events, including segmental duplication and tandem duplication, are important for the expansion of the gene family during the process of the evolution (Cannon et al., 2004). In this study, a total of 83 *StPRX* gene pairs were identified from the phylogenetic and comparative analysis (**Figure 4**), among which 57 pairs were found to be involved in the segmental duplication events, and 26 pairs were confirmed to be tandem duplicated genes (**Figure 4** and **Supplementary Table S3**). The number of segmental duplication gene pairs was twice as many as that of the tandem duplicated, and most of the tandem duplicated gene pairs were densely distributed at the end of chromosomes 5, 7, 10, 11, and 12.

Expression Patterns of *StPRX* Genes

To further explore the expression patterns of the *StPRX* genes, the transcript data of major tissues was obtained from the public genome database, including root, shoot, petal, carpel, sepal, stamen, tuber, leaf, flower and petiole. A heatmap was generated based on the transcript data of 80 *StPRX* genes, and the other 22 genes were excluded from the heat map analysis due to the low expression level (FPKM < 0.5) or lack of expression in all tissues (**Figure 5**). As shown in **Figure 5**, some *AtPRX* genes exhibited distinct tissue-specific expression patterns, while others were active in the whole plant. The 80 *StPRX* genes were grouped into four groups (**Figure 5**). Six genes (*StPRX41*/*19*/*28*/*25*/*40*/*21*) were included in group IV, which had especially abundant expression level in all of the developmental stages, suggesting that these genes might play important basic roles in all development stages of plant. A total of 19 genes (*StPRX51*/*13*/*20*/*30*/*29*/*55*/*39*/*62*/*23*/*2*/*3*/*34*/*9*/*69*/*83*/*68*/*35*/*57*/*33*) were included in group III, which had high expression levels in most of the analyzed tissues. In contrast, 44 genes in group II exhibited low expression or no expression in the most of the tissues analyzed. However, most of them showed relatively high expression in root than in other tissues. This was also observed in the *PRX* family of maize (Wang et al., 2015) and *Arabidopsis* (Tognolli et al., 2002). Therefore, the expression patterns of *StPRX* genes may reflect the correlation with their functions.

Expression of *StPRX* Genes in Response to Abiotic Stress

To further investigate the response of the *StPRX* genes subjected to different stresses, including heat, salt and drought, the relative expression levels among the tissues were measured based on the expression FPKM values (stress/control) (**Figure 6**). The expression levels of different genes showed great variation under the various types of treatments. Under heat stress, most of the *StPRX* genes in groups III and IV were significant upregulated, such as *StPRX93*/*33*/*39*/*74*/*31*/100/36/22, whereas most of the *StPRX* genes in groups I and II exhibited significant downregulation, such as *StPRX 28*, *-18* and *-17*. Most of the *StPRX* genes showed downregulation under the stress of drought, in contrast to the performance under heat stress, and some *StPRX* genes in groups II and III were extremely sensitive to the drought stress and exhibited significant downregulation, such as *StPRX4*/*13*/*31*/*38*/*46*/*57*/*51*/*58*/*74*/*77*/*89*/*91*. In terms of saline stress, the genes exhibited diverse responses, the genes in group IV were significantly upregulated, and the genes in group II showed significant downregulation, while the genes in groups I and III only showed slight changes, implying the functional dissimilation among the *StPRX* genes.

Seven *StPRX* genes (*StPRX19*/*28*/*33*/*35*/*40*/*41*/*57*) with high expression levels in all organs were selected for further qRT-PCR analysis under different abiotic stresses. The 7 genes showed different levels of response to the three abiotic stress treatments (**Figure 7**). All of the selected genes were up-regulated under heat stress, and the expression change of *StPRX57* was over twofold. For saline stress, most of the selected genes were up-regulated, and the upregulation of *StPRX33 and StPRX57* was more than threefold. In response to drought stress, 5 genes (*StPRX19*/*28*/*40*/*41*/*57*) were upregulated, whereas *StPRX33 and StPRX35* appeared to be slightly down-regulated. Notably, 2 genes (*StPRX41 and StPRX57*) were extremely sensitive to drought stress. Their expression levels increased 4 folds compared to that of control.

DISCUSSION

PRXs are plant-specific enzymes that have multiple functions in the growth and development of plants. It has been confirmed that *PRX* genes are widely involved in stress response in many species (Gray and Montgomery, 2003; Xue et al., 2008). To date, the comprehensive genome-wide analysis of the PRX gene family has been performed in many plant species, including *Arabidopsis* (Tognolli et al., 2002), rice (Passardi et al., 2004a), and maize (Wang et al., 2015). In this study, a total of 102 *StPRX* genes were identified. The number was more than that in *Arabidopsis* (containing 73 *PRX* genes), but slightly less than that of maize (119) and rice (138). According to the phylogenetic tree of the 288 PRX family members from *Arabidopsis*, maize and potato, we found that some groups only contained members from one or two species. In addition, orthologs between potato and *Arabidopsis* (82 pairs) were much more than those between potato and maize (4 pairs), suggesting that the *PRX* gene family underwent a specific expansion after the three species diverged each other in the evolutionary path of speciation, especially after the divergence of monocots and eudicots.

We have seen that the StPRX proteins showing similar domain architectures and motif constitutions were usually grouped in the same subfamily (**Figure 3**), and the structural constitutions of the *StPRX* genes in each group were basically in consistence with the result of phylogenetic analysis (**Figure 1**). Similar phenomenon have also been found in many other species, such as maize (Wang et al., 2015) and rice (Passardi et al., 2004a). As protein function is mostly determined by its domain structure, these results imply that the StPRX proteins with similar domain architectures and motif constitutions could probably perform similar functions.

It is known that the structural diversity of genes drives the evolution of multigene families. The intron/exon structure variation is a cause of gene diversity. Many studies have shown that introns are specifically inserted into and retained in the genome during evolution (Rogozin et al., 2003). The loss and

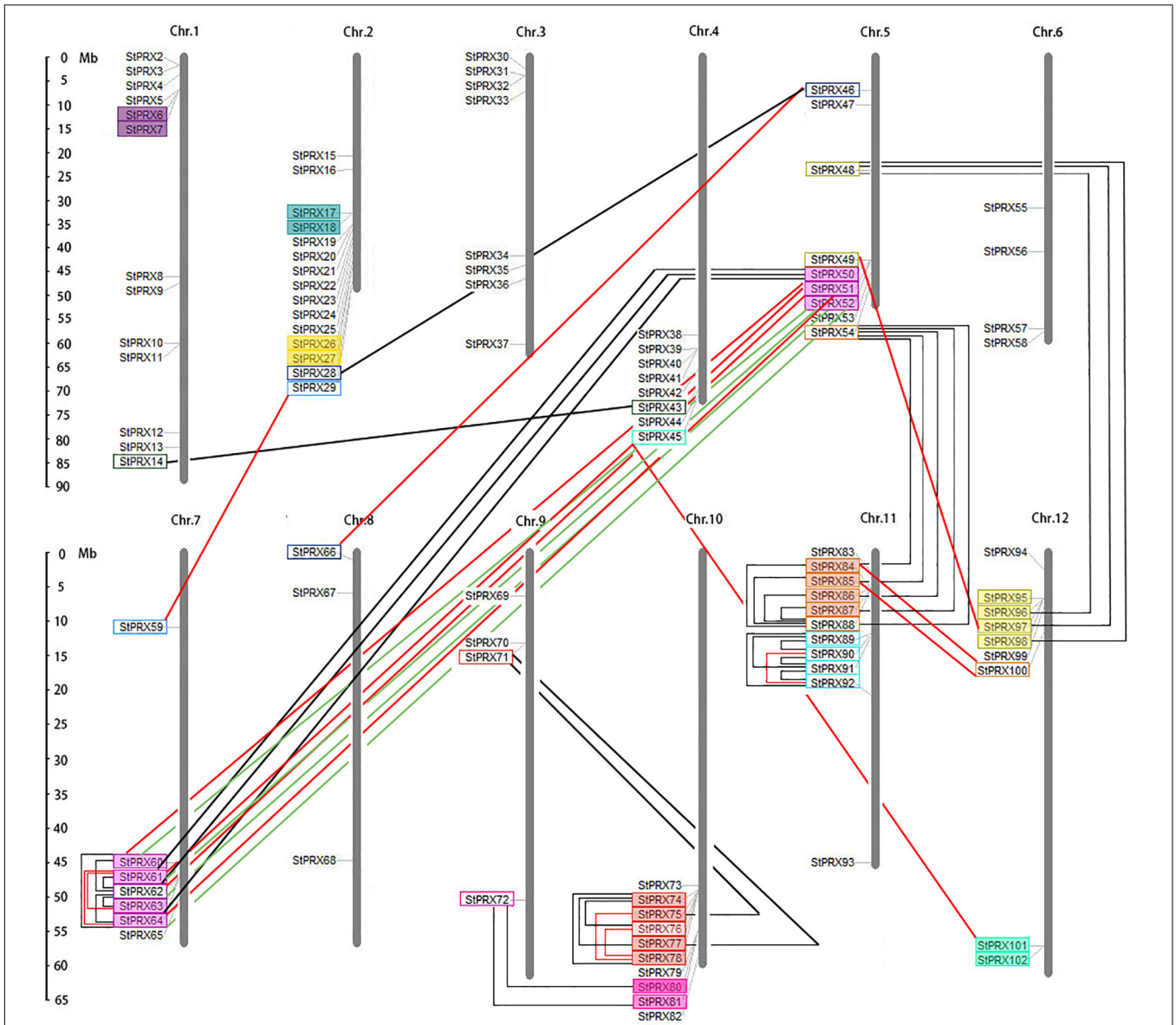


FIGURE 4 | Chromosomal locations of potato PRX genes. The scale on the left presents the length of potato chromosomes (Mb). Tandem duplicated gene pairs are displayed with boxes of the same color. Segmented duplicated gene pairs are displayed with boxes connected by lines.

insertion of new introns appear to be frequent events, which may result in diverse functional consequences in gene evolution. The duplication from an ancient gene formed by shuffling of small exons could be the reason that resulted in the genes with a relatively high number of introns (Gilbert et al., 1997). In maize, it is speculated that the intron variation among *PRX* genes might result from the depletion and duplication of single introns in the course of evolution (Wang et al., 2015). In this study, we found that the number of introns in *StPRX* genes was also quite variable (varying from 0 to 7), and the proteins in the same subfamily were not completely identical in terms of their intron/exon structure and motifs (**Figures 1**, **3**), implying that exon shuffling might be a main pattern of *StPRX* evolution, which might be the main contributors to the functional diversity of the potato PRX family. Among the 102 *StPRX* genes, more than half of them consist of 3 introns and 4 exons. This 3-intron/4-exon model is also represented a significant proportion in *Arabidopsis* (Tognolli et al., 2002) and in rice (Passardi et al., 2004a), suggesting that it is an ancestral intronic model of *PRX* genes.

Gene duplication events, including segmental and tandem duplication are important for the expansion of the gene family during the process of the evolution (Cannon et al., 2004). Gene duplication events can theoretically produce two gene copies, and one or both copies can acquire the novel gene functions for adaptation under a smaller selective pressure of evolution (Van de Peer et al., 2009). Each paralog is specialized

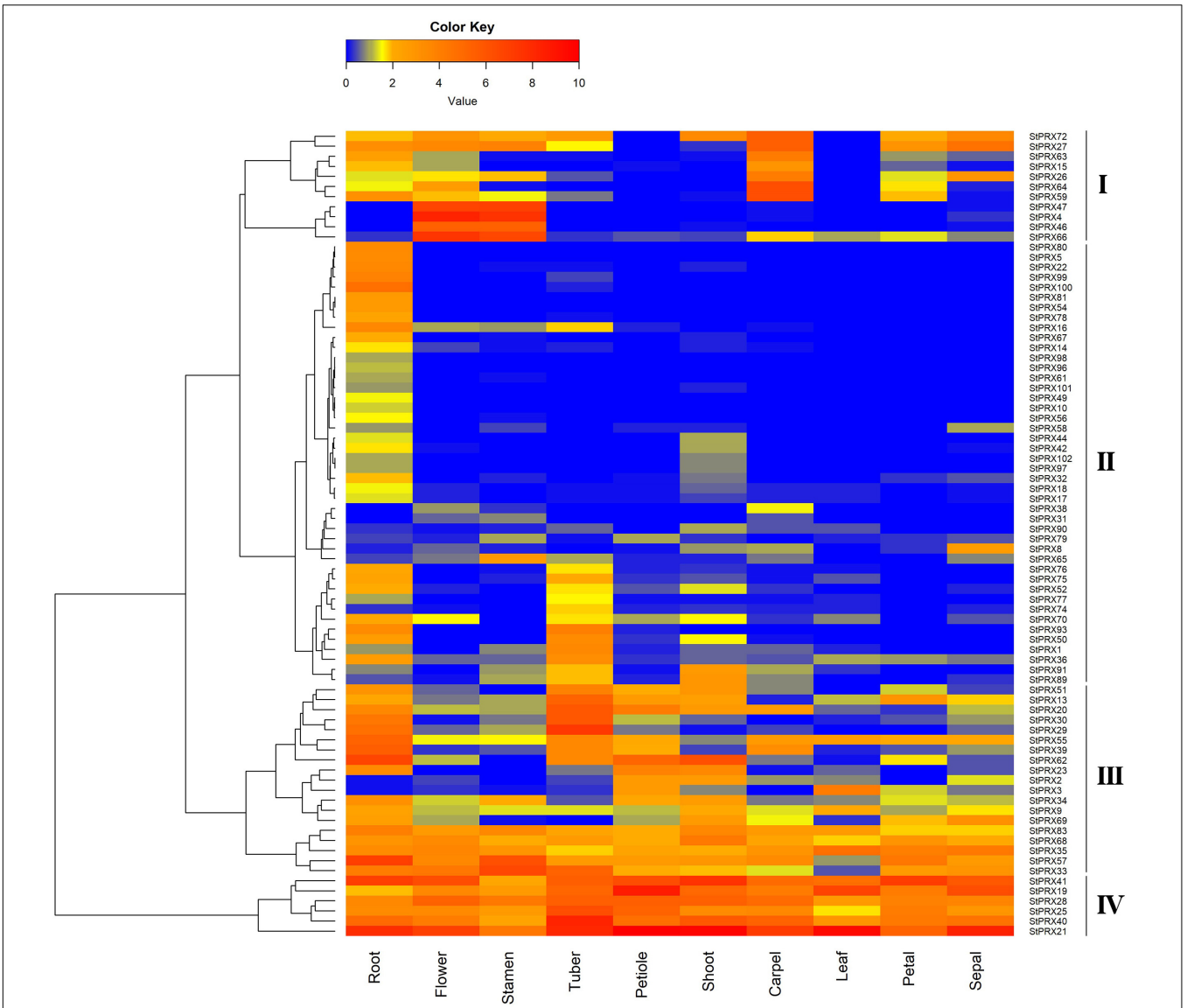


FIGURE 5 | Expression profiles of *StPRX* genes in different tissues. FPKM values for *StPRX* genes were transformed by $\log_2$(FPKM+1). The heatmap was generated using the R package of the heatmap function.

for a specific functional assignment (Zhang, 2003), which often leads to the expansion of gene family (Cannon et al., 2004). The segment duplication event refers to the duplication of large fragments of the genome, which may have derived from segmental, chromosomal or whole genome duplications with many losses and rearrangements (Zhang, 2003). Tandem duplication affects a limited number of genes (one or more neighboring genes); it often derives from unequal crossing-over (Achaz et al., 2000) and multiple episodes of unequal crossovers. In addition, the retrotransposition event of cDNA also contributes to the expansion of gene family, which is characterized by the loss of all introns and related regulatory sequences and by a random insertion within the genome. In our study, a total of 83 duplicated gene pairs were identified in potato *PRX* gene family, including 57 segmental duplication gene pairs and 26 tandem duplicated genes pairs (**Supplementary Table S3**), which were much more than those in maize (28 duplicated gene pairs; Wang et al., 2015) and rice (Passardi et al., 2004a). The segmental duplication gene pairs were twice as many as the tandem duplicated gene pairs, indicating that segmental duplication might play the dominant role in the expansion of the potato PRX family. Most of the tandem duplicated gene pairs were densely distributed in telomeric regions of chromosomes (such as on chromosomes 5 and 7), and many tandem duplicated genes shared high similarity with the same segmental duplication genes, implying that most of the tandem duplicated genes might appear after the segmental duplication events. Notably, two gene pairs (*StPRX28/46, StPRX46/66*) met the criteria of segmental

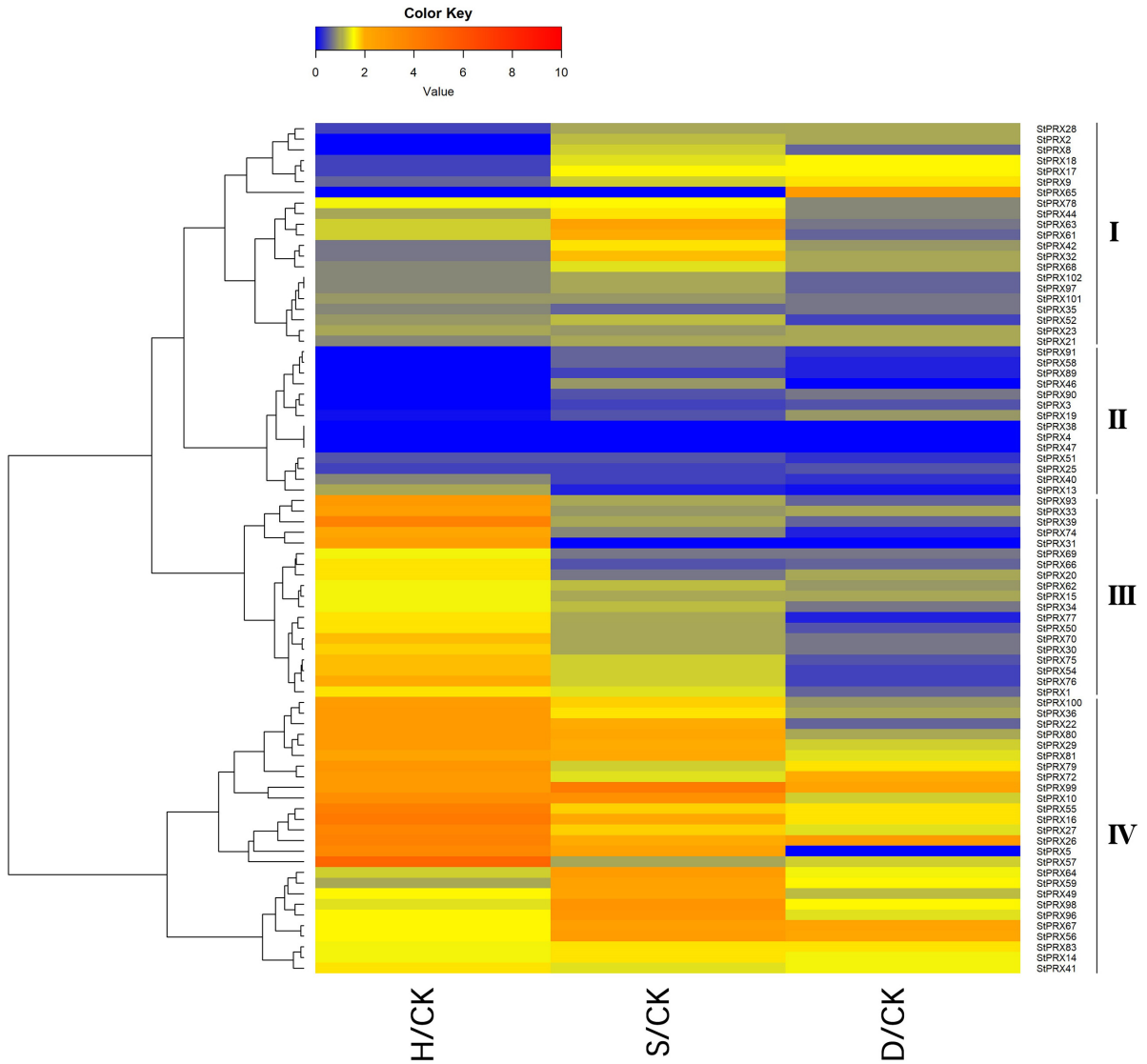


FIGURE 6 | Expression changes of StPRX genes under heat (H), salt (S) and drought (D) stresses. The expression change is indicated by the ratio of FPKM value of the treatment to that of the control (CK). The heatmap was generated using the R package of the heatmap function.

duplication gene pairs but had no introns, suggesting that they might likely be generated by retrotransposition. Strictly speaking, therefore, they might not be segmental duplication gene pairs.

Gene expression pattern is an important aspect related to gene function. In this study, among the 102 *StPRX* genes, except for 22 with weak or without expression in all of the tissues examined, the rest all exhibited distinct patterns of tissue-specific expression (**Figure 5**) and response to stress (**Figure 6**), indicating the functional dissimilation of StPRX proteins. This is consistent with the results of phylogenetic and protein motif analyses. Several genes in group IV were expressed in all organs (**Figure 5**), suggesting that they might play basic roles for the plant. Notably, the largest number of *StPRX* genes with high expression levels was found in root (**Figure 5**). Similar observations were also reported in maize (Wang et al., 2015), *Arabidopsis* (Tognolli et al., 2002), and rice (Passardi et al., 2004a), suggesting that the *PRX* family might be critical for root function in plants. In maize, there are many cell wall or membrane-bound PRXs in root (Mika et al., 2008; Šukaloviæ et al., 2015); several *ZmPRX* genes from roots are regulated by methyl jasmonate, salicylic acid and pathogen elicitors (Mika et al., 2010); and some genes (*ZmPRX26/42/71/75/78*) highly expressed in root show significant responses to H_2O_2, SA, NaCl, and PEG treatments (Wang et al., 2015). In *Arabidopsis*, two *AtPRX* genes (*AtPrx33/34*) are associated with root elongation (Passardi et al., 2006). Interestingly, the five *ZmPRX* genes (*ZmPRX26/42/71/75/78*) and two *AtPRX* genes (*AtPrx33/34*) were all clustered in grouped G in this study (**Figure 2**). Most *StPRX* genes in group G (**Figure 2**) were included in the subfamily I (**Figures 1**, **3**) with similar gene structure and motif

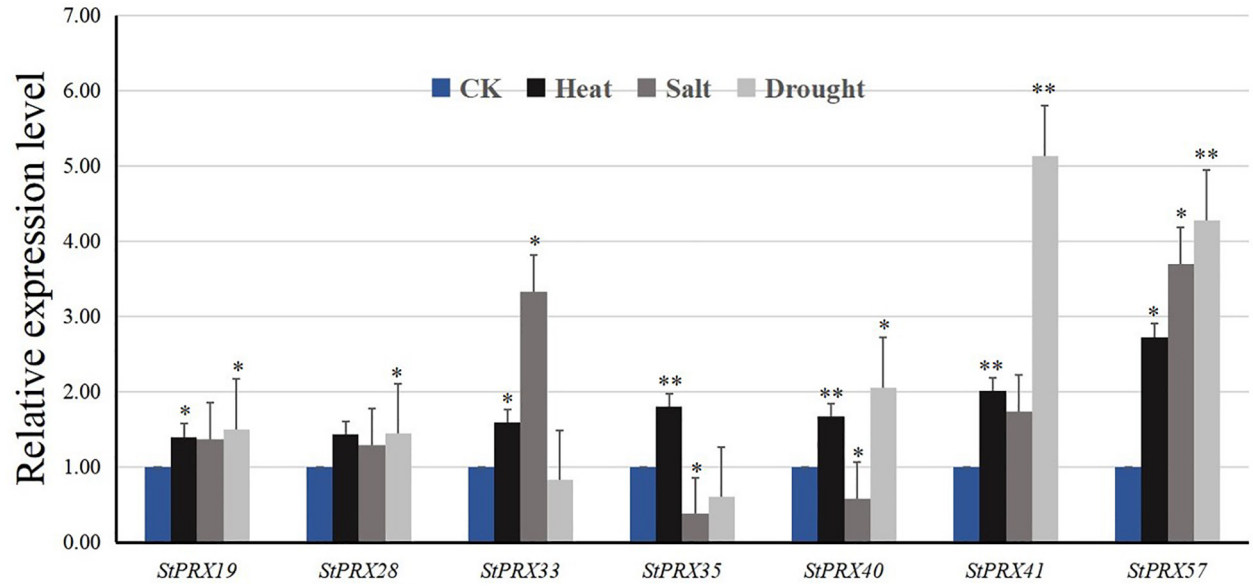


FIGURE 7 | Relative expression levels of 7 *StPRX* genes in response to heat, salt and drought stresses after 6 h treatment compared with that of control (CK). Error bars are standard deviations of three biological replicates. $^{*}P < 0.05$, $^{**}P < 0.01$.

components. The results imply that the *StPRX* genes clustered in group G might also function in root similar to their counterparts in maize and Arabidopsis. Therefore, the results of our study may provide a basis for the functional exploration of the potato *PRX* gene family members.

To analyze the trend of the gene expression derived from qRT-PCR (**Figure 7**) and the FPKM values, we compared the results from these two different platforms (**Figure 6**). Overall, similar propensity of gene expression was found between the two different approaches. However, the results of qRT-PCR did not totally agree with the pattern of the gene expression from RNA-seq data. There could be several reasons for the discrepancy. First, the genotypes of potato varieties used in the two experiments were different. A doubled monoploid potato variety (DM) was used in RNA-seq (Xu et al., 2011), whereas an autotetraploid cultivar Zhongshu 3 was used for qRT-PCR analysis. Second, the experimental treatments were different. The plantlets for qRT-PCR were grown under a photoperiod of 16 h light/8 h dark environment, while the materials for RNA-seq were grown in the dark (Xu et al., 2011).

CONCLUSION

In this study, a genome-wide investigation and comprehensive analysis of the *PRX* gene family in potato was conducted. The structural diversity of *StPRXs* may reflect their functional diversity. The analysis of expression patterns of *StPRX* genes showed that these genes were expressed distinctly in different tissues of potato, and some might be linked to stress responses. It is important to thoroughly investigate the biological functions of *StPRX* genes, especially the roles in the resistance to abiotic stresses. Our results provide the vital information for the exploration of the functional aspect of the gene family.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

XY and JYa performed the experiments and analyzed the data. WL collected plant materials and performed the experiments. MQ participated in handling figures and tables. XX designed this research and wrote the manuscript. WW and JYu helped to draft the manuscript. All authors read and approved the manuscript.

FUNDING

This study was supported by the National Natural Science Foundation of China (No. 31501085) and the Fujian Agriculture and Forestry University innovation foundation (No. CXZX2019052G).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.593577/full#supplementary-material

REFERENCES

Achaz, G., Coissac, E., Viari, A., and Netter, P. (2000). Analysis of intrachromosomal repeats in yeast Saccharomyces cerevisiae: a possible model for their origin. *Mol. Biol. Evol.* 17, 1268–1275. doi: 10.1093/oxfordjournals.molbev.a026410

Allison, S. D., and Schultz, J. C. (2004). Differential activity of peroxidase isozymes in response to wounding, gypsy moth, and plant hormones in northern red oak (*Quercus rubra* L.). *J. Chem. Ecol.* 30, 1363–1379. doi: 10.1023/b:joec.0000037745.66972.3e

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., and Miller, W. (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

Bailey, T. L., Boden, M., Buske, F. A., Frith, A., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37, 202–208.

Barcelo, A. R., and Pomar, F. (2001). Oxidation of cinnamyl alcohols and aldehydes by a basic peroxidase from lignifying *Zinnia elegans* hypocotyls. *Phytochemistry* 57, 1105–1113. doi: 10.1016/s0031-9422(01)00050-4

Bindschedler, L. V., Julia, D., Julie, M., Blee, K. A., Julia, P., Dewi, R. D., et al. (2006). Peroxidase-dependent apoplastic oxidative burst in *Arabidopsis* required for pathogen resistance. *Plant J.* 47, 851–863. doi: 10.1111/j.1365-313x.2006.02837.x

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

Cannon, S. B., Mitra, A., Baumgarten, A., Young, N. D., and May, G. (2004). The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biol.* 4:10. doi: 10.1186/1471-2229-4-10

Duroux, L., and Welinder, K. G. (2003). The peroxidase gene family in plants: a phylogenetic overview. *J. Mol. Evol.* 57, 397–407. doi: 10.1007/s00239-003-2489-3

Erman, J. E., and Vitello, L. B. (2002). Yeast cytochrome c peroxidase: mechanistic studies via protein engineering. *Biochim. Biophys. Acta* 1597, 193–220. doi: 10.1016/s0167-4838(02)00317-5

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D., and Bairoch, A. (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* 31, 3784–3788. doi: 10.1093/nar/gkg563

Gazaryan, I. G., Lagrimini, L. M., Ashby, G. A., and Thorneley, R. N. (1996). Mechanism of indole-3-acetic acid oxidation by plant peroxidases: anaerobic stopped-flow spectrophotometric studies on horseradish and tobacco peroxidases. *Biochem. J.* 313, 841–847. doi: 10.1042/bj3130841

Gilbert, W., de Souza, S. J., and Long, M. (1997). Origin of genes. *Proc. Natl. Acad. Sci. U.S.A.* 94, 7698–7703.

Gray, J. S., and Montgomery, R. (2003). Purification and characterization of a peroxidase from corn steep water. *J. Agric. Food Chem.* 51, 1592–1601. doi: 10.1021/jf025883n

Gu, Z. L., Cavalcanti, A., Chen, F. C., Bouman, P., and Li, W. H. (2002). Extent of gene duplication in the genomes of *Dro-sophila*. *Nematode Yeast. Mol. Biol. Evol.* 19, 256–262. doi: 10.1093/oxfordjournals.molbev.a004079

Guo, A. Y., Zhu, Q. H., and Chen, X. (2007). GSDS: a gene structure display server. *Hereditas* 29, 1023–1026. doi: 10.1360/yc-007-1023

Hammer, P. A., Tibbitts, T. W., Langhans, R. W., and Mcfarlane, J. C. (1978). Baseline growth studies of 'Grand Rapids' lettuce in controlled environments. *J. Amer. Soc. Hor. Sci.* 103, 649–655.

Herrero, J., Fernández-Pérez, F., Yebra, T., Novo-Uzal, E., Pomar, F., Pedreño, M. Á, et al. (2013). Bioinformatic and functional characterization of the basic peroxidase 72 from *Arabidopsis thaliana* involved in lignin biosynthesis. *Planta* 237, 1599–1612. doi: 10.1007/s00425-013-1865-5

Hiraga, S., Ichinose, C., Onogi, T., Niki, H., and Yamazoe, M. (2000). Bidirectional migration of SeqA-bound hemimethylated DNA clusters and pairing of oriC copies in *Escherichia coli*. *Genes Cells* 5, 327–341. doi: 10.1046/j.1365-2443.2000.00334.x

Hiraga, S., Sasaki, K., Ito, H., Ohashi, Y., and Matsui, H. (2001). A large family of class III plant peroxidases. *Plant Cell Physiol.* 42, 462–468. doi: 10.1093/pcp/pce061

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

Li, X. X., Duan, X., Jiang, H. X., Sun, Y. J., Tang, Y. P., Yuan, Z., et al. (2006). Genome-wide analysis of basic/helix-loop-helix transcription factor family in rice and *Arabidopsis*. *Plant Physiol.* 141, 1167–1184. doi: 10.1104/pp.106.080580

Liszkay, A., Vander, Z. E., and Schopfer, P. (2004). Production of reactive oxygen intermediates (O2-, H2O2, and (.)OH) by maize roots and their role in wall loosening and elongation growth. *Plant Physiol.* 136, 3114–3123. doi: 10.1104/pp.104.044784

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

Loukili, A., Limam, F., Ayadi, A., Boyer, N., and Ouelhazi, L. (1999). Purification and characterization of a neutral peroxidase induced by rubbing tomato internodes. *Physiol. Plant.* 105, 24–31. doi: 10.1034/j.1399-3054.1999.105105.x

Marchler, B., A., Bo, Y., Han, L., He, J., Lanczycki C, J., Lu S, N., et al. (2017). CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* 45, D200–D203. doi: 10.1093/nar/gkw1129

Mei, W., Qin, Y., and Song, W. (2009). Cotton GhPOX1 encoding plant class III peroxidase may be responsible for the high level of reactive oxygen species production that is related to cotton fiber elongation. *Genet. Genom.* 36, 141–150. doi: 10.1016/s1673-8527(08)60101-0

Mika, A., Boenisch, M. J., Hopff, D., and Lüthje, S. (2010). Membrane-bound guaiacol peroxidases from maize (*Zea mays* L.) roots are regulated by methyl jasmonate, salicylic acid, and pathogen elicitors. *J. Exp. Bot.* 61, 831–841. doi: 10.1093/jxb/erp353

Mika, A., Buck, F., and Lüthje, S. (2008). Membrane-bound class III peroxidases: identification, biochemical properties and sequence analysis of isoenzymes purified from maize (*Zea mays* L.) roots. *J. Proteome.* 71, 412–424. doi: 10.1016/j.jprot.2008.06.006

Pang, W. Y. (2019). The significance and implementation of the potato main graining strategy. *Grain Process.* 44, 59–61.

Passardi, F., Longet, D., Penel, C., and Dunand, C. (2004a). The class III peroxidase multigenic family in rice and its evolution in land plants. *Phytochemistry* 65, 1879–1893. doi: 10.1016/j.phytochem.2004.06.023

Passardi, F., Penel, C., and Dunand, C. (2004b). Performing the paradoxical: how plant peroxidases modify the cell wall. *Trends Plant Sci.* 9, 534–540. doi: 10.1016/j.tplants.2004.09.002

Passardi, F., Tognolli, M., and De Meyer, M. (2006). Two cell wall associated peroxidases from *Arabidopsis* influence root elongation. *Planta* 223, 965–974. doi: 10.1007/s00425-005-0153-4

Piontek, K., Smith, A. T., and Blodig, W. (2001). Lignin peroxidase structure and function. *Biochem. Soc. Trans.* 29, 111–116. doi: 10.1042/bst0290111

Rogozin, I. B., Wolf, Y. I., Sorokin, A. V., Mirkin, B. G., and Koonin, E. V. (2003). Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr. Biol.* 13, 1512–1517. doi: 10.1016/s0960-9822(03)00558-x

Schopfer, P., Liszkay, A., Bechtold, M., Frahry, G., and Wagner, A. (2002). Evidence that hydroxyl radicals mediate auxin-induced extension growth. *Planta* 214, 821–828. doi: 10.1007/s00425-001-0699-8

Shigeoka, S., Ishikawa, T., Tamoi, M., Miyagawa, Y., Takeda, T., Yabuta, Y., et al. (2002). Regulation and function of ascorbate peroxidase isoenzymes. *J. Exp. Bot.* 53, 1305–1319. doi: 10.1093/jexbot/53.372.1305

Šukaloviæ, V. H.-T., Vuletiæ, M., Markoviæ, K., Antiæ, T. C., and Vuèiniæ, Ž (2015). Comparative biochemical characterization of peroxidases (class III) tightly bound to the maize root cell walls and modulation of the enzyme properties as a result of covalent binding. *Protoplasma* 252, 335–343. doi: 10.1007/s00709-014-0684-2

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997). The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25, 4876–4882. doi: 10.1093/nar/25.24.4876

Tognolli, M., Penel, C., Greppin, H., and Simon, P. (2002). Analysis and expression of the class III peroxidase large gene family in *Arabidopsis thaliana*. *Gene* 288, 129–138. doi: 10.1016/s0378-1119(02)00465-1

Van de Peer, Y., Maere, S., and Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* 10, 725–732. doi: 10.1038/nrg2600

Wang, L. Q., Guo, K., Li, Y., Tu, Y. Y., Hu, H. Z., Wang, B. R., et al. (2010). Expression profiling and integrative analysis of the CESA/CSL superfamily in rice. *BMC Plant Biol.* 10:282. doi: 10.1186/1471-2229-10-282

Wang, R. Q., Zhao, P., Kong, N. N., Lu, R. Z., Pei, Y., Huang, C. X., et al. (2018). Genome-wide identification and characterization of the potato bHLH transcription factor family. *Genes* 9:54. doi: 10.3390/genes9010054

Wang, Y., Wang, Q. Q., Zhao, Y., Han, G. M., and Zhu, S. W. (2015). Systematic analysis of maize class III peroxidase gene family reveals a conserved subfamily involved in abiotic stress response. *Gene* 566, 95–108. doi: 10.1016/j.gene.2015.04.041

Warnes, G., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., et al. (2016). *Various R Programming Tools for Plotting Data*. Available online at: https://cran.r-project.org/

Welinder, K. G. (1993). "Plant peroxidases: structure–function relationships," in *Plant Peroxidases*, eds C. Penel, T. Gaspar, and H. Greppin (Geneva: University of Geneva), 1–24. doi: 10.1385/1-59259-232-5:1

Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D., Ni, P., et al. (2011). Potato Genome Sequencing Consortium. Genome sequence and analysis of the tuber crop potato. *Nature* 475, 189–195. doi: 10.1038/nature10158

Xue, Y. J., Tao, L., and Yang, Z. M. (2008). Aluminum-induced cell wall peroxidase activity and lignin synthesis are differentially regulated by jasmonate and nitric oxide. *J. Agric. Food Chem.* 56, 9676–9684. doi: 10.1021/jf802001v

Yang, S. H., Zhang, X. H., Yue, J. X., Tian, D. C., and Chen, J. Q. (2008). Recent duplications dominate NBS-encoding gene expansion in two woody species. *Mol. Genet. Genomics* 280, 187–198. doi: 10.1007/s00438-008-0355-0

Zhang, J. Z. (2003). Evolution by gene duplication: an update. *Trends Ecol. Evol.* 18, 292–298. doi: 10.1016/s0169-5347(03)00033-8

Zheng, X. H., and Van Huystee, R. B. (1991). Oxidation of tyrosine by peroxidase isozymes derived from peanut suspension culture medium and by isolated cell walls. *Plant Cell Tissue Organ. Cult.* 25, 35–43. doi: 10.1007/bf00033910

Zhou, X. J., Gao, Y. X., and Wang, Z. B. (2010). Study on the characteristics of "Black Beauty" potato peroxidase. *Food Ferment. Ind.* 36, 55–63.

Zhu, H. S. (2017). "Cloning and analysis of the family of common melon peroxidase (POD) gene," in *Proceedings of the 13th National Congress of the Chinese Horticultural Society and the 2017 Annual Conference of The Chinese Horticultural Society*, Kunming.

Zhu, X. P., and Hu, H. (2013). Advances in the study of browning in potato processing. *Agro Process.* 8, 60–62.

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.

Copyright © 2020 Yang, Yuan, Luo, Qin, Yang, Wu and Xie. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 03 December 2020
doi: 10.3389/fgene.2020.550677

Identification and Expression Analysis of the *CsMYB* Gene Family in Root Knot Nematode-Resistant and Susceptible Cucumbers

Chunyan Cheng[1], Qingrong Li[1], Xing Wang[2], Ying Li[3], Chuntao Qian[1], Ji Li[1], Qunfeng Lou[1], Molly Jahn[4] and Jinfeng Chen[1]**

[1] College of Horticulture, Nanjing Agricultural University, Nanjing, China, [2] Hebei University of Engineering, Handan, China, [3] Nanjing Vegetable Science Research Institute, Nanjing, China, [4] Jahn Research Group, Department of Agronomy, University of Wisconsin–Madison, Madison, WI, United States

MYB (myeloblastosis) transcription factors (TFs) play important roles in controlling various physiological processes in plants, such as responses to biotic and abiotic stress, metabolism, and defense. A previous study identified a gene, *Csa6G410090*, encoding a plant lipid transfer protein (LTP), as a possible regulator in cucumber (*Cucumis sativus* L.) of the resistance response to root-knot nematode (RKN) [*Meloidogyne incognita* Kofoid and White (Chitwood)]. Myb-type DNA-binding TFs were presumed to regulate downstream genes expression, including LTPs, however, the regulation mechanism remained unclear. To elucidate whether and which MYB TFs may be involved in regulation of the resistance response, this study identified 112 genes as candidate members of the *CsMYB* gene family by combining CDD and SMART databases, using the Hidden Markov Model (HMM) and manual calibration. Within this group, ten phylogenetic subgroups were resolved according to sequence-based classification, consistent with results from comprehensive investigation of gene structure, conserved motifs, chromosome locations, and *cis*-element analysis. Distribution and collinearity analysis indicated that amplification of the *CsMYB* gene family in cucumber has occurred mainly through tandem repeat events. Spatial gene expression analysis showed that 8 *CsMYB* genes were highly expressed at differing levels in ten different tissues or organs. The roots of RKN-resistant and susceptible cucumbers were inoculated with *M. incognita*, finding that *CsMYB* (*Csa6G538700*, *Csa1G021940*, and *Csa5G641610*) genes showed up-regulation coincident with upregulation of the "hub" gene *LTP* (*Csa6G410090*) previously implicated as a major gene in the resistance response to RKN in cucumber. Results of this study suggest hypotheses regarding the elements and regulation of the resistant response as well as possible RKN resistance-enhancing strategies in cucumber and perhaps more broadly in plants.

Keywords: cucumber, resistant introgression line, MYB, root-knot nematode, resistant mechanism

OPEN ACCESS

Edited by:
Xiaoming Song,
North China University of Science and Technology, China

Reviewed by:
Tariq Mukhtar,
Pir Mehr Ali Shah Arid Agriculture University, Pakistan
Yunyan Kang,
South China Agricultural University, China
Moses Muraya,
Chuka University, Kenya

****Correspondence:***
Molly Jahn
molly.jahn@jahnresearchgroup.net
Jinfeng Chen
jfchen@njau.edu.cn

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

Received: *10 April 2020*
Accepted: *10 September 2020*
Published: *03 December 2020*

Citation:
Cheng C, Li Q, Wang X, Li Y, Qian C, Li J, Lou Q, Jahn M and Chen J (2020) Identification and Expression Analysis of the CsMYB Gene Family in Root Knot Nematode-Resistant and Susceptible Cucumbers. Front. Genet. 11:550677. doi: 10.3389/fgene.2020.550677

INTRODUCTION

Myeloblastosis proteins, named for their discovery as transcription regulators in myeloblastosis, function in plants as widely distributed and functionally diverse plant genetic transcription factors (TFs). In plants, MYB TFs play crucial roles in regulating plant physiological and biochemical processes (e.g., Daniel et al., 1999; Mengiste et al., 2003 Zhang et al., 2012; Sara et al., 2018). The MYB domain, consisting of a region of about 52 amino acids, bind DNA in a sequence-specific manner (Jin and Martin, 1999). The MYB gene family can be divided into four classes according to the number and type of MYB domains known as 1RMYB (or MYB-related), 2RMYB, 3RMYB, and 4RMYB, respectively. The 2RMYB class is specific to plants and the largest (Dubos et al., 2010), having been annotated genome-wide in many plants. Previous studies showed that there are 190 members of the 2RMYB family in *Arabidopsis* (Riechmann et al., 2000), 102 members in *Oryza sativa* (Chen et al., 2006), 117 members in *Vitis vinifera* (Matus et al., 2008), 192 members in *Populus trichocarpa* (Wilkins et al., 2009), 244 members in *Glycine max* (Du et al., 2012), and more than 200 members in *Zea mays* (Dias et al., 2003). Only 55 *2RMYB* genes, however, have been identified in the most recent version of cucumber genome sequence (*Cucumis sativus* L.), the smallest number identified in any plant species to date, owing to the absence of recent gene duplication events (Li et al., 2012).

Functions of these MYB TFs have been explored in many studies on plant defense and response to various biotic and abiotic stresses (Raffaele et al., 2008; Cominelli and Tonelli, 2009; Zhang et al., 2019; Wang H. H. et al., 2020). For example, *BOS1* in *Arabidopsis* encodes a *2RMYB* TF that regulates anther development and also responses to biotic stresses, such as biotrophic pathogens (Ogata et al., 1994 Mengiste et al., 2003). *PbrMYB5* plays an active role in enhancing tolerance to chilling stresses in tobacco (*Nicotiana tabacum*) (Xing et al., 2019). *AtMYB60* and *AtMYB96* in *Arabidopsis* are required for ABA signaling, stomatal movement, drought stress and disease resistance (Cominelli et al., 2005; Seo et al., 2009, 2011; Seo and Park, 2010). In tomato (*Solanum lycopersicum* L.), R3-MYB inhibits anthocyanin production (Colanero et al., 2018). In kiwifruit (*Actinidia chinensis*), *MYB7* regulates transcriptional activation of metabolic pathway genes to modulate carotenoid and chlorophyll pigment accumulation (Ampomah-Dwamena et al., 2019). *McMYB10* is involved in anthocyanin biosynthesis in crabapple during continuous light treatments (Li et al., 2018). In cotton, *GhMYB7*, *GhMYB9*, and *GhMYB109* were involved in cotton fiber initial cells, as well as elongating fibers (Suo et al., 2003).

Although new insights regarding the mechanisms by which MYB proteins control various plant processes and several core genes have been identified, the TFs in question are usually encoded by very large multigene families confounding detailed mechanistic studies, consistent with the likelihood that many of the complex regulatory roles played by *MYB* genes-encoding-proteins remain unknown (Riechmann et al., 2000). Our previous comparative transcriptomic study showed that MYB TFs may participate in regulating the hub gene *LTP* (lipid transfer protein) of cucumber in resistance to root-knot nematode (RKN) [*Meloidogyne incognita* Kofoid and White (Chitwood)] (Wang et al., 2018). LTP, a member of plant pathogenesis-related protein (PRs), is involved in lipid transport in giant cells (GCs) membranes and receives signals stimulated by exposure to nematodes (Wang X. et al., 2020). These results from cucumber are consistent with results from *Arabidopsis*, where *LTP3* is positively regulated by the TF *MYB96* to mediate freezing and drought stress (Guo et al., 2013). However, the mechanism by which MYB TF(s) regulates LTP in cucumber is still unknown. In the present study, we carried out a systematic analysis of *CsMYB* genes in cucumber, showing that this family members are highly variable and apparently play a positive role in responding to *M. incognita* in cucumber. Our results provide more evidence regarding whether and how *CsMYB* genes regulate *LTP*, thereby potentially affecting resistance to RKN in cucumber.

MATERIALS AND METHODS

Identification of Cucumber MYB Gene Family

The cucumber whole genome sequence was downloaded from the cucumber genome database[1]. The MYB binding domains (PF00249) in the sequence were identified and confirmed using the Hidden Markov Model (HMM) profile from PFam database[2] applied as a query to previously annotated cucumber MYB-containing sequences, further identified by HMMSEARCH and manual screened using default parameter ($E \leq 1.2 \times 10^{-8}$). HMMRESEARCH was applied to establish the cucumber MYB (*CsMYB*) gene family-specific model (Lozano et al., 2015), which was visualized online by SKylign[3]. Then all predicted cucumber MYB proteins with an E-value below 0.001 were selected using the cucumber MYB-specific HMM. The candidate genes for the *CsMYB* family were identified according to cucumber-specific MYB HMM.

All the candidate *CsMYB* genes initially obtained above were uploaded to BlastP by CDD (Conserved Domain Database) (Marchler-Bauer et al., 2017) and SMART (Simple Modular Architecture Research Tool) (Marchler-Bauer et al., 2015; Letunic and Bork, 2018) using the default parameter $E \leq 0.01$. Four classes (1RMYB, 2RMYB, 3RMYB, and 4RMYB) of *CsMYB* genes were separated according to characteristic conserved domains within the MYB sequence.

All the predicted protein sequences of the candidate *CsMYB* family members were submitted to ExPASy[4] to calculate the number of amino acids (aa), molecular weight (Mw) and isoelectric point (pI). The candidate *CsMYB* genes were renamed according to the repeat number of conserved domains and the physical position on the chromosome. The subcellular localization analysis of the candidate *CsMYB* family protein sequences were carried out by CELLO v.2.5

[1]http://cucurbitgenomics.org/
[2]http://pfam.xfam.org/
[3]http://skylign.org/
[4]http://web.expasy.org/protparam/

programme (subCELlular LOcalization predictor) at Molecular Bioinformatics Center[5] (Yu et al., 2004, 2006).

Gene Structure and Analysis of Conserved Motifs Among Cucumber MYB Gene Family Members

The exon and intron structures of the candidate *CsMYB* genes were obtained by comparison with the corresponding full-length genome sequences. Conserved motifs in the gene family were determined using the following parameters: arbitrary number of repetitions, up to 10; length, 6–200 amino acid residues via MEME software and predicted protein sequence (Bailey et al., 2009; Letunic et al., 2015). The gene structures and conserved motifs of *CsMYB* genes were displayed by TBTools (Chen et al., 2020).

Sequence Alignment and Phylogenetic Analysis of *CsMYB*

The sequences of candidate *CsMYB* proteins obtained were used to create Clustal V multiple sequence alignment using MEGA7 with *AtMYBs* in *Arabidopsis thaliana*. MEGA7 was used to analyze the phylogeny and evolution of all the predicted protein sequences by using the adjacency method (neighbor-joining, NJ) criterion. According to the results of the alignment, all the nodes were analyzed by bootstrap under the parameters, No. of Differences, Uniform rates, Complete deletion, and then repeateds 1000 times. The maximum likelihood (maximum likelihood, ML) method was used to verify the phylogenetic tree. Besides, phylogenetic tree between cucumber and other species was constructed based on MYB gene sequences from Gcorn database[6].

Orthologous Analysis and Chromosome Localization

The candidate *CsMYB* genes were located on cucumber chromosomes by Mapchart (Voorrips, 2002). The duplication events within the gene family were analyzed by Multiple Collinearity Scan toolkit (MCScanX) (Wang et al., 2012) and Circos (Krzywinski et al., 2009). In order to reveal the homology of MYB gene between cucumber and *Arabidopsis thaliana*, the homology analysis of MYB gene between *Arabidopsis thaliana* and cucumber was constructed by using the Dual Synteny Plotter software.

Cis-Elements Analysis and Prediction of CsMYB

The upstream sequence (1.5 kb) of *CsMYB* coding sequences were extracted from the cucumber genome sequence and submitted to PlantCARE[7] (Magali et al., 2002) for identification of *cis*-elements. The Gene Structure Display Server (GSDS)[8] (Li et al., 2012; Hu et al., 2014) is used to visualize all the regulatory elements involved in defense and stress response or defense factors, and to further analyze the regulatory elements related to resistance defense.

Response to *Meloidogyne incognita*

In order to study changes in gene expression correlated with expression of resistance to *Meloidogyne incognita* in cucumber, gene expression in a RKN-resistant line, IL10-1, and RKN-susceptible line, "Beijingjietou" (CC3), that display marked differences in against *M. incognita* (Ye et al., 2011), were obtained directly from NCBI by SRP125669 (Wang et al., 2018) at 0, 1, 2 and 3 days after inoculation (dpi) with approximately 50 fresh and motile second-stage juveniles (J2) *M. incognita*. The readings per kilobyte million enzyme (RPKM) for each MYB gene were downloaded and calculated from NCBI[9]. The RPKM values were converted through log2 transformation value (Li et al., 2011), and the cucumber genome database HEATMAP was used to draw the heatmap online.

RNA Isolation and RT-qPCR Analyses

Seeds of IL10-1 and CC3, provided by the State Key Lab of Cucurbit Genetics and Germplasm Enhancement of Nanjing Agricultural College, were surface sterilized (70% ethanol for 15 s and 1% sodium hypochlorite solution for 10 min, then rinsing with distilled water three times). Sterilized seeds were germinated on wet filter paper in a growth chamber at 28°C. After 3–4 days, seedlings were transplanted onto pluronic F-127 (Sigma-Aldrich, United States) gel to assess RKN infectivity (Wang et al., 2009), under controlled condition (28°C/16 h light, 24°C/8 h dark).

Inoculated root samples were harvested at 0, 1, 2, and 3 dpi as follows. Root tips from five different seedlings were mixed into a single biological sample, then 3 replicates were made from each sample. Total RNA from cucumber roots was extracted using the TRIzol reagent (Invitrogen, Carlsbad, CA, United States). Extracted RNA was quantified by Thermo NanoDrop 2000 (Thermo Fisher Scientific, MA, United States). cDNA was synthesized from the total RNA used for RNA-seq using the PrimeScript RT reagent Kit (Takara, Japan), and then qPCR, repeated three times, was performed on Bio-Rad CFX96 thermocycler using the 2 × SYBR green PCR master mix (Applied Biosystems). Seven DEGs were selected to verify the accuracy of the transcriptome results by qPCR. A housekeeping gene *Csa6M484600* encoding ACTIN was used as the internal control (Wilson et al., 2015). The primers for qPCR listed in **Supplementary Table S3** were designed using Primer 6 software. The relative expression of each gene was calculated using the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001). Each expression profile was independently verified in three replicated experiments performed under identical conditions. Primer sequences used are listed in **Supplementary Table S3**.

Tissue-Specific Expression

In order to study the gene-specific expression of *CsMYB* genes in different tissues of cucumber, the serial number PRJNA80169 was used from NCBI to obtain cucumber RNA samples from

[5]http://cello.life.nctu.edu.tw/

[6]www.plant.osakafu-u.ac.jp/~kagiana/gcorn/p

[7]http://bioinformatics.psb.ugent.be/webtools/plantcare/html/

[8]http://gsds.cbi.pku.edu.cn/

[9]https://www.ncbi.nlm.nih.gov/sra/

TABLE 1 | The characteristics of *CsMYB* -encoded proteins in cucumber.

Rename	gene ID	Chr	length	MV (Da)	pI	Predicted location	Classification
Cs1RMYB-1	Csa1G004120	1	89	10203.8	9.08	Extracellular	MYB_related
Cs1RMYB-2	Csa1G024160	1	248	27448.6	4.89	Nuclear	MYB_related
Cs1RMYB-3	Csa1G033200	1	280	31195.2	5.57	Nuclear	MYB_related
Cs1RMYB-4	Csa1G042350	1	252	27974.9	5.29	Nuclear	MYB_related
Cs1RMYB-5	Csa1G071840	1	313	34927.7	5.17	Nuclear	MYB_related
Cs1RMYB-6	Csa1G420880	1	299	33745.2	5.5	Nuclear	MYB_related
Cs1RMYB-7	Csa1G488750	1	103	11966.7	10.01	Chloroplast	MYB_related
Cs1RMYB-8	Csa2G007440	2	164	18840.4	4.34	Nuclear	MYB_related
Cs1RMYB-9	Csa2G035350	2	305	34400.3	7.44	Nuclear	MYB_related
Cs1RMYB-10	Csa2G049890	2	117	13463.8	9.42	Plasma Membrane	MYB_related
Cs1RMYB-11	Csa2G169770	2	160	18460.9	6.6	Cytoplasmic	MYB_related
Cs1RMYB-12	Csa2G229940	2	240	28644.2	10.61	Nuclear	MYB_related
Cs1RMYB-13	Csa2G229950	2	62	6881.8	8.55	Nuclear	MYB_related
Cs1RMYB-14	Csa2G352940	2	90	10047.6	9.89	Nuclear	MYB_related
Cs1RMYB-15	Csa2G370470	2	660	70029.6	8.11	Nuclear	MYB_related
Cs1RMYB-16	Csa3G045200	3	59	6782.7	9.69	Nuclear	MYB_related
Cs1RMYB-17	Csa3G122510	3	95	10392.6	5.2	Nuclear	MYB_related
Cs1RMYB-18	Csa3G134010	3	109	12886.9	10.84	Mitochondrial	MYB_related
Cs1RMYB-19	Csa3G135050	3	100	11579.5	10.09	Nuclear	MYB_related
Cs1RMYB-20	Csa3G165690	3	93	10737.2	9.34	Nuclear	MYB_related
Cs1RMYB-21	Csa3G740800	3	116	13176.1	9.07	Nuclear	MYB_related
Cs1RMYB-22	Csa3G742860	3	205	24306.1	6.94	Nuclear	MYB_related
Cs1RMYB-23	Csa3G828990	3	109	12185.8	9.32	Nuclear	MYB_related
Cs1RMYB-24	Csa3G848130	3	280	30712.4	9.85	Nuclear	MYB_related
Cs1RMYB-25	Csa3G891670	3	386	42250.2	8.67	Nuclear	MYB_related
Cs1RMYB-26	Csa4G000700	4	329	36786.4	8.62	Cytoplasmic	MYB_related
Cs1RMYB-27	Csa4G290190	4	674	76072.4	10.06	Nuclear	MYB_related
Cs1RMYB-28	Csa4G645320	4	51	5651.51	8.54	Nuclear	MYB_related
Cs1RMYB-29	Csa5G171690	5	109	12436.6	9.67	Mitochondrial	MYB_related
Cs1RMYB-30	Csa5G524690	5	158	17462.2	4.87	Nuclear	MYB_related
Cs1RMYB-31	Csa5G524720	5	83	9198.4	5.54	Plasma Membrane	MYB_related
Cs1RMYB-32	Csa6G105150	6	68	8080.3	11.56	Nuclear	MYB_related
Cs1RMYB-33	Csa6G136580	6	275	31613.5	4.44	Nuclear	MYB_related
Cs1RMYB-34	Csa6G303240	6	191	22007.6	4.85	Nuclear	MYB_related
Cs1RMYB-35	Csa6G311520	6	107	12124.7	9.75	Nuclear	MYB_related
Cs1RMYB-36	Csa6G495710	6	307	33757.5	10.25	Cytoplasmic	MYB_related
Cs1RMYB-37	Csa6G496960	6	56	6084	8.75	Cytoplasmic	MYB_related
Cs1RMYB-38	Csa6G499890	6	315	35014.2	6.03	Nuclear	MYB_related
Cs1RMYB-39	Csa7G179630	7	559	63088.6	6.91	Nuclear	MYB_related
Cs2RMYB-1	Csa1G004900	1	472	54883.7	6.3	Nuclear	R2R3-MYB
Cs2RMYB-2	Csa1G008430	1	368	41942.7	6.19	Nuclear	R2R3-MYB
Cs2RMYB-3	Csa1G009700	1	332	37603.3	9.14	Nuclear	R2R3-MYB
Cs2RMYB-4	Csa1G021940	1	308	33313	7.41	Nuclear	R2R3-MYB
Cs2RMYB-5	Csa1G029570	1	233	26589.2	8.84	Nuclear	R2R3-MYB
Cs2RMYB-6	Csa1G046820	1	378	42093.5	6.92	Nuclear	R2R3-MYB
Cs2RMYB-7	Csa1G109320	1	319	36229	8.2	Nuclear	R2R3-MYB
Cs2RMYB-8	Csa1G529120	1	432	48469.4	5.67	Nuclear	R2R3-MYB
Cs2RMYB-9	Csa1G561370	1	342	37757.4	5.94	Nuclear	R2R3-MYB
Cs2RMYB-10	Csa1G575180	1	249	27776.7	6.73	Nuclear	R2R3-MYB
Cs2RMYB-11	Csa1G682630	1	517	59414.2	6.97	Nuclear	R2R3-MYB
Cs2RMYB-12	Csa2G100550	2	265	29982.6	8.85	Nuclear	R2R3-MYB
Cs2RMYB-13	Csa2G174670	2	108	12650.5	10.2	Nuclear	R2R3-MYB
Cs2RMYB-14	Csa2G270220	2	232	26160.6	7.61	Nuclear	R2R3-MYB

(Continued)

TABLE 1 | Continued

Rename	gene ID	Chr	length	MV (Da)	pI	Predicted location	Classification
Cs2RMYB-15	Csa2G302300	2	211	23544.3	10.95	Nuclear	R2R3-MYB
Cs2RMYB-16	Csa2G352410	2	205	23037	8.29	Nuclear	R2R3-MYB
Cs2RMYB-17	Csa2G355030	2	291	31804.4	9.57	Nuclear	R2R3-MYB
Cs2RMYB-18	Csa2G360620	2	243	27199.1	9.27	Nuclear	R2R3-MYB
Cs2RMYB-19	Csa2G427310	2	302	32296.8	8.13	Nuclear	R2R3-MYB
Cs2RMYB-20	Csa2G439230	2	300	33891.8	6.37	Nuclear	R2R3-MYB
Cs2RMYB-21	Csa3G076520	3	520	57033.1	7.78	Nuclear	R2R3-MYB
Cs2RMYB-22	Csa3G168940	3	306	33693.1	5.69	Nuclear	R2R3-MYB
Cs2RMYB-23	Csa3G182040	3	300	34029.5	10.12	Nuclear	R2R3-MYB
Cs2RMYB-24	Csa3G199590	3	220	25251.4	6.78	Nuclear	R2R3-MYB
Cs2RMYB-25	Csa3G264750	3	355	40501.9	9.6	Nuclear	R2R3-MYB
Cs2RMYB-26	Csa3G303630	3	268	30096.4	5.81	Nuclear	R2R3-MYB
Cs2RMYB-27	Csa3G386830	3	170	19348.4	4.48	Nuclear	R2R3-MYB
Cs2RMYB-28	Csa3G535090	3	225	25603.6	7.86	Nuclear	R2R3-MYB
Cs2RMYB-29	Csa3G592130	3	220	24887.1	6.61	Nuclear	R2R3-MYB
Cs2RMYB-30	Csa3G732480	3	269	30010.8	9.76	Extracellular	R2R3-MYB
Cs2RMYB-31	Csa3G812750	3	349	39131.2	9.07	Nuclear	R2R3-MYB
Cs2RMYB-32	Csa3G816030	3	295	32662.3	6.53	Nuclear	R2R3-MYB
Cs2RMYB-33	Csa3G824850	3	234	26130.2	9.77	Nuclear	R2R3-MYB
Cs2RMYB-34	Csa3G826690	3	314	35125.9	6.35	Nuclear	R2R3-MYB
Cs2RMYB-35	Csa3G901090	3	105	12050.7	8.76	Nuclear	R2R3-MYB
Cs2RMYB-36	Csa3G914600	3	254	28517.6	9.38	Nuclear	R2R3-MYB
Cs2RMYB-37	Csa4G022940	4	315	36057.3	6.51	Nuclear	R2R3-MYB
Cs2RMYB-38	Csa4G305350	4	378	40500	5.31	Nuclear	R2R3-MYB
Cs2RMYB-39	Csa4G308500	4	300	34358.2	6.81	Nuclear	R2R3-MYB
Cs2RMYB-40	Csa4G638510	4	230	25192.6	9.56	Nuclear	R2R3-MYB
Cs2RMYB-41	Csa4G641690	4	264	29140.4	8.35	Nuclear	R2R3-MYB
Cs2RMYB-42	Csa5G148680	5	261	28837.7	4.74	Nuclear	R2R3-MYB
Cs2RMYB-43	Csa5G152790	5	324	36414.6	6.44	Nuclear	R2R3-MYB
Cs2RMYB-44	Csa5G160120	5	259	29457.7	6.17	Nuclear	R2R3-MYB
Cs2RMYB-45	Csa5G198240	5	296	33799.6	8.84	Mitochondrial	R2R3-MYB
Cs2RMYB-46	Csa5G209490	5	285	31234.3	6.88	Nuclear	R2R3-MYB
Cs2RMYB-47	Csa5G353650	5	296	31215.5	6.92	Nuclear	R2R3-MYB
Cs2RMYB-48	Csa5G425400	5	301	33770.6	5.11	Nuclear	R2R3-MYB
Cs2RMYB-49	Csa5G579040	5	287	30876.7	6.17	Nuclear	R2R3-MYB
Cs2RMYB-50	Csa5G604410	5	465	52285.8	7.27	Nuclear	R2R3-MYB
Cs2RMYB-51	Csa5G605730	5	370	41397.2	7.08	Nuclear	R2R3-MYB
Cs2RMYB-52	Csa5G637660	5	399	44517.7	6.99	Nuclear	R2R3-MYB
Cs2RMYB-53	Csa5G641610	5	231	27272.2	9.77	Nuclear	R2R3-MYB
Cs2RMYB-54	Csa5G651640	5	295	33129	8.51	Nuclear	R2R3-MYB
Cs2RMYB-55	Csa6G040640	6	287	32771	6.87	Nuclear	R2R3-MYB
Cs2RMYB-56	Csa6G046240	6	234	26439.4	8.87	Nuclear	R2R3-MYB
Cs2RMYB-57	Csa6G046460	6	149	16821.1	7.68	Cytoplasmic	R2R3-MYB
Cs2RMYB-58	Csa6G121970	6	270	31333.9	6.92	Nuclear	R2R3-MYB
Cs2RMYB-59	Csa6G187960	6	316	35118.2	9.91	Nuclear	R2R3-MYB
Cs2RMYB-60	Csa6G344210	6	310	35826.2	9.56	Nuclear	R2R3-MYB
Cs2RMYB-61	Csa6G491690	6	302	32449.4	9.71	Nuclear	R2R3-MYB
Cs2RMYB-62	Csa6G495660	6	240	26841.1	10.28	Nuclear	R2R3-MYB
Cs2RMYB-63	Csa6G538700	6	290	33192	6.95	Nuclear	R2R3-MYB
Cs2RMYB-64	Csa7G043580	7	501	54622.8	6.31	Nuclear	R2R3-MYB
Cs2RMYB-65	Csa7G045590	7	255	29446.4	5.21	Nuclear	R2R3-MYB
Cs2RMYB-66	Csa7G046120	7	338	37544.7	6.39	Nuclear	R2R3-MYB
Cs2RMYB-67	Csa7G375780	7	255	29393.9	4.78	Nuclear	R2R3-MYB

(Continued)

TABLE 1 | Continued

Rename	gene ID	Chr	length	MV (Da)	pI	Predicted location	Classification
Cs2RMYB-68	Csa7G413890	7	277	31884.7	4.97	Nuclear	R2R3-MYB
Cs2RMYB-69	Csa7G431360	7	1503	169045	4.87	Nuclear	R2R3-MYB
Cs3RMYB-1	Csa2G375240	2	1000	111544	5.54	Nuclear	R1R2R3-MYB
Cs3RMYB-2	Csa3G734230	3	551	61163.7	8.53	Nuclear	R1R2R3-MYB
Cs3RMYB-3	Csa3G889150	3	961	105806	5.3	Nuclear	R1R2R3-MYB
Cs4RMYB	Csa3G113280	3	1005	113684	7.06	Nuclear	4RMYB

different tissues and organs (leaves, stems, female flowers, male flowers, enlarged ovary (unfertilized), dilated ovary (fertilized), ovary, root, RNA-Seq data for tendrils and the base of tendrils (Harris et al., 2008).

RESULTS

Identification of the Cucumber *CsMYB* Gene Family

A cucumber-specific model (**Supplementary Figure S1**) was built for typical MYB or MYB-like domains using the HMM profile (PF00249) of the MYB-conserved domain. A total of 197 *CsMYB* genes were selected from the cucumber genome database[10]. Combining the CDD and SMART databases using manual calibration, 215 genes were identified for the *CsMYB* gene family, further divided into the four subclasses according to the number of repeats of the conserved protein domain. Predicted CsMYB proteins (listed in **Table 1**) showed considerable redundancy, e.g., 33, 67, and 3 redundant proteins in the group MYB-related, R2R3-MYB, and R1R2R3-MYB, respectively. The number of unique proteins is 112, corresponding to the actual number of 112 genes.

Predicted physicochemical properties of the *CsMYB* gene family proteins, including molecular weight (M_W), isoelectric point (pI) and length (aa), were analyzed through ExPASy online. *CsMYB* genes encoded proteins ranging from 56 to 1503 aa in length, pI values ranging from 4.34 to 11.56 and molecular weights from 6084 to 169044.6 Da. Protein subcellular location was predicted through CELLO program (v.2.5). It was found that the vast majority genes (89.72%) encoded proteins that were predicted to be intranuclear, while others were predicted to be located in mitochondria, chloroplasts, cytoplasm, and the plasma membrane (**Table 1**). Members of the *CsMYB* gene family were then renamed according to their classification (subclasses 1–4) and their chromosome location (*Cs1RMYB-1* to *Cs1RMYB-39*, *Cs2RMYB-1* to *Cs2RMYB-69*, *Cs3RMYB-1* to *Cs3RMYB-3*, and *Cs4RMYB*).

Evolutionary Relationships, Gene Structure and Conserved Motif Analyses of the *CsMYB* Gene Family in Cucumber

Clustal V was applied to generate multiple sequence alignment of the conserved structural domains of all cucumber *CsMYB* genes (**Figure 1**). Results were used to construct a phylogenetic tree (**Figure 2**), illustrating distribution and extent of conserved structural elements (**Figure 3**, **Supplementary Table S1**, and **Supplementary Figure S2**). As shown in **Figure 3**, the number of exons in *CsMYB* gene family members varied from 1 to 19 (18 with one exon, 34 with two exons, 41 with three exons, 6 with four exons, 1 with five exons, 3 with six exons, 3 with seven exons, 3 with eleven exons, 1 with twelve exons, 1 with thirteen exons and 1 with nineteen exons). Intron-exon structure of *CsMYB* genes was also analyzed after exton identification (**Supplementary Figure S4**). The results showed that exon 1 (133 bp) and exon 2 (130 bp) appeared to be more consistent in length, while exon 3 was more variable (31–850 bp). Phylogenetic analysis of the predicted CsMYB protein sequences revealed differences between cucumber and *Arabidopsis*, with eight subgroups in common. Two cucumber subgroups were missing from *Arabidopsis*; ten sub-groups from Arabidopsis were absent in the cucumber genome, indicating *CsMYB* genes may evolve or be lost in a given taxon, following divergence. Some cucumber CsMYB proteins were clustered into *Arabidopsis* functional clades, which provided an excellent reference to explore the putative functions of the cucumber *CsMYB* genes. The comparative evolutionary analysis between *Arabidopsis* and cucumber suggests both species are derived from a common ancestor, no longer extant, and represent independent paths defined by duplication, divergence and loss of gene family members (**Figure 4**).

The obvious redundancy of *CsMYB* gene at the level of amino acid sequence is partly due to the similar molecular functions of *CsMYB* gene, but different biological phenotypes can still be displayed by these loci when mutated, due to different temporal or spatial expression characteristics. Studies have shown that homologous members of this gene family are not redundant in terms of plant development (Jin and Martin, 1999).

Distribution and Collinearity Analysis of *CsMYB* Genes in Cucumber

Members of the CsMYB gene family were mapped to all seven chromosomes with Mapchart (**Figure 5**). Results showed that chromosome 3 had the largest number of gene family members, 29 genes, while chromosome 7 had the fewest number of genes, containing only seven. The other five chromosomes 1, 2, 4, 5 and 6 each contained 18, 18, 8, 16, and 16 gene family members, respectively. The chromosome distribution of *CsMYB* gene family members is non-regular and does not correlate with overall chromosome length.

[10]http://cucurbitgenomics.org/

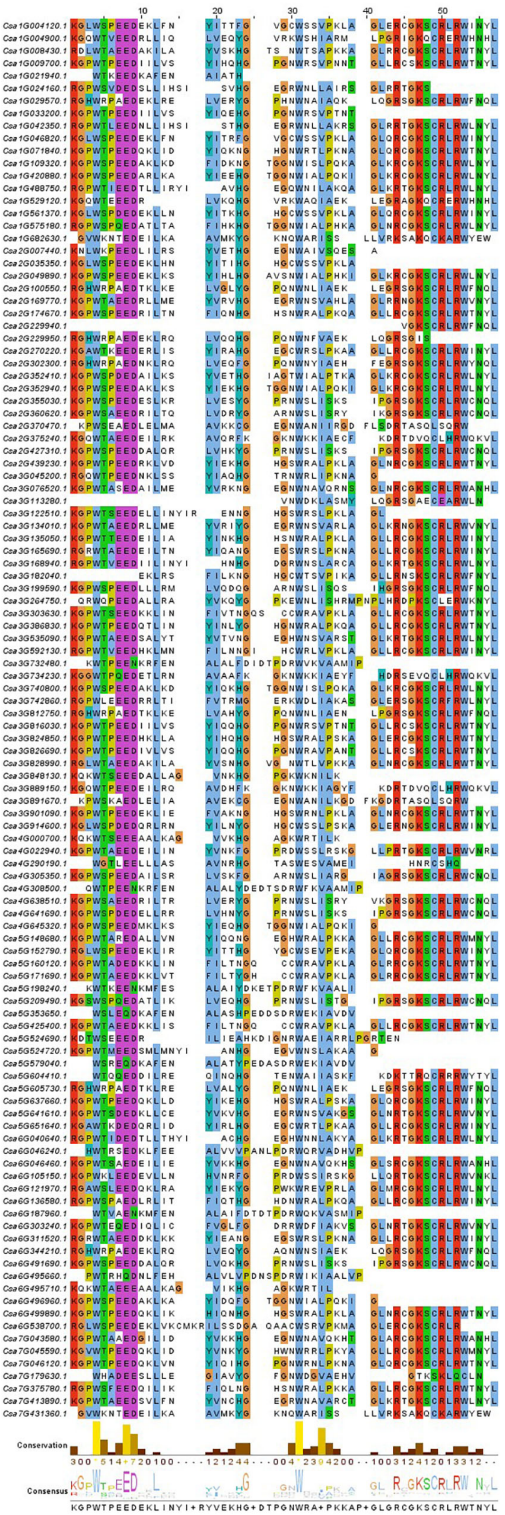


FIGURE 1 | Alignment, degree of conservation and consensus of the conserved domain of *CsMYB* genes. The sequence logos of the cucumber special model was built based on full-length alignments of the MYB-conserved domain. The overall height of each stack indicates the conservation of the sequence at that position, whereas the height of letters within each stack represents the relative frequency of the corresponding amino acid.

To further analyze divergence of *CsMYB* gene family members in cucumber, the tandem repeat events and fragment repeat events of the gene family were analyzed by using KAKS_Calculator2.0. In this analysis, if two or more genes are present within 200 kb, the elements are considered a tandem repeat event. In total, 23 *CsMYB* genes were involved in 11 tandem repeats as follows: (*Csa1G008430.1* & *Csa1G009700.1*, *Csa1G021940.1* & *Csa1G024160.1*, *Csa2G229940.1* & *Csa2G229950.1*, *Csa2G352410.1* & *Csa2G352940.1*, *Csa3G740800.1* & *Csa3G742860.1*, *Csa4G305350.1* & *Csa4G308500.1*, *Csa5G524690.1* & *Csa5G524720.1*, *Csa5G604410.1* & *Csa5G605730.1*, *Csa6G046240.1* & *Csa6G046460.1*, *Csa6G495660.1* & *Csa6G495710.1* & *Csa6G496960.1*, *Csa7G045590.1* & *Csa7G046120.1*). Gene family fragment duplication events were determined by BLASTP, which showed that only one pair of genes (*Csa3G740800.1*, *Csa2G352940.1*) had experienced fragment duplication. Based on the results above, it is inferred that tandem repeat events play an important role in the amplification of gene families rather than duplication of gene fragments.

In order to uncover the linkage relationships of *CsMYB* genes throughout the cucumber genome, collinearity analysis was carried out between cucumber and *Arabidopsis thaliana* MYB genes (**Figure 6**) showing that high homology of the MYB gene family existed between the two species. Twenty-five pairs of genes showed collinearity and conserved linkage relationships (**Supplementary Figure S3**).

Analysis of *cis*-Elements of *CsMYB* Genes in Cucumber

In order to study associated *cis*-elements of *CsMYB* promoters, the sequences (1.5 kb) upstream of *CsMYB* coding sequence were selected from cucumber genome, and then were submitted to PlantCARE for further identification of regulatory elements. Using the online site Genetic Structure Display Server (GSDS), visualization all regulatory elements was conducted involving TF binding sites in defense and stress responses or defense responses (**Figure 7**). Three genes, *Csa6G538700, Csa1G021940*, and *Csa5G641610* which had shown higher expression level in responses of cucumber to *M. incognita*, showed specific differences in the upstream *cis*-elements associated with the homolog in question. All three genes show associated upstream sequence that contains the *cis*-element G-box (TACGTG) and *cis*-element ABRE (ACGTG). The former sequence is associated with light-regulated gene expression; the latter is involved in regulation of gene expression by abscisic acid.

Expression Analysis of MYB Gene Detected From Transcriptome

In order to analyze tissue-specific expression of the *CsMYB* gene family in cucumber, expression levels of gene family members were assayed in 10 tissues and organs obtained by PRJNA80169 (**Figure 8**). Results showed that most of the genes were expressed in all the different tissues sampled, with similar level of expression between tissues. Some genes, such as *Csa6G495710*, *Csa3G848130*, *Csa2G355030*, *Csa4G308500*,

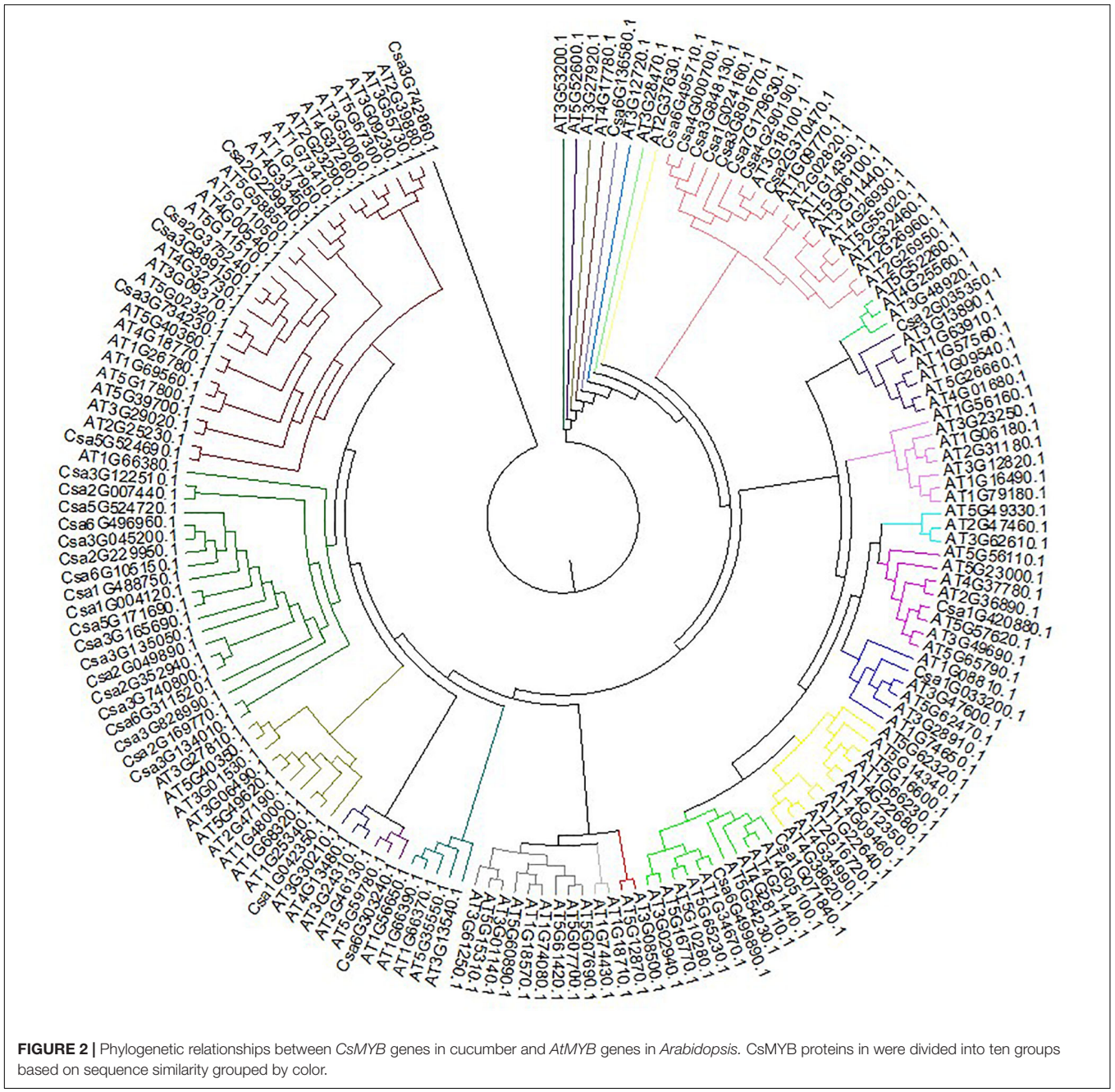


FIGURE 2 | Phylogenetic relationships between *CsMYB* genes in cucumber and *AtMYB* genes in *Arabidopsis*. CsMYB proteins in were divided into ten groups based on sequence similarity grouped by color.

Csa5G198240, *Csa5G579040*, *Csa6G491690*, and *Csa5G353650*, were highly expressed in all 10 tissues and organs. Expression of *Csa3G076520*, *Csa2G229940*, *Csa2G229950*, *Csa1G529120*, and *Csa1G008430* were low in almost all tissues. *Csa6G303240*, *Csa5G524720*, *Csa5G160120*, *Csa6G136580*, *Csa3G168940*, *Csa6G496960*, *Csa2G007440*, *Csa3G824850*, *Csa1G029570*, *Csa1G488750*, *Csa2G169770*, *Csa5G152790*, *Csa5G605730*, *Csa3G740800*, *Csa6G344210*, *Csa6G495660*, *Csa2G049890*, *Csa6G499890*, *Csa6G121970*, *Csa3G134010*, *Csa1G420880*, *Csa6G538700*, *Csa4G645320*, and *Csa2G3522410* (21.4%) were only highly expressed in the root, showing lower levels of expression in other tissues and organs.

The post-RKN-inoculation transcriptome data extracted from resistant IL10-1 and susceptible CC3 lines were acquired using SRP125669 to ascertain whether members of the *CsMYB* gene family were subject to altered regulation during susceptible or resistant response to RKN (**Figure 9**). Six genes were randomly picked to validate the RNA-seq data. Additionally, to gain more insight into the role of *CsMYB* genes in the response to RKN in cucumber, relative real-time qPCR was performed to evaluate the transcript abundance of these *CsMYB* genes previously identified as differentially regulated during the resistant response to *M. incognita* (**Supplementary Figure S5**), which showed the similar expression trends with their RNA-seq data, respectively.



FIGURE 3 | Conserved protein motifs and gene structure of *CsMYB* genes in cucumber. The left column shows conserved motifs in all the proteins encoded by *CsMYB* gene. There are 10 different color boxes with numbers 1–10 indicted 10 subclasses defined by conserved motifs. The right column shows exons, introns, and UTRs (untranslated regions) structures with yellow boxes, single lines and green boxes, respectively. Introns phases 0, 1 and 2 are indicated by numbers 0, 1 and 2, respectively. The length of each *CsMYB* gene can be estimated using the scale at the bottom.

The homolog *Csa6G303240* was highly expressed in both resistant and susceptible cucumber genotypes, but at different time points, suggesting that expression of this gene was not specifically correlated with resistance to RKN in cucumber. Further comparative analysis of the transcriptomes from RKN-resistant and susceptible genotypes revealed that the MYB gene family member, *Csa5G641610*, shown to affect expression of the core gene in resistance to RKN, *LTP*, gradually decreased in CC3, while in IL10-1, transcript levels initially increased, then decreased. The expression level of *Csa5G641610* was analyzed, similar to the expression patterns of *Csa1G021940*, and *Csa6G538700*. Thus, it is speculated that *Csa6G538700*, *Csa1G021940*, and *Csa5G641610* may regulate gene *LTP* to some extent, further affecting resistance to RKN in cucumber.

DISCUSSION

Organization and Characterization of the Cucumber *CsMYB* Gene Family

The MYB family is the largest TF gene family in plants and has been shown to play an important role in secondary metabolite formation, sexual differentiation, cell differentiation, plant morphogenesis, and biotic and abiotic stress (Jung et al., 2003; Hsu et al., 2005; Guo et al., 2013; Li et al., 2018). The MYB TFs are ubiquitous to all plant species and have been surveyed in several species with sequenced genomes, such as *Arabidopsis* (Riechmann et al., 2000), *grape* (Matus et al., 2008), *poplar* (Wilkins et al., 2009), *soybean* (Du et al., 2012), *maize* (Dias et al., 2003). Additionally, 55 members of the cucumber *R2R3MYB* gene family were identified and characterized from annotated genes in the 26682 cucumber genome (Li et al., 2012). In the current study, 112 *CsMYB* genes were identified in the last and previous versions of the genome assembly of the cucumber "Chinese Long" (v3). Compared to the draft genome sequence of *Cucumis sativus* var. *sativus* L. used by Li et al. (2012), the reference genome sequence has been improved, and therefore, our 112 *CsMYB* genes family, included those 55 *R2R3MYB* genes in which 36 genes show one to one correspondence (**Supplementary Table S2**).

The 112 *CsMYB* genes family was consist of 39 1RMYB, 69 R2R3-MYB, 3 R1R2R3-MYB and 1 4RMYB according to the number and type of MYB domains. All the details of the gene structure, conserved motifs, and *cis*-element were analyzed and illustrated. Structural analysis found that the lengths of exons 1 and 2 are very conserved, similar to previous results reported for cucumber *R2R3MYB* work (Li et al., 2012). However, in our study, exon 1 and 2 were not as highly conserved, perhaps because we investigated all the MYB genes instead of just the subclass, *R2R3MYB*, and the other subclasses of MYB genes increased the variation in structural conservation. All 112 *CsMYB* genes have been located on seven chromosomes

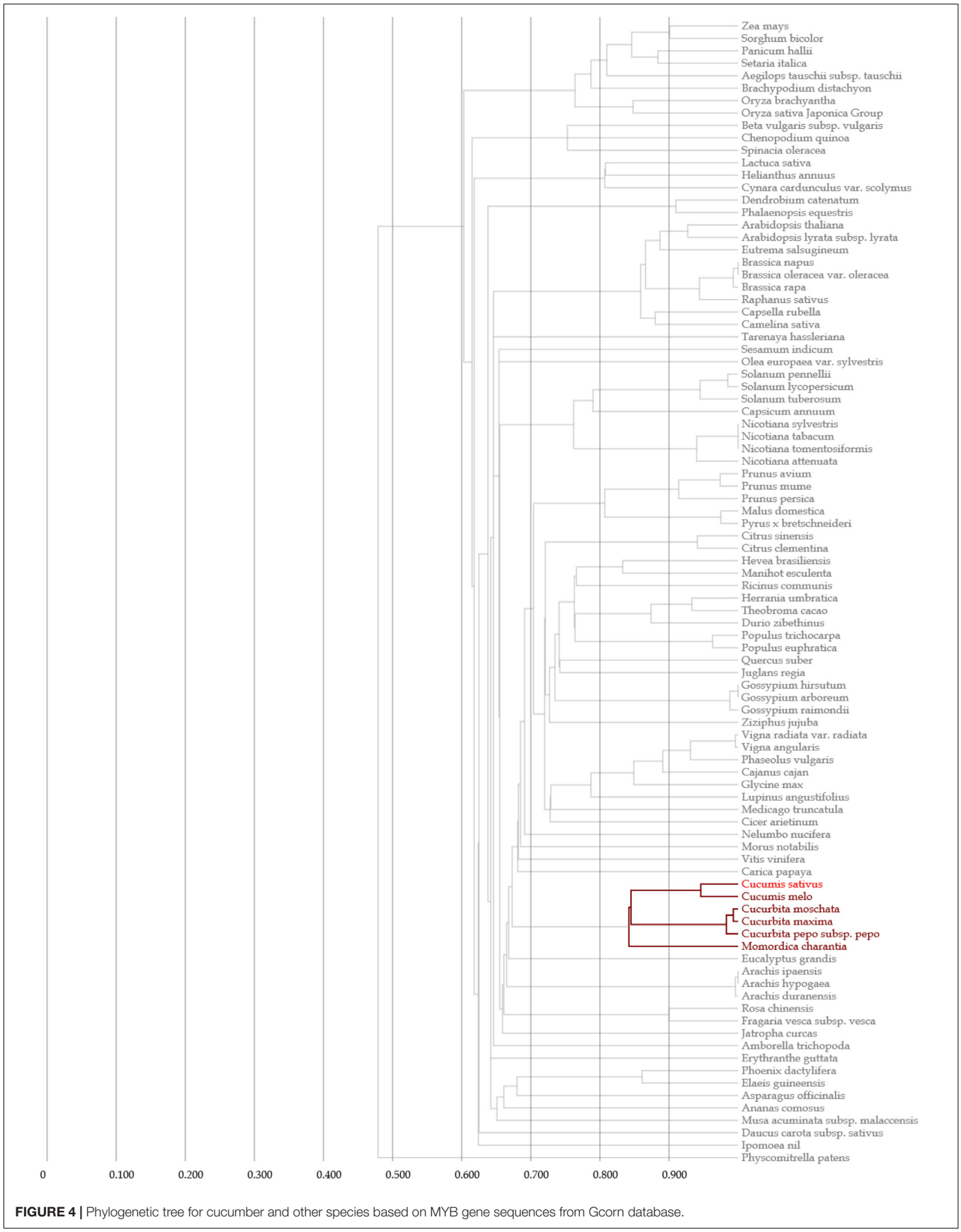


FIGURE 4 | Phylogenetic tree for cucumber and other species based on MYB gene sequences from Gcorn database.

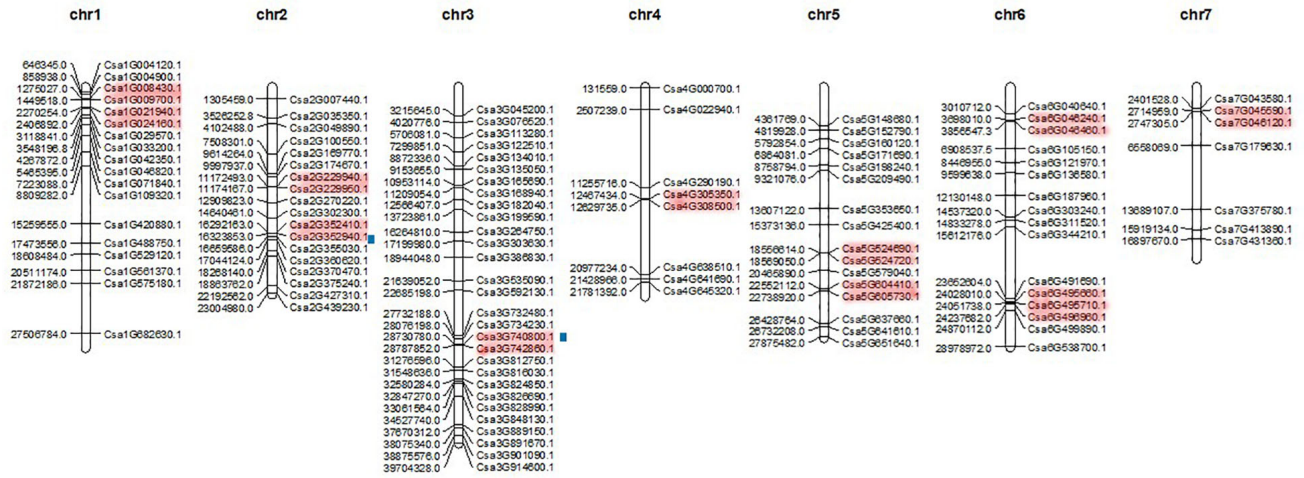


FIGURE 5 | The chromosomal distribution of *CsMYB* genes positions. Chromosomal positions of all the *CsMYB* genes are indicated by the full name of the genes (all the genes assigned in **Table 1**). The red color regions in chromosomes indicate tandem duplication; Blue squares along chromosome indicate segmental duplication.

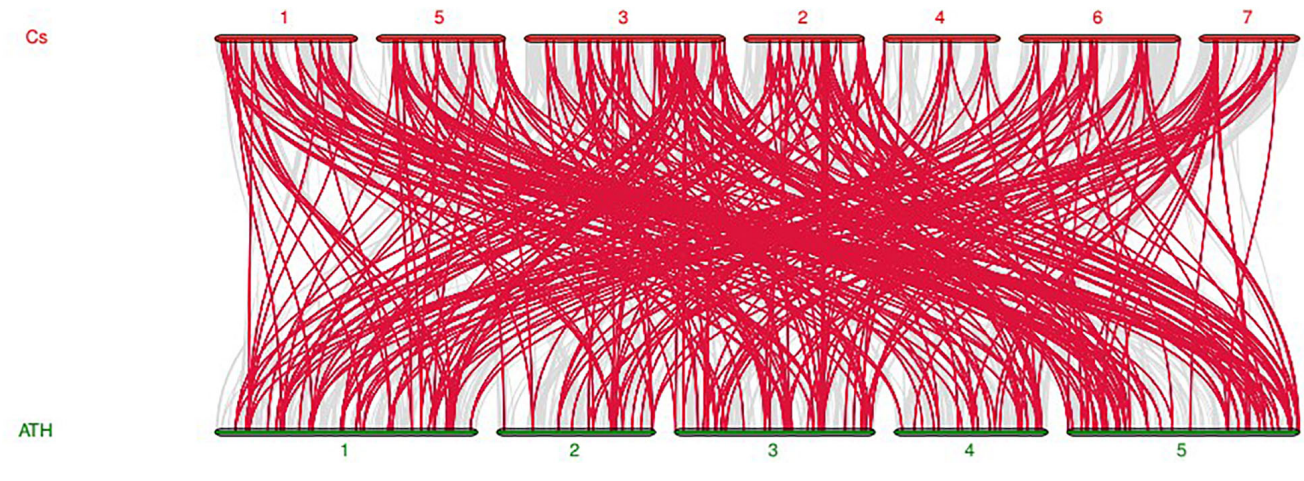


FIGURE 6 | Synteny analysis of *MYB* genes between cucumber and *Arabidopsis thaliana.* The red lines indicate *MYB* collinear genes, and the gray indicated the collinear genes of both species. Cs represents cucumber, while ATH represents *Arabidopsis thaliana*.

with uneven distribution. The largest number of homologs are found chromosome 3 and the smallest number on chromosome 7 (**Figure 5**). The locations are generally consistent with the more limited results in a previous study (Li et al., 2012). It is interesting to note that after the correspondence was identified between the published MYB gene (*R2R3-MYB*_v1) (Li et al., 2012) and *CsMYB*_v2 (**Supplementary Table S2**), two inversions were found in chromosome 5 and 7, (*Csa 5G148680.1*, *Csa5G152790.1* and *Csa5G160120.1* on chr.5, *Csa7G046120.1* and *Csa7G045590.1* on chr.7), compared with the chromosomal locations of Li et al. (2012). These chromosomal inversions might due to the differences between the improved cucumber genome and the old version released in 2009, which been reported (Lou et al., 2013).

Although whole genome duplication events occur commonly in many angiosperms, especially the recent gene duplications which are crucial for the rapid expansion and evolution of gene families (Cannon et al., 2004; Taylor and Raes, 2004), whole-genome duplication events and tandem duplications are rare in cucumber genome (Huang et al., 2009). In this study of the *CsMYB* gene family, tandem repeat events were detected more frequently than duplication of gene fragments in further analysis of the divergence of *CsMYB* gene family members, thus duplication may be the main driving force in the amplification of gene families.

Expression Analysis of *CsMYB* Genes Expressed in Response to *M. incognita*

Studies show that selection of favorable interactions between stress tolerant genes and their TFs can be an effective method to improve the comprehensive stress tolerance (Sun et al., 2010).

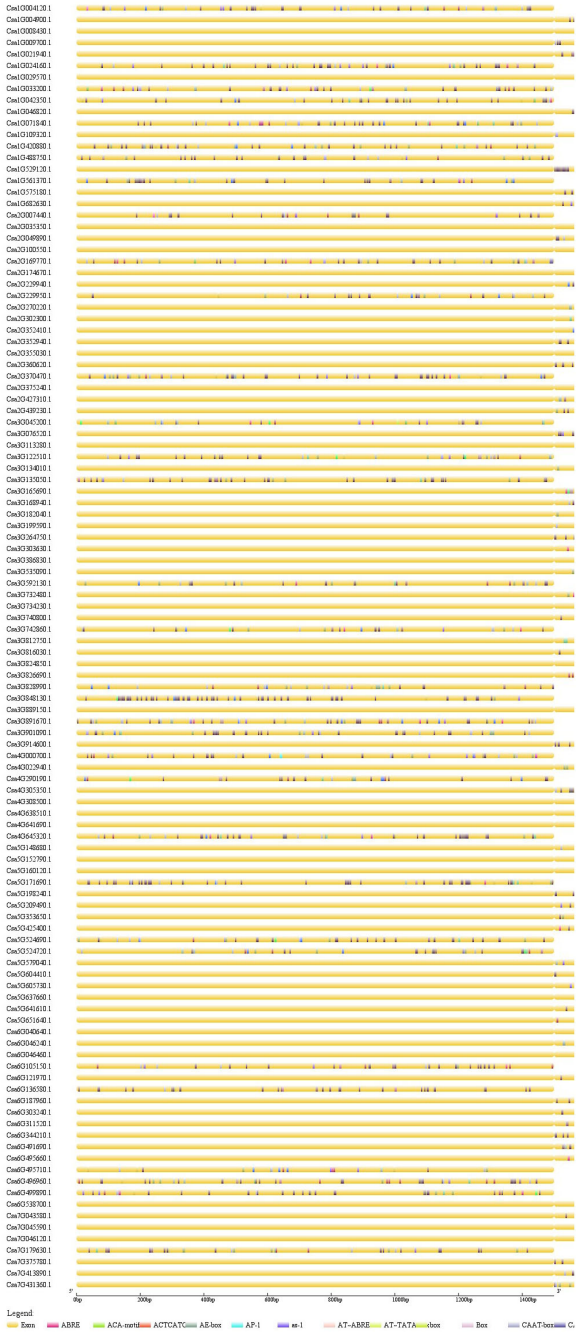


FIGURE 7 | Analysis of *cis*-elements in the upstream regions of *CsMYB* genes. Predicted *cis*-elements in CsMYB promoters. Promoter sequences (–1500 bp) of 112 *CsMYB* genes were analysis by PlantCARE for further identification of regulatory elements. The upstream length to the translation start site can be inferred according to the scale at the bottom.

The *MYBS1* gene of *Arabidopsis thaliana* is involved in salt stress of plants in response to various plant hormones (such as auxin, gibberellin, jasmonic acid, salicylic acid). In cucumber, the TF, *Csa1G021940*, is the homolog of *MYBS1*, and also responds to various phytohormones, suggesting that, similar to its homolog in Arabidopsis, this gene could condition a response to biological stress in cucumber, such as RKN infection (Gutierrez et al., 2009; Molinari et al., 2014; Zhao et al., 2018). In Arabidopsis, the *MYB67*TF (cucumber homolog, *Csa6G538700)* is involved in cell differentiation (Gheysen and Mitchum, 2011), thus in cucumber, perhaps the homolog may also affect the differentiation of giant cells which would hinders the development of *M. incognita*, thereby causing resistance to *M. incognita* in IL10-1. Finally, in Arabidopsis, *AT3G30210.1* (cucumber homolog, *Csa5G641610*) is a TF that responds to abscisic acid (ABA) which in cucumber, could inhibit cell elongation, affecting giant cell elongation, thereby enhancing the resistance to RKN. The results from our previous study of physiological observations of abnormal development of giant cells (GCs) between IL10-1 (resistant line) and CC3 (susceptible line) indicated that a series of reactions may occurred after the infection of RKN in the resistant line. Also, multi-omics study showed high levels of expression of the *LTP* gene which has been identified as the "hub gene" involved in lipid transport in GCs membranes and receives signals stimulated by nematodes (Wang et al., 2019).

In this study, the expression data for all 112 *CsMYB* genes in the two lines, IL10-1 and CC3, contrasting for RKN resistance were obtained directly from NCBI by SRP125669 (Wang et al., 2018) at 0, 1, 2 and 3 days after inoculation (dpi) with *M. incognita*. Further comparative analysis of the transcriptomes from IL10-1 (RKN-resistant) and CC3 (RKN-susceptible) revealed that three genes (*Csa5G641610*, *Csa1G021940*, and *Csa6G538700*) in *CsMYB* gene family, showed the same expression pattern where they initially increased, then decreased in IL10-1, but only decreased through time in the susceptible line, CC3. It further affected the expression of *LTP*, which assumed to be the core gene in response to RKN (Wang et al., 2018). In our previous comparative study on transcriptional events, combining with the physiological responses that occurs in resistant line IL10-1 and susceptible line CC3 during *M. incognita* infection, potential effector-targeted host genes were identified, while their regulation network was illuminated in the IL10-1 against *M. incognita*. Through the construction of a co-expression network, 2 out of 8 genes with the highest edges, considered the core genes of the network and assumed to have consequential roles in the resistant response against *M. incognita* perhaps as a result of abnormal development of GCs, were related to plant lipid transfer proteins (LTPs). One of the two most significantly induced *LTP* genes, which had the most edges, was considered as the hub gene in the resistance to RKN (Wang et al., 2018). More and more studies have indicated that LTPs are associated with plant resistance and defense (Thoma et al., 1994; Jung et al., 2003), for instance, LTPs can inhibit α-amylase, thereby contributing to resistance to biotic stress in wheat (Oda et al., 1997). However, reports on the relationships between LTPs and plant nematodes are still limited. As a result, further studies are necessary to elucidate the exact role of LTPs in plant resistance to nematodes. Studies in Arabidopsis show that the Myb-type DNA-binding transcriptional factor, MYB96, positively regulated the downstream gene *LTP3* to mediate freezing and drought stress (Guo et al., 2013). Whether there is a correlation between MYB

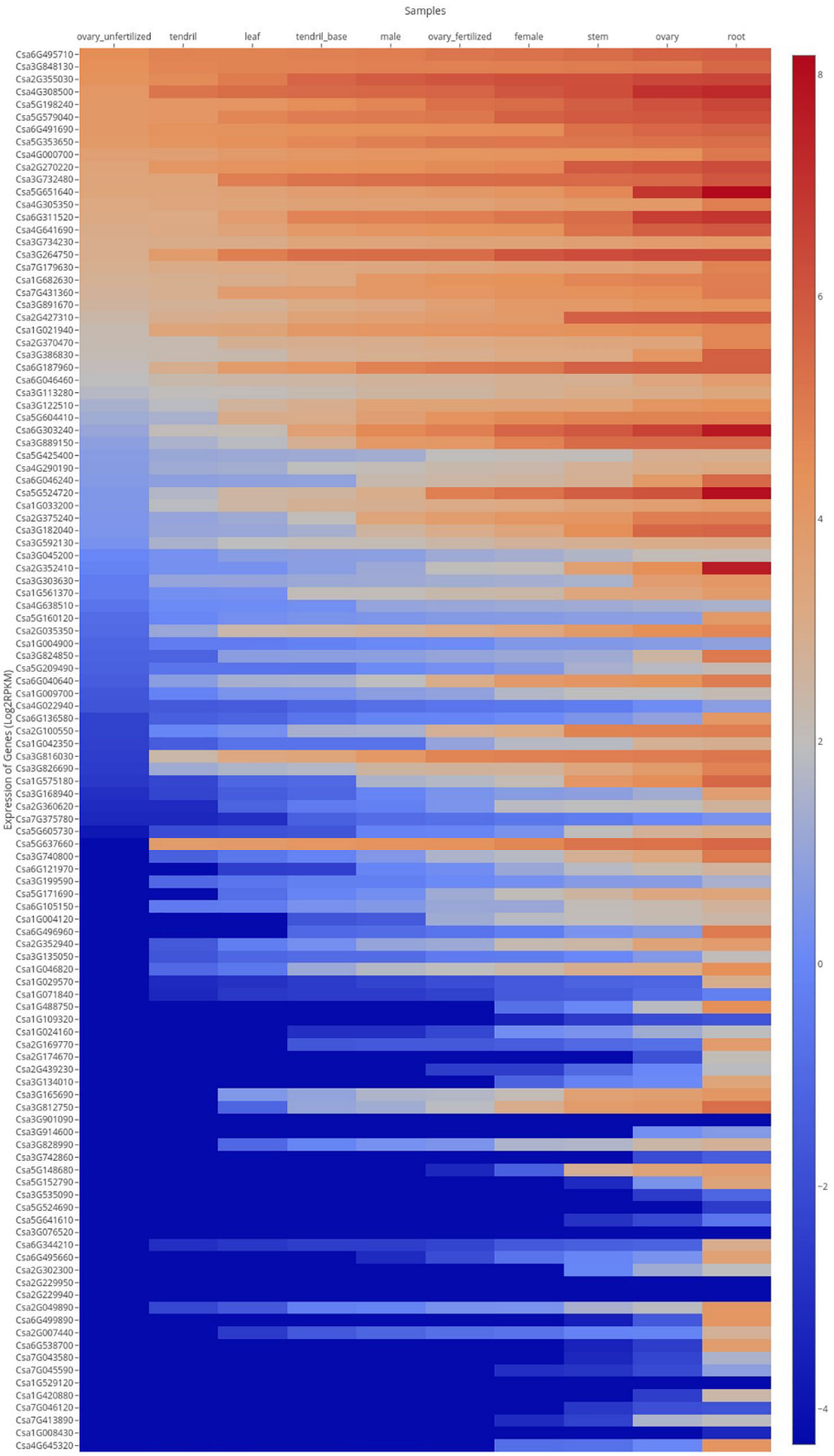


FIGURE 8 | Expression profiles of *CsMYB* genes in different tissues or organs. FPKM values of *CsMYB* genes were transformed by log2 and a heatmap was constructed using the cucumber genome database HEATMAP to draw the heatmap online. Red color indicates a high level. Blue color indicates a low level of gene expression.

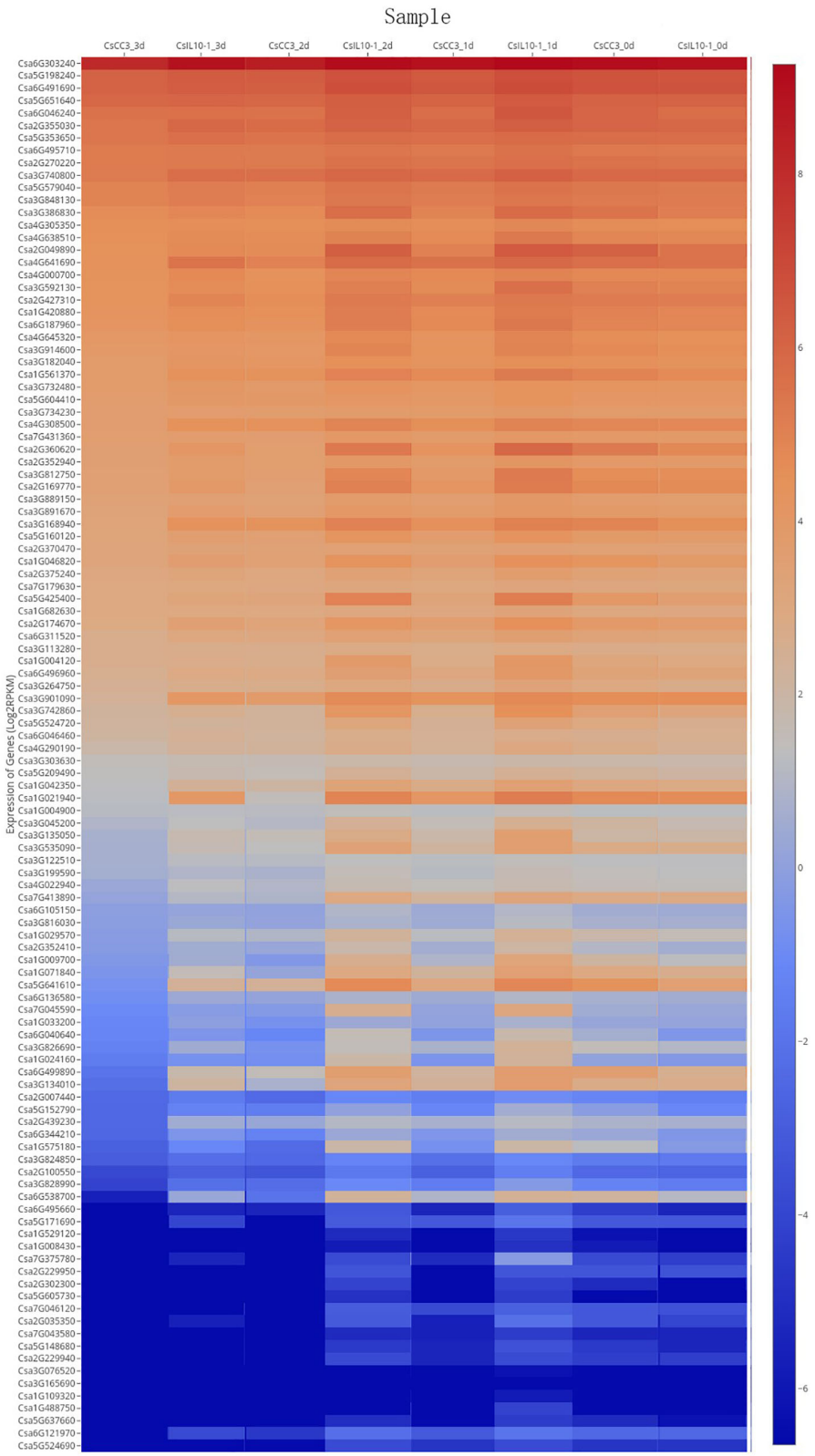


FIGURE 9 | Expression profile of *CsMYB* genes through time after infection with *M. incognita* in resistant and susceptible cucumber lines, IL10-1 and CC3. Expression profile of *CsMYB* genes in response to the inoculation of *M. incognita* on published data RNA-seq SRP125669 (Wang et al., 2018). The color scale represents relative gene expression levels.

TFs and LTP genes during plant-pathogen interaction process remains unknown.

CONCLUSION

In summary, this paper carried out the basic bioinformatics analysis including the characterization, the phylogenetic analysis and evolutionary analysis of the 112 *CsMYB* genes newly identified in cucumber. Additionally, comparative analysis of the transcriptomes discovered three *CsMYB* genes (*Csa5G641610*, *Csa1G021940*, and *Csa6G538700*) might positively regulate predicted hub gene *LTP* (*Csa6G410090*) in resistance to RKN in cucumber. The exact relationship between these three MYB TFs and the LTP gene during the plant-pathogen interactions, as well as the mechanisms how *CsMYB* genes regulate LTP to work will require further study.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at NCBI under the accession PRJNA80169, https://www.ncbi.nlm.nih.gov/search/all/?term=PRJNA80169.

AUTHOR CONTRIBUTIONS

JC and MJ designed and managed the project. QL, JL, and XW handled the samples and performed the experiments. QL and CC performed the data analyses and wrote the manuscript. All authors reviewed and approved this submission.

FUNDING

This work was supported by the Key Project of Natural Science Foundation of China (Nos. 31902007 and 31672168), National Key Research and Development Program of China (2018YFD1000804), the Fundamental Research Funds for the Central Universities of China (KJQN202030), and Jiangsu Agricultural Innovation of New Cultivars Innovation and application of cucumber cultivars with high quality and multi disease resistances (No. PZCZ201719).

ACKNOWLEDGMENTS

We thank Dr. Wan Hongjian (Associate researcher of Institute of Vegetable, Zhejiang Academy of Agricultural Sciences) for his insightful comments and pertinent suggestions to improve earlier version of the manuscript. We also thank Professor Wei Lihui (Researcher of Institute of Plant Protection, Jiangsu Academy of Agricultural Sciences) for providing the origin source of southern root-knot nematode.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.550677/full#supplementary-material

Supplementary Figure 1 | Special Hidden Markov Model (HMM) of *CsMYB* gene in cucumber. The special HMM of *CsMYB* gene in cucumber was built according to the MYB HMM (PF00249) from PFam with the parameter ($E \leq 1.2 \times 10^{-8}$). Then the logo of special model was visualizated online by SKylign.

Supplementary Figure 2 | Sequence logo of conserved Motif 1 to Motif 10.

Supplementary Figure 3 | Circos diagram of *CsMYB* genes in cucumber. Twenty-five pairs of *CsMYB* genes showed collinearity and linkage relationships in cucumber.

Supplementary Figure 4 | Exon length distribution analysis of the cucumber MYB genes. **(A)** Exon length values were analyzed using Boxplot. Each box represents the exon size range in which 50% of the values for a particular exon are grouped. The mean value is shown as a dotted line and the median as a continuous line. The number of exons in *CsMYB* gene family members varied from 1 to 19. **(B)** First, second and third exon lengths distribution of *CsMYB* genes using 3D Scatter Plot.

Supplementary Figure 5 | Relative RT-qPCR assay of six CsMYB family genes randomly picked from RNA-seq data in CC3 and IL10-1 at 0, 1, 2, 3 days post inoculation with *M. incognita*. **(A–F)** represented *Cs1RMYB-15*, *Cs2RMYB-7*, *Cs2RMYB-4*, *Cs2RMYB-65*, *Cs1RMYB-18*, and *Cs2RMYB-63*, respectively. The default expression value for each gene was Actin gene. The default expression value for each gene was from 0 dpi plants. Error bars indicate the standard deviation from the mean (three replicates).

Supplementary Table 1 | The sequence and characterize of motif.

Supplementary Table 2 | Correspondence between the published MYB gene (*R2R3-MYB*_v1) in cucumber (Li et al., 2012) and identified genes (*CsMYB*_v2) in cucumber.

Supplementary Table 3 | Primers Sequencs of the three random picked genes.

REFERENCES

Ampomah-Dwamena, C., Thrimawithana, A. H., Dejnoprat, S., Lewis, D., and Espley, R. V. (2019). A kiwifruit (Actinidia deliciosa) R2R3-MYB transcription factor modulates chlorophyll and carotenoid accumulation. *New Phytol.* 221, 309–325. doi: 10.1111/nph.15362

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: tools for motif discovery and searching. *Nucl. Acids Res.* 37, 202–208. doi: 10.1093/nar/gkp335

Cannon, S. B., Mitra, A., Baumgarten, A., Young, N. D., and May, G. (2004). The roles of segmental and tandem gene duplication in the evolution of large gene families in Arabidopsis thaliana. *BMC Plant Biol.* 4:10. doi: 10.1186/1471-2229-4-10

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y. H., 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

Chen, Y. H., Yang, X. Y., He, K., Liu, M. H., Li, J. G., Gao, Z. F., et al. (2006). The MYB transcription factor superfamily of Arabidopsis: expression analysis and phylogenetic comparison with the rice MYB family. *Plant Mol. Biol.* 60, 107–124. doi: 10.1007/s11103-005-2910-y

Colanero, S., Perata, P., and Gonzali, S. (2018). The atroviolacea gene encodes an R3-MYB protein repressig anthocyanin syntheis in tomato plants. *Front. Plant Sci.* 9:830. doi: 10.3389/fpls.2018.00830

Cominelli, E., and Tonelli, C. (2009). A new role for plant R2R3-MYB transcription factors in cell cycle regulation. *Cell Res.* 19, 1231–1232. doi: 10.1038/cr.2009.123

Cominelli, E., Galbiati, M., Vavasseur, A., Conti, L., Sala, T., Vuylsteke, M., et al. (2005). A Guard-Cell-Specific MYB transcription factor regulates stomatal movements and plant drought tolerance. *Curr. Biol.* 15, 1196–1200. doi: 10.1016/j.cub.2005.05.048

Daniel, X., Lacomme, C., Morel, J. B., and Roby, D. (1999). A novel myb oncogene homologue in Arabidopsis thaliana related to hypersensitive cell death. *Plant J.* 20, 57–66. doi: 10.1046/j.1365-313x.1999.00578.x

Dias, A. P., Braun, E. L., McMullen, M. D., and Grotewold, E. (2003). Recently Duplicated Maize R2R3 Myb Genes Provide Evidence for Distinct Mechanisms of Evolutionary Divergence after Duplication. *Plant Physiol.* 131, 610–620. doi: 10.1104/pp.012047

Du, H., Yang, S. S., Feng, B. R., Liu, L., Huang, Y. B., and Tang, Y. X. (2012). Genome-wide analysis of the MYB transcription factor superfamily in soybean. *BMC Plant Biol.* 12:106. doi: 10.1186/1471-2229-12-106

Dubos, C., Stracke, R., Grotewold, E., Weisshaar, B., Martin, C., and Lepiniec, L. (2010). MYB transcription factors in Arabidopsis. *Trends Plant Sci.* 15, 573–581. doi: 10.1016/j.tplants.2010.06.005

Gheysen, G., and Mitchum, M. G. (2011). How nematodes manipulate plant development pathways for infection. *Curr. Opin. Plant Biol.* 14, 415–421. doi: 10.1016/j.pbi.2011.03.012

Guo, L., Yang, H., Zhang, X., and Yang, S. H. (2013). Lipid transfer protein 3 as a target of MYB96 mediates freezing and drought stress in Arabidopsis. *J Exp. Bot.* 64, 1755–1767. doi: 10.1093/jxb/ert040

Gutierrez, O. A., Wubben, M. J., Howard, M., Roberts, B., Hanlon, E., and Wilkinson, J. R. (2009). The role of phytohormones ethylene and auxin in plant-nematode interactions. *Russ. J Plant Physiol.* 56, 1–5. doi: 10.1134/s1021443709010014

Harris, T. D., Buzby, P. R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., et al. (2008). Single-molecule DNA sequencing of a viral genome. *Science* 320, 106–109.

Hsu, C. Y., Jenkins, J. N., Saha, S., and Ma, D. P. (2005). Transcriptional regulation of the lipid transfer protein gene LTP3 in cotton fibers by a novel MYB protein. *Plant Sci.* 168, 167–181. doi: 10.1016/j.plantsci.2004.07.033

Hu, B., Jin, J., Guo, A. Y., Zhang, H., and Gao, G. (2014). GSDS 2.0*: An upgraded gene feature visualization server*. *Bioinformatics* 31, 1296–1297. doi: 10.1093/bioinformatics/btu817

Huang, S. W., Li, R. Q., Zhang, Z. H., Li, L., Gu, X. F., Fan, W., et al. (2009). The genome of the cucumber. *Cucumis sativus L. Nat Genet.* 475, 1–7.

Jin, H., and Martin, C. (1999). Multifunctionality and diversity within the plant MYB-gene family. *Plant Mol. Biol.* 41, 577–585. doi: 10.1023/A:1006319732410

Jung, H. W., Kim, W., and Hwang, B. K. (2003). Three pathogen-inducible genes encoding lipid transfer protein from pepper are differentially activated by pathogens, abiotic, and environmental stresses. *Plant Cell Environ.* 26:915. doi: 10.1046/j.1365-3040.2003.01024.x

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

Letunic, I., and Bork, P. (2018). 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res*. 46.

Letunic, I., Doerks, T., and Bork, P. (2015). SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res*. 43.

Li, K. T., Zhang, J., Kang, Y. H., Chen, C. M., Song, T. T., Geng, H., et al. (2018). McMYB10 modulates the expression of a Ubiquitin Ligase. *McCOP1 during leaf coloration in crabapple*. *Front. Plant Sci.* 9, 704–717. doi: 10.3389/fpls.2018.00704

Li, Q., Zhang, C., Li, J., Wang, L. N., and Ren, Z. H. (2012). Genome-wide identification and characterization of R2R3MYB family in Cucumis sativus. *PLoS One.* 7:e47576. doi: 10.1371/journal.pone.0047576

Li, Z., Zhang, Z. H., Yan, P. C., Huang, S. W., and Lin, K. (2011). RNA-Seq improves annotation of protein-coding genes in the cucumber genome. *BMC Genomics.* 12, 540–550. doi: 10.1186/1471-2164-12-540

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

Lou, Q. F., He, Y. H., Cheng, C. Y., Zhang, Z. H., Li, J., Huang, S. W., et al. (2013). Integration of high-resolution physical and genetic map reveals differential recombination frequency between chromosomes and the genome assembling quality in cucumber. *PLoS One.* 8:e62676. doi: 10.1371/journal.pone.0062676

Lozano, R., Hamblin, M. T., Prochnik, S., and Jannink, J. (2015). Identification and distribution of the NBS-LRR gene family in the Cassava genome. *BMC Genomics* 16:360. doi: 10.1186/s12864-015-1554-1559

Magali, L., Patrice, D., Gert, T., et al. (2002). PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucl. Acids Res.* 30, 325–327. doi: 10.1093/nar/30.1.325

Marchler-Bauer, A., Bo, Y., Han, L. Y., He, J., Lanczycki, C. J., Lu, S. N., et al. (2017). CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucl. Acids Res.* 45, 200–203. doi: 10.1093/nar/gkw1129

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S. N., Chitsaz, F., Geer, L. Y., et al. (2015). CDD: NCBI's conserved domain database. *Nucl. Acids Res.* 43, 222–226. doi: 10.1093/nar/gku1221

Matus, J. T., Aquea, F., and Arce-Johnson, P. (2008). Analysis of the grape MYB R2R3 subfamily reveals expanded wine quality-related clades and conserved gene structure organization across Vitis and Arabidopsis genomes. *BMC Plant Biol.* 8:83. doi: 10.1186/1471-2229-8-83

Mengiste, T., Chen, X., Salmeron, J., and Dietrich, R. (2003). The BOTRYTIS SUSCEPTIBLE1 Gene Encodes an R2R3MYB transcription factor protein that is required for biotic and abiotic stress responses in Arabidopsis. *Plant Cell.* 15, 2551–2565. doi: 10.1105/tpc.014167

Molinari, S., Fanelli, E., and Leonetti, P. (2014). Expression of tomato salicylic acid (SA)-responsive pathogenesis-related genes in Mi-1-mediated and SA-induced resistance to root-knot nematodes. *Mol. Plant Pathol.* 15, 255–264. doi: 10.1111/mpp.12085

Oda, Y., Matsunaga, T., Fukuyama, K., Miyazaki, T., and Morimoto, T. (1997). Tertiary and quaternary structures of 0.*19* alpha-amylase inhibitor from wheat kernel determined by X-ray analysis at 2.06 A resolution. *Biochemistry* 36:13503. doi: 10.1021/bi971307m

Ogata, K., Morikawa, S., Nakamura, H., Sekikawa, A., Inoue, T., Kanai, H., et al. (1994). Solution structure of a specific DNA complex of the Myb DNA-binding domain with cooperative recognition helices. *Cell* 79, 639–648. doi: 10.1016/0092-8674(94)90549-5

Raffaele, S., Vailleau, F., Léger, A., Joubès, J., Miersch, O., Huard, C., et al. (2008). A MYB transcription factor regulates very-long-chain fatty acid biosynthesis for activation of the hypersensitive cell death response in *Arabidopsis*. *Plant Cell* 20, 752–767. doi: 10.1105%2Ftpc.107.054858

Riechmann, J. L., Heard, J., Martin, G., Reuber, L., Jiang, C., Keddie, J., et al. (2000). Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science* 290, 2105–2110. doi: 10.1126/science.290.5499.2105

Sara, C., Pierdomenico, P., and Silvia, G. (2018). The atroviolacea gene encodes an R3-MYB protein repressing anthocyanin synthesis in tomato plants. *Front. Plant Sci.* 9, 830–846.

Seo, P. J., Lee, S. B., Suh, M. C., Park, M. J., Go, Y. S., and Park, C. M. (2011). The MYB96 Transcription factor regulates cuticular wax biosynthesis under drought conditions in Arabidopsis. *Plant Cell.* 23, 1138–1152. doi: 10.1105/tpc.111.083485

Seo, P. J., and Park, C. M. (2010). MYB96-mediated abscisic acid signals induce pathogen resistance response by promoting salicylic acid biosynthesis in Arabidopsis. *New Phytol.* 186, 471–483. doi: 10.1111/j.1469-8137.2010.03183.x

Seo, P. J., Xiang, F., Qiao, M., Park, J. Y., and Park, C. M. (2009). The MYB96 Transcription Factor Mediates Abscisic Acid Signaling during Drought Stress Response in Arabidopsis. *Plant Physiol.* 151, 275–289. doi: 10.1104/pp.109.144220

Sun, X. L., Shao, W. C., Cai, H., and Zhu, Y. M. (2010). Regulation of AtbZIP1 gene expression by t-dna insertion in Arabidopsis thaliana. *J. North. Agricul. Unive.* 17, 19–23.

Suo, J., Liang, X., Pu, L., Zhang, Y. S., and Xue, Y. B. (2003). Identification of GhMYB109 encoding a R2R3MYB transcription factor that expressed specifically in fiber initials and elongating fibers of cotton (Gossypium hirsutum L.). *Biochim. Biophys. Acta* 1630, 25–34. doi: 10.1016/j.bbaexp.2003.08.009

Taylor, J. S., and Raes, J. (2004). Duplication and divergence: The evolution of new genes and old ideas. *Annu. Rev. Genet.* 38, 615–643. doi: 10.1146/annurev.genet.38.072902.092831

Thoma, S., Hecht, U., Kippers, A., Botella, J., Vries, S. D., and Somerville, C. (1994). Tissue-specific expression of a gene encoding a cell wall-localized lipid transfer protein from Arabidopsis. *Plant Physiol.* 105:35. doi: 10.1104/pp.105.1.35

Voorrips, R. E. (2002). MapChart: software for the graphical presentation of linkage maps and QTLs. *J. Hered.* 93, 77–78. doi: 10.1093/jhered/93.1.77

Wang, C. L., Lower, S., and Williamson, V. M. (2009). Application of Pluronic gel to the study of root-knot nematode behaviour. *Nematology* 11, 453–464. doi: 10.1163/156854109x447024

Wang, H. H., Wang, X. Q., Yu, C. Y., Wang, C. T., and Zhang, H. X. (2020). MYB transcription factor PdMYB118 directly interacts with bHLH transcription factor PdTT8 to regulate wound-induced anthocyanin biosynthesis in poplar. *BMC Plant Biol* 20:173. doi: 10.1186/s12870-020-02389-1

Wang, X., Cheng, C. Y., Li, Q. R., Zhang, K. J., and Chen, J. F. (2019). Multi-omics analysis revealed that MAPK signaling and flavonoid metabolic pathway contributed to resistance against Meloidogyne incognita in the introgression line cucumber. *J Proteomics.* 220:103675. doi: 10.1016/j.jprot.2020.103675

Wang, X., Cheng, C. Y., Zhang, K. J., Tian, Z., Yang, S. Q., Lou, Q. F., et al. (2018). Comparative transcriptomics reveals suppressed expression of genes related to auxin and the cell cycle contributes to the resistance of cucumber against Meloidogyne incognita. *BMC Genomics.* 19, 583–596. doi: 10.1186/s12864-018-4979-0

Wang, X., Li, Q. R., Cheng, C. Y., Zhang, K. J., Lou, Q. F., Li, J., et al. (2020). Genome-wide analysis of putative lipid transfer protein LTP_2 gene family reveals CsLTP_2 genes involved in response of cucumber against root-knot nematode (Meloidogyne incognita). *Genome* 63, 225–238. doi: 10.1139/gen-2019-0157

Wang, Y. P., Tang, H. B., Debarry, J. D., Tan, X., Li, J. P., Wang, X. Y., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucl. Acids Res.* 40:e49. doi: 10.1093/nar/gkr1293

Wilkins, O., Nahal, H., Foong, J., Provart, N. J., and Campbell, M. M. (2009). Expansion and diversification of the Populus R2R3-MYB family of transcription factors. *Plant Physiol.* 149, 981–993. doi: 10.1104/pp.108.132795

Wilson, H. T., Xu, K., and Taylor, A. G. (2015). Transcriptome analysis of gelatin seed treatment as a biostimulant of cucumber plant growth. *Sci. World J.* 2015:391234. doi: 10.1155/2015/391234

Xing, C. H., Liu, Y., Zhao, L. Y., Zhang, S. L., and Huang, X. S. (2019). A novel MYB transcription factor regulates ascorbic acid synthesis and affects cold tolerance. *Plant Cell Environ*. 42, 832–845.

Ye, D. Y., Qian, C. T., and Chen, J. F. (2011). Screening and identification of cucumber-sour cucumber introgression lines resistant to the root-knot nematode Meloidogyne incognita. *Acta Horticult. Sin.* 12:16.

Yu, C. S., Chen, Y. C., Lu, C. H., and Hwang, J. K. (2006). Prediction of protein subcellular localization. *Prot. Struct. Funct. Bioinform.* 64, 643–651. doi: 10.1110/ps.03479604

Yu, C. S., Lin, C. J., and Hwang, J. K. (2004). Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. *Prot. Sci.* 13, 1402–1406. doi: 10.1110/ps.03479604

Zhang, L. C., Zhao, G. Y., Jia, J. Z., Liu, X., and Kong, X. Y. (2012). Molecular characterization of 60 isolated wheat MYB genes and analysis of their expression during abiotic stress. *J. Exp. Bot.* 63, 203–214. doi: 10.1093/jxb/err264

Zhang, Y. L., Zhang, C. L., Wang, G. L., Wang, Y. X., and Hao, Y. J. (2019). The R2R3 MYB transcription factor MdMYB30 modulates plant resistance against pathogens by regulating cuticular wax biosynthesis. *BMC Plant Biol.* 19:362. doi: 10.1186/s12870-019-1918-4

Zhao, W., Zhou, X., Lei, H., Fan, J., Yang, R., Li, Z., et al. (2018). Transcriptional evidence for crosstalk between JA and ET or SA during root-knot nematode invasion in tomato. *Physiol. Genomics.* 50, 197–207. doi: 10.1152/physiolgenomics.00079.2017

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.

Copyright © 2020 Cheng, Li, Wang, Li, Qian, Li, Lou, Jahn and Chen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 08 December 2020
doi: 10.3389/fgene.2020.584817

Comparative Transcriptome Reveals the Genes' Adaption to Herkogamy of *Lumnitzera littorea* (Jack) Voigt

Ying Zhang[1,2†], Yukai Chen[3†], Yan Zhou[1], Jingwen Zhang[3], He Bai[3] and Chunfang Zheng[2]*

[1] School of Life Sciences and Technology, Lingnan Normal University, Zhanjiang, China, [2] National and Local Joint Engineering Research Center of Ecological Treatment Technology for Urban Water Pollution, College of Life and Environmental Science, Wenzhou University, Wenzhou, China, [3] Ministry of Education Key Laboratory for Ecology of Tropical Islands, College of Life Sciences, Hainan Normal University, Haikou, China

Lumnitzera littorea (Jack) Voigt is among the most endangered mangrove species in China. The morphology and evolution of *L. littorea* flowers have received substantial attention for their crucial reproductive functions. However, little is known about the genomic regulation of flower development in *L. littorea*. In this study, we characterized the morphology of two kinds of *L. littorea* flowers and performed comparative analyses of transcriptome profiles of the two different flowers. Morphological observation showed that some flowers have a column embedded in the petals while others produce a stretched flower style during petal unfolding in flowering. By using RNA-seq, we obtained 138,857 transcripts that were assembled into 82,833 unigenes with a mean length of 1055.48 bp. 82,834 and 34,997 unigenes were assigned to 52 gene ontology (GO) functional groups and 364 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, respectively. A total of 4,267 differentially expressed genes (DEGs), including 1,794 transcription factors (TFs), were identified between two types of flowers. These TFs are mainly involved in bHLH, B3, bZIP, MYB-related, and NAC family members. We further validated that 12 MADS-box genes, including 4 MIKC-type and 8 M-type TFs, were associated with the pollinate of *L. littorea* by herkogamy. Our current results provide valuable information for genetic analysis of *L. littorea* flowering and may be useful for illuminating its adaptive evolutionary mechanisms.

Keywords: transcriptome, *Lumnitzera littorea*, floral organ, MADS-box, mangrove

OPEN ACCESS

Edited by:
Rong Zhou,
Aarhus University, Denmark

Reviewed by:
Guohua Chai,
Qingdao Agricultural University, China
Cheng Qin,
Zunyi Vocational and Technical College, China

****Correspondence:***
Chunfang Zheng
20195101@wzu.edu.cn

† These authors have contributed equally to this work

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

Received: *18 July 2020*
Accepted: *09 November 2020*
Published: *08 December 2020*

Citation:
Zhang Y, Chen Y, Zhou Y, Zhang J, Bai H and Zheng C (2020) Comparative Transcriptome Reveals the Genes' Adaption to Herkogamy of Lumnitzera littorea (Jack) Voigt. Front. Genet. 11:584817. doi: 10.3389/fgene.2020.584817

INTRODUCTION

Lumnitzera littorea (Jack) Voigt. (Combretaceae, *Lumnitzera* genus) is a non-viviparous Indo-West Pacific mangrove species. *L. littorea* is sparsely distributed in India, Sri Lanka, Myanmar, Thailand, Malaysia and Indonesia, and China (Zhou et al., 2018). Based on IUCN (International Union for Conservation of Nature) Red List Categories and Criteria, *L. littorea* was listed as a least concern

Abbreviations: bZIP, basic leucine zipper; COG, clusters of orthologous groups of proteins; DEGs, differentially expressed genes; ERF, ethylene-responsive factor; FLC, TF flowering locus C; GO, gene ontology; HLH, helix-loop-helix; NR, RefSeq non-redundant proteins; KEGG, Kyoto Encyclopedia of Genes and Genomes; MYB, v-myb avian myeloblastosis viral oncogene homolog; PDB, protein data bank; MADS, mini chromosome maintenance 1, agamous, deficiens, and serum response factor; PIR, protein information resource; PRF, protein research foundation; SEP, sepallata; TFs, transcription factors.

(LC) species (Polidoro et al., 2010). In China, the wild plant number was only 359 in 2006, and rapidly declined to 9 in 2018. The narrow distribution of it was only in Sanya Tielu harbor and Lingshui Dadun village of Hainan Island. In 2018, all of the wild *L. littorea* growing in Lingshui Dadun village died (Fan and Chen, 2006; Zhang et al., 2018). During 13 years of field observation, no seedlings or young trees were observed due to high (76%) seed abortive rate (Zhang et al., 2017). The protection of this species is facing a great challenge and the mangrove *L. littorea* is therefore is listed as a plant under state protection (category II) (Zhong et al., 2011; Zhang et al., 2013).

L. littorea has relatively low genetic diversity and gene flow in China (Su, 2004), possibly because of the limited number of wild individuals and distribution area (Zhang et al., 2018). According to the pollen-ovule ratio (P/O) and hybridization index (OCI) analyses, *L. littorea* is classified as a typical cross-pollinated plant with red petals and erectly terminal inflorescence. Most of the *L. littorea* flowers can only be pollinated from the same tree or even from the same flower (Li et al., 2016; Zhang et al., 2016, 2017). The pollen viability *L. littorea* in China was lower than 10% (Zhang et al., 2013, 2016). The heavy abortion of *L. littorea* seeds resulted from the high empty embryo rate (Zhang et al., 2018). In woody perennials, there are several studies on the regulation of flowering (Chen et al., 2018), but the underlying molecular mechanisms of floral dynamics and breeding systems in the development of *L. littorea* remain poorly understood.

MADS-box transcription factors play crucial roles in floral organ formation, embryo and reproductive development and flowering time control (Angenent, 1995; Moon et al., 2003; Arora et al., 2007; Irish, 2010; Mohanty and Joshi, 2018; Zhang et al., 2018). Extensive studies of *Arabidopsis* mutants show several genetic models of floral organ formation (Coen and Meyerowitz, 1991; Theissen et al., 2016). The ABCDE model involves five subgroups of the MADS family. *AP1* (*APETALA 1*) and *AP2* belong to A-class; *AP3* and *PI* (*PISITTALA*) belong to B-class; *AGAMOUS* (*AG*) belongs to C-class; *SEEDSTICK* (*STK*) belongs to D-class; and *SEP1* (*SEPALLATA 1*), *SEP2, SEP3*, and *SEP4* belong to E-class (Bowman and Meyerowitz, 1991). The combinations of MADS-box proteins determine the tetrameric complexes (Theiβen and Saedler, 2001). For instance, class A+B+E genes control petal development, B+E+C genes determine stamen development, C+E specify carpels, and D+E are necessary for ovule development (Wellmer et al., 2014). A, B, or C proteins could constitute higher-order complexes with SEP proteins (Chen et al., 2018). The *sep1/2/3/4* mutant displays indeterminate flowers composed of leaf-like organs and sepal development, indicating the role of SEP proteins in control flower development (Favaro et al., 2003). Importantly, the "ABCDE" model key genes are conservative in the control of petal and style development in *Soybean*, *Impatiens* and *Marcgravia* (Geuten et al., 2006; Litt and Kramer, 2010; Huang et al., 2014).

TF flowering locus C is a convergence point for environmental and endogenous pathways that regulate flowering time in *Arabidopsis* (Mateos et al., 2015). *AGL27* mutants flower earlier in a dosage dependent manner while transgenic plants carrying *AGL27* overexpression cassettes are delayed in flowering (Scortecci et al., 2001; Yun et al., 2011). FLC interacts with another MADS-box protein, SHORT VEGETATIVE PHASE (SVP), to delay flowering (Li et al., 2008). The function of the MADS-box gene has been verified in the flower development of many species, but it has not been reported in *L. littorea*.

Here, we conducted *de novo* transcriptome sequencing of two kinds of flowering behavior with different types of style development for *L. littorea* in order to investigate gene expression patterns associated with special style development morphology. To our knowledge, this is the first comprehensive transcriptomic study of flower development for *L. littorea*, providing important bioinformatic resources for the investigation of genes involved in flower development, and building a foundation for investigating the role of these genes and gene networks in the evolution of floral diversity across *L. littorea.*

MATERIALS AND METHODS

Plant Material

The *L. littorea* trees live in Sanya Tielu Bay, Hainan, China (18°15′-18°17′N, 109°42′-109°44′E). Flowers with columns embedded in the petals (L-1) or with stretched styles (L-2) were collected from one florescence of one tree at 9 am on July 20, 2017. For each type, at least three flowers were selected. The materials were immersed into liquid nitrogen and stored at −80°C for subsequent research. Three biological replicates were prepared for sequencing.

RNA Extraction and Deep Sequencing

RNA isolation was performed with TRIzol® Reagent (Invitrogen, United States) following the manufacturer's instructions. RNA samples with an RNA integrity number (RIN) >9.5 were used for purification and subsequent cDNA construction with the TruSeq RNA sample RNA prep kit (Illumina, United States). After synthesis of the first-strand cDNA, the second-strand cDNA was produced using buffer, dNTPs, RNase H, and DNA polymerase I. The double-strand cDNA was purified using the QIAquick PCR extraction kit (QIAGEN, Germany) and washed with EB buffer for end repair and single nucleotide adenine (A) addition. After PCR amplification for 15 cycles, the products were loaded onto flow cell channels at 12 pM for paired-end 150 bp × 2 sequencing with the Illumina HiSeq 4000 platform (Majorbio, Shanghai, China).

De novo Assembly and Analysis of Illumina Reads

Clean reads were obtained by (1) removing the adapters and reads without fragmentation; (2) cutting the low quality bases (quality score less than 20) at the 3′ end of the sequence and then, if the quality of the residual sequence is still less than 10, removing the entire sequence, while sequences with a quality greater than 10 are retained; (3) removing reads that contain too many Ns (≥10%); and (4) removing reads less than 20 bp long after adapter discarding and quality control. Analysis tools:

SeqPrep[1] and Sickle[2]. The *de novo* assembly was conducted with the Trinity software[3] (Grabherr et al., 2011). The raw data have been uploaded to NCBI SRA under accession numbers SRR6429108 to SRR6429113.

Transcriptome Annotation

BlastX was used to perform sequence alignments between the transcriptome and sequence data from the NR, String, SwissProt and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases. Alignments with *E*-values less than $1e^{-5}$ were chosen. NCBI_NR is a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, the Protein Information Resource (PIR), the Protein Research Foundation (PRF), and the Protein Data Bank (PDB). Via GO (gene ontology) annotation, the database standardizes the biological terms of genes and gene products and unifies the definitions and descriptions of gene and protein functions (Grabherr et al., 2011). The Clusters of Orthologous Groups of proteins (COG) database[4] is an orthologous protein cluster database that depends on the phylogenetic relationships of complete protein sequences from 66 selected strains. Functional annotation, classification and protein evolution analysis can be performed by comparing sequences with the COG database (Zhang et al., 2014). Pathway assignments were performed according to the KEGG[5] pathway database (Grabherr et al., 2011; Huang et al., 2015) with BlastX and an *E*-value threshold of $1e^{-5}$.

Identification of Differentially Expressed Transcripts

EdgeR[6] was used for differential expression analysis. Gene read count data were calculated as the input of EdgeR or DESeq2. This analysis method is based on the negative binomial distribution model. The screening criteria of significant DEGs were as follows: FDR < 0.05 and $|\log2FC| >= 1$ (Anders and Huber, 2010).

Annotation and Phylogenetic Analysis

To identify the TFs represented in the *L. littorea* transcriptomes, all DEGs were searched against the plant TF database PlantTFDB 4.0 (Jin et al., 2014)[7]. BlastN searches of the Phytozome database, using *Arabidopsis* genes as queries, were used to identify flower and floral organ development-related genes in *L. littorea*. The CDS sequences of all MADS-box genes from *Arabidopsis*, *Hevea brasiliensis*, and *Fragaria ananassa* were downloaded from GenBank. Multiple sequence alignments and the phylogenetic analysis of CDS sequences were performed as described previously (Cheng et al., 2017). An unrooted phylogenetic tree was created with the neighbor-joining method by using MEGA-X, and a bootstrap test was set to 1000 replicates (Kumar et al., 2018).

Real-Time PCR Analysis

Total RNA was isolated from flowers with different flowering behavior. Three biological replicates were set. qRT-PCR assays were conducted using the ABI PRISM 7300 Sequence Detection System (Applied Biosystems) with SYBR Green PCR Master Mix (Applied Biosystems). The housekeeping gene *ACTIN* (c14139_g1) was used for normalization in each qRT-PCR run. The relative expression levels of target genes are presented as $2^{-\Delta\Delta CT}$ (Livak and Schmittgen, 2001). Primers used in this study are listed in **Supplementary Table S1**.

RESULTS

Floral Structure Morphogenesis of *L. littorea* Flowers

The *L. littorea* flowers are hermaphroditic with red, erect petals and a deep, curved calyx tube with abundant nectar (**Figures 1A,B**). The diameter of a single flower is approximately 6.70 ± 0.04 mm. The five petals per flower are 4.20 ± 0.43 mm long. The pistils and stamens are approximately the same length and as long as 8 mm. Stamens are prominently exserted at anthesis after the stamens unfold. During flower opening, two kinds of flowers can be found before the petals are uncovered. One kind retains stigma within the petals and was named as L-1 (**Figure 1C**); the other kind (L-2) are herkogamy flowers and has columns stretched beyond the petals before flowering (**Figure 1D**). In L-1, no stylar canal was found on the stigma (**Figure 1E**). In L-2, there is an obvious stylar canal in the stigma (**Figure 1F**), the filaments are kept folded, and the anthers are kept intact.

RNA-Seq and *de novo* Assembly

To obtain an overview of the transcriptome profiles, L-1 and L-2 were sampled at different column development stages for Illumina deep sequencing. 66.83 (L-1) and 65.40 (L-2) million raw reads were yielded. A total of 138,857 transcripts with a GC percent of 38.91%, average length of 1665.35 and an N50 size of 3049, were obtained (**Table 1**). 82,833 unigenes with a GC percent of 38.59, an average length of 1055.48 and an N50 size of 2270, were produced.

Functional Annotation of Unigenes

The unigene sets obtained from the *L. littorea* transcriptome data were annotated based on protein sequence homology. All transcripts and unigenes produced were searched against the NCBI NR, SwissProt, String, KEGG and Pfam databases with an *E*-value threshold $< 1e^{-5}$. As a result, 138,857 transcripts and 82,833 unigenes were annotated (**Supplementary Figure S1**). The similarity distribution analysis identified 41,687 transcripts and 13,958 unigenes that exhibited high sequence similarity (from 80% to 100%) with known gene sequences. Regarding species distribution, the NR database queries showed

[1] https://github.com/jstjohn/SeqPrep
[2] https://github.com/najoshi/sickle
[3] https://github.com/trinityrnaseq/trinityrnaseq/releases
[4] http://www.ncbi.nlm.nih.gov/COG/
[5] http://www.genome.jp/kegg/
[6] http://www.bioconductor.org/packages/2.12/bioc/html/edgeR.html
[7] http://plantregmap.gao-lab.org/



FIGURE 1 | The structure of flowers in *Lumnitzera littorea*. **(A,B)** Hermaphroditic flower; **(C)** L-1: column into a petal; **(D)** L-2: column stretched beyond the petals; **(E)** No pistillar chord on stigma; **(F)** Pistillar chord on stigma.

that 42.6% of the *L. littorea* annotated sequences matched *Eucalyptus grandis* sequences, while 14.14, 13.52, and 8.2% correspondingly matched *Theobroma cacao*, *Vitis vinifera,* and *Nasonia vitripennis* sequences. Characteristics of the homology search of *L. littorea* unigenes against the NR database are shown in **Supplementary Figure S2**.

Based on the BLASTX results against the NR database, we assigned GO terms to the assembled unigenes to obtain

TABLE 1 | Summary of Illumina transcriptome sequencing.

Type	Unigene	Transcripts
Total sequence number	82833	138857
Total sequence base	87428846	231245876
Percent GC	38.59%	38.91%
Largest	17599	17599
Smallest	201	201
Average unigene length (bp)	1055.48	1665.35
N50 length (bp)	2270	3049

TABLE 2 | MADS-box gene names and attributes of *Lumnitzera littorea*.

Gene Name	Transcriptome ID	Strand	Protein length	Subfamily
LliMADS1	c14522_g1_i1	−	175	AP3
LliMADS2	c16514_g1_i1	−	175	SVP/AGL24
LliMADS3	c17232_g1_i1	+	175	Mα
LliMADS4	c17574_g1_i1	−	175	SVP/AGL24
LliMADS5	c19549_g1_i1	+	165	MIKC*
LliMADS6	c19801_g1_i3	−	122	MIKC^c
LliMADS7	c21067_g7_i3	−	104	SVP/AGL24
LliMADS8	c23460_g1_i1	−	173	MIKC*
LliMADS9	c24786_g2_i1	−	174	AP3
LliMADS10	c25298_g6_i2	+	174	SEP
LliMADS11	c44184_g1_i1	−	157	Mα
LliMADS12	c44474_g1_i1	+	142	Mγ

GO functional annotations and categorizations. All of the unigenes were used to query the GO database in order to classify their predicted functions (**Supplementary Figure S3** and **Supplementary Table S2**). In the "biological process" category (34,918 unigenes), macromolecule metabolic process (4,389) was the largest subcategory. In the "cell component" (31,040 unigenes) and "molecular function" (16,876 unigenes) categories, intracellular (4,429) and nucleotide binding (2,755) were the most abundant GO terms, respectively. The GO analysis indicated that a high number of unigenes were associated with the various biological processes and molecular functions in *L. littorea* floral tissues.

The annotated sequences were further applied to a search against the clusters of orthologous groups of proteins (COG) and clusters of orthologous groups for eukaryotic complete genomes (KOG) databases for functional prediction and classification. As a result, each annotated unigenes was assigned 25 COG and 25 KOG terms. Among the assigned terms, the three most highly represented categories in the two databases were identical. (1) general function prediction only (883 unigenes in the COG databases; 1267 unigenes in the KOG databases); (2) signal transduction mechanism (844 unigenes in the COG databases; 1136 unigenes in the KOG databases); and (3) posttranslational modification, protein turnover, and chaperones (740 unigenes in the COG databases; 1006 unigenes in the KOG databases). The smallest group was "cell motility," with 3 unigenes in the COG databases and 1 unigene in the KOG databases (**Supplementary Figure S4**).

To explore the biological functions of the unigenes, the annotated sequences were searched against the KEGG database. 42.3% (34,997/82,833) of unigenes were assigned to 364 KEGG pathways. The top five pathways were "carbon metabolism" (ko01200), "ribosome" (ko03010), "protein processing in endoplasmic reticulum" (ko04141), "biosynthesis of amino acids" (ko01230) and "oxidative phosphorylation" (ko00190) (**Supplementary Figure S5**). These results provide valuable information for gene discovery and functional characterization.

Comparation of DEGs Between Two Types of *L. littorea* Flower

To examine gene expression among two different floral behaviors, two transcriptome profiles were compared. We found 4,267 distinct unigene sequences that were significantly different between L-1 and L-2. Of these, 1,874 were upregulated and 2,393 were downregulated in herkogamy flowers (L-2) (**Supplementary Table S3**). Floral homeotic protein DEFICIENS-like gene (c44184_g1), SVP-like floral repressor gene (c16514_g1), receptor-like kinase in flowers (c18462_g3), MYB family genes (MYB16, c24005_g10; MYB86-like, c22851_g2; MYBP, c14148_g1, c22970_g1, c16951_g1; MYB1R1, c20799_g1; R2R3-MYB, c14888_g1; MYB32-like, c14253_g1; MYBJ6, c2873_g1; MYB124, c22682_g1; MYB-like protein, c23232_g17, and MYB5, c10786_g1), MADS-box family genes (STAMADS11, c21067_g7; K-box, c14522_g1, and MADS-box 24?, c25298_g6) and embryo development related genes (MEDEA 18-1, c24129_g1 and LEA, c23402_g1). The GO annotation analysis classified groups of genes with significantly differential expression into three categories: the biological process, cellular component and molecular function categories (**Supplementary Figure S6**). To identify the unigenes involved in metabolic or signal transduction pathways that were significantly enriched, all of the DEGs were used to query the KEGG database. A total of 1,215 pathways from the KEGG database were enriched (**Supplementary Table S4**). These significant pathways were classified into environmental information processing (EIP), genetic information processing (GIP), cellular processes (CP) and metabolism (M) categories. The top three significant pathways were "flavonoid biosynthesis" (KO00941), "plant hormone signal transduction" (KO04075) and "sesquiterpenoid and triterpenoid biosynthesis" (KO00909) (**Supplementary Figure S7**).

Transcription Factors in DEGs Modulating the Herkogamy of *L. littorea* Flowers

Transcription factors (TFs) play key regulatory roles in floral development by binding to specific motifs in the promoters of target genes (Chen et al., 2018). Here, we showed that 41% (1,749/4,267) of DEGs between L-1 and L-2 belong to TFs (**Supplementary Table S4**). These TFs were classed into 54 categories with bHLH, B3, bZIP, MYB-related, NAC, C2H2, C3H, ERF, WRKY, and MYB being the most highly represented (**Figure 2A**). It is noted that the *MADS*, *bZIP*, *bHLH*, and *MYB* genes play key roles in the regulation of flower development and

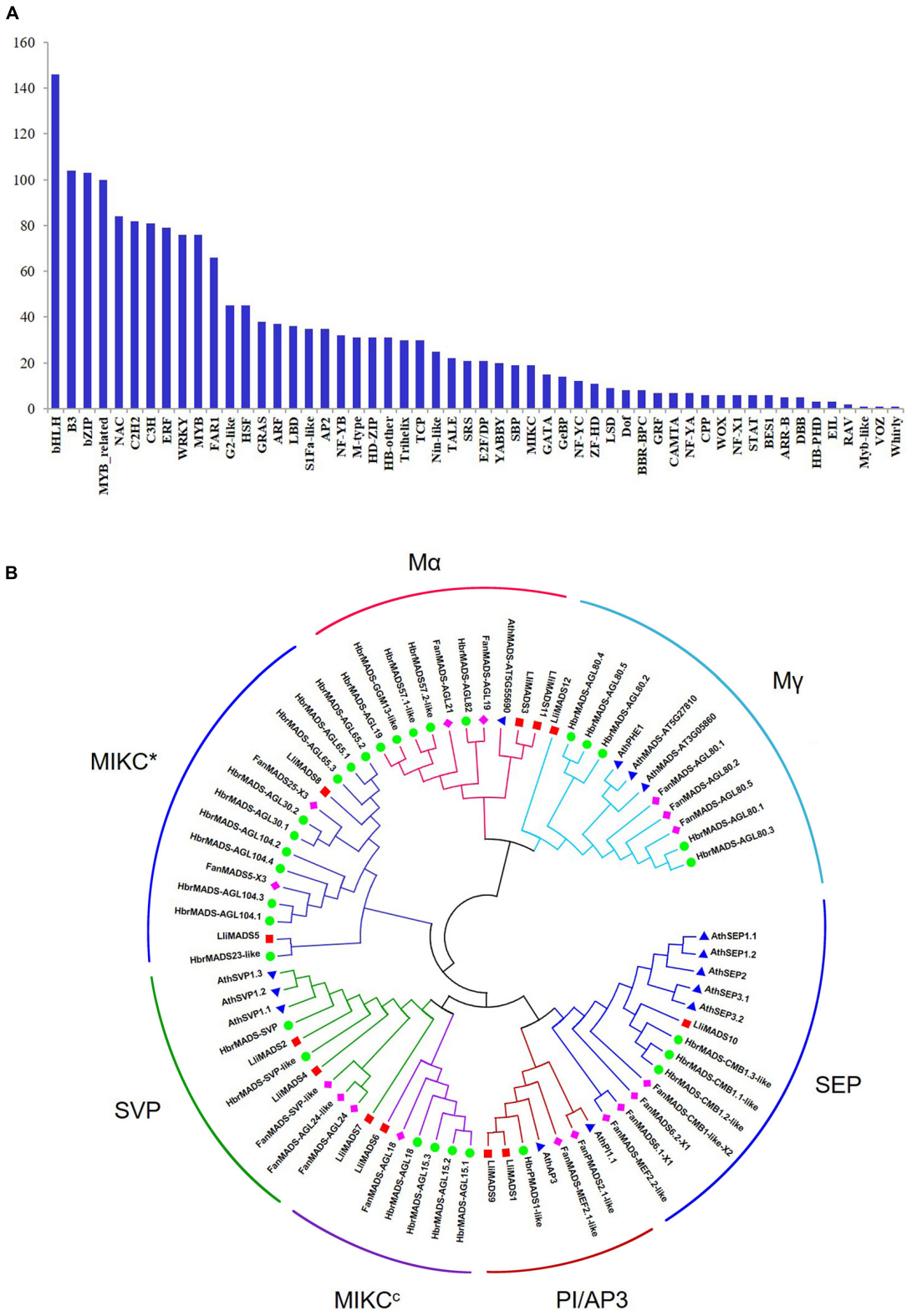


FIGURE 2 | Identification of TFs and MADS-box genes in DEGs of *Lumnitzera littorea*. **(A)** Classification of TF families. **(B)** Phylogenetic relationships of MADS-box TF CDSs from *L. littorea*, *Arabidopsis*, *Fragaria ananassa*, and *Hevea brasiliensis*. The unrooted phylogenetic tree was created with MEGA-X by the neighbor-joining method, and the bootstrap test was performed with 1,000 iterations. Red, blue, green, and purple dots indicate *L. littorea*, *Arabidopsis*, *Hevea brasiliensis*, and *Fragaria ananassa* genes, respectively. The outer circle shows the identified subfamily in MADS proteins.

flowering time (Streisfeld et al., 2013; Wang et al., 2013; Rocheta et al., 2014).

Phylogenetic Analysis of MADS-Box Genes Associated With the Herkogamy of *L. littorea* Flowers

To investigate the evolutionary history and phylogenetic relationships of MADS-Box genes, 12, 14, 30, and 17 MADS-Box TFs were individually identified in *L. littorea, Arabidopsis*, rubber tree, and strawberry based on the PlantTFDB 4.0 database (**Figure 2B** and **Supplementary Table S5**). A neighbor-joining phylogenetic tree was then generated by alignment of these MADS-box proteins. As shown in **Figure 2B**, these proteins were classified into seven groups, similar to the description in a previous report (Parenicova et al., 2003). In each subgroup, MADS proteins in *L. littorea* were more closely related to those in rubber tree, except in the Mα subfamily. Twelve

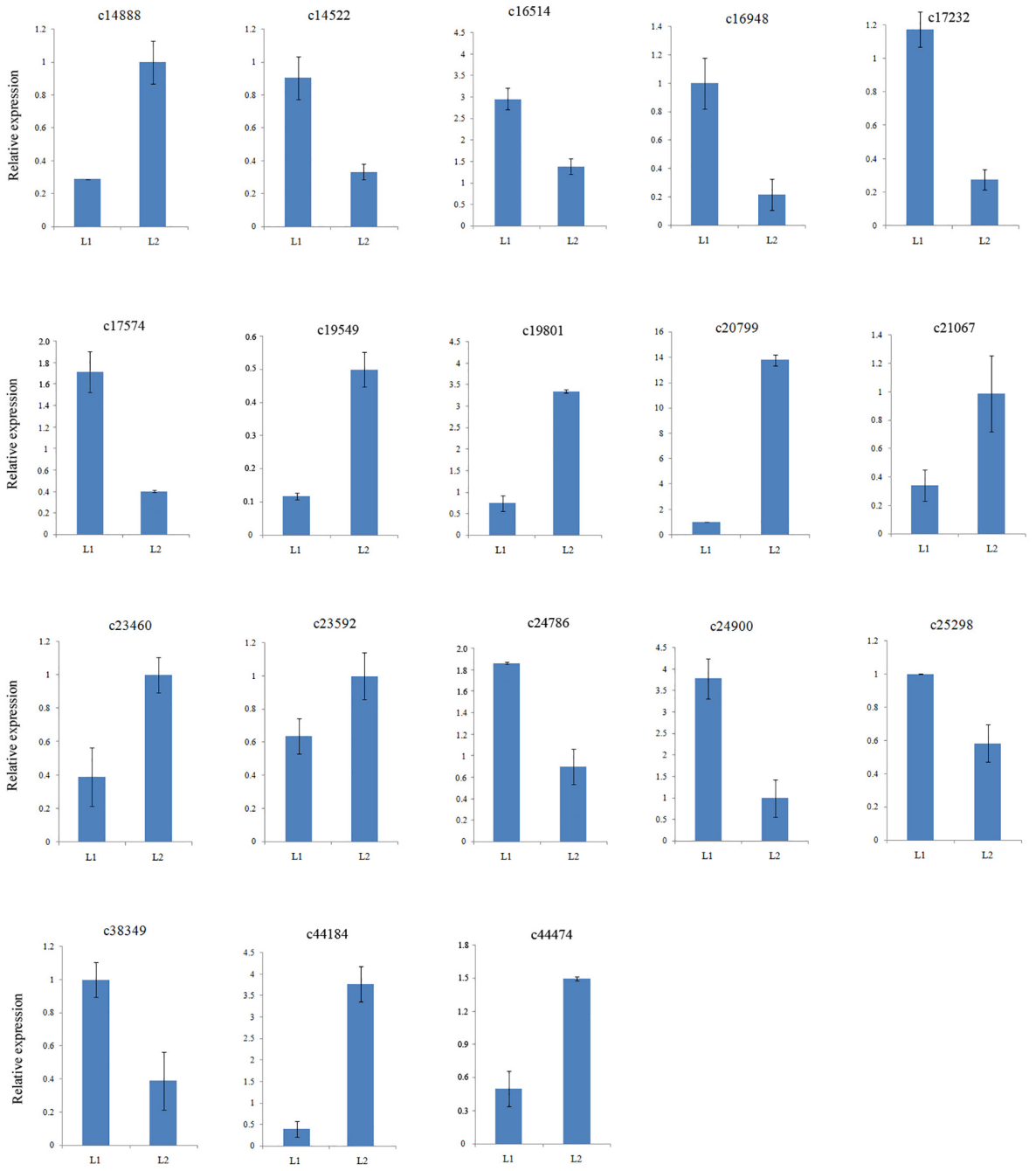


FIGURE 3 | Validation of assembled unigenes by qPCR.

differentially expressed MADS proteins between L-1 and L-2 appeared in each putative functional group. Of them, *LliMADS3* and *LliMADS11* are the most homologous with AT5G55690, and belong to the Mα subdivision of type I MADS-box genes. *LliMADS12* and *AthPHERES1* (*PHE1*) are on the same branch in the Mγ subdivision of type I MADS-box genes. *LliMADS1* and *LliMADS9* belong to the B class of the AP3 subfamily, and *LliMADS10* belongs to the SEP subfamily.

RNA-Seq Expression Validation by Real-Time PCR

To confirm the gene expression patterns identified by RNA-Seq data, the transcript levels of twelve MADS-box genes together with six other DEGs were examined by qRT-PCR. All 18 selected DEGs were successfully amplified with single bands of the expected sizes. The expression of 10 genes were down regulated and that of 7 genes were up regulated (**Supplementary Table S6**), consistent with those of the RNA-Seq data (**Figure 3**). Therefore, the DEGs obtained from the assembled transcriptome were accurate and reliable.

DISCUSSION

Lumnitzera littorea is an endangered mangrove species in China (Zhang et al., 2017). Herkogamy is found in almost 60% of *L. littorea* flowers, but approximately 40% of the flowers have a column embedded in the petals when the petals unfold during florescence. Almost all those flowers have empty seeds, which are speculated by the results of forced self-pollination. Thus, the breeding system of *L. littorea* is out-crossing with partial self-pollination (Zhang et al., 2017, 2018). Out-crossing is obligate in unisexual flowers and selfing can occur in hermaphrodite flowers, but a self-compatible level can be strongly selected by herkogamy, i.e., the spatial separation of anthers and stigmas within a flower (Mertens et al., 2018). In addition, geitonogamous selfing is not prevented within or between inflorescences on a plant when flowers are at different sexual phases in *L. littorea* (Zhang and Wolfe, 2016; Zhang et al., 2017). The temporal separation of male and female phases is a common floral feature in hermaphrodite species (Rosas-Guerrero et al., 2017). In such a bad survival trend, only the herkogamy flower morphogenesis of *L. littorea* could hardly improve the natural reproduction rate. Therefore, flower morphogenesis suitable for geitonogamous selfing as L-1 may be a positive adaption for this survival condition, even though most of them failed in seed production.

In the present study, we analyzed the transcriptome profiles of flowers with columns embedded in the petals (L-1) and with stretched styles (L-2) and identified a total of 82,833 unigenes and 138,857 transcripts, respectively. Using the stringent criteria of both FDR < 0.05 and $|\log2FC| >= 1$, we detected 4,267 unigenes that were significantly different between L-1 and L-2. These results imply a diverse and complex mechanism of column development gene expression in *L. littorea*.

Gene ontology functional enrichment revealed that a high number of genes were associated with various biological processes and molecular functions in *L. littorea* floral tissues. KEGG pathway analysis showed that many DEGs were involved in secondary metabolite biosynthesis, including carbon metabolism, ribosome, protein processing in the endoplasmic reticulum and amino acid biosynthesis. Furthermore, plant oxidative phosphorylation that plays an important role in plant floral development (Liu et al., 2019) is also activated.

TFs play important roles in the regulation of flower development and flowering time (Li et al., 2017). In this study, we identified 1,749 differentially expressed TFs with 54 diverse categories. The different expression of these TFs indicated their possible different roles in modulating the formation of herkogamy flowers in *L. littorea*. Of these TFs, MADS-box family genes function in flower development (Parenicova et al., 2003). MADS-box proteins are generally divided into types I and II. Type I is categorized into Mα, Mβ, Mγ, and Mδ clades and type II is classified into $MIKC^c$ and MIKC* (Mohanty and Joshi, 2018). Several type I MADS box TFs are shown to be involved in reproductive development in *Arabidopsis* (Luo et al., 2008). The empty embryo rate of 70% may be related to the upregulated expression of Mγ genes in the flowers of *L. littorea* in China. *SVP* has been identified to delay flowering by modulating the biosynthesis of gibberellin in *Arabidopsis* (Andrés et al., 2014) and *Jatropha curcas* (Hui et al., 2018). In the flowers of *L. littorea*, two downregulated *SVP* genes (*LliMADS2* and *LliMADS4*), and one upregulated *SVP* genes (*LliMADS7*) affect dichogamy morphogenesis, which may also be caused by the modulating the biosynthesis of gibberellin in florescence. In *Arabidopsis*, the petal, stamen and carpel of *sep1sep2sep3* mutants are switched to sepal (Ditta et al., 2004). *LliMADS10* and *AthSEP* were found in the same branch covering SEPs (**Figure 2B**). In the flowers of *L. littorea*, *LliMADS10* was downregulated and may contribute to dichogamy morphogenesis. These genes should be paid more attention in further 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 below: https://www.ncbi.nlm.nih.gov/, SRP127706.

AUTHOR CONTRIBUTIONS

YZa and CZ designed the study and modified the manuscript. YZa, YC, YZo, JZ, and HB collected the samples and acquired the data. YZa and YC drafted the manuscript. CZ, JZ, and HB helped to draft the manuscript. All authors read and approved the final manuscript.

FUNDING

This study was funded by the National Natural Science Foundation of China (41776148 and 31360173) and the project of Zhejiang Provincial Natural Science Foundation (LY18C030001).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.584817/full#supplementary-material

Supplementary Figure 1 | Transcript and unigene annotation with the NCBI NR, SwissProt, String, KEGG and Pfam databases of *L. littorea* transcriptome data.

Supplementary Figure 2 | Species distribution with BLAST hits to the annotated unigenes of *L. littorea*.

Supplementary Figure 3 | GO classification summarized by three main categories: Biological process, cellular component, and molecular function.

Supplementary Figure 4 | COG and KOG functional classification of the unigenes of *L. littorea* with NR annotation.

Supplementary Figure 5 | The biological pathways of *L. littorea* flower unigenes against the KEGG database.

Supplementary Figure 6 | DEGs in different column development stages between L-1 and L-2.

Supplementary Figure 7 | Functional KEGG pathway annotation of the DEGs of *L. littorea*.

Supplementary Table 1 | Primers used in this study.

Supplementary Table 2 | Classify the predicted functions of the query GO database.

Supplementary Table 3 | DEGs in the materials L-2 vs. L-1.

Supplementary Table 4 | KEGG pathway enrichment analysis.

Supplementary Table 5 | Sequence of the MADS-Box genes in *L. littorea*.

Supplementary Table 6 | Validation of assembled unigenes by RNA-Seq.

REFERENCES

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11:R106. doi: 10.1186/gb-2010-11-10-r106

Andrés, F., Porri, A., Torti, S., Mateos, J., Romera-Branchat, M., García-Martínez, J. L., et al. (2014). SHORT VEGETATIVE PHASE reduces gibberellin biosynthesis at the *Arabidopsis* shoot apex to regulate the floral transition. *Proc. Natl. Acad. Sci. U.S.A.* 111, 2760–2769. doi: 10.1073/pnas.1409567111

Angenent, G. (1995). A novel class of MADS box genes is involved in ovule development in petunia. *Plant Cell* 7, 1569–1582. doi: 10.1105/tpc.7.10.1569

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

Bowman, J. L., and Meyerowitz, E. M. (1991). Genetic control of pattern formation during flower development in Arabidopsis. *Symp. Soc. Exp. Biol.* 45, 89–115.

Chen, Z., Rao, P., Yang, X., Su, X., Zhao, T., Gao, K., et al. (2018). A global view of transcriptome dynamics during male floral bud development in Populus tomentosa. *Sci. Rep.* 8:722. doi: 10.1038/s41598-017-18084-5

Cheng, Z., Ge, W., Li, L., Hou, D., Ma, Y., Liu, J., et al. (2017). Analysis of MADS-Box gene family reveals conservation in floral organ ABCDE model of moso bamboo (*Phyllostachys edulis*). *Front. Plant Sci.* 8:656. doi: 10.3389/fpls.2017.00656

Coen, E. S., and Meyerowitz, E. M. (1991). The war of the whorls: genetic interactions controlling flower development. *Nature* 353, 31–37. doi: 10.1038/353031a0

Ditta, G., Pinyopich, A., Robles, P., Pelaz, S., and Yanofsky, M. F. (2004). The SEP4 gene of Arabidopsis thaliana functions in floral organ and meristem identity. *Curr. Microbiol.* 14, 1935–1940. doi: 10.1016/j.cub.2004.10.028

Fan, H., and Chen, L. (2006). Current distribution of endangered mangrove *Lumnitzera littorea* (Jack.) Voigt in China. *Guangxi Sci.* 13, 226–227. doi: 10.13656/j.cnki.gxkx.2006.03.019

Favaro, R., Pinyopich, A., Battaglia, R., Kooiker, M., Borghi, L., Ditta, G., et al. (2003). MADS-box protein complexes control carpel and ovule development in *Arabidopsis*. *Plant Cell* 15, 2603–2611. doi: 10.1105/tpc.015123

Geuten, K., Becker, A., Kaufmann, K., Caris, P., Janssens, S., Viaene, T., et al. (2006). Petaloidy and petal identity MADS-box genes in the balsaminoid genera Impatiens and Marcgravia. *Plant J.* 47, 501–518. doi: 10.1111/j.1365-313X.2006.02800.x

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Huang, F., Xu, G., Chi, Y., Liu, H., Xue, Q., Zhao, T., et al. (2014). A soybean MADS-box protein modulates floral organ numbers, petal identity and sterility. *BMC Plant Biol.* 14:89. doi: 10.1186/1471-2229-14-89

Huang, J. Z., Lin, C. P., Cheng, T. C., Chang, B. C., Cheng, S. Y., Chen, Y.-W., et al. (2015). A de novo floral transcriptome reveals clues into Phalaenopsis orchid flower development. *PLoS One* 10:e0123474. doi: 10.1371/journal.pone.0123474

Hui, W.-K., Wang, Y., Chen, X.-Y., Zayed, M. Z., and Wu, G.-J. (2018). Analysis of transcriptional responses of the inflorescence meristems in *Jatropha curcas* following gibberellin treatment. *Int. J. Mol. Sci.* 19:432. doi: 10.3390/ijms19020432

Irish, V. F. (2010). The flowering of Arabidopsis flower development. *Plant J.* 61, 1014–1028. doi: 10.1111/j.1365-313X.2009.04065.x

Jin, J., Zhang, H., Kong, L., Gao, G., and Luo, J. (2014). PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Res.* 42, D1182–D1187. doi: 10.1093/nar/gkt1016

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096

Li, D., Liu, C., Shen, L., Wu, Y., Chen, H., Robertson, M., et al. (2008). A repressor complex governs the integration of flowering signals in *Arabidopsis*. *Dev. Cell* 15, 110–120. doi: 10.1016/j.devcel.2008.05.002

Li, W., Zhang, L., Ding, Z., Wang, G., Zhang, Y., Gong, H., et al. (2017). De novo sequencing and comparative transcriptome analysis of the male and hermaphroditic flowers provide insights into the regulation of flower formation in and romonoecious Taihangia rupestris. *BMC Plant Biol.* 17:54. doi: 10.1186/s12870-017-0990-x

Li, Y. H., Yang, Y., and Zhang, Y. (2016). Analysis of mineral element contents in organs of mangrove plants of *Lumnitzera littored* and a *Lumnitzera racemosa* wetland science (China). *China Acad. J.* 14, 433–438. doi: 10.13248/j.cnki.wetlandsci.2016.03.020

Litt, A., and Kramer, E. M. (2010). The ABC model and the diversification of floral organ identity. *Semin. Cell Dev. Biol.* 21, 129–137. doi: 10.1016/j.semcdb.2009.11.019

Liu, B., Ou, C., Chen, S., Cao, Q., Zhao, Z., Miao, Z., et al. (2019). Differentially expressed genes between carrot petaloid cytoplasmic male sterile and maintainer during floral development. *Sci. Rep.* 9:17384. doi: 10.1038/s41598-019-53717-x

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

Luo, M., Luo, M.-Z., Buzas, D., Finnegan, J., Helliwell, C., Dennis, E. S., et al. (2008). UBIQUITIN-SPECIFIC PROTEASE 26 is required for seed development and the repression of PHERES1 in Arabidopsis. *Genetics* 180, 229–236. doi: 10.1534/genetics.108.091736

Mateos, J. L., Madrigal, P., Tsuda, K., Rawat, V., Richter, R., Romera-Branchat, M., et al. (2015). Combinatorial activities of SHORT VEGETATIVE PHASE and FLOWERING LOCUS C define distinct modes of flowering regulation in *Arabidopsis*. *Genome Biol.* 16:31. doi: 10.1186/s13059-015-0597-1

Mertens, A., Brys, R., Schouppe, D., and Jacquemyn, H. (2018). The impact of floral morphology on genetic differentiation in two closely related biennial plant species. *AoB Plants* 10:ly051. doi: 10.1093/aobpla/ply051

Mohanty, J. N., and Joshi, R. K. (2018). Molecular cloning, characterization and expression analysis of MADS-box genes associated with reproductive development in Momordica dioica Roxb. *3 Biotech* 8:150. doi: 10.1007/s13205-018-1176-4

Moon, J., Suh, S. S., Lee, H., Choi, K. R., Hong, C. B., Paek, N.-C., et al. (2003). The SOC1 MADS-box gene integrates vernalization and gibberellin signals for flowering in *Arabidopsis*. *Plant J.* 35, 613–623. doi: 10.1046/j.1365-313x.2003.01833.x

Parenicova, L., De Folter, S., Kieffer, M., Horner, D. S., Favalli, C., Busscher, J., et al. (2003). Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in *Arabidopsis*: new openings to the MADS world. *Plant Cell* 15, 1538–1551. doi: 10.1105/tpc.011544

Polidoro, B. A., Carpenter, K. E., Collins, L., Duke, N. C., Ellison, A. M., Ellison, J. C., et al. (2010). The loss of species: mangrove extinction risk and geographic areas of global concern. *PLoS One* 5:e10095. doi: 10.1371/journal.pone.0010095

Rocheta, M., Sobral, R., Magalhaes, J., Amorim, M. I., Ribeiro, T., Pinheiro, M., et al. (2014). Comparative transcriptomic analysis of male and female flowers of monoecious Quercus suber. *Front. Plant Sci.* 5:599. doi: 10.3389/fpls.2014.00599

Rosas-Guerrero, V., Hernandez, D., and Cuevas, E. (2017). Influence of pollen limitation and inbreeding depression in the maintenance of incomplete dichogamy in Salvia elegans. *Ecol. Evol.* 7, 4129–4134. doi: 10.1002/ece3.2827

Scortecci, K. C., Michaels, S. D., and Amasino, R. M. (2001). Identification of a MADS-box gene, FLOWERING LOCUS M, that represses flowering. *Plant J.* 26, 229–236. doi: 10.1046/j.1365-313x.2001.01024.x

Streisfeld, M. A., Young, W. N., and Sobel, J. M. (2013). Divergent selection drives genetic differentiation in an R2R3-MYB transcription factor that contributes to incipient speciation in Mimulus aurantiacus. *PLoS Genet.* 9:e1003385. doi: 10.1371/journal.pgen.1003385

Su, G. H. (2004). *Study on Genetic Diversity in Lumnitzera of Mangrove*. Ph. D. Thesis, Sun Yat-sen University, Guangzhou.

Theissen, G., Melzer, R., and Ruempler, F. (2016). MADS-domain transcription factors and the floral quartet model of flower development: linking plant development and evolution. *Development* 143, 3259–3271. doi: 10.1242/dev.134080

Wang, Y., Li, L., Ye, T., Lu, Y., Chen, X., and Wu, Y. (2013). The inhibitory effect of ABA on floral transition is mediated by ABI5 in *Arabidopsis*. *J. Exp. Bot.* 64, 675–684. doi: 10.1093/jxb/ers361

Wellmer, F., Graciet, E., and Luis Riechmann, J. (2014). Specification of floral organs in *Arabidopsis*. *J. Exp. Bot.* 65, 1–9. doi: 10.1093/jxb/ert385

Yun, H., Hyun, Y., Kang, M. J., Noh, Y. S., Noh, B., and Choi, Y. (2011). Identification of regulators required for the reactivation of FLOWERING LOCUS C during *Arabidopsis* reproduction. *Planta* 234, 1237–1250. doi: 10.1007/s00425-011-1484-y

Zhang, Q., Liu, L., Zhu, F., Ning, Z., Hincke, M. T., Yang, N., et al. (2014). Integrating de novo transcriptome assembly and cloning to obtain chicken Ovocleidin-17 full-length cDNA. *PLoS One* 9:e93452. doi: 10.1371/journal.pone.0093452

Zhang, X. N., Zhong, C. R., Yan, T. L., and Zhang, Y. (2016). The germplasm resource rescue of endangered mangrove (Lumnitzera littorea (Jack.) Voigt) by artificial pollination. *Ecol. Sci.* 35, 38–42. doi: 10.14108/j.cnki.1008-8873.2016.05.006

Zhang, Y., Li, Y. H., Zhang, X. N., and Yang, Y. (2017). Flower phenology and breeding system of endangered mangrove Lumnitzera littorea (Jack.) Voigt. *Chin. J. Appl. Environ. Biol.* 23, 77–81. doi: 10.3724/SP.J.1145.2016.03021

Zhang, Y., Zhong, C., Li, S., Yan, T., and Guan, W. (2013). Endangered species of mangrove plants: *Lumnitzera littore*. *Forest Resour. Manag.* 5, 103–107. doi: 10.13466/j.cnki.lyzygl.2013.05.026

Zhang, Y., Zhong, C. R., Yang, Y., Zhong, H. B., Zeng, Z. P., Zhang, J., et al. (2018). Rescue of germplasm resources of endangered mangrove plant *Lumnitzera littorea*. *Mol. Plant Breed.* 14, 4112–4118. doi: 10.13271/j.mpb.016.004112

Zhong, C. R., Li, S. C., Guan, W., Li, H. L., Lin, X. Y., Bao-wen, L., et al. (2011). Current distributions of three endangered mangrove species in China. *Ecol. Sci.* 30, 431–435.

Zhou, Q. J., Chen, Y. M., Wu, W., Zhou, R. Z., and Zhang, Y. (2018). The complete chloroplast genome sequence of an endangered mangrove tree *Lumnitzera littorea* (Combretaceae). *Conserv. Genet. Resour.* 10, 911–913. doi: 10.1007/s12686-017-0929-4

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.

Copyright © 2020 Zhang, Chen, Zhou, Zhang, Bai and Zheng. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 09 December 2020
doi: 10.3389/fgene.2020.560368

Initial Complete Chloroplast Genomes of *Alchemilla* (Rosaceae): Comparative Analysis and Phylogenetic Relationships

Peninah Cheptoo Rono[1,2,3†], Xiang Dong[1,2,3†], Jia-Xin Yang[1,2†], Fredrick Munyao Mutie[1,2,3], Millicent A. Oulo[1,2,3], Itambo Malombe[4], Paul M. Kirika[4], Guang-Wan Hu[1,2] and Qing-Feng Wang[1,2]*

[1] CAS Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan, China, [2] Sino-Africa Joint Research Center, Chinese Academy of Sciences, Wuhan, China, [3] University of Chinese Academy of Sciences, Beijing, China, [4] East African Herbarium, National Museums of Kenya, Nairobi, Kenya

OPEN ACCESS

Edited by:
Zhiyong Liu, Shenyang Agricultural University, China

Reviewed by:
Joel Sharbrough, New Mexico Institute of Mining and Technology, United States
Anjil Kumar Srivastava, Durham University, United Kingdom

***Correspondence:**
Guang-Wan Hu
Huguangwanhu@wbgcas.cn; guangwanhu@sohu.com

[†] These authors have contributed equally to this work

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

Received: *08 May 2020*
Accepted: *15 October 2020*
Published: *09 December 2020*

Citation:
Rono PC, Dong X, Yang J-X, Mutie FM, Oulo MA, Malombe I, Kirika PM, Hu G-W and Wang Q-F (2020) Initial Complete Chloroplast Genomes of Alchemilla (Rosaceae): Comparative Analysis and Phylogenetic Relationships. Front. Genet. 11:560368. doi: 10.3389/fgene.2020.560368

The genus *Alchemilla* L., known for its medicinal and ornamental value, is widely distributed in the Holarctic regions with a few species found in Asia and Africa. Delimitation of species within *Alchemilla* is difficult due to hybridization, autonomous apomixes, and polyploidy, necessitating efficient molecular-based characterization. Herein, we report the initial complete chloroplast (cp) genomes of *Alchemilla*. The cp genomes of two African (*Afromilla*) species *Alchemilla pedata* and *Alchemilla argyrophylla* were sequenced, and phylogenetic and comparative analyses were conducted in the family Rosaceae. The cp genomes mapped a typical circular quadripartite structure of lengths 152,438 and 152,427 base pairs (bp) in *A. pedata* and *A. argyrophylla*, respectively. *Alchemilla* cp genomes were composed of a pair of inverted repeat regions (IRa/IRb) of length 25,923 and 25,915 bp, separating the small single copy (SSC) region of 17,980 and 17,981 bp and a large single copy (LSC) region of 82,612 and 82,616 bp in *A. pedata* and *A. argyrophylla*, respectively. The cp genomes encoded 114 unique genes including 88 protein-coding genes, 37 transfer RNA (tRNA) genes, and 4 ribosomal RNA (rRNA) genes. Additionally, 88 and 95 simple sequence repeats (SSRs) and 37 and 40 tandem repeats were identified in *A. pedata* and *A. argyrophylla*, respectively. Significantly, the loss of group II intron in *atpF* gene in *Alchemilla* species was detected. Phylogenetic analysis based on 26 whole cp genome sequences and 78 protein-coding gene sequences of 27 Rosaceae species revealed a monophyletic clustering of *Alchemilla* nested within subfamily Rosoideae. Based on a protein-coding region, negative selective pressure (Ka/Ks < 1) was detected with an average Ka/Ks value of 0.1322 in *A. argyrophylla* and 0.1418 in *A. pedata*. The availability of complete cp genome in the genus *Alchemilla* will contribute to species delineation and further phylogenetic and evolutionary studies in the family Rosaceae.

Keywords: *Alchemilla argyrophylla*, *A. pedata*, chloroplast genome, phylogenetic, Rosaceae

INTRODUCTION

The genus *Alchemilla* L. (Rosaceae), composed of more than 1,000 species, is important for its ornamental and medicinal values (Kaya et al., 2012; Boroja et al., 2018). It is widely distributed in the Holarctic region with high species richness in west Eurasia and few species found in montane tropical Asia, Madagascar, South Africa, and East African mountains (Izmailow, 1981; Fröhner, 1995; Gehrke et al., 2008). *Alchemilla*, together with *Aphanes* and *Lachemilla*, is classified under subfamily Rosoideae, tribe Potentilleae and subtribe Alchemillinae (Rothmaler, 1937; Kalkman, 2004; Notov and Kusnetzova, 2004; Soják, 2008; Xiang et al., 2016). Initially, *Alchemilla* was categorized under tribe Sanguisorbinae by Hutchinson (1964) due to superficial similarity of floral traits but was later reclassified as Potentilleae following Schulze-Menz's (1964) observational concern on the anther structure and further confirmation through molecular characterization using nuclear ribosomal DNA and *trnL/F* region of chloroplast (cp) DNA (Eriksson et al., 2003). Morphologically, *Alchemilla* is distinguished from other Rosaceae genera by the silvery-silky white hair covering on the stems and the leaf surface, achene fruits that are hidden within the calyx tube, inconspicuous individuality, and small but fairly showy inflorescence (Graham, 1960; Faghir et al., 2014; Gehrke et al., 2016). Notwithstanding, circumscription within the genus remain poorly understood due to hybridization, autonomous apomixes, and polyploidization dominant in *Alchemilla* species (Izmailow, 1982; Czapik, 1996; Gehrke et al., 2008). The genus *Alchemilla* is commonly used as perfect example of apomictic traits combined with morphological polymorphism in Rosaceae (Czapik, 1996; Bicknell and Koltunow, 2004; Hayirhoglu-Ayaz et al., 2006; Salamone et al., 2013). This has resulted in the existence of diverse micro-species and species complexes with variable indumentum, unstable flower characteristic structure, and heteroblastic plasticity, making *Alchemilla* a taxonomically difficult group (Hörandl, 2004; Hayirhoglu-Ayaz et al., 2006; Lundberg et al., 2009). This necessitated the need for complete cp genome in the genus *Alchemilla* to help resolve the taxonomic and phylogenetic uncertainties between its species.

Chloroplast (cp) is found in the cytoplasmic matrix of a plant cell and plays significant roles in photosynthesis, carbon fixation, and synthesis of starch, fatty acids, and amino acids (Daniell et al., 2016). It is a semi-autonomous organelle, similar to the nuclei and the mitochondria, essential in the transfer and expression of the plant's genetic material (Wolfe et al., 1987). The cp has its own double-stranded circular genome whose size in most terrestrial plants ranges between 120 and 180 kb, encoding about 110–130 different genes in a highly conserved order (Shinozaki et al., 1986; Liu et al., 2017). Based on its genome content and the ultrastructure features, the cp traces its origin to free-living cyanobacteria through a single event of endosymbiosis (Gray, 1989; Keeling, 2004). The cp genome has a characteristic quadripartite structure comprising two identical copies of inverted repeat (IRa/IRb) regions separated by the large single copy (LCS) and small single copy (SSC) regions (Zhao M.-L. et al., 2018). The two IR regions are paramount in defining the size and structure of the cp genome of land plants (Palmer et al., 1988). Variations in genome size could consequently be due to expansion/contraction or loss of one of the IR regions in some species. For instance, the loss of one copy of IR in *Taxus chinensis* var. *mairea* resulted in the reduction of its genome size (Zhang et al., 2014). The uniqueness of the cp is evident in its maternal inheritance, small size, conserved sequences, and simple structure (Kress et al., 2005; Parks et al., 2009; Liu et al., 2017). The evolutionary process in angiosperms is dependent on the conserved structure, gene content, and organization of the cp genome (Doyle et al., 1992; Saski et al., 2005). This makes it an appropriate candidate for plant taxonomy, and comparative genomic and evolutionary studies. The advent of the next-generation DNA sequencing technology (Shendure and Ji, 2008) has magnified the rate of cp genome sequencing reports since the process is simpler, cost-effective and fast, resulting in the expansion of the cp genetics and genomics (Daniell et al., 2016). Since the report of the first cp genome sequence of tobacco (*Nicotiana tabacum*) (Shinozaki et al., 1986), several species in the Rosaceae have had their complete cp genome deposited to the National Center for Biotechnology Information (NCBI) organelle genome database[1] (Salamone et al., 2013; Zhao Y. et al., 2018; Wang et al., 2020). However, none of the *Alchemilla* species genome has been sequenced to date.

In this study, the initial cp genomes of the genus *Alchemilla* are reported in *Alchemilla pedata* and *Alchemilla argyrophylla*. First, we obtained the complete cp genome of the two species and characterized the structure, gene content, and organization of each genome. Second, we establish the codon usage frequencies, simple sequence repeats (SSRs), regions of high sequence divergence, and nucleotide substitution rates. Finally, phylogenetic position was evaluated by comparative analysis based on 24 complete cp genomes and 78 protein-coding gene (PCG) sequences of Rosaceae species. Our results provide a reference for the resolution of *Alchemilla* species classification and facilitate elucidation of evolutionary and phylogenetic relationships in Rosaceae.

MATERIALS AND METHODS

DNA Extraction and Chloroplast Genome Sequencing

Plant materials of two *Alchemilla* species of *Alchemilla pedata* (voucher number SAJIT-001337) and *Alchemilla argyrophylla* (SAJIT-002399) were collected from Mt. Kenya, Kenya. Young leaves were sampled and immediately preserved using silica gel in plastic bags (Chase and Hills, 1991). The voucher specimens were deposited at the East African Herbarium (EA) in the National Museums of Kenya and at the Herbarium of Wuhan Botanical Garden, CAS (HIB) (China). The total genomic DNA of the two species was extracted from 0.5 g of the silica dried leaves using modified cetyltrimethylammonium bromide (CTAB) protocol

[1] https://www.ncbi.nlm.nih.gov/

(Doyle, 1991). Results were then sequenced based on the Illumina paired end technology platform at the Novogen Company (Beijing, China).

Genome Assembly and Genome Annotation

Genome assembly was performed using GetOrganelle v1.6.2d with default parameters (Jin et al., 2020). The GetOrganelle first filtered plastid-like reads, conducted the *de novo* assembly, purified the assembly, and finally generated the complete plastid genomes (Camacho et al., 2009; Bankevich et al., 2012; Langmead and Salzberg, 2012). *K*-mer gradients for a mean and maximum 150-bp reads were set to as "-k 21, 45, 65, 85,105" for both species. Bandage (Wick et al., 2015) was used to visualize the final assembly graphs to authenticate the automatically generated plastid genome. The best fit k-mer of 45 was selected for use in a subsequent analysis of the genomes. The quality of the newly assembled genomes was evaluated on read level basis by aligning the trimmed raw reads to the *de novo* assemblies using Geneious mapper, Geneious version 9.1.4 (Kearse et al., 2012) with medium- to low-sensitivity option and iteration up to five times (Köhler et al., 2020). Gene annotation was conducted using Plastid Genome Annotator (PGA) (Qu et al., 2019) with an annotated plastome *Amborella trichopoda* (GenBank accession no. GCA_000471905.1) as the initial reference genome. Further annotation confirmation was performed with published genomes of *Fragaria virginiana* (JN884817) and *Fragaria vesca* subsp. *vesca* (KC507760) in the Rosacea family. Geneious was used to manually correct and complement problematic annotations. The whole genome circular map was drawn using Organelle Genome DRAW software (Lohse et al., 2013).

Comparative Genome Analysis and Sequence Divergence

The cp genome sequence of the *A. pedata* and *A. argyrophylla* was compared with that of eight other Rosaceae species. This included *Dasiphora fruticosa* (MF683841), *Fragaria iinumae* (KC507759), *Fragaria nipponica* (KY769125), *Fragaria pentaphylla* (KY434061), *Fragaria orientalis* (KY769126), *Fragaria chiloensis* (JN884816), *F. virginiana* (JN884817), *Fragaria mandshurica* (KC507760), *F. vesca* (KC507760), *Rosa multiflora* (NC_039989), *Rosa odorata* var. *gigantea* (KF753637), and *Hagenia abyssinica* (KX008604) retrieved from GenBank database. The mVISTA (Mayor et al., 2000) program under the shuffle-LAGAN alignment strategy (Frazer et al., 2004) was applied to compare all the complete genomes with *A. argyrophylla* as reference. The contraction and expansion of the IR boundaries of five Rosaceae species were visualized using the IRscope software (Amiryousefi et al., 2018).

Repeat Analysis and Codon Usage

REPuter online program (Kurtz et al., 2001) was used to identify long repeat sequences (forward, reverse, complementary, and palindromic). Repeat sequence locations and sizes in cp genomes were visualized with a minimal criterion of 30 bp, a hamming distance of 3, and less than 90% identity between two repeat copies. Tandem repeat finder (Benson, 1999) was used to identify tandem repeats in the two species *A. pedata* and *A. argyrophylla* cp genomes with default alignment parameters of match, mismatch, and insertions and deletions (indels) of 2, 7, and 7, respectively. SSRs were detected using the perl script MicroSAtellite (MISA) (Thiel et al., 2003) with a size motif of one to six nucleotides and a threshold of 10, 5, 5, 3, and 3 for mono-, di-, tri-, tetra-, penta-, and hexa-nucleotide, respectively. The codon usage frequency and relative synonymous codon usage (RSCU) of the two species were conducted based on 88 PCGs using MEGA 5 (Tamura et al., 2011).

Adaptive Evolution and Substitution Rate Analysis

To evaluate the evolutionary rate variation within the *Alchemilla* species, 78 protein-coding regions within the cp genomes were explored with *F. virginiana* as reference species. PCGs were extracted using Geneious var. 9.1.4. Gaps and stop codons were manually removed, and the seven sequences were separately aligned using MAFFT v7. 308 (Katoh and Standley, 2013). The aligned files were converted into AXT format using the parseFastaIntoAXT.pl Perl script[2]. The non-synonymous (Ka) and synonymous (Ks) substitution rates as well as Ka/Ks ratio of each gene were estimated using the software KaKs_calculator 1.2 using the default model averaging (MA) method (Zhang et al., 2006). Taking into consideration that KaKs_calculator uses model averaging estimates in site selection, we implored the impact of site selection in 78 genes of five species phylogenetically related to *Alchemilla*. Positive selective pressure within shared genes of the seven species of subfamily Rosoideae was evaluated using PAML v4.7 (Yang, 2007) package implemented in EasyCodeML software (Gao et al., 2019). Non-synonymous (dN) and synonymous substitution (dS) substitution rates, and their ratio (ω = dN/dS) were calculated based on four site-specific models (M0 vs. M3, M1a vs. M2a, M7 vs. M8, and M8a vs. M8) with likelihood ratio test (LRT) threshold of $p < 0.05$ elucidating adaptation signatures within the genome. The models permit dN/dS variation within sites while keeping the ω ratio fixed within branches. Selective pressure analysis was conducted along ML tree in plain Newick format based on protein-coding sites used in the generation of phylogenetic relationship of the selected seven species. Here, individual coding DNA sequences (CDSs) were aligned in correspondence to their amino acids and their selection evaluated based on both ω and LRT values.

Phylogenetic Analysis

Phylogenetic relationship analysis was done based on two data sets: (1) complete cp genome sequences (genomic tree) and (2) PCGs (CDS tree). The complete cp genomes of *A. pedata* and *A. argyrophylla* obtained from this study and other 24 cp genomes of Rosaceae species downloaded from NCBI database were inferred for the genomic tree (**Supplementary Table S1**). Multiple sequence alignment was performed using MAFFT

[2]http://code.google.com/p/kaks-calculator/

v7.308 with default parameters setting. Phylogenetic relationship reconstructions were performed based on maximum-likelihood (ML) analysis using the program IQ-Tree v.6.1 (Nguyen et al., 2015) with 1,000 bootstrap replications. The best fit model TVM+I+G4 (Kalyaanamoorthy et al., 2017) was chosen according to Bayesian information criterion (BIC). The CDS tree was constructed by ML, PhyML, and BI methods based on 78 PCGs shared by all the 27 species under comparative evaluation. Gene sequences were extracted and aligned individually using Mega 7 and concatenated into a single file using PhyloSuite (Zhang et al., 2020). The ML CDS tree phylogenies were inferred using IQ-Tree with the best fit model GTR+F+R2 from ModelFinder in accordance to Akaike information criterion (AIC). Bayesian inference (BI) phylogenetic relationship of our taxa was constructed using MrBayes 3.2.6 (Ronquist et al., 2012) in PhyloSiute under GT+F+I+G4 best fit model from ModelFinder (two runs, 200,000 generations) following the AIC. The online program PhyMl vl 3.0 (Guindon et al., 2010) was used to infer phylogenetic relationship following GTR+G+I model selected by sms (Lefort et al., 2017). The constructed tree was visualized using FigTree version 1.4.4 (Rambaut, 2018).

RESULTS

The Chloroplast Genome Structure and Content

The total number of assembled reads was 3,383,185 and 1,158,253 with an average genome coverage depth of 3,299.6 and 1,133.5 in *Alchemilla pedata* and *Alchemilla argyrophylla*, respectively. The complete cp genomes display a typical circular structure with DNA sizes 152,438 and 152,427 bp for *A. pedata* and *A. argyrophylla*, respectively (**Figure 1A**). The quadripartite structure is composed of 82,612 bp in the LSC region, 17,980 bp in the SSC region, and the two IR regions made up of 25,923 bp each in *A. pedata*. *A. argyrophylla* has 82,616 bp in the LSC, 17,981 bp in the SSC region, and the two IR regions, each having 25,915 bp in length. The overall guanine–cytosine (GC) content for both *Alchemilla* species are 37% with 42.7, 34.9, and 30.4% in the IRs, LSC, and SSC, respectively. These are similar to other complete cp genome sequences of Rosaceae species obtained from NCBI database ranging between 36.6 and 37.1% (**Table 1**). To assess for any mis-assemblies, the raw reads were aligned against the *de novo* assembled genomes. The high per-base read coverages, plotted against the genomes position for each of the assemblies (**Figures 1B,C**), reveal the quality of our assembly. The annotated cp genomes were deposited in the GenBank database with the following accession numbers; *A. argyrophylla* MT382661 and *A. pedata* MT382662.

A total of 114 unique genes were annotated including 88 PCGs, 37 transfer RNA (tRNA) genes, and 4 ribosomal RNA (rRNA) genes in both *A. argyrophylla* and *A. pedata* (**Table 1**). Based on their gene functional category, 59 genes were associated with self-replication while 44 genes are responsible for photosynthesis (**Table 2**). Similar gene order and genome structure were reported in both *Alchemilla* species (**Figure 1A**). The IR regions (IRa and IRb) had 18 duplicate genes comprising seven PCGs (*rpl2*, *rpl23*, *ycf2*, *ycf15*, *ndhB*, *rps7*, and *rps12*), seven tRNA (*trnI-CAU*, *trnL-CAA*, *trnV-GAC*, *trnI-GAC*, *trnA-UGC*, *trnR-ACG*, and *trnN-GUU*), and four rRNAs (*rrn16*, *rrn23*, *rrn4.5*, and *rrn5*). The SSC region had 13 genes of which 12 were PCGs and one tRNA, whereas the LSC had 62 PCGs and 22 tRNA (**Figure 1A**). In total, 15 genes (*trnK-UUU*, *rps16*, *trnG-UCC*, *rpoC1*, *trnL-AUU*, *trnV-UAC*, *petB*, *petD*, *rpl16*, *rpl2*, *ndhB*, *trnA-UGC*, *ndhA*, *trnA-UGC*, and *trnI-GAU*) had one intron with *rpl2* and *ndhB* duplicated in the IR, whereas two genes, *clpP* and *ycf3*, had two introns. The *rps12* is a trans-spliced gene, with one exon shared between two introns, in which the 3′ exons were duplicated in the IR regions and the 5′ exon end situated in the LSC region (**Table 2**). Among the 133 genes, three instances of overlapping sequences were detected in *Alchemilla*. The *psbD* and *psbC* genes shared coding regions (53 bp); *ycf68* gene was embedded within *trnI-GAU* in one of the inverted region (IRa); and *matK*, which has the longest intron (2,523 bp in *A. argyrophylla* and 2,528 bp in *A. pedata*), was embedded in *trnK-UUU* in the SSC region (**Figure 1A**).

Codon Usage Analysis

The CDSs of the cp genomes were used to estimate the frequency of codon usage of both *A. pedata* and *A. argyrophylla*. A total of 22,948 and 22,984 codons encoding 88 genes were detected in *A. argyrophylla* and *A. pedata*, respectively. Of all the codons, leucine reported the highest amino acid usage frequency of 10.57% (2,426) in *A. argyrophylla* and 10.58% (2,432) in *A. pedata*, whereas cysteine had the lowest amino acid usage frequency of 1.09% (251) in *A. argyrophylla* and 1.08% (249) in *A. pedata* (**Supplementary Figure S1**). The most frequently used codons are AUU (993 and 996), AAA (938 and 940), and GAA (922 and 921) encoding isoleucine, lysine, and glutamic acid in *A. argyrophylla* and *A. pedata*, respectively (**Supplementary Table S2**). Furthermore, RSCU value was estimated in the 88 CDSs in both *Alchemilla* species. RSCU is the ratio between the expected frequency of use and the actual frequency usage of a particular codon. Codons reporting RSCU value < 1 indicates lower frequency usage than expected, while a score > 1 signifies higher usage frequency (Sharp and Li, 1987; Munyao et al., 2020). In *Alchemilla* species, apart from the stop codon (UGA), isoleucine (I) codon AUA and leucine (L) codon (CUA) with RSCU value below 1, all the other codons with synonymous codons usage (RSCU > 1) preferred to end with A or U in both *A. pedata* and *A. argyrophylla* signifying their preferential codon use. Exceptionally, codon UUG encoding leucine recorded higher bias of RSCU = 1.19 in both species despite ending with G in the third position than did other codons of low frequencies (RSCU < 1) that ended in C or G. Codons AUG (M) and UGG (W) encoding methionine and tryptophan showed no bias (RSCU = 1) (**Supplementary Table S2**). Our findings are consistent with majority cp genomes of land plants (Cheng et al., 2017). Due to usage frequency variation, RSCU values of the cp genome form a valuable source of evolutionary signature traits resulting from mutation and selection that are

essential in studying the evolution of an organism (Morton, 2003; Wang et al., 2018).

Repeat Structure and Simple Sequence Repeats

Repeat motifs are significant in the computation of phylogenetic and genomic rearrangement (Cavalier-Smith, 2002). In this report, *A. pedata* reported 36 long repeats comprising four palindromic (P), 14 reverse (R), and 18 forward (F) repeats, whereas *A. argyrophylla* recorded 42 long repeats composed of 19 P, 8 R, and 15 F repeats (**Figure 2**). In both species, complementary repeats were not found. This is similar to finding obtained in other Rosaceae species (Gichira et al., 2017). The comparative analysis results revealed that most repeats were between 30 and 40 bp. The longest repeat was a palindromic repeat having 71 bp located in the intergenic spacer (IGS) of the LSC region between *trnM-CAU* and *atpE* in both species (**Supplementary Table S3**). Most repeats were distributed in LSC (non-coding) region, whereas some were found in genes including *ndhA*, *ycf3*, *ycf1*, *rpoC1*, *rpl16*, and *ndhB*. Six repeats were found exclusively in the IR regions, three of them relating to *ndhB* gene (**Supplementary Table S3**). The number of tandem repeats

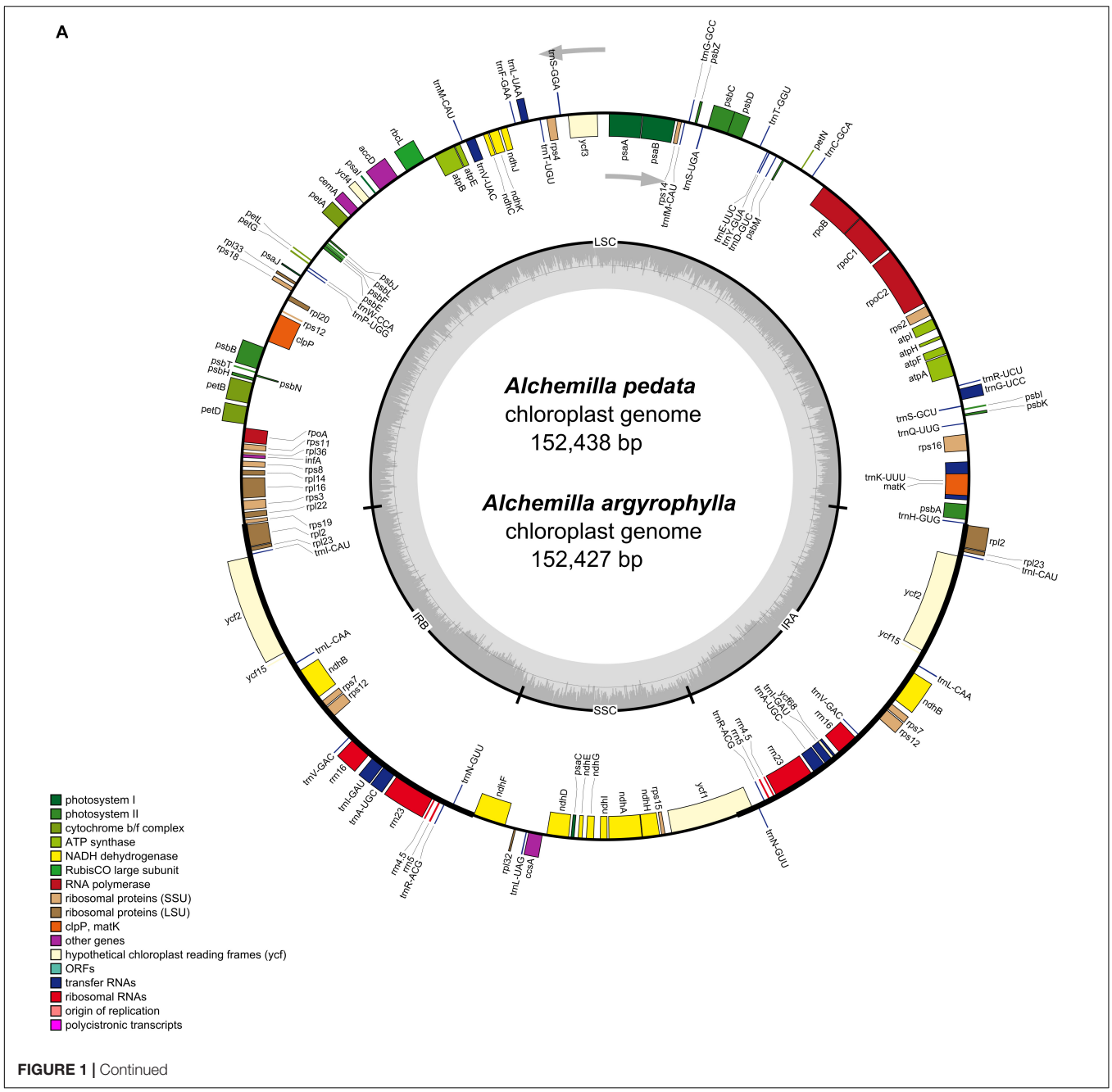


FIGURE 1 | Continued

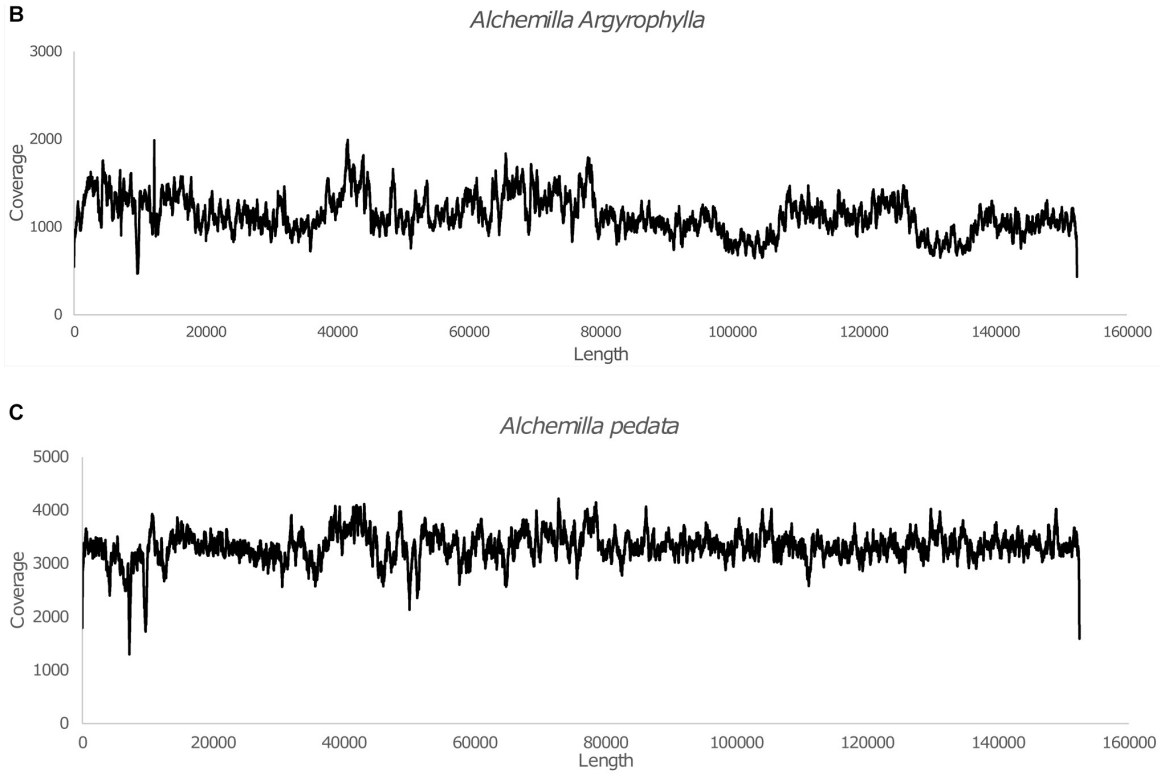


FIGURE 1 | (A) The genome map of *Alchemilla pedata* and *Alchemilla argyrophylla* drawn using Organelle Genome DRAW software (Lohse et al., 2013). The genes outside the circle are transcribed in the counterclockwise direction, while those inside are transcribed in the clockwise direction. The colored bars indicate different functional groups. The lighter gray color denotes the AT content, while the gray area in the inner circle corresponds to the guanine–cytosine (GC) content of the genome. IRA, inverted repeat region A; IRB, inverted repeat region B; LSC, large singe copy; SSC, small single copy. **(B)** *A. argyrophylla* and **(C)** *A. pedata* graphs of read level coverage plots. The *de novo* assembly reads were trimmed, aligned, and mapped to their chloroplast genome after initial alignment with Bowtie in GetOrganelle using Geneious mapper (Kearse et al., 2012).

TABLE 1 | Comparison of chloroplast genomes of 14 Rosaceae species.

Species	GenBank Acc. No.	Size (bp)	GC (%)	Protein	rRNA	tRNA	Gene
Alchemilla argyrophylla	MT382661	152,427	37	88	8	37	133
Alchemilla pedata	MT382662	152,438	37	88	8	37	133
Dasiphora fruticosa	MF683841	152,931	37.2	84	8	37	130
Fragaria iinumae	KC507759	155,554	37.1	85	8	37	130
Fragaria nipponica	KY769125	148,592	37.6	84	8	35	127
Fragaria pentaphylla	KY434061	155,640	37.3	85	8	35	129
Fragaria orientalis	KY769126	147,835	37.6	84	8	36	128
Fragaria chiloensis	JN884816	155,603	37.2	85	8	37	130
Fragaria virginiana	JN884817	155,621	37.1	85	8	37	130
Fragaria mandshurica	KC507760	155,596	37.2	85	8	37	130
Fragaria vesca	KC507760	155,691	37.2	85	8	37	130
Rosa multiflora	NC_039989	156,592	37.2	90	8	37	136
Rosa odorata var. *gigantea*	KF753637	156,634	37.2	88	8	40	139
Hagenia abyssinica	KX008604	154,961	37.1	85	8	37	129

GC, guanine–cytosine; rRNA, ribosomal RNA; tRNA, transfer RNA.

was 40 in *A. argyrophylla* and 37 repeats in *A. pedata* (**Figure 2**). In *A. argyrophylla*, only four repeats were > 30 bp in length, while the rest were between 1 and 28 bp. Of these repeat units, 22 repeats had mismatches, and 15 had indels (**Supplementary Table S4a**). In *A. pedata*, two repeats were > 30 bp, and the rest were between 9 and 28 bp in length. Sixteen repeat units reported mismatches, and 11 had indels (**Supplementary Table S4b**).

TABLE 2 | Genes present and functional gene category in *Alchemilla pedata* and *Alchemilla argyrophylla* chloroplast genome.

Functional category	Group of genes	Name of genes
Transcription and translation related genes (self-replication)	DNA-dependent RNA polymerase	*rpoA*, *rpoB*, **rpoC1*, *rpoC2*
	Ribosomal proteins (SSU)	*rps2*, *rps3*, *rps4*, ***rps7***, *rps8*, *rps11*, **rps12*, *rps14*, *rps15*, **rps16*, *rps18*, *rps19*
	Ribosomal protein (LSU)	***rpl2***, *rpl14*, **rpl16*, *rpl20*, *rpl22*, ***rpl23***, *rpl33*, *rpl32*, *rpl36*
RNA genes (self-replication)	Ribosomal RNA	***rrn4.5***, ***rrn5***, ***rrn16***, ***rrn23***
	Transfer RNA	***trnA-UGC***, *trnC-GCA*, *trnD-GUC*, *trnE-UUC*, *trnF-GAA*, *trnfM-CAU*, *trnG-GCC*, **trnG-UCC*, *trnH-GUG*, *trnI-CAU*, ***trnI-GAU***, **trnK-UUU*, ***trnL-CAA***, **trnL-UAA*, *trnL-UAG*, *trnM-CAU*, ***trnN-GUU***, *trnP-UGG*, *trnQ-UUG*, ***trnR-ACG***, *trnR-UCU*, *trnS-GCU*, *trnS-GGA*, *trnS-UGA*, *trnT-GGU*, *trnT-UGU*, ***trnV-GAC***, **trnV-UAC*, *trnW-CCA*, *trnY-GUA*
Photosynthesis related	Rubisco	*rbcL*
	Photosystem I	*psaA*, *psaB*, *psaC*, *psaI*, *psaJ*
	Photosystem II	*psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbN*, *psbT*, *psbZ*
	ATP synthase	*atpA*, *atpB*, *atpE*, *atpF*, *atpH*, *atpI*
	Cytochrome b/f complex	*petA*, **petB*, **petD*, *petG*, *petL*, *petN*
	NADPH dehydrogenase	**ndhA*, ****ndhB***, *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *ndhJ*, *ndhK*
Other genes	Maturase	*matK*
	cytochrome *c*-type synthesis	*ccsA*
	Carbon metabolism	*cemA*
	Fatty acid synthesis	*accD*
	Transfer initiation factor	*infA*
	Proteolysis	***clpP*
Genes of unknown function/Hypothetical Protein RF	Conserved open reading frames	*ycf1*, ***ycf2***, ***ycf3*, *ycf4*, ***ycf68, ycf15***

Genes in bold are located within the inverted repeat (IR) and therefore in duplicates except ycf68.
**Genes have one intron.*
***Genes have two introns.*

SSRs, also called short tandem repeats or microsatellites, are repeating sequences of about 1–6 bp that are uniparentally inherited and widely distributed in the whole cp genome (Cheng et al., 2015). SSRs are ideally co-dominant, having the highest degree of intraspecific polymorphism (Weber, 1990), high mutation rates, locus specificity, and multi-allelism (Kuang et al., 2011; Asaf et al., 2017). Thus, the microsatellites are valuable markers ideal for molecular breeding (Rafalski and Tingey, 1993), population genetics (Powell et al., 1995), gene mapping, and genetic linkage analysis (Pugh et al., 2004; Xue et al., 2012). In our study, a total 95 SSRs were identified in *A. argyrophylla* composed of 70 mononucleotides, 16 dinucleotides, 5 trinucleotides, and four tetranucleotides (**Figure 3** and **Table 3**). Similarly, *A. pedata* cp genome had 88 SSRs composed of 62 mononucleotides, 17 dinucleotides, 8 trinucleotides, and 3 tetranucleotides (**Figure 3** and **Table 3**). In both species, mononucleotides were the most abundant repeat types (*A. argyrophylla* 73.68% and *A. pedata* 70.45%). Pentanucleotides and hexanucleotides were not detected in both species (**Figure 3**). Apart from one mononucleotide, all the other SSRs were rich in A and T (**Table 3**). These findings are consistent with contention that SSRs are typically composed of polyadenine (PolyA) and polythyamine (PolyT) repeats in line with previous reports (Cheng et al., 2015; Shen et al., 2016). This perpetually contributes to biasness in base composition of the whole cp genome, where A/T content in the reported *Alchemilla* species is 62.98% in *A. argyrophylla* and 62.99% in *A. pedata* compared with the GC content represented by 37.0% in both species.

Comparative Genome Analysis

For further analysis of the cp genome of *Alchemilla* species, 12 whole cp genome sequences of the Rosaceae species were downloaded from NCBI, and the basic genomic characteristics were compared (**Table 1**). A high similarity was observed in all genome sequences. The genome size ranged from 148,592 bp in *Fragaria nipponica* to 156,634 in *Rosa odorata* var. *gigantea* and was concurrent with the lowest and highest number of genes that ranged from 127 to 139, respectively. The number of PCGs ranged from 84 in *Fragaria orientalis* to 90 in *Rosa multiflora*, whereas tRNA genes ranged from 35 in *F. nipponica* to *Fragaria pentaphylla* to 40 in *R. odorata* var. *gigantea.* The cp genome structure and gene arrangement are conserved in all the species evaluated (**Table 1**).

To ascertain divergence within the cp genome, multiple alignment analysis of *Alchemilla* species and seven other Rosacea cp genomes were conducted using mVISTA program. Results showed that IR regions have higher similarity than the SC regions (**Supplementary Figure S2**). Higher conservation was observed in the coding regions than the non-coding regions, which is a common phenomenon in most angiosperms (Cheng et al., 2017). Significantly, the most conserved regions were observed in the tRNA and rRNA regions across all species. High variation

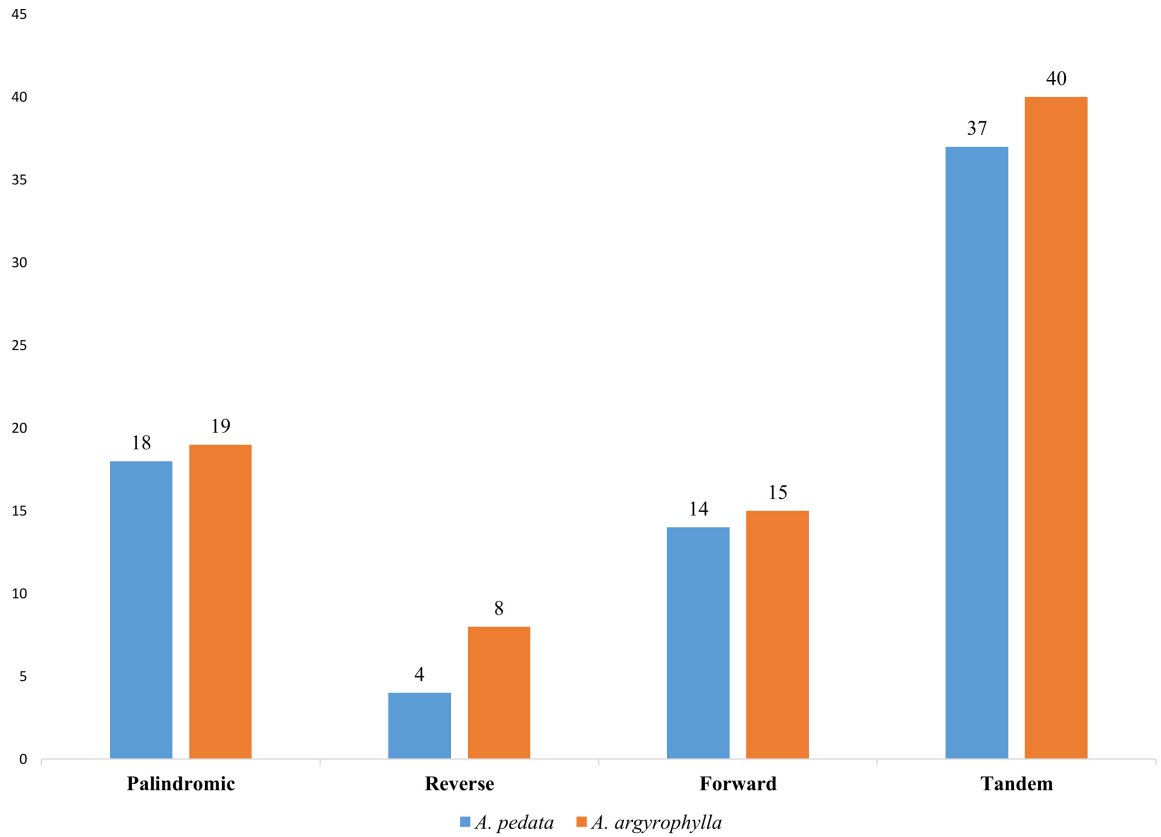


FIGURE 2 | Number of long repeat sequence types in *Alchemilla pedata* and *Alchemilla argyrophylla* in the distribution of palindromic, reverse, forward, and tandem repeats.

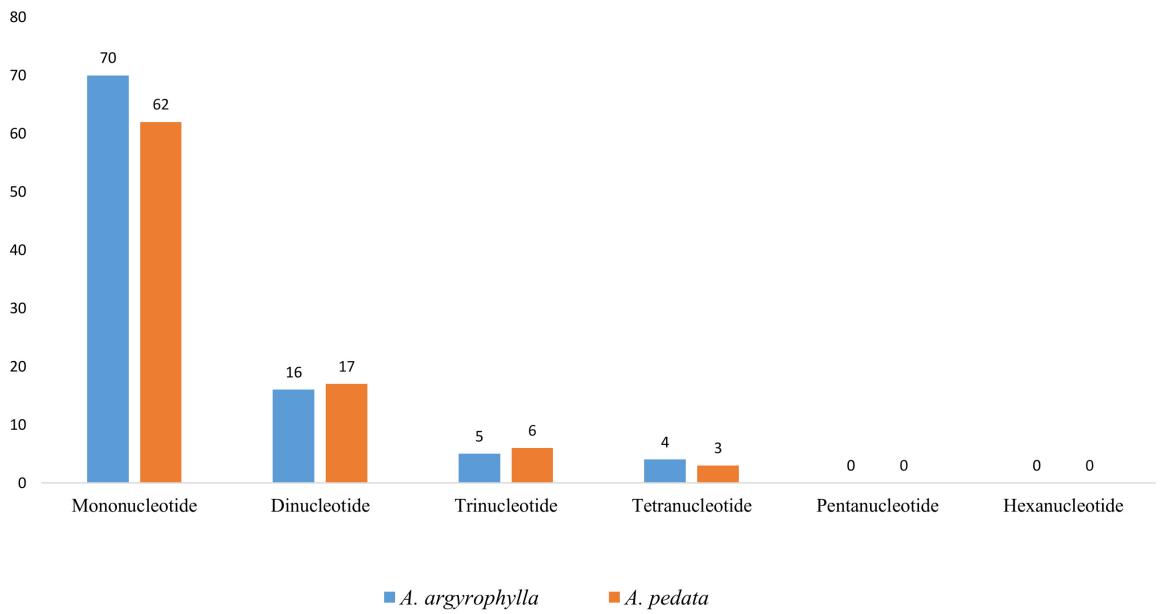


FIGURE 3 | Number of different simple sequence repeat (SSR) units in *Alchemilla argyrophylla* and *Alchemilla pedata* comprising mono-, di-, tri, tetra-, penta-, and hexa-nucleotide repeats.

TABLE 3 | Number of different simple sequence repeats (SSRs) in *Alchemilla argyrophylla* and *Alchemilla pedata*.

SSRs	*A. argyrophylla*	*A. pedata*
A/T	69	61
C/G	1	1
AT/TA	16	17
AAT/ATT	5	6
AAAT/ATTT	3	2
AATT/AATT	1	1
Total no. of SSRs	95	88

SSRs, simple sequence repeats.

was observed in the IGS regions of *trnH-GUG–psbA*, *trnK-UUU–rps16*, *rps16–trnQ-UUG*, *trnR-UCU–atpA*, *ndhC–trnV-UAC*, *atpF–atpH*, *trnC-GCA–petN*, *petN–psbM*, *trnD-GUC–trnY-GUA*, *psaA–ycf3*, *rps4–trnT-UGU*, *trnL-UAA–trnF-GAA*, *ndhC–trnV-UAC*, *petD–rpoA*, *rps11–rpl36*, *trnL-CAA–ndhB*, and *PetA–psbJ* in the LSC. Non-coding regions in *ndhF–rpl32*, *rpl32–trnL-UAG*, *ccsA–ndhD*, and *rps15–ycf1* reported high divergence in the SSC. Coding regions with the highest variation include *ycf1*, *ndhF*, *rpoA*, *infA*, *accD*, *rpoC2*, and *matK*. Divergence was also detected in introns of *trnK-UUU*, *rps16*, *ycf3*, *petD*, *rpl16*, *clpP*, *rps12*, and *ndhA*. These are regions of rapid evolutionary changes and therefore are essential sites for the development of molecular markers that could be useful in population genetics and phylogenetic studies. Our results are consistent with findings of other Rosaceae species (Shen et al., 2016; Jian et al., 2018). Generally, *A. pedata* and *A. argyrophylla* are highly similar to *Dasiphora* and *Fragaria* species and most divergent from *Pentactina* and *Prinsepia* species among the seven evaluated species (**Supplementary Figure S2**).

Expansion and Contraction of the Inverted Repeat Regions

The IR boundaries of *A. pedata* and *A. argyrophylla* were compared with those of five other species of Rosaceae to analyze probable expansion or contraction in the IR (**Figure 4**). Despite the IRs being the most conserved region of the cp genome, constant variation in the position of the IR/SC boundary and their associated adjacent genes observed in plant lineages has been because of the contraction and expansion of the IR region, which subsequently acts as an evolutionary indicator (Wang et al., 2008). Our results showed that the different species had varied IR sizes, ranging from 25,311 bp in *Dasiphora fruticosa* to 26,053 bp in *R. odorata* (**Figure 4**). The *rpl22* and *rps19* genes lied exclusively in the LSC region adjacent to the LSC/IRb junction, while the *rpl19* shifted away from the LSC/IRb boundary with gap of 4–14 bp. The *ndhF* gene in the analyzed species was located entirely in the SSC region having varied distances from the IRa/SSC boarder (JBL). However, *Hagenia abyssinica* had the *ndhF* gene stretching 21 bp into IRb region. The *ycf1* gene stretched through the SSC/IRa boarder (JSA) in all the species at varied lengths. The *trnH* gene located entirely in the LSC region stretched 3–34 bp away from the IRa/LSC junction in all the analyzed species. Generally, the *trnH* gene in monocots is located in the IR region, while that in dicots is located in the LSC region (Asano, 2004; Cheng et al., 2017).

Adaptive Evaluation Analysis

Non-synonymous (Ka) and synonymous (Ks) substitutions and their proportional ratios (Ka/Ks) similarly referred to as (dN/dS) have been used to evaluate nucleotide's natural selection pressure and evolution rates (Ninio, 1984; Yang and Nielsen, 2000). In most protein-coding regions, occurrence of synonymous substitutions has been reported more frequently than occurrence of non-synonymous substitutions (Makałowski and Boguski, 1998). The synonymous substitutions normally do not alter the amino acid chain unlike the non-synonymous substitutions that change the amino acid sequence. In this study, Ka and Ks values were estimated in 78 genes of the *A. pedata* and *A. argyrophylla*, computed against a close relative *Fragaria virginiana* (**Supplementary Table S5**). In our evaluation, none of the genes reported Ka value above 1 of which *ycf1* PCG reported the highest value (Ka = 0.0416) in *A. pedata* and (Ka = 0.0411) in *A. argyrophylla*. On the other hand, the highest Ks value was recorded in photosynthesis gene *psbD* (Ks = 1.2010) in *A. pedata* and *petL* (Ks = 0.2642) in *A. argyrophylla*.

The Ka/Ks value indicates the intensity of selective pressure imposed on a particular gene. Neutral selection is denoted by a Ka/Ks value of 1, Ka/Ks ratio < 1 signifies negative (purifying) selection, and Ka/Ks ratio > 1 indicates positive (adaptive) selection (Nei and Kumar, 2000). Purifying selection is common in many protein-coding regions (Nielsen, 2005). In this study, most of the genes had Ka/Ks ratio of less than 0.5, accounting for over 90% of the analyzed genes. However, the high Ka/Ks values were noted in *rps7* (Ka/Ks = 50), *rpl23* (Ka/Ks = 50), and *psbJ* (Ka/Ks = 47) in *A. argyrophylla* and *rpl32* (Ka/Ks = 50) and *psbJ* (Ka/Ks = 47) in *A. pedata* due to very low Ks value < 0.001 implying 0 synonymous changes in the genes. This means that there was very low or no substitutions (NA) between the aligned gene sequences (Mo et al., 2020). We therefore replaced the high Ka/Ks values in these genes with 0. The average Ka/Ks value was found to be 0.1322 in *A. argyrophylla* and 0.1418 in *A. pedata*, signifying an overall negative selection pressure of the genes (**Supplementary Table S5**). Genes with Ka/Ks > 0.5 included *petN*, *psbL*, and *psbN* in *A. argyrophylla* and *petN*, *psbD*, *psbL*, and *PsbN in A. pedata*. In both *Alchemilla* species, the least Ka/Ks value (0.0010) was recorded in photosynthesis-related genes (*atpH*, *ndhI*, *petD*, *petD*, *petG*, *petL*, *psaC*, *psbA*, *psbH*, *psbI*, *psbM*, and *psbT*) and self-replicating genes (*rps2*, *rps19*, and *rps36*), indicating significant purifying selection (**Supplementary Table S5**). The same functional protein-coding sequences in seven Rosoideae species were used to detect sites of positive selection. Among the four models, comparative LRT of M7 vs. M8 was positive in determining p-value of chi square < 0.05 and the selection strength. Bayes empirical Bayes (BEB) (Yang et al., 2005) and naïve empirical Bayes (NEB) analyses were implemented in model M8. In the BEB method, three sites were detected as site of positive selection, which represented one photosynthesis-related gene *ndhB*, self-replication gene *rpoC1*, and hypothetical gene *ycf2* (**Table 4**). NEB method on the other hand detected 59 sites that coded for 17 genes under selective pressure. Among

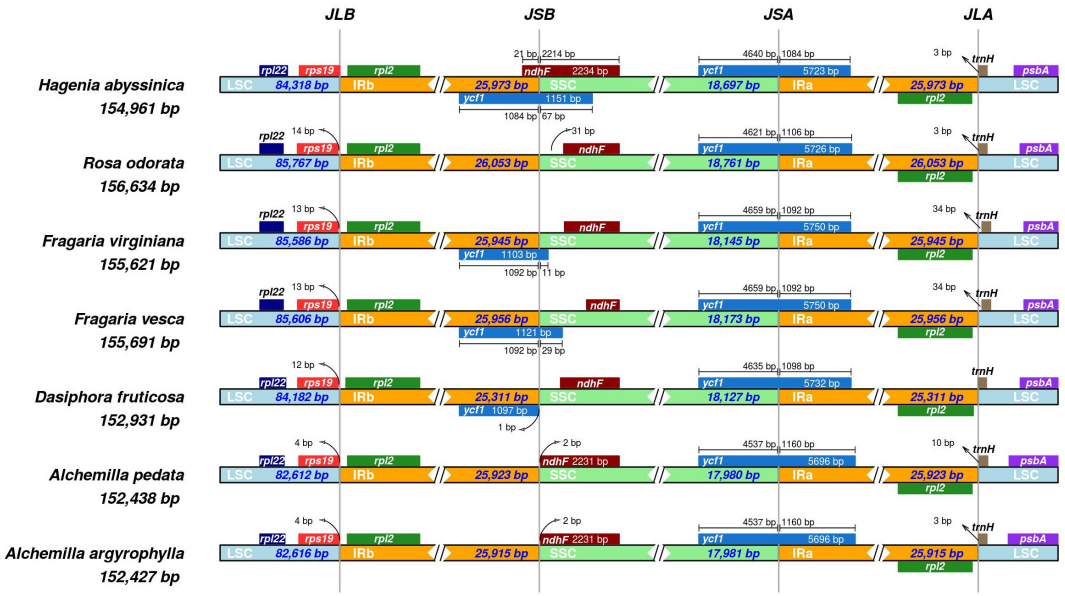


FIGURE 4 | Comparative distance between the boundaries of the two inverted repeats (IRa/IRb), small single copy (SSC), and large single copy (LSC) regions and adjacent genes among the chloroplast genomes of seven Rosaceae species using IRscope software (Amiryousefi et al., 2018). The species name and their corresponding genome lengths are shown on the left side of the figure. JLB, JSB, JSA, and JLA correspond to LSC/IRb, IRb/SSC, SSC/IRa, and IRa/LSC junctions, respectively. Genes drawn above the track indicate direct transcription, and genes below the track indicate complement transcription. The arrows indicate the distance from the beginning or ending coordinate of the specific gene from the corresponding junction; AT bar above or below the gene extending into two regions shows to what extent in base pairs it has stretched. The figure is not drawn to scale based on sequence length but only shows the relative change near or at the IR/SC junctions.

the genes, *rpoc2*, *ycf2*, and *ndhB* had $p > 0.99\%$. The generally slow evolutionary rates and subsequent low Ka/Ks ratio observed in *Alchemilla* species is a common attribute of the cp genome. The varying results of Ka/Ks ratio obtained in our study give evidence that evolutionary rates of cp genomes vary among genes. Similar conclusions were drawn by Menezes et al. (2018) in the cp genome analysis of Malpighiales.

Phylogenetic Analysis

In order to understand the evolutionary relationship among Rosaceae species, complete cp genome sequences and 78 PCGs in 27 species of subfamily Rosoideae and Amygdaloideae were used to infer the phylogenetic position of *A. argyrophylla* and *A. pedata* with *Elaeagnus macrophylla* (Elaeagnaceae) and *Morus indica* (Moraceae) as outgroups. The cp genome sequences and its PCGs provide precise and systematic genomic information for phylogenetic and evolutionary relationship reconstruction (Yang X. et al., 2020). Phylogenetic analysis was conducted using ML, BI, and PhyML methods (**Figures 5, 6**). A comparison between the CDS tree and the genome tree revealed an overall similar topology with few incongruences observed in subfamily Amygdaloideae (**Figure 6**). In both trees, *Alchemilla* species were found to be closely related to species of genus *Fragaria* and *Dasiphora* with strong bootstrap support in the subfamily Rosoideae. The *Alchemilla* species clustered together, assuming a monophyletic clade. Our findings are consistent with previous phylogenies reconstructed including representatives of the genus *Alchemilla* using molecular markers (Eriksson et al., 2003; Xiang et al., 2016; Zhang et al., 2017).

TABLE 4 | Positively selected sites detected in the chloroplast genome of subfamily Rosoideae based of Bayes empirical Bayes (BEB) method.

	M8	
Gene name	**Selected sites**	**Pr (w > 1)**
rpoC1	3,755 P	0.994**
ycf2	5,727 I	0.983*
ndhB	13,616 I	0.964*

*$p < 0.05$.
**$p < 0.01$.

DISCUSSION

Alchemilla Chloroplast Genome, Conserved Genome Structure, and Gene Content

Our study is the initial report and analysis of the complete cp genome of the *Alchemilla* species. They are represented by

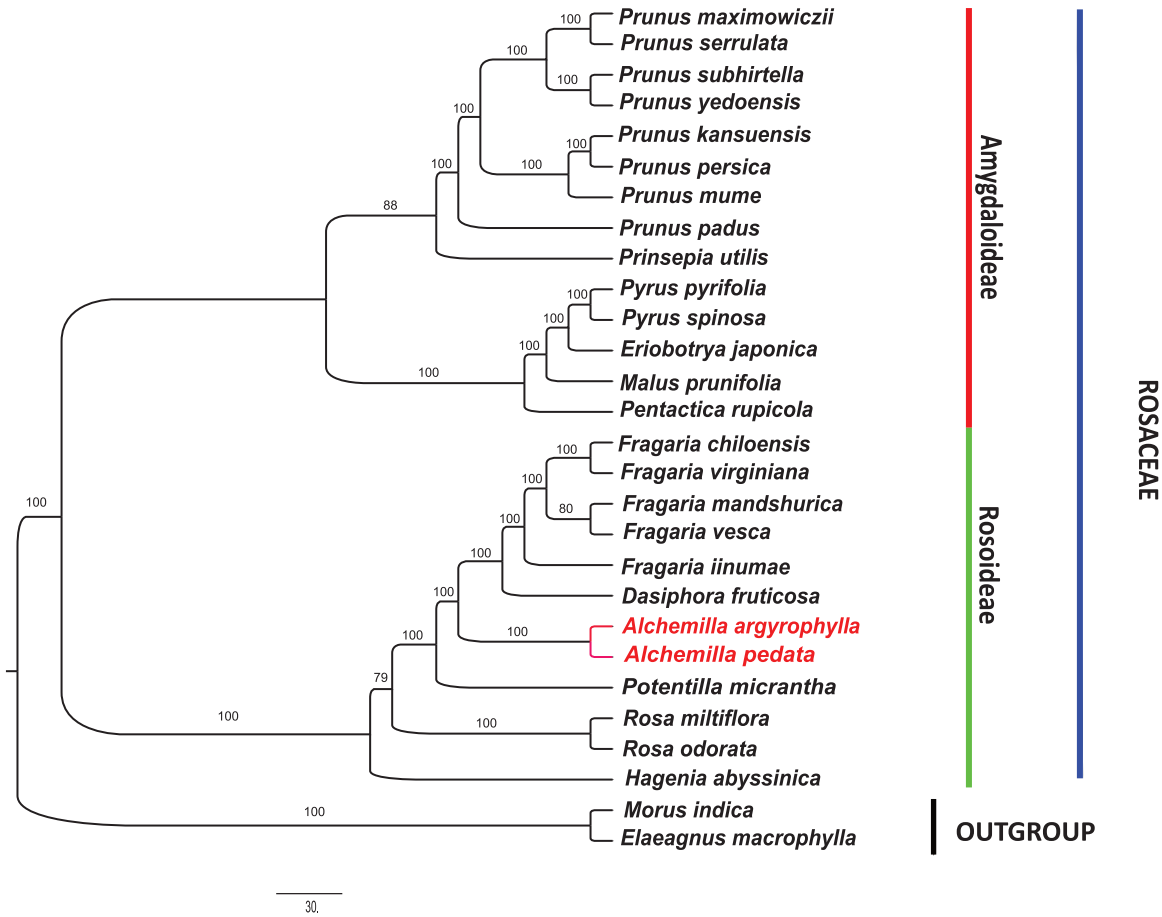


FIGURE 5 | Phylogenetic tree reconstruction of Rosaceae species based on maximum-likelihood (ML) analysis using the program IQ-Tree v.6.1 (Nguyen et al., 2015) with 1,000 bootstrap replications in 26 complete chloroplast genome sequences. *Morus indica* and *Elaeagnus macrophylla* were used as outgroups.

Alchemilla pedata and *Alchemilla argyrophylla*, a herb and a shrub, respectively, of the *Afromilla* clade found in the cooler mountainous regions of East Africa at altitudes of 2,250–4,500 m above sea level (Graham, 1960). The sequenced cp genome comprises 114 unique genes with slight variation in genome size of 152,427 bp in *A. argyrophylla* and 152, 438 bp in *A. pedata* (**Figure 1A**). Compared with other species, the size variation could be a result of the expansion and contraction of the IR region (Palmer et al., 1988). Similar to other cp genomes of higher plants, the genome annotation of *Alchemilla* species revealed an LSC-IR–SSC-IR arrangement as well as a systemic gene category and functional classification. Comparative analysis with other Rosaceae species reveals the conserved structural and organizational nature with slight variation in gene content and genome length (**Table 1**). Furthermore, the hypothetical *ycf68* embedded within *trnI-GUA*, previously not annotated in the most Rosacea species, was detected in the two annotated *Alchemilla* species (**Figure 1A**). Nucleotide substitution rates in the cp genome of angiosperms are considered lower than those of the nuclear genome (Wolfe et al., 1987). Therefore, the low gene substitution rate reported by the PCGs in *Alchemilla* is consistent with other cp genomes of higher plants. Concomitantly, the low rate of nucleotide substitutions in the PCG could be accentuated by recombination between the IRs. This is primarily due to the recurrent intra-chromosomal recombination event interplay between the two identical IR regions of the cp genome. As a result, selective constraints are imposed on both the structural stability and the sequence homogeneity (Wolfe et al., 1987; Palmer et al., 1988). The ratios between non-synonymous and synonymous substitutions (Ka/Ks) are fundamental in elucidating natural selection pressure (Nei and Kumar, 2000). In *Alchemilla*, *petN*, *psbL*, and *psbN* with Ka/Ks $>$ 0.5 and *rpoC2*, *ycf2*, and *ndhB* with $p > 0.99\%$ are essential in unfolding evolutionary history of the genus (**Table 4** and **Supplementary Table S5**). The genomic information in this study will be fundamental in the phylogenetic studies as well as the generation of molecular markers of not only the African *Alchemilla* clade but also genus *Alchemilla*.

The Loss of *atpF* Group II Intron

Introns are generally conserved regions among land plants; and therefore, instances of intron loss or gain in the cp genome could signify an evolutionary event (Daniell et al., 2016). In the annotated *Alchemilla* cp genomes, we report the absence of an intron in *atpF* belonging to group II introns. This is

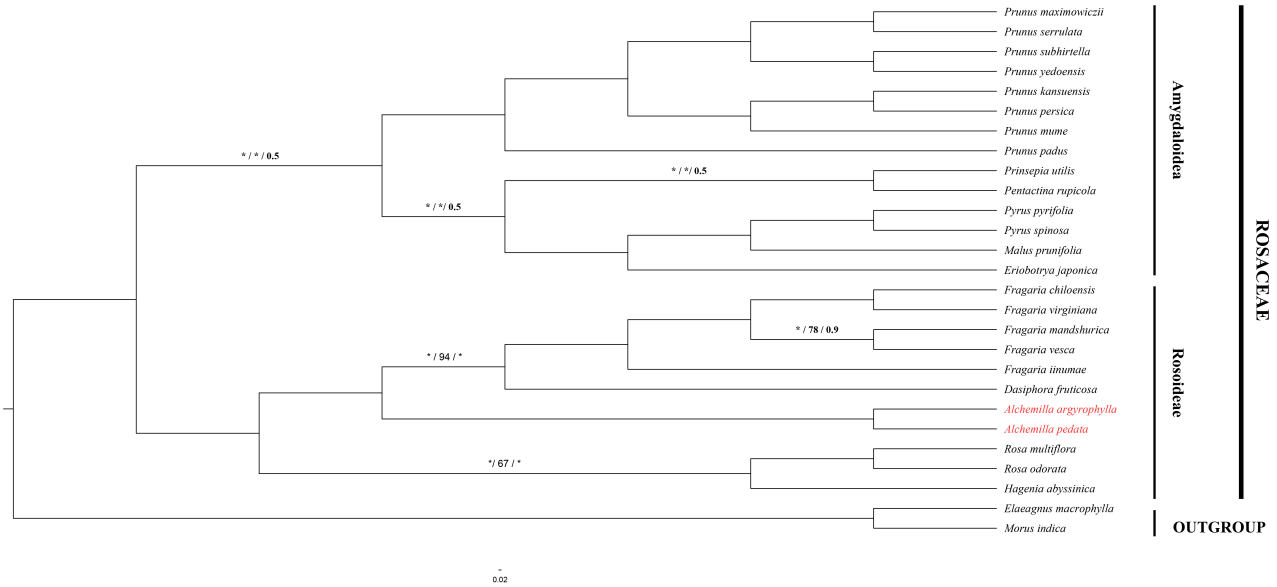


FIGURE 6 | Phylogenetic tree reconstruction based on maximum-likelihood IQ-Tree in PhyloSuite (Zhang et al., 2020) with 1,000 replications, PhyML (Lefort et al., 2017) and BI based on MrBayers (Ronquist et al., 2012) in 200,000 generation using 78 protein-coding genes common in 27 Rosaceae species. The numbers above the branch represent bootstrap support value for ML/PhyML/BI methods, where the asterisk signifies maximum support value of 100 in IQ and 1 BI. Blank branches signify 100% support value.

a rare phenomenon in land plant besides Euphorbiaceae and Malpighiales (Daniell et al., 2008). A similar observation was made in *Fragaria vesca* for the first time in Rosoideae (Shulaev et al., 2011), which we found to be closely related to our species based on phylogenetic analysis (**Figure 5**). Subsequent absence of the *atpF* intron has been evident in species of *Rosa*, *Potentilla*, *Rubus*, and *Fragaria* (Yang J. et al., 2020). However, a comparison with relative species of subfamily Amygdaloidea indicates the retention of introns within *atpF* genes (**Supplementary Table S6**). Based on the present phylogenetic framework, the loss of intron within *atpF* genes seems to have taken place once within the Rosoideae subfamily. The loss of intron in *atpF* genes is yet to be determined in other species of subfamily Rosoideae, genera in the Rosaceae, and the broader families of Rosids (Yang J. et al., 2020). Introns are broadly classified as either group I or group II and are considered as the mobile genetic elements in the cp genome. The cp genome of land plants has 17 to 20 introns classified under group II within tRNA and PCGs (Daniell et al., 2008). The only group I intron present within *trnL-AUU* gene is considered most ancient since it is also found in cyanobacteria other than in land plants and algae's cp genome (Vogel et al., 1999). The splicing function of *trnL* and other RNA transcripts for genes such as *trnA*, *rpl2*, *rps12*, and *atpF* is fully dependent on the function of maturases. The *matK* is the only maturase encoded protein present in the cp genome of land plants, yet it lacks the reverse transcriptase domain and hence cannot promote intron mobility (Hao et al., 2010; Liu et al., 2016). This causes the splicing of group II introns, including *atpF* intron, to be dependent on the host encoded splicing factors through lariat formation (Barkan, 2004). The loss of *atpF* intron among *Alchemilla* species and its close relatives could be used as a new classification basis of structural change within Rosaceae. Furthermore, the presence or absence of introns maybe insightful in understanding phylogenetic relationships as well as providing a potentially resourceful marker for evolutionary lineages in angiosperms.

Phylogenetic Analysis of *Alchemilla*

Over the previous century, the taxonomic treatment of Rosaceae has yielded varied results pertaining to its family position among the angiosperms (Potter et al., 2007). Furthermore, classification within the family has always differed depending on the treatment imposed (Hutchinson, 1964; Schulze-Menz, 1964; Kalkman, 2004; Potter et al., 2007). Morgan et al. (1994) resolved that Rosaceae be composed of monophyletic groups based on phylogenetic analysis of *rbcL* sequences across the family. This acclamation was later supported by phylogenetic studies based on *ndhF*, *matK*, and *trnL-trnF* sequences (Potter et al., 2002; Zhang et al., 2017). However, there exist uncertainties within genera resulting from unresolved tree portions having weak support (Potter et al., 2007). Notably in the genus *Alchemilla*, the distinction between the African clade (*Afromilla*) and the Eurasian (*Eualchemilla*) clade is geographically significant but remain unresolved (Gehrke et al., 2016). The contention requires the separation of the *Alchemilla* into *Eualchemilla* and *Afromilla* clade, which has been restrained due to lack of substantial morphological borderlines despite extensive and significant studies (Notov and Kusnetzova, 2004; Gehrke et al., 2008; Lundberg et al., 2009). The emergence and rapid development in cp genome sequencing technologies have been essential in providing resourceful genomic information that has been

fundamental for the reconstruction of lower and higher plant phylogenies as well as evolutionary trends (Daniell et al., 2016; Yang X. et al., 2020). Significantly, the cp genome sequences have been proven effective in eliminating phylogenetic incongruences arising from incomplete lineage sorting (ILS) and hybridization (Lundberg et al., 2009; Morales-Briones et al., 2018). In this study, *A. pedata* and *A. argyrophylla* nested into a monophyletic clade with 100% bootstrap support. They share a recent ancestry with *Dasiphora* and *Fragaria* species (**Figure 5**). There is need for more complete cp genomes of *Alchemilla* species for higher precision in phylogenetic conclusion. This study is therefore resourceful for further delimitation of species in *Alchemilla* and phylogenetic studies of the genus.

DATA AVAILABILITY STATEMENT

The assembled chloroplast genome sequences have been uploaded to and deposited in GenBank and we have received the accession numbers which are shown in **Table 1** in the manuscript.

AUTHOR CONTRIBUTIONS

PR and XD performed the experiment. PR, XD, and J-XY performed the data analysis. PR drafted the manuscript. FM, MO, IM, and PK revised the manuscript. G-WH and Q-FW designed and supervised the experiment. All the authors contributed to and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (31970211) and Sino-Africa Joint Research Center, CAS (SAJC201614).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.560368/full#supplementary-material

REFERENCES

Amiryousefi, A., Hyvönen, J., and Poczai, P. (2018). IRscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics* 34, 3030–3031. doi: 10.1093/bioinformatics/bty220

Asaf, S., Khan, A. L., Khan, M. A., Waqas, M., Kang, S.-M., Yun, B.-W., et al. (2017). Chloroplast genomes of *Arabidopsis halleri* ssp. *gemmifera* and *Arabidopsis lyrata* ssp. *petraea*: structures and comparative analysis. *Sci. Rep.* 7:7556. doi: 10.1038/s41598-017-07891-5

Asano, T. (2004). Complete nucleotide sequence of the sugarcane (*Saccharum Officinarum*) chloroplast genome: a comparative analysis of four monocot chloroplast genomes. *DNA Res.* 11, 93–99. doi: 10.1093/dnares/11.2.93

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Barkan, A. (2004). "Intron splicing in plant organelles," in *Molecular Biology and Biotechnology of Plant Organelles*, eds H. Daniell and C. D. Chase (Dordrecht: Springer), 295–322.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi: 10.1093/nar/27.2.573

Bicknell, R. A., and Koltunow, A. M. (2004). Understanding apomixis: recent advances and remaining conundrums. *Plant Cell* 16(Suppl. 1), S228–S245. doi: 10.1105/tpc.017921

Boroja, T., Mihailović, V., Katanić, J., Pan, S. P., Nikles, S., Imbimbo, P., et al. (2018). The biological activities of roots and aerial parts of *Alchemilla vulgaris* L. *S. Afr. J. Bot.* 116, 175–184. doi: 10.1016/j.sajb.2018.03.007

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421

Cavalier-Smith, T. (2002). Chloroplast evolution: secondary symbiogenesis and multiple losses. *Curr. Biol.* 12, R62–R64. doi: 10.1016/S0960-9822(01)00675-3

Chase, M. W., and Hills, H. H. (1991). Silica gel: an ideal material for field preservation of leaf samples for DNA studies. *Taxon* 40, 215–220. doi: 10.2307/1222975

Cheng, B., Zheng, Y., and Sun, Q. (2015). Genetic diversity and population structure of *Taxus cuspidata* in the Changbai Mountains assessed by chloroplast DNA sequences and microsatellite markers. *Biochem. Syst. Ecol.* 63, 157–164. doi: 10.1016/j.bse.2015.10.009

Cheng, H., Li, J., Zhang, H., Cai, B., Gao, Z., Qiao, Y., et al. (2017). The complete chloroplast genome sequence of strawberry (*Fragaria* × *ananassa* Duch.) and comparison with related species of Rosaceae. *PeerJ* 5:e3919. doi: 10.7717/peerj.3919

Czapik, R. (1996). Problems of apomictic reproduction in the FamiliesCompositae andRosaceae. *Folia Geobot. Phytotaxonom.* 31, 381–387. doi: 10.1007/BF02815382

Daniell, H., Lin, C.-S., Yu, M., and Chang, W.-J. (2016). Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biol.* 17:134. doi: 10.1186/s13059-016-1004-2

Daniell, H., Wurdack, K. J., Kanagaraj, A., Lee, S.-B., Saski, C., and Jansen, R. K. (2008). The complete nucleotide sequence of the cassava (*Manihot esculenta*) chloroplast genome and the evolution of *atpF* in Malpighiales: RNA editing and multiple losses of a group II intron. *Theor. Appl. Genet.* 116, 723–737. doi: 10.1007/s00122-007-0706-y

Doyle, J. (1991). "DNA protocols for plants," in *Molecular Techniques in Taxonomy*, eds G. M. Hewitt, A. W. B. Johnston, and J. P. W. Young (Berlin: Springer), 283–293. doi: 10.1007/978-3-642-83962-7_18

Doyle, J. J., Davis, J. I., Soreng, R. J., Garvin, D., and Anderson, M. J. (1992). Chloroplast DNA inversions and the origin of the grass family (Poaceae). *Proc. Natl. Acad. Sci. U.S.A.* 89, 7722–7726. doi: 10.1073/pnas.89.16.7722

Eriksson, T., Hibbs, M. S., Yoder, A. D., Delwiche, C. F., and Donoghue, M. J. (2003). The Phylogeny of Rosoideae (Rosaceae) based on sequences of the internal transcribed spacers (ITS) of nuclear ribosomal DNA and the *trnL/F* region of chloroplast DNA. *Int. J. Plant Sci.* 164, 197–211. doi: 10.1086/346163

Faghir, M. B., Chaichi, K. K., and Shahvon, R. S. (2014). Foliar epidermis micromorphology of the genus *Alchemilla* (Rosaceae) in Iran. *Phytol. Balcanica* 20, 215–225.

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., and Dubchak, I. (2004). VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* 32, W273–W279. doi: 10.1093/nar/gkh458

Fröhner, S. E. (1995). "Alchemilla," in *Hegi: Illustrierte Flora von Mitteleuropa, Vol. 4 Teil 2B*, eds H. Scholz, H. J. Conert, E. J. Jäger, J. W. Kadereit, W. Schultze-Motel, G. Wagenitz, et al. (Berlin: Verlag Paul Parey), 13–242.

Gao, F., Chen, C., Arab, D. A., Du, Z., He, Y., and Ho, S. Y. W. (2019). EasyCodeML: a visual tool for analysis of selection using CodeML. *Ecol. Evol.* 9, 3891–3898. doi: 10.1002/ece3.5015

Gehrke, B., Bräuchler, C., Romoleroux, K., Lundberg, M., Heubl, G., and Eriksson, T. (2008). Molecular phylogenetics of *Alchemilla*, *Aphanes* and *Lachemilla* (Rosaceae) inferred from plastid and nuclear intron and spacer DNA sequences,

with comments on generic classification. *Mol. Phylogenet. Evol.* 47, 1030–1044. doi: 10.1016/j.ympev.2008.03.004

Gehrke, B., Kandziora, M., and Pirie, M. D. (2016). The evolution of dwarf shrubs in alpine environments: a case study of *Alchemilla* in Africa. *Ann. Bot.* 117, 121–131. doi: 10.1093/aob/mcv159

Gichira, A. W., Li, Z., Saina, J. K., Long, Z., Hu, G., Gituru, R. W., et al. (2017). The complete chloroplast genome sequence of an endemic monotypic genus *Hagenia* (Rosaceae): structural comparative analysis, gene content and microsatellite detection. *PeerJ* 5:e2846. doi: 10.7717/peerj.2846

Graham, R. A. (1960). "Rosaceae," in *Flora of Tropical East Africa*, eds C. E. Hubbard and E. Milne-Redhead (London: Crown Agents), 1–61.

Gray, M. W. (1989). The evolutionary origins of organelles. *Trends Genet.* 5, 294–299.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Hao, D. C., Chen, S. L., and Xiao, P. G. (2010). Molecular evolution and positive Darwinian selection of the chloroplast maturase matK. *J. Plant Res.* 123, 241–247. doi: 10.1007/s10265-009-0261-5

Hayirhoglu-Ayaz, S., İnceer, H., and Frost-Olsen, P. (2006). Chromosome counts in the genus *Alchemilla* (Rosaceae) from SW Europe. *Folia Geobot.* 41, 335–344.

Hörandl, E. (2004). Comparative analysis of genetic divergence among sexual ancestors of apomictic complexes using isozyme data. *Int. J. Plant Sci.* 165, 615–622. doi: 10.1086/386557

Hutchinson, J. (1964). *The Genera of Flowering Plants. Dicotyledons*, Vol. 1. (London: Oxford University Press), 195–208.

Izmailow, R. (1981). Karyological studies in species of *Alchemilla* L. from the series Calycinae Bus. (Section Brevicaulon Rothm.). *Acta Biol. Cracov. Ser. Bot.* 23, 117–130.

Izmailow, R. (1982). Further Karyological studies in species of *Alchemilla* L. from the series Calycinae Bus.(section Brevicaulon Rothm.). *Acta Biol. Cracov. Ser. Bot.* 24, 117–130.

Jian, H.-Y., Zhang, Y.-H., Yan, H.-J., Qiu, X.-Q., Wang, Q.-G., Li, S.-B., et al. (2018). The complete chloroplast genome of a key ancestor of modern roses, *Rosa chinensis* var. *Spontanea*, and a Comparison with Congeneric Species. *Molecules* 23:389. doi: 10.3390/molecules23020389

Jin, J.-J., Yu, W.-B., Yang, J.-B., Song, Y., dePamphilis, C. W., Yi, T.-S., et al. (2020). GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* 21:241. doi: 10.1186/s13059-020-02154-5

Kalkman, C. (2004). "Rosaceae," in *The Families and Genera of Vascular Plants, vol. 6, Flowering Plants - Dicotyledons: Celastrales, Oxalidales, Rosales, Cornales, Ericales*, ed. K. Kubitzki (Berlin: Springer), 343–386. doi: 10.1007/978-3-662-07257-8_39

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth.4285

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kaya, B., Menemen, Y., and Saltan, F. (2012). Flavonoid compounds identified in *Alchemilla* l. species collected in the North-eastern Black sea region of Turkey. *Afr. J. Trad. Complement. Altern. Med.* 9, 418–425. doi: 10.4314/ajtcam.v9i3.18

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649. doi: 10.1093/bioinformatics/bts199

Keeling, P. J. (2004). Diversity and evolutionary history of plastids and their hosts. *Am. J. Bot.* 91, 1481–1493. doi: 10.3732/ajb.91.10.1481

Köhler, M., Reginato, M., Souza-Chies, T. T., and Majure, L. C. (2020). Insights into chloroplast genome evolution across opuntioideae (Cactaceae) reveals robust yet sometimes conflicting phylogenetic topologies. *Front. Plant Sci.* 11:729. doi: 10.3389/fpls.2020.00729

Kress, W. J., Wurdack, K. J., Zimmer, E. A., Weigt, L. A., and Janzen, D. H. (2005). Use of DNA barcodes to identify flowering plants. *Proc. Natl. Acad. Sci. U.S.A.* 102, 8369–8374. doi: 10.1073/pnas.0503123102

Kuang, D.-Y., Wu, H., Wang, Y.-L., Gao, L.-M., Zhang, S.-Z., and Lu, L. (2011). Complete chloroplast genome sequence of *Magnolia kwangsiensis* (Magnoliaceae): implication for DNA barcoding and population genetics. *Genome* 54, 663–673. doi: 10.1139/g11-026

Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* 29, 4633–4642. doi: 10.1093/nar/29.22.4633

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Lefort, V., Longueville, J.-E., and Gascuel, O. (2017). SMS: smart model selection in PhyML. *Mol. Biol. Evol.* 34, 2422–2424. doi: 10.1093/molbev/msx149

Liu, C., Zhu, H., Xing, Y., Tan, J., Chen, X., Zhang, J., et al. (2016). *Albino Leaf 2* is involved in the splicing of chloroplast group I and II introns in rice. *J. Exp. Bot.* 67, 5339–5347. doi: 10.1093/jxb/erw296

Liu, L.-X., Li, R., Worth, J. R. P., Li, X., Li, P., Cameron, K. M., et al. (2017). The complete chloroplast genome of chinese bayberry (*Morella rubra*, Myricaceae): implications for understanding the evolution of Fagales. *Front. Plant Sci.* 8:968. doi: 10.3389/fpls.2017.00968

Lohse, M., Drechsel, O., Kahlau, S., and Bock, R. (2013). OrganellarGenomeDRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Res.* 41, W575–W581. doi: 10.1093/nar/gkt289

Lundberg, M., Töpel, M., Eriksen, B., Nylander, J. A. A., and Eriksson, T. (2009). Allopolyploidy in *Fragariinae* (Rosaceae): comparing four DNA sequence regions, with comments on classification. *Mol. Phylogenet. Evol.* 51, 269–280. doi: 10.1016/j.ympev.2009.02.020

Makałowski, W., and Boguski, M. S. (1998). Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl. Acad. Sci. U.S.A.* 95, 9407–9412. doi: 10.1073/pnas.95.16.9407

Mayor, C., Brudno, M., Schwartz, J. R., Poliakov, A., Rubin, E. M., Frazer, K. A., et al. (2000). VISTA: Visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* 16, 1046–1047. doi: 10.1093/bioinformatics/16.11.1046

Menezes, A. P. A., Resende-Moreira, L. C., Buzatti, R. S. O., Nazareno, A. G., Carlsen, M., Lobo, F. P., et al. (2018). Chloroplast genomes of *Byrsonima* species (Malpighiaceae): comparative analysis and screening of high divergence sequences. *Sci. Rep.* 8:2210. doi: 10.1038/s41598-018-20189-4

Mo, Z., Lou, W., Chen, Y., Jia, X., Zhai, M., Guo, Z., et al. (2020). The chloroplast genome of *Carya illinoinensis*: genome structure, adaptive evolution, and phylogenetic analysis. *Forests* 11:207. doi: 10.3390/f11020207

Morales-Briones, D. F., Romoleroux, K., Kolář, F., and Tank, D. C. (2018). Phylogeny and evolution of the neotropical radiation of *Lachemilla* (Rosaceae): uncovering a history of reticulate evolution and implications for infrageneric classification. *Syst. Bot.* 43, 17–34. doi: 10.1600/036364418X696897

Morgan, D. R., Soltis, D. E., and Robertson, K. R. (1994). Systematic and evolutionary implications of *rbc* L sequence variation in Rosaceae. *Am. J. Bot.* 81, 890–903. doi: 10.1002/j.1537-2197.1994.tb15570.x

Morton, B. R. (2003). The role of context-dependent mutations in generating compositional and codon usage bias in grass chloroplast DNA. *J. Mol. Evol.* 56, 616–629. doi: 10.1007/s00239-002-2430-1

Munyao, J. N., Dong, X., Yang, J.-X., Mbandi, E. M., Wanga, V. O., Oulo, M. A., et al. (2020). Complete Chloroplast Genomes of *Chlorophytum comosum* and *Chlorophytum gallabatense*: genome structures, comparative and phylogenetic analysis. *Plants* 9:296. doi: 10.3390/plants9030296

Nei, M., and Kumar, S. (2000). *Molecular Evolution and Phylogenetics*. Oxford: Oxford university press.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Nielsen, R. (2005). Molecular signatures of natural selection. *Annu. Rev. Genet.* 39, 197–218. doi: 10.1146/annurev.genet.39.073003.112420

Ninio, J. (1984). The neutral theory of molecular evolution: edited by Mooto Kimura Cambridge University Press. Cambridge, 1983 366 pages. *FEBS Lett.* 170, 210–211. doi: 10.1016/0014-5793(84)81411-8

Notov, A. A., and Kusnetzova, T. V. (2004). Architectural units, axiality and their taxonomic implications in *Alchemillinae*. *Wulfenia* 11, 85–130.

Palmer, J. D., Jansen, R. K., Michaels, H. J., Chase, M. W., and Manhart, J. R. (1988). Chloroplast DNA variation and plant phylogeny. *Ann. Mi. Bot. Gard.* 75:1180. doi: 10.2307/2399279

Parks, M., Cronn, R., and Liston, A. (2009). Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biol.* 7:84. doi: 10.1186/1741-7007-7-84

Potter, D., Eriksson, T., Evans, R. C., Oh, S., Smedmark, J. E. E., Morgan, D. R., et al. (2007). Phylogeny and classification of Rosaceae. *Plant Syst. Evol.* 266, 5–43. doi: 10.1007/s00606-007-0539-9

Potter, D., Gao, F., Bortiri, P. E., Oh, S.-H., and Baggett, S. (2002). Phylogenetic relationships in Rosaceae inferred from chloroplast mat K and trn L- trn F nucleotide sequence data. *Plant Syst. Evol.* 231, 77–89. doi: 10.1007/s006060200012

Powell, W., Morgante, M., McDevitt, R., Vendramin, G. G., and Rafalski, J. A. (1995). Polymorphic simple sequence repeat regions in chloroplast genomes: Applications to the population genetics of pines. *Proc. Natl. Acad. Sci. U.S.A.* 92, 7759–7763. doi: 10.1073/pnas.92.17.7759

Pugh, T., Fouet, O., Risterucci, A. M., Brottier, P., Abouladze, M., Deletrez, C., et al. (2004). A new cacao linkage map based on codominant markers: development and integration of 201 new microsatellite markers. *Theor. Appl. Genet.* 108, 1151–1161. doi: 10.1007/s00122-003-1533-4

Qu, X.-J., Moore, M. J., Li, D.-Z., and Yi, T.-S. (2019). PGA: a software package for rapid, accurate, and flexible batch annotation of plastomes. *Plant Methods* 15:50. doi: 10.1186/s13007-019-0435-7

Rafalski, J. A., and Tingey, S. V. (1993). Genetic diagnostics in plant breeding: RAPDs, microsatellites and machines. *Trends Genet.* 9, 275–280. doi: 10.1016/0168-9525(93)90013-8

Rambaut, A. (2018). *FigTree v1.4.4. Institute of Evolutionary Biology*. Edinburgh: University of Edinburgh.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., et al. (2012). MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. doi: 10.1093/sysbio/sys029

Rothmaler, W. (1937). Systematische Vorarbeiten zu einer Monographie der Gattung *Alchemilla* (L.) Scop. VII. Aufteilung der Gattung und Nomenklatur. *Repertorium Novarum Specierum Regni Vegetabilis* 42, 164–173. doi: 10.1002/fedr.19370421106

Salamone, I., Govindarajulu, R., Falk, S., Parks, M., Liston, A., and Ashman, T.-L. (2013). Bioclimatic, ecological, and phenotypic intermediacy and high genetic admixture in a natural hybrid of octoploid strawberries. *Am. J. Bot.* 100, 939–950. doi: 10.3732/ajb.1200624

Saski, C., Lee, S.-B., Daniell, H., Wood, T. C., Tomkins, J., Kim, H.-G., et al. (2005). Complete chloroplast genome sequence of *Glycine max* and comparative analyses with other legume genomes. *Plant Mol. Biol.* 59, 309–322. doi: 10.1007/s11103-005-8882-0

Schulze-Menz, G. K. (1964). "Rosaceae," in *A. Engler's Syllabus der Pflanzenfamilien*, Vol. II, ed. H. Melchoir (Berlin: Borntraeger), 209–218.

Sharp, P. M., and Li, W.-H. (1987). The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15, 1281–1295. doi: 10.1093/nar/15.3.1281

Shen, L., Guan, Q., Amin, A., Zhu, W., Li, M., Li, X., et al. (2016). Complete plastid genome of *Eriobotrya japonica* (Thunb.) Lindl and comparative analysis in Rosaceae. *Springerplus* 5:2036. doi: 10.1186/s40064-016-3702-3

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. *Nat. Biotechnol.* 26, 1135–1145. doi: 10.1038/nbt1486

Shinozaki, K., Ohme, M., Tanaka, M., Wakasugi, T., Hayashida, N., Matsubayashi, T., et al. (1986). The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *EMBO J.* 5, 2043–2049. doi: 10.1002/j.1460-2075.1986.tb04464.x

Shulaev, V., Sargent, D. J., Crowhurst, R. N., Mockler, T. C., Folkerts, O., Delcher, A. L., et al. (2011). The genome of woodland strawberry (*Fragaria vesca*). *Nat. Genet.* 43, 109–116. doi: 10.1038/ng.740

Soják, J. (2008). Notes on Potentilla XXI. A new division of the tribe Potentilleae (Rosaceae) and notes on generic delimitations. *Bot. Jahrbücher* 127, 349–358.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28, 2731–2739. doi: 10.1093/molbev/msr121

Thiel, T., Michalek, W., Varshney, R., and Graner, A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor. Appl. Genet.* 106, 411–422. doi: 10.1007/s00122-002-1031-0

Vogel, J., Börner, T., and Hess, W. R. (1999). Comparative analysis of splicing of the complete set of chloroplast group II introns in three higher plant mutants. *Nucleic Acids Res.* 27, 3866–3874. doi: 10.1093/nar/27.19.3866

Wang, L., Xing, H., Yuan, Y., Wang, X., Saeed, M., Tao, J., et al. (2018). Genome-wide analysis of codon usage bias in four sequenced cotton species. *PLoS One* 13:e0194372. doi: 10.1371/journal.pone.0194372

Wang, Q., Niu, Z., Li, J., Zhu, K., and Chen, X. (2020). The complete chloroplast genome sequence of the Chinese endemic species *Sorbus setschwanensis* (Rosaceae) and its phylogenetic analysis. *Nord. J. Bot.* 38:njb.02532. doi: 10.1111/njb.02532

Wang, R.-J., Cheng, C.-L., Chang, C.-C., Wu, C.-L., Su, T.-M., and Chaw, S.-M. (2008). Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. *BMC Evol. Biol.* 8:36. doi: 10.1186/1471-2148-8-36

Weber, J. L. (1990). Human DNA polymorphisms and methods of analysis. *Curr. Opin. Biotechnol.* 1, 166–171. doi: 10.1016/0958-1669(90)90026-H

Wick, R. R., Schultz, M. B., Zobel, J., and Holt, K. E. (2015). Bandage: interactive visualization of *de novo* genome assemblies: Fig. 1. *Bioinformatics* 31, 3350–3352. doi: 10.1093/bioinformatics/btv383

Wolfe, K. H., Li, W. H., and Sharp, P. M. (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc. Natl. Acad. Sci. U.S.A.* 84, 9054–9058. doi: 10.1073/pnas.84.24.9054

Xiang, Y., Huang, C.-H., Hu, Y., Wen, J., Li, S., Yi, T., et al. (2016). Evolution of Rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. *Mol. Biol. Evol.* 34, 262–281. doi: 10.1093/molbev/msw242

Xue, J., Wang, S., and Zhou, S.-L. (2012). Polymorphic chloroplast microsatellite loci in Nelumbo (Nelumbonaceae). *Am. J. Bot.* 99, e240–e244. doi: 10.3732/ajb.1100547

Yang, J., Kang, G.-H., Pak, J.-H., and Kim, S.-C. (2020). Characterization and comparison of two complete plastomes of Rosaceae Species (*Potentilla dickinsii* var. *Glabrata* and *Spiraea insularis*) Endemic to Ulleung Island, Korea. *Int. J. Mol. Sci.* 21:4933. doi: 10.3390/ijms21144933

Yang, X., Xie, D.-F., Chen, J.-P., Zhou, S.-D., Yu, Y., and He, X.-J. (2020). Comparative analysis of the complete chloroplast genomes in *Allium* Subgenus *Cyathophora* (Amaryllidaceae): phylogenetic relationship and adaptive evolution. *Biomed Res. Int.* 2020:1732586. doi: 10.1155/2020/1732586

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Yang, Z., and Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* 17, 32–43. doi: 10.1093/oxfordjournals.molbev.a026236

Yang, Z., Wong, W. S., and Nielsen, R. (2005). Bayes empirical bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* 22, 1107–1118. doi: 10.1093/molbev/msi097

Zhang, D., Gao, F., Jakovlić, I., Zou, H., Zhang, J., Li, W. X., et al. (2020). PhyloSuite: An integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. *Mol. Ecol. Resour.* 20, 348–355. doi: 10.1111/1755-0998.13096

Zhang, S.-D., Jin, J.-J., Chen, S.-Y., Chase, M. W., Soltis, D. E., Li, H.-T., et al. (2017). Diversification of Rosaceae since the late cretaceous based on plastid phylogenomics. *New Phytol.* 214, 1355–1367. doi: 10.1111/nph.14461

Zhang, Y., Ma, J., Yang, B., Li, R., Zhu, W., Sun, L., et al. (2014). The complete chloroplast genome sequence of *Taxus chinensis* var. *mairei* (Taxaceae): loss of an inverted repeat region and comparative analysis with related species. *Gene* 540, 201–209. doi: 10.1016/j.gene.2014.02.037

Zhang, Z., Li, J., Zhao, X.-Q., Wang, J., Wong, G. K.-S., and Yu, J. (2006). KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics* 4, 259–263. doi: 10.1016/S1672-0229(07)60007-2

Zhao, M.-L., Song, Y., Ni, J., Yao, X., Tan, Y.-H., and Xu, Z.-F. (2018). Comparative chloroplast genomics and phylogenetics of nine *Lindera* species (Lauraceae). *Sci. Rep.* 8:8844. doi: 10.1038/s41598-018-27090-0

Zhao, Y., Lu, D., Han, R., Wang, L., and Qin, P. (2018). The complete chloroplast genome sequence of the shrubby cinquefoil *Dasiphora fruticosa* (Rosales: Rosaceae). *Conserv. Genet. Resour.* 10, 675–678. doi: 10.1007/s12686-017-0899-6

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.

Copyright © 2020 Rono, Dong, Yang, Mutie, Oulo, Malombe, Kirika, Hu and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 11 January 2021
doi: 10.3389/fgene.2020.597890

UDP-Glucose Dehydrogenases: Identification, Expression, and Function Analyses in Upland Cotton (*Gossypium hirsutum*)

Tingting Jia[1†], Qun Ge[1†], Shuya Zhang[1], Zhen Zhang[1], Aiying Liu[1], Senmiao Fan[1], Xiao Jiang[1], Yulong Feng[2], Lipeng Zhang[1], Doudou Niu[1], Shen Huang[3], Wankui Gong[1], Youlu Yuan[1,2]* and Haihong Shang[1,2]**

[1] State Key Laboratory of Cotton Biology, Key Laboratory of Biological and Genetic Breeding of Cotton, The Ministry of Agriculture, Institute of Cotton Research, Chinese Academy of Agricultural Sciences, Anyang, China, [2] Zhengzhou Research Base, State Key Laboratory of Cotton Biology, Zhengzhou University, Zhengzhou, China, [3] Zhengzhou University of Light Industry College of Food and Bioengineering, Zhengzhou, China

UDP-glucose dehydrogenase (UGD; EC1.1.1.22) is a NAD^+-dependent enzyme that catalyzes the two-fold oxidation of UDP-glucose (UDP-Glc) to produce UDP-glucuronic acid and plays an important role in plant cell wall synthesis. A total of 42 *UGD* genes from four *Gossypium* genomes including *G. hirsutum*, *G. arboretum*, *G. barbadense*, and *G. raimondii* were identified and found that the *UGD* gene family has conservative evolution patterns in gene structure and protein domain. The growth of fibers can be effectively promoted after adding the UDP-Glc to the medium, and the *GhUGD* gene expression enhanced. In addition, the transgenic *Arabidopsis* lines over-expressing *GH_D12G1806* had longer root lengths and higher gene expression level than the wild-type plants of *Columbia-0*. These results indicated that *UGD* may play important roles in cotton fiber development and has a guiding significance for dissecting fiber development mechanism.

Keywords: cotton fiber, UDP-glucose dehydrogenase, UDP-glucuronic acid, transgenic *Arabidopsis*, fiber development

OPEN ACCESS

Edited by:
Hongjian Wan, ZheJiang Academy of Agricultural Sciences, China

Reviewed by:
Longbiao Guo, China National Rice Research Institute, Chinese Academy of Agricultural Sciences, China
Abdul Qayyum Rao, University of the Punjab, Pakistan

***Correspondence:**
Wankui Gong
gongwankui@caas.cn; wkgong@aliyun.com
Youlu Yuan
yuanyoulu@caas.cn
Haihong Shang
shanghaihong@caas.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *22 August 2020*
Accepted: *27 November 2020*
Published: *11 January 2021*

Citation:
Jia T, Ge Q, Zhang S, Zhang Z, Liu A, Fan S, Jiang X, Feng Y, Zhang L, Niu D, Huang S, Gong W, Yuan Y and Shang H (2021) UDP-Glucose Dehydrogenases: Identification, Expression, and Function Analyses in Upland Cotton (Gossypium hirsutum). Front. Genet. 11:597890. doi: 10.3389/fgene.2020.597890

INTRODUCTION

Cotton fiber is the most important textile raw material. Its primary cell wall components, including pectin, hemicellulose, and cellulose, are mainly derived from a common biochemical precursor, UDP-glucuronic acid (UDP-GlcA; Hofte and Voxeur, 2017), which is responsible for the derivation of about 50% of the cell wall biomass (Zablackis et al., 1995). Plants have evolved two independent pathways for the synthesis of UDP-GlcA. One pathway consists of conversion of UDP-glucose (UDP-Glc) to UDP-GalA catalyzed by UDP-glucose dehydrogenase (UGD, which belongs to the family of NAD^+-linked oxidoreductase; EC 1.1.1.22; Tenhaken and Thulke, 1996). The other is the more complex myo-inositol pathway, in which cleavage of inositol into D-GlcA catalyzed

Abbreviations: UDP-Glc, UDP-α-D-glucose; UDP-GalA, UDP-α-D-galacturonic acid; ABA, abscisic acid; UGD, UDP-glucose dehydrogenase; MIOX, inositol oxygenase; GlcAK, glucuronokinase; USP, UDP-sugar pyrophosphorylase; DPA, days post anthesis; FPKM, fragments per kilobase of exon per million reads; BLAST, Basic Local Alignment Search Tool; GSDS, Gene Structure Display Server; PI, isoelectric point; Mw, molecular weight; cDNA, complementary DNA.

by myoinositol oxygenase (MIOX; EC 1.13.99.1) is involved. Then, D-GlcA is activated to UDP-GlcA (Loewus et al., 1962; Seitz et al., 2000; Ute et al., 2004; Kanter et al., 2005; Pieslinger et al., 2010). Studies ever demonstrated that down-regulation of *MIOX* genes in the *miox1/2/4/5*-mutants has no effect on cell wall composition in *Arabidopsis thaliana*. Further analysis revealed that an alternative pathway to UDP-GlcA via UDP-Glc is compensatorily up-regulated in these mutants (Endres and Tenhaken, 2011).

The enzyme UGD is a key factor in converting UDP-Glc to UDP-GlcA (Tenhaken and Thulke, 1996). It catalyzes irreversible loss of 4-electrons in UDP-Glc and finally oxidizes it to UDP-GlcA, with a concomitant production of two NADH molecules from NAD^+. UDP-GlcA is subsequently converted to UDP-D-xylose (UDP-Xyl), of which the corresponding sugar moieties provide abundant components for the synthesis of pectins and hemicelluloses (Reiter, 2008). *UGD* was cloned in *eukaryotes* for the first time (Campbell et al., 1997) and subsequently cloned in higher plants such as soybean (Tenhaken and Thulke, 1996), *A. thaliana* (Klinghammer and Tenhaken, 2007), tobacco (Bindschedler et al., 2005), and poplar (Johansson et al., 2002). *UGDs* have different locations and expression levels in different species and tissues. In soybean, the *UGDs* have high expression level in radicle, while low in epicotyl and spire. The expression pattern of the *UGDs* in different developmental stages indicated that UGD is a key regulator for the availability of hemicellulose precursors (Tenhaken and Thulke, 1996). Analyses of four *UGD* genes in *A. thaliana* revealed that *AtUGD1* is a pseudogene, and *AtUGD2* has the highest expression level in cotyledons and hypocotyls but less in root and *AtUGD3* has the highest enzyme activity (Klinghammer and Tenhaken, 2007). *UGD* plays a key role in the expression process from UDP-glucose to UDP-Xyl in tobacco (Bindschedler et al., 2005). The different expression level of *UGDs* in different parts of poplar development indicated that it played a key role in the spire and immature xylem development (Johansson et al., 2002).

Although a few functional studies of *UGD* genes were performed in some plant species, *UGD* behavior and function in *Gossypium* species is still poorly understood. In this study, a series of bioinformatics analyses including physicochemical properties, evolutionary relationships and gene structure, conserved motifs, and *UGD* domain features were performed. Expression analyses of *UGD* genes, including biochemical experiments, subcellular localization, and functional verification of transgenic *A. thaliana*, were also performed. These results indicated that *UGDs* play an important role in the conversion of UDP-Glc to UDP-GlcA and cotton fiber development, which provides fundamental basis for fiber quality improvement.

MATERIALS AND METHODS

Plant Materials

The plants of upland cotton cultivars 0-153, an upland cotton germplasm line with high-quality fibers, and sGK9708, an upland cotton cultivar with low-quality fibers, were grown under experimental field conditions in Zhengzhou, China (Zhang Z. et al., 2020). The flowers were tagged on the day of anthesis as 0 day post anthesis (DPA) flowers. Developing bolls from the tagged flowers at 5, 10, 15, and 20 DPA were harvested and peeled out with a sterile scalpel to get fiber samples. Fiber samples were immediately dipped into liquid nitrogen and then stored at −80°C conditions for subsequent real-time quantitative PCR (qRT-PCR) assays. Plants of *Nicotiana benthamiana* were grown in the greenhouse with a 16/8 h light/dark cycle at 22°C, 60% relative humidity, and light intensity of 100–120 μmol m^{-2} S^{-1}.

Transcriptomic Data Analysis

The transcriptome data of TM-1, a standard genetic line of upland cotton, was downloaded from NCBI SPA database (accession number PRJNA490626, https://www.ncbi.nlm.nih.gov/bioproject/PRJNA490626; Hu et al., 2019). RNA-Seq raw data of sGK9708 and 0-153 derived from Zhang Z. et al. (2020) were reanalyzed as described previously by Zou et al. (2019). The mapping value of fragments per kilobase of exon per million reads (FPKM) was calculated using the Cufflinks program. Gene expression levels were analyzed by the $\log_2$ (FPKM + 1) method. FPKM value greater than five was used as a threshold to declare a gene expression. Gene expression patterns were visualized with heat maps prepared using the R (3.3.0) software package[1].

RNA Isolation and cDNA Synthesis

Total RNA samples were extracted as described in the kit instructions of the RNAprep Pure Plant Kit (TIANGEN, Beijing, China) and treated with RNase-free DNase I. RNA concentration was calibrated using NanoDrop 2000 (Thermo Scientific), and RNA quality was monitored by agarose gel electrophoresis. The absorbance ratio of A260/280 was used as an indicator of protein contamination. The RNA samples that ultimately meet the minimum eligibility criteria were applied to reverse transcription. First cDNA strand was synthesized using the TransScript All-in-One First-Strand cDNA Synthesis SuperMix for qPCR (Trans, Beijing, China). The concentration of cDNA was calibrated to 100 ng/μl for qRT-PCR experiments.

The Real-Time Quantitative PCR Analysis

Quantitative real-time PCR analysis of candidate gene expression in cotton tissues was performed using SYBR Green (Roche, 9115 Hague Road, Indianapolis, United States) on a LightCycler® 480 System (Roche). Upland cotton *GhHistone3* (*AF024716*) was used as an internal reference gene (Xu et al., 2004). Primers for candidate genes and internal references are detailed in **Supplementary Table 3**. Each reaction was carried out in a 384-well plate in a volume of 10 μl. The amplification was performed in a two-step cyclic reaction process as described by Ren et al. (2017). Relative expression levels were analyzed using the LightCycler® 480 Gene Scanning Software. Gene expression levels were calculated according to the $2^{-\Delta\Delta Ct}$ method based on three independent PCR amplifications (Livak and Schmittgen, 2001).

[1]https://CRAN.R-project.org/package=pheatmap

Ovule Culture and Chemical Treatment

Ovule culture was performed using ovules from sGK9708. Cotton flowers were tagged on the day of anthesis as 0 DPA flowers, and ovules were sampled from the developing bolls from the tagged flowers at 1 DPA. After harvest, ovules were immediately sterilized and cultured in BT medium (Beasley, 1973) that contains 0 (as control), 0.5, 1, and 5 μM nucleotide sugar UDP-Glc, respectively. The culture was incubated at 30°C in the dark. The nucleotide sugar UDP-Glc (CAS: 28053-08-9) was purchased from Sigma-Aldrich (Sigma Corporation of America). UDP-Glc was prepared with double-distilled water (ddH_2O) to a stock concentration of 5 mM and sterilized by filtration through a 0.22-μm Millex filter (Millipore). In the medium preparation, the stock solution was diluted to the required working concentration corresponding to the treatments. At 7 DPA, the ovules in culture were sampled to evaluate the phenotypes of fiber development. Because the fibers are too fragile to stand any mechanical processing in the early stage of culture, the phenotypes of fiber development were measured intact using digital images in the form of fiber mass areas to elucidate the effect of precursor substances on fiber development. Digital images of the fibers intact on the cultured ovules were recorded with Leica DFC7000 T stereo microscope (Leica M165 FC) under the same parameters of shooting distance, light source, resolution, and magnification. The area of the fiber mass was measured and calculated using ImageJ software (version 1.52v). All experiments were in three independent replications, and 30 ovules per row were selected for fiber mass area assessment.

Genetic Sequence Retrieval and Phylogenetic Analysis

To identify candidate gene family members, the hidden Markov model (Pfam: PF00984\PF03720\PF03721) was downloaded from the Pfam website and renamed as UGD.hmm. The gene sequence retrieval was performed with the UGD.hmm model using the hmmer search program in HMMER 3.0 software (Finn et al., 2011). The UGD gene family sequences of *A. thaliana* were used to retrieve the genome data of nine species. The preliminary identification of UGD protein ID was searched in the following published protein sequences: *Gossypium hirsutum* (Hu et al., 2019)[2], *Gossypium arboretum* (Du et al., 2018)[3], *Gossypium barbadense* (Hu et al., 2019)[4], and *Gossypium raimondii* (Paterson et al., 2012)[5]. The genomes for comparison analysis, including *Theobroma cacao*, *Oryza sativa*, *A. thaliana*, *Populus trichocarpa*, and *Brachypodium distachyon* genomes were downloaded from ePhytozome 2.1 database[6]. The Basic Local Alignment Search Tool (BLAST) was performed to search the protein sequences from the four species using protein–protein BLAST (BLASTP)[7] program. The results obtained were compared to the results of the previous hmmer search. Since the genes in the annotation files were corresponded to multiple transcripts, redundant transcripts were manually discarded.

Phylogenetic Tree Construction, Motif, and Exon/Intron Structure Analyses

Phylogenetic analysis was performed using the conservative regions that were obtained through multiple sequence alignments. The phylogenetic tree was constructed using the MEGA 7 program (Tamura et al., 2013) using the neighbor-joining (NJ) method with 1,000 lead replicates.

Structural information of the *UGD* genes was obtained from the GFF3 file, and the predicted coding sequence was aligned with the corresponding genomic sequence using the Gene Structure Display Server (GSDS; Hu et al., 2015) to search for the exon/intron structure. The conservative motifs of candidate genes were determined using the online program MEME[8], and the identified motifs were annotated using the InterProScan program (Hu et al., 2015). TBtools[9] was used to visualize gene motifs. The isoelectric point (PI) and molecular weight (Mw) were calculated online the ExPASy website[10].

The values of non-synonymous substitution rate (Ka) to synonymous substitution rate (Ks) were evaluated using KaKs Calculator 2.0 software to analyze the selection pressure of UGD genes.

Promoter Region Retrieval and Analyses

Promoter regions of about 2,000 bp upstream of the start codon of the candidate gene were extracted from the genomic sequence of *G. hirsutum*, and the *cis*-acting regulatory elements (CARE) were thus analyzed using the online PlantCARE software[11] (Lescot et al., 2002). In addition, all CAREs in the promoter region of the gene of interest were classified according to their efficacies in plant responses and development mechanisms.

Subcellular Localization Assay

The coding region of *GH_D12G1806* was isolated from the cDNA template, which was derived from the mixture of fiber samples of 5DPA, 10DPA, 15DPA, and 20DPA developing bolls of sGK9708, by hi-fi PCR amplification using primer pair of F: 5′-TCTAGAATGGTGAAGATCTGTTGCATCG-3′ and R: 5′-TTAATTAATGCCACAGCTGGCATGTCC-3′. The amplified product was cloned into the GFP-containing vector *pCambia2300* in an expression cassette driven by the 35S promoter according to Witte et al. (2004). Tobacco (*N. benthamiana*) leaves were co-infiltrated with *Agrobacterium* strains containing the GFP coupled with *pCambia2300* according to Witte et al. (2004). After 48 h of infection following the procedure described previously (Maiti et al., 1988; Goodin et al., 2002), leaves were observed using an Olympus FV1200 (Olympus, Tokyo, Japan) confocal laser scanning microscope. Fluorescent images of EGFP were collected to detect the subcellular localization of *UGD*.

[2]https://cottonfgd.org/about/download/assembly/genome.Ghir.ZJU.fa.gz
[3]https://cottonfgd.org/about/download/assembly/genome.Garb.CRI.fa.gz
[4]https://cottonfgd.org/about/download/assembly/genome.Gbar.ZJU.fa.gz
[5]https://cottonfgd.org/about/download/assembly/genome.Grai.JGI.fa.gz
[6]http://www.phytozome.net
[7]https://blast.ncbi.nlm.nih.gov/Blast.cgi
[8]http://meme-suite.org/
[9]https://github.com/CJ-Chen/TBtools
[10]http://web.expasy.org/compute_pi/
[11]http://bioinformatics.psb.ugent.be/webtools/plantcare/html/

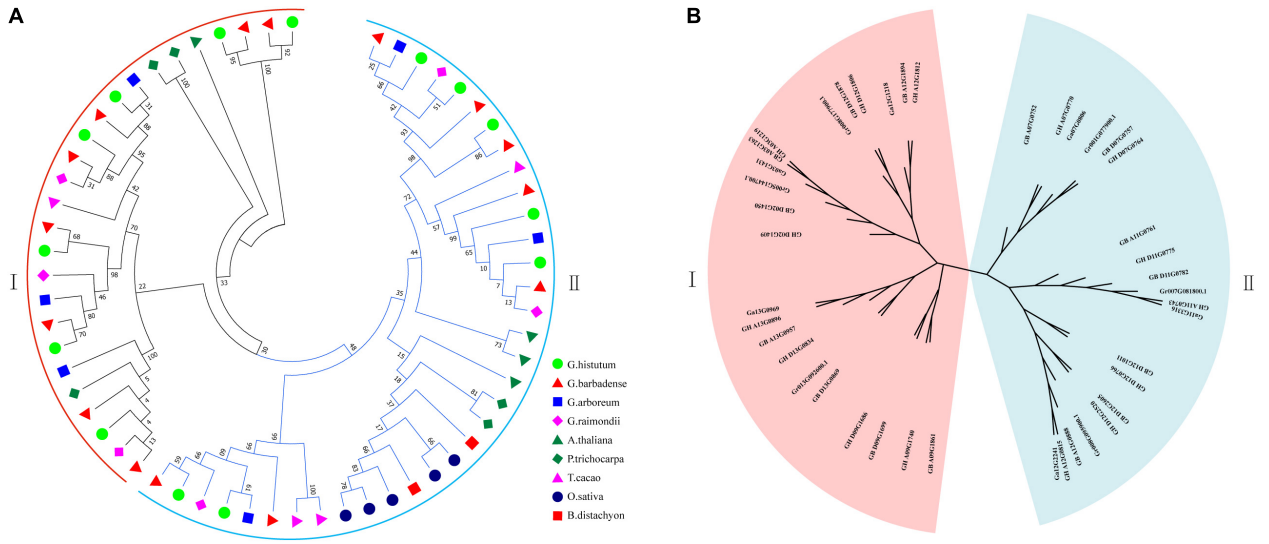


FIGURE 1 | Phylogenetic trees of *UGDs*. **(A)** Phylogenetic tree of 61 *UGDs* from 9 species, including *G. hirsutum*, *G. barbadense*, *G. arboreum*, *G. raimondii*, *A. thaliana*, *T. cacao*, *P. trichocarpa*, *B. distachyon*, and *O. sativa*; **(B)** phylogenetic tree of 42*UGDs* from four *Gossypium* species.

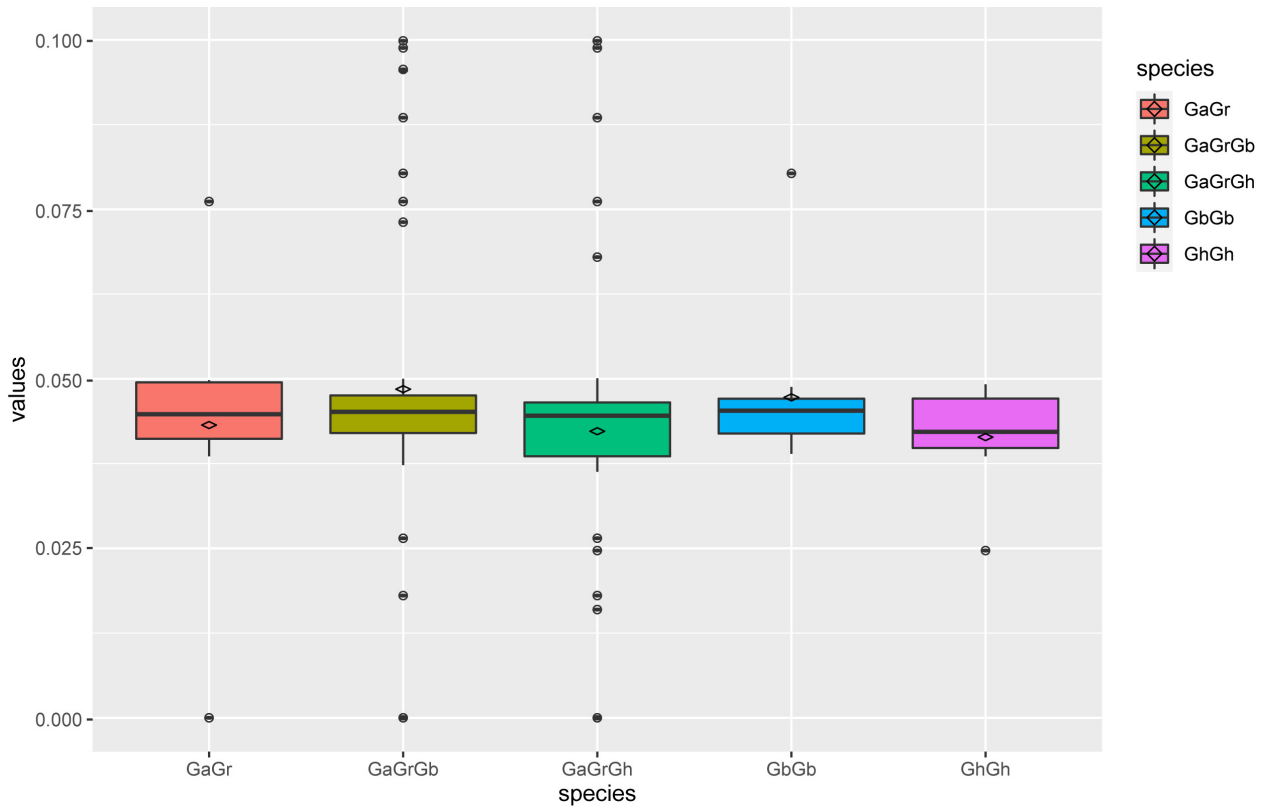


FIGURE 2 | Multiple comparison of Ka/Ks values of homologous pairs of UGD family in four *Gossypium* species.

Transformation of *A. thaliana*

GH_D12G1806 was isolated using the same cDNA template and procedures as described in the *Subcellular Localization Assay* section with primer pairs of forward: 5′-TCTAGAGGTGAAGATCTGTTGCATTGGA-3′ and reverse: 5′-GAGCTCTGCAGGCATGTCCTTGAGC-3′. The amplified product was cloned into vector *pCambia2300* in an expression cassette driven by the 35S promoter. Then, the constructed

vector was introduced into *Agrobacterium tumefaciens GV3101* (Clough and Bent, 1998). Floral dip method was used for *Agrobacterium*-mediated transformation into *A. thaliana*. Phenotypic identifications were performed in T_3 generation plants of transgenic *A. thaliana*.

RESULTS

Identification of the *UGD* Gene Family Numbers

A total of 42 total *UGD* genes from *G. hirsutum*, *G. arboreum*, *G. barbadense*, and *G. raimondii* genomes were identified, including 15 *GhUGDs*, 15 *GbUGDs*, 6 *GaUGDs*, and 6 *GrUGDs* (**Supplementary Table 1**). In order to elucidate their evolutionary and phylogenetic relationships, *UGD* genes from extra five species, including four in *A. thaliana*, four in *T. cacao*, four in *P. trichocarpa*, two in *B. distachyon*, and five in *O. sativa*, were identified and analyzed (**Supplementary Table 2**).

Phylogenetic Analysis, Gene Structure, and Protein Domain of the *UGD* Gene Family

According to the topological structure of the phylogenetic tree, the *UGD* genes in the nine species can be assorted into two subfamilies, *UGD*-I and *UGD*-II, containing 28 and 33 members, respectively, (**Figure 1A**). The result of phylogenetic analysis indicates that the *UGDs* have a closer evolutionary relationship among the four *Gossypium* species as compared with those among other species. Further phylogenetic analysis of *UGDs* in the four cotton species indicates that the *UGD*-I are assorted into four subgroups, while the *UGD*-II three (**Figure 1B**). Each subgroup of *Gossypium UGDs* consists of six members including one from A genome (*G. arboreum*; Du et al., 2018; see text footnote 3), one from D genome (*G. raimondii*; Paterson et al., 2012; see text footnote 5), two from *G. hirsutum*, and two from *G. barbadense*. As both *G. hirsutum* and *G. barbadense* are consisted of A_t and D_t subgenomes, each subgenome provides one member in a sub-group of *UGD* family. There is one subgroup in *UGD*-I that has four *UGDs*, in which there is no *UGD* from *G. arboreum* and *G. raimondii* identified. The *UGD*-II has one subgroup that has eight *UGDs* including three from *G. hirsutum* and three from *G. barbadense* (**Figure 1B**). The PI values and Mws of the four cotton *UGDs* range from 5.00 to 7.39 and from 29.41 to 56.50 kD, respectively. CDS lengths of most *UGD* genes are between 777 and 2,544 bp. The protein sequences that these genes encode consist of 258–847 amino acids (aa; **Supplementary Table 1**). The CDS and protein lengths of the *UGD* genes suggest that the *UGD* genes are in a conservative mode in genome evolution among the four *Gossypium* species. Ka/Ks value analysis reveals that the Ka/Ks values of most UGD homologous pairs are below 0.1 (**Figure 2**), indicating purifying selection pressure of UGD gene family during its evolution (Ge et al., 2020).

In order to comprehensively study the phylogenetic relationship among *UGD* genes in four cotton species, we performed analyses of gene structures and protein domains. The

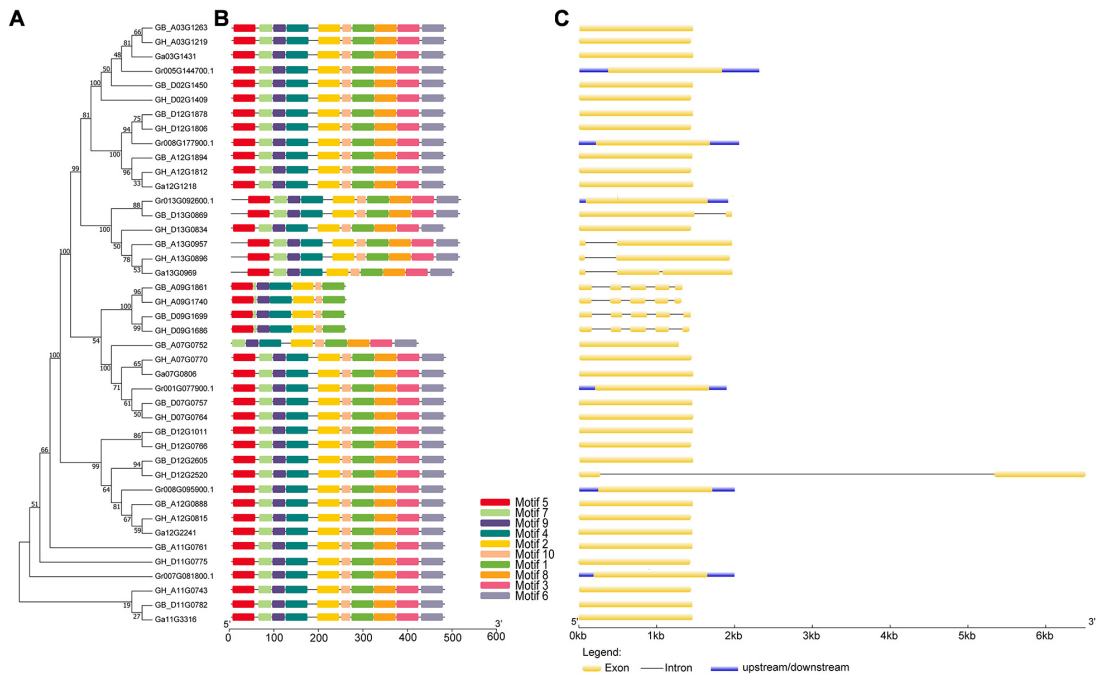


FIGURE 3 | Evolutionary relationship, motif, and gene structure analysis of the UGD family in *Gossypium* species. **(A)** Phylogenetic relationships; **(B)** conserved motifs of UGD proteins. Motif compositions were determined using MEME. **(C)** Gene structure of UGD genes. The neighbor-joining tree was constructed with MEGA 7. I and II indicate the two subfamilies of UGDs. Yellow and blue bars represent CDS and UTRs, respectively.

results reveal that *UGD* genes share similar motif and genetic structures (**Figures 3B,C**). It demonstrates that the exon number variations are exclusively identified in *UGD*-I subfamily, in which its members contain one–five exons, and most genes only have only one exon. UTR structures are identified exclusively in *UGD* genes of *G. raimondii*. These results suggest that *UGD* family in *Gossypium* is evolutionarily conservative, while some *UGD* genes in *G. raimondii* may be evolutionarily active (**Figure 3C**).

Cis-Acting Regulatory Elements in the Promoter Regions of *UGD* Genes

Analysis of the CAREs within the 2,000 bp upstream region from the start codon (ATG) of *GhUGDs* revealed that each gene promoter region contains multiple CAREs. Totally, 14 potential CAREs are identified in *UGD* promoter regions. The composition and distribution of CAREs vary significantly across the whole UGD gene family. The homologous *UGD* genes have same or similar CAREs composition and distribution in the same subfamily. Detailed analyses reveal these CAREs mainly consist of light response, hormone response, and abiotic stress responses (including drought, low temperature, defense, and stress) and specific organ/tissue development elements (**Figure 4**). Light-responsive elements include AT1-motif, TCT-motif, GT1-motif, AAAC-motif, TCCC-motif, GATA-motif, GA-motif, GA-motif, TGA-element, I-box, MRE, Box4, G-Box, ACE, chs-CMA, and Gap-box. Hormone-responsive elements include salicylic acid-responsive element (TCA-element), gibberellin-responsive elements (including P-box, GARE-motif, and TATC-box), abscisic acid-responsive element (ABRE), and auxin-responsive elements (AuxRR-core, TGA-element). Stress response elements include TC-rich repeats (as defense and stress responsive elements), MBS (drought-inducibility elements), and LTR (low-temperature-responsive element). Overall, these results suggest that the distribution and composition of these diverse CAREs may shape the expression of *UGD* family genes in responses to various stimuli of light, hormones, abiotic stresses, defense signal transduction, and cotton plant development.

Expression Pattern Analyses of *UGD* Genes in Developing Fibers

To explore the potential function of *UGD* genes in developing fibers of upland cotton, we analyzed the gene expression

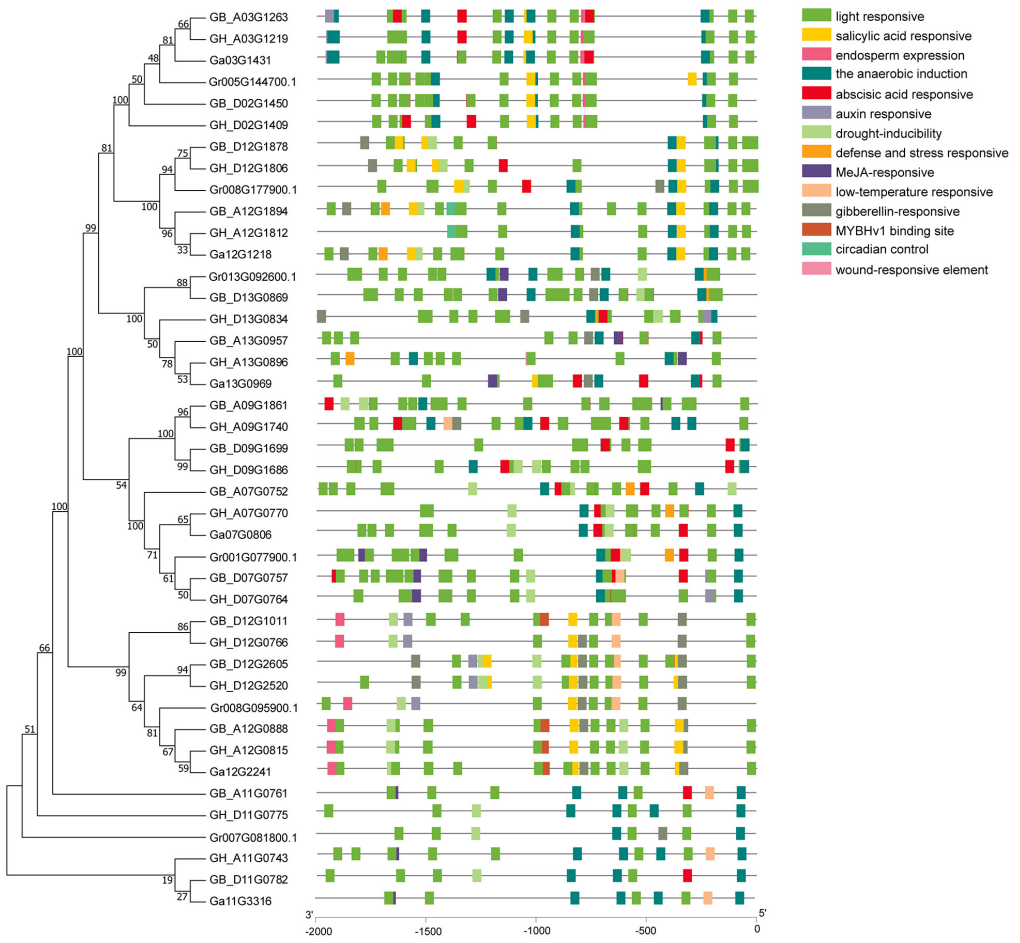


FIGURE 4 | *Cis*-acting regulatory elements analysis of cotton *UGD* genes.

patterns of the *UGD* gene families using two cotton fiber transcriptome datasets. One is dataset of *TM-1* from NCBI SPA database (accession number PRJNA490626, https://www.ncbi.nlm.nih.gov/bioproject/PRJNA490626; Hu et al., 2019) and the other is RNA-Seq raw dataset of sGK9708 and 0-153 derived from Zhang Z. et al. (2020). The results of expression analysis show that *GH_D12G0766*, *GH_D12G1806*, *GH_A12G1812*, and *GH_A11G0743* from *G. hirsutum* have significant FPKM values in developing fibers both in *TM-1* (**Figure 5A** and **Supplementary Table 3**) and in current commercial cultivars sGK9708 and 0-153 (**Figure 5B** and **Supplementary Table 3**). The result indicated that these genes had a similar expression profiles across various materials in our study. qRT-PCR verifications further confirmed their expression specificities in developing fibers of sGK9708 and 0-153 (**Figure 5C** and **Supplementary Table 4**) and their expression consistencies with transcriptome analysis. Promoter response elements affect downstream gene expression (Li and Zhang, 2014). In these genes, by observing the promoter elements (**Figure 4**), it was found that only the *GH_D12G1806* promoter region contained abscisic acid (ABA) response elements. Studies have shown that shoot ABA application promotes root growth by accelerating both cell division and elongation of the root tips (Lin and Kao, 2001). ABA synthesis regulates fiber development (Zhang G. et al., 2020). Therefore, we chose *GH_D12G1806* as our candidate genes for the next step of functional verification.

Effect of UDP-Glc on Fiber Development

UDP-glucose is the substrate of *UGD* to synthesize UDP-GlcA, which is a precursor of pectin (Hofte and Voxeur, 2017). To monitor the effects of UDP-Glc on fiber development, different concentrations of UDP-Glc were added to the media of ovule culture. The areas of fiber cluster under ovule culture were evaluated, and the result shows that the effects of UDP-Glc are significantly different at different concentrations at 7 DPA after ovule culture as compared to that of controls (**Figures 6A,B**). It shows that 1 μM of UDP-Glc has the most significant effect on fiber development. The growth of fibers can be effectively promoted in the ovules that are cultured on the medium containing UDP-Glc. qRT-PCR analysis

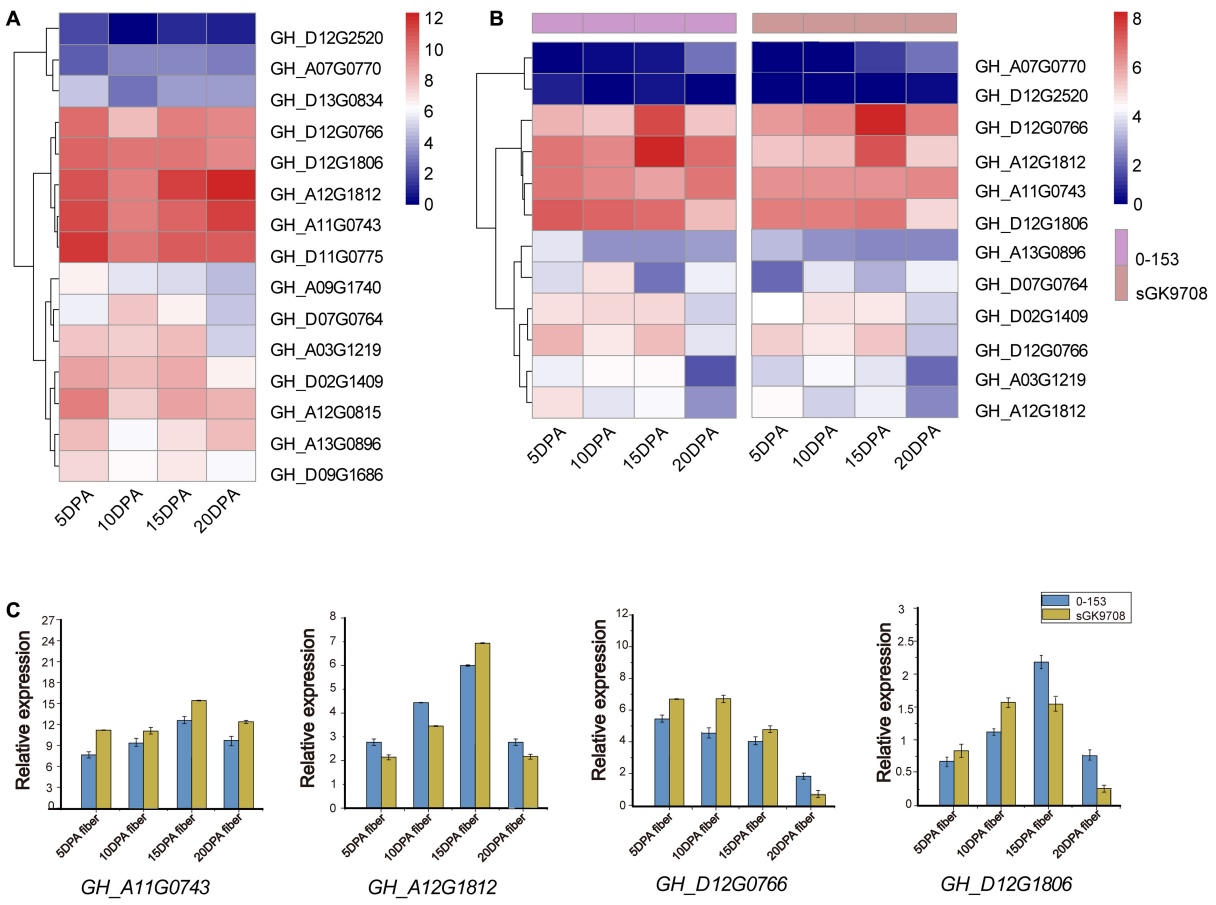


FIGURE 5 | Gene expression patterns among different varieties. Heat maps of gene expression levels in cotton fibers are shown for fibers (indicated at the bottom of each lane). **(A)** Heat map of *UGD* gene expression levels in fiber samples collected on 5, 10, 15, and 20 DPAs in TM-1 (Hu et al., 2019). The expression level was scaled using the $\log_2$ (FPKM + 1) method. **(B)** Heat map of *UGD* gene expression levels in fiber samples collected on 5, 10, 15, and 20 DPAs in two upland cotton varieties 0-153 and sGK9708. **(C)** qRT-PCR analysis of some *UGD* members in 0-153 and sGK9708. Expression levels are shown relative to the reference gene *GhHis3*. Error bars represent the standard deviations of three independent experiments.

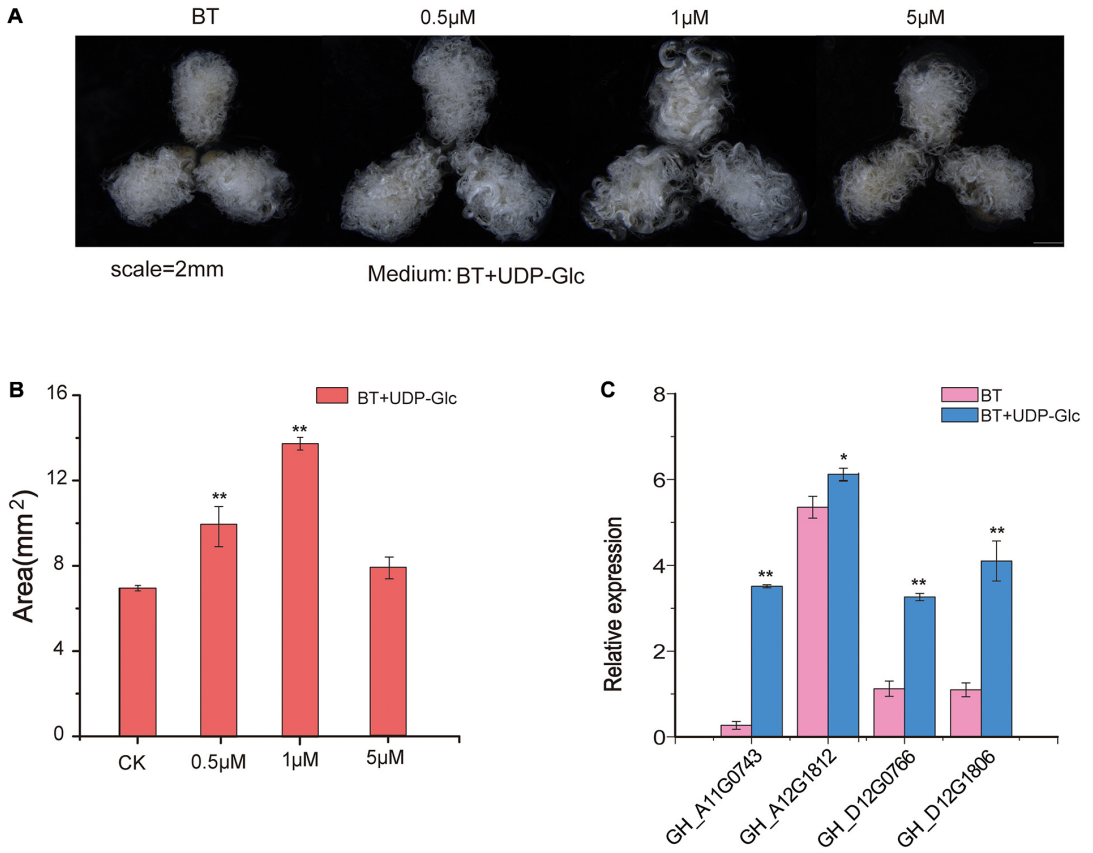


FIGURE 6 | Phenotypic evaluation of fibers after addition of UDP-Glc precursors to ovule culture. **(A)** Phenotypic variations of the area of fiber mass in different concentrations of UDP-Glc precursors applied in the culture media. **(B)** Statistical analysis of the phenotypic variations of the area of the fiber mass in different concentrations. $^{**}P < 0.01$ **(C)** The qRT-PCR verification of relative expression levels of *GH_D12G1806* ovules at 7 DPA after UDP-Glc treatments. CK represents control in which UDP-Glc is not applied and ovules only grows on BT medium. $^{*}P < 0.05$, $^{**}P < 0.01$.

demonstrates that the expression patterns of *GhUGD* genes are significantly up-regulated in UDP-Glc treatment as compared to those in control, which shows a consistency with our speculations (**Figure 6C**).

Subcellular Localization of the Expression Product of *GH_D12G1806*

To elucidate the subcellular localization of *GhUGDs*, the coding region of *GH_D12G1806* was isolated and constructed into a co-expression construct of *GH_D12G1806-GFP*. Leaves of *N. benthamiana* were infiltrated with *A. tumefaciens* transformed with the co-expression construct to detect its transient expression. Confocal imaging of the permeated tobacco leaves shows that the *GhUGD*-GFP signal are present on the cell membrane, indicating that the *UGD* genes are expressed on the cell membrane (**Figure 7**).

Over-Expression of *GH_D12G1806* in *A. thaliana*

To verify *GhUGD* genes function in *A. thaliana*, the coding region of *GH_D12G1806* was constructed in an over-expression cassette driven by 35S promoter. When the over-expression cassette was introduced into *A. thaliana* (*Columbio-0, Col-0*), totally 10 tranformants were obtained in T_0 generation. Genetic screening in T_2 generation showed seven single-copy insertion lines were obtained. In T_3 generation, the phenotypic variations of two transgenic lines L1 and L2 were evaluated (**Figure 8**). At flowering stage, the number of trichomes on the sepals of transgenic lines L1 and L2 increased significantly, which were 155% and 160% of that of the control plants, respectively, (**Figures 8A,D**). Compared with the *Col-0*, most of the trichomes on the main stem of L1 and L2 showed bifurcated phenotypes (**Figure 8B**). The results show that root lengths of L1 and L2 increase significantly by 121% and 125%, respectively, as compared with those of the control *Col-0* plant at 7 days after germination (**Figures 8C,E**). To confirm these phenotypic variations, the expressions of *GH_D12G1806* in both the transgenic lines (L1 and L2) and wild-type *Arabidopsis* plants were verified through qRT-PCR. The results of qRT-PCR verification demonstrate that there is no detectable expression of *GH_D12G1806* in the control plant and that expression levels in both L1 and L2 are significantly high (**Figure 8F**). These results indicate that over-expression of *GH_D12G1806* in transgenic lines may promote trichome initiation and root length in the transgenic *Arabidopsis* plant.

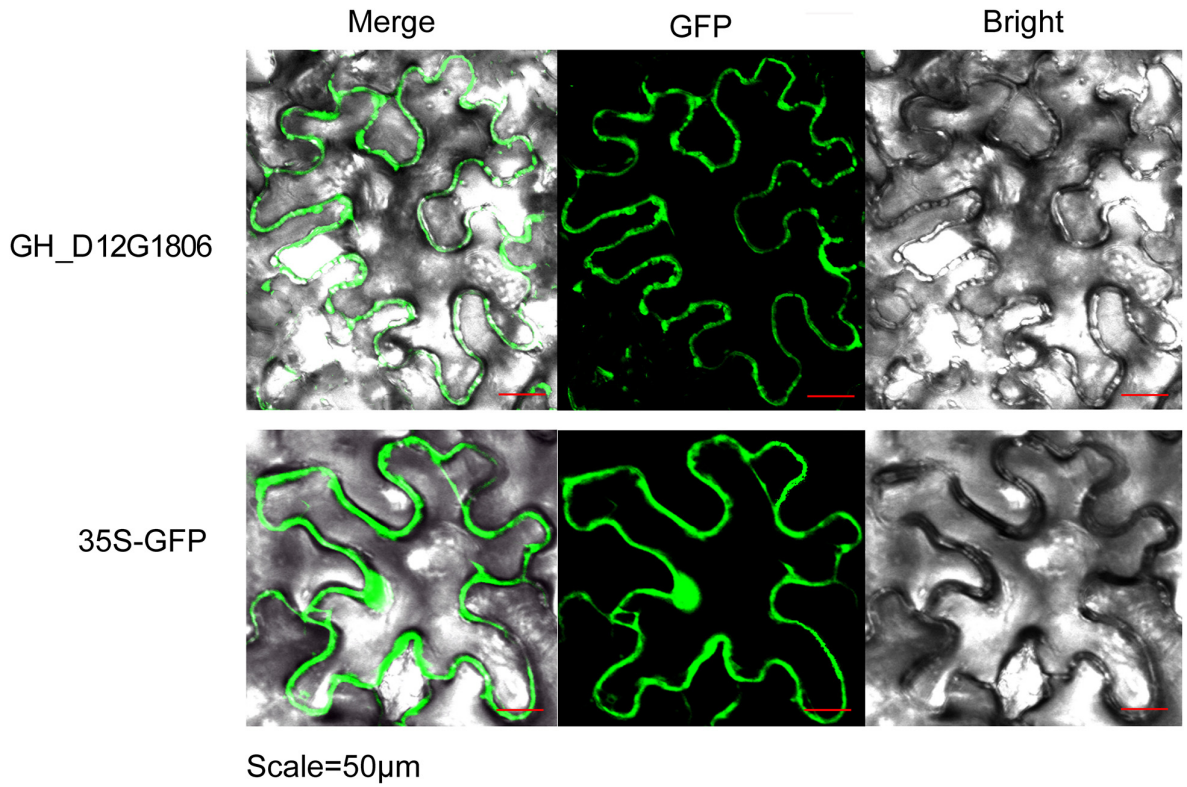


FIGURE 7 | Subcellular localization of *GH_D12G1806* in bright field, fluorescence (GFP), and merged images of tobacco leaves infiltrated with *Agrobacterium*.

DISCUSSION

UDP-glucose dehydrogenase catalyzes the NAD^+-dependent two-fold oxidation of UDP-Glc to UDP-GlcA, which is a cell wall-specific biochemical precursor required for the synthesis of much of the cell wall biomass including hemicellulose and pectin (Endres and Tenhaken, 2011), especially in newly formed plant cell walls (Tenhaken and Thulke, 1996; Johansson et al., 2002; Griffith et al., 2004). The *UGDs* had been extensively studied in pathogenic microbes where it played multiple functions including keeping cell wall integrity of *C. neoformans* (Bar-Peled et al., 2004) and maintain virulence of *Cryptococcus neoformans*. In maize, UGDs play an important role in cell wall pentose biosynthesis. However, the characterization and functional analysis of the UGD family has not been performed in *G. hirsutum*.

In this work, *UGD* gene family kept a certain stability in different diploid species including *A. thaliana* (in which four *UGD* genes were identified), *T. cacao* (four *UGD* genes), and *G. arboretum* (six *UGD* genes), and *G. raimondii* (six *UGD* genes; **Figure 1A**). We also simultaneously identified *UGD* genes in four representative cotton species, two tetraploids of *G. hirsutum* and *G. barbadense* and two diploids of *G. arboreum* and *G. raimondii*, from the ancestors of which the tetraploids were formed 1.5 million years ago (Hu et al., 2019). We identified 15 *UGD* genes each in *G. hirsutum* and *G. barbadense* and 6 *UGD* genes each in *G. arboreum* and *G. raimondii* (**Figure 1B**). These results indicate that UGD gene family in *G. hirsutum* and *G. barbadense* may expand through duplications during their evolutions, which is a common phenomenon in plant genome evolutions (Schaper and Anisimova, 2015; Ge et al., 2020). On the other hand, gene losses in both allotetraploid diploid species are also observed (Li et al., 2015; Hu et al., 2019). However, how these forces drive the formation of *UGD* gene family is still open to elucidation.

The protein sequences of UGDs are highly conservative in the plant, which can be inferred from the observations that more than 90% sequences were identical between different isoforms (Klinghammer and Tenhaken, 2007). In the current study, bioinformatics analysis of the *GhUGD* gene family revealed conserved motifs and gene structures (**Figure 3**). Together with purifying selection pressure of UGD homologous pairs, they further illustrate that the *GhUGD* gene family is conservative in evolution. Promoter region analysis revealed that ABRE, a *cis*-acting regulatory element of ABA-promoting ethylene synthesis, was identified in some *GhUGDs*. Ethylene participates in the regulation of specific types of cell growth and promotes elongation of fibroblasts *in vitro* (Shi et al., 2006; Pang et al., 2010). In *A. thaliana*, most genes containing ABRE elements are highly expressed in leaves and roots (Kanter et al., 2005). In the current study, the genes specifically expressed in developing fibers also harbored ABRE in their promoter regions. Therefore, the ABRE effects in *GhUGD* expression in developing fibers are worth of further studies.

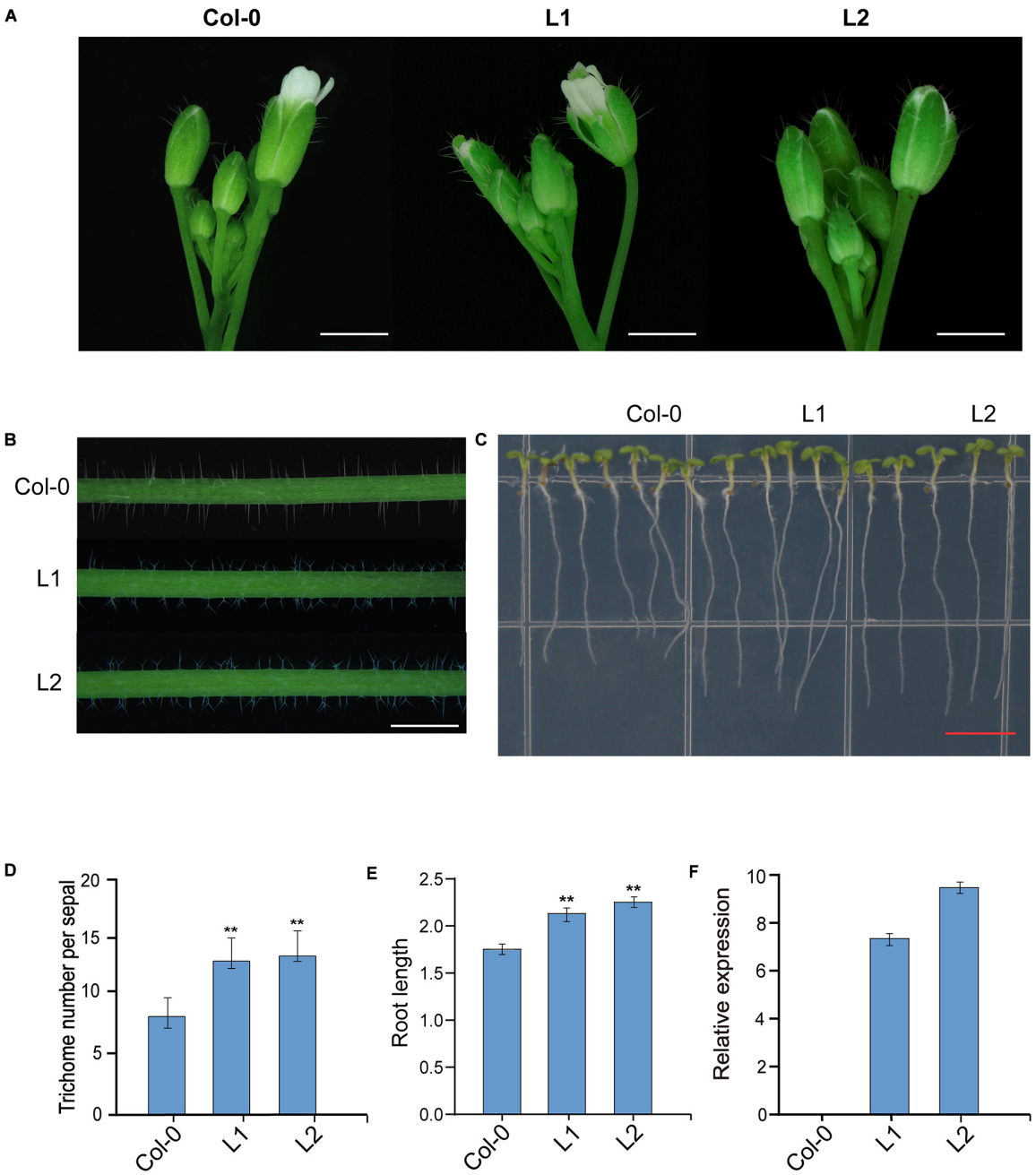


FIGURE 8 | Overexpression of the *GH_D12G1806* gene in *A. thaliana*. **(A)** Trichome initiation on the sepals of wild-type *Col-0* and transgenic lines L1 and L2. Scale bars = 0.1 mm. **(B)** Trichomes on the main inflorescence stems of *Col-0*, L1, and L2 plants. Scale bars = 1.0 mm. **(C)** Root phenotypes of *Col-0*, L1, and L2. Scale bars = 1 cm. **(D)** Statistical analysis of the phenotypic variations of the densities of trichomes in the sepals from *Col-0* and L1 and L2. Error bars represent the SD of three biological replicates. $^{**}P < 0.01$. Scale bars = 2.0 mm. **(E)** Statistical analysis of the phenotypic variations of root length. Error bars represent the SD of three biological replicates ($n = 30$ each) $^{**}P < 0.01$. **(F)** qRT-PCR verification of the relative expression of *GH_D12G1806* in *Col-0*, T1, and T2.

UDP-GlcA can be synthesized alternatively by the MIOX or UGD in *A. thaliana* (Seitz et al., 2000; Kanter et al., 2005). However, biochemical labeling experiments of cell cultures with *A. Thaliana* suggested a dominance of the UGD pathway for UDP-GlcA formation (Sharples and Fry, 2007). Down-regulation of UDP-GlcA biosynthesis leads to swollen plant cell walls and severe developmental defects associated with changes in pectic polysaccharides (Reboul et al., 2011). Also, recent research has demonstrated that the double mutant *ugd2ugd3* shows defects in seedling development, slow growth, dwarfism, and low seed-set rate, suggesting that UGD enzymes are of great importance for the integrity of the cell wall. Although *UGD* genes are mostly

expressed in roots and leaves in *A. thaliana*, the results of current study revealed *UGD* genes that are also highly expressed in cotton fiber (**Figure 5**). The potential functions of these genes are still open to further analysis and verifications. Some studies suggested that on genome-wide scale, homoeologous gene pairs may not show expression bias (Zhang et al., 2015); however, in a specific tissue or developmental stage, the biased expression might be correlated to subfunctionalization of these homoeologous genes (Zhang et al., 2015). These expression profiling might indicate the functional relevance. A previous study indicated that the exogenous application of UDP-Glc into ovule culture solution significantly promoted cotton fiber elongation and increased *UGD* gene expression (Pang et al., 2010). The transcriptome analysis in the current study at least identified four *GhUGD* genes highly expressed in developing fibers. Subcellular localization of *GH_D12G1806* from *G. hirsutum* revealed that its functional site was mainly located in the cell membrane, which inferred that the enzyme UGD may play an important role in cotton fiber development. When UDP-Glc was applied to the ovule culture solution, fiber mass growth was promoted in the culture with UDP-Glc and *UGD* gene expressions increased (**Figure 6**), which is consistent with the results of Pang et al. (2010). The results also suggest that different UDP-Glc concentrations may modulate its final effect on the development of fibers.

Plant growth is largely determined by cell elongation, a process that is strongly controlled at the level of the cell wall. Down-regulation of UGD-GlcA reduces pentose contents in the cell walls (Klinghammer and Tenhaken, 2007), which suggests that UGD as a key enzyme for the synthesis of UDP-GlcA may play an important role in cell wall synthesis. Examination of UGD function in *A. thaliana* revealed that the antisense *GbUGD6* lines had shorter roots, deferred blossoming, and activities of associated enzymes were also affected by UGD reduction, compared to wild-type plants (Han et al., 2016). In this study, we transformed *GH_D12G1806* into *A. thaliana* (*Columbia-0, Col-0*), and the over-expression transgenic lines L1 and L2 had longer root and more trichomes on sepals of flowers than control plant of *Col-0* (**Figures 8A–E**), indicating that the promoted expression of *GH_D12G1806* may enhance root cell growth and trichome initiation. However, what the mechanism is between *GH_D12G1806* expression and cotton fiber development is still unclear. As trichomes and fibers are all single epidermic cells, they share similar initiation and growth mechanisms (Wang et al., 2019). Possibly, *GH_D12G1806* works the same way it does on trichomes.

CONCLUSION

In this study, a total of 42 *UGD* genes based on the genome information of *G. raimondii*, *G. arboreum*, *G. hirsutum*, and *G. barbadense* were identified. The family could be divided into two groups based on the phylogenetic tree and gene structures. UGD gene family in *G. hirsutum* and *G. barbadense* may expand through duplications during their evolutions, which is a common phenomenon in plant genome evolutions. When UDP-Glc was applied to the ovule culture solution, fiber mass growth was promoted in the culture with UDP-Glc and UGD gene expressions increased. The over-expression transgenic lines L1 and L2 had longer root and more trichomes on sepals of flowers than control plant of *Col-0*, indicating that the promoted expression of *GH_D12G1806* may enhance root cell growth and trichome initiation. These results indicate that UGD may play an important role in cell wall synthesis and its mechanism on cotton fiber development and fiber quality formation is worth of further dissecting.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

HS and YY designed the experiments and provided the resources for all experiments. TJ coordinated the experiments. TJ, QG, SH, and WG analyzed the data and wrote the manuscript. SF and QG contributed in the plant material preparation and RNA extraction. SZ and XJ contributed to multiple sequences alignments and phylogeny analysis. YF, LZ, and DN contributed to qRT-PCR analysis. ZZ, AL, and WG discussed the results and commented on the manuscript. All authors read, edited, and approved the current version of the manuscript.

FUNDING

This work was funded by the Natural Science Foundation of China (No. 31621005), Central Public-interest Scientific Institution Basal Research Fund (No. 1610162019010101), the State Key Laboratory of Cotton Biology Open Fund (CB2017A05), and the Agricultural Science and Technology Innovation Program for CAAS (CAAS-ASTIP-ICRCAAS). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.597890/full#supplementary-material

Supplementary Table 1 | Chromosomal locus ID and length of UGD proteins in four cotton species.

Supplementary Table 2 | *UGD* gene family in nine species.

Supplementary Table 3 | *UGD* gene family and their PFKM values in TM-1, 0-153, and sGK9708.

Supplementary Table 4 | Primer pair sequence for qRT-PCR analysis of UGD genes and internal reference *GhHis3* gene.

REFERENCES

Bar-Peled, M., Griffith, C. L., Ory, J. J., and Doering, T. L. (2004). Biosynthesis of UDP-GlcA, a key metabolite for capsular polysaccharide synthesis in the pathogenic fungus *Cryptococcus neoformans*. *Biochem. J.* 381(Pt 1), 131–136. doi: 10.1042/bj20031075

Beasley, C. A. (1973). Hormonal regulation of growth in unfertilized cotton ovules. *Science (New York, NY)* 179, 1003–1005. doi: 10.1126/science.179.4077.1003

Bindschedler, L. V., Wheatley, E., Gay, E., Cole, J., Cottage, A., and Bolwell, G. P. (2005). Characterisation and expression of the pathway from UDP-glucose to UDP-xylose in differentiating tobacco tissue. *Plant Mol. Biol.* 57, 285–301. doi: 10.1007/s11103-004-7795-7

Campbell, R. E., Sala, R. F., van de Rijn, I., and Tanner, M. E. (1997). Properties and kinetic analysis of UDP-glucose dehydrogenase from group A streptococci. Irreversible inhibition by UDP-chloroacetol. *J. Biol. Chem.* 272, 3416–3422. doi: 10.1074/jbc.272.6.3416

Clough, S. J., and Bent, A. F. (1998). Floral dip: a simplified method for Agrobacterium-mediated transformation of *Arabidopsis thaliana*. *Plant J. Cell Mol. Biol.* 16, 735–743. doi: 10.1046/j.1365-313x.1998.00343.x

Du, X., Huang, G., He, S. P., Yang, Z. E., Sun, G. F., Ma, X. F., et al. (2018). Resequencing of 243 diploid cotton accessions based on an updated A genome identifies the genetic basis of key agronomic traits. *Nat. Genet.* 50, 796–802. doi: 10.1038/s41588-018-0116-x

Endres, S., and Tenhaken, R. (2011). Down-regulation of the myo-inositol oxygenase gene family has no effect on cell wall composition in *Arabidopsis*. *Planta* 234, 157–169. doi: 10.1007/s00425-011-1394-z

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39, W29–W37. doi: 10.1093/nar/gkr367

Ge, Q., Cui, Y., Li, J., Gong, J., Lu, Q., Li, P., et al. (2020). Disequilibrium evolution of the Fructose-1,6-bisphosphatase gene family leads to their functional biodiversity in *Gossypium* species. *BMC Genomics* 21:379. doi: 10.1186/s12864-020-6773-z

Goodin, M. M., Dietzgen, R. G., Schichnes, D., Ruzin, S., and Jackson, A. O. (2002). pGD vectors: versatile tools for the expression of green and red fluorescent protein fusions in agroinfiltrated plant leaves. *Plant J. Cell Mol. Biol.* 31, 375–383. doi: 10.1046/j.1365-313x.2002.01360.x

Griffith, C. L., Klutts, J. S., Zhang, L., Levery, S. B., and Doering, T. L. (2004). UDP-glucose dehydrogenase plays multiple roles in the biology of the pathogenic fungus *Cryptococcus neoformans*. *J. Biol. Chem.* 279, 51669–51676. doi: 10.1074/jbc.M408889200

Han, J., Pan, Y., Wang, X., Zhang, Y., and Ma, Z. (2016). Antisense expression of *Gossypium barbadense* UGD6 in *Arabidopsis thaliana* significantly alters cell wall composition. *Sci. China Life Sci.* 59, 213–218. doi: 10.1007/s11427-016-5004-y

Hofte, H., and Voxeur, A. (2017). Plant cell walls. *Curr. Biol. CB* 27, R865–R870. doi: 10.1016/j.cub.2017.05.025

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

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

Johansson, H., Sterky, F., Amini, B., Lundeberg, J., and Kleczkowski, L. A. (2002). Molecular cloning and characterization of a cDNA encoding poplar UDP-glucose dehydrogenase, a key gene of hemicellulose/pectin formation. *Biochim. Biophys. Acta* 1576, 53–58. doi: 10.1016/s0167-4781(02)00292-0

Kanter, U., Usadel, B., Guerineau, F., Li, Y., Pauly, M., and Tenhaken, R. (2005). The inositol oxygenase gene family of *Arabidopsis* is involved in the biosynthesis of nucleotide sugar precursors for cell-wall matrix polysaccharides. *Planta* 221, 243–254. doi: 10.1007/s00425-004-1441-0

Klinghammer, M., and Tenhaken, R. (2007). Genome-wide analysis of the UDP-glucose dehydrogenase gene family in *Arabidopsis*, a key enzyme for matrix polysaccharides in cell walls. *J. Exp. Bot.* 58, 3609–3621. doi: 10.1093/jxb/erm209

Lescot, M., Dehais, 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

Li, F., Fan, G., Lu, C., Xiao, G., Zou, C., Kohel, R. J., et al. (2015). Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. *Nat. Biotechnol.* 33, 524–530. doi: 10.1038/nbt.3208

Li, J., and Zhang, Y. (2014). Relationship between promoter sequence and its strength in gene expression. *Eur. Phys. J. E Soft. Matter.* 37:44. doi: 10.1140/epje/i2014-14086-1

Lin, C. C., and Kao, C. H. (2001). Abscisic acid induced changes in cell wall peroxidase activity and hydrogen peroxide level in roots of rice seedlings. *Plant Sci.* 160, 323–329. doi: 10.1016/s0168-9452(00)00396-4

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

Loewus, F. A., Kelly, S., and Neufeld, E. F. (1962). Metabolism of myo-inositol in plants: conversion to pectin, hemicellulose, d-xylose, and sugar acids. *Proc. Natl. Acad. Sci. U.S.A.* 48, 421–425. doi: 10.1073/pnas.48.3.421

Maiti, I. B., Hunt, A. G., and Wagner, G. J. (1988). Seed-transmissible expression of mammalian metallothionein in transgenic tobacco. *Biochem. Biophys. Res. Commun.* 150, 640–647. doi: 10.1016/0006-291x(88)90440-8

Pang, C. Y., Wang, H., Pang, Y., Xu, C., Jiao, Y., Qin, Y. M., et al. (2010). Comparative proteomics indicates that biosynthesis of pectic precursors is important for cotton fiber and *Arabidopsis* root hair elongation. *Mol. Cell. Proteomics MCP* 9, 2019–2033. doi: 10.1074/mcp.M110.000349

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

Pieslinger, A. M., Hoepflinger, M. C., and Tenhaken, R. (2010). Cloning of Glucuronokinase from *Arabidopsis thaliana*, the last missing enzyme of the myo-inositol oxygenase pathway to nucleotide sugars. *J. Biol. Chem.* 285, 2902–2910. doi: 10.1074/jbc.M109.069369

Reboul, R., Geserick, C., Pabst, M., Frey, B., Wittmann, D., Lütz-Meindl, U., et al. (2011). Down-regulation of UDP-glucuronic acid biosynthesis leads to swollen plant cell walls and severe developmental defects associated with changes in pectic polysaccharides. *J. Biol. Chem.* 286, 39982–39992. doi: 10.1074/jbc.M111.255695

Reiter, W. D. (2008). Biochemical genetics of nucleotide sugar interconversion reactions. *Curr. Opin. Plant Biol.* 11, 236–243. doi: 10.1016/j.pbi.2008.03.009

Ren, Z., Yu, D., Yang, Z., Li, C., Qanmber, G., Li, Y., et al. (2017). Genome-wide identification of the MIKC-type MADS-box gene family in *Gossypium hirsutum* L. Unravels their roles in flowering. *Front. Plant Sci.* 8:384. doi: 10.3389/fpls.2017.00384

Schaper, E., and Anisimova, M. (2015). The evolution and function of protein tandem repeats in plants. *New Phytol.* 206, 397–410. doi: 10.1111/nph.13184

Seitz, B., Klos, C., Wurm, M., and Tenhaken, R. (2000). Matrix polysaccharide precursors in *Arabidopsis* cell walls are synthesized by alternate pathways with organ-specific expression patterns. *Plant J. Cell Mol. Biol.* 21, 537–546. doi: 10.1046/j.1365-313x.2000.00696.x

Sharples, S. C., and Fry, S. C. (2007). Radioisotope ratios discriminate between competing pathways of cell wall polysaccharide and RNA biosynthesis in living plant cells. *Plant J. Cell Mol. Biol.* 52, 252–262. doi: 10.1111/j.1365-313X.2007.03225.x

Shi, Y. H., Zhu, S. W., Mao, X. Z., Feng, J. X., Qin, Y. M., Zhang, L., et al. (2006). Transcriptome profiling, molecular biological, and physiological studies reveal a major role for ethylene in cotton fiber cell elongation. *Plant Cell* 18, 651–664. doi: 10.1105/tpc.105.040303

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst197

Tenhaken, R., and Thulke, O. (1996). Cloning of an enzyme that synthesizes a key nucleotide-sugar precursor of hemicellulose biosynthesis from soybean: UDP-glucose dehydrogenase. *Plant Physiol.* 112, 1127–1134. doi: 10.1104/pp.112.3.1127

Ute, K., Thomas, P., Victor, R., Jones, D. A., Hedley, C. L., and Andreas, R. (2004). myo-Inositol and sucrose concentrations affect the accumulation of raffinose family oligosaccharides in seeds. *J. Exp. Bot.* 55, 1981–1987. doi: 10.1093/jxb/erh216

Wang, Z., Yang, Z., and Li, F. (2019). Updates on molecular mechanisms in the development of branched trichome in *Arabidopsis* and nonbranched in cotton. *Plant Biotechnol. J.* 17, 1706–1722. doi: 10.1111/pbi.13167

Witte, C. P., Noel, L. D., Gielbert, J., Parker, J. E., and Romeis, T. (2004). Rapid one-step protein purification from plant material using the eight-amino acid StrepII epitope. *Plant Mol. Biol.* 55, 135–147. doi: 10.1007/s11103-004-0501-y

Xu, Y. H., Wang, J. W., Wang, S., Wang, J. Y., and Chen, X. Y. (2004). Characterization of GaWRKY1, a cotton transcription factor that regulates the sesquiterpene synthase gene (+)-delta-cadinene synthase-A. *Plant Physiol.* 135, 507–515. doi: 10.1104/pp.104.038612

Zablackis, E., Huang, J., Muller, B., Darvill, A. G., and Albersheim, P. (1995). Characterization of the cell-wall polysaccharides of *Arabidopsis thaliana* leaves. *Plant Physiol.* 107, 1129–1138. doi: 10.1104/pp.107.4.1129

Zhang, G., Yue, C., Lu, T., Sun, L., and Hao, F. (2020). Genome-wide identification and expression analysis of NADPH oxidase genes in response to ABA and abiotic stresses, and in fibre formation in Gossypium. *Peer J.* 8:e8404. doi: 10.7717/peerj.8404

Zhang, T., Hu, Y., Jiang, W., Fang, L., Guan, X., Chen, J., et al. (2015). Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. *Nat. Biotechnol.* 33, 531–537.

Zhang, Z., Li, J., Jamshed, M., Shi, Y., Liu, A., Gong, J., et al. (2020). Genome-wide quantitative trait loci reveal the genetic basis of cotton fibre quality and yield-related traits in a *Gossypium hirsutum* recombinant inbred line population. *Plant Biotechnol. J.* 18, 239–253. doi: 10.1111/pbi.13191

Zou, X., Liu, A., Zhang, Z., Ge, Q., Fan, S., Gong, W., et al. (2019). Co-expression network analysis and hub gene selection for high-quality fiber in upland cotton (*Gossypium hirsutum*) using RNA sequencing analysis. *Genes* 10:119. doi: 10.3390/genes10020119

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.

Copyright © 2021 Jia, Ge, Zhang, Zhang, Liu, Fan, Jiang, Feng, Zhang, Niu, Huang, Gong, Yuan and Shang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

BRIEF RESEARCH REPORT
published: 13 January 2021
doi: 10.3389/fgene.2020.590535

Genome-Wide Identification and Comparative Analysis of *ARF* Family Genes in Three Apiaceae Species

Qiaoying Pei[1†], Nan Li[1†], Qihang Yang[1], Tong Wu[1], Shuyan Feng[1], Xuehuan Feng[2], Zange Jing[3], Rong Zhou[4], Ke Gong[1], Tong Yu[1], Zhen Wang[5] and Xiaoming Song[1,2,6]*

[1] College of Life Sciences/Center for Genomics and Bio-computing, North China University of Science and Technology, Tangshan, China, [2] Food Science and Technology Department, University of Nebraska–Lincoln, Lincoln, NE, United States, [3] College of Agriculture and Life Science, Kunming University, Kunming, China, [4] Department of Food Science, Aarhus University, Aarhus, Denmark, [5] Suzhou Polytechnic Institute of Agriculture, Suzhou, China, [6] School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, China

OPEN ACCESS

Edited by:
Million Tadege, Oklahoma State University, United States

Reviewed by:
Guang-Long Wang, Huaiyin Institute of Technology, China
Hui Wang, China Agricultural University, China

***Correspondence:**
Xiaoming Song
songxm@ncst.edu.cn;
songxiaoming116@163.com

[†] *These authors have contributed equally to this work*

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

Received: *01 August 2020*
Accepted: *27 November 2020*
Published: *13 January 2021*

Citation:
Pei Q, Li N, Yang Q, Wu T, Feng S, Feng X, Jing Z, Zhou R, Gong K, Yu T, Wang Z and Song X (2021) Genome-Wide Identification and Comparative Analysis of ARF Family Genes in Three Apiaceae Species. Front. Genet. 11:590535. doi: 10.3389/fgene.2020.590535

The family Apiaceae includes many important vegetables and medicinal plants. Auxin response factors (ARFs) play critical roles in regulating plant growth and development. Here, we performed a comprehensive analysis of the *ARF* gene family in three Apiaceae species, celery, coriander, and carrot, and compared the results with the *ARF* gene family of lettuce, Arabidopsis, and grape. We identified 156 *ARF* genes in all six species and 89 genes in the three Apiaceae species, including 28, 34, and 27 in celery, coriander, and carrot, respectively. The paralogous gene number in coriander was far greater than that in carrot and celery. Our analysis revealed that *ARF* genes of the three Apiaceae species in 34 branches of the phylogenetic tree underwent significant positive selection. Additionally, our findings indicated that whole-genome duplication played an important role in *ARF* gene family expansion. Coriander contained a greater number of *ARF* genes than celery and carrot because of more gene duplications and less gene losses. We also analyzed the expression of *ARF* genes in three tissues by RNA-seq and verified the results by quantitative real-time PCR. Furthermore, we found that several paralogous genes exhibited divergent expression patterns. Overall, this study provides a valuable resource for exploring how *ARF* family genes regulate plant growth and development in other plants. Since this is the first report of the *ARF* gene family in Apiaceae, our results will serve as a guide for comparative and functional analyses of *ARF* and other gene families in Apiaceae.

Keywords: *ARF* gene family, orthologous and paralogous genes, gene duplication and loss, expression pattern, Apiaceae species, phylogenetic analysis

INTRODUCTION

Celery (*Apium graveolens*), coriander (*Coriandrum sativum*), and carrot (*Daucus carota*) are three typical members of Apiaceae family. They grow all over the world and are especially famous for their fragrance and medicinal value. Celery is a multipurpose plant, which used as a vegetable and medicinal herb for treating diseases (Ballmer-Weber et al., 2002; Maljaei et al., 2019).

Coriander contains many bioactive phytochemicals and has also been used as a traditional medicine (Prachayasittikul et al., 2018). Carrot is one of the most important root vegetables around the world and is valued for its high content of beta-carotene, an essential component of vitamin A (Ahmad et al., 2019). Above all, these three Apiaceae crops are of high economical value, given their medicinal and edible properties, and therefore are a major source of income for growers.

The phytohormone auxin plays a very significant role in regulating not only plant developmental processes, such as apical dominance, later root initiation, and vascular differentiation, but also in cellular processes, including cell division, expansion, and differentiation (Wang et al., 2007; Xing et al., 2011). Auxin Response Factors (ARFs) regulate the expression of auxin-responsive genes during plant growth. The auxin response signal is transmitted to related response genes via auxin response elements (AuxREs) (Wan et al., 2014).

ARF1 was the first *ARF* gene identified in *Arabidopsis thaliana* using a yeast one-hybrid screen (Ulmasov et al., 1997). A total of 23 *ARF* genes have been identified in Arabidopsis to date (Okushima et al., 2005). With the sequencing of genomes, more *ARF* family genes have been detected in plants. For example, 39 *ARF* genes were identified in poplar (*Populus trichocarpa*) (Kalluri et al., 2007), 25 in rice (*Oryza sativa*) (Wang et al., 2007), 19 in grape (*Vitis vinifera*) (Wan et al., 2014), 19 in sweet orange (*Citrus* × *sinensis*) (Li et al., 2015), 31 in Chinese cabbage (*Brassica rapa* subsp. *pekinensis*) (Mun et al., 2012), 20 in pineapple (*Ananas comosus*) (Su et al., 2017), 20 in barley (*Hordeum vulgare*) (Tombuloglu, 2019), 31 in maize (*Zea mays*) (Xing et al., 2011), 17 in tomato (*Solanum lycopersicum*) (Kumar et al., 2011), 31 in apple (*Malus domestica*) (Luo et al., 2014), 17 in physic nut (*Jatropha curcas*) (Tang et al., 2018), and 19 in pepper (*Capsicum annuum*) (Zhang et al., 2017). However, a comprehensive analysis of *ARF* gene family in Apiaceae species has not yet been reported. Recently, more sequencing data of Apiaceae have been released, thus serving as a valuable resource for further analysis of the *ARF* gene family (Iorizzo et al., 2016; Song et al., 2020b).

In this study, we performed a comprehensive and systematic analysis of the *ARF* gene family in three representative species of Apiaceae (celery, coriander, and carrot), with the aim to (i) identify *ARF* gene family members; (ii) classify these members based on phylogenetic relationship; (iii) map *ARF* genes to chromosomes; (iv) identify the paralogous and orthologous genes; (v) explore *ARF* gene loss and duplication; and (vi) explore the gene expression patterns in three tissues of celery and coriander.

MATERIALS AND METHODS

Genome Sequence Retrieval and ARF Gene Identification

Whole genome sequences of coriander and celery were retrieved from the Coriander Genome Database (CGDB[1]) (Song et al., 2020a). The genome sequence of Arabidopsis was downloaded from the TAIR database[2]. Genome sequences of carrot (v2.0), lettuce (v5.0), and grape (Genoscope.12X) were downloaded from Phytozome (Jaillon et al., 2007; Iorizzo et al., 2016; Reyes-Chin-Wo et al., 2017).

The Pfam database was used to identify *ARF* family genes using the identifier PF06507 with an *e*-value < 1e-4 (Punta et al., 2012). Furthermore, the conserved domain database (CDD) and simple modular architecture research tool (SMART) were used to verify the identified genes (Marchler-Bauer et al., 2009; Letunic et al., 2012).

Phylogenetic Analysis of ARFs

Amino acid sequences of ARFs of celery, coriander, carrot, Arabidopsis, lettuce, and grape were used for phylogenetic analysis. First, sequences were aligned using ClustalW (Li, 2003). Then, the multiple sequence alignment was restored in the PHYLIP format. Finally, a phylogenetic tree was constructed using the maximum likelihood (ML) method with IQ-TREE (Nguyen et al., 2015), based on JTT + F + R8 model, with 1,000 bootstrap replications.

Chromosomal Location, Gene Structure, and Conserved Motif Analysis of ARF Genes

The chromosomal location of each *ARF* gene was retrieved from general feature format (gff) file, and the chromosome number, start position and end position of each gene were extracted using a Perl script. The chromosome information file was submitted to the MapChart to display the distribution of each gene (Voorrips, 2002).

The structure of *ARF* genes was drawn using Gene Structure Displayer Server 2.0 (GSDS) (Hu et al., 2015). The gff file of *ARF* gene was submitted to GSDS to illustrate the positions of exons, introns, and untranslated regions (UTRs). Conserved motif analysis was performed using the Multiple Expression motifs (Em) for Motif Elicitation (MEME) (Bailey et al., 2009).

Analysis of ARF Gene Duplication and Loss

Orthologous and paralogous genes were identified using the OrthoMCL, with an *e*-value of 1e-5 (Li et al., 2003). The relationship among *ARF* gene orthologs and paralogs was illustrated using the Circos (Krzywinski et al., 2009). Gene duplication and gene loss analyses were performed using the Notung2.9 (Stolzer et al., 2012).

Analysis of Collinearity and Duplication Type

The MCScanX was used to conduct collinearity analysis (Wang et al., 2012). Amino acid sequences were analyzed using the Blastp, with an *e*-value set at 1.0×10^{-5}.

[1]http://cgdb.bio2db.com/

[2]http://www.arabidopsis.org/

Then, collinear blocks were detected by submitting the whole genome gff and Blastp result to the MCScanX. The duplicate_gene_classifier sub-program was used to identify the duplication type.

Evolutionary Analysis of ARF Genes

Coding sequences (CDSs) of orthologous *ARF* gene pairs were aligned using the ClustalW, and the alignment file was transformed into the axt format file. The synonymous (Ks) and non-synonymous (Ka) substitution rates were calculated using the Ka/Ks_calculator 2.0 (Wang et al., 2010). The divergence time (T) was estimated using the equation, T = Ks/2r. The "r" represents neutral substitution rate (5.2×10^{-9} substitutions per site per year) (Song et al., 2020b).

Selective Pressure Analysis of ARF Gene Family

The selective pressure analysis was performed using the PAML4.9 (Yang, 2007). The ML method and codon substitution models were adopted to test the likelihood rate of positive selection. Firstly, CDSs of *ARF* genes were aligned using the ClustalW. Then, each branch of the phylogenetic tree constructed by the PhyML3.0 using amino acid sequences was analyzed to speculate ω (the ratio of non-synonymous to synonymous distances) (Guindon et al., 2010). The M0, M1, M7, and M8 models were used to calculate variation sites.

ARF Gene Expression Analysis Using RNA-Seq

RNA-seq data of *ARF* gene expression in three tissues (root, petiole, and leaf) of celery and coriander (each with three replicates) were obtained from our previous study (Song et al., 2020a). The RNA-seq data were deposited in the Genome Sequence Archive (GSA) of the BIG Data Center[3] under the accession numbers CRA001996 and CRA001658. The expression data expressed as Fragments Per Kilobase of transcript sequence per Millions base pairs (FPKM) were log2-transformed for cluster analysis, as described previously (Song et al., 2014b). Hierarchical clustering analysis was conducted using the TBtools (Chen et al., 2020).

Verification of RNA-Seq Data

The RNA-seq data of celery and coriander *ARF* genes were verified by quantitative real-time PCR (qRT-PCR). Total RNA was extracted from each sample using the RNA Kit (Tiangen, Beijing, China), and the mRNA was transcribed into cDNA using the PrimeScript cDNA Synthesis Kit (TaKaRa, Dalian, China). The resulting cDNA was used as a template for qRT-PCR, which was performed on the CFX96^{TM} Real-Time System (Bio-Rad, Beijing, China) using sequence-specific primers (**Supplementary Table 1**), with three replicates for each gene as described previously (Song et al., 2014a, 2016).

[3]http://bigd.big.ac.cn/gsa

RESULTS

Identification, Phylogenetic Analysis, and Classification of ARF Genes

We identified 28, 34, and 27 *ARF* genes in celery, coriander, and carrot, respectively, and renamed these genes according to their order on chromosomes (**Supplementary Tables 2, 3**). Lettuce shows the closest relationship with Apiaceae, and its genome sequence has been released (Reyes-Chin-Wo et al., 2017). Additionally, the *ARF* gene family of Arabidopsis and grape has been analyzed in detail (Okushima et al., 2005; Wei et al., 2006; Wan et al., 2014). Comparative analysis of the *ARF* genes of celery, coriander, and carrot with the Arabidopsis, lettuce, and grape genomes revealed 22, 26, and 19 *ARF* genes in the latter three species, respectively.

To explore the evolutionary history and relationship of the *ARF* gene family, we constructed a phylogenetic tree using 156 ARF amino acid sequences from six species, including Arabidopsis, lettuce, grape, celery, coriander, and carrot (**Figure 1**). According to the phylogenetic analysis, all ARFs were divided into four classes (I − IV), based on the topology and classification in grape. Classes I and III contained a greater number of ARFs than classes II and IV. Interestingly, the number of Arabidopsis ARFs in one branch of class I was notably higher than that of other species, and eight Arabidopsis ARFs (AtARF9, AtARF12–15, AtARF20–22) clustered together.

We further constructed another phylogenetic tree using ARF sequences of celery, coriander, and carrot (**Supplementary Figure 1**). ARFs were also divided into four classes. The classification of ARFs among three Apiaceae species was highly consistent with that of six plant species (**Figure 1**). Additionally, classification within the Apiaceae showed that class III contained the highest number of ARFs (31), followed by class IV (23), class I (21), and class II (14).

ARF Gene Structure Analysis and Conserved Motif Identification

To further explore the conservation of *ARF* family genes, we analyzed the gene structure and motifs. Full-length cDNA sequences of *ARF* genes were compared with the corresponding genomic sequences using the GSDS program (**Supplementary Figure 2A**). The number of exons in *ARF* genes varied from 1 − 15. In class I, most *ARF* genes contained 10–14 exons, with the exception of *AgARF1* and *CsARF16*. In class III, most *ARF* genes harbored 10–15 exons, except *AgARF26* and *CsARF29*. In class IV, *ARF* genes contained only 3–4 exons, which was far less than the number of exons in other classes. We found that the position and number of exons and introns in genes belonging to the same class or subclass were similar. This finding supports the phylogenetic relationship of *ARF* family genes.

We also analyzed conserved motifs in ARF genes using MEME. Eight motifs were detected (**Supplementary Figure 2B**), of which four (motifs 1, 2, 3, and 6) were common to almost all *ARF* genes. In classes I–III, most *ARF* genes contained motifs 1–4 and motifs 6–8. Interestingly, *AgARF1* contained only motif 4. In class IV, all *ARF* genes carried motifs 1–4 and motifs 6–8, except

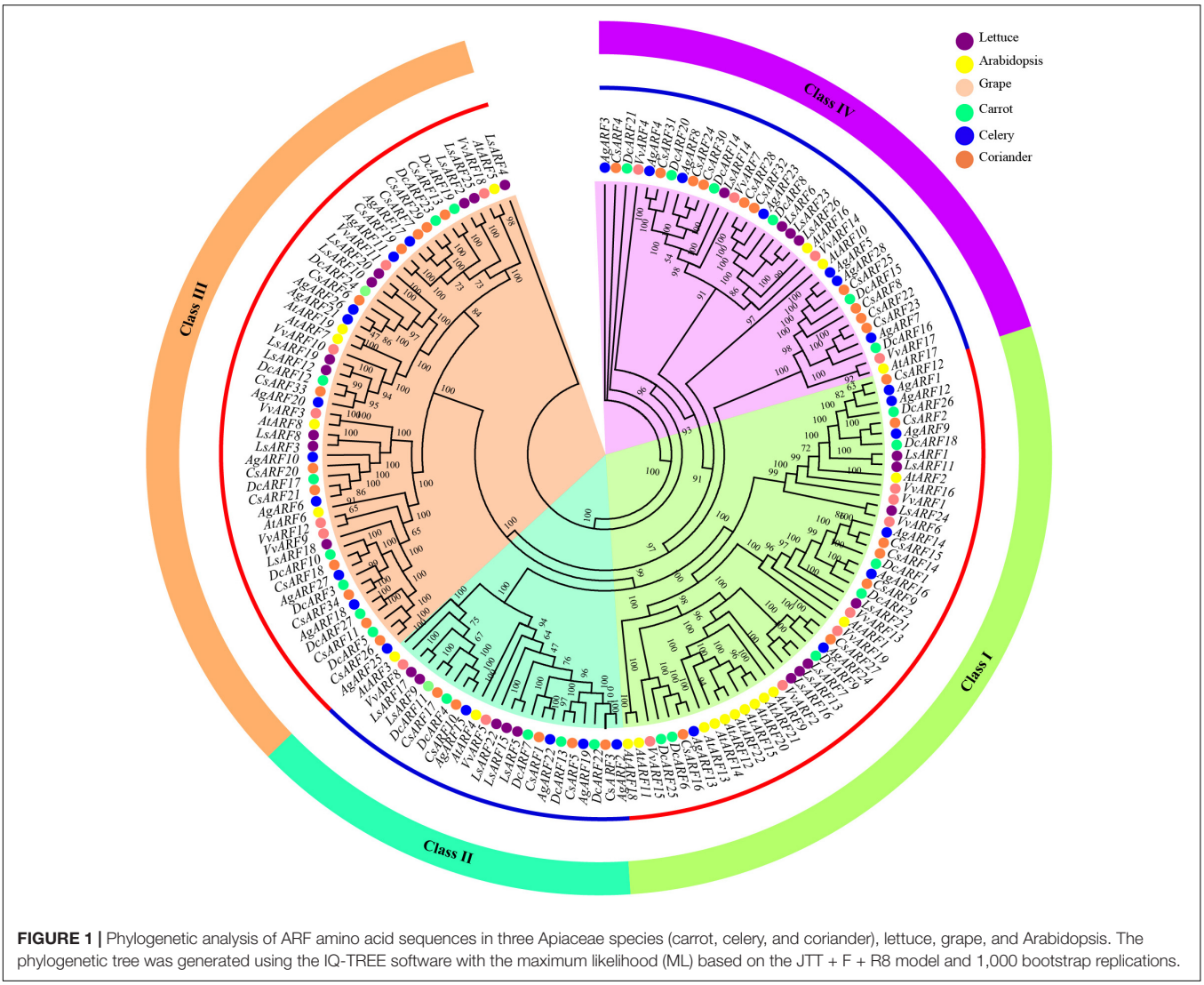


FIGURE 1 | Phylogenetic analysis of ARF amino acid sequences in three Apiaceae species (carrot, celery, and coriander), lettuce, grape, and Arabidopsis. The phylogenetic tree was generated using the IQ-TREE software with the maximum likelihood (ML) based on the JTT + F + R8 model and 1,000 bootstrap replications.

CsARF30. However, motif 5 was absent from almost all genes in class IV. In conclusion, motif 5 was lost in most *ARF* genes of celery, carrot, and coriander, while motifs 1–4 and motifs 6–8 were highly conserved in three Apiaceae species.

Chromosomal Distribution of ARF Family Genes

In celery, 25 out of 28 *ARF* genes were unevenly distributed on 10 chromosomes, and three genes could not be mapped to any chromosome (**Supplementary Figure 3A**). Chromosomes 3 and 11 harbored the highest number of *ARF* genes (four genes), but no gene was detected on chromosome 8. Several genes, such as *AgARF4*, *AgARF8*, *AgARF9*, *AgARF12,* and *AgARF25*, were located at the end of the chromosomes 2, 3, 4, 5, and 11, respectively.

In coriander, 28 out of 34 *ARF* genes were unevenly distributed on eight chromosomes, while six *ARF* genes could not be mapped to any chromosome (**Supplementary Figure 3B**). Chromosome 10 contained the highest number of *ARF* genes (nine genes), but no *ARF* gene was found on chromosomes 6 and 8. Two *ARF* gene clusters were detected at the ends of chromosome 10, which might be caused by gene duplication. *CsARF3*, *CsARF14,* and *CsARF20* were located very close to *CsARF4*, *CsARF15,* and *CsARF21*, respectively.

In carrot, all 27 *ARF* genes were mapped to one of the eight chromosomes (**Supplementary Figure 3C**). Chromosome 2 carried the highest number of *ARF* genes (seven genes), followed by chromosomes 4 and 5, whereas chromosomes 3 and 8 harbored only one *ARF* gene each.

Identification of Orthologous and Paralogous Gene Pairs

Forty orthologous gene pairs were detected between celery and coriander, while eight gene pairs could not be mapped to any chromosome. Twenty-six orthologous genes were found between celery and carrot, of which one pair could not be mapped. Totally, 29 orthologous gene pairs were identified between coriander and carrot (**Figure 2A** and **Supplementary Table 4**). Next, we

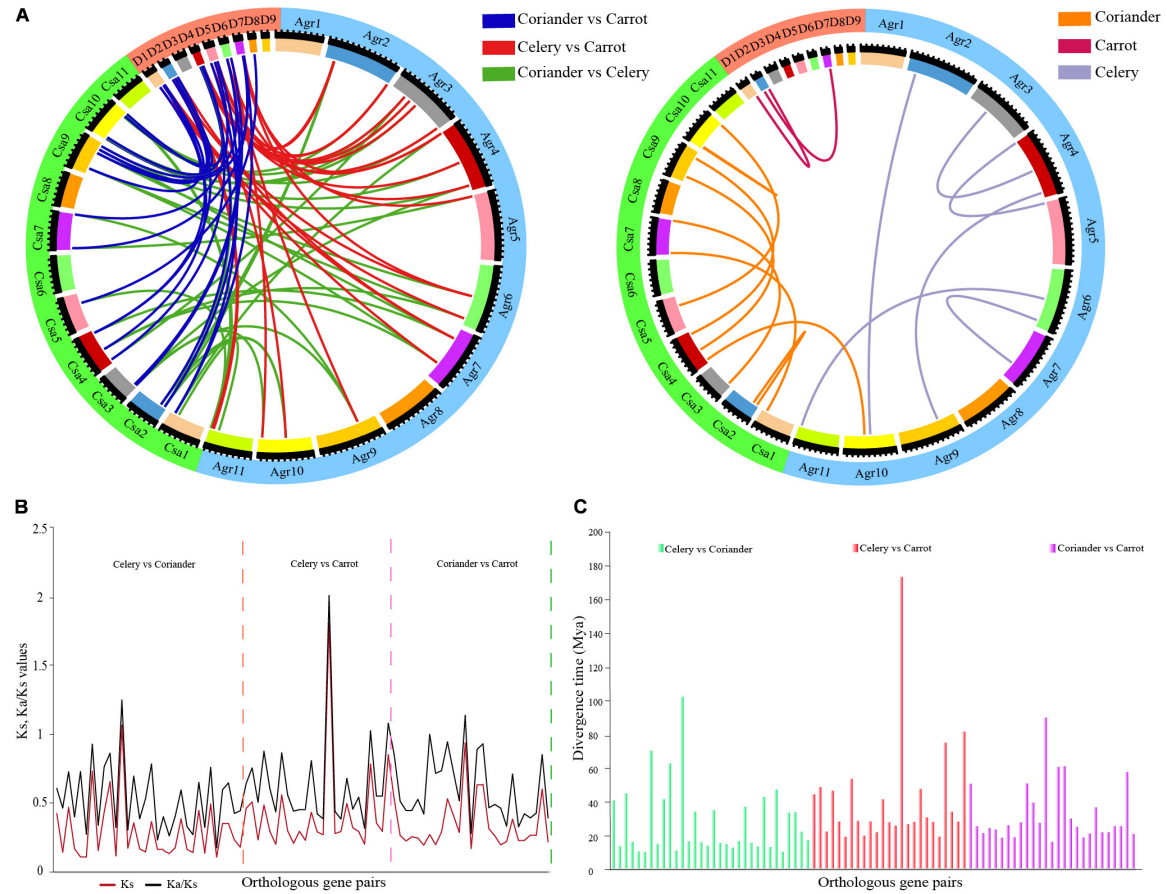


FIGURE 2 | Analysis of paralogous and orthologous *ARF* gene pairs. **(A)** Circos plot of *ARF* gene paralogs and orthologs in three Apiaceae species. **(B)** Ks and Ka/Ks values of orthologous *ARF* gene pairs between any two of three Apiaceae species. **(C)** Divergence time of orthologous *ARF* gene pairs between any two of three Apiaceae species.

identified paralogous gene pairs to explore the relationship of *ARF* genes within species (**Supplementary Table 5**). A total of eight and 12 paralogous gene pairs were detected in celery and coriander, respectively; however, only two paralogous gene pairs were identified in carrot.

We calculated the Ka/Ks ratios (**Figure 2B** and **Supplementary Table 6**) and divergence time for orthologous genes using Ks values (**Figure 2C** and **Supplementary Table 6**). The results showed the divergence time of orthologous gene pairs varied from 10.52–102.39 million years between celery and coriander, 19.38–173.48 million years between celery and carrot and 16.40–90.09 million years between coriander and carrot.

Whole-Genome Duplication (WGD) Played a Leading Role in the Expansion of ARF Gene Family in Apiaceae

Five gene duplication types were detected, including singleton, dispersed, proximal, tandem, and WGD or segmental duplication (**Figure 3A** and **Supplementary Tables 7, 8**). The results indicated that WGD or segmental duplication played a significant role in *ARF* gene family expansion in Apiaceae species. In celery, coriander, and carrot, 52.0, 57.1, and 74.1% of *ARF* genes, respectively, arose by WGD or segmental duplication. No singleton or tandem duplication was detected in *ARF* gene family of these species. Moreover, most *ARF* genes formed collinear blocks within each genome (**Supplementary Table 9**).

More Genes Were Lost in Celery and Carrot Than in Coriander After Gene Duplication

In celery, the number of *ARF* genes lost was greater than that duplicated (8 vs. 2), whereas in coriander, the number of *ARF* genes lost and duplicated was equal (**Figure 3B** and **Supplementary Figure 2**). Interestingly, in carrot, eight genes were lost, but no gene duplication was detected compared with other species. This phenomenon indicates that more genes were lost after the genome duplication in Apiaceae species. In the common ancestor of coriander and celery, one *ARF* gene duplication and two gene losses were detected. In the common ancestor of coriander, celery, and carrot, 19 genes were duplicated while three genes were lost. This phenomenon indicates that there were more gene duplications in the common

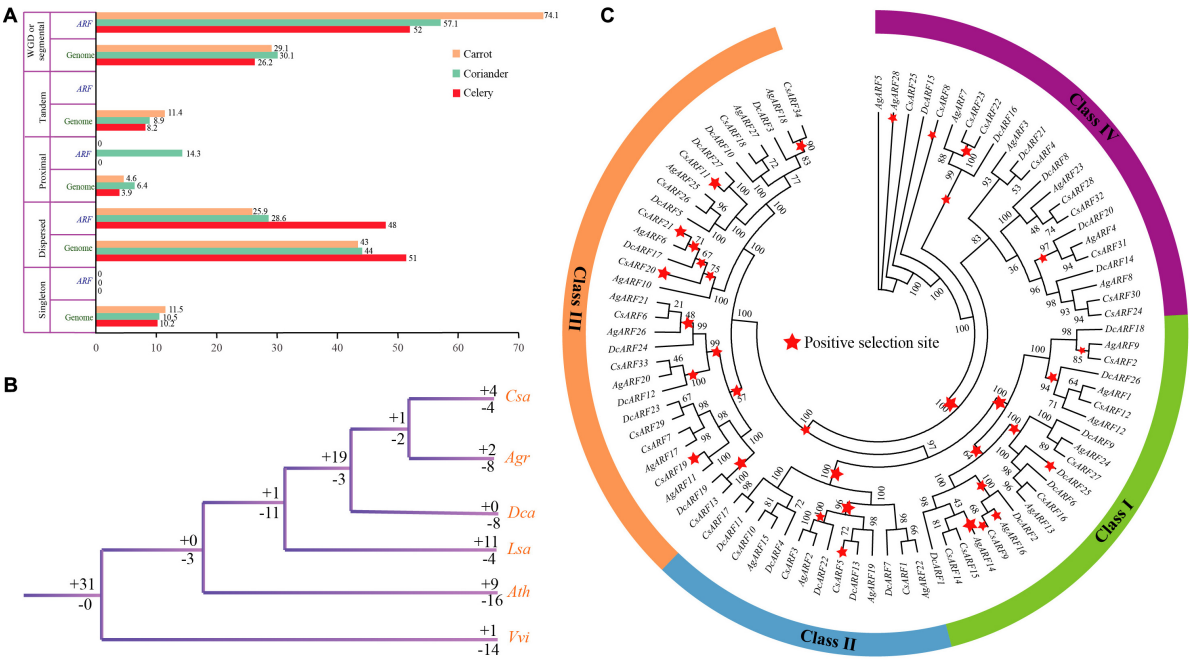


FIGURE 3 | Duplication, loss, and positive selection analysis of *ARF* genes in three Apiaceae species. **(A)** Percentage of different duplication types for *ARF* genes and all other genes. **(B)** Analysis of *ARF* gene duplication and loss. The " + " and "–" symbols indicate *ARF* gene duplication and loss, respectively, and the values after these symbols represent gene number. **(C)** Positive selection of *ARF* genes in celery, coriander, and carrot. Red stars represent positive selection branches.

ancestor of Apiaceae species compared with the ancestor of lettuce and Apiaceae.

Most Apiaceae ARF Genes Underwent Positive Selection

Strong positive selection was observed at major nodes of the phylogenetic tree, which may have contributed to the divergence of Apiaceae species. A total of 10, 4, 13, and 5 positive selection branches were detected in class I, II, III, and IV, respectively (**Figure 3C**). The number of positive selection sites was the highest in class III, indicating that *ARF* genes in class III were under stronger natural selection than those in other classes. Overall, we found that most branches underwent positive selection, which indicates that *ARF* genes played an important role in the evolution of Apiaceae.

Expression Analysis of ARF Genes

We analyzed the expression patterns of *ARF* genes in three different tissues of celery and coriander. In celery, 16, 9, and 2 *ARF* genes showed higher expression levels in the root, petiole, and leaf, respectively (**Supplementary Table 10** and **Supplementary Figure 5A**). The expression level (FPKM values) of three celery genes (*AgARF9*, *AgARF12*, *AgARF24*) was over 100 in the root. Among all *AgARF* genes, *AgARF9* showed the highest expression level in the petiole. However, *AgARF26* showed no expression in any of the three tissues, while *AgARF1* and *AgARF7* showed no expression in the leaf. Overall, genes in the same phylogenetic group or subgroup showed a similar expression patterns.

In coriander, 9, 22, and 1 *ARF* genes exhibited higher expression in the root, petiole, and leaf, respectively (**Supplementary Table 10**). Most coriander *ARF* genes were expressed to higher levels in the petiole than in the other two tissues (**Supplementary Figure 5B**). *CsARF27* showed the highest expression level in the root, while *CsARF12* showed the highest expression level in petiole and leaf. However, *CsARF22* and *CsARF23*, which clustered within the same phylogenetic group, were not expressed in any tissue.

To validate the RNA-seq data, the expression of six *ARF* genes was analyzed by qRT-PCR. The results were consistent with the transcriptome results (**Supplementary Figure 6**), indicating that our RNA-seq data were reliable.

ARF Gene Paralogs Exhibit Notably Different Expression Patterns

Next, we explored the expression patterns of paralogous genes in celery and coriander (**Figure 4** and **Supplementary Figure 7**). Although most paralogous genes showed similar expression patterns, there were several exceptions. For example, the expression level of *AgARF19* was notably higher than that of *AgARF2* in all three tissues (**Figure 4**). Similarly, the expression of *AgARF24* was notably higher than that of *AgARF13* in three tissues. In coriander, *CsARF5*, *CsARF7*, *CsARF10*, and *CsARF32* were expressed to higher levels than *CsARF3*, *CsARF13*, *CsARF17*, and *CsARF28*, respectively (**Supplementary Figure 7**). These results suggest that some paralogous *ARF* genes diverged during the evolution of Apiaceae.

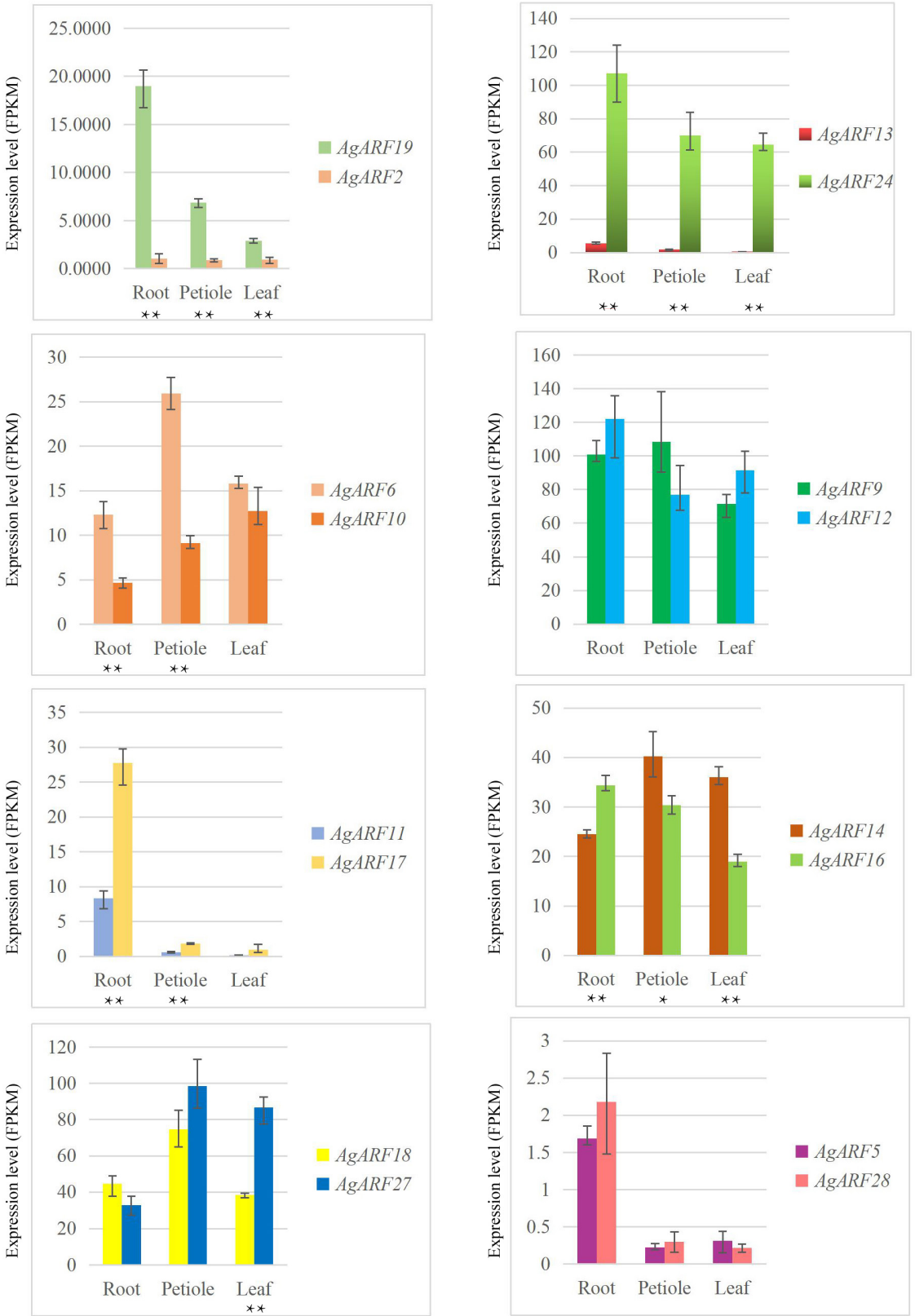


FIGURE 4 | Relative expression level (FPKM) of celery *ARF* paralogous gene pairs in the root, petiole, and leaf. Asterisks indicate significant differences (*$P < 0.05$, **$P < 0.01$).

DISCUSSION

Functions of ARF Genes in Apiaceae

ARF genes exist in most plant species and play key roles in plant growth and development (Li et al., 2016). Most ARF proteins contain three domains: a B3-type DNA-binding domain (DBD) at N-terminus, a variable middle region comprising an activation domain (AD) and a repression domain (RD), and a C-terminal dimerization domain (Tombuloglu, 2019). These domains facilitate interactions with Aux/IAA proteins (Mun et al., 2012; Li et al., 2015). Depending on these domains, ARFs act as transcriptional activators or repressors (Finet et al., 2013).

In Arabidopsis, the function of most *ARFs* has been studied. *ARF1* regulates flower development (Ellis et al., 2005); *ARF2* connects the ethylene and auxin signaling pathways to regulate hypocotyl bending (Okushima et al., 2005); *ARF7* and *ARF19* regulate lateral root formation (Okushima et al., 2007). The characterization of *ARF* gene function in Arabidopsis facilitates the functional analysis of *ARF* genes in Apiaceae. For example, the Apiaceae *ARF* genes in class II clustered together with *AtARF3* and *AtARF4*. Therefore, we speculate that most genes in class II might be related to the floral meristem and reproductive organs. Here, we used bioinformatics to predict the function on a large scale. However, further experimentation is needed to determine the *ARF* gene function in Apiaceae.

Here, we found that some *ARF* gene paralogs exhibit notably different expression patterns, indicating the functions these paralogs diverged during the evolution of Apiaceae. This result is consistent with previous reports on the functional divergence of paralogous genes (Wang et al., 2013a; Soria et al., 2014). In fact, gene duplicates undergo one of four evolutionary fates, including conservation, subfunctionalization, neofunctionalization, and specialization, according to previous reports (Blanc and Wolfe, 2004; Assis and Bachtrog, 2013; Wang et al., 2016). Therefore, this study lays a foundation for further studies on the function of paralogous *ARF* genes in Apiaceae.

ARFs Interact With Aux/IAA Proteins to Regulate Auxin-Responsive Gene Expression

Both ARF and Aux/IAA proteins act as transcription factors that regulate the expression of auxin-responsive genes (Kumar et al., 2015). *Aux/IAA* genes contain several highly conserved domains, and the structure of ARF and Aux/IAA proteins is similar at the C-terminal. Both ARFs and Aux/IAAs harbor the CTD, domain III, and domain IV. Auxin responses rely on ARF–Aux/IAA interactions, which are mediated by CTD (Wang et al., 2013b).

Aux/IAA proteins function as transcriptional repressors, while ARFs function as either activators or repressors for regulating auxin-responsive genes (Liscum and Reed, 2002). The function of ARF proteins is determined by variable middle region (Liscum and Reed, 2002). ARFs are released from Aux/IAAs to repress/activate the expression of auxin-responsive genes (Wu et al., 2017). Here, we identified all *ARF* genes present in three Apiaceae species. This will enable the analysis of interactions between ARF and Aux/IAA proteins in Apiaceae.

Systematic and Comprehensive Analysis of ARF Gene Family in Apiaceae Species

Although the *ARF* gene family was previously investigated in many plants, there was no report of it in Apiaceae. Whole-genome sequences of celery, coriander, and carrot were released recently (Iorizzo et al., 2016; Song et al., 2020b), which greatly facilitated this study.

To understand the evolution of *ARF* genes, we constructed a phylogenetic tree using ARF amino acid sequences of celery, coriander, carrot, Arabidopsis, lettuce, and grape. Gene structure and conserved motif analyses revealed that genes in the same group or subgroup showed similar features. The number of paralogous *ARF* gene pairs in celery (8) and coriander (12) was notably more than that in carrot (2). Moreover, based on collinearity analysis, we found that WGD had a significant impact on *ARF* gene family expansion in Apiaceae.

In conclusion, we conducted a comprehensive analysis of the *ARF* gene family in three Apiaceae species. Our results provide a strong foundation for comparative and functional analyses of the *ARF* gene family in plants.

DATA AVAILABILITY STATEMENT

The RNA-seq data were deposited in the Genome Sequence Archive (GSA) of the BIG Data Center (http://bigd.big.ac.cn/gsa) under the accession numbers CRA001996 and CRA001658.

AUTHOR CONTRIBUTIONS

XS conceived the project and was responsible for the project initiation. XS and QP supervised and managed the project and research. XS, QP, NL, QY, TW, SF, KG, and TY led the data collection and bioinformatics analyses. XS, QP, NL, XF, ZJ, RZ, and ZW organized, wrote, and revised the manuscript. All authors read and revised the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (31801856), the Hebei Province Higher Education Youth Talents Program (BJ2018016), the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No. 18KJB210010), and the Agricultural Science and Technology Innovation Project of Science and Technology Plan of Suzhou (No. SNG2017057).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.590535/full#supplementary-material

REFERENCES

Ahmad, T., Cawood, M., Iqbal, Q., Arino, A., Batool, A., Tariq, R. M. S., et al. (2019). Phytochemicals in *Daucus carota* and Their Health Benefits-Review Article. *Foods* 8:424. doi: 10.3390/foods8090424

Assis, R., and Bachtrog, D. (2013). Neofunctionalization of young duplicate genes in Drosophila. *Proc. Natl. Acad. Sci. U S A* 110, 17409–17414. doi: 10.1073/pnas.1313759110

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–W208.

Ballmer-Weber, B. K., Hoffmann, A., Wuthrich, B., Luttkopf, D., Pompei, C., Wangorsch, A., et al. (2002). Influence of food processing on the allergenicity of celery: DBPCFC with celery spice and cooked celery in patients with celery allergy. *Allergy* 57, 228–235. doi: 10.1034/j.1398-9995.2002.1o3319.x

Blanc, G., and Wolfe, K. H. (2004). Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell* 16, 1679–1691. doi: 10.1105/tpc.021410

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* 2020:009.

Ellis, C. M., Nagpal, P., Young, J. C., Hagen, G., Guilfoyle, T. J., and Reed, J. W. (2005). AUXIN RESPONSE FACTOR1 and AUXIN RESPONSE FACTOR2 regulate senescence and floral organ abscission in *Arabidopsis thaliana*. *Development* 132, 4563–4574. doi: 10.1242/dev.02012

Finet, C., Berne-Dedieu, A., Scutt, C. P., and Marletaz, F. (2013). Evolution of the ARF gene family in land plants: old domains, new tricks. *Mol. Biol. Evol.* 30, 45–56. doi: 10.1093/molbev/mss220

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

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

Iorizzo, M., Ellison, S., Senalik, D., Zeng, P., Satapoomin, P., Huang, J., et al. (2016). A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. *Nat. Genet.* 48, 657–666. doi: 10.1038/ng.3565

Jaillon, O., Aury, J. M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463–467. doi: 10.1038/nature06148

Kalluri, U. C., Difazio, S. P., Brunner, A. M., and Tuskan, G. A. (2007). Genome-wide analysis of Aux/IAA and ARF gene families in *Populus trichocarpa*. *BMC Plant Biol.* 7:59. doi: 10.1186/1471-2229-7-59

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

Kumar, R., Agarwal, P., Pareek, A., Tyagi, A. K., and Sharma, A. K. (2015). Genomic Survey. Gene Expression, and Interaction Analysis Suggest Diverse Roles of ARF and Aux/IAA Proteins in Solanaceae. *Plant Mole. Biol. Rep.* 33, 1552–1572. doi: 10.1007/s11105-015-0856-z

Kumar, R., Tyagi, A. K., and Sharma, A. K. (2011). Genome-wide analysis of auxin response factor (ARF) gene family from tomato and analysis of their role in flower and fruit development. *Mol. Genet. Genom.* 285, 245–260. doi: 10.1007/s00438-011-0602-7

Letunic, I., Doerks, T., and Bork, P. (2012). SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res.* 40, D302–D305.

Li, K. B. (2003). ClustalW-MPI: ClustalW analysis using distributed and parallel computing. *Bioinformatics* 19, 1585–1586. doi: 10.1093/bioinformatics/btg192

Li, L., Stoeckert, C. J. Jr., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189. doi: 10.1101/gr.1224503

Li, S. B., OuYang, W. Z., Hou, X. J., Xie, L. L., Hu, C. G., and Zhang, J. Z. (2015). Genome-wide identification, isolation and expression analysis of auxin response factor (ARF) gene family in sweet orange (*Citrus sinensis*). *Front. Plant Sci.* 6:119. doi: 10.3389/fpls.2015.00119

Li, S. B., Xie, Z. Z., Hu, C. G., and Zhang, J. Z. (2016). A Review of Auxin Response Factors (ARFs) in Plants. *Front. Plant Sci.* 7:47. doi: 10.3389/fpls.2016.00047

Liscum, E., and Reed, J. W. (2002). Genetics of Aux/IAA and ARF action in plant growth and development. *Plant Mol. Biol.* 49, 387–400. doi: 10.1007/978-94-010-0377-3_10

Luo, X. C., Sun, M. H., Xu, R. R., Shu, H. R., Wang, J. W., and Zhang, S. Z. (2014). Genomewide identification and expression analysis of the ARF gene family in apple. *J. Genet.* 93, 785–797. doi: 10.1007/s12041-014-0462-0

Maljaei, M. B., Moosavian, S. P., Mirmosayyeb, O., Rouhani, M. H., Namjoo, I., and Bahreini, A. (2019). Effect of Celery Extract on Thyroid Function; Is Herbal Therapy Safe in Obesity? *Int. J. Prev. Med.* 10:55. doi: 10.4103/ijpvm.ijpvm_209_17

Marchler-Bauer, A., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., Fong, J. H., et al. (2009). CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res.* 37, D205–D210.

Mun, J. H., Yu, H. J., Shin, J. Y., Oh, M., Hwang, H. J., and Chung, H. (2012). Auxin response factor gene family in *Brassica rapa*: genomic organization, divergence, expression, and evolution. *Mol. Genet. Genomics* 287, 765–784. doi: 10.1007/s00438-012-0718-4

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Okushima, Y., Fukaki, H., Onoda, M., Theologis, A., and Tasaka, M. (2007). ARF7 and ARF19 regulate lateral root formation via direct activation of LBD/ASL genes in Arabidopsis. *Plant Cell* 19, 118–130. doi: 10.1105/tpc.106.047761

Okushima, Y., Overvoorde, P. J., Arima, K., Alonso, J. M., Chan, A., Chang, C., et al. (2005). Functional genomic analysis of the AUXIN RESPONSE FACTOR gene family members in *Arabidopsis thaliana*: unique and overlapping functions of ARF7 and ARF19. *Plant Cell* 17, 444–463. doi: 10.1105/tpc.104.028316

Prachayasittikul, V., Prachayasittikul, S., Ruchirawat, S., and Prachayasittikul, V. (2018). Coriander (*Coriandrum sativum*): A promising functional food toward the well-being. *Food Res. Int.* 105, 305–323. doi: 10.1016/j.foodres.2017.11.019

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., et al. (2012). The Pfam protein families database. *Nucleic. Acids. Res.* 40, 290–301. doi: 10.1093/nar/gkr1065

Reyes-Chin-Wo, S., Wang, Z., Yang, X., Kozik, A., Arikit, S., Song, C., et al. (2017). Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. *Nat. Commun.* 8:14953.

Song, X. M., Liu, T. K., Duan, W. K., Ma, Q. H., Ren, J., Wang, Z., et al. (2014b). Genome-wide analysis of the GRAS gene family in Chinese cabbage (*Brassica rapa* ssp. *pekinensis*). *Genomics* 103, 135–146. doi: 10.1016/j.ygeno.2013.12.004

Song, X., Liu, G., Duan, W., Liu, T., Huang, Z., Ren, J., et al. (2014a). Genome-wide identification, classification and expression analysis of the heat shock transcription factor family in Chinese cabbage. *Mol. Genet. Genomics* 289, 541–551. doi: 10.1007/s00438-014-0833-5

Song, X., Liu, G., Huang, Z., Duan, W., Tan, H., Li, Y., et al. (2016). Temperature expression patterns of genes and their coexpression with LncRNAs revealed by RNA-Seq in non-heading Chinese cabbage. *BMC Genomics* 17:297. doi: 10.1186/s12864-016-2625-2

Song, X., Nie, F., Chen, W., Ma, X., Gong, K., Yang, Q., et al. (2020a). Coriander Genomics Database: a genomic, transcriptomic, and metabolic database for coriander. *Horticul. Res.* 7:55.

Song, X., Wang, J., Li, N., Yu, J., Meng, F., Wei, C., et al. (2020b). Deciphering the high-quality genome sequence of coriander that causes controversial feelings. *Plant Biotechnol. J.* 18, 1444–1456. doi: 10.1111/pbi.13310

Soria, P. S., McGary, K. L., and Rokas, A. (2014). Functional divergence for every paralog. *Mol. Biol. Evol.* 31, 984–992. doi: 10.1093/molbev/msu050

Stolzer, M., Lai, H., Xu, M., Sathaye, D., Vernot, B., and Durand, D. (2012). Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics* 28, 409–415.

Su, Z., Wang, L., Li, W., Zhao, L., Huang, X., Azam, S. M., et al. (2017). Genome-Wide Identification of Auxin Response Factor (ARF) Genes Family and its Tissue-Specific Prominent Expression in Pineapple (*Ananas comosus*). *Tropical Plant Biol.* 10, 86–96. doi: 10.1007/s12042-017-9187-6

Tang, Y., Bao, X., Liu, K., Wang, J., Zhang, J., Feng, Y., et al. (2018). Genome-wide identification and expression profiling of the auxin response factor (ARF) gene family in *physic nut*. *PLoS One* 13:e0201024. doi: 10.1371/journal.pone.0201024.t001

Tombuloglu, H. (2019). Genome-wide analysis of the auxin response factors (ARF) gene family in barley (*Hordeum vulgare* L.). *J. Plant Biochem. Biotechnol.* 28, 14–24. doi: 10.1007/s13562-018-0458-6

Ulmasov, T., Hagen, G., and Guilfoyle, T. J. (1997). ARF1, a transcription factor that binds to auxin response elements. *Science* 276, 1865–1868. doi: 10.1126/science.276.5320.1865

Voorrips, R. E. (2002). MapChart: software for the graphical presentation of linkage maps and QTLs. *J. Hered.* 93, 77–78. doi: 10.1093/jhered/93.1.77

Wan, S., Li, W., Zhu, Y., Liu, Z., Huang, W., and Zhan, J. (2014). Genome-wide identification, characterization and expression analysis of the auxin response factor gene family in *Vitis vinifera*. *Plant Cell Rep.* 33, 1365–1375. doi: 10.1007/s00299-014-1622-7

Wang, D., Pei, K., Fu, Y., Sun, Z., Li, S., Liu, H., et al. (2007). Genome-wide analysis of the auxin response factors (ARF) gene family in rice (*Oryza sativa*). *Gene* 394, 13–24. doi: 10.1016/j.gene.2007.01.006

Wang, D., Zhang, Y., Zhang, Z., Zhu, J., and Yu, J. (2010). KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genom. Proteomics Bioinform.* 8, 77–80. doi: 10.1016/s1672-0229(10)60008-3

Wang, J., Marowsky, N. C., and Fan, C. (2013a). Divergent Evolutionary and Expression Patterns between Lineage Specific New Duplicate Genes and Their Parental Paralogs in *Arabidopsis thaliana*. *PLoS One* 8:e72362. doi: 10.1371/journal.pone.0072362

Wang, J., Tao, F., Marowsky, N. C., and Fan, C. (2016). Evolutionary Fates and Dynamic Functionalization of Young Duplicate Genes in Arabidopsis Genomes. *Plant Physiol.* 172, 427–440. doi: 10.1104/pp.16.01177

Wang, S., Hagen, G., and Guilfoyle, T. J. (2013b). ARF-Aux/IAA interactions through domain III/IV are not strictly required for auxin-responsive gene expression. *Plant signal. Behav.* 8, e24526–e24526.

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40, e49. doi: 10.1093/nar/gkr1293

Wei, H. B., Cui, B. M., Ren, Y. L., Li, J. H., Liao, W. B., Xu, N. F., et al. (2006). Research Progresses on Auxin Response Factors. *J. Int. Plant Biol.* 48, 622–627. doi: 10.1111/j.1744-7909.2006.00280.x

Wu, W., Liu, Y., Wang, Y., Li, H., Liu, J., Tan, J., et al. (2017). Evolution Analysis of the Aux/IAA Gene Family in Plants Shows Dual Origins and Variable Nuclear Localization Signals. *Int. J. Mol. Sci.* 18:2107. doi: 10.3390/ijms18102107

Xing, H., Pudake, R. N., Guo, G., Xing, G., Hu, Z., Zhang, Y., et al. (2011). Genome-wide identification and expression profiling of auxin response factor (ARF) gene family in maize. *BMC Genomics* 12:178. doi: 10.1186/1471-2164-12-178

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Zhang, H., Cao, N., Dong, C., and Shang, Q. (2017). Genome-wide Identification and Expression of ARF Gene Family during Adventitious Root Development in Hot Pepper (*Capsicum annuum*). *Horticultural Plant J.* 3, 151–164. doi: 10.1016/j.hpj.2017.07.001

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.

Copyright © 2021 Pei, Li, Yang, Wu, Feng, Feng, Jing, Zhou, Gong, Yu, Wang and Song. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 25 January 2021
doi: 10.3389/fgene.2020.581993

Integration of Transcriptional and Post-transcriptional Analysis Revealed the Early Response Mechanism of Sugarcane to Cold Stress

Xing Huang[1,2†], Yongsheng Liang[3†], Baoqing Zhang[2], Xiupeng Song[2], Yangrui Li[1,2], Zhengqiang Qin[2], Dewei Li[2], Rongfa Chen[2], Zhongfeng Zhou[2], Yuchi Deng[2], Jiguang Wei[1] and Jianming Wu[2*]*

[1] College of Agriculture, Guangxi University, Nanning, China, [2] Key Laboratory of Sugarcane Biotechnology and Genetic Improvement (Guangxi), Guangxi Key Laboratory of Sugarcane Genetic Improvement, Sugarcane Research Institute, Guangxi Academy of Agricultural Sciences, Ministry of Agriculture, Nanning, China, [3] Nanning Institute of Agricultural Sciences, Nanning, China

Cold stress causes major losses to sugarcane production, yet the precise molecular mechanisms that cause losses due to cold stress are not well-understood. To survey miRNAs and genes involved in cold tolerance, RNA-seq, miRNA-seq, and integration analyses were performed on *Saccharum spontaneum*. Results showed that a total of 118,015 genes and 6,034 of these differentially expressed genes (DEGs) were screened. Protein–protein interaction (PPI) analyses revealed that ABA signaling via protein phosphatase 2Cs was the most important signal transduction pathway and late embryogenesis abundant protein was the hub protein associated with adaptation to cold stress. Furthermore, a total of 856 miRNAs were identified in this study and 109 of them were differentially expressed in sugarcane responding to cold stress. Most importantly, the miRNA–gene regulatory networks suggested the complex post-transcriptional regulation in sugarcane under cold stress, including 10 miRNAs−42 genes, 16 miRNAs−70 genes, and three miRNAs−18 genes in CT vs. LT0.5, CT vs. LT1, and CT0.5 vs. LT1, respectively. Specifically, key regulators from 16 genes encoding laccase were targeted by novel-Chr4C_47059 and Novel-Chr4A_40498, while five *LRR-RLK* genes were targeted by Novel-Chr6B_65233 and Novel-Chr5D_60023, 19 PPR repeat proteins by Novel-Chr5C_57213 and Novel-Chr5D_58065. Our findings suggested that these miRNAs and cell wall-related genes played vital regulatory roles in the responses of sugarcane to cold stress. Overall, the results of this study provide insights into the transcriptional and post-transcriptional regulatory network underlying the responses of sugarcane to cold stress.

Keywords: miRNAs, integrative analysis, cold stress, sugarcane, miRNA targeted genes

OPEN ACCESS

Edited by:
Rong Zhou,
Aarhus University, Denmark

Reviewed by:
Kranthi Varala,
Purdue University, United States
Himanshu Sharma,
National Agri-Food Biotechnology Institute, India
Paula Barbim Donate,
University of São Paulo, Brazil

***Correspondence:**
Jiguang Wei
jiguangwei@gxu.edu.cn
Jianming Wu
wujianming2004@126.com

[†]These authors have contributed equally to this work

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

Received: *10 July 2020*
Accepted: *11 December 2020*
Published: *25 January 2021*

Citation:
Huang X, Liang Y, Zhang B, Song X, Li Y, Qin Z, Li D, Chen R, Zhou Z, Deng Y, Wei J and Wu J (2021) Integration of Transcriptional and Post-transcriptional Analysis Revealed the Early Response Mechanism of Sugarcane to Cold Stress. Front. Genet. 11:581993. doi: 10.3389/fgene.2020.581993

INTRODUCTION

Sugarcane planting, yield, and quality are affected by abiotic factors including low-temperature stress. Previous research has shown that the response of sugarcane to cold stress is regulated by sophisticated mechanisms at the morphological, anatomical, physiological, and biochemical levels. Once subjected to cold stress, sugarcane leaves dry up, growth points and lateral buds become frostbitten, and the tilling rate and number of effective perennial stems are affected, especially at the seedling stage. Cold stress markedly inhibits sugarcane seedling growth and survival rates fall (Rasheed et al., 2010; Pang et al., 2015). The thickness of sugarcane leaves, the area of vesicular cells, the thickness of thick-walled vascular bundle tissue, and electrolyte extravasation rates are all known to be positively correlated with cold resistance (Zhu et al., 2019). Indeed, at the seedling stage, the degree of peroxidation of the leaf lipid membrane under cold stress is enhanced and occurs in concert with membrane permeability and relative conductivity. The content of soluble sugar, proline, and 3,4-methylenedioxyamphetamine (MDA) also increases (Ding et al., 2006; Wang et al., 2016a), while chlorophyll content decreases with the decrease in temperature and increased duration of cold temperature treatment. Cold stress also inhibits the activity of ribulose-1,5-diphosphate carboxylase (Rubisco) and phosphoenolpyruvate carboxylase (PEP), while SOD (superoxide dismutase) and POD (peroxidase) increase. Traditional genetic and molecular analyses have identified that COLD1-CRLK1/2-MEKK-MPK-ICE-CBF-CORs and CRPK1-14-3-3-CBF are important cold signal sensing pathways in rice and *Arabidopsis* (Guo et al., 2018; Shi et al., 2018). Recent RNA-seq research has revealed the presence of regulated mechanisms that are more complex in other plants (Li et al., 2016). Almost 50% of paper mulberry transcriptional factors are involved in cold stress responses (Peng et al., 2015). In addition, 1,841 differentially expressed genes (DEGs) are involved in cold stress response in maize (Li et al., 2017). A total of 47 genes are upregulated in wheat, whereas just six are downregulated in the ICE1-CBF-COR pathway as a component of cold stress response. Ten genes in the ABA-dependent pathway are upregulated, while four are downregulated across these four genotypes, confirming that ABA accumulates when plants are exposed to low temperatures (Li et al., 2018).

The regulation of gene expression via microRNA (miRNA) is a negative mechanism that occurs in response to cold stress in plants at post-transcriptional levels. In plants, miRNAs are divided into two groups, conserved, and species-specific (Esposito et al., 2020). In recent years, numerous cold-responsive miRNAs have been identified in different plant species, and some of them are involved in the CBF-dependent pathway, triggering the elimination of ROS, modifying cellular antioxidant capacity, and in the auxin signaling pathway (Megha et al., 2018). miR319 is downregulated by cold stress in rice while the overexpression of miR319 or target RNA interference leads to the upregulation of cold-responsive genes and improved cold tolerance (Thiebaut et al., 2012; Yang et al., 2013; Wang et al., 2014). miR160 and miR167 have been identified as temperature-responsive miRNAs in cotton. The miR160 family also exhibits lower expression levels at 12°C compared with the control at 25°C (Wang et al., 2016b). One hundred and seventy-two miRNAs belonging to 39 families and 126 unique examples were identified in *Nelumbo nucifera* (Zou et al., 2020). 172 miRNA–target pairs have been identified as being involved in plant temperature responses by integrating miRNA with mRNA expression profiles. Three conserved and 25 predicted miRNAs have also been found to be involved in response to cold stress in *Brachypodium* (Zhang et al., 2009).

Prior to the sequencing of the complete sugarcane genome, expressed sequence tags (ESTs) and RNA-seq were both used to determine the mechanisms involved in cold stress response. In one example, 34 cold-inducible ESTs were identified in 2003 from 1,536 ESTs in sugarcane (*Saccharum* sp. cv SP80-3280) exposed to cold stress (Nogueira et al., 2003). Transcriptome profiles from the low temperature-tolerant *S. spontaneum* clone IND 00-1037 are also analyzed, as well as 2,583 genes that are upregulated and 3,302 that are downregulated in a low temperature stress treatment (Dharshini et al., 2016). A total of 170 cold-responsive TFs in 30 families are also differentially regulated (Selvarajan et al., 2018). Transcriptome profiles of *S. spontaneum* under cold stress indicate that certain genes involved in transmembrane transporter activities are absent from cold-susceptible sugarcane (Jong-Won et al., 2015). Additional work on small RNA transcriptomes from sugarcane has also been possible via deep sequencing (Sternes and Moyle, 2015). However, analysis of miRNA–genes under cold stress has not been undertaken since publication of the complete sugarcane genome (Zhang et al., 2018).

The wild relative of sugarcane, *Saccharum spontaneum* L., is known for its adaptability to environmental stressors, particularly cold stress conditions. Thus, to gain a comprehensive understanding of the molecular mechanisms involved in the response of sugarcane seedlings to cold stress, we studied the expression profiles and relevant genes in *S. spontaneum* GX87-16 seedlings treated with cold stress (4°C). In addition, we integrated a DEG and differentially expressed miRNA analysis and constructed a regulation network of miRNA–mRNA. The findings from this study provide insight into the transcriptional and post-transcriptional regulation of cold tolerance and low temperature stress in sugarcane.

Abbreviations: MEKK, mitogen-activated protein/ERK kinase kinases; MPK, MAPK phosphatase; CBF, C-repeat cis-element binding factor; COR, cold-regulated; CRPK, calcium-regulated protein kinase; GSH, glutathione; ROS, reactive oxygen species; ERF, ethylene response factor; ARF, auxin response factors; ABA, abscisic acid; SOD, superoxide dismutase; MDA, malondialdehyde; CT, control; LT, low temperature; TF, transcriptional factor; ESTs, expressed sequence tags; FPKM, number of fragments per kilobase of transcript sequence per million base pairs sequenced; COG, clusters of orthologous groups of proteins; PP2C, protein phosphatase 2C; KOG, eukaryotic ortholog groups; KEGG, Kyoto Encyclopedia of Genes and Genomes; GO, gene ontology; DEG, differentially expressed gene; LRR-RLK, leucine rich repeats-receptor-like protein kinase; LEA, late embryogenesis abundant protein; TPM, transcripts per million.

METHODS

Plant Materials and Cold Stress Treatment

Stems of *S. spontaneum* GX87-16 were planted in pots and grown in an artificial incubator at 28°C under a 16-h/8-h (light/

dark) photoperiod. After 2 weeks of growth, 2 h after entering the light culture, seedlings from six pots were transferred to an artificial incubator, which had been set to 4°C for cold stress treatment. Mature leaves were collected from treatment plants at 0.5 h and 1 h after cold treatment and denoted as LT0.5 and LT1. Meanwhile, another three potted plants were sampled at the time when the six pots were transferred to a 4°C artificial incubator; these three repeats were considered as controls (CT), namely, the 0 h of cold stress. Collected leaves were frozen in liquid nitrogen and stored at −80°C prior to RNA extraction.

Construction and Sequencing of mRNA-Seq and Small RNA Libraries

Total RNA was extracted using the Trizol reagent (Invitrogen), and RNA integrity was assessed using an RNA Nano 6000 Assay Kit in an Agilent Bioanalyzer 2100 system (Agilent Technologies, CA, United States). RNA concentrations were measured using a Qubit® RNA Assay Kit in a Qubit® 2.0 Fluorometer (Life Technologies, CA, United States). Equal amounts of RNA from all samples were used for the construction of mRNA-Seq and small RNA libraries.

A series of mRNA-Seq libraries were prepared using an Illumina TruSeq RNA Sample PrepKit following the manufacturer's protocols. Sequencing was performed via paired-end reads (2 × 151 base pairs, bp) using an Illumina X Ten sequencing platform. Bands of RNA between 18 and 40 nt in length were isolated for small RNA sequencing. Libraries were prepared following the Small RNA Sample Preparation Protocol (Illumina) and sequenced on an Illumina Hiseq-2500 sequencing platform using SE50.

Processing of mRNA-Seq Data

To preprocess miRNA-Seq data, adaptor sequences were removed alongside low-quality (<Q20) bases at 5′ and 3′ ends using the software Trimmomatic (v0.30) (Bolger et al., 2014). Reads longer than 70 bp were then used for further experiments. These reads were mapped onto the sugarcane genome (Zhang et al., 2018) using the software Bowtie 2 (2.1.0) (Langmead and Salzberg, 2012) with default parameters set after the preprocessing of mRNA-Seq data. The software TopHat2 was then used to perform a sequence comparison between clean reads and the reference genome to obtain positional information for the reference genome or gene and sequence information unique to the sample. Gene expression levels were presented as fragments per kilo base per million reads (FPKM); thus, genes with FPKM values >1 were retained for further analysis.

A false discovery rate (FDR) <0.05 was used to detect DEGs. After processing with the software DEseq, we applied a more than 2-fold change as the criterion to classify DEGs between two pairs (www.bioconductor.org). Resultant *P*-values were then adjusted using Benjamini and Hochberg's approaches for controlling the FDR.

Functional annotations were determined using the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses. Default parameters in the software AmiGO were used to obtain GO terms for each gene as well as analyze functional enrichment via hypergeometric tests with FDR correction. This procedure enabled us to obtain an adjusted $p < 0.01$ between test gene groups and the whole annotation dataset. DEGs in the KEGG pathway were analyzed using the software Cytoscape (Ideker, 2011) using the ClueGO plugin (Wouter et al., 2014).

Processing miRNA-Seq Data

In this step, clean reads were screened from raw data by removing adaptors, poly A sequences, and low-quality bases at both ends of small RNA reads. Reads were then trimmed and cleaned by removing sequences smaller than 18 nt or longer than 40 nt. Values for the sequence duplication level, GC content, Q20, and Q30 in clean data were then calculated. All downstream analyses were based on high quality clean data.

To identify known and novel miRNAs, high-quality clean reads were mapped to the *S. spontaneum* genome using the software Bowtie 2 (2.1.0) (Langmead and Salzberg, 2012) and searched against NCBI, Rfam, and Repbase databases to remove known classes of RNAs (i.e., mRNA, rRNA, tRNA, snRNA, and snoRNA) and repeats. The remaining reads were then used to detect both known and novel (potentially species-specific) miRNAs among the 5,940 mature miRNAs in plants using mirBase Release 19 (Ana and Sam, 2011).

We performed a differential expression analysis under two conditions using the DESeq R package (1.10.1). This approach provides a statistical analysis for determining differential expression in digital miRNA expression data using a model based on the negative binomial distribution. Thus, miRNAs with an adjusted $p < 0.01$ were assigned as differentially expressed.

Predicting Potential miRNA Targets: A Joint Analysis of miRNA and mRNA

The psRNATarget tool (http://plantgrn.noble.org/psRNATarget/) and software TargetFinder were used to predict *S. spontaneum* mRNA targets of putative miRNAs. The software packages BLAST and KOBAS2.0 were then used to compare and carry out a functional annotation of discovered target genes vs. the NR, Swiss-Prot, GO, COG, KOG, Pfam, and KEGG databases. Finally, miRNA and mRNA were related to one another using their interaction in the STRING protein database (http://string-db.org/).

All clean sequencing data used in this study are available in the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra/) under accession numbers PRJNA636260 for transcriptomics and PRJNA635765 for miRNA-seq data.

Validation of mRNA and miRNA Relative Expression by QPCR

The RNA used for RNA-seq was used for the QPCR validation of the mRNA expression level. The first-strand cDNA for each sample was prepared from 500 ng of total RNA using SuperScript II reverse transcriptase (Takara, Dalian, China) following the manufacturer's instructions, and template cDNAs were diluted 10-fold for QPCR. Primers were designed using the Primer 5 program and are listed in **Supplementary Table 1**. Samples and standards were run in triplicate on each plate using the SYBR Premix Ex Taq™ II kit (Takara, Dalian, China) on a

StepOne™ real-time PCR system (Applied Biosystems, Foster City, USA) following the manufacturer's recommendations. QPCR was performed in a 20-μL reaction volume containing 6.8 μL of ddH_2O, 10 μL of SYBR Premix Ex Taq II, 0.4 μL of ROX Reference Dye II, 0.4 μL of forward primer (10 μmol/L), 0.4 μL of reverse primer (10 μmol/L), and 2 μL of template cDNA. The PCR programs were run as follows: 30 s of pre-denaturation at 95°C, 40 cycles of 5 s at 95°C, and 30 s at 60°C, followed by steps for dissociation curve generation. Dissociation curves for each amplicon were carefully examined to confirm the lack of multiple amplicons at different melting temperatures. Relative transcript levels for each sample were obtained using the comparative cycle threshold method using the cycle threshold value of transcript Sspon.005D0014601 encoding a glyceraldehyde-3-phosphate dehydrogenase was used as a control. The $2^{-\Delta\Delta Ct}$ method was used to calculate the data.

The RNA used for miRNA-Seq was used for the QPCR validation of miRNA expression. No more than 500 ng of each miRNA sample was used for miRNA first-strand preparation according to Stem-loop miRNA cDNA Synthesis kit (EnzyValley, Lot: T435A). Stem-loop primers were listed in **Supplementary Table 1**. QPCR was performed (Stem-loop miRNA qPCR SYBR kit, Lot: T436A) in a 20-μL reaction volume containing 3.5 μL of ddH_2O, 10 μL of 2× Stem-loop miRNA qPCR SYBR ProMix, 0.5 μL of ROX Reference Dye II, 0.5 μL of forward primer (10 μmol/L), 0.5 μL of Uni-SL reverse primer (10 μmol/L), and 5 μL of template cDNA. The PCR programs were run as follows: 10 min of pre-denaturation at 95°C, 40 cycles of 5 s at 95°C, and 30 s at 60°C, followed by steps for dissociation curve generation. The $2^{-\Delta\Delta Ct}$ method was used to calculate the data. The U6 snRNA was taken as the control.

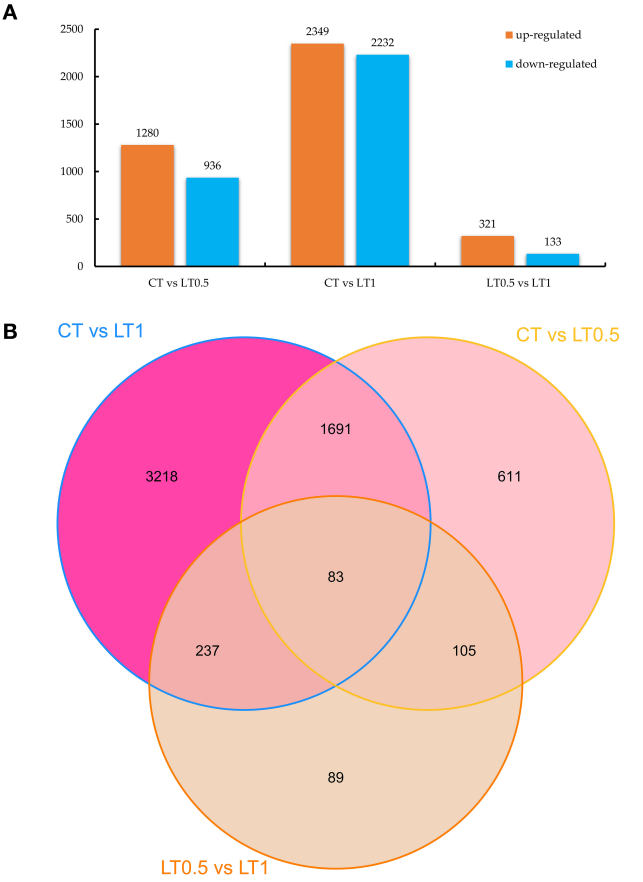


FIGURE 1 | DEG statistics for comparisons. **(A)** Number of downregulated and upregulated DEGs in CT vs. LT0.5, CT vs. LT1, and LT0.5 vs. LT1, respectively. **(B)** Venn diagram of DEGs distributed in every sample.

RESULTS

Gene Identification and Their Functional Annotation Analysis in This Study

A total of nine RNA-seq libraries from low temperature (LT)-treated and control (CT) seedlings were constructed for high-throughput sequencing. After the removal of low-quality reads and adapter sequences, 346,617,812 clean paired-end reads were obtained from raw data, with between 33,119,4345 and 44,605,663 obtained from each sample (**Supplementary Table 2**). Nearly 80% of these paired-end reads could be mapped to the reference genome, and 118,015 potential genes were identified with FPKM values between 0 and 89,290.4 (Sspon.007A0027210) (**Supplementary Table 3**). In total, 34,052 (28.85%), 56,311 (47.71%), 78,517 (66.53%), 80,524 (68.23%), 87,333 (74.00%), 101,433 (85.95%), and 108,262 (91.73%) genes were annotated in the COG, KOG, Swissprot, Pfam, GO, TrEMBL, and nr databases, respectively. Approximately 10.07% (11,887) of genes were annotated in the KEGG database (**Supplementary Tables 3**, **4**, **Supplementary Figure 1**), with most genes significantly matched with *Sorghum bicolor* (70.32%).

Identifying Cold-Responsive Genes in Sugarcane

A total of 6,034 DEGs (10.19% of total genes) were detected via DEGseq analysis (p-value <0.005 and $|\log_2$ (fold change)| > 1) under cold stress conditions. Compared with CT, the LT0.5 and LT1 treatments had 2,216 and 4,581 DEGs, respectively, including 1,280 and 936; 2,349 and 2,232 upregulated and downregulated DEGs for each compared pair. When LT1 was compared with LT0.5, 454 genes, including 321 upregulated and 133 downregulated DEGs, were identified (**Figure 1A**, **Supplementary Tables 5**–**7**). The resultant Venn diagram suggests that 83 core DEGs were shared across all compared pairs, and the largest number of unique DEGs was 3,218 in the CT vs. LT1 comparison (**Figure 1B**). The heat map for all samples indicates that biological repeats reached their expected effect and that four dominant expression profile clades were found via this dynamic hierarchical clustering approach (**Figure 2**).

Functional annotations of all DEGs for each compared pair were presented in **Supplementary Tables 5**–**8** and **Supplementary Figure 2**. Comparisons between CT and

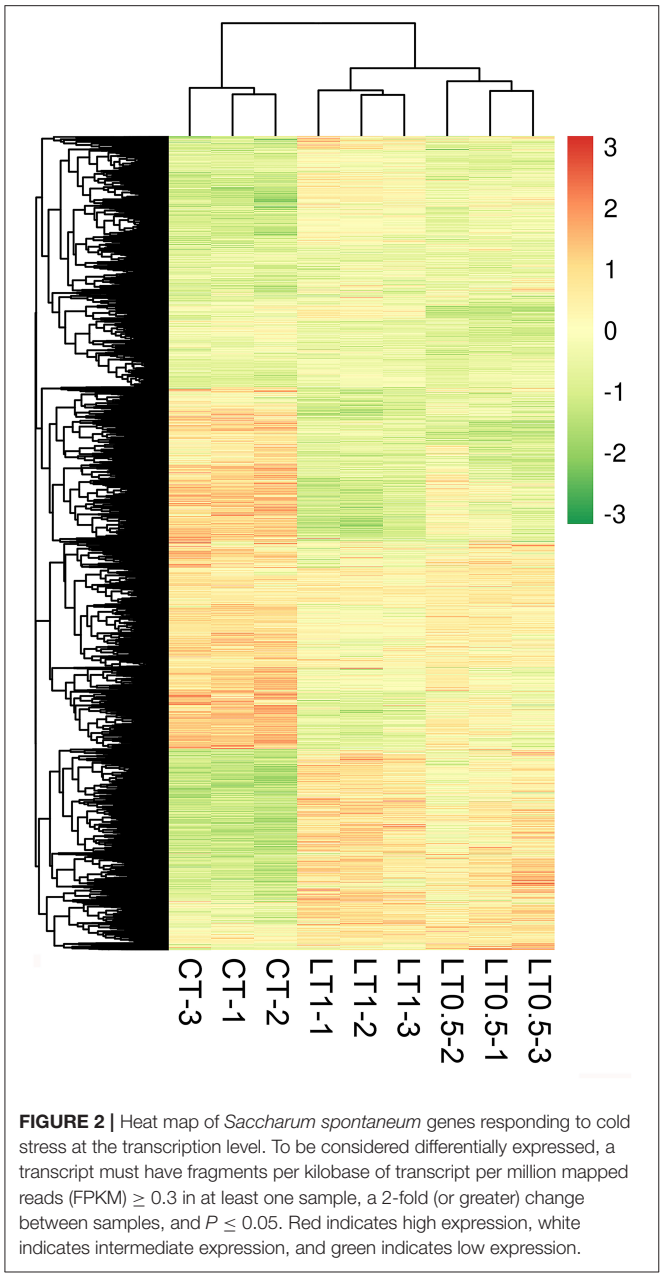


FIGURE 2 | Heat map of *Saccharum spontaneum* genes responding to cold stress at the transcription level. To be considered differentially expressed, a transcript must have fragments per kilobase of transcript per million mapped reads (FPKM) $\geq$ 0.3 in at least one sample, a 2-fold (or greater) change between samples, and $P \leq 0.05$. Red indicates high expression, white indicates intermediate expression, and green indicates low expression.

LT0.5 shows that 1,744 DEGs can be assigned to GO terms; of these, the term "regulation of defense response" was the most highly enriched among both upregulated and downregulated DEGs (**Figure 3A**). In the 1-h cold stress treatment, 3,645 DEGs could be GO annotated and the chloroplast was most affected.

The resultant KEGG pathway classification (**Supplementary Table 8** and **Figure 3B**) reveals that most significantly changed transcripts belong to metabolic pathways (especially those for starch, sucrose, pyruvate, and carbon fixation), signal transduction (plant hormones), environmental adaptation (plant–pathogen interaction), and the biosynthesis of secondary metabolites (mainly related to phenylpropanoid biosynthesis).

Protein–Protein Interaction Network Analysis in the mRNA Level Under Cold Stress

We constructed an interaction network based on the DEGs retrieved from known PPI databases. In terms of comparisons between CT and LT0.5, 1,186 DEGs can be linked via 1,100 known relationships, including activation, binding, and inhibition (**Supplementary Figure 3A**). The gene Sspon.004C0004740 can be identified as the core hub encoding an ATP synthase delta subunit, which was downregulated in the LT0.5 treatment. Sspon.006B0008980 (peptidyl-prolyl cis-trans isomerase CYP38), Sspon.008C0004432 (glycosyl hydrolases), Sspon.001D0013480 (fibrillarin), and Sspon.005A0004900 (ribosomal protein L23) are also important candidate DEGs that function in response to a half an hour cold stress.

After a 1-h cold stress, a total of 2,507 DEGs could be linked via 6,230 known relationships (**Figure 4**). Significantly, three DEGs (i.e., Sspon.004B0016440, Sspon.004B0016450, and Sspon.004B0015590) encoding the protein phosphatase 2C and the cyclic nucleotide-binding/kinase domain-containing protein (PP2C family) can all be proposed as potential hubs. Similarly, Sspon.003B0011973 (ribosomal protein L13) and Sspon.006B0008980 (peptidyl-prolyl cis-trans isomerase CYP38) also appear to be of importance to the PP2C protein family.

In terms of comparisons between CT0.5 and LT1, 230 DEGs could be linked via 1,600 known relationships (**Supplementary Figure 3B**). Four TIFY family proteins, including Sspon.001B0025400, Sspon.001D0021420, Sspon.001B0025420, and Sspon.001C0021660, were calculated as the core regulated protein that interacted with other genes through 13 linkages. TIFY is a plant-specific family involved in the regulation of plant-specific biologic processes, stress, and responses to phytohormones.

Conserved and Novel miRNAs Identified in Sugarcane in This Study

To identify the miRNAs that respond to cold stress, we constructed nine structural RNA (sRNA) libraries from nine samples. These libraries yielded 110,555,326 clean reads by high-throughput sequencing, which was ~11.05 million clean reads per library (**Supplementary Table 9**). Mapped sRNAs were then annotated and classified into sRNA, known miRNAs, precursor RNAs, intergenic RNAs, and unknown sRNAs. To verify these results, we analyzed Pearson correlation coefficients of the miRNA expression between biological duplicates (**Supplementary Figure 4**); all the Pearson values for miRNA expression were >0.82.

We then obtained 20 conserved and 836 novel miRNAs belonging to 13 families from nine libraries (**Supplementary Tables 10**, **11**). Expression profiles of identified miRNAs reveal great variation in TPM values, ranging from 83,065.85 to <10. Four miRNAs (Novel-Chr4B_41904, Novel-Chr4C_44656, Novel-Chr4C_44658, and Novel-tig00009572_93271) belong to the miR396 family and were highly expressed in all samples. Almost of all the identified miRNAs (835 of 856) were shared by all samples, indicating that

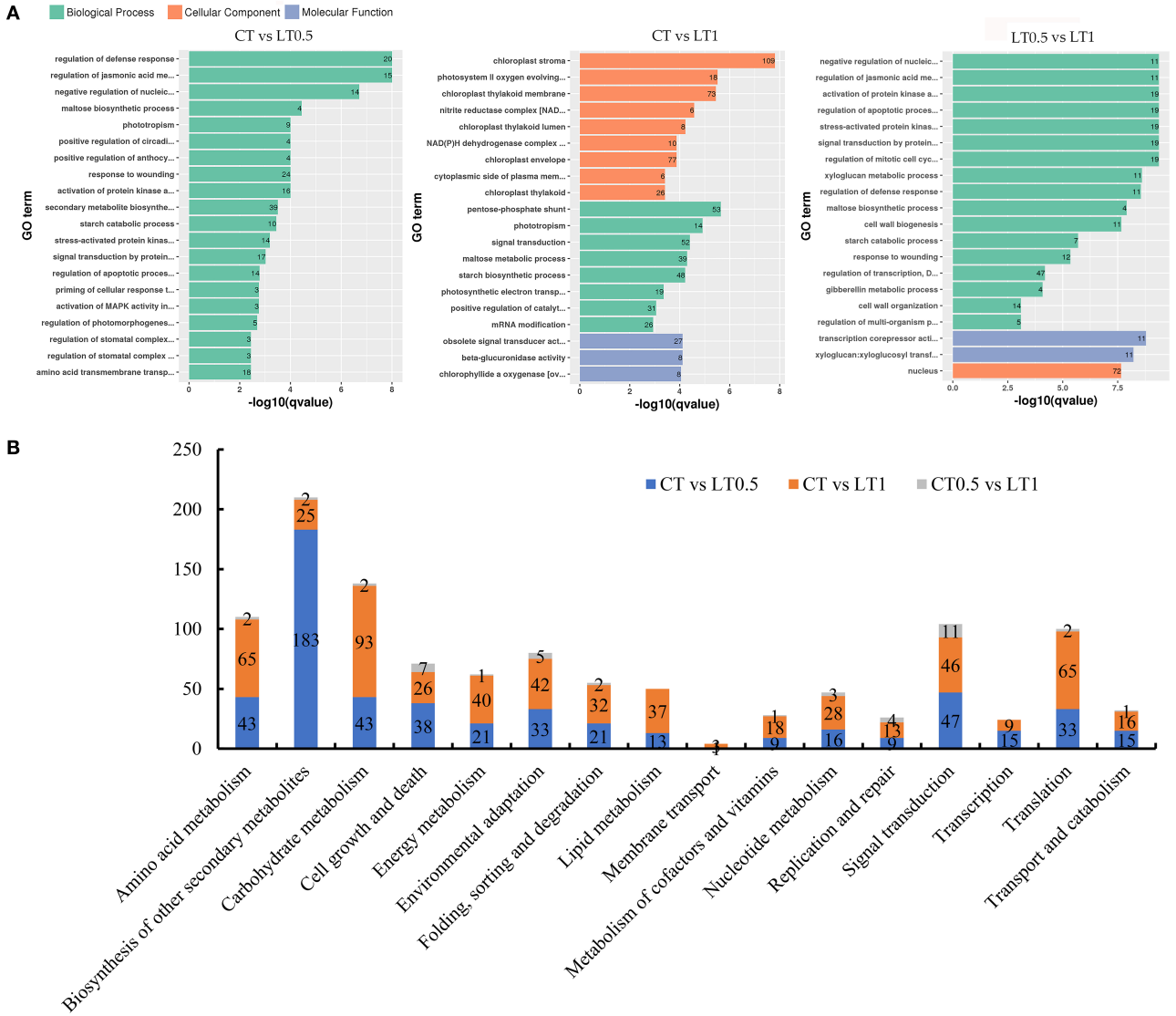


FIGURE 3 | Annotation and functional classification of DEGs. **(A)** Histogram of most enriched Gene Ontology (GO) classifications of *S. spontaneum* DEGs under cold stress. The ordinate is the enriched GO term; the abscissa is the enriched *q* value of the term and the number of differential genes on the column. Different colors are used to distinguish biological processes, cellular components, and molecular functions. **(B)** Statistical analysis of DEGs mapped to KEGG pathways including carbohydrates, energy, and lipid metabolisms.

these were stable in *S. spontaneum*, even under cold stress. sRNAs range from 18 nucleotides (nt) to 30 nt long, while the majority of mature miRNAs are 21 nt in plants. The distribution of the length of miRNA in *S. spontaneum* showed that 399 miRNAs were 21 nt long and 277 were 24 nt long (**Figure 5A**). All 16 known miRNAs were 21 nt long, and the first nucleotide was always U, which is different from the length of novel miRNAs, which were 24 nt long (**Supplementary Figure 5**).

Prediction of Targeted mRNAs in Sugarcane

A total of 496 genes were targeted by 423 miRNAs (**Supplementary Tables 12**, **13**). A total of 15 known and 290 novel miRNAs corresponded to 65 and 318 targeted genes, respectively. The miRNA Novel-Chr5B_53523 has the largest number of targeted genes 27. A total of 453 targeted genes (accounting for 91.33%) were annotated in KEGG, GO, and other databases (**Supplementary Tables 14–16**). In the GO annotation, regulation of transcription (35), oxidation–reduction process (34), lignin catabolic process (14), auxin-activated signaling pathway (12), lignin biosynthesis (11), methylation (11), and cell differentiation processes (10) were enriched in biological processes. The nucleus (60), integral components of membrane (46), apoplast (15), cytoplasm (9), and plasma membrane (8) were enriched in the cellular component, and DNA binding (37), ATP binding (24), copper ion binding (14),

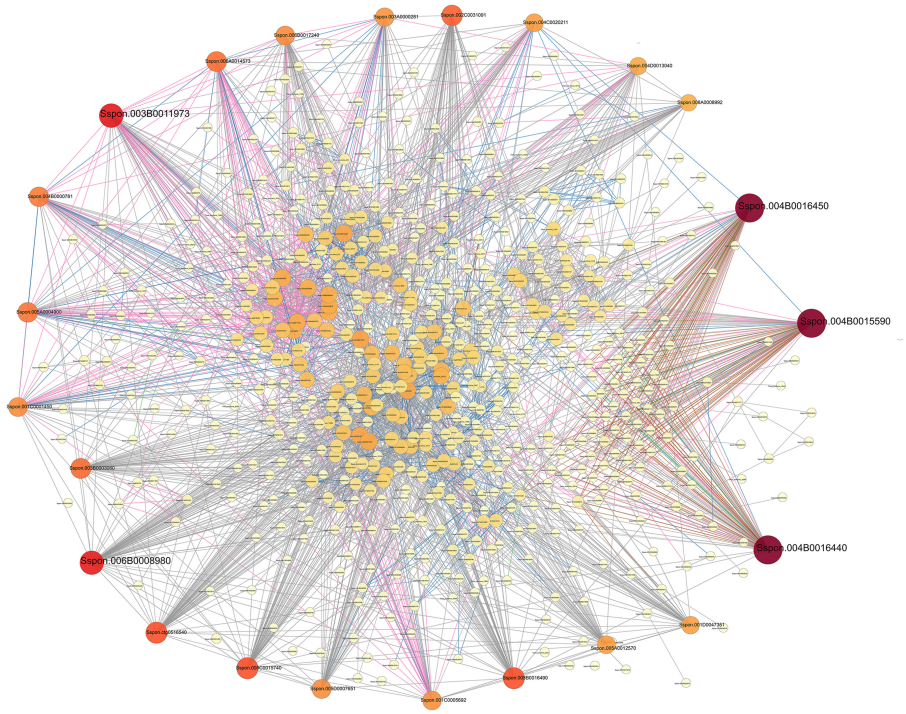


FIGURE 4 | A PPI (protein–protein interaction) network of DEGs from CT vs. LT1. The interaction relationship in the STRING protein interaction database (http://string-db.org/) was then used to analyze the DEG coding protein interaction network. The size of a node in this interaction network is proportional to the degree of this node; thus, the more edges connected to a node, the greater its degree and the larger the node, and the darker the color. Nodes may be in a more central position in the network, and lines of different colors represent varied interactions. The thicker the line, the higher the score value.

oxygen oxidoreductase activity (14), oxidoreductase activity and oxidizing metal ions (14), iron ion binding (13), heme binding (12), lipid binding (11), DNA-binding transcription factor activity (10), O-methyltransferase activity (10), protein dimerization activity (10), and zinc ion binding (10) were the top enriched go terms for molecular function. A total of 251 genes were mapped to 132 pathways within the KEGG database; plant hormone signal transduction (29), stilbenoid diarylheptanoid and gingerol biosynthesis (12), and plant–pathogen interactions (11) were the top three annotated pathways (**Supplementary Table 16**).

Differentially Expressed miRNAs and Their Targets Under Cold Stress

A total 109 differentially expressed miRNAs exhibiting a more than 2-fold change were identified in *S. spontaneum* after cold stress, including four known and 105 novel miRNAs, respectively (**Supplementary Tables 17–19**). Compared with the control, the numbers of differentially expressed miRNAs after cold stress treatment for 0.5 h and 1 h were 53 (38 upregulated and 15 downregulated) and 94 (46 upregulated and 48 downregulated). When LT1 was compared with LT0.5, 28 downregulated miRNAs were identified (**Figure 5B**). The Venn diagram from this analysis suggests the presence of 13 core miRNAs shared by all compared pairs, including 42 unique miRNAs to CT vs. LT1, the largest number (**Figure 5C**).

TargetFinder was used to predict candidate targets of differentially expressed miRNAs (**Supplementary Tables 20–22**). In a comparison of CT and LT0.5, 10 miRNAs targeting 42 genes were identified; of these genes, 14 were targeted by miRNAs Novel-tig00084004_92871, referring to signal transduction, transcription, and energy production, such as MYB and LRR-RLK (leucine-rich repeat receptor-like protein kinase). Besides, there were three LRR-RLK genes (Sspon.006A0015270, Sspon.006B0014350, Sspon.003D0010720, and Sspon.005B0004571) were targeted by miRNA Novel-Chr6B_65233, indicating an important function in the regulation of cold stress signal transduction. Furthermore, a total of four laccase genes (Sspon.003D0006520, Sspon.003B0007420, Sspon.003C0020690, and Sspon.003C0010170) were targeted by Novel-Chr4A_40498. These 18 miRNAs were upregulated by cold stress.

A total of 16 miRNAs targeting 70 genes were identified from the comparison of LT1 and CT. The targeted genes include 14 PPR repeat family members (all targeted by Novel-Chr5C_57213), five LRR-RLKs, six scarecrow-like proteins, and four laccases.

Three miRNAs targeting 18 genes were identified in the comparison between CT0.5 and LT1. Among these, total of 11 genes encoding laccase were targeted by Novel-Chr4C_47059 which was downregulated by cold stress.

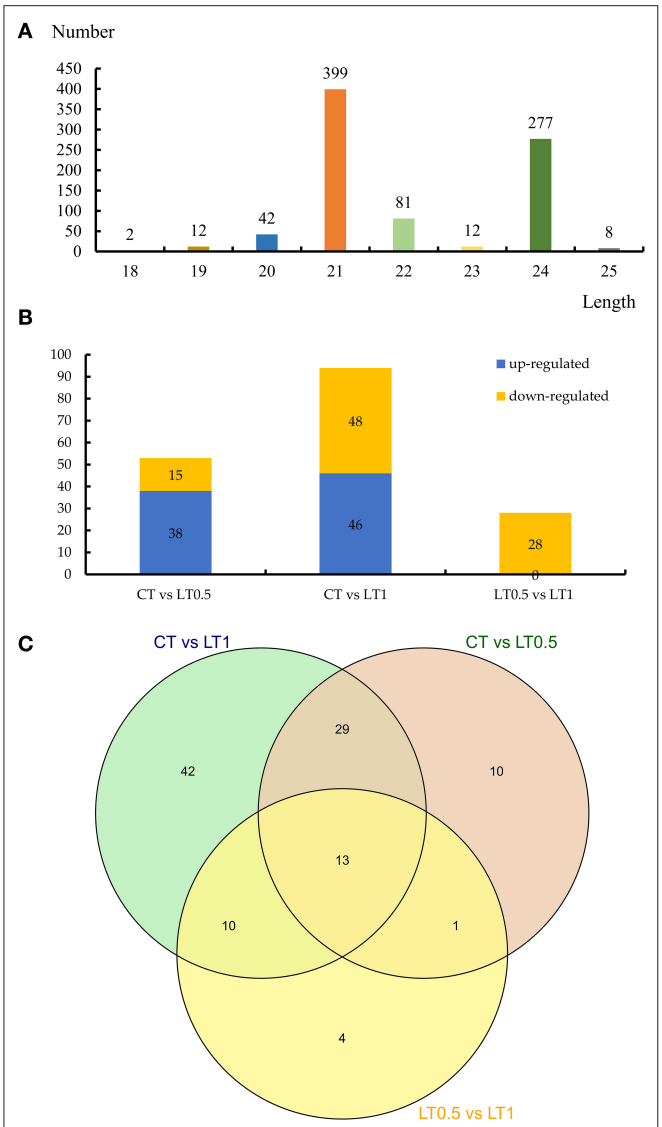


FIGURE 5 | Statistics of expressed miRNAs in this study. **(A)** The length distribution of miRNAs identified. **(B)** Number of downregulated and upregulated differentially expressed miRNAs in CT vs. LT0.5, CT vs. LT1, and LT0.5 vs. LT1. **(C)** Venn diagram of differentially expressed miRNAs distributed in samples.

Validation of Differentially Expressed mRNAs and miRNAs by Real-time QPCR Analysis

The expression of 10 cold-responsive miRNAs and 25 DEGs was detected by QPCR. The comparisons between QPCR and mRNA-seq (or miRNA-seq) provided in **Supplementary Figure 6** suggested that the result of mRNA-seq and miRNA-seq was reliable. This was also confirmed by calculating their Pearson correlation coefficients which were 0.847 and 0.852, respectively (**Figure 6**).

DISCUSSION

Cold stress is a major abiotic environmental factor that negatively affects plants by decelerating seed germination, root development, chlorophyll synthesis, and photosynthesis. Considering this, several protective mechanisms are employed by plants to manage adverse conditions. In sugarcane fields in low latitude and complex mountain and plain environments, cold damage has a serious impact on normal sucrose synthesis. To improve cold resistance in sugarcane, genetics and breeding research has emphasized high-quality cold-resistant varieties. A total of 6,034 DEGs were identified in this analysis as responsive to cold stress (4°C). This number is similar to that of previous reports, which proposed the presence of 2,583 upregulated and 3,302 downregulated genes from a low-temperature (10°C) treatment of *S. spontaneum* clone IND 00-1037 (Dharshini et al., 2016).

Transcriptional Regulatory Network Underlying Cold Stress at the Seedling Stage

Cold sensing occurs via membrane fluidity, calcium, and lipid signaling genes as well as via MAP kinases and phytohormone signaling (Zhao et al., 2017). Over the past several decades, various reports have confirmed that many transcription factors, including CBF, MYB, CAMAT, and NAC, are either directly or indirectly involved in cold stress tolerance; these were crossed with kinase, hormones, circadian cycles, Ca^{2+}, and light to accomplish cold signal transduction (Peng et al., 2015; Li et al., 2018; Shi et al., 2018). A total of 4% of upregulated genes were related to signaling, while MYB was downregulated 2.4-fold (Dharshini et al., 2016). In maize seedlings under various stress conditions, 43 TF families containing 403 differentially expressed TFs belong to the ERF, MYB, bZIP, bHLH, WRKY, and NAC TF families (Li et al., 2017). In our study, we found that 56 DEGs belong to MYB, bZIP, GATA, and MADS-box TFs in the CT vs. LT0.5 comparison, while 179 other DEGs related to PP2C, CBL-IPK, CDPK, AMPK, calcium, cAMP, MAPK, phosphatidylinositols, plant hormones, and sphingolipids were the top enriched signaling pathways. These results emphasized that TFs, kinase, and plant hormones played vital roles in the cold signal transduction of *S. spontaneum*.

A total of 14 PP2Cs and 16 PP2Cs in LT0.5 and LT1 compared with CT, respectively, were either downregulated or upregulated under cold stress conditions. Three DEGs (Sspon.004B0016440, Sspon.004B0016450, and Sspon.004B0015590) encoding PP2C proteins are predicted to be potential hubs via string analysis (**Figure 4**). The first plant PP2C protein (ABI1) was identified as a negative regulator of ABA signaling (Schweighofer et al., 2004). An increasing number of PP2Cs have been defined in different species, and their functions have been subsequently studied. Thus, PP2C proteins were thought to be core members involved in the ABA signaling pathway (Zhu, 2016); however, more recent research showed that the PYL-PP2C-SnRK2 core ABA signaling module activates a MAPK cascade comprised of MAP3Ks MAP3K17/18, MAP2K MKK3, and MAPKs MPK1/2/7/14. These

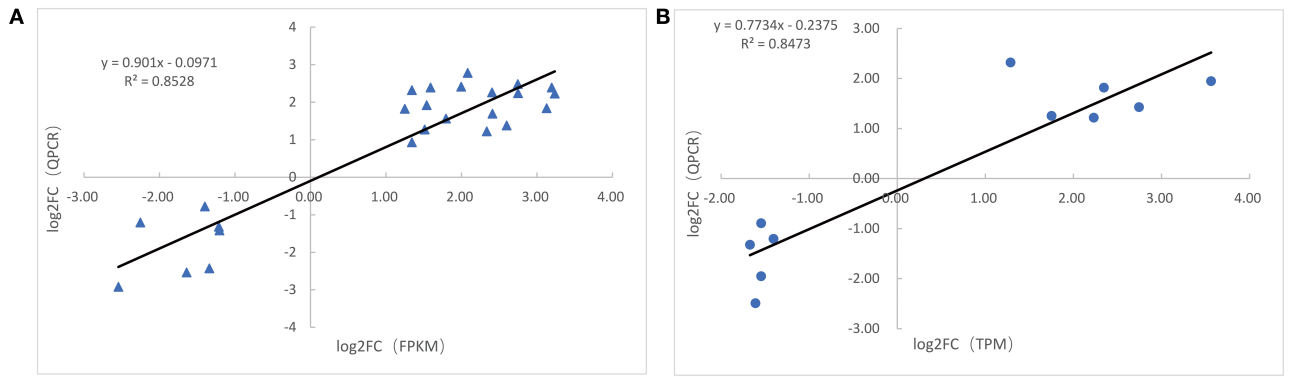


FIGURE 6 | Correlation analysis of validation. **(A)** Correlation analysis of gene expression between qPCR and RNA-seq data. **(B)** Correlation analysis of miRNA expression between qPCR and miRNA-seq data. The X-axis represents RNA-seq (miRNA-seq) data, and the Y-axis represents the qPCR data.

genes might regulate a range of ABA effector proteins via phosphorylation (De Zelicourt et al., 2016) and may also play major roles in the regulation of cell growth as well as cellular biotic and abiotic stress signaling (Lammers and Lavi, 2007; Fan et al., 2019). Proteins belonging to MAPK cascades have also been shown to be involved in the response of plants to cold stress (Markus et al., 2004; Tak et al., 2019), while PTP1 is known to downregulate cold stress (Liu et al., 2016). Thus, we propose that the ABA signaling pathway via PP2C is most important in terms of the cold stress responses of *S. spontaneum* at the seedling stage; however, future research should be conducted on this pathway.

Research has shown that both the regeneration rate and carboxylation efficiency of RuBP decline under chilling stress (Hardigan et al., 2016; Jin et al., 2017). DEG pathways shown to be enriched in GO and KEGG analyses in CT vs. LT1 have emphasized carbohydrate metabolism including how chloroplasts are related to the photosystem and starch and sucrose biosynthesis. These could also be verified by the functional annotation and PPI analysis (**Supplementary Figure 3**).

Plant secondary metabolism is closely related to abiotic stress responses (Zhang et al., 2017b). Flavonoids play a role in the response of higher plants to abiotic stressors (Jin et al., 2017). For example, in *E. nutans*, seven genes involved in phenylpropanoid biosynthesis are known to be specifically related to cold stress response (Fu et al., 2016). In terms of enzymes involved in phenylpropanoid biosynthesis, phenylalanine ammonia lyase (PAL) is one of the most relevant (Vogt, 2010). The RNA-seq data presented here show that the number of DEGs related to phenylpropanoid pathways accounts for the largest proportion of secondary metabolic pathways. Therefore, we inferred that lignin might play a key adaptive role in cold stress response.

Cold Stress-Responsive miRNAs: Sugarcane Regulatory Network

Yang et al. (2017) found that 4,002 targets were predicted for 259 miRNAs, which may be due to the absence of a sugarcane genome leading to inaccuracies in prediction. Using the published sugarcane genome as our starting point, we found 365 genes from 305 miRNAs, which is similar to reports on *Astragalus membranaceus* (Abla et al., 2019), *Nelumbo nucifera* (Zou et al., 2020), potato tubers (Ou et al., 2014), sweet potato (Xie et al., 2017), and other plants (Gupta et al., 2014).

Several previous studies on various plant species have confirmed the involvement of miRNAs in cold stress responses (Tang and Chu, 2017; Zeng et al., 2018; Song et al., 2019; Sun et al., 2019; Esposito et al., 2020). A total of 412 sugarcane miRNAs have been isolated from two cultivars, including 261 known and 151 novel miRNAs. These include nodes for ROC22 (relative cold sensitivity) and FN39 (relative cold tolerance); data indicate that 62 of these miRNAs exhibit significant cold stress-induced or depressed expression (Yang et al., 2017), which is less than that found in our study. We found that 109 miRNAs belonging to 11 families responded to cold stress, while miR444 had the most differentially expressed miRNAs (52) followed by miR169-1 (18) (22) and miR396 (13). Compared to previous studies, we also found that miR156, miR168, and miR408 are upregulated under cold stress in sugarcane (Yang et al., 2017). The miR444 in rice has been reported as a key factor relaying antiviral signaling from virus infections to *OsRDR1* expression (Wang et al., 2016a) as well as the promotion of BR biosynthesis (Jiao et al., 2019). This gene is also involved in the nitrate signaling pathway (Yan et al., 2014). Increased expression of miR408 could also lead to improved *Arabidopsis* tolerance to cold and oxidative stress, as shown by a lower ROS level and the induction of genes related to antioxidative function (Ma et al., 2015).

We found 28, 44, and 13 miR444s in CT vs. LT0.5, CT vs. LT1, and CT0.5 vs. LT1, respectively. This is especially the case for Novel-Chr5C_55773, which exhibited induced expression after a 1-h cold stress and targeted three DEGs encoding the apyrase 3 family (Sspon.005D0023491, Sspon.005B0020100, and Sspon.005C0019251). It has been shown that these can maintain plasma membrane integrity and reduce electrolyte leakage under cold stress (Deng et al., 2015). Four LRR-RLKs were also targeted by miR444 (Novel-Chr6B_65223) and miR168 (Novel-Chr5D_60023) after a 1-h cold stress. We propose that at an early stage, cold stress reduces plasma membrane integrity by decreasing the expression of genes encoding apyrase 3 through

the regulation of miR444, which also acts as membrane location receptors. These four LRR-RLK genes play an important function in cold stress signal transduction.

An integrated miRNA–gene/mRNA regulation network suggests that 25 miRNAs belong to nine families that play important roles in response to cold stress. Among these nine families, Novel-tig00084004_47811 (miR169), Novel-Chr4C_24049 (miR396), Novel-Chr5C_29143 (miR396), and Novel-tig00005469_46467 (miR159) were significant regulators. Specifically, miR169 and miR396 have been reported to play an indirect role in cold stress response by regulating the TFs (Wang et al., 2017; Sombir et al., 2020). The OsGRF4 transcription factor is targeted by miRNA396, whereas mutations disturb the *GRF4*-miR396 stress response network and results in the development of enlarged grains and cold tolerance enhancements in rice (Chen et al., 2019). Among 14 targets, two genes encoding MYB TFs are thought to be involved in transcriptional regulation. LRR-RLKs and phosphatidylinositol-4-phosphate 5-kinase have also been identified as regulators that function in early cold stress signal transduction.

Gene miR396 in *Poncirus trifoliata* functions in cold tolerance by regulating acc oxidase gene expression and modulating ethylene-polyamine homeostasis (Zhang et al., 2016). This gene also acts as a stress-responsive regulator by conferring tolerance to abiotic stressors and susceptibility to mold infection in tobacco (Chen et al., 2015). A total of 17 genes that encode pentatricopeptide repeat (PPR) family proteins were targeted by Novel-Chr5C_29143 (miR396); these genes play critical roles in all aspects of organelle RNA metabolism including development and stress defense (Zhang et al., 2017a). Another miR396 family member, Novel-Chr4C_24049, was upregulated after a 0.5-h cold stress and is thought to regulate 11 laccases−13 genes, especially Sspon.003B0004220, Sspon.003A0003410, and Sspon.007A0025870. These 11 laccases−13 genes are also differentially expressed under cold stress. Our results suggested that MiR444, miR169, and miR396 play key roles in the regulation of signal transduction, transcriptional, and lignin metabolism in sugarcane responding to cold stress.

CONCLUSIONS

The results of this study show that the multiple regulated mechanisms are involved in the response of sugarcane to cold stress. The signaling pathway mediated by PP2C and LRR-RLK is thought to play a key role in the response of sugarcane to cold stress and has not been previously reported. Most importantly, genes involved in the biosynthesis of polyunsaturated fatty acids and LEA protein, lignin, and pectin pathways were found to be directly involved in cold stress responses targeted by miR444, miR169, and miR396. These key candidate genes and miRNAs will be useful for future research on cold stress response mechanisms. This study provides new insight into the regulatory network at transcriptional and post-transcriptional levels, thus allowing us to better understand the response of sugarcane to cold stress.

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

YLia, XS, and BZ: data curation. BZ, ZQ, and DL: formal analysis. RC, ZZ, and YD: investigation. YLia and JWu: funding acquisition. JWe and JWu: project administration. XH and JWu: writing—original draft. XH, YLi, and XS: writing—review and editing. All authors approved the final manuscript.

FUNDING

This research was supported by grants from the National Natural Science Foundation of China (31660356, 31901594, and 31360312), the Development and Regulation of Economically Important Traits in Tropical Crops (2018YFD1000500), the Guangxi Post-doctoral Science Foundation, Guangxi Science and Technology Base and Talents Special Project (GuikeAD17195100), the Fund for Guangxi Innovation Teams of Modern Agriculture Technology (nycytxgxcxtd-03-01), and the Fund Science and Technology Development of Guangxi Academy of Agricultural Sciences (Guinongke2021YT013, Guinongke2017JZ19, and Guinongke2016JM06).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.581993/full#supplementary-material

Supplementary Figure 1 | Functional annotation of all genes isolated from *S. spontaneum* in this study. **(A)** Statistics for the functional annotation for all identified genes. **(B)** GO assignment for all identified genes. **(C)** Statistics of the NR annotation results mapped to other plants. **(D)** Unigenes were aligned to the COG (Clusters of Orthologous Groups) database to predict and categorize possible functions.

Supplementary Figure 2 | Annotation statistics for DEGs in the second level of the GO database **(A–C)** and Annotation statistics for functional categories for DEGs from the COG database **(D,E)**. **(A,D)** Statistics of DEGs from CT vs. LT0.5. **(B,E)** Statistics of DEGs from CT vs. LT1. **(C,F)** Statistics of DEGs from CT0.5 vs. LT1.

Supplementary Figure 3 | DEG PPI network. Interaction relationship in the STRING protein interaction database (http://string-db.org/) to analyze DEGs coding in the protein interaction network. **(A)** PPI map for CT vs. LT0.5. **(B)** Map of PPI for CT 0.5 vs. LT1. The size of a node in this interaction network is proportional to the degree of this node; thus, the more edges connected to a node, the greater its degree, the larger the node, and the darker the color. Nodes may be in a more central position in the network and lines of different colors represent varied interactions. The thicker the line, the higher the score.

Supplementary Figure 4 | Analysis of miRNA expression correlation among samples and differential expression clustering. **(A)** Clusters of expression patterns of differentially expressed miRNAs. The abscissa represents the sample name and the clustering result of the sample, while the ordinate represents differential miRNA and miRNA clustering results. Different columns represent different samples, while different rows represent different miRNAs. The color in this Figure represents miRNA expression level log2 (TPM) in the sample after normalization (scale

number). **(B)** Correlation heat map of the expression levels of two samples. Pearson's Correlation Coefficient was used as an evaluation index for biological repeat correlations. The closer *r* is to 1, the stronger the correlation between the two replicate samples.

Supplementary Figure 5 | Base preferences for miRNA. **(A)** Base bias of known miRNAs. **(B)** Base bias of novel miRNAs. The abscissa is the base position of miRNA while the ordinate is the percentage of bases A/U/C/G appearing in the miRNA at that position.

Supplementary Figure 6 | QPCR validation for mRNA and miRNA. **(A–C)** The validation of mRNA in CT vs. LT0.5, CT vs. LT1, and LT0.5 vs. LT1. **(D–F)** The validation of miRNA in CT vs. LT0.5, CT vs. LT1, and LT0.5 vs. LT1. The Y-axis represents the value of relative expression which is conversed by Log_2 (Fold change), the X-axis represents the ID of mRNA or miRNA.

Supplementary Table 1 | List of primers used for QPCR validation.

Supplementary Table 2 | Statistics of mRNA-seq data and the comparison between sequencing data and reference genome.

Supplementary Table 3 | The FPKM value and Functional annotation of all identified genes from *S. spontaneum* in this study.

Supplementary Table 4 | Statistics of functional annotation for all identified genes in this study.

Supplementary Table 5 | List of DEGs in comparison between CT and LT0.5.

Supplementary Table 6 | List of DEGs in comparison between CT and LT1.

Supplementary Table 7 | List of DEGs in comparison between LT0.5 and LT1.

Supplementary Table 8 | Statistics of KEGG annotation in comparisons.

Supplementary Table 9 | Statistics of sRNA-seq reads.

Supplementary Table 10 | The family and TPM value of every miRNA in each sample.

Supplementary Table 11 | Statistics of miRNAs in each sample.

Supplementary Table 12 | Statistics of miRNA with target and target genes.

Supplementary Table 13 | List of target genes for every miRNA.

Supplementary Table 14 | The function annotation of target genes.

Supplementary Table 15 | List of annotation for target genes against GO database.

Supplementary Table 16 | List of annotation for target genes against KEGG database.

Supplementary Table 17 | The differentially expressed miRNAs in comparison of CT vs. LT 0.5.

Supplementary Table 18 | The differentially expressed miRNAs in comparison of CT vs. LT1.

Supplementary Table 19 | The differentially expressed miRNAs in comparison of LT 0.5 vs. LT 1.

Supplementary Table 20 | Differentially expressed miRNAs under cold stress and their TPM value.

Supplementary Table 21 | Corresponding miRNA-mRNA and the functional annotation of mRNA from CT vs. LT0.5.

Supplementary Table 22 | Corresponding miRNA-mRNA and the functional annotation of mRNA from CT vs. LT1.

REFERENCES

Abla, M., Sun, H., Li, Z., Wei, C., Gao, F., Zhou, Y., et al. (2019). Identification of miRNAs and their response to cold stress in *Astragalus membranaceus*. *Biomolecules* 9:182. doi: 10.3390/biom9050182

Ana, K., and Sam, G.-J. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* 39(Database issue), D152–D157. doi: 10.1093/nar/gkq1027

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

Chen, L., Luan, Y., and Zhai, J. (2015). Sp-miR396a-5p acts as a stress-responsive genes regulator by conferring tolerance to abiotic stresses and susceptibility to *Phytophthora nicotianae* infection in transgenic tobacco. *Plant Cell Rep.* 34, 2013–2025. doi: 10.1007/s00299-015-1847-0

Chen, X., Jiang, L., Zheng, J., Chen, F., Wang, T., Wang, M., et al. (2019). A missense mutation in large grain size 1 increases grain size and enhances cold tolerance in rice. *J. Exp. Bot.* 70, 3851–3866. doi: 10.1093/jxb/erz192

De Zelicourt, A., Colcombet, J., and Hirt, H. (2016). The role of MAPK modules and ABA during abiotic stress signaling. *Trends Plant Sci.* 21, 677–685. doi: 10.1016/j.tplants.2016.04.004

Deng, S., Sun, J., Zhao, R., Ding, M., and Chen, S. (2015). Populus euphratica APYRASE2 enhances cold tolerance by modulating vesicular trafficking and extracellular atp in *Arabidopsis* plants. *Plant Physiol.* 169, 530–548. doi: 10.1104/pp.15.00581

Dharshini, S., Chakravarthi, M., Ashwin Narayan, J, Manoj, V. M., Naveenarani, M., Kumar, R., et al. (2016). *De novo* sequencing and transcriptome analysis of a low temperature tolerant *Saccharum spontaneum* clone IND 00-1037. *J. Biotechnol.* 231, 280–294. doi: 10.1016/j.jbiotec.2016.05.036

Ding, C., Yang, Q., Li, F., and Weifu, L. (2006). Effect of cold stress on the content of free-proline in leaf of *Saccharum spontaneum* L. and sclerostachya (II). *J. Anhui Agric. Sci.* 34, 30–33. doi: 10.13989/j.cnki.0517-6611.2006.05.010

Esposito, S., Aversano, R., Bradeen, J. M., Di Matteo, A., Villano, C., and Carputo, D. (2020). Deep-sequencing of *Solanum commersonii* small RNA libraries reveals riboregulators involved in cold stress response. *Plant Biol.* 22, 133–142. doi: 10.1111/plb.12955

Fan, K., Yuan, S., Chen, J., Chen, Y., Li, Z., Lin, W., et al. (2019). Molecular evolution and lineage-specific expansion of the PP2C family in *Zea mays*. *Planta* 250, 1521–1538. doi: 10.1007/s00425-019-03243-x

Fu, J., Miao, Y., Shao, L., Hu, T., and Yang, P. (2016). *De novo* transcriptome sequencing and gene expression profiling of *Elymus nutans* under cold stress. *BMC Genomics* 17:870. doi: 10.1186/s12864-016-3222-0

Guo, X., Liu, D., and Kang, C. (2018). Cold signaling in plants: insights into mechanisms and regulation. *J. Integr. Plant Biol.* 60, 745–756. doi: 10.1111/jipb.12706

Gupta, O., Meena, N., Sharma, I., and Sharma, P. (2014). Differential regulation of microRNAs in response to osmotic, salt and cold stresses in wheat. *Mol. Biol. Rep.* 41, 4623–4629. doi: 10.1007/s11033-014-3333-0

Hardigan, M. A., Crisovan, E., Hamilton, J. P., Kim, J., Laimbeer, P., Leisner, C. P., et al. (2016). Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated *Solanum tuberosum*. *Plant Cell* 28, 388–405. doi: 10.1105/tpc.15.00538

Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27, 431–432. doi: 10.1093/bioinformatics/btq675

Jiao, X., Wang, H., Yan, J., Kong, X., and Yan, Y. (2019). Promotion of BR biosynthesis by miR444 is required for ammonium-triggered inhibition of root growth. *Plant Physiol.* 182, 1454–1466. doi: 10.1104/pp.19.00190

Jin, J., Zhang, H., Zhang, J., Liu, P., Chen, X., Li, Z., et al. (2017). Integrated transcriptomics and metabolomics analysis to characterize cold stress responses in *Nicotiana tabacum*. *BMC Genomics* 18:496. doi: 10.1186/s12864-017-3871-7

Jong-Won, P., Benatti, T. R., Thiago, M., Yu, Q., Nora, S. G., Victoria, M., et al. (2015). Cold responsive gene expression profiling of sugarcane and *Saccharum spontaneum* with functional analysis of a cold inducible saccharum homolog of NOD26-Like intrinsic protein to salt and water stress. *PLoS ONE* 10:e0125810. doi: 10.1371/journal.pone.0125810

Lammers, T., and Lavi, S. (2007). Role of type 2C protein phosphatases in growth regulation and in cellular stress signaling. *Crit. Rev. Biochem. Mol. Biol.* 42, 437–461. doi: 10.1080/10409230701693342

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Li, H.-B., Li, N., Yang, S.-Z., Peng, H.-Z., Wang, L.-L., Wang, Y., et al. (2016). Transcriptomic analysis of *Casuarina equisetifolia* L. *in responses to cold stress. Tree Genet. Genomes* 13:7. doi: 10.1007/s11295-016-1090-z

Li, P., Cao, W., Fang, H., Xu, S., Yin, S., Zhang, Y., et al. (2017). Transcriptomic profiling of the maize (*Zea mays* L.) leaf response to abiotic stresses at the seedling stage. *Front. Plant Sci.* 8:290. doi: 10.3389/fpls.2017.00290

Li, Q., Byrns, B., Badawi, M. A., Diallo, A. B., Danyluk, J., Sarhan, F., et al. (2018). Transcriptomic insights into phenological development and cold tolerance of wheat grown in the field. *Plant Physiol.* 176, 2376–2394. doi: 10.1104/pp.17.01311

Liu, S., Chen, H., Li, X., and Zhang, W. (2016). A low-temperature-responsive element involved in the regulation of the *Arabidopsis thaliana* At1g71850/At1g71860 divergent gene pair. *Plant Cell Rep.* 35, 1757–1767. doi: 10.1007/s00299-016-1994-y

Ma, C., Burd, S., and Lers, A. (2015). miR408 is involved in abiotic stress responses in *Arabidopsis*. *Plant J.* 84, 169–187. doi: 10.1111/tpj.12999

Markus, T., Elisabeth, S., Thomas, E., Dóczi, R., Kazuya, I., Kazuo, S., et al. (2004). The MKK2 pathway mediates cold and salt stress signaling in *Arabidopsis*. *Mol. Cell* 15, 141–152. doi: 10.1016/j.molcel.2004.06.023

Megha, S., Basu, U., and Kav, N. N. V. (2018). Regulation of low temperature stress in plants by microRNAs. *Plant Cell Environ.* 41, 1–15. doi: 10.1111/pce.12956

Nogueira, F. T., De Rosa, V. E. Jr, Menossi, M., Ulian, E. C., and Arruda, P. (2003). RNA expression profiles and data mining of sugarcane response to low temperature. *Plant Physiol.* 132, 1811–1824. doi: 10.1104/pp.102.017483

Ou, Y., Liu, X., Xie, C., Zhang, H., Lin, Y., Li, M., et al. (2014). Genome-wide identification of microRNAs and their targets in cold-stored potato tubers by deep sequencing and degradome analysis. *Plant Mol. Biol. Rep.* 33, 584–597. doi: 10.1007/s11105-014-0771-8

Pang, X., Zhu, P., Zhou, Q., Huang, Q., Tan, Q., and Chun, L. (2015). Research progress on sugarcane response to cold stress. *Guizhou Agric. Sci.* 43, 39–43.

Peng, X., Wu, Q., Teng, L., Tang, F., Pi, Z., and Shen, S. (2015). Transcriptional regulation of the paper mulberry under cold stress as revealed by a comprehensive analysis of transcription factors. *BMC Plant Biol.* 15:108. doi: 10.1186/s12870-015-0489-2

Rasheed, R., Wahid, A., Ashraf, M., and Basra, S. M. A. (2010). Role of proline and glycinebetaine in improving chilling stress tolerance in sugarcane buds at sprouting. *Int. J. Agric. Biol.* 12, 1560–8530. doi: 10.3763/ijas.2009.0478

Schweighofer, A., Hirt, H., and Meskiene, I. (2004). Plant PP2C phosphatases: emerging functions in stress signaling. *Trends Plant Sci.* 9, 236–243. doi: 10.1016/j.tplants.2004.03.007

Selvarajan, D., Mohan, C., Dhandapani, V., Nerkar, G., Jayanarayanan, A. N., Vadakkancherry Mohanan, M., et al. (2018). Differential gene expression profiling through transcriptome approach of *Saccharum spontaneum* L. *under low temperature stress reveals genes potentially involved in cold acclimation. 3 Biotech* 8:195. doi: 10.1007/s13205-018-1194-2

Shi, Y., Ding, Y., and Yang, S. (2018). Molecular regulation of cbf signaling in cold acclimation. *Trends Plant Sci.* 23, 623–637. doi: 10.1016/j.tplants.2018.04.002

Sombir, R., Sonia, B., Sarita, J., and Saloni, M. (2020). Novel insights into expansion and functional diversification of MIR169 family in tomato. *Planta* 251, 1–17. doi: 10.1007/s00425-020-03346-w

Song, X., Li, Y., Cao, X., and Qi, Y. (2019). MicroRNAs and their regulatory roles in plant-environment interactions. *Annu. Rev. Plant Biol.* 70, 489–525. doi: 10.1146/annurev-arplant-050718-100334

Sternes, P. R., and Moyle, R. L. (2015). Deep sequencing reveals divergent expression patterns within the small RNA transcriptomes of cultured and vegetative tissues of sugarcane. *Plant Mol. Biol. Rep.* 33, 931–951. doi: 10.1007/s11105-014-0787-0

Sun, X., Lin, L., and Sui, N. (2019). Regulation mechanism of microRNA in plant response to abiotic stress and breeding. *Mol. Biol. Rep.* 46, 1447–1457. doi: 10.1007/s11033-018-4511-2

Tak, H., Negi, S., Rajpurohit, Y. S., Misra, H. S., and Ganapathi, T. R. (2019). MusaMPK5, a mitogen activated protein kinase is involved in regulation of cold tolerance in banana. *Plant Physiol. Biochem.* 146, 112–123. doi: 10.1016/j.plaphy.2019.11.012

Tang, J., and Chu, C. (2017). MicroRNAs in crop improvement: fine-tuners for complex traits. *Nat. Plants* 3:17077. doi: 10.1038/nplants.2017.77

Thiebaut, F., Rojas, C. A., Almeida, K. L., Grativol, C., Domiciano, G. C., Lamb, C. R. C., et al. (2012). Regulation of miR319 during cold stress in sugarcane. *Plant Cell Environ.* 35, 502–512. doi: 10.1111/j.1365-3040.2011.02430.x

Vogt, T. (2010). Phenylpropanoid biosynthesis. *Mol. Plant* 3, 2–20. doi: 10.1093/mp/ssp106

Wang, H., Jiao, X., Kong, X., Humaira, S., Wu, Y., Chen, X., et al. (2016a). A signaling cascade from miR444 to RDR1 in rice antiviral RNA silencing pathway. *Plant Physiol.* 170, 2365–2377. doi: 10.1104/pp.15.01283

Wang, L. L., Zhao, T. L., Jin-Tao, G. E., and Liu, X. M. (2017). Application prospects of plant cold-stress-responsive miRNAs in cold resistance research of plants. *Acta Agric. Shanghai* 33, 129–134. doi: 10.15955/j.issn1000-3924.2017.06.25

Wang, Q., Liu, N., Yang, X., Tu, L., and Zhang, X. (2016b). Small RNA-mediated responses to low- and high-temperature stresses in cotton. *Sci. Rep.* 6:35558. doi: 10.1038/srep35558

Wang, S.-T., Sun, X.-L., Hoshino, Y., Yu, Y., Jia, B., Sun, Z.-W., et al. (2014). MicroRNA319 positively regulates cold tolerance by targeting OsPCF6 and OsTCP21 in rice (*Oryza sativa* L.). *PLoS ONE* 9:e91357. doi: 10.1371/journal.pone.0091357

Wouter, W., Carmen, A. A., Gertjan, K., Johannes, P. V., Marrije, R. B., Gemma, G. K., et al. (2014). Label-free LC-MSe in tissue and serum reveals protein networks underlying differences between benign and malignant serous ovarian tumors. *PLoS ONE* 9:e108046. doi: 10.1371/journal.pone.0108046

Xie, Z., Wang, A., Li, H., Yu, J., Jiang, J., Tang, Z., et al. (2017). High throughput deep sequencing reveals the important roles of microRNAs during sweetpotato storage at chilling temperature. *Sci. Rep.* 7:16578. doi: 10.1038/s41598-017-16871-8

Yan, Y., Wang, H., Hamera, S., Chen, X., and Fang, R. (2014). miR444a has multiple functions in the rice nitrate-signaling pathway. *Plant J.* 78, 44–55. doi: 10.1111/tpj.12446

Yang, C., Li, D., Mao, D., Liu, X., Ji, C., Li, X., et al. (2013). Overexpression of microRNA319 impacts leaf morphogenesis and leads to enhanced cold tolerance in rice (*Oryza sativa* L.). *Plant Cell Environ.* 36, 2207–2218. doi: 10.1111/pce.12130

Yang, Y., Zhang, X., Su, Y., Zou, J., Wang, Z., Xu, L., et al. (2017). miRNA alteration is an important mechanism in sugarcane response to low-temperature environment. *BMC Genomics* 18:833. doi: 10.1186/s12864-017-4231-3

Zeng, X., Xu, Y., Jiang, J., Zhang, F., Ma, L., Wu, D., et al. (2018). Identification of cold stress responsive microRNAs in two winter turnip rape (*Brassica rapa* L.) by high throughput sequencing. *BMC Plant Biol.* 18:52. doi: 10.1186/s12870-018-1242-4

Zhang, J., Xiao, J., Li, Y., Su, B., Xu, H., Shan, X., et al. (2017a). PDM3, a pentatricopeptide repeat-containing protein, affects chloroplast development. *J. Exp. Bot.* 68, 5615–5627. doi: 10.1093/jxb/erx360

Zhang, J., Xu, Y., Huan, Q., and Chong, K. (2009). Deep sequencing of brachypodium small RNAs at the global genome level identifies microRNAs involved in cold stress response. *BMC Genomics* 10:449. doi: 10.1186/1471-2164-10-449

Zhang, J., Zhang, X., Tang, H., Zhang, Q., 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

Zhang, X., Teixeira Da Silva, J. A., Niu, M., Li, M., He, C., Zhao, J., et al. (2017b). Physiological and transcriptomic analyses reveal a response mechanism to cold stress in *Santalum album* L. leaves. *Sci. Rep.* 7:42165. doi: 10.1038/srep42165

Zhang, X., Wei, W., Ming, W., Zhang, H.-Y., and Liu, J.-H. (2016). The miR396b of poncirus trifoliata functions in cold tolerance by regulating ACC oxidase gene expression and modulating ethylene–polyamine homeostasis. *Plant Cell Physiol.* 57, 1865–78. doi: 10.1093/pcp/pcw108

Zhao, C., Wang, P., Si, T., Hsu, C. C., Wang, L., Zayed, O., et al. (2017). MAP kinase cascades regulate the cold response by modulating

ICE1 protein stability. *Dev. Cell* 43, 618–629. doi: 10.1016/j.devcel.2017.09.024

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

Zhu, P., Pang, X., Tan, Q., Liang, C., Yan, L., Zhou, Q., et al. (2019). Effects of chilling on photosynthesis and photosynthetic pigment contents in leaves of different sugarcane varieties. *Chin. J. Trop. Crops* 40, 51–57.

Zou, Y., Chen, G., Jin, J., Wang, Y., Xu, M., Peng, J., et al. (2020). Small RNA and transcriptome sequencing reveals miRNA regulation of floral thermogenesis in nelumbo nucifera. *Int. J. Mol. Sci.* 21:3324. doi: 10.3390/ijms21093324

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.

Copyright © 2021 Huang, Liang, Zhang, Song, Li, Qin, Li, Chen, Zhou, Deng, Wei and Wu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 28 January 2021
doi: 10.3389/fgene.2021.552454

Profiling of MicroRNAs and Their Targets in Roots and Shoots Reveals a Potential MiRNA-Mediated Interaction Network in Response to Phosphate Deficiency in the Forestry Tree *Betula luminifera*

Junhong Zhang*, Yan Lin, Fangmin Wu, Yuting Zhang, Longjun Cheng, Menghui Huang and Zaikang Tong*

State Key Laboratory of Subtropical Silviculture, School of Forestry and Bio-Technology, Zhejiang A&F University, Hangzhou, China

OPEN ACCESS

Edited by:
Rong Zhou,
Aarhus University, Denmark

Reviewed by:
Xiping Yang,
Guangxi University, China
Ru Gang Yu,
Huaibei Normal University, China

****Correspondence:***
Junhong Zhang
zjh6719036@163.com;
zhangjunhong@zafu.edu.cn
Zaikang Tong
zktong@zafu.edu.cn

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

Received: *16 April 2020*
Accepted: *06 January 2021*
Published: *28 January 2021*

Citation:
Zhang J, Lin Y, Wu F, Zhang Y, Cheng L, Huang M and Tong Z (2021) Profiling of MicroRNAs and Their Targets in Roots and Shoots Reveals a Potential MiRNA-Mediated Interaction Network in Response to Phosphate Deficiency in the Forestry Tree Betula luminifera. Front. Genet. 12:552454. doi: 10.3389/fgene.2021.552454

Inorganic phosphate (Pi) is often lacking in natural and agro-climatic environments, which impedes the growth of economically important woody species. Plants have developed strategies to cope with low Pi (LP) availability. MicroRNAs (miRNAs) play important roles in responses to abiotic stresses, including nutrition stress, by regulating target gene expression. However, the miRNA-mediated regulation of these adaptive responses and their underlying coordinating signals are still poorly understood in forestry trees such as *Betula luminifera*. Transcriptomic libraries, small RNA (sRNA) libraries, and a mixed degradome cDNA library of *B. luminifera* roots and shoots treated under LP and normal conditions (CK) were constructed and sequenced using next-generation deep sequencing. A comprehensive *B. luminifera* transcriptome derived from its roots and shoots was constructed, and a total of 76,899 unigenes were generated. Analysis of the transcriptome identified 8,095 and 5,584 differentially expressed genes in roots and shoots, respectively, under LP conditions. sRNA sequencing analyses indicated that 66 and 60 miRNAs were differentially expressed in roots and shoots, respectively, under LP conditions. A total of 109 and 112 miRNA–target pairs were further validated in the roots and shoots, respectively, using degradome sequencing. Kyoto Encyclopedia of Genes and Genomes pathway enrichment analysis of differential miRNA targets indicated that the "ascorbate and aldarate metabolism" pathway responded to LP, suggesting miRNA-target pairs might participating in the removing of reactive oxidative species under LP stress. Moreover, a putative network of miRNA–target interactions involved in responses to LP stress in *B. luminifera* is proposed. Taken together, these findings provide useful information to decipher miRNA functions and establish a framework for exploring P signaling networks regulated by miRNAs in *B. luminifera* and other woody plants. It may provide new insights into the genetic engineering of high use efficiency of Pi in forestry trees.

Keywords: abiotic stress, degradome, transcriptome, miRNA, Pi deficiency, *Betula luminifera*

INTRODUCTION

Phosphate (P), one of the essential macronutrients in plants, is often not readily accessible to plants because of the low availability of inorganic phosphate (Pi), the major acquired form of P, in the soil (Marschner, 1995). Therefore, plants are frequently exposed to conditions of Pi limitation. To maintain plant productivity, large quantities of Pi fertilizers are applied, which is not only expensive but also harmful to the environment (Cordell et al., 2009). Genetic engineering of plants with enhanced Pi starvation tolerance is a promising approach to address these challenges. Therefore, a great deal of effort has been directed toward understanding the mechanisms of plant adaptation to Pi deficiency.

Plants have evolved the ability to acclimatize to Pi deficiency by triggering the Pi starvation response (PSR), which includes a series of morphological, physiological, and biochemical/metabolic adaptations. For example, extending the root's surface area for phosphate absorption underlies the morphological adaptations of Pi starved plants (Vance et al., 2003); increased synthesis of anthocyanins (Takahashi et al., 1991) and the synthesis of acid phosphatase enzymes (APases) is a universal response by plants to Pi deficiency (Duff et al., 1994; Tran et al., 2010). The PSR is partially achieved by the coordinated induction of hundreds of Pi starvation-inducible (PSI) genes, which increase Pi use efficiency and enhance Pi acquisition from the environment (Lin et al., 2009; Nilsson et al., 2010). These PSI genes are involved in Pi uptake, allocation, or remobilization, and include transcription factors, signaling molecules, and other upstream regulators (Lin et al., 2009; Zeng et al., 2016). Pi uptake by roots occurs mainly through Pi transporters, encoded by Phosphate Transporter1 (*PHT1*), of which PHT1;1 and PHT1;4 play major roles in acquiring Pi from the soil in *Arabidopsis* (Shin et al., 2004). Phosphate1 (*PHO1*) is located in root stelar cells, and participates in Pi transfer from roots to shoots (Hamburger et al., 2002). Under Pi starvation, WRKY6 and WRKY42 are degraded *via* 26S proteasome-mediated proteolysis (Chen et al., 2009; Su et al., 2015). Phosphate Response Ubiquitin E3 Ligase1 (*PRU1*) modulates Pi homeostasis by regulating the abundance of WRKY6 in response to low-Pi (LP) stress in *Arabidopsis* (Ye et al., 2018). The recent development of next-generation sequencing technology has facilitated the studies of the transcriptome at a whole-genome scale with single-base resolution (Trapnell et al., 2012), triggering a large number of Pi-responsive genes have been identified in several plants. For example, Secco et al. (2013) analyzed 126 paired-end RNA sequencing (RNA-Seq) libraries, and presented comprehensive spatiotemporal transcriptome analysis of responses to Pi stress in rice, revealing a large number of potential key regulators of Pi homeostasis in plants. In soybean, a total of 1,612 genes were differentially expressed in roots after exposure to Pi deficiency for 7 days (Zeng et al., 2016), and 2,055 genes exhibited differential expression patterns between Pi sufficient and deficient conditions in nodule (Xue et al., 2018). White lupin (*Lupinus albus*) has evolved unique adaptations for growth in Pi-deficient soil, and 2,128 sequences were shown to be differentially expressed in response to Pi deficiency (O'Rourke et al., 2013). Huang et al. (2019) identify 237 genes differentially expressed in the vasculature of *Plantago major* upon 24 h of Pi deficiency using RNA-Seq, of which only 27 have been previously identified to be specifically expressed in the vasculature tissues in other plant species.

Recent studies have also emphasized the importance of posttranscriptional mechanisms in the control of PSI gene expression. There is accumulating evidence that microRNAs (miRNAs) play critical roles in sensing Pi abundance, controlling Pi uptake and phloem-mediated long-distance transport, and Pi homeostasis (Hsieh et al., 2009; Zeng et al., 2013). MiRNAs negatively regulate gene expression by modulating both mRNA degradation and translational inhibition (Jones-Rhoades et al., 2006). Their expression profiles are significantly altered under conditions of stress, suggesting that attenuated plant growth and development under stress may be controlled by stress-responsive miRNAs (Sunkar et al., 2012). Recent studies have shown that abiotic stresses induce aberrant expression of many miRNAs, suggesting that miRNAs may be new targets for genetically improving plant tolerance to certain types of stress (Zhang, 2015). Mounting evidence shows that miRNAs play key roles in regulating nutrient metabolism in plants (Liang et al., 2015). Many miRNAs are involved in plant responses to nutrient stresses by acting as systemic signals to coordinate physiological activities that help plants to respond and survive these stresses (Zeng et al., 2013). A series of Pi starvation-responsive miRNAs were identified *via* high-throughput sequencing of small RNAs (sRNAs), which mediate the regulation of Pi uptake, transport, and assimilation in plants by targeting a set of genes (Kumar et al., 2017). For example, Hsieh et al. (2009) reported that the expression of miR156, miR399, miR778, miR827, and miR2111 was induced, whereas the expression of miR169, miR395, and miR398 was repressed upon Pi deprivation in *Arabidopsis* using sRNA sequencing. Xu et al. (2013) found 25 miRNAs were induced and 11 miRNAs were repressed by Pi starvation in soybean using sRNA sequencing, and 154 target genes of miRNAs were predicted *via* degradome sequencing. The induction of miR399 and miR398 and the downregulation of miR156, miR159, miR160, miR171, and miR2111 were observed under Pi deficiency in alfalfa (Li et al., 2018). Several conserved families, such as miR156, miR159, miR166, miR169, miR319, miR395, miR397, miR398, miR399, and miR827 have been reported to be commonly responsive to Pi starvation in various plant species, suggesting that these miRNAs are coordinately involved in the conserved Pi signaling pathways in plants (Sunkar et al., 2012). Of these miRNAs, miR399 has attracted the most attention. miR399 is induced by Pi starvation and regulates Pi homeostasis through Pi acquisition, allocation, and remobilization by downregulating its target gene *PHO2* (Fujii et al., 2005; Chiou et al., 2006; Pant et al., 2008). Phosphate Starvation Response 1 (*PHR1*), PHR1-LIKE 1 (*PHL1*), *MYB2*, and several other Pi starvation-responsive genes act upstream to induce increased miR399 expression (Zeng et al., 2013). Hsieh et al. (2009) suggest that miR169 members are downregulated by LP, and some have PHR1 binding sites in their upstream regions. Li et al. (2016) also show that the promoter regions of miR169c/o/q harbored PHO-like binding sites and/or W-box elements in soybean. miR827 is also specifically induced by Pi starvation and plays a

critical role in regulating Pi homeostasis by downregulating its target gene Nitrogen Limitation Adaptation (*NLA*) in *Arabidopsis* (Hsieh et al., 2009; Kant et al., 2011). miR827 is regulated by *OsPHR2* (an ortholog of *AtPHR1*), and overexpression of miR827 leads to increased Pi contents in leaves by decreasing Pi remobilization from old to young leaves (Wege et al., 2016). Wheat TaemiR408 was induced upon Pi starvation, and tobacco lines with TaemiR408 overexpression showed increased P accumulation upon Pi deprivation, suggesting the role of miR408 in improving plant Pi uptake (Bai et al., 2018). Khandal et al. (2020) show a regulatory role for the miR397b-*LAC2* module in root lignification during phosphate deficiency. However, some Pi-responsive miRNAs are species-specific and may participate in exclusive pathways by regulating specific target genes. It suggests that identification of species-specific miRNA and the corresponding targets would be meaningful for improving the ability to adapt to Pi deficiency.

Betula luminifera is a broadleaf tree species belonging to the Betulaceae family that is indigenous to China and naturally distributed in more than 14 provinces as part of the subtropical forest ecosystems in China (Huang et al., 2014; Zhang J.H. et al., 2016). Its desirable wood properties and fast-growing features make it the most common afforestation species in southern China. However, *B. luminifera* frequently encounters conditions of P limitation. Although the total P content in soil is generally high, the abundance of the orthophosphate form (Pi) available to plants is often limited in acidic soil due to adsorption by soil particles, precipitation, or conversion to organic forms (Zeng et al., 2016). In the present study, we aimed to uncover the complex transcriptional and posttranscriptional responses of *B. luminifera* to Pi stress, which might aid the development and selection of trees with increased P use efficiency. First, we generated a comprehensive and integrated dataset that provides insight into the molecular responses of the *B. luminifera* transcriptome using next-generation RNA-Seq. Second, to investigate the miRNA-regulated network response to Pi deficiency, deep sequencing of the small RNAome and degradome were performed to determine the responsive miRNAs and the gene targets of specific miRNAs. Then, these analyses were integrated to elucidate the molecular network of the posttranscriptional response to Pi deficiency in *B. luminifera*. These results will enable future researchers to further elucidate transcriptional and posttranscriptional responses of *B. luminifera* to Pi deficiency, which may help to create plants proficient to offset Pi deficiency.

MATERIALS AND METHODS

Plant Materials and Growth Conditions

The *B. luminifera* clone designated G49$^{\#}$ was propagated *in vitro* in the laboratory at 25 ± 2°C under fluorescent light with a 16-h photoperiod, and the seedlings were grown in Murashige and Skoog (MS) medium without growth regulators. Two-month-old tissue culture seedlings of G49$^{\#}$ were subsequently planted in pots with washed and sterilized sand, with three holes at the bottom of the pot for drainage. Then, seedlings were grown in an incubator with a light intensity of 150 μmol/m^2/s and temperature of 25/20°C (day/night), which were irrigated with modified liquid MS medium containing normal Pi for 2 weeks, and irrigated with deionized water for 2 weeks. Then, 54 uniformly developed plants were divided into two groups, and a completely randomized block design with three replicates was used. For the control (CK) treatment, MS medium contained 1.25 mM KH_2PO_4, without KH_2PO_4 (the only source of Pi) for the phosphate deficiency treatment (i.e., LP), and adjusted the K^+ concentration accordingly by adding K_2SO_4 (Jain et al., 2009; Zeng et al., 2016). The pH of the medium was adjusted to 5.7, and medium was used for irrigation every day. Shoots and roots were harvested after days 3, 7, and 15 of the phosphate deficiency treatment.

Determination of P Concentration, Anthocyanin, and Acid Phosphatase

Fine powder (∼100 mg dry weight) was digested in a mixture of 5 mL 98% H_2SO_4 and 1 mL H_2O_2, and the P content was determined spectrophotometrically at 700 nm based on the molybdenum blue method (Sun et al., 2012). The anthocyanin concentrations in the shoots under Pi starvation were analyzed spectrophotometrically according to the method of Lei et al. (2011). APase activity in the roots and shoots were analyzed using the modified *p*-nitrophenyl phosphate (NPP) method (Liu et al., 2004).

RNA Extraction, Library Construction, and Sequencing

Shoots and roots were sampled after 7 days of the phosphate deficiency treatment, which were used to RNA sequencing, according to Zeng et al. (2016). Specifically, total RNA was isolated using Trizol reagent (Invitrogen, Carlsbad, CA, United States). Total RNA quality and quantity were analyzed using a Bioanalyzer 2100 and RNA 6000 Nano LabChip Kit (Agilent Technologies, Santa Clara, CA, United States) with RNA integrity number (RIN) > 7.0. Under CK and LP treatment, each mixed from three biological replicates were used for library construction and sequencing, thus root RNA was used to construct a root library, and equal quantities of RNA from stems and leaves were pooled to generate a shoot library. For transcriptome sequencing, approximately 10 μg of total RNA was subjected to poly(A) mRNA isolation using poly-T oligo-attached magnetic beads (Invitrogen). Following purification, the mRNA was fragmented into small pieces using divalent cations under elevated temperature conditions. The cleaved RNA fragments were then reverse transcribed to generate the final cDNA library using the mRNA-Seq sample preparation kit (Illumina), and the average insert size for the paired-end libraries was 300 (±50) bp. Paired-end sequencing was then performed on an Illumina Hiseq2500 at the facilities of LC-BIO (Hangzhou, China).

Approximately 1 μg of total RNA was used to construct sRNA library using a TruSeq Small RNA Sample Prep Kit in accordance with the manufacturer's protocol (Illumina, San Diego, CA, United States). Single-end sequencing (36 bp) was then performed on an Illumina HiSeq2500 at the facilities of LC-BIO (Hangzhou, China).

Sequence Data Analysis

For transcriptome data, prior to assembly, the low-quality reads, including reads containing sequencing adaptors and sequencing primers, nucleotides with quality score <20 were removed, and cleaned sequences were obtained. Sequencing reads were assembled *de novo* using Trinity software with the default parameters. All unigene sequences were compared against the protein databases (Nr, SwissProt, KEGG, COG) using BLASTx (*E*-value < 0.00001). Gene ontology (GO) functional annotation was obtained from SwissProt annotation. To investigate the expression level of each unigene in all samples, all paired-end reads for each sample were aligned back to the final assembly using Perl scripts in Trinity with the default parameters. The alignment produced digital expression levels for each contig and these were then normalized by RESM-based algorithm using Perl scripts in the Trinity package to obtain RPKM values. Based on expression levels, significantly differentially expressed mRNAs (DE mRNAs) among the four samples were identified by $P \leq 0.05$ and $\log_2$ fold-change ($\log_2$FC) ≥ 1. Clustering of DE mRNAs was performed using the common Perl and R scripts. DE mRNAs were identified using the GOseq R package (Young et al., 2010), which is based on the Wallenius non-central hyper-geometric distribution, with the threshold $P \leq 0.05$. In addition, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of the DE mRNAs was performed using KOBAS (2.0) with false discovery rate (FDR) ≤ 0.05 (Mao et al., 2005).

For sRNA data, the raw reads were subjected to the Illumina pipeline filter (Solexa 3.0), and the dataset was further processed using an in-house program (ACGT101-miR; LC Sciences, Houston, TX, United States) (Li et al., 2010) to remove adaptors, junk, regions of low complexity, common RNA families (rRNA, tRNA, snRNA, and snoRNA), and repeats. As described in our previous report (Zhang J.H. et al., 2016), the filtered sequencing reads were compared to known miRNAs in miRBase 22.0[1] (Kozomara and Griffiths-Jones, 2013), and the remaining sRNA sequences were used to identify potential novel miRNAs based on the criteria for the miRNA definition described by Meyers et al. (2008). miRNA differential expression based on normalized deep-sequencing counts was analyzed by selectively using Fisher's exact test, with the significance threshold set to 0.05. The resulting network was visualized by Cytoscape 3.7.0 (Shannon et al., 2003).

Degradome Library Construction and Target Identification

The putative target genes in *B. luminifera* were predicted using TargetFinder software. Mismatched pairs or single nucleotide bulges were each scored as one and G:U pairs were scored as 0.5. Mismatched and G:U pair scores were doubled within the core segment (nucleotide pairs at positions 2–13) and score ≤ 4 (Jones-Rhoades and Bartel, 2004). The clean sequences of the high-throughput degradome sequencing were compared to the integrated transcriptome dataset. Then, identification and classification of the sliced miRNA target categories were performed according to the CleaveLand 3.0 pipeline (Charles Addo-Quaye et al., 2008).

To better elucidate the potential miRNA regulatory roles, we subjected all identified target genes to GO analysis. GO annotations of all targets were performed using AgriGO[2] (Du et al., 2010). Differentially expressed miRNAs (DE miRNAs) with their target lists were investigated for cognate mRNA targets in their respective DE mRNA list to delineate miRNA–mRNA functional interactions using ACGT101-CORR1.1, as described previously (Zhang G. et al., 2016).

Validation of MiRNA Target Genes

5′RLM-RACE verification was performed to further validate the target genes, using the GeneRacer Kit (Invitrogen) as described previously (Sunkar et al., 2005). Total RNA was extracted as described above, and the primers used are listed in **Supplementary Table 1**. The PCR products were cloned, and at least 10 independent clones were sequenced to analyze the cleavage sites.

Transient overexpression of *MIR169* precursors in *B. luminifera* leaves was conducted to validate the corresponding target genes. Briefly, the agrobacterium GV3101 containing *35S:MIR169a* and *35S:MIR169c* was overnight propagated, collected, and resuspended in mixture containing 10 mM $MgCl_2$, 10 mM MES, pH 5.6, and 150 μM acetosyringone, respectively. Then agrobacterium suspension was injected into three-week-old *B. luminifera* leaves. Leaves were harvested after 3 days of injection, then the expression levels of *NFYA1* and *NFYA10* were determined by qRT-PCR.

Verification by qRT-PCR

Total RNA was extracted from roots, stems, and leaves after 3, 7, and 15 days under CK and LP conditions as described above. RNA from roots was used to detect the expression patterns of genes and miRNAs, and equal quantities of RNA from stems and leaves were pooled to monitor their expression levels. qRT-PCR of genes was performed using a PrimeScript RT Reagent Kit and SYBR Premix EX Taq Kit (TaKaRa Biotechnology), in accordance with the manufacturer's instructions. The expression levels of target genes were normalized relative to that of Malate Dehydrogenase (*MDH*, KP245810) and α-tubulin (*TUA*, KP245813) in roots and shoots, respectively. The primers of target genes and their reference genes are listed in **Supplementary Table 2**. Reactions were performed using a CFX96 Real-time PCR Detection System (Bio-Rad, Hercules, CA, United States) with three replicates per sample. After PCR, dissociation curves and 2% agarose gel electrophoresis were performed to verify the specificity of amplification.

miRNA qRT-PCR was performed using the RNA as the template and Mir-X miRNA qRT-PCR SYBR Kits (Takara-Clontech, Shiga, Japan), in accordance with the manufacturer's instructions. The forward primers of selected miRNAs were designed based on the mature miRNA sequence, as listed in **Supplementary Table 3**, and the reverse primer was provided with the kit. All expression levels were

[1] http://www.mirbase.org

[2] http://bioinfo.cau.edu.cn/agriGO/

normalized relative to that of U6 (forward primer: 5′-TCGGGGACATCCGATAAAATTGGAA-3′, reverse primer: 5′-GGACCATTTCTCGATTTATGCGTGTCA-3′), and the reactions were performed as described above.

Statistical Analysis

Statistical tests were performed using SPSS 17.0 (SPSS Inc., Chicago, IL, United States). Data were considered to be significantly different at $P < 0.05$ in the *t*-test. Correlation analysis between the miRNAs and target genes was conducted using bivariate correlation in SPSS.

RESULTS

Pi Deprivation Alters Physiological Characteristics

After day 7 of Pi starvation, the P concentration was decreased significantly both in shoots and roots, and the anthocyanin content in shoots was increased significantly. However, APase activity in roots and shoots increased significantly under Pi deprivation. The seedlings of *B. luminifera* showed attenuated growth and increased root biomass after day 15 of Pi starvation (**Figure 1**).

The Pi Deficiency-Responsive Transcriptome in *Betula luminifera*

Overall, more than 216 million raw reads were obtained (GSM4105637–GSM4105640). After removing reads that failed to meet quality control standards, valid reads were used for *de novo* assembly, resulting in 154,340 transcripts and 76,899 unigenes. Unigene expression levels were calculated using transcripts per million (TPM). Under conditions of Pi starvation, 4,002 DE mRNAs were upregulated and 4,093 DE mRNAs were downregulated in roots, whereas 2,617 and 2,967 DE mRNAs were upregulated and downregulated, respectively, in shoots (**Figure 2A**). A total of 1,467 DE mRNAs were common to roots and shoots (**Figure 2B**).

Gene ontology enrichment analysis showed that 26 and 38 GO terms were significantly enriched in roots and shoots, respectively, at a threshold of $P < 0.001$. In roots, the three most enriched GO terms were "nucleus" (GO: 0005634), "chloroplast" (GO:0009507), and "response to salt stress" (GO:0009651). However, the three most enriched GO terms in shoots were distinct from those enriched in roots, i.e., "extracellular region" (GO:0005576), "DNA binding" (GO:0003677), and "cell cycle" (GO: 0007049).

Kyoto Encyclopedia of Genes and Genomes enrichment analysis showed that 10 and 23 terms were significantly enriched in roots and shoots, respectively, with the threshold of $P < 0.05$ (**Supplementary Figure 1**). In roots, the most enriched KEGG terms were "plant hormone signal transduction" (ko04075), "glycosphingolipid biosynthesis – lacto and neolacto series" (ko00601), and "ubiquitin mediated proteolysis" (ko04120) in roots. A total of 229 DE mRNAs involved in signaling of various hormones, including auxin, ethylene, abscisic acid (ABA), and jasmonic acid, were identified in Pi-deficient roots, of which 126 DE mRNAs were downregulated. Nineteen auxin-related genes were found to be responsive to Pi

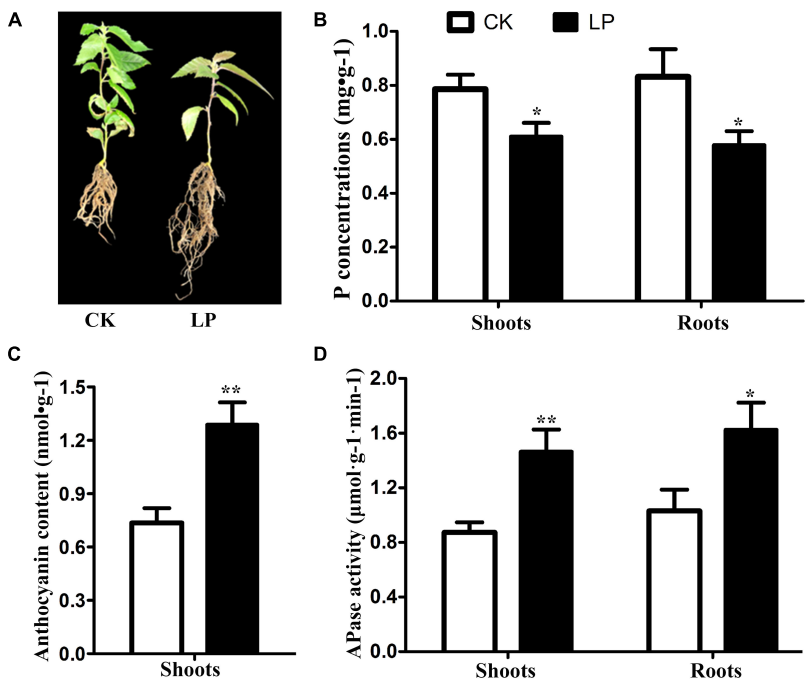


FIGURE 1 | Changes in phenotype **(A)**, P concentration **(B)**, anthocyanin content **(C)**, and APase activity **(D)** in *B. luminifera* after 7 days of Pi starvation. Bars indicate means ± SE ($n = 3$). *P*-values were obtained from *t*-tests between LP and CK conditions. $^{**}P < 0.01$; $^{*}P < 0.05$.

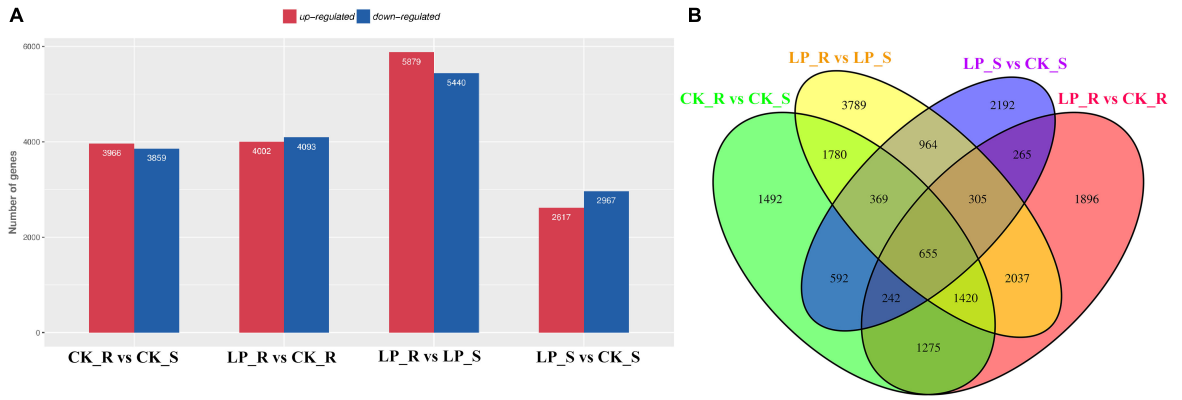


FIGURE 2 | Expression profiling of –Pi-responsive DE mRNAs in *B. luminifera*. **(A)** The numbers of DE mRNAs in roots and shoots after 7 days of Pi starvation compared with control plants were determined by $P \leq 0.05$ and $\log_2 FC \geq 1$. Red and blue columns represent upregulated and downregulated DE mRNAs, respectively. **(B)** Venn diagrams showing the numbers of overlapping and specific DE mRNAs obtained across four comparisons. R, roots; S, shoots.

starvation, with six genes showing repression and 13 genes showing induction. However, the most enriched KEGG terms were "phenylpropanoid biosynthesis" (ko00940), "flavonoid biosynthesis" (ko00941), and "starch and sucrose metabolism" (ko00500) in shoots (**Supplementary Figure 1B**). Totals of 121 and 46 DE mRNAs were involved in phenylpropanoid and flavonoid biosynthesis, respectively, and more than half of the DE mRNAs were repressed under conditions of Pi starvation.

To validate the expression levels identified by RNA-seq, 16 differentially expressed genes were arbitrarily selected for expression analyses after 3, 7, and 15 days of LP treatment using qRT-PCR. The qRT-PCR results showed that the expression levels of the examined genes, with the exception of *PER57* in shoots, and *WRKY41* in roots, exhibited similar trends as those seen the RNA-seq results after 7 days of LP treatment (**Supplementary Figure 2**).

Analysis of sRNA Populations Under Conditions of Pi Deficiency

A total of approximately 47 million raw reads were obtained from sRNA libraries (GSM4105633–GSM4105636). After removing reads that failed to meet quality control standards, there were 31,980,253 redundant reads (10,503,545 non-redundant reads). The sequences were aligned to known RNAs in RFam, mRNA, Repeat-repbase, and other known RNA sequences; 16,203,000 redundant reads (9,688,171 non-redundant sequences) were processed for further analysis.

For the redundant sequences, there were major peaks at 24 nucleotides (nt) in the shoot libraries. Unexpectedly, a major peak at 21 nt was observed in the roots under LP conditions (**Figure 3A**). For the non-redundant sequences, an overwhelming peak was seen at 24 nt in the shoot sRNA libraries (**Figure 3B**), while the sRNA length was preferentially 18–21 nt under LP conditions. Further analysis showed that tRNA ratios with sizes of 18 and 19 nt were significantly increased in LP-treated roots, especially for 18 nt, whereas the corresponding ratios were significantly decreased in LP-treated shoots (**Supplementary Table 4**).

Identification of Pi Deficiency-Responsive MiRNAs in *Betula luminifera*

A total of 164 known miRNAs from 56 families were identified based on a homology search for known miRNAs using

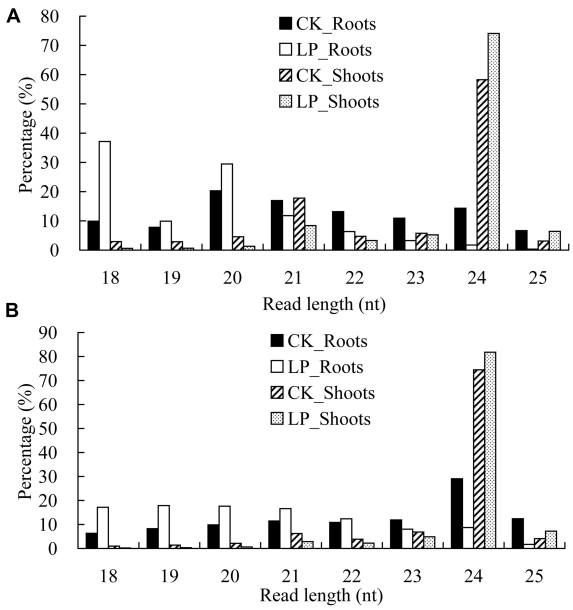


FIGURE 3 | Global profiling of processed small RNAs (sRNAs) in *B. luminifera*. **(A)** Frequencies of processed sRNAs are expressed as percentages of the total number of redundant sequences. **(B)** Frequencies of processed sRNAs are expressed as percentages of the total number of non-redundant sequences. Four samples including CK_Roots (black bars), LP_Roots (white bars), CK_Shoots (striped bars), and LP_Shoots (dotted bars) in roots and shoots under normal and phosphate-deficient (–Pi) conditions.

miRBase (release 22.0) as a reference set. Of these miRNAs, 82 were identified unambiguously, as these miRNAs had corresponding precursor sequences, whereas 82 miRNAs without corresponding precursor sequences were identified ambiguously (**Supplementary Table 5**). These miRNAs were named using the initials of the species (*blu* for *B. luminifera*) followed by the number (and letter when applicable) assigned to the same miRNA in other species. Furthermore, 19 novel *MIRNA* genes were identified with both 5′ and 3′ arms, and 21 were considered as candidate miRNAs with 5′ or 3′ arms (**Supplementary Table 6**).

Totals of 66 miRNAs were differentially expressed in Pi-deficient roots, including 20 that were upregulated and 20 that were downregulated, as well as 9 and 17 miRNAs that were specific to LP-treated and control roots, respectively (**Supplementary Table 7** and **Supplementary Figure 3A**). Sixty miRNAs were differentially expressed in Pi-deficient shoots, consisting of 20 that were upregulated and 33 that were downregulated, and five and two miRNAs were specific to LP-treated and control shoots, respectively (**Supplementary Table 8** and **Supplementary Figure 3B**). Of these DE miRNAs, 17 shared the same expression trends between roots and shoots, whereas six miRNAs showed the opposite expression patterns between roots and shoots. Furthermore, 80 of the DE miRNAs were Pi-responsive and tissue-specific (43 and 37 in roots and shoots, respectively). Specifically, in the −Pi roots, miR1511a/b, miR399c, miR397a/b/c, miR398a/b, miR858, and miR169b were induced, whereas miR319b-5′, miR530a, miR169c, miR396b, and miR164 were repressed. In the −Pi shoots, miR397a/b, miR398b, miR156i, miR408a, miR396d-3′, and six novel miRNAs were upregulated, whereas miR319a-5′, miR169a/b, miR159c, miR395b, miR828b, and miR172a were downregulated.

To validate the expression patterns identified by sRNA sequencing, 20 differentially expressed miRNAs after 3, 7, and 15 days of Pi deficiency were arbitrarily selected for expression analyses using qRT-PCR. The qRT-PCR results showed that the expression of most examined miRNAs, except for miR828b and miR858 in shoots, and miR482b in roots, exhibited similar trends after 7 days of Pi deficiency as those revealed by sRNA-seq (**Figure 4**).

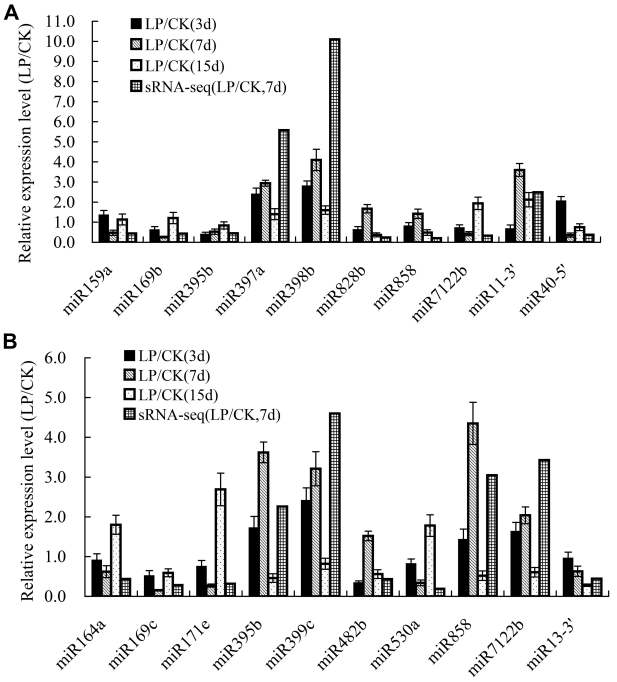


FIGURE 4 | Validation of the expression profiles of miRNAs in *B. luminifera* shoots and roots identified by sRNA-seq using qRT-PCR. **(A)** Relative expression of ten LP-responsive miRNA in shoots after 3, 7, and 15 days of LP treatment; **(B)** Relative expression of ten LP-responsive miRNA in roots after 3, 7, and 15 days of LP treatment. Expression level is represented by the ratio of LP treatment to CK. The value of relative expression level above one indicates that miRNAs are upregulated by Pi starvation, and the value below one shows that miRNAs are downregulated by Pi starvation. The experiments were repeated three times. Error bars indicate standard deviation.

Target Genes Regulated by DE MiRNAs

In roots, 579 unigenes were predicted as targets of 60 DE miRNAs using the TargetFinder software, consisting of 610 miRNA–target pairs (**Supplementary Table 9**), whereas 407 unigenes were predicted as targets of 51 DE miRNAs in shoots, consisting of 461 miRNA–target pairs (**Supplementary Table 10**). Some miRNAs were predicted to target multiple genes, exemplified by one novel miRNA (blu-miR40-5p) that had 112 potential target genes. To validate these target genes, a mixed degradome library was sequenced, yielding 45,697,220 raw fragments, of which 45,437,576 reads were mapped to referenced transcriptomic data; thus 54,744 transcripts were processed to target identification. Therefore, 109 and 112 miRNA–target pairs were further validated in the roots (**Supplementary Table 9**) and shoots (**Supplementary Table 10**), respectively. The identified targets were further classified into five categories (0–4) according to the relative abundance of tags at the target sites. For representative miRNAs of the five categories and their corresponding targets, red lines indicate the cleavage site of each transcript in **Supplementary Figure 4**. These targets of known miRNAs were mainly transcription factors (**Table 1**), suggesting that these miRNAs may play important roles in gene regulation. In addition, many target genes involved in specific pathways were also identified, for example, ATP SULFURYLASE 1 (*APS1*) by miR395b, and six *LAC* genes by miR397 (**Table 1**). We also identified a series of novel target genes that were targeted by isomiRNAs with lengths of 18 or 19 nt, which were only differentially expressed in roots, including a PHOSPHOLIPASE D DELTA (*PLDD*) and one unknown gene by miR399c (19 nt), seven targets including *WRKY24* by miR395c-5p (18 nt), as well as seven targets by miR482-5′ (**Supplementary Table 9**).

We then performed 5′RLM-RACE for five of the identified miRNA targets for validation; one target of miR164, two potential targets of miR169, and two targets of miR397 were confirmed (**Figure 5A**). Furthermore, two *NFYA* genes were cleaved by miR169, as shown in a transient assay, in which the relative expression levels of *NFYA1* and *NFYA10* was decreased in the *B. luminifera* leaves transiently overexpressing *BlMIR169a* and *BlMIR169c* (**Figure 5B**).

TABLE 1 | Main target genes of DE miRNAs in *B. luminifera*.

miRNA	Targets	miRNA	Targets
miR156	*SPL14*	miR319	Two *TCP* genes
miR159	*GAMYB*	miR396	Four *GRF* genes
miR160	*ARF18*	miR536, blu-miR40-5p	Two zinc finger protein genes
miR164	*NAC21*	miR395b	*APS1*
miR166	five *HD-ZIP* genes	miR397	Six laccases
miR169	two *NFYA* genes	miR482b	Two disease resistance proteins
miR171	*SCL6*	miR828, miR858	Six *MYB* genes
miR172	three *AP2* genes		

Gene ontology enrichment analysis of targets for DE miRNAs revealed that 11 categories including "oxidoreductase activity, oxidizing metal ions," "transcription factor activity, sequence-specific DNA binding" and "ATP binding" pathways responded to −Pi stress in roots; 17 GO categories, such as "transcription factor activity, sequence-specific DNA binding," and "DNA binding," and "cell differentiation" were enriched for targets of DE miRNAs in shoots (**Supplementary Table 11**). KEGG pathway enrichment analysis showed that "ascorbate and aldarate metabolism" was enriched for targets of DE miRNAs in both of roots and shoots, whereas "circadian rhythm – plant" was specifically enriched in shoots (**Supplementary Table 12**).

Correlation Between DE MiRNAs and DE mRNAs of *Betula luminifera* in Response to LP Treatment

To determine the complex relationship between miRNA regulation and transcriptional response to Pi starvation, miRNA–mRNA interactions were predicted using sRNA-Seq, RNA-Seq, and degradome-Seq data. A total of 109 miRNA–target pairs were analyzed in roots, involving 38 DE miRNAs and 101 target genes validated by degradome-Seq. Thus, 30 pairs, including 14 DE miRNAs and 28 DE mRNAs, were identified, of which 14 miRNA–mRNA pairs showed opposite expression patterns under LP conditions (**Table 2**). For example, miR169c was downregulated significantly with $\log_2$FC of −1.81, and the target gene *NFYA10* was significantly induced with $\log_2$FC of 2.30 under –Pi conditions, which was predicted to participate in aminoacyl-tRNA biosynthesis (ko00970).

In shoots, 112 pairs involving 38 DE miRNAs and 87 target genes were subjected to further analysis. Thus, 23 pairs including 13 DE miRNAs and 16 DE mRNAs were identified, of which 16 miRNA–mRNA pairs showed a negative correlation under Pi starvation (**Table 2**). For example, miR159a was significantly repressed with $\log_2$FC of −1.18, and the corresponding target gene GAMYB was significantly induced with $\log_2$FC of 1.25 under LP conditions. Three miR397 members were upregulated, and four *Laccase* genes were significantly repressed. Unexpectedly, only blu-miR40-5′ and miR319b-5′ were common to roots and shoots, and were downregulated under LP conditions, whereas the corresponding target E3 ubiquitin-protein ligase MIEL1-like gene and one putative mitochondrial protein gene were induced. KEGG analysis showed that E3 ubiquitin-protein ligase MIEL1-like gene participated in "ubiquitin mediated proteolysis" (ko04120).

As summarized in **Figure 6**, the shoots and roots of *B. luminifera* showed unique expression profiles of miRNA-target pairs when subjected to LP conditions. These observations suggest that miRNAs may show distinct responses to Pi starvation in different tissues/organs. Of these miRNA-target pairs, three miR397 members showed negative relationship with four *Laccase* genes in shoots, while miR169c was showed negative correlation with two *NFYA* genes in roots, and one novel miRNA blu-miR40-5p regulated distinct target genes in roots and shoots, respectively. However, some miRNAs showed similar expression patterns with corresponding target genes. For example, miR319a and two targets were inhibited in shoots, and miR160c and three targets were downregulated in roots under Pi starvation.

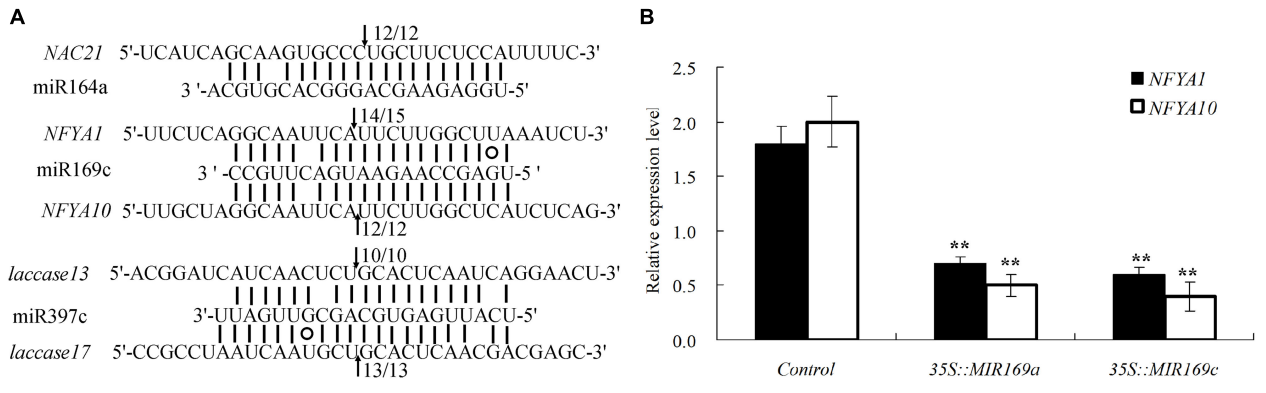


FIGURE 5 | Validation of miRNA target genes. **(A)** Validation of five miRNA targets by 5′ RACE. Matches between miRNAs and their corresponding target genes are indicated by straight lines. G:U wobbles are represented by circles. Arrows indicate cleavage sites, and the clone frequencies are shown. **(B)** Two *NFYA* genes were validated by transiently overexpressing *BlMIR169a* and *BlMIR169c* in *B. luminifera* leaves.

TABLE 2 | Correlation analysis between differentially expressed miRNAs and their target genes under low phosphate conditions.

miR_name	Log2(FC)	Gene_ID	Annotation	log2(FC)
In roots				
blu-miR40-5′	-inf [a]	TRINITY_DN23871_c0_g5	L484_015979 (*Morus notabilis*)	2.75
		TRINITY_DN25037_c1_g1	E3 ubiquitin-protein ligase MIEL1-like (*Malus domestica*)	2.71
		TRINITY_DN29288_c2_g1	Unknown	2.55
miR169b	1.07	TRINITY_DN26121_c0_g1	LOC109003207 (*J. regia*)	−2.61
miR169c	−1.81	TRINITY_DN23675_c2_g2	*NFYA1* (*J. regia*)	3.14
		TRINITY_DN24688_c1_g2	*NFYA10* (*J. regia*)	2.30
miR171e	−1.64	TRINITY_DN28131_c2_g2	*SCL27* (*J. regia*)	1.42
miR172e-5′	−1.16	TRINITY_DN27324_c1_g1	LOC108991439 (*J. regia*)	1.65
miR319b-5′	−3.04	TRINITY_DN28715_c2_g3	Putative mitochondrial protein (*Glycine soja*)	1.11
miR395c-5′	-inf[a]	TRINITY_DN24269_c0_g5	Transmembrane 9 superfamily member 8-like (*J. regia*)	1.04
miR403b	1.74	TRINITY_DN23405_c0_g6	LOC108990185 (*J. regia*)	−1.18
miR482b	−1.21	TRINITY_DN22938_c0_g10	Unknown	3.18
miR482b-5′	inf[b]	TRINITY_DN26208_c1_g4	Serine/threonine-protein kinase At5g01020-like (*J. regia*)	−1.76
miR7122b	1.78	TRINITY_DN27260_c2_g9	Pentatricopeptide repeat-containing protein At1g62914 (*Nelumbo nucifera*)	−1.01
In shoots				
blu-miR40-5′	−1.41	TRINITY_DN25037_c1_g1	E3 ubiquitin-protein ligase MIEL1-like (*M. domestica*)	1.92
blu-miR40-5′	−1.41	TRINITY_DN29286_c0_g3	Sucrose synthase 1 (*B. luminifera*)	1.50
miR159a	−1.18	TRINITY_DN24093_c0_g4	Transcription factor GAMYB (*J. regia*)	1.25
miR159b	-inf[a]	TRINITY_DN24093_c0_g4	Transcription factor GAMYB (*J. regia*)	1.25
miR319b-5′	−2.28	TRINITY_DN28715_c2_g3	Putative mitochondrial protein (*G. soja*)	1.77
miR397a	2.48	TRINITY_DN26138_c0_g6	Laccase 1 (*B. platyphylla*)	−2.67
miR397a	2.48	TRINITY_DN27015_c0_g1	Laccase-17-like (*J. regia*)	−1.54
miR397a	2.48	TRINITY_DN26934_c1_g1	Laccase-17-like (*J. regia*)	−1.83
miR397a	2.48	TRINITY_DN28629_c3_g4	Laccase-7-like (*J. regia*)	−2.37
miR397b	2.05	TRINITY_DN26934_c1_g1	Laccase-17-like (*J. regia*)	−1.83
miR397b	2.05	TRINITY_DN27015_c0_g1	Laccase-17-like (*J. regia*)	−1.54
miR397b	2.05	TRINITY_DN28629_c3_g4	Laccase-7-like (*J. regia*)	−2.37
miR397c	1.86	TRINITY_DN27015_c0_g1	Laccase-17-like (*J. regia*)	−1.54
miR397c	1.86	TRINITY_DN26934_c1_g1	Laccase-17-like (*J. regia*)	−1.83
miR397c	1.86	TRINITY_DN28629_c3_g4	Laccase-7-like (*J. regia*)	−2.37
miR403a	1.50	TRINITY_DN27428_c1_g4	*AGO2* (*J. regia*)	−1.05

-inf[a] indicates that no read was detected in LP_R or LP_SL sample, while the counts were observed in CK_R or CK_SL sample; inf[b] indicates that no read was detected in CK_R sample, while the counts were observed in LP_R sample.

The English in this document has been checked by at least two professional editors, both native speakers of English. For a certificate, please see: http://www.textcheck.com/certificate/Mh4hCj.

Expression Patterns of MiRNAs and Their Target Genes Under Conditions of Pi Deficiency

To confirm the RNA-Seq data and the dynamic correlation between miRNAs and their target genes under conditions of Pi deficiency, the expression changes in eight selected miRNAs and their corresponding target genes after 3, 7, and 15 days of Pi deficiency were assayed using qRT-PCR (**Figure 7**). The trends in expression patterns of the eight miRNAs and target genes obtained were similar between qRT-PCR and RNA-Seq, and an overall negative correlation between the expression of the miRNAs and their corresponding targets was observed. Specifically, in the shoots, miR159a was induced non-significantly after 3 days, and was then downregulated significantly after 7 days, but was again non-significantly induced after 15 days of Pi deficiency. The target *GAMYB* was upregulated after 3 and 7 days, and then downregulated after 15 days. miR397a was induced significantly under LP condition, and the corresponding target *laccase1* was significantly downregulated, showing a significant negative correlation. BlmiR40-5′ was upregulated after 3 days, and then inhibited after 7 and 15 days of Pi deficiency, and the target *MIEL1-like* showed a negative expression pattern. In the roots, miR169c was inhibited under LP condition, and *NFYA1* was significantly upregulated, showing a significant negative correlation; miR171e was downregulated after 3 and 7 days of Pi deficiency, while the expression level was induced significantly after 15 days of Pi deficiency, and the target *SCL27* showed the opposite expression pattern; three miRNAs including miR7122b, miR395b, and miR399c showed similar expression pattern, exemplified by they were upregulated after 3 and 7 days of Pi deficiency, but were inhibited after 15 days, whereas the corresponding target *PPR*, *APS1*, and *PLDD* were downregulated after 3 and 7 days of LP treatment, of

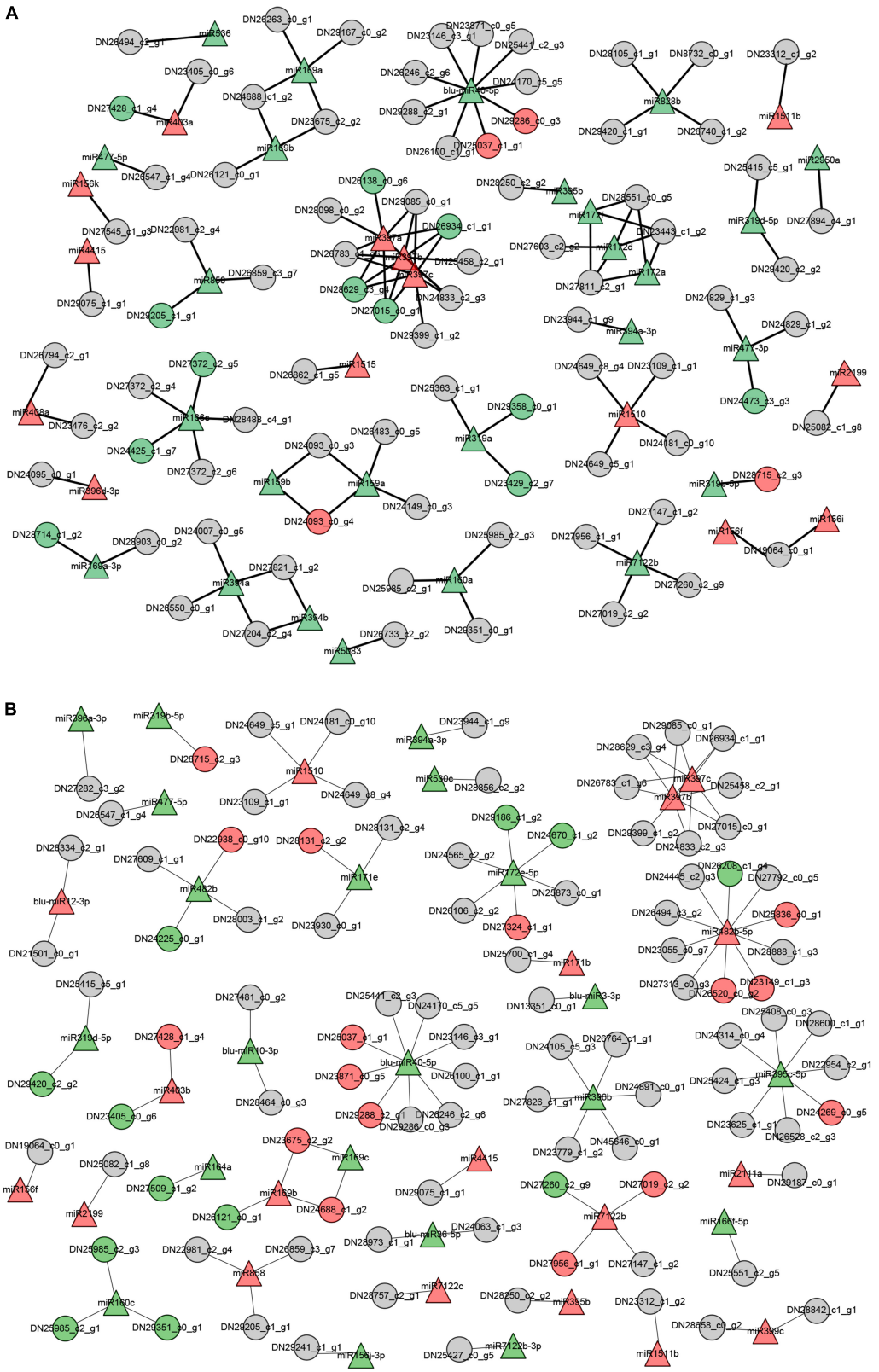


FIGURE 6 | The putative functional networks for –Pi-responsive miRNAs and their corresponding target genes in *B. luminifera*. The network depicts the relationship of –Pi-responsive miRNAs and target genes from their expression patterns upon Pi deficiency in shoots **(A)** and roots **(B)**. Triangle indicates miRNAs and circles are target genes. Red represents upregulated, green indicates downregulated, and gray means unchanged under Pi starvation.

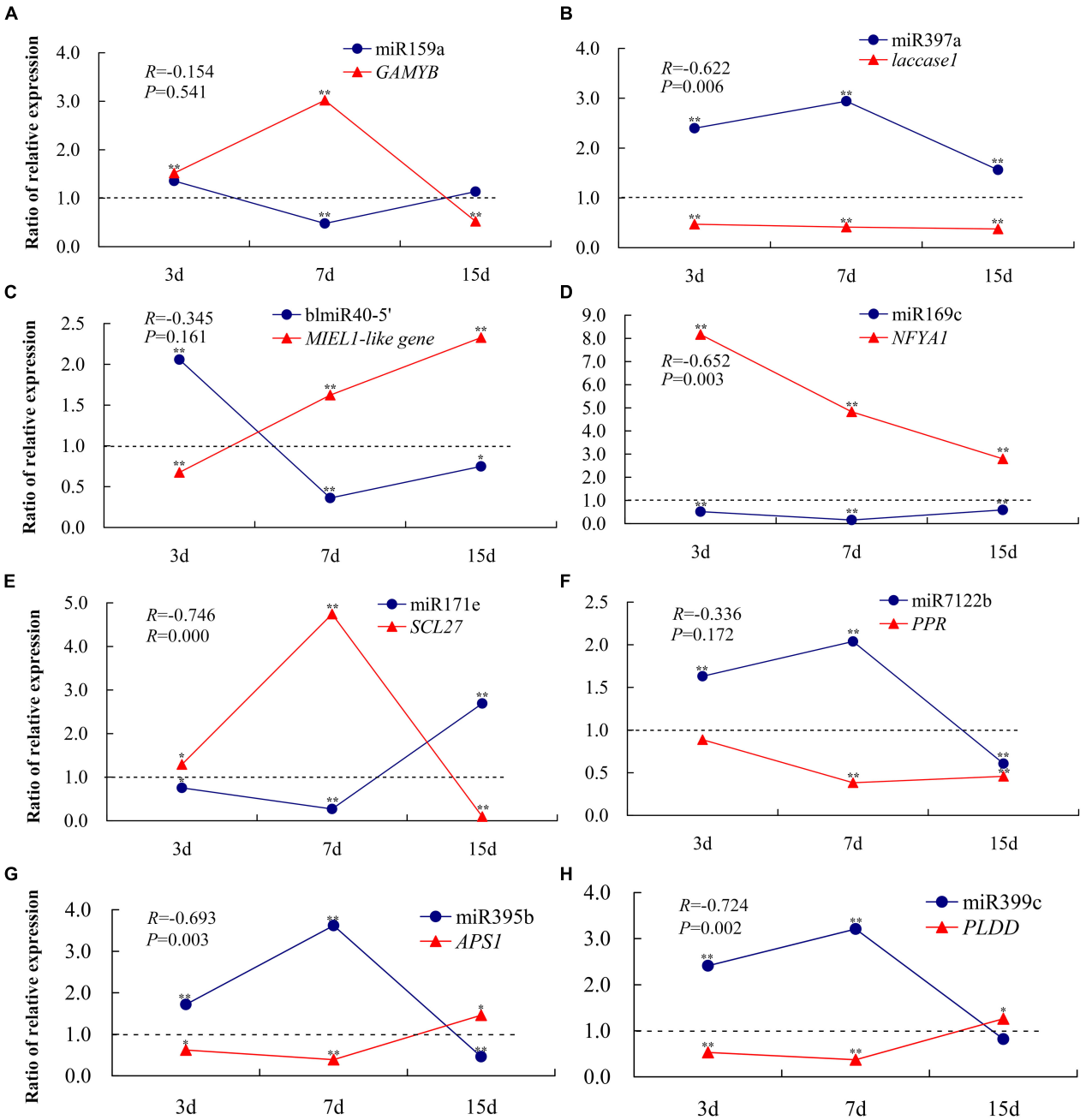


FIGURE 7 | qRT-PCR-derived expression analyses of eight miRNAs and their target genes under –Pi stress. To show the expression patterns of miRNAs and their corresponding target genes more directly, the ratio (K value) is used in the figure instead of individual relative expression levels. K is the ratio of the relative expression levels of LP (–Pi treatment) and CK (normal treatment); i.e., K = LP/CK. Each experimental time has a corresponding control group. A K-value > 1 indicates upregulated expression compared to the control, while a K-value < 1 indicates downregulated expression compared to the control. *P < 0.05; **P < 0.01. The expression levels of miR159a-*GAMYB* **(A)**, miR397a-*laccase1* **(B)**, and BlmiR40-5′-*MIEL1-like* **(C)** were detected in the shoots, and miR169c-*NFYA1* **(D)**, miR171e-*SCL27* **(E)**, miR7122b-*PPR* **(F)**, miR395b-*APS1* **(G)**, and miR399c-*PLDD* **(H)** in the roots were measured. All expression levels were normalized to *TUA* and *MDH* in shoots and roots, respectively. The experiments were repeated three times. Error bars indicate standard deviation.

which miR395b and miR399c were significantly and negatively correlated with their target genes.

Furthermore, a putative model has been proposed (**Figure 8**). Several miRNA-target pairs showed common and unique expression profiles in the shoots and roots, when the seedlings of *B. luminifera* encounter the Pi deficiency stress. Specifically, the expression of miR169a/b and miR169c were inhibited in shoots and roots, respectively, while their corresponding target *NFYA1/10* were induced; miR395b was downregulated and was induced in shoots and roots, respectively, while the target

APS1 showed negative relationship in roots; miR397a/b/c were upregulated, and their targets *Laccase 1/7/17* were inhibited in shoots; miR399c was induced in roots, and target *PLDD* was downregulated upon Pi deficiency stress.

DISCUSSION

Plants, particularly forestry species, frequently face LP availability in soils and have thus evolved a series of responses and adaptive mechanisms to cope with conditions of Pi deficiency. Understanding the underlying molecular mechanisms is important for selecting and/or producing P-efficient cultivars for modern forestry. The miRNA-mediated posttranscriptional regulation of plant resistance to nutritional deficiency has been described in several plants. In contrast, the molecular mechanisms underlying the responses of forestry trees to Pi deficiency and the posttranscriptional regulation triggered by miRNAs are poorly understood. Based on miRNA, degradome sequencing, and RNA-Seq analyses, we investigated the regulatory roles of miRNAs in roots and shoots of *B. luminifera* under conditions of Pi deficiency. Our results indicated that the genes involved in miRNA-mediated posttranscriptional regulation play an important role in the root and shoot responses of *B. luminifera* to Pi deficiency stress. However, our findings provide only initial insights into this topic, and a great deal of further effort will be required to gain a full understanding of the posttranscriptional regulation related to Pi deficiency.

sRNA Length Patterns and Potential Roles in *Betula luminifera*

Many annual herbaceous plants, such as *A. thaliana* and *Oryza sativa*, contain substantially more 24-nt sRNAs than 21-nt sRNAs (Jones-Rhoades et al., 2006; Morin et al., 2008a), whereas the major peak of redundant sRNAs occurs at 21 nt in woody plants, such as Chinese white poplar (Chen et al., 2012), and *B. luminifera* (Zhang J.H. et al., 2016). However, further studies have shown that sRNA compositions are dependent on developmental stages, organs, and/or external stressors (Slotkin et al., 2009; Zhang et al., 2013; Pan et al., 2017). We found that the length distribution pattern of sRNAs in roots differed significantly from that of shoots, as exemplified by the major peak of redundant sequences at 20 nt in CK roots, in contrast to the major peak at 18 nt and minor peak at 20 nt in LP roots (**Figure 3A**). Pei et al. (2013) suggested that the abundance of sRNAs may be related to Pi tolerance, as exemplified by the lower relative abundances of 21-nt and 24-nt sRNAs in the LP-tolerant mutant 99038 compared to wild-type Qi319 maize under LP conditions. Further analysis of the components of sRNAs with lengths 18 nt and 19 nt revealed that a large proportion of 18–20-nt sequences originated from tRNAs. The tRNA ratios with sizes of 18 and 19 nt were significantly increased in LP-treated roots, especially for 18 nt, whereas the corresponding ratios were significantly decreased in LP-treated shoots (**Supplementary Table 4**). Previous studies revealed the accumulation of tRNA fragments in roots of P-starved *Arabidopsis* (Hsieh et al., 2009) and in nitrate-starved seedlings (Liang et al., 2012). Hsieh et al. (2009) reported that the 19-nt RNAs from Gly-tRNA^{TCC} and Asp-tRNA^{GTC} corresponding to the 5′ end of tRNA were highly accumulated in roots but were present at much lower levels in shoots, indicating a spatial and temporal expression pattern of sRNAs derived from specific cleavage of tRNAs rather than random degradation. Several studies have indicated that tRNA-derived sRNAs accumulate in specific tissues or under specific stress conditions (Thompson et al., 2008; Fu et al., 2009). Furthermore, Zhang et al. (2009) reported that tRNA halves detected in phloem sap of pumpkin (*Cucurbita maxima*) inhibited translational activity *in vitro*, suggesting that these tRNA halves may act as long-distance signals to coordinate the metabolic status between source and sink tissues. Recent studies have shown that tRNA fragments are incorporated into Argonaute (AGO) complexes, and are likely to regulate gene expression post-transcriptionally in a manner similar to miRNAs (Loss-Morais et al., 2013).

Apart from tRNA fragments, the isomiR accumulation was observed in our study. It exemplified by several isomiRs with a length bias of 18 or 19 nt were only differentially expressed in roots and had more diversified targets, such as two targets for miR399c, eight targets for miR395c-5′, and 10 targets for miR482-5′ (**Supplementary Table 9**). Recent evidence suggests that isomiR production is differentially modulated at different developmental stages and by adverse environmental signals (Hackenberg et al., 2013; Balyan et al., 2020; Fard et al., 2020). For example, Shao et al. (2015) found that the abundance of iso-osa-miR528-5p in AGO1 complexes was higher than that of its canonical osa-miR528-5p in rice. The 19 nt isomiR of pvu-miR171a was exclusively detected in flowers, whereas the 20 nt isomiR of pvu-miR479 was highly expressed in roots (Pelaez et al., 2012). The isomiRs that are shorter than their canonical miRNAs could have been subjected to nucleotide truncation by exonucleolytic digestion (Lu et al., 2009), or they could simply have been synthesized to a shorter length compared with their dominant miRNAs (Morin et al., 2008b). The trimming of the 3′ end shortens the length of canonical miRNAs and increases their stability in AGO protein complexes (Ameres et al., 2010). Subtle differences in miRNA sequences can affect their AGO sorting, even for isomiRs derived from the same miRNA precursor (Iki et al., 2018); the alteration in the lengths of isomiRs could therefore be a cellular strategy to diversify miRNA target specificity by differential loading into AGO complexes. On the other hand, different isomiRs may regulate specific targets. For example, Naya et al. (2010) showed that the 20 nt isomiR of miR156 cleaved a new WD40-like protein target. Therefore, the diversified target genes triggered by the isomiR accumulation might be an important mechanism in response to Pi deficiency in *B. luminifera.*

Divergent Expression of MiRNAs Between Roots and Shoots Under Conditions of Pi Deficiency

There is accumulating evidence that miRNA-mediated genes are involved in basic metabolic processes, root and shoot development, stress responses, and Pi uptake (Li et al., 2018).

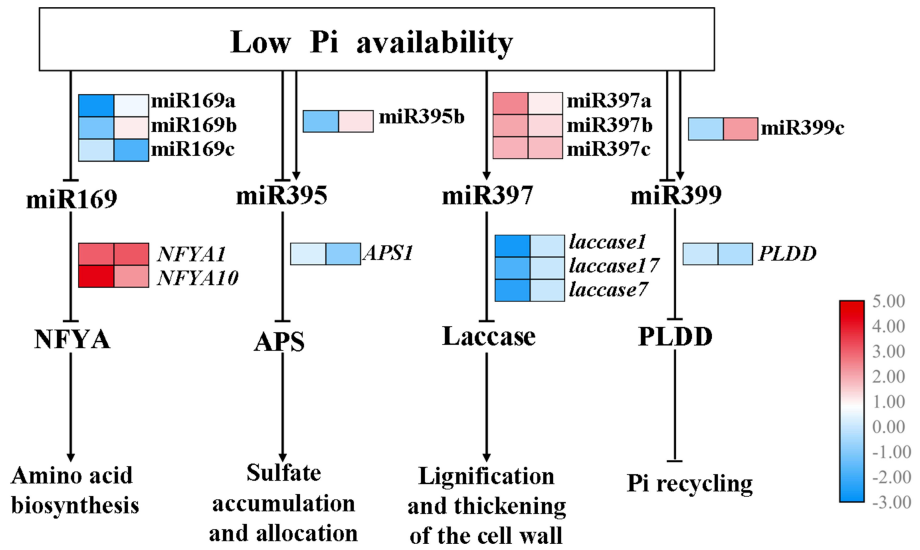


FIGURE 8 | Hypothetical model for the functions of miR169, miR395, miR397, and miR399 in response to Pi starvation. miR169a/b/c, miR395b, miR397a/b/c, and miR399c showed common and unique expression profiles in the shoots and roots when subjected to low Pi availability, then specifically inhibited expression of various target genes, thus affecting downstream biological processes. Arrows and blunted lines indicate positive and negative interactions, respectively. The left and right columns represent the log2(LP/CK) of the transcriptional levels in shoots and roots, respectively.

However, many studies have reported that changes in the expression of genes, including *MIRNA* genes, were organ-dependent (Aceituno et al., 2008; Xu et al., 2013). Aceituno et al. (2008) reported that 10-fold more genes were differentially expressed between organs as compared to any other experimental variable. Xu et al. (2013) showed that a series of miRNAs showed leaf- or root-specific expression under Pi-sufficient and -deficient conditions in soybean. Similar phenomena were observed in cucumber, with some miRNAs preferentially expressed in certain tissues (Mao et al., 2012). In the present study, 43 and 37 miRNAs were specifically differentially expressed in roots and shoots, respectively (**Supplementary Figure 3**). Therefore, the divergent expressions of miRNAs were detected among different organs in response to Pi deficiency. Further GO enrichment analysis of targets for DE miRNAs showed that –Pi responsive miRNA-target pairs were mainly classified into transcription regulation, oxidoreductase activity, ATP binding, copper ion binding, defense response to bacterium, lignin biosynthetic process, and leaf development categories. It suggested that these miRNA-target pairs trigger removing of H_2O_2 because of Pi stress, activating defense response, and regulating growth. Similar phenomenon was observed in maize, which is exemplified by miRNA-target pairs alter the morphology, physiology or metabolism of plants upon prolonged Pi deficiency (Li et al., 2016).

The putative functional networks of miRNA-target pairs presented that miRNAs show distinct responses to Pi starvation in different tissues/organs (**Figure 6**). A large proportion of DE miRNAs showed stress-related expression. For example, miR397 and miR398 were upregulated in both roots and shoots of Pi-starved *B. luminifera*. This is in contrast to studies in soybean roots where miR397a was induced by −Pi conditions in soybean roots, whereas miR398a, miR398b, miR398c, and miR408 were repressed in Pi-depleted leaves (Xu et al., 2013); miR397a, miR397b, miR398a, and miR398b were co-downregulated in both the leaves and roots of maize (Li et al., 2016), indicating divergent roles for these stress-related miRNA families between plant tissues and/or species. In *B. luminifera*, three miR397 members showed significant negative correlations with four *laccase* genes under −Pi conditions in shoots, but not in roots (**Table 2**). Previous studies have shown that *laccase* genes are related to lignification and thickening of the cell wall in secondary cell growth (Constabel et al., 2000). Khandal et al. (2020) aslo showed a regulatory role for the miR397b-*LAC2* module in root lignification during Pi deficiency. Therefore, we speculate that miR397 may mainly regulate *laccase* genes in shoots, and increased miR397 expression may contribute to the decreased *laccase* mRNA levels, participating in the shoot lignification and thickening of cell wall under −Pi conditions in *B. luminifera.*

miR169 members showed differential changes in expression in roots and shoots of maize under LP conditions, as four miR169 members (miR169f/g/h/l) were upregulated in leaves but miR169a was repressed only in roots after 4 days of treatment (Li et al., 2016). In the present study, miR169b was induced and miR169c was repressed in roots, whereas miR169a/b was inhibited in shoots. Similarly, soybean miR169c/r were repressed both in leaves and roots, while miR169q was upregulated in leaves under LP conditions (Xu et al., 2013). Previous studies showed miR169-*NYFAs* participate in nitrogen and drought stress responses (Zhao et al., 2011). In *B. luminifera*, miR169c was downregulated in −Pi-treated roots, and the corresponding targets *NFYA1* and *NFYA10* were significantly induced (**Table 2**). *NFYA10* was predicted to participate in aminoacyl-tRNA biosynthesis. Previous studies have shown

that aminoacyl-tRNA synthetases (aaRSs) ensure the fidelity of translation of the genetic code, covalently attaching appropriate amino acids to the corresponding nucleic acid adaptor molecule–tRNA (Klipcan and Safro, 2004). The aminoacylation reaction catalyzed by aaRSs is the most fundamental step for amino acid biosynthesis. In addition, Pant et al. (2008) indicated that miR169 is a potential long-distance signal that can report shoot Pi status to the roots, similar to the role of miR399. Therefore, miR169-*NFYAs* might have a multifunctional role in response to Pi starvation in *B. luminifera*.

Previous studies showed that miR395 was induced by low sulfate treatment, involved in the regulation of sulfate accumulation and allocation by targeting *APS* genes (Liang et al., 2010). miR395b was specifically induced in −Pi-treated roots, and the target *APS1* was downregulated in *B. luminifera* (**Figure 7**). Several studies have indicated that miR395 and miR399 are transported from shoots to roots *via* the phloem (Buhtz et al., 2010), which it was consistent with the decreasing level of miR395b in *B. luminifera* shoots. Therefore, miR395 may act as a mobile signal and travel long distances to induce systemic silencing of *APS* and other targets, thus contributing to adaptation to −Pi environments in *B. luminifera*.

miR399 is responsive to Pi starvation, which demonstrates its important regulatory role in Pi homeostasis (Fujii et al., 2005; Chiou et al., 2006; Pant et al., 2008). Four miR399 members were identified in our library, which cleave *PHO2* mRNA based on our degradome data. However, only miR399c, a 19 nt isomiR that contains two truncated nucleotides at the 3′ end of canonical miR399a, was induced significantly in LP-treated *B. luminifera* roots. A similar phenomenon was observed in barley (*H. vulgare*), in which several isomiRs of miR399 and miR827 were also significantly upregulated under conditions of Pi starvation (Hackenberg et al., 2013). In our study, *PLDD* was firstly found to be a cleavage target of miR399c. *PLDD*, which belongs to the *PLD* family, was inhibited in a manner typical for PLD family proteins. PLD is involved in multiple plant processes, ranging from abiotic and biotic stress responses to plant development. *PLDD*-knockout mutants are tolerant to severe drought stress, indicating that *PLDD* acts as a negative regulator in response to drought conditions (Distefano et al., 2015). This suggests that *PLDD* might play a negative regulatory role in *B. luminifera* under LP conditions. Therefore, the induction of isomiRs, such as miR399c, in LP-treated roots of *B. luminifera* to reduce the *PLDD* abundance, which may improve the tolerance to Pi starvation.

miR827-*NLA* pair was reported to be involved in regulating Pi homeostasis (Hsieh et al., 2009; Kant et al., 2011). However, miR827 was absent in both of present sRNA libraries and our previous analyses of *B. luminifera* (Zhang J.H. et al., 2016), except for five reads detected after 4 h of heat stress (Pan et al., 2017). Because of the lack of its precursor sequence in the draft genome, the presence of miR827 in *B. luminifera* is inconclusive. Its absence could be explained either by loss of *MIR827* during evolution or by failed ligation during sRNA library construction. Similar phenomena were observed in three legumes and *C. papaya*, in which the *MIR827* locus was lost during evolution (Lin et al., 2018).

CONCLUSION

Integrated analysis of mRNA-Seq, miRNA-Seq, and degradome-Seq was performed to elucidate the molecular mechanisms underlying the response of *B. luminifera* to conditions of Pi deficiency. Totals of 8,095 and 5,584 DE mRNAs were identified in –Pi-treated roots and shoots, respectively. sRNA sequencing showed that 66 and 60 miRNAs were identified as −Pi-responsive miRNAs, and 109 and 112 miRNA–target pairs were further validated in roots and shoots, respectively, thus a network was proposed. Taken together, these findings provide useful information to decipher miRNA functions and establish a framework for exploring P signaling networks regulated by miRNAs in *B. luminifera* and other woody plants.

DATA AVAILABILITY STATEMENT

We have uploaded our raw data to the NCBI and the accession number of transcriptome data were GSM4105637–GSM4105640, and the accession number of sRNA data were GSM4105633–GSM4105636.

AUTHOR CONTRIBUTIONS

JZ and ZT conceived and designed the experiments. JZ, YL, and LC analyzed the high-throughput sequencing data. YL, FW, and MH contributed reagents and materials. YZ, YL, and LC performed the qRT-PCR experiment. JZ and ZT wrote and revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was financially supported by State Key Laboratory of Tree Genetics and Breeding (Northeast Forestry University), the Zhejiang Science and Technology Major Program on Agricultural New Variety Breeding (Grant No. 2016C02056-2), and 111 Project (D18008).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.552454/full#supplementary-material

Supplementary Figure 1 | KEGG pathway enrichment analysis of DE mRNAs under −Pi stress. The 20 most-enriched KEGG pathways among the DE mRNAs identified in **(A)** roots and **(B)** shoots under −Pi stress using KOBAS (2.0) with FDR $\leq$ 0.05.

Supplementary Figure 2 | Validation of the expression profiles of genes in *B. luminifera* shoots and roots identified by RNA-seq using qRT-PCR. **(A)** Relative expression of eight LP-responsive genes in shoots after 3, 7, and 15 days of LP treatment; **(B)** Relative expression of eight LP-responsive genes in roots after 3, 7, and 15 days of LP treatment. Expression level is represented by the ratio of LP treatment to CK. The value of relative expression level above one indicates that

genes are upregulated by Pi starvation, and the value below one shows that genes are downregulated by Pi starvation. All expression levels were normalized to *TUA* and *MDH* in shoots and roots, respectively. The experiments were repeated three times. Error bars indicate standard deviation.

Supplementary Figure 3 | Identification of DE miRNAs in shoots and roots after 7 days of Pi starvation. DE miRNAs in response to 7 days of Pi starvation in **(A)** shoots and **(B)** roots were identified using Fisher's exact test, with the significance threshold set to 0.05. Only significantly differentially expressed miRNAs with a fold change greater than 2 are shown. **(C)** Venn diagrams showing the numbers of common DE miRNAs and the overlapping sets obtained across four comparisons.

Supplementary Figure 4 | Validation of miRNA target genes. Target plots (t-plots) of representative miRNA targets confirmed by degradome-Seq. The abundance of each signature is plotted as a function of its position in the transcript. The representative miRNAs and their corresponding targets are shown; the red line indicates the cleavage site of each transcript. *ARF17*, auxin response factor 17; *NAC21*, (NAM, ATAF, CUC 21); *NFYA1*, nuclear transcription factor Y subunit A1; *APS1*, ATP sulfurylase 1; *MYB108,* Myb domain protein 108.

Supplementary Figure 5 | Possible functional networks for −Pi-responsive miRNAs and their corresponding target genes in *B. luminifera*. Relationships between −Pi-responsive miRNAs and their target genes shown based upon putative physiological functions. Red arrows represent upregulation, whereas green arrows represent downregulation.

Supplementary Table 1 | 5′RLM-RACE primers for validation of miRNA targets.

Supplementary Table 2 | Gene primers used in qRT-PCR.

Supplementary Table 3 | miRNA primers used in qRT-PCR.

Supplementary Table 4 | tRNA identification in sRNA libraries.

Supplementary Table 5 | Identification of known miRNAs in *B. luminifera*.

Supplementary Table 6 | Identification of novel miRNAs in *B. luminifera*.

Supplementary Table 7 | Differences in expression of miRNAs between control and low phosphate (LP) treatment in roots.

Supplementary Table 8 | Differences in expression of miRNAs between control and LP treatment in shoots.

Supplementary Table 9 | miRNA targets in roots.

Supplementary Table 10 | miRNA targets in shoots.

Supplementary Table 11 | Gene ontology (GO) enrichment analysis for targets of DE miRNAs under conditions of phosphate (Pi) starvation.

Supplementary Table 12 | Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis for targets of DE miRNAs under conditions of Pi starvation.

REFERENCES

Aceituno, F. F., Moseyko, N., Rhee, S. Y., and Gutierrez, R. A. (2008). The rules of gene expression in plants: organ identity and gene body methylation are key factors for regulation of gene expression in *Arabidopsis thaliana*. *BMC Genom.* 9:438. doi: 10.1186/1471-2164-9-438

Ameres, S. L., Horwich, M. D., Hung, J. H., Xu, J., Ghildiyal, M., Weng, Z., et al. (2010). Target RNA-directed trimming and tailing of small silencing RNAs. *Science* 328, 1534–1539.

Bai, Q. Q., Wang, X. Y., Chen, X., Shi, G. Q., Liu, Z. P., Guo, C. J., et al. (2018). Wheat miRNA TaemiR408 acts as an essential mediator in plant tolerance to Pi deprivation and salt stress via modulating stress-associated physiological processes. *Front. Plant Sci.* 9:499. doi: 10.3389/fpls.2018.00499

Balyan, S., Joseph, S. V., Jain, R., Mutum, R. D., Functional, S. R. J., and Genomics, I. (2020). Investigation into the miRNA/5 ' isomiRNAs function and drought-mediated miRNA processing in rice. *Funct. Integr. Genom.* 20, 509–522. doi: 10.1007/s10142-020-00731-2

Buhtz, A., Pieritz, J., Springer, F., and Kehr, J. (2010). Phloem small RNAs, nutrient stress responses, and systemic mobility. *BMC Plant Biol.* 10:64. doi: 10.1186/1471-2229-10-64

Charles Addo-Quaye, T., Eshoo, W., Bartel, D. P., and Michael, J. A. (2008). Endogenous siRNA and miRNA targets identified by sequencing of the *Arabidopsis* degradome. *Curr. Biol.* 18, 758–762. doi: 10.1016/j.cub.2008.04.042

Chen, L., Ren, Y., Zhang, Y., Xu, J., Sun, F., Zhang, Z., et al. (2012). Genome-wide identification and expression analysis of heat-responsive and novel microRNAs in *Populus tomentosa*. *Gene* 504, 160–165. doi: 10.1016/j.gene.2012.05.034

Chen, Y. F., Li, L. Q., Xu, Q., Kong, Y. H., Wang, H., and Wu, W. H. (2009). The WRKY6 transcription factor modulates PHOSPHATE1 expression in response to low Pi stress in *Arabidopsis*. *Plant Cell* 21, 3554–3566. doi: 10.1105/tpc.108.064980

Chiou, T. J., Aung, K., Lin, S. I., Wu, C. C., Chiang, S. F., and Su, C. L. (2006). Regulation of phosphate homeostasis by microRNA in *Arabidopsis*. *Plant Cell* 18, 412–421. doi: 10.1105/tpc.105.038943

Constabel, C. P., Yip, L., Patton, J. J., and Christopher, M. E. (2000). Polyphenol oxidase from hybrid poplar. Cloning and expression in response to wounding and herbivory. *Plant Physiol.* 124, 285–295. doi: 10.1104/pp.124.1.285

Cordell, D., Drangert, J.-O., and White, S. (2009). The story of phosphorus: global food security and food for thought. *Glob. Environ. Chang.* 19, 292–305. doi: 10.1016/j.gloenvcha.2008.10.009

Distefano, A. M., Valinas, M. A., Scuffi, D., Lamattina, L., Ten Have, A., Garcia-Mata, C., et al. (2015). Phospholipase D delta knock-out mutants are tolerant to severe drought stress. *Plant Signal. Behav.* 10:e1089371. doi: 10.1080/15592324.2015.1089371

Du, Z., Zhou, X., Ling, Y., Zhang, Z., and Su, Z. (2010). agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Res.* 38, W64–W70.

Duff, S. M. G., Sarath, G., and Plaxton, W. C. (1994). The role of acid phosphatases in plant phosphorus metabolism. *Physiol. Plant.* 90, 791–800. doi: 10.1111/j.1399-3054.1994.tb02539.x

Fard, E. M., Moradi, S., Salekdeh, N. N., Bakhshi, B., Ghaffari, M. R., Zeinalabedini, M., et al. (2020). Plant isomiRs: origins, biogenesis, and biological functions. *Genomics* 112, 3382–3395. doi: 10.1016/j.ygeno.2020.06.019

Fu, H., Feng, J., Liu, Q., Sun, F., Tie, Y., Zhu, J., et al. (2009). Stress induces tRNA cleavage by angiogenin in mammalian cells. *FEBS Lett.* 583, 437–442. doi: 10.1016/j.febslet.2008.12.043

Fujii, H., Chiou, T. J., Lin, S. I., Aung, K., and Zhu, J. K. (2005). A miRNA involved in phosphate-starvation response in *Arabidopsis*. *Curr. Biol.* 15, 2038–2043. doi: 10.1016/j.cub.2005.10.016

Hackenberg, M., Shi, B. J., Gustafson, P., and Langridge, P. (2013). Characterization of phosphorus-regulated miR399 and miR827 and their isomirs in barley under phosphorus-sufficient and phosphorus-deficient conditions. *BMC Plant Biol.* 13:214. doi: 10.1186/1471-2229-13-214

Hamburger, D., Rezzonico, E., Macdonald-Comber Petetot, J., Somerville, C., and Poirier, Y. (2002). Identification and characterization of the *Arabidopsis* PHO1 gene involved in phosphate loading to the xylem. *Plant Cell* 14, 889–902. doi: 10.1105/tpc.000745

Hsieh, L. C., Lin, S. I., Shih, A. C., Chen, J. W., Lin, W. Y., Tseng, C. Y., et al. (2009). Uncovering small RNA-mediated responses to phosphate deficiency in *Arabidopsis* by deep sequencing. *Plant Physiol* 151, 2120–2132. doi: 10.1104/pp.109.147280

Huang, H. H., Jiang, C., Tong, Z. K., Cheng, L. J., Zhu, M. Y., and Lin, E. P. (2014). Eight distinct cellulose synthase catalytic subunit genes from *Betula luminifera* are associated with primary and secondary cell wall biosynthesis. *Cellulose* 21, 2183–2198. doi: 10.1007/s10570-014-0261-z

Huang, J., Huang, Z., Zhou, X., Xia, C., Imran, M., Wang, S., et al. (2019). Tissue-specific transcriptomic profiling of Plantago major provides insights for the involvement of vasculature in phosphate deficiency responses. *Mol. Genet. Genom.* 294, 159–175. doi: 10.1007/s00438-018-1496-4

Iki, T., Clery, A., Bologna, N. G., Sarazin, A., Brosnan, C. A., Pumplin, N., et al. (2018). Structural flexibility enables alternative maturation, ARGONAUTE sorting and activities of miR168, a global gene silencing regulator in plants. *Mol. Plant* 11, 1008–1023. doi: 10.1016/j.molp.2018.05.006

Jain, A., Poling, M. D., Smith, A. P., Nagarajan, V. K., Lahner, B., Meagher, R. B., et al. (2009). Variations in the composition of gelling agents affect

morphophysiological and molecular responses to deficiencies of phosphate and other nutrients. *Plant Physiol.* 150, 1033–1049. doi: 10.1104/pp.109.136184

Jones-Rhoades, M. W., and Bartel, D. P. (2004). Computational identification of plant MicroRNAs and their targets, including a stress-induced miRNA. *Mol. Cell* 14, 787–799. doi: 10.1016/j.molcel.2004.05.027

Jones-Rhoades, M. W., Bartel, D. P., and Bartel, B. (2006). MicroRNAs and their regulatory roles in plants. *Annu. Rev. Plant Biol.* 57, 19–53. doi: 10.1146/annurev.arplant.57.032905.105218

Kant, S., Peng, M., and Rothstein, S. J. (2011). Genetic regulation by NLA and microRNA827 for maintaining nitrate-dependent phosphate homeostasis in *Arabidopsis*. *PLoS Genet.* 7:e1002021. doi: 10.1371/journal.pgen.1002021

Khandal, H., Singh, A. P., and Chattopadhyay, D. J. P. P. (2020). MicroRNA397b-LACCASE2 module regulates root lignification under water- and phosphate deficiency. *Plant Physiol.* 182:921.

Klipcan, L., and Safro, M. (2004). Amino acid biogenesis, evolution of the genetic code and aminoacyl-tRNA synthetases. *J. Theor. Biol.* 228, 389–396. doi: 10.1016/j.jtbi.2004.01.014

Kozomara, A., and Griffiths-Jones, S. (2013). miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* 2013:gkt1181.

Kumar, S., Verma, S., and Trivedi, P. K. (2017). Involvement of small RNAs in phosphorus and sulfur sensing, signaling and stress: current update. *Front. Plant Sci.* 8:285. doi: 10.3389/fpls.2017.00285

Lei, M., Liu, Y., Zhang, B., Zhao, Y., Wang, X., Zhou, Y., et al. (2011). Genetic and genomic evidence that sucrose is a global regulator of plant responses to phosphate starvation in *Arabidopsis*. *Plant Physiol.* 156, 1116–1130. doi: 10.1104/pp.110.171736

Li, M., Xia, Y., Gu, Y., Zhang, K., Lang, Q., Chen, L., et al. (2010). MicroRNAome of porcine pre- and postnatal development. *PLoS One* 5:e11541. doi: 10.1371/journal.pgen.0011541

Li, Z., Xu, H., Li, Y., Wan, X., Ma, Z., Cao, J., et al. (2018). Analysis of physiological and miRNA responses to Pi deficiency in alfalfa (*Medicago sativa* L.). *Plant Mol. Biol.* 96, 473–492. doi: 10.1007/s11103-018-0711-3

Li, Z., Zhang, X., Liu, X., Zhao, Y., Wang, B., and Zhang, J. (2016). miRNA alterations are important mechanism in maize adaptations to low-phosphate environments. *Plant Sci.* 252, 103–117. doi: 10.1016/j.plantsci.2016.07.009

Liang, G., Ai, Q., and Yu, D. Q. (2015). Uncovering miRNAs involved in crosstalk between nutrient deficiencies in *Arabidopsis*. *Sci. Rep.* 5:11813.

Liang, G., He, H., and Yu, D. (2012). Identification of nitrogen starvation-responsive microRNAs in *Arabidopsis thaliana*. *PLoS One* 7:e48951. doi: 10.1371/journal.pone.0048951

Liang, G., Yang, F., and Yu, D. (2010). MicroRNA395 mediates regulation of sulfate accumulation and allocation in *Arabidopsis thaliana*. *Plant J.* 62, 1046–1057.

Lin, W. Y., Lin, S. I., and Chiou, T. J. (2009). Molecular regulators of phosphate homeostasis in plants. *J. Exper. Bot.* 60, 1427–1438. doi: 10.1093/jxb/ern303

Lin, W. Y., Lin, Y. Y., Chiang, S. F., Syu, C., Hsieh, L. C., and Chiou, T. J. (2018). Evolution of microRNA827 targeting in the plant kingdom. *New Phytol.* 217, 1712–1725. doi: 10.1111/nph.14938

Liu, Y., Mi, G., Chen, F., Zhang, J., and Zhang, F. (2004). Rhizosphere effect and root growth of two maize (*Zea mays* L.) genotypes with contrasting P efficiency at low P availability. *Plant Sci.* 167, 217–223. doi: 10.1016/j.plantsci.2004.02.026

Loss-Morais, G., Waterhouse, P. M., and Margis, R. (2013). Description of plant tRNA-derived RNA fragments (tRFs) associated with argonaute and identification of their putative targets. *Biol. Direct.* 8:6.

Lu, S., Sun, Y. H., and Chiang, V. L. (2009). Adenylation of plant miRNAs. *Nucleic Acids Res.* 37, 1878–1885. doi: 10.1093/nar/gkp031

Mao, W., Li, Z., Xia, X., Li, Y., and Yu, J. (2012). A combined approach of high-throughput sequencing and degradome analysis reveals tissue specific expression of microRNAs and their targets in cucumber. *PLoS One* 7:e33040. doi: 10.1371/journal.pgen.10033040

Mao, X., Cai, T., Olyarchuk, J. G., and Wei, L. (2005). Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics* 21, 3787–3793. doi: 10.1093/bioinformatics/bti430

Marschner, H. (1995). *Mineral Nutrition on Higher Plants*. London: Academic Press.

Meyers, B. C., Axtell, M. J., Bartel, B., Bartel, D. P., Baulcombe, D., Bowman, J. L., et al. (2008). Criteria for annotation of plant MicroRNAs. *Plant Cell* 20, 3186–3190. doi: 10.1105/tpc.108.064311

Morin, R. D., Aksay, G., Dolgosheina, E., Ebhardt, H. A., Magrini, V., Mardis, E. R., et al. (2008a). Comparative analysis of the small RNA transcriptomes of *Pinus contorta* and *Oryza sativa*. *Genome Res.* 18, 571–584. doi: 10.1101/gr.6897308

Morin, R. D., O'connor, M. D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A. L., et al. (2008b). Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res.* 18, 610–621. doi: 10.1101/gr.7179508

Naya, L., Khan, G. A., Sorin, C., Hartmann, C., Crespi, M., and Lelandais-Briere, C. (2010). Cleavage of a non-conserved target by a specific miR156 isoform in root apexes of *Medicago truncatula*. *Plant Signal Behav.* 5, 328–331. doi: 10.4161/psb.5.3.11190

Nilsson, L., Muller, R., and Nielsen, T. H. (2010). Dissecting the plant transcriptome and the regulatory responses to phosphate deprivation. *Physiol. Plant* 139, 129–143. doi: 10.1111/j.1399-3054.2010.01356.x

O'Rourke, J. A., Yang, S. S., Miller, S. S., Bucciarelli, B., Liu, J. Q., Rydeen, A., et al. (2013). An RNA-Seq transcriptome analysis of orthophosphate-deficient white lupin reveals novel insights into phosphorus acclimation in plants. *Plant Physiol.* 161, 705–724. doi: 10.1104/pp.112.209254

Pan, Y., Niu, M. Y., Liang, J. S., Lin, E. P., Tong, Z. K., and Zhang, J. H. (2017). Identification of heat-responsive miRNAs to reveal the miRNA-mediated regulatory network of heat stress response in *Betula luminifera*. *Trees Struct. Funct.* 31, 1635–1652. doi: 10.1007/s00468-017-1575-x

Pant, B. D., Buhtz, A., Kehr, J., and Scheible, W. R. (2008). MicroRNA399 is a long-distance signal for the regulation of plant phosphate homeostasis. *Plant J.* 53, 731–738. doi: 10.1111/j.1365-313X.2007.03363.x

Pei, L., Jin, Z., Li, K., Yin, H., Wang, J., and Yang, A. (2013). Identification and comparative analysis of low phosphate tolerance-associated microRNAs in two maize genotypes. *Plant Physiol. Biochem.* 70, 221–234. doi: 10.1016/j.plaphy.2013.05.043

Pelaez, P., Trejo, M. S., Iniguez, L. P., Estrada-Navarrete, G., Covarrubias, A. A., Reyes, J. L., et al. (2012). Identification and characterization of microRNAs in *Phaseolus vulgaris* by high-throughput sequencing. *BMC Genom.* 13:83. doi: 10.1186/1471-2164-13-83

Secco, D., Jabnoune, M., Walker, H., Shou, H., Wu, P., Poirier, Y., et al. (2013). Spatio-temporal transcript profiling of rice roots and shoots in response to phosphate starvation and recovery. *Plant Cell* 25, 4285–4304. doi: 10.1105/tpc.113.117325

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

Shao, H., Wang, H., and Tang, X. (2015). NAC transcription factors in plant multiple abiotic stress responses: progress and prospects. *Front Plant Sci.* 6:902. doi: 10.3389/fpls.2015.00902

Shin, H., Shin, H. S., Dewbre, G. R., and Harrison, M. J. (2004). Phosphate transport in Arabidopsis: Pht1;1 and Pht1;4 play a major role in phosphate acquisition from both low- and high-phosphate environments. *Plant J.* 39, 629–642.

Slotkin, R. K., Vaughn, M., Borges, F., Tanurdzic, M., Becker, J. D., Feijo, J. A., et al. (2009). Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell* 136, 461–472. doi: 10.1016/j.cell.2008.12.038

Su, T., Xu, Q., Zhang, F. C., Chen, Y., Li, L. Q., Wu, W. H., et al. (2015). WRKY42 modulates phosphate homeostasis through regulating phosphate translocation and acquisition in *Arabidopsis*. *Plant Physiol.* 167, 1579–U1717.

Sun, S., Gu, M., Cao, Y., Huang, X., Zhang, X., Ai, P., et al. (2012). A constitutive expressed phosphate transporter, OsPht1;1, modulates phosphate uptake and translocation in phosphate-replete rice. *Plant Physiol.* 159, 1571–1581. doi: 10.1104/pp.112.196345

Sunkar, R., Girke, T., Jain, P. K., and Zhu, J. K. (2005). Cloning and characterization of microRNAs from rice. *Plant Cell* 17, 1397–1411. doi: 10.1105/tpc.105.031682

Sunkar, R., Li, Y. F., and Jagadeeswaran, G. (2012). Functions of microRNAs in plant stress responses. *Trends Plant Sci.* 17, 196–203. doi: 10.1016/j.tplants.2012.01.010

Takahashi, A., Takeda, K., and Ohnishi, T. (1991). Light-induced anthocyanin reduces the extent of damage to DNA in UV-irradiated *Centaura cyanus* cells in culture. *Plant Cell Physiol.* 32, 541–547.

Thompson, D. M., Cheng, L., Green, P. J., and Parker, R. (2008). tRNA cleavage is a conserved response to oxidative stress in eukaryotes. *RNA* 14, 2095–2103. doi: 10.1261/rna.1232808

Tran, H. T., Hurley, B. A., and Plaxton, W. C. (2010). Feeding hungry plants: the role of purple acid phosphatases in phosphate nutrition. *Plant Sci.* 179, 14–27. doi: 10.1016/j.plantsci.2010.04.005

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

Vance, C. P., Uhde-Stone, C., and Allan, D. L. (2003). Phosphorus acquisition and use: critical adaptations by plants for securing a nonrenewable resource. *New Phytol.* 157, 427–447.

Wege, S., Khan, G. A., Jung, J. Y., Vogiatzaki, E., Pradervand, S., Aller, I., et al. (2016). The EXS domain of PHO1 participates in the response of shoots to phosphate deficiency via a root-to-shoot signal. *Plant Physiol.* 170, 385–400. doi: 10.1104/pp.15.00975

Xu, F., Liu, Q., Chen, L., Kuang, J., and Liao, H. (2013). Genome-wide identification of soybean microRNAs and their targets reveals their organ-specificity and responses to phosphate starvation. *BMC Genom.* 14:66. doi: 10.1186/1471-2164-14-66

Xue, Y., Zhuang, Q., Zhu, S., Xiao, B., Liang, C., Liao, H., et al. (2018). Genome wide transcriptome analysis reveals complex regulatory mechanisms underlying phosphate homeostasis in soybean nodules. *Int. J. Mol. Sci.* 19:2924. doi: 10.3390/ijms19102924

Ye, Q., Wang, H., Su, T., Wu, W. H., and Chen, Y. F. (2018). The ubiquitin E3 Ligase PRU1 regulates WRKY6 degradation to modulate phosphate homeostasis in response to low-Pi stress in *Arabidopsis*. *Plant Cell* 30, 1062–1076. doi: 10.1105/tpc.17.00845

Young, M. D., Wakefield, M. J., Smyth, G. K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* 11:R14.

Zeng, H., Wang, G., Hu, X., Wang, H., Du, L., and Zhu, Y. (2013). Role of microRNAs in plant responses to nutrient stress. *Plant Soil* 374, 1005–1021. doi: 10.1007/s11104-013-1907-6

Zeng, H. Q., Wang, G. P., Zhang, Y. Q., Hu, X. Y., Pi, E. X., Zhu, Y. Y., et al. (2016). Genome-wide identification of phosphate-deficiency-responsive genes in soybean roots by high-throughput sequencing. *Plant Soil* 398, 207–227. doi: 10.1007/s11104-015-2657-4

Zhang, B. (2015). MicroRNA: a new target for improving plant tolerance to abiotic stress. *J. Exper. Bot.* 66, 1749–1761. doi: 10.1093/jxb/erv013

Zhang, G., Yin, S., Mao, J., Liang, F., Zhao, C., Li, P., et al. (2016). Integrated analysis of mRNA-seq and miRNA-seq in the liver of *Pelteobagrus vachelli* in response to hypoxia. *Sci. Rep.* 6:22907.

Zhang, J. H., Huang, M. H., Liang, J. S., Pan, Y., Cheng, L. J., Wu, J., et al. (2016). Genome-wide mining for microRNAs and their targets in *Betula luminifera* using high-throughput sequencing and degradome analyses. *Tree Genet. Genom.* 12:99.

Zhang, J. H., Wu, T., Li, L., Han, S. Y., Li, X. M., Zhang, S. G., et al. (2013). Dynamic expression of small RNA populations in larch (*Larix leptolepis*). *Planta* 237, 89–101. doi: 10.1007/s00425-012-1753-4

Zhang, S., Sun, L., and Kragler, F. (2009). The phloem-delivered RNA pool contains small noncoding RNAs and interferes with translation. *Plant Physiol.* 150, 378–387. doi: 10.1104/pp.108.134767

Zhao, M., Ding, H., Zhu, J. K., Zhang, F., and Li, W. X. (2011). Involvement of miR169 in the nitrogen-starvation responses in *Arabidopsis*. *New Phytol.* 190, 906–915. doi: 10.1111/j.1469-8137.2011.03647.x

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.

Copyright © 2021 Zhang, Lin, Wu, Zhang, Cheng, Huang and Tong. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

REVIEW
published: 03 February 2021
doi: 10.3389/fgene.2021.583888

Transcriptomics of Biostimulation of Plants Under Abiotic Stress

*Susana González-Morales[1], Susana Solís-Gaona[2], Marin Virgilio Valdés-Caballero[2], Antonio Juárez-Maldonado[3], Araceli Loredo-Treviño[4] and Adalberto Benavides-Mendoza[5]**

[1] CONACYT-Universidad Autónoma Agraria Antonio Narro, Saltillo, Mexico, [2] UPL, Saltillo, Mexico, [3] Departamento de Botánica, Universidad Autónoma Agraria Antonio Narro, Saltillo, Mexico, [4] Departamento de Alimentos, Universidad Autónoma de Coahuila, Saltillo, Mexico, [5] Departamento de Horticultura, Universidad Autónoma Agraria Antonio Narro, Saltillo, Mexico

Plant biostimulants are compounds, living microorganisms, or their constituent parts that alter plant development programs. The impact of biostimulants is manifested in several ways: via morphological, physiological, biochemical, epigenomic, proteomic, and transcriptomic changes. For each of these, a response and alteration occur, and these alterations in turn improve metabolic and adaptive performance in the environment. Many studies have been conducted on the effects of different biotic and abiotic stimulants on plants, including many crop species. However, as far as we know, there are no reviews available that describe the impact of biostimulants for a specific field such as transcriptomics, which is the objective of this review. For the commercial registration process of products for agricultural use, it is necessary to distinguish the specific impact of biostimulants from that of other legal categories of products used in agriculture, such as fertilizers and plant hormones. For the chemical or biological classification of biostimulants, the classification is seen as a complex issue, given the great diversity of compounds and organisms that cause biostimulation. However, with an approach focused on the impact on a particular field such as transcriptomics, it is perhaps possible to obtain a criterion that allows biostimulants to be grouped considering their effects on living systems, as well as the overlap of the impact on metabolism, physiology, and morphology occurring between fertilizers, hormones, and biostimulants.

Keywords: gene expression, PGPRs, macroalgae, peptides, humic acid, chitosan, selenium, silicon

OPEN ACCESS

Edited by:
Rong Zhou,
Aarhus University, Denmark

Reviewed by:
Antonio Ferrante,
University of Milan, Italy
Shabir Hussain Wani,
Sher-e-Kashmir University of Agricultural Sciences and Technology, India

***Correspondence:**
Adalberto Benavides-Mendoza
abenmen@gmail.com

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *15 July 2020*
Accepted: *06 January 2021*
Published: *03 February 2021*

Citation:
González-Morales S, Solís-Gaona S, Valdés-Caballero MV, Juárez-Maldonado A, Loredo-Treviño A and Benavides-Mendoza A (2021) Transcriptomics of Biostimulation of Plants Under Abiotic Stress. Front. Genet. 12:583888. doi: 10.3389/fgene.2021.583888

INTRODUCTION

Biostimulation has been described as a general biological phenomenon dependent on the interactions between the cell's molecular structures and the external physical, chemical and biological stimuli, or impulses. The biostimulation results in the alteration of metabolic processes that allows the most efficient use of environmental resources, substantially increased growth or yield, and increased tolerance to adverse environmental factors (Juárez-Maldonado et al., 2019). It would be expected that starting from the definition of biostimulation that the definition of biostimulant would be derived; but in reality, that of biostimulant preceded that of biostimulation as a result of the need for regulation in the agricultural sector, which has found a very relevant niche in biostimulants (du Jardin, 2015).

Indeed, at present, biostimulants have gained importance in agriculture from ecological and commercial perspectives. Biostimulants are labeled as ecologically benign since most are considered biodegradable, non-toxic, non-polluting, and non-hazardous for various organisms (Yakhin et al., 2017).

Currently, there is no agreed-upon legal or academic definition of biostimulant; thus, there are several definitions of this term. One academic definition of a plant biostimulant refers to any substance or microorganism applied to plants to improve nutrient use efficiency, stress tolerance, and/or quality traits of plants, regardless of their nutrient content (du Jardin, 2015). Another definition defines biostimulants as materials that, when applied in small amounts, are capable of promoting growth in plants (Juárez-Maldonado et al., 2019).

It should be noted that, from the biological point of view, biostimulants are all those biotic and abiotic factors that induce biostimulation (including physical factors such as UV radiation, physicochemical factors such as nanomaterials, and biological factors such as pests, pathogens, competitors, and beneficial organisms and microorganisms) (Juárez-Maldonado et al., 2019). However, for purposes of developing this review with a focus in agriculture, we start from the biostimulants definition and categorization developed by du Jardin (2015). In his paper, the author classified biostimulants into seven groups: (1) beneficial bacteria, (2) beneficial fungi, (3) algae and botanical extracts, (4) protein hydrolysates and other nitrogen-containing compounds, (5) humic acid and fulvic acid, (6) chitosan and other biopolymers, and (7) inorganic compounds.

Currently, the advancement of omics sciences through the development of genome sequencing technologies, as well as the reduction in the costs of these techniques, have revolutionized the understanding of the identification of metabolic pathways in plants (Owen et al., 2017). The review is based on the supposition that through transcriptomic studies, it is possible to identify molecular markers (Harper et al., 2016) that are associated with common responses in a wide range of crop species; these molecular markers can, in turn, be used to understand the mode of action of biostimulants, their interaction with the environment, and the genotype of plants (De Palma et al., 2019). The present review contributes to the description, in terms of transcriptomics, of the impact of the different categories of biostimulants on agricultural crop species, with an emphasis on corn, rice, wheat, tomato, and *Arabidopsis*. It is intended to describe the transcriptomic landscape induced by biostimulants in plants under abiotic stress. Abiotic stress, such as salinity, drought or temperature variation, can decrease productivity and cause considerable losses in crop yields by more than 60% (Singhal et al., 2016). The following sections are included in this manuscript: (1) beneficial bacteria; (2) beneficial fungi; (3) algae and botanical extracts; (4) protein hydrolysates and other nitrogen-containing compounds; (5) humic acid and fulvic acid; (6) chitosan and other biopolymers; and (7) inorganic compounds. In each section, we present works that describe the genes that whose transcript level changes in plants in response to the application of a biostimulant (in a specific category) in plants under abiotic stress. In total, 56 recent works (published from 2010 to the present) from the seven biostimulant category groups were included.

BENEFICIAL BACTERIA

Microorganisms are widely used to produce biostimulants (Xavier and Boyetchko, 2002; Sofo et al., 2014; Colla et al., 2015; Matyjaszczyk, 2015). Biostimulants based on microorganisms are preparations that include living and/or non-living microorganisms and their metabolites. Within this category, the most studied microorganisms are those defined as plant growth-promoting bacteria (PGPBs) (Calvo et al., 2014). Among the PGPBs, the genres most studied for use as biostimulants include *Bacillus* (Adesemoye et al., 2010), *Azospirillum*, *Pseudomonas*, *Streptomyces*, *Achromobacter*, and *Rhizobium* (Calvo et al., 2014). The plant-promoting effects of PGPBs are mostly explained by the release of metabolites directly stimulating growth. The direct mechanism includes those in which the bacteria produce growth regulators [auxins, gibberellins, cytokinins, ethylene, and abscisic acid (ABA)] that are ultimately incorporated into the plant system and thus affect the balance of plant growth regulators, or they act as a sink for plant-released hormones that induce changes in plant metabolism, promoting the overall growth of the plant (Govindasamy et al., 2010). The indirect mechanisms include biological nitrogen fixation, phosphate solubilization, production of siderophores, antibiotics, and other metabolites, production of defense enzymes and modulation of plant stress markers, induction of systemic resistance (ISR) and competition within the rhizosphere (Kumar et al., 2020).

Genes and their function are described below using the results of 10 studies in which beneficial bacteria-based biostimulants were applied to plants under abiotic stress (**Table 1**).

The effect of a biostimulant from *Pseudomonas fluorescens* applied to reproductive-stage rice plants under drought stress (water restriction for 15 days) was analyzed (Saakre et al., 2017), and six differentially expressed genes (DEGs), *COX1*, *PKDP*, *bZIP1*, *AP2-EREBP*, *Hsp20*, and *COC1*, were reported in leaves of 65-day-old plants. Differentially expressed transcripts correspond to the cytochrome oxidase subunit 1 (*COX1*) gene, protein kinase domain protein (*PKDP*), bZIP protein family (*bZIP1*) members, APETALA2-ethylene-responsive element-binding protein (*AP2-EREBP*), heat-shock protein 20 (*Hsp20*), and the circadian oscillator component (*COC1*). *bZIP1*, *AP2-EREBP*, *COC1*, and *PKDP* genes are considered regulatory proteins. *bZIP1*, *AP2-EREBPs*, and *COC1* genes are considered stress-induced transcription factors (TFs) involved in the ABA-dependent signaling pathway (Reeves et al., 2011; Seung et al., 2012). ABA is generally defined as a stress hormone since it intervenes in the regulation of the response to biotic and abiotic stress (Vishwakarma et al., 2017); it also plays a vital role in physiological processes such as seed germination, dormancy, and stomatal closure (Akpinar et al., 2012). In the ABA-dependent signaling pathway, protein receptors (PYR/PIL/RCAR) bind to ABA by inhibiting PP2C activity, allowing SnRK2 activation by autophosphorylation. Subsequently, active SnRK2 can phosphorylate TFs such as

TABLE 1 | Description of genes whose transcript level changed in response to the application of bacteria-based biostimulants in plants under abiotic stress.

Gene/gene expression time	Biostimulant type	Abiotic stress type	Crop/gene expression organ	References
COX1, *PKDP*, *bZIP1*, *AP2-EREBP*, *Hsp20*, *COC1*/65 days old	*Pseudomonas fluorescens*	Drought	Rice/leaves	Saakre et al., 2017
ACS, *AC*/45 days after stress	*Ochrobactrum pseudogrignonense*, *Pseudomonas* sp., and *Bacillus subtilis*	Drought	*Vigna mungo* and *Pisum sativum* L./leaves and roots	Saikia et al., 2018
OsNAM, *OsGRAM*/not specified	*Bacillus amyloliquefaciens*	Salinity	Rice/roots	Chauhan et al., 2019
UP1, *NAS*, *MT*, *MTR-1-P*, *VITs*, *P450*, *HRP*, *UP2*/not specified	*Arthrobacter nitroguajacolicus*	Salinity	Wheat/roots	Safdarian et al., 2019
RD29A, *RD29B*, *RD20*, *RD22*, *KIN1*, *ERD1*/not specified	*Bacillus oryzicola*	Salinity	*Arabidopsis*/shoots and roots	Baek et al., 2020
IA24, *IA41*, *IA19*/after 6 h at heat stress, after 6 h at cold stress, after 5 days without water	*Bacillus amyloliquefaciens* subsp. *plantarum*	Heat, cold, drought	Wheat/leaves	Abd El-Daim et al., 2018
APX, *CAT*, *SOD*, *RBCS*, *RBCL*, H^+-*PPase*, *HKT1*, *NHX1*, *NCED*/24 days after transplantation	*Serratia liquefaciens*	Salinity	Maize/leaves	El-Esawi et al., 2018b
P5CS1, *SOD*, *APX1*, *CAT*, *SHSP*/45 days from sowing	*Bacillus cereus*, *Providencia rettgeri* and *Myroides odoratimimus*	Chromium and heat stresses	Sorghum/leaves	Bruno et al., 2020
CAPIP2, *CaKR1*, *CaOSM1*, *CAChi2*/7 days after inoculation	*Bacillus fortis*	Salinity	Capsicum/leaves	Yasin et al., 2018
APX1, *CAT1*, *SOD2*, *SOD4*, *APX2*, *PR1*, *PRP2*, *PRP4*, *HSP70*/32 days after transplant	*Azospirillum brasilense* and *Rhizobium tropici*	Salinity	Maize/leaves and roots	Fukami et al., 2018

AREB/ABF, DREB2A/2B, RD22BP1, and MYC/MYB TFs, which bind to their corresponding cis-elements and modulate the expression of response genes (de Zelicourt et al., 2016; Bulgakov and Wu, 2019). The *PKDP* gene encodes an enzyme with kinase activity that participates in phosphorylation reactions; this enzyme is a key component of signal transduction (Mohanta et al., 2015) and is involved in multiple cellular processes in plants such as growth and development, carbon and nitrogen metabolism, the formation of the cytoskeleton, senescence and cell death, hormone signal transduction, regulation of ion channels, and the defense response (Lei et al., 2007). *Hsp20* and *COX1* genes are considered functional proteins. The *Hsp20* gene encodes a small (12–42 kDa) heat-shock protein (sHsp) with chaperone activity; its function is to degrade denatured proteins, and it has an ATP-dependent chaperonin function. These proteins can bind partially folded or denatured proteins and avoid permanent unfolding or aggregation (Reddy et al., 2015; Nagaraju et al., 2020). The *COX1* gene encodes an enzyme with cytochrome oxidase activity, which is related to the process of apoptosis, a type of programmed cell death (PCD) in plants. PCD is triggered in plants during biotic and abiotic stress since recycling and mobilizing nutrients are carried out under adverse conditions; this process can be triggered by reactive oxygen species (ROS) (Foyer and Noctor, 2005). Moreover, autophagy (another type of PCD) is related to protein turnover and removal of damaged organelle proteins, which can lead to changes in cell behavior (Akpinar et al., 2012).

As a strategy to improve *Vigna mungo* L. and *Pisum sativum* L. plants under conditions of water deficit (polyethylene glycol-6000), (Saikia et al., 2018) a PGPB consortium (*Ochrobactrum pseudogrignonense*, *Pseudomonas* spp., and *Bacillus subtilis*) was applied. The plants treated with the consortium presented, in addition to increased growth, a better adaptation to resist stress due to downregulation of the *ACS* and *ACO* genes at 45 days after stress. In the plants that had only stress, the opposite occurred, and the expression of these genes was increased. The *ACS* gene encodes the ACC synthase enzyme, and the *ACO* gene encodes the ACC oxidase enzyme. ACC synthase is responsible for producing ACC, and ACC oxidase is involved in the production of ethylene (Li et al., 2020b). Ethylene is a hormone related to plant growth and development as well as processes such as fruit ripening and abscission, leaf senescence, seed germination, and organogenesis (Trujillo-Moya and Gisbert, 2012); however, a sudden increase in ethylene concentration in the plant during stress causes negative effects and leads to senescence (Czarny et al., 2006). The mechanism by which the expression of the *ACS* and *ACO* genes is reduced is triggered because the consortium of bacteria are producers of hormones, specifically indoleacetic acid (IAA), which triggers the production of relatively high concentrations of ACC and, subsequently, feedback inhibition of IAA synthesis (Glick, 2012). Additionally, some rhizobial strains produce the enzyme ACC deaminase, which removes some of the ACC (the immediate precursor to ethylene) before it can be converted to ethylene (Ma et al., 2002).

In rice plants under salinity stress (100 mM NaCl), the biostimulant effect of *Bacillus amyloliquefaciens*-SN13 was evaluated (Chauhan et al., 2019). During stress, *B. amyloliquefaciens* increased the relative water content, biomass, proline content and total soluble sugar content in plants while decreasing the electrolyte leakage and lipid peroxidation. Alterations in gene expression were also observed in the transcriptome of rice roots under salt stress with *B. amyloliquefaciens.* Changes in the expression of a considerable number of genes related to the stress response,

hormones, photosynthesis, lipid metabolism and cell wall were induced. Additionally, DEGs in functional categories such as the stress response (*Os11g26750*, *Os10g38600*, *Os05g28740*, and *Os05g45480*), metabolism (*Os03g08860*, *Os04g27060*, *Os01g43750*, *Os05g41460*, *Os04g47360*, and *Os03g12510*), transporters (*Os07g15460*), and regulation (*Os04g56990*, *Os06g41770*, *Os09g39650*, *Os09g31031*, *Os03g14590*, *Os11g03370*, and *Os05g10620*) were quantified. The functional validation in *Saccharomyces cerevisiae* of the *OsNAM* and *OsGRAM* genes with 5- and 3-fold induction, respectively, in transcriptome of rice roots with *B. amyloliquefaciens* under salinity was analyzed. Transformed yeast cells expressing *OsNAM* and *OsGRAM* showed enhanced tolerance under osmotic stress; moreover, the transformed cells were able to survive at relatively high temperatures, and they displayed increase tolerance to arsenite and arsenate. *OsNAM* belongs to the NAC family of TFs, which is a large plant-specific gene family that is related to the regulation of tissue development and the response to abiotic stress (Shan et al., 2020). *OsGRAM* is important for the ABA response, hormone signaling under stress conditions, and the perception of environmental stimuli and regulation of the response to those stimuli (Mauri et al., 2016).

The effect of *Arthrobacter nitroguajacolicus* (seeds inoculated with 10^8 CFU mL^{-1}) on wheat tolerance to salinity stress (200 mM NaCl) has been studied (Safdarian et al., 2019). Transcriptomic analysis revealed upregulation of 152 genes, and 5 genes were significantly downregulated. Inoculated roots of plants under salt stress presented differential expression of many genes involved in stilbenoids, diarylheptanoid metabolism, phenylpropanoids, flavonoids, terpenoid, porphyrin, and chlorophyll metabolism. RNA-seq results, 11 DEGs (including *UP1*, *NAS*, *MT*, *MTR-1-P*, *VITs*, *P450*, *UP2*) were quantified. *UP1* is a functional cis-element responsible for germination-associated gene expression (Tatematsu et al., 2008). *NAS* is induced by salt stress and is related to niacinamide, the synthetic precursor of plant iron (Johnson et al., 2011). *MT* gene encodes a metal chelatin protein involved in the detoxification of heavy metals and in the homeostasis of intracellular metal ions (Ahn et al., 2012). *MTR-1-P* encodes a 5-methylthioribose 1-phosphate related to ethylene biosynthesis (Arraes et al., 2015). VIT proteins are involved in the transport and storage of iron, can maintain iron within the optimal physiological range and prevent cellular toxicity (Cao, 2019). Cytochrome P450 proteins are related to signals for growth and development, are responsible for protecting plants from different stresses, and are involved in in redox reactions and a large number of biosynthetic pathways (Bak et al., 2011; Safdarian et al., 2019). *UP2* encodes an uncharacterized conserved protein and is related to cell structure (Duan et al., 2017).

In *Arabidopsis thaliana* plants under salt stress (80 and 100 mM NaCl), the biostimulant effect of *Bacillus oryzicola* was studied (Baek et al., 2020). Compared with control plants, plants with *B. oryzicola* while under salinity showed increased numbers of lateral roots, increased fresh weight, and increased chlorophyll content, and they accumulated less salt-induced malondialdehyde and Na^+. Additionally, plants with *B. oryzicola* while under salinity presented enhanced transcription of the *RD29A*, *RD29B*, *RD20*, *RD22*, *KIN1*, and *ERD1* genes in the shoots and roots. The *RD29A* (which is responsive to desiccation) gene encodes a protein strongly related to the *Brassicaceae* family. Its function is similar to that of LEA proteins (Kaur and Gupta, 2005). *RD29B* encodes a protein that is induced in response to water deprivation, low temperature, salinity, and desiccation, and its response is mediated by ABA (Bihmidine et al., 2013). *RD20* is a stress-inducible gene that belongs to the caleosin family and is dependent on ABA signaling. *RD20* plays a role in drought tolerance through stomatal control under water-deficit conditions (Aubert et al., 2010). The *RD22* (which is responsive to desiccation) gene can be induced by stress caused by drought and salinity and by exogenous applications of ABA, but its role in the stress response is unknown (Wang et al., 2012). The *KIN1* gene encodes a protein induced in response to cold, dehydration, osmotica and ABA, possibly functioning as an antifreeze protein (Wang et al., 1995). The *ERD1* (early responsive to dehydration stress 1) protein functions in chloroplasts by degrading unassembled or misfolded proteins (Peltier et al., 2001).

Using wheat plants under heat (6 h at 45°C), cold (6 h at −5°C) and drought stress (5 days without water), the effect of *B. amyloliquefaciens* subsp. *plantarum* was studied (Abd El-Daim et al., 2018). cDNA-AFLP analysis revealed differential expression of more than 200 transcript-derived fragments (TDFs) in wheat leaves. Five TDFs (*IA20*, *IA24*, *IA35*, *IA41*, *IA19*) were selected for RT-PCR analysis. All the TDFs changed in expression pattern in plants treated with *B. amyloliquefaciens* subsp. *plantarum* and with the different stresses. *IA19*, *IA20*, and *IA24* are ABA responsive homologous. *IA20* encodes to TF WRKY20, this TF enhance drought tolerance regulating ABA signaling (Luo et al., 2013).

The effect of *Serratia liquefaciens* on maize plants under salinity stress (0, 80, and 160 mM NaCl) was studied (El-Esawi et al., 2018b). *S. liquefaciens* inoculation significantly reduced oxidative stress markers but increased the maize growth and biomass production along with better antioxidant defense system, leaf gas exchange, osmoregulation, and nutrient uptake under salinity. Also, it was found the upregulation of stress-related genes (*APX*, *CAT*, *SOD*, *RBCS*, *RBCL*, H^+-*PPase*, *HKT1*, *NHX1*) and downregulation *NCED* gene. The *APX* gene encodes ascorbate peroxidase, which catalyzes the reduction of H_2O_2 and the oxidation of ascorbate, generating monodehydroascorbate (Eltelib et al., 2012). The *CAT* gene encodes catalase, and this enzyme is important in the removal of H_2O_2 generated in peroxisomes by oxidases involved in the β-oxidation of fatty acids, photorespiration and purine catabolism (Gill and Tuteja, 2010). The *SOD* gene encodes an enzyme that is crucial in the first line of antioxidant defense since it can convert highly reactive O_2^- radicals into H_2O_2 and O_2 in response to oxidative stress (Wang W. et al., 2016). The *RBCS* gene encodes the small Rubisco subunit and is part of a multigenic family. The Rubisco enzyme catalyzes the assimilation of CO_2 in plants and is related to the photosynthesis process (Atkinson et al., 2017). *RBCL* is involved in photosynthesis, encodes the large subunit of the primary CO_2 fixation enzyme Rubisco. This enzyme serves

as the primary engine of carbon assimilation being the most abundant protein on earth (Igamberdiev, 2015). The *NCED3* (nine-cis-epoxycarotenoid dioxygenase 3) gene is related to the pathways of ABA biosynthesis and is highly expressed at the root level (Li et al., 2020a). *H^+-PPase* ecodes H^+-pumping pyrophosphate and *HKT1* encodes high-affinity K^+ transporter 1 related with ion balance regulators. *NHX1* encodes Na^+/H^+ antiporter, which participates in Na^+ sequestration, export and recirculation (Chen et al., 2016). Using sorghum plants under chromium (200 mg $K_2Cr_2O_7$ kg^{-1} in soil) and heat stress (42°C during day and 28°C at night), the effect of *Bacillus cereus*, *Providencia rettgeri*, and *Myroides odoratimimus* (chromium reducing-thermotolerant CRT-PGPB) was studied (Bruno et al., 2020). Inoculation with CRT-PGPB increased plant growth, antioxidant status and decreased malondialdehyde and proline contents in plants under stress. Also, gene expression studies down-regulated the expression of *P5CS1* gene, and up-regulated the expression of *SOD*, *APX1*, *CAT* and *SHSP*. All genes were already mentioned.

The effect of *Bacillus fortis* (halotolerant HPGPR) on capsicum plants under salt stress (1 and 2 g NaCl kg^{-1} soil) was studied (Yasin et al., 2018). HPGPR promote growth attributes, chlorophyll, protein content and water use efficiency on capsicum plants under salinity. Also, up-regulated the expression profiles of stress related genes including *CAPIP2*, CaKR1, CaOSM1, and CAChi2. *CAPIP2* encodes plasma membrane intrinsic protein related in transportation of smaller neutral solutes and water (already mentioned). *CaKR1* is a member of ankyrin repeat zinc finger protein TF family, the over-expressed of this TF exhibited enhanced antioxidant metabolism (Seong et al., 2007). *CaOSM1* is a osmotin gene related with resistance against biotic and abiotic stress (Choi et al., 2013). *CAChi2* encodes chitinase class II is related with osmotic stress tolerance (Hong and Hwang, 2006).

Using maize plants under salinity stress (170 mM NaCl), the effect of *Azospirillum brasilense* and *Rhizobium tropici* inoculation was studied (Fukami et al., 2018). Inoculation affected antioxidant enzymes, proline, and MDA contents in leaves and roots. The expression of genes related to antioxidant activity were up-regulated (*APX1*, *CAT1*, *SOD2*, *SOD4* in leaves, and *APX2* in roots), while the expression of pathogenesis-related genes *PR1*, *prp2*, *prp4i* and *hsp70* were down-regulated in leaves and roots. Antioxidant genes and were already mentioned. PR1 (a member of a multigene family) is a salicylic acid inducible marker gene for systemic acquired resistance (SAR) (Hussain et al., 2018). The gene *prp2* encodes β-1, 3-glucanase, and *prp4* encodes chitinase family (Fukami et al., 2017). The gene *hsp70* were already mentioned.

Examination of **Table 1** allows determining the absence of transcriptomic responses that were common among the 10 studies. One possible explanation, in addition to the small number of studies available, is that the interactions between the plant–bacteria are specific. There were no analogous responses between the different species studied. If to the above it is added that not all the studies were carried out under the same type of stress, it can be concluded that the possibility of demonstrating that a certain class of biostimulant can induce common responses in plants depends on carrying out studies with several species of plants treated with the same type of biostimulant under the same stress condition. However, common genes in some studies were related to ABA, TFs, and antioxidants.

BENEFICIAL FUNGI

This category includes fungi and arbuscular mycorrhizal fungi (AMFs), which are considered microbial inoculants along with bacteria (Dodd and Ruiz-Lozano, 2012). Among AMFs, those of the *Glomus* genus have been widely studied (Calvo et al., 2014). Fungi of the genera *Neotyphodium*, *Curvularia*, *Colletotrichum*, *Fusarium*, *Alternaria*, and *Trichoderma* have been studied as biostimulants in plants, with *Trichoderma* being the most studied (Calvo et al., 2014; Jardin, 2015; De Palma et al., 2019).

Genes and their functions are described below using the results of eight studies in which beneficial fungi and AMFs were applied to plants under abiotic stress (**Table 2**).

In lettuce plants (Jahromi et al., 2008), researchers studied the effect of applying *Glomus intraradices* AMFs under salinity stress at two levels (50 and 100 mM NaCl). Compared with plants without the application of AMFs while with stress, plants treated with *G. intraradices* while under salinity stress presented increased growth and an increase in the relative water content. Additionally, the expression of the gene of an intrinsic plasma membrane protein (*LsPIP1*) increased 60 days after transplanting; *LsPIP1* encodes a type of aquaporin in the plasma membrane and is related to the transport of water and small solutes with no charge (Wang et al., 2019).

Trichoderma longibrachiatum inoculated (10^8 CFU mL^{-1}) onto wheat seedlings under salt stress (150 mM NaCl) (Zhang et al., 2016) can enhance the plant stress tolerance. The relative water content in the leaves and roots, the chlorophyll content, the proline content and the root activity were increased, but the content of leaf malondialdehyde under salinity stress was decreased. The antioxidant enzymes SOD, POD, and CAT were increased, and the relative expression of the *SOD*, *POD*, and *CAT* genes was upregulated. The possible mechanisms by which salinity suppresses the negative effect on wheat may be due to the improvement of the antioxidant defense system. *SOD* and *CAT* are discussed above. The *POD* gene encodes a peroxidase with a variety of biological functions, including hydrogen peroxide detoxification, hormone signaling, lignin biosynthesis and stress responses (Gao et al., 2010).

In tomato seedlings under water deficit, inoculation of a *Trichoderma harzianum* biostimulant was evaluated (Mastouri et al., 2012). The enhanced redox state of inoculated plants could be explained by the increased activity of antioxidant enzymes. Additionally, *T. harzianum* modulated the expression of genes encoding antioxidant enzymes. The *MDHAR1*, *MDHAR2*, *DHARc*, *DHARp*, *GRc*, *GRp*, *Fe-SODp*, *Cu/Zn-SODp*, and *APXc* genes showed increased expression in the shoots, whereas only the *APXc* gene showed increased expression in the roots. The *MDHAR* gene encodes a monodehydroascorbate reductase, and together with *DHAR* (dehydroascorbate reductase) and *GR* (glutathione reductase), it participates

TABLE 2 | Description of genes whose transcript level changed in response to the application of fungi and AMF-based biostimulants based to plants under abiotic stress.

Gene/gene expression time	Biostimulant type	Abiotic stress type	Crop/gene expression organ	References
LsPIP1/60 days after transplant	*Glomus intraradices* (AMF)	Salinity	Lettuce/roots	Jahromi et al., 2008
MIOX1, *GLX1*, *CSD1*, *TT5*/after 0, 1, 3, and 6 days of stress	*Glomus mosseae* (AMF)	Drought	*Poncirus trifoliata*/leaves	Fan and Liu, 2011
SOD, *POD*, *CAT*/15 days of NaCl treatment	*Trichoderma longibrachiatum* (fungi)	Salinity	Wheat/leaves	Zhang et al., 2016
MDHAR1, *MDHAR2*, *DHARc*, *DHARp*, *GRc*, *GRp*, *Fe-SODp*, *Cu/Zn-SODp*, *APXc*/3 days after 50% of seeds germinated	*Trichoderma harzianum* (fungi)	Drought	Tomato/shoots and roots	Mastouri et al., 2012
RBCL, *PPH*, *Cu-Zn SOD*, *CAT*, *APX*, *GR*/6 days after stress	*Funneliformis mosseae* (AMF)	Salinity-alkalinity	Watermelon/leaves	Ye et al., 2019
LePT1, *LePT2*, *LePT3*, *LePT4*, *LePT5*/not specified	*F. mosseae* and *Rhizophagus intraradices* (AMF)	Drought	Tomato/roots	Volpe et al., 2018
XM_020345004.1, *XM_020296045.1*, *AY973229.1*/8 weeks after transplanting germinated seeds	*F. mosseae* (AMF)	Drought	Wheat/roots	Moradi Tarnabi et al., 2020
23660, *26140*, *20188*, *18003*, *4379*, *20176*, *6553*, *23129*, *29567*/10 days after stress	*Rhizophagus irregularis* (AMF)	Salinity	Asparagus/leaves	Zhang et al., 2019

in the ascorbate-glutathione cycle. This cycle plays an important role in the efficient removal of excess ROS. *APX* and *SOD* are discussed above.

The effect of *Glomus mosseae* (AMF) inoculation on the drought (3 days of water depletion) tolerance of *Poncirus trifoliata* seedlings was studied (Fan and Liu, 2011). Plants inoculated with *G. mosseae* showed an increase in growth and increased relative water and chlorophyll contents. Under drought, the inoculated plants showed increased levels of proline and increased activity of antioxidant enzymes (SOD, POD). Additionally, four genes (*MIOX1*, *GLX1*, *CSD1*, *TT5*) involved in ROS homeostasis and counteracting oxidative stress presented increased expression in inoculated plants. The *MIOX1* gene encodes myo-inositol oxygenase, which is involved in the biosynthesis of ascorbic acid; ascorbic acid has been shown to be an important antioxidant protecting plants against oxidative damage (Fan and Liu, 2011). The *GLX1* gene encodes glyoxalase I, a key enzyme involved in the glutathione-based detoxification of methylglyoxal, a product of lipid and carbohydrate metabolism. The *CSD1* gene encodes a copper/zinc SOD, and this gene is discussed above. *TT5* (transparent testa 5) encodes a chalcone isomerase involved in flavonoid synthesis. Flavonoids have an important role in the modulation of ROS levels (Fan and Liu, 2011).

Using watermelon plants under salinity-alkalinity stresses (irrigation with 400 mL 60 mM salinity-alkalinity solution [NaCl, Na_2SO_4, $NaHCO_3$, Na_2CO_3]), the effect of *Funneliformis mosseae* (AMF) was studied (Ye et al., 2019). The photosynthesis related parameters were alleviated after incubation of AMF. Under stress, the relative expression level of *RBCL*, *Cu-Zn SOD*, *CAT*, *APX*, *GR* were increased after AMF treatment. *PPH* was reduced after AMF treatment. RBCL, *Cu-Zn SOD*, *CAT*, *APX*, *GR* genes are involved in antioxidant metabolism and are discussed above.

The effect of *F. mosseae* and *R. intraradices* (AMF) on tomato plants under water stress (leaf water potential of about −0.9 and −1.0 MPa) was studied (Volpe et al., 2018). Gene expression analysis involved in inorganic phosphate uptake and transport was made. *LePT1*, *LePT2*, *LePT3*, *LePT4*, and *LePT4* changed their expression under water deficit and AMF treatment. P uptake, transfer and delivery are improved in AM roots. P fertilization can increase stress tolerance and productivity in several plant species (Sawers et al., 2017).

Using wheat plants under water deficit stress (each pot three times weekly irrigating, 25 ml per time), the effect of *F. mosseae* (AMF) was studied (Moradi Tarnabi et al., 2020). The results showed that symbiotic association between plant and AMF and irrigation not only affect transcription profile of the plant growth, but also membrane components and cell wall. The most DEGs were observed in lipid and carbohydrate metabolic process, membrane transports, cellulose synthase activity, chitinase activity, and nitrogen compound metabolic process related genes. The expression of three randomly selected DEGs were examined. The selected genes were associated to chitinase activity (*AY973229.1*), cellulose biosynthetic process (*XM_020345004.1*), and beta-glucosidase BoGH3B-like (*XM_020296045.1*). Chitin-related genes such as chitinase detect chitin molecules of fungus as a signal to trigger a defense response an increase the plant tolerance (Moradi Tarnabi et al., 2020). A stress-response related role can be considered for cellulose, since cellulose microfibrils and the other factor that lead the direction of cell growth can be regulated by water availability (Wang T. et al., 2016). Beta-glucosidase have important in cell wall biogenesis, which strongly provides protection against biotic and abiotic stress (Moradi Tarnabi et al., 2020).

The effect of *Rhizophagus irregularis* (AMF) on asparagus plants under salinity stress (100 mM NaCl) was studied (Zhang et al., 2019). The authors conducted a transcriptome analysis on leaves of garden asparagus to identify gene expression under salinity stress. 455 DEGs were identified in plants with salinity to plants with AMF and salinity. The expression profiles of 9 DEGs (23660, 26140, 20188, 18003, 4379, 20176, 6553, 23129, 29567) were by qRT-PCR. Their putative functions involved in nitrogen

metabolism, synthesis of secondary metabolites, ion homeostasis, osmotic adjustment, and scavenging ROS, among others.

The review of **Table 2** again, as in **Table 1**, reveals a small number of studies available. Except for the expression of antioxidant enzymes, where there were analogous responses between the different species studied, more studies are needed with the same species of beneficial fungus in different species of plants, under the same stress condition, to effectively demonstrate that the determination of the transcriptomic landscape will contribute to the omic definition of biostimulant.

ALGAL AND BOTANICAL EXTRACTS

Recently, marine algal extracts have been used as biostimulants, as their use in agriculture had previously been limited to being a fertilizer or a source of organic matter (Jardin, 2015). Marine macroalgae are divided into three broad groups (brown, red, and green), of which brown algae are the most widely used in agriculture; brown algae include species such as *Fucus* spp., *Sargassum* spp., *Laminaria* spp., *Turbinaria* spp., and *Ascophyllum nodosum*, this last of which is the most studied (Blunden and Gordon, 1986; Ugarte et al., 2006; Hong et al., 2007). The biostimulant function of algal extracts is commonly associated with the content of hormones such as cytokinins, auxins, abscisic acid, gibberellins, and other classes of hormone-like compounds such as sterols and polyamines (Craigie, 2011; Wally et al., 2013). However, they also contain compounds such as polysaccharides (laminarin), alginates, carrageenans, macro- and micronutrients, and nitrogenous compounds such as betaines (Khan et al., 2009; Craigie, 2011).

In the case of botanical extracts, substances extracted from plants (seeds, leaves, roots, and exudates) of various families have generally been used in agriculture as pesticides; in terms their functions as biostimulants, there is little research, thus representing an important area of opportunity (Ertani et al., 2013; Ziosi et al., 2013; Yakhin et al., 2017).

Genes and their functions are described below for 14 studies where biostimulants based on extracts of algae and botanicals were applied to plants under abiotic stress (**Table 3**).

In *A. thaliana* plants under freezing stress (−2°C), the effect of extracts (lipophilic fraction) of the brown macroalga species *A. nodosum* (1 g L^{-1}) was studied (Rayirath et al., 2009), where the expression of three cold-responsive genes (*CBF3*, *RD29A*, and *COR15A*) increased when the plants were at −2°C. The authors pointed out that in response to the application of the extract of *A. nodosum* to seedlings of *A. thaliana*, freezing tolerance was obtained by protecting the integrity of the membrane; also, compared with the untreated seedlings, the treated seedlings presented a smaller decrease in the concentration of chlorophyll as a result of stress. The *CBF3* (C-repeat-binding factor) gene is a TF that plays an important role in tolerance to low temperatures in *Arabidopsis* (Takuhara et al., 2011). *CBF3* is activated in response to low temperatures and dehydration and is independent of ABA; *CBF3* binds to the cis-element DRE-CRT (C-repeat/dehydration-responsive element) that is present in the promoter region of the *RD29A* and *COR15A* genes in *Arabidopsis*, the *WCS120* gene in wheat and the *BN115* gene in *Brassica napus* (Medina et al., 2011). Moreover, there are indications that *CBF3* regulation is carried out through responses to light quality and the circadian rhythm (Fowler et al., 2005; Franklin and Whitelam, 2007). The *RD29A* gene is discussed above. The *COR15A* (cold-regulated) gene regulates cold tolerance by stabilizing chloroplast membranes (Thalhammer et al., 2014).

In *A. thaliana* plants to which lipophilic components of an extract of *A. nodosum* (ANE), the effect of tolerance to stress by freezing (−2°C for 24 h) was studied (Nair et al., 2012). Expression of the *P5CS1*, *P5CS2*, and *ProDH* genes was detected; the expression of *P5CS1* and *P5CS2* increased, while that of *ProDH* decreased. These genes are involved in proline synthesis (*P5CS1* and *P5CS2*) and degradation (*ProDH*). *P5CS1* and *P5CS2* encode delta1-pyrroline-5-carboxylate synthase enzymes that regulates the rate of proline biosynthesis. This gene is expressed in some tissues under normal conditions and throughout the plant under conditions of water deficit, in addition to being induced by ABA and salt stress (Kesari et al., 2012). *ProDH* catalyzes the degradation of proline to produce glutamic acid; this gene is related to decreasing the oxidative burst and cell death associated with the hypersensitive response (Cecchini et al., 2011).

In *A. thaliana* plants under drought stress (absence of a hydroponic solution for 4 days), the effect of an extract of *A. nodosum* was evaluated (Santaniello et al., 2017); the extract positively influenced the survival of the plants. The plants under stress and treated with the extract presented increased expression of the *NCED3*, *MYB60*, *RAB18*, *RD29A*, *RbCS1A*, *RCA*, *PIP1;2*, β*CA1*, *PsbS*, *VDE*, *DFR*, *APX2*, and *SOD* genes 4 days after stress. These genes are involved in the pathways of the antioxidant system and are ABA dependent. The *MYB60* gene is related to the regulation of stomatal movement, and the expression of this gene increases with low levels of ABA; additionally, in an initial drought state, this gene can induce root growth. In contrast, in a severe state of drought, its expression is inhibited, resulting in stomatal closure and a decrease in root growth (Oh et al., 2011). The *RAB18* gene (which is sensitive to ABA) encodes a glycine-rich hydrophilic protein (Hoque et al., 2012) that belongs to the LEA family of proteins and has a dehydrin function (Hernández-Sánchez et al., 2019). *RAB18* protects membranes under dehydration conditions by binding to anionic phospholipids through electrostatic forces (Eriksson and Harryson, 2011), in addition to binding to other proteins to prevent their denaturation (Graether and Boddington, 2014). The *RD29A* gene is discussed above. Similarly, the RCA gene encodes the Rubisco activase enzyme, which is a chloroplastic enzyme encoded in the nucleus and participates in the activation of Rubisco (Hasse et al., 2015); this gene is also related to jasmonate-induced leaf senescence (Elizabete Carmo-Silva and Salvucci, 2013). Two genes related to the regulation of mesophilic diffusion restriction include *PIP1;2* (discussed above) and β*CA1* (β-carbonic anhydrase 1), both of which participate in carboxylation or decarboxylation reactions related to photosynthesis and respiration; as such, they play an important role in the catalysis of CO_2 and water to form protons and bicarbonate (Hu et al., 2015). The *PsbS* (photosystem II subunit

TABLE 3 | Description of genes whose transcript level changed in response to the application of biostimulants based on extracts of algae and botanicals to plants under abiotic stress.

Gene/gene expression time	Biostimulant type	Abiotic stress type	Crop/gene expression organ	References
AtDREB2a, *AtRD29*, *AtNFYA1*, *AtNFYA2*, *AtUBC24*, *AtWAK2*, *AtSYG1*, *At3g27150*/6 and 12 h after treatment	Extract of *A. nodosum*	Salinity	*Arabidopsis*/whole plant	Shukla et al., 2018
LEA3, *LEA1*, *CCA1*, *HVA22d*, *LTP6*, *DREB1A*, *DREB1C*, *LEA2*, *VA22b*, *Di21*, *SnRK2*, *CIPK25*, *LTP*, *RING*, *MYB*/1 and 5 days after treatment	Extract of *A. nodosum*	Salinity	*Arabidopsis*/whole plant	Jithesh et al., 2018
ANN1, *ANN2*, *PIP1*, *P5CS1*, *CHS*, *APX1*, *GPX3*/not specified	Extract of *A. nodosum* and 5-aminolevulinic acid (foliar application)	Salinity	*Asparagus aethiopicus*/leaves	Al-Ghamdi and Elansary, 2018
COR15A, *RD29A*, *CBF3*/when plants were exposed at −2°C	Extract of *Ascophyllum nodosum*	Cold	*Arabidopsis*/leaves	Rayirath et al., 2009
P5CS1, *P5CS2*, *PRODH*/not specified	*A. nodosum* lipophilic components	Freezing	*Arabidopsis*/leaves	Nair et al., 2012
NCED3, *MYB60*, *RAB18*, *RD29A*, *RbCS1A*, *RCA*, *PIP1;2*, B*ca1*, *PsbS*, *VDE*, *DFR*, *APX2*, *SOD*/From the moment of stress until the fourth day	Extract of *A. nodosum*	Drought	*Arabidopsis*/leaves	Santaniello et al., 2017
GmCYP707A1a, *GmCYP707A3b*, *GmRD22*, *GmDREB1B*, *GmERD1*, *FIB1a*, *GmPIP1b*, *GmGST*, *GmBIP*, *GmTp55*/75 h and 59 h after treatment	Extract of *A. nodosum*	Drought	Soybean/leaves	Shukla et al., 2017
TAS14/7th day of stress	Extract of *A. nodosum* (foliar application)	Drought	Tomato/leaves	Goñi et al., 2018
TaNCED3.1, *TaNCED3.2*, *TaHAI1*, *DREB family*, *TaRap2-4*, *TaMYB31*, *TaAREB3*, *dehydrin*, *TaP5CS1*, *TaNCED4*, *TaNCED2*, *TaERD2*/6 days after stress	Extract of *Gracilaria dura*	Drought	Wheat/leaves	Sharma et al., 2019
GRMZM2G439784, *GRMZM2G324221*, *GRMZM2G164129*, *GRMZM2G163866*/10 days after treatment	Extract of *Kappaphycus alvarezii*	Drought	Maize/roots	Kumar et al., 2019
SOD, *CAT*, *APX*, *DAHR*, *GR*, *PrxQ*/not specified	Licorice root extract	Salinity	Pea (*Pisum sativum* L.)/leaves	Desoky et al., 2019
DtDREB2A, *DtMYB30*, *DtNAC019*, *DtNAC72*, *DtNAC19*, *DtNAC69*, *DtZIP63*, *DtABF3*, *DtHB12*, *DtHB7*/2, 4, 6, 9, and 24 h from exposure to stress	Extract of borage	Salinity	Wild rocket/leaves	Franzoni et al., 2019
CuZnSOD, *MnSOD*, *CAT*, *FeSOD*/24 and 48 h after stress	Lignin derivatives, plant-derived aminoacids and molybdenum	Heat	Cucumber/seeds	Campobenedetto et al., 2020
ObOLP/3-months old plants	*Moringa oleifera* extract	Salinity	*Ocimum basilicum*/leaves	Alkuwayti et al., 2020

S) and *VDE* (violaxanthin depoxidase) genes are involved in the photoprotection mechanism to avoid damage caused by oxidative stress in plants due to excess energy from sunlight (Fufezan et al., 2012; Ciszak et al., 2015). The *DFR* (dihydroflavonol reductase) gene is involved in the biosynthesis of anthocyanins, which protect plants from stress through their activity of ROS detoxification (Cui et al., 2014). The *APX2* and *SOD* genes are discussed above.

Researchers (Shukla et al., 2017) studied the effect of an extract of *A. nodosum* on soybean plants under drought stress (without irrigation). Compared with the untreated plants, the treated plants had higher relative water content, antioxidant activity, and stomatal conductance under drought stress. In addition, there were changes in the expression of the *GmCYP707A1a*, *GmCYP707A3b*, *GmRD22*, *GmDREB1B*, *GmERD1*, *GmNFYA3*, *FIB1a*, *GmPIP1b*, *GmGST*, *GmBIP*, and *GmTp55* genes at 75 h (stress) and 89 h (recovery) after treatment. The *GmCYP707A1a* and *GmCYP707A3b* genes encode ABA 8′-hydroxylases, which participate in the regulation of ABA levels during dehydration and rehydration (Umezawa et al., 2006; Zheng et al., 2012a). *GmRD22* is discussed above. The *GmDREB1B* (dehydration response element-binding) gene belongs to a family of TFs induced by drought and salinity, and this gene is ABA dependent. This TF binds to the cis-element DRE-CRT, which is present in the promoter of the *COR* and *RD29A* genes, both of which are related to the abiotic stress response (Tuteja, 2007). *GmERD1* is discussed above. The *FIB1a* gene improves the phototolerance of photosystem II (PSII) (Mutava et al., 2015), and the *GmPIP1b* gene is also discussed above. The *GmGST* gene, via its glutathione reduction potential, detoxifies ROS and protects cells from oxidative damage (Mcgonigle et al., 2000). The *GmBIP* gene encodes a chaperone protein that is related to delayed senescence in leaves and therefore increases tolerance to drought stress (Carvalho et al., 2014). The *GmTp55* gene

encodes an aldehyde dehydrogenase enzyme that reduces reactive aldehydes derived from lipid peroxidation under oxidative stress (Wang et al., 2017).

Using tomato plants under drought stress, researchers (Goñi et al., 2018) studied the effect of several commercial products based on extracts of *A. nosodum*. All of the ANEs affected drought stress tolerance but to different degrees. Regulation of the *TAS14* gene was reported 7 days after stress, and this gene was differentially overexpressed in response to applications of all extracts. *TAS14* encodes a group 2 LEA protein called dehydrin, which is induced by osmotic stress and ABA (Godoy et al., 1994). When this gene is overexpressed, long-term tolerance to drought and salinity is achieved through the reduction in osmotic potential and the accumulation of sugars and potassium (Muñoz-Mayor et al., 2012).

Using wheat plants under water stress (no water for 10 days), researchers (Sharma et al., 2019) evaluated foliar applications of *Gracilaria dura* (red algae) sap. The expression levels of the *TaNCED3.1*, *TaNCED3.2*, *TaHAI1*, *DREB*, *TaRap2-4*, *TaMYB31*, *TaAREB3*, *dehydrin*, *TaP5CS1*, *TaNCED4*, *TaNCED2*, and *TaCla013224* genes were analyzed, which increased on the sixth day of the onset of stress. *TaNCED3.1*, *TaNCED3.2*, *TaNCED4*, and *TaNCED2* belong to the *NCED* family of genes, which encode 9-cis-epoxycarotenoid dioxygenases; these are key enzymes involved in ABA biosynthesis and are regulated in response to drought and salinity (Behnam et al., 2013). *TaHAI1* encodes a member of the PP2C family, whose members includes class A and type 2C phosphatase proteins and are related to the downregulation of osmotic stress and ABA signaling (Nguyen et al., 2019). *DREBs* are discussed above. *TaRap2-4* encodes a DREB-subfamily TF related to light mediation and ethylene signaling (Lin et al., 2008). *TaMYB31* belongs to a subfamily of TFs, is ABA dependent and plays an important role in the development of and defense response in plants (Yanhui et al., 2006; Tuteja, 2007). *TaAREB3* is also a TF that is sensitive to ABA and is related to stomatal movement and ROS generation in response to ABA (Wang et al., 2013). *Dehydrins* are LEA-like proteins (as mentioned above), and *TaP5CS1* participates in proline biosynthesis (as mentioned above). *TaERD2* encodes an HSP70-type chaperonin that is synthesized in response to stress and is the main chaperone in maintaining protein homeostasis (Rowarth et al., 2019).

The effect of *Kappaphycus alvarezii* extract applied to the soil was analyzed on maize plants under drought stress (no irrigation for 10 days) (Kumar et al., 2019). To obtain a global view of the effect, the authors analyzed the transcriptome of the roots of the plants 10 days after the application of the treatment. A total of 896 upregulated genes and 533 downregulated genes were differentially expressed. However, only 4 genes were overexpressed in response to the application of the extract and stress compared to the application of the extract only, and 18 genes were repressed. The overexpressed genes included *GRMZM2G439784*, *GRMZM2G324221*, *GRMZM2G164129*, and *GRMZM2G163866*. The *GRMZM2G439784* gene encodes an LRR-type kinase, which belong to the receptor kinase subfamily, and its function lies in communication between cells to transmit signals during development and before environmental stimuli to activate defense; it is of the utmost importance in resistance against pathogens (Antolín-Llovera et al., 2014; Dievart et al., 2015). *GRMZM2G324221* encodes a structural protein of the small ribosomal subunit (40S) and participates in the regulation of response to virus infections (Li, 2019). The function of *GRMZM2G164129* has not been characterized, whereas *GRMZM2G163866* is a high-affinity nitrate transporter; overexpression of the latter could translate into an improvement in nitrogen metabolism (Liu X. et al., 2014).

Using *A. thaliana* plants under saline conditions (150 mM NaCl), researchers (Shukla et al., 2018) studied the effects of an extract of *A. nodosum* by measuring the expression of several microRNAs (miRNAs) and their target genes. The miRNA *miR396a-5p* was downregulated, which repressed the expression of the *AtGRF7* gene. In turn, this repression positively regulated the expression of *AtDREB2a* and *AtRD29*; the greater expression of the *AtDREB2a* and *AtRD29* genes resulted in tolerance to salt stress. *AtDREB2a* and *AtRD29* are discussed above. The expression of the miRNA *ath-miR169g-5P* also increased, and as a result, the expression of the target nuclear factors *AtNFYA1* and *AtNFYA2* increased. *AtNFYA1* is associated with hypersensitivity to salt stress and ABA during the early stages of post-germination growth (Li et al., 2013). *AtNFYA2* participates in nitrogen metabolism, regulation of light signaling, and chloroplast biogenesis (Laloum et al., 2013; Petroni et al., 2013). The expression of the miRNAs *ath-miR399*, ath-*miR827*, and *ath-miR2111b* as well as their target genes *AtUBC24*, *AtWAK2*, *AtSYG1*, and *At3g27150* was also altered, suggesting a role of the extract of *A. nodosum* in phosphate homeostasis (Liu et al., 2012; Shukla et al., 2018).

Using *A. thaliana* plants under salinity stress (150 mM NaCl), researchers (Jithesh et al., 2018) studied the effect of extracts of *A. nodosum* subfractionated with ethyl acetate. The transcriptome of the plants was analyzed under salinity and in response to the application of the extract on the first and fifth days after salinity. On the first day, the expression of the *LEA3*, *CCA1*, *LEA1*, *HVA22d*, *LTP6*, *ATGOLS3*, *ATGOLS2*, *DREB1A*, and *DREB1C* genes increased. The products of *LEA* genes belong to a family of hydrophilic proteins with a protective function of functional proteins (Tunnacliffe and Wise, 2007) to prevent their aggregation (Chakrabortee et al., 2007). The *LEA3* gene is induced by drought and salinity (Du et al., 2016), and *LEA1* is induced by wounds and mild stress (Dunaeva and Adamska, 2001). *CCA1* has a regulatory function in the circadian rhythm and controls various processes, such as the stomatal opening (Hassidim et al., 2017). *HVA22d* is induced by ABA and by stress and is related to the regulation of autophagy (Chen et al., 2002, 2009). *LTP6* is predicted to encode a protein related to pathogenesis (PR) (Le et al., 2014). *ATGOLS3* and *ATGOLS2* encode the enzyme galactinol synthase, which is key to the biosynthesis of the oligosaccharides of the raffinose family and are related to tolerance to drought, salinity, and cold stress (Taji et al., 2002; Nishizawa et al., 2008). *DREB1A* and *DREB1C* are TFs of the DREB family, which is discussed above. On the fifth day after salinity, the expression of the *LEA1*, *LEA2*, *VA22b*, *Di21*, *SnRK2*, *CIPK25*, *LTP*, *RING*, and *MYB* genes increased. *VA22b* and *Di21* regulate stress and are induced by

ABA (Jithesh et al., 2018). *SnRK2* is discussed above. *CIPK25* is a serine-threonine protein kinase that interacts with calcineurin (a calcium sensor); both of these are related to signal transduction in response to environmental stress (Kanwar et al., 2014). *LTP* is discussed above. *RING* and *MYBs* are TFs; *RING* acts as a ligase and participates in the regulation of gene expression in response to environmental or hormone signals (Qin et al., 2008), and *MYBs* are discussed above.

Using *Asparagus aethiopicus* plants subjected to salinity stress (2,000 and 4,000 mg L^{-1} NaCl), researchers (Al-Ghamdi and Elansary, 2018) studied the synergistic effects of the application of a commercial product based on *A. nodosum* and 5-aminolevulinic acid (ALA) applied via foliar treatment. ALA is a precursor of porphyrins, so it has effects on the photosynthetic apparatus, thus stimulating the defense system in plants (Wu et al., 2018). The application of both products caused a synergistic effect on plant growth associated with increased expression of the *ANN1*, *ANN2*, and *PIP1* genes, which are associated with the transport of water and Ca^{+2}; the *P5CS1* and *CHS* genes, which are related to the production of secondary metabolites; and the *APX1* and *GPX3* genes, which are associated with antioxidant metabolism in plants. *ANN1* and *ANN2* encode a family of proteins called annexins, which are Ca^{+2} transporter permeases attached to the plasma membrane. They participate in processes such as stomatal closure, adaptation to stress, and cell signaling (Laohavisit et al., 2012; Wang J. et al., 2018). *PIP1* and *P5CS1* are discussed above. *CHS* encodes a chalcone synthase, which is a key enzyme involved in flavonoid biosynthesis and is also involved in auxin transport (Brown et al., 2001; Dana et al., 2006). *APX1* and *GPX3* play a role in antioxidant metabolism, and *GPX3* acts as a redox transducer whose function is similar to that of *APX1* in H_2O_2 homeostasis and is related to ABA signal transduction during stress (Miao et al., 2006).

Using pea plants under salinity stress (150 mM NaCl for 2 weeks), researchers (Desoky et al., 2019) studied the effect of applying licorice (*Glycyrrhiza glabra*) root extract on pea seeds. Saline stress reduced seedling growth and increased oxidative stress; however, in pretreated seedlings, mitigation of these effects was observed. Treatment with the extract also increased the transcription of the *CAT*, *SOD*, *APX*, *GR*, *DHAR*, and *PrxQ* genes, decreasing oxidative stress. *CAT*, *SOD*, *APX*, *GR*, *DHAR*, and *PrxQ* (peroxiredoxin) participate in the cellular antioxidant system, which maintains ROS homeostasis to mitigate oxidative damage. However, ROS are essential for maintaining metabolic flow and activating acclimation responses to stress through systemic signaling (Ahmad et al., 2010; Suzuki et al., 2012).

Using wild rocket (*Diplotaxis tenuifolia* L.) plants under salinity stress (200 mM NaCl), researchers (Franzoni et al., 2019) studied the effect of a foliar application of a borage (*Borago officinalis*) extract. The expression of several TFs related to salinity stress was studied at 2, 4, 6, 9, and 24 h after exposure to stress. TFs such as *DtDREB2A*, *DtMYB30*, *DtNAC019*, *DtNAC72*, *DtNAC19*, *DtNAC69*, *DtZIP63*, *DtABF3*, *DtHB12*, and *DtHB7* presented positive regulation. *DtDREB2A* is discussed above. *DtMYB30* is an ABA-sensitive TF that participates in processes such as germination and response to stress (Zheng et al., 2012b). *DtNAC019*, *DtNAC72*, *DtNAC19*, and *DtNAC69* compose a family of TFs related to development and stress responses in plants, respond to ABA, and promote the antioxidant system (Xu et al., 2015). *DtZIP63* is a TF that regulates the circadian cycle through a low-energy response and is activated by a kinase (SnRK1) (Frank et al., 2018). *DtABF3* is a TF induced by ABA and osmotic stress (Bogamuwa and Jang, 2014), and *DtHB12* and *DtHB7* are ABA-dependent TFs and act by mediating the growth response to water stress (Olsson et al., 2004).

Using cucumber plants (*Cucumis sativus* L.) subjected to heat stress (35°C), researchers (Campobenedetto et al., 2020) studied the effect of a seed application of a biostimulant based on lignin derivatives and containing plant-derived amino acids and molybdenum (KIEM®). The application of the biostimulant increased the percent germination, fresh biomass, and increased in expression levels *RBOHD*, *CuZnSOD*, *MnSOD*, *CAT*, and *GST* genes, while *FeSOD* gene was decreased. *CuZnSOD*, *MnSOD*, *CAT*, *GST*, and *FeSOD* are ROS-scavenging enzymes and are discussed above. *RBOHD* is a ROS-producing enzyme, H_2O_2 is a relatively long-lived ROS, and its accumulation is caused by the induction of membrane-bound respiratory burst oxidase homolog proteins (Rboh), which are important players in abiotic stress responses (Huang et al., 2014).

The effect of *Moringa oleifera* leaves extract applied on *Ocimum basilicum* plants under salt stress (1,000 mg L^{-1}) was studied (Alkuwayti et al., 2020). The application of *M. oleifera* extract altered the expression of *ObOLP* gene and was positively correlated with the plant growth and yield enhancement. *ObOLP* is a osmotin-like proteins, members of the pathogenesis-related protein 5 (PR-5), which are produced in plants under different abiotic and biotic stresses (Mayer Weber et al., 2014). Under salinity and drought stress, OLP maintains cellular osmolarity by compartmentalization of solutes or by structural and metabolic alterations (Chowdhury et al., 2017).

Examination of **Table 3** allows determining a few of transcriptomic responses that were common among the 14 studies. Common genes in some studies were related to ABA, TFs, antioxidants, and LEA proteins. Most of the studies were carried out in *A. nodosum* in different types of stress. Regarding algae extracts, most of the studies were carried out on *A. nodosum* under different types of stress. However, in the case of botanical extracts, there are few related studies. More studies are needed with extracts of the same species of algal and botanical in different species of plants, under the same stress condition.

PROTEIN HYDROLYSATES AND OTHER NITROGEN-CONTAINING COMPOUNDS

The two main categories of protein-based products are divided into: (1) protein hydrolysates consisting of mixtures of peptides and amino acids that can be of animal or plant origin and (2) individual amino acids such as proline, glutamate, glutamine, and glycine betaine (Calvo et al., 2014). Mixtures of amino acids and peptides can be obtained by chemical, enzymatic, or thermal hydrolysis from byproducts of plant and animal origin (Calvo et al., 2014; Jardin, 2015). Chemical synthesis can be used to produce amino acids or mixtures of these. This

TABLE 4 | Description of genes whose transcript level changed in response to the application of biostimulants based on protein hydrolysates and other nitrogen-containing compounds to plants under abiotic stress.

Gene/gene expression time	Biostimulant type	Abiotic stress type	Crop/gene expression organ	References
Solyc02g084840, *Solyc03g025810*/3 to 11 days after application of treatment	A mixture of amino acids, proteins, vitamins, and betaines	Drought	Tomato/leaves	Petrozza et al., 2014a
ZmPAL/12 days after stress	Hydrolysate of alfalfa plants, triacontanol, and indole 3-acetic acid	Salinity	Maize/leaves	Ertani and Schiavon, 2013
ZmNRT2.1, *ZmNRT2.2*, *ZmNRT2.3*, *ZmNAR2.2*, *ZmNRT1.1*, *ZmNRT1b*, *ZmNRT*, *ZmSOD1A*, *ZmRbohA*, *ZmRbohC*/3 days after stress	Protein hydrolysate	Hypoxia, salinity, nutrient deficiency	Maize/roots	Trevisan et al., 2019
BI-1, *MAPK1*, *WRKY53*, *CAT1*, *GPX*, *SOD*, *ATG1*, *ATG2*, *ATG4*, *ATG6*, *ATG7*, *ATG8*, *ATG9*, *ATG10*, *ATG13*/not specified	Panchagavya	Salinity	Rice/leaves	Khan et al., 2018
CAT, *MnSOD*, *WRKY53*, *BI-1*/not specified	Panchagavya	Salinity	Rice/leaves	Khan et al., 2017
Cla018095, *Cla010664*, *Cla004567*, *Cla009820*, *Cla012125*, *Cla003187*, *Cla021039*, *Cla007826*, *Cla013224*, *Cla001877*, *Cla001590*, *Cla006037*/not specified	Melatonin	Vanadium toxicity	Watermelon/roots	Nawaz et al., 2018
HEMA1, *CHLH*, *POR*/10 days after treatment	5-Aminolevulinic acid	Salinity	Cucumber/roots	Wu et al., 2018

category also includes other nitrogenous molecules, such as polyamines, betaines, and non-protein amino acids (Vranova et al., 2011), which are diverse in higher plants but are poorly characterized in terms of their physiological and ecological functions (Vranova et al., 2011).

Genes and their function are described below based on the results of seven studies in which biostimulants based on protein hydrolysates and other nitrogen-containing compounds were applied to plants under abiotic stress (**Table 4**).

Using tomato plants under drought stress (no water after the development of the third or fourth leaf), researchers (Petrozza et al., 2014b) evaluated the foliar application of a commercial product based on a complex mixture of vitamins, amino acids, proteins, and betaines. The results indicated that, compared with the untreated plants, the plants treated with the product while under stress had higher growth and stress tolerance. Additionally, the expression of the *Solyc02g084840* and *Solyc03g025810* genes was studied for 14 days after treatment; these genes were expressed between days 3 and 11. The *Solyc02g084840* gene is an ortholog of the *Arabidopsis RAB18* gene that is discussed above. *Solyc03g025810* is an ortholog of the *RD29B* gene, which encodes a protein that is induced in response to water deprivation, low temperature, salinity, and desiccation, and its response is mediated by ABA (Bihmidine et al., 2013).

Using maize plants under salinity stress (25, 75, and 150 mM NaCl), researchers (Ertani and Schiavon, 2013) evaluated a biostimulant based on a hydrolysate of alfalfa (*Medicago sativa* L.) plants, triacontanol (TRIA), and indole-3-acetic acid (IAA) added to the irrigation water for 48 h. Compared with untreated plants under stress, the treated plants under stress had greater biomass and presented greater activity of enzymes related to nitrogen metabolism. Plants treated with the biostimulant while under one of 3 levels of salinity stress presented increased expression of the *ZmPAL* gene at 12 days after stress. The *ZmPAL* gene encodes the enzyme phenylalanine ammonium lyase (PAL), which is a key enzyme that catalyzes the first step of the phenylpropanoid pathway, producing precursors of a wide variety of vital secondary metabolites related to plant defense, such as lignin, flavonoids, isoflavonoids, coumarin, and stilbenes (Huang et al., 2010).

Using maize plants treated with a commercial biostimulant based on a protein hydrolysate added to the hydroponic solution, researchers (Trevisan et al., 2019) evaluated the tolerance to three types of abiotic stress: hypoxia (deprivation of air bubbles in the liquid hydroponic solution), salinity (25 mM NaCl), nutrient deficiency (only distilled water was supplied in the hydroponic solution) and the combination of these. The treated plants had increased root and shoot growth and increased tolerance to single and combined stress conditions. Additionally, genes related to nitrate transport (*ZmNRT2.1*, *ZmNRT2.2*, *ZmNRT2.3*, *ZmNAR2.2*, *ZmNRT1.1*, *ZmNRT1b*, and *ZmNRT*) and ROS metabolism (*ZmSOD1A*) were expressed. The *ZmNRT2.1*, *ZmNRT2.2*, *ZmNRT2.3*, and *ZmNAR2.2* genes belong to the high-affinity nitrate transport system, and the *ZmNRT1.1*, *ZmNRT1b*, *ZmNRT* genes belong to the low-affinity nitrate transport system. An increase in the expression of these genes occurs since the application of protein hydrolysates in plants can modulate the expression of critical genes involved in the assimilation of nitrogen (transporters) (Sestili et al., 2018). *ZmSOD1A* is discussed above.

Using two rice cultivars (one susceptible and one tolerant to salinity) under salinity stress (100 mM NaCl), researchers (Khan et al., 2018) studied the effects of a natural biostimulant called panchagavya (a mixture of milk, butter, curd, urine, and cow dung) applied as a soil drench. The results indicated that the treated plants under salinity stress showed an improvement in

the physiological and biochemical characteristics and presented increased expression of several genes: the expression of *BI-1*, *MAPK1*, and *WRKY53* increased in the tolerant variety; the expression of *CAT1*, *GPX*, and *SOD* increased in the susceptible variety; and the expression of *ATG1*, *ATG2*, *ATG4*, *ATG6*, *ATG7*, *ATG8*, *ATG9*, *ATG10*, and *ATG13* increased in both varieties. The *BI-1* gene is expressed during senescence and under various stress conditions, and its expression gradually decreases throughout the cell death process (Ishikawa et al., 2015). *MAPK1* belongs to a family of kinases (mitogen-activated protein kinases) that regulate the response to abiotic and biotic stress via signaling cascades (Neupane et al., 2019). *WRKY53* is an early response factor to water deficit; its expression regulates the stomatal response (Sun and Yu, 2015). The *CAT1*, *GPX*, and *SOD* genes are discussed above. The *ATG1*, *ATG2*, *ATG4*, *ATG6*, *ATG7*, *ATG8*, *ATG9*, *ATG10*, and *ATG13* genes are related to the autophagy process; together with PCD, this process regulates responses to stress since it is essential to degrade oxidized proteins during oxidative stress (Xiong et al., 2007).

The effect of the panchagavya biostimulant amendment to the soil drench applied to rice plants under salinity stress (100 mM NaCl) has been studied previously (Khan et al., 2017), and the results showed that stressed and treated plants presented upregulated *CAT*, *MnSOD*, *WRKY53*, and *BI-1* gene expression. All these genes are discussed above.

In watermelon plants under stress from exposure to vanadium (V) (50 mg L^{-1}), the application of melatonin added to the irrigation solution was studied (Nawaz et al., 2018). The results indicated that treatment with melatonin reduced the concentrations of V in the leaves and stems and reduced the concentrations of H_2O_2 and malondialdehyde (MDA). The expression of the *Cla018095*, *Cla010664*, *Cla004567*, *Cla009820*, *Cla012125*, *Cla003187*, *Cla021039*, *Cla007826*, *Cla013224*, *Cla001877*, and *Cla001590* genes also increased. The *Cla018095* gene is related to chlorophyll biosynthesis. *Cla010664* encodes an O-methyl transferase, and *Cla004567* encodes an S-methyl transferase, which participates in the biosynthesis of melatonin. *Cla010664* can eliminate ROS or activate antioxidant enzymes such as SOD and CAT (Manchester et al., 2015). The *Cla009820* (superoxide dismutase), *Cla012125* (superoxide dismutase), *Cla003187* (peroxidase), *Cla021039* (glutathione peroxidase), *Cla013224* and *Cla007826* (glutathione S-transferase) genes encode enzymes with antioxidant activity. *Cla001877* is a respiratory burst oxidase and can integrate Ca^{+2} signaling and protein phosphorylation with ROS production, the last being key to the regulation of growth, development, responses to environmental stimuli, and cell death (Suzuki et al., 2011). *Cla001590* encodes a V-dependent haloperoxidase that may be related to the absorption of inorganic forms of iodine in plants, although its function has not been fully defined (Smoleñ et al., 2019).

Using cucumber plants under salinity stress (50 mmol L^{-1} NaCl), researches (Wu et al., 2018) evaluated a biostimulant based on ALA applied via foliar sprays. Plants exhibited increased photosynthesis (increased plant height and leaf area, increased gas exchange capacity, increased the use of light by photosystem II and improved chlorophyll biosynthesis) in response to the application of ALA and low stress conditions. There was an increase in the expression of the *HEMA1*, *CHLH*, and *POR* genes at 10 days after treatment. *HEMA1* encodes glutamyl-tRNA reductase, *CHLH* encodes Mg-chelatase, and *POR* encodes protochlorophyllide oxidoreductase. All three genes are involved in chlorophyll biosynthesis (Stephenson and Terry, 2008). As the chlorophyll content increased in response to the application of ALA, the tolerance of cucumber plants to stress also increased.

The review of **Table 4** reveals a small number of studies available (7). Except for the expression of antioxidant enzymes, where there were analogous responses between the different species studied, more studies are needed with the same proteins, hydrolysates, and other nitrogen-containing compounds in different species of plants, under the same stress condition.

HUMIC ACID AND FULVIC ACID

Humic substances are natural components of soil organic matter resulting from the decomposition of animal, plant, and microbial waste (Jardin, 2015). They are considered the main components of soil organic matter and are the most abundant natural organic compounds on Earth (Calvo et al., 2014). Humic substances are heterogeneous compounds classified by their molecular weight and solubility in (1) humic acids, which are soluble in basic media, (2) fulvic acids, which are soluble in alkaline and acidic media, and (3) humins, which are not extractable from the soil (Berbara and García, 2014). In addition to their use as biostimulants, humic substances are related to key processes in the soil and plants, such as carbon and oxygen exchange between the soil and the atmosphere, availability of nutrients, and the detoxification and transport of toxic substances (Piccolo and Spiteller, 2003).

In this category, six studies related to humic and fulvic acid substances was found, where gene expression was analyzed (**Table 5**).

Using maize plants, researchers (Schiavon et al., 2010) evaluated a high-molecular-weight (>3,500 Da) humic fraction from *Nicodrilus caliginosus* feces, which was added for 48 h hydroponically. The effect on phenylpropanoid metabolism was subsequently evaluated, where the expression of the *ZmPAL1* gene increased considerably in plants treated with the humic fraction at three different concentrations. The *ZmPAL1* gene is discussed above.

Using *Arabidopsis* plants under heat stress (45°C), the application of humic acid (commercial product) was studied (Cha et al., 2020). The authors performed a transcriptomic analysis to identify the HA-prompted molecular mechanisms. Gene ontology analysis indicated that humic acid up-regulates diverse genes related in the response to stress. Heat stress causes induction in gene families such as heat-shock protein (HSP), coding genes including *HSP101*, *HSP81.1*, *HSP26.5*, *HSP23.6*, and *HSP17.6A*. HSPs function as molecular chaperones to protect against thermal denaturation of substrates and stimulate refolding of denatured substrates, also play an important role in maintaining cell membrane integrity,

TABLE 5 | Description of genes whose transcript level changed in response to the application of biostimulants based on humic acids and chitosan and other biopolymers to plants under abiotic stress.

Gene/gene expression time	Biostimulant type	Abiotic stress type	Crop/gene expression organ	References
ZmPAL1/48 h after treatment	Hummus of *Nicodrilus caliginosus*	None	Maize/leaves	Schiavon et al., 2010
HSP101, *HSP81.1*, *HSP26.5*, *HSP23.6*, *HSP17.6A*/9h after application of humic acid	Humic acids	Heat	*Arabidopsis*/not specified	Cha et al., 2020
SnRK2.2, *MDH*, *WRKY DNA-binding transcription factors family*/6 days after stress	Humic acids (vermicompost)	Weak acids	Maize/root	Baía et al., 2020
SETIT_021707mg, *SETIT_016840mg*, *SETIT_ 015030mg*, *SETIT_004913mg*, *SETIT_016654mg*/5 days after stress	Humic acid	Drought	Foxtail millet/leaves	Shen et al., 2020
GME, *AO*, *ALDH*, *GST*, *G6PDH*, *CYS-GYL*, *C4H*, *CHS*, *F3′5′H*, *F3H*/4 and 8 days under stress	Fulvic acid	Drought	*Camellia sinensis*/shoots	Sun et al., 2020
CCoAOMT, *CAB37*, *AGT1*, *At4g26520*/0, 4, 8, and 12 days after stress	Fulvic acid	Drought	*Paeonia ostii*/leaves	Fang et al., 2020
SOD, *JA*/48 h after treatment	nCu-chitosan-PVA and chitosan-PVA complex	Salinity	Tomato/leaves	Hernández-Hernández et al., 2018
AOX1/3 days after treatment	Chitosan	Salinity	Maize/leaves	Turk, 2019
CAT3/24 h after treatment	Xyloglucan oligosaccharides	Salinity	*Arabidopsis*/leaves	González-Pérez et al., 2018
GH3/24 h after treatment	Chitosan oligosaccharide	Cold	*Camellia sinensis*/leaves	Li et al., 2020e
CAT, *APX*, *POD*, *SOD*, *GmSALT3*, *CHS*/12 days of growth	Chitosan modified biochar	Salinity	Soybean/plant tissue	Mehmood et al., 2020

ROS scavenging and production of antioxidants, osmolytes (Khan and Shahwar, 2020).

Using maize seedlings under weak acids stress (acetic and salicylic acids), the application of humic acids extracted from vermicompost produced with cattle manure was studied (Baía et al., 2020). Humic acids decrease the intracellular pH and produce high level of *SnRK2.2* and *MDH* genes, and low level of *WRKY* TFs family. *SnRK2.2* is discussed above. *MDH* encodes malate dehydrogenase, this enzyme is critical in malate metabolism and is related in ROS producing genes (Akbar et al., 2020). *WRKY* TF family are associated to transduction of stress signaling and play a major role in plant defense to abiotic and biotic stress (Li et al., 2020c).

The effect of humic acid in foxtail millet plants under drought conditions (natural simulation conditions) was studied (Shen et al., 2020). Transcriptome sequencing and RT-qPCR was performed on plants. Humic acid caused a significant increase in the yield, dry weight and root-shoot ratio. *SETIT_021707mg*, *SETIT_016840mg*, and *SETIT_ 015030mg* genes were significantly up-regulated, while *SETIT_004913mg* and *SETIT_016654mg* genes were significantly down-regulated in the plants treated with humic acid and drought. These genes are related with metabolic pathways, secondary metabolite biosynthesis and starch and sucrose metabolism.

The effect of fulvic acid in tea plants (*Camellia sinensis*) under drought stress was studied (Sun et al., 2020). The authors examined the transcriptomics and metabolomics profiles, 604 and 3331 differentially expressed metabolite genes (DEGs) were found in plants at 4 and 8 days under drought respectively. DEGs are related in diverse biological processes such as ascorbate metabolism (*GME*, *AO*, *ALDH*), glutathione metabolism (*GST*, *G6PDH*, *CYS-GYL*), and flavonoids biosynthesis (*C4H*, *CHS*, *F3′5′H*, *F3H*). Ascorbic acid functions as an enzymatic cofactor and antioxidant plays roles in maintenance of ROS homeostasis (Conklin and Barth, 2004). Glutathione is one of the important endogenous antioxidants in plants, which functions as a substrate in antioxidative defense mechanisms by scavenging free radicals, conjugating to toxic electrophilic compounds, and reducing peroxides (Anderson and Davis, 2004). Flavonoids are polyphenol compounds with antioxidant activities, the accumulation of flavonoids could be a key step in development of plant tolerance to different stresses (Wang P. et al., 2018).

Using *Paeonia ostii* plants under natural drought stress (it was mainly characterized by low soil water content, and the roots of plants cannot absorb enough water to compensate for the consumption of transpiration) the effect of fulvic acid was studied (Fang et al., 2020). The fulvic acid treatment increased the leaf water content and antioxidant enzyme activities and decrease proline content, ROS accumulation, and relative electrical conductivity. Also, increased the expression level of drought-tolerant genes, like *CCoAOMT*, *CAB37*, *AGT1*, *At4g26520*. *CCoAOMT* ecodes caffeoyl-coenzyme A O-methyltransferase, this enzyme is involved in monolignol synthesis that affects the efficiency of lignification and lignin composition (Rakoczy et al., 2018), and the changes of lignin composition may serve in stress resistance. *CAB37* is a important regulatory site of photosynthesis under drought stress (Li et al., 2020d). *AGT1* is related to photosynthetic processing and photorespiration (Fang et al., 2020). *At4g26520* is ABA dependent signaling pathway in drought response (Fang et al., 2020).

Examination of **Table 5** allows determining the absence of transcriptomic responses that were common among the six

studies of humic and fulvic acids. One possible explanation, in addition to the small number of studies available, is the presence of chemical or physicochemical differences between the humic substances, or the fact that not all the studies were carried out under the same type of stress.

CHITOSAN AND OTHER BIOPOLYMERS

Chitosan is a polymer obtained by the deacetylation of chitin extracted from crustaceans, fungi, and insects. Chitosan is composed of N-acetyl-D-glucosamine and D-glucosamine units that have different degrees of deacetylation (Riaz Rajoka et al., 2020). The physiological effects of chitosan in plants are related to the ability of this polycationic compound to bind to a wide range of cellular components, such as the plasma membrane, cell wall components, and DNA, in addition to binding to specific receptors involved in plant defense (Katiyar et al., 2015). Other natural and synthetic polymers can be used in agriculture, including polyacrylates, polyacrylamides, and polysaccharides (Ekebafe et al., 2011).

Genes and their function are described below on the basis of the results of five studies in which chitosan-based biostimulants and other biopolymers were applied to plants under abiotic stress (**Table 5**).

Using tomato plants, researches (Hernández-Hernández et al., 2018) studied the application of hydrogels of nCu-chitosan-PVA and chitosan-PVA as promoters of tolerance to salt stress (100 mM NaCl). The treated and stressed plants presented improved growth and increased expression of *JA* and *SOD* genes at 48 h after stress. The *JA* gene was related to the biosynthesis of jasmonic acid, which has been shown to improve tolerance to osmotic and oxidative stress in plants under salinity stress (Zhao et al., 2014). *SOD* is discussed above.

Using maize plants under salinity stress (100 mM NaCl), researchers (Turk, 2019) studied the ability of foliar applications of chitosan (0.1%) to mitigate this stress. The treated plants presented increased growth and higher expression of the *AOX1* gene on the third day after the treatment; the *AOX1* gene encodes the mitochondrial alternative oxidase enzyme. This enzyme is involved in the alternative dissipative flow of the electron transport chain, in addition to optimizing the metabolism of respiration under normal and stress conditions (Erdal and Turk, 2016). *AOX1* also plays an essential role in modulating the balance between carbon and nitrogen (Hu et al., 2017).

Using *A. thaliana* seedlings grown *in vitro* under salinity stress (100 mM NaCl), researchers (González-Pérez et al., 2018) studied the effect of a biostimulant based on xyloglucan oligosaccharides extracted from *Tamarindus indica* L. (0.1 mg L^{-1}) applied to the growth media. An increase in the expression of the *CAT3* gene was reported 24 h after stress, and the *CAT* gene is discussed above.

The effect of chitosan oligosaccharide (COS) on tea plants (*C. sinensis*) under cold stress (−4°C) was studied (Li et al., 2020e). The activity of SOD and POD, content of soluble sugar and chlorophyll in COS-treated tea plant were increased. The tea plants also were analyzed by transcriptomics with RNA-sequencing. There were identified 4503 DEGs between the control and COS under cold stress. By RT-qPCR, the GH3 gene expression was significantly higher in COS-treated plants (under stress). *GH3* (indole-3-acetic acid) is an important response gene of auxin-responsive protein, encode a class of IAA-amido synthetase related for balancing endogenous free IAA content, and play an important role in plant growth and development (Feng et al., 2015).

Using soybean under salt-stress, the effect of chitosan modified biochar (CMB) was studied (Mehmood et al., 2020). CMB treatment with salt-stress increased plant growth, root architecture characteristics, biomass yield, nutrients acquisition, chlorophyll, soluble protein and sugar contents. The gene expression levels triggered by salinity but with the application of CMB significantly increased the expression profile of *CAT*, *APX*, *POD*, *SOD*, *GmSALT3*, and *CHS* genes. *CAT*, *APX*, *POD*, and *SOD* encodes antioxidant enzyme and are discussed above. *GmSALT3* (salt tolerance-associated gene on chromosome 3) is associated with limiting the accumulation of sodium ions in shoots and a substantial enhancement in salt tolerance in soybean (Guan et al., 2014). *CHS* encode a chalcone synthase, this enzyme is the first key enzyme in the biosynthesis of flavonoids (El-Esawi et al., 2018a).

The review of **Table 5** allows determining the absence of transcriptomic responses that were common among the five studies of chitosan and other biopolymers, except for the expression of antioxidant enzymes. More studies are needed with the same chitosan and others biopolymers in different species of plants, under the same stress condition.

INORGANIC COMPOUNDS

Beneficial elements are chemical elements that are not essential for all plants and are related to growth promotion (Pilon-Smits et al., 2009). The main beneficial elements include Si, Se, Co, Al, and Na. These are present in soils and in plants as different inorganic salts and as insoluble forms (Jardin, 2015).

The following describes the genes and their function obtained from the results of six studies in which biostimulants based on inorganic compounds and mixtures with other types of biostimulants were applied to plants under abiotic stress (**Table 6**).

Using two rice cultivars under drought stress (no irrigation for 10 days), researchers (Khattab et al., 2014) studied the application of sodium selenate (0.03 mM) and potassium silicate (1.5 mM) to the seeds to promote stress tolerance. Increased expression of genes such as *DREB2A*, *NAC5*, *RDCP1*, *CMO*, and *RAB16b* was reported. *DREB2A* and *NAC5* genes encode TFs. *NAC5* TFs activate genes related to the production of osmolytes, redox homeostasis, detoxification, and formation of macromolecules (Hu et al., 2008). *RDCP1* participates in a set of physiological responses to counteract dehydration stress (Bae et al., 2011). *CMO* encodes a choline monooxygenase that participates in glycine betaine biosynthesis, resulting in increased tolerance to salinity stress (Luo et al., 2012). *RAB16B* encodes the dehydrin-like LEA protein, whose expression which is induced

TABLE 6 | Description of genes whose transcript level changed in response to the application of biostimulants based on inorganic compounds and mixtures to plants under abiotic stress.

Gene/gene expression time	Biostimulant type	Abiotic stress type	Crop/gene expression organ	References
OsDREB2A, *OsNAC5*, *OsRDCP1*, *OsCMO*, *OsRab16b*/not specified	Selenium and silicon	Drought	Rice/seedlings	Khattab et al., 2014
SlPIP1;5, *SlPIP2;6*/3 and 5 days after treatment and stress	Silicon	Drought	Tomato/roots	Shi et al., 2016
SbPIP1;3/1;4, *SbPIP1;6*, *SbPIP2;2*, *SbPIP2;6*, *SbPIP2;3*/4 and 24 h after stress	Silicon	Drought	Sorghum/roots	Liu P. et al., 2014
MT2a, *MT2b*, *PCS1*, *CSD1*, *CSD2*/3 days after treatment	Silicon	Copper toxicity	*Arabidopsis*/leaves	Khandekar and Leisner, 2011
Os08g02630 (PsbY), *Os05g48630 (PsaH)*, *Os07g37030 (PetC)*, *Os03g57120 (PetH)*, *Os09g26810*, *Os04g38410*/72 h after treatment	Silicon	Zinc toxicity	Rice/leaves	Song et al., 2014
RAB18, *P5CS1*, *ERF1A*, *MYB75*/field capacity 0.1 (10%)	Three different biostimulants (*A. nodosum* extract, amino acids, and potassium phosphite)	Drought	*Arabidopsis*/leaves	Fleming et al., 2019

in response to abiotic stress, and this protein is ABA dependent (Bies-Ethève et al., 2008).

Using tomato plants under drought stress (induced with 10% polyethylene glycol-6,000 added to the irrigation solution), researchers (Shi et al., 2016) studied the effect of potassium silicate (2.5 mM) added to the irrigation solution. The results showed that Si improved the growth, photosynthesis, and water status of plants under stress. The expression of the *SlPIP1;5* and *SlPIP2;6* genes increased at 5 and 6 days, respectively. Both genes are discussed above.

Using sorghum plants, researchers (Liu P. et al., 2014) studied the effect of applying sodium silicate (1.67 mM) to counteract water-deficit stress (10% PEG-6000). Increased expression of *SbPIP1;3/1;4*, *SbPIP1;6*, *SbPIP2;2*, *SbPIP2;6*, and *SbPIP2;3* was reported at 4 and 24 h after stress, and this group of genes encodes aquaporins and is discussed above.

Using *A. thaliana* plants under stress caused by copper toxicity (30 μM), researchers (Khandekar and Leisner, 2011) evaluated the effect of potassium metasilicate (1.5 mM). The *MT2a*, *MT2b*, *PCS1*, *CSD1*, and *CSD2* genes increased at 3 days after treatment. The *MT2a* and *MT2b* genes encode metallothioneins, which are proteins that bind to Cu to regulate its concentration in the cell (Guo et al., 2008). *PCS1* encodes a phytochelatin, which, like metallothioneins, binds to heavy metals and transports them to the vacuole (Adams et al., 2011). *CSD1* and *CSD2* encode the Cu/Zn SOD enzymes and are discussed above.

Using rice plants, researchers (Song et al., 2014) studied the effect of silicic acid (1.5 mM) added to the irrigation solution to increase stress tolerance caused by Zn toxicity (2 mM irrigation solution). The levels of the genes *Os08g02630 (PsbY)*, *Os05g48630 (PsaH)*, *Os07g37030 (PetC)*, *Os03g57120 (PetH)*, *Os09g26810*, and *Os04g38410* increased in response to the application of Si and stress at 72 h after the start of the treatments. The *Os08g02630 (PsbY)* gene encodes a photosystem II polyprotein, which, when its expression increases, subsequently increases both the activity of photosystem II and the amount of chlorophyll (Von Sydow et al., 2016). *Os05g48630 (PsaH)* encodes a photosystem I subunit (Zhang and Scheller, 2004), and *Os07g37030 (PetC)* encodes a polypeptide that binds to the Rieske FeS center of cytochrome bf and regulates the electron transport of the b_6f complex during photosynthesis (C Breyton et al., 1994). *Os03g57120 (PetH)* encodes a ferredoxin-$NADP^+$ reductase, which regulates the level of reduced glutathione in the cell (Song et al., 2014).

Using *Arabidopsis* plants under drought stress, researchers (Fleming et al., 2019) evaluated three commercial products based on *A. nodosum*, amino acids, and potassium phosphite. An increase in the expression of the *RAB18*, *P5CS1*, *ERF1A*, and *MYB75* genes was reported in response to the application of the products and at a field capacity of 0.1 (10%). The *ERF1A* gene encodes a TF of the ERF (ethylene response factor) family, with ethylene being a highly important hormone in plants (Trujillo-Moya and Gisbert, 2012). The other genes are discussed above.

Examination of **Table 6** allows determining a few transcriptomic responses that were common among the six studies. Common genes in some studies were related to TFs, antioxidants and aquaporins. More studies are needed with the same inorganic compound in different species of plants, under the same stress condition.

As a summary of the collection of all the genes differentially expressed in response to the application of biostimulants to plants under abiotic stress, **Figure 1** shows an outlook where the different types of biostimulants are presented according to the categorization described in du Jardin (2015). In addition to the 7 categories, mixtures based on 2 or more different types of biostimulants were added. Subsequently, the 7 different types of abiotic stress described in this review are shown. Finally, under each type of stress, DEGs are listed that may be part of the transcriptional modification of plants and may be involved in abiotic stress tolerance. Biostimulants based on algal and botanical extracts have been extensively studied in plants under different types of abiotic stress, such as drought, salinity, cold and heat. On the other hand, biostimulants based on protein hydrolysates and compounds with N have been

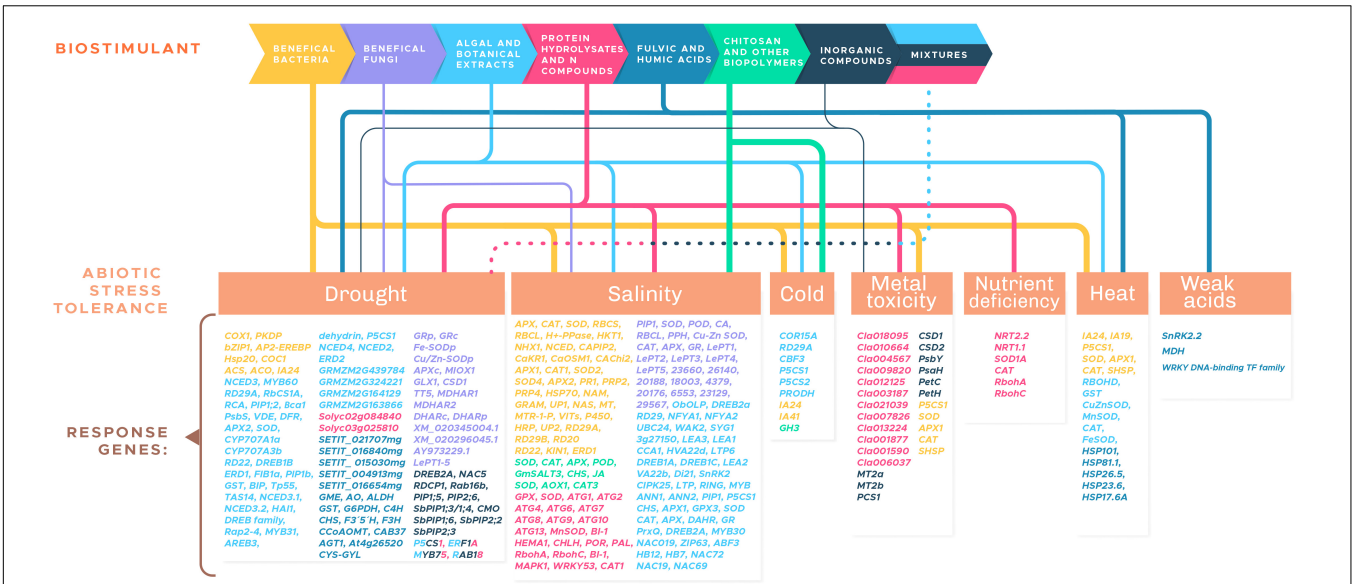


FIGURE 1 | Global view of the differentially expressed genes in plants treated with different types of biostimulants that are involved in the tolerance of various types of abiotic stress.

studied in response to only various types of abiotic stress, such as drought, salinity, metal toxicity, and nutrient deficiency. The most studied is drought stress, followed by salinity, metal toxicity, low and high temperature, weak acids, and nutrient deficiency. However, considering that high-temperature stress (in the form of heatwaves) has become increasingly common in tropical and subtropical regions as climate change progresses, it is necessary to carry out additional studies to verify the usefulness of different biostimulants under extreme temperature conditions.

In the present review, emphasis was placed on presenting DEGs in plants treated with biostimulants, as well as presenting the cellular function of these genes to try to decipher the transcriptional mechanism underlying the tolerance to abiotic stress. **Figure 2** depicts how biostimulants are perceived in cells through receptors. Such recognition triggers a series of downstream events where kinase-like proteins and TFs (regulatory proteins) drive signal transduction, after which specific response genes are ultimately activated to produce functional proteins to counteract the effects caused by abiotic stress. In the genes presented in **Figure 2**, overlap in the functional categories of functional-type proteins can be observed, in which some genes are homologous since they have the same function in different plant species. Some orthologous genes are also presented in different plant species with similar functions due to a common ancestor. However, to support this information, it is necessary to carry out a phylogenetic analysis of the sequences. Among the functional categories that present the most DEGs are proteins with a role in antioxidant metabolism, followed by functions involving photosynthesis, biosynthesis and signaling of ABA, PCD, aquaporins, osmoprotectants, LEA-like proteins, nitrogen metabolism, heat-shock proteins, biosynthesis of phytohormones, phosphorus metabolism, PR proteins, secondary metabolite biosynthesis, metal detoxification, flavonoids metabolism, photoprotection, carbohydrate metabolism, Ca^{+2} transport, ascorbate metabolism, glutathione metabolism, stomatal regulation, metabolism of phenylpropanoids, iron metabolism, ROS, cell wall, ribosomes, dehydrins, energy metabolism, senescence, and germination. It is difficult to know if the abovementioned categories correspond to the spectrum of biological responses induced by biostimulants or if the spectrum is an effect of the sampling of the works used in this study (a result of the choice of plant species, type of stress, and the genes chosen in the different works). More studies of specific metabolic pathways that represent plant status against stresses, for example, antioxidant metabolism, biosynthesis and signaling of ABA, PCD, photosynthesis, water transport, and C or N metabolism, are needed. A massive study of gene expression is also proposed through the analysis of transcriptomes or complete proteomes.

With the information analyzed in this review, in **Table 7** some putative molecular markers are proposed that can be used to test the potential of different types of biostimulants under different stress conditions. The genes selected as molecular markers are functional proteins that could be associated with tolerance to abiotic stress in plants.

CONCLUSION AND PERSPECTIVES

Given the information described above, genes that can act as molecular markers in abiotic stress tolerance can be globally visualized by applying biostimulants. The genes described encode functional proteins as well as regulatory proteins.

Most related studies have focused on the use of algal extracts, especially those of *A. nodosum*, and the effects of drought and salinity stress have been the most explored. Thus, there is an area

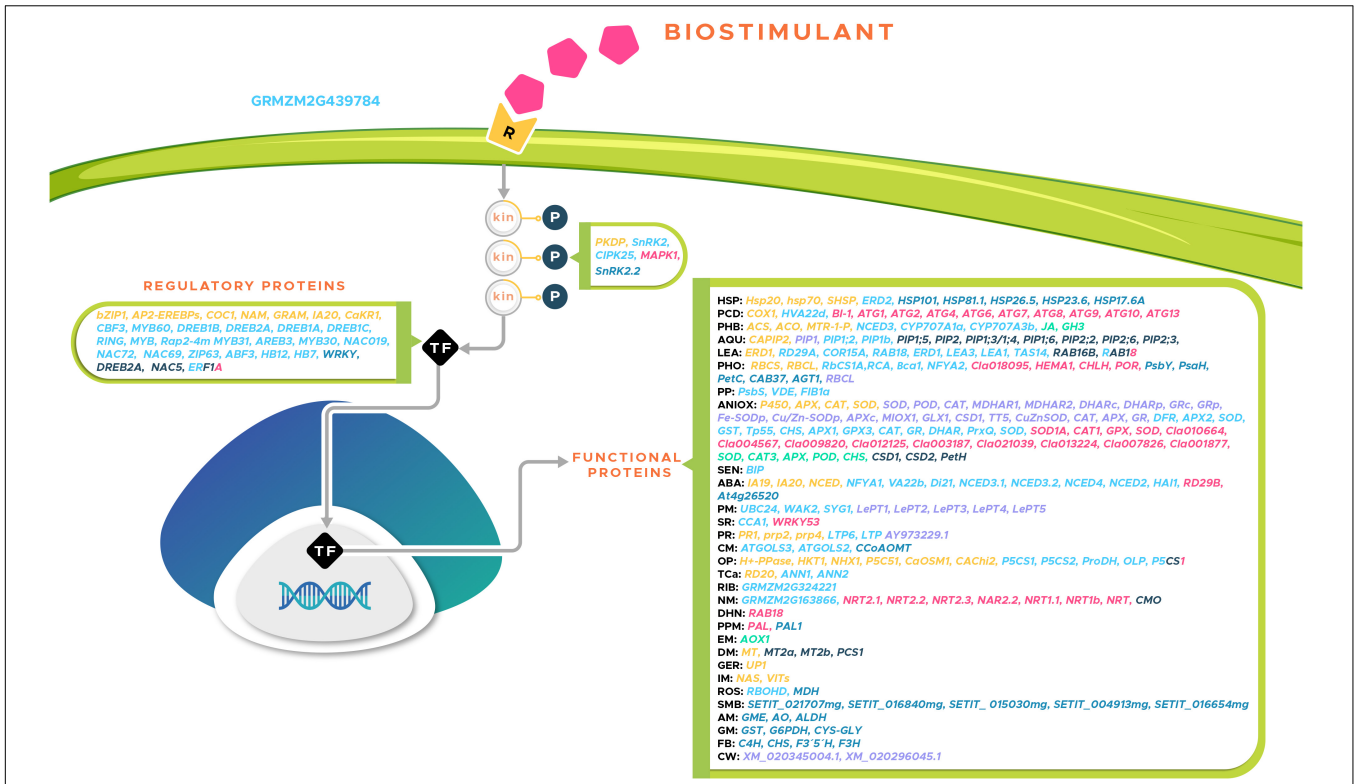


FIGURE 2 | Cellular representation of the functional categories of genes differentially expressed in plants in response to the application of different types of biostimulants. R: receptor, kin: kinases, P: phosphate groups, TF: transcription factor, HSP: heat-shock protein, PCD: programmed cell death, PHB: phytohormone biosynthesis, AQU: aquaporin, LEA: late embryogenesis abundant, PHO: photosynthesis, PP: photoprotection, ANTIOX: antioxidant metabolism, SEN: senescence, ABA: biosynthesis and signaling of ABA, FOS: phosphorous metabolism, SR: stomatal regulation, PR: PR protein, CM: carbohydrate metabolism, OP: osmoprotectant, TCa: CaC2 transport, RIB: ribosome, NM: nitrogen metabolism, DHN: dehydrin, PPM: phenylpropanoid metabolism, EM: energy metabolism, DM: metal detoxification, GER: germination, IM: iron metabolism, ROS: reactive oxygen species, SMB: secondary metabolite biosynthesis, AM: ascorbate metabolism, GM: glutathione metabolism, FB: flavonoids metabolism, CW: cell wall. Gene colors: beneficial bacteria, beneficial fungi, algal and botanical extracts, protein hydrolysates and N compounds, fulvic acids and humic acids, chitosan and other biopolymers, inorganic compounds, biostimulant mixtures.

TABLE 7 | Putative molecular markers for different potential biostimulants in varying abiotic stress conditions.

Molecular markers	Function	Biostimulant type	Abiotic stress type
CAT, SOD, POD, APX, GPX, GR, CHS, CSD1, DHAR	Antioxidant metabolism	Beneficial bacteria, beneficial fungi, algal and botanical extracts, protein hydrolysates and N compounds, chitosan and other biopolymers, inorganic compounds	Drought, salinity, metal toxicity, nutrient deficiency, heat
HSP70	Heat-shock protein (HSP)	Beneficial bacteria, algal and botanical extracts	Drought, salinity
PIP1, PIP2	Aquaporin	Beneficial bacteria, beneficial fungi, algal and botanical extracts, inorganic compounds	Drought, salinity
ERD1, RAB18	Late embryogenesis abundant (LEA)	Beneficial bacteria, algal and botanical extracts, mixtures of biostimulants	Drought, salinity
NCED	Biosynthesis and signaling of ABA	Beneficial bacteria, algal and botanical extracts	Drought, salinity
P5CS1	Osmoprotectant	Beneficial bacteria, algal and botanical extracts, mixtures of biostimulants	Drought, salinity, cold, metal toxicity, heat
PAL	Phenylalanine ammonium lyase (PAL)	Protein hydrolysates and N compounds, fulvic and humic acids	Salinity

of opportunity to study the different types of biostimulants and abiotic stresses that have been studied little.

However, to focus efforts of future research on the analysis of DEGs in plants in response to the application of biostimulants (those with greater accessibility to producers such as humic substances, vegetable extracts, hydrolysates, and elements such as silicon), it is necessary to take into account several factors, such as the type of biological model (study species with high nutritional importance, such as potato, wheat, maize, rice and tomato), the class of biostimulant, the concentration of the

biostimulant (focus on commercial recommendations), the type of application (depending on the crop species and the type of biostimulant), the study organ (include leaves and, if possible, the organ of commercial interest, such as the roots, tubers and fruits), and the time of analysis of gene expression (establish comparable standards of sampling times at 12, 24, and 36 h after the application of the biostimulant).

AUTHOR CONTRIBUTIONS

AB-M: conceptualization. SG-M: writing of the original draft and editing. SG-M, SS-G, MV-C, AJ-M, AL-T, and AB-M: responsible for processing and organizing information. All authors have read and approved the final manuscript.

FUNDING

This research received external funding from the UPL Limited.

ACKNOWLEDGMENTS

We acknowledge the UPL for funding to publish this review.

REFERENCES

Abd El-Daim, I. A., Bejai, S., Fridborg, I., and Meijer, J. (2018). Identifying potential molecular factors involved in *Bacillus amyloliquefaciens* 5113 mediated abiotic stress tolerance in wheat. *Plant Biol.* 20, 271–279. doi: 10.1111/plb.12680

Adams, J. P., Adeli, A., Hsu, C.-Y., Harkess, R. L., Page, G. P., Depamphilis, C. W., et al. (2011). Poplar maintains zinc homeostasis with heavy metal genes HMA4 and PCS1. *J. Exp. Bot.* 62, 3737–3752. doi: 10.1093/jxb/err025

Adesemoye, A., Torbert, H., and Ecology, J. K.-A. S. (2010). Increased plant uptake of nitrogen from 15N-depleted fertilizer using plant growth-promoting rhizobacteria. *Appl. Soil Ecol.* 46, 54–58.

Ahmad, P., Jaleel, C. A., Salem, M. A., Nabi, G., and Sharma, S. (2010). Roles of enzymatic and nonenzymatic antioxidants in plants during abiotic stress. *Crit. Rev. Biotechnol.* 30, 161–175. doi: 10.3109/07388550903524243

Ahn, Y. O., Kim, S. H., Lee, J., Ran Kim, H., Lee, H. S., and Kwak, S. S. (2012). Three *Brassica rapa* metallothionein genes are differentially regulated under various stress conditions. *Mol. Biol. Rep.* 39, 2059–2067. doi: 10.1007/s11033-011-0953-5

Akbar, S., Wei, Y., Yuan, Y., Khan, M. T., Qin, L., Powell, C. A., et al. (2020). Gene expression profiling of reactive oxygen species (ROS) and antioxidant defense system following Sugarcane mosaic virus (SCMV) infection. *BMC Plant Biol.* 20:532. doi: 10.1186/s12870-020-02737-1

Akpinar, B. A., Avsar, B., Lucas, S. J., and Budak, H. (2012). Plant abiotic stress signaling. *Plant Signal. Behav.* 7, 1450–1455. doi: 10.4161/psb.21894

Al-Ghamdi, A. A., and Elansary, H. O. (2018). Synergetic effects of 5-aminolevulinic acid and *Ascophyllum nodosum* seaweed extracts on *Asparagus phenolics* and stress related genes under saline irrigation. *Plant Physiol. Biochem.* 129, 273–284. doi: 10.1016/j.plaphy.2018.06.008

Alkuwayti, M. A., El-Sherif, F., Yap, Y. K., and Khattab, S. (2020). Foliar application of Moringa oleifera leaves extract altered stress-responsive gene expression and enhanced bioactive compounds composition in *Ocimum basilicum*. *S. Afr. J. Bot.* 129, 291–298. doi: 10.1016/j.sajb.2019.08.001

Anderson, J. V., and Davis, D. G. (2004). Abiotic stress alters transcript profiles and activity of glutathione S-transferase, glutathione peroxidase, and glutathione reductase in *Euphorbia esula*. *Physiol. Plant.* 120, 421–433. doi: 10.1111/j.0031-9317.2004.00249.x

Antolín-Llovera, M., Petutsching, E. K., Ried, M. K., Lipka, V., Nürnberger, T., Robatzek, S., et al. (2014). Knowing your friends and foes – plant receptor-like kinases as initiators of symbiosis or defence. *New Phytol.* 204, 791–802.

Arraes, F. B. M., Beneventi, M. A., Lisei de Sa, M. E., Paixao, J. F. R., Albuquerque, E. V. S., Marin, S. R. R., et al. (2015). Implications of ethylene biosynthesis and signaling in soybean drought stress tolerance. *BMC Plant Biol.* 15:213. doi: 10.1186/s12870-015-0597-z

Atkinson, N., Leitão, N., Orr, D. J., Meyer, M. T., Carmo-Silva, E., Griffiths, H., et al. (2017). Rubisco small subunits from the unicellular green alga Chlamydomonas complement Rubisco-deficient mutants of *Arabidopsis*. *New Phytol.* 214, 655–667. doi: 10.1111/nph.14414

Aubert, Y., Vile, D., Pervent, M., Aldon, D., Ranty, B., Simonneau, T., et al. (2010). RD20, a stress-inducible caleosin, participates in stomatal control, transpiration and drought tolerance in *Arabidopsis thaliana*. *Plant Cell Physiol.* 51, 1975–1987. doi: 10.1093/PCP/PCQ155

Bae, H., Kim, S. K., Cho, S. K., Kang, B. G., and Kim, W. T. (2011). Overexpression of OsRDCP1, a rice RING domain-containing E3 ubiquitin ligase, increased tolerance to drought stress in rice (*Oryza sativa* L.). *Plant Sci.* 180, 775–782. doi: 10.1016/j.plantsci.2011.02.008

Baek, D., Rokibuzzaman, M., Khan, A., Kim, M. C., Park, H. J., Yun, D., et al. (2020). Plant-growth promoting *Bacillus oryzicola* YC7007 modulates stress-response gene expression and provides protection from salt stress. *Front. Plant Sci.* 10:1646. doi: 10.3389/fpls.2019.01646

Baía, D. C., Olivares, F. L., Zandonadi, D. B., de Paula Soares, C., Spaccini, R., and Canellas, L. P. (2020). Humic acids trigger the weak acids stress response in maize seedlings. *Chem. Biol. Technol. Agric.* 7:31. doi: 10.1186/s40538-020-00193-5

Bak, S., Beisson, F., Bishop, G., Hamberger, B., Höfer, R., Paquette, S., et al. (2011). Cytochromes P450. *Arab. B* 9:e0144. doi: 10.1199/tab.0144

Behnam, B., Iuchi, S., Fujita, M., Fujita, Y., Takasaki, H., Osakabe, Y., et al. (2013). Characterization of the promoter region of an *Arabidopsis* gene for 9-cis-epoxycarotenoid dioxygenase involved in dehydration-inducible transcription. *DNA Res.* 20, 315–324. doi: 10.1093/dnares/dst012

Berbara, R. L. L., and García, A. C. (2014). "Humic substances and plant defense metabolism," in *Physiological Mechanisms and Adaptation Strategies in Plants Under Changing Environment*, eds P. Ahmad, M. R. Wani (New York, NY: Springer), 297–319.

Bies-Ethève, N., Gaubier-Comella, P., Debures, A., Lasserre, E., Jobet, E., Raynal, M., et al. (2008). Inventory, evolution and expression profiling diversity of the LEA (late embryogenesis abundant) protein gene family in *Arabidopsis thaliana*. *Plant Mol. Biol.* 67, 107–124. doi: 10.1007/s11103-008-9304-x

Bihmidine, S., Lin, J., Stone, J. M., Awada, T., Specht, J. E., and Clemente, T. E. (2013). Activity of the *Arabidopsis* RD29A and RD29B promoter elements in soybean under water stress. *Planta* 237, 55–64. doi: 10.1007/s00425-012-1740-9

Blunden, G., and Gordon, S. M. (1986). Betaines and their sulphonio analogues in marine algae. *Prog. Phycol. Res.* 4, 39–80.

Bogamuwa, S. P., and Jang, J.-C. (2014). Tandem CCCH zinc finger proteins in plant growth, development and stress response. *Plant Cell Physiol.* 55, 1367–1375. doi: 10.1093/pcp/pcu074

Breyton, C., de Vitry, C., and Popot, J. L. (1994). Membrane association of cytochrome b6f subunits. The Rieske iron-sulfur protein from *Chlamydomonas reinhardtii* is an extrinsic protein. *J. Biol. Chem.* 269, 7597–7602.

Brown, D. E., Rashotte, A. M., Murphy, A. S., Normanly, J., Tague, B. W., Peer, W. A., et al. (2001). Flavonoids act as negative regulators of auxin transport in vivo in *Arabidopsis*. *Plant Physiol.* 126, 524–535. doi: 10.1104/pp.126.2.524

Bruno, L. B., Karthik, C., Ma, Y., Kadirvelu, K., Freitas, H., and Rajkumar, M. (2020). Amelioration of chromium and heat stresses in Sorghum bicolor by Cr6+ reducing-thermotolerant plant growth promoting bacteria. *Chemosphere* 244:125521. doi: 10.1016/j.chemosphere.2019.125521

Bulgakov, V., and Wu, H.-C. (2019). Coordination of ABA and chaperone signaling in plant stress responses. *Trends Plant Sci.* 24, 636–651. doi: 10.1016/j.tplants.2019.04.004

Calvo, P., Nelson, L., and Kloepper, J. W. (2014). Agricultural uses of plant biostimulants. *Plant Soil* 383, 3–41. doi: 10.1007/s11104-014-2131-8

Campobenedetto, C., Grange, E., Mannino, G., van Arkel, J., Beekwilder, J., Karlova, R., et al. (2020). A biostimulant seed treatment improved heat stress tolerance during cucumber seed germination by acting on the antioxidant system and glyoxylate cycle. *Front. Plant Sci.* 11:836. doi: 10.3389/fpls.2020.00836

Cao, J. (2019). Molecular evolution of the vacuolar iron transporter (VIT) family genes in 14 plant species. *Genes* 10:144. doi: 10.3390/genes10020144

Carvalho, H. H., Brustolini, O. J. B., Pimenta, M. R., Mendes, G. C., Gouveia, B. C., Silva, P. A., et al. (2014). The molecular chaperone binding protein BiP prevents leaf dehydration-induced cellular homeostasis disruption. *PLoS One* 9:e86661. doi: 10.1371/journal.pone.0086661

Cecchini, N. M., Monteoliva, M. I., and Alvarez, M. E. (2011). Proline dehydrogenase is a positive regulator of cell death in different kingdoms. *Plant Signal. Behav.* 6, 1195–1197. doi: 10.4161/psb.6.8.15791

Cha, J., Kang, S., Ali, I., Lee, S., Ji, M., Reports, S. J.-S., et al. (2020). Humic acid enhances heat stress tolerance via transcriptional activation of heat-shock proteins in *Arabidopsis*. *Sci. Rep.* 10, 1–12.

Chakrabortee, S., Boschetti, C., Walton, L. J., Sarkar, S., Rubinsztein, D. C., and Tunnacliffe, A. (2007). Hydrophilic protein associated with desiccation tolerance exhibits broad protein stabilization function. *Proc. Natl. Acad. Sci. U.S.A.* 104, 18073–18078. doi: 10.1073/pnas.0706964104

Chauhan, P. S., Lata, C., Tiwari, S., Chauhan, A. S., Mishra, S. K., Agrawal, L., et al. (2019). Transcriptional alterations reveal *Bacillus amyloliquefaciens*-rice cooperation under salt stress. *Sci. Rep.* 9, 1–13. doi: 10.1038/s41598-019-48309-8

Chen, C. N., Chu, C. C., Zentella, R., Pan, S. M., and Ho, T. H. D. (2002). AtHVA22 gene family in *Arabidopsis*: phylogenetic relationship, ABA and stress regulation, and tissue-specific expression. *Plant Mol. Biol.* 49, 633–644. doi: 10.1023/A:1015593715144

Chen, C. N. N., Chen, H. R., Yeh, S. Y., Vittore, G., and Ho, T. H. D. (2009). Autophagy is enhanced and floral development is impaired in AtHVA22d RNA interference *Arabidopsis*. *Plant Physiol.* 149, 1679–1689. doi: 10.1104/pp.108.131490

Chen, L., Liu, Y., Wu, G., Veronican Njeri, K., Shen, Q., Zhang, N., et al. (2016). Induced maize salt tolerance by rhizosphere inoculation of *Bacillus amyloliquefaciens* SQR9. *Physiol. Plant.* 158, 34–44. doi: 10.1111/ppl.12441

Choi, D. S., Hong, J. K., and Hwang, B. K. (2013). Pepper osmotin-like protein 1 (CaOSM1) is an essential component for defense response, cell death, and oxidative burst in plants. *Planta* 238, 1113–1124. doi: 10.1007/s00425-013-1956-3

Chowdhury, S., Basu, A., and Kundu, S. (2017). Overexpression of a new osmotin-like protein gene (SindOLP) confers tolerance against biotic and abiotic stresses in sesame. *Front. Plant Sci.* 8:410. doi: 10.3389/fpls.2017.00410

Ciszak, K., Kulasek, M., Barczak, A., Grzelak, J., Maækowski, S., and Karpiński, S. (2015). PsbS is required for systemic acquired acclimation and post-excess-light-stress optimization of chlorophyll fluorescence decay times in *Arabidopsis*. *Plant Signal. Behav.* 10:e982018. doi: 10.4161/15592324.2014.982018

Colla, G., Rouphael, Y., Di Mattia, E., El-Nakhel, C., and Cardarelli, M. (2015). Co-inoculation of Glomus intraradices and Trichoderma atroviride acts as a biostimulant to promote growth, yield and nutrient uptake of vegetable crops. *J. Sci. Food Agric.* 95, 1706–1715. doi: 10.1002/jsfa.6875

Conklin, P. L., and Barth, C. (2004). Ascorbic acid, a familiar small molecule intertwined in the response of plants to ozone, pathogens, and the onset of senescence. *Plant Cell Environ.* 27, 959–970. doi: 10.1111/j.1365-3040.2004.01203.x

Craigie, J. S. (2011). Seaweed extract stimuli in plant science and agriculture. *J. Appl. Phycol.* 23, 371–393. doi: 10.1007/s10811-010-9560-4

Cui, L. G., Shan, J. X., Shi, M., Gao, J. P., and Lin, H. X. (2014). The miR156-SPL9-DFR pathway coordinates the relationship between development and abiotic stress tolerance in plants. *Plant J.* 80, 1108–1117. doi: 10.1111/tpj.12712

Czarny, J. C., Grichko, V. P., and Glick, B. R. (2006). Genetic modulation of ethylene biosynthesis and signaling in plants. *Biotechnol. Adv.* 24, 410–419. doi: 10.1016/j.biotechadv.2006.01.003

Dana, C. D., Bevan, D. R., and Winkel, B. S. J. (2006). Molecular modeling of the effects of mutant alleles on chalcone synthase protein structure. *J. Mol. Model.* 12, 905–914. doi: 10.1007/s00894-005-0071-1

De Palma, M., Salzano, M., Villano, C., Aversano, R., Lorito, M., Ruocco, M., et al. (2019). Transcriptome reprogramming, epigenetic modifications and alternative splicing orchestrate the tomato root response to the beneficial fungus *Trichoderma harzianum*. *Hortic. Res.* 6, 1–15. doi: 10.1038/s41438-018-0079-1

de Zelicourt, A., Colcombet, J., and Hirt, H. (2016). The role of MAPK modules and ABA during abiotic stress signaling. *Trends Plant Sci.* 21, 677–685. doi: 10.1016/j.tplants.2016.04.004

Desoky, E. S. M., ElSayed, A. I., Merwad, A. R. M. A., and Rady, M. M. (2019). Stimulating antioxidant defenses, antioxidant gene expression, and salt tolerance in *Pisum sativum* seedling by pretreatment using licorice root extract (LRE) as an organic biostimulant. *Plant Physiol. Biochem.* 142, 292–302. doi: 10.1016/j.plaphy.2019.07.020

Dievart, A., Perin, C., Hirsch, J., Bettembourg, M., Lanau, N., Artus, F., et al. (2015). The phenome analysis of mutant alleles in Leucine-Rich Repeat Receptor-Like Kinase genes in rice reveals new potential targets for stress tolerant cereals. *Plant Sci.* 242, 240–249. doi: 10.1016/j.plantsci.2015.06.019

Dodd, I. C., and Ruiz-Lozano, J. M. (2012). Microbial enhancement of crop resource use efficiency. *Curr. Opin. Biotechnol.* 23, 236–242. doi: 10.1016/j.copbio.2011.09.005

Du, J., Wang, L., Zhang, X., Xiao, X., Wang, F., Lin, P., et al. (2016). Heterologous expression of two Physcomitrella patens group 3 late embryogenesis abundant protein (LEA3) genes confers salinity tolerance in *Arabidopsis*. *J. Plant Biol.* 59, 182–193. doi: 10.1007/s12374-016-0565-7

du Jardin, P. (2015). Plant biostimulants: definition, concept, main categories and regulation. *Sci. Hortic.* 196, 3–14. doi: 10.1016/j.scienta.2015.09.021

Duan, M., Wang, J., Zhang, X., Yang, H., Wang, H., Qiu, Y., et al. (2017). Identification of optimal reference genes for expression analysis in radish (*Raphanus sativus* L.) and its relatives based on expression stability. *Front. Plant Sci.* 8:1605. doi: 10.3389/fpls.2017.01605

Dunaeva, M., and Adamska, I. (2001). Identification of genes expressed in response to light stress in leaves of *Arabidopsis thaliana* using RNA differential display. *Eur. J. Biochem.* 268, 5521–5529. doi: 10.1046/j.1432-1033.2001.02471.x

Ekebafe, L. O., Ogbeifun, D. E., and Okieimen, F. (2011). Polymer applications in agriculture. *Biokemistri* 23, 81–89.

El-Esawi, M. A., Alaraidh, I. A., Alsahli, A. A., Alamri, S. A., Ali, H. M., and Alayafi, A. A. (2018a). *Bacillus firmus* (SW5) augments salt tolerance in soybean (*Glycine max* L.) by modulating root system architecture, antioxidant defense systems and stress-responsive genes expression. *Plant Physiol. Biochem.* 132, 375–384. doi: 10.1016/j.plaphy.2018.09.026

El-Esawi, M. A., Alaraidh, I. A., Alsahli, A. A., Alzahrani, S. M., Ali, H. M., Alayafi, A. A., et al. (2018b). *Serratia liquefaciens* KM4 improves salt stress tolerance in maize by regulating redox potential, ion homeostasis, leaf gas exchange and stress-related gene expression. *Int. J. Mol. Sci.* 19:3310. doi: 10.3390/ijms19113310

Elizabete Carmo-Silva, A., and Salvucci, M. E. (2013). The regulatory properties of rubisco activase differ among species and affect photosynthetic induction during light transitions. *Plant Physiol.* 161, 1645–1655. doi: 10.1104/pp.112.213348

Eltelib, H. A., Fujikawa, Y., and Esaka, M. (2012). Overexpression of the acerola (Malpighia glabra) monodehydroascorbate reductase gene in transgenic tobacco plants results in increased ascorbate levels and enhanced tolerance to salt stress. *S. Afr. J. Bot.* 78, 295–301. doi: 10.1016/j.sajb.2011.08.005

Erdal, S., and Turk, H. (2016). Cysteine-induced upregulation of nitrogen metabolism-related genes and enzyme activities enhance tolerance of maize seedlings to cadmium stress. *Environ. Exp. Bot.* 132, 92–99. doi: 10.1016/j.envexpbot.2016.08.014

Eriksson, S. K., and Harryson, P. (2011). *Dehydrins: Molecular Biology, Structure and Function*. Berlin: Springer, 289–305.

Ertani, A., and Schiavon, M. (2013). Alfalfa plant-derived biostimulant stimulate short-term growth of salt stressed *Zea mays* L. *Plants* 364, 145–158. doi: 10.1007/s11104-012-1335-z

Ertani, A., Schiavon, M., Muscolo, A., and Nardi, S. (2013). Alfalfa plant-derived biostimulant stimulate short-term growth of salt stressed *Zea mays* L. plants. *Plant Soil* 364, 145–158.

Fan, Q. J., and Liu, J. H. (2011). Colonization with arbuscular mycorrhizal fungus affects growth, drought tolerance and expression of stress-responsive genes in *Poncirus trifoliata*. *Acta Physiol. Plant.* 33, 1533–1542. doi: 10.1007/s11738-011-0789-6

Fang, Z., Wang, X., Zhang, X., Zhao, D., and Tao, J. (2020). Effects of fulvic acid on the photosynthetic and physiological characteristics of *Paeonia ostii* under drought stress. *Plant Signal. Behav.* 15:1774714. doi: 10.1080/15592324.2020.1774714

Feng, S., Yue, R., Tao, S., Yang, Y., Zhang, L., Xu, M., et al. (2015). Genome-wide identification, expression analysis of auxin-responsive GH3 family genes

in maize (*Zea mays* L.) under abiotic stresses. *J. Integr. Plant Biol.* 57, 783–795. doi: 10.1111/jipb.12327

Fleming, T. R., Fleming, C. C., Levy, C. C. B., Repiso, C., Hennequart, F., Nolasco, J. B., et al. (2019). Biostimulants enhance growth and drought tolerance in *Arabidopsis thaliana* and exhibit chemical priming action. *Ann. Appl. Biol.* 174, 153–165. doi: 10.1111/aab.12482

Fowler, S. G., Cook, D., and Thomashow, M. F. (2005). Low temperature induction of *Arabidopsis* CBF1, 2, and 3 is gated by the circadian clock. *Plant Physiol.* 137, 961–968. doi: 10.1104/pp.104.058354

Foyer, C. H., and Noctor, G. (2005). Redox homeostasis and antioxidant signaling: a metabolic interface between stress perception and physiological responses. *Plant Cell* 17, 1866–1875. doi: 10.1105/tpc.105.033589

Frank, A., Matiolli, C. C., Viana, A. J. C., Hearn, T. J., Kusakina, J., Belbin, F. E., et al. (2018). Circadian entrainment in *Arabidopsis* by the sugar-responsive transcription factor bZIP63. *Curr. Biol.* 28, 2597.e–2606.e. doi: 10.1016/j.cub.2018.05.092

Franklin, K. A., and Whitelam, G. C. (2007). Light-quality regulation of freezing tolerance in *Arabidopsis thaliana*. *Nat. Genet.* 39, 1410–1413. doi: 10.1038/ng.2007.3

Franzoni, G., Cocetta, G., Trivellini, A., and Ferrante, A. (2019). Transcriptional regulation in rocket leaves as affected by salinity. *Plants* 9:20. doi: 10.3390/plants9010020

Fufezan, C., Simionato, D., and Morosinotto, T. (2012). Identification of key residues for pH dependent activation of violaxanthin de-epoxidase from *Arabidopsis thaliana*. *PLoS One* 7:e35669. doi: 10.1371/journal.pone.0035669

Fukami, J., de la Osa, C., Ollero, F. J., Megías, M., and Hungria, M. (2018). Co-inoculation of maize with *Azospirillum brasilense* and *Rhizobium tropici* as a strategy to mitigate salinity stress. *Funct. Plant Biol.* 45:328. doi: 10.1071/FP17167

Fukami, J., Ollero, F. J., Megías, M., and Hungria, M. (2017). Phytohormones and induction of plant-stress tolerance and defense genes by seed and foliar inoculation with Azospirillum brasilense cells and metabolites promote maize growth. *AMB Express* 7, 1–13. doi: 10.1186/s13568-017-0453-7

Gao, C., Wang, Y., Liu, G., Wang, C., Jiang, J., and Yang, C. (2010). Cloning of ten peroxidase (POD) genes from Tamarix Hispida and characterization of their responses to abiotic stress. *Plant Mol. Biol. Report.* 28, 77–89. doi: 10.1007/s11105-009-0129-9

Gill, S. S., and Tuteja, N. (2010). Reactive oxygen species and antioxidant machinery in abiotic stress tolerance in crop plants. *Plant Physiol. Biochem.* 48, 909–930. doi: 10.1016/j.plaphy.2010.08.016

Glick, B. R. (2012). Plant growth-promoting bacteria: mechanisms and applications. *Scientifica* 2012, 1–15. doi: 10.6064/2012/963401

Godoy, J. A., Lunar, R., Torres-Schumann, S., Moreno, J., Rodrigo, R. M., and Pintor-Toro, J. A. (1994). Expression, tissue distribution and subcellular localization of dehydrin TAS14 in salt-stressed tomato plants. *Plant Mol. Biol.* 26, 1921–1934. doi: 10.1007/BF00019503

Goñi, O., Quille, P., and O'Connell, S. (2018). Ascophyllum nodosum extract biostimulants and their role in enhancing tolerance to drought stress in tomato plants. *Plant Physiol. Biochem.* 126, 63–73. doi: 10.1016/j.plaphy.2018.02.024

González-Pérez, L., Páez-Watson, T., Álvarez-Suarez, J. M., Obando-Rojas, M. C., Bonifaz-Arcos, E., Viteri, G., et al. (2018). Application of exogenous xyloglucan oligosaccharides affects molecular responses to salt stress in *Arabidopsis thaliana* seedlings. *J. Soil Sci. Plant Nutr.* 18, 1187–1205. doi: 10.4067/S0718-95162018005003301

Govindasamy, V., Senthilkumar, M., Magheshwaran, V., Kumar, U., Bose, P., Sharma, V., et al. (2010). *Bacillus and Paenibacillus spp.: Potential PGPR for Sustainable Agriculture*. Berlin: Springer, 333–364. doi: 10.1007/978-3-642-13612-2_15

Graether, S. P., and Boddington, K. F. (2014). Disorder and function: a review of the dehydrin protein family. *Front. Plant Sci.* 5:576. doi: 10.3389/fpls.2014.00576

Guan, R., Qu, Y., Guo, Y., Yu, L., Liu, Y., Jiang, J., et al. (2014). Salinity tolerance in soybean is modulated by natural variation in GmSALT3. *Plant J.* 80, 937–950. doi: 10.1111/tpj.12695

Guo, W. J., Meetam, M., and Goldsbrough, P. B. (2008). Examining the specific contributions of individual *Arabidopsis metallothioneins* to copper distribution and metal tolerance. *Plant Physiol.* 146, 1697–1706. doi: 10.1104/pp.108.115782

Harper, A. L., McKinney, L. V., Nielsen, L. R., Havlickova, L., Li, Y., Trick, M., et al. (2016). Molecular markers for tolerance of European ash (*Fraxinus excelsior*) to dieback disease identified using Associative Transcriptomics. *Sci. Rep.* 6, 19335. doi: 10.1038/srep19335

Hasse, D., Larsson, A. M., and Andersson, I. (2015). Structure of *Arabidopsis thaliana* Rubisco activase. *Acta Crystallogr. Sect. D Biol. Crystallogr.* 71, 800–808. doi: 10.1107/S1399004715001182

Hassidim, M., Dakhiya, Y., Turjeman, A., Hussien, D., Shor, E., Anidjar, A., et al. (2017). CIRCADIAN CLOCK ASSOCIATED1 (CCA1) and the circadian control of stomatal aperture. *Plant Physiol.* 175, 1864–1877. doi: 10.1104/pp.17.01214

Hernández-Hernández, H., Juárez-Maldonado, A., Benavides-Mendoza, A., Ortega-Ortiz, H., Cadenas-Pliego, G., Sánchez-Aspeytia, D., et al. (2018). Chitosan-PVA and copper nanoparticles improve growth and overexpress the SOD and JA genes in tomato plants under salt stress. *Agronomy* 8:175. doi: 10.3390/agronomy8090175

Hernández-Sánchez, I. E., Maruri-López, I., Molphe-Balch, E. P., Becerra-Flora, A., Jaimes-Miranda, F., and Jiménez-Bremont, J. F. (2019). Evidence for in vivo interactions between dehydrins and the aquaporin AtPIP2B. *Biochem. Biophys. Res. Commun.* 510, 545–550. doi: 10.1016/j.bbrc.2019.01.095

Hong, D. D., Hien, H. M., and Son, P. N. (2007). Seaweeds from Vietnam used for functional food, medicine and biofertilizer. *J. Appl. Phycol.* 19, 817–826. doi: 10.1007/s10811-007-9228-x

Hong, J., and Hwang, B. (2006). Promoter activation of pepper class II basic chitinase gene, CAChi2, and enhanced bacterial disease resistance and osmotic stress tolerance in the CAChi2. *Planta* 223:433.

Hoque, T. S., Uraji, M., Tuya, A., Nakamura, Y., and Murata, Y. (2012). Methylglyoxal inhibits seed germination and root elongation and up-regulates transcription of stress-responsive genes in ABA-dependent pathway in *Arabidopsis*. *Plant Biol.* 14, 854–858. doi: 10.1111/j.1438-8677.2012.00607.x

Hu, H., Rappel, W. J., Occhipinti, R., Ries, A., Böhmer, M., You, L., et al. (2015). Distinct cellular locations of carbonic anhydrases mediate carbon dioxide control of stomatal movements. *Plant Physiol.* 169, 1168–1178. doi: 10.1104/pp.15.00646

Hu, H., You, J., Fang, Y., Zhu, X., Qi, Z., and Xiong, L. (2008). Characterization of transcription factor gene SNAC2 conferring cold and salt tolerance in rice. *Plant Mol. Biol.* 67, 169–181. doi: 10.1007/s11103-008-9309-5

Hu, W. H., Yan, X. H., and Yu, J. Q. (2017). Importance of the mitochondrial alternative oxidase (AOX) pathway in alleviating photoinhibition in cucumber leaves under chilling injury and subsequent recovery when leaves are subjected to high light intensity. *J. Hortic. Sci. Biotechnol.* 92, 31–38. doi: 10.1080/14620316.2016.1219239

Huang, J., Gu, M., Lai, Z., Fan, B., Shi, K., Zhou, Y. H., et al. (2010). Functional analysis of the *Arabidopsis* PAL gene family in plant growth, development, and response to environmental stress. *Plant Physiol.* 153, 1526–1538. doi: 10.1104/pp.110.157370

Huang, Y. W., Zhou, Z. Q., Yang, H. X., Wei, C. X., Wan, Y. Y., Wang, X. J., et al. (2014). Glucose application protects chloroplast ultrastructure in heat-stressed cucumber leaves through modifying antioxidant enzyme activity. *Biol. Plant.* 59, 131–138. doi: 10.1007/s10535-014-0470-1

Hussain, R. M. F., Sheikh, A. H., Haider, I., Quareshy, M., and Linthorst, H. J. M. (2018). *Arabidopsis* WRKY50 and TGA transcription factors synergistically activate expression of PR1. *Front. Plant Sci.* 9:930. doi: 10.3389/fpls.2018.00930

Igamberdiev, A. U. (2015). Control of Rubisco function via homeostatic equilibration of CO2 supply. *Front. Plant Sci.* 6:106. doi: 10.3389/fpls.2015.00106

Ishikawa, T., Aki, T., Yanagisawa, S., Uchimiya, H., and Kawai-Yamada, M. (2015). Overexpression of BAX INHIBITOR-1 links plasma membrane microdomain proteins to stress. *Plant Physiol.* 169, 1333–1343. doi: 10.1104/pp.15.00445

Jahromi, F., Aroca, R., Porcel, R., and Ruiz-Lozano, J. M. (2008). Influence of salinity on the in vitro development of Glomus intraradices and on the in vivo physiological and molecular responses of mycorrhizal lettuce plants. *Microb. Ecol.* 55, 45–53. doi: 10.1007/s00248-007-9249-7

Jardin, P. (2015). Scientia horticulturae plant biostimulants?: definition, concept, main categories and regulation. *Sci. Hortic.* 196, 3–14.

Jithesh, M. N., Shukla, P. S., Kant, P., Joshi, J., Critchley, A. T., and Prithiviraj, B. (2018). Physiological and transcriptomics analyses reveal that ascophyllum nodosum extracts induce salinity tolerance in *Arabidopsis* by regulating the

expression of stress responsive genes. *J. Plant Growth Regul.* 38, 463–478. doi: 10.1007/s00344-018-9861-4

Johnson, A. A. T., Kyriacou, B., Callahan, D. L., Carruthers, L., Stangoulis, J., Lombi, E., et al. (2011). Constitutive overexpression of the OsNAS gene family reveals single-gene strategies for effective iron- and zinc-biofortification of rice endosperm. *PLoS One* 6:e0024476. doi: 10.1371/journal.pone.0024476

Juárez-Maldonado, A., Ortega-Ortíz, H., Morales-Díaz, A. B., González-Morales, S., Morelos-Moreno, Á, Cabrera-De la Fuente, M., et al. (2019). Nanoparticles and nanomaterials as plant biostimulants. *Int. J. Mol. Sci.* 20:162. doi: 10.3390/ijms20010162

Kanwar, P., Sanyal, S. K., Tokas, I., Yadav, A. K., Pandey, A., Kapoor, S., et al. (2014). Comprehensive structural, interaction and expression analysis of CBL and CIPK complement during abiotic stresses and development in rice. *Cell Calcium* 56, 81–95. doi: 10.1016/j.ceca.2014.05.003

Katiyar, D., Hemantaranjan, A., and Singh, B. (2015). Chitosan as a promising natural compound to enhance potential physiological responses in plant: a review. *Indian J. Plant Physiol.* 20, 1–9. doi: 10.1007/s40502-015-0139-6

Kaur, N., and Gupta, A. K. (2005). Signal transduction pathways under abiotic stresses in plants. *Curr. Sci.* 88, 1771–1780. doi: 10.2307/24110354

Kesari, R., Lasky, J. R., Villamor, J. G., Des Marais, D. L., Chen, Y. J. C., Liu, T. W., et al. (2012). Intron-mediated alternative splicing of *Arabidopsis* P5CS1 and its association with natural variation in proline and climate adaptation. *Proc. Natl. Acad. Sci. U.S.A.* 109, 9197–9202. doi: 10.1073/pnas.1203433109

Khan, M. S., Akther, T., and Hemalatha, S. (2017). Impact of panchagavya on *Oryza sativa* L. grown under saline stress. *J. Plant Growth Regul.* 36, 702–713. doi: 10.1007/s00344-017-9674-x

Khan, M. S., Pandey, M. K., and Hemalatha, S. (2018). Comparative studies on the role of organic biostimulant in resistant and susceptible cultivars of rice grown under saline stress - organic biostimulant alleviate saline stress in tolerant and susceptible cultivars of rice. *J. Crop Sci. Biotechnol.* 21, 459–467. doi: 10.1007/s12892-018-0089-0

Khan, W., Rayirath, U. P., Subramanian, S., Jithesh, M. N., Rayorath, P., Hodges, D. M., et al. (2009). Seaweed extracts as biostimulants of plant growth and development. *J. Plant Growth Regul.* 28, 386–399. doi: 10.1007/s00344-009-9103-x

Khan, Z., and Shahwar, D. (2020). "Role of heat shock proteins (HSPs) and heat stress tolerance in crop plants," in *Sustainable Agriculture in the Era of Climate Change*, eds M. Hasanuzzaman, R. Roychowdhury, S. Choudhury, and S. Srivastava (Berlin: Springer International Publishing), 211–234.

Khandekar, S., and Leisner, S. (2011). Soluble silicon modulates expression of *Arabidopsis thaliana* genes involved in copper stress. *J. Plant Physiol.* 168, 699–705. doi: 10.1016/j.jplph.2010.09.009

Khattab, H. I., Emam, M. A., Emam, M. M., Helal, N. M., and Mohamed, M. R. (2014). Effect of selenium and silicon on transcription factors NAC5 and DREB2A involved in drought-responsive gene expression in rice. *Biol. Plant.* 58, 265–273. doi: 10.1007/s10535-014-0391-z

Kumar, A., Kumar, R., Kumari, M., and Goldar, S. (2020). Enhancement of plant growth by using PGPR for a sustainable agriculture: a review. *Int. J. Curr. Microbiol. App. Sci.* 9, 152–165. doi: 10.20546/ijcmas.2020.902.019

Kumar, R., Trivedi, K., Anand, K. G. V., and Ghosh, A. (2019). Science behind biostimulant action of seaweed extract on growth and crop yield: insights into transcriptional changes in roots of maize treated with Kappaphycus alvarezii seaweed extract under soil moisture stressed conditions. *J. Appl. Phycol.* 32, 599–613. doi: 10.1007/s10811-019-01938-y

Laloum, T., De Mita, S., Gamas, P., Baudin, M., and Niebel, A. (2013). CCAAT-box binding transcription factors in plants: y so many? *Trends Plant Sci.* 18, 157–166. doi: 10.1016/j.tplants.2012.07.004

Laohavisit, A., Shang, Z., Rubio, L., Cuin, T. A., Véry, A. A., Wang, A., et al. (2012). *Arabidopsis* annexin1 mediates the radical-activated plasma membrane Ca2+-and K+-permeable conductance in root cells. *Plant Cell* 24, 1522–1533. doi: 10.1105/tpc.112.097881

Le, M. H., Cao, Y., Zhang, X. C., and Stacey, G. (2014). LIK1, a CERK1-interacting kinase, regulates plant immune responses in *Arabidopsis*. *PLoS One* 9:e102245. doi: 10.1371/journal.pone.0102245

Lei, Z., Chen, J., and Technol, X. C.-J. (2007). *The Physiological Functions of Calcium-Dependent Protein Kinases in Plant Calcium Signal Transduction. en.cnki.com.cn*. Available online at: http://en.cnki.com.cn/Article_en/CJFDTotal-FJLK200703061.htm (accessed July 30, 2019).

Li, H., Gao, Z., Chen, Q., Li, Q., Luo, M., Wang, J., et al. (2020a). Grapevine ABA receptor VvPYL1 regulates root hair development in transgenic *Arabidopsis*. *Plant Physiol. Biochem.* 149, 190–200. doi: 10.1016/j.plaphy.2020.02.008

Li, L., Shuai, L., Sun, J., Li, C., Yi, P., Zhou, Z., et al. (2020b). The Role of 1-Methylcyclopropene in the regulation of ethylene biosynthesis and ethylene receptor gene expression in *Mangifera indica* L. (Mango Fruit). *Food Sci. Nutr.* 8, 1284–1294. doi: 10.1002/fsn3.1417

Li, S. (2019). Regulation of ribosomal proteins on viral infection. *Cells* 8:508. doi: 10.3390/cells8050508

Li, W., Pang, S., Lu, Z., and Jin, B. (2020c). Function and mechanism of WRKY transcription factors in abiotic stress responses of plants. *Plants* 9:1515. doi: 10.3390/plants9111515

Li, X., Zhu, Y., Chen, C., Geng, Z., Li, X., Reports, T. Y.-S., et al. (2020d). cloning and characterization of two chlorophyll A/B binding protein genes and analysis of their gene family in *Camellia sinensis*. *Sci. Rep.* 10, 1–9.

Li, Y. J., Fang, Y., Fu, Y. R., Huang, J. G., Wu, C. A., and Zheng, C. C. (2013). NFYA1 is involved in regulation of postgermination growth arrest under salt stress in *Arabidopsis*. *PLoS One* 8:e0061289. doi: 10.1371/journal.pone.0061289

Li, Y., Zhang, Q., Ou, L., Ji, D., Liu, T., Lan, R., et al. (2020e). Response to the cold stress signaling of the tea plant (*Camellia sinensis*) elicited by chitosan oligosaccharide. *Agronomy* 10:915. doi: 10.3390/agronomy10060915

Lin, R.-C., Park, H.-J., and Wang, H.-Y. (2008). Role of *Arabidopsis* RAP2.4 in regulating light- and ethylene-mediated developmental processes and drought stress tolerance. *Mol. Plant* 1, 42–57. doi: 10.1093/mp/ssm004

Liu, P., Yin, L., Deng, X., Wang, S., Tanaka, K., and Zhang, S. (2014). Aquaporin-mediated increase in root hydraulic conductance is involved in silicon-induced improved root water uptake under osmotic stress in *Sorghum bicolor* L. *J. Exp. Bot.* 65, 4747–4756. doi: 10.1093/jxb/eru220

Liu, T. Y., Huang, T. K., Tseng, C. Y., Lai, Y. S., Lin, S. I., Lin, W. Y., et al. (2012). PHO2-dependent degradation of PHO1 modulates phosphate homeostasis in *Arabidopsis*. *Plant Cell* 24, 2168–2183. doi: 10.1105/tpc.112.096636

Liu, X., Huang, D., Tao, J., Miller, A. J., Fan, X., and Xu, G. (2014). Identification and functional assay of the interaction motifs in the partner protein OsNAR2.1 of the two-component system for high-affinity nitrate transport. *New Phytol.* 204, 74–80. doi: 10.1111/nph.12986

Luo, D., Niu, X., Yu, J., Yan, J., Gou, X., Lu, B. R., et al. (2012). Rice choline monooxygenase (OsCMO) protein functions in enhancing glycine betaine biosynthesis in transgenic tobacco but does not accumulate in rice (*Oryza sativa* L. ssp. japonica). *Plant Cell Rep.* 31, 1625–1635. doi: 10.1007/s00299-012-1276-2

Luo, X., Bai, X., Sun, X., Zhu, D., Liu, B., Ji, W., et al. (2013). Expression of wild soybean WRKY20 in *Arabidopsis* enhances drought tolerance and regulates ABA signalling. *J. Exp. Bot.* 64, 2155–2169. doi: 10.1093/jxb/ert073

Ma, W., Penrose, D. M., and Glick, B. R. (2002). Strategies used by rhizobia to lower plant ethylene levels and increase nodulation. *Can. J. Microbiol.* 48, 947–954. doi: 10.1139/w02-100

Manchester, L. C., Coto-Montes, A., Boga, J. A., Andersen, L. P. H., Zhou, Z., Galano, A., et al. (2015). Melatonin: an ancient molecule that makes oxygen metabolically tolerable. *J. Pineal Res.* 59, 403–419. doi: 10.1111/jpi.12267

Mastouri, F., Björkman, T., and Harman, G. E. (2012). Trichoderma harzianum enhances antioxidant defense of tomato seedlings and resistance to water deficit. *Mol. Plant Microbe Interact.* 25, 1264–1271. doi: 10.1094/MPMI

Matyjaszczyk, E. (2015). Products containing microorganisms as a tool in integrated pest management and the rules of their market placement in the European Union. *Pest Manag. Sci.* 71, 1201–1206. doi: 10.1002/ps.3986

Mauri, N., Fernández-Marcos, M., Costas, C., Desvoyes, B., Pichel, A., Caro, E., et al. (2016). GEM, a member of the GRAM domain family of proteins, is part of the ABA signaling pathway. *Sci. Rep.* 6, 1–11. doi: 10.1038/srep22660

Mayer Weber, R. L., Wiebke-Strohm, B., Bredemeier, C., Margis-Pinheiro, M., de Brito, G. G., Rechenmacher, C., et al. (2014). Expression of an osmotin-like protein from Solanum nigrum confers drought tolerance in transgenic soybean. *BMC Plant Biol.* 14:343. doi: 10.1186/s12870-014-0343-y

Mcgonigle, B., Keeler, S. J., Lau, S.-M. C., Koeppe, M. K., and O'keefe, D. P. (2000). *A Genomics Approach to the Comprehensive Analysis of the Glutathione S-Transferase Gene Family in Soybean and Maize*. Available online at: www.plantphysiol.org (accessed February 20, 2020).

Medina, J., Catalá, R., and Salinas, J. (2011). The CBFs: three *Arabidopsis* transcription factors to cold acclimate. *Plant Sci.* 180, 3–11. doi: 10.1016/j.plantsci.2010.06.019

Mehmood, S., Ahmed, W., Ikram, M., Imtiaz, M., Mahmood, S., Tu, S., et al. (2020). Chitosan modified biochar increases soybean (*Glycine max* L.) resistance to salt-stress by augmenting root morphology, antioxidant defense mechanisms and the expression of stress-responsive genes. *Plants* 9:1173. doi: 10.3390/plants9091173

Miao, Y., Lv, D., Wang, P., Wang, X. C., Chen, J., Miao, C., et al. (2006). An *Arabidopsis* glutathione peroxidase functions as both a redox transducer and a scavenger in abscisic acid and drought stress responses. *Plant Cell* 18, 2749–2766. doi: 10.1105/tpc.106.044230

Mohanta, T. K., Arora, P. K., Mohanta, N., Parida, P., and Bae, H. (2015). Identification of new members of the MAPK gene family in plants shows diverse conserved domains and novel activation loop variants. *BMC Genomics* 16:58. doi: 10.1186/s12864-015-1244-7

Moradi Tarnabi, Z., Iranbakhsh, A., Mehregan, I., and Ahmadvand, R. (2020). Impact of arbuscular mycorrhizal fungi (AMF) on gene expression of some cell wall and membrane elements of wheat (*Triticum aestivum* L.) under water deficit using transcriptome analysis. *Physiol. Mol. Biol. Plants* 26, 143–162. doi: 10.1007/s12298-019-00727-8

Muñoz-Mayor, A., Pineda, B., Garcia-Abellán, J. O., Antón, T., Garcia-Sogo, B., Sanchez-Bel, P., et al. (2012). Overexpression of dehydrin tas14 gene improves the osmotic stress imposed by drought and salinity in tomato. *J. Plant Physiol.* 169, 459–468. doi: 10.1016/j.jplph.2011.11.018

Mutava, R. N., Prince, S. J. K., Syed, N. H., Song, L., Valliyodan, B., Chen, W., et al. (2015). Understanding abiotic stress tolerance mechanisms in soybean: a comparative evaluation of soybean response to drought and flooding stress. *Plant Physiol. Biochem.* 86, 109–120. doi: 10.1016/j.plaphy.2014.11.010

Nagaraju, M., Reddy, P. S., Kumar, S. A., Kumar, A., Rajasheker, G., Rao, D. M., et al. (2020). Genome-wide identification and transcriptional profiling of small heat shock protein gene family under diverse abiotic stress conditions in *Sorghum bicolor* (L.). *Int. J. Biol. Macromol.* 142, 822–834. doi: 10.1016/j.ijbiomac.2019.10.023

Nair, P., Kandasamy, S., Zhang, J., Ji, X., Kirby, C., Benkel, B., et al. (2012). Transcriptional and metabolomic analysis of *Ascophyllum nodosum* mediated freezing tolerance in *Arabidopsis thaliana*. *BMC Genomics* 13:643. doi: 10.1186/1471-2164-13-643

Nawaz, M. A., Jiao, Y., Chen, C., Shireen, F., Zheng, Z., Imtiaz, M., et al. (2018). Melatonin pretreatment improves vanadium stress tolerance of watermelon seedlings by reducing vanadium concentration in the leaves and regulating melatonin biosynthesis and antioxidant-related gene expression. *J. Plant Physiol.* 220, 115–127. doi: 10.1016/j.jplph.2017.11.003

Neupane, S., Schweitzer, S., Neupane, A., Andersen, E., Fennell, A., Zhou, R., et al. (2019). Identification and characterization of mitogen-activated protein kinase (MAPK) genes in sunflower (*Helianthus annuus* L.). *Plants* 8:28. doi: 10.3390/plants8020028

Nguyen, N. H., Jung, C., and Cheong, J. J. (2019). Chromatin remodeling for the transcription of type 2C protein phosphatase genes in response to salt stress. *Plant Physiol. Biochem.* 141, 325–331. doi: 10.1016/j.plaphy.2019.06.012

Nishizawa, A., Yabuta, Y., and Shigeoka, S. (2008). Galactinol and raffinose constitute a novel function to protect plants from oxidative damage. *Plant Physiol.* 147, 1251–1263. doi: 10.1104/pp.108.122465

Oh, J. E., Kwon, Y., Kim, J. H., Noh, H., Hong, S. W., and Lee, H. (2011). A dual role for MYB60 in stomatal regulation and root growth of *Arabidopsis thaliana* under drought stress. *Plant Mol. Biol.* 77, 91–103. doi: 10.1007/s11103-011-9796-7

Olsson, A. S. B., Engström, P., and Söderman, E. (2004). The homeobox genes ATHB12 and ATHB7 encode potential regulators of growth in response to water deficit in *Arabidopsis*. *Plant Mol. Biol.* 55, 663–677. doi: 10.1007/s11103-004-1581-4

Owen, C., Patron, N. J., Huang, A., and Osbourn, A. (2017). Harnessing plant metabolic diversity. *Curr. Opin. Chem. Biol.* 40, 24–30. doi: 10.1016/j.cbpa.2017.04.015

Peltier, J. B., Ytterberg, J., Liberles, D. A., Roepstorff, P., and Van Wijk, K. J. (2001). Identification of a 350-kDa ClpP protease complex with 10 different Clp isoforms in chloroplasts of *Arabidopsis thaliana*. *J. Biol. Chem.* 276, 16318–16327. doi: 10.1074/jbc.M010503200

Petroni, K., Kumimoto, R. W., Gnesutta, N., Calvenzani, V., Fornari, M., Tonelli, C., et al. (2013). The promiscuous life of plant NUCLEAR FACTOR Y transcription factors. *Plant Cell* 24, 4777–4792. doi: 10.1105/tpc.112.105734

Petrozza, A., Santaniello, A., Summerer, S., Di Tommaso, G., Di Tommaso, D., Paparelli, E., et al. (2014a). *Physiological Responses to Megafol® Treatments in Tomato Plants Under Drought Stress: A Phenomic and Molecular Approach*. Amsterdam: Elsevier.

Petrozza, A., Santaniello, A., Summerer, S., Di Tommaso, G., Di Tommaso, D., Paparelli, E., et al. (2014b). Physiological responses to Megafol® treatments in tomato plants under drought stress: a phenomic and molecular approach. *Sci. Hortic.* 174, 185–192. doi: 10.1016/j.scienta.2014.05.023

Piccolo, A., and Spiteller, M. (2003). Electrospray ionization mass spectrometry of terrestrial humic substances and their size fractions. *Anal. Bioanal. Chem.* 377, 1047–1059. doi: 10.1007/s00216-003-2186-5

Pilon-Smits, E. A., Quinn, C. F., Tapken, W., Malagoli, M., and Schiavon, M. (2009). Physiological functions of beneficial elements. *Curr. Opin. Plant Biol.* 12, 267–274. doi: 10.1016/j.pbi.2009.04.009

Qin, F., Sakuma, Y., Tran, L. S. P., Maruyama, K., Kidokoro, S., Fujita, Y., et al. (2008). *Arabidopsis* DREB2A-interacting proteins function as Ring E3 ligases and negatively regulate plant drought stress-responsive gene expression. *Plant Cell* 20, 1693–1707. doi: 10.1105/tpc.107.057380

Rakoczy, M., Femiak, I., Alejska, M., Figlerowicz, M., and Podkowinski, J. (2018). Sorghum CCoAOMT and CCoAOMT-like gene evolution, structure, expression and the role of conserved amino acids in protein activity. *Mol. Genet. Genomics* 293, 1077–1089. doi: 10.1007/s00438-018-1441-6

Rayirath, P., Benkel, B., Mark Hodges, D., Allan-Wojtas, P., MacKinnon, S., Critchley, A. T., et al. (2009). Lipophilic components of the brown seaweed, *Ascophyllum nodosum*, enhance freezing tolerance in *Arabidopsis thaliana*. *Planta* 230, 135–147. doi: 10.1007/s00425-009-0920-8

Reddy, P. S., Sharma, K. K., Vadez, V., and Reddy, M. K. (2015). Molecular cloning and differential expression of cytosolic class I small Hsp gene family in *Pennisetum glaucum* (L.). *Appl. Biochem. Biotechnol.* 176, 598–612. doi: 10.1007/s12010-015-1598-y

Reeves, W. M., Lynch, T. J., Mobin, R., and Finkelstein, R. R. (2011). Direct targets of the transcription factors ABA-insensitive(ABI)4 and ABI5 reveal synergistic action by ABI4 and several bZIP ABA response factors. *Plant Mol. Biol.* 75, 347–363. doi: 10.1007/s11103-011-9733-9

Riaz Rajoka, M. S., Mehwish, H. M., Wu, Y., Zhao, L., Arfat, Y., Majeed, K., et al. (2020). Chitin/chitosan derivatives and their interactions with microorganisms: a comprehensive review and future perspectives. *Crit. Rev. Biotechnol.* 40, 365–379. doi: 10.1080/07388551.2020.1713719

Rowarth, N. M., Dauphinee, A. N., Denbigh, G. L., and Gunawardena, A. H. (2019). Hsp70 plays a role in programmed cell death during the remodelling of leaves of the lace plant (*Aponogeton madagascariensis*). *J. Exp. Bot.* 71, 907–918. doi: 10.1093/jxb/erz447

Saakre, M., Baburao, T. M., Salim, A. P., Ffancies, R. M., Achuthan, V. P., Thomas, G., et al. (2017). Identification and characterization of genes responsible for drought tolerance in rice mediated by *Pseudomonas* fluorescens. *Rice Sci.* 24, 291–298. doi: 10.1016/j.rsci.2017.04.005

Safdarian, M., Askari, H., Shariati, J. V., and Nematzadeh, G. (2019). Transcriptional responses of wheat roots inoculated with *Arthrobacter nitroguajacolicus* to salt stress. *Sci. Rep.* 9, 1–12. doi: 10.1038/s41598-018-38398-2

Saikia, J., Sarma, R. K., Dhandia, R., Yadav, A., Bharali, R., Gupta, V. K., et al. (2018). Alleviation of drought stress in pulse crops with ACC deaminase producing rhizobacteria isolated from acidic soil of Northeast India. *Sci. Rep.* 8, 1–16. doi: 10.1038/s41598-018-21921-w

Santaniello, A., Scartazza, A., Gresta, F., Loreti, E., Biasone, A., Di Tommaso, D., et al. (2017). *Ascophyllum nodosum* seaweed extract alleviates drought stress in *Arabidopsis* by affecting photosynthetic performance and related gene expression. *Front. Plant Sci.* 8:1362. doi: 10.3389/fpls.2017.01362

Sawers, R. J. H., Svane, S. F., Quan, C., Grønlund, M., Wozniak, B., Gebreselassie, M. N., et al. (2017). Phosphorus acquisition efficiency in arbuscular mycorrhizal maize is correlated with the abundance of root-external hyphae and the accumulation of transcripts encoding PHT1 phosphate transporters. *New Phytol.* 214, 632–643. doi: 10.1111/nph.14403

Schiavon, M., Pizzeghello, D., and Muscolo, A. (2010). High molecular size humic substances enhance phenylpropanoid metabolism in maize (*Zea mays* L .). *J. Chem. Ecol.* 36, 662–669. doi: 10.1007/s10886-010-9790-6

Seong, E. S., Cho, H. S., Choi, D., Joung, Y. H., Lim, C. K., Hur, J. H., et al. (2007). Tomato plants overexpressing CaKR1 enhanced tolerance to salt and oxidative stress. *Biochem. Biophys. Res. Commun.* 363, 983–988. doi: 10.1016/j.bbrc.2007.09.104

Sestili, F., Rouphael, Y., Cardarelli, M., Pucci, A., Bonini, P., Canaguier, R., et al. (2018). Protein hydrolysate stimulates growth in tomato coupled with N-dependent gene expression involved in N assimilation. *Front. Plant Sci.* 9:1233. doi: 10.3389/fpls.2018.01233

Seung, D., Risopatron, J. P. M., Jones, B. J., and Marc, J. (2012). Circadian clock-dependent gating in ABA signalling networks. *Protoplasma* 249, 445–457. doi: 10.1007/s00709-011-0304-3

Shan, Z., Jiang, Y., Li, H., Guo, J., Dong, M., Zhang, J., et al. (2020). Genome-wide analysis of the NAC transcription factor family in broomcorn millet (*Panicum miliaceum* L.) and expression analysis under drought stress. *BMC Genomics* 21:96. doi: 10.1186/s12864-020-6479-2

Sharma, S., Chen, C., Khatri, K., Rathore, M. S., and Pandey, S. P. (2019). Gracilaria dura extract confers drought tolerance in wheat by modulating abscisic acid homeostasis. *Plant Physiol. Biochem.* 136, 143–154. doi: 10.1016/j.plaphy.2019.01.015

Shen, J., Guo, M., Wang, Y., Yuan, X., Dong, S., Song, X., et al. (2020). An investigation into the beneficial effects and molecular mechanisms of humic acid on foxtail millet under drought conditions. *PLoS One* 15:e0234029. doi: 10.1371/journal.pone.0234029

Shi, Y., Zhang, Y., Han, W., Feng, R., Hu, Y., Guo, J., et al. (2016). Silicon enhances water stress tolerance by improving root hydraulic conductance in *Solanum lycopersicum* L. *Front. Plant Sci.* 7:196. doi: 10.3389/fpls.2016.00196

Shukla, P. S., Borza, T., Critchley, A. T., Hiltz, D., Norrie, J., and Prithiviraj, B. (2018). Ascophyllum nodosum extract mitigates salinity stress in *Arabidopsis thaliana* by modulating the expression of miRNA involved in stress tolerance and nutrient acquisition. *PLoS One* 13:e0206221. doi: 10.1371/journal.pone.0206221

Shukla, P. S., Shotton, K., Norman, E., Neily, W., Critchley, A. T., and Prithiviraj, B. (2017). Seaweed extract improve drought tolerance of soybean by regulating stress-response genes. *AoB Plants* 10:plx051. doi: 10.1093/aobpla/plx051

Singhal, P., Jan, A. T., Azam, M., and Haq, Q. M. R. (2016). Plant abiotic stress: a prospective strategy of exploiting promoters as alternative to overcome the escalating burden. *Front. Life Sci.* 9:52–63. doi: 10.1080/21553769.2015.1077478

Smoleń, S., Kowalska, I., Halka, M., Ledwożyw-Smoleń, I., Grzanka, M., Skoczylas, Ł., et al. (2019). Selected aspects of iodate and iodosalicylate metabolism in lettuce including the activity of vanadium dependent haloperoxidases as affected by exogenous vanadium. *Agronomy* 10:1. doi: 10.3390/agronomy10010001

Sofo, A., Nuzzaci, M., Vitti, A., Tataranni, G., and Scopa, A. (2014). "Control of biotic and abiotic stresses in cultivated plants by the use of biostimulant microorganisms," in *Improvement of Crops in the Era of Climatic Changes*, eds P. Ahmad, M. Wani, M. Azooz, and L. S. Tran (New York, NY: Springer).

Song, A., Li, P., Fan, F., Li, Z., and Liang, Y. (2014). The effect of silicon on photosynthesis and expression of its relevant genes in rice (*Oryza sativa* L.) under high-zinc stress. *PLoS One* 9:e0113782. doi: 10.1371/journal.pone.0113782

Stephenson, P. G., and Terry, M. J. (2008). Light signalling pathways regulating the Mg-chelatase branchpoint of chlorophyll synthesis during de-etiolation in *Arabidopsis thaliana*. *Photochem. Photobiol. Sci.* 7, 1243–1252. doi: 10.1039/b802596g

Sun, J., Qiu, C., Ding, Y., Wang, Y., Sun, L., Sun, L., et al. (2020). Fulvic acid ameliorates drought stress-induced damage in tea plants by regulating the ascorbate metabolism and flavonoids biosynthesis. *BMC Genomics* 21:411. doi: 10.1186/s12864-020-06815-4

Sun, Y., and Yu, D. (2015). Activated expression of AtWRKY53 negatively regulates drought tolerance by mediating stomatal movement. *Plant Cell Rep.* 34, 1295–1306. doi: 10.1007/s00299-015-1787-8

Suzuki, N., Koussevitzky, S., Mittler, R., and Miller, G. (2012). ROS and redox signalling in the response of plants to abiotic stress. *Plant Cell Environ.* 35, 259–270. doi: 10.1111/j.1365-3040.2011.02336.x

Suzuki, N., Miller, G., Morales, J., Shulaev, V., Torres, M. A., and Mittler, R. (2011). Respiratory burst oxidases: the engines of ROS signaling. *Curr. Opin. Plant Biol.* 14, 691–699. doi: 10.1016/j.pbi.2011.07.014

Taji, T., Ohsumi, C., Iuchi, S., Seki, M., Kasuga, M., Kobayashi, M., et al. (2002). Important roles of drought- and cold-inducible genes for galactinol synthase in stress tolerance in *Arabidopsis thaliana*. *Plant J.* 29, 417–426. doi: 10.1046/j.0960-7412.2001.01227.x

Takuhara, Y., Kobayashi, M., and Suzuki, S. (2011). Low-temperature-induced transcription factors in grapevine enhance cold tolerance in transgenic *Arabidopsis* plants. *J. Plant Physiol.* 168, 967–975. doi: 10.1016/j.jplph.2010.11.008

Tatematsu, K., Kamiya, Y., and Nambara, E. (2008). Co-regulation of ribosomal protein genes as an indicator of growth status. *Plant Signal. Behav.* 3, 450–452. doi: 10.4161/psb.3.7.5577

Thalhammer, A., Bryant, G., Sulpice, R., and Hincha, D. K. (2014). Disordered cold regulated15 proteins protect chloroplast membranes during freezing through binding and folding, But do not stabilize chloroplast enzymes in vivo. *Plant Physiol.* 166, 190–201. doi: 10.1104/pp.114.245399

Trevisan, S., Manoli, A., and Quaggiotti, S. (2019). A novel biostimulant, belonging to protein hydrolysates, mitigates abiotic stress effects on maize seedlings grown in hydroponics. *Agronomy* 9:28. doi: 10.3390/agronomy9010028

Trujillo-Moya, C., and Gisbert, C. (2012). The influence of ethylene and ethylene modulators on shoot organogenesis in tomato. *Plant Cell. Tissue Organ Cult.* 111, 41–48. doi: 10.1007/s11240-012-0168-z

Tunnacliffe, A., and Wise, M. J. (2007). The continuing conundrum of the LEA proteins. *Naturwissenschaften* 94, 791–812. doi: 10.1007/s00114-007-0254-y

Turk, H. (2019). Chitosan-induced enhanced expression and activation of alternative oxidase confer tolerance to salt stress in maize seedlings. *Artic. Plant Physiol. Biochem.* 141, 415-422. doi: 10.1016/j.plaphy.2019.06.025

Tuteja, N. (2007). Abscisic acid and abiotic stress signaling. *Plant Signal. Behav.* 2, 135–138. doi: 10.4161/psb.2.3.4156

Ugarte, R. A., Sharp, G., and Moore, B. (2006). "Changes in the brown seaweed Ascophyllum nodosum (L.) Le Jol. plant morphology and biomass produced by cutter rake harvests in southern New Brunswick, Canada," in *Proceedings of the Eighteenth International Seaweed Symposium*, (Dordrecht: Springer), 125–133.

Umezawa, T., Fujita, M., Fujita, Y., Yamaguchi-Shinozaki, K., and Shinozaki, K. (2006). Engineering drought tolerance in plants: discovering and tailoring genes to unlock the future. *Curr. Opin. Biotechnol.* 17, 113–122. doi: 10.1016/j.copbio.2006.02.002

Vishwakarma, K., Upadhyay, N., Kumar, N., Yadav, G., Singh, J., Mishra, R. K., et al. (2017). Abscisic acid signaling and abiotic stress tolerance in plants: a review on current knowledge and future prospects. *Front. Plant Sci.* 8:161. doi: 10.3389/fpls.2017.00161

Volpe, V., Chitarra, W., Cascone, P., Volpe, M. G., Bartolini, P., Moneti, G., et al. (2018). The association with two different arbuscular mycorrhizal fungi differently affects water stress tolerance in tomato. *Front. Plant Sci.* 9:1480. doi: 10.3389/fpls.2018.01480

Von Sydow, L., Schwenkert, S., Meurer, J., Funk, C., Mamedov, F., and Schröder, W. P. (2016). The PsbY protein of *Arabidopsis* photosystem II is important for the redox control of cytochrome b559. *Biochim. Biophys. Acta Bioenerg.* 1857, 1524–1533. doi: 10.1016/j.bbabio.2016.05.004

Vranova, V., Rejsek, K., Skene, K. R., and Formanek, P. (2011). Non-protein amino acids: plant, soil and ecosystem interactions. *Plant Soil* 342, 31–48. doi: 10.1007/s11104-010-0673-y

Wally, O. S. D., Critchley, A. T., Hiltz, D., Craigie, J. S., Han, X., Zaharia, L. I., et al. (2013). Regulation of phytohormone biosynthesis and accumulation in *Arabidopsis* following treatment with commercial extract from the marine macroalga *Ascophyllum nodosum*. *J. Plant Growth Regul.* 32, 324–339. doi: 10.1007/s00344-012-9301-9

Wang, H., Datla, R., Georges, F., Loewen, M., and Cutler, A. J. (1995). Promoters from kin1 and cor6.6, two homologous *Arabidopsis thaliana* genes: transcriptional regulation and gene expression induced by low temperature, ABA, osmoticum and dehydration. *Plant Mol. Biol.* 28, 605–617. doi: 10.1007/BF00021187

Wang, H., Zhou, L., Fu, Y., Cheung, M.-Y., Wong, F.-L., Phang, T.-H., et al. (2012). Expression of an apoplast-localized BURP-domain protein from soybean (GmRD22) enhances tolerance towards abiotic stress. *Plant. Cell Environ.* 35, 1932–1947. doi: 10.1111/j.1365-3040.2012.02526.x

Wang, J., Song, J., Clark, G., and Roux, S. J. (2018). ANN1 and ANN2 function in post-phloem sugar transport in root tips to affect primary root growth. *Plant Physiol.* 178, 390–401. doi: 10.1104/pp.18.00713

Wang, P., Xue, L., Batelli, G., Lee, S., Hou, Y. J., Van Oosten, M. J., et al. (2013). Quantitative phosphoproteomics identifies SnRK2 protein kinase substrates and reveals the effectors of abscisic acid action. *Proc. Natl. Acad. Sci. U.S.A.* 110, 11205–11210. doi: 10.1073/pnas.1308974110

Wang, P., Zhang, L., Jiang, X., Dai, X., Xu, L., Li, T., et al. (2018). Evolutionary and functional characterization of leucoanthocyanidin reductases from *Camellia sinensis*. *Planta* 247, 139–154. doi: 10.1007/s00425-017-2771-z

Wang, T., McFarlane, H. E., and Persson, S. (2016). The impact of abiotic factors on cellulose synthesis. *J. Exp. Bot.* 67, 543–552. doi: 10.1093/jxb/erv488

Wang, W., Jiang, W., Liu, J., Li, Y., Gai, J., and Li, Y. (2017). Genome-wide characterization of the aldehyde dehydrogenase gene superfamily in soybean and its potential role in drought stress response. *BMC Genomics* 18:518. doi: 10.1186/s12864-017-3908-y

Wang, W., Xia, M. X., Chen, J., Yuan, R., Deng, F. N., and Shen, F. F. (2016). Gene expression characteristics and regulation mechanisms of superoxide dismutase and its physiological roles in plants under stress. *Biochemistry* 81, 465–480. doi: 10.1134/S0006297916050047

Wang, X., Gao, F., Bing, J., Sun, W., Feng, X., Ma, X., et al. (2019). Overexpression of the jojoba aquaporin gene, ScPIP1, enhances drought and salt tolerance in transgenic *Arabidopsis*. *Int. J. Mol. Sci.* 20:153. doi: 10.3390/ijms20010153

Wu, Y., Jin, X., Liao, W., Hu, L., Dawuda, M. M., Zhao, X., et al. (2018). 5-aminolevulinic acid (ALA) alleviated salinity stress in cucumber seedlings by enhancing chlorophyll synthesis pathway. *Front. Plant Sci.* 9:635. doi: 10.3389/fpls.2018.00635

Xavier, L. J. C., and Boyetchko, S. M. (2002). Arbuscular mycorrhizal fungi as biostimulants and bioprotectants of crops. *Appl. Mycol. Biotechnol.* 2, 311–340.

Xiong, Y., Contento, A. L., Nguyen, P. Q., and Bassham, D. C. (2007). Degradation of oxidized proteins by autophagy during oxidative stress in *Arabidopsis*. *Plant Physiol.* 143, 291–299. doi: 10.1104/pp.106.092106

Xu, Z., Gongbuzhaxi, Wang, C., Xue, F., Zhang, H., and Ji, W. (2015). Wheat NAC transcription factor TaNAC29 is involved in response to salt stress. *Plant Physiol. Biochem.* 96, 356–363. doi: 10.1016/j.plaphy.2015.08.013

Yakhin, O. I., Lubyanov, A. A., Yakhin, I. A., and Brown, P. H. (2017). Biostimulants in plant science: a global perspective. *Front. Plant Sci.* 7:2049. doi: 10.3389/fpls.2016.02049

Yanhui, C., Xiaoyuan, Y., Kun, H., Meihua, L., Jigang, L., Zhaofeng, G., et al. (2006). The MYB transcription factor superfamily of *Arabidopsis*: expression analysis and phylogenetic comparison with the rice MYB family. *Plant Mol. Biol.* 60, 107–124. doi: 10.1007/s11103-005-2910-y

Yasin, N. A., Akram, W., Khan, W. U., Ahmad, S. R., Ahmad, A., and Ali, A. (2018). Halotolerant plant-growth promoting rhizobacteria modulate gene expression and osmolyte production to improve salinity tolerance and growth in *Capsicum annum* L. *Environ. Sci. Pollut. Res.* 25, 23236–23250. doi: 10.1007/s11356-018-2381-8

Ye, L., Zhao, X., Bao, E., Cao, K., and Zou, Z. (2019). Effects of arbuscular mycorrhizal fungi on watermelon growth, elemental uptake, antioxidant, and photosystem ii activities and stress-response gene expressions under salinity-alkalinity stresses. *Front. Plant Sci.* 10:863. doi: 10.3389/fpls.2019.00863

Zhang, S., Gan, Y., and Xu, B. (2016). Application of plant-growth-promoting fungi *Trichoderma longibrachiatum* T6 enhances tolerance of wheat to salt stress through improvement of antioxidative defense system and gene expression. *Front. Plant Sci.* 07:1405. doi: 10.3389/fpls.2016.01405

Zhang, S., and Scheller, H. V. (2004). Light-harvesting complex II binds to several small subunits of photosystem I. *J. Biol. Chem.* 279, 3180–3187. doi: 10.1074/jbc.M311640200

Zhang, X., Han, C., Gao, H., and Cao, Y. (2019). Comparative transcriptome analysis of the garden asparagus (*Asparagus officinalis* L.)reveals the molecular mechanism for growth with arbuscular mycorrhizal fungi under salinity stress. *Plant Physiol. Biochem.* 141, 20–29. doi: 10.1016/j.plaphy.2019.05.013

Zhao, Y., Dong, W., Zhang, N., Ai, X., Wang, M., Huang, Z., et al. (2014). A wheat allene oxide cyclase gene enhances salinity tolerance via jasmonate signaling. *Plant Physiol.* 164, 1068–1076. doi: 10.1104/pp.113.227595

Zheng, Y., Huang, Y., Xian, W., Wang, J., and Liao, H. (2012a). Identification and expression analysis of the Glycine max CYP707A gene family in response to drought and salt stresses. *Ann. Bot.* 110, 743–756. doi: 10.1093/aob/mcs133

Zheng, Y., Schumaker, K. S., and Guo, Y. (2012b). Sumoylation of transcription factor MYB30 by the small ubiquitin-like modifier E3 ligase SIZ1 mediates abscisic acid response in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U.S.A.* 109, 12822–12827. doi: 10.1073/pnas.1202630109

Ziosi, V., Zandoli, R., Di Nardo, A., Biondi, S., Antognoni, F., and Calandriello, F. (2013). Biological activity of different botanical extracts as evaluated by means of an array of in vitro and in vivo bioassays. *Acta Hortic.* 1009, 61–66. doi: 10.17660/ActaHortic.2013.1009.5

Conflict of Interest: SS-G and MV-C were employed by company UPL Mexico.

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

Copyright © 2021 González-Morales, Solís-Gaona, Valdés-Caballero, Juárez-Maldonado, Loredo-Treviño and Benavides-Mendoza. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 04 February 2021
doi: 10.3389/fgene.2021.584886

Genome-Wide Characterization of HSP90 Gene Family in Cucumber and Their Potential Roles in Response to Abiotic and Biotic Stresses

*Kaijing Zhang[1], Shuaishuai He[1], Yihu Sui[1], Qinghai Gao[1], Shuangshuang Jia[1], Xiaomin Lu[1] and Li Jia[2]**

[1]College of Agriculture, Anhui Science and Technology University, Fengyang, China, [2]Key Laboratory of Genetic Improvement and Ecophysiology of Horticultural Crop, Institute of Horticulture, Anhui Academy of Agricultural Sciences, Hefei, China

OPEN ACCESS

Edited by:
Hongjian Wan, Zhejiang Academy of Agricultural Sciences, China

Reviewed by:
Peng Wu, Yangzhou University, China
Si Shen, China Agricultural University, China

***Correspondence:**
Li Jia
jiali820@aaas.org.cn

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

Received: *18 July 2020*
Accepted: *14 January 2021*
Published: *04 February 2021*

Citation:
Zhang K, He S, Sui Y, Gao Q, Jia S, Lu X and Jia L (2021) Genome-Wide Characterization of HSP90 Gene Family in Cucumber and Their Potential Roles in Response to Abiotic and Biotic Stresses. Front. Genet. 12:584886. doi: 10.3389/fgene.2021.584886

Heat shock protein 90 (HSP90) possesses critical functions in plant developmental control and defense reactions. The HSP90 gene family has been studied in various plant species. However, the HSP90 gene family in cucumber has not been characterized in detail. In this study, a total of six HSP90 genes were identified from the cucumber genome, which were distributed to five chromosomes. Phylogenetic analysis divided the cucumber HSP90 genes into two groups. The structural characteristics of cucumber HSP90 members in the same group were similar but varied among different groups. Synteny analysis showed that only one cucumber HSP90 gene, Csa1G569290, was conservative, which was not collinear with any HSP90 gene in *Arabidopsis* and rice. The other five cucumber HSP90 genes were collinear with five *Arabidopsis* HSP90 genes and six rice HSP90 genes. Only one pair of paralogous genes in the cucumber HSP90 gene family, namely one pair of tandem duplication genes (Csa1G569270/Csa1G569290), was detected. The promoter analysis showed that the promoters of cucumber HSP90 genes contained hormone, stress, and development-related *cis*-elements. Tissue-specific expression analysis revealed that only one cucumber HSP90 gene Csa3G183950 was highly expressed in tendril but low or not expressed in other tissues, while the other five HSP90 genes were expressed in all tissues. Furthermore, the expression levels of cucumber HSP90 genes were differentially induced by temperature and photoperiod, gibberellin (GA), downy mildew, and powdery mildew stimuli. Two cucumber HSP90 genes, Csa1G569270 and Csa1G569290, were both differentially expressed in response to abiotic and biotic stresses, which means that these two HSP90 genes play important roles in the process of cucumber growth and development. These findings improve our understanding of cucumber HSP90 family genes and provide preliminary information for further studies of cucumber HSP90 gene functions in plant growth and development.

Keywords: cucumber, HSP90, abiotic stress, biotic stress, big data, expression profiles

INTRODUCTION

Plants are usually exposed to different kinds of environmental stress conditions during their growth and development. The primary environmental stresses included abiotic stresses such as high temperature, drought, chilling, salinity, and chemical pollutants and biotic stresses such as fungi, bacteria, viruses, and nematodes, which always cause cell injury and produce secondary stresses such as osmotic and oxidative stresses (Wang et al., 2004). In recent years, with the continuous development of global warming, high temperature stress has become one of the main factors affecting the normal growth and development of plants (Ahuja et al., 2010).

Heat shock proteins (HSPs) are a group of proteins induced by heat shock, which are widely distributed in various organisms from prokaryotes to eukaryotes (De Maio, 1999). Under the condition of high temperature stress, the organisms will activate and accumulate a large amount of the HSPs to participate in the heat shock response to maintain the stability of cells (Richter et al., 2010). Based on the molecular weight, HSPs are classified into five major families: sHSP, HSP60, HSP70, HSP90, and HSP100 (Wang et al., 2015; Muthusamy et al., 2017; Genest et al., 2019). HSP90 is a heat shock protein family with a molecular weight of about 90 kD, which is composed of three structural domains, including the N-terminal region containing ATP binding and hydrolysis sites, the middle region (M) and the C-terminal region containing dimerization regions (Rizzolo et al., 2014). HSP90 proteins are mainly distributed in cytoplasm, mitochondria, chloroplasts, and the endoplasmic reticulum (Ishiguro et al., 2002; Cao et al., 2003; Prassinos et al., 2008). HSP90 is an abundant and highly conserved molecular chaperone that is essential for viability in eukaryotes, which fulfills housekeeping functions on important biological processes such as signal transduction and cell cycle (Johnson and Brown, 2009).

The essential cellular functions of the molecular chaperone HSP90 have been intensively investigated in fungal and mammalian model systems (Sangster and Queitsch, 2005). The HSP90 genes have also been studied in some important plants, such as *Arabidopsis thaliana*, rice, *Populus*, tomato, pepper, and so on. In *Arabidopsis thaliana*, seven HSP90 family members were identified (Krishna and Gloor, 2001), which was the same number of HSP90 family members identified in tomato (Liu et al., 2014) and pepper (Wang et al., 2020). Nine and 10 HSP90 genes were identified in rice (Hu et al., 2009) and *Populus* (Zhang et al., 2013), respectively. Recently, the HSP90 gene family was also identified in *Nicotiana tabacum* (Song et al., 2019), *Aeluropus littoralis* (Hashemi-petroudi et al., 2019), *Hordeum vulgare* (Chaudhary et al., 2019), *Camellia sinensis* (Chen et al., 2018), and 44 plants (covering the stages from lower plants to higher plants, including algae, moss, ferns, gymnosperms, and angiosperms; Li et al., 2020). The diverse biological function of the HSP90 gene has also been well characterized in some plants. For example, in *Arabidopsis*, *AtHsp90.5* was important for chloroplast biogenesis and embryogenesis (Feng et al., 2014; Oh et al., 2014). In *Brassica napus*, HSP90 was found to play a role in the process of seed development and germination (Reddy et al., 1998). In cotton, HSP90 families played a crucial role in cotton fiber differentiation and development by maintaining cellular homeostasis (Sable et al., 2018). Plant HSP90 genes also played important roles in response to abiotic stresses. For instance, the HSP90-SGT1 chaperone system was required for the plant response to high temperature (Wang et al., 2016). The accumulation of HSP90 transcripts in *Brassica napus* was induced in response to the cold condition (Krishna et al., 1995). In rice, *OsHSP50.2* positively regulated drought stress tolerance probably by modulating ROS homeostasis and osmotic adjustment (Xiang et al., 2018). Overexpression of *AtHsp90.2*, *AtHsp90.5*, and *AtHsp90.7* in *Arabidopsis thaliana* enhanced plant sensitivity to salt and drought stresses (Song et al., 2009). Overexpression of *AtHsp90.3* in *Arabidopsis thaliana* impaired plant tolerance to heavy metal stress (Song et al., 2012). With the exception of HSP90 genes responding to abiotic stresses, HSP90 genes also played crucial roles in response to biotic stresses. *Arabidopsis* cytosolic HSP90 associated with and modulated the RPM1 disease resistance protein (Hubert et al., 2003). The suppression expression of *TaHsp90.2* or *TaHsp90.3* genes compromised the hypersensitive resistance response of wheat to stripe rust fungus (Wang et al., 2011). The above studies indicated that the HSP90 family genes are multifunctional in the life cycle of plants.

Cucumber (*Cucumis sativus* L.) is the first vegetable crop that finished the complete genome sequencing project (Huang et al., 2009), and is widely cultivated around the world [UN Food and Agriculture Organization Corporate Statistical Database (FAOSTAT), 2017]. With the high-quality genome information, a lot of gene families have been identified in cucumber, such as WRKY (Ling et al., 2011), MADS-box (Hu and Liu, 2012), NBS (Wan et al., 2013), bZIP (Baloglu et al., 2014), and so on. However, the identification of HSP90 gene family has not been reported so far, which largely limits the research of the biological function of HSP90 genes in cucumber. Only a few studies reported the functions of cucumber HSP genes, such as chitosan oligosaccharides treatment which resulted in upregulated HSP70 and HSP45.9 expression during cold storage (Ru et al., 2020) and ABA triggered the expression of HSP70 in an apoplastic H_2O_2-dependent manner under heat stress (Li et al., 2014). To the best of our knowledge, the functions of cucumber HSP90 genes were not reported up to now. Therefore, this study will identify the HSP90 gene family in the cucumber genome with bioinformatics methods, and analyze the physicochemical characteristics, chromosomal location, phylogenetic tree, gene structure, conserved motifs, homologous gene pairs, synteny, and *cis*-elements in the promoter of cucumber HSP90 genes. Furthermore, with the big data of cucumber transcriptome sequencing, the tissue-specific expression analysis of cucumber HSP90 family genes and their expression pattern analysis in response to the different stresses will be carried out. This study will preliminarily explore the relevant roles of cucumber HSP90 genes in response to abiotic and biotic stresses, which will create an important foundation for further research on the biological function of cucumber HSP90 genes, and provide a clear reference for cucumber resistance molecular breeding.

MATERIALS AND METHODS

Genome-Wide Identification of HSP90 Genes in Cucurbitaceae Crops

The whole genome (dna, cds, pep, and gff3) of eight Cucurbitaceae crops including cucumber, watermelon, melon, *Cucurbita maxima*, bottle gourd, wax gourd, luffa, and bitter gourd was downloaded from the Cucurbit Genomic Database[1] and CNSA database.[2] The Hidden Markov Model (HMM) profile of the HSP90 domain (PF00183) was downloaded from the Pfam database.[3] This model was used as a query to search for HSP90 proteins in the eight Cucurbitaceae crop protein files based on an expected value (E-value) cutoff of 1×10^{-5} in HMMER 3.0 (Finn et al., 2011). Subsequently, all the obtained HSP90 protein sequences were validated with the search results from the SMART[4] and NCBI CDD[5] databases. The protein sequences containing HSP90 and HATPase_c structure domains were finally determined as HSP90 proteins. The protein sequences of confirmed HSP90 family members were analyzed with the online tool ExPASy[6] to predict their physicochemical characteristics. To obtain chromosomal distribution of cucumber HSP90 genes, the DNA sequence of each HSP90 gene was used in Blastn search of the cucumber genome in the Cucurbit Genomics Database. The cucumber HSP90 genes were then mapped onto chromosomes using the software TBtools (Chen et al., 2020).

Structural Analysis and Phylogenetic Tree Construction of HSP90 Genes in Cucumber

The exon/intron structures of the HSP90 genes identified in the eight Cucurbitaceae crops were generated using the online tool Gene Structure Display Server[7] (Hu et al., 2015). The conserved motifs in the HSP90 proteins identified in the eight Cucurbitaceae crops were analyzed with MEME server[8] (Bailey et al., 2006) using the following parameters: maximum number of motifs, 10; minimum motif width, 6; and maximum motif width, 100. The conserved motifs were annotated with the Pfam database. Based on the results of HSP90 genes identified in *Arabidopsis* (Krishna and Gloor, 2001) and rice (Hu et al., 2009), seven *Arabidopsis* HSP90 proteins and nine rice HSP90 proteins were downloaded from the *Arabidopsis* Information Resource[9] and the Rice Genome Annotation Project[10], respectively. Multiple alignments of HSP90 protein sequences of cucumber, *Arabidopsis*, and rice were performed by Muscle in MEGA X (Kumar et al., 2018) with default parameters. Then, a phylogenetic tree was constructed on the basis of the alignment results using the maximum likelihood method with the following parameters: LG+G model, partial deletion, and 1,000 bootstrap tests. Another phylogenetic tree was constructed with the HSP90 proteins identified in the eight Cucurbitaceae crops using the maximum likelihood method.

Synteny Analysis of HSP90 Genes in Cucumber, *Arabidopsis*, and Rice

To understand the syntenic relationships among the cucumber HSP90 genes, *Arabidopsis* HSP90 genes, and rice HSP90 genes, synteny analysis of HSP90 genes in cucumber, *Arabidopsis*, and rice was conducted. MCScanX software (Wang et al., 2012) was used to analyze the syntenic relationships of the HSP90 genes in cucumber, *Arabidopsis*, and rice. Circos software (Krzywinski et al., 2009) was used to visualize the syntenic relationships of the HSP90 genes in cucumber, *Arabidopsis*, and rice.

Cis-Regulatory Element Prediction for Cucumber HSP90 Gene Promoters

The promoter sequence (1.5 kb DNA sequence upstream of the start codon) of each HSP90 gene was extracted from the cucumber genome. *Cis*-acting regulatory elements in the promoters of each cucumber HSP90 gene were then analyzed using the PlantCARE database[11] (Lescot et al., 2002). The predicted *cis*-acting regulatory elements were classified according to their regulatory functions.

Expression Profile Analyses of HSP90 Genes in Different Tissues of Cucumber

The RNA-seq data (PRJNA80169; Li et al., 2011) for cucumber gene expression in different tissues (root, stem, male flower, female flower, ovary, leaf, and tendril) were deposited into the Cucurbit Genomic Database. The expression profiles, as fragments per kilobase per million reads (RPKM), of all cucumber HSP90 genes were retrieved from PRJNA80169. An expression heatmap of cucumber HSP90 genes in different tissues was drawn using the TBtools software with Euclidean distances and the complete linkage method of hierarchical clustering.

Expression Profile Analyses of Cucumber HSP90 Genes in Response to Heat Treatment

Cucumber seeds were soaked overnight at room temperature. The germinated seeds were sow in the nursery tray at the same time. Seedlings were grown in a growth room at 25°C and 14 h light/10 h dark cycle for 2 weeks. For heat stress, uniform-sized seedlings were transferred to a growth chamber that maintained the temperature at 45°C. The leaves were collected at 0 (control), 3, 6, and 12 h (three repeats, the first true leaf was taken for each repeat). The collected samples were used for qRT-PCR analysis. The primers used in qRT-PCR for the six cucumber HSP90 genes are shown in **Supplementary Table S1**.

[1]http://www.icugi.org/
[2]https://db.cngb.org/cnsa/home/
[3]http://pfam.xfam.org/family/PF00183/hmm
[4]http://smart.embl.de/smart/batch.pl
[5]https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi
[6]http://web.expasy.org/protparam/
[7]http://gsds.cbi.pku.edu.cn/
[8]http://memesuite.org/
[9]http://www.arabidopsis.org/index.jsp
[10]http://rice.plantbiology.msu.edu/cgi-bin/ORF_infopage.cgi
[11]http://bioinformatics.psb.ugent.be/webtools/plantcare/html/

Expression Profile Analyses of Cucumber HSP90 Genes in Response to Abiotic and Biotic Stresses With Publicly Available Transcriptome Data

For abiotic and biotic treatments, RNA-seq data available in the Cucurbit Genomic Database under the series accession numbers of PRJNA376073 (GA treatment; Zhang et al., 2017), PRJNA285071 (downy mildew infection; Burkhardt and Day, 2016), PRJNA321023 (powdery mildew infection; Xu et al., 2017), and the GEO database under the series accession number of GSE117867 (temperature and photoperiod treatment; Zhang et al., 2018) were selected to analysis the biological functions of cucumber HSP90 genes. The expression profiles of cucumber HSP90 genes were retrieved from different RNA-seq data, respectively. The expression heatmaps of cucumber HSP90 genes under different stresses were drawn with the TBtools software.

RESULTS

Genome-Wide Identification and Chromosomal Location of HSP90 Family Genes in Cucumber

Based on the published cucumber genome information (ChineseLong_V2), a total of six members of the HSP90 gene family were identified using bioinformatics methods. The length of the encoded HSP90 proteins ranged from 631 (Csa1G569290) to 817 (Csa2G368880) amino acids, with CDS sizes of 1896–2,454 bp, and a molecular weight of 72.31–93.59 kD. The aliphatic index varied from 79.61 (Csa6G111370) to 85.08 (Csa1G569270). The theoretical isoelectric point (pI) of the six HSP90 proteins ranged from 4.86 (Csa5G135340) to 5.47 (Csa6G111370), indicating that all the cucumber HSP90 proteins were acidic (pI < 7.0). The instability indexes of Csa5G135340 and Csa6G111370 were greater than 40, suggesting that Csa5G135340 and Csa6G111370 were stable proteins. While the instability indexes of other HSP90 proteins were less than 40, which suggested that the other four HSP90 proteins were unstable. The grand average of hydropathicity (GRAVY) values of all HSP90 proteins were negative, indicating that all the cucumber HSP90 proteins were hydrophilic (**Table 1**).

Moreover, the HSP90 gene family in the other seven Cucurbitaceae crops including watermelon, melon, *Cucurbita maxima*, bottle gourd, wax gourd, luffa, and bitter gourd were also identified. Most Cucurbitaceae crops (diploid) had similar HSP90 members (5–7), which were similar to the HSP90 gene members in cucumber (6; **Table 1**). Whereas *Cucurbita maxima* was tetraploid, which had 10 HSP90 gene members. From the numbers of HSP90 gene family in different Cucurbitaceae crops, the HSP90 gene family in Cucurbitaceae crops was conservative.

Based on the chromosomal positions of the HSP90 genes annotated in the cucumber genome, the chromosomal locations of six HSP90 genes were marked on the physical maps of cucumber. In the cucumber genome, the six HSP90 family genes were distributed on five chromosomes of cucumber. Two HSP90 genes were located on chromosome 1, which contained the most HSP90 genes, whereas chromosomes 2, 3, 5, and 6 each contained only one HSP90 gene (**Supplementary Figure S1**).

Phylogenetic, Gene Structure, and Conserved Motifs Analysis of Cucumber HSP90 Proteins

Based on the unrooted phylogenetic tree of the six HSP90 protein sequences, the cucumber HSP90 gene family was classified into two major subgroups, namely subgroup A clustered with Csa1G569270, Csa1G569290, and Csa3G183950 and subgroup B clustered with Csa2G368880, Csa5G135340, and Csa6G111370 (**Supplementary Figure S2**). The exon/intron distribution showed that the number of exons in the six cucumber HSP90 genes varied from 1 (Csa1G569270) to 20 (Csa6G111370). The average number of exons in the HSP90 genes in subgroup A was the lowest, only 3.7, while the average number of exons in the HSP90 genes in subgroup B was the highest, 18 (**Supplementary Figure S2**). It is noteworthy that closely related HSP90 genes in the phylogenetic tree showed similar distribution of exon/intron, suggesting that the functions of HSP90 genes in the same subgroup were similar. A total of 10 conserved motifs were identified in all six cucumber HSP90 proteins using the MEME web server. Based on the annotation information on the Pfam database, motif 6 was annotated as HATPase_c, motif 4 had no annotation, and the other eight motifs were all annotated as HSP90 (**Supplementary Table S2**). In subgroup A, in addition to 10 motifs, each HSP90 protein contained one more motif 8 at the 3′ end, and the order of all motifs was consistent. In subgroup B, each HSP90 protein contained 10 motifs, and the order of all motifs was consistent (**Supplementary Figure S2**). These results revealed that the type, order, and number of motifs were similar in the HSP90 proteins within the same subgroup, indicating functional similarities among members of the same subgroup. Whereas the differences in motif distribution in HSP90 proteins among different subgroups revealed that the functions of these genes might have diverged during evolution.

Evolutionary Relationships and Classification of Cucumber HSP90 Family Genes

To examine the evolutionary relationships and classification of cucumber HSP90 genes, a maximum likelihood phylogenetic tree was generated based on the multiple sequence alignment of 22 full-length HSP90 protein sequences from cucumber (6), *Arabidopsis* (7), and rice (9). The evolutionary tree was divided into three subgroups: A, B, and C. There were, respectively, three cucumber HSP90 genes in the A and B subgroups (**Figure 1**), which were consistent with the results of the cluster analysis of HSP90 family genes in the cucumber (**Supplementary Figure S2**). Whereas subgroup C only contained one rice HSP90 gene (**Figure 1**), indicating that this gene was significantly different from the HSP90 genes in cucumber and *Arabidopsis*. Homologous genes were often clustered, indicating that cucumber HSP90 genes have a closer evolutionary relationship to those of *Arabidopsis* than to those of rice.

TABLE 1 | The physiochemical characteristics of HSP90 genes identified in the 8 Cucurbitaceae crops.

Organism	Gene ID	CDS size/bp	Number of amino acids/aa	Molecular weight/kD	pI	Instability index	Aliphatic index	Grand average of hydropathicity
Cucumber	Csa1G569270	2,100	699	79.93	4.99	35.19	85.08	−0.56
	Csa1G569290	1896	631	72.31	5.01	38.69	83.90	−0.57
	Csa2G368880	2,454	817	93.59	4.89	35.25	79.95	−0.72
	Csa3G183950	2,112	703	80.93	4.97	38.33	83.20	−0.61
	Csa5G135340	2,352	783	88.64	4.86	43.92	82.21	−0.53
	Csa6G111370	2,376	791	89.52	5.47	43.51	79.61	−0.53
Melon	MELO3C027461.2	792	263	29.86	4.73	30.94	86.77	−0.56
	MELO3C015295.2	261	86	9.54	4.23	39.66	79.42	−0.69
	MELO3C011021.2	2,271	756	87.02	5.21	32.54	83.66	−0.68
	MELO3C006935.2	2,295	764	87.89	5.06	38.15	86.78	−0.51
	MELO3C005179.2	2,352	783	88.95	4.90	44.14	82.09	−0.55
	MELO3C025976.2	1758	585	66.48	6.21	40.74	80.32	−0.51
Watermelon	Cla97C01G007690	2,370	789	89.56	4.89	43.38	81.81	−0.54
	Cla97C02G035700	2,364	787	89.15	5.58	38.27	79.90	−0.55
	Cla97C03G067590	2,100	699	80.03	5.01	34.70	84.94	−0.57
	Cla97C03G067600	2,100	699	80.03	5.01	34.70	84.94	−0.57
	Cla97C03G067620	780	259	29.56	4.84	43.24	82.86	−0.61
	Cla97C03G067630	1,296	431	49.47	5.22	35.21	84.13	−0.57
	Cla97C08G157760	2,454	817	93.45	4.88	33.33	79.95	−0.72
Cucurbita maxima	CmaCh02G002220	4,380	1,459	165.95	6.08	46.40	84.74	−0.42
	CmaCh04G021830	2,379	792	89.73	4.89	43.17	81.62	−0.50
	CmaCh05G003150	2,451	816	93.19	4.94	33.20	80.31	−0.68
	CmaCh06G006950	2097	698	79.79	4.97	34.47	85.20	−0.56
	CmaCh11G011980	2,106	701	80.27	4.99	33.59	84.15	−0.58
	CmaCh12G004010	2,559	852	97.73	5.00	33.39	81.37	−0.67
	CmaCh15G008490	2,370	789	89.70	4.94	43.32	81.67	−0.54
	CmaCh16G000980	2,115	704	80.87	4.98	30.09	82.26	−0.62
	CmaCh16G009910	2,103	700	80.29	4.97	37.85	84.10	−0.60
	CmaCh20G006760	1,122	373	41.71	8.94	47.81	84.75	−0.30
Bottle gourd	Lsi06G000940	3,243	1,080	122.20	8.47	41.19	86.03	−0.51
	Lsi06G000960	2,100	699	79.99	5.01	35.48	84.94	−0.57
	Lsi08G013090	2,562	853	97.63	4.92	32.96	80.23	−0.70
	Lsi09G007770	2,445	814	92.72	4.84	41.52	81.45	−0.53
	Lsi10G000540	4,416	1,471	167.61	6.10	46.52	83.20	−0.49
Wax gourd	Bhi02M001715	1,326	441	51.11	5.11	40.66	89.09	−0.52
	Bhi02M001717	1,266	421	48.74	5.10	39.67	88.93	−0.52
	Bhi04M000114	2,451	816	93.40	4.97	33.20	79.69	−0.72
	Bhi05M001813	2,112	703	80.92	5.01	36.89	82.65	−0.62
	Bhi10M001028	2,349	782	88.52	5.58	35.41	81.29	−0.52
	Bhi12M000916	2,352	783	88.85	4.95	43.46	82.55	−0.54
Luffa	Lcy01g016200	2,457	818	93.38	4.94	33.15	79.38	−0.71
	Lcy03g001820	2,100	699	79.97	5.00	32.79	84.81	−0.57
	Lcy03g001830	2,100	699	79.95	5.00	33.42	84.81	−0.56
	Lcy07g011100	2,388	795	90.11	5.52	36.90	80.82	−0.54
	Lcy09g008390	2,379	792	89.80	4.93	43.01	81.50	−0.54
	Lcy11g001270	2,112	703	80.87	4.99	38.20	81.82	−0.63
Bitter gourd	MC02g0401	2,199	732	83.24	5.30	41.58	80.85	−0.54
	MC04g1065	2,379	792	89.77	4.88	44.12	80.63	−0.56
	MC08g0753	2,454	817	93.21	4.89	30.40	82.01	−0.69
	MC10g0106	2,118	705	81.02	4.95	39.92	82.70	−0.62
	MC00g0429	2,106	701	80.10	4.96	34.64	83.87	−0.58
	MC00g0432	2,193	730	83.50	5.03	33.59	87.62	−0.48

The HSP90 genes with a evolutionary relationship were similar in gene structure and function, which suggested that the biological function of cucumber HSP90 genes could be predicted based on similar genes in *Arabidopsis*.

A maximum likelihood phylogenetic tree was also generated based on the multiple sequence alignment of 52 full-length HSP90 protein sequences from cucumber and the other seven Cucurbitaceae crops (**Figure 2**). The phylogenetic tree showed that the cucumber HSP90 proteins were closely clustered with melon HSP90 proteins. The subgroup A HSP90 proteins were functionally conservative in different Cucurbitaceae crops, which contained 1–2 subgroup A HSP90 proteins in each Cucurbitaceae crops. The subgroup B HSP90 proteins were functionally differentiated in different Cucurbitaceae crops. The HSP90 proteins in the same subgroup showed similar conserved motifs and similar distribution of exon/intron.

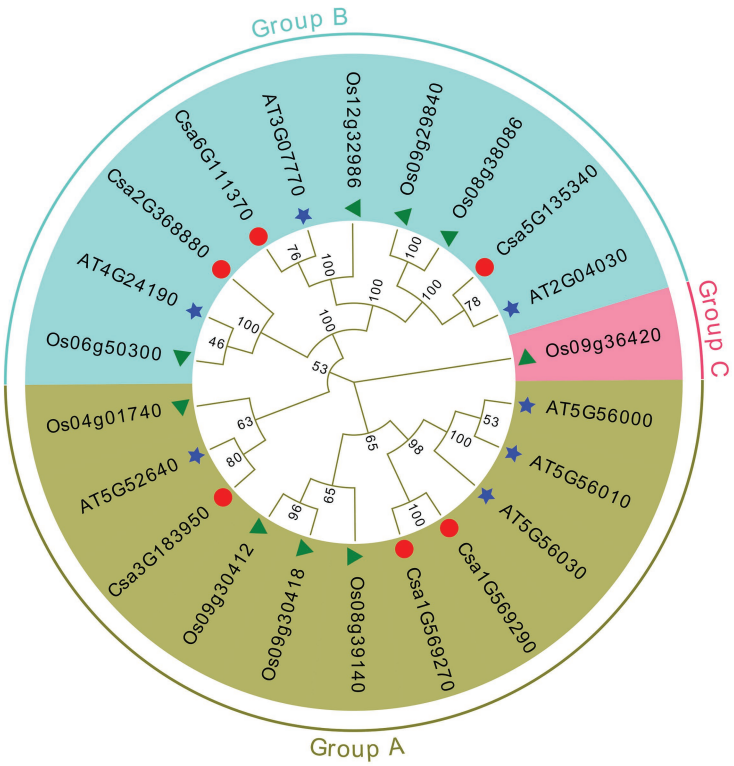


FIGURE 1 | Phylogenetic analysis of HSP90 proteins from cucumber, *Arabidopsis*, and rice.

Homologous Gene Pairs and Synteny Analysis of HSP90 Family Genes

Tandem and segmental duplications are reported to be the two main mechanisms underlying gene family expansion (Cannon et al., 2004). The analysis of cucumber HSP90 gene duplication events showed that there was only one pair of paralogous genes, namely one pair of tandem duplication genes (Csa1G569270/Csa1G569290) in the cucumber HSP90 gene family, and segmental duplications were not found. Orthologous gene pairs can provide effective information about evolutionary relationships between species (Wei et al., 2015). We therefore investigated the orthologous genes of the HSP90 family between cucumber and *Arabidopsis*, and cucumber and rice with synteny analysis. The results revealed that five cucumber HSP90 genes and five *Arabidopsis* HSP90 genes were orthologous genes with seven syntenic relationships. Five cucumber HSP90 genes and six rice HSP90 genes were orthologous genes with 10 syntenic relationships (**Figure 3**). Only one cucumber HSP90 gene Csa1G569290 did not form the syntenic relationships with neither *A. thaliana* nor rice, which suggested that Csa1G569290 was conservative in the cucumber HSP90 gene family.

Analysis of the Promoter Sequences of Cucumber HSP90 Genes

For cucumber HSP90 genes, 14 main types of *cis*-elements were identified in their promoter sequences. The number of various *cis*-elements in the promoters of each cucumber HSP90 gene is shown in **Figure 4A**. The largest number of *cis*-elements observed across the six HSP90 genes was associated with light-responsiveness, such as ACE, G-box, and MRE. Most notably, light-responsive *cis*-elements constituted more than half (up to 52%) of the presumptive *cis*-elements (**Figure 4B**). In addition, various *cis*-elements involved in hormone response (e.g., MeJA, gibberellins, abscisic acid, auxin, and salicylic acid), stress response (e.g., drought, defense, and stress), anaerobic induction, meristem expression, zein metabolism regulation, cell cycle regulation, endosperm expression, and circadian control were also identified in the promoter sequences of cucumber HSP90 genes. The different members of *cis*-elements were present in the promoter regions of different HSP90 genes, indicating that cucumber HSP90 genes performed multiple functions in the process of plant growth and development.

Tissue-Specific Expression Profile Analyses of Cucumber HSP90 Genes

Based on the published transcriptome sequencing data from different tissues of cucumber (PRJNA80169; Li et al., 2011), the expression heatmap of HSP90 family genes in different tissues of cucumber was drawn (**Figure 5**). The results revealed that the HSP90 gene Csa3G183950 was only highly expressed in tendril, but was low or not expressed in other tissues. Three

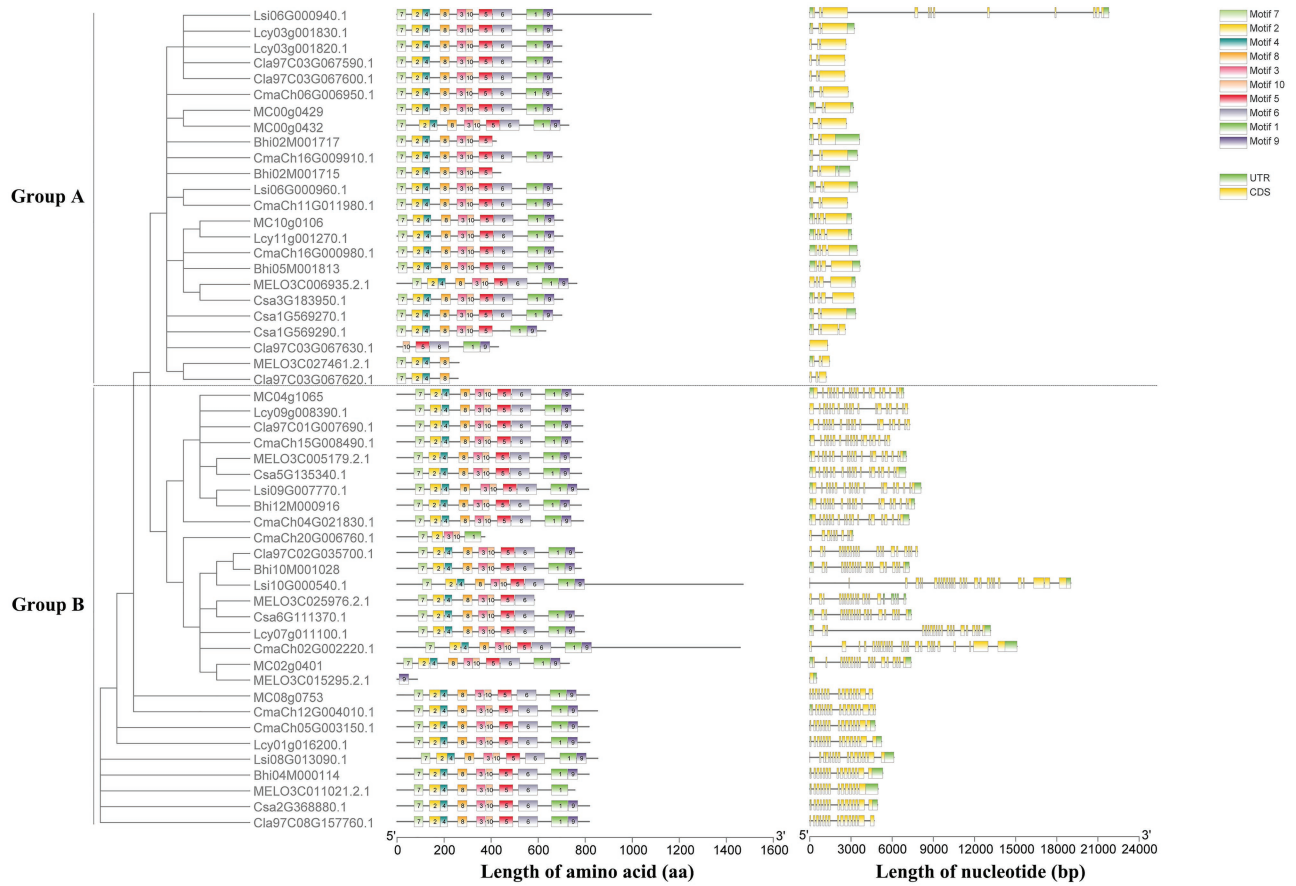


FIGURE 2 | Phylogenetic tree, conserved motifs of HSP90 proteins, and exon/intron structures of HSP90 genes in the eight Cucurbitaceae crops.

HSP90 genes, Csa1G569270, Csa1G569290, and Csa2G368880, were highly expressed in root, stem, male flower, female flower, ovary, leaf, and tendril, indicating that these three HSP90 genes played an important role in cucumber growth and development. The expression levels of HSP90 genes Csa5G135340 and Csa6G111370 were relatively low in male flower, but were high in other tissues.

Expression Profile Analyses of Cucumber HSP90 Genes Under Heat Stress

The qRT-PCR analysis of the six cucumber HSP90 genes showed that all of the six cucumber HSP90 genes responded to heat stress (**Figure 6**). Among them, Csa1G569270, Csa1G569290, Csa2G368880, and Csa3G1839050 were highly expressed after 3 h of heat stress, and then downregulated along with the time. The two cucumber HSP90 genes Csa5G13540 and Csa6G111370 were highly expressed after 3 h of heat stress, downregulated at 6 h, and then continued to upregulate at 12 h of heat stress. After the 12 h treatment of heat stress, the expression levels of the six cucumber HSP90 genes were still higher than that of blank control. The results indicated that all the cucumber HSP90 genes quickly responded to heat stress with a long response time.

Expression Profile Analyses of Cucumber HSP90 Genes Under Temperature and Photoperiod Treatment

Based on the published cucumber transcriptome sequencing data under temperature and photoperiod treatment (GSE117867; Zhang et al., 2018), the expression heatmap of the cucumber HSP90 gene family under temperature and photoperiod treatment was drawn (**Figure 7**). It was found that the expression levels of Csa1G569270 and Csa3G183950 at high temperature were higher than that at low temperature, indicating that these two HSP90 genes responded to high temperature stress. The expression level of the Csa1G569290 gene at high temperature was lower than that at low temperature, suggesting that the HSP90 gene Csa1G569290 responded to low temperature stress. The expression levels of Csa1G569270 and Csa1G569290 genes under high temperature and long day treatment were higher than those under high temperature and short day treatment, and the expression levels of these two HSP90 genes under low temperature and long day treatment were lower than those under low temperature and short day treatment, which indicated that these two HSP90 genes were associated with photoperiod response. The expression levels of Csa2G368880, Csa6G111370, and Csa5G135340 under temperature and photoperiodic stresses

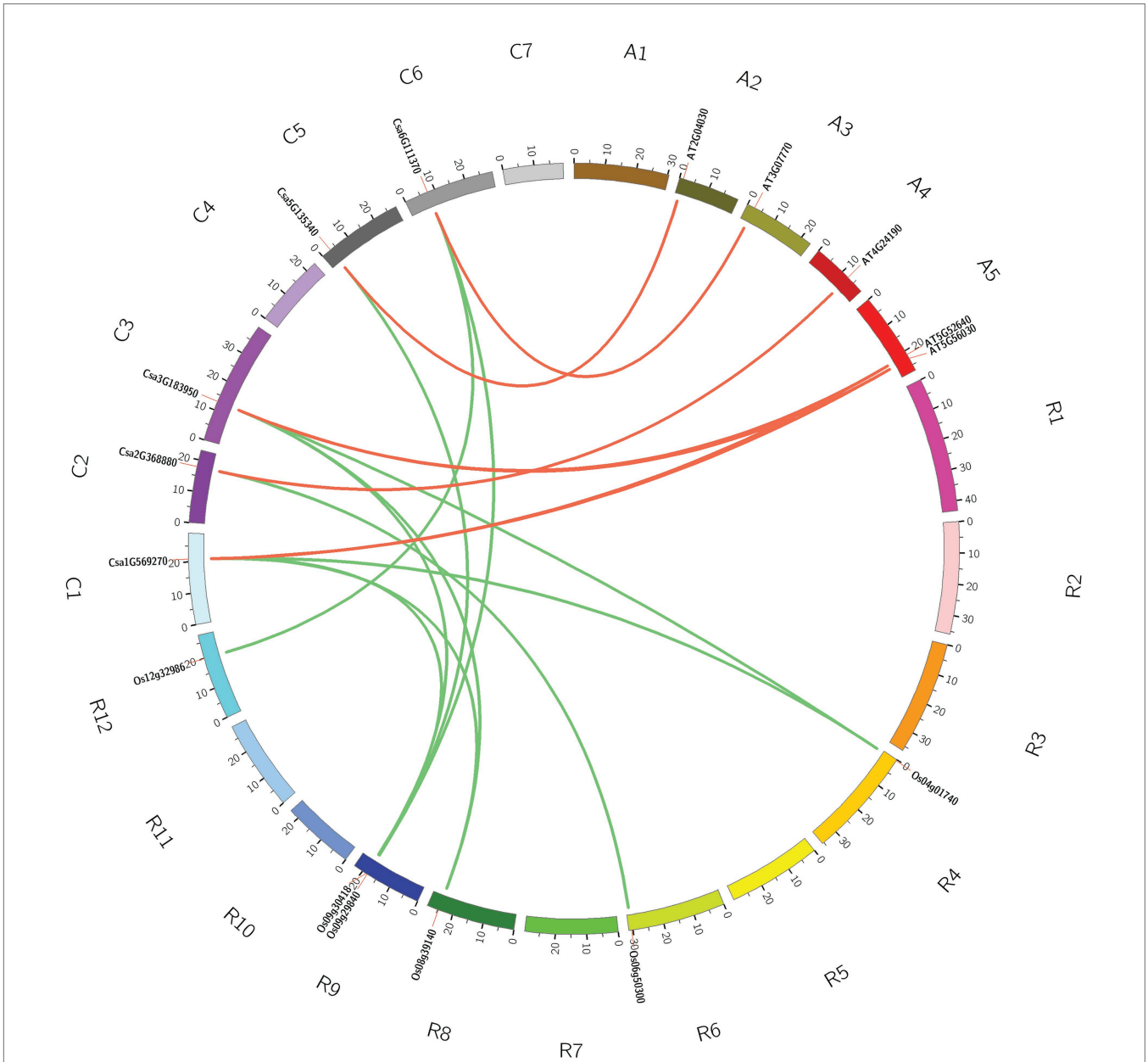


FIGURE 3 | Syntenic relationships of HSP90 gene family in cucumber, *Arabidopsis*, and rice.

were not changed, indicating that these three HSP90 genes did not respond to temperature and photoperiodic stresses.

Expression Profile Analyses of Cucumber HSP90 Genes Under GA Treatment

Based on the published transcriptome sequencing data of cucumber treated with gibberellin (GA) at different times (PRJNA376073; Zhang et al., 2017), the expression heatmap of the cucumber HSP90 gene family under GA treatment was plotted (**Figure 8**). The results revealed that the expression levels of all the six cucumber HSP90 family genes were significantly downregulated under GA treatment when compared with the control materials. With the extension of GA treatment time, the expression levels of the six HSP90 genes were all gradually decreased, indicating that these six cucumber HSP90 family genes were closely related to GA signal transduction.

Expression Profile Analyses of Cucumber HSP90 Genes in Response to Downy Mildew Infection

Based on the published transcriptome sequencing data of resistant and susceptible cucumber plants inoculated with downy mildew

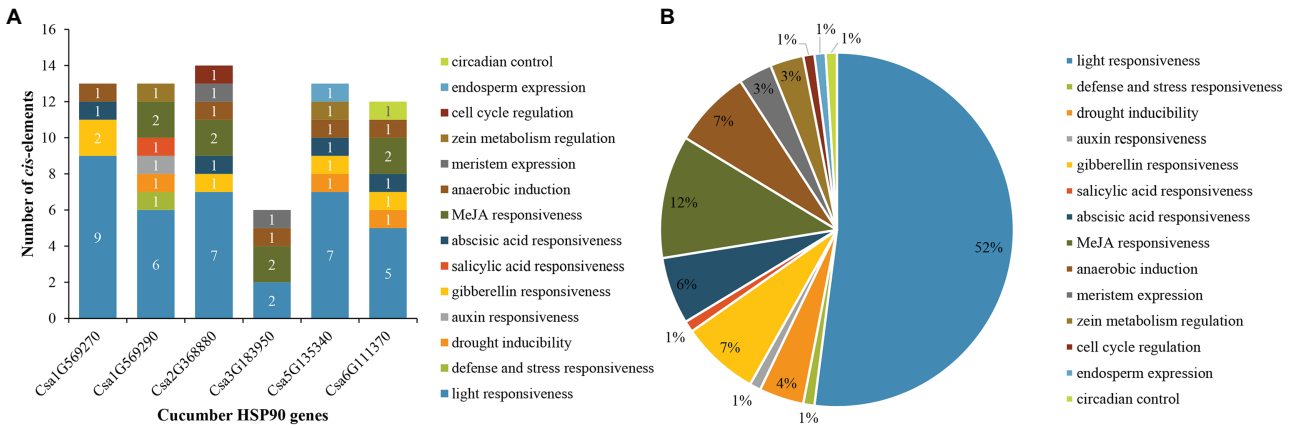


FIGURE 4 | Distribution of *cis*-elements in the promoters of cucumber HSP90 genes. **(A)** The number of various *cis*-elements in the promoters of each cucumber HSP90 gene. **(B)** The relative proportions of different *cis*-elements in the promoters of cucumber HSP90 genes are indicated by the pie chart. *Cis*-elements sharing identical or similar functions are represented by the same color.

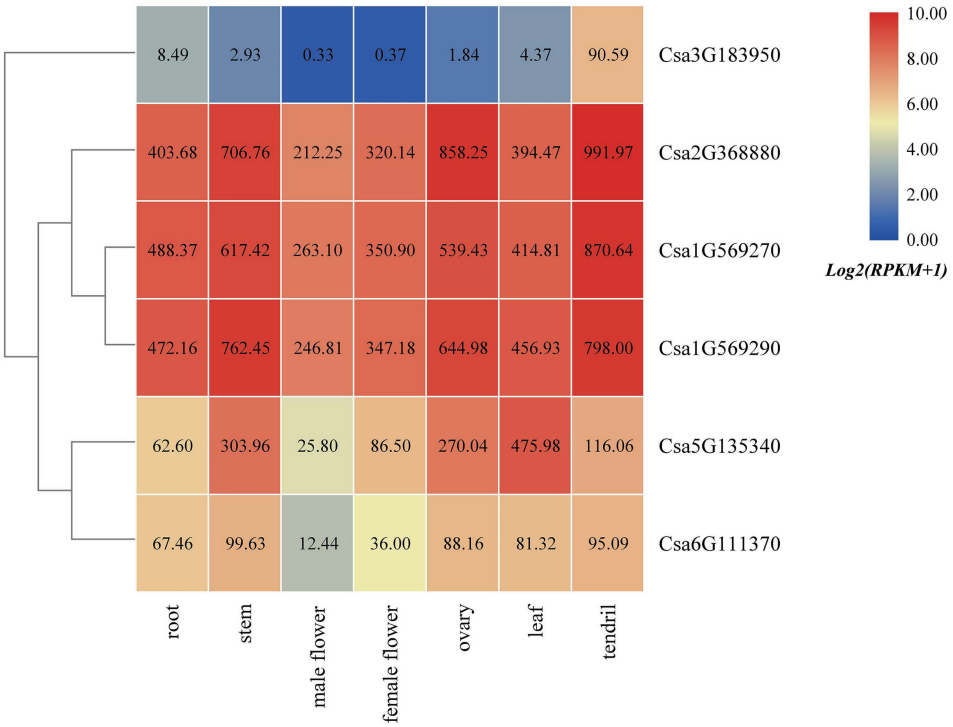


FIGURE 5 | The expression heatmap of the HSP90 gene family in different tissues of cucumber. The data in the boxes indicate original RPKM values.

for different time points (PRJNA285071; Burkhardt and Day, 2016), the expression heatmap of cucumber HSP90 genes under the treatment of downy mildew inoculation was drawn (**Figure 9**). The results showed that the Csa3G183950 gene was not expressed in resistant and susceptible cucumber materials, which may be related to the low or nil expression level of the Csa3G183950 gene in the leaf tissue. The expression levels of the Csa6G111370 gene in resistant and susceptible cucumber plants were not changed significantly after inoculation with the downy mildew pathogen, indicating that the Csa6G111370 gene did not function in response to downy mildew inoculation. The expression level of the Csa5G135340 gene was slightly upregulated in resistant cucumber material after inoculation with downy mildew, but was significantly downregulated in susceptible cucumber material after inoculation with downy mildew, which suggested that the expression patterns of the Csa5G135340 gene were different

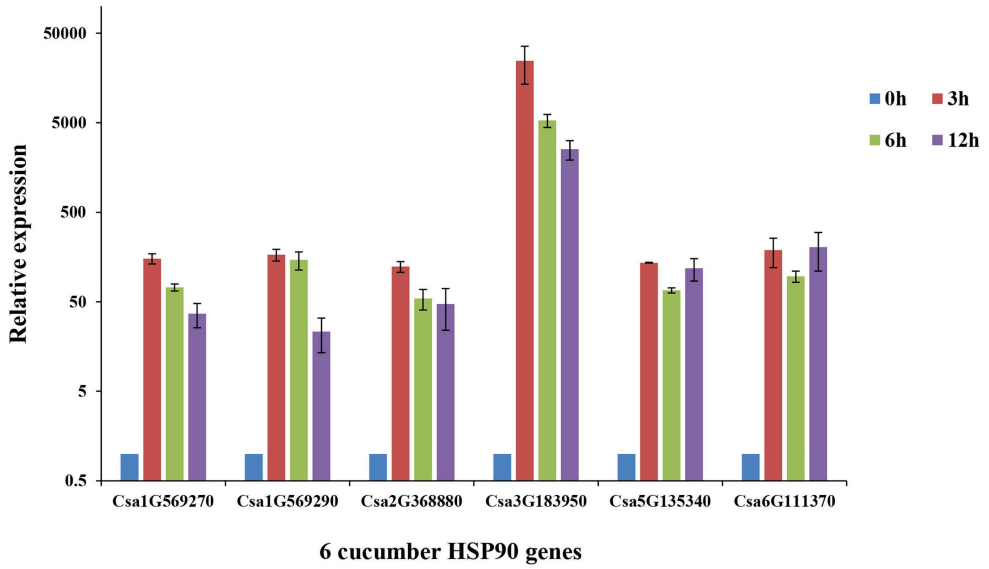


FIGURE 6 | Relative expression levels of six cucumber HSP90 gene after 0, 3, 6, and 12 h treatment of heat stress in cucumber seedlings by qRT-PCR analysis. Data are displayed using the *CsActin* gene as an internal control with three biological and three technical replicates. Values are the mean ± SD.

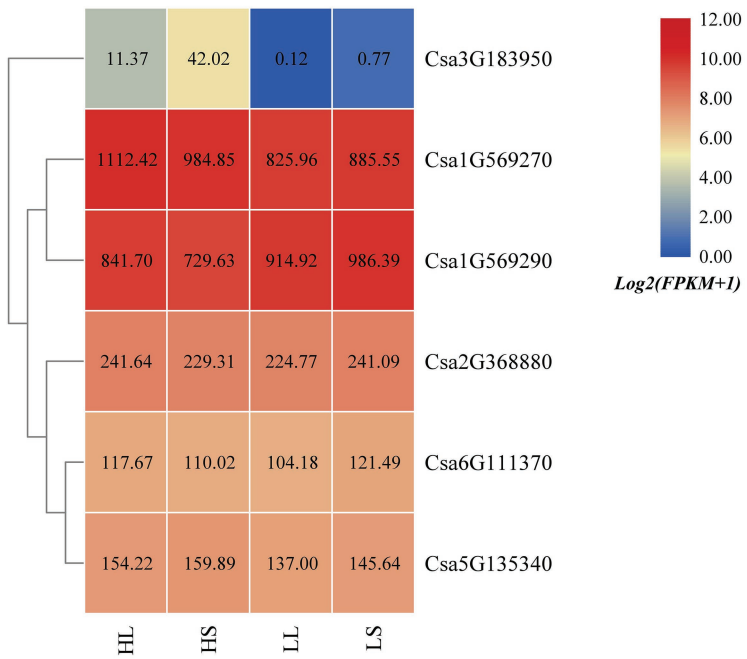


FIGURE 7 | The expression heatmap of the cucumber HSP90 gene family under temperature and photoperiod treatment. HL is high temperature and long day; HS is high temperature and short day; LL is low temperature and long day; LS is low temperature and short day. The data in the boxes indicate original FPKM values.

in the resistant and susceptible cucumber plants after inoculation with downy mildew. After inoculation with downy mildew, the expression levels of the Csa2G368880 gene in resistant and susceptible cucumber plants were all upregulated at first, and then decreased slowly, but the expression level on the 6th day was still seven times higher than that of the mock plant. After inoculation with downy mildew, the expression levels of the Csa1G569270 and Csa1G569290 genes in resistant and susceptible cucumber plants were upregulated first, and then decreased slowly. Six days post inoculation with downy mildew, the expression levels of the Csa1G569270 and Csa1G569290 genes in resistant cucumber plants were still about three times

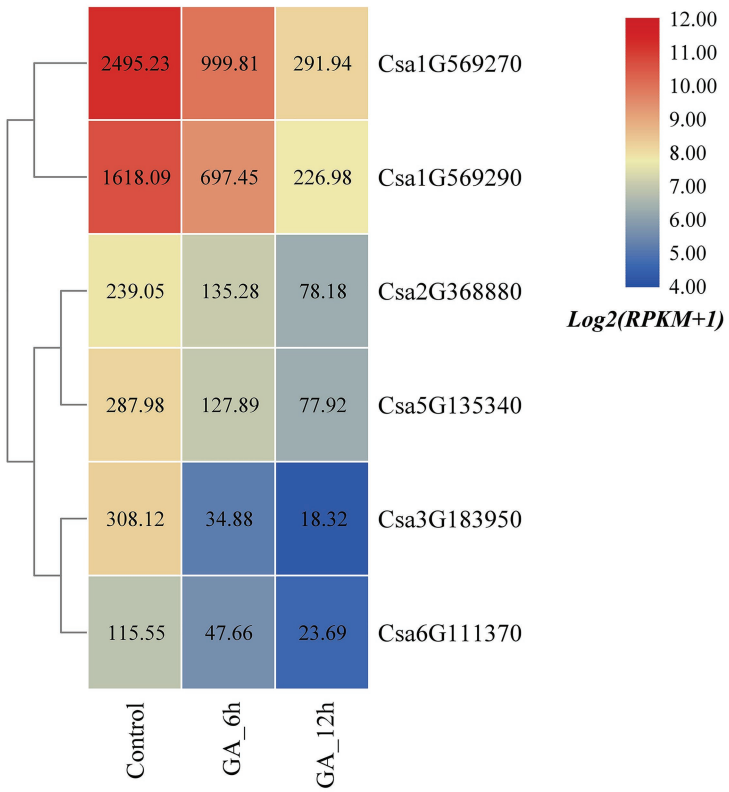


FIGURE 8 | The expression heatmap of the cucumber HSP90 gene family under GA treatment. Control is the control sample; GA_6h is the sample treated with GA for 6 h; GA_12h is the sample treated with GA for 12 h. The data in the boxes indicate original RPKM values.

the expression level in the control plant, whereas the expression levels of these two genes in susceptible cucumber plants were almost consistent with the expression levels in the control plant, which indicated that these two HSP90 genes may be involved in the regulation pathways related to cucumber downy mildew resistance.

Expression Profile Analyses of Cucumber HSP90 Genes in Response to Powdery Mildew Infection

Based on the published transcriptome sequencing data of resistant and susceptible cucumber plants inoculated with powdery mildew for different time points (PRJNA321023; Xu et al., 2017), the expression heatmap of cucumber HSP90 genes under the treatment of powdery mildew inoculation was drawn (**Figure 10**). The results showed that the expression levels of the Csa1G569270, Csa1G569290, and Csa3G183950 genes in resistant and susceptible cucumber plants were all significantly upregulated after inoculation with powdery mildew, and the differential expression ratio in the resistant cucumber was significantly higher than that in the susceptible cucumber, indicating that these three HSP90 genes may be associated with cucumber powdery mildew resistance. After inoculation with powdery mildew, the expression level of Csa5G135340 doubled in the resistant cucumber plant, but was not changed in the susceptible cucumber plant, suggesting that this HSP90 gene may also be involved in the regulation pathway related to cucumber powdery mildew resistance. The expression level of the Csa2G368880 gene in resistant and susceptible cucumber plants was not changed after inoculation with powdery mildew. The expression levels of the Csa6G111370 gene in resistant and susceptible cucumber plants were low, and the difference was not significant. These results revealed that these two HSP90 genes Csa2G368880 and Csa6G111370 did not function in response to powdery mildew infection in cucumber.

DISCUSSION

HSP90 is a class of chaperone protein that is highly conserved in prokaryotes and all eukaryotes, which is widely involved in signal transduction, cell cycle, and other important biological processes in organisms (Pearl and Prodromou, 2006). To date, the function of HSP90 genes have been widely investigated in animal and fungus systems, but lagged behind in plants. Recently, with the continuous development and extensive use of genome sequencing technology, more and more plant genome information has been published (Kersey, 2019). Many important gene families have been identified in plants. The HSP90 gene family has been successively identified in the model plants *Arabidopsis*

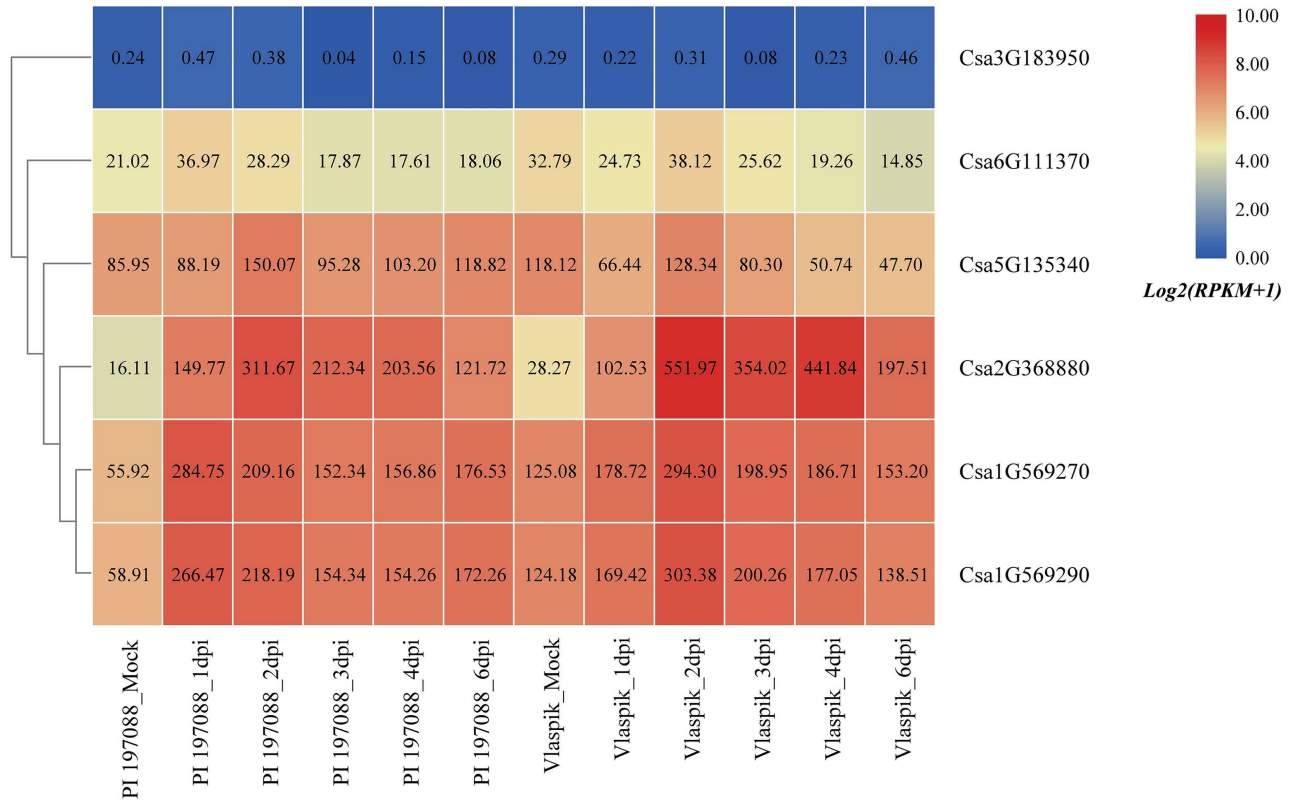


FIGURE 9 | The expression heatmap of the cucumber HSP90 gene family under the treatment of downy mildew inoculation. PI 197088 is the downy mildew resistant cucumber plant; Vlaspik is the downy mildew susceptible cucumber plant. Mock is the control sample. dpi means the days post inoculation. The data in the boxes indicate original RPKM values.

(Krishna and Gloor, 2001) and rice (Hu et al., 2009). Subsequently, the HSP90 gene family was also reported in some important vegetable crops, such as tomato (Liu et al., 2014) and pepper (Wang et al., 2020). Cucumber is an important vegetable crop widely cultivated in the world, whose cultivated area is next only to tomato and onion [UN Food and Agriculture Organization Corporate Statistical Database (FAOSTAT), 2017]. The cucumber genome was sequenced as early as 2009, which included high-quality genomic information (Huang et al., 2009). However, the genome-wide identification of the HSP90 gene family and their expression patterns analysis has not been conducted in cucumber. This greatly limited the study of the biological functions of HSP90 genes in cucumber. Therefore, in this study, with the cucumber genomic information and transcriptomic sequencing big-data, the HSP90 family genes were identified in cucumber, and the tissue-specific expression patterns of the cucumber HSP90 genes and expression profiles of the cucumber HSP90 genes in response to different stresses were analyzed. The results will provide vital references for further study of the biological functions of cucumber HSP90 genes and provide a theoretical basis for cucumber resistance molecular breeding.

In this study, a total of six HSP90 family genes were identified in cucumber, which was less than the members of the HSP90 gene family identified in *Arabidopsis thaliana* (7; Krishna and Gloor, 2001), rice (9; Hu et al., 2009), poplar (10; Zhang et al., 2013), tomato (7; Liu et al., 2014), and pepper (7; Wang et al., 2020). This may be one of the reasons for the differences in heat resistance among different plants. We also identified the HSP90 gene family in the other seven Cucurbitaceae crops including watermelon, melon, *Cucurbita maxima*, bottle gourd, wax gourd, luffa, and bitter gourd. Most Cucurbitaceae crops (diploid) had similar HSP90 members (5–7) with cucumber, except for the tetraploid crop *Cucurbita maxima* (10). It indicated that the HSP90 gene family in Cucurbitaceae crops was conservative. However, the number of cucumber HSP90 genes identified in the report of Li et al. (2020) was only four, which may be due to the different criterion for identification for the HSP90 gene family. Physicochemical characteristics analysis showed that all six HSP90 proteins in cucumber were acidic. This finding was consistent with the HSP90 proteins in other plants, such as *Arabidopsis*, rice, *Populus*, and tomato. While, except for one HSP90 protein, most of the HSP90 proteins in pepper were acidic. This indicated that the HSP90 proteins were conserved in different plant species. Phylogenetic tree analysis of 22 HSP90 genes derived from *Arabidopsis*, rice, and cucumber were divided into three subgroups: A, B, and C. The six HSP90 genes were classified into A and B subgroups, which was similar to the results of phylogenetic analysis for HSP90

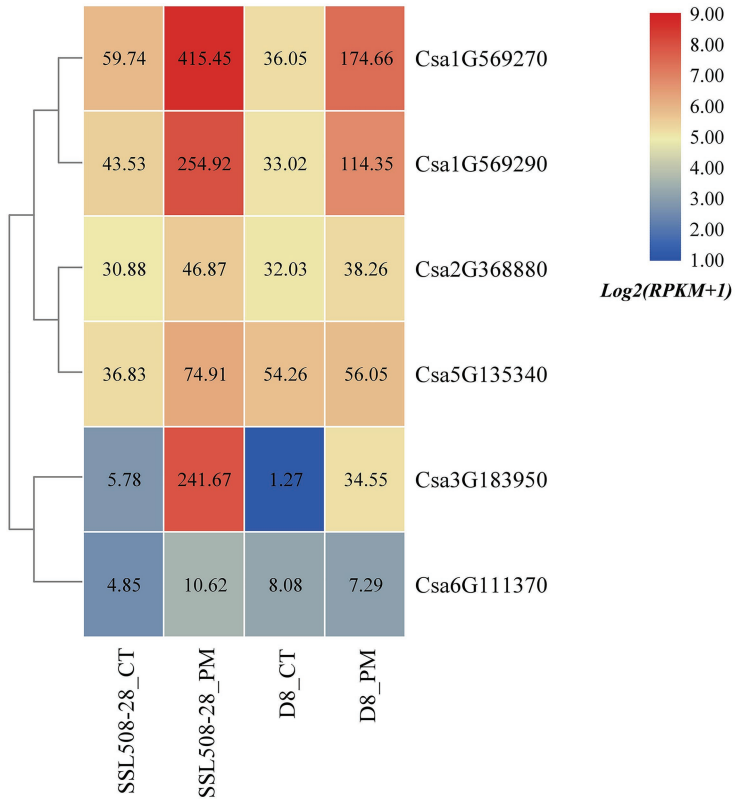


FIGURE 10 | The expression heatmap of the cucumber HSP90 gene family under the treatment of powdery mildew inoculation. SSL508-28 is the powdery mildew resistant cucumber plant; D8 is the powdery mildew susceptible cucumber plant. CT is the control sample. PM means powdery mildew inoculation. The data in the boxes indicate original RPKM values.

genes in *Arabidopsis* (Krishna and Gloor, 2001), rice (Zhang et al., 2016), and *Populus* (Hu et al., 2009). According to the phylogenetic and gene structure analysis, most of the HSP90 genes in each subgroup showed a similar exon/intron structure and conserved motif, the gene structure of HSP90 genes between subgroups A and B were significantly different, which indicated that the evolution might not only affect the gene function, but also the gene structure (Babenko et al., 2004; Roy and Penny, 2007). The similar gene structure in the same subgroup and the different gene structure between different subgroups were also identified in the HSP90 gene family of other plants, such as pepper and *Populus*. Only one rice HSP90 gene was classified into the C subgroup, which was consistent with the results of phylogenetic analysis of rice HSP90 genes (Zhang et al., 2016).

According to the synteny analysis of HSP90 family genes in *Arabidopsis*, rice, and cucumber, only one pair of paralogous genes in the cucumber HSP90 gene family, namely one pair of tandem duplication genes (Csa1G569270/Csa1G569290), was detected, and segmental duplications were not found, indicating that the expansion of the HSP90 gene in cucumber was mainly caused by tandem duplication genes. Only one cucumber HSP90 gene, Csa1G569290, was conservative and did not have collinearity with the HSP90 genes in *Arabidopsis* and rice. Whereas the other five cucumber HSP90 genes showed a variety of collinearity with the five HSP90 genes in *Arabidopsis* and the six HSP90 genes in rice, which were orthologous genes. These results indicated that the expansion of the HSP90 gene family in each species was performed in a specific way. This phenomenon was also common in the studies of other plant gene families (Zhang et al., 2005; Jain et al., 2006). The promoter analysis showed that the promoter of cucumber HSP90 genes contained various regulation elements, such as *cis*-acting regulatory elements essential for hormone response, stress response, meristem expression, cell cycle regulation, and so on, which further suggested that HSP90 genes play key roles in signal transduction, cell-cycle control, protein degradation, genomic silencing, and protein trafficking (Richter and Buchner, 2001; Young et al., 2001).

In recent years, the rapid development of high-throughput sequencing technologies has dramatically increased the throughput of sequence generation and decreased the overall cost (Wang et al., 2010). Many researchers have conducted a large number of cucumber transcriptome sequencing studies, forming the big data of cucumber transcriptome sequencing. The big data of cucumber transcriptome sequencing have been validated with the qRT-PCR analysis and peer-reviewed, which could be considered as reliable data. Therefore, the effective utilization of these big data of cucumber transcriptome sequencing can

not only reduce the research cost, but also facilitate the deep mining of the data of each transcriptome sequence. Besides, integrated analysis of the big data of cucumber transcriptome sequencing under different treatments is beneficial for studying the biological functions of cucumber genes. In this study, the published cucumber transcriptome sequencing big data were used to analyze the tissue-specific expression of six identified cucumber HSP90 family genes and their expression patterns in response to different stresses. The tissue-specific expression analysis showed that only cucumber HSP90 gene Csa3G183950 was highly expressed in tendril and low or not expressed in the other tissues, while the other five HSP90 genes were expressed or even highly expressed in all tissues. These results indicated that the functions of cucumber HSP90 genes were diverse in different tissues. Through the expression profile analyses of cucumber HSP90 family genes in response to abiotic stresses, two HSP90 genes, Csa1G569270 and Csa3G183950, were identified to respond to high temperature stress, which was consistent with the functions of *Arabidopsis* homologous HSP90 genes AT5G56010 and AT5G52640 (Sangster et al., 2007). The expression profiles of Csa1G569270 and Csa3G183950 genes under heat stress were also validated by the qRT-PCR analysis. While the expression levels of some other cucumber HSP90 genes were not changed in the publicly available transcriptome data, but upregulated in the qRT-PCR analysis. This may be due to the complex treatments in the transcriptome data which combined temperature and photoperiod. Cucumber HSP90 gene Csa1G569290 responded to low temperature stress, which showed similar functions with *Arabidopsis* HSP90 genes HSP90.2 (AT5G56010) and HSP90.3 (AT5G56030; Bao et al., 2014). Two cucumber HSP90 genes, Csa1G569270 and Csa1G569290, were related to photoperiod regulation. It had also been reported that *Arabidopsis* HSP90 genes were involved in photoperiod regulation (Ma, 2014). All the cucumber HSP90 genes were downregulated after GA treatment. This finding has not been reported in the other plant HSP90 genes. However, it has been reported that jasmonic acid (JA) signal was involved in the formation of the SGT1b-HSP70-HSP90 chaperone complex, and the SGT1b protein is necessary for plant response to JA, IAA, and GA stresses (Zhang et al., 2015). In addition, we found that the promoter sequence of cucumber HSP90 genes contained GA response elements. Therefore, it is believed that HSP90 genes were also involved in response to GA stress. Through the expression patterns analysis of cucumber HSP90 family genes in response to biotic stresses, two cucumber HSP90 genes, Csa1G569270 and Csa1G569290, were involved in the regulatory pathway related to cucumber downy mildew resistance. Zhang et al. (2019) reported that seven heat shock proteins (not clear whether they were HSP90) were differentially accumulated in cucumber under *Pseudoperonospora cubensis* infection. Previous studies had also found that wheat HSP90 genes were associated with disease resistance (Wang et al., 2011). Four cucumber HSP90 genes, Csa1G569270, Csa1G569290, Csa3G183950, and Csa5G135340, were involved in powdery mildew resistance in cucumber. It was also found that HSP90 genes performed the function of powdery mildew resistance in barley (Hein et al., 2005). Overall, it was found that two cucumber HSP90 genes, Csa1G569270 and Csa1G569290, were differentially expressed in response to abiotic stresses and biotic stresses, indicating that these two genes played an important role in cucumber growth and development.

CONCLUSION

In this study, HSP90 family genes were identified and characterized in the cucumber for the first time. A total of six HSP90 genes in cucumber were identified and characterized by the systematic analyses of physicochemical characteristics, chromosomal location, gene structure, conserved motif, phylogenetic tree, homologous gene pairs, synteny, and *cis*-elements in the promoters, which showed a clear evolutionary history for this family in cucumber. The expression patterns of cucumber HSP90 genes in different tissues and stresses responses were different, which coordinately regulated the growth and development of cucumber. Two cucumber HSP90 genes, Csa1G569270 and Csa1G569290, were both differentially expressed in response to abiotic and biotic stresses, which means these two HSP90 genes play important roles in the process of cucumber growth and development. This study could not only provide a scientific foundation for the comprehensive understanding of the cucumber HSP90 gene family, but could also be helpful in screening candidate genes for breeding new cucumber varieties with a high yield and stresses resistance.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

LJ conceived and designed the project. KZ, SH, and YS conducted the bioinformatics analysis. KZ, QG, and SJ performed the analysis of cucumber transcriptome sequencing big-data. KZ and XL wrote the paper. All authors reviewed and approved the final manuscript.

FUNDING

This research was supported by the Talent Foundation of Anhui Science and Technology University (100005) and the College Students' Innovative Entrepreneurial Training Plan Programs (202010879108 and X202010879087).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.584886/full#supplementary-material

REFERENCES

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

Babenko, V. N., Rogozin, I. B., Mekhedov, S. L., and Koonin, E. V. (2004). Prevalence of intron gain over intron loss in the evolution of paralogous gene families. *Nucleic Acids Res.* 32, 3724–3733. doi: 10.1093/nar/gkh686

Bailey, T. L., Williams, N., Misleh, C., and Li, W. W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* 34, W369–W373. doi: 10.1093/nar/gkl198

Baloglu, M. C., Eldem, V., Hajyzadeh, M., and Unver, T. (2014). Genome-wide analysis of the bZIP transcription factors in cucumber. *PLoS One* 9:e96014. doi: 10.1371/journal.pone.0096014

Bao, F., Huang, X., Zhu, C., Zhang, X., Li, X., and Yang, S. (2014). *Arabidopsis* HSP90 protein modulates RPP4-mediated temperature-dependent cell death and defense responses. *New Phytol.* 202, 1320–1334. doi: 10.1111/nph.12760

Burkhardt, A., and Day, B. (2016). Transcriptome and small RNAome dynamics during a resistant and susceptible interaction between cucumber and downy mildew. *Plant Genome* 9, 1–19. doi: 10.3835/plantgenome2015.08.0069

Cannon, S. B., Mitra, A., Baumgarten, A., Young, N. D., and May, G. (2004). The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biol.* 4:10. doi: 10.1186/1471-2229-4-10

Cao, D. S., Froehlich, J. E., Zhang, H., and Cheng, C. L. (2003). The chlorate-resistant and photomorphogenesis-defective mutant *cr88* encodes a chloroplast-targeted HSP90. *Plant J.* 33, 107–118. doi: 10.1046/j.1365-313X.2003.016011.x

Chaudhary, R., Baranwal, V. K., Kumar, R., Sircar, D., and Chauhan, H. (2019). Genome-wide identification and expression analysis of Hsp70, Hsp90, and Hsp100 heat shock protein genes in barley under stress conditions and reproductive development. *Funct. Integr. Genomic.* 19, 1007–1022. doi: 10.1007/s10142-019-00695-y

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

Chen, J., Gao, T., Wan, S., Zhang, Y., Yang, J., Yu, Y., et al. (2018). Genome-wide identification, classification and expression analysis of the HSP gene superfamily in tea plant (*Camellia sinensis*). *Int. J. Mol. Sci.* 19:2633. doi: 10.3390/ijms19092633

De Maio, A. (1999). Heat shock proteins: facts, thoughts, and dreams. *Shock* 11, 1–12. doi: 10.1097/00024382-199901000-00001

Feng, J., Fan, P., Jiang, P., Lv, S., Chen, X., and Li, Y. (2014). Chloroplast-targeted Hsp90 plays essential roles in plastid development and embryogenesis in *Arabidopsis* possibly linking with VIPP1. *Physiol. Plant.* 150, 292–307. doi: 10.1111/ppl.12083

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39, W29–W37. doi: 10.1093/nar/gkr367

Genest, O., Wickner, S., and Doyle, S. M. (2019). Hsp90 and Hsp70 chaperones: collaborators in protein remodeling. *J. Biol. Chem.* 294, 2109–2120. doi: 10.1074/jbc.REV118.002806

Hashemi-petroudi, S. H., Nematzadeh, G., Mohammadi, S., and Kuhlmann, M. (2019). Analysis of expression pattern of genome and analysis of HSP90 gene family in Aeluropus littoralis under salinity stress. *J. Crop Breed.* 11, 134–143.

Hein, I., Pacak, M. B., Hrubikova, K., Williamson, S., Dinesen, M., Soenderby, I. E., et al. (2005). Virus-induced gene silencing-based functional characterization of genes associated with powdery mildew resistance in barley. *Plant Physiol.* 138, 2155–2164. doi: 10.1104/pp.105.062810

Hu, W., Hu, G., and Han, B. (2009). Genome-wide survey and expression profiling of heat shock proteins and heat shock factors revealed overlapped and stress specific response under abiotic stresses in rice. *Plant Sci.* 176, 583–590. doi: 10.1016/j.plantsci.2009.01.016

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

Hu, L., and Liu, S. (2012). Genome-wide analysis of the MADS-box gene family in cucumber. *Genome* 55, 245–256. doi: 10.1139/g2012-009

Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., Fan, W., et al. (2009). The genome of the cucumber, *Cucumis sativus* L. *Nat. Genet.* 41, 1275–1281. doi: 10.1038/ng.475

Hubert, D. A., Tornero, P., Belkhadir, Y., Krishna, P., Takahashi, A., Shirasu, K., et al. (2003). Cytosolic HSP90 associates with and modulates the *Arabidopsis* RPM1 disease resistance protein. *EMBO J.* 22, 5679–5689. doi: 10.1093/emboj/cdg547

Ishiguro, S., Watanabe, Y., Ito, N., Nonaka, H., Takeda, N., Sakai, T., et al. (2002). SHEPHERD is the *Arabidopsis* GRP94 responsible for the formation of functional CLAVATA proteins. *EMBO J.* 21, 898–908. doi: 10.1093/emboj/21.5.898

Jain, M., Tyagi, A. K., and Khurana, J. P. (2006). Genome-wide analysis, evolutionary expansion, and expression of early auxin-responsive SAUR gene family in rice (*Oryza sativa*). *Genomics* 88, 360–371. doi: 10.1016/j.ygeno.2006.04.008

Johnson, J. L., and Brown, C. (2009). Plasticity of the Hsp90 chaperone machine in divergent eukaryotic organisms. *Cell Stress Chaperones* 14, 83–94. doi: 10.1007/s12192-008-0058-9

Kersey, P. J. (2019). Plant genome sequences: past, present, future. *Curr. Opin. Plant Biol.* 48, 1–8. doi: 10.1016/j.pbi.2018.11.001

Krishna, P., and Gloor, G. (2001). The Hsp90 family of proteins in *Arabidopsis thaliana*. *Cell Stress Chaperones* 6, 238–246. doi: 10.1379/1466-1268(2001)006<0238:THFOPI>2.0.CO;2

Krishna, P., Sacco, M., Cherutti, J. F., and Hill, S. (1995). Cold-induced accumulation of hsp90 transcripts in *Brassica napus*. *Plant Physiol.* 107, 915–923. doi: 10.1104/pp.107.3.915

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

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096

Lescot, M., Dehais, 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

Li, W., Chen, Y., Ye, M., Wang, D., and Chen, Q. (2020). Evolutionary history of the heat shock protein 90 (Hsp90) family of 43 plants and characterization of Hsp90s in *Solanum tuberosum*. *Mol. Biol. Rep.* 47, 6679–6691. doi: 10.1007/s11033-020-05722-x

Li, H., Liu, S., Yi, C., Wang, F., Zhou, J., Xia, X., et al. (2014). Hydrogen peroxide mediates abscisic acid-induced HSP70 accumulation and heat tolerance in grafted cucumber plants. *Plant Cell Environ.* 37, 2768–2780. doi: 10.1111/pce.12360

Li, Z., Zhang, Z., Yan, P., Huang, S., Fei, Z., and Lin, K. (2011). RNA-Seq improves annotation of protein-coding genes in the cucumber genome. *BMC Genomics* 12:540. doi: 10.1186/1471-2164-12-540

Ling, J., Jiang, W., Zhang, Y., Yu, H., Mao, Z., Gu, X., et al. (2011). Genome-wide analysis of WRKY gene family in *Cucumis sativus*. *BMC Genomics* 12:471. doi: 10.1186/1471-2164-12-471

Liu, Y., Wan, H., Yang, Y., Wei, Y., Li, Z., Ye, Q., et al. (2014). Genome-wide identification and analysis of heat shock protein 90 in tomato. *Hereditas* 36, 1043–1052. doi: 10.3724/SP.J.1005.2014.1043. (in Chinese)

Ma, Z. (2014). *Hsp90 mediates temperature regulation on the Arabidopsis circadian clock*. Bonn: Universitäts-und Landesbibliothek.

Muthusamy, S. K., Dalal, M., Chinnusamy, V., and Bansal, K. C. (2017). Genome-wide identification and analysis of biotic and abiotic stress regulation of small heat shock protein (HSP20) family genes in bread wheat. *J. Plant Physiol.* 211, 100–113. doi: 10.1016/j.jplph.2017.01.004

Oh, S. E., Yeung, C., Babaei-Rad, R., and Zhao, R. (2014). Cosuppression of the chloroplast localized molecular chaperone HSP90.5 impairs plant development and chloroplast biogenesis in *Arabidopsis*. *BMC Res. Notes* 7:643. doi: 10.1186/1756-0500-7-643

Pearl, L. H., and Prodromou, C. (2006). Structure and mechanism of the Hsp90 molecular chaperone machinery. *Annu. Rev. Biochem.* 75, 271–294. doi: 10.1146/annurev.biochem.75.103004.142738

Prassinos, C., Haralampidis, K., Milioni, D., Samakovli, D., Krambis, K., and Hatzopoulos, P. (2008). Complexity of Hsp90 in organelle targeting. *Plant Mol. Biol.* 67, 323–334. doi: 10.1007/s11103-008-9322-8

Reddy, R. K., Chaudhary, S., Patil, P., and Krishna, P. (1998). The 90 kDa heat shock protein (hsp90) is expressed throughout *Brassica napus* seed development and germination. *Plant Sci.* 131, 131–137. doi: 10.1016/S0168-9452(97)00254-9

Richter, K., and Buchner, J. (2001). Hsp90: chaperoning signal transduction. *J. Cell. Physiol.* 188, 281–290. doi: 10.1002/jcp.1131

Richter, K., Haslbeck, M., and Buchner, J. (2010). The heat shock response: life on the verge of death. *Mol. Cell* 40, 253–266. doi: 10.1016/j.molcel.2010.10.006

Rizzolo, K., Wong, P., Tillier, E. R. M., and Houry, W. A. (2014). "The interaction network of the Hsp90 molecular chaperone" in *The molecular chaperones interaction networks in protein folding and degradation. Interactomics and systems biology*. ed. W. A. Houry (New York: Springer), 111–131.

Roy, S. W., and Penny, D. (2007). On the incidence of intron loss and gain in paralogous gene families. *Mol. Biol. Evol.* 24, 1579–1581. doi: 10.1093/molbev/msm082

Ru, L., Jiang, L., Wills, R. B. H., Golding, J. B., Hou, Y., Yang, H., et al. (2020). Chitosan oligosaccharides induced chilling resistance in cucumber fruit and associated stimulation of antioxidant and HSP gene expression. *Sci. Hortic.* 264:109187. doi: 10.1016/j.scienta.2020.109187

Sable, A., Rai, K. M., Choudhary, A., Yadav, V. K., Agarwal, S. K., and Sawant, S. V. (2018). Inhibition of heat shock proteins HSP90 and HSP70 induce oxidative stress, suppressing cotton fiber development. *Sci. Rep.* 8:3620. doi: 10.1038/s41598-018-21866-0

Sangster, T. A., Bahrami, A., Wilczek, A., Watanabe, E., Schellenberg, K., McLellan, C., et al. (2007). Phenotypic diversity and altered environmental plasticity in *Arabidopsis thaliana* with reduced Hsp90 levels. *PLoS One* 2:e648. doi: 10.1371/journal.pone.0000648

Sangster, T. A., and Queitsch, C. (2005). The HSP90 chaperone complex, an emerging force in plant development and phenotypic plasticity. *Curr. Opin. Plant Biol.* 8, 86–92. doi: 10.1016/j.pbi.2004.11.012

Song, Z., Pan, F., Yang, C., Jia, H., Jiang, H., He, F., et al. (2019). Genome-wide identification and expression analysis of HSP90 gene family in *Nicotiana tabacum*. *BMC Genet.* 20:35. doi: 10.1186/s12863-019-0738-8

Song, H. M., Wang, H. Z., and Xu, X. B. (2012). Overexpression of *AtHsp90.3* in *Arabidopsis thaliana* impairs plant tolerance to heavy metal stress. *Biol. Plant.* 56, 197–199. doi: 10.1007/s10535-012-0042-1

Song, H., Zhao, R., Fan, P., Wang, X., Chen, X., and Li, Y. (2009). Overexpression of *AtHsp90.2*, *AtHsp90.5* and *AtHsp90.7* in *Arabidopsis thaliana* enhances plant sensitivity to salt and drought stresses. *Planta* 229, 955–964. doi: 10.1007/s00425-008-0886-y

UN Food and Agriculture Organization Corporate Statistical Database (FAOSTAT) (2017). Production of cucumbers and gherkins in 2016.

Wan, H., Yuan, W., Bo, K., Shen, J., Pang, X., and Chen, J. (2013). Genome-wide analysis of NBS-encoding disease resistance genes in *Cucumis sativus* and phylogenetic study of NBS-encoding genes in Cucurbitaceae crops. *BMC Genomics* 14:109. doi: 10.1186/1471-2164-14-109

Wang, L., Li, P., and Brutnell, T. P. (2010). Exploring plant transcriptomes using ultra high-throughput sequencing. *Brief. Funct. Genomics* 9, 118–128. doi: 10.1093/bfgp/elp057

Wang, J., Tan, F., Liang, C., Zhang, X., Ou, L., Niran, J., et al. (2020). Genome-wide identification and analysis of HSP90 gene family in pepper. *Acta Horticulturae Sinica* 47, 665–674, (in Chinese).

Wang, Y., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). *MCScanX*: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40:e49. doi: 10.1093/nar/gkr1293

Wang, W., Vinocur, B., Shoseyov, O., and Altman, A. (2004). Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response. *Trends Plant Sci.* 9, 244–252. doi: 10.1016/j.tplants.2004.03.006

Wang, G., Wei, X., Fan, R., Zhou, H., Wang, X., Yu, C., et al. (2011). Molecular analysis of common wheat genes encoding three types of cytosolic heat shock protein 90 (Hsp90): functional involvement of cytosolic Hsp90s in the control of wheat seedling growth and disease resistance. *New Phytol.* 191, 418–431. doi: 10.1111/j.1469-8137.2011.03715.x

Wang, A., Yu, X., Mao, Y., Liu, Y., Liu, G., Liu, Y., et al. (2015). Overexpression of a small heat-shock-protein gene enhances tolerance to abiotic stresses in rice. *Plant Breed.* 134, 384–393. doi: 10.1111/pbr.12289

Wang, R., Zhang, Y., Kieffer, M., Yu, H., Kepinski, S., and Estelle, M. (2016). HSP90 regulates temperature-dependent seedling growth in *Arabidopsis* by stabilizing the auxin co-receptor F-box protein TIR1. *Nat. Commun.* 7:10269. doi: 10.1038/ncomms10269

Wei, X., Wang, L., Yu, J., Zhang, Y., Li, D., and Zhang, X. (2015). Genome-wide identification and analysis of the MADS-box gene family in sesame. *Gene* 569, 66–76. doi: 10.1016/j.gene.2015.05.018

Xiang, J., Chen, X., Hu, W., Xiang, Y., Yan, M., and Wang, J. (2018). Overexpressing heat-shock protein OsHSP50.2 improves drought tolerance in rice. *Plant Cell Rep.* 37, 1585–1595. doi: 10.1007/s00299-018-2331-4

Xu, Q., Xu, X., Shi, Y., Qi, X., and Chen, X. (2017). Elucidation of the molecular responses of a cucumber segment substitution line carrying *Pm5.1* and its recurrent parent triggered by powdery mildew by comparative transcriptome profiling. *BMC Genomics* 18:21. doi: 10.1186/s12864-016-3438-z

Young, J. C., Moarefi, I., and Hartl, F. U. (2001). Hsp90: a specialized but essential protein-folding tool. *J. Cell Biol.* 154, 267–273. doi: 10.1083/jcb.200104079

Zhang, S., Chen, C., Li, L., Meng, L., Singh, J., Jiang, N., et al. (2005). Evolutionary expansion, gene structure, and expression of the rice wall-associated kinase gene family. *Plant Physiol.* 139, 1107–1124. doi: 10.1104/pp.105.069005

Zhang, X., Lai, Y., Zhang, W., Ahmad, J., Qiu, Y., Zhang, X., et al. (2018). MicroRNAs and their targets in cucumber shoot apices in response to temperature and photoperiod. *BMC Genomics* 19:819. doi: 10.1186/s12864-018-5204-x

Zhang, J., Li, J., Liu, B., Zhang, L., Chen, J., and Lu, M. (2013). Genome-wide analysis of the *Populus* Hsp90 gene family reveals differential expression patterns, localization, and heat stress responses. *BMC Genomics* 14:532. doi: 10.1186/1471-2164-14-532

Zhang, H., Li, L., Ye, T., Chen, R., Gao, X., and Xu, Z. (2016). Molecular characterization, expression pattern and function analysis of the OsHSP90 family in rice. *Biotechnol. Biotec. Eq.* 30, 669–676. doi: 10.1080/13102818.2016.1184588

Zhang, X., Millet, Y. A., Cheng, Z., Bush, J., and Ausubel, F. M. (2015). Jasmonate signalling in *Arabidopsis* involves SGT1b-HSP70-HSP90 chaperone complexes. *Nat. Plants* 1:15049. doi: 10.1038/nplants.2015.49

Zhang, Y., Zhao, G., Li, Y., Mo, N., Zhang, J., and Liang, Y. (2017). Transcriptomic analysis implies that GA regulates sex expression via ethylene-dependent and ethylene-independent pathways in cucumber (*Cucumis sativus* L.). *Front. Plant Sci.* 8:10. doi: 10.3389/fpls.2017.00010

Zhang, P., Zhu, Y., Shen, C., and Zhou, S. (2019). Proteome analysis of cucumber responses to *Pseudoperonospora cubensis* infection. *J. Plant Pathol.* 101, 917–925. doi: 10.1007/s42161-019-00290-x

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.

Copyright © 2021 Zhang, He, Sui, Gao, Jia, Lu and Jia. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 04 March 2021
doi: 10.3389/fgene.2021.590830

Comparative Transcriptome Analysis of Early- and Late-Bolting Traits in Chinese Cabbage (*Brassica rapa*)

Xiaochun Wei[1†], Md. Abdur Rahim[2†], Yanyan Zhao[1], Shuangjuan Yang[1], Zhiyong Wang[1], Henan Su[1], Lin Li[1], Liujing Niu[1], Md. Harun-Ur-Rashid[2], Yuxiang Yuan[1] and Xiaowei Zhang[1]**

[1] Institute of Horticulture, Henan Academy of Agricultural Sciences, Zhengzhou, China, [2] Department of Genetics and Plant Breeding, Sher-e-Bangla Agricultural University, Dhaka, Bangladesh

Chinese cabbage is one of the most important and widely consumed vegetables in China. The developmental transition from the vegetative to reproductive phase is a crucial process in the life cycle of flowering plants. In spring-sown Chinese cabbage, late bolting is desirable over early bolting. In this study, we analyzed double haploid (DH) lines of late bolting ("Y410-1" and "SY2004") heading Chinese cabbage (*Brassica rapa* var. *pekinensis*) and early-bolting Chinese cabbage ("CX14-1") (*B. rapa* ssp. *chinensis* var. *parachinensis*) by comparative transcriptome profiling using the Illumina RNA-seq platform. We assembled 721.49 million clean high-quality paired-end reads into 47,363 transcripts and 47,363 genes, including 3,144 novel unigenes. There were 12,932, 4,732, and 4,732 differentially expressed genes (DEGs) in pairwise comparisons of Y410-1 vs. CX14-1, SY2004 vs. CX14-1, and Y410-1 vs. SY2004, respectively. The RNA-seq results were confirmed by reverse transcription quantitative real-time PCR (RT-qPCR). A Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis of DEGs revealed significant enrichment for plant hormone and signal transduction as well as starch and sucrose metabolism pathways. Among DEGs related to plant hormone and signal transduction, six unigenes encoding the indole-3-acetic acid-induced protein ARG7 (BraA02g009130), auxin-responsive protein SAUR41 (BraA09g058230), serine/threonine-protein kinase BSK11 (BraA07g032960), auxin-induced protein 15A (BraA10g019860), and abscisic acid receptor PYR1 (BraA08g012630 and BraA01g009450), were upregulated in both late bolting Chinese cabbage lines (Y410-1 and SY2004) and were identified as putative candidates for the trait. These results improve our understanding of the molecular mechanisms underlying flowering in Chinese cabbage and provide a foundation for studies of this key trait in related species.

Keywords: comparative transcriptome, early bolting, late bolting, differentially expressed genes, Chinese cabbage

OPEN ACCESS

Edited by:
Zhiyong Liu,
Shenyang Agricultural University, China

Reviewed by:
Pengtao Ma,
Yantai University, China
Shengnan Huang,
Shenyang Agricultural University, China

***Correspondence:**
Yuxiang Yuan
yuxiangyuan126@126.com
Xiaowei Zhang
xiaowei5737@163.com

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *03 August 2020*
Accepted: *18 January 2021*
Published: *04 March 2021*

Citation:
Wei X, Rahim MA, Zhao Y, Yang S, Wang Z, Su H, Li L, Niu L, Harun-Ur-Rashid M, Yuan Y and Zhang X (2021) Comparative Transcriptome Analysis of Early- and Late-Bolting Traits in Chinese Cabbage (Brassica rapa). Front. Genet. 12:590830. doi: 10.3389/fgene.2021.590830

INTRODUCTION

Brassica belongs to the Brassicaceae family, which consists of several economically important vegetable crops consumed worldwide (Cheng et al., 2014). Vegetables in this genus exhibits various morphotypes, including leafy heads (cabbage, *Brassica oleracea* var. *capitata*; Chinese cabbage, *Brassica rapa* var. *pekinensis*), enlarged inflorescences (cauliflower, *B. oleracea* var. *botrytis*; broccoli,

B. oleracea var. *italica*), enlarged axillary buds (Brussels sprouts, *B. oleracea* var. *gemmifera*), enlarged stems (kohlrabi, *B. oleracea* var. *gongylodes*), and enlarged roots (turnip, *B. rapa* subsp. *rapa*) (Zhao et al., 2005; Cheng et al., 2014, 2016). Furthermore, these vegetables contain various health-promoting secondary metabolites, including glucosinolates, phenolics, carotenoids, flavonoids, and anthocyanins, which have protective effects against inflammation, cardiovascular diseases, and age-related diseases (Manchali et al., 2012).

Brassica rapa subspecies show high morphological diversity (Zhao et al., 2005). In particular, the heading Chinese cabbage (*B. rapa* L. subsp. *pekinensis*) is one of the most important *Brassica* vegetables in Asian countries, especially in China, Korea, and Japan (Bong et al., 2012; Cheng et al., 2014). It is a major vegetable crop produced in China, where it is considered as a key source of mineral nutrition (Wu et al., 2008). Moreover, Chinese cabbage is the main ingredient of the most popular traditional Korean side dish kimchi (Bong et al., 2012; Lee et al., 2014). The flowering Chinese cabbage (*B. rapa* ssp. *chinensis* var. *parachinensis*) produces elongated, tender, and thick stalks with rapid bolting as edible organs (Cheng et al., 2014; Huang et al., 2017). Pak choi (*B. rapa* subsp. *chinensis*) produces smooth dark green leaves with a prominent white midrib instead of forming a leafy head (Zhao et al., 2005).

In angiosperms, flowering is the most important developmental transition in the plant life cycle (Song et al., 2015). This transition is controlled by endogenous and environmental signals (Song et al., 2015; Zhang et al., 2015). More than 180 genes identified by functional analyses are associated with flowering time in *Arabidopsis* (Fornara et al., 2010). Several of these genes form a complex regulatory network involving six key pathways, including photoperiod, vernalization, ambient temperature, age, autonomy, and gibberellin pathways (Fornara et al., 2010; Zhang et al., 2015; Huang et al., 2017). However, photoperiod and vernalization related to day length and low temperatures, respectively, have also been identified as the major pathways for the regulation of flowering time (reviewed in Song et al., 2013). Several homologs of *Arabidopsis* genes related to flowering in *B. rapa* and other plant species have been identified (Andersen et al., 2004; Greenup et al., 2009; Cheng et al., 2011; Duan et al., 2015; Xu et al., 2015). Mutations in the *CONSTANS* (*CO*), *GIGANTEA* (*GI*), and *FLOWERING LOCUS T* (*FT*), related to the photoperiod pathway, result in delayed flowering, but short days do not affect flowering time, unlike wild-type *Arabidopsis* (Suárez-López et al., 2001; Simpson and Dean, 2002; Moon et al., 2003). The *CO* gene encodes a zinc finger transcription factor that triggers the transcriptional upregulation of downstream floral integrator genes, including *FT* and *SUPPRESSOR OF OVEREXPRESSION OF CO1* (*SOC1*), in leaves under long-day conditions (Samach et al., 2000; Srikanth and Schmid, 2011; Zhang et al., 2015). Furthermore, *FT* and *SOC1* act as activators of floral meristem identity genes, such as *LEAFY* (*LFY*) and *APETALA 1* (*AP1*) (Amasino, 2005). In *Arabidopsis*, *FRIGIDA* (*FRI*; encoding two coiled−coil motif-containing protein) and *FLOWERING LOCUS C* (*FLC*; encoding a MADS−box transcription factor) are involved in the annual winter habit (late flowering) (Michaels and Amasino, 1999; Johanson et al., 2000; Moon et al., 2003; Amasino, 2005). Mutations in these genes result in early flowering in *Arabidopsis* (Michaels and Amasino, 1999; Johanson et al., 2000; Choi et al., 2011). *FRI* positively regulates the transcription of the flowering repressor, *FLC* (Choi et al., 2011). In contrast, vernalization results in early flowering via suppressing the *FLC* (Michaels and Amasino, 1999; Sheldon et al., 1999). Three genes, *VIN3* (*VERNALIZATION INSENSITIVE 3*), *VRN2* (*VERNALIZATION 2*), and *VRN1* (*VERNALIZATION 1*), are required for the cold-mediated repression of *FLC* (Kobayashi et al., 1999; Gendall et al., 2001; Sung and Amasino, 2004). Su et al. (2018) reported that *BrVIN3.1* and *BrFLC1* are the important genetic determinants of bolting time variation in *B. rapa*. Recent reports have revealed that there are various orthologs of *FLC* and *FT* for flowering time variation in *B. rapa* and *B. oleracea* (Schiessl et al., 2017). In *B. rapa*, there are two copies of *FT* on chromosomes A02 and A07 and four copies of *FLC* on chromosomes A02, A03, and A10 (reviewed in Schiessl et al., 2017). Moreover, Shu et al. (2018) reported a major quantitative trait locus (QTL) (*Ef2.1*) for early flowering in "broccoli × cabbage" and identified *BolGRF6* as a putative candidate for early flowering in broccoli. *BrSDG8* code for encoding a histone methyltransferase is associated with bolting in *B. rapa* ssp. *pekinensis* (Fu et al., 2020). Huang et al. (2020) found that histone methyltransferase *CURLY LEAF* (*CLF*; Bra032169) controls the expression of flowering-related genes, and mutation in the Bra032169 caused early bolting in Chinese cabbage.

Gibberellic acid (GA), a plant hormone, plays an important role in the regulation of flowering time (Bernier, 1988). In *Arabidopsis*, the application of exogenous GA promotes flowering under short-day conditions (Langridge, 1957; Chandler and Dean, 1994). Mutations in genes related to GA biosynthesis and signaling can alter flowering time in *Arabidopsis* (Wilson et al., 1992; Sun and Kamiya, 1994; Jacobsen et al., 1996). In *Brassica napus*, flower initiation, flowering time, and shoot elongation are regulated by endogenous GA (Dahanayake and Galwey, 1999). Exogenous GA can stimulate flower development, while an inhibitor of GA delays or inhibits flowering in *Brassica* (Rood et al., 1989; Zanewich et al., 1990).

In flowering Chinese cabbage, aerial vegetative parts and floral buds are consumed. The early-bolting flowering Chinese cabbage double haploid (DH) line "CX14-1" is characterized by rapid flower stalk development and early flowering. In contrast, late-bolting heading Chinese cabbage exhibits a long period of vegetative growth before flower bud initiation. In *B. rapa*, long-day conditions together with vernalization promote reproductive growth over vegetative growth; consequently, plants bolt before reaching the harvesting stage, which is a serious problem in this crop (Zhang et al., 2015). Therefore, late-bolting traits are desirable than early flowering for the cultivation of spring-sown Chinese cabbage.

In this study, we performed transcriptome sequencing of flower bud samples from early-bolting flowering Chinese cabbage ("CX14-1") (*B. rapa* ssp. *chinensis* var. *parachinensis*) and late-bolting ("Y410-1" and "SY2004") heading Chinese cabbage (*B. rapa* var. *pekinensis*) DH lines using the Illumina platform, with a focus on genes related to plant hormone signaling.



FIGURE 1 | Phenotypes of early- and late-bolting Chinese cabbage double haploid lines used in the study.

MATERIALS AND METHODS

Plant Materials

The late-bolting DH heading Chinese cabbage (*B. rapa* L. ssp. *pekinensis*) lines "Y410-1" and "SY2004" and a vernalization-independent early-bolting flowering Chinese cabbage (*B. rapa* ssp. *chinensis* var. *parachinensis*) DH line "CX14-1" were used (**Figure 1**). The plants were grown in the experimental field at Yuanyang (113° 97' E and 35° 5' N), Henan Academy of Agricultural Sciences, China. The flower primordia of these lines were used for RNA sequencing. The samples were collected, immediately frozen in liquid nitrogen, and stored at -80°C.

Total RNA Isolation, Library Construction, RNA Sequencing, and Assembly

The frozen flower bud samples were ground into a powder in liquid nitrogen. Thereafter, total RNA was isolated from 100 mg of powder using the RNeasy Mini Kit (Qiagen, Valencia, CA, United States) according to the manufacturer's guidelines. The concentration, integrity, and purity of the RNA samples were determined using a NanoDrop spectrophotometer (NanoDrop Technologies, Wilmington, DE, United States) and Agilent 2100 Bio Analyzer (Agilent Technologies, Palo Alto, CA, United States). RNA samples with more than seven RNA integrity numbers (RINs) were used for RNA-seq library preparation with biological replications. Nine RNA-seq libraries were constructed using the NEBNext Ultra RNA Library Prep Kit for Illumina (New England Biolabs, Beverly, MA, United States) following the manufacturer's recommendations. The clustering of the index-coded samples was performed on a cBot Cluster Generation System using the TruSeq PE Cluster Kit v4-cBot-HS (Illumina, San Diego, CA, United States) according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on the Illumina high-throughput sequencing-by-synthesis technology platform to generate paired-end reads. The adapter sequences and low-quality reads were removed. The clean reads were then mapped to the *B. rapa* (Chiifu-401) v. 3.0 reference genome [*Brassica* database (BRAD)][1] using HISAT2[2].

Functional Annotation and Novel Gene Discovery

The functions of assembled genes and novel gene discovery were performed by BLAST searches against the NR [National Center for Biotechnology Information (NCBI) non-redundant protein sequences[3]], Swiss-Prot[4], gene ontology (GO)[5], Clusters of Orthologous Groups of proteins (COG)[6], Pfam[7], and Kyoto

[1]http://brassicadb.org/brad/
[2]http://ccb.jhu.edu/software/hisat2/index.shtml
[3]ftp://ftp.ncbi.nih.gov/blast/db/
[4]http://www.uniprot.org/
[5]http://www.geneontology.org/
[6]http://www.ncbi.nlm.nih.gov/COG/
[7]http://pfam.xfam.org/

Encyclopedia of Genes and Genomes (KEGG)[8] databases. After prediction of the amino acid sequences of the new genes, HMMER (Eddy, 1998) was used for comparisons with the Pfam database to obtain annotation information.

Quantification of Gene Expression and Identification of Differentially Expressed Genes

Gene expression levels were estimated by fragments per kilobase of transcript per million fragments mapped (FPKM). The FPKM value for each gene was quantified according to the length of the gene and read count mapped to this gene. Differentially expressed genes (DEGs) between early-bolting and late-bolting Chinese cabbage lines were identified using the R package DEGseq (Wang et al., 2010). The resulting *p*-values were adjusted using the Benjamini and Hochberg method to control the false discovery rate. Genes with an adjusted *p*-value < 0.01 found by DEseq were identified as DEGs.

Gene Ontology and Kyoto Encyclopedia of Genes and Genomes Pathway Enrichment Analysis

Gene ontology enrichment analysis of the DEGs was implemented by the GOseq R packages based on the Wallenius non-central hyper-geometric distribution (Young et al., 2010), which can adjust for gene length bias in DEGs. KEGG (Kanehisa et al., 2008) is a database resource for understanding high-level functions and utilities of the biological system from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies (see text foot note 8). We used KOBAS (Mao et al., 2005) software to test the statistical enrichment of differential expression genes in KEGG pathways. A GO enrichment analysis of DEGs was performed using the GOseq R package based on the Wallenius non-central hyper-geometric distribution (Young et al., 2010), which can adjust for gene length bias in DEGs.

cDNA Synthesis and qPCR Validation

A total of 1 μg of RNA was used for cDNA synthesis using SuperScript III following the manufacturer's protocol (Invitrogen, Gaithersburg, MD, United States). Then, the FPKM values for nine randomly selected genes were validated by reverse transcription quantitative real-time PCR (RT-qPCR) using the LightCycler 480II (Roche, Mannheim, Germany). A total of 45 ng of cDNA was used as a template for RT-qPCR with gene-specific primers (**Supplementary Table 1**) using 2 × SyGreen Mix (qPCRBIO Lo-ROX) (PCR Biosystems, London, United Kingdom). Thermocycling conditions were 95°C for 5 min, 45 cycles of 95°C for 10 s, 60°C for 10 s, and 72°C for 15 s. At the end of the PCR cycles, the Ct values were analyzed using LightCycler 480II software (Roche). The efficiency of each gene-specific primer was determined using pooled cDNA samples. The expression of each gene was normalized using the comparative $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001) with *BrActin* as a reference gene.

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

RESULTS

Transcriptome Sequencing of Early- and Late-Bolting Chinese Cabbage Floral Buds

We performed transcriptome analyses of two late-bolting heading Chinese cabbage (*B. rapa* L. ssp. *pekinensis*) DH lines (Y410-1 and SY2004) and one early-bolting flowering Chinese cabbage DH line (CX14-1) (*B. rapa* ssp. *chinensis* var. *parachinensis*) with three biological replications using Illumina sequencing technology. The raw reads generated by RNA-seq were deposited in the NCBI "Sequence Reads Archive" (SRA) under the accession number PRJNA605481. After removal of low-quality sequences, adapters, and ambiguous reads, a total of 721.49 million clean paired-end reads were obtained (**Table 1**). The clean reads for each sample totaled 8.81 Gb, and the Q30 base percentage was 92.71% or greater (**Table 1**). The clean reads were aligned with the *B. rapa* (Chiifu-401) v. 3.0 reference genome (BRAD, see text foot note 1), and the efficiency of the alignment ranged from 86.20 to 89.54%. Furthermore, variable splicing prediction, a gene structure analysis, and new gene discovery were performed based on the comparison. The clean reads were then assembled into 47,363 transcripts and 47,363 genes, of which 3,144 were predicted as novel genes (**Supplementary Table 2**).

Identification of Differentially Expressed Genes

A total of 30,375 DEGs (**Supplementary Tables 3–5**) were identified in pairwise comparisons (Y410-1 vs. SY2004, Y410-1 vs. CX14-1, and SY2004 vs. CX14-1) (**Figure 2A**). The most DEGs were found in Y410-1 vs. CX14-1 (12,932), with 7,913 upregulated and 5,019 downregulated genes (**Figure 2**). The fewest DEGs were found in Y410-1 vs. SY2004 (4732), with 2,925 and 1,807 up- and downregulated genes, respectively (**Figure 2B**). Overall, 1,445 DEGs were common to all comparisons. Volcano plots (**Figure 3**) and MA plots (**Figure 4**) were generated to summarize significant DEGs. The upregulated and downregulated genes in comparisons between each pair of early-bolting Chinese cabbage lines were determined by hierarchical clustering based on FPKM values (**Figure 5**).

Functional Annotation and Classification

The putative functions of assembled genes were annotated by searches against public databases, including GO, COG, KEGG Orthology (KOG), KEGG, eggNOG, PFAM, NR, and SWISS-PROT. Among them, 47,214, 40,497, 35, 083, 32,944, and 24,390 genes were annotated to the NR, eggNOG, PFAM, SWISS-PROT, and KOG databases, respectively (**Table 2**). In addition, 2,564 of 3,144 novel genes were functionally annotated (**Supplementary Table 2**). Furthermore, we performed GO, COG, and KEGG pathway analyses to illustrate the biological functions of the Chinese cabbage floral bud transcriptomes.

TABLE 1 | Overview of the Chinese cabbage transcriptome sequencing and assembly.

Samples	Y410-1			SY2004			CX14-1		
	T1	T2	T3	T4	T5	T6	T7	T8	T9
Clean reads	84,649,952	110,937,646	86,119,332	61,878,880	58,846,612	62,047,916	82,928,554	89,056,338	85,030,426
Clean bases	12.62	16.56	12.85	9.23	8.81	9.27	12.39	13.26	12.67
GC content (%)	47.56%	47.65%	47.75%	47.56%	47.47%	47.60%	47.39%	47.21%	47.25%
Q30 (%)	94.57%	94.75%	94.41%	92.76%	92.81%	92.71%	95.22%	94.56%	94.75%
Mapped reads	73,679,588 (87.04%)	97,117,048 (87.54%)	75,349,891 (87.49%)	54,591,125 (88.22%)	52,692,240 (89.54%)	55,186,834 (88.94%)	72,209,297 (87.07%)	76,873,854 (86.32%)	73,294,494 (86.20%)
Uniquely mapped reads	72,019,424 (85.08%)	94,886,164 (85.53%)	73,661,268 (85.53%)	53,399,380 (86.30%)	51,512,996 (87.54%)	53,981,346 (87.00%)	70,403,562 (84.90%)	74,887,781 (84.09%)	71,481,831 (84.07%)
Multiple map reads	1,660,164 (1.96%)	2,230,884 (2.01%)	1,688,623 (1.96%)	1,191,745 (1.93%)	1,179,244 (2.00%)	1,205,488 (1.94%)	1,805,735 (2.18%)	1,986,073 (2.23%)	1,812,663 (2.13%)
Reads map to "+"	36,562,440 (43.19%)	48,204,505 (43.45%)	37,387,481 (43.41%)	27,067,998 (43.74%)	26,126,686 (44.40%)	27,361,985 (44.10%)	35,865,686 (43.25%)	38,167,001 (42.86%)	36,400,366 (42.81%)
Reads map to "-"	36,694,556 (43.35%)	48,376,741 (43.61%)	37,538,457 (43.59%)	27,220,998 (43.99%)	26,275,023 (44.65%)	27,524,467 (44.36%)	35,975,214 (43.38%)	38,287,253 (42.99%)	36,504,731 (42.93%)

Q30 is the percentage of bases within a Phred value > 30.

A GO term enrichment analysis was performed to identify terms in three general categories, biological process (BP), molecular function (MF), and cellular component (CC) (Berardini et al., 2004) (**Figure 6A**). A total of 37,394 genes were assigned to 54 main functional groups. The cellular process (GO:0009987), metabolic process (GO:0008152), and single-organism process (GO:0044699) were the most important in the BP category. Cell (GO:0005623), cell part (GO:0044464), and organelle (GO:0043226) were highly enriched in CC. Binding (GO:0005488) and catalytic activity (GO:0003824) were the most important GO terms in MF (**Supplementary Table 6–8** and **Figure 6A**). These results indicated that early and late bolting might be associated with DEGs in these functional subgroups. According to a COG functional annotation analysis, 14,411 genes were classified into 25 COG categories (**Figure 6B**). The predominant COG categories represented in the Chinese cabbage floral bud transcriptomes were G (carbohydrate transport and metabolism), R (general function prediction only), and T (signal transduction mechanisms) (**Figure 7B**). Several genes in these categories were differentially expressed, which might contribute to flowering time differences between the early- and late-bolting Chinese cabbage DH lines.

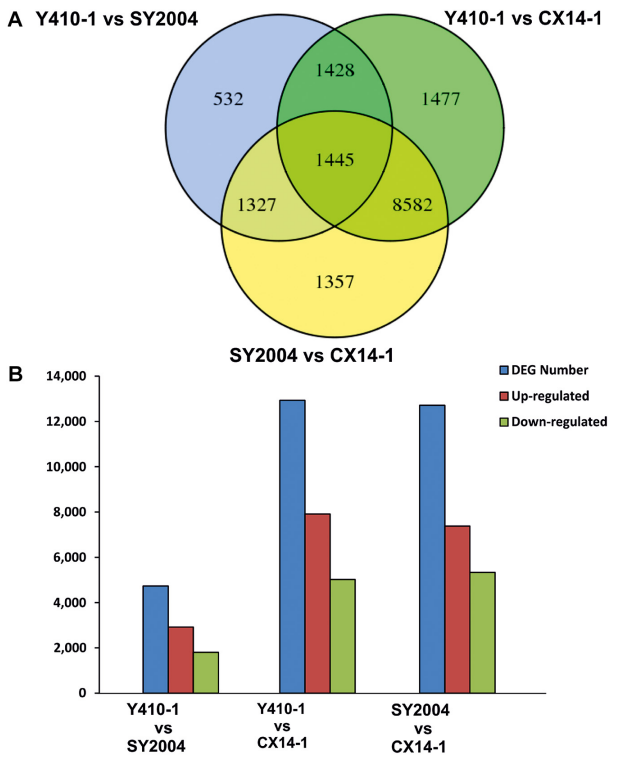


FIGURE 2 | Differentially expressed genes (DEGs) between early- and late-bolting Chinese cabbage double haploid lines. **(A)** Venn diagram DEGs identified through pairwise comparisons. **(B)** Number of up- and downregulated genes in each comparison. Venn diagram was generated using the freely available VENNY 2.1 online tool (http://bioinfogp.cnb.csic.es/tools/venny/).

Kyoto Encyclopedia of Genes and Genomes Pathway Enrichment Analysis of the Differentially Expressed Genes

Different gene products interact with each other to exert biological functions, and pathway annotation analyses are therefore useful to predict the functions of gene. KEGG is a resource for the systematic analysis of gene function and genomic information databases, providing information about genes and expression patterns from a whole network perspective (Kanehisa et al., 2004). We therefore analyzed the metabolic pathways of the DEGs between early- and late-bolting Chinese cabbage lines using the KEGG database. Total 2,430 and 2,417 DEGs were assigned to 124, 128, and 127 pathways for Y410-1 vs. SY2004, Y410-1 vs. CX14-1, and SY2004 vs. CX14-1, respectively (**Supplementary Table 9**). Moreover, these pathways involved 15,408 genes, which were different from the DEGs

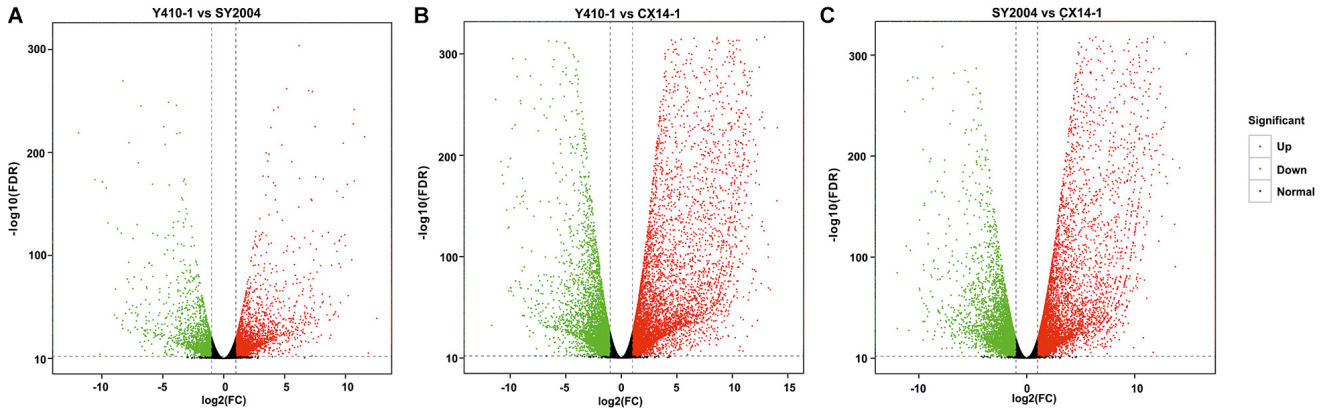


FIGURE 3 | Volcano plots of differentially expressed genes (DEGs) for each comparison in flower bud of early- and late-bolting Chinese cabbage double haploid lines. The *x* and *y* axes indicate logarithm fold change [$\log_2$(FC)] of the difference in the expression of a gene between two samples and the negative logarithm of the statistical significance [$\log_2$(FDR)] of the change in gene expression, respectively. The red and green dots represent significantly up- and downregulated genes, respectively; while black dots represent non-differentiated genes. FC, fold change; FDR, false discovery rate.

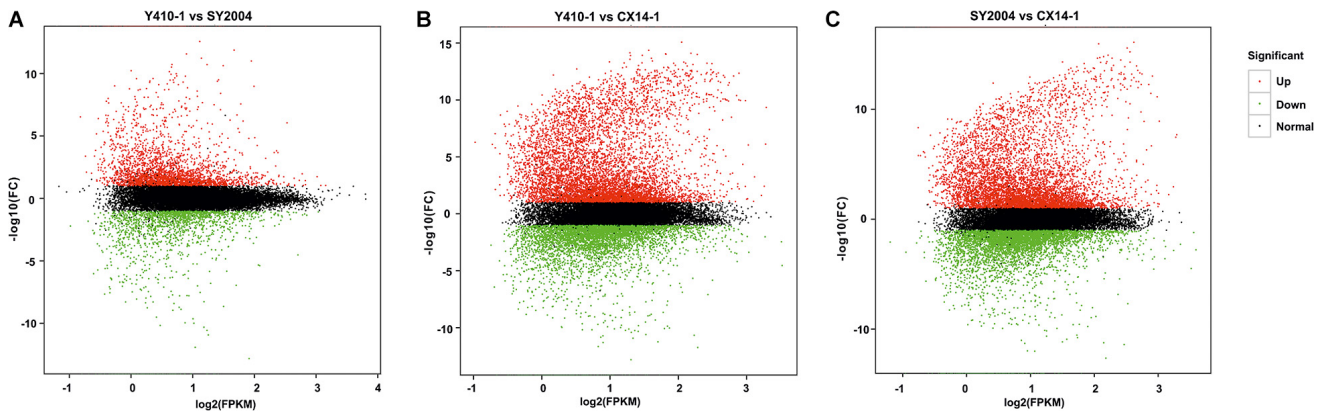


FIGURE 4 | MA plot of differential expression for two samples. Each point represents the differentially expressed genes (DEGs). The *x* axis indicates $\log_2$(FPKM), and the *y* axis indicates $\log_2$(FC). The red and green dots represent significantly up- and downregulated genes, respectively; while black dots represent non-differentiated genes. FPKM, fragments per kilobase of transcript per million fragments mapped; FC, fold change.

assigned to pathways, indicating that some genes contribute to more than one KEGG pathway (**Table 2** and **Supplementary Table 9**). For example, the novel gene predicted in this study "Brassica_rapa_newGene_2848" is involved in homologous recombination, mismatch repair, nucleotide excision repair, and DNA replication. However, six pathways, including plant–pathogen interaction, starch and sucrose metabolism, carbon metabolism, biosynthesis of amino acids, phenylpropanoid biosynthesis, and plant hormone signal transduction, contained over 100 DEGs between early- and late-bolting lines (SY410-1 vs. CX14-1 and SY2004 vs. CX14-1) (**Figure 6C**). Further, a KEGG pathway enrichment analysis of DEGs between early-and late-bolting lines revealed that plant hormone and signal transduction (197 genes) (Ko04075), starch and sucrose metabolism (158 genes) (Ko00500), and phenylalanine biosynthesis (117 genes) (Ko00940) are highly enriched for Y410-1 vs. CX14-1, while carbon metabolism (177 genes) (Ko01200), plant hormone and signal transduction (191 genes) (Ko04075), starch and sucrose metabolism (151 genes) (Ko00500), and phenylalanine biosynthesis (117 genes) (Ko00940) are highly enriched for SY2004 vs. CX14-1 (**Figure 7** and **Supplementary Table 8**). Plant hormone and signal transduction as well as starch and sucrose metabolism were common to both comparisons. Among genes related to plant hormone and signal transduction, 98 and 99 genes were up- and downregulated in Y410-1 vs. CX14-1 (**Supplementary Table 10**), and 89 and 102 genes were up- and downregulated in SY2004 vs. CX14-1 (**Supplementary Table 10**).

Expression Patterns of Differentially Expressed Genes Related to Plant Hormone and Signal Transduction

The expression levels of most of the DEGs related to ABA receptors were downregulated, except *PYL8* and *PYL9*, while most of the DEGs encoding ABA-insensitive five-like protein

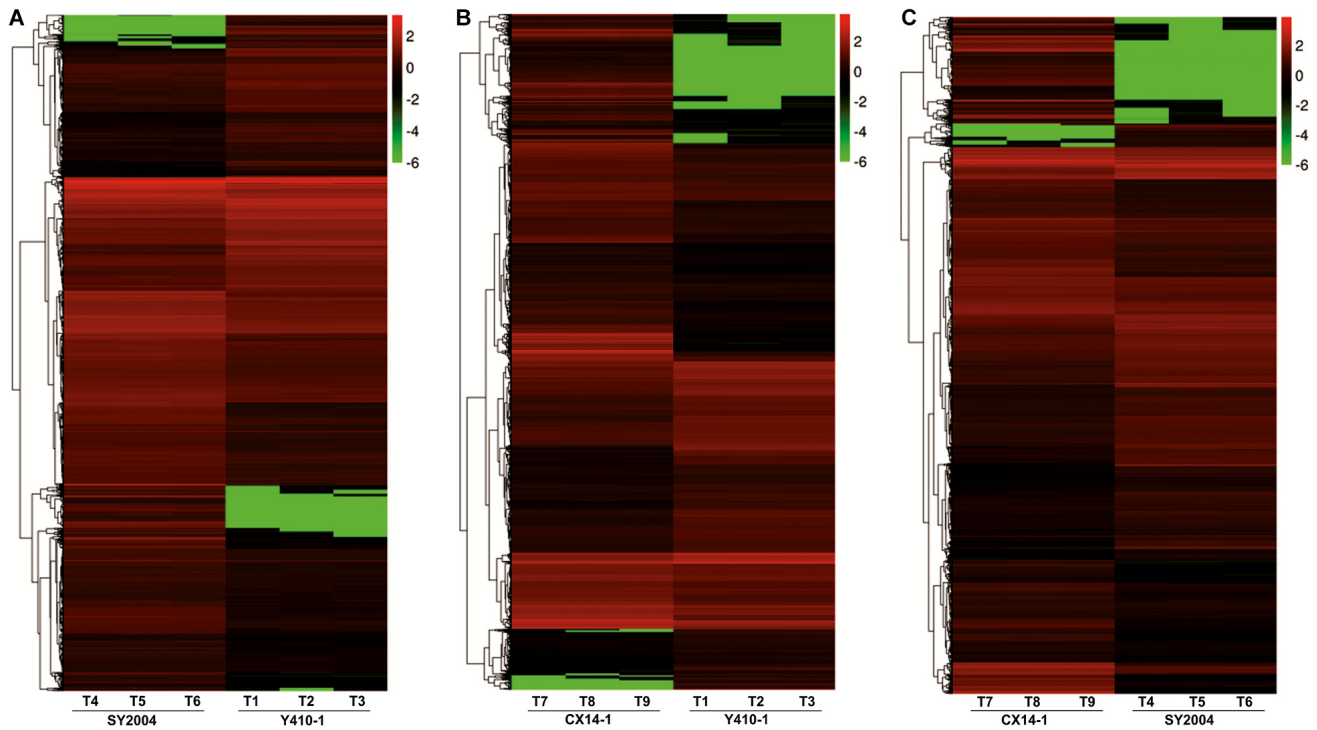


FIGURE 5 | Heatmap representation of differentially expressed genes (DEGs) in the floral bud of early- and late-bolting Chinese cabbage double haploid lines. The expression patterns are based on the $\log_2$ of FPKM (fragments per kilobase of transcript per million fragments mapped for each gene) used for hierarchical clustering for each sample. The different columns represent different samples, with different rows representing different genes. Biological replicates: T1, T2, and T3 (Y410-1); T4, T5, and T6 (SY2004); and T7, T8, and T9 (CX14-1).

were upregulated in late-bolting Y410-1 compared with early-bolting flowering Chinese cabbage DH lines (**Figure 8** and **Supplementary Table 9**). Similarly, the expression levels of most of the DEGs related to ABA receptors were downregulated, except for *PYL8* and *PYL9*. In contrast, most of the DEGs related to ABA-insensitive 5-like proteins were upregulated in late-bolting compared with early-bolting flowering Chinese cabbage DH lines (**Figure 8** and **Supplementary Table 10**).

TABLE 2 | Summary of functional annotation and classification of assembled genes.

Database annotated	No. of genes	Length (300 ≤ 1,000)	Length (≥1,000)
COG	14,411	4,341	9,839
GO	37,394	15,993	19,336
KEGG	15,408	6,204	8,422
KOG	24,390	9,639	13,624
Pfam	35,083	14,009	20,085
Swiss-Prot	32,944	13,405	18,021
eggnog	40,497	17,383	20,916
NR	47,214	21,202	22,487
All	47,363	21,268	22,505

COG, Clusters of Orthologous Groups of proteins; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; KOG, KEGG Orthology.

The expression levels of DEGs encoding gibberellin receptors GID1A and GID1B were upregulated in both late-bolting (Y410-1 and SY2004) heading Chinese cabbage compared with the early-bolting flowering Chinese cabbage DH line (**Figure 8** and **Supplementary Table 10**).

The expression levels of JA signaling genes, like *jasmonate-amino synthetase* (*JAR1*, a GH3 family of protein), were downregulated in both late-bolting heading Chinese cabbage DH lines (Y410-1 and SY2004) compared with the early-bolting flowering Chinese cabbage DH line (**Figure 8** and **Supplementary Table 10**). Moreover, the expression levels of DEGs encoding TIFY proteins were upregulated in the early-bolting flowering Chinese cabbage DH line (CX14-1) compared with both late-bolting heading Chinese cabbage lines (**Figure 8** and **Supplementary Table 10**).

The expression levels of DEGs encoding brassinazole-resistant 2 proteins (*BraA09g056680*, *BraA06g014960*, and *BraA08g028240*) and brassinosteroid-insensitive 1-associated receptor kinase 1 (*BraA08g016610*) were upregulated, whereas BRI1 kinase inhibitor 1 (*BraA02g030240*) was downregulated in both late-bolting heading Chinese cabbage lines compared with early-bolting flowering Chinese cabbage DH lines (**Figure 8** and **Supplementary Table 9**). Furthermore, two DEGs related to DELLA protein RGA1 (*BraA06g040430*) and RGA2 (*BraA09g023210*) were upregulated in both late-bolting heading Chinese cabbage lines.

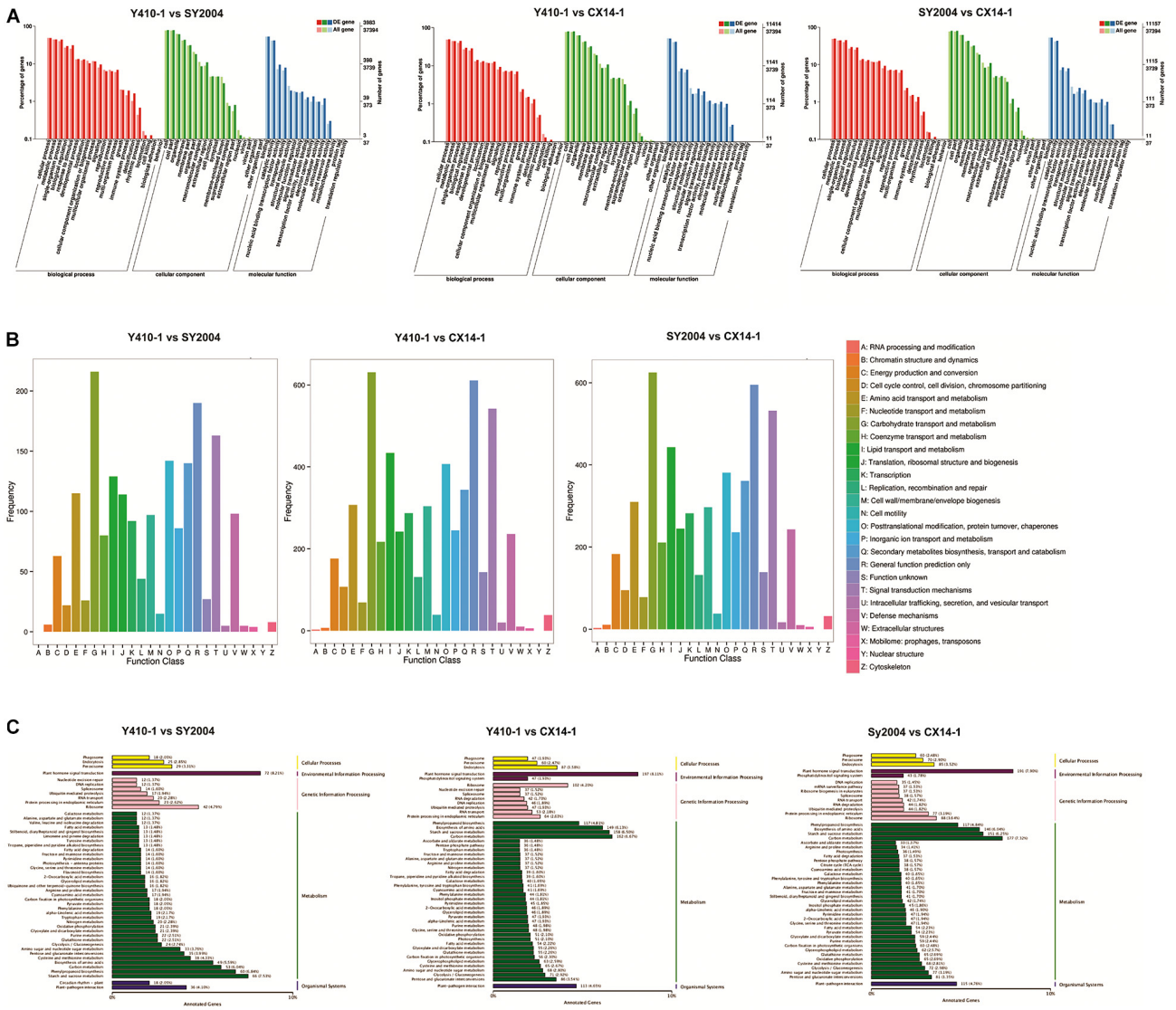


FIGURE 6 | Functional classification of differentially expressed genes (DEGs) among early- and late-bolting Chinese cabbage double haploid lines. **(A)** Gene ontology (GO), **(B)** Clusters of Orthologous Groups of proteins (COG), and **(C)** Kyoto Encyclopedia of Genes and Genomes (KEGG) classification of DEGs.

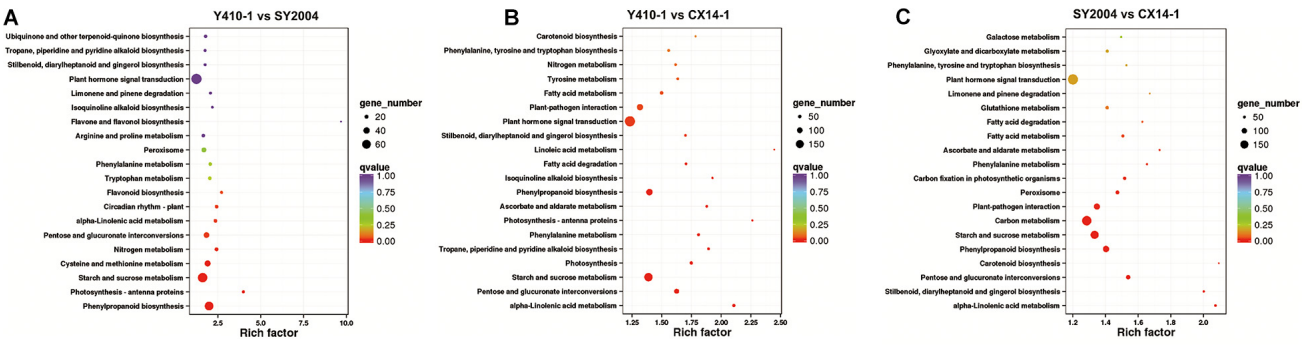


FIGURE 7 | Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment scatter plot of differentially expressed genes. Rich factor is the ratio of the differentially expressed gene number to the total gene number in a certain pathway. The dot size and color indicate the number of genes and the range of the false discovery rate (FDR) value, respectively.

A Down-regulating DEGs in late-bolting heading Chinese cabbage DH lines

#GeneID	Description (Swiss-Prot)	Late-bolting						Earl-bolting		
		Y410-1			SY2004			CX14-1		
		T1	T2	T3	T4	T5	T6	T7	T8	T9
BraA01g000880	Serine/threonine-protein kinase SRK2I	1.251	0.817	0.565	2.493	2.646	2.278	36.654	39.505	37.912
BraA01g003550	Auxin-responsive protein SAUR50	6.468	4.940	4.720	5.227	4.943	6.194	26.726	28.605	19.424
BraA01g004270	Serine/threonine-protein kinase SRK2E	0.114	0.119	0.107	3.429	2.013	1.586	3.279	3.719	3.093
BraA01g008880	Auxin-responsive protein IAA11	3.675	3.951	3.076	3.751	4.683	3.753	12.017	12.651	12.915
BraA01g009450	Abscisic acid receptor PYR1	0.049	0.090	0.000	0.124	0.049	0.065	2.899	2.178	2.625
BraA01g013230	Auxin-responsive protein SAUR36	3.003	3.019	2.970	5.538	4.290	3.501	28.976	31.826	27.492
BraA01g034810	Protein TIFY 6B	5.743	5.903	6.063	10.498	10.857	11.591	29.996	27.946	28.470
BraA01g036620	Auxin-responsive protein IAA19	3.533	2.556	3.335	1.852	1.775	1.902	38.038	34.046	32.533
BraA02g004740	Protein TIFY 9	0.154	0.106	0.430	0.964	1.538	1.147	12.948	18.402	14.882
BraA02g004820	Indole-3-acetic acid-amido synthetase GH3.17	0.071	0.124	0.149	0.021	0.088	0.094	1.939	2.148	1.932
BraA02g009130	Indole-3-acetic acid-induced protein ARG7	0.000	0.049	0.011	0.029	0.003	0.043	11.810	12.624	9.634
BraA02g011810	Two-component response regulator ARR18	0.012	0.232	0.130	0.041	0.093	0.219	0.841	0.814	1.152
BraA02g014480	Auxin-responsive protein SAUR32	1.624	1.778	1.157	5.337	3.155	5.245	7.433	8.849	6.856
BraA02g020000	Protein TIFY 7	6.382	5.989	5.566	20.231	21.946	16.605	142.491	143.596	151.010
BraA02g021020	Protein TIFY 11B	17.025	16.664	17.917	23.389	24.465	23.364	43.812	45.564	47.714
BraA02g024300	Auxin transporter-like protein 3	5.332	5.349	4.742	3.553	3.020	3.059	8.972	11.026	10.753
BraA02g030240	BRI1 kinase inhibitor 1	1.698	1.373	2.728	0.762	1.160	1.764	5.191	5.372	4.394
BraA03g005650	Protein TIFY 9	1.704	2.109	1.580	1.844	2.711	1.208	28.067	25.587	28.200
BraA03g008300	Auxin-responsive protein SAUR21	0.914	0.576	1.393	0.525	0.622	0.182	4.845	5.824	4.074
BraA03g009810	Indole-3-acetic acid-induced protein ARG7	0.000	0.000	0.000	0.031	0.000	0.000	8.341	9.963	7.727
BraA03g020010	Abscisic acid receptor PYL4	9.959	11.184	10.566	6.031	3.880	5.347	27.573	25.107	28.956
BraA03g021430	Abscisic acid receptor PYL6	1.702	0.960	1.687	1.287	1.803	1.410	9.603	9.357	9.297
BraA03g021580	Two-component response regulator ARR16	2.048	0.989	1.512	2.886	3.628	2.517	15.861	15.179	17.358
BraA03g027900	Auxin-responsive protein SAUR36	0.000	0.000	0.000	0.000	0.000	0.000	27.939	30.427	29.741
BraA03g028670	GRR1-like protein 1	25.808	26.215	25.887	19.882	20.143	19.079	46.627	44.123	45.717
BraA03g036990	Auxin-responsive protein IAA19	8.631	8.862	11.936	3.638	4.659	3.641	45.786	50.115	47.277
BraA03g043530	Probable inactive receptor-like kinase BSK12	0.000	0.000	0.000	0.000	0.000	0.000	1.189	1.159	1.223
BraA03g047820	Abscisic acid receptor PYR1	47.433	42.106	46.642	46.300	49.599	42.861	119.131	112.876	119.553
BraA03g053600	Indole-3-acetic acid-amido synthetase GH3.5	0.919	1.026	1.134	2.455	2.484	3.275	5.633	5.881	5.347
BraA03g058530	ABSCISIC ACID-INSENSITIVE 5-like protein 6	2.688	2.348	2.958	0.000	0.000	0.000	4.496	6.344	5.881
BraA04g001580	Auxin-responsive protein SAUR36	1.911	2.723	2.056	2.761	3.594	3.013	5.921	6.346	5.935
BraA04g003480	Regulatory protein NPR6	0.181	0.467	0.411	0.173	0.223	0.163	1.438	1.603	1.128
BraA04g008290	Auxin-responsive protein SAUR15	0.000	0.000	0.000	0.030	0.000	0.573	42.133	39.443	38.716
BraA04g025810	Protein ABSCISIC ACID-INSENSITIVE 5	0.198	0.071	0.160	0.273	0.239	0.436	9.315	9.083	10.151
BraA04g032040	Jasmonic acid-amido synthetase JAR1	0.736	0.571	0.527	0.869	0.917	0.842	2.409	2.370	2.679
BraA05g005770	Abscisic acid receptor PYL6	2.476	4.167	1.950	3.141	4.106	4.806	13.098	13.869	14.675
BraA05g007500	Auxin transporter protein 1	34.950	33.231	35.162	66.922	60.690	60.859	136.666	125.715	133.807
BraA05g029670	Protein TIFY 6B	2.808	2.457	2.735	10.109	8.981	9.641	24.826	23.297	24.989
BraA05g031540	Auxin-responsive protein IAA19	15.070	13.580	12.953	12.650	15.103	13.376	76.891	77.942	72.812
BraA05g035520	Protein ph phutase 2C 37	33.993	34.934	36.364	36.879	38.327	39.191	75.216	71.638	73.144
BraA06g001750	Auxin-responsive protein IAA6	0.269	0.473	0.324	0.036	0.000	0.106	7.561	6.538	6.167
BraA06g005530	Transcription factor TGA9	0.030	0.077	0.093	1.680	1.369	1.564	4.200	4.818	4.746
BraA06g007450	4-substituted benzoates-glutamate ligase GH3.12	0.000	0.000	0.000	0.000	0.000	0.000	11.629	14.475	13.244
BraA06g007460	4-substituted benzoates-glutamate ligase GH3.12	1.718	1.431	1.681	2.599	2.814	1.908	7.474	8.749	8.728
BraA06g012490	Auxin-responsive protein SAUR41	3.131	3.547	3.293	8.955	6.345	8.526	13.117	13.815	14.527
BraA06g022270	Indole-3-acetic acid-induced protein ARG7	1.998	1.459	2.112	1.457	1.576	1.354	35.396	39.363	30.994
BraA07g002570	Indole-3-acetic acid-amido synthetase GH3.17	0.162	0.237	0.370	0.214	0.277	0.246	0.567	0.966	1.328
BraA07g002710	Auxin-responsive protein SAUR50	0.000	0.043	0.000	0.554	0.537	1.553	3.600	3.985	4.098
BraA07g011840	Indole-3-acetic acid-amido synthetase GH3.17	0.201	0.310	0.411	0.576	0.611	0.443	7.477	9.300	7.382
BraA07g016220	Auxin-responsive protein SAUR50	1.298	1.261	1.156	2.436	4.169	3.229	12.111	13.581	16.787
BraA07g022670	Two-component response regulator ARR9	2.248	1.859	2.205	1.997	3.060	3.290	11.017	11.186	10.912
BraA07g024750	Abscisic acid receptor PYL9	0.066	0.000	0.048	0.404	0.088	0.379	5.333	4.246	3.271
BraA07g027690	Auxin-responsive protein SAUR50	0.557	0.238	0.877	2.620	3.253	2.988	5.039	6.320	4.765
BraA07g032960	Serine/threonine-protein kinase BSK11	0.032	0.083	0.030	0.211	0.182	0.146	6.473	6.839	6.562
BraA07g038660	Protein TIFY 10B	29.840	29.015	29.815	38.852	36.407	32.973	61.298	60.373	63.571
BraA08g010620	Auxin-responsive protein IAA14	1.300	1.360	1.067	0.000	0.000	0.000	6.587	6.531	6.542
BraA08g012630	Abscisic acid receptor PYR1	0.943	0.163	0.175	0.264	1.814	0.891	45.476	49.108	55.843
BraA08g029290	Protein TIFY 11A	13.363	11.542	13.327	26.691	26.705	26.289	109.713	112.715	108.944
BraA08g034940	Ethylene response sensor 2	6.291	5.068	5.554	2.779	3.246	3.188	16.004	14.754	15.221
BraA09g001190	GRR1-like protein 1	0.000	0.103	0.054	0.000	0.000	0.000	2.021	1.796	1.951
BraA09g010220	Probable indole-3-acetic acid-amido synthetase GH3.1	0.211	0.180	0.293	0.800	0.618	0.716	15.547	13.765	15.128
BraA09g019950	Auxin-responsive protein IAA4	29.799	24.946	27.532	23.825	29.324	19.797	46.200	46.908	46.784
BraA09g022370	Transcription factor bHLH28	0.065	0.326	0.092	0.451	0.428	0.206	1.942	1.465	1.947
BraA09g026350	Auxin-responsive protein SAUR36	0.786	1.404	1.876	1.569	2.094	2.432	25.500	26.664	31.083
BraA09g047720	Two-component response regulator ARR9	1.075	0.952	0.755	0.521	0.210	0.524	4.215	4.408	4.198
BraA09g050440	Auxin-responsive protein SAUR36	0.886	0.585	0.502	0.602	0.805	0.076	2.099	2.150	2.135
BraA09g055720	Auxin-responsive protein SAUR50	0.560	1.129	1.898	0.527	0.848	1.063	17.311	19.376	17.963
BraA09g056830	Protein TIFY 10A	0.682	0.906	0.454	14.068	14.530	12.963	44.827	43.710	48.438
BraA09g058230	Auxin-responsive protein SAUR41	0.000	0.083	0.285	0.139	0.166	0.515	52.157	50.939	53.365
BraA09g062830	Transcription factor TGA9	0.000	0.000	0.000	0.141	0.298	0.151	1.679	2.077	2.359
BraA09g063330	Protein ph phutase 2C 3	0.182	0.049	0.000	0.242	0.176	0.306	3.475	3.015	4.242
BraA10g002610	Auxin-responsive protein IAA10	2.337	2.316	2.860	3.510	3.242	2.654	8.296	9.679	9.806
BraA10g002790	Auxin-responsive protein IAA17	6.829	7.025	5.928	4.682	5.524	4.082	22.961	25.171	23.995
BraA10g002840	Ethylene response sensor 2	1.780	2.345	1.967	2.946	2.571	2.119	52.362	52.069	53.536
BraA10g016990	Probable protein ph phutase 2C 78	1.167	1.553	1.580	9.419	7.631	7.407	19.973	20.867	21.595
BraA10g019860	Auxin-induced protein 15A	0.172	0.499	0.213	0.251	0.440	0.224	33.545	35.104	24.368
BraA10g025420	Indole-3-acetic acid-amido synthetase GH3.17	0.000	0.000	0.000	0.000	0.000	0.000	12.336	12.656	10.708
BraA10g025540	Protein TIFY 9	3.153	3.305	3.316	6.803	6.505	5.005	113.757	113.060	116.578
BraA10g028350	Abscisic acid receptor PYL5	4.218	4.424	4.683	9.498	7.314	6.656	14.542	15.593	14.553
Brassica_rapa_newGene_4124	Putative two-component response regulator-like APRR6	0.000	0.000	0.000	0.000	0.000	0.000	2.246	2.785	3.273

B Up-regulating DEGs in late-bolting heading Chinese cabbage DH lines

#GeneID	Description (Swiss-Prot)	Late-bolting						Early-bolting		
		Y410-1			SY2004			CX14-1		
		T1	T2	T3	T4	T5	T6	T7	T8	T9
BraA01g004130	Cyclin-D3-1	18.126	13.514	15.589	13.993	14.627	16.311	6.459	5.287	5.317
BraA01g014680	Auxin response factor 9	6.539	5.781	5.953	7.070	5.427	6.279	0.662	0.635	0.512
BraA01g016840	Regulatory protein NPR1	17.784	18.295	19.000	15.986	15.435	17.098	3.692	3.210	3.760
BraA01g016870	Regulatory protein NPR2	13.587	13.237	12.721	14.427	12.499	12.785	1.704	1.373	1.751
BraA01g023270	Cyclin-D3-3	38.007	35.572	34.668	39.430	37.187	38.106	6.625	8.107	7.187
BraA01g030990	Ethylene receptor 2	25.759	26.761	35.446	14.055	12.595	14.013	4.360	4.293	5.049
BraA01g032170	Histidine-containing ph photransfer protein 1	15.946	16.221	15.958	19.607	19.407	16.381	2.193	3.872	2.468
BraA02g003280	Transcription factor TGA4	7.236	6.747	6.073	7.043	7.632	7.727	1.328	1.890	1.003
BraA02g013860	Indole-3-acetic acid-amido synthetase GH3.6	29.132	29.359	27.958	21.980	22.348	23.036	6.750	5.840	6.247
BraA02g033740	Ethylene-responsive transcription factor 2	65.731	73.027	71.087	31.575	33.802	27.967	6.868	7.486	8.140
BraA03g017680	Auxin-responsive protein IAA13	10.228	9.716	8.418	11.095	9.645	10.673	2.678	2.697	2.178
BraA03g019900	Auxin transporter protein 1	3.462	3.506	3.674	5.404	4.136	5.188	0.167	0.144	0.025
BraA03g021190	Auxin transporter-like protein 2	6.380	5.809	6.625	8.264	7.435	8.926	1.624	1.510	1.307
BraA03g025290	ETHYLENE INSENSITIVE 3-like 1 protein	29.125	29.449	28.720	23.972	23.633	24.416	6.531	7.501	7.427
BraA03g025870	Auxin transporter-like protein 2	13.173	12.692	12.836	23.217	19.474	25.370	3.761	3.636	3.713
BraA03g038820	ABSCISIC ACID-INSENSITIVE 5-like protein 7	2.668	1.897	2.337	1.838	1.604	1.820	0.280	0.334	0.430
BraA03g042590	Pathogenesis-related protein 1	61.470	41.122	39.868	23.027	28.342	27.360	0.000	0.000	0.000
BraA03g045010	Cyclin-D3-3	3.816	3.936	3.255	5.369	4.932	4.704	1.433	1.696	1.393
BraA03g046670	Two-component response regulator ARR2	2.280	1.750	1.625	2.392	2.438	2.923	0.669	0.513	0.766
BraA03g057090	Two-component response regulator ARR10	18.027	17.582	16.923	15.884	16.820	18.474	5.824	6.029	6.085
BraA03g058630	Cyclin-D3-1	30.907	26.632	27.185	19.442	18.377	18.136	3.915	3.141	3.127
BraA03g059070	Auxin-responsive protein SAUR20	4.703	4.614	3.959	5.856	6.681	5.275	0.756	0.797	0.679
BraA04g000380	Gibberellin receptor GID1B	1.446	1.178	1.054	1.381	1.347	1.233	0.364	0.379	0.121
BraA05g002190	Regulatory protein NPR5	4.600	5.416	5.059	2.932	3.580	3.391	0.710	0.960	1.053
BraA05g013460	Ethylene-responsive transcription factor 15	16.248	18.163	17.296	7.492	9.350	8.014	0.470	0.221	0.203
BraA05g027980	ABSCISIC ACID-INSENSITIVE 5-like protein 7	18.065	17.408	18.363	16.845	19.214	17.547	4.058	5.270	4.933
BraA05g040050	Gibberellin receptor GID1A	35.347	45.975	43.943	37.547	45.288	38.148	15.541	11.744	14.423
BraA06g003820	ABSCISIC ACID-INSENSITIVE 5-like protein 4	10.857	10.866	11.143	11.815	13.328	11.233	1.128	1.714	1.653
BraA06g007440	4-substituted benzoates-glutamate ligase GH3.12	5.327	5.541	5.732	1.459	1.802	1.798	0.725	0.590	0.518
BraA06g014960	Protein BRASSINAZOLE-RESISTANT 2	126.641	126.771	125.539	125.897	136.438	117.043	39.542	38.808	42.769
BraA06g015450	Auxin response factor 5	12.360	11.264	11.435	12.483	13.511	14.027	3.736	3.216	3.810
BraA06g033000	Auxin-responsive protein SAUR20	31.496	25.096	27.039	16.152	16.278	14.569	2.086	1.478	1.141
BraA06g041230	Ethylene-responsive transcription factor 2	24.536	25.061	25.961	19.190	10.480	10.814	2.038	2.314	2.091
BraA06g041630	Transcription factor bHLH28	1.860	1.579	1.551	1.216	1.591	1.707	0.486	0.498	0.641
BraA07g016670	Cyclin-D3-2	17.426	18.424	18.465	10.281	10.710	13.435	1.606	1.879	2.446
BraA07g022730	Regulatory protein NPR6	1.762	3.711	3.478	1.296	1.816	0.895	0.167	0.070	0.112
BraA07g025310	Protein TRANSPORT INHIBITOR RESPONSE 1	15.922	14.813	15.025	13.360	14.642	13.658	4.723	4.650	4.628
BraA07g026840	Transcription factor TGA7	3.643	3.254	2.834	5.984	6.290	4.533	0.817	1.354	1.276
BraA07g026930	Auxin transporter-like protein 3	21.772	20.767	20.113	22.118	18.693	20.140	6.764	7.299	7.381
BraA07g033750	Transcription factor PERIANTHIA	4.223	4.684	4.519	3.305	3.888	2.893	0.528	0.340	0.339
BraA08g015730	Auxin-responsive protein SAUR50	16.054	13.813	11.363	11.629	21.682	12.334	1.278	2.971	1.663
BraA08g019880	Regulatory protein NPR1	4.903	5.303	5.731	6.613	6.313	6.408	1.909	1.988	1.871
BraA08g019900	Protein ph phutase 2C 56	8.519	9.131	8.536	10.229	9.358	9.403	2.501	1.893	2.564
BraA08g026970	Transcription factor TGA3	2.360	2.674	1.956	11.293	12.264	11.463	0.265	0.378	0.410
BraA09g005750	Auxin-responsive protein IAA28	5.472	3.965	4.336	11.983	10.419	10.025	1.342	2.052	1.441
BraA09g013710	Regulatory protein NPR1	5.177	4.683	4.516	5.555	4.825	4.758	0.465	0.416	0.553
BraA09g022350	Transcription factor bHLH28	5.375	6.108	5.678	3.350	3.362	3.161	0.259	0.280	0.294
BraA09g022610	Ethylene-responsive transcription factor 2	13.607	14.844	12.673	13.326	13.464	14.612	0.925	1.654	1.746
BraA09g036560	Indole-3-acetic acid-amido synthetase GH3.17	0.822	0.797	0.805	0.519	0.532	0.858	0.064	0.000	0.028
BraA09g040870	Transcription factor TGA3	37.371	40.257	30.125	27.282	40.349	38.212	10.056	9.962	10.361
BraA09g047460	ABSCISIC ACID-INSENSITIVE 5-like protein 2	3.605	5.152	4.919	5.659	5.861	3.145	2.501	1.941	0.997
BraA09g051620	Auxin-responsive protein IAA30	2.031	1.731	1.842	2.151	2.831	3.624	0.488	0.207	0.131
BraA09g052310	Gibberellin receptor GID1B	8.855	9.136	8.671	12.555	10.956	12.760	2.150	2.814	2.689
BraA09g055800	Auxin transporter-like protein 2	20.626	18.843	18.036	25.319	22.410	25.141	4.612	4.694	4.612
BraA09g056680	Protein BRASSINAZOLE-RESISTANT 2	2.627	6.216	2.803	1.261	1.342	2.507	0.625	0.521	0.519
BraA09g062360	Transcription factor PIF3	11.347	11.880	11.601	15.468	17.568	15.352	1.845	1.717	1.541
BraA09g065000	Auxin-responsive protein IAA3	20.978	22.866	24.115	20.025	20.505	19.592	2.113	2.449	2.945
BraA10g000590	Abscisic acid receptor PYL9	26.083	27.010	29.357	30.971	29.968	26.422	6.692	6.969	6.084
BraA10g001990	Histidine-containing ph photransfer protein 5	18.863	19.807	19.233	18.851	21.099	16.610	5.298	6.561	7.063
BraA10g003090	Auxin-responsive protein IAA12	11.297	12.479	10.530	9.925	10.847	8.307	1.552	0.868	1.687
BraA10g006100	Protein ph phutase 2C 3	7.671	9.076	9.258	5.468	5.788	5.679	1.022	0.673	1.033
BraA10g011520	Abscisic acid receptor PYL8	5.017	4.042	5.337	6.579	8.097	7.660	1.450	1.355	1.211
BraA10g014950	Protein ph phutase 2C 77	22.757	23.277	22.981	26.275	27.280	25.508	2.512	3.254	2.286
BraA10g015660	Xyloglucan endotransglue ylase/hydrolase protein 22	11.069	11.490	6.279	4.659	6.422	4.321	1.165	1.582	0.955

FIGURE 8 | Heatmap representation of the expression pattern [fragments per kilobase of transcript per million fragments mapped (FPKM) values] of differentially expressed genes (DEGs) related to plant hormone and signal transduction. **(A)** Downregulated and **(B)** upregulated DEGs in both late-bolting heading Chinese cabbage double haploid (DH) lines.

In case of ethylene signaling genes, DEGs encoding ethylene response sensors, such as *ethylene response sensor 2* and *ethylene-insensitive 3-like 3*, were downregulated, while *ethylene-responsive transcription factor 2* (*BraA06g041230*, *BraA02g033740*, and *BraA09g022610*), *ethylene-responsive transcription factor 15* (*BraA05g013460* and *BraA04g022530*), and ethylene receptors, such as *ethylene-insensitive 3-like 1* (*BraA03g025290*), were upregulated in both late-bolting heading Chinese cabbage DH lines compared with the early-bolting flowering Chinese cabbage DH line (**Figure 8** and **Supplementary Table 10**).

The unigenes related to auxin signal transduction, including auxin-responsive proteins, auxin transporter-like proteins, auxin responsive factor (ARF), and auxin-induced proteins, were differentially expressed between late-bolting heading Chinese cabbage and early-bolting flowering Chinese cabbage DH lines. The expression levels of 34 of these DEGs were upregulated in late-bolting heading Chinese cabbage Y410-1 compared with early-bolting flowering Chinese cabbage DH lines. Likewise, 27 and 33 DEGs were down- and upregulated, respectively, in another late-bolting DH line, SY2004 (**Figure 8** and **Supplementary Table 10**). However, among these DEGs, 27 upregulated and 15 downregulated were common to both late-bolting lines (**Supplementary Table 10**).

Expression Patterns of Differentially Expressed Genes Related to Starch and Sucrose Metabolism

Among the DEGs related to starch and sucrose metabolism, trehalose-phosphate phosphatase (A, B, G, and J), trehalase, alpha-amylase, beta-glucosidase, beta-fructofuranosidase, pectinesterase/pectinesterase inhibitor, sucrose-phosphate synthase, UDP-glucuronic acid decarboxylase, inactive beta-amylase, galacturonosyltransferase, UDP-glucuronate 4-epimerase, polygalacturonase, beta-D-xylosidase (2, 3, 5), alpha-glucosidase, UTP-glucose-1-phosphate uridylyltransferase 1, fructokinase (4, 5, 7), galacturonosyltransferase, polygalacturonate 4-alpha-galacturonosyltransferase, 1,4-alpha-glucan-branching enzyme, hexokinase, acid beta-fructofuranosidase, endoglucanase, alpha-trehalose-phosphate synthase, and exopolygalacturonase were downregulated in

both late-bolting heading Chinese cabbage DH lines (Y410-1 and SY2004) compared with the early-bolting flowering Chinese cabbage (**Figure 9**). On the other hand, trehalose-phosphate phosphatase (I, H), glucan endo-1,3-beta-glucosidase, UTP-glucose-1-phosphate uridylyltransferase 2, alpha-glucan phosphorylase (1, 2), fructokinase-1, glucose-1-phosphate adenylyltransferase large subunit, beta-D-xylosidase (1, 4), starch synthase 1, and cellulose synthase-like protein D5 were

A Down-regulating DEGs in late-bolting heading Chinese cabbage DH lines

#GeneID	Description (Swiss-Prot)	Late-bolting						Early-bolting		
		Y410-1			SY2004			CX14-1		
		T1	T2	T3	T4	T5	T6	T7	T8	T9
Bra.A01g013200	Probable trehalose-phosphate phosphatase G	2.107	1.568	2.100	6.357	7.704	6.886	29.641	30.477	29.693
Bra.A01g014770	Trehalase	0.282	0.103	0.103	0.445	0.594	0.579	5.788	6.130	6.746
Bra.A01g022020	Alpha-amylase 1	1.103	1.312	1.078	1.882	1.086	1.188	2.723	2.522	2.529
Bra.A01g029610	Beta-glucosidase 46	0.000	0.011	0.000	0.027	0.037	0.000	0.653	0.580	0.290
Bra.A01g037790	Beta-fructofuranosidase, insoluble isoenzyme CWINV1	1.079	0.775	0.502	1.645	1.191	1.315	1.545	1.534	1.732
Bra.A01g040060	Probable pectinesterase/pectinesterase inhibitor 25	0.623	0.246	0.396	4.215	2.568	3.804	1.248	1.415	1.331
Bra.A01g041790	Probable pectinesterase/pectinesterase inhibitor 23	0.057	0.034	0.042	0.000	0.000	0.000	39.601	39.767	35.816
Bra.A01g042480	Probable pectinesterase/pectinesterase inhibitor 21	0.015	0.017	0.000	0.000	0.000	0.000	14.167	14.396	15.642
Bra.A02g002330	Pectinesterase PPME1	0.000	0.000	0.046	0.000	0.000	0.000	326.472	300.827	292.380
Bra.A02g006030	UDP-glucose 6-dehydrogenase 3	34.168	35.216	34.521	27.055	22.845	25.997	56.922	56.382	55.824
Bra.A02g008800	Sucrose-phosphate synthase 1	2.102	1.916	1.858	9.986	9.314	9.936	15.195	15.145	15.926
Bra.A02g011100	UDP-glucuronic acid decarboxylase 3	65.992	70.859	68.934	42.862	31.544	51.725	167.738	165.473	169.701
Bra.A02g011800	Glucan endo-1,3-beta-glucosidase 6	0.055	0.152	0.136	1.726	1.055	1.297	1.640	1.485	1.925
Bra.A02g013180	Inactive beta-amylase 4, chloroplastic	5.107	5.739	5.322	4.630	4.571	4.772	15.360	13.875	15.003
Bra.A02g013720	Probable galacturonosyltransferase 12	4.178	3.881	3.551	3.277	3.380	3.317	11.040	11.573	11.727
Bra.A02g023150	Beta-glucosidase 20	0.000	0.000	0.000	0.000	0.000	0.000	1.220	1.350	1.906
Bra.A02g024520	Trehalose-phosphate phosphatase B	0.161	0.164	0.069	0.083	0.072	0.127	53.157	55.856	60.750
Bra.A02g028000	UDP-glucuronate 4-epimerase 5	4.436	4.548	4.693	2.337	2.458	2.801	7.059	5.913	7.134
Bra.A02g028970	Probable sucrose-phosphate synthase 4	0.308	0.404	0.314	1.540	0.959	1.416	62.951	66.042	67.045
Bra.A02g038580	Pectinesterase 31	0.245	0.699	0.823	4.898	5.250	4.266	2.784	2.402	1.830
Bra.A02g039170	Polygalacturonase	0.000	0.000	0.066	0.000	0.000	0.000	32.885	63.035	82.592
Bra.A02g039400	Polygalacturonase	0.000	0.000	0.000	0.000	0.000	0.000	77.063	74.541	60.776
Bra.A02g039420	Polygalacturonase	0.691	0.000	0.000	0.000	0.000	0.031	247.895	265.477	241.336
Bra.A02g040020	Probable pectinesterase/pectinesterase inhibitor 58	0.441	0.310	0.564	0.154	0.045	0.058	218.048	203.104	199.026
Bra.A02g044660	Beta-D-xylosidase 4	14.173	11.997	13.394	4.789	3.984	4.203	31.744	30.099	31.135
Bra.A03g001860	Probable pectinesterase/pectinesterase inhibitor 47	0.000	0.024	0.000	6.344	5.544	4.034	3.597	3.636	3.903
Bra.A03g004560	Probable sucrose-phosphate synthase 2	1.420	1.637	1.615	1.098	1.418	1.225	5.682	6.088	6.233
Bra.A03g004780	Alpha-glucosidase	12.053	12.619	11.656	10.943	9.910	10.884	22.438	22.242	22.694
Bra.A03g005050	Beta-fructofuranosidase, insoluble isoenzyme CWINV6	0.042	0.000	0.000	0.000	0.000	0.000	1.504	2.014	1.817
Bra.A03g007910	UTP--glucose-1-phosphate uridylyltransferase 1	26.087	26.040	24.519	23.241	22.256	23.308	59.338	55.307	55.291
Bra.A03g011770	Beta-glucosidase 3	0.484	0.552	0.493	1.046	0.728	0.808	11.023	10.696	11.009
Bra.A03g011780	Beta-glucosidase 4	0.446	0.800	0.657	0.403	0.153	0.256	11.057	11.350	10.180
Bra.A03g015310	Probable fructokinase-7	0.000	0.010	0.000	0.036	0.129	0.016	3.283	4.356	3.958
Bra.A03g018880	Beta-fructofuranosidase, insoluble isoenzyme CWINV4	0.214	0.184	0.257	0.023	0.161	0.181	2.134	2.389	1.908
Bra.A03g018900	Beta-fructofuranosidase, insoluble isoenzyme CWINV4	0.160	0.166	0.211	0.028	0.100	0.085	1.714	1.745	1.598
Bra.A03g023110	Polygalacturonase	0.000	0.000	0.071	0.000	0.000	0.000	1.219	1.057	1.037
Bra.A03g023690	UDP-glucuronate 4-epimerase 4	22.994	21.304	22.034	16.049	14.497	16.229	28.351	26.623	26.302
Bra.A03g027170	Probable fructokinase-5	0.000	0.050	0.111	0.078	0.000	0.169	8.489	5.253	5.182
Bra.A03g030880	Probable galacturonosyltransferase 9	15.069	10.954	15.149	51.579	47.734	35.021	72.304	70.335	70.693
Bra.A03g032200	Probable pectinesterase/pectinesterase inhibitor 21	0.236	0.306	0.334	0.021	0.185	0.014	138.763	126.141	126.187
Bra.A03g032950	Probable pectinesterase/pectinesterase inhibitor 23	0.000	0.000	0.000	0.043	0.000	0.000	17.413	18.867	15.856
Bra.A03g040900	UDP-glucuronate 4-epimerase 6	80.860	84.536	85.775	92.072	83.510	84.428	91.219	92.432	91.690
Bra.A03g041030	Beta-amylase 1, chloroplastic	3.306	3.767	3.958	3.831	4.494	3.449	18.223	17.042	18.105
Bra.A03g041470	Galacturonosyltransferase 8	42.576	42.032	44.150	104.059	103.546	104.200	61.077	59.733	59.249
Bra.A03g050290	Probable trehalose-phosphate phosphatase G	0.773	0.831	1.385	0.633	0.345	0.768	7.399	6.251	6.147
Bra.A03g051630	Trehalase	0.000	0.000	0.000	0.000	0.000	0.000	0.699	0.803	0.690
Bra.A04g000560	UDP-glucuronic acid decarboxylase 2	29.025	33.974	33.052	42.885	45.106	39.356	34.316	45.302	45.446
Bra.A04g000640	Probable beta-D-xylosidase 2	0.189	0.273	0.226	0.033	0.035	0.049	127.997	127.815	115.522
Bra.A04g001350	Polygalacturonase 4-alpha-galacturonosyltransferase	65.643	71.887	70.032	33.299	34.869	31.168	43.441	43.848	45.129
Bra.A04g002030	Beta-glucosidase 30	0.000	0.000	0.000	0.083	0.029	0.134	4.075	4.436	3.966
Bra.A04g002040	Beta-glucosidase 27	0.020	0.000	0.000	0.035	0.047	0.011	0.680	0.653	0.765
Bra.A04g003240	Polygalacturonase ADPG1	2.457	11.738	3.583	0.000	0.000	0.011	3.261	3.550	3.354
Bra.A04g006780	Beta-fructofuranosidase, insoluble isoenzyme CWINV2	4.467	3.911	3.846	0.000	0.000	0.000	25.021	23.742	23.149
Bra.A04g020590	UDP-glucuronic acid decarboxylase 6	1.001	0.782	1.290	1.968	2.768	1.845	18.458	17.162	20.283
Bra.A04g020960	Beta-glucosidase 15	1.077	0.625	0.593	0.648	0.242	0.312	5.866	6.003	5.764
Bra.A04g025740	Beta-fructofuranosidase, insoluble isoenzyme CWINV4	0.018	0.014	0.034	0.012	0.012	0.000	153.265	153.591	158.916
Bra.A04g025900	L-4-alpha-glucan-branching enzyme 2-1	5.150	5.889	5.491	20.885	20.010	21.022	12.448	12.724	12.920
Bra.A04g031090	Beta-glucosidase 27	0.102	0.123	0.100	0.881	0.841	0.623	0.080	0.900	0.821
Bra.A04g031100	Beta-glucosidase 32	0.075	0.011	0.008	5.484	5.859	4.162	0.000	0.044	0.229
Bra.A05g004350	Beta-glucosidase 15	19.522	19.555	18.444	11.634	11.834	11.363	30.283	29.811	31.709
Bra.A05g004360	Beta-glucosidase 28	0.000	0.000	0.000	0.000	0.000	0.000	4.631	4.665	4.429
Bra.A05g012860	Beta-glucosidase 33	0.093	0.020	0.102	1.290	1.028	1.540	0.824	0.691	1.000
Bra.A05g012870	Beta-glucosidase 33	0.000	0.000	0.067	0.217	0.214	0.112	1.158	1.132	1.158
Bra.A05g015010	Sucrose synthase 5	1.434	1.204	1.184	3.439	2.835	4.417	2.713	2.754	2.610
Bra.A05g016640	Pectinesterase 2	4.000	4.054	4.618	1.683	1.771	1.990	26.808	26.743	28.502
Bra.A05g019040	Hexokinase-3	4.229	4.197	3.747	2.302	1.587	1.962	8.049	6.446	6.185
Bra.A05g027680	Probable beta-D-xylosidase 5	1.020	0.717	0.676	0.183	0.122	0.257	1.811	1.313	1.328
Bra.A05g027690	Probable beta-D-xylosidase 5	1.623	1.576	1.874	0.486	0.430	0.767	3.902	3.486	3.697
Bra.A05g032750	Pectinesterase/pectinesterase inhibitor 3	184.892	193.803	193.874	91.071	80.384	95.374	164.588	154.603	167.668
Bra.A05g033220	Beta-fructofuranosidase, insoluble isoenzyme CWINV5	0.007	0.021	0.000	0.111	0.267	0.175	12.862	14.523	13.496
Bra.A05g033960	Beta-glucosidase 10	3.686	3.817	3.255	5.708	5.172	5.040	72.719	74.531	74.177
Bra.A05g038690	Probable pectinesterase/pectinesterase inhibitor 23	0.016	0.000	0.000	0.000	0.000	0.000	32.404	35.139	31.845
Bra.A05g039640	Putative pectinesterase/pectinesterase inhibitor 22	0.730	1.030	0.787	1.752	1.646	2.041	1.802	1.581	2.135
Bra.A05g040960	Probable galacturonosyltransferase 13	9.505	8.071	9.007	6.657	5.299	4.932	11.162	11.508	12.988
Bra.A06g000740	Beta-fructofuranosidase, insoluble isoenzyme CWINV3	10.914	12.113	11.414	10.821	11.805	11.100	37.277	34.778	37.080
Bra.A06g009420	Acid beta-fructofuranosidase 4, vacuolar	2.176	2.111	2.173	1.149	0.778	1.368	13.288	30.108	30.085
Bra.A06g021930	Endoglucanase 16	0.037	0.015	0.000	0.000	0.000	0.000	72.193	70.863	68.240
Bra.A06g022220	Probable pectinesterase/pectinesterase inhibitor 32	1.322	1.352	1.129	5.657	6.199	5.942	31.862	30.118	29.541
Bra.A06g023300	Endoglucanase 16	0.687	0.171	0.165	0.000	0.027	0.026	96.450	94.838	92.613
Bra.A06g028030	Beta-D-xylosidase 4	29.268	27.033	29.999	21.523	19.469	22.473	42.118	39.982	41.468
Bra.A06g028610	Probable trehalose-phosphate phosphatase J	0.412	0.391	0.523	2.442	3.031	3.064	2.458	2.258	1.892
Bra.A06g033760	Putative pectinesterase/pectinesterase inhibitor 28	0.013	0.020	0.019	0.000	0.000	0.000	62.940	58.045	56.219
Bra.A06g035800	Polygalacturonase	0.392	0.473	0.525	0.000	0.225	0.192	553.523	594.218	527.478
Bra.A07g000960	Hexokinase-1	49.822	49.612	47.476	49.166	47.466	48.152	169.493	175.290	181.912
Bra.A07g000810	Beta-amylase 1, chloroplastic	20.160	20.438	19.118	42.070	43.783	45.152	64.077	66.475	66.439
Bra.A07g019060	UDP-glucuronic acid decarboxylase 6	2.324	2.030	1.420	4.865	4.230	6.447	44.235	47.559	48.036
Bra.A07g022910	Polygalacturonase ADPG1	0.017	0.002	0.000	0.000	0.000	0.000	12.217	11.876	12.113
Bra.A07g023920	Probable fructokinase-4	0.000	0.000	0.027	0.160	0.110	0.000	53.925	50.202	51.041
Bra.A07g030960	Alpha,alpha-trehalose-phosphate synthase 6	21.054	21.749	21.096	18.294	18.415	21.008	33.817	32.804	35.092
Bra.A07g032510	Beta-D-xylosidase 3	0.000	0.007	0.000	0.000	0.000	0.000	1.139	0.959	0.782
Bra.A07g041220	Trehalose-phosphate phosphatase B	0.806	0.841	0.920	1.692	2.066	2.831	17.755	19.848	21.350
Bra.A08g011580	Putative pectinesterase/pectinesterase inhibitor 43	0.000	0.058	0.000	0.000	0.053	0.000	14.041	12.955	13.486
Bra.A08g016700	Probable pectinesterase/pectinesterase inhibitor 44	29.698	29.158	29.311	27.272	26.847	28.203	44.230	45.646	46.068
Bra.A08g024270	Exopolygalacturonase clone GBGE184	0.318	0.626	0.352	0.267	0.098	0.030	1087.615	1198.597	1055.682
Bra.A08g024280	Exopolygalacturonase clone GBGE184	0.723	0.203	0.893	0.027	0.236	0.000	407.574	407.829	388.031
Bra.A08g025770	Beta-glucosidase 40	0.356	0.270	0.018	1.129	0.635	0.953	7.976	8.922	8.414
Bra.A08g031660	Pectinesterase/pectinesterase inhibitor 18	0.000	0.000	0.000	0.000	0.000	0.000	1.865	1.166	2.165
Bra.A08g035580	Probable pectinesterase/pectinesterase inhibitor 7	0.000	0.000	0.000	0.022	0.068	0.000	1.698	1.786	2.207
Bra.A08g035590	Exopolygalacturonase clone GBGE184	1.099	1.065	1.408	0.056	0.206	0.069	1023.005	1101.986	972.900
Bra.A09g000780	UDP-glucuronate 4-epimerase 3	33.905	34.186	34.613	27.260	29.870	23.942	50.460	49.462	46.027
Bra.A09g001860	Probable pectinesterase/pectinesterase inhibitor 39	0.000	0.000	0.000	0.000	0.000	0.000	1.799	1.603	1.552
Bra.A09g001880	Sucrose synthase 3	0.211	0.243	0.150	0.340	0.272	0.226	14.794	14.757	15.393
Bra.A09g003750	Pectinesterase 31	0.337	0.332	0.028	0.629	0.824	0.381	14.069	15.765	14.027
Bra.A09g005220	Putative pectinesterase/pectinesterase inhibitor 28	0.080	0.007	0.053	0.000	0.000	0.000	76.655	69.593	67.541
Bra.A09g011560	Probable alpha,alpha-trehalose-phosphate synthase 11	6.681	6.826	6.365	10.232	8.368	10.150	12.644	12.385	10.776
Bra.A09g012100	Hexokinase-1	28.078	28.756	29.281	26.142	23.074	26.888	81.584	81.122	84.972
Bra.A09g015710	Acid beta-fructofuranosidase 3, vacuolar	0.201	0.079	0.088	0.225	0.376	0.206	9.689	7.683	9.290
Bra.A09g039970	Probable phosphoglucomutase, cytoplasmic 1	81.975	84.779	83.516	71.236	73.814	68.074	130.684	132.605	133.768
Bra.A09g048090	Polygalacturonase ADPG1	0.032	0.041	0.000	0.044	0.000	0.000	354.761	390.534	390.872
Bra.A09g049980	Beta-glucosidase 30	0.000	0.000	0.015	0.000	0.000	0.000	1.244	1.150	1.121
Bra.A09g052040	Beta-glucosidase 10	4.824	5.363	5.333	9.019	9.544	8.686	11.711	12.481	14.129
Bra.A09g060900	Acid beta-fructofuranosidase 4, vacuolar	8.908	10.142	9.465	4.674	5.492	5.310	73.229	75.167	77.624
Bra.A09g065620	Exopolygalacturonase clone GBGE184	0.000	0.000	0.000	0.000	0.000	0.000	47.299	50.243	44.370
Bra.A10g001430	Exopolygalacturonase clone GBGE184	0.824	0.918	1.410	0.230	0.479	0.161	968.653	1023.004	912.621
Bra.A10g001440	Probable pectinesterase/pectinesterase inhibitor 7	0.298	0.216	0.299	0.800	0.762	0.586	1.978	1.924	1.821
Bra.A10g001490	Beta-glucosidase 11	0.000	0.000	0.000	0.393	0.265	0.174	52.586	55.289	54.945
Bra.A10g003470	Probable sucrose-phosphate synthase 3	0.319	0.286	0.288	1.833	1.908	0.408	24.640	25.613	25.375
Bra.A10g005470	Probable galacturonosyltransferase 6	7.211	6.614	6.252	7.484	7.610	6.980	13.374	13.396	12.929
Bra.A10g010090	Trehalose-phosphate phosphatase A	0.657	0.615	0.638	1.296	1.456	1.348	7.965	6.226	6.144
Bra.A10g012660	Putative beta-glucosidase 41	2.150	2.114	2.329	3.004	3.112	3.136	17.993	20.821	18.604
Bra.A10g012780	Probable galacturonosyltransferase 12	0.657	0.501	0.566	1.353	0.951	1.252	12.175	12.179	12.110
Bra.A10g019840	Sucrose synthase 1	174.922	170.422	176.473	193.812	172.569	185.577	365.793	362.252	375.345
Bra.A10g020290	Sucrose-phosphate synthase 1	38.342	39.966	39.431	54.097	47.593	55.647	65.859	67.088	68.669
Bra.A10g029820	Pectinesterase PPME1	0.321	0.216	0.282	0.069	0.205	0.000	988.046	986.193	896.576
Bra.A10g031330	Beta-D-xylosidase 3	2.602	2.415	3.652	3.864	5.543	2.837	9.617	7.606	7.728
Brassica_rapa_newGene_1642	Beta-glucosidase 3	0.848	1.100	1.624	3.648	3.598	3.375	2.835	3.447	3.186
Brassica_rapa_newGene_4986	Endoglucanase 1	0.000	0.012	0.000	0.000	0.000	0.000	3.913	4.299	3.959
Brassica_rapa_newGene_5312	Galacturonosyltransferase 8	0.000	0.096	0.035	0.000	0.038	0.000	1.555	1.274	1.348

B Up-regulating DEGs in late-bolting heading Chinese cabbage DH lines

#GeneID	Description (Swiss-Prot)	Late-bolting						Early-bolting		
		Y410-1			SY2004			CX14-1		
		T1	T2	T3	T4	T5	T6	T7	T8	T9
Bra.A01g000660	Probable trehalose-phosphate phosphatase H	6.182	6.066	6.064	13.713	16.457	17.604	0.260	0.334	0.314
Bra.A01g004740	Probable pectinesterase/pectinesterase inhibitor 44	16.343	14.937	14.486	6.158	5.880	5.283	5.702	5.004	6.002
Bra.A01g006680	Glucan endo-1,3-beta-glucosidase 5	24.033	25.820	25.668	31.913	27.431	29.022	9.904	10.226	11.201
Bra.A01g019450	Beta-amylase 3, chloroplastic	11.885	10.849	11.470	36.213	38.065	35.567	1.345	1.446	1.796
Bra.A01g034680	Beta-glucosidase 44	342.549	334.880	343.212	144.103	141.129	145.445	55.924	59.781	64.083
Bra.A01g034690	Beta-glucosidase 44	16.100	17.486	17.152	8.640	8.620	8.663	1.120	1.110	1.688
Bra.A01g037350	Pectinesterase/pectinesterase inhibitor 3	199.434	201.875	203.821	146.840	137.706	147.742	44.719	43.390	46.559
Bra.A01g037360	Pectinesterase/pectinesterase inhibitor 3	12.227	12.248	12.400	3.546	3.282	3.205	1.254	1.466	1.401
Bra.A01g043570	Probable inactive beta-glucosidase 25	64.374	68.066	69.157	41.619	39.497	37.394	24.490	24.367	24.896
Bra.A01g043720	UTP--glucose-1-phosphate uridylyltransferase 2	8.892	10.206	9.061	7.658	8.381	7.984	0.906	1.136	0.960
Bra.A01g044420	Probable galacturonosyltransferase 9	40.998	37.740	40.383	17.545	13.355	13.889	21.254	21.765	20.447
Bra.A02g015830	Trehalose-phosphate phosphatase A	18.357	17.041	16.390	10.956	9.166	11.092	5.007	5.619	5.861
Bra.A02g016980	Glucan endo-1,3-beta-glucosidase 2	56.935	53.996	51.846	32.476	30.245	33.428	7.082	5.820	5.971
Bra.A02g034940	Glucan endo-1,3-beta-glucosidase 3	70.263	72.924	73.648	49.327	44.207	50.508	19.911	21.770	22.177
Bra.A02g038730	Alpha-glucan phosphorylase 1	208.806	218.758	212.236	154.233	155.654	154.692	23.011	20.981	23.460
Bra.A02g045000	Probable trehalose-phosphate phosphatase J	1.224	1.409	1.270	10.035	10.906	9.464	1.239	1.057	1.143
Bra.A03g004140	Probable trehalose-phosphate phosphatase I	1.053	0.596	0.898	3.032	2.487	3.070	0.509	0.607	0.610
Bra.A03g006880	UDP-glucose 6-dehydrogenase 3	12.930	12.817	12.316	4.650	4.125	3.966	4.871	4.946	6.083
Bra.A03g016690	Probable fructokinase-1	237.636	261.166	255.043	269.880	256.101	263.303	57.975	51.194	55.135
Bra.A03g036380	Pectinesterase/pectinesterase inhibitor 3	47.334	47.593	48.685	57.414	48.256	56.262	5.856	5.658	5.985
Bra.A03g047370	Beta-amylase 3, chloroplastic	221.994	260.081	260.636	414.026	397.506	367.971	96.010	95.439	99.457
Bra.A03g047730	Alpha,alpha-trehalose-phosphate synthase 5	43.619	47.637	49.831	49.058	54.437	56.234	22.628	19.524	21.239
Bra.A03g049730	Beta-glucosidase 47	14.896	15.684	14.471	6.166	8.069	6.577	0.435	0.509	0.503
Bra.A03g055790	UDP-glucuronate 4-epimerase 1	91.102	91.446	94.796	136.689	124.682	132.782	53.434	50.869	50.334
Bra.A04g011590	UDP-glucose 6-dehydrogenase 4	146.468	148.426	143.974	106.662	84.237	92.261	132.826	127.003	132.955
Bra.A04g015780	Probable glucose-1-phosphate adenylyltransferase large subunit	57.260	60.795	60.858	44.470	44.758	44.337	17.642	16.641	17.068
Bra.A04g023640	Beta-glucosidase 33	0.556	0.371	0.456	2.829	2.927	2.852	2.806	2.588	2.519
Bra.A04g031140	Beta-glucosidase 15	66.387	63.856	63.926	40.771	38.981	36.426	15.637	14.829	16.875
Bra.A05g004950	Probable pectinesterase/pectinesterase inhibitor 17	0.147	0.148	0.036	0.755	1.097	0.790	0.163	0.152	0.163
Bra.A05g008810	1,4-alpha-glucan-branching enzyme 2-1	128.022	137.070	132.195	52.368	49.188	52.624	10.182	12.430	11.197
Bra.A05g015060	Beta-glucosidase 42	3.093	3.688	3.982	0.607	0.703	0.650	0.136	0.165	0.061
Bra.A05g026800	1,4-alpha-glucan-branching enzyme 3	9.861	8.869	9.098	14.380	13.551	15.481	4.779	4.919	5.490
Bra.A06g019860	Beta-D-xylosidase 1	20.617	18.823	19.335	17.918	18.525	18.055	0.246	0.306	0.495
Bra.A06g019890	Beta-D-xylosidase 1	21.123	21.305	21.504	18.549	19.580	19.781	5.278	4.966	5.643
Bra.A06g019940	Alpha-glucan phosphorylase 2, cytosolic	280.553	306.694	309.708	382.933	360.435	370.979	126.786	129.826	134.012
Bra.A06g028790	Probable fructokinase-7	0.565	0.839	0.668	2.321	2.450	3.362	0.029	0.000	0.019
Bra.A06g031630	Starch synthase 1, chloroplastic/amyloplastic	92.715	94.093	94.726	87.652	83.651	85.950	25.614	24.055	23.074
Bra.A06g036570	Pectinesterase 31	25.714	26.516	25.383	28.797	30.874	29.843	11.984	11.427	11.004
Bra.A06g044220	Probable trehalose-phosphate phosphatase H	1.034	0.502	0.737	1.485	1.093	0.968	0.021	0.104	0.000
Bra.A06g044440	Glucose-1-phosphate adenylyltransferase large subunit 3	145.852	170.007	156.160	207.971	204.309	187.260	55.537	56.278	58.835
Bra.A07g007620	Glucose-1-phosphate adenylyltransferase small subunit	245.926	248.841	244.647	309.902	287.349	314.496	63.118	65.150	68.679
Bra.A07g025240	UDP-glucuronic acid decarboxylase 2	20.930	22.293	22.379	13.917	14.104	15.820	6.885	6.273	7.241
Bra.A07g028740	Sucrose synthase 6	2.030	1.458	1.730	4.516	4.364	5.364	3.679	3.649	3.678
Bra.A07g037520	Sucrose synthase 6	1.813	1.562	1.494	3.573	4.001	4.127	3.582	3.715	3.657
Bra.A07g039610	Probable alpha-amylase 2	1.761	2.261	1.855	7.854	8.654	8.848	5.170	6.103	5.107
Bra.A08g002960	Hexokinase-3	1.907	1.974	2.349	0.112	0.456	0.699	0.717	0.741	0.569
Bra.A08g022190	Probable trehalose-phosphate phosphatase H	10.240	7.951	8.104	20.305	19.452	17.798	0.706	0.514	0.559
Bra.A08g031380	Glucan endo-1,3-beta-glucosidase 1	69.638	68.984	68.550	59.755	56.895	60.041	12.843	12.928	13.309
Bra.A09g000540	Beta-amylase 2, chloroplastic	61.206	63.520	62.833	120.149	119.456	110.521	12.241	11.615	12.100
Bra.A09g001840	Probable pectinesterase/pectinesterase inhibitor 41	10.166	9.267	9.286	28.774	31.060	29.236	1.678	1.963	2.132
Bra.A09g044500	UDP-glucuronic acid decarboxylase 1	18.371	17.096	18.570	37.848	36.665	37.996	29.715	29.336	31.292
Bra.A09g049020	Probable galacturonosyltransferase 15	16.440	17.834	15.149	22.663	25.095	21.551	8.579	9.106	10.516
Bra.A09g049950	Beta-gluc

upregulated in both late-bolting heading Chinese cabbage lines (**Figure 9**).

Validation of RNA-seq Data by RT-qPCR

We further tested the reliability of FPKM expression patterns (determined by RNA-seq) by RT-qPCR. Eight DEGs were randomly selected, and their relative expression levels were quantified using the same RNA samples extracted from early- and late-bolting Chinese cabbage lines for RNA sequencing. The results confirmed that the expression patterns of the analyzed unigenes were consistent with the FPKM expression pattern obtained by RNA-seq (**Figure 10**).

DISCUSSION

Late bolting is an important economic trait in spring-sown lines, and late-bolting varieties have been developed and characterized for off-season production to meet the annual demand for Chinese cabbage (Yang et al., 2007; Su et al., 2018). Nevertheless, little is

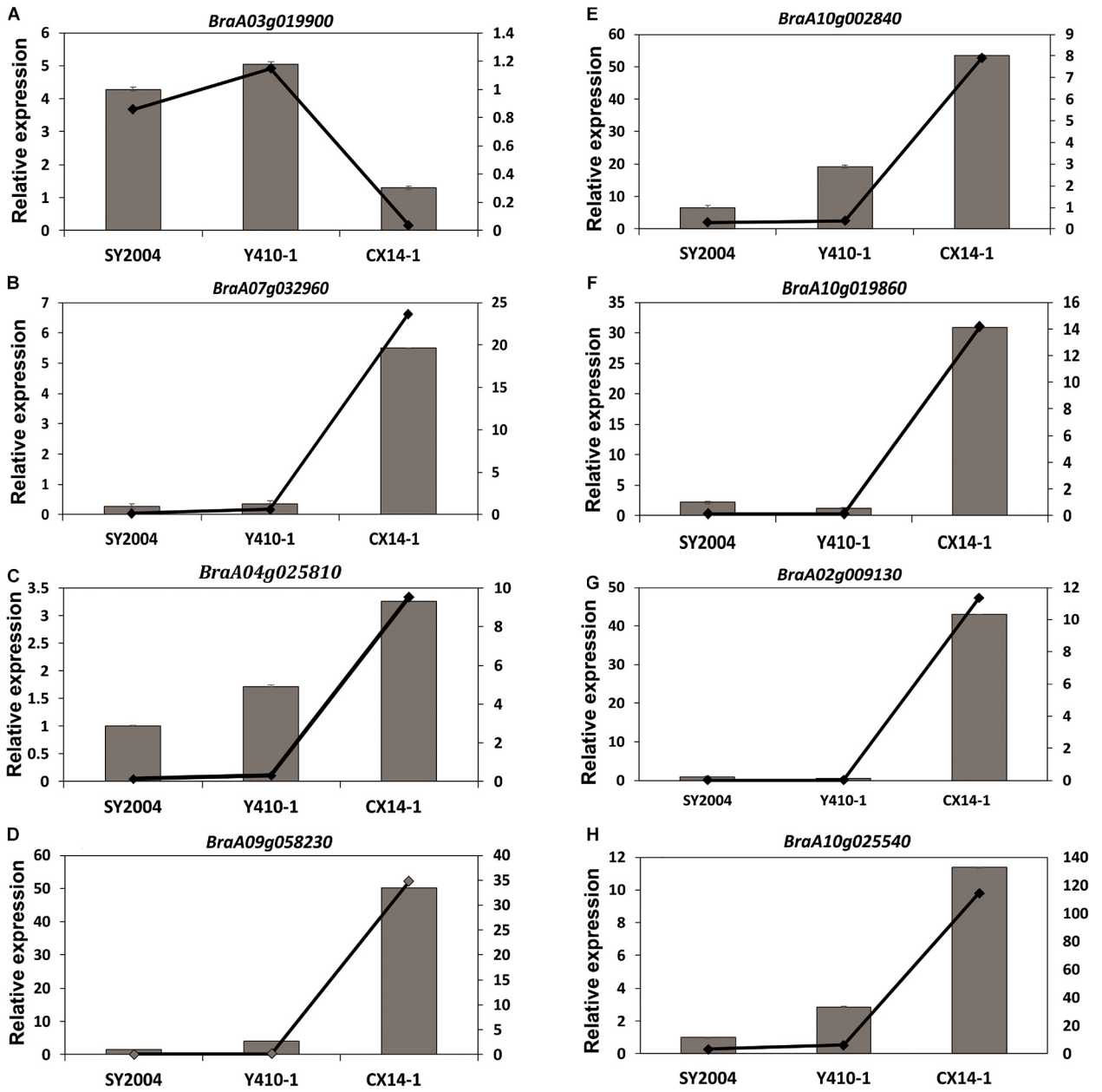


FIGURE 10 | Relative expression of eight unigenes in the floral bud of early- and late-bolting Chinese cabbage double haploid lines. Error bar represents ± SE of the means of triplicates. The superimposed line graph represents the RNA-seq expression profiles [fragments per kilobase of transcript per million fragments mapped (FPKM)].

known about the genes and pathways associated with flowering time in flowering Chinese cabbage (*B. rapa* ssp. *chinensis* var. *parachinensis*) and late-bolting heading Chinese cabbage (*B. rapa* var. *pekinensis*). RNA sequencing technology has been effectively utilized for transcriptome analyses in a wide range of texa (Wang et al., 2009; Lu et al., 2010). This high-throughput sequencing technology can be used to determine global gene expression differences between populations or species with phenotypic differences and responses to environmental stress (Mortazavi et al., 2008; Wang et al., 2009; Miao and Luo, 2013). In the present study, we performed a comparative transcriptome analysis of early-bolting Chinese cabbage DH lines by RNA-seq. We obtained 721.49 million clean paired-end reads assembled into 47,363 unigenes, including 3,144 genes predicted as novel (**Table 1** and **Supplementary Table 2**). We detected 12,932 (Y410-1 vs. CX14-1) and 12,711 (SY2004 vs. CX14-1) DEGs between two sets of early- and late-bolting Chinese cabbage (**Figure 2A** and **Supplementary Tables 4, 5**).

A functional analysis of assembled unigenes revealed that the predominant COG categories were carbohydrate transport and metabolism, general function prediction only, and signal transduction mechanisms (**Figure 6B**). Several unigenes in these categories were differentially expressed and might explain the difference in flowering times in early-bolting Chinese cabbage DH lines. Moreover, we identified 3,144 novel unigenes, of which 2,564 were functionally annotated (**Supplementary Table 2**).

A KEGG pathway enrichment analysis of DEGs between early- and late-bolting DH lines indicated that plant hormone and signal transduction (Ko04075) and starch and sucrose metabolism (Ko00500) were the most enriched pathways in both comparisons (**Supplementary Table 9**). These results suggest that unigenes related to plant hormones, signal transduction, and sucrose metabolism are involved in the regulation of flowering time in these Chinese cabbage DH lines.

The "*no hydrotropic response*" (*nhr1*) *Arabidopsis* mutants show high levels of ABA, resulting in late flowering (Quiroz-Figueroa et al., 2010), whereas ABA-deficient (*aba2* and *aba3*) or ABA-insensitive (*abai4*) mutants show early flowering than wild types (Martínez-Zapater et al., 1994; Matsoukas, 2014). Moreover, Zhu et al. (2013) demonstrated that *ABA-insensitive 5* (*ABI5*, encoding a bZIP transcription factor) delayed flowering time in *Arabidopsis* under long-day conditions. Our results also indicated that the expression levels of most genes encoding ABI5-like proteins (ABA-insensitive 5-like protein 2, ABA-insensitive 5-like protein 4, ABA-insensitive 5-like protein 5, ABA-insensitive 5-like protein 6, and ABA-insensitive 5-like protein 7) were higher in late-bolting Chinese cabbage DH lines than in early-bolting flowering Chinese cabbage (**Figure 8** and **Supplementary Table 10**).

Previous studies of mutants related to GA biosynthesis or signal transduction have revealed that GA can alter the flowering time (Wilson et al., 1992; Peng and Harberd, 1993; Sun and Kamiya, 1994; Jacobsen et al., 1996; Peng et al., 1997; Andres et al., 2014). Moreover, the DELLA domain protein RGA (repressor of ga1-3), GAI (GA insensitive), and RGA-like1 (RGL1 and RGL2) act as negative regulators of the GA signaling pathway (Yu et al., 2004; Silverstone et al., 2007). Our results revealed that two DEGs encoding DELLA proteins RGA1 (BraA06g040430) and RGA2 (BraA09g023210) are upregulated in late-bolting heading Chinese cabbage "Y410-1" (**Figure 8** and **Supplementary Table 10**) and downregulated at the flowering stage compared with the vegetative stages in early-bolting flowering Chinese cabbage. These results suggest that the downregulation of negative regulators of the GA signaling pathway might increase GA levels, and the lack of function of these genes may trigger early bolting. Similar results have been reported by Huang et al. (2017), who reported that *RGA1* and *RGA2* are downregulated in flowering Chinese cabbage.

Several previous studies have indicated that the phytohormone jasmonic acid (JA) also regulates flowering time in *Arabidopsis* (Zhai et al., 2015). The F-box protein COI1 (coronatine insensitive 1) degrades JAZ (contains TIFY and Jas domains) repressors (Hoo and Howe, 2009). Transcript levels of *NaJAZd* and *NaJAZh* are upregulated in the early floral stages of *NaJAZi*-silenced plants due to the high JA content in the flowers (Li et al., 2017). Our results also showed that DEGs encoding TIFY proteins are more highly expressed in the early-bolting flowering Chinese cabbage DH line (CX14-1) than in both late-bolting heading Chinese cabbage lines (**Figure 8** and **Supplementary Table 10**). These results suggest that the elevated expression of these genes in early flowering Chinese cabbage might be related to a high JA content.

Another steroidal phytohormone, brassinosteroid (BR), promotes flower induction in plants (Matsoukas, 2014). The BR signaling genes *BZR1* (*brassinazole-resistant1*) and *BES1* suppress the expression of important genes related to BR biosynthesis (*CPD*, *constitutive photomorphogenesis*, *and dwarfism*; *DWF4*, and *DWARF4*) by binding to their promoter regions in *Arabidopsis* (Wei et al., 2017). Moreover, BR-deficient/BR-insensitive mutants show delayed flowering time (Domagalska et al., 2007; Zhu et al., 2013). We also found that DEGs related to *BZR1*, *BZR2*, and *BRI1* (*brassinosteroid insensitive 1*) were more highly expressed in both late-bolting heading Chinese cabbage DH lines than in early-bolting flowering Chinese cabbage (**Figure 8** and **Supplementary Table 10**). These results suggest that the upregulation of BR signaling genes might affect BR biosynthesis and cause late bolting in late-bolting heading Chinese cabbage DH lines.

Starch and sucrose play important roles in flowering (Turnbull, 2011; Cho et al., 2018). Previous reports revealed that trehalose-6-phosphate acts as a signal molecule for flowering initiation in different plant species, including *Arabidopsis thaliana* (Sheen, 2014), grape (Caspari et al., 1998), and citrus (Shalom et al., 2014). Besides, Micallef et al. (1995) found that an increased level of endogenous sucrose promotes flowering in tomato. Wahl et al. (2013) reported that *TREHALOSE-6-PHOSPHATE SYNTHASE 1* (*TPS1*) is required for the regulation of flowering time in *A. thaliana*. In the present study, *alpha,alpha-trehalose-phosphate synthase 5* (BraA03g047730) was upregulated in both late-bolting heading Chinese cabbage lines compared with early-bolting flowering Chinese cabbage lines (**Figure 9**).

Razi et al. (2008) demonstrated that *BoFLC* alleles segregate independently from flowering time alleles in *Brassica oleracea*. However, Yuan et al. (2009) reported that variation in *BrFLC1*

is linked to flowering time in *B. rapa*. Moreover, Xie et al. (2015) performed a QTL analysis and showed that *Br2* is the key determinant of *BrFLC2* and a candidate flowering time locus in *B. rapa*. A transposon insertion in the coding sequence of *BrFT2* located on a QTL on chromosome A07 (region Br5) causes late flowering (Zhang et al., 2015). Therefore, we further analyzed the flowering time-related genes, especially *FLC* and *FT*, based on FPKM expression values. Unigenes, such as *FLK* (*BraA03g031700*) and *FLD* (*BraA03g034300*), showed higher expression levels in both late-bolting lines than in early-bolting lines, while *FLT* (*BraA02g016700*) showed the opposite pattern (**Supplementary Table 11**). With regard to *FT* genes, the expression levels of three unigenes encoding proteins related to flowering time control, such as FCA (BraA01g020520), FY (BraA02g004930), and FPA (BraA09g036880), were higher in both late-bolting lines than in early-bolting lines (**Supplementary Table 11**). Nonetheless, none of these unigenes were differentially expressed between early- and late-bolting lines.

In late-bolting Chinese cabbage, 98 and 89 unigenes related to plant hormone and signal transduction were upregulated in Y410-1 vs. CX14-1 and SY2004 vs. CX14-1, respectively (**Figure 8** and **Supplementary Table 9**). Among these, 19 and 12 unigenes showed $\log_2$ fold change values of > 5.0 in Y410-1 vs. CX14-1 and SY2004 vs. CX14-1, respectively; while six unigenes, including *BraA02g009130* (indole-3-acetic acid-induced protein ARG7), *BraA09g058230* (Auxin-responsive protein SAUR41), *BraA07g032960* (serine/threonine-protein kinase BSK11), *BraA10g019860* (auxin-induced protein 15A), *BraA08g012630* (abscisic acid receptor PYR1), and *BraA01g009450* (abscisic acid receptor PYR1), were common in both late-bolting heading Chinese cabbage lines (Y410-1 and SY2004). These unigenes are candidates for earl-bolting and late-bolting traits in these DH Chinese cabbage lines and could be useful for the development of molecular markers for the detection of early- and late-bolting cultivars.

DATA AVAILABILITY STATEMENT

Data is available at NCBI SRA accession: PRJNA605481.

AUTHOR CONTRIBUTIONS

XZ and YY conceived and designed the experiments. YZ, SY, and ZW performed the experiments. HS, LL, LN, and MH-U-R prepared the figures and tables. XW and MR drafted the work or revised it critically for important content. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (31801874), the Programs for Science and Technology Development of Henan Province (212102110124 and 212102110126), the Modern Agro-Industry Technology Research System (CARS-23-G-16), and the Zhongyuan Scholar Program (202101510003).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.590830/full#supplementary-material

REFERENCES

Amasino, R. M. (2005). Vernalization and flowering time. *Curr. Opin. Biotech*. 16, 154–158. doi: 10.1016/J.COPBIO.2005.02.004

Andersen, C. H., Jensen, C. S., and Petersen, K. (2004). Similar genetic switch systems might integrate the floral inductive pathways in dicots and monocots. *Trends Plant Sci*. 9, 105–107. doi: 10.1016/J.TPLANTS.2004.01.002

Andres, F., Porri, A., Torti, S., Mateos, J., Romera-Branchat, M., García-Martínez, J. L., et al. (2014). SHORT VEGETATIVE PHASE reduces gibberellin biosynthesis at the *Arabidopsis* shoot apex to regulate the floral transition Fernando Andrés. *Proc. Natl. Acad. Sci. U.S.A*. 111, E2760–E2769. doi: 10.1073/pnas.1409567111

Berardini, T. Z., Mundodi, S., Reiser, L., Huala, E., Garcia-Hernandez, M., Zhang, P., et al. (2004). Functional annotation of the *Arabidopsis* genome using controlled vocabularies. *Plant Physiol*. 135, 745–755. doi: 10.1104/pp.104.040071

Bernier, G. (1988). The control of floral evocation and morphogenesis. *Ann. Rev. Plant Physiol. Plant Mol. Biol*. 39, 175–219. doi: 10.1146/annurev.pp.39.060188.001135

Bong, Y. S., Shin, W. J., Gautam, M. K., Jeong, Y. J., Lee, A. R., Jang, C. S., et al. (2012). Determining the geographical origin of Chinese cabbages using multielement composition and strontium isotope ratio analyses. *Food Chem*. 135, 2666–2674. doi: 10.1016/J.FOODCHEM.2012.07.045

Caspari, H. W., Lang, A., and Alspach, P. (1998). Effects of girdling and leaf removal on fruit set and vegetative growth in grape. *Am. J. Enol. Viti*. 49, 359–366.

Chandler, J., and Dean, C. (1994). Factors influencing the vernalization response and flowering time of late flowering mutants of *Arabidopsis thaliana* (L.) Heynh. *J. Exp. Bot*. 45, 1279–1288. doi: 10.1093/jxb/45.9.1279

Cheng, F., Liu, S., Wu, J., Fang, L., Sun, S., Liu, B., et al. (2011). BRAD, the genetics and genomics database for *Brassica* plants. *BMC Plant Biol*. 11:136. doi: 10.1186/1471-2229-11-136

Cheng, F., Sun, R., Hou, X., Zheng, H., Zhang, F., Zhang, Y., et al. (2016). Subgenome parallel selection is associated with morphotype diversification and convergent crop domestication in *Brassica rapa* and *Brassica oleracea*. *Nat. Genet*. 48, 1218–1224. doi: 10.1038/ng.3634

Cheng, F., Wu, J., and Wang, X. (2014). Genome triplication drove the diversification of *Brassica* plants. *Hort. Res*. 1:14024. doi: 10.1038/hortres.2014.24

Cho, L. H., Pasriga, R., Yoon, J., Jeon, J. S., and An, G. (2018). Roles of sugars in controlling flowering time. *J. Plant Biol*. 61, 121–130. doi: 10.1007/s12374-018-0081-z

Choi, K., Kim, J., Hwang, H. J., Kim, S., Park, C., Kim, S. Y., et al. (2011). The FRIGIDA complex activates transcription of FLC, a strong flowering repressor in *Arabidopsis*, by recruiting chromatin modification factors. *Plant Cell* 23, 289–303. doi: 10.1105/tpc.110.075911

Dahanayake, S. R., and Galwey, N. W. (1999). Effects of Interactions between low-temperature treatments, gibberellin (GA3), and photoperiod on flowering and stem height of spring rape (*Brassica napus* var. *annua*). *Ann. Bot*. 84, 321–327. doi: 10.1006/ANBO.1999.0920

Domagalska, M. A., Schomburg, F. M., Amasino, R. M., Vierstra, R. D., Nagy, F., and Davis, S. J. (2007). Attenuation of brassinosteroid signaling enhances FLC

expression and delays flowering. *Development* 134, 2841–2850. doi: 10.1242/dev.02866

Duan, W., Song, X., Liu, T., Huang, Z., Ren, J., Hou, X., et al. (2015). Genome-wide analysis of the MADS-box gene family in *Brassica rapa* (Chinese cabbage). *Mol. Genet. Genom.* 290, 239–255. doi: 10.1007/s00438-014-0912-7

Eddy, S. R. (1998). Profile hidden markov models. *Bioinformatics* 14, 755–763.

Fornara, F., de Montaigu, A., and Coupland, G. (2010). Snapshot: control of flowering in *Arabidopsis*. *Cell* 141, 550–550. doi: 10.1016/j.cell.2010.04.024

Fu, W., Huang, S., Gao, Y., Zhang, M., Qu, G., Wang, N., et al. (2020). Role of BrSDG8 on bolting in Chinese cabbage (*Brassica rapa*). *Theor. Appl. Genet.* 133, 2937–2948. doi: 10.1007/s00122-020-03647-4

Gendall, A. R., Levy, Y. Y., Wilson, A., and Dean, C. (2001). The VERNALIZATION 2 gene mediates the epigenetic regulation of vernalization in *Arabidopsis*. *Cell* 107, 525–535.

Greenup, A., Peacock, W. J., Dennis, E. S., and Trevaskis, B. (2009). Molecular biology of seasonal flowering-responses in *Arabidopsis* and cereals. *Ann. Bot.* 103, 1165–1172. doi: 10.1093/aob/mcp063

Hoo, S. C., and Howe, G. A. (2009). A critical role for the TIFY motif in repression of jasmonate signaling by a stabilized splice variant of the JASMONATE ZIM-domain protein JAZ10 in *Arabidopsis*. *Plant Cell* 21, 131–145. doi: 10.1105/tpc.108.064097

Huang, S., Hou, L., Fu, W., Liu, Z., Li, C., Li, X., et al. (2020). An insertion mutation in Bra032169 encoding a histone methyltransferase is responsible for early bolting in chinese cabbage (*Brassica rapa* L. ssp. *pekinensis*). *Front. Plant Sci.* 11:547. doi: 10.3389/fpls.2020.00547

Huang, X., Lei, Y., Guan, H., Hao, Y., Liu, H., Sun, G., et al. (2017). Transcriptomic analysis of the regulation of stalk development in flowering Chinese cabbage (*Brassica campestris*) by RNA sequencing. *Sci. Rep.* 7:15517. doi: 10.1038/s41598-017-15699-6

Jacobsen, S. E., Binkowski, K. A., and Olszewski, N. E. (1996). SPINDLY, a tetratricopeptide repeat protein involved in gibberellin signal transduction in *Arabidopsis*. *Proc. Natl. Acad. Sci. U.S.A.* 93, 9292–9296. doi: 10.1073/pnas.93.17.9292

Johanson, U., West, J., Lister, C., Michaels, S., Amasino, R., and Dean, C. (2000). Molecular analysis of FRIGIDA, a major determinant of natural variation in *Arabidopsis* flowering time. *Science* 290, 344–347.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., et al. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* 36, D480–D484. doi: 10.1093/nar/gkm882

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32, 277–280. doi: 10.1093/nar/gkh063

Kobayashi, Y., Kaya, H., Goto, K., Iwabuchi, M., and Araki, T. (1999). A pair of related genes with antagonistic roles in mediating flowering signals. *Science* 286, 1960–1962. doi: 10.1126/science.286.5446.1960

Langridge, J. (1957). Effect of day-length and gibberellic acid on the flowering of *Arabidopsis*. *Nature* 180, 36–37. doi: 10.1038/180036a0

Lee, M. K., Chun, J. H., Byeon, D. H., Chung, S. O., Park, S. U., Park, S., et al. (2014). Variations in glucosinolates in 62 varieties of Chinese cabbage (*Brassica rapa* L. ssp. *pekinensis*) and their antioxidant activity. *LWT Food Sci. Technol.* 58, 93–101. doi: 10.1016/J.LWT.2014.03.001

Li, R., Wang, M., Wang, Y., Schuman, M. C., Weinhold, A., Schäfer, M., et al. (2017). Flower-specific jasmonate signaling regulates constitutive floral defenses in wild tobacco. *Proc. Natl. Acad. Sci. U.S.A.* 114, E7205–E7214. doi: 10.1073/pnas.1703463114

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

Lu, T., Lu, G., Fan, D., Zhu, C., Li, W., Zhao, Q., et al. (2010). Functional annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome Res.* 20, 1238–1249. doi: 10.1101/gr.106120.110

Manchali, S., Chidambara Murthy, K. N., and Patil, B. S. (2012). Crucial facts about the health benefits of popular cruciferous vegetables. *J. Funct. Foods* 4, 94–106. doi: 10.1016/J.JFF.2011.08.004

Mao, X., Cai, T., Olyarchuk, J. G., and Wei, L. (2005). Automated genome annotation and pathway identification using KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics* 21, 3787–3793. doi: 10.1093/bioinformatics/bti430

Martínez-Zapater, J., Coupland, G., Dean, C., and Koornneef, M. (1994). Transition to flowering in *Arabidopsis*. *Cold Spring Harb. Monogr. Arch.* 27, 403–433.

Matsoukas, I. G. (2014). Interplay between sugar and hormone signaling pathways modulates floral signal transduction. *Front. Genet.* 5:218. doi: 10.3389/fgene.2014.00218

Miao, X., and Luo, Q. (2013). Genome-wide transcriptome analysis between small-tail Han sheep and the Surabaya fur sheep using high-throughput RNA sequencing. *Reproduction* 145, 587–596. doi: 10.1530/REP-12-0507

Micallef, B. J., Haskins, K. A., Vanderveer, P. J., Roh, K. S., Shewmaker, C. K., and Sharkey, T. D. (1995). Altered photosynthesis, flowering, and fruiting in transgenic tomato plants that have an increased capacity for sucrose synthesis. *Planta* 196, 327–334. doi: 10.1007/BF00201392

Michaels, S. D., and Amasino, R. M. (1999). FLOWERING LOCUS C encodes a novel MADS domain protein that acts as a repressor of flowering. *Plant Cell* 11, 949–956.

Moon, J., Suh, S. S., Lee, H., Choi, K. R., Hong, C. B., Paek, N. C., et al. (2003). The *SOC1* MADS-box gene integrates vernalization and gibberellin signals for flowering in *Arabidopsis*. *Plant J.* 35, 613–623. doi: 10.1046/j.1365-313X.2003.01833.x

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628. doi: 10.1038/nmeth.1226

Peng, J., Carol, P., Richards, D. E., King, K. E., Cowling, R. J., and Murphy, G. P. (1997). The *Arabidopsis GAI* gene defines a signaling pathway that negatively regulates gibberellin responses. *Genes Devel.* 11, 3194–3205. doi: 10.1101/gad.11.23.3194

Peng, J., and Harberd, N. P. (1993). Derivative alleles of the *Arabidopsis* gibberellin-insensitive (gai) mutation confer a wild type phenotype. *Plant Cell* 5, 351–360. doi: 10.1105/tpc.5.3.351

Quiroz-Figueroa, F., Rodríguez-Acosta, A., Salazar-Blas, A., Hernández-Domínguez, E., Campos, M. E., Kitahata, N., et al. (2010). Accumulation of high levels of ABA regulates the pleiotropic response of the nhr1 *Arabidopsis* mutant. *J. Plant Biol.* 53, 32–44. doi: 10.1007/s12374-009-9083-1

Razi, H., Howell, E. C., Newbury, H. J., and Kearsey, M. J. (2008). Does sequence polymorphism of FLC paralogues underlie flowering time QTL in *Brassica oleracea*? *Theor. Appl. Genet.* 116, 179–192. doi: 10.1007/s00122-007-0657-3

Rood, S. B., Mandel, R., and Pharis, R. P. (1989). Endogenous gibberellins and shoot growth and development in *Brassica napus*. *Plant Physiol.* 89, 269–273. doi: 10.1104/PP.89.1.269

Samach, A., Onouchi, H., Gold, S. E., Ditta, G. S., Schwarz-Sommer, Z., Yanofsky, M. F., et al. (2000). Distinct roles of CONSTANS target genes in the reproductive development of *Arabidopsis*. *Science* 288, 1613–1616.

Schiessl, S. V., Huettel, B., Kuehn, D., Reinhardt, R., and Snowdon, R. J. (2017). Flowering time gene variation in *Brassica* species shows evolutionary principles. *Front. Plant Sci.* 8:1742. doi: 10.3389/fpls.2017.01742

Shalom, L., Samuels, S., Zur, N., Shlizerman, L., Doron-Faigenboim, A., Blumwald, E., et al. (2014). Fruit load induces changes in global gene expression and in abscisic acid (ABA) and indole acetic acid (IAA) homeostasis in citrus buds. *J. Exp. Bot.* 65, 3029–3044. doi: 10.1093/jxb/eru148

Sheen, J. (2014). Master regulators in plant glucose signaling networks. *J. Plant Biol.* 57, 67–79. doi: 10.1007/s12374-014-0902-7

Sheldon, C. C., Burn, J. E., Perez, P. P., Metzger, J., Edwards, J. A., Peacock, W. J., et al. (1999). The FLF MADS box gene: a repressor of flowering in *Arabidopsis* regulated by vernalization and methylation. *Plant Cell* 11, 445–458. doi: 10.1105/tpc.11.3.445

Shu, J., Liu, Y., Zhang, L., Li, Z., Fang, Z., Yang, L., et al. (2018). QTL-seq for rapid identification of candidate genes for flowering time in broccoli× cabbage. *Theor. Appl. Genet.* 131, 917–928. doi: 10.1007/s00122-017-3047-5

Silverstone, A. L., Tseng, T. S., Swain, S. M., Dill, A., Sun, Y. J., Olszewski, N. E., et al. (2007). Functional analysis of spindly in gibberellin signaling in *Arabidopsis*. *Plant Physiol.* 143, 987–1000. doi: 10.1104/pp.106.091025

Simpson, G. G., and Dean, C. (2002). *Arabidopsis*, the rosetta stone of flowering time? *Science* 296, 285–289. doi: 10.1126/science.296.5566.285

Song, X., Duan, W., Huang, Z., Liu, G., Wu, P., Liu, T., et al. (2015). Comprehensive analysis of the flowering genes in Chinese cabbage and examination of evolutionary patterns of CO-like genes in the plant kingdom. *Sci. Rep.* 5:14631. doi: 10.1038/srep14631

Song, Y. H., Ito, S., and Imaizumi, T. (2013). Flowering time regulation: photoperiod and temperature sensing in leaves. *Trends Plant Sci.* 18, 575–583. doi: 10.1016/J.TPLANTS.2013.05.003

Srikanth, A., and Schmid, M. (2011). Regulation of flowering time: all roads lead to Rome. *Cell. Mol. Life Sci.* 68, 2013–2037. doi: 10.1007/s00018-011-0673-y

Su, T., Wang, W., Li, P., Zhang, B., Li, P., Xin, X., et al. (2018). A genomic variation map provides insights into the genetic basis of spring Chinese cabbage (*Brassica rapa* ssp. *pekinensis*) selection. *Mol. Plant* 11, 1360–1376. doi: 10.1016/j.molp.2018.08.006

Suárez-López, P., Wheatley, K., Robson, F., Onouchi, H., Valverde, F., and Coupland, G. (2001). CONSTANS mediates between the circadian clock and control of flowering in *Arabidopsis*. *Nature* 410, 1116–1120. doi: 10.1038/35074138

Sun, T. P., and Kamiya, Y. (1994). The *Arabidopsis* GA1 locus encodes the cyclase ent-kaurene synthetase a of gibberellin biosynthesis. *Plant Cell* 6, 1509–1518. doi: 10.1105/TPC.6.10.1509

Sung, S., and Amasino, R. M. (2004). Vernalization in *Arabidopsis thaliana* is mediated by the PHD finger protein VIN3. *Nature* 427, 159–164. doi: 10.1038/nature02195

Turnbull, C. (2011). Long-distance regulation of flowering time. *J. Exp. Bot.* 62, 4399–4413. doi: 10.1093/jxb/err191

Wahl, V., Ponnu, J., Schlereth, A., Arrivault, S., Langenecker, T., Franke, A., et al. (2013). Regulation of flowering by trehalose-6-phosphate signaling in *Arabidopsis thaliana*. *Science* 339, 704–707. doi: 10.1126/science.1230406

Wang, L., Feng, Z., Wang, X., Wang, X., and Zhang, X. (2010). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* 26, 136–138. doi: 10.1093/bioinformatics/btp612

Wang, Y., Li, L., Ye, T., Lu, Y., Chen, X., and Wu, Y. (2013). The inhibitory effect of ABA on floral transition is mediated by ABI5 in *Arabidopsis*. *J. Exp. Bot.* 64, 675–684. doi: 10.1093/jxb/ers361

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics *Nat. Rev. Genet*. 10, 57–63. doi: 10.1038/nrg2484

Wei, Z., Yuan, T., Tarkowská, D., Kim, J., Nam, H. G., Novák, O., et al. (2017). Brassinosteroid biosynthesis is modulated via a transcription factor cascade of COG1, PIF4, and PIF5. *Plant Physiol.* 174, 1260–1273. doi: 10.1104/pp.16.01778

Wilson, R. N., Heckman, J. W., and Somerville, C. R. (1992). Gibberellin is required for flowering in *Arabidopsis thaliana* under short days. *Plant Physiol.* 100, 403–408.

Wu, J., Yuan, Y. X., Zhang, X. W., Zhao, J., Song, X., Li, Y., et al. (2008). Mapping QTLs for mineral accumulation and shoot dry biomass under different Zn nutritional conditions in Chinese cabbage (*Brassica rapa* L. ssp. *pekinensis*). *Plant Soil* 310, 25–40. doi: 10.1007/s11104-008-9625-1

Xie, Q., Lou, P., Hermand, V., Aman, R., Park, H. J., Yun, D. J., et al. (2015). Allelic polymorphism of GIGANTEA is responsible for naturally occurring variations in circadian period in *Brassica rapa*. *Proc. Natl. Acad. Sci. U.S.A.* 112, 3829–3834. doi: 10.1073/pnas.1421803112

Xu, L., Hu, K., Zhang, Z., Guan, C., Chen, S., Hua, W., et al. (2015). A genome-wide association study revealed the genetic architecture of flowering time in rapeseed (*Brassica napus* L.). *DNA Res.* 23:dsv035. doi: 10.1093/dnares/dsv035

Yang, X., Yu, Y. J., Zhang, F. L., Zou, Z. R., Zhao, X. Y., Zhang, D. S., et al. (2007). Linkage map construction and quantitative trait loci analysis for bolting based on a double haploid population of *Brassica rapa*. *J. Int. Plant Biol.* 49, 664–671. doi: 10.1111/j.1744-7909.2007.00447.x

Young, M. D., Wakefield, M. J., Smyth, G. K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* 11:R14. doi: 10.1186/gb-2010-11-2-r14

Yu, H., Ito, T., Zhao, Y., Peng, J., Kumar, P., and Meyerowitz, E. M. (2004). Floral homeotic genes are targets of gibberellin signaling in flower development. *Proc. Natl. Acad. Sci. U.S.A.* 101, 7827–7832. doi: 10.1073/pnas.0402377101

Yuan, Y. X., Wu, J., Sun, R. F., Zhang, X. W., Xu, D. H., Bonnema, G., et al. (2009). A naturally occurring splice site mutation in the *Brassica rapa* FLC1 gene is associated with variation in flowering time. *J. Exp. Bot.* 60, 1299–1308. doi: 10.1093/jxb/erp010

Zanewich, K. P., Rood, S. B., and Williams, P. H. (1990). Growth and development of *Brassica* genotypes differing in endogenous gibberellin content I. Leaf and reproductive development. *Physiol. Plant* 79, 673–678. doi: 10.1111/j.1399-3054.1990.tb00043.x

Zhai, Q., Zhang, X., Wu, F., Feng, H., Deng, L., Xu, L., et al. (2015). Transcriptional mechanism of the jasmonate receptor COI1-mediated delay of flowering time in *Arabidopsis*. *Plant Cell* 27, 2814–2828. doi: 10.1105/tpc.15.00619

Zhang, X., Meng, L., Liu, B., Hu, Y., Cheng, F., Liang, J., et al. (2015). A transposon insertion in FLOWERING LOCUS T is associated with delayed flowering in *Brassica rapa*. *Plant Sci.* 241, 211–220. doi: 10.1016/J.PLANTSCI.2015.10.007

Zhao, J., Wang, X., Deng, B., Lou, P., Wu, J., Sun, R., et al. (2005). Genetic relationships within *Brassica rapa* as inferred from AFLP fingerprints. *Theor. Appl. Genet.* 110, 1301–1314.

Zhu, J. Y., Sae-Seaw, J., and Wang, Z. Y. (2013). Brassinosteroid signaling. *Development* 140, 1615–1620. doi: 10.1242/dev.060590

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.

Copyright © 2021 Wei, Rahim, Zhao, Yang, Wang, Su, Li, Niu, Harun-Ur-Rashid, Yuan and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 31 March 2021
doi: 10.3389/fgene.2021.596749

Identify of Fast-Growing Related Genes Especially in Height Growth by Combining QTL Analysis and Transcriptome in *Salix matsudana* (Koidz)

*Guoyuan Liu[1†], Qingshan Yang[2†], Junfeng Gao[1], Yuwei Wu[1], Zhicong Feng[1], Jingke Huang[1], Hang Zou[1], Xingzhao Zhu[1], Yanhong Chen[1], Chunmei Yu[1], Bolin Lian[1], Fei Zhong[1] and Jian Zhang[1]**

[1] Key Laboratory of Landscape Plant Genetics and Breeding, School of Life Sciences, Nantong University, Nantong, China, [2] Shandong Academy of Forestry, Jinan, China

OPEN ACCESS

Edited by:
Zhiyong Liu,
Shenyang Agricultural University, China

Reviewed by:
Vijay Gahlaut,
Institute of Himalayan Bioresource Technology (CSIR), India
Zhuo Renying,
Chinese Academy of Forestry, China

***Correspondence:**
Jian Zhang
ntdxylzw@163.com

[†] These authors have contributed equally to this work

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

Received: 20 August 2020
Accepted: 03 March 2021
Published: 31 March 2021

Citation:
Liu G, Yang Q, Gao J, Wu Y, Feng Z, Huang J, Zou H, Zhu X, Chen Y, Yu C, Lian B, Zhong F and Zhang J (2021) Identify of Fast-Growing Related Genes Especially in Height Growth by Combining QTL Analysis and Transcriptome in Salix matsudana (Koidz). Front. Genet. 12:596749. doi: 10.3389/fgene.2021.596749

The study on the fast-growing traits of trees, mainly valued by tree height (TH) and diameter at breast height (DBH), is of great significance to promote the development of the forest industry. Quantitative trait locus (QTL) mapping based on high-density genetic maps is an efficient approach to identify genetic regions for fast-growing traits. In our study, a high-density genetic map for the F_1 population was constructed. The genetic map had a total size of 5,484.07 centimorgan (cM), containing 5,956 single nucleotide polymorphisms (SNPs) based on Specific Length Amplified Fragment sequencing. Six fast-growing related stable QTL were identified on six chromosomes, and five stable QTL were identified by a principal component analysis (PCA). By combining the RNA-seq analysis for the two parents and two progenies with the qRT-PCR analysis, four candidate genes, annotated as *DnaJ, 1-aminocyclopropane-1-carboxylate oxidase 1 (ACO1), Caffeic acid 3-O-methyltransferase 1 (COMT1),* and *Dirigent protein 6 (DIR6)*, that may regulate height growth were identified. Several lignin biosynthesis-related genes that may take part in height growth were detected. In addition, 21 hotspots in this population were found. The results of this study will provide an important foundation for further studies on the molecular and genetic regulation of TH and DBH.

Keywords: *Salix matsudana Koidz.*, fast-growing, height growth, RNA-seq, quantitative trait locus

INTRODUCTION

Forests are the most important source of natural raw material for industries and the environment. The study on the fast-growing traits of trees is of great significance to promote the development of the forest industry. Fast-growing traits, including height growth, diameter growth, volume of wood, growth period, dry weight, and biomass, are complex quantitative traits controlled by multiple genes and environmental factors. High fast-growing trees produce enough raw material to satisfy the industry. Developing high fast-growing cultivars with good wood quality remains a challenge for forest breeding. Till now, a series of studies have focused on genes that play an important

role during tree development. However, their roles in genetically controlling fast-growing traits are poorly understood.

Fast-growing traits are both complex and quantitative. They are not only controlled by environmental factors such as light, temperature, water, and fertilizer, but are also controlled by genetic variation, the effect of which is much greater. Among the fast-growing related traits, TH and DBH growth are the most important characteristics of forest trees. Many studies have identified important genes that are involved in the development of TH and DBH. For example, in the stems of woody dicotyledonous plants, the procambium further differentiates outward into phloem cells and inward into xylem cells. In the process of differentiation, the cambium in the bundle between the xylem and phloem still has its meristematic ability (Rose, 2016). About 60 mm from the top of the stem, the stem begins to thicken gradually and the secondary growth of vascular tissue begins (Murmanis, 1970). Studies in poplars have shown that *WUSCHEL*, *CLAVATA*, *SHOOT MERISTEMLESS,* and members of these gene families are involved in cambium primordial cell activity and regulate tree growth (Brand et al., 2000; Gross-Hardt and Laux, 2003; Sarkar et al., 2007). It has also been found that *PXY*, a gene encoding CLV-like LRR-kinase in poplars, plays an important role in maintaining the normal polarity of procambium cells and the structure of the vascular development space (Fisher and Turner, 2007; Fukuda et al., 2007; Hirakawa et al., 2008). In *Arabidopsis thaliana*, it was found that the cambium activity of the mutant hca gene and the secondary growth of the whole plant were hampered, and the expanded secondary growth changed the structure of the stem's vascular tissue (Pineau et al., 2005). The *Arabidopsis COV1* gene encodes a membrane protein with an unknown function and regulates the proliferation of procambium/cambium. The number of xylems and phloems in the stem of a *COV1* mutant increases along with the number of vascular bundles near the base of the stem (Parker et al., 2003). The *Arabidopsis STM* gene and *BP* gene are the main regulators of stem cell maintenance and cell differentiation in the apical meristem (Groover et al., 2006; Du et al., 2009). The *Arabidopsis HD-ZIP III* gene is involved in regulating the growth of the apical meristem (Floyd et al., 2006) and its homologous gene *PCN* can also slow down the growth of the poplar (Schrader et al., 2004).

Due to its strong tolerance to salt, water, heavy metals, the cold, diseases, and pests, *Salix matsudana* Koidz. is widely distributed around the earth, especially in China. Additionally, the willow has a high biomass, is easy to reproduce, and is rich in variety. It is widely used in artificial forests and its wood is an important raw material in papermaking, gunpowder, construction equipment, particleboard, and other industries. Researchers have already studied gene expression on biomass, salt stress, response to heavy metals, and so on. However, the genetic relationship with fast-growing traits such as TH and DBH is unclear.

In our previous study, an F_1 population was developed from two *Salix matsudana* Koidz. cultivars with significant differences in TH and DBH (Zhang et al., 2016). The objectives of this study were (1) to locate TH- and DBH-related QTL in the F_1 population based on its high-density genetic map and a reference genome and (2) to combine the above results with RNA-seq analysis and qRT-PCR analysis to identify candidate genes within stable TH QTL regions.

MATERIALS AND METHODS

Plant Materials and Tissue Collection

Two *Salix matsudana* Koidz. with significantly different TH and DBH were chosen as parents: "9901," the male parent with the taller TH and bigger DBH and "Yanjiang," the female with the shorter TH and smaller DBH. The two parents were hybridized to produce F_1 in 2014. The F_1 population of 195 plants was grown for DNA extraction at the experimental forest farm in Nantong, Jiangsu, China, in 2015 (Zhang et al., 2016). The branches of 195 F_1 progenies and the two parents were clipped at 10 cm lengthwise and 1 cm thick and cut into nutrient soil in Nantong University, in March 2020. The clipped branches of a high F_1 progeny (named as "FH"), a short F_1 progeny (named as "FS"), and the two parents were cut into nutrient soil in three biological replications for RNA-seq, additionally. The terminals of the stems (0–5 cm) were collected from each replication. The excised stem terminals were immediately frozen in liquid nitrogen and stored at −80°C until use.

Measurement of Fast-Growing Traits

In November 2018 and 2019, the TH and DBH of the F_1 population were investigated because the two parents significantly differed in those two traits. In November 2020, only TH of the cuttings of F_1 population were measured, while little difference was identified on DBH for these cuttings. Additionally, the growing speed of TH (height per year, HPY) and DBH (DBH per year, DPY) were calculated; the phenotypic traits were listed in **Supplementary Table S1**. The statistical analysis, i.e., the correlation analysis and PCA, were performed by R language.

RNA Sequencing and Library Construction

Total RNA was extracted from each replication using the Plant RNA Reagent kit (Tiangen, China), according to the manufacturer's instructions. All RNA samples were quantified using a Nanodrop ND 2000 spectrophotometer (NanoDrop, Thermo, United States). Three RNA samples of each biological replications from stem terminal sample of "9901," "Yanjian," "FH", and "FS" were stored at −80°C. Finally, Illumina sequencing technology (Illumina, San Diego, CA, United States) was employed to perform RNA sequencing by Majorbio (Shanghai, China).

Analysis of Sequencing Data

The transcriptome reads were processed into clean, full-length reads by removing the low-quality and adapter reads (Chen et al., 2020). The assembled *Salix matsudana* Koidz. ("Yanjiang") genome sequence was selected as the reference for paired-end

reads mapping (Zhang et al., 2020). The clean reads were aligned to genes of the reference genome using the HiSAT2 software[1] with default parameters (Kim et al., 2019). Then StringTie[2] was used to detect new transcripts (Pertea et al., 2015). RSEM[3] was chosen to calculate the fragments per kilobase transcriptome per million mapped reads (FPKM) by normalizing for the length of the gene and for the number of mapped reads. The differentially expressed genes (DEGs) were judged based on the following standards using the Deseq2 package: (1) the FPKM of DEGs should show at least 2-fold changes in expression level between different libraries and (2) the p-adjust value of false discovery rate (FDR) correction with Benjamini–Hochberg should be less than 0.05 (Li et al., 2017; Chen et al., 2020). A BLASTx search was then performed against the NCBI non-redundant protein database to annotate the DEGs. The top three hits ($p < 0.001$) were chosen as the annotation of the potential function of each predicted target. Finally, gene ontology (GO) categories and KEGG pathway analysis were performed to evaluate the potential functions of the targets using Blast2GO (Young et al., 2010). The data that support the findings of this study have been deposited into CNGB Sequence Archive (CNSA) of China National GeneBank DataBase (CNGBdb)[4] with accession number CNP0001576.

Linkage Map Construction and Mapping of Fast-Growing Traits

The DNA of 195 F_1 progeny were extracted, constructed, and sequenced by the Specific Length Amplified Fragment sequencing (SALF-seq) in our previous research. After removing the low-quality reads, the clean reads from each sample were then aligned to the reference genome using Burrows-Wheeler Aligner (BWA) software (set at mem -t 4 -k 32 -M -R) (Li and Durbin, 2009). GATK software was used to call SNPs for all of the samples (McKenna et al., 2010). SNP markers with segregation patterns of ab × cd, ef × eg, hk × hk, nn × np, lm × ll in the parents were used to construct a linkage map. SNP markers with no more than 15% missing data in the F_1 population and a *p*-value of segregation distortion of less than 0.05 were selected to construct a linkage map (Liu et al., 2019). The SNP markers were first divided into 38 groups according to the position mapped on the 38 chromosomes of the reference genome of "Yanjiang." JoinMap 4.0 was used for the linkage map construction (van Ooijen, 2006). Interval mapping (IM) method was employed to detect TH-, DBH-, and PCA-related QTL using MapQTL 6 (Bokore et al., 2019). The parameters were set to 1 cM of the step and 1,000 permutations were taken as the LOD threshold. QTL were named according to McCouch et al. (1997). MareyMap was applied to construct a recombination map, which displayed a smooth curve with the Loess method (Rezvoy et al., 2007). Regions no less than 50 cM/Mb were regarded as recombination hotspots (Liu et al., 2019).

[1]https://daehwankimlab.github.io/hisat2/

[2]http://ccb.jhu.edu/software/stringtie/

[3]http://deweylab.github.io/RSEM/

[4]https://db.cngb.org/cnsa/

Identification of Candidate Genes and Expression Pattern Analysis

The DEGs and TH QTL were co-localized onto the reference genome based on a BLAST search. Total RNA of stem terminals from "9901," "Yanjian," "FH," and "FS" were extracted. The One-Step SYBR Primer Script Plus RT-PCR kit (Takara, Beijing, China) was used according to the manufacturer's instructions to conduct a qRT-PCR analysis of the candidate genes. The *Actin* gene was used as an internal control (Chen et al., 2020). All primers are listed in **Supplementary Table S3**.

RESULTS

Determination of Fast-Growing Traits in the F1 Population

The traits of the F_1 population are summarized in **Table 1**. As shown, there is a significant difference between the TH and DBH of the two parents. Both TH and DBH exhibited transgressive segregation in the segregating population. The heritability of HPY and DPY were 0.877 and 0.853, respectively. For the lack of replicates in each environment, the heritability of TH and DBH were not performed. Furthermore, TH, DBH, HPY, and DPY were significantly correlated (**Figure 1A**), indicating possible pleiotropic effects of the same QTL for these fast-growing traits. According to the PCA, which was performed to detect the common factors underlying trait variation, all traits showed high positive loadings on PCA1, which can explain 78.8% of the variance of traits (**Figure 1B**). This result suggests that F_1 plants with high PCA1 scores in this population exhibited tall TH and high DBH. This corresponds to a trade-off relationship between TH and DBH. The PCA2 only explained a 9.8% variance. The loading on different environments was different, suggesting that PCA2 is representative of a different environment. Additionally, the result also showed that TH and DBH were stable in different years, which is consistent with the correlation analysis.

High-Throughput Linkage Map Construction

A high-density linkage map containing a total of 5,956 SNPs divided into 38 linkage groups (LGs) was constructed. The total genetic distance was 5,484.07 cM and the average distance between adjacent markers was 0.92 cM (**Table 2**). The genetic distances of the 38 LGs ranged from to 96.02 cM (B09) to 252.34 cM (A16). The SNP markers mapped to each LG varied from 72 (A19) to 332 (A16). Among the 38 LGs, the genetic distance of chromosome A16 and D16 were longer than 200 cM and only A17 and A19 contained no more than 100 SNP markers. Chromosome B10 contained the largest gap (19.89 cM). Overall, the markers were randomly distributed on 38 chromosomes, which means the linkage map is suitable to perform further QTL analysis on the whole genome.

TABLE 1 | Performance and analysis of fast-growing traits in two parents and F_1 population.

Trait	Environment	Parents			F_1 population			
		Male	Female	Significant	Range	Means	Kurtosis	Skewness
		Parent (9901)	Parent (Yanjiang)					
DBH(cm)	2018	3.66	3.40	*	0.60–8.20	3.59	0.20	0.68
	2019	5.49	4.99	*	3.20–13.10	5.11	0.53	0.78
Height(cm)	2018	584.72	442.23	**	134.00–663.00	498.59	0.68	−1.11
	2019	768.47	600.94	**	198.11–1208.45	695.51	−0.13	−0.07
	2020	249.94	145.37	**	90.00–290.00	190.42	−0.13	0.07
DPY(cm)	Combine	1.01	0.92	*	0.15–2.34	0.93	0.27	0.68
HPY(cm)	Combine	103.62	81.27	**	33.50–199.97	128.66	0.08	−0.64

*Significant difference at $p < 0.01$ and $p < 0.05$ were marked as ** and *, respectively.*

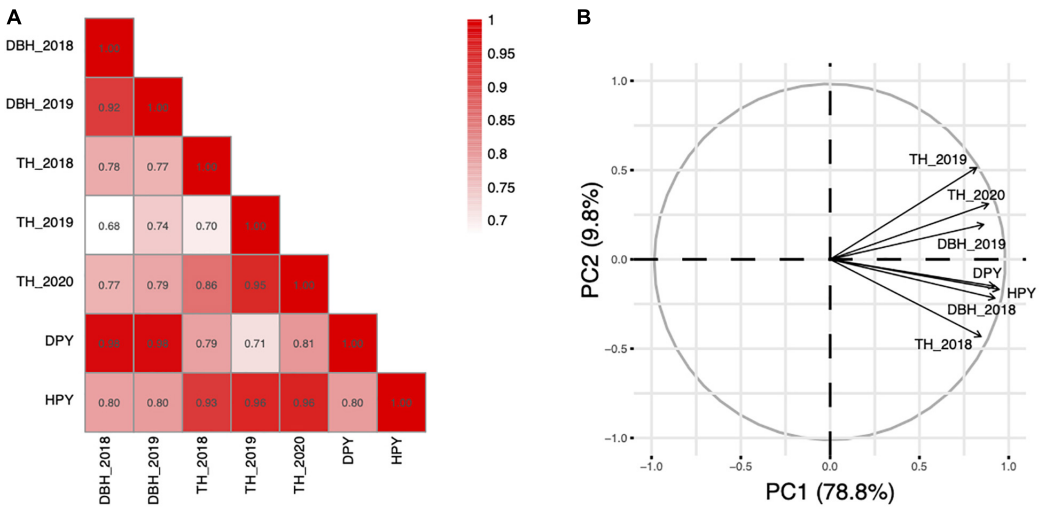


FIGURE 1 | Correlation and PCA analysis among the fast-growing related traits. **(A)** Correlation analysis among the fast-growing related traits. **(B)** PCA analysis of the fast-growing related traits.

Collinearity was measured to assess the quality of this genetic map (**Figure 2** and **Table 2**). The results indicated that most LGs in this newly constructed linkage map have high collinearity with the physical map of the reference genome "Yanjiang."

We also measured the distribution of genome-wide variation of recombination rates. As shown in **Figure 2**, the recombination rate was found to vary across chromosomes. Twenty-one recombination spikes were found in the centromeric regions of A01, A02, A05, A08, A12, A13, A14, A16, A18, A19, B01, B03, B07, B08, B15, B17, and B18. The average recombination rate for each chromosome also showed significant differences, ranging from 8.01 to 15.41 cM/Mb (**Table 2**), with an overall genome-wide recombination rate of 10.29 cM/Mb. On average, the recombination rate was similar between the At and Bt sub-genomes. Most recombination spikes were induced by chromosome segment inversions, such as A01, A02, A08, A13, A16, B01, B03, B07, B17, B18, and B19. Several recombination hotspots were detected in both homoeologous chromosomes, i.e., A01 and B01 and A08 and B08. Except for the segment inversion, the average genome-wide recombination rate was not random, as the distal chromosomal regions showed higher recombination rates than the proximal regions in most chromosomes.

QTL Mapping of Fast-Growing Traits in the F1 Population

Based on the high-density genetic map and the PCA, a total of 29 fast-growing and PC1-related QTL were identified on 15 chromosomes: A02, A03, A10, A11, A13, A15, A16, B02, B04, B08, B10, B13, B16, B18, and B19 (**Figure 3**). These 29 QTL contained 10 DBH QTL, 11 TH QTL, and 8 PCA QTL, each explaining 5.8–10% of the phenotypic variation (PV). Among the 29 QTL, six and seven stable QTL related to DBH and TH in at least two environments, respectively, were detected (**Table 3**). Moreover, seven regions of the genome contained both TH QTL and DBH QTL, which indicates that these six multi-effect regions could affect both TH and DBH. According to the distribution of these 29 QTLs, chromosomes A02, A15, A16, B02, B04, and B18 contained both TH- and DBH-related genes. However, we also detected QTLs on chromosomes A03, A10, A11, B13, and

TABLE 2 | Detailed information on the high-density genetic map.

Chromosome	Length (cM)	Number of mapped markers	Average interval (cM)	Largest gap (cM)	Recombination rate (cM/Mb)	Collinearity ratio
A01	156.42	103	1.52	8.58	13.45	99.02%
A02	174.32	218	0.80	13.78	10.35	98.36%
A03	137.38	164	0.84	5.29	11.04	99.56%
A04	153.93	142	1.08	6.80	10.37	98.35%
A05	151.98	197	0.77	7.76	10.53	96.97%
A06	182.24	223	0.82	10.83	8.45	98.48%
A07	126.16	152	0.83	8.59	8.86	97.37%
A08	142.91	162	0.88	4.74	11.59	98.64%
A09	103.07	110	0.94	7.62	9.31	97.65%
A10	155.46	151	1.03	6.73	9.37	98.47%
A11	121.46	100	1.21	13.03	10.53	97.36%
A12	129.50	114	1.14	6.16	11.78	95.19%
A13	138.26	171	0.81	6.23	11.12	97.37%
A14	121.48	120	1.01	5.14	13.08	94.85%
A15	111.51	120	0.93	6.82	9.93	94.14%
A16	252.34	332	0.76	7.30	8.88	98.67%
A17	110.12	79	1.39	11.64	12.06	97.34%
A18	113.58	104	1.09	11.38	9.42	89.44%
A19	104.84	72	1.46	9.09	11.66	95.86%
At sub-genome	2686.95	2834	0.95	13.78	10.33	97.01%
B01	165.44	201	0.82	8.02	8.94	98.98%
B02	175.73	193	0.91	6.81	9.76	97.40%
B03	144.46	135	1.07	9.72	10.79	95.20%
B04	166.48	232	0.72	5.23	9.74	98.11%
B05	175.40	294	0.60	10.47	9.07	97.32%
B06	174.28	257	0.68	6.8	8.01	98.04%
B07	137.23	136	1.01	16.77	11.94	97.72%
B08	144.56	135	1.07	7.55	11.91	98.48%
B09	96.02	113	0.85	5.94	8.65	97.65%
B10	158.97	135	1.18	19.89	13.56	97.80%
B11	129.25	109	1.19	7.67	11.08	96.89%
B12	106.10	125	0.85	7.48	8.72	94.36%
B13	150.69	188	0.80	8.33	10.71	96.60%
B14	129.30	105	1.23	7.38	14.69	93.98%
B15	148.81	118	1.26	9.95	15.41	92.36%
B16	240.00	278	0.86	8.52	8.60	99.28%
B17	123.62	152	0.81	6.39	9.26	97.19%
B18	123.86	112	1.11	12.62	11.52	97.17%
B19	106.93	104	1.03	10.54	10.82	94.36%
Bt sub-genome	2797.12	3122	0.90	19.89	10.25	96.78%
Total	5484.07	5956	0.92	19.89	10.29	94.70%

B16 only related with TH, and QTLs on chromosomes A13, B08, B10, and B19 that only regulate DBH. These chromosomes were detected related with TH or DBH, respectively. The result may provide scientific basis for genetic engineering and molecular marker-assistant breeding of TH and DBH. The eight PC1-related QTL were detected on eight chromosomes. We then compared whether these eight regions overlapped with the seven multi-effect regions. As a result, five of the eight (62.50%) QTL were located in the same region as the seven multi-effect regions on the reference genome, and all five QTL could be detected in the two environments. The result indicated that the PCA is powerful enough to detect multi-effect QTL for high correlated traits. Based on the PCA, five stable QTL that could affect both DBH and TH were detected, which may explain the significant positive relationship between TH and DBH. The five stable QTL were located on chromosomes A02, B02, B04, A15, and A16 (**Figure 3**).

Fifteen of the 29 QTL were located on the At sub-genome and the remaining 14 QTL were located on the Bt sub-genome. Both TH QTL and DBH QTL were detected on chromosomes A02, B02, B04, A15, A16, and B18. Interestingly, for the other nine chromosomes, the At sub-genome chromosomes (A03, A10, and A11) contained more TH QTL, while most DBH QTL were

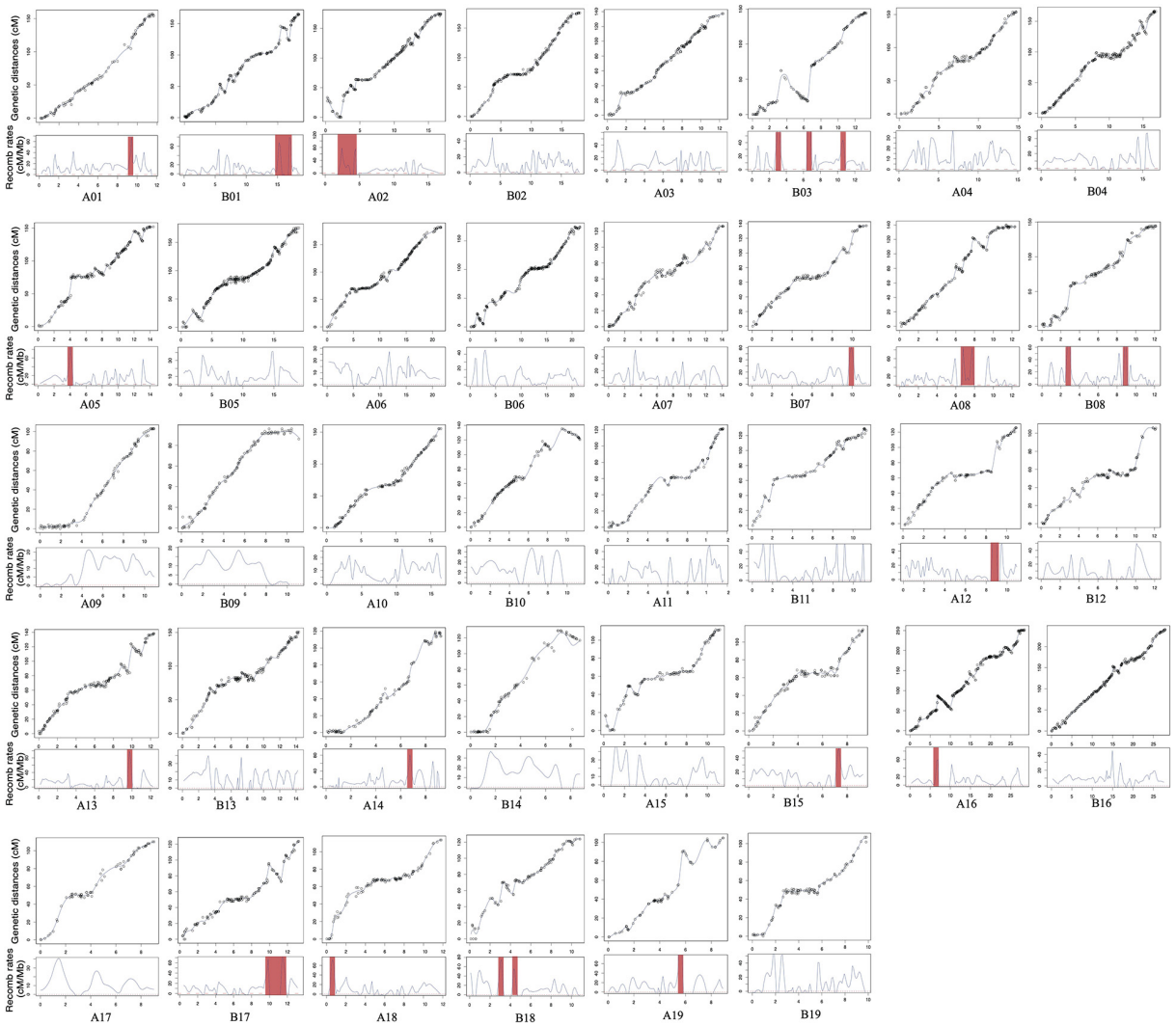


FIGURE 2 | Correlation among genetic and physical maps, estimated local recombination rates, and their distribution in chromosomal rearrangement regions. The red shadow represents recombination hotspots.

located on the Bt sub-genome chromosomes (B08, B10, and B19). TH QTL on A03, A10, A11, and A15 could not be identified on its homoeologous chromosomes in the Bt sub-genome. Similarly, chromosomes A08, A10, and A19 also did not contain DBH QTL. By comparing the QTL distribution on the two sub-genomes, we could draw a conclusion that for this population, one-third of the QTL region could affect both TH and DBH and for the single effect QTL, the At sub-genome contains more TH-related QTL, while the Bt Sub-genome contains more DBH-related QTL.

Comparative Transcriptome Analysis of TH

We performed a transcriptome analysis to understand the molecular mechanism of TH. By sequencing the stem terminal of "FH," "FS," "9901," and "Yanjiang," a total of 39,327 genes were detected to be expressed in at least one library. This result suggests that 67.99% of the predicted genes (a total of 57,841 gene models in "Yanjiang") are expressed in the stem terminals. After trimming off the adapter sequences and removing the low-quality reads, we obtained 42,080,544–58,249,766 clean reads for the 12 libraries, with a single read length of 90 bp and Q20 and Q30 percentage (percentage of sequences with sequencing error rates lower than 1 and 0.1%) over 97 and 93%, respectively (**Supplementary Table S2**). The DEGs in both parents and the two progenies were then identified using the threshold FDR $\leq$ 0.05 and the absolute value of log2-fold change $\geq$1 (Li et al., 2017; Chen et al., 2020). As a result, a total of 10,635 and 1,209 DEGs were identified between two parents and two progenies, respectively. Among the 10,635 DEGs, 4,953 genes expressed higher in "9901" than "Yanjiang" and 5,682 genes expressed higher in "Yanjiang" than "9901." For the two progenies, 531 genes expressed higher in "FH" than "FS" and 678 genes expressed higher in "FS" than "FH". According to the Venn

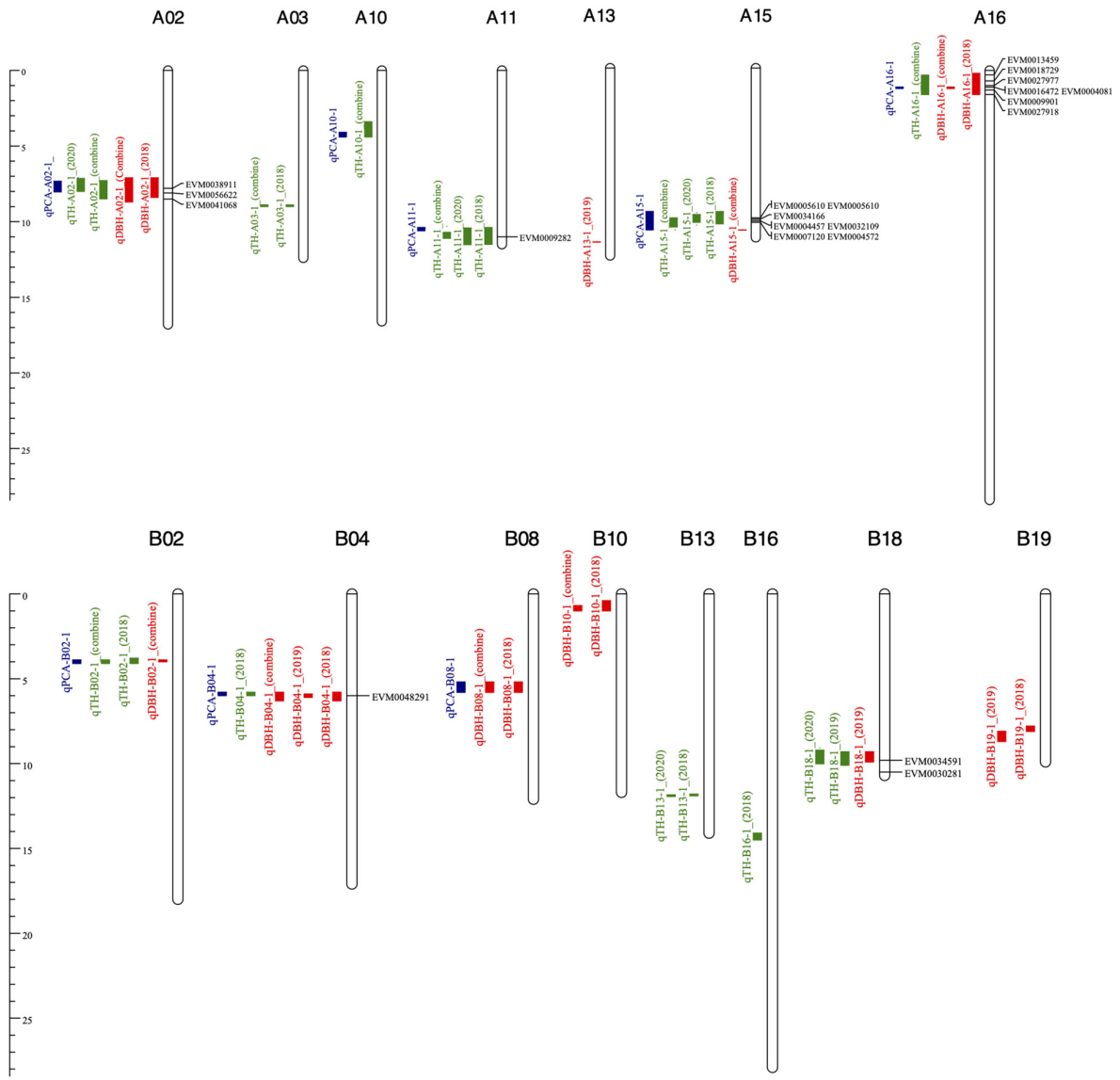


FIGURE 3 | QTL of TH, DBH, and PCA on the physical map. TH, DBH, and PCA QTL are marked in green, red and blue, respectively.

analysis, 116 DEGs showed high expression levels in both "9901" and "FH" and another 207 DEGs expressed higher in "Yanjiang" and "FS" (**Figure 4A**). Since most DEGs expressed higher in low fast-growing cultivar, we may infer that most DEGs have a negative relationship with fast-growing traits.

The 323 DEGs (116 up-regulated and 207 down-regulated) in "9901" and "FH" then underwent the GO and KEGG enrichment analysis. As shown in **Figure 4B**, most DEGs were enriched in the molecular function term. ADP binding was the most enriched, followed by adenyl nucleotide binding, squalene monooxygenase activity, and adenyl ribonucleotide binding, among others. For the biological process, the sterol biosynthetic process was the most enriched. A cell wall and an external encapsulating structure were detected in the cellular components. For the KEGG enrichment analysis, most identified pathways were enriched in metabolism (**Figure 4C**). Only sesquiterpenoid and triterpenoid biosynthesis, steroid biosynthesis, monoterpenoid biosynthesis, nitrogen metabolism, and phenylpropanoid biosynthesis pathways were identified with significant enrichment levels (**Figure 4D**). The results of the GO and KEGG analysis indicate that terpenoid-, steroid- and phenylpropanoid-related pathways may play important roles in TH growth.

Identification of Candidate Genes for TH

To further understand the genetic mechanism of *Salix matsudana* Koidz. on fast-growing traits, we co-located the DEGs and 11 TH QTL on the reference genome. Our hypothesis is that only these differentially expressed genes that are located in QTL regions may be genetically associated with TH.

TABLE 3 | Detailed information of 29 QTL for fast-growing in F_1 population.

Traits	QTL	Environment	Chromosome	LOD	R^2 (%)	95% confidence interval (cM)	Physical interval (Mb)
DBH	qDBH-A02-1	2018	A2	3.62	9	67.84–84.53	7.11–8.41
		Combine	A2	3.48	8.87	67.84–85.09	7.11–8.69
	qDBH-B02-1	Combine	B2	2.5	5.8	55.40–55.40	3.86–4.00
	qDBH-B04-1	2018	B4	3.37	7.7	61.58–73.34	5.80–6.32
		2019	B4	2.75	7.3	65.58–69.48	5.85–6.05
		Combine	B4	3.97	9.3	60.58–73.34	5.78–6.32
	qDBH-B08-1	2018	B8	3.67	8.6	72.78–75.78	5.24–5.76
		Combine	B8	3.31	7.8	73.78–75.28	5.24–5.76
	qDBH-B10-1	2018	B10	2.94	6.8	6.53–11.10	0.39–0.96
		Combine	B10	2.74	6.3	7.53–10.10	0.74–0.96
	qDBH-A13-1	2019	A13	2.73	8.5	126.54–127.54	11.45–11.50
	qDBH-A15-1	Combine	A15	2.73	6.3	108.92–110.42	10.68–10.70
	qDBH-A16-1	2018	A16	2.73	7.7	0–8.91	0.15–1.57
		Combine	A16	2.5	5.9	0.15–1.57	1.10–1.12
	qDBH-B18-1	2019	B18	2.78	7.8	118.00–120.45	9.29–9.91
	qDBH-B19-1	2018	B19	2.58	5.9	71.05–78.29	7.83–8.12
		2019	B19	2.72	6.9	78.06–87.63	8.12–8.72
PCA	qPCA-A01-1	PCA	A2	3.02	7.2	68.00–80.13	7.11–7.79
	qPCA-B02-1	PCA	B2	2.51	5.8	55.37–55.40	3.86–4.06
	qPCA-B04-1	PCA	B4	3.24	7.7	62.58–72.71	5.82–6.02
	qPCA-B08-1	PCA	B8	2.92	6.9	73.78–75.28	5.24–5.76
	qPCA-A10-1	PCA	A10	2.52	6.9	38.99–42.52	4.10–4.44
	qPCA-A11-1	PCA	A11	2.94	7.3	97.62–103.25	10.35–10.55
	qPCA-A15-1	PCA	A15	2.73	6.9	89.97–109.42	9.46–10.70
	qPCA-A16-1	PCA	A16	2.51	5.9	0.15–1.57	1.10–1.12
TH	qTH-A02-1	2020	A2	4.06	10	66.46–77.58	7.09–7.67
		Combine	A2	2.65	6.2	71.97–74.28	7.26–7.46
	qTH-B02-1	2018	B2	2.61	6	53.82–55.88	3.80–4.06
		Combine	B2	2.57	5.9	55.37–55.40	3.86–4.06
	qTH-A03-1	2018	A3	2.66	6.6	102.79–105.15	8.91–9.03
		Combine	A3	2.67	6.7	102.79–105.15	8.91–9.03
	qTH-B04-1	2018	B4	3.02	7.9	63.58–72.71	5.83–6.02
	qTH-A10-1	Combine	A10	3.13	7.7	33.37–42.52	3.42–4.44
	qTH-A11-1	2018	A11	2.82	7	98.07–119.51	10.36–11.50
		2020	A11	3.37	8.4	98.27–105.57	10.36–11.04
		Combine	A11	3.23	7.9	95.62–104.90	10.34–10.70
	qTH-B13-1	2018	B13	2.66	6.2	106.14–114.08	10.53–10.82
		2020	B13	3.08	7.7	105.34–117.11	10.53–10.82
	qTH-A15-1	2018	A15	3.01	7.7	89.97–104.17	9.46–10.32
		2020	A15	3.54	8.8	95.71–103.07	9.79–10.27
		Combine	A15	2.9	6.7	106.14–109.42	10.53–10.70
	qTH-A16-1	Combine	A16	2.59	6.2	0.15–1.57	1.10–1.57
	qTH-B16-1	2018	B16	2.83	6.5	124.90–128.80	14.11–14.45
	qTH-B18-1	2019	B18	2.85	7.4	118.00–121.26	9.29–10.06
		2020	B18	3.82	9.5	112.78–120.12	8.98–9.91

Based on the reference genome, 228 of the 323 DEGs were mapped on 38 chromosomes and 95 of the 323 DEGs were mapped on scaffolds. Of the 228 DEGs, 18 could be co-located with seven stable TH QTL on chromosomes A02, A11, A15, A16, B04, and B18. The 18 co-located DEGs were annotated as plastid-lipid-associated proteins: *DnaJ*, Berberine bridge enzyme, *ACO1*, *COMT1,* and NAC and WAKY transcription factors, among others (**Supplementary Table S4**). Most (14) of these 18 genes showed higher expression levels in "Yanjiang" and "FS" than "9901" and "FH," which means that these genes have a negative relationship with TH. For example, *ACO1, COMT1, DnaJ*, and several transcription factors showed a negative relationship with TH. The remaining four genes showed a positive relationship with

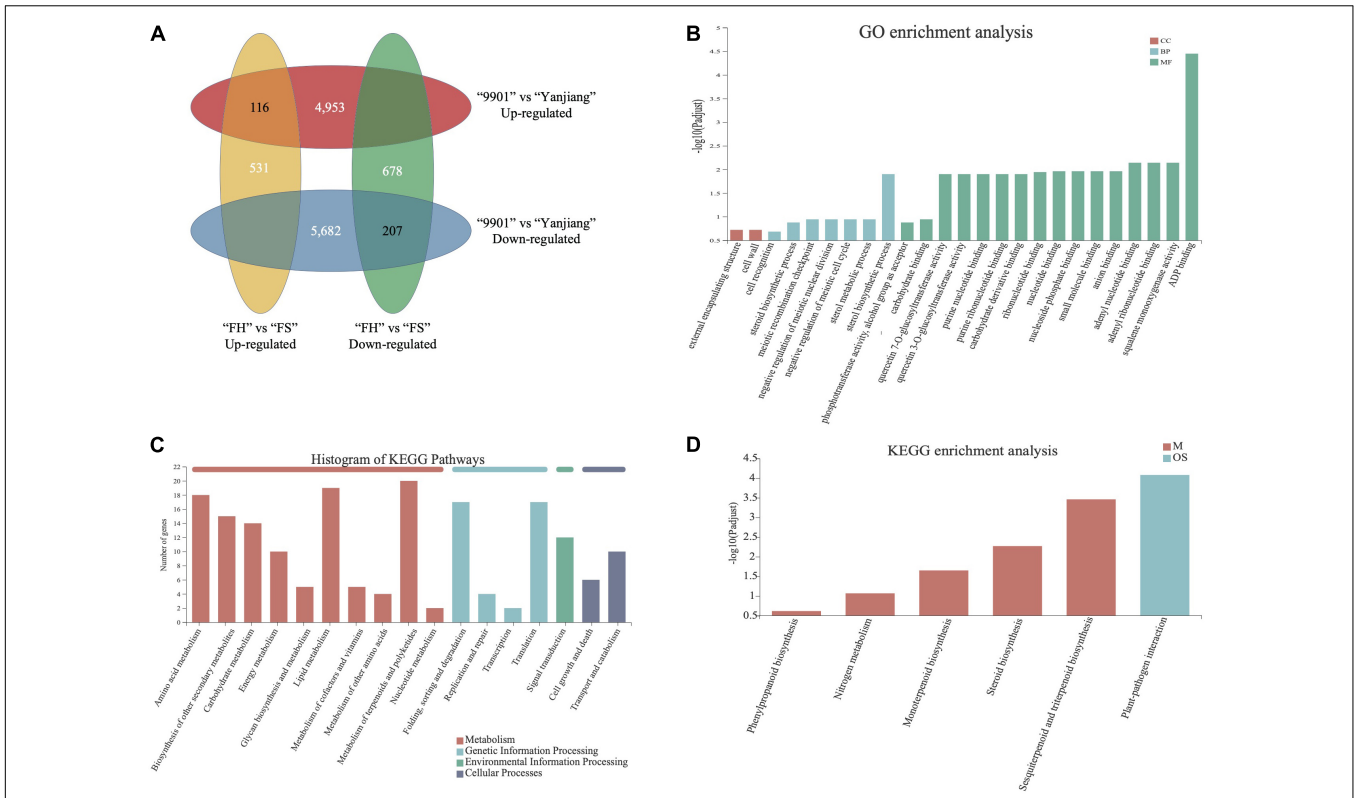


FIGURE 4 | GO and KEGG analysis of differentially expressed genes (DEGs). **(A)** Venn analysis of DEGs. **(B)** GO enrichment analysis of 323 DEGs. **(C)** Statistics analysis of KEGG pathway among the 323 DEGs. **(D)** KEGG enrichment analysis of 323 DEGs.

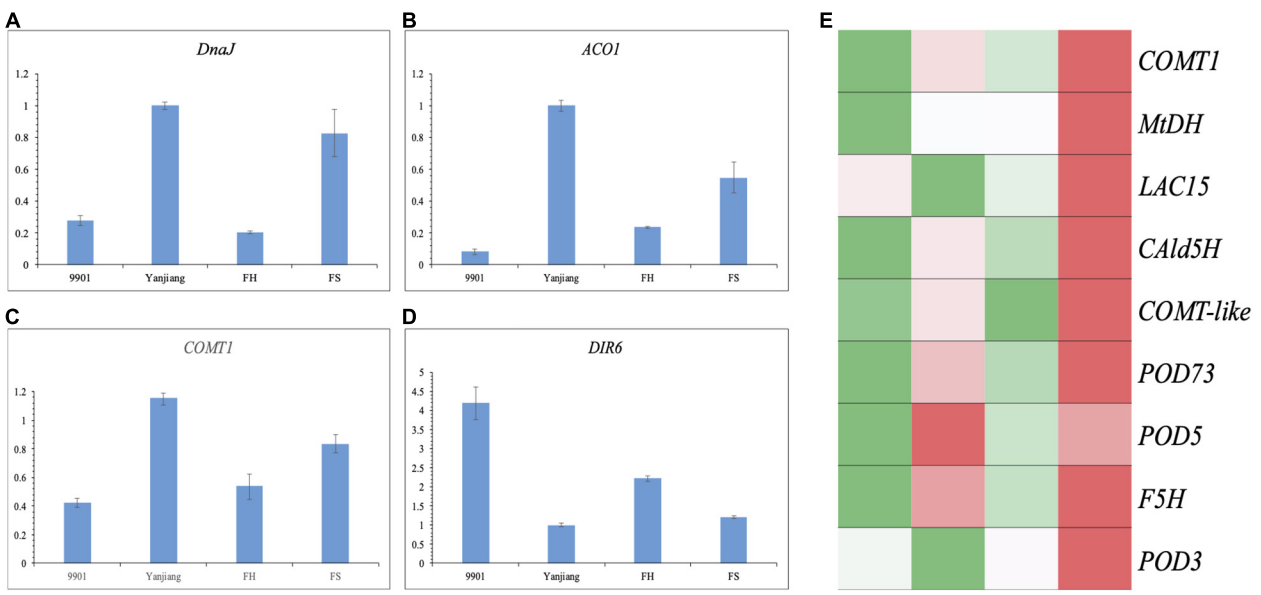


FIGURE 5 | The expression trend of candidate genes and lignin biosynthesis related genes in the two parents and two progenies. **(A–D)** qRT–PCR analysis of the transcript levels of candidate genes in the two parents and two progenies. The Y-axis represents the relative expression level, and the X-axis represents different cultivars. The error bars denote the standard error (SE). **(E)** Heatmap analysis of lignin biosynthesis related genes in the two parents and two progenies.

TH, including genes encoding DIR6 and plastid-lipid-associated protein, among others.

According to the annotations, four genes were selected as possible candidate genes to perform a qRT-PCR analysis. The *DnaJ* gene for *qTH-A15-1*, the *ACO1* gene for *qTH-A11-1,* and the *COMT1* gene for *qTH-A16-1* showed higher levels of expression in the stem terminals of "Yanjiang" and "FS" than those of "9901" and "FH". The result indicates that the expression of these genes may suppress the TH growth (**Figures 5A–C**). Another gene encoding *DIR6* for *qTH-A02-1* was expressed at a higher level in "9901" and "FH" than in "Yanjiang" and "FS" (**Figure 5D**), displaying a positive relationship with TH. Interestingly, both *DIR6* and *COMT1* were associated with phenylpropanoid or lignin biosynthesis. However, these two genes showed an opposite expression trend. Another candidate gene, *ACO1,* has already been verified to regulate the biosynthesis of ethylene in *Arabidopsis, Popolus,* and *Solanum*, among other plants. Previous studies also found that ethylene could affect the biosynthesis of lignin through activating the downstream transcription factor (EIN and ERF/AP2, among others) or other proteins. We then verified the genes on the pathway of lignin biosynthesis from the 323 DEGs, according to the annotation. As a result, nine genes were identified, annotated as mannitol dehydrogenase (MtDH), ferulic acid 5-hydroxylase (F5H), coniferyl aldehyde 5-hydroxylase 2 (CAld5H), laccase (LAC), and peroxidase (POD), among others. Most of them were expressed higher in "Yanjiang" and "FS" than in "9901" and "FH" (**Figure 5E**), which is consistent with the expression of *ACO1* and *COMT1*. Lignin is known to take part in cell wall formation, and so the regulation of lignin biosynthesis may affect the development of cells and bring the stage of secondary growth forward.

DISCUSSION

TH and DBH in *Salix matsudana* Koidz

Salix is known for its versatile use in industries (papermaking, gunpowder, and particleboard, among others) and for ecological purposes, such as afforestation in the city and coastal beachlands (Zhang et al., 2017). The physiological and biochemical traits on *Salix*, such as nitrogen economy, leaf senescence, bud burst, enzymolysis saccharify, salicin, and insect resistance, have already been studied (Sulima et al., 2009; Brereton et al., 2010; Höglund et al., 2012; Berlin et al., 2014; Ghelardini et al., 2014). Both *Salix* and *Populus* belong to the family of *Salicaceae*. Many researchers have already studied the wood growth of *Populus* (Dubois et al., 2018). However, there remains a lack of information on the wood growth of *Salix matsudana* Koidz. *Salix matsudana* Koidz. is a tetraploid forest tree and has a much more complex genetic mechanism than other diploid forest trees. It is also an ideal model system for studying plant polyploidization (Zhang et al., 2020). In this study, TH and DBH were measured in the F_1 population during the fourth and fifth year after seeding and 8-month- old cuttings. However, few differences were identified on DBH for these cuttings. To avoid the environmental effect, HPY and DPY were also measured and showed a significant correlation with TH and DBH. While calculating the heritability, for the lack of replicates in each environment, only the heritability of HPY and DPY were measured. Both traits showed high heritability. It is understandable that a bigger DBH could provide stronger mechanical support for trees and result in a taller TH. However, it was found that DBH and TH are determined by different growth patterns. DBH is mainly determined by secondary growth, which includes secondary xylem and phloem thickening, cell anticlinal division, and cell wall thickening (Chaffey et al., 2002; Helariutta and Bhalerao, 2003). TH is mainly determined by stem apical meristem (SAM) cell growth and division in primary growth. Stem cells in the central region of SAM produce various types of vascular cells through continuous division, which in turn promote primary growth (Altamura et al., 2001; Little et al., 2002; Ye et al., 2002). The relationship between TH and DBH is also determined by primary growth and secondary growth. To further understand the relationship between TH and DBH and discover the genetic mechanism difference on TH and DBH, we first analyzed the genetic mechanism on TH and DBH by combining the PCA. A total of seven QTL regions that could affect both TH and DBH were determined, which could explain the high positive correlation between TH and DBH. The QTL mapping result of the PCA is highly consistent with the multi-effect QTL, which indicates that the PCA is able to understand positively correlated traits (Yano et al., 2019). However, there still remains the question of whether multi-effect genes are located in these seven QTL regions or TH-related genes and DBH-related genes are located closely on the genome. To solve this question, more experimental data and a finer mapping of TH and DBH are needed.

QTL of Fast-Growing and Recombination Hotspots in *Salix matsudana* Koidz

Based on the reference genome of "Yanjiang," we re-analyzed the genetic map of the F_1 population. Only the SNPs that could be mapped onto the chromosomes of the genome were selected to construct the genetic map. According to the phenotypes of fast-growing traits, we identified 21 QTL, including 10 DBH QTL and 11 TH QTL. For these QTLs, the PV were ranged from 5.8 to 10%, which is not treated as major QTL in most crops. It might be because the population of *Salix matsudana* Koidz is an F_1 population, which has a lower additive genetic variance than F_2 or RIL populations. The PV of this population is similar to *Populus* F_1 populations previously reported (Mousavi et al., 2016; Liu et al., 2017).

Additionally, six multi-effect QTL regions were identified; the At and Bt sub-genomes contained the same numbers of multi-effect QTL regions. Except for these six multi-effect QTL, three DBH QTL were located on the Bt sub-genome, while only one DBH QTL was located on the At sub-genome. For TH, the At and Bt sub-genomes contained three and two QTL, respectively. We also compared whether QTL is located on homoeologous chromosomes. As a result, only A02 and B02, A10 and B10, A13 and B13, and A16 and B16 showed partial

consistency on QTL distribution. While both sub-genomes contained several stable QTL on TH and DBH, the At sub-genome contained more TH QTL and more DBH QTL were detected on the Bt sub-genome. The results indicated that some DBH QTL on the At sub-genome and some TH QTL on the Bt sub-genome were lost under the evolution and selection. We first compared the difference in QTL distribution between the two sub-genomes in *Salix matsudana* Koidz. Although more genetic and natural populations are needed to confirm the differentiation of the sub-genomes, this study proved that the At and Bt genomes were different under artificial selection and natural selection.

Based on the high-density genetic map and the reference genome, we identified 21 recombination hotspots on 17 chromosomes. Previous studies have reported that recombination in *Arabidopsis* (Osman et al., 2011), maize (Schnable et al., 2009), and *Triticum aestivum* (Arbeithuber et al., 2015) usually occurs in intergenic regions and around transposons. This study identified the genome-wide recombination rates, which is helpful to understand the genetic variation. Additionally, lots of segment insertions were detected to be distributed on most chromosomes in this population. The most likely reason for this is the genotypic difference between the two parents.

A Potential Functional Network of Genes Associated With TH

According to the transcriptomes, 10,635 and 1,209 DEGs were identified between the two parents and two progenies. "Yanjiang" and "9901" were different varieties while "FH" and "FS" contained similar genetic backgrounds. By overlapping the two DEGs, we narrowed the DEGs into 323 genes. The present study identified four DEGs, co-localized with TH QTL, that may be associated with TH growth. The *ACO1* gene located in *qTH-A11-1* was annotated to directly regulate the biosynthesis of ethylene. Several studies have already reported that ethylene could affect cell expansion and the activity of cambium (Dubois et al., 2018). Many ethylene biosynthesis-related enzymes and transcription factors in *Populus* have been proven to express wood tissue (Seyfferth et al., 2018). Other studies have shown that ethylene participates in cell growth and differentiation in cambium and affects the secondary growth of stems (Pesquet and Tuominen, 2011; Felten et al., 2018; Harkey et al., 2019). The content of ethylene could regulate the cell division in meristems and affect the formation of tension wood in *Populus* (Andersson-Gunneras et al., 2003; Love et al., 2009). The up-regulation of the *ACO1* gene could promote ethylene production and induce cotton fiber cell elongation (Shi et al., 2006). According to the expression pattern of *ACO1* in the two parents and two progenies, the high expression level of *ACO1* may reduce the TH growth of *Salix matsudana* Koidz. by regulating the ethylene biosynthesis. A gene encoding the DnaJ protein was detected to be differentially expressed in *qTH-A15-1*. Previously, *DnaJ* was mostly identified as a response to heat stress or other environmental stimulation (Orme et al., 2001; Walsh et al., 2004; Liberek et al., 2008). Our results showed that the *DnaJ* gene may negatively regulate TH growth, due to the high expression of this gene in "Yanjiang" and "FS". In yeast, the DnaJ protein was reported to regulate the formation of tubulin (Oka et al., 1998), which could determine the direction of cell expansion. Interestingly, 17 of the 18 candidate genes were located in multi-effect QTLs, indicating that these genes might also regulate DBH. For most of the single TH-related chromosomes, we did not detect many DEGs. This may be because the genetic background of "FH" and "FS" were similar and other loci on chromosomes, such as miRNAs and long non-coding RNAs, may also regulate TH. We also identified several lignin-related genes that were differentially expressed. Moreover, two of them could co-localize in stable TH QTL, as seen in *COMT1* in *qTH-A16-1* and *DIR6* in *qTH-A02-1*. *COMT1* was annotated as a catalyst phenylalanine for coniferyl alcohol (CA) in a phenylalanine pathway, which was the first step to lignin biosynthesis. DIR could then catalyze CA into secoisolariciresinol (SECO) rather than normal lignans (Boerjan et al., 2003; Mazzaferro et al., 2015; Xiao et al., 2015). Our results showed that the down-regulation of *COMT1* and up-regulation of *DIR6* could result in a low expression level of normal lignin. Our study suggests that the reduction of lignin biosynthesis might promote TH growth. We then measured the expression level of other phenylalanine pathway genes. Most lignin biosynthesis genes (*Mannitol dehydrogenase, F5H, peroxidase, etc.*) slowed a low expression level in "9901" and "FH", which is consistent with the hypothesis. Although further experiments, such as overexpression or CRISPR-Cas9 silencing, are needed to confirm roles in TH, the present QTL analysis, comparative transcriptome analysis, and qRT-PCR analysis showed that these several lignig-related genes may be important for TH and DBH.

CONCLUSION

In our study, two varieties were chosen as two parents to construct F_1 populations to detect fast-growing related (TH and DBH) QTL and genes. Based on the high-density genetic map, five DBH- and eight TH-related stable QTL were identified. A PCA was undertaken to confirm the multi-function stable QTL. The further use of our previous RNA-seq data for the two parents and two F_1 progenies, followed by the qRT-PCR analysis, detected four candidate genes that may regulate TH. Several lignin biosynthesis-related genes were found to be differentially expressed. Furthermore, 21 recombination hotspots were identified in our population. These results provide an important foundation for further studies on the molecular and genetic regulation of the fast-growing traits of forest trees.

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: "CNGB Sequence Archive (CNSA) of China National GeneBank DataBase (CNGBdb) (https://db.cngb.org/cnsa/) with accession number CNP0001576."

AUTHOR CONTRIBUTIONS

JZ and GL conceived and designed the experiments. GL, JG, YW, ZF, JH, HZ, and XZ performed the experiments. GL, YC, CY, BL, and FZ analyzed the data. GL, QY, and JZ wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

The research was supported by grants from the National Natural Science Foundation of China (31971681), Natural Science Foundation of Jiangsu Province (BK20200963), Science and Technology Plan of Nantong City (JC2020157), Nantong University Scientific Research Start-up project for Introducing Talents (135419609070), and the Jiangsu Provincial Key Projects of Students Innovation and Entrepreneurship Training Program (202001030402OZ).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.596749/full#supplementary-material

REFERENCES

Andersson-Gunneras, S., Hellgren, J. M., Bjorklund, S., Regan, S., Moritz, T., and Sundberg, B. (2003). Asymmetric expression of a poplar ACC oxidase controls ethylene production during gravitational induction of tension wood. *Plant J.* 34, 339–349. doi: 10.1046/j.1365-313x.2003.01727.x

Altamura, M. M., Possenti, M., Matteucci, A., Baima, S., Ruberti, I., and Morelli, G. (2001). Development of the vascular system in the inflorescence stem of *Arabidopsis*. *New Phytol.* 151, 381–389. doi: 10.1046/j.0028-646x.2001.00188.x

Arbeithuber, B., Betancourt, A. J., Ebner, T., and Tiemann-Boege, I. (2015). Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proc. Natl. Acad. Sci. U.S.A.* 112, 2109–2114.

Berlin, S., Ghelardini, L., Bonosi, L., Weih, M., and Rönnberg-Wästljung, A. C. (2014). QTL mapping of biomass and nitrogen economy traits in willows (*Salix* spp.) grown under contrasting water and nutrient conditions. *Mol. Breed.* 34, 1987–2003. doi: 10.1007/s11032-014-0157-5

Boerjan, W., Ralph, J., and Baucher, M. (2003). Lignin biosynthesis. *Annu. Rev. Plant Biol.* 54, 519–546. doi: 10.1146/annurev.arplant.54.031902.134938

Bokore, F., Cuthbert, R., Knox, R., Singh, A., Campbell, H., Pozniak, C., et al. (2019). Mapping quantitative trait loci associated with common bunt resistance in a spring wheat (*Triticum aestivum* L.) variety Lillian. *TAG Theor. Appl. Genet.* 132, 3023–3033. doi: 10.1007/s00122-019-03403-3

Brand, U., Fletcher, J. C., Hobe, M., Meyerowitz, E. M., and Simon, R. (2000). Dependence of stem cell fate in *Arabidopsis* on a feedback loop regulated by CLV3 activity. *Science* 289, 617–619. doi: 10.1126/science.289.5479.617

Brereton, N. J. B., Pitre, F. E., Hanley, S. J., Ray, M. J., Karp, A., and Murphy, R. J. (2010). QTL mapping of enzymatic saccharification in short rotation coppice willow and its independence from biomass yield. *Bioenergy Res.* 3, 251–261. doi: 10.1007/s12155-010-9077-3

Chaffey, N., Cholewa, E., Regan, S., and Sundberg, B. (2002). Secondary xylem development in *Arabidopsis: a model for wood formation*. *Physiol. Plant.* 114, 594–600. doi: 10.1034/j.1399-3054.2002.1140413.x

Chen, Y., Jiang, Y., Chen, Y., Feng, W., Liu, G., Yu, C., et al. (2020). Uncovering candidate genes responsive to salt stress in *Salix matsudana* (Koidz) by transcriptomic analysis. *PLoS One* 15:e0236129. doi: 10.1371/journal.pone.0236129

Du, J., Mansfield, S. D., and Groover, A. T. (2009). The *Populus* homeobox gene ARBORKNOX2 regulates cell differentiation during secondary growth. *Plant J.* 60, 1000–1014. doi: 10.1111/j.1365-313X.2009.04017.x

Dubois, M., van den Broeck, L., and Inze, D. (2018). The pivotal role of ethylene in plant growth. *Trends Plant Sci.* 23, 311–323. doi: 10.1016/j.tplants.2018.01.003

Felten, J., Vahala, J., Love, J., Gorzsas, A., Ruggeberg, M., Delhomme, N., et al. (2018). Ethylene signaling induces gelatinous layers with typical features of tension wood in hybrid aspen. *New Phytol.* 218, 999–1014. doi: 10.1111/nph.15078

Fisher, K., and Turner, S. (2007). PXY, a receptor-like kinase essential for maintaining polarity during plant vascular-tissue development. *Curr. Biol.* 17, 1061–1066. doi: 10.1016/j.cub.2007.05.049

Floyd, S. K., Zalewski, C. S., and Bowman, J. L. (2006). Evolution of class III homeodomain-leucine zipper genes in streptophytes. *Genetics* 173, 373–388. doi: 10.1534/genetics.105.054239

Fukuda, H., Hirakawa, Y., and Sawa, S. (2007). Peptide signaling in vascular development. *Curr. Opin. Plant Biol.* 10, 477–482. doi: 10.1016/j.pbi.2007.08.013

Ghelardini, L., Berlin, S., Weih, M., Lagercrantz, U., Gyllenstrand, N., and Ronnberg-Wastljung, A. C. (2014). Genetic architecture of spring and autumn phenology in *Salix*. *BMC Plant Biol.* 14:31. doi: 10.1186/1471-2229-14-31

Groover, A. T., Mansfield, S. D., DiFazio, S. P., Dupper, G., Fontana, J. R., Millar, R., et al. (2006). The *Populus* homeobox gene ARBORKNOX1 reveals overlapping mechanisms regulating the shoot apical meristem and the vascular cambium. *Plant Mol. Biol.* 61, 917–932. doi: 10.1007/s11103-006-0059-y

Gross-Hardt, R., and Laux, T. (2003). Stem cell regulation in the shoot meristem. *J. Cell Sci.* 116(Pt 9), 1659–1666. doi: 10.1242/jcs.00406

Harkey, A. F., Yoon, G. M., Seo, D. H., DeLong, A., and Muday, G. K. (2019). Light modulates ethylene synthesis, signaling, and downstream transcriptional networks to control plant development. *Front. Plant Sci.* 10:1094. doi: 10.3389/fpls.2019.01094

Helariutta, Y., and Bhalerao, R. (2003). Between xylem and phloem: the genetic control of cambial activity in plants. *Plant Biol.* 5, 465–472. doi: 10.1055/s-2003-44780

Hirakawa, Y., Shinohara, H., Kondo, Y., Inoue, A., Nakanomyo, I., Ogawa, M., et al. (2008). Non-cell-autonomous control of vascular stem cell fate by a CLE peptide/receptor system. *Proc. Natl. Acad. Sci. U.S.A.* 105, 15208–15213. doi: 10.1073/pnas.0808444105

Höglund, S., Rönnberg-Wästljung, A. C., Lagercrantz, U., and Larsson, S. (2012). A rare major plant QTL determines non-responsiveness to a gall-forming insect in willow. *Tree Genet. Genomes* 8, 1051–1060. doi: 10.1007/s11295-012-0485-8

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

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, X., Wu, M., Liu, G., Pei, W., and Yu, S. (2017). Identification of candidate genes for fiber length quantitative trait loci through rna-seq and linkage and physical mapping in cotton. *BMC Genomics* 18:427. doi: 10.1186/s12864-017-3812-5

Liberek, K., Lewandowska, A., and Zietkiewicz, S. (2008). Chaperones in control of protein disaggregation. *EMBO J.* 27, 328–335. doi: 10.1038/sj.emboj.7601970

Little, C., Macdonald, J., and Olsson, O. (2002). Involvement of indole-3-acetic acid in fascicular and interfascicular cambial growth and interfascicular extraxylary fiber differentiation in *Arabidopsis thaliana* inflorescence stems. *Int. J. Plant Sci.* 163, 519–529. doi: 10.1086/339642

Liu, F., Tong, C., Tao, S., Wu, J., Chen, Y., Yao, D., et al. (2017). Mvqtlcim: composite interval mapping of multivariate traits in a hybrid f 1 population of outbred species. *BMC Bioinformatics* 18:515. doi: 10.1186/s12859-017-1908-1

Liu, G., Pei, W., Li, D., Ma, J., Cui, Y., Wang, N., et al. (2019). A targeted QTL analysis for fiber length using a genetic population between two introgressed

backcrossed inbred lines in upland cotton (*Gossypium hirsutum*). *Crop J.* 7, 273–282. doi: 10.1016/j.cj.2018.11.005

Love, J., Björklund, S., Vahala, J., Hertzberg, M., Kangasjärvi, J., and Sundberg, B. (2009). Ethylene is an endogenous stimulator of cell division in the cambial meristem of *Populus*. *Proc. Natl. Acad. Sci. U.S.A.* 106, 5984–5989. doi: 10.1073/pnas.0811660106

Mazzaferro, L. S., Huttel, W., Fries, A., and Muller, M. (2015). Cytochrome P450-catalyzed regio- and stereoselective phenol coupling of fungal natural products. *J. Am. Chem. Soc.* 137, 12289–12295. doi: 10.1021/jacs.5b06776

McCouch, S. R., Chen, X. L., Panaud, O., Temnykh, S., Xu, Y. B., Cho, Y. G., et al. (1997). Microsatellite marker development, mapping and applications in rice genetics and breeding. *Plant Mol. Biol.* 35, 89–99.

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

Mousavi, M., Tong, C., Liu, F., Tao, S., Wu, J., Li, H., et al. (2016). De novo snp discovery and genetic linkage mapping in poplar using restriction site associated DNA and whole-genome sequencing technologies. *BMC Genomics* 17:656. doi: 10.1186/s12864-016-3003-9

Murmanis, L. (1970). Locating the initial in the vascular cambium of *Pinus strobus* l. by electron microscopy. *Wood Sci. Technol.* 4, 1–14. doi: 10.1007/BF00356233

Oka, M., Nakai, M., Endo, T., Lim, C. R., Kimata, Y., and Kohno, K. (1998). Loss of Hsp70-Hsp40 chaperone activity causes abnormal nuclear distribution and aberrant microtubule formation in M-phase of *Saccharomyces cerevisiae*. *J. Biol. Chem.* 273, 29727–29737. doi: 10.1074/jbc.273.45.29727

Orme, W., Walker, A. R., Gupta, R., and Gray, J. C. (2001). A novel plastid-targeted J-domain protein in *Arabidopsis thaliana*. *Plant Mol. Biol.* 46, 615–626. doi: 10.1023/a:1010665702621

Osman, K., Higgins, J. D., Sanchez-Moran, E., Armstrong, S. J., and Franklin, F. C. H. (2011). Pathways to meiotic recombination in *Arabidopsis thaliana*. *New Phytol.* 190, 523–544. doi: 10.1111/j.1469-8137.2011.03665.x

Parker, G., Schofield, R., Sundberg, B., and Turner, S. (2003). Isolation of COV1, a gene involved in the regulation of vascular patterning in the stem of *Arabidopsis*. *Development* 130, 2139–2148. doi: 10.1242/dev.00441

Pertea, M., Pertea, G., Antonescu, C., Chang, T., Mendell, J., and Salzberg, S. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295. doi: 10.1038/nbt.3122

Pesquet, E., and Tuominen, H. (2011). Ethylene stimulates tracheary element differentiation in *Zinnia elegans* cell cultures. *New Phytol.* 190, 138–149. doi: 10.1111/j.1469-8137.2010.03600.x

Pineau, C., Freydier, A., Ranocha, P., Jauneau, A., Turner, S., Lemonnier, G., et al. (2005). hca: an *Arabidopsis* mutant exhibiting unusual cambial activity and altered vascular patterning. *Plant J.* 44, 271–289. doi: 10.1111/j.1365-313X.2005.02526.x

Rezvoy, C., Charif, D., Gueguen, L., and Marais, G. A. B. (2007). MareyMap: an R-based tool with graphical interface for estimating recombination rates. *Bioinformatics* 23, 2188–2189. doi: 10.1093/bioinformatics/btm315

Rose, R. J. (2016). *Molecular Cell Biology of the Growth and Differentiation of Plant Cells*. Boca Raton, FL: CRC Press.

Sarkar, A. K., Luijten, M., Miyashima, S., Lenhard, M., Hashimoto, T., Nakajima, K., et al. (2007). Conserved factors regulate signalling in *Arabidopsis thaliana* shoot and root stem cell organizers. *Nature* 446, 811–814. doi: 10.1038/nature05703

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F. S., Pasternak, S., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science* 326, 1112–1115. doi: 10.1126/science.1178534

Schrader, J., Nilsson, J., Mellerowicz, E., Berglund, A., Nilsson, P., Hertzberg, M., et al. (2004). A high-resolution transcript profile across the wood-forming meristem of poplar identifies potential regulators of cambial stem cell identity. *Plant Cell* 16, 2278–2292. doi: 10.1105/tpc.104.024190

Seyfferth, C., Wessels, B., Jokipii-Lukkari, S., Sundberg, B., Delhomme, N., Felten, J., et al. (2018). Ethylene-Related gene expression networks in wood formation. *Front. Plant Sci.* 9:272. doi: 10.3389/fpls.2018.00272

Shi, Y. H., Zhu, S. W., Mao, X. Z., Feng, J. X., Qin, Y. M., Zhang, L., et al. (2006). Transcriptome profiling, molecular biological, and physiological studies reveal a major role for ethylene in cotton fiber cell elongation. *Plant Cell* 18, 651–664. doi: 10.1105/tpc.105.040303

Sulima, P., Przyborowski, J. A., and Zauski, D. (2009). RAPD markers reveal genetic diversity in *Salix purpurea* l. *Crop Sci.* 49, 857–863. doi: 10.2135/cropsci2008.07.0397

van Ooijen, J. (2006). *JoinMap® 4, Software for the Calculation of Genetic Linkage Maps in Experimental Populations*, Vol. 33. Wageningen: Kyazma BV.

Walsh, P., Bursac, D., Law, Y. C., Cyr, D., and Lithgow, T. (2004). The J-protein family: modulating protein assembly, disassembly and translocation. *EMBO Rep.* 5, 567–571. doi: 10.1038/sj.embor.7400172

Xiao, Y., Ji, Q., Gao, S., Tan, H., Chen, R., Li, Q., et al. (2015). Combined transcriptome and metabolite profiling reveals that IiPLR1 plays an important role in lariciresinol accumulation in *Isatis indigotica*. *J. Exp. Bot.* 66, 6259–6271. doi: 10.1093/jxb/erv333

Yano, K., Morinaka, Y., Wang, F., Huang, P., Takehara, S., Hirai, T., et al. (2019). GWAS with principal component analysis identifies a gene comprehensively controlling rice architecture. *Proc. Natl. Acad. Sci. U.S.A.* 116, 21262–21267. doi: 10.1073/pnas.1904964116

Ye, Z. H., Freshour, G., Hahn, M. G., Burk, D. H., and Zhong, R. (2002). Vascular development in *Arabidopsis*. *Int. Rev. Cytol. A Survey Cell Biol.* 220, 225–256. doi: 10.1016/S0074-7696(02)20007-8

Young, M., Wakefield, M., Smyth, G., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* 11:R14. doi: 10.1186/gb-2010-11-2-r14

Zhang, J., Yuan, H., Li, M., Li, Y., Wang, Y., Ma, X., et al. (2016). A high-density genetic map of tetraploid *Salix matsudana* using specific length amplified fragment sequencing (SLAF-seq). *PLoS One* 11:e0157777. doi: 10.1371/journal.pone.0157777

Zhang, J., Yuan, H., Li, Y., Chen, Y., Liu, G., Ye, M., et al. (2020). Genome sequencing and phylogenetic analysis of allotetraploid *Salix matsudana* Koidz. *Hortic. Res.* 7:201.

Zhang, J., Yuan, H., Yang, Q., Li, M., Wang, Y., Li, Y., et al. (2017). The genetic architecture of growth traits in *Salix matsudana* under salt stress. *Hortic. Res.* 4:17024. doi: 10.1038/hortres.2017.24

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.

Copyright © 2021 Liu, Yang, Gao, Wu, Feng, Huang, Zou, Zhu, Chen, Yu, Lian, Zhong and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership