EVOLUTION OF CROP GENOMES AND EPIGENOMES

EDITED BY: Hai Du, Viktor Demko, Zhe Liang and Wei Hu

PUBLISHED IN: Frontiers in Plant Science

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-83250-414-7
DOI 10.3389/978-2-83250-414-7

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

EVOLUTION OF CROP GENOMES AND EPIGENOMES

Topic Editors:
Hai Du, Southwest University, China
Viktor Demko, Comenius University, Slovakia
Zhe Liang, Biotechnology Research Institute, Chinese Academy of Agricultural Sciences, China
Wei Hu, Institute of Tropical Bioscience and Biotechnology, Chinese Academy of Tropical Agricultural Sciences, China

Citation: Du, H., Demko, V., Liang, Z., Hu, W., eds. (2022). Evolution of Crop Genomes and Epigenomes. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-83250-414-7

Table of Contents

05 ***Editorial: Evolution of Crop Genomes and Epigenomes***
Hai Du, Wei Hu, Viktor Demko and Zhe Liang

08 ***Variation and Evolution of the Whole Chloroplast Genomes of* Fragaria *spp. (Rosaceae)***
Chenxin Li, Chaonan Cai, Yutian Tao, Zhongshuai Sun, Ming Jiang, Luxi Chen and Junmin Li

22 ***Family-Wide Evaluation of Multiple C2 Domain and Transmembrane Region Protein in* Gossypium hirsutum**
Qianqian Hu, Mengting Zeng, Miao Wang, Xiaoyu Huang, Jiayi Li, Changhui Feng, Lijie Xuan, Lu Liu and Gengqing Huang

37 ***Silencing of* GhKEA4 *and* GhKEA12 *Revealed Their Potential Functions Under Salt and Potassium Stresses in Upland Cotton***
Yi Li, Zhen Feng, Hengling Wei, Shuaishuai Cheng, Pengbo Hao, Shuxun Yu and Hantao Wang

57 ***The Complete Chloroplast Genome Sequences of Eight* Fagopyrum *Species: Insights Into Genome Evolution and Phylogenetic Relationships***
Yu Fan, Ya'nan Jin, Mengqi Ding, Yu Tang, Jianping Cheng, Kaixuan Zhang and Meiliang Zhou

74 ***Chloroplast Phylogenomic Analyses Reveal a Maternal Hybridization Event Leading to the Formation of Cultivated Peanuts***
Xiangyu Tian, Luye Shi, Jia Guo, Liuyang Fu, Pei Du, Bingyan Huang, Yue Wu, Xinyou Zhang and Zhenlong Wang

87 **TALE *Transcription Factors in Sweet Orange* (Citrus sinensis*): Genome-Wide Identification, Characterization, and Expression in Response to Biotic and Abiotic Stresses***
Weiye Peng, Yang Yang, Jing Xu, Erping Peng, Suming Dai, Liangying Dai, Yunsheng Wang, Tuyong Yi, Bing Wang, Dazhi Li and Na Song

103 ***Phylogenetic Analysis of the SQUAMOSA Promoter-Binding Protein-Like Genes in Four* Ipomoea *Species and Expression Profiling of the* IbSPLs *During Storage Root Development in Sweet Potato (*Ipomoea batatas*)***
Haoyun Sun, Jingzhao Mei, Weiwei Zhao, Wenqian Hou, Yang Zhang, Tao Xu, Shaoyuan Wu and Lei Zhang

122 ***Comparative Analysis the Complete Chloroplast Genomes of Nine* Musa *Species: Genomic Features, Comparative Analysis, and Phylogenetic Implications***
Weicai Song, Chuxuan Ji, Zimeng Chen, Haohong Cai, Xiaomeng Wu, Chao Shi and Shuo Wang

137 ***Genome-Wide Characterization of Serine/Arginine-Rich Gene Family and Its Genetic Effects on Agronomic Traits of* Brassica napus**
Meili Xie, Rong Zuo, Zetao Bai, Lingli Yang, Chuanji Zhao, Feng Gao, Xiaohui Cheng, Junyan Huang, Yueying Liu, Yang Li, Chaobo Tong and Shengyi Liu

158 ***Genome-Wide Association Mapping of Hulless Barely Phenotypes in Drought Environment***
Jie Li, Xiaohua Yao, Youhua Yao, Likun An, Zongyun Feng and Kunlun Wu

169 *Cassava (*Manihot esculenta*) Slow Anion Channel (*MeSLAH4*) Gene Overexpression Enhances Nitrogen Assimilation, Growth, and Yield in Rice*
Linhu Song, Xingmei Wang, Liangping Zou, Zakaria Prodhan, Jiaheng Yang, Jianping Yang, Li Ji, Guanhui Li, Runcong Zhang, Changyu Wang, Shi Li, Yan Zhang, Xiang Ji, Xu Zheng, Wanchen Li and Zhiyong Zhang

184 *Transcriptome Analysis of Moso Bamboo (*Phyllostachys edulis*) Reveals Candidate Genes Involved in Response to Dehydration and Cold Stresses*
Zhuo Huang, Peilei Zhu, Xiaojuan Zhong, Jiarui Qiu, Wenxin Xu and Li Song

frontiers | Frontiers in Plant Science

TYPE Editorial
PUBLISHED 22 September 2022
DOI 10.3389/fpls.2022.1027698

OPEN ACCESS

EDITED AND REVIEWED BY
Jim Leebens-Mack,
University of Georgia, United States

*CORRESPONDENCE
Hai Du
haidu81@126.com;
dh20130904@swu.edu.cn

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

RECEIVED 25 August 2022
ACCEPTED 08 September 2022
PUBLISHED 22 September 2022

CITATION
Du H, Hu W, Demko V and Liang Z (2022) Editorial: Evolution of crop genomes and epigenomes. *Front. Plant Sci.* 13:1027698. doi: 10.3389/fpls.2022.1027698

Editorial: Evolution of crop genomes and epigenomes

Hai Du[1,2,3]*, Wei Hu[4], Viktor Demko[5,6] and Zhe Liang[7]

[1]College of Agronomy and Biotechnology, Chongqing Engineering Research Center for Rapeseed, Southwest University, Chongqing, China, [2]Integrative Science Center of Germplasm Creation in Western China (CHONGQING) Science City and Southwest University, College of Agronomy and Biotechnology, Southwest University, Chongqing, China, [3]Engineering Research Center of South Upland Agriculture, Ministry of Education, Chongqing, China, [4]Institute of Tropical Bioscience and Biotechnology, Chinese Academy of Tropical Agricultural Sciences, Haikou, China, [5]Faculty of Natural Sciences, Comenius University in Bratislava, Bratislava, Slovakia, [6]Plant Science and Biodiversity Center, Slovak Academy of Sciences, Bratislava, Slovakia, [7]Biotechnology Research Institute, Chinese Academy of Agricultural Sciences, Beijing, China

KEYWORDS

multi-omics, crops, genome, epigenome, evolution, epigenomics

Editorial on the Research Topic
Evolution of crop genomes and epigenomes

In the last two decades, the advances in large-scale sequencing technology (e.g., the second- and third-generation sequencing technologies) have generated increasingly available omics datasets (e.g. genomics, chloroplast, epigenomics, transcriptomics) of plant organisms, offering valuable resource to address the various biological questions more effectively. Available genome resources allow us to uncover yet hidden mysteries of plant genomes and their evolutionary stories at genome-wide level and to screen for agronomically significant genes in crops. In this Research Topic, we aim to present novel and important findings on all aspect of genome and epigenetics in plants, especially crops. This topic includes 12 original research papers focusing on the research areas highlighted above, viewed more than 15,263 times by the time of this Editorial. We organize this collection of studies into two major groups based on common themes as described below.

Genome-wide evolution analysis and gene resource screen

Hu et al. present a systematic identification and evolutionary analysis of the multiple C2 domain and transmembrane region proteins (MCTPs) in upland cotton, and also conducted a phylogenetic analysis with the homologs in another 17 plant species including algae, fern, moss, monocotyledons and dicotyledons. Based on global expression analyses using transcriptome dataset and qRT-PCR assay, they identify

three candidate genes (*GhMCTP7*, *GhMCTP12* and *GhMCTP17*) and demonstrate the interaction of GhMCTP7/12/17 with GhKNAT1/2 proteins to regulate cotton shoot meristem development in an integrated multiple signal pathway manner.

Sun et al. conducted a genome-wide study on the plant-specific SQUAMOSA promoter-binding protein-like (SPL) transcription factor family in four *Ipomoea* species, including gene and protein structure characteristics, gene duplication, selective pressure, expansion pattern, spatiotemporal and exogenous phytohormone induction expression profiles, etc. This study revealed that segmental duplication is the main driving force for gene expansion in *Ipomoea* species. The data suggest that most of the *Ipomoea SPL* genes are miR156 targets with seven miR156-SPL interaction relationships were verified by degradome sequencing. They finally identify four *SPL* genes with putative function in sweet potato storage root development.

Li et al. identified and analyzed the K^+ efflux antiporter (KEA) genes in four cotton species and seven other plants (*Arabidopsis thaliana*, *Oryza sativa*, *Zea mays*, *Populus trichocarpa*, *Sorghum bicolor*, *Triticum aestivum* and *Glycine max*) at genome-wide level. The candidate genes were classified into three subfamilies with similar motif compositions and gene structure characteristics in each family. They found that segmental replication and purifying selection play key role in the evolution of the *KEA* gene family in upland cotton. The global expression profiles of *KEA* genes in various tissues and under multiple stress conditions (e.g. salt, drought, and low potassium treatments) were comprehensively analyzed in upland cotton. Virus-induced gene silencing (VIGS) experiment demonstrated that two candidate genes (*GhKEA4* and *GhKEA12*) are involved in salt and potassium stresses response in upland cotton.

Song et al. performed an in-depth investigation on the slow type anion channels (SLAHs) gene family in Cassava, including phylogenetic relationship with other related organisms, chromosomal localization and genome-wide expression analysis. *MeSLAH4* gene was identified as a potential nitrogen-responsive gene, and overexpression analysis in rice demonstrated its capability to enhance the nitrogen assimilation, root growth, and grain yield, indicating its vital role in enhancing nitrogen utilization efficiency and yield.

Peng et al. investigated the three-amino-acid-loop-extension (TALE) transcription factor encoding genes in sweet orange genome, accompanied by systematic analysis regarding their phylogeny, evolution, gene and protein structure, *cis*-acting regulatory element, and protein–protein interaction. They revealed that segmental duplication and purifying selection are the major driving force in the evolution of this gene family in sweet orange. The biological functions of this gene family in sweet orange in response to biotic and abiotic stresses were elucidated by global biotic/abiotic- stress-induced expression pattern analyses (e.g. high temperature, salt, wounding and pathogen stresses). Then, the authors confirmed the transcriptional activity and protein interaction networks of several candidate TALE proteins using yeast two-hybridization assay system

Xie et al. performed a genome-wide analyses of the serine/arginine-rich (SR) gene family in *Brassica napus*, and demonstrated that the genes in each subfamily have conserved structures and motifs and distinct expression patterns. They found that this gene family had a widespread alternative splicing pattern including the paralogous gene pairs. They identified 12 *SR* genes that were potentially associated with specific agronomic traits in *B. napus* by association mapping analysis.

Huang et al. generated a moso bamboo transcriptome dataset under dehydration and cold treatments using RNA-seq technology. This study identified a series of differentially expressed genes and dehydration- and cold-responsive genes. The authors selected a dehydration-responsive gene, *PeLEA14*, as the candidate gene, and proved the roles of *PeLEA14* in response to abiotic stress tolerance including salt stress by overexpression analysis in tobacco.

Li et al. performed genome-wide association study for drought-resistance traits in hulless barley. They evaluated various quantitative traits and field phenotypes of 269 hulless barley lines under either normal or drought conditions. They obtained a total of 8,936,130 highly consistent population SNP markers. Eight candidate genes were suggested to be involved in drought resistance in this genus.

Chloroplast genome sequencing and evolution analysis

Fan et al. reported the chloroplast (cp) genomes of three *Fagopyrum* species, and presented an integrated analysis with five published *Fagopyrum* cp genomes. The evidence of sequence differentiation, repeated sequences, gene number, gene order, codon usage and phylogenetic relationship were presented. This study detected six variable regions and 66 SSR types in the eight species with potential applications in plant genetic relationship and taxonomic status identification. They supported the unique taxonomic status in *Fagopyrum* in the Polygonaceae.

Song et al. assembled and compared the complete cp genomes of nine new *Musa* plants with focus on the genomic features and phylogenetic implications. They found that the studied *Musa* chloroplast genomes commonly encoded 135 functional genes that were composed of photosynthesis-related genes, chloroplast self-replication genes, and other genes. This study identified six non-coding sites and three genes that can be used for DNA barcoding and phylogenetic analysis. They divided the nine *Musa* species into two groups based on phylogenetic analyses.

Li et al. provided 27 complete cp genomes of 11 wild *Fragaria* species accompanied by in-depth variation and evolutionary analyses. They found that the genome structure is highly conserved among these cp genomes, and non-coding regions were relative more variable than coding regions. In addition, the authors revealed that the contraction and expansion of the inverted repeat (IR) regions contributed to different cp genome size in these *Fragaria* species. Five variable loci that could be developed as DNA barcoding for *Fragaria* species were identified. The authors divided the studied species into two groups: Group A distributed in western China and Group B originating from Europe and Americas. The results also revealed allopolyploid origins of the octoploid and tetraploid *Fragaria* species.

Tian et al., present plastome sequencing of 33 peanuts, and then explore their taxonomic status and evolutionary relationship. The authors divided the studied species into two lineages: Lineage I containing all the cultivated species and Lineage II possessing diverse genome types. Next, the authors suggested that all allotetraploid cultivated peanut species were derived from a maternal hybridization event with one of the diploid *Arachis duranens* accessions having a AA sub-genome ancestor, and *Arachis monticola* that represent transitional tetraploid wild species of all the cultivated peanuts.

Author contributions

HD wrote the first draft of the editorial. ZL, WH, and VD all commented on the draft. All authors contributed to the editorial and approved it for publication.

Funding

This work was supported by the National Natural Science Foundation of China (32072094), the Major Science and Technology Program of Hainan Province (ZDKJ202001), the Central Public-interest Scientific Institution Basal Research Fund for Chinese Academy of Tropical Agricultural Sciences (1630052022017 and 1630052020006), and Slovak Research and Development Agency grant APVV-17-0570.

Acknowledgments

The editors would like to thank all the authors and reviewers who have participated in this Research Topic.

Conflict of interest

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

Publisher's note

frontiers in Plant Science

ORIGINAL RESEARCH
published: 14 October 2021
doi: 10.3389/fpls.2021.754209

Variation and Evolution of the Whole Chloroplast Genomes of *Fragaria* spp. (Rosaceae)

*Chenxin Li[1,2], Chaonan Cai[2,3], Yutian Tao[3], Zhongshuai Sun[2,3], Ming Jiang[2], Luxi Chen[2] and Junmin Li[2,3]**

[1]College of Life Sciences and Medicine, Zhejiang Sci-Tech University, Hangzhou, China, [2]Zhejiang Provincial Key Laboratory of Plant Evolutionary Ecology and Conservation, Taizhou University, Taizhou, China, [3]School of Advanced Study, Taizhou University, Taizhou, China

Species identification is vital for protecting species diversity and selecting high-quality germplasm resources. Wild *Fragaria* spp. comprise rich and excellent germplasm resources; however, the variation and evolution of the whole chloroplast (cp) genomes in the genus *Fragaria* have been ignored. In the present study, 27 complete chloroplast genomes of 11 wild *Fragaria* species were sequenced using the Illumina platform. Then, the variation among complete cp genomes of *Fragaria* was analyzed, and phylogenetic relationships were reconstructed from those genome sequences. There was an overall high similarity of sequences, with some divergence. According to analysis with mVISTA, non-coding regions were more variable than coding regions. Inverted repeats (IRs) were observed to contract or expand to different degrees, which resulted in different sizes of cp genomes. Additionally, five variable loci, *trnS-trnG*, *trnR-atpA*, *trnC-petN*, *rbcL-accD*, and *psbE-petL*, were identified that could be used to develop DNA barcoding for identification of *Fragaria* species. Phylogenetic analyses based on the whole cp genomes supported clustering all species into two groups (A and B). Group A species were mainly distributed in western China, while group B contained several species from Europe and Americas. These results support allopolyploid origins of the octoploid species *F. chiloensis* and *F. virginiana* and the tetraploid species *F. moupinensis* and *F. tibetica*. The complete cp genomes of these *Fragaria* spp. provide valuable information for selecting high-quality *Fragaria* germplasm resources in the future.

Keywords: *Fragaria*, chloroplast genome, comparative analysis, wild species, phylogenetic

OPEN ACCESS

Edited by:
Wei Hu,
Chinese Academy of Tropical Agricultural Sciences, China

Reviewed by:
Xiwen Li,
China Academy of Chinese Medical Sciences, China
Yong Qi Zheng,
Chinese Academy of Forestry, China
Guanglong Hu,
Beijing Academy of Agricultural and Forestry Sciences, China

***Correspondence:**
Junmin Li
lijm@tzc.edu.cn

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

Received: *06 August 2021*
Accepted: *21 September 2021*
Published: *14 October 2021*

Citation:
Li C, Cai C, Tao Y, Sun Z, Jiang M, Chen L and Li J (2021) Variation and Evolution of the Whole Chloroplast Genomes of Fragaria spp. (Rosaceae). Front. Plant Sci. 12:754209. doi: 10.3389/fpls.2021.754209

INTRODUCTION

The genus *Fragaria* Linnaeus belongs to the family Rosaceae and is comprised of 25 species, including 13 diploids (2n), five tetraploids (4n), one pentaploid (5n), one hexaploid (6n), three octoploids (8n), and two decaploids (10n; Staudt, 1962, 1989, 2009; Davis et al., 2010; Hummer, 2012; Lei et al., 2017). Most *Fragaria* species are wild, except for *F.* × *ananassa*, which is a cultivated species and an economically important crop (Staudt, 1962; Potter et al., 2000; Cheng et al., 2017). China has been recognized as an important distribution center of wild strawberry resources in the world, as it has 14 wild *Fragaria* species including nine diploid species and

five tetraploid species (Staudt, 1989, 2006; Lei et al., 2017). Compared with cultivated species, wild species have several advantages, including stronger resistance (Diamanti et al., 2012; Guo et al., 2018), unique fruit aromas (Urrutia et al., 2017), and richness in nutrients (Capocasa et al., 2008a,b; Tulipani et al., 2008). Therefore, elucidating the relationships among *Fragaria* spp. is vital to developing high-quality strawberry varieties.

As well as being an important plastid functioning in photosynthesis and carbon fixation (Neuhaus and Emes, 2000), the chloroplast (cp) contains a genome that can provide valuable information for taxonomic and phylogenic purposes, with the key advantages of its relatively small size, more conservative structure, and low nucleotide substitution rate (Palmer, 1985; Wolfe et al., 1987; Jansen et al., 2005; Wei et al., 2006; Wicke et al., 2011; Terakami et al., 2012; Liu et al., 2019). Furthermore, the cp genome is haploid and uniparentally inherited, so it is helpful for tracing source populations and conducting phylogenetic studies to resolve complex evolutionary relationships (Jansen et al., 2007; Parks et al., 2009; Ruhsam et al., 2015). To date, phylogenetic relationships based on complete cp genomes of many angiosperms have been fully studied in many genera (Hu et al., 2016; Zhang et al., 2016; Zhao et al., 2018). Njuguna et al. (2013) systematically studied the phylogenetic relationships of *Fragaria* based on cp genomes, but there were only 10 sequences assembled from total genomic data, and the coverage of PCR amplified sequences ranged from 60 to 90%, which may mean the assembled chloroplast sequences were incomplete. To the best of our knowledge, there is only one study that has focused on the molecular phylogenetic analysis of *Fragaria* genus based on whole complete chloroplast genome sequences (Sun et al., 2021), which revealed that 20 *Fragaria* species were clustered into northern group (eight species), southern group (11 species), and an oldest extant species (one species) based on whole cp genomes. However, this previous study focused on molecular clock analysis and ignored the variation and evolution of whole cp genomes of *Fragaria*, including, for example, the contraction and expansion of inverted repeats (IRs) regions (Plunkett and Downie, 2000; Kode et al., 2005; Ma et al., 2014; Lei et al., 2016).

In addition, although some *Fragaria* species can be identified by their morphological characteristics, some species that have similar morphological structures, such as *F. mandshurica* and *F. orientalis* (Staudt, 2003; Lei et al., 2017), are likely to be inaccurately identified if morphological indexes are not collected (Yang et al., 2020). Furthermore, Yang et al. (2020) found that most wild *Fragaria* in Yunnan were difficult to accurately identify without flowers and fruits of collected species. In total, owing to the stage of species development and subjective factors (such as differences in personal knowledge and experience), traditional morphology classification is often inconsistent and unreliable, which can also affect the results of species identification (Yang et al., 2020). Over the years, many molecular analyses have provided insights into the species taxonomy and identification. Currently, DNA barcoding such as *rbcL*, *matK*, *psbA-trnH* and ITS sequences (Potter et al., 2007; Fazekas et al., 2008; Chen et al., 2013; Xin et al., 2013; Tegally et al., 2019; Phi et al., 2020; Islam et al., 2021) has been used for species identification. In *Fragaria*, the study of DNA barcoding dates back to the phylogenetic analysis conducted by Potter et al. (2000) using ITS and *trnL-trnF* sequences, but the authors were unable to completely distinguish among species in this genus. Njuguna and Bassil (2011) analyzed *psbA-trnH* and ITS sequences and reported that the two DNA barcoding sequences are not suitable for species identification in *Fragaria*. Moreover, more and more studies demonstrated that the four universal barcoding sequences were problematic with low bootstrap support and inability to distinguish between species in land plants (Xin et al., 2013; Tegally et al., 2019; Phi et al., 2020; Islam et al., 2021). Therefore, it is urgent to excavate superior DNA barcoding for special land plants, including *Fragaria* species, utilizing complete cp genomes, which contain more important variation information for taxonomic and phylogenic purposes (Huang et al., 2014).

In the present study, we sequenced 27 complete cp genomes of 11 wild *Fragaria* species from different collection sites in China and downloaded the whole cp genome sequences of seven more *Fragaria* species. This research had the following objectives: (1) to describe the characteristics of *Fragaria* cp genomes; (2) to infer the phylogenetic relationships among *Fragaria* spp.; (3) to detect the variations among *Fragaria* cp genomes and infer the evolution of the whole cp genomes of *Fragaria* species; and (4) to provide candidate DNA barcodes for *Fragaria* species identification. These results provide new insights into the interspecies relationships and evolution of *Fragaria* spp. as well as basic reference material for the application of *Fragaria* germplasm resources.

MATERIALS AND METHODS

Plant Material Collection and Genome Data Sources

Twenty-seven individuals belonging to 11 *Fragaria* species were included in the present study, as summarized in **Table 1**. All the sampled plants were cultured in the greenhouse facilities of Taizhou University under conditions of 70% relative humidity with temperatures of 25°C in the day and 20°C at night. The plants were identified by Professor Beifen Yang of Taizhou University. The specimens were stored in Zhejiang Provincial Key Laboratory of Plant Evolutionary Ecology and Conservation, Taizhou University, China.

The complete cp genome sequences of *F. orientalis* (NC_035501), *Fragaria chiloensis* (NC_019601), and *F. virginiana* (NC_019602) were downloaded from the National Center for Biotechnology Information (NCBI). *Fragaria nipponica* (KY769125) and *F. iinumae* (KC507759) were excluded for there is no supporting of publication reference. *Fragaria* × *ananassa* (KY358226) was also downloaded from NCBI to test the accuration of the sequencing and *de novo* assembly of cp genome.

Raw sequence data from three species, including *F. gracilis* (BOP214815), *F. tibetica* (BOP214818), and *F. moschata* (BOP214819; Sun et al., 2021), were downloaded from NCBI

TABLE 1 | *Fragaria* species collection information.

Voucher	Ploidy	Locality	Latitude (N)	Longitude (E)	Altitude (m)	Genbank accession
*F. nilgerrensis*_1	2	Yunnan, China	27.01°	100.12°	3,403.50	MZ851761
*F. nilgerrensis*_2	2	Yunnan, China	25.32°	100.13°	2,159.26	MZ851762
*F. mandshurica*_JL	2	Jilin, China	42.23°	128.17°	1,047.00	MZ851758
*F. mandshurica*_HLJ	2	Heilongjiang, China	52.41°	125.14°	350.00	MZ851757
*F. corymbosa*_JL	4	Jilin, China	42.09°	128.00°	1,530.00	MZ851750
*F. corymbosa*_XZ	4	Tibet, China	29.61°	94.70°	4,082.81	MZ851751
*F. corymbosa*_GS	4	Gansu, China	35.77°	103.96°	2,833.09	MZ851749
*F. moupinensis*_XZ	4	Tibet, China	29.76°	94.73°	3,381.00	MZ851760
*F. moupinensis*_SC	4	Sichuan, China	29.21°	94.22°	3,062.85	MZ851759
*F. pentaphylla*_1	2	Qinghai, China	36.65°	101.48°	2,399.77	MZ851764
*F. pentaphylla*_2	2	Qinghai, China	36.98°	102.43°	2,327.00	MZ851765
*F. pentaphylla*_3	2	Tibet, China	28.07°	86.00°	3,261.00	MZ851766
*F. pentaphylla*_4	2	Gansu, China	35.83°	104.12°	1,966.78	MZ851767
F. nubicola	2	Tibet, China	28.06°	85.99°	3,353.00	MZ851763
*F. daltoniana*_1	2	Tibet, China	28.07°	86.00°	3,261.00	MZ851752
*F. daltoniana*_2	2	Tibet, China	28.07°	86.00°	3,261.00	MZ851753
*F. daltoniana*_3	2	Tibet, China	28.02°	85.98°	2,724.00	MZ851754
*F. daltoniana*_4	2	Tibet, China	28.03°	85.98°	2,956.00	MZ851755
*F. daltoniana*_5	2	Tibet, China	28.03°	85.98°	2,956.00	MZ851756
F. viridis	2	Segovia, Spain	41.42°	−3.76°	1,170.00	MZ851772
F. vesca ssp. *bracteata*	2	California, United States	38.77°	−120.45°	1,044.00	MZ851768
F. vesca ssp. *vesca*_1	2	Valcea, Romania	46.43°	23.76°	355.61	MZ851769
F. vesca ssp. *vesca*_2	2	Sichuan, China	30.04°	101.82°	3,868.00	MZ851770
F. vesca ssp. *vesca*_3	2	Xinjiang, China	43.87°	85.38°	1,468.00	MZ851771
*F. chinensis*_1	2	Shaanxi, China	33.27°	108.30°	1,186.95	MZ851747
*F. chinensis*_2	2	Shaanxi, China	33.27°	108.30°	1,186.95	MZ851748
F. × *ananassa*	8	Taizhou, China	28.66°	121.39°	22.69	MZ851773

and *de novo* assembly was performed according to the following processes. The other seven species listed by Sun et al. (2021) were excluded for duplication of our own sequencing species.

Genome Sequencing and Assembly

Fresh and clean leaves of each sampled species were collected and frozen in liquid nitrogen immediately. The samples were used to extract the total DNA by the modified CTAB method (Doyle and Doyle, 1987). Then, paired-end sequencing (Insert size: 350 bp) was performed using the Illumina Novaseq 6000 platform (Illumina, San Diego, CA, United States). Raw reads were filtered to obtain high-quality clean data. Then, *de novo* assembly was performed with the GetOrganelle software package (Jin et al., 2020).

Chloroplast Genome Annotation

Thirty cp genomes were annotated using the online GeSeq tool (Tyagi et al., 2020)[1] with default parameters and *F. chiloensis* (NC_019601) was used as the reference to predict protein-coding genes (PCGs), transfer RNA (tRNA) genes, and ribosomal RNA (rRNA) genes. Then, the cp genome sequences were manually modified using Geneious Prime 2021.1.1 (Biomatters Ltd., Auckland, New Zealand). A circular diagram of the cp genomes was generated using the online OrganellarGenome DRAW tool (OGDRAW; Lohse et al., 2013; Bai et al., 2017).

[1]https://chlorobox.mpimp-golm.mpg.de/geseq.html

Comparative Genome Analyses

The boundaries of the large single-copy region (LSC), short single-copy region (SSC), and IRs of 12 newly assembled complete cp genomes of *Fragaria* species (*F. nilgerrensis*_2, *F. mandshurica*_JL, *F. corymbosa*_GS, *F. moupinensis*_SC, *F. pentaphylla*_3, *F. nubicola*, *F. daltoniana*_1, *F. viridis*, *F. vesca* ssp. *bracteata*, *F. vesca* ssp. *vesca*_1, *F. chinensis*_1, *F.* × *ananassa*) in this study together with six complete cp genomes of *F. gracilis* (BOP214815), *F. tibetica* (BOP214818), *F. moschata* (BOP214819), *F. orientalis* (NC_035501), *F. chiloensis* (NC_019601), and *F. virginiana* (NC_019602) from NCBI were compared using IRscope software (Ali et al., 2018). The level of divergence among the 18 sequences was visualized with Shuffle-LAGAN mode (Cheng et al., 2017; Tyagi et al., 2020) in mVISTA software (Kawabe et al., 2018; Jeon and Kim, 2019) with the default settings, and *F. gracilis* was used as a reference sequence.

Hypervariable Site Identification

All 18 sequences were aligned using the Geneious prime 2021.1.1 plugin MAFFT v.7.450 (Katoh and Standley, 2013), and the alignment was manually adjusted. Then, we performed a sliding window analysis using DnaSP v6 software (Rozas et al., 2017; Jeon and Kim, 2019) to analyze nucleotide diversity (π) in order to detect hypervariable sites among *Fragaria* cp genomes. The window length was set to 600 bp, and the step size was set to 200 bp (Sun et al., 2021).

Phylogenetic Analysis

Thirty-four complete cp genome sequences were used to reconstruct the phylogenetic relationships among *Fragaria* spp. based on maximum likelihood (ML) using RAxML 8.2.10 (Stamatakis, 2014), with 1,000 bootstrap replicates employed for estimating node support. *Potentilla fruticosa* (NC_036423) and *Drymocallis saviczii* (NC_050966) were downloaded from Genbank and used as outgroups (Eriksson et al., 1998, 2003; Potter et al., 2000; Feng et al., 2017; Dimeglio et al., 2014). In total, 36 cp genome sequences were aligned using a Geneious prime 2021.1.1 plugin MAFFT v.7.450 (Katoh and Standley, 2013), and the alignment was manually adjusted when necessary. Modeltest 3.7 (Liu et al., 2013) was used to select the best-fit evolutionary model of *Fragaria* cp genome sequence evolution for reconstruction of the phylogenetic relationships of *Fragaria*.

RESULTS

General Chloroplast Genome Characteristics

The total genome sequence lengths of *Fragaria* species ranged from 155,479 bp (*F. viridis*) to 155,832 bp (*F. daltoniana*_3). The cp genomes presented a typical quadripartite structure including a pair of IR regions with lengths of 51,872 bp (*F. vesca* ssp. *bracteata*) to 51,948 bp (*F. corymbosa*_JL), separated by a LSC region from 85,471 bp (*F. viridis*) to 85,726 bp (*F. daltoniana*_3) and a SSC region from 18,116 bp (*F. viridis*) to 18,219 bp (*F. moupinensis*_XZ). The GC contents ranged from 37.2 to 37.3%. Overall, the cp genome of *Fragaria* encodes a total of 130 genes, including 85 PCGs, 37 tRNA genes, and eight rRNA genes (**Figure 1**; **Table 2**). Among these genes, 15 contained a single intron (*ndhA*, *ndhB*, *petB*, *petD*, *rpl2*, *rpl16*, *rpoC1*, *rps12*, *rps16*, *trnA-UGC*, *trnG-GCC*, *trnI-CAU*, *trnK-UUU*, *trnL-UAA*, and *trnV-UAC*), while two harbored two introns (*ycf3* and *clpP*). The *trnK-UUU* gene had the longest intron (2,492–2,496 bp), which contained the *matK* gene, whereas *trnL-UAA* was smallest (420 bp; **Table 3**).

Comparative Analyses

The detailed comparisons of LSC, SSC, and IR boundaries in 19 *Fragaria* complete cp genomes are shown in **Figure 2**, revealing differences at boundary regions. There were five genes around borders of these regions, including *rps19*, *rpl2*, *ycf1*, *ndhF*, and *trnH*, in which *rpl2* and *ycf1* had two copies. Overall, *F. moschata*, *F. mandshurica*_JL, *F. viridis*, *F. vesca* ssp. *vesca*_1, *F. vesca* ssp. *bracteata*, and *F.* × *ananassa* showed higher structural similarity, with *rps19*, *rpl2*, and *trnH* genes found at the same distances from the boundaries. The *ycf1* gene in *Fragaria* species spanned 1,092 bp from the SSC to IRb region, resulting in a non-functional *ψycf1* fragment gene with the same length in IRa. Additionally, *ψycf1* extended to the SSC region, with different distances ranging from 3 bp (*F. orientalis*) to 37 bp (*F. daltoniana*_1), with the cp genomes of *F. mandshurica*_JL, *F. viridis*, *F. vesca* ssp. *vesca*_1, *F. vesca* ssp. *bracteata*, *F.* × *ananassa*, *F. chiloensis*, and *F. virginiana* all showing the same gap of 13 bp. Moreover, *ψycf1* and *ndhF* had more length polymorphisms at the IRa/SSC border (**Figure 2**).

Annotation of the *F. gracilis* cp genome was used for plotting the total sequence identity of the 18 *Fragaria* cp genomes in mVISTA (**Figure 3**). Overall, there was a high similarity among *Fragaria* species, except *F. chinensis*_1, *F. viridis*, and *F. orientalis*. Furthermore, these results showed non-coding regions are more divergent than coding regions. In our research, the most divergent coding regions in the *Fragaria* cp genomes analyzed were *rpoC2*, *psbJ*, *ycf1*, and *ndhA*. In addition, the representative most divergent non-coding regions were *rps16-trnQ*, *trnR-atpA*, *petN-psbM*, *trnT-trnL*, *ndhC-trnV*, *petA-psbJ*, *trnP-psaJ*, *rpl32-trnL*, and *rps15-ycf1*.

Nucleotide diversity (π) was calculated to evaluate the sequence variability level in *Fragaria* cp genomes using DnaSP 6.0 software (**Figure 4**). The values ranged from 0 to 0.01016, and the average value was 0.00167. Four more polymorphic regions ($\pi > 0.007$) were identified, including *trnS-trnG*, *trnR-atpA*, *trnC-petN*, *rbcL-accD*, and *psbE-petL*, in which *rbcL-accD* had the highest π value, and all of these regions were located in the LSC region.

Phylogenetic Analysis

The GTR + I + G model was selected as the best-fit evolutionary model by using Modeltest 3.7. The phylogenetic tree was constructed using 34 complete cp sequences of *Fragaria* species, with *P. fruticosa* and *D. saviczii* as outgroups. As shown in **Figure 5**, *Fragaria* species can be clustered into two groups, A and B, with 100% bootstrap support. Group A included two subgroups, A1 (*F. chinensis* and *F. daltoniana*) and A2. Subgroup A2 contained six *Fragaria* species (*F. nubicola*, *F. pentaphylla*, *F. corymbosa*, *F. moupinensis*, *F. gracilis* and *F. tibetica*) that are mainly distributed in western China. Group B was composed of the remaining species, which included *F. nilgerrensis*, *F. mandshurica*, *F. viridis*, *F. orientalis*, *F. moschata*, *F.* × *ananassa*, *F. chiloensis*, *F. virginiana*, *F. vesca* ssp. *vesca* and *F. vesca* ssp. *bracteata*. These latter species are from Europe and America, apart from *F. nilgerrensis*, which is mainly distributed in southeast Asia, *F. mandshurica*, which is distributed in northeast China, and *F. orientalis*, which is mainly found in northern Asia.

DISCUSSION

Variations and Evolution of Whole Cp Genomes of *Fragaria* spp.

In this study, 27 cp genomes of *Fragaria* species were sequenced and found to range in size from 155,479 to 155,832 bp, which falls within the cp genome size range for angiosperms but tends to be smaller than the cp genomes of other Rosaceae species (Palmer, 1985; Cheng et al., 2017). LSC regions showed the most difference in size, ranging from 85,471 to 85,726 bp. Additionally, the inferred structures and gene contents were in accordance with previous research (Sun et al., 2021).

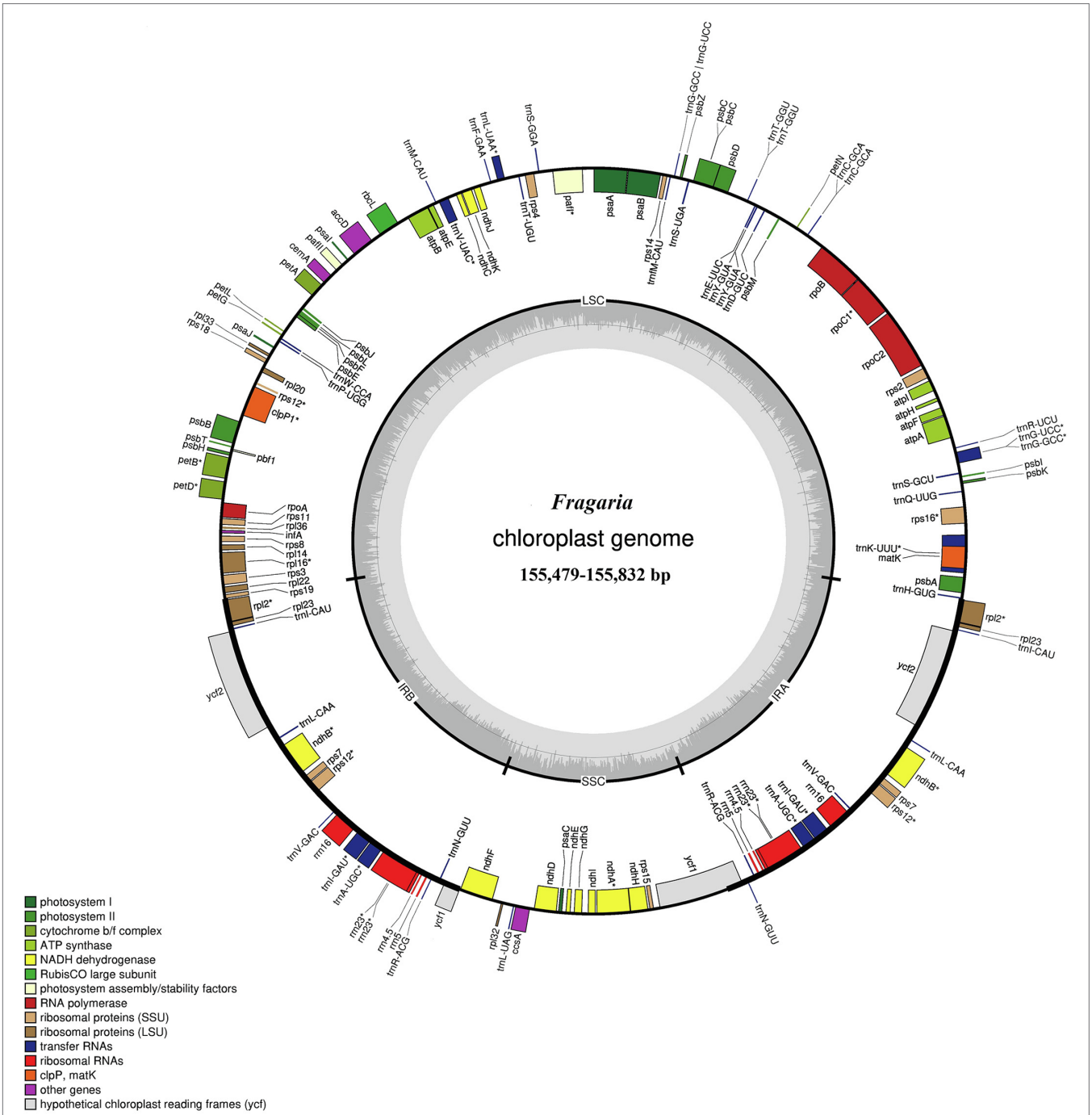

FIGURE 1 | Gene map of the *Fragaria* chloroplast genome. Genes drawn inside the circle are transcribed clockwise, and those outside are transcribed counter clockwise. Color coding is used to show genes belonging to different functional groups. The inner circle indicates the range of the long single-copy region (LSC), short single-copy region (SSC), and inverted repeat regions (IRs) and also shows a GC content graph of the genome. In the GC content graph, the dark gray lines indicate GC content, while light gray lines indicate the AT content of each locus.

Overall, *Fragaria* cp genomes were highly conservative, both in sequence and structure. Analysis with mVISTA showed that there is high similarity among *Fragaria* species apart from *F. chinensis*_1, *F. viridis*, and *F. orientalis*. We also observed that most variable regions were located in LSC, and non-coding regions were more variable than coding regions. This is a common phenomenon in the cp genomes of most angiosperms (Nazareno et al., 2015; Cheng et al., 2017; Asaf et al., 2018; Tyagi et al., 2020). Additionally, some of the most divergent regions of *ycf1*, *rps16-trnQ*, *petN-psbM*, and *rpl32-trnL*, as shown in **Figure 5**, were consistent with previous research (Cheng et al., 2017), indicating that these regions indeed evolve rapidly in *Fragaria*.

TABLE 2 | Summary of 34 complete chloroplast genomes for *Fragaria* species.

Species	Size (bp)				Gene number				GC content (%)	References
	Total	LSC	SSC	IR	Total	PCG	tRNA	rRNA		
*F. nilgerrensis*_1	155,783	85,712	18,165	25,953	130	85	37	8	37.3	This study
*F. nilgerrensis*_2	155,675	85,602	18,147	25,963	130	85	37	8	37.2	This study
*F. mandshurica*_JL	155,559	85,504	18,161	25,947	130	85	37	8	37.2	This study
*F. mandshurica*_HLJ	155,556	85,507	18,155	25,947	130	85	37	8	37.2	This study
*F. corymbosa*_JL	155,684	85,538	18,198	25,974	130	85	37	8	37.2	This study
*F. corymbosa*_XZ	155,683	85,544	18,217	25,961	130	85	37	8	37.2	This study
*F. corymbosa*_GS	155,696	85,557	18,217	25,961	130	85	37	8	37.2	This study
*F. moupinensis*_XZ	155,630	85,487	18,219	25,962	130	85	37	8	37.2	This study
*F. moupinensis*_SC	155,626	85,489	18,215	25,961	130	85	37	8	37.2	This study
*F. pentaphylla*_1	155,626	85,510	18,192	25,962	130	85	37	8	37.2	This study
*F. pentaphylla*_2	155,640	85,524	18,192	25,962	130	85	37	8	37.2	This study
*F. pentaphylla*_3	155,628	85,511	18,193	25,962	130	85	37	8	37.2	This study
*F. pentaphylla*_4	155,666	85,549	18,193	25,962	130	85	37	8	37.2	This study
F. nubicola	155,608	85,512	18,174	25,961	130	85	37	8	37.2	This study
*F. daltoniana*_1	155,829	85,723	18,170	25,968	130	85	37	8	37.2	This study
*F. daltoniana*_2	155,829	85,723	18,170	25,968	130	85	37	8	37.2	This study
*F. daltoniana*_3	155,832	85,726	18,170	25,968	130	85	37	8	37.2	This study
*F. daltoniana*_4	155,827	85,721	18,170	25,968	130	85	37	8	37.2	This study
*F. daltoniana*_5	155,827	85,721	18,170	25,968	130	85	37	8	37.2	This study
F. viridis	155,479	85,471	18,116	25,946	130	85	37	8	37.2	This study
F. vesca ssp. *bracteata*	155,564	85,541	18,151	25,936	130	85	37	8	37.2	This study
F. vesca ssp. *vesca*_1	155,607	85,520	18,191	25,948	130	85	37	8	37.3	This study
F. vesca ssp. *vesca*_2	155,638	85,559	18,173	25,953	130	85	37	8	37.2	This study
F. vesca ssp. *vesca*_3	155,564	85,521	18,147	25,948	130	85	37	8	37.2	This study
*F. chinensis*_1	155,806	85,696	18,184	25,963	130	85	37	8	37.2	This study
*F. chinensis*_2	155,797	85,688	18,183	25,963	130	85	37	8	37.2	This study
F. × *ananassa*	155,549	85,532	18,145	25,936	130	85	37	8	37.2	This study
F. × *ananassa*	155,549	85,532	18,145	25,936	130	85	37	8	37.2	Cheng et al., 2017
F. orientalis	147,835	83,233	13,386	25,608	128	84	36	8	37.6	Han et al., 2018
F. chiloensis	155,603	85,567	18,146	25,945	130	85	37	8	37.2	Salamone et al., 2013
F. virginiana	155,621	85,586	18,145	25,945	130	85	37	8	37.2	Salamone et al., 2013
F. gracilis	155,684	85,538	18,198	25,974	130	85	37	8	37.2	Sun et al., 2021
F. tibetica	155,643	85,498	18,219	25,963	130	85	37	8	37.2	Sun et al., 2021
F. moschata	155,601	85,572	18,127	25,951	130	85	37	8	37.2	Sun et al., 2021

LSC, large single-copy region; SSC, small single-copy region; IR, inverted repeat; PCG, Protein-coding gene.

TABLE 3 | Genes in the *Fragaria* chloroplast genome.

Category	Gene group	Gene name
Photosynthesis	Subunits of photosystem I	*psaA*, *pasB*, *psaC*, *psaI*, *psaJ*
	Subunits of photosystem II	*psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *pabL*, *psbM*, *psbN*, *psbT*, *psbZ*
	Subunits of NADH dehydrogenase	*ndhA*[a], *ndhB*[a,c], *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *ndhJ*, *ndhK*
	Subunits of cytochrome b/f complex	*petA*, *petB*[a], *petD*[a], *petG*, *petL*, *petN*
	Subunits of ATP synthase	*atpA*, *atpB*,*atpE*,*atpF*, *atpH*, *atpI*
	Large subunit of rubisco	*rbcL*
Self-replication	Large subunit of ribosome	*rpl2*[a,c], *rpl14*, *rpl16*[a], *rpl20*, *rpl22*[a], *rpl23*[c], *rpl32*, *rpl33*, *rpl36*
	Small subunit of ribosome	*rps2*, *rps3*, *rps4*, *rps7*[c], *rps8*, *rps11*, *rps12*[a,c], *rps14*, *rps15*, *rps16*[a], *rps18*, *rps19*
	Subunits of RNA polymerase	*rpoA*, *rpoB*, *rpoC1*[a], *rpoC2*
	Ribosomal RNA genes	*rrn16*[c], *rrn23*[c], *rrn4.5*[c], *rrn5*[c]
	Transfer RNA genes	*trnA-UGC*[a,c], *trnC-GCA*, *trnD-GUC*, *trnE-UUC*, *trnF-GAA*, *trnG-GCC*[a], *trnG-UCC*, *trnH-GUG*, *trnI-CAU*[c], *trnI-GAU*[a,c], *trnK-UUU*[a], *trnL-CAA*[c], *trnL-UAA*[a], *trnL-UAG*, *trnfM-CAU*, *trnM-CAU*, *trnN-GUU*[c], *trnP-UGG*, *trnQ-UUG*, *trnR-ACG*[c], *trnR-UCU*, *trnS-GCU*, *trnS-GGA*, *trnS-UGA*, *trnT-GGU*, *trnT-UGU*, *trnV-GAC*[c], *trnV-UAC*[a], *trnW-CCA*, *trnY-GUA*
Other genes	Maturase	*matK*
	Protease	*clpP*[b]
	Envelope membrane protein	*cemA*
	Acetyl-CoA carboxylase	*accD*
	c-type cytochrome synthesis gene	*ccsA*
Genes of unknown function	Conserved open reading frames	*ycf1*[c], *ycf2*[c], *ycf3*[b], *ycf4*

[a]*Gene with one intron.*
[b]*Gene with two introns.*
[c]*Genes with two copies.*

The size differences among cp genomes in angiosperms may be caused by both the contraction and expansion of IR regions (Raubeson et al., 2007; Zhao et al., 2015, 2018). To elucidate this mechanism in *Fragaria*, we compared IR/SC boundaries of *Fragaria* cp genomes. In general, the distribution of border genes is conserved, but the distances between genes and the borders do differ somewhat. The distances between genes and IR/SC borders of *F. virginiana*, *F. orientalis*, and *F.* × *ananassa* are in accordance with the prior report by Cheng et al. (2017). *Fragaria mandshurica*, *F. viridis*, *F. moschata*, *F.* × *ananassa*, *F. vesca* ssp. *vesca* and *F. vesca* ssp. *bracteata* showed the same gap between *rps19*, *rpl2*, *ycf1*, *trnH*, and IR/SC junctions, which may explain why these cp genomes are more conserved. Additionally, some species also exhibited the same distances between the *rps19* and LSC/IRa border, including *F. pentaphylla*, *F. moupinensis*, and *F. tibetica*. These results also indicated a low level of molecular divergence in the genus *Fragaria*.

Additionally, *rpl2* and *rps19* differed in their distances from the LSC/IRa border, which may be owing to IR contraction and expansion. Compared with *F.* × *ananassa*, the IR regions of *F. pentaphylla*, *F. daltoniana*, *F. nubicola*, *F. chinensis*, *F. corymbosa*, *F. moupinensis*, *F. orientalis*, *F. gracilis*, *F. tibetica*, *F. virginiana*, and *F. chiloensis* expanded to different degrees. Therefore, IR regions of these species are longer than that of *F.* × *ananassa*. For the other species in the family Rosaceae, such as *Malus* (Terakami et al., 2012), *Prinsepia* (Wang et al., 2013), *Pyrus* (Li et al., 2018), and *Prunus* (Kim et al., 2019), *rps19* crossed the LSC region and IR region. Additionally, *rps19* extended to the IRa region, resulting in the presence of *ψrps19* having the same length within IRb region. Notably, we found that *rps19* was located inside the LSC region. The contraction of *rps19* inside the LSC region would result in the size of the *Fragaria* cp genomes being smaller than that of other species in Rosaceae.

Phylogenetic Analysis

The similar morphology of most *Fragaria* spp. and the geographical overlap in ranges may lead to taxonomic confusion among collected specimens (Johnson et al., 2014) and finally result in misjudgment of phylogenetic relationship. To explore the evolutionary relationships among *Fragaria* species, we constructed a phylogenetic tree of *Fragaria* species based on whole cp genomes with more than 99% bootstrap support across all nodes. Our results showed high consistency with previous results, especially for species clustering in group B (Njuguna et al., 2013; Sun et al., 2021). Additionally, our results include multiple individuals of each species from different collection sites, which increases the reliability of the phylogenetic analysis.

The phylogenetic analysis showed that *F.* × *ananassa* and two octoploids, *F. chiloensis* and *F. virginiana*, formed a group, which was in accordance with the inferred origin of *F.* × *ananassa* from a hybridization between *F. virginiana* and *F. chiloensis* (Staudt, 1962, 1989). In addition, the two octoploids were considered to share a common maternal ancestor that may be *F. vesca* or *F. mandshurica* (Harrison et al., 1997; Rousseau-Gueutina et al., 2009; Dimeglio et al., 2014). Fortunately, we observed that *F. vesca* ssp. *bracteata* is evolutionarily closely related to the two octoploids. Similar results were also presented by Sun et al. (2021). Additionally, Njuguna et al. (2013) and Edger et al. (2019) suggested that *F. vesca* ssp. *bracteata* was probably the last diploid progenitor to the octoploid species. Thus, the conclusion that *F. vesca* ssp. *bracteata* is an ancestor of octoploid species was strengthened. However, the other ancestors of the octoploid species remain uncertain.

The evolutionary relationships of *F. nubicola*, *F. pentaphylla*, *F. chinensis*, *F. daltoniana*, *F. corymbosa*, *F. moupinensis*, *F. gracilis*,

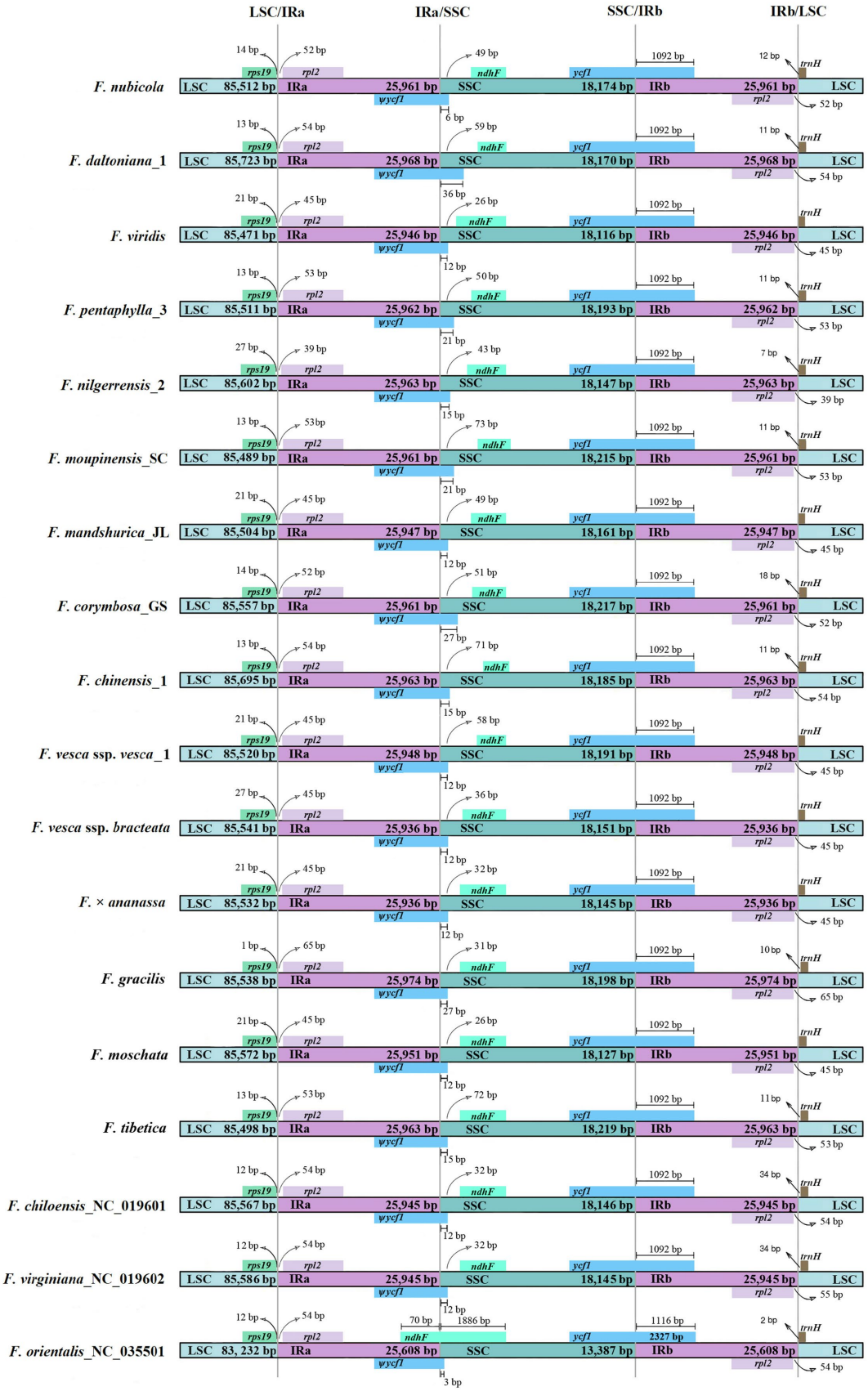

FIGURE 2 | Comparison of the LSC, SSC, and IR border regions among 18 *Fragaria* chloroplast genomes. Ψ is used to indicate pseudogenes.

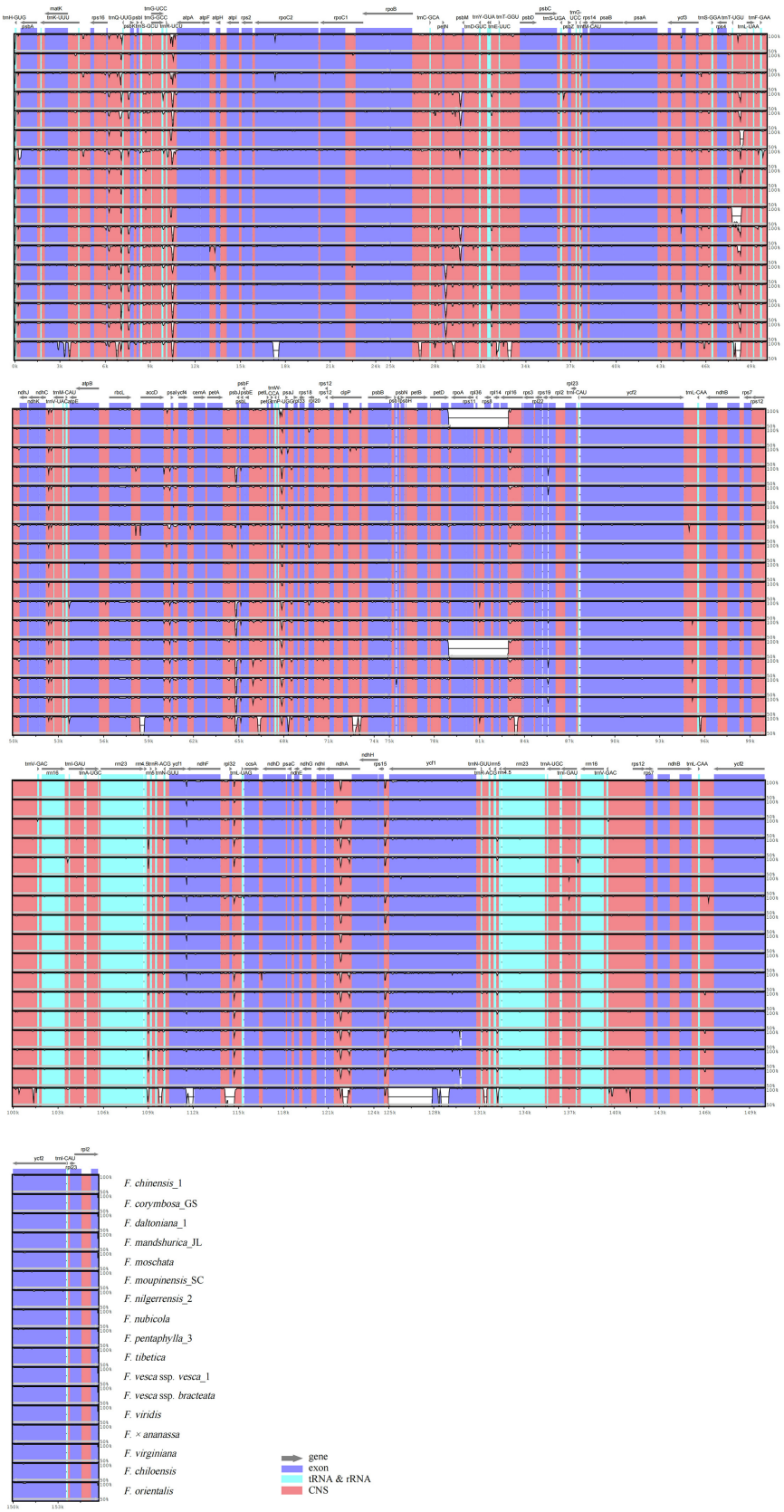

FIGURE 3 | Continued

FIGURE 3 | Visualized alignment of the *Fragaria* chloroplast genome sequences with annotations, using mVISTA. Each horizontal row shows the graph for the pairwise sequence identity with the *Fragaria gracilis* chloroplast genome sequence. The *x*-axis represents the base sequence of the alignment, and the *y*-axis represents the pairwise percent identity ranging from 50 to 100%. Gray arrows indicate the position and direction of each gene. Red indicates non-coding sequences (CNS); blue indicates the exons of protein-coding genes (exons); lime green indicates ribosomal RNA (rRNA) or transfer RNA (tRNA) sequences.

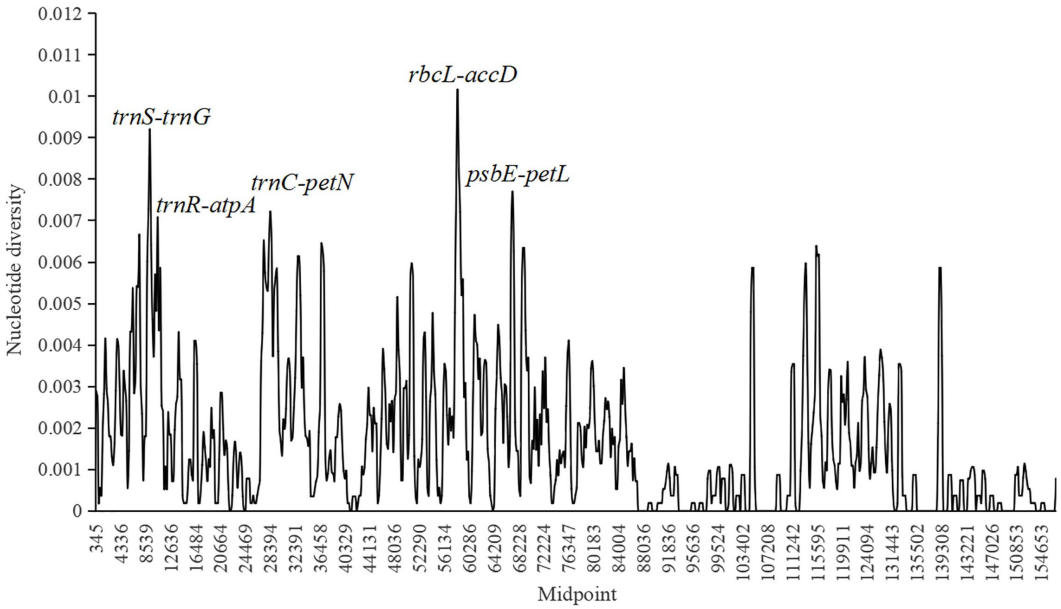

FIGURE 4 | Sliding window analysis of 18 complete chloroplast (cp) genomes of *Fragaria* species (window length, 600 bp; step size, 200 bp; *x*-axis, position of the midpoint of a window; *y*-axis, nucleotide diversity of each window). Each highly polymorphic region of the *Fragaria* cp genomes is annotated on the graph.

and *F. tibetica* have never been clear (Rousseau-Gueutina et al., 2009; Sun et al., 2021). These species are mainly distributed in Western China (Lei et al., 2017) and, in our results, they were clustered into group A. In group A2, diploid *F. pentaphylla* was sister to the tetraploids *F. moupinensis* and *F. tibetica* with 99.9% bootstrap support (**Figure 5**), which was consistent with previous research (Njuguna et al., 2013; Sun et al., 2021). In addition, *F. pentaphylla* had been suggested to be the diploid ancestor to *F. moupinensis* (Rousseau-Gueutina et al., 2009; Kamneva et al., 2017). Lei et al. (2017) hypothesized that *F. tibetica* is a descendant of *F. pentaphylla* based on their runner branching and number of leaflets. Thus, it can be hypothesized that the tetraploid species *F. moupinensis* and *F. tibetica* may share the same female parent of *F. pentaphylla*, which is supported by their more similar morphological characteristics (Lei et al., 2017) and overlapping distribution in Southwestern China (Staudt, 1989; Johnson et al., 2014; Lei et al., 2017).

In our study, we revealed a sister relationship between *F. corymbosa* and *F. gracilis*, which was consistent with Rousseau-Gueutina et al. (2009). And *F. corymbosa* and *F. gracilis* have some similar morphological characteristics, such as runners are filiform and monopodial, petioles and peduncles have spreading hairs, fruits are red and tasteless, calyx is reflexed, etc. (Lei et al., 2017). So our results strengthened the point that *F. corymbosa* and *F. gracilis* may have the same ancestor (Rousseau-Gueutina et al., 2009). In addition, *F. corymbosa* and *F. gracilis* may be the descendent of *F. chinensis* (Staudt, 2009; Yang and Davis, 2017). Notably, in this study, all accessions of *F. chinensis* and *F. daltoniana* were clustered into group A1, in contrast with the findings of previous studies (Yang and Davis, 2017; Sun et al., 2021). Therefore, future research should explore the relationship among *F. chinensis*, *F. corymbosa*, and *F. daltoniana*. In the future, multiple molecular markers, including cp and nuclear sequence data from more samples from different geographical populations, should be combined with geographical distribution data to analyse ancestral state reconstruction to clarify their phylogenetic ancestor relationships among *F. moupinensis* and *F. tibetica*; *F. chinensis*, *F. corymbosa*, and *F. daltoniana*; *F. vesca* ssp. *bracteata* and the other octoploid species.

Chloroplast capture is an important process of plant evolution (Okuyama et al., 2005). Hybridization and repeated backcross, the cytoplasm of one species is replaced by the cytoplasm of another species through gene flow infiltration, so that the genetic components of the species not only have nuclear genome components inherited from parents, but also capture new chloroplast gene components (Fehrer et al., 2007). More and more studies have proved the phenomenon of organelle DNA introgression (Du et al., 2011), and the phenomenon of chloroplast introgression between plant species has also been observed in previous studies on hazelnut (Hu et al., 2020). In this study, the phylogeographical relationships among *Fragaria* species were not declared for the lack of geographical population collections.

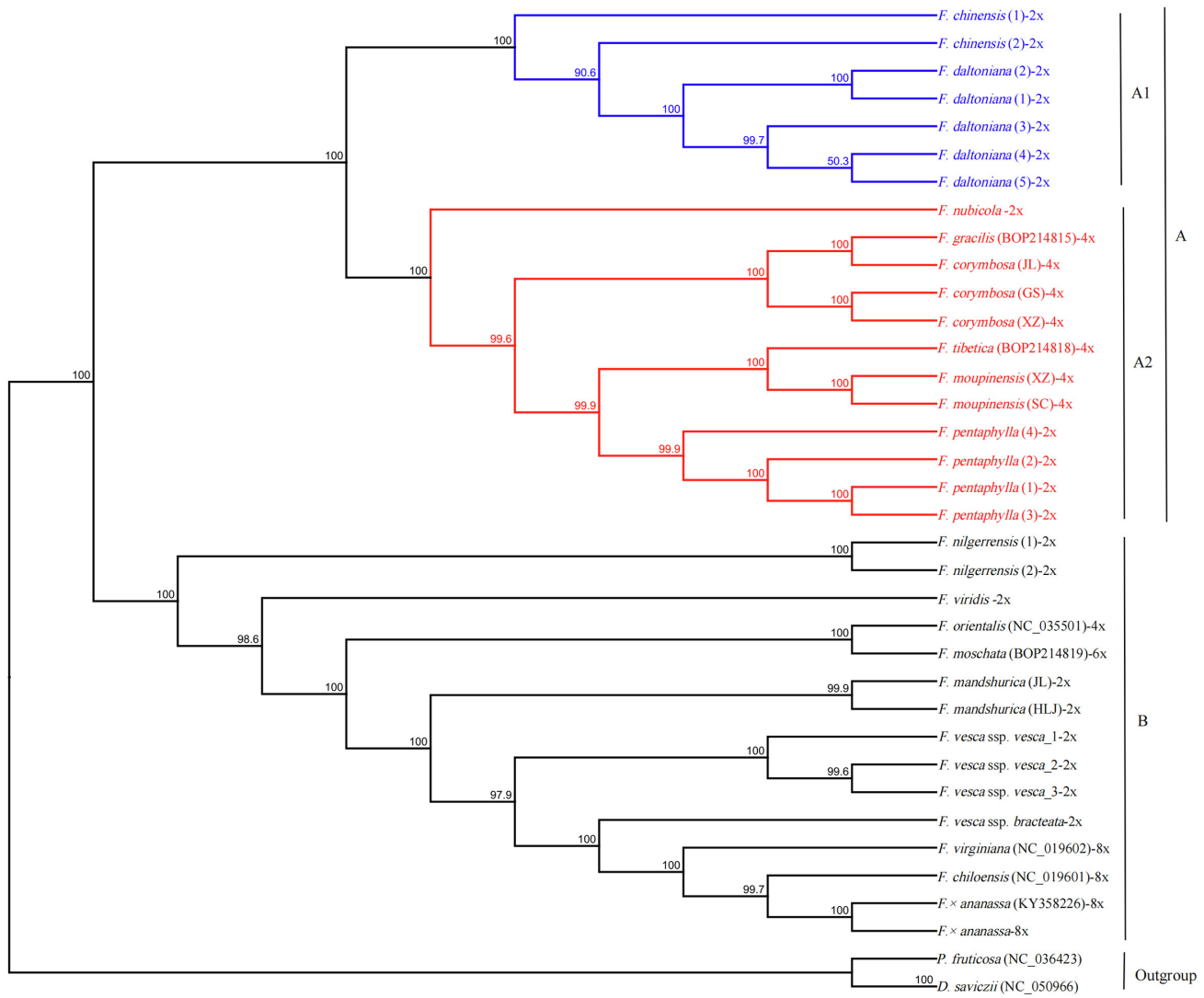

FIGURE 5 | A maximum likelihood phylogenetic tree was reconstructed based on 34 *Fragaria* cp genomes. *Potentilla fruticosa* and *Drymocallis saviczii* were used as outgroups, and −2x, −4x, and −8x represent the different ploidies of *Fragaria* spp. Nodes marked with capital letters are discussed in the text.

However, clear geographical patterns (Western China, Southwestern China, and Northeastern China) have been clearly inferred in phylogenic tree figure. Chloroplast capture could be another explanation why the chloroplast genome analysis does not appear to reflect the species phylogeographical relationship (Tsitrone et al., 2003). Further research should be conducted by combining multiple molecular tools (e.g., nuclear DNA sequences) together with more comprehensive sampling to clarify if chloroplast capture does occur in *Fragaria* genus.

Candidate Barcoding Sequences for *Fragaria*

Species identification based on morphology is affected by season, environment, and human factors, which may cause results to be unreliable (Yang et al., 2020). In recent years, DNA barcoding has been widely used to promote accurate species identification owing to its clear advantages (Tegally et al., 2019; Phi et al., 2020; Islam et al., 2021). The ideal DNA barcode would be a single locus that could be universally amplified and sequenced across a broad range of taxa and provide sufficient variation to reliably distinguish among closely related species (Song et al., 2017). Many introns, coding regions, and intergenic regions, such as *trnL-trnF* (Potter et al., 2000), *accD-psaI* (Wang et al., 2017), *ycf1-ndhF* (Amar, 2020), *matK*, and *trnK* (Hilu et al., 2008), have been used as barcodes for constructing phylogenetic relationships. In this study, with the threshold of nucleotide diversity as 0.007, the five intergenic regions, *trnS-trnG*, *trnR-atpA*, *trnC-petN*, *rbcL-accD*, and *psbE-petL* were found to be the most divergent and provide potential information for species identification and phylogenetic analyses of *Fragaria*. Among them, three intergenic regions, including *trnS-atpA*, *rbcL-accD*, and *psbE-petL* were also suggested in Sun et al. (2021). However,

the threshold of nucleotide diversity used in Sun et al. (2021) was only 0.006, resulting in a low nucleotide diversity of the selected candidate barcoding sequences. In addition, the amplified fragments of the selected intergenic regions *trnS-atpA* more than 2,000 bp in length would reduce the success of the sequencing. In our study, *trnS-trnG* and *trnR-atpA* are located in *trnS-atpA* regions and the length of these two regions are about 700 bp, which will result in the high success of sequencing. Therefore, *trnS-trnG* and *trnR-atpA* are more suitable than *trnS-atpA* to be the potential candidate barcoding sequence.

CONCLUSION

This study provides 27 complete cp genome sequences of 11 wild *Fragaria* species. Comparative analysis of cp genomes of *Fragaria* species revealed that their genome structure is highly conserved. However, IR expansion or contraction was observed among different *Fragaria* cp genomes, resulting in cp genomes of different sizes. Five identified highly variable gene regions (*trnS-trnG*, *trnR-atpA*, *trnC-petN*, *rbcL-accD*, and *psbE-petL*) showed strong potential for species identification and phylogenetic relationship construction in the genus *Fragaria*. Phylogenetic analysis indicated that *F. vesca* ssp. *bracteata* may be one of the progenitor species of octoploids. Similarly, we hypothesize that *F. pentaphylla* is one of the progenitors of *F. corymbosa* and *F. tibetica*. The analysis of multiple molecular markers combined with morphological characters would be helpful for future research to test this hypothesis.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository and accession number(s) can be found below: NCBI repository, accession numbers MZ851747 and MZ851773.

AUTHOR CONTRIBUTIONS

JL designed the research. CL, CC, YT, ZS, and MJ performed the research and analyzed the data. CL wrote the first draft of the manuscript. All authors commented on previous versions of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was financially supported by the Ten Thousand Talent Program of Zhejiang Province (No. 2019R52043) and the National Natural Science Foundation of China (No. 31261120580).

ACKNOWLEDGMENTS

The authors would like to thank Beifen Yang of Taizhou University (Taizhou, China) for the help of identification of *Fragaria* species.

REFERENCES

Ali, A., Jaakko, H., and Peter, P. (2018). IRscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics* 34, 3030–3031. doi: 10.1093/bioinformatics/bty220

Amar, M. H. (2020). *ycf*1-*ndh*F genes, the most promising plastid genomic barcode, sheds light on phylogeny at low taxonomic levels in *Prunus persica*. *J. Genet. Eng. Biotechnol.* 18:42. doi: 10.1186/s43141-020-00057-3

Asaf, S., Khan, A. L., Khan, M. A., Shahzad, R., and Lee, I. J. (2018). Complete chloroplast genome sequence and comparative analysis of loblolly pine (*Pinustaeda* L.) with related species. *PLoS One* 13:e0192966. doi: 10.1371/journal.pone.0192966

Bai, L., Ye, Y., Chen, Q., and Tang, H. R. (2017). The complete chloroplast genome sequence of the white strawberry *Fragaria pentaphylla*. *Conserv. Genet. Resour.* 9, 659–661. doi: 10.1007/s12686-017-0713-5

Capocasa, F., Diamanti, J., Tulipani, S., and Battino, M. (2008a). Breeding strawberry (*Fragaria* × *ananassa* Duch) to increase fruit nutritional quality. *Biofactors* 34, 67–72. doi: 10.1002/biof.5520340107

Capocasa, F., Scalzo, J., Mezzetti, B., and Battino, M. (2008b). Combining quality and antioxidant attributes in the strawberry: the role of genotype. *Food Chem.* 111, 872–878. doi: 10.1016/j.foodchem.2008.04.068

Chen, X. C., Liao, B. S., Song, J. Y., Pang, X. H., Han, J. P., and Chen, S. L. (2013). A fast SNP identification and analysis of intraspecific variation in the medicinal *Panax* species based on DNA barcoding. *Gene* 530, 39–43. doi: 10.1016/j.gene.2013.07.097

Cheng, H., Li, J. F., Zhang, H., Cai, B. H., Gao, Z. H., Qiao, Y. S., 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

Davis, T. M., Shields, M. E., Reinhard, A. E., Reavey, P. A., Lin, J., Zhang, H., et al. (2010). Chloroplast DNA inheritance, ancestry, and sequencing in *Fragaria*. *Acta Hortic.* 859, 221–228. doi: 10.17660/ActaHortic.2010.859.25

Diamanti, J., Capocasa, F., Balducci, F., Battino, M., Hancock, J., and Mezzetti, B. (2012). Increasing strawberry fruit sensorial and nutritional quality using wild and cultivated germplasm. *PLoS One* 7:e46470. doi: 10.1371/journal.pone.0046470

Dimeglio, L. M., Staudt, G., Yu, H., and Davis, T. M. (2014). A phylogenetic analysis of the genus *Fragaria* (strawberry) using intron-containing sequence from the *ADH-1* gene. *PLoS One* 9:e102237. doi: 10.1371/journal.pone.0102237

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

Du, F. K., Peng, X. L., Liu, J. Q., Lascoux, M., Hu, F. S., and Petit, R. J. (2011). Direction and extent of organelle DNA introgression between two spruce species in the Qinghai-Tibetan plateau. *New Phytol.* 192, 1024–1033. doi: 10.1111/j.1469-8137.2011.03853.x

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

Eriksson, T., Donoghue, M. J., and Hibbs, M. S. (1998). Phylogenetic analysis of *Potentilla* using DNA sequences of nuclear ribosomal internal transcribed spacers (ITS), and implications for the classification of Rosoideae (Rosaceae). *Plant Syst. Evol.* 211, 155–179. doi: 10.1007/BF00985357

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 *trn*L/F region of chloroplast DNA. *Int. J. Plant Sci.* 164, 197–211. doi: 10.1086/346163

Fazekas, A. J., Burgess, K. S., Kesanakurti, P. R., Graham, S. W., Newmaster, S. G., Husband, B. C., et al. (2008). Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PLoS One* 3:e2802. doi: 10.1371/journal.pone.0002802

Fehrer, J., Gemeinholzer, B., Chrtek, J. J., and Braütigam, S. (2007). Incongruent plastid and nuclear DNA phylogenies reveal ancient intergeneric hybridization

in *Pilosella* hawkweeds (*Hieracium*, Cichorieae, Asteraceae). *Mol. Phylogenet. Evol.* 42, 347–361. doi: 10.1016/j.ympev.2006.07.004

Feng, T., Moore, M. J., Yan, M. H., Sun, Y. X., Zhang, H. J., Meng, A. P., et al. (2017). Phylogenetic study of the tribe Potentilleae (Rosaceae), with further insight into the disintegration of *Sibbaldia*. *J. Syst. Evol.* 55, 177–191. doi: 10.1111/jse.12243

Guo, R. X., Xue, L., Luo, G. J., Zhang, T. C., and Lei, J. J. (2018). Investigation and taxonomy of wild *Fragaria* resources in Tibet, China. *Genet. Resour. Crop Evol.* 65, 405–415. doi: 10.1007/s10722-017-0541-1

Han, Y., Wu, H. B., and Liu, Y. (2018). The complete chloroplast genome sequence of *Fragaria orientalis* (Rosales: Rosaceae). *Mitochondrial DNA B Resour.* 3, 127–128. doi: 10.1080/23802359.2018.1424578

Harrison, R. E., Luby, J. J., and Furnier, G. R. (1997). Chloroplast DNA restriction fragment variation among strawberry (*Fragaria* spp.) taxa. *J. Am. Soc. Hortic. Sci.* 122, 63–68.

Hilu, K. W., Black, C., Diouf, D., and Burleigh, J. G. (2008). Phylogenetic signal in *mat*K vs. *trn*K: a case study in early diverging eudicots (angiosperms). *Mol. Phylogenet. Evol.* 48, 1120–1130. doi: 10.1016/j.ympev.2008.05.021

Hu, G. L., Cheng, L. L., Huang, W. G., Cao, Q. C., Zhou, L., Jia, W. S., et al. (2020). Chloroplast genomes of seven species of Coryloideae (Betulaceae): structures and comparative analysis. *Genome* 63, 337–348. doi: 10.1139/gen-2019-0153

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, H., Shi, C., Liu, Y., Mao, S. Y., and Gao, L. Z. (2014). Thirteen camellia chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. *BMC Evol. Biol.* 14:151. doi: 10.1186/1471-2148-14-151

Hummer, K. E. (2012). A new species of *Fragaria* (Rosaceae) from Oregon. *J. Bot. Res. Inst. Texas* 6, 9–15.

Islam, S. U., Dar, T., Khuroo, A. A., Bhat, B. A., and Malik, A. H. (2021). DNA barcoding aids in identification of adulterants of *trillium govanianum* Wall. ex D. Don. *J. Appl. Res. Med. Aroma* 23:100305. doi: 10.1016/j.jarmap.2021.100305

Jansen, R. K., Cai, Z., Raubeson, L. A., Daniell, H., and Boore, J. L. (2007). Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proc. Natl. Acad. Sci. U. S. A.* 104, 19369–19374. doi: 10.1073/pnas.0709121104

Jansen, R. K., Raubeson, L. A., Boore, J. L., Depamphilis, C. W., and Cui, L. (2005). Methods for obtaining and analyzing whole chloroplast genome sequences. *Methods Enzymol.* 395, 348–384. doi: 10.1016/S0076-6879(05)95020-9

Jeon, J. H., and Kim, S. C. (2019). Comparative analysis of the complete chloroplast genome sequences of three closely related east-Asian wild roses (Rosa sect. Synstylae; Rosaceae). *Genes* 10:23. doi: 10.3390/genes10010023

Jin, J. J., Yu, W. B., Yang, J. B., Song, Y., de Pamphilis, 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

Johnson, A. L., Govindarajulu, R., and Ashman, T. (2014). Bioclimatic evaluation of geographical range in *Fragaria* (Rosaceae): consequences of variation in breeding system, ploidy and species age. *Bot. J. Linn. Soc.* 176, 99–114. doi: 10.1111/boj.12190

Kamneva, O. K., Syring, J., Liston, A., and Rosenberg, N. A. (2017). Evaluating allopolyploid origins in strawberries (*Fragaria*) using haplotypes generated from target capture sequencing. *BMC Evol. Biol.* 17:180. doi: 10.1186/s12862-017-1019-7

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

Kawabe, A., Nukii, H., and Furihata, H. Y. (2018). Exploring the history of chloroplast capture in *Arabis* using whole chloroplast genome sequencing. *Int. J. Mol. Sci.* 19:602. doi: 10.3390/ijms19020602

Kim, H. T., Kim, J. S., Lee, Y. M., Mun, J. H., and Kim, J. H. (2019). Molecular markers for phylogenetic applications derived from comparative plastome analysis of Prunus species. *J. Syst. Evol.* 57, 15–22. doi: 10.1111/jse.12453

Kode, V., Mudd, E. A., Iamtham, S., and Day, A. (2005). The tobacco plastid *accD* gene is essential and is required for leaf development. *Plant J.* 44, 237–244. doi: 10.1111/j.1365-313X.2005.02533.x

Lei, W. J., Ni, D. P., Wang, Y. J., Shao, J. J., Wang, X. C., Yang, D., et al. (2016). Intraspecific and heteroplasmic variations, gene losses and inversions in the chloroplast genome of *Astragalus membranaceus*. *Sci. Rep.* 6:21669. doi: 10.1038/srep21669

Lei, J. J., Xue, L., Guo, R. X., and Dai, H. P. (2017). The *Fragaria* species native to China and their geographical distribution. *Acta Hortic.* 1156, 37–46. doi: 10.17660/ActaHortic.2017.1156.5

Li, Y., Zhang, Z. R., Yang, J. B., and Lv, G. H. (2018). Complete chloroplast genome of seven *Fritillaria* species, variable DNA markers identification and phylogenetic relationships within the genus. *PLoS One* 13:e0194613. doi: 10.1371/journal.pone.0194613

Liu, X. L., Wen, J., Ni, Z. L., Johnson, G., Liang, Z. S., and Chang, Z. Y. (2013). Polyphyly of the Padus group of *Prunus* (Rosaceae) and the evolution of biogeographic disjunctions between eastern Asia and eastern North America. *J. Plant Res.* 126, 351–361. doi: 10.1007/s10265-012-0535-1

Liu, E. X., Yang, C. Z., Liu, J. D., Jin, S. R., Harijati, N., Hu, Z. L., et al. (2019). Comparative analysis of complete chloroplast genome sequences of four major *Amorphophallus* species. *Sci. Rep.* 9:809. doi: 10.1038/s41598-018-37456-z

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

Ma, J., Yang, B. X., Zhu, W., Sun, L. L., Tian, J. K., and Wang, X. M. (2014). The complete chloroplast genome sequence of *Mahonia bealei* (Berberidaceae) reveals a significant expansion of the inverted repeat and phylogenetic relationship with other angiosperms. *Gene* 528, 120–131. doi: 10.1016/j.gene.2013.07.037

Nazareno, A. G., Carlsen, M., and Lohmann, L. G. (2015). Complete chloroplast genome of *Tanaecium tetragonolobum*: the first bignoniaceae plastome. *PLoS One* 10:e0129930. doi: 10.1371/journal.pone.0129930

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

Njuguna, W., and Bassil, N. V. (2011). DNA barcoding: unsuccessful for species identification in *Fragaria* L. *Acta Hortic.* 918, 349–356. doi: 10.17660/ActaHortic.2011.918.45

Njuguna, W., Listonb, A., Cronn, R., Ashman, T. L., and Bassil, N. (2013). Insights into phylogeny, sex function and age of *Fragaria* based on whole chloroplast genome sequencing. *Mol. Phylogenet. Evol.* 66, 17–29. doi: 10.1016/j.ympev.2012.08.026

Okuyama, Y., Fujii, N., Wakabayashi, M., Kawakita, A., Ito, M., Watanabe, M., et al. (2005). Nonuniform concerted evolution and chloroplast capture: heterogeneity of observed introgression patterns in three molecular data partition phylogenies of Asian *Mitella* (Saxifragaceae). *Mol. Biol. Evol.* 22, 285–296. doi: 10.1093/molbev/msi016

Palmer, J. D. (1985). Comparative organization of chloroplast genomes. *Annu. Rev. Genet.* 19, 325–354. doi: 10.1146/annurev.ge.19.120185.001545

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

Phi, T. C. M., Chu, H. H., Le, N. T., and Nguyen, D. B. (2020). Phylogenetic relationship of *Paramignya trimera* and its relatives: an evidence for the wide sexual compatibility. *Sci. Rep.* 10:21662. doi: 10.1038/s41598-020-78448-2

Plunkett, G. M., and Downie, S. R. (2000). Expansion and contraction of the chloroplast inverted repeat in Apiaceae subfamily Apioideae. *Syst. Bot.* 25, 648–667. doi: 10.2307/2666726

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., Luby, J. J., and Harrison, R. E. (2000). Phylogenetic relationships among species of *Fragaria* (Rosaceae) inferred from non-coding nuclear and chloroplast DNA sequences. *Syst. Bot.* 25, 337–348. doi: 10.2307/2666646

Raubeson, L. A., Peery, R., Chumley, T. W., Dziubek, C., Fourcade, H. M., Boore, J. L., et al. (2007). Comparative chloroplast genomics: analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. *BMC Genomics* 8:174. doi: 10.1186/1471-2164-8-174

Rousseau-Gueutina, M., Gastona, A., Aïnoucheb, A., Aïnoucheb, M. L., Olbricht, K., Staudt, G., et al. (2009). Tracking the evolutionary history of polyploidy in *Fragaria* L. (strawberry): new insights from phylogenetic analyses of low-

copy nuclear genes. *Mol. Phylogenet. Evol.* 51, 515–530. doi: 10.1016/j.ympev.2008.12.024

Rozas, J., Ferrer-Mata, A., Sánchez-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.* 34, 3299–3302. doi: 10.1093/molbev/msx248

Ruhsam, M., Rai, H. S., Mathews, S., Ross, T. G., Graham, S. W., Raubeson, L. A., et al. (2015). Does complete plastid genome sequencing improve species discrimination and phylogenetic resolution in *Araucaria*? *Mol. Ecol. Resour.* 15, 1067–1078. doi: 10.1111/1755-0998.12375

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

Song, Y., Wang, S. J., Ding, Y. M., Xu, J., Li, M. F., Zhu, S. F., et al. (2017). Chloroplast genomic resource of *Paris* for species discrimination. *Sci. Rep.* 7:3427. doi: 10.1038/s41598-017-02083-7

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

Staudt, G. (1962). Taxonomic studies in the genus *Fragaria*. *Can. J. Bot.* 40, 869–886. doi: 10.1139/b62-081

Staudt, G. (1989). The species of *Fragaria*, their taxonomy and geographical distribution. *Acta Hortic.* 265, 23–33. doi: 10.17660/ActaHortic.1989.265.1

Staudt, G. (2003). Notes on Asiatic *Fragaria* species: III. *Fragaria orientalis* Losinsk. and *Fragaria mandshurica* spec. nov. *Bot. Jahrb.* 124, 397–419. doi: 10.1127/0006-8152/2003/0124-0397

Staudt, G. (2006). Himalayan species of *Fragaria* (Rosaceae). *Bot. Jahrb.* 126, 483–508. doi: 10.1127/0006-8152/2006/0126-0483

Staudt, G. (2009). Strawberry biogeography, genetics and systematics. *Acta Hortic.* 842, 71–84. doi: 10.17660/ActaHortic.2009.842.1

Sun, J., Sun, R., Liu, H., Chang, L. L., Li, S. T., Zhao, M. Z., et al. (2021). Complete chloroplast genome sequencing of ten wild *Fragaria* species in China provides evidence for phylogenetic evolution of *Fragaria*. *Genomics* 113, 1170–1179. doi: 10.1016/j.ygeno.2021.01.027

Tegally, A., Jaufeerally-Fakim, Y., and Dulloo, M. E. (2019). Molecular characterisation of *Solanum melongena* L. and the crop wild relatives, *S. violaceum* Ortega and *S. torvum* Sw., using phylogenetic/DNA barcoding markers. *Genet. Resour. Crop. Evol.* 66, 1625–1634. doi: 10.1007/s10722-019-00827-0

Terakami, S., Matsumura, Y., Kurita, K., Kanamori, H., Katayose, Y., Yamamoto, T., et al. (2012). Complete sequence of the chloroplast genome from pear (*Pyrus pyrifolia*): genome structure and comparative analysis. *Tree Genet. Genomes* 8, 841–854. doi: 10.1007/s11295-012-0469-8

Tsitrone, A., Kirkpatrick, M., and Levin, D. A. (2003). A model for chloroplast capture. *Evolution* 57, 1776–1782. doi: 10.1111/j.0014-3820.2003.tb00585.x

Tulipani, S., Mezzetti, B., Capocasa, F., Bompadre, S., Beekwilder, J., de Voset, C. H. A., et al. (2008). Antioxidants, phenolic compounds, and nutritional quality of different strawberry genotypes. *J. Agric. Food Chem.* 56, 696–704. doi: 10.1021/jf0719959

Tyagi, S., Jung, J. A., Kim, J. S., and Won, S. Y. (2020). Comparative analysis of the complete chloroplast genome of mainland *Aster spathulifolius* and other *Aster* species. *Plants* 9:568. doi: 10.3390/plants90505680

Urrutia, M., Rambla, J. L., Alexiou, K. G., Granell, A., and Monfort, A. (2017). Genetic analysis of the wild strawberry (*Fragaria vesca*) volatile composition. *Plant Physiol. Biochem.* 121, 99–117. doi: 10.1016/j.plaphy.2017.10.015

Wang, S., Shi, C., and Gao, L. Z. (2013). Plastid genome sequence of a wild woody oil species, *Prinsepia utilis*, provides insights into wvolutionary and mutational patterns of Rosaceae chloroplast genomes. *PLoS One* 8:e73946. doi: 10.1371/journal.pone.0073946

Wang, Y. H., Qu, X. J., Chen, S. Y., Li, D. Z., and Yi, T. S. (2017). Plastomes of Mimosoideae: structural and size variation, sequence divergence, and phylogenetic implication. *Tree Genet. Genomes* 13:41. doi: 10.1007/s11295-017-1124-1

Wei, W., Zheng, Y. L., Chen, L., Wei, Y. M., Yan, Z. H., and Yang, R. W. (2006). PCR-RFLP analysis of cpDNA and mtDNA in the genus *Houttuynia* in some areas of China. *Hereditas* 142, 24–32. doi: 10.1111/j.1601-5223.2005.01704.x

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

Wolfe, K. H., Li, W., and Sharp, P. M. (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc. Nati. Acad. Sci. U. S. A.* 84, 9054–9058. doi: 10.1073/pnas.84.24.9054

Xin, T. Y., Yao, H., Gao, H. H., Zhou, X. Z., Ma, X. C., Xu, C. Q., et al. (2013). Super food *Lycium barbarum* (Solanaceae) traceability via an internal transcribed spacer 2 barcode. *Food Res. Int.* 54, 1699–1704. doi: 10.1016/j.foodres.2013.10.007

Yang, Y., and Davis, T. M. (2017). A new perspective on polyploid *Fragaria* (strawberry) genome composition based on large-scale, multi-locus phylogenetic analysis. *Genome Biol. Evol.* 9, 3433–3448. doi: 10.1093/gbe/evx214

Yang, J. Y., Wei, S. J., Su, D. B., Chen, S. Y., Luo, Z. W., Shen, X. M., et al. (2020). Molecular identification and evolutionary characteristics of *Fragaria nilgerrensis* in Yunnan based on nrDNA ITS and cpDNA *psb*A-*trn*H sequence analysis. *J. South Agric.* 51, 748–757.

Zhang, Y., Du, L., Liu, A., Chen, J. J., Wu, L., Hu, W. M., 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

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. B., Yin, J. L., Guo, H. Y., Zhang, Y. Y., Xiao, W., Sun, C., et al. (2015). The complete chloroplast genome provides insight into the evolution and polymorphism of *Panax ginseng*. *Front. Plant Sci.* 5:696. doi: 10.3389/fpls.2014.00696

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

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

frontiers in Plant Science

ORIGINAL RESEARCH
published: 25 October 2021
doi: 10.3389/fpls.2021.767667

Family-Wide Evaluation of Multiple C2 Domain and Transmembrane Region Protein in *Gossypium hirsutum*

Qianqian Hu[1†], Mengting Zeng[1†], Miao Wang[1], Xiaoyu Huang[1], Jiayi Li[2], Changhui Feng[3], Lijie Xuan[2], Lu Liu[2] and Gengqing Huang[1,4]**

[1]Hubei Key Laboratory of Genetic Regulation and Integrative Biology, School of Life Sciences, Central China Normal University, Wuhan, China, [2]Joint Center for Single Cell Biology, School of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai, China, [3]Institute of Cash Crops, Hubei Academy of Agricultural Sciences, Wuhan, China, [4]Xinjiang Key Laboratory of Special Species Conservation and Regulatory Biology, College of Life Science, Xinjiang Normal University, Ürümqi, China

OPEN ACCESS

Edited by:
Hai Du,
Southwest University, China

Reviewed by:
Yang Zhu,
Zhejiang University, China
Hantao Wang,
Cotton Research Institute, Chinese Academy of Agricultural Sciences (CAAS), China
Zuo Ren Yang,
Institute of Cotton Research, Chinese Academy of Agricultural Sciences (CAAS), China

***Correspondence:**
Gengqing Huang
gqhuang@mail.ccnu.edu.cn
Lu Liu
lu.liu@sjtu.edu.cn

[†]These authors have contributed equally to this work

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

Received: 31 August 2021
Accepted: 27 September 2021
Published: 25 October 2021

Citation:
Hu Q, Zeng M, Wang M, Huang X, Li J, Feng C, Xuan L, Liu L and Huang G (2021) Family-Wide Evaluation of Multiple C2 Domain and Transmembrane Region Protein in Gossypium hirsutum. Front. Plant Sci. 12:767667. doi: 10.3389/fpls.2021.767667

Multiple C2 domain and transmembrane region proteins (MCTPs) are a group of evolutionarily conserved proteins and show emerging roles in mediating protein trafficking and signaling transduction. Although, several studies showed that MCTPs play important roles during plant growth and development, their biological functions in cotton remain largely unknown. Here, we identify and characterize 33 *GhMCTP* genes from upland cotton (*Gossypium hirsutum*) and reveal the diverse expression patterns of *GhMCTPs* in various tissues. We also find that *GhMCTP7*, *GhMCTP12*, and *GhMCTP17* are highly expressed in the main stem apex, suggesting their possible roles in shoot development. Through analyzing different cotton species, we discover plant heights are closely related to the expression levels of *GhMCTP7*, *GhMCTP12*, and *GhMCTP17*. Furthermore, we silence the expression of *GhMCTP* genes using virus-induced gene silencing (VIGS) system in cotton and find that *GhMCTP7*, *GhMCTP12*, and *GhMCTP17* play an essential role in shoot meristem development. GhMCTPs interact with GhKNAT1 and GhKNAT2 and regulate meristem development through integrating multiple signal pathways. Taken together, our results demonstrate functional redundancy of GhMCTPs in cotton shoot meristem development and provide a valuable resource to further study various functions of GhMCTPs in plant growth and development.

Keywords: cotton, main stem apex development, GhMCTP, gene expression, protein interaction, KNOX family protein

INTRODUCTION

The development of multicellular organisms is regulated by various signaling pathways. These signaling transduction events are mediated by extensive intercellular and intracellular membrane trafficking processes, which control the signal perception and the trafficking of signal molecules from one compartment to another. Thus, membrane trafficking regulates the processing,

modification, and secretion of signal molecules, or conversely, their translocation to the nucleus (Shilo and Schejter, 2011; Van Norman et al., 2011).

C2 domain is one of the most prevalent eukaryotic lipid-binding domains and could serve as a docking module that targets proteins to a specific intracellular membrane (Nalefski and Falke, 1996; Cho and Stahelin, 2006). A large number of C2 domain-containing proteins have been identified, and most of them are involved in membrane trafficking and signal transduction (Corbalan-Garcia and Gómez-Fernández, 2014). Multiple C2 domain and transmembrane region proteins (MCTPs) are evolutionarily conserved in eukaryotic organisms, containing 3–4 C2 domains at the N-terminus and transmembrane regions at the C-terminus (Shin et al., 2005; Lek et al., 2012). MCTPs mediate the trafficking of key regulators and are essential for signaling transduction in diverse species, thus regulating various developmental processes (Liu et al., 2013; Genç et al., 2017). MCTPs also function as unique membrane tethers controlling endoplasmic reticulum (ER)-plasma membrane (PM) contact specifically at plasmodesmata and regulate cell-to-cell communication (Brault et al., 2019).

The function of MCTP was first identified in *Caenorhabditis elegans* in a high-throughput RNAi screening. Genetic mutants of *MCTP* were embryonic lethal (Maeda et al., 2001). In addition, different alleles of *mctp* mutants in *Drosophila* were isolated and showed to regulate various developmental processes, including larval development and neurotransmission, suggesting multiple roles of MCTPs in different developmental stages (Tunstall et al., 2012; Genç et al., 2017). However, the molecular functions of MCTPs in regulating these processes were still largely unknown.

The invertebrate organisms *C. elegans* and *Drosophila melanogaster* contain a single *MCTP* gene. In all plant lineages, the number of MCTP repertoire significantly increases, and each of MCTPs exhibits distinct or overlapping patterns of gene expression and subcellular protein localization (Liu et al., 2018a; Hao et al., 2020; Zhu et al., 2020), suggesting more diverse and specific functions of MCTPs in regulating multiple cellular and developmental processes. Several members of MCTP proteins have been identified in plants to mediate the intercellular and intracellular trafficking of various macromolecules. QUIRKY (QKY) and FT-INTERACTING PROTEIN 1 (FTIP1) belong to the *MCTP* family, mediate the trafficking of florigen protein FLOWERING LOCUS T (FT) from companion cells to sieve elements, thus regulating flowering time in *Arabidopsis* (Liu et al., 2012, 2019). QKY also interacts with and stabilizes a leucine-rich repeat receptor-like kinase SCRAMBLED (SCM), and is required for proper cell-type patterning and organogenesis (Trehin et al., 2013; Vaddepalli et al., 2014; Song et al., 2019; Mergner et al., 2020). In addition, two other MCTP proteins, FTIP3 and FTIP4, interact with key meristem regulator SHOOTMERISTEMLESS (STM) and control its subcellular localization and intercellular trafficking in the shoot apex, thus determining the fate of shoot apical meristem (Liu et al., 2018b).

Multiple C2 domain and transmembrane region proteins also regulate multiple developmental processes in other plant species. OsFTIP1, the closet ortholog of FTIP1 in rice, mediates the flowering transition by affecting the trafficking of RICE FLOWERING LOCUS T1 (RFT1) from companion cells to sieve elements (Song et al., 2017). OsFTIP1 also determines the nuclear localization of rice MOTHER OF FT AND TFL1 (OsMFT1) and promotes drought tolerance (Chen et al., 2021). Another MCTP protein in rice OsFTIP7 facilitates nuclear translocation of a homeodomain transcription factor, *Oryza sativa* homeobox 1 (OSH1), to determine the auxin-mediated anther dehiscence in rice (Song et al., 2018). ZmCpd33, the closet homolog of QKY in maize, promotes sucrose export from companion cells into sieve elements (Tran et al., 2019). DOFTIP1, the orchid orthologs of FTIP1, plays an important role in promoting flowering in the orchid *Dendrobium Chao Praya Smile* (Wang et al., 2017). These results demonstrate that MCTPs are involved in diverse protein trafficking events and regulate plant development.

Upland cotton (*Gossypium hirsutum*) is an important economic crop in the world, and it is the primary fiber crop and an important oil crop (John and Crow, 1992). The architecture of cotton plants is determined primarily by their plant heights, shoot branching patterns, and flowering patterns, all of which directly affect cotton planting strategies, yield, planting area, mechanized harvesting suitability, and cotton planting costs (Reinhardt and Kuhlemeier, 2002; Su et al., 2018). In *G. hirsutum*, *GhMCTPs* have been genome-widely identified and the gene expression patterns have been analyzed (Hao et al., 2020); however, the biological functions of GhMCTPs in *G. hirsutum* are still largely unknown.

In this study, we systematically investigated MCTPs in cotton and analyzed their gene expressions in various developing cotton tissues. We further characterized the function of *GhMCTP7*, *GhMCTP12*, and *GhMCTP17* in shoot development. GhMCTP7, GhMCTP12, and GhMCTP17 show distinct or overlapping subcellular localization patterns in *Nicotiana benthamiana* leaf epidermal cells. They interact with GhKNAT1 and GhKNAT2 and regulate shoot apical meristem development through integrating multiple signaling pathways. Our results demonstrate the functional redundancy of GhMCTPs in shoot development and reveal a gene regulatory framework that determines the meristem fate, providing a valuable resource for cotton architecture improvement.

MATERIALS AND METHODS

Plant Materials

The seeds of upland cotton (*G. hirsutum "coker 312"*) were surface sterilized with 70% (v/v) ethanol for 1 min, and then with 10% hydrogen peroxide for 2 h, followed by washing with sterile water several times. The sterilized seeds were germinated on one-half strength Murashige and Skoog (MS) medium (12-h-light/12-h-dark cycle, 28°C), and seedlings were transplanted to the soil for further growth. The roots, stems, main stem apex (MSA, apex length is about 5 mm), young leaves of three-leaf stage cotton plants, and 10 DPA (days post-anthesis) cotton fiber after flowering were harvested for RNA extraction.

The shoot apexes (about 1 cm) of eight upland cotton cultivars (*G. hirsutun* Okra, Emian JD1718, Lumian 1, Jimian 958, Emian JB2150, Emian SJA146, Emian JC1751, Zaosong2) were harvested for RNA extraction when these plants were flowering. The height of plants was calculated for each cultivar ($n = 50$) when cotton bolls were open.

Sequence Analysis

A BLASTP search was performed using the protein sequences of 16 *Arabidopsis* MCTPs as query sequences on the website of COTTONGEN[1] by chosen the *G. hirsutum* (AD1) ZJUv2.1 proteins (totally 72,761 protein sequences). The proteins with high sequence similarity (E-value = 0) were selected as putative GhMCTPs. All GhMCTP sequences were then manually searched against MotifScan,[2] InterProScan,[3] and SMART,[4] to confirm the sequence containing the C2 domain and PRT_C domain. Finally, a total of 33 GhMCTP members were identified. The chromosomal location, amino acid length, protein molecular mass, and isoelectric point of the 33 GhMCTPs were analyzed using COTTONGEN[5] and ExPASy ProtParam.[6] DNA and protein sequences were analyzed using DNASTAR software (DNAStar, MD, United States).

Phylogenetic Analysis

The protein sequences of 33 GhMCTPs were used as queries to identify GhMCTP homologs in *Gossypium raimondii* and *Gossypium arboretum* from COTTONGEN,[7] and different plant species from Phytozome v12.[8] A phylogenetic tree of deduced GhMCTP amino acid sequences was constructed using the neighbor-joining algorithm with default parameters, with 1,000 bootstrap replicates in MEGA_X_10.2.4 (Kumar et al., 2016).[9]

Gene Structure and Chromosomal Mapping

The Gene Structure Display Server Program[10] was used to draw the exon-intron structure of *GhMCTP* genes based on the full-length genome sequence and the corresponding coding sequences. Domain analysis of all MCTP proteins from various plant species was performed by InterProScan. The chromosomal location information of all MCTP genes was derived from the annotation information downloaded on the COTTONGEN (see footnote 5; **Supplementary Table 1**).[11] Based on the location information of *GhMCTPs*, we manually drew the chromosome map using Photoshop software.

[1]https://www.cottongen.org/blast/protein/protein
[2]http://myhits.isb-sib.ch/cgi-bin/motif_scan
[3]http://www.ebi.ac.uk/Tools/pfa/iprscan
[4]http://smart.emblheidelberg.de/
[5]https://www.cottongen.org/find/genes
[6]http://us.expasy.org/tools/protparam.html
[7]https://www.cottongen.org/
[8]https://phytozome.jgi.doe.gov/pz/portal.html
[9]https://www.megasoftware.net
[10]http://gsds.cbi.pku.edu.cn/
[11]https://www.cottongen.org/species/Gossypium_hirsutum/ZJU-AD1_v2.1

Heat-Map Analysis of Gene Expression

The reads per kb per million reads (RPKM) values denoting the expression levels of *GhMCTP* genes were obtained from a comprehensive profile of the TM-1 transcriptome data (Trapnell et al., 2012; Zhang et al., 2015),[12] and the expression data of main stem apex were generated in this study. A heat-map analysis was performed using *TBtools* (Chen et al., 2020).[13]

Expression Analysis

Total RNA was extracted from cotton roots, stems, leaves, ovules, and the 10 days fiber after flowering using the RNAprep Pure Plant kit (TIANGEN, Beijing, China) and reverse transcribed using Moloney Murine Leukemia Virus Reverse Transcriptase (Promega, Madison, Wisconsin, United States) according to the manufacturer's instructions. Quantitative real-time PCR was performed using MJ Research DNA Engine Option 2 detection system with the fluorescent intercalating dye SYBR-Green (Toyobo). The relative expression levels were normalized to a cotton polyubiquitin gene (*GhUBI1*, GenBank accession no. EU604080).

A two-step PCR procedure was performed in all experiments using the previously described method (Huang et al., 2013). The expressions of the putative target genes were determined using the comparative cycle threshold method. To achieve optimal amplification, PCR conditions for each set of primers were optimized for annealing temperature and Mg^{2+} concentration. Data presented in the quantitative real-time PCR (qRT-PCR) analysis are the mean and SD of three biological replicates of plant materials and three technical replicates in each biological sample using gene-specific primers. Primers used for qRT-PCR are designed to target *GhMCTP-A* and *GhMCTP-D* simultaneously and are listed in **Supplementary Table 2**.

Yeast Two-Hybrid Assay

The N-terminal fragments of *GhMCTP7*, *GhMCTP12*, and *GhMCTP17* devoid of the sequences encoding the transmembrane regions were amplified and cloned into pGADT7 (Prey vector, Clontech). The coding sequences of *GhKNAT1* and *GhKNAT2* were amplified and cloned into pGBKT7 (Bait vector, Clontech). The prey and bait vectors were transformed into AH109 and Y187 cells, respectively. After mating, all transformed cells were grown on a Synthetic Defined-Ade/-His/-Trp/-Leu medium for interaction tests. Primers used are listed in **Supplementary Table 2**.

Luciferase Complementation Imaging Assay

The coding sequences of *GhMCTP7*, *GhMCTP12*, and *GhMCTP17* were amplified and cloned into the N-terminal Luciferase fusion vector JW771-N-terminal luciferase fragment (nLUC). The coding sequences of *GhKNAT1* and *GhKNAT2* were amplified and cloned into the C-terminal Luciferase fusion vector JW772-C-terminal luciferase fragment (cLUC). The resulting plasmids

[12]http://structuralbiology.cau.edu.cn/gossypium
[13]https://github.com/CJ-Chen/TBtools

were transformed into *Agrobacterium* cells, which were infiltrated into *N. benthamiana* leaves. The luciferase complementation imaging (LCI) assay was performed as described previously (Chen et al., 2008). The LUC luminescence signal was visualized using a cryogenically cooled CCD camera (Night Shade LB985, Berthold Technologies) and indiGO software. Primers used are listed in **Supplementary Table 2**.

Agrobacterium tumefaciens-Mediated VIGS

Overnight *Agrobacterium* cultures with the desired tobacco rattle virus (TRV) vectors, including *TRV2:GhMCTP7*, *TRV2:GhMCTP12*, *TRV2:GhMCTP17*, and *TRV2:GhKNAT1/2*, were infiltrated into two fully expanded cotyledons of 10-day-old cotton plants grown at 22–24°C as described previously (Gao et al., 2013). At least 15 plants were inoculated for each construct. The *TRV2:GhCLA1* construct was included as a visual marker for virus-induced gene silencing (VIGS) efficiency. To simultaneously silence the expression of *GhMCTP7*, *GhMCTP12*, and *GhMCTP17*, *Agrobacterium* cultures containing *TRV2:GhMCTP7*, *TRV2:GhMCTP12*, and *TRV2:GhMCTP17* were mixed and infiltrated. Primers used are listed in **Supplementary Table 2**.

RESULTS

Identification of MCTP Genes in Cotton

As MCTPs were first identified in *Arabidopsis* in plants (Liu et al., 2018a), we used 16 MCTP protein sequences in *Arabidopsis* as queries to search against the protein database of *G. hirsutum* to identify putative GhMCTPs in cotton using the BLASTP program in COTTONGEN [*G. hirsutum* (AD1) ZJUv2.1 proteins (72761)]. The identified MCTP members were further used as queries to search for other possible MCTP proteins in *G. hirsutum*. All these putative GhMCTP proteins were subjected for domain analysis to confirm that all these candidates contain 3–4 N-terminal C2 domains and phosphoribosyltransferase C-terminal region (PRT_C). Finally, we identified a total of 33 GhMCTP proteins in *G. hirsutum* (**Figure 1**).

We classified these 33 GhMCTPs into seven clades through phylogenetic analysis based on multiple sequence alignment (**Figure 1A**). In addition, we aligned all these GhMCTP protein sequences and identified 16 pairs of MCTP protein with high sequence similarity. The members from each pair are originated from the cotton A subgenome (At, where 't' stands from tetraploid) and D subgenome (Dt), respectively. The GhMCTPs were therefore classified as GhMCTP-A and GhMCTP-D.

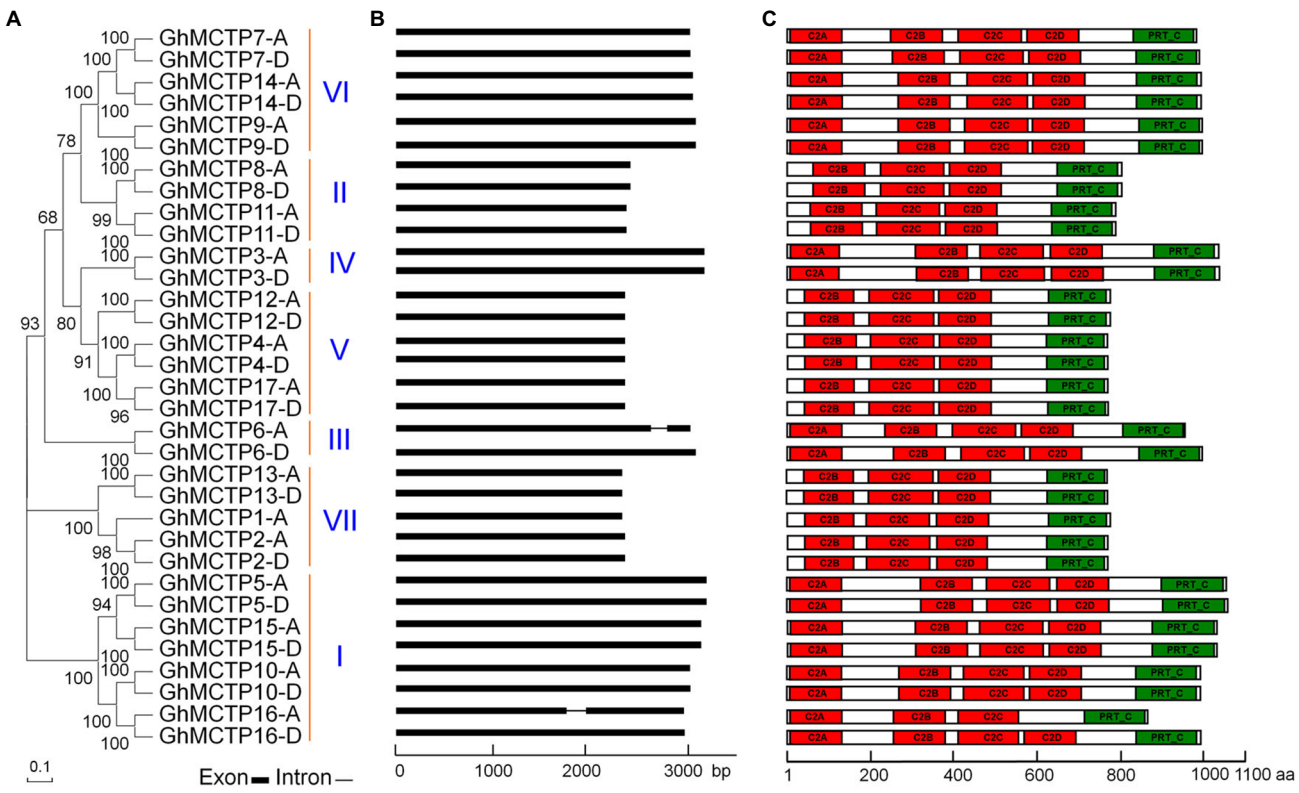

FIGURE 1 | Characterization of multiple C2 domain and transmembrane region proteins (*MCTP*) family proteins in upland cotton (*Gossypium hirsutum*). **(A)** Thirty-three GhMCTP proteins are classified into seven groups based on phylogenetic analysis of MCTP proteins in *G. hirsutum*. The phylogenetic tree was generated with MEGA X using the neighbor-joining algorithm. Numbers on the major branches indicate bootstrap values (>50%) in 1,000 replicates. **(B)** Schematic diagrams showing the gene structures of *GhMCTP* genes. The coding regions are indicated by black boxes. The intron is represented by a black line. **(C)** Protein motif analysis of GhMCTPs. The prediction of protein motifs is based on InterProScan. C2 domain and Phosphoribosyltransferase C-terminal are labeled as red and green boxes, respectively. bp, base pair; aa, amino acids.

The members in these two subgroups were further designated according to their locations on the chromosome (**Supplementary Figure 1**; **Supplementary Table 1**), according to the naming principle in other cotton genome studies (Paterson et al., 2012; Huang et al., 2020).

We analyzed the gene structures of all *GhMCTPs* and found most of *GhMCTP* genes do not contain introns (**Figure 1B**). Similar to MCTPs in *Arabidopsis*, most GhMCTP proteins contain 1–4 C-terminal transmembrane regions except for GhMCTP3 (**Supplementary Figure 2**), suggesting the functional conservation and divergence among MCTPs. We also searched the MCTP proteins in the diploid cotton *Gossypium arboreum* and *Gossypium raimondii* (Paterson et al., 2012; Huang et al., 2020), and identified 17 and 18 MCTP members from *G. arboreum* genome (A-genome) and *G. raimondii* genome (D-genome), respectively (**Supplementary Table 1**).

Phylogenetic Analysis of MCTP Homologs in Plants

To identify MCTP homologs in plant genomes, we performed BLASTP or TBLASTN search in protein and genome databases in Phytozome 12 using MCTP1 as a query sequence. We obtained MCTP homologs in different land plants and algae including dicotyledons (*G. hirsutum*, *G. arboreum*, *G. raimondii*, *Arabidopsis*, *Aquilegia coerulea*, *Amaranthus hypochondriacus*, *olanum lycopersicum*, *Eucalyptus grandis*, *Populus trichocarpa*, and *Medicago truncatula*), monocotyledons (*Ananas comosus*, *Oryza sativa*, and *Zea mays*), moss (*Physcomitrella patens*), fern (*Selaginella moellendorffii*), and algaes (*Chlamydomonas reinhardtii* and *Micromonas pusilla*; **Supplementary Table 3**). To understand the evolutionary relationships among MCTPs in plants, MCTP homologs in different species were analyzed in detail using neighbor-joining and maximum likelihood methods, and the unrooted phylogenetic tree was constructed (**Figure 2A**).

The MCTP homologs could be classified into seven clades, namely clade I to clade VII. Each clade contains MCTPs from eudicot plants (**Figure 2A**). The clade I was the largest branch containing 41 members, which are from different plant species including eudicots, monocots, fern, and moss. Clade III is the smallest one with only eight members. Furthermore, clade VII is specific for eudicots, and clade I is the only clade containing MCTPs from moss (*P. patens*) and fern (*S. moellendorffii*; **Figure 2A**). It is noteworthy that MCTPs exist in many plant lineages (**Figure 2B**), suggesting the fundamental roles of MCTPs in plant development. The number of MCTPs is greatly expanded in dicotyledons and monocotyledons (**Figure 2B**), indicating that the family members of MCTPs among species increase substantially after several rounds of whole-genome duplication, and may evolve to generate functional specialized MCTPs to respond to changing environmental stimuli.

Expression Analysis of *GhMCTPs* in Upland Cotton

To identify the potential roles of GhMCTPs in cotton development, the expression patterns of all *GhMCTPs* were investigated in various cotton tissues, including root, stem, leaf, main stem apex, torus, calycle, pistil petal, and stamen, through analyzing previously published transcriptome datasets (You et al., 2017). We also analyzed the expression of all *MCTPs* in ovules and fibers at several developmental stages (**Figure 3A**). Based on their expression profiles, *GhMCTPs* were generally divided into two groups. One group of *GhMCTPs* was highly expressed in almost all tissues, including *GhMCTP3-A/D*, *GhMCTP4-A/D*, *GhMCTP5-A/D*, *GhMCTP7-A/D*, *GhMCTP12-A/D*, *GhMCTP14-A/D*, and *GhMCTP17-A/D*. Members of the other groups were tissue-preferred genes (**Figure 3A**). For example, *GhMCTP1-A* and *GhMCTP2-A/D* were preferentially expressed in petals. *GhMCTP6-D*, *GhMCTP9-A/D*, *GhMCTP10-A/D*, and *GhMCTP16-A/D* were highly expressed in the main stem apex and ovules at early stages (**Figure 3A**). This tissue/organ-preferred expression pattern indicates that these MCTPs may function in specific developmental stages.

We further isolated different cotton tissues, including root, stem, leaf, main stem apex, and fibers at 10-day post-anthesis, and carried out qRT-PCR to investigate expression profiles of all *GhMCTPs* (**Figure 3B**). The results revealed that *GhMCTP* genes were expressed in all organs and tissues detected. Furthermore, we found that *GhMCTPs* from different subfamilies exhibited different expression patterns (**Figure 3**), suggesting that *GhMCTPs* might be involved in different cotton developmental processes.

GhMCTP7, GhMCTP12, and GhMCTP17 Function Additively to Regulate Cotton Shoot Meristem Development

The majority of *GhMCTP* genes were highly detected in the main stem apex (**Figures 3A,B**), suggesting their possible roles in meristem development. In *Arabidopsis*, two MCTP proteins, FT INTERACTING PROTEIN 3 (FTIP3) and FTIP4 have been reported to play an essential role in mediating shoot meristem development, thus determining the overall plant architecture (Liu et al., 2018b). GhMCTP12-A and GhMCTP12-D were the closest homologs of FTIP3 and FTIP4 in *G. hirsutum* (**Figure 2A**). In cotton, GhMCTP7-A/D and GhMCTP17-A/D showed sequence similarity with GhMCTP12-A/D (**Figure 2A**), and *GhMCTP7*, *GhMCTP12*, and *GhMCTP17* were all highly expressed in the main stem apex (**Figure 3**). These results suggest that they are potential candidates to regulate meristem development.

The effect of different cotton cultivars on plant height was highly significant. We chose eight allotetraploid cotton cultivars and divided them into two groups based on their plant heights (**Figure 4A**). To understand the roles of GhMCTPs in shoot development, we examined the expressions of *GhMCTP7*, *GhMCTP12*, and *GhMCTP17* in the shoot apex of selected allotetraploid cotton cultivars and found that the expression levels of *GhMCTP7*, *GhMCTP12*, and *GhMCTP17* were correlated with their plant heights (**Figures 4B–D**).

Virus-induced gene silencing is a powerful reverse genetic technology for quick functional characterizations of plant genes, which serves as an alternative to mutant collection or creating stable transgenic lines (Gao et al., 2013; Lange et al., 2013;

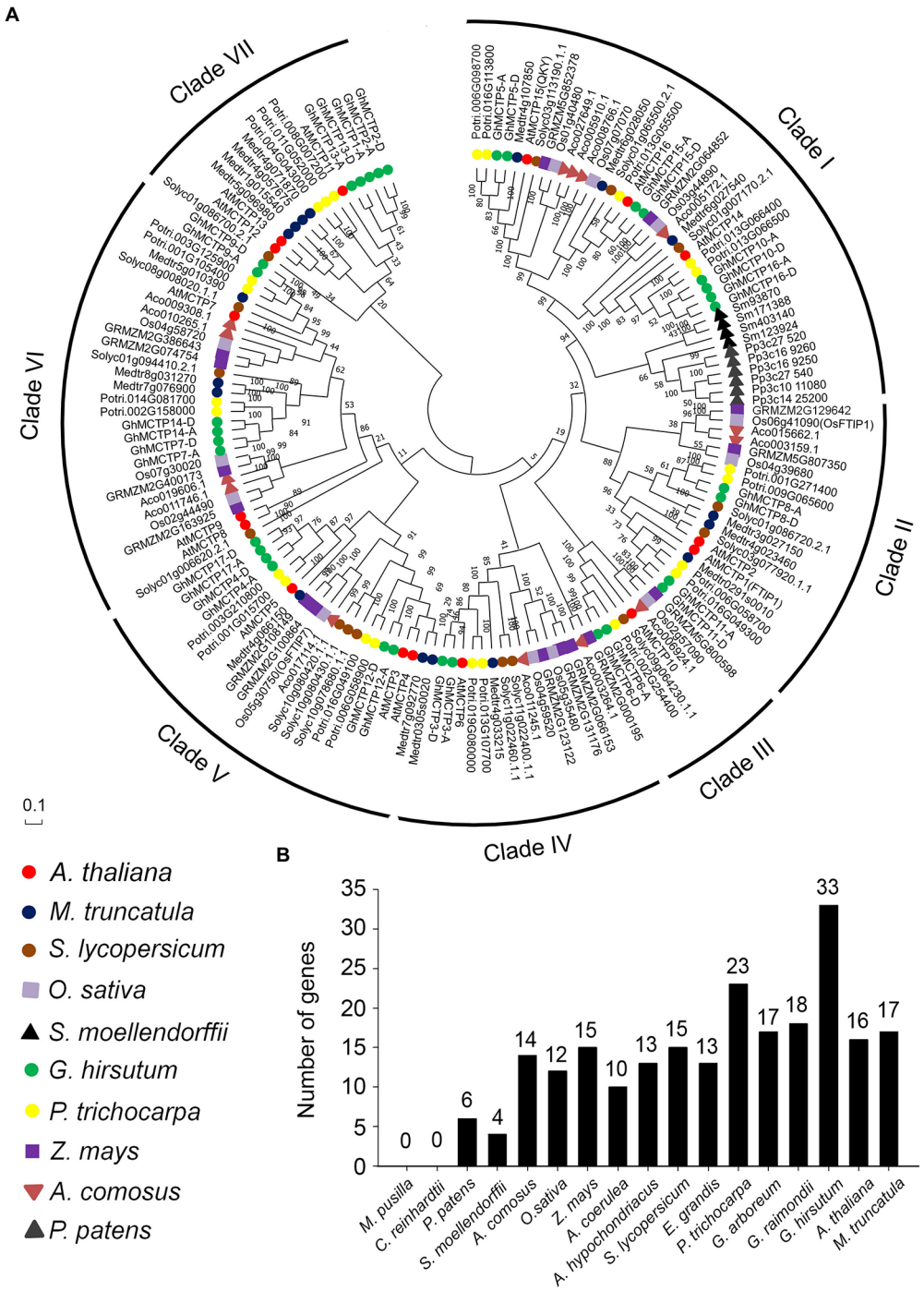

FIGURE 2 | Polygenetic relationships of MCTP homologs in different plant species. **(A)** Proteins from 10 different species (*Arabidopsis thaliana*, *Medicago truncatula*, *Solanum lycopersicum*, *Oryza sativa*, *Selaginella moellendorffii*, *G. hirsutum*, *Populus trichocarpa*, *Zea mays*, *Ananas comosus*, and *Physcomitrella patens*) are indicated by different icons and are classified into seven groups. All available gene names are also indicated. The level of statistical support was conducted by neighbor-joining method, and numbers on the major branches indicate bootstrap values. **(B)** Numbers of *MCTP* genes in different species.

McGarry et al., 2016). To further understand the biological functions of *GhMCTPs* in shoot development, we silenced the gene expression of *GhMCTP7*, *GhMCTP12*, and *GhMCTP17* using TRV-based VIGS technique. Real-time quantitative PCR assays showed that the expressions of *GhMCTPs* were downregulated in cotton main stem apex of *TRV2:GhMCTP* plants compared to negative control plants treated with *TRV2:00* (**Supplementary Figure 3**); however, most of the VIGS-mediated gene silencing lines exhibited mild dwarf phenotype compared to those of negative control plants (**Figures 5A–C**). We reasoned

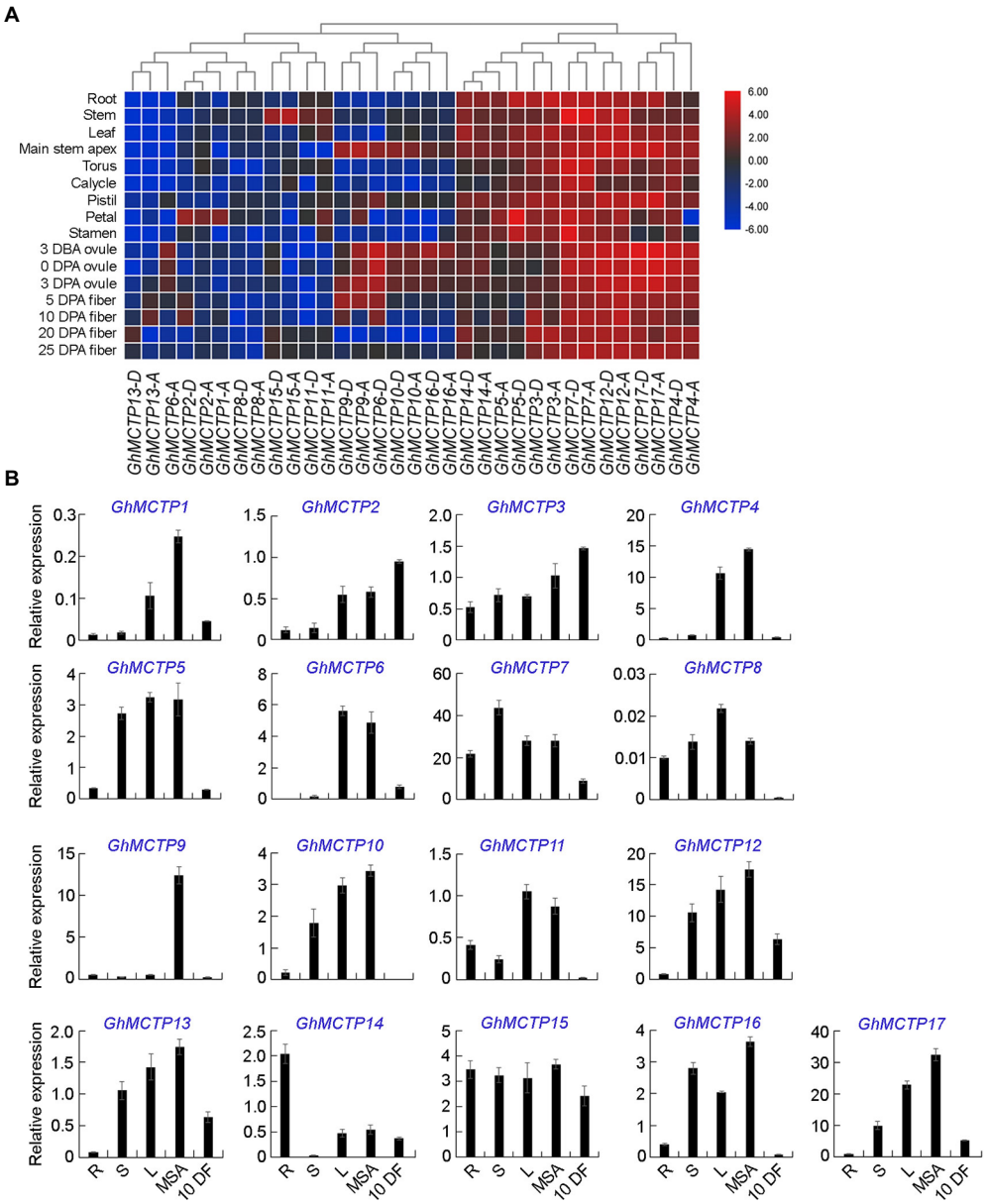

FIGURE 3 | Expression patterns of *GhMCTPs* in upland cotton. **(A)** Heat map analysis of *GhMCTP* gene expressions in different organs of upland cotton. The relative fold changes in gene expression for all *GhMCTP* genes were compared. The color from blue to red indicates low to high expression. DBA, days before-anthesis; DPA, days post-anthesis. **(B)** Quantitative real-time PCR (qRT-PCR) analysis of 17 *GhMCTPs* in various tissues of upland cotton. Results were normalized against the expression level of *GhUBI1*. R, root; S, stem; L, leaf; MSA, main stem apex; and 10 DF, fibers at 10-day post-anthesis. Error bars indicate SD.

that GhMCTP7, GhMCTP12, and GhMCTP17 might play a redundant role to regulate meristem development. Thus, we conducted a VIGS assay to simultaneously silence the expression of *GhMCTP7*, *GhMCTP12*, and *GhMCTP17* (**Supplementary Figure 3B**). The resulting plants exhibited a dwarf phenotype, with down-curly leaves and occasionally shoot branching (**Figure 5D**; **Supplementary Figure 4**). Furthermore, the longitudinal sections showed that *TRV2:GhMCTP7/12/17* plants had a narrow dome-shaped meristem, and the organs generated at the flanking of the meristems were also malformed and disordered positioned (**Figures 5E–H**). These results substantiate that *GhMCTP7*, *GhMCTP12*, and *GhMCTP17* are essential for shoot meristem development.

Subcellular Localization of GhMCTP7, GhMCTP12, and GhMCTP17

To determine the subcellular localization of GhMCTP7, GhMCTP12, and GhMCTP17, we transiently expressed their full-length open reading frames fused with the green fluorescent protein (GFP)-MCTPs reporter in *N. benthamiana* leaf epidermal cells and observed two different types of subcellular localization patterns (**Figure 6**; **Supplementary Figure 5**). GhMCTP7 was

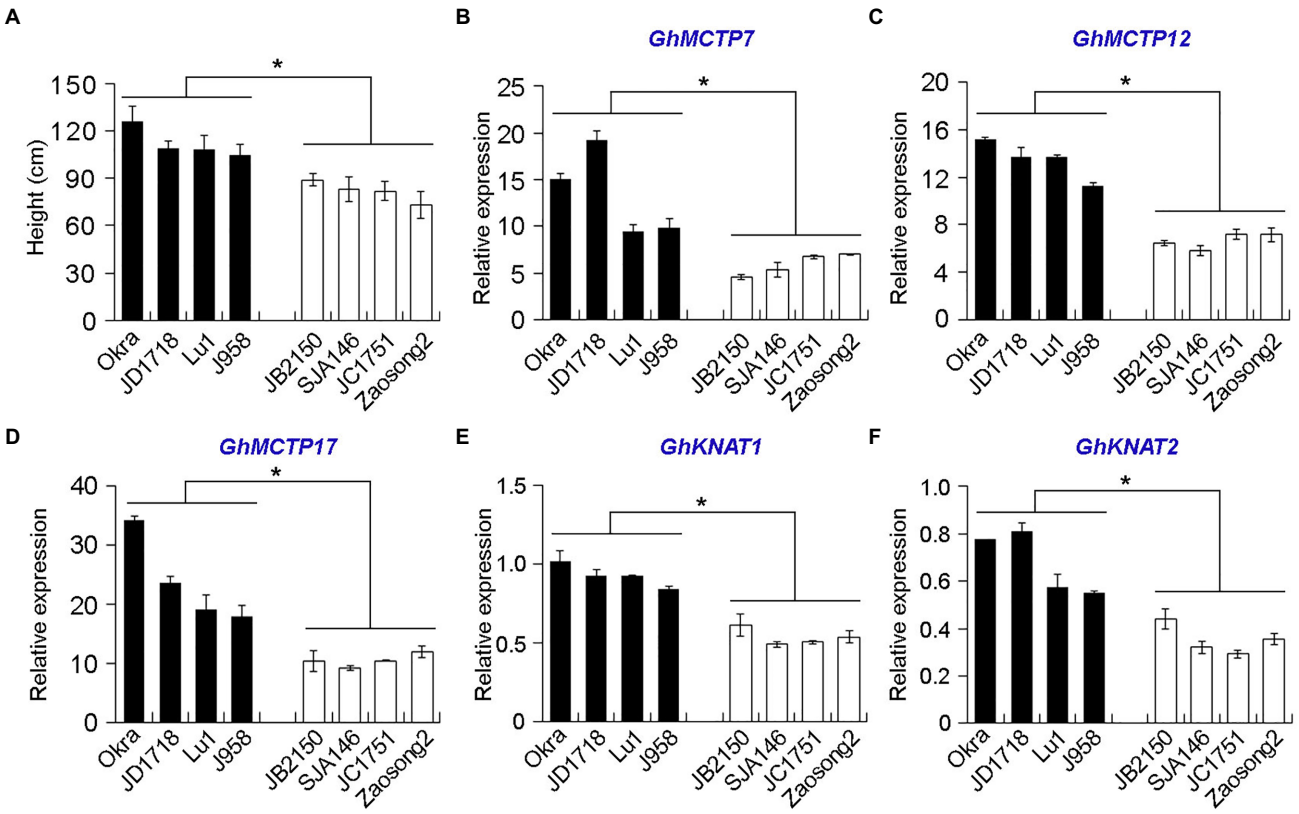

FIGURE 4 | The expression levels of *GhMCTP7/12/17* and *GhKNAT1/2* are correlated with the average plant height (cm) of cotton varieties. **(A)** Comparison of plant height of eight allotetraploid cotton cultivars (*G. hirsutun*). The heights of plants are calculated when cotton bolls were opening. The cotton varieties are divided into two groups according to their plant heights. The varieties over 100 cm in height or under 100 cm are grouped and are shown in black bars or white bars, respectively. **(B-F)** Quantitative real-time PCR analysis of *GhMCTP7* **(B)**, *GhMCTP12* **(C)**, *GhMCTP17* **(D)**, *GhKNAT1* **(E)**, and *GhKNAT2* **(F)** in main stem apex of different cotton species. The X-axis represents different upland cotton species, while the Y-axis represents gene relative expressions. Results were normalized against the expression level of *GhUBI1*. Error bars indicate SD of three biological replicates (*$p < 0.05$).

localized to puncta-like structures in the cytosol within cells (**Figure 6A**; **Supplementary Figure 5**). We then coexpressed *35S:GFP-GhMCTP7* with the fluorescence-tagged endosome marker *35S:RFP-RabF2b* (Jaillais et al., 2006), and observed GhMCTP7 was partially localized in endosomal compartments (**Figure 6D**). GhMCTP12 and GhMCTP17 were localized in whole cells and substantially colocalized with an ER marker, RFP-HDEL (**Figures 6B,C,E**; Nelson et al., 2007). These results suggest that GhMCTP7, GhMCTP12, and GhMCTP17 function coordinately to regulate intercellular signaling and control cotton development.

GhMCTP7/12/17 Interact With GhKNAT1/2 to Regulate Shoot Development

FTIP3 and FTIP4, two MCTP proteins in *Arabidopsis*, are required for shoot apical meristem development through mediating subcellular localization and intercellular trafficking of STM (Liu et al., 2018b). GhMCTP7, GhMCTP12, and GhMCTP17 shared high sequence similarities with FTIP3 and FTIP4 in *Arabidopsis* (**Figure 2A**). Our finding on the tissue expression patterns of *GhMCTP7*, *GhMCTP12*, and *GhMCTP17* in the shoot apex and their roles in cotton meristem development prompted us to investigate whether GhMCTP7, GhMCTP12, and GhMCTP17 interact with KNOTTED1 (KN1)-like homeobox (KNOX) family proteins in cotton, like their counterparts in *Arabidopsis* (Liu et al., 2018b).

First, we examined the expression profiles of all *GhKNAT* genes in different tissue of upland cotton. The results showed that *GhKNAT1* and *GhKNAT2* were highly expressed in the main stem apex (**Supplementary Figure 6**). We then investigated whether GhMCTP7/12/17 interact with GhKNAT1/2. We conducted a detailed analysis of protein interaction between GhMCTP7/12/17 and GhKNAT1/2. Yeast two-hybrid assays revealed that the GhMCTP7/12/17$^{\Delta TM}$, a truncated GhMCTP7/12/17 devoid of the transmembrane region, interacted with GhKNAT1/2 (**Figure 7A**). To test the interaction between GhMCTP7/12/17 and GhKNAT1/2 in planta, we performed luciferase complementation imaging (LCI) assays. We coexpressed nLUC-GhMCTP7/12/17 and cLUC-GhKNAT1/2 and detected fluorescence signals in *N. benthamiana* leaves (**Figures 7B,C**). These results demonstrate the protein interactions between GhMCTP7/12/17 and GhKNAT1/2 in plants.

Comparing the expression levels of *GhKNAT1* and *GhKNAT2* in the shoot apex of selected allotetraploid cotton cultivars, we also

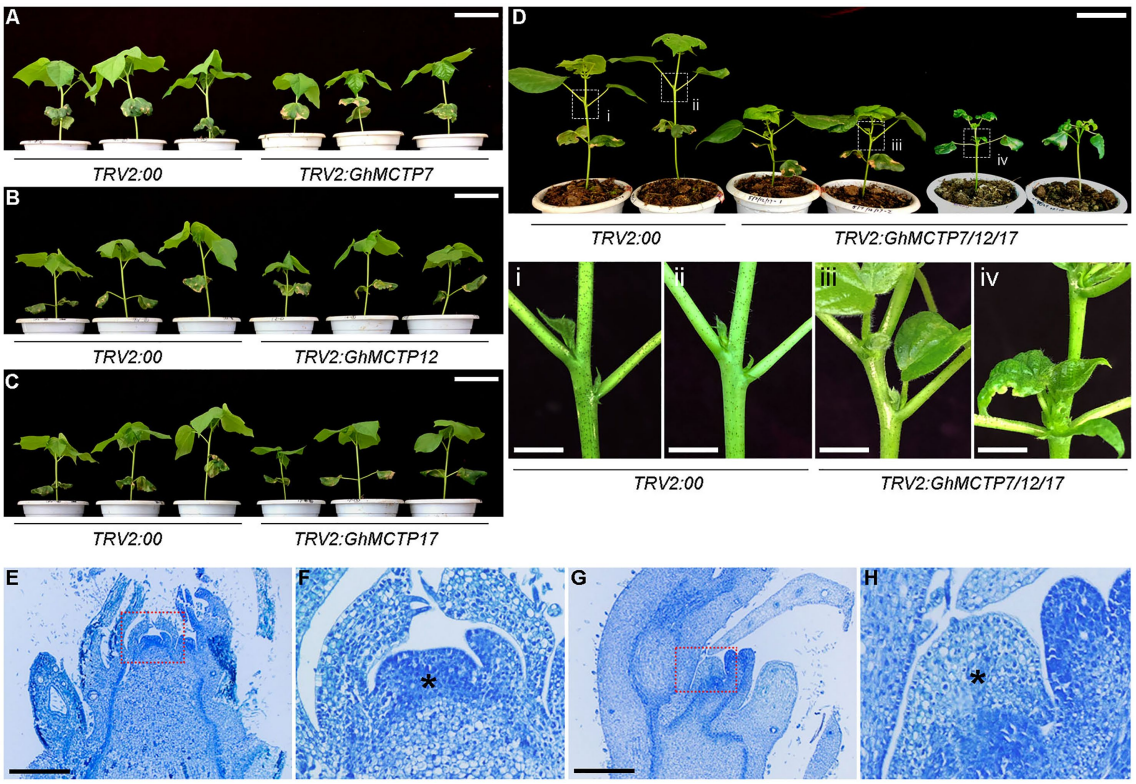

FIGURE 5 | Phenotypic analyses of *TRV2:GhMCTPs* plants. **(A–C)** Plant height comparison of 3–4 leaf stage *TRV2:GhMCTP7* **(A)**, *TRV2:GhMCTP12* **(B)**, and *TRV2:GhMCTP17* **(C)** plants with *TRV2:00* plants (negative control group). Scale bars = 5 cm. **(D)** Plant height comparison of four leaf stage *TRV2:GhMCTP7/12/17* with *TRV2:00* plants (upper panel). The lower panels show the magnified view of boxes indicated in the upper panel. Scale bars = 5 cm (top) and 1 cm (bottom). **(E–H)** Median longitudinal section of main inflorescence shoot apices of *TRV2:00* **(E)** and *TRV2:GhMCTP7/12/17* **(G)** plants. Scale bars = 1 μm. **(F,H)** The magnified views of boxes are indicated in **(G)** and **(H)**, respectively. Asterisks indicate main inflorescence meristems.

found that the expression levels of *GhKNAT1* and *GhKNAT2* were higher in the taller cotton variants (**Figures 4E,F**), suggesting that GhKNAT1 and GhKNAT2 are the potential regulators for meristem development. Then, we further investigated the effects of GhKNAT1 and GhKNAT2 on meristem development. We silenced the expression of *GhKNAT1* and *GhKNAT2* using the VIGS in soil-grown upland cotton. Most of these *TRV2:GhKNAT1/2* plants exhibited a dwarf phenotype (**Figure 8A**). Examination of some selected plants revealed that the expressions of *GhKNAT1* and *GhKNAT2* were downregulated in these *TRV2:GhKNAT1/2* plants (**Figure 8B**), suggesting that GhKNAT1 and GhKNAT2 regulate meristem development.

GhMCTP7/12/17 Are Involved in the Regulation of Multiple Signal Pathways

KNOTTED1-like homeobox transcription factors promote meristem function through directly targeting various transcription factors and genes participating in hormone pathways (Bolduc et al., 2012). Given that GhMCTP7/12/17 may regulate meristem development through mediating the function of GhKNAT1/2, we further conducted a detailed expression analysis to investigate the involvement of GhMCTP7/12/17 in regulating key regulators in meristem development and hormone signaling pathway.

To further characterize the functions of GhMCTP7/12/17 and GhKNAT1/2, the expressions of selected genes, which counterparts in *Arabidopsis* are key regulators in shoot meristem development and genes in various hormone signaling pathways, were analyzed. First, key regulators with high expression in the main stem apex were selected based on the public expression data (**Supplementary Figures 7–10**). We then carried out qRT-PCR to examine the expressions of the selected genes in the main stem apex of *TRV2:00*, *TRV2:GhMCTP7/12/17*, and *TRV2:GhKNAT1/2* plants (**Figure 9**). *Gh AGAMOUS-LIKE 8-1/-2* (*GhAGL8-1/-2*), the homologs of *APETALA1* (*AP1*) in cotton, regulate plant height and early maturity of cotton (Su et al., 2018). GENERAL REGULATORY FACTORs (GRFs) play important roles in flowering regulation and meristem development (Sang et al., 2021). We found that the expressions of *GhAGL8-1/2* and *GhGRF6* were downregulated in both *TRV2:GhMCTP7/12/17* and *TRV2:GhKNAT1/2*, suggesting the delayed determination of floral meristem identity (**Figures 9A,B**).

KNOTTED1-like homeobox protein promotes meristem development through repressing gibberellin (GA) biosynthesis and activating cytokinin (CK) pathway (Jasinski et al., 2005). KNOX protein also regulates the expressions of auxin-related genes (Bolduc et al., 2012), suggesting that KNOX protein integrates multiple phytohormone signaling pathways to regulate

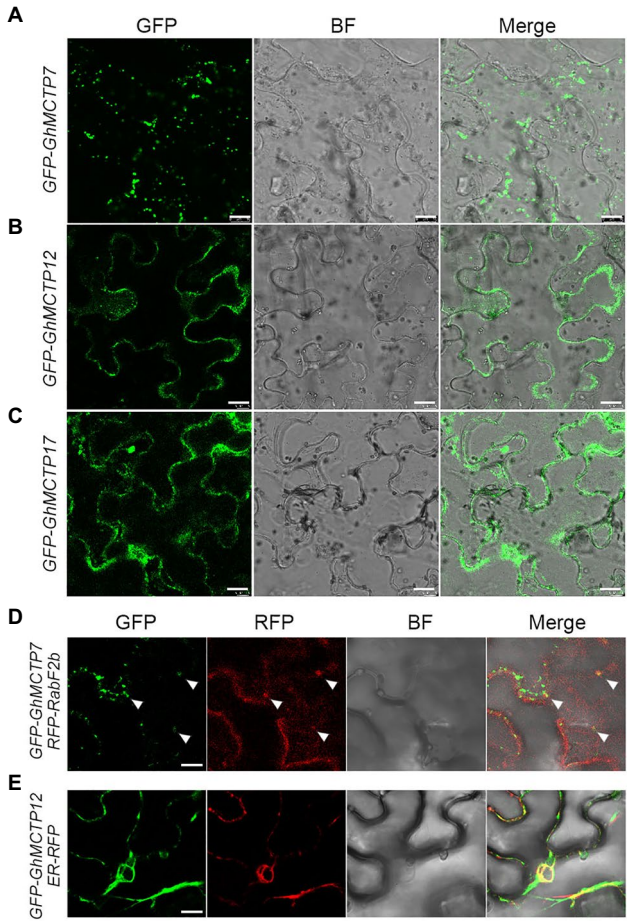

FIGURE 6 | Subcellular localization of green fluorescent protein (GFP)-GhMCTP7/12/17 in *Nicotiana benthamiana* leaf epidermal cells. **(A–C)** Representative GFP fluorescence images of GFP-GhMCTP7 **(A)**, GFP-GhMCTP12 **(B)**, and GFP-GhMCTP17 **(C)**. Scale bars = 50 μm. **(D)** Colocalization of GFP-GhMCTP7 with an endosome marker (RFP-RabF2b). **(E)** Co-expression of GFP-GhMCTP12 with an ER marker (ER-RFP). Scale bars = 10 μm.

meristem development. We studied the expression of cytokinin biosynthesis and signaling genes, auxin biosynthesis and signaling genes, GA biosynthesis and catabolism genes in both *TRV2:GhMCTP7/12/17* and *TRV2:GhKNAT1/2*. We observed that the expression of GA catabolic enzyme *GhGA2OX1-1* was elevated in both *TRV2:GhMCTP7/12/17* and *TRV2:GhKNAT1/2*, whereas GA biosynthesis gene *GhGA20OX2-1*, CK biosynthesis gene *Gh ISOPENTENYLTRANSFERASE 1* (*GhIPT1*), and auxin biosynthesis gene *Gh YUCCA3* (*GhYUC3*) were consistently downregulated in *TRV2:GhMCTP7/12/17* and *TRV2:GhKNAT1/2* (**Figures 9E–G**). These results suggest that MCTP7/12/17 regulate meristem development partially through GhKNAT1/2-mediated regulatory pathway.

We also observed that some other regulators in meristem development and phytohormone signaling pathway in main stem apex of *TRV2:00*, *TRV2:GhMCTP7/12/17*, and *TRV2:GhKNAT1/2* plants (**Supplementary Figure 11**). Stem cell regulators *Gh AINTEGUMENTA-1/-2/-3* (*GhANT-1/-2/-3*), *Gh GROWTH-REGULATING FACTOR 1* (*GhGRF1*), and *Gh GRF1-INTERACTING FACTOR 3* (*GhGIF3*) were downregulated in *TRV2:GhMCTP7/12/17* plants (Mudunkothge and Krizek, 2012; Kim and Tsukaya, 2015); however, their expressions in *TRV2:GhKNAT1/2* plants were not changed (**Supplementary Figures 11A–E**). In addition, we observed that, although the expressions of *Gh CLAVATA3/ESR-RELATED 27* (*GhCLE27*) and *Gh KNOTTED1-LIKE HOMEOBOX GENE 4* (*GhKNAT4*) were consistently downregulated in both *TRV2:GhMCTP7/12/17* and *TRV2:GhKNAT1/2* plants, the downregulation of *GhKNAT4* and *GhCLE27* was more obvious in *TRV2:GhMCTP7/12/17* plants (Fletcher, 2020; **Supplementary Figures 11F,G**). Auxin efflux regulator *Gh PIN-FORMED 3/-2* (*GhPIN3/-2*) and *Gh PHOTOSYSTEM I LIGHT HARVESTING COMPLEX GENE2* (*Gh LHCA2*) has been reported to regulate cotton height (Su et al., 2018; Ma et al., 2019). We observed that the expressions of *GhPIN3/-2* and *GhLHCA2* were downregulated in *TRV2:GhMCTP7/12/17* plants, but minor altered in *TRV2:GhKNAT1/2* plants (**Supplementary Figures 11H–J**). Furthermore, the expression of *GhARR5D*, which functions in the cytokinin signaling pathway, was increased in *TRV2:GhMCTP7/12/17* plants but not in *TRV2:GhKNAT1/2* plants (**Supplementary Figure 11L**). These results demonstrate that *GhMCTP7*, *GhMCTP12*, and *GhMCTP17* also regulate meristem development independent of *GhKNAT1* and *GhKNAT2*, possible through other *KNOX* family members.

DISCUSSION

The development of multicellular organisms relies on the coordination of a variety of specialized cell types through intercellular communication. *MCTP* family proteins in *Arabidopsis* and its orthologs in several plant species have been shown to play an important role in protein intercellular movement and are essential for plant development (Liu et al., 2013, 2019; Hao et al., 2020; Zhu et al., 2020). MCTP proteins have been identified in the cotton genome (Hao et al., 2020); however, their biological functions are still largely unknown. In this study, we identified 33 *GhMCTP* genes from the upland cotton genome and analyzed their evolutionary relationships. Through examining the expression patterns of all *GhMCTPs* in different tissues of upland cotton, we found that *GhMCTP7*, *GhMCTP12*, and *GhMCTP17* are highly expressed in the main stem apex and play a key role in shoot development. GhMCTP7/12/17 interacted with GhKNAT1/2 and modulated the expression of multiple shoot meristem regulators in a *GhKNAT1/2*-dependent and independent manner. Our findings suggest that GhMCTP proteins are evolutionarily conserved in upland cotton and play conserved roles in meristem development.

We have systematically characterized 33 *GhMCTP* genes in *G. hirsutum*, which are grouped into seven clades based on the phylogenetic analysis (**Figure 1A**). The key feature of MCTPs is the presence of multiple C2 domain at the N-terminus and PRT_C domain at the C-terminus (**Figure 1C**). All MCTPs in *Arabidopsis* contain the C-terminal transmembrane region, which anchors the MCTP in the intracellular membrane. In

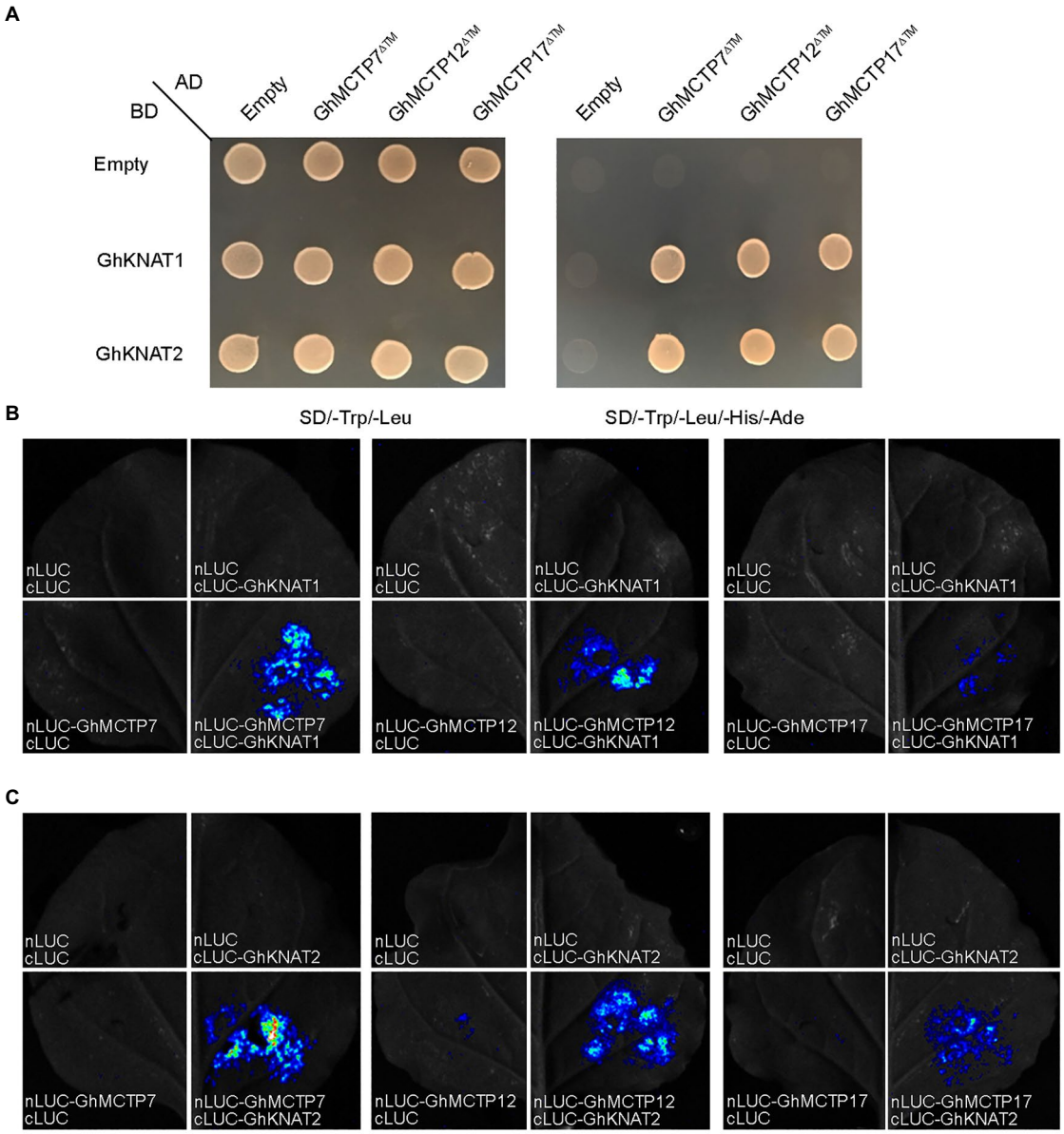

FIGURE 7 | GhMCTP7/12/17 interact with GhKNAT1/2. **(A)** Yeast two-hybrid assay showing the interaction between GhMCTP7/12/17$^{\Delta TM}$ and GhKNAT1/2. Transformed yeast cells were grown on SD -Ade/-His/-Leu/-Trp medium. **(B,C)** Luciferase complementation imaging (LCI) assays showing that GhMCTP7/12/17 interact with GhKNAT1 **(B)** and GhKNAT2 **(C)** in *N. benthamiana* leaves. Fluorescence signal intensities represent their protein interaction intensities.

upland cotton, GhMCTP3 does not contain the C-terminal transmembrane region, and the transmembrane regions of some GhMCTPs are located in the N-terminal C2 domain region (**Supplementary Figure 2**), suggesting the functional divergence among MCTPs in different species. Comparing with the limited number of MCTPs in animals, a large number of MCTPs are identified in cotton and other plant lineages (**Figure 2A**), implying that plants have evolved functional specialized MCTPs to regulate distinct biological processes. We also noticed that all GhMCTPs exist in cotton A subgenome and D subgenome, demonstrating their fundamental roles in cotton development. Chromosome distribution of *GhMCTP* shows that *GhMCTP* genes are dispersed across the chromosome but not in a cluster pattern (**Supplementary Figure 1**), indicating that the *GhMCTP* gene family does not simply arise from chromosome region duplication but also are involved in the extensive reshuffling and divergent evolution. It is also noteworthy that *MCTP* family proteins exist in most plant lineages (**Figure 2B**), suggesting the fundamental roles of *MCTPs* in plant development. *GhMCTPs* exhibit distinct or overlapping expression patterns in various tissues at different developmental stages (**Figure 3**), demonstrating *MCTPs* are functionally specialized during plant evolution.

In *Arabidopsis*, different *MCTPs* showed distinct patterns in various tissues, and no *MCTPs* exhibited the identical

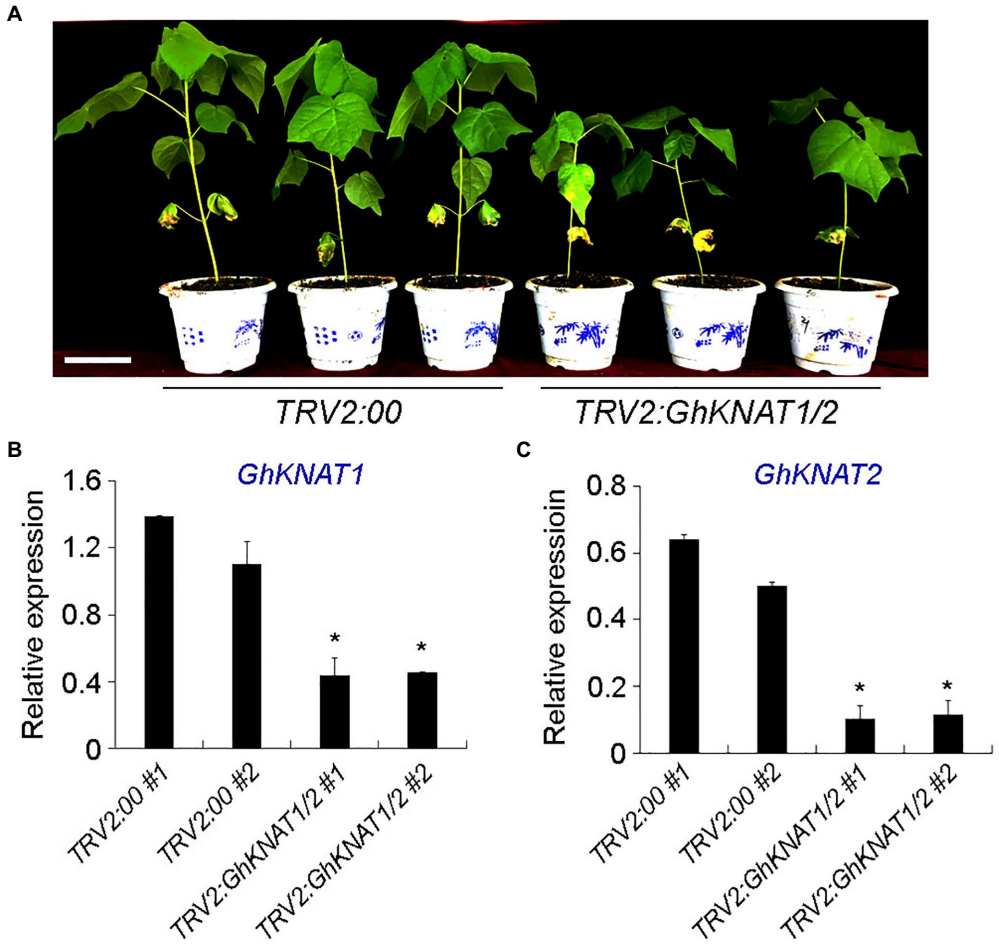

FIGURE 8 | Phenotypic analyses of *TRV2:GhKNAT1/2* plants. **(A)** Plant height comparison of 5–6 leaf stage *TRV2:GhKNAT1/2* plants with *TRV2:00* plants (negative control group). **(B,C)** Expression analysis of *GhKNAT1* and *GhKNAT2* in main stem apex of *TRV2:00* **(B)** and *TRV2:GhKNAT1/2* **(C)** plants, respectively. Results were normalized against the expression level of *GhUBI1*. Error bars indicate SD of three biological replicates (*$p < 0.05$).

expression pattern at both vegetative and reproductive tissues (Liu et al., 2018a), suggesting that *MCTPs* might be differentially regulated and play different roles in various tissues. Additionally, some *MCTPs* share similar expression patterns in several tissues (Liu et al., 2018a), indicating that *MCTPs* might function redundantly during plant development. To understand the biological function of *GhMCTPs* in cotton, we examined the expression patterns of all *GhMCTP* in various tissues and revealed their distinct or overlapping expressions in various tissues (**Figure 3**). *GhMCTP7*, *GhMCTP12*, and *GhMCTP17* are highly expressed in the shoot apex, and their expression levels are correlated with the plant heights in different allotetraploid cotton cultivars. Furthermore, plants with the downregulation of *GhMCTP7*, *GhMCTP12*, and *GhMCTP17* exhibit dwarf phenotype, demonstrating that they might function redundantly to regulate meristem development.

Multiple C2 domain and transmembrane region proteins have been shown to regulate multiple developmental processes by mediating the trafficking of various macromolecules. In *Arabidopsis*, FTIP3 and FTIP4 interact with and regulate STM intercellular and intracellular trafficking, thus affecting the protein distribution within the meristem (Liu et al., 2018b). Regulation of STM trafficking at subcellular and tissue levels causes early termination of shoot apices and continuously generation secondary shoots, resulting in dwarf and bushy phenotypes. The position of leaves and branches, timing of the flowering, and relative position of reproductive structures are traits that affect cotton productivity. Genetic engineering the cotton architecture is crucial for cotton domestication and will benefit crop production. GhMCTP7, GhMCTP12, and GhMCTP17 show high sequence similarity with FTIP3 and FTIP4, hinting that *GhMCTP7*, *GhMCTP12*, and *GhMCTP17* might regulate meristem development through *KNOX* family proteins. Considering the protein interactions between GhMCTP7/12/17 and GhKNAT1/2 (**Figures 7, 8**), we reasoned that GhMCTP7, GhMCTP12, and GhMCTP17 might regulate the function of GhKNAT1 and GhKNAT2 to modulate meristem development. Consistently, gene silencing of *GhKNAT1* and *GhKNAT2* in *TRV2:GhKNAT1/2* results in a dwarf phenotype (**Figure 7**), similar to *TRV2:GhMCTP7/12/17*.

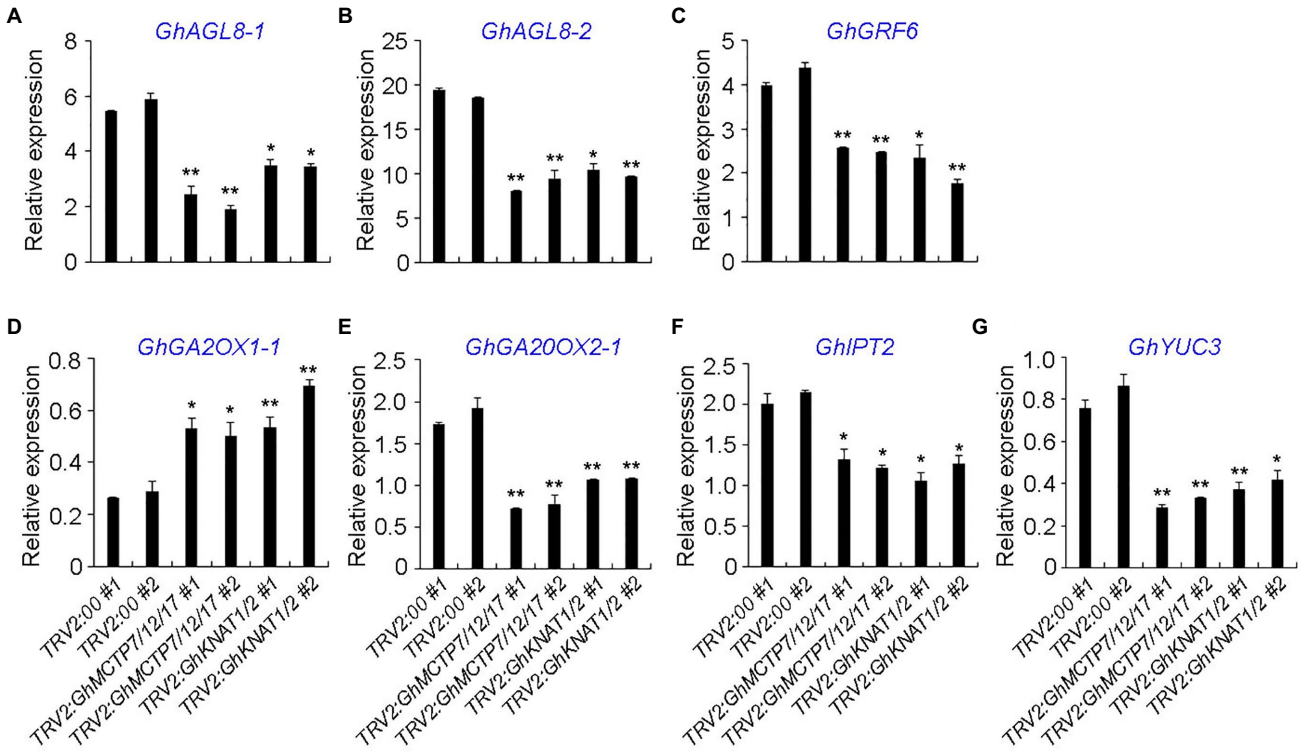

FIGURE 9 | Quantitative analysis of multiple shoot meristem regulators in main stem apex of *TRV2:00*, *TRV2:GhMCTP7/12/17*, and *TRV2:GhKNAT1/2* plants. The expression levels of *GhALG8-1* **(A)**, *GhAGL8-2* **(B)**, *GhGRF6* **(C)**, *GhGA2OX1-1* **(D)**, *GhGA20OX2-1* **(E)**, *GhIPT2* **(F)**, and *GhYUC3* **(G)** in *TRV2:GhMCTP7/12/17* and *TRV2:GhKNAT1/2* plants. Results were normalized against the expression level of *GhUBI1*. Error bars indicate SD of three biological replicates (*$p < 0.05$; **$p < 0.01$).

KNOTTED1-like homeobox-like transcription factors promote meristem function through regulating various transcription factors and manipulating phytohormone pathways (Jasinski et al., 2005; Bolduc et al., 2012). We examined some of the putative downstream targets of *GhKNAT1* and *GhKNAT2* in the main apex of *TRV2:00*, *TRV2:GhMCTP7/12/17*, and *TRV2:GhKNAT1/2* plants (**Figure 9**). The expression levels of several *GhKNAT1/2* putative targets are consistently upregulated or downregulated in both *TRV2:GhMCTP7/12/17* and *TRV2:GhKNAT1/2* plants (**Figure 9**). We also observed that genes with altered expression in *TRV2:GhMCTP7/12/17* are not similarly affected in *TRV2:GhKNAT1/2*, suggesting that *GhMCTP7*, *GhMCTP12*, and *GhMCTP17* also regulate meristem development independent of *GhKNAT1* and *GhKNAT2* (**Supplementary Figure 11**). Taken together, these results demonstrate that *GhMCTP7*, *GhMCTP12*, and *GhMCTP17* function redundantly to regulate meristem development partially through *GhKNAT1* and *GhKNAT2*.

Plant development requires cell-cell communication and the coordination of various specialized cell types. Non-cell-autonomous signals are one of the regulatory mechanisms to integrate various signals and coordinate plant development. Through the combination of different approaches, an increasing number of mobile signals, such as transcription factors and peptides, have been discovered to regulate plant growth. Severe mutations of key regulators in plants always lead to embryonic lethality or obvious growth defect, which precludes evaluation of later phenotypes and prevents its application for crop breeding (Paaby and Rockman, 2013). Genetic engineering the *cis*-elements of key regulators will bypass its lethality effect and expose multiple pleiotropic roles of this gene (Hendelman et al., 2021). Regulation of protein trafficking will be another alternative approach to make dysfunction of key regulators. MCTP-mediated regulation of proteins at sub-cellular and tissue levels provides a chance to modulate the protein function at the desired degree, which could generate a phenotype that different from the severe mutants and provide important implications for biotechnological application for crop breeding.

CONCLUSION

In conclusion, our systematic analysis of *GhMCTPs* in upland cotton illustrates their diverse expression patterns. We find that *GhMCTP7*, *GhMCTP12*, and *GhMCTP17* are highly expressed in the main stem apex, and they function redundantly to regulate the development of the main stem apex and affect cotton architecture. We also show that the expression levels of *GhMCTP7*, *GhMCTP12*, and *GhMCTP17* in the shoot apex of eight selected allotetraploid cotton cultivars are correlated with their plant heights. In addition, VIGS plants silenced for

GhMCTP7, *GhMCTP12*, and *GhMCTP17* show a dwarf phenotype. Taken together, this study provides important clues for studying the function of *GhMCTPs*, deepens our understanding of cotton apex regulation, and establishes a resource for cotton breeding.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

GH, QH, and MZ conceived and designed the experiment and performed most of the experiments. MW, XH, and JL performed some of the experiments and assisted in data analysis. CF analyzed some data. QH, LX, LL, and GH analyzed the data and wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (grant no. 31271317), Outstanding Youth Science Fund of Xinjiang Uygur Autonomous Region (grant no. 2021D01E17), Fundamental Research Funds for the Central Universities (grant nos. CCNU16A02047 and CCNU19TS064), and the Shanghai Pujiang Program (20PJ1405200).

ACKNOWLEDGMENTS

The authors would like to thank the Germplasm Bank of Cotton Institute of Shanxi Academy of Agricultural Sciences for providing the cotton seeds (Zaosong 2).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.767667/full#supplementary-material

REFERENCES

Bolduc, N., Yilmaz, A., Mejia-Guerra, M. K., Morohashi, K., O'Connor, D., Grotewold, E., et al. (2012). Unraveling the KNOTTED1 regulatory network in maize meristems. *Genes Dev.* 26, 1685–1690. doi: 10.1101/gad.193433.112

Brault, M. L., Petit, J. D., Immel, F., Nicolas, W. J., Glavier, M., Brocard, L., et al. (2019). Multiple C2 domains and transmembrane region proteins (MCTPs) tether membranes at plasmodesmata. *EMBO Rep.* 20:e47182. doi: 10.15252/embr.201847182

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, Y., Shen, J., Zhang, L., Qi, H., Yang, L., Wang, H., et al. (2021). Nuclear translocation of OsMFT1 that is impeded by OsFTIP1 promotes drought tolerance in rice. *Mol. Plant* 14, 1297–1311. doi: 10.1016/j.molp.2021.05.001

Chen, H., Zou, Y., Shang, Y., Lin, H., Wang, Y., Cai, R., et al. (2008). Firefly luciferase complementation imaging assay for protein-protein interactions in plants. *Plant Physiol.* 146, 323–324. doi: 10.1104/pp.107.111740

Cho, W., and Stahelin, R. V. (2006). Membrane binding and subcellular targeting of C2 domains. *Biochim. Biophys. Acta* 1761, 838–849. doi: 10.1016/j.bbalip.2006.06.014

Corbalan-Garcia, S., and Gómez-Fernández, J. C. (2014). Signaling through C2 domains: more than one lipid target. *Biochim. Biophys. Acta* 1838, 1536–1547. doi: 10.1016/j.bbamem.2014.01.008

Fletcher, J. C. (2020). Recent advances in *Arabidopsis* CLE peptide signaling. *Trends Plant Sci.* 25, 1005–1016. doi: 10.1016/j.tplants.2020.04.014

Gao, W., Long, L., Zhu, L.-F., Xu, L., Gao, W.-H., Sun, L.-Q., et al. (2013). Proteomic and virus-induced gene silencing (VIGS) analyses reveal that gossypol, brassinosteroids, and jasmonic acid contribute to the resistance of cotton to Verticillium dahliae. *Mol. Cell. Proteomics* 12, 3690–3703. doi: 10.1074/mcp.M113.031013

Genç, Ö., Dickman, D. K., Ma, W., Tong, A., Fetter, R. D., and Davis, G. W. (2017). MCTP is an ER-resident calcium sensor that stabilizes synaptic transmission and homeostatic plasticity. *eLife* 6:e22904. doi: 10.7554/eLife.22904

Hao, P., Wang, H., Ma, L., Wu, A., Chen, P., Cheng, S., et al. (2020). Genome-wide identification and characterization of multiple C2 domains and transmembrane region proteins in *Gossypium hirsutum*. *BMC Genomics* 21:445. doi: 10.1186/s12864-020-06842-1

Hendelman, A., Zebell, S., Rodriguez-Leal, D., Dukler, N., Robitaille, G., Wu, X., et al. (2021). Conserved pleiotropy of an ancient plant homeobox gene uncovered by cis-regulatory dissection. *Cell* 184, 1724.e16–1739.e16. doi: 10.1016/j.cell.2021.02.001

Huang, G.-Q., Gong, S.-Y., Xu, W.-L., Li, W., Li, P., Zhang, C.-J., et al. (2013). A fasciclin-like arabinogalactan protein, GhFLA1, is involved in fiber initiation and elongation of cotton. *Plant Physiol.* 161, 1278–1290. doi: 10.1104/pp.112.203760

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

Jaillais, Y., Fobis-Loisy, I., Miège, C., Rollin, C., and Gaude, T. (2006). AtSNX1 defines an endosome for auxin-carrier trafficking in *Arabidopsis*. *Nature* 443, 106–109. doi: 10.1038/nature05046

Jasinski, S., Piazza, P., Craft, J., Hay, A., Woolley, L., Rieu, I., et al. (2005). KNOX action in *Arabidopsis* is mediated by coordinate regulation of cytokinin and gibberellin activities. *Curr. Biol.* 15, 1560–1565. doi: 10.1016/j.cub.2005.07.023

John, M. E., and Crow, L. J. (1992). Gene expression in cotton (*Gossypium hirsutum* L.) fiber: cloning of the mRNAs. *Proc. Natl. Acad. Sci. U. S. A.* 89, 5769–5773. doi: 10.1073/pnas.89.13.5769

Kim, J. H., and Tsukaya, H. (2015). Regulation of plant growth and development by the growth-regulating factor and GRF-interacting factor duo. *J. Exp. Bot.* 66, 6093–6107. doi: 10.1093/jxb/erv349

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

Lange, M., Yellina, A. L., Orashakova, S., and Becker, A. (2013). Virus-induced gene silencing (VIGS) in plants: an overview of target species and the virus-derived vector systems. *Methods Mol. Biol.* 975, 1–14. doi: 10.1007/978-1-62703-278-0_1

Lek, A., Evesson, F. J., Sutton, R. B., North, K. N., and Cooper, S. T. (2012). Ferlins: regulators of vesicle fusion for auditory neurotransmission, receptor trafficking and membrane repair. *Traffic* 13, 185–194. doi: 10.1111/j.1600-0854.2011.01267.x

Liu, L., Li, C., Liang, Z., and Yu, H. (2018a). Characterization of multiple C2 domain and transmembrane region proteins in *Arabidopsis*. *Plant Physiol.* 176, 2119–2132. doi: 10.1104/pp.17.01144

Liu, L., Li, C., Song, S., Teo, Z. W. N., Shen, L., Wang, Y., et al. (2018b). FTIP-dependent STM trafficking regulates shoot meristem development in *Arabidopsis*. *Cell Rep.* 23, 1879–1890. doi: 10.1016/j.celrep.2018.04.033

Liu, L., Li, C., Teo, Z. W. N., Zhang, B., and Yu, H. (2019). The MCTP-SNARE complex regulates florigen transport in *Arabidopsis*. *Plant Cell* 31, 2475–2490. doi: 10.1105/tpc.18.00960

Liu, L., Liu, C., Hou, X., Xi, W., Shen, L., Tao, Z., et al. (2012). FTIP1 is an essential regulator required for florigen transport. *PLoS Biol.* 10:e1001313. doi: 10.1371/journal.pbio.1001313

Liu, L., Zhu, Y., Shen, L., and Yu, H. (2013). Emerging insights into florigen transport. *Curr. Opin. Plant Biol.* 16, 607–613. doi: 10.1016/j.pbi.2013.06.001

Ma, J., Pei, W., Ma, Q., Geng, Y., Liu, G., Liu, J., et al. (2019). QTL analysis and candidate gene identification for plant height in cotton based on an interspecific backcross inbred line population of *Gossypium hirsutum*×*Gossypium barbadense*. *Theor. Appl. Genet.* 132, 2663–2676. doi: 10.1007/s00122-019-03380-7

Maeda, I., Kohara, Y., Yamamoto, M., and Sugimoto, A. (2001). Large-scale analysis of gene function in caenorhabditis elegans by high-throughput RNAi. *Curr. Biol.* 11, 171–176. doi: 10.1016/S0960-9822(01)00052-5

McGarry, R. C., Prewitt, S. F., Culpepper, S., Eshed, Y., Lifschitz, E., and Ayre, B. G. (2016). Monopodial and sympodial branching architecture in cotton is differentially regulated by the *Gossypium hirsutum* SINGLE FLOWER TRUSS and SELF-PRUNING orthologs. *New Phytol.* 212, 244–258. doi: 10.1111/nph.14037

Mergner, J., Frejno, M., List, M., Papacek, M., Chen, X., Chaudhary, A., et al. (2020). Mass-spectrometry-based draft of the *Arabidopsis* proteome. *Nature* 579, 409–414. doi: 10.1038/s41586-020-2094-2

Mudunkothge, J. S., and Krizek, B. A. (2012). Three *Arabidopsis* AIL/PLT genes act in combination to regulate shoot apical meristem function. *Plant J.* 71, 108–121. doi: 10.1111/j.1365-313X.2012.04975.x

Nalefski, E. A., and Falke, J. J. (1996). The C2 domain calcium-binding motif: structural and functional diversity. *Protein Sci.* 5, 2375–2390. doi: 10.1002/pro.5560051201

Nelson, B. K., Cai, X., and Nebenführ, A. (2007). A multicolored set of in vivo organelle markers for co-localization studies in *Arabidopsis* and other plants. *Plant J.* 51, 1126–1136. doi: 10.1111/j.1365-313X.2007.03212.x

Paaby, A. B., and Rockman, M. V. (2013). The many faces of pleiotropy. *Trends Genet.* 29, 66–73. doi: 10.1016/j.tig.2012.10.010

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

Reinhardt, D., and Kuhlemeier, C. (2002). Plant architecture. *EMBO Rep.* 3, 846–851. doi: 10.1093/embo-reports/kvf177

Sang, N., Liu, H., Ma, B., Huang, X., Zhuo, L., and Sun, Y. (2021). Roles of the 14-3-3 gene family in cotton flowering. *BMC Plant Biol.* 21:162. doi: 10.1186/s12870-021-02923-9

Shilo, B.-Z., and Schejter, E. D. (2011). Regulation of developmental intercellular signalling by intracellular trafficking. *EMBO J.* 30, 3516–3526. doi: 10.1038/emboj.2011.269

Shin, O. H., Han, W., Wang, Y., and Sudhof, T. C. (2005). Evolutionarily conserved multiple C2 domain proteins with two transmembrane regions (MCTPs) and unusual Ca2+ binding properties. *J. Biol. Chem.* 280, 1641–1651. doi: 10.1074/jbc.M407305200

Song, S., Chen, Y., Liu, L., See, Y. H. B., Mao, C., Gan, Y., et al. (2018). OsFTIP7 determines auxin-mediated anther dehiscence in rice. *Nat. Plants* 4, 495–504. doi: 10.1038/s41477-018-0175-0

Song, S., Chen, Y., Liu, L., Wang, Y., Bao, S., Zhou, X., et al. (2017). OsFTIP1-mediated regulation of florigen transport in rice is negatively regulated by the ubiquitin-like domain kinase OsUbDKgamma4. *Plant Cell* 29, 491–507. doi: 10.1105/tpc.16.00728

Song, J. H., Kwak, S.-H., Nam, K. H., Schiefelbein, J., and Lee, M. M. (2019). QUIRKY regulates root epidermal cell patterning through stabilizing SCRAMBLED to control CAPRICE movement in *Arabidopsis*. *Nat. Commun.* 10:1744. doi: 10.1038/s41467-019-09715-8

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

Tran, T. M., McCubbin, T. J., Bihmidine, S., Julius, B. T., Baker, R. F., Schauflinger, M., et al. (2019). Maize carbohydrate partitioning defective33 encodes an MCTP protein and functions in sucrose export from leaves. *Mol. Plant* 12, 1278–1293. doi: 10.1016/j.molp.2019.05.001

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

Trehin, C., Schrempp, S., Chauvet, A., Berne-Dedieu, A., Thierry, A. M., Faure, J. E., et al. (2013). QUIRKY interacts with STRUBBELIG and PAL OF QUIRKY to regulate cell growth anisotropy during *Arabidopsis* gynoecium development. *Development* 140, 4807–4817. doi: 10.1242/dev.091868

Tunstall, N. E., Herr, A., de Bruyne, M., and Warr, C. G. (2012). A screen for genes expressed in the olfactory organs of *Drosophila melanogaster* identifies genes involved in olfactory behaviour. *PLoS One* 7:e35641. doi: 10.1371/journal.pone.0035641

Vaddepalli, P., Herrmann, A., Fulton, L., Oelschner, M., Hillmer, S., Stratil, T. F., et al. (2014). The C2-domain protein QUIRKY and the receptor-like kinase STRUBBELIG localize to plasmodesmata and mediate tissue morphogenesis in *Arabidopsis thaliana*. *Development* 141, 4139–4148. doi: 10.1242/dev.113878

Van Norman, J. M., Breakfield, N. W., and Benfey, P. N. (2011). Intercellular communication during plant development. *Plant Cell* 23, 855–864. doi: 10.1105/tpc.111.082982

Wang, Y., Liu, L., Song, S., Li, Y., Shen, L., and Yu, H. (2017). DOFT and DOFTIP1 affect reproductive development in the orchid Dendrobium Chao Praya smile. *J. Exp. Bot.* 68, 5759–5772. doi: 10.1093/jxb/erx400

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

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

Zhu, M., Yan, B., Hu, Y., Cui, Z., and Wang, X. (2020). Genome-wide identification and phylogenetic analysis of rice FTIP gene family. *Genomics* 112, 3803–3814. doi: 10.1016/j.ygeno.2020.03.003

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

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

frontiers in Plant Science

ORIGINAL RESEARCH
published: 07 December 2021
doi: 10.3389/fpls.2021.789775

Silencing of *GhKEA4* and *GhKEA12* Revealed Their Potential Functions Under Salt and Potassium Stresses in Upland Cotton

Yi Li[1], Zhen Feng[1], Hengling Wei[1], Shuaishuai Cheng[1], Pengbo Hao[1], Shuxun Yu[1] and Hantao Wang[1,2]**

[1] State Key Laboratory of Cotton Biology, Institute of Cotton Research of Chinese Academy of Agricultural Sciences, Anyang, China, [2] Zhengzhou Research Base, State Key Laboratory of Cotton Biology, Zhengzhou University, Zhengzhou, China

The K^+ efflux antiporter (KEA) mediates intracellular K^+ and H^+ homeostasis to improve salt tolerance in plants. However, the knowledge of KEA gene family in cotton is largely absent. In the present study, 8, 8, 15, and 16 putative KEA genes were identified in *Gossypium arboreum*, *G. raimondii*, *G. hirsutum,* and *G. barbadense*, respectively. These KEA genes were classified into three subfamilies, and members from the same subfamilies showed similar motif compositions and gene structure characteristics. Some hormone response elements and stress response elements were identified in the upstream 2000 bp sequence of *GhKEAs*. Transcriptome data showed that most of the *GhKEAs* were highly expressed in roots and stems. The quantificational real-time polymerase chain reaction (qRT-PCR) results showed that most of the *GhKEAs* responded to low potassium, salt and drought stresses. Virus-induced gene silencing (VIGS) experiments demonstrated that under salt stress, after silencing genes *GhKEA4* and *GhKEA12*, the chlorophyll content, proline content, soluble sugar content, peroxidase (POD) activity and catalase (CAT) activity were significantly decreased, and the Na^+/K^+ ratio was extremely significantly increased in leaves, leading to greater salt sensitivity. Under high potassium stress, cotton plants silenced for the *GhKEA4* could still maintain a more stable Na^+ and K^+ balance, and the activity of transporting potassium ions from roots into leaves was reduced silenced for *GhKEA12*. Under low potassium stress, silencing the *GhKEA4* increased the activity of transporting potassium ions to shoots, and silencing the *GhKEA12* increased the ability of absorbing potassium ions, but accumulated more Na^+ in leaves. These results provided a basis for further studies on the biological roles of KEA genes in cotton development and adaptation to stress conditions.

Keywords: upland cotton, K^+ efflux antiporter (KEA), salt and potassium stresses, virus-induced gene silencing, K^+ transport

OPEN ACCESS

Edited by:
Wei Hu,
Institute of Tropical Bioscience and Biotechnology, Chinese Academy of Tropical Agricultural Sciences, China

Reviewed by:
Longbiao Guo,
China National Rice Research Institute, Chinese Academy of Agricultural Sciences (CAAS), China
Quanwei Lu,
Anyang Institute of Technology, China

****Correspondence:***
Shuxun Yu
ysx195311@163.com
Hantao Wang
w.wanghantao@163.com

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

Received: *05 October 2021*
Accepted: *09 November 2021*
Published: *07 December 2021*

Citation:
Li Y, Feng Z, Wei H, Cheng S, Hao P, Yu S and Wang H (2021) Silencing of GhKEA4 and GhKEA12 Revealed Their Potential Functions Under Salt and Potassium Stresses in Upland Cotton. Front. Plant Sci. 12:789775. doi: 10.3389/fpls.2021.789775

INTRODUCTION

With changes in the global environment, crops are facing abiotic stress environments such as soil salinization, drought and extreme temperatures in the process of production. Salt stress is one of the most important abiotic stress factors, that seriously affect the growth, development and survival of plants (Munns, 2002; Liu et al., 2010). Although cotton is considered to be a salt-tolerant

and drought-tolerant crop, the restrictions on cotton growth and yield caused by high salt and drought stress cannot be ignored. Research has shown that when the salt concentration exceeds a certain threshold, the normal physiological function and material metabolism of cotton are significantly affected, and the growth and development of cotton are inhibited, which leads to a decrease in yield and a deterioration in fiber quality (Sharif et al., 2019). On the one hand, when the salt concentration is high, the content of sodium ions is significantly higher than the content of potassium ions, resulting in a higher Na^+/K^+ ratio, which destroys the water balance in the plant cells (Hauser and Horie, 2010); on the other hand, salt stress can lead to plant membrane damage, enzyme activity inhibition, metabolic disorders, etc., resulting in plant growth inhibition and even death (Munns and Tester, 2008). K^+ is one of a large number of mineral elements needed by plants, and it also plays an important role in plant salt tolerance (Britto and Kronzucker, 2008). Plants balance excessive Na^+ by accumulating K^+ to reduce the damage caused by salt stress (Cuin et al., 2003). Therefore, it is very important for plants to absorb and transport potassium ions effectively.

In plants, there are many proteins that consume energy to absorb potassium ions from the outside environment, that is, proteins that carry out active transport. These proteins, called potassium transporters, are divided into three families according to their structures and functions: the KUP/HAK/KT family, the HKT family and the CPA family (Gierth and Maser, 2007). The CPA (Cation Proton Antiporter) family mediates the homeostasis of ions and pH in cells, maintains osmotic balance, and regulates plant growth and development and signal transduction. The CPA family is divided into two subfamilies: CPA1 and CPA2. The CPA1 subfamily is mainly NHX (Na^+/H^+ exchanger) transporters, and the CPA2 subfamily includes KEA (K^+ efflux antiporter) and CHX (Cation/H^+ exchanger) transporters (Chanroj et al., 2011). The numbers of CHX gene families have increased dramatically from charophyte algae to flowering plants (Aranda-Sicilia et al., 2012). Fan et al. (2020) studied the molecular evolution and expansion of KUP gene family in *Gossypium hirsutum* and *G. barbadense*, and found that KUP family genes showed different expression levels under various stress treatment. In our previous study, KUP/HAK/KT gene family and NHX gene family have been identified and found to be involved in abiotic stress response (Fu et al., 2020; Yang et al., 2021). Compared with the diverse CHX gene family, the number of KEA family members varies from plant to plant. Seven, 4, 4, 12, 6, and 24 KEA family members were identified in *Populus trichocarpa*, *Oryza sativa*, *Sorghum bicolor*, *Glycine max*, *Zea mays,* and *Triticum aestivum*, respectively (Ye et al., 2013; Sze and Chanroj, 2018; Sharma et al., 2020). The KEA family in higher plant was first reported in *Arabidopsis thaliana* and originated from bacterial glutathione-regulated K^+ efflux antiporters KefB and KefC with an N-terminal Na^+/H^+ exchanger domain and a C-terminal KTN-NAD (H)-binding domain (also known as the TrkA-N domain) (Maser et al., 2001; Choe, 2002; Chanroj et al., 2012). The AtKEA family was divided into KEAI and KEAII. KEAI was divided into Ia and Ib according to its N-terminal domain and had a complete C-terminal KTN domain, which was closely related to EcKefB and EcKefC proteins (Booth, 2003; Fujisawa et al., 2007; Chanroj et al., 2012). KEAII lost the KTN domain at the C-terminus and had high homology with cyanobacteria, which was similar to the transmembrane coiled coil protein 3 (TMCO3) (Chanroj et al., 2012). The first research on the KEA family showed that *AtKEA2* was located on the chloroplast, and was involved in the regulation of K^+ and the pH of the plastid (Aranda-Sicilia et al., 2012). The inner envelope *AtKEA1* and thylakoid *AtKEA3* transporters were reported to be involved in chloroplast function, osmotic regulation, photosynthesis and pH regulation, and resisted high potassium and high hygromycin in yeast (Zheng et al., 2013; Kunz et al., 2014). In addition, genetic analysis showed that *AtKEA4*, *AtKEA5,* and *AtKEA6* had similar tissue expression patterns, and they cooperated with endosomes *NHX5* and *NHX6* to promote the pH homeostasis and salt tolerance (Zhu et al., 2018). Therefore, the KEA gene family not only plays a significant role in K^+ transport but also may be involved in abiotic stress responses in plants.

However, to date, there have been no reports on the genome-wide identification and characterization of the cotton KEA gene family members. With the publication of the cotton genome sequence and transcriptome data, it is possible to comprehensively identify and analyze the KEA gene family, which is also a key step in studying the function of cotton KEA genes. In this study, members of the KEA gene family in *G. arboreum*, *G. raimondii*, *G. hirsutum,* and *G. barbadense* were identified. The physical and chemical properties, chromosome distributions, gene structures, evolutionary relationships, gene replications and expression patterns were comprehensively analyzed. The functions of *GhKEA4* and *GhKEA12* under salt and potassium stresses were preliminarily explored by virus-induced gene silencing (VIGS) experiments. This research provides basic data for further study on the function of KEA genes in cotton.

MATERIALS AND METHODS

Identification of the K^+ Efflux Antiporter Gene Family

The genome databases of *Gossypium arboreum* (accession number: PRJNA382310) (Du et al., 2018), *Gossypium raimondii* (accession number: PRJNA171262) (Paterson et al., 2012), *Gossypium hirsutum* (accession number: PRJNA433615) (Wang M. et al., 2019) and *Gossypium barbadense* (accession number: PRJNA433615) (Wang M. et al., 2019) were obtained from the CottonGen database[1] (Yu et al., 2014). In addition, the genome databases of *Arabidopsis thaliana* (accession number: PRJNA10719), *Oryza sativa* (accession number: PRJNA448171), *Zea mays* (Schnable et al., 2009), *Populus trichocarpa* (Tuskan et al., 2006), *Sorghum bicolor* (accession number: PRJNA374837), *Triticum aestivum* (Mayer et al., 2014) and *Glycine max* (accession number: ACUP00000000) were downloaded from the phytozome database[2]. The protein sequences of AtKEA1-AtKEA6 were used to construct the hidden markov model (HMM) of the conserved domain of a specific KEA gene family. The HMMER 3.0 program and the constructed

[1]https://www.cottongen.org/

[2]https://phytozome-next.jgi.doe.gov/

HMM model were used to search the predicted KEAs from the above plant genomes. The default parameter of the e-value threshold was set at 1e-50. Then, the NCBI Conserved Domain Database[3] and SMART database[4] were used to confirm whether the candidate protein sequences contain the special domain of the KEA family (Letunic et al., 2015). The protein sequence length, molecular weight (Mw), isoelectric point (pI), grand average of hydropathicity (GRAVY) and subcellular localization of the identified KEA members were predicted from the ExPasy website[5] and ProtComp 9.0[6] (Artimo et al., 2012).

Sequence Alignments and Phylogenetic Analysis

All the identified KEA protein sequences from *G. arboreum*, *G. raimondii*, *G. hirsutum*, *G. barbadense*, *Arabidopsis thaliana*, *Oryza sativa*, *Zea mays*, *Populus trichocarpa*, *Sorghum bicolor*, *Triticum aestivum,* and *Glycine max* were aligned by ClustalX 2.0 (Larkin et al., 2007). The phylogenetic tree was constructed using the neighbor-joining (NJ) method of MEGA 7.0 with the p-distance model and 1000 bootstrap replications (Kumar et al., 2016).

Locations of K^+ Efflux Antiporter Gene on Cotton Chromosomes and Gene Duplication Analysis

The chromosome physical locations of the KEA gene family were extracted from the genome annotation file information of *G. hirsutum*, *G. raimondii*, *G. barbadense,* and *G. arboreum*, and the positions of KEA genes on chromosomes were visualized with Map Chart 2.2 software (Voorrips, 2002). The replication gene pairs of *G. hirsutum*, *G. raimondii,* and *G. arboreum* were detected by MCScanX software (Wang et al., 2012), and gene replication was confirmed according to the following conditions: the coverage of the alignment sequence was ≥80% of the longer gene; the similarity of the regions on the alignment was ≥80%; tightly linked genes on the same chromosome were considered tandem duplications. Circos was adopted to plot the diagram of segmental duplication events on chromosomes (Krzywinski et al., 2009). KaKs_Calculator 2.0 software was used to calculate the non-synonymous mutation rate (Ka) and synonymous mutation rate (Ks) of KEA gene replication (Wang et al., 2010).

Gene Structure and Conserved Motif Analysis

The GhKEA proteins were used for multiple sequence alignment by ClustalX 2.0. The exon-intron structures of upland cotton KEA genes were analyzed on the Gene Structure Display Server (GSDS 2.0[7]) (Hu et al., 2015). The conserved motifs of KEA proteins were identified by the MEME program. The optimization parameters were set as follows: size distribution, zero or once per sequence; motif count, 10; pattern width, between 6 and 50 residues (Bailey et al., 2006).

Promoter Region *Cis*-Acting Element Analysis

DNA sequences 2000 bp upstream of the KEA start codon (ATG) were retrieved from the *G. hirsutum* genome database and submitted to PlantCARE[8] for analysis of *cis*-acting elements (Lescot et al., 2002).

Gene Expression Pattern Analysis

The raw RNA-sequencing data of *G. hirsutum* TM-1 in different tissues were obtained from previously reported transcriptome data (accession number: PRJNA248163) (Zhang et al., 2015). TBtools was used to draw a heatmap, using row-scale and zero to one scale methods, showing the expression patterns of *GhKEAs* (Chen et al., 2020).

Plant Materials and Treatments

In this study, *G. hirsutum* Texas Marker-1 (TM-1) was cultivated by hydroponics with hoagland nutrient solution by Solarbio Biology Co., Ltd., and grown in a climate-controlled chamber with a light/dark cycle of 16 h at 28°C/8 h at 22°C. When the third true leaf was unfolded (approximately 4 weeks), it was treated with NaCl (300 mMol/L), KCl (0.03 mMol/L), PEG6000 (30%) and control group. Samples were taken at 0, 1, 3, 6, 12, and 24 h respectively. All the samples were immediately frozen in liquid nitrogen and stored at –80°C.

Construction of the Virus-Induced Gene Silencing Vector and Determination of Physiological Parameters

The *GhKEA4* (*Ghir_D12G011600.1*) and *GhKEA12* (*Ghir_A06G009360.1*) fragments of 300 nt were introduced by primers, respectively. The fragments of the above genes were then ligated into the pYL156 vector. The primers used for vector construction are listed in **Table 1**. The recombinant vector was transformed into *Agrobacterium tumefaciens* LBA4404. According to the method mentioned by Gao et al. (2016), we injected LBA4404 bacterial solution carrying pYL156 (empty vector), pYL156-GhKEA4, pYL156-GhKEA12, pYL156-CLA1 (positive control) and pYL192 (helper vector) into the cotyledons of TM-1. After 24 h of dark treatment, the cotton plants were moved to a greenhouse with 12 h of light/12 h of darkness for approximately 15 days, and then treated with NaCl. Before salt treatment, 5–6 leaves of control plants and VIGS plants were taken and weighed immediately. Then the leaves were placed in a petri dish with filter paper, placed in a 28°C incubator, set to three repeats, and weighed every other hour. Water loss rate of isolated leaves = (leaf fresh weight-leaf dry weight)/leaf fresh weight × 100%. The experiment was repeated three times independently.

After 24 h of 300 mM NaCl treatment, 0.1 g of sample powder mixed by at least 20 cotton plants were taken to determine the total chlorophyll content, soluble sugar content and proline

[3]https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi

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

[5]https://web.expasy.org/protparam/

[6]http://linux1.softberry.com/berry.phtml?topic=protcomppl&group=programs&subgroup=proloc

[7]http://gsds.gao-lab.org/

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

TABLE 1 | List of the primers used to construct PYL-156 vectors in present study.

ID	Name	Forward primer	Reverse primer
Ghir_D12G011600.1	GhKEA4-156	AAGGTTACCGAATTCTCTAGAATCAAATTTTCCTGTCATTGC	GAGCTCGGTACCGGATCCACATCGTGCAGCTCAAAA
Ghir_A06G009360.1	GhKEA12-156	AAGGTTACCGAATTCTCTAGAATTTGCTTGTGCTGGACAAC	GAGCTCGGTACCGGATCCAGAAATACACCAACAAATACACC

(Pro) content, as well as catalase (CAT) activity and peroxidase activity (POD). And three biological replicates were performed. Chlorophyll was extracted with a ratio of ethanol to acetone (1:1), and the absorbance was measured at 663 and 645 nm (Sun et al., 2013). Other indicators used the kit developed by Solarbio Biology Co., Ltd., and the specific operation steps were guided according to the operating instructions.

Measurement of K^+ and Na^+ Concentration

First, control plants and plants silenced for the target gene were sampled before and after treatment with 300 mM NaCl (high salt), 0.2 mM KCl (low potassium) and 10 mM KCl (high potassium). The samples were then quickly put at 105°C for 30 min to kill and dried at 80°C for 48 h until the weight was unchanged. 0.5 g of plant sample was weighed and used for determination of Na^+ and K^+ contents (Bao, 2005).

RNA Extraction and Quantificational Real-Time Polymerase Chain Reaction Analysis

Total RNA was isolated from the collected samples using the RNA-prep Pure Plant Kit (TIANGEN, Beijing, China). One microgram of RNA was reverse transcribed into cDNA using the Prime Script RT Reagent kit (Takara, Japan), and the system was diluted fivefold upon completion of reverse transcription for the next experiment. SYBR Premix Ex Taq (Takara) and the ABI 7500 Real-time PCR system (Applied Biosystems) were used to carry out quantificational real-time polymerase chain reaction (qRT-PCR) experiments. The protocol was performed as follows: step 1: 95°C for 30 s; step 2: 40 cycles of 95°C for 5 s, and 60°C for 34 s; and step 3: melting curve analysis. For each sample, three biological repeats were performed to obtain reliable results (Livak and Schmittgen, 2001). The specificity of the qRT-PCR primers was demonstrated by ePCR and melting curves. The cotton histone-3 gene (GenBank accession number AF024716) was used as an internal reference gene to normalize the expression level of the target gene (Liu et al., 2017). The data were calculated according to the $2^{-\Delta\Delta Ct}$ formula (Livak and Schmittgen, 2001). Gene specific primers for qRT-PCR were designed by Oligo 7 (**Table 2**).

RESULTS

Identification and Characteristics of K^+ Efflux Antiporters in *Gossypium* spp.

Based on the HMM model of the KEA specific protein conserved domain constructed by the ATKEA1–ATKEA6 protein sequences, a total of 8, 8, 15, and 16 KEA members were identified from *G. arboreum*, *G. raimondii*, *G. hirsutum*, and *G. barbadense*, respectively (**Table 3**). All these putative genes were detected to contain the typical Na^+/H^+ exchanger domain (pfam: PF00999) and some members contained the TrkA-N domain (pfam: PF02254) of the KEA gene family in CDD and SMART databases. The confirmed members of the KEA gene family were named *GaKEA1* to *GaKEA8, GrKEA1* to *GrKEA8*, *GhKEA1* to *GhKEA15* and *GbKEA1* to *GbKEA16* according to the size of the e-values screened. These putative GaKEAs

TABLE 2 | List of the primers used for quantitative real-time PCR in present study.

ID	Name	Forward primer	Reverse primer
Ghir_A08G013810.1	GhKEA1	CGTGCACTGGACCTTCCTGTTT	CTGCAGCACATGCTCTTTCAGC
Ghir_A12G011370.1	GhKEA2	GGGGATTTCAATGGCCCTCACA	ATCAAGGGCAACGAATGG
Ghir_D08G014670.1	GhKEA3	TAGAAAAGGCTGGTGCTACGGC	CGTTGATCGTTGCCGCAATCTC
Ghir_D12G011600.1	GhKEA4	TTTAGGTCTCGCCATCTTGCGG	TTATCGTCCGAAGAATCCGGCG
Ghir_D02G024430.1	GhKEA5	AGGGCTGCTGATTTCGTTGACA	CAGAAGCAAGCGGAGTCGACAA
Ghir_A03G023000.1	GhKEA6	AAGCCGGTGCAACAGATGCAAT	AGCCTTATCGATTCTCGCCTGC
Ghir_D13G021510.1	GhKEA7	ATCGAGCAGATGATGCACCGAC	AAAAGCAATGCCACCGCATGTT
Ghir_A13G020690.1	GhKEA8	TTTGCTGCTCTTTTCCTCGCCA	GTTGTGCCCAGAAGTAGCAGGT
Ghir_A07G004510.1	GhKEA9	GCGAAGGGAGCATTGCCAAAAT	GCTACAGTCTCCAGTACAGCTTGC
Ghir_D06G009670.1	GhKEA10	CAACATGCTTCACGGCCAAGTC	AGGAAGGCCTGTACGGGACAAT
Ghir_A08G025770.1	GhKEA11	TCTTAGTTTGGTGACTACT	GAATGGCAATGCGGCATC
Ghir_A06G009360.1	GhKEA12	ACCGAAGGACGGTACTTTTGCC	GTCTTCACTCTGGCCACGGTTT
Ghir_D08G026630.1	GhKEA13	GATGCTCATTTTCCGGGT	TTAAGAAATTGGCAGTAAAAGT
Ghir_D07G004540.1	GhKEA14	TGGCGATTTGCTCCTCGAGA	GCCGGTACCCGTCACCGA
Ghir_A07G015190.1	GhKEA15	AGGTTAAGTTCCATGAAG	ACTTAGTTACCTGCAGGA

TABLE 3 | Basic information for KEAs in *Arabidopsis thaliana*, *G. arboreum*, *G. raimondii*, *G. hirsutum*, and *G. barbadense*.

ID	NAME	Chr.	Length	Mw(kDa)	pI	GRAVY	Subcellular localization	Location
AT1G01790	ATKEA1	1	1193	128.034	5.22	0.08	Chloroplast	284350–291203
AT4G00630	ATKEA2	4	1185	127.605	5.11	0.086	Chloroplast	261246–268097
AT4G04850	ATKEA3	4	776	83.7907	5.53	0.369	Chloroplast	2452664–2457767
AT2G19600	ATKEA4	2	592	64.2494	5.91	0.589	Plasma membrane	8478970–8483854
AT5G51710	ATKEA5	5	568	61.5984	5.84	0.613	Plasma membrane	21004251–21008849
AT5G11800	ATKEA6	5	597	64.3917	7.1	0.599	Plasma membrane	3803315–3808273
Ga08G1558.1	GaKEA1	8	1205	130.33	5.37	0.03	Chloroplast	104013215–104021677(–)
Ga12G1877.1	GaKEA2	12	1209	129.982	5.23	0.12	Chloroplast	30757828-30767684(+)
Ga03G2727.1	GaKEA3	3	972	106.397	8.58	0.199	Chloroplast	135169648–135176726(+)
Ga07G1590.1	GaKEA4	7	867	94.8015	5.99	− 0.089	Chloroplast	32599799–32639902(+)
Ga13G2411.1	GaKEA5	13	599	64.6596	6.03	0.557	Plasma membrane	119381339–119388157(+)
Ga07G0481.1	GaKEA6	7	1119	122.991	6.28	0.24	Plasma membrane	5200681–5212745(–)
Ga08G2879.1	GaKEA7	8	968	104.966	8.39	0.113	Plasma membrane	128629936–128641870(–)
Ga06G0996.1	GaKEA8	6	595	64.6759	5.57	0.559	Plasma membrane	32202261–32208138(–)
Gbar_A08G014340.1	GbKEA1	A08	1180	127.738	5.39	0	Chloroplast	96262469–96271708(–)
Gbar_A12G011370.1	GbKEA2	A12	1209	130.115	5.26	0.117	Chloroplast	73274768–73284631(–)
Gbar_D08G015210.1	GbKEA3	D08	1205	130.231	5.44	0.057	Chloroplast	47645632–47654679(+)
Gbar_D12G011550.1	GbKEA4	D12	1209	129.909	5.26	0.092	Chloroplast	36785162–36794816(–)
Gbar_D07G015620.1	GbKEA5	D07	920	99.0304	5.16	0.294	Chloroplast	24522424–24536031(+)
Gbar_D02G025040.1	GbKEA6	D02	791	86.0319	6.24	0.284	Chloroplast	67204008–67209499(+)
Gbar_A03G023150.1	GbKEA7	A03	789	86.0191	6.36	0.302	Chloroplast	104812638–104817596(+)
Gbar_D13G021490.1	GbKEA8	D13	599	64.6225	6.03	0.548	Plasma membrane	55916493–55924208(+)
Gbar_A13G021130.1	GbKEA9	A13	599	64.6596	6.03	0.557	Plasma membrane	104841961–104849912(+)
Gbar_A07G004290.1	GbKEA10	A07	575	62.205	5.91	0.661	Plasma membrane	5132112–5138369(–)
Gbar_D07G004590.1	GbKEA11	D07	575	62.1859	5.91	0.657	Plasma membrane	4798294–4804446(–)
Gbar_A08G026670.1	GbKEA12	A08	557	60.2731	5.82	0.715	Plasma membrane	119100202–119107770(–)
Gbar_D06G009660.1	GbKEA13	D06	600	65.2856	5.57	0.561	Plasma membrane	18540497–18547146(+)
Gbar_D08G027360.1	GbKEA14	D08	478	51.4877	5.35	0.753	Plasma membrane	65273584–65280951(–)
Gbar_A06G009270.1	GbKEA15	A06	563	61.3481	5.68	0.575	Plasma membrane	29086781–29093508(–)
Gbar_A07G015190.1	GbKEA16	A07	441	48.0017	6.5	0.242	Chloroplast	32524616–32547893(+)
Ghir_A08G013810.1	GhKEA1	A08	1205	130.202	5.31	0.03	Chloroplast	99333580–99342035(–)
Ghir_A12G011370.1	GhKEA2	A12	1210	130.22	5.18	0.143	Chloroplast	78653991–78663846(–)
Ghir_D08G014670.1	GhKEA3	D08	1110	119.902	5.97	0.099	Chloroplast	50279059–50287371(+)
Ghir_D12G011600.1	GhKEA4	D12	1209	129.825	5.26	0.086	Chloroplast	39774617–39784377(–)
Ghir_D02G024430.1	GhKEA5	D02	791	85.9718	6.24	0.279	Chloroplast	69322443–69327418(+)
Ghir_A03G023000.1	GhKEA6	A03	794	86.6378	6.67	0.288	Chloroplast	112523990–112529630(+)
Ghir_D13G021510.1	GhKEA7	D13	599	64.6225	6.03	0.548	Plasma membrane	59326338–59334263(+)
Ghir_A13G020690.1	GhKEA8	A13	599	64.6596	6.03	0.557	Plasma membrane	103926397–103934287(+)
Ghir_A07G004510.1	GhKEA9	A07	575	62.2781	5.91	0.653	Plasma membrane	5212492–5219763(–)
Ghir_D06G009670.1	GhKEA10	D06	596	64.794	5.57	0.554	Plasma membrane	19282756–19289573(+)
Ghir_A08G025770.1	GhKEA11	A08	588	63.8421	5.62	0.69	Plasma membrane	121887172–121893956(–)
Ghir_A06G009360.1	GhKEA12	A06	595	64.6178	5.54	0.575	Plasma membrane	31118011–31124755(–)
Ghir_D08G026630.1	GhKEA13	D08	545	58.6921	5.43	0.714	Plasma membrane	68121405–68136594(–)
Ghir_D07G004540.1	GhKEA14	D07	523	56.316	5.76	0.693	Plasma membrane	4811397–4822915(–)
Ghir_A07G015190.1	GhKEA15	A07	121	12.7251	6.24	0.908	Chloroplast	33266007–33266823(+)
Gorai.004G157200.1	GrKEA1	4	1205	130.313	5.36	0.055	Chloroplast	44542807–44551407(–)
Gorai.008G118000.1	GrKEA2	8	1209	129.88	5.26	0.084	Chloroplast	35331031–35340753(–)
Gorai.005G262000.1	GrKEA3	5	791	86.0218	6.24	0.28	Chloroplast	63648334–63653813(+)
Gorai.001G170200.1	GrKEA4	1	1011	109.902	5.85	0.085	Chloroplast	24364847–24371649(+)
Gorai.013G228100.1	GrKEA5	13	599	64.6225	6.03	0.548	Plasma membrane	54744028–54751788(+)
Gorai.004G283900.1	GrKEA6	4	573	61.7578	5.69	0.718	Plasma membrane	61514964–61521493(–)
Gorai.001G047000.1	GrKEA7	1	575	62.1759	5.91	0.648	Plasma membrane	4455718–4462880(–)
Gorai.010G101600.1	GrKEA8	10	596	64.794	5.57	0.554	Plasma membrane	18371755–18378385(+)

TABLE 4 | The KEA member ID and their names of seven species.

Arabidopsis thaliana ID	Name	*Oryza sativa* ID	Name 2	*Zea mays* ID	Name 3	*Populus trichocarpa* ID	Name 4	*Sorghum bicolor* ID	Name 5	*Triticum aestivum* ID	Name 6	*Glycine max* ID	Name 7
AT1G01790	ATKEA1	LOC_Os04g58620.1	OsKEA1	GRMZM2G093643_P01	ZmKEA1	Potri.002G157200.1.p	PtKEA1	Sobic.006G271800.1.p	SbKEA1	Traes_2AL_A2CBDB5F7.1	TaKEA1	Glyma.03G014000.1.p	GmKEA1
AT4G00630	ATKEA2	LOC_Os12g42300.1	OsKEA2	GRMZM2G009715_P01	ZmKEA2	Potri.014G080800.1.p	PtKEA2	Sobic.008G173800.1.p	SbKEA2	Traes_2BL_BA71DD685.1	TaKEA2	Glyma.07G073700.1.p	GmKEA2
AT4G04850	ATKEA3	LOC_Os06g36590.1	OsKEA3	GRMZM2G169114_P01	ZmKEA3	Potri.009G080800.2.p	PtKEA3	Sobic.010G168900.1.p	SbKEA3	Traes_2DL_053E73CFE.1	TaKEA3	Glyma.09G262000.1.p	GmKEA3
AT2G19600	ATKEA4	LOC_Os03g03590.1	OsKEA4	GRMZM2G171031_P01	ZmKEA4	Potri.006G230400.1.p	PtKEA4	Sobic.001G522100.1.p	SbKEA4	Traes_5DS_49CF8A4C4.1	TaKEA4	Glyma.18G230100.1.p	GmKEA4
AT5G51710	ATKEA5			GRMZM2G058948_P01	ZmKEA5	Potri.012G130500.1.p	PtKEA5			Traes_5BS_6759CB1AD.1	TaKEA5	Glyma.08G064800.1.p	GmKEA5
AT5G11800	ATKEA6			GRMZM2G040158_P01	ZmKEA6	Potri.018G054100.2.p	PtKEA6			Traes_5AS_AF90452F5.1	TaKEA6	Glyma.07G184800.1.p	GmKEA6
				GRMZM2G474078_P01	ZmKEA7	Potri.015G132400.1.p	PtKEA7			Traes_7DL_51216CD52.1	TaKEA7	Glyma.05G222500.1.p	GmKEA7
										Traes_7AL_23AA36A1B.1	TaKEA8	Glyma.08G029300.1.p	GmKEA8
										Traes_4BL_723A7E539.1	TaKEA9	Glyma.17G229700.1.p	GmKEA9
										Traes_7BL_B01B8C760.1	TaKEA10	Glyma.14G093900.1.p	GmKEA10
										Traes_5AL_91132361F.1	TaKEA11	Glyma.16G203200.1.p	GmKEA11
										Traes_4DL_59D5A8BCA.1	TaKEA12	Glyma.09G152300.1.p	GmKEA12
										Traes_4DL_806404E38.1	TaKEA13		

encoded proteins ranging from 595 amino acids (aa) (GaKEA8) to 1209 aa (GaKEA2), while GrKEAs encoded ranging from 573 aa (GrKEA6) to 1209 aa (GrKEA2), GhKEAs encoded 121 aa (GhKEA15) to 1210 aa (GhKEA2) and GbKEAs encoded 441 aa (GbKEA16) to 1209 aa (GbKEA2 and GbKEA4). Most members of the KEA family have transmembrane domains, which means that these proteins may be located on the membrane. The location of these proteins on chromosomes and the predicted molecular weight (MW), isoelectric point (pI) and grand average of hydropathicity (GRAVY) are shown in **Table 3**.

The results of subcellular localization prediction indicated that the proteins in KEAI were found to be located in the chloroplast (**Table 3**). *AtKEA1*, *-2*, *-3* in *Arabidopsis* belong to the KEAI clade were shown to be subcellular localized in chloroplasts (Aranda-Sicilia et al., 2012; Kunz et al., 2014; Sheng et al., 2014). Online tool prediction showed that the proteins in the KEAII clade were located in the plasma membrane (**Table 3**). However, *KEA4*, *-5*, *-6* in *Arabidopsis* belong to the KEAII clade were shown to be subcellular localized to the Golgi, trans-Golgi reticulum, and the prevacuolar compartment/multivesicular bodies (Zhu et al., 2018; Wang Y. et al., 2019). This is inconsistent with the online website prediction results, and the subcellular localization results of KEA in upland cotton require further validation.

Phylogenetic Analysis

Using the same method, a total of 4, 7, 7, 4,13, and 12 members were identified from *Oryza sativa*, *Zea mays*, *Populus trichocarpa*, *Sorghum bicolor*, *Triticum aestivum,* and *Glycine max*, respectively (**Table 4**). To examine the evolutionary relationship of KEA proteins, an unrooted phylogenetic tree was constructed using 94 KEA protein sequences from 11 species (excluding TaKEA8, TaKEA12, TaKEA13, ZmKEA6, ZmKEA7, OsKEA4 short sequences, as they did not meet the requirement of 1000 bootstrap replicates) (**Figure 1**). The phylogenetic tree was divided into three main categories namely KEAIa, KEAIb, and KEAII, which were consistent with the members of the AtKEA gene family (Aranda-Sicilia et al., 2012; Chanroj et al., 2012). The 11 species had distributed members in all three classifications. Among these members, the KEAII branch was the largest group, containing 46 members, while branch KEAIb was the smallest, with only 16 members. However, in the two large categories of KEAI and KEAII, the distributions of members were basically uniform. For example, 7 and 8 members of upland cotton were distributed in branch KEAI and branch KEAII, respectively. By checking the sequence characteristics, we determined that the sequence of KEAIa was the longest (except *GhKEA15* and *GbKEA16*), and the length of KEAII was the shortest.

Chromosome Distribution and Gene Replication Events

To determine the evolutionary relationship of KEA genes in cotton, the number and location of genes on the chromosome were analyzed (**Figure 2**). Each chromosome contained only one or two KEA genes. All *GaKEAs* and *GrKEAs* were distributed on 6 chromosomes, while *GhKEAs* and *GbKEAs* were distributed on six chromosomes of A_t subgenomes and six chromosomes

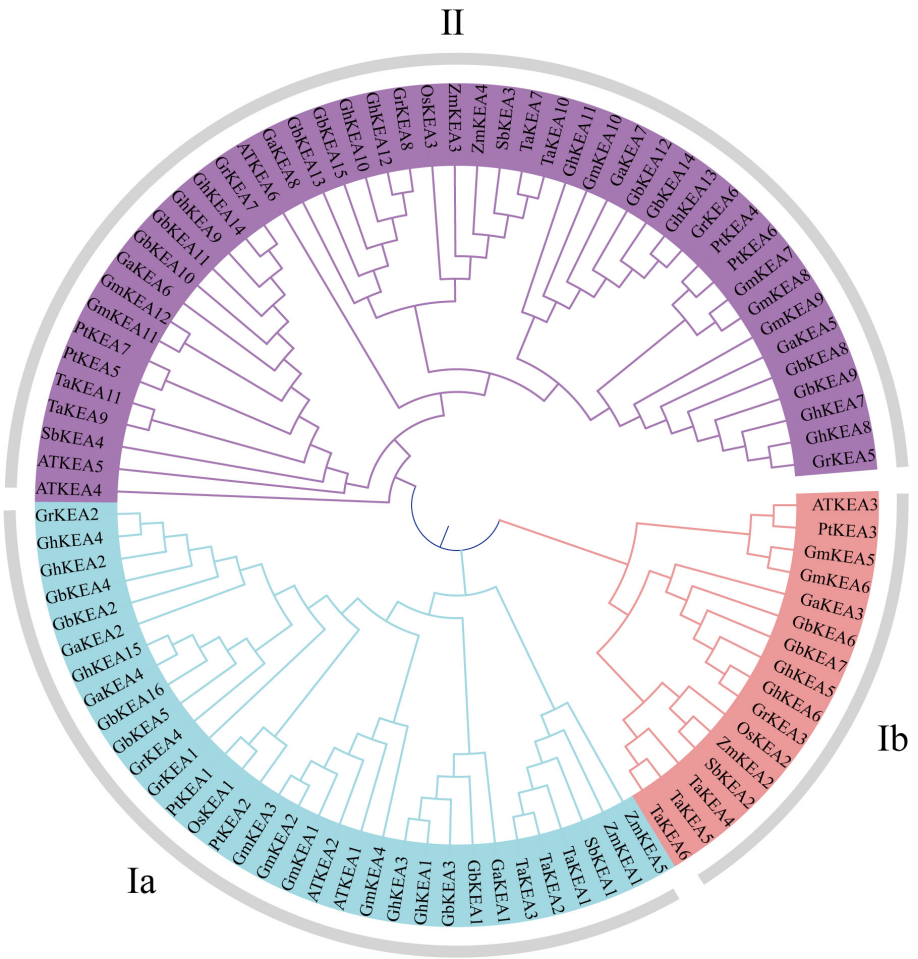

FIGURE 1 | Phylogenetic relat ionship of KEA gene family. The analysis included full-length protein sequences from *G. hirsutum*, *G. raimondii*, *G. barbadense*, *G. arboreum*, *Arabidopsis thaliana*, *Oryza sativa*, *Zea mays*, *Populus trichocarpa*, *Sorghum bicolor*, *Triticum aestivum,* and *Glycine max*. Using MEGA software, 1000 bootstrap repetitive phylogenetic trees were constructed by neighbor-joining method.

of D_t subgenomes. Comparing the location and number of chromosomes on which the *GaKEAs* were located with the At subgenomic chromosomes on which the *GhKEAs* and the *GbKEAs* were located, we found that the location and number of genes distributed on these chromosomes of the KEA genes were basically the same. This phenomenon suggested that the distributions of KEA genes in the cotton genome were relatively conservative.

Gene replication events are very important for the expansion of the gene family. In general, gene replication events include tandem repeats and segmental repeats (Cannon et al., 2004; Xu et al., 2012). In this study, tandem repeat genes were defined as adjacent homologous genes on a single chromosome, and there was no more than one intermediate gene. A total of 9, 12, and 10 gene duplication pairs were identified between the A_t and D_t subgenomes of *G. hirsutum* and their corresponding ancestral A and D diploid genomes, respectively (**Table 5**). The data showed that all members of the KEA gene family were amplified only by segmental replication, which meant that segmental replication played a key role in the evolution of the KEA gene family. The syntenic relationships of putative KEA genes among two diploid genomes (*G. arboreum* and *G. raimondii*) and subgenomes in cultivated allotetraploid (*G. hirsutum*) were shown in **Figure 3**. The results showed that *GaKEAs* and *GrKEAs* were distributed among 5 and 6 chromosomes of A geneme and D genome, respectively, whereas *GhKEAs* were distributed among 6 and 6 chromosomes of A_t and D_t subgenomes, respectively. The Ka:Ks ratio can be used to judge whether the homologous gene is under positive selection pressure (Ka:Ks $>$ 1) or purification selection pressure (Ka:Ks $<$ 1). The results showed that the Ka:Ks ratios of GhKEA gene pairs were less than 1, indicating that KEA homologous gene pairs had undergone purifying selection during evolution and may have similar functions (**Table 5**).

Analysis of Gene Structure and Conservative Motif Distribution

Phylogenetic analysis showed that the GhKEA gene family was divided into three groups, containing 8, 2, and 5 members (**Figure 4A**). The sequences of most members of each subfamily were similar, indicating their close evolutionary relationship. The

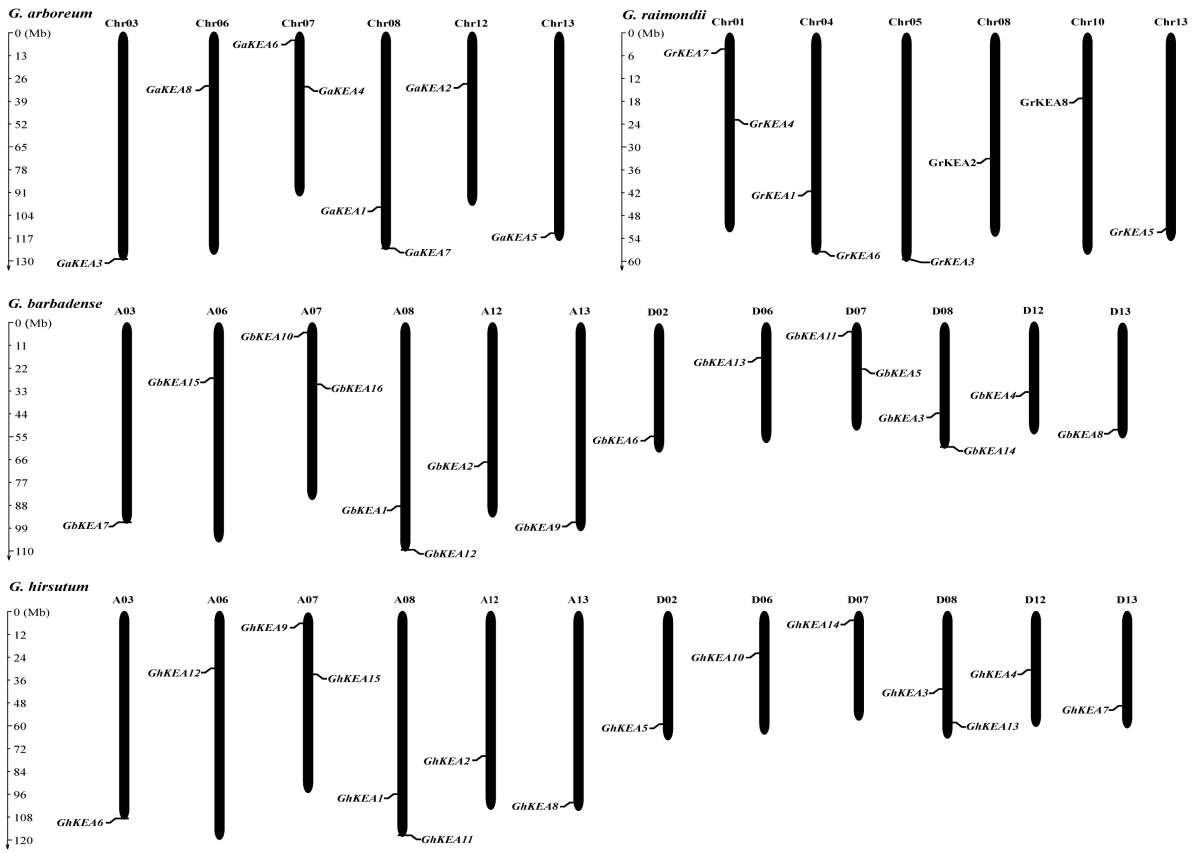

FIGURE 2 | Chromosome distribution of KEAs in *G. arboreum*, *G. raimondii*, *G. hirsutum,* and *G. barbadense*. The scale stands for mega-base (Mb). The chromosome numbers are shown above each vertical line.

gene length of *GhKEA15* was the shortest and that of *GhKEA13* was the longest. The exon numbers of *GhKEAs* were ranged from 4 to 20, but most of the genes contained at least 15 exons, except that *GhKEA15* contained only 4 exons (**Figure 4B**). The GhKEA proteins were further analyzed by the MEME program and 10 conserved motifs were identified (**Figure 4C**). Most GhKEA members contained multiple motifs, except GhKEA15 (1 motif), and motif 1, motif 2, motif 3, motif 4, motif 7, and motif 10 were widely distributed in these members. Motif 5 existed in the KEAIb and KEAII classes, motif 6 only did not exist in KEAIb, and motif 8 and motif 9 existed only in the KEAII class. Members of the same subfamily have similar motif characteristics, exon-intron structures and gene lengths, supporting a close evolutionary relationship.

Analysis of *Cis*-Elements in Putative GhKEA Promoter Regions

To analyze the *cis*-elements that may be involved in the regulation of *GhKEAs*, the upstream 2000 bp sequence from the start codon (ATG) of each GhKEA gene was extracted for analysis. The *cis*-elements were classified into hormone response elements, stress response elements and plant growth and development elements (**Figure 5**). The hormone response elements included mainly salicylic acid (SA), methyl jasmonate (MeJA), gibberellin (GA), auxin (IAA) and abscisic acid (ABA). Most of the *GhKEAs* promoter regions contained 2–4 hormone response elements, except *GhKEA11* (1 ABA response element). Of these, *GhKEA5*, *GhKEA6*, *GhKEA8,* and *GhKEA12* contained the most hormone response elements, while *GhKEA14* contained the largest number. Among these hormone response elements located in the promoter regions of *GhKEAs*, the largest number is the MeJA response element, followed by the ABA response element (**Figure 5A**). Previous studies have shown that MeJA and ABA were the main plant hormone signaling molecules under stress (Mantyla et al., 1995; Pichersky and Gershenzon, 2002), so we speculated that *GhKEAs* might be involved in various stress responses of upland cotton.

The stress response elements and plant growth and development elements of the *GhKEAs* were shown in **Figure 5B**. *GhKEA1* contained most kinds of elements, including zein metabolism regulation elements (O2-site), low temperature response elements (LTR), endosperm expression regulatory elements (GCN4_motif), drought response elements (MBS), defense and stress elements (TC-rich repeats) and anaerobic induction elements (ARE). In addition, *GhKEA7* contained the largest number of components, while *GhKEA9* contained only one ARE element. Among the elements, the content of anaerobic

TABLE 5 | The Ka/Ks ratio of repetitive gene pairs between upland cotton A and D subgenomes and their corresponding ancestor An and D diploid genomes.

Gene 1	Gene 2	Ka	Ks	Ka/Ks	Duplicate
Ghir_A08G013810.1	Ghir_D08G014670.1	0.016473214	0.041483262	0.397105071	Segmental
Ghir_A12G011370.1	Ghir_D12G011600.1	0.019168907	0.044895994	0.426962539	Segmental
Ghir_A07G015180.1	Ghir_D08G014670.1	0.448822637	1.042093598	0.43069321	Segmental
Ghir_A03G023000.1	Ghir_D02G024430.1	0.010157691	0.044162866	0.23000525	Segmental
Ghir_A08G025770.1	Ghir_D13G021510.1	0.078953692	0.37203022	0.212223868	Segmental
Ghir_A13G020690.1	Ghir_D13G021510.1	0.00294515	0.030413524	0.096836869	Segmental
Ghir_A07G004510.1	Ghir_D07G004540.1	0.012239361	0.041468956	0.295145138	Segmental
Ghir_A06G009360.1	Ghir_D06G009670.1	0.005201259	0.040182906	0.129439599	Segmental
Ghir_A08G025770.1	Ghir_D08G026630.1	0.032581423	0.098093352	0.332147107	Segmental
Ghir_A08G013810.1	Ga07G1589.1	0.221236177	0.412846079	0.535880533	Segmental
Ghir_A08G013810.1	Ga08G1558.1	0.002555524	0.002299293	1.111439207	Segmental
Ghir_A08G013810.1	Ga12G1877.1	0.110829884	0.357277495	0.310206731	Segmental
Ghir_A12G011370.1	Ga07G1589.1	0.227066948	0.444887909	0.510391368	Segmental
Ghir_A12G011370.1	Ga08G1558.1	0.121462593	0.359390179	0.337968593	Segmental
Ghir_A12G011370.1	Ga12G1877.1	0.010670819	0.008999527	1.185708825	Segmental
Ghir_A13G020690.1	Ga08G2879.1	0.090571368	0.401754117	0.2254398	Segmental
Ghir_A13G020690.1	Ga13G2411.1	0	0.002297972	0	Segmental
Ghir_A08G025770.1	Ga08G2879.1	0.043530715	0.059247855	0.734722203	Segmental
Ghir_A08G025770.1	Ga13G2411.1	0.08243898	0.379636239	0.217152556	Segmental
Ghir_A06G009360.1	Ga06G0996.1	0.002966263	0.002308582	1.2848851	Segmental
Ghir_A07G015190.1	Ga07G1590.1	0.027526025	0.029670124	0.927735441	Segmental
Ghir_D08G014670.1	Gorai.001G170200.1	0.159009405	0.324757398	0.489625195	Segmental
Ghir_D12G011600.1	Gorai.004G157200.1	0.111218686	0.34816119	0.319445962	Segmental
Ghir_D12G011600.1	Gorai.008G118000.1	0.00145059	0.004628751	0.313386833	Segmental
Ghir_D02G024430.1	Gorai.005G262000.1	0.001678635	0.01382926	0.121382884	Segmental
Ghir_D13G021510.1	Gorai.004G283900.1	0.046007176	0.349474957	0.13164656	Segmental
Ghir_D13G021510.1	Gorai.013G228100.1	0	0.004594195	0	Segmental
Ghir_D06G009670.1	Gorai.010G101600.1	0	0.004611852	0	Segmental
Ghir_D08G026630.1	Gorai.004G283900.1	0.030198482	0.036981567	0.816581995	Segmental
Ghir_D08G026630.1	Gorai.013G228100.1	0.068394632	0.369738838	0.184980924	Segmental
Ghir_D07G004540.1	Gorai.001G047000.1	0.008809839	0.019963568	0.441295826	Segmental

inducible elements (AREs) was the highest, followed by Zein metabolic regulatory elements (O2 sites). In previous study, zein and its hydrolyzates have antioxidant activity (Diaz-Gomez et al., 2018), so we inferred that *GhKEAs* could have high antioxidant activity and antioxidant stability. These results further indicated that the KEA genes might be involved in the stress response of upland cotton.

Organ Expression Pattern Analysis of *GhKEAs*

To determine the expressions of *GhKEAs* in various tissues of upland cotton, the transcription data of *GhKEAs* in different tissues (anther, filament, pistil, petal, root, leaf, and stem) in TM-1 were analyzed. As shown in **Figure 6**, *GhKEAs* were widely expressed in different tissues, and the same gene was highly expressed in several different tissues. The expressions of *GhKEAs* in these tissues could be divided into three groups. Group a contained *GhKEA1*, *GhKEA3*, *GhKEA5,* and *GhKEA6*, which were expressed in anthers, petals and leaves. Group b consisted of *GhKEA9*, *GhKEA10*, *GhKEA11,* and *GhKEA13*, which were highly expressed in the pistils. The last group c was composed of *GhKEA2*, *GhKEA4*, *GhKEA7*, *GhKEA8*, *GhKEA12,* and *GhKEA14*, these genes were highly expressed in the pistils, roots and stems. The multiple expression patterns indicated that the functions of *GhKEAs* had been differentiated in long-term evolution.

GhKEAs Expression Patterns Under Multiple Stresses

According to the analysis of *cis*-elements in the promoter region and previous studies on the KEA genes in other plants, *GhKEAs* might be involved in the stress response. To test this hypothesis, we used the available transcriptome data to analyze the expression profiles of 15 *GhKEAs* under salt and drought treatments, and further verified by qRT-PCR experiment. As shown in **Figure 7A**, *GhKEAs* were regulated by PEG treatments. The transcriptome data showed that the expression levels of *GhKEA2*, *GhKEA4*, *GhKEA9*, *GhKEA10*, *GhKEA12*, *GhKEA13,* and *GhKEA14* were significantly upregulated under PEG stress. However, the expression levels of *GhKEA1*, *GhKEA3*, *GhKEA5*, *GhKEA6,* and *GhKEA7* were significantly downregulated. In addition, the expression levels of *GhKEA8* and *GhKEA11* were

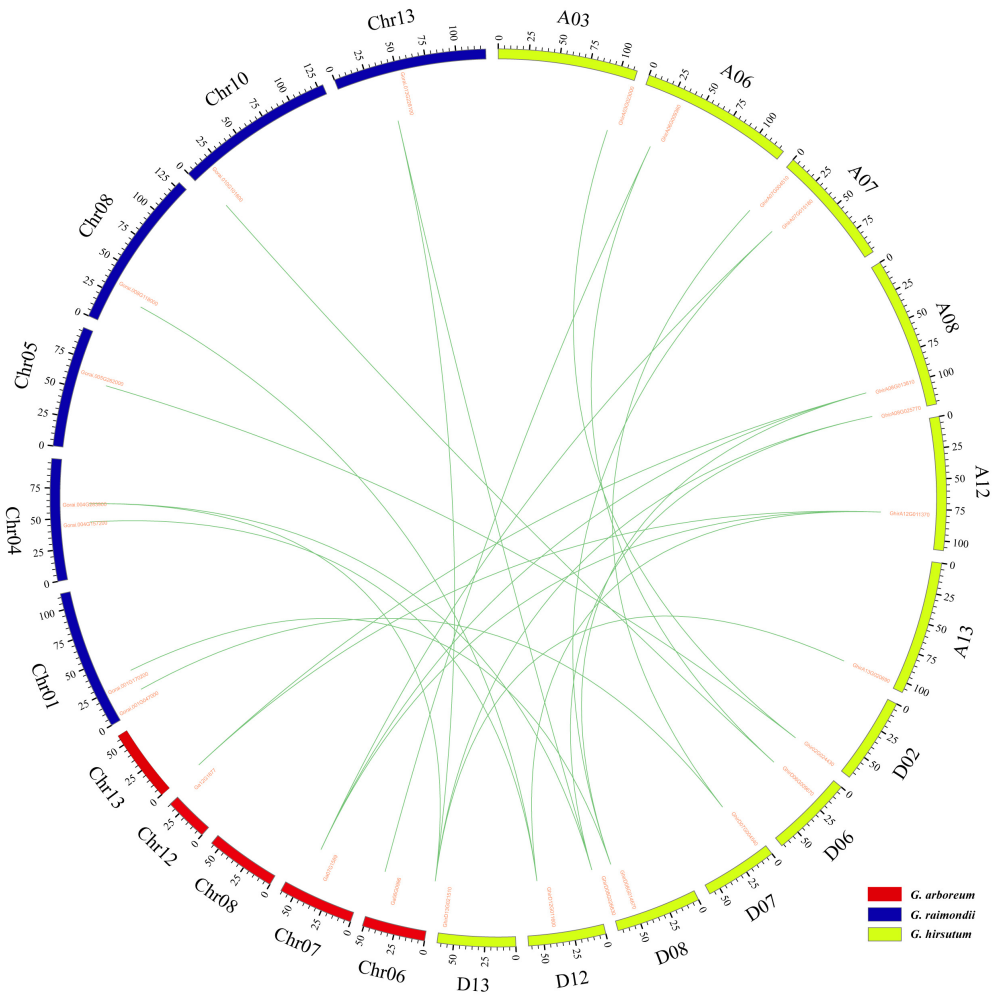

FIGURE 3 | Syntenic relationships among KEA genes of two diploid (*G. arboreum*, *G. raimondii*) and one allotetraploid (*G. hirsutum*) cotton was visualized in a circos plot. The chromosomes of *G. arboreum*, *G. raimondii*, and *G. hirsutum* were shaded with red, blue, and green colors, respectively.

decreased at first and then increased. Then, we selected five genes from each of the above upregulated and downregulated groups to carry out qRT-PCR experiments and further verify their response to drought stress. The data showed that *GhKEAs* expressions could be regulated by PEG treatment (**Figure 7B**). After PEG treatment, the expressions of *GhKEA2*, *GhKEA4*, *GhKEA10*, *GhKEA12,* and *GhKEA14* were increased, while the expressions of *GhKEA1*, *GhKEA3*, *GhKEA5*, *GhKEA6,* and *GhKEA7* were decreased, which was basically consistent with the results in the transcriptome database.

At the same time, we also analyzed the gene expression pattern under salt conditions in publicly available RNA-seq data, and the results showed that the expressions of *GhKEAs* were induced by salt stress (**Figure 8A**). The expressions of *GhKEA9*, *GhKEA10,* and *GhKEA12* were upregulated, while the expressions of *GhKEA5*, *GhKEA6,* and *GhKEA11* were downregulated. What's more, the expressions of *GhKEA7*, *GhKEA8*, *GhKEA13,* and *GhKEA14* were upregulated at first and then downregulated, and reached the highest level at 1 h after treatment, while the expressions of *GhKEA1*, *GhKEA2*, *GhKEA3,* and *GhKEA4* were also upregulated and then downregulated. The difference was that the expressions of these 4 genes reached the highest level at 3–6 h. The inconsistent expression patterns of *GhKEAs* under the same stress may be due to their different promoter elements, resulting in their possible regulation by different upstream genes and thus affecting their expression patterns. We selected 10 *GhKEAs* from the above groups to further investigate the effect of salt stress on their expressions by qRT-PCR (**Figure 8B**). The results were basically consistent with the RNA-seq data; the expression levels of *GhKEA2*, *GhKEA7*, *GhKEA8*, *GhKEA13,* and *GhKEA14* were upregulated at first and then downregulated, *GhKEA5* and *GhKEA11* were downregulated, and *GhKEA9*, *GhKEA10,* and *GhKEA12* were upregulated.

To verify the potential roles of *GhKEAs* in potassium absorption and transport, qRT-PCR experiments were used to observe whether the *GhKEAs* responded to low potassium treatment. The results are shown in **Figure 9**. The expressions of *GhKEAs* responded to low potassium treatment, and 15 *GhKEAs*

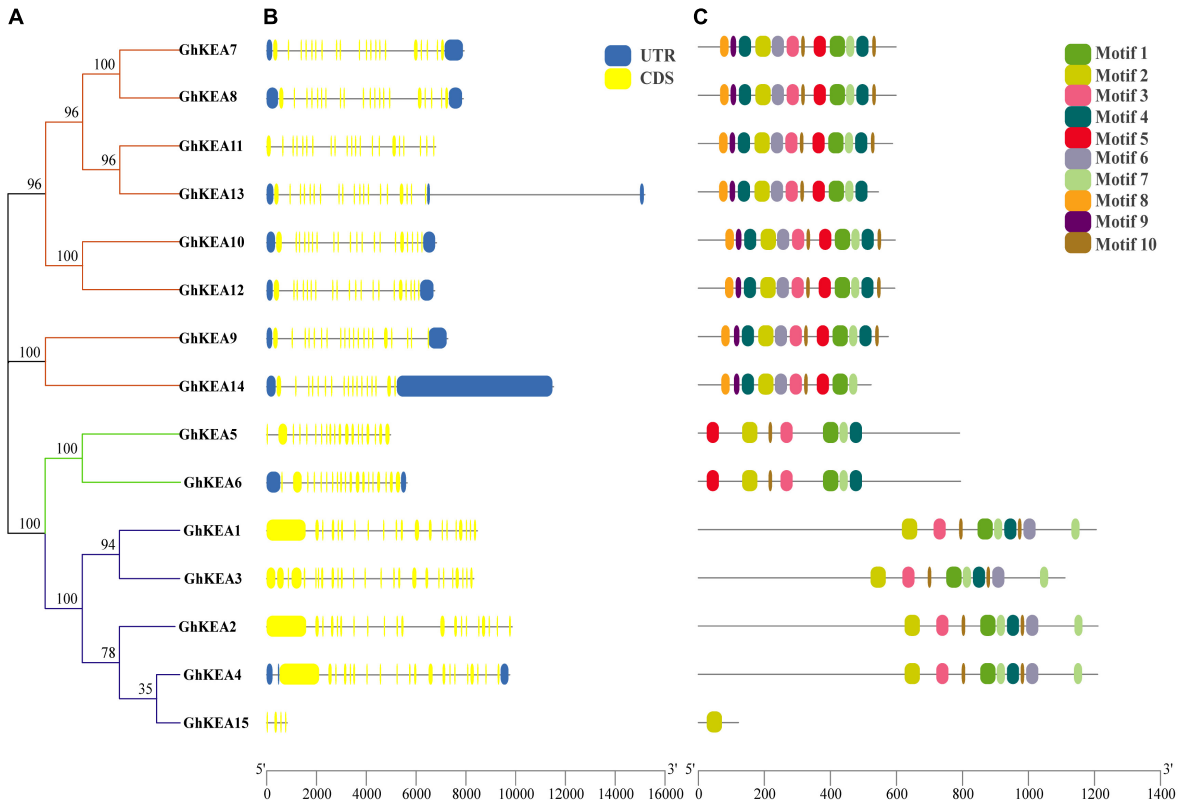

FIGURE 4 | Phylogenetic relationship, exon-intron structure, and conserved motif of KEA genes in *G. hirsutum*. **(A)** The rootless tree is constructed in MEGA, and the three subfamilies are represented by different colors. **(B)** The number, length, and location of exons and introns in *GhKEAs*. Boxes and lines represent exons and introns, respectively. **(C)** Distribution of conserved motifs in the GhKEA gene family.

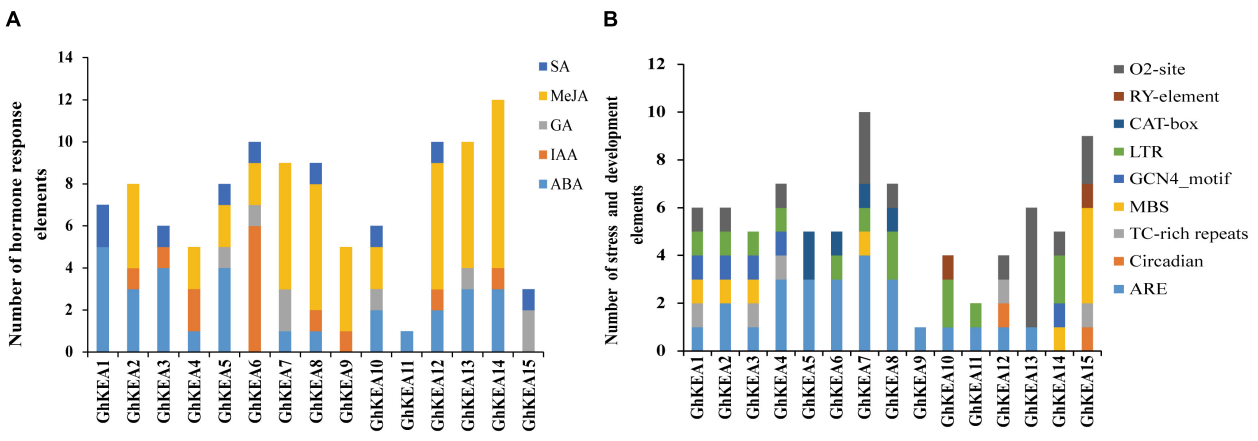

FIGURE 5 | The number of plant hormone *cis*-acting elements and induced stress *cis*-acting elements in the promoter fragment of *GhKEAs*. **(A)** The sum of hormone response *cis*-acting elements in each member of *GhKEAs*. **(B)** The sum of *cis*-acting elements of stress and development in each member of *GhKEAs*.

were divided into three groups according to their expression characteristics. In group a, the expressions of *GhKEAs* were downregulated after low potassium treatment. The expressions of *GhKEA5*, *GhKEA6,* and *GhKEA15* in group b were upregulated at first, then downregulated, and then upregulated. The gene expressions in group c first increased and then decreased. Among them, the expressions of *GhKEA7*, *GhKEA8*, *GhKEA9,* and *GhKEA12* reached the highest level at 1 h after low potassium treatment, while the expressions of *GhKEA2*, *GhKEA4,* and *GhKEA11* reached the highest level at 12 h, and only *GhKEA10* reached the highest level at 3 h. These results indicated that *GhKEAs* might regulate the abiotic stress response to

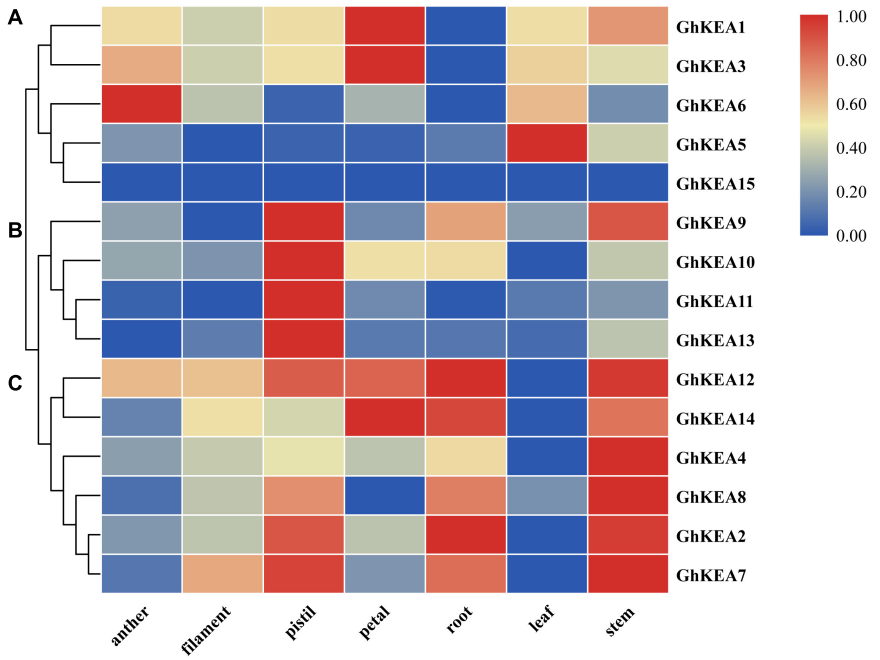

FIGURE 6 | Hierarchical clustering of *GhKEAs* expression levels in ten different tissues of TM-1. The genes are displayed on the right side of each line, and the phylogenetic relationship is shown on the left **(A–C)**.

low potassium ions in the environment and participate in the absorption and transport of potassium ions in cotton.

Silencing of *GhKEA4* and *GhKEA12* Compromise the Tolerance of Cotton to Salt Stress

Analysis of the promoter regions of *GhKEAs* revealed that the promoter regions of both *GhKEA4* and *GhKEA12* contain ABA and MeJA hormone response elements as well as antioxidant response elements ARE. *GhKEA12* was significantly differentially expressed under low potassium stress and significantly up-regulated by drought and salt stress. The expression of *GhKEA4* was significantly up-regulated under drought stress, while it also responded to low potassium and high salt stress. Then, *GhKEA4* and *GhKEA12* were selected for the VIGS experiments and salt treatments. The albino phenotype on the pYL156-*CLA1* cotton plant ensured the success of the VIGS experiment (**Figure 10A**). As shown in **Figure 10B**, the expressions of the *GhKEA4* and *GhKEA12* in the corresponding VIGS plants leaves were significantly lower than the expressions in the pYL156 empty vector plants, indicating that the genes had been silenced successfully. Then these plants were treated with salt to observe the phenotype. **Figure 10A** showed that after salt stress, cotton plants silenced for *GhKEA4* and *GhKEA12* showed obvious wilting compared with control cotton plants. Subsequently, the leaves of these plants were treated with drought *in vitro*, to calculate the water loss rate of detached leaves. The results showed that the water loss rates of cotton leaves were significantly higher than that of control plants after silencing *GhKEA4* and *GhKEA12* genes, indicating that the water holding capacity of leaves decreased after silencing the target gene (**Figure 10E**). In order to further investigate the effects of salt stress on the physiological and biochemical characteristics of plants, the chlorophyll content, proline content, soluble sugar content, peroxidase (POD) activity and catalase (CAT) activity in VIGS cotton leaves were measured after salt stress. The results showed that after salt treatment, CAT and POD in control plants were extremely significantly and significantly higher than those in plants silenced for the target gene (**Figure 10F**). Furthermore, the soluble sugar, proline and total chlorophyll contents in the leaves of control plants were significantly higher than those of plants silenced for the target gene (**Figures 10G–I**). These results indicated that the salt tolerance of cotton plants decreased after silencing *GhKEA4* and silencing *GhKEA12*.

In addition, the results of the expression levels of related genes encoding K and Na transporters showed that *GhAKT2* was down-regulated, *GhHKT1*, *GhPOT11*, *GhNHX1*, and *GhNHX6* were up-regulated in plants silenced for the target gene, while *GhNHX2* was up-regulated in plants silenced for *GhKEA4* and down-regulated in plants silenced for *GhKEA12* (**Figures 10C,D**).

K^+ and Na^+ Contents in Virus-Induced Gene Silencing Cotton Plants Under High Salt, High Potassium, and Low Potassium Stress

In order to preliminarily characterize the potassium ion transport activity of KEA gene, the K^+ and Na^+ content in cotton silenced for *GhKEA4* and *GhKEA12* before and after high salt, high

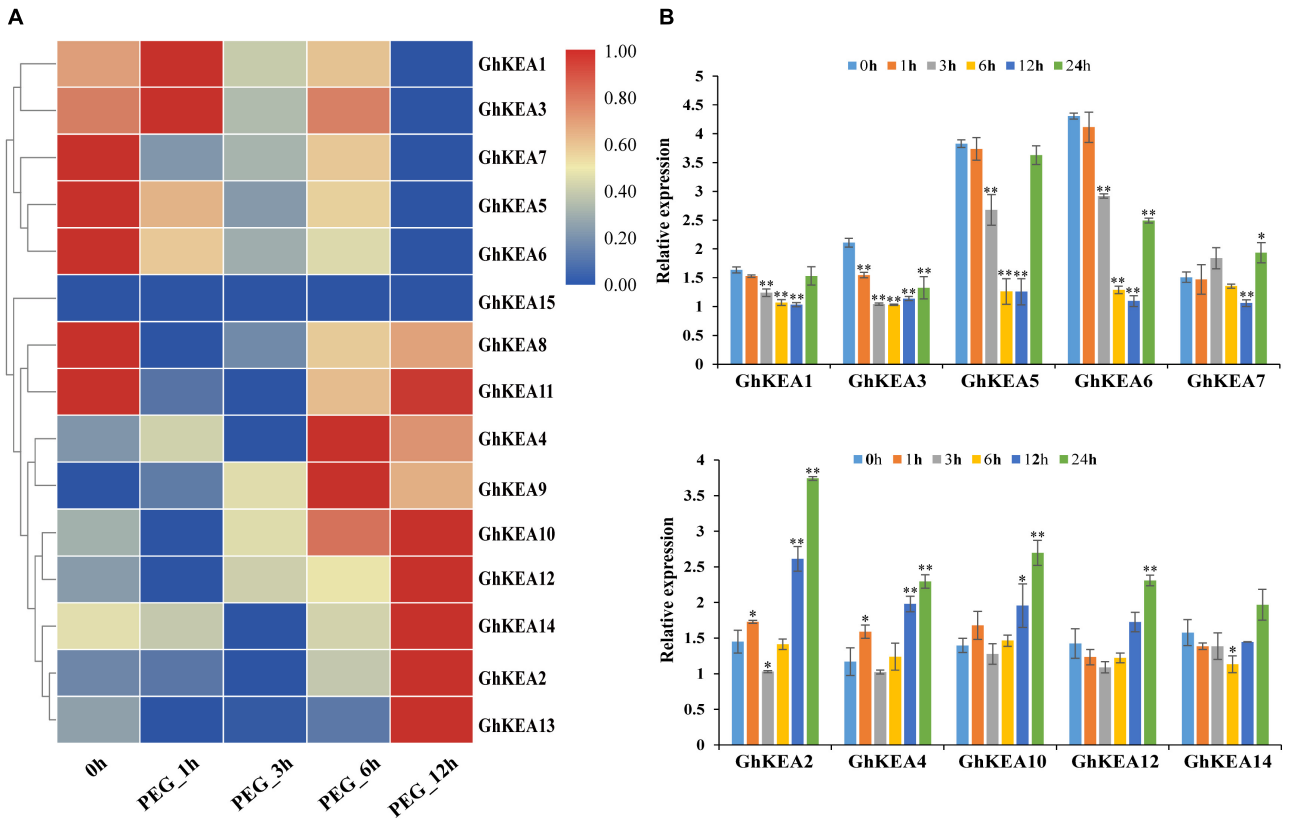

FIGURE 7 | Expression profiles of *GhKEAs* under drought stress. **(A)** Hierarchical clustering of 15 *GhKEAs* expression levels under PEG treatment. **(B)** Quantificational real-time polymerase chain reaction (qRT-PCR) results of 10 *GhKEAs* under PEG treatment. 0, 1, 3, 6, 12, and 24 h represent hours after PEG treatment. The experiments were repeated three times and provided consistent results. The gene names are shown on the right. Color blocks represent the relative expression levels of *GhKEAs*. Error bars indicates SD (**$p < 0.01$, $0.01 < *p < 0.05$, $n = 3$).

potassium, and low potassium treatments were measured. The results showed that under control conditions (untreated), after silencing *GhKEA4* in cotton plants, the Na^+ content in stems and roots was extremely significantly higher than that in empty vector control plants, the K^+ content in the entire plant was significantly increased, but only the Na^+/K^+ ratio in roots was extremely significantly increased (**Figures 11B,C**). After silencing the *GhKEA12* gene in cotton plants, the Na^+ content in leaves was extremely significantly higher than that in empty vector control plants, the K^+ content in the entire plant was extremely significantly increased, and only the Na^+/K^+ ratio in stems was extremely significantly decreased (**Figure 11A**). The results showed that the Na^+/K^+ balance could be basically maintained after silencing *GhKEA4* and *GhKEA12* genes.

Under salt stress, after silencing the *GhKEA4* gene in cotton plants, the Na^+ content in leaves and roots was significantly higher than that in unloaded control plants, the K^+ content was significantly decreased in leaves and stems and significantly increased in roots, and the Na^+/K^+ ratio was significantly increased in leaves and significantly decreased in roots (**Figure 11**), indicating that silencing the *GhKEA4* gene decreased the potassium ion transport activity from the lower ground to the shoot of plants under salt stress, resulting in a higher Na^+/K^+ ratio in leaves. After silencing *GhKEA12*, the Na^+ content in the plants was extremely significantly increased, the K^+ content was extremely significantly increased in the roots, significantly decreased in the leaves and stems, and the Na^+/K^+ ratio in the plants was extremely significantly increased (**Figure 11**), indicating that the potassium transport activity was decreased and the sodium transport activity was increased in the plants silenced *GhKEA12*, resulting in the plants with a higher Na^+/K^+ ratio.

Under high potassium treatment, the contents of Na^+ and K^+ in plants silenced with *GhKEA4* increased significantly, but their Na/K ratio did not change significantly (**Figure 11**), indicating that cotton silenced with *GhKEA4* could still maintain a more stable Na^+ and K^+ balance. Additionally, Na^+/K^+ ratio was significantly increased in leaves and stems and significantly decreased in roots in *GhKEA12*-silenced plants, indicating that potassium ion transport activity from roots to leaves was reduced in plants with *GhKEA12* gene silencing.

Under low potassium stress, the Na^+ content in leaves and roots of plants silenced for the *GhKEA4* was extremely significantly decreased and significantly increased in stems. K^+ content was significantly increased in leaves and stems and extremely significantly decreased in roots. Na^+/K^+ ratio was extremely significantly decreased in leaves and stems and significantly increased in roots (**Figure 11**), indicating that

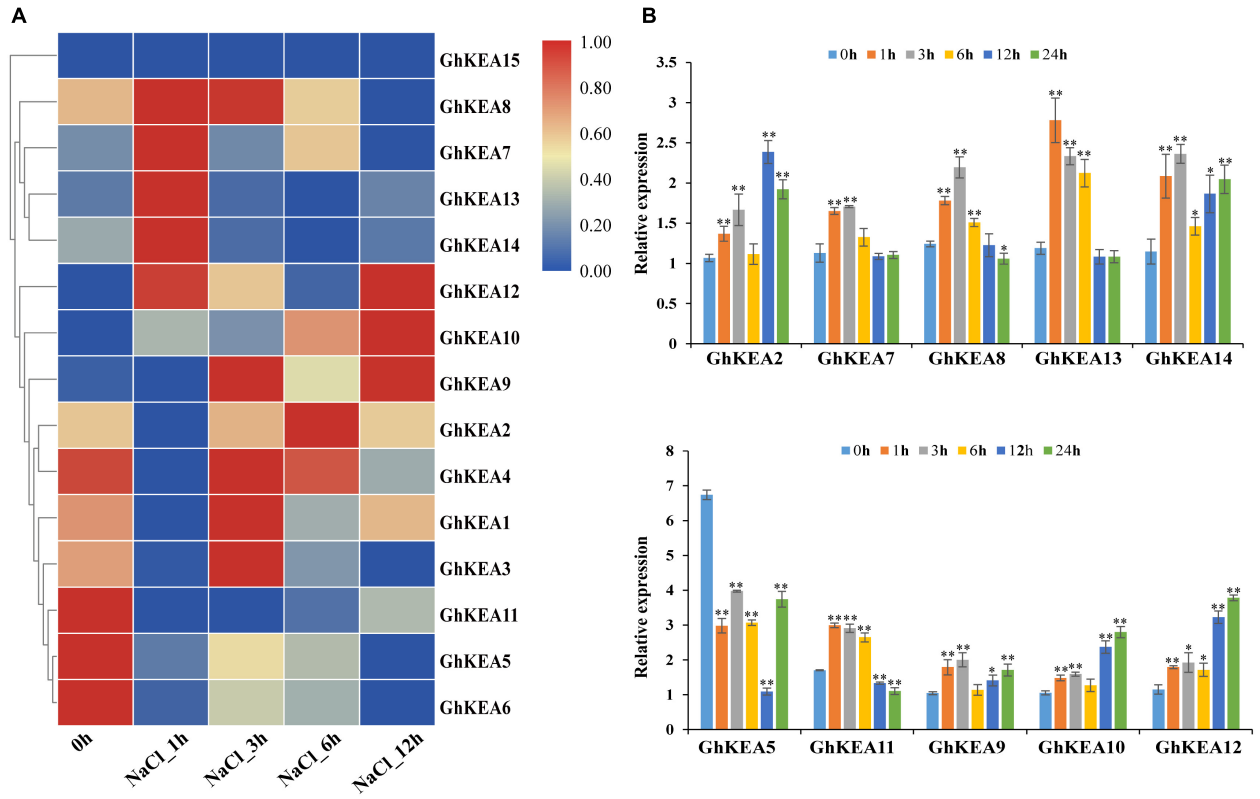

FIGURE 8 | Expression profiles of *GhKEAs* under salt stress. **(A)** Hierarchical clustering of 15 *GhKEAs* expression levels under NaCl treatment. **(B)** qRT-PCR results of 10 *GhKEAs* under NaCl treatment. 0, 1, 3, 6, 12, and 24 h represent hours after NaCl treatment. The experiments were repeated three times and provided consistent results. The gene names are shown on the right. Color blocks represent the relative expression levels of *GhKEAs*. Error bars indicates SD (**$p < 0.01$, $0.01 < *p < 0.05$, $n = 3$).

silencing of *GhKEA4* gene increased the activity of plants to transport potassium to shoots under low potassium conditions; K^+ content in plants silenced with *GhKEA12* was extremely significantly increased, the Na^+ content in leaves and roots was extremely significantly increased in stems, Na^+/K^+ was extremely significantly increased in leaves, and significantly decreased in stems and roots, indicating that silencing with *GhKEA12* gene improved the ability of plants to absorb potassium, but accumulated more Na^+ in leaves, resulting in higher Na^+/K^+ in leaves.

DISCUSSION

The *AtKEAs* are homologous to EcKefB and EcKefC of *Escherichia coli* (Chanroj et al., 2012). When EcKefB/EcKefC binds to its helper proteins EcKefF and glutathione, the conformation of the KTN domain changes, which turns on the potassium ion transport switch of EcKefB/EcKefC (Miller et al., 2000; Roosild et al., 2002, 2009, 2010). Some studies have shown that the expressions of *AtKEA1*, *AtKEA3,* and *AtKEA4* were enhanced under low potassium stress, and the expressions of *AtKEA2* and *AtKEA5* were enhanced under sorbitol and abscisic acid treatment (Aranda-Sicilia et al., 2012; Kunz et al., 2014; Zhu et al., 2018). The CPA family in some plants has been identified and verified (Maser et al., 2001; Chanroj et al., 2012; Ye et al., 2013; Zhou et al., 2016; Sharma et al., 2020), but there are few studies to identify the CPA family in cotton, especially the KEA family, which is a subfamily of the CPA family. In the current study, we identified the members of the KEA family in cotton by sequence similarity, and then carried out a comprehensive bioinformatics analysis of the KEA gene in cotton. The comprehensive analysis of the characteristics of the cotton KEA gene family will provide a basis for further research.

Evolution and Characterization of the K^+ Efflux Antiporter Gene Family in Cotton Species

Genome-wide doubling events occurring in the process of plant evolution have had a lasting and far-reaching impact on plants, and some plants have even experienced whole-genome doubling events repeatedly (Tuskan et al., 2006). Approximately 130 million years ago, the common ancestor of dicotyledons experienced a genome-wide triploid event (Jaillon et al., 2007). Then, cotton independently experienced a genome-wide pentaploid event (Paterson et al., 2012; Wang et al., 2016). Allotetraploid upland cotton enlarged the number of genes

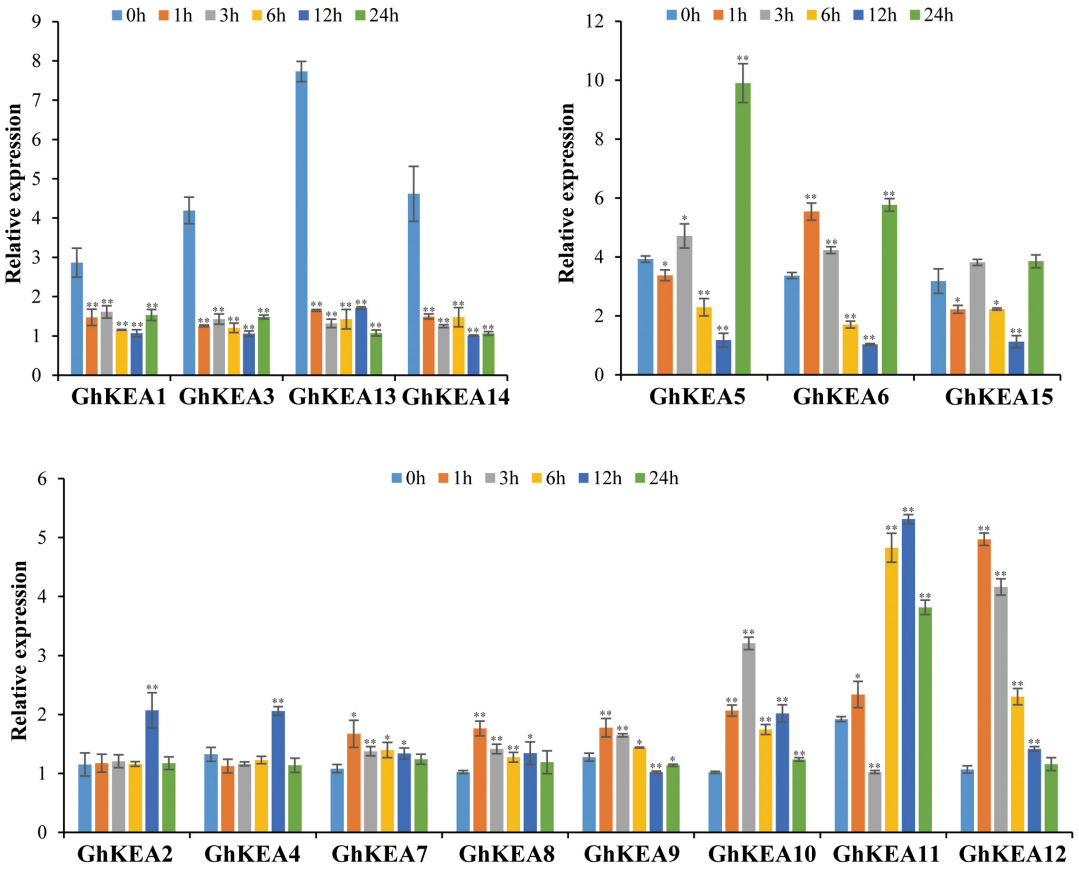

FIGURE 9 | The qRT-PCR results of *GhKEAs* under low potassium treatment. 0, 1, 3, 6, 12, and 24 h represent hours after low potassium treatment. Error bars show the standard deviation of three biological repeats.

after multiple replication events. In this study, 8, 8, 15, and 16 KEA genes were identified in *G. raimondii*, *G. arboreum*, *G. hirsutum,* and *G. barbadense*, respectively. The number of KEA genes in tetraploid cotton is approximately the sum of KEA genes in *G. arboreum* and *G. raimondii*. The unbalanced distribution of KEA genes on each chromosome number proved the existence of genetic variation in the process of evolution (Paterson et al., 2012; Zhu and Li, 2013; Yu et al., 2014). Based on the classification of the KEA gene family in *Arabidopsis* (Maser et al., 2001), all putative KEA genes in this study can be divided into three subfamilies. The lineal homologous genes of monocotyledons tend to form lineal homologous gene pairs at the end of the branches of phylogenetic trees, while the *KEA* genes of dicotyledons tend to be clustered together, possibly due to the different functions of KEA proteins in monocotyledons and dicotyledons (Li et al., 2016). Furthermore, the gene members of each subfamily not only have similar gene structure, sequence length and motif structure, but also have the same results of subcellular location prediction. These results suggested that the members of the KEA gene family may show relatively conservative functions in the growth of upland cotton, especially those of the same subfamily (Palusa et al., 2007). The similarities and differences in the gene structure, domain and motif of *GhKEAs* may be related to the long evolutionary history and gene replication of cotton (He and Zhang, 2005). In the process of gene family evolution, tandem replication and segmented replication contributed to the emergence of gene families to a certain extent. We found that *GhKEAs* could be amplified only by segmental replication, indicating that segmental replication played a key role in the evolution of the GhKEA gene family. Collinear analysis showed that most of the KEA homologous gene pairs between the A_t and D_t subgenomes of *G. hirsutum* and their corresponding A and D diploid genomes were located in the collinear region. Based on these results, we speculated that whole genome replication was the main driving force for the expansion of the KEA gene from diploid to allotetraploid.

GhKEAs May Play an Important Role in Facilitating K^+ Homeostasis

Maintaining the homeostasis of intracellular ions is not only the basic cellular activity needed for plant growth, but also the basis for regulating plant growth and development and coping with environmental stress (Yang et al., 2019). Previous studies have shown that *AtKEA1*, *-3*, and *-4* are induced by low potassium stress (Zheng et al., 2013) and that the AtKEA gene family plays a key role in K^+ homeostasis and osmoregulation

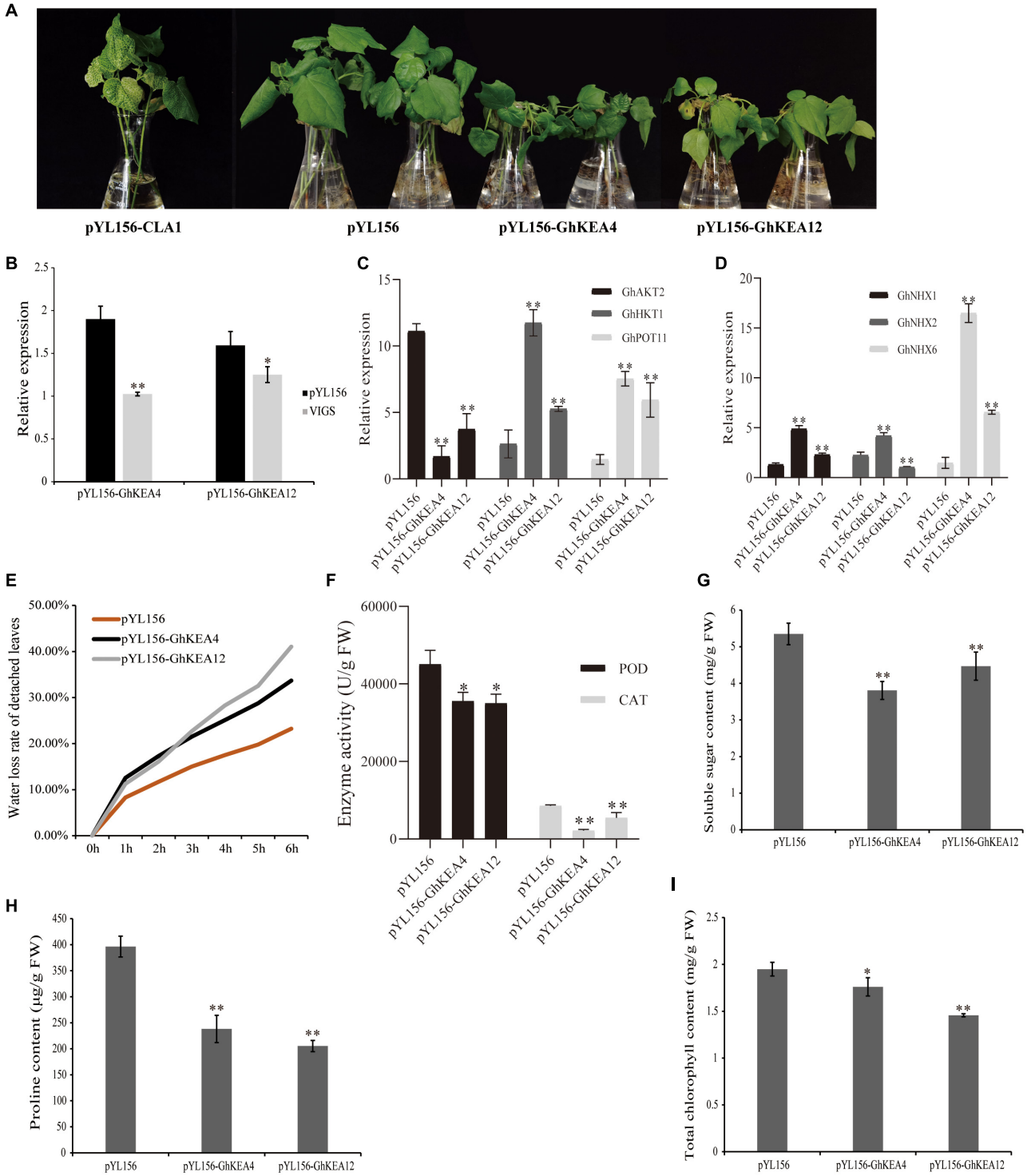

FIGURE 10 | Silencing *GhKEA4* and Silencing *GhKEA12* in cotton can destroy the tolerance of cotton to salt stress. **(A)** The phenotypes of positive control, empty control, and virus-induced gene silencing (VIGS) plants under salt treatment. **(B)** The expression levels of *GhKEA4* and *GhKEA12* in pYL156 and VIGS plants leaves. **(C,D)** Relative expression levels of K and Na transport related genes in pYL156 and VIGS plants leaves. **(E)** The water loss rate of isolated leaves of control plants and VIGS plants leaves. **(F)** The activities of CAT and POD in empty control and VIGS plants leaves after salt treatment. **(G–I)** After salt treatment, the contents of soluble sugar, proline and chlorophyll in empty control and VIGS plants leaves. Error bars indicate SD (**$p < 0.01$, $0.01 < {}^{*}p < 0.05$, $n = 3$).

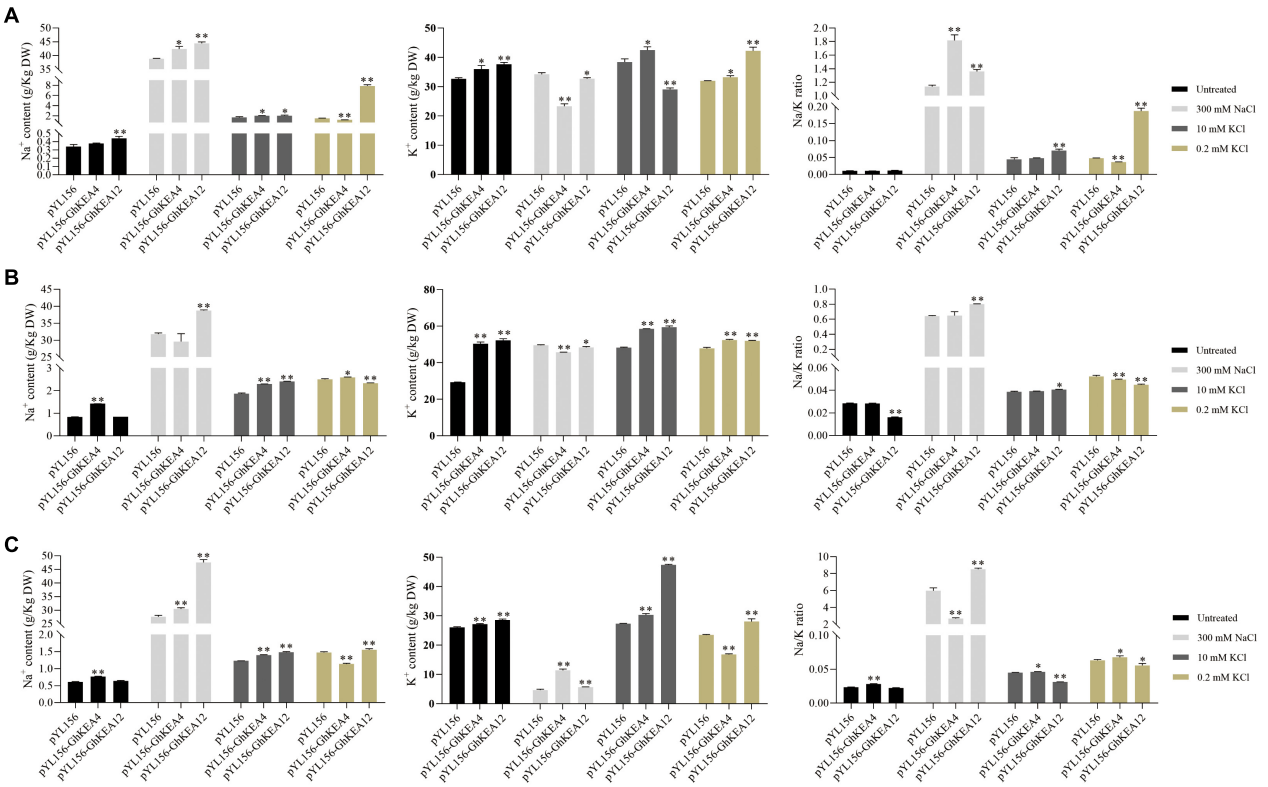

FIGURE 11 | Changes in Na^+, K^+ content and Na^+/K^+ ratio in plants injected with empty vector and in plants silenced for *GhKEA4* and silenced for *GhKEA12* under high salt, high and low potassium stress. **(A)** Leaf. **(B)** Stem. **(C)** Root. Error bars indicate SD (**$p < 0.01$, $0.01 < *p < 0.05$, $n = 3$).

(Zheng et al., 2013; Kunz et al., 2014; Zhu et al., 2018). In this report, the expression characteristics of *GhKEAs* under low potassium stress were characterized for the first time, and most of the *GhKEAs* were induced by low potassium stress. For example, *GhKEA12*, as an ortholog of *AtKEA4*, showed significant changes in its expression levels under low potassium stress. Many studies have demonstrated that the KEA family with Na^+/H^+ exchanger domains and NAD-binding (KTN) domains participates in the absorption and transport of potassium ions (Qiu et al., 2003; Bölter et al., 2020). Therefore, to preliminarily explore the involvement of *GhKEAs* in K^+ transport, the contents of K^+ and Na^+ in the leaves of silenced *GhKEA4* cotton plants and silenced *GhKEA12* cotton plants were examined. In addition, silencing *GhKEA4* and *GhKEA12* decreased the activity of transporting potassium ions in leaves under high salt condition, resulting in higher Na/K ratio in leaves. Silencing *GhKEA12* inhibited K^+ transport activity and increased Na^+ content in cotton leaves under high potassium stress. Silencing *GhKEA12* increased K^+ transport activity and Na^+ absorption capacity under low potassium stress. However, the effect of silencing *GhKEA4* gene on K^+ transport activity in plant leaves was very low under high and low potassium stresses. *GhKEA4* is an ortholog of *AtKEA2*, and it has been shown that *AtKEA2* is involved in K^+ homeostasis in chloroplasts or plastids (Aranda-Sicilia et al., 2012), which is consistent with our results. Alternatively, it has also been shown that *AtKEA2* has an important function in maintaining local osmotic pressure, ionic and pH homeostasis, and the formation of thylakoid membranes (Kunz et al., 2014; Ali, 2016; Aranda-Sicilia et al., 2016). While *GhKEA12* is an ortholog of *AtKEA6*, *AtKEA6* likewise plays a role in maintaining ion homeostasis (Zhu et al., 2018), as indicated by the K^+ uptake system (Tsujii et al., 2019). Overall, our results demonstrated that the *GhKEAs* were involved in regulating the dynamic balance of intracellular K^+ during the growth and development of cotton.

It has been shown that *AKT1* is involved in K^+ uptake in the micromolar concentration range (Spalding et al., 1999); *HKT1* and *KUP7* have an important role in regulating K homeostasis in plants (Han et al., 2016; Wang et al., 2018); and Na^+ and H^+ exchange rates are significantly increased in vacuoles of plants overexpressing the *AtNHX1* gene (Apse et al., 1999). We found significant changes in the expression levels of these genes in cotton silenced for *GhKEA4* or *GhKEA12* genes. These results suggest that there may be a "genetic compensation" mechanism for potassium ion absorption and transport in cotton.

GhKEAs Regulate Cotton Response to Salt Stress

We analyzed the *cis*-acting elements in the promoter region of the upland cotton KEA gene family and found a variety

of stress-responsive *cis*-acting elements such as TC-rich repeats and MBS. Studies have shown that ABA and MeJA played an important role in regulating plant stress responses (Reyes and Chua, 2007; Tavallali and Karimi, 2019). ABA response elements (ABREs) and MeJA response elements (TGACG-elements and CGTCA-motifs) were obtained in the *cis*-acting elements of the GhKEA gene family. Moreover, qRT-PCR results showed that the GhKEA gene family responded positively to stress in the early or late stages of salt and drought treatments. When plants are subjected to salt stress, a large number of reactive oxygen species accumulate, leading to lipid peroxidation and interfering with the normal physiological process (Jithesh et al., 2006). POD and CAT are important antioxidant enzymes for scavenging reactive oxygen species in plants (Harb et al., 2010). At the same time, proline and soluble sugar, as important substances regulating plant cell osmotic potential, promote the scavenging of intracellular reactive oxygen species to some extent (Ullah et al., 2018). In this study, *GhKEA4* and *GhKEA12*, which were sensitive to salt treatment, were selected for VIGS experiments. The results showed that the water loss rate of detached leaves, the contents of proline and soluble sugar, and the activities of POD and CAT in VIGS plants were lower than those in the blank control to varying degrees. *AtKEA2* and *AtKEA6*, paralogs of *GhKEA4* and *GhKEA12*, have been reported to confer tolerance to high Na^+ stress in *Arabidopsis* (Aranda-Sicilia et al., 2012; Wang Y. et al., 2019). It has been reported that salt stress-induced production of ROS can promote K^+ entry into the cytoplasm, thereby reducing the Na^+/K^+ ratio (Ma et al., 2012). We speculated that excessive accumulation of ROS in plants under salt stress may further promote *GhKEAs* to transport K^+, reduce the Na^+/K^+ ratio, and then promote the accumulation of osmoregulatory substances to improve tolerance to abiotic stress, and the conclusion needs further verification.

CONCLUSION

Under salt stress, plants accumulate a large amount of Na^+ and inhibit the absorption of K^+, resulting in an imbalance in the ion dynamic balance. Ion transporters can maintain ion homeostasis in plant inner membrane systems. In the present study, K^+ efflux transporters were identified in *Gossypium* spp. Then, the distribution, sequence structure and expression pattern of the KEA gene family in cotton were analyzed in detail at the whole genome level, as well as its potential function in cotton growth and development and response to abiotic stress. In addition, VIGS experiments were used to be verified that the *GhKEAs* could maintain a relatively stable Na/K ratio in upland cotton under high salt, high potassium and low potassium stresses, as well as play an important function in salt stress. The comprehensive analysis of KEA genes in this study lays a foundation for future functional research on cotton KEA genes.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at the National Center for Biotechnology Information Search database (https://www.ncbi.nlm.nih.gov/) under accession numbers PRJNA382310, PRJNA171262, PRJNA433615, PRJNA10719, PRJNA448171, PRJNA374837, ACUP00000000, and PRJNA248163.

AUTHOR CONTRIBUTIONS

YL: conceptualization, methodology, data curation, software, original draft, and writing – review and editing. ZF: methodology and software. HWe, SC, and PH: methodology. HWa: conceptualization, supervision, and writing – review and editing. SY: conceptualization and supervision. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (32072112), the Agricultural Science and Technology Innovation Program of Chinese Academy of Agricultural Sciences and the China Agriculture Research System (CARS-15-06).

REFERENCES

Ali, M. A. (2016). *A study of the role of KEA1* and *KEA2* K^+/H^+ antiporters in chloroplast development and division in *arabidopsis thaliana*. dissertation/master's thesis. Granada: Universidad de Granada.

Apse, M. P., Aharon, G. S., Snedden, W. A., and Blumwald, E. (1999). Salt tolerance conferred by overexpression of a vacuolar Na^+/H^+ antiport in *Arabidopsis*. *Science* 285, 1256–1258. doi: 10.1126/science.285.5431.1256

Aranda-Sicilia, M. N., Aboukila, A., Armbruster, U., Cagnac, O., Schumann, T., Kunz, H. H., et al. (2016). Envelope K^+/H^+ antiporters *AtKEA1* and *AtKEA2* function in plastid development. *Plant Physiol.* 172, 441–449. doi: 10.1104/pp.16.00995

Aranda-Sicilia, M. N., Cagnac, O., Chanroj, S., Sze, H., Rodriguez-Rosales, M. P., and Venema, K. (2012). *Arabidopsis KEA2*, a homolog of bacterial KefC, encodes a K^+/H^+ antiporter with a chloroplast transit peptide. *Biochim. Biophys. Acta* 1818, 2362–2371. doi: 10.1016/j.bbamem.2012.04.011

Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., De Castro, E., et al. (2012). ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res.* 40, W597–W603. doi: 10.1093/nar/gks400

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

Bao, S. D. (2005). *Agricultural and Chemistry Analysis of Soil*. Beijing: China Agriculture Press.

Bölter, B., Mitterreiter, M. J., Schwenkert, S., Finkemeier, I., and Kunz, H.-H. (2020). The topology of plastid inner envelope potassium cation efflux antiporter *KEA1* provides new insights into its regulatory features. *Photosynth Res.* 145, 43–54. doi: 10.1007/s11120-019-00700-2

Booth, I. R. (2003). Bacterial ion channels. *Genet Eng.* 25, 91–111. doi: 10.1007/978-1-4615-0073-5_5

Britto, D. T., and Kronzucker, H. J. (2008). Cellular mechanisms of potassium transport in plants. *Physiol. Plant.* 133, 637–650. doi: 10.1111/j.1399-3054.2008.01067.x

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

Chanroj, S., Lu, Y., Padmanaban, S., Nanatani, K., Uozumi, N., Rao, R., et al. (2011). Plant-specific cation/H^+ exchanger 17 and its homologs are endomembrane K^+ transporters with roles in protein sorting. *J. Biol. Chem.* 286, 33931–33941. doi: 10.1074/jbc.M111.252650

Chanroj, S., Wang, G., Venema, K., Zhang, M. W., Delwiche, C. F., and Sze, H. (2012). Conserved and diversified gene families of monovalent cation/h^+ antiporters from algae to flowering plants. *Front. Plant Sci.* 3:25. doi: 10.3389/fpls.2012.00025

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

Choe, S. (2002). Potassium channel structures. *Nat. Rev. Neurosci.* 3, 115–121. doi: 10.1038/nrn727

Cuin, T. A., Miller, A. J., Laurie, S. A., and Leigh, R. A. (2003). Potassium activities in cell compartments of salt-grown barley leaves. *J. Exp. Bot.* 54, 657–661. doi: 10.1093/jxb/erg072

Diaz-Gomez, J. L., Ortiz-Martinez, M., Aguilar, O., Garcia-Lara, S., and Castorena-Torres, F. (2018). Antioxidant activity of zein hydrolysates from zea species and their cytotoxic effects in a hepatic Cell Culture. *Molecules* 23:2. doi: 10.3390/molecules23020312

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

Fan, K., Mao, Z., Zheng, J., Chen, Y., Li, Z., Lin, W., et al. (2020). Molecular evolution and expansion of the KUP family in the allopolyploid cotton species *Gossypium hirsutum* and *Gossypium barbadense*. *Front. Plant Sci.* 11:545042. doi: 10.3389/fpls.2020.545042

Fu, X., Lu, Z., Wei, H., Zhang, J., Yang, X., Wu, A., et al. (2020). Genome-wide identification and expression analysis of the NHX (sodium/hydrogen antiporter) gene family in cotton. *Front. Genet.* 11:964. doi: 10.3389/fgene.2020.00964

Fujisawa, M., Ito, M., and Krulwich, T. A. (2007). Three two-component transporters with channel-like properties have monovalent cation/proton antiport activity. *Proc. Natl. Acad. Sci. U S A* 104, 13289–13294. doi: 10.1073/pnas.0703709104

Gao, W., Long, L., Xu, L., Lindsey, K., Zhang, X., and Zhu, L. (2016). Suppression of the homeobox gene *HDTF1* enhances resistance to *Verticillium dahliae* and *Botrytis cinerea* in cotton. *J. Integr. Plant Biol.* 58, 503–513. doi: 10.1111/jipb.12432

Gierth, M., and Maser, P. (2007). Potassium transporters in plants - Involvement in K^+ acquisition, redistribution and homeostasis. *Febs Lett.* 581, 2348–2356. doi: 10.1016/j.febslet.2007.03.035

Han, M., Wu, W., Wu, W. H., and Wang, Y. (2016). Potassium transporter KUP7 is involved in K^+ acquisition and translocation in *Arabidopsis* root under K^+-limited conditions. *Mol. Plant* 9, 437–446. doi: 10.1016/j.molp.2016.01.012

Harb, A., Krishnan, A., Ambavaram, M. M., and Pereira, A. (2010). Molecular and physiological analysis of drought stress in *Arabidopsis* reveals early responses leading to acclimation in plant growth. *Plant Physiol.* 154, 1254–1271. doi: 10.1104/pp.110.161752

Hauser, F., and Horie, T. (2010). A conserved primary salt tolerance mechanism mediated by HKT transporters: a mechanism for sodium exclusion and maintenance of high K^+/Na^+ ratio in leaves during salinity stress. *Plant Cell Environ.* 33, 552–565. doi: 10.1111/j.1365-3040.2009.02056.x

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

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

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

Jithesh, M. N., Prashanth, S. R., Sivaprakash, K. R., and Parida, A. (2006). Monitoring expression profiles of antioxidant genes to salinity, iron, oxidative, light and hyperosmotic stresses in the highly salt tolerant grey mangrove, *Avicennia marina* (Forsk.) Vierh. by mRNA analysis. *Plant Cell Rep.* 25, 865–876. doi: 10.1007/s00299-006-0127-4

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

Kunz, H. H., Gierth, M., Herdean, A., Satoh-Cruz, M., Kramer, D. M., Spetea, C., et al. (2014). Plastidial transporters KEA1, -2, and -3 are essential for chloroplast osmoregulation, integrity, and pH regulation in *Arabidopsis*. *Proc. Natl. Acad. Sci. U S A* 111, 7480–7485. doi: 10.1073/pnas.1323899111

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

Letunic, I., Doerks, T., and Bork, P. (2015). SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res.* 43, D257–D260. doi: 10.1093/nar/gku949

Li, W., Shang, H., Ge, Q., Zou, C., Cai, J., Wang, D., et al. (2016). Genome-wide identification, phylogeny, and expression analysis of pectin methylesterases reveal their major role in cotton fiber development. *BMC Genomics* 17:1000. doi: 10.1186/s12864-016-3365-z

Liu, J., Guo, W. Q., and Shi, D. C. (2010). Seed germination, seedling survival, and physiological response of sunflowers under saline and alkaline conditions. *Photosynthetica* 48, 278–286. doi: 10.1007/s11099-010-0034-3

Liu, Z., Ge, X., Yang, Z., Zhang, C., Zhao, G., Chen, E., et al. (2017). Genome-wide identification and characterization of SnRK2 gene family in cotton (*Gossypium hirsutum* L.). *BMC Genet.* 18:54. doi: 10.1186/s12863-017-0517-3

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

Ma, L., Zhang, H., Sun, L., Jiao, Y., Zhang, G., Miao, C., et al. (2012). NADPH oxidase AtrbohD and AtrbohF function in ROS-dependent regulation of Na^+/K^+ homeostasis in *Arabidopsis* under salt stress. *J. Exp. Bot.* 63, 305–317. doi: 10.1093/jxb/err280

Mantyla, E., Lang, V., and Palva, E. T. (1995). Role of abscisic-acid in drought-induced freezing tolerance, cold-acclimation, and accumulation of Lt178 and Rab18 proteins in *Arabidopsis-thaliana*. *Plant Physiol.* 107, 141–148. doi: 10.1104/pp.107.1.141

Maser, P., Thomine, S., Schroeder, J. I., Ward, J. M., Hirschi, K., Sze, H., et al. (2001). Phylogenetic relationships within cation transporter families of *Arabidopsis*. *Plant Physiol.* 126, 1646–1667. doi: 10.1104/pp.126.4.1646

Mayer, K. F. X., Rogers, J., Dolezel, J., Pozniak, C., Eversole, K., Feuillet, C., et al. (2014). A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* 345:6194. doi: 10.1126/science.1251788

Miller, S., Ness, L. S., Wood, C. M., Fox, B. C., and Booth, I. R. (2000). Identification of an ancillary protein, YabF, required for activity of the KefC glutathione-gated potassium efflux system in *Escherichia coli*. *J. Bacteriol.* 182, 6536–6540. doi: 10.1128/jb.182.22.6536-6540.2000

Munns, R. (2002). Comparative physiology of salt and water stress. *Plant Cell Environ.* 25, 239–250. doi: 10.1046/j.0016-8025.2001.00808.x

Munns, R., and Tester, M. (2008). Mechanisms of salinity tolerance. *Annu. Rev. Plant Biol.* 59, 651–681. doi: 10.1146/annurev.arplant.59.032607.092911

Palusa, S. G., Golovkin, M., Shin, S. B., Richardson, D. N., and Reddy, A. S. (2007). Organ-specific, developmental, hormonal and stress regulation of expression of putative pectate lyase genes in *Arabidopsis*. *New Phytol.* 174, 537–550. doi: 10.1111/j.1469-8137.2007.02033.x

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

Pichersky, E., and Gershenzon, J. (2002). The formation and function of plant volatiles: perfumes for pollinator attraction and defense. *Curr. Opin. Plant Biol.* 5, 237–243. doi: 10.1016/S1369-5266(02)00251-0

Qiu, Q. S., Barkla, B. J., Vera-Estrella, R., Zhu, J. K., and Schumaker, K. S. (2003). Na^+/H^+ exchange activity in the plasma membrane of *Arabidopsis*. *Plant Physiol.* 132, 1041–1052. doi: 10.1104/pp.102.010421

Reyes, J. L., and Chua, N. H. (2007). ABA induction of miR159 controls transcript levels of two MYB factors during *Arabidopsis* seed germination. *Plant J.* 49, 592–606. doi: 10.1111/j.1365-313X.2006.02980.x

Roosild, T. P., Castronovo, S., Healy, J., Miller, S., Pliotas, C., Rasmussen, T., et al. (2010). Mechanism of ligand-gated potassium efflux in bacterial pathogens. *Proc. Natl. Acad. Sci. USA* 107, 19784–19789. doi: 10.1073/pnas.1012716107

Roosild, T. P., Castronovo, S., Miller, S., Li, C., Rasmussen, T., Bartlett, W., et al. (2009). KTN (RCK) domains regulate K^+ channels and transporters by controlling the dimer-hinge conformation. *Structure* 17, 893–903. doi: 10.1016/j.str.2009.03.018

Roosild, T. P., Miller, S., Booth, I. R., and Choe, S. (2002). A mechanism of regulating transmembrane potassium flux through a ligand-mediated conformational switch. *Cell* 109, 781–791. doi: 10.1016/s0092-8674(02)00768-7

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

Sharif, I., Aleem, S., Farooq, J., Rizwan, M., Younas, A., Sarwar, G., et al. (2019). Salinity stress in cotton: effects, mechanism of tolerance and its management strategies. *Physiol. Mol. Biol. Plants* 25, 807–820. doi: 10.1007/s12298-019-00676-2

Sharma, H., Taneja, M., and Upadhyay, S. K. (2020). Identification, characterization and expression profiling of cation-proton antiporter superfamily in *Triticum aestivum* L. and functional analysis of *TaNHX4*-B. *Genomics* 112, 356–370. doi: 10.1016/j.ygeno.2019.02.015

Sheng, P., Tan, J., Jin, M., Wu, F., Zhou, K., Ma, W., et al. (2014). Albino midrib 1, encoding a putative potassium efflux antiporter, affects chloroplast development and drought tolerance in rice. *Plant Cell Rep.* 33, 1581–1594. doi: 10.1007/s00299-014-1639-y

Spalding, E. P., Hirsch, R. E., Lewis, D. R., Qi, Z., Sussman, M. R., and Lewis, B. D. (1999). Potassium uptake supporting plant growth in the absence of *AKT1* channel activity: Inhibition by ammonium and stimulation by sodium. *J. Gen. Physiol.* 113, 909–918. doi: 10.1085/jgp.113.6.909

Sun, X. L., Ji, W., Ding, X. D., Bai, X., Cai, H., Yang, S. S., et al. (2013). *GsVAMP72*, a novel Glycine soja R-SNARE protein, is involved in regulating plant salt tolerance and ABA sensitivity. *Plant Cell Tiss. Org.* 113, 199–215. doi: 10.1007/s11240-012-0260-4

Sze, H., and Chanroj, S. (2018). Plant endomembrane dynamics: studies of K^+/H^+ antiporters provide insights on the effects of pH and ion homeostasis. *Plant Physiol.* 177, 875–895. doi: 10.1104/pp.18.00142

Tavallali, V., and Karimi, S. (2019). Methyl jasmonate enhances salt tolerance of almond rootstocks by regulating endogenous phytohormones, antioxidant activity and gas-exchange. *J. Plant Physiol.* 234-235, 98–105. doi: 10.1016/j.jplph.2019.02.001

Tsujii, M., Kera, K., Hamamoto, S., Kuromori, T., Shikanai, T., and Uozumi, N. (2019). Evidence for potassium transport activity of *Arabidopsis KEA1-KEA6*. *Sci. Rep.* 9:10040. doi: 10.1038/s41598-019-46463-7

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313, 1596–1604. doi: 10.1126/science.1128691

Ullah, A., Sun, H., Hakim, Yang, X., and Zhang, X. (2018). A novel cotton WRKY gene, *GhWRKY6*-like, improves salt tolerance by activating the ABA signaling pathway and scavenging of reactive oxygen species. *Physiol. Plant* 162, 439–454. doi: 10.1111/ppl.12651

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, 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. *GPB* 8, 77–80. doi: 10.1016/S1672-0229(10)60008-3

Wang, L., Liu, Y., Feng, S., Wang, Z., Zhang, J., Zhang, J., et al. (2018). *AtHKT1* gene regulating K^+ state in whole plant improves salt tolerance in transgenic tobacco plants. *Sci. Rep.* 8:16585. doi: 10.1038/s41598-018-34660-9

Wang, M., Tu, L., Yuan, D., Zhu, 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, X., Guo, H., Wang, J., Lei, T., Liu, T., Wang, Z., et al. (2016). Comparative genomic de-convolution of the cotton genome revealed a decaploid ancestor and widespread chromosomal fractionation. *New Phytol.* 209, 1252–1263. doi: 10.1111/nph.13689

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., Tang, R. J., Yang, X., Zheng, X., Shao, Q., Tang, Q. L., et al. (2019). Golgi-localized cation/proton exchangers regulate ionic homeostasis and skotomorphogenesis in *Arabidopsis*. *Plant Cell Environ.* 42, 673–687. doi: 10.1111/pce.13452

Xu, G., Guo, C., Shan, H., and Kong, H. (2012). Divergence of duplicate genes in exon-intron structure. *Proc. Natl. Acad. Sci. U S A* 109, 1187–1192. doi: 10.1073/pnas.1109047109

Yang, X., Zhang, J., Wu, A., Wei, H., Fu, X., Tian, M., et al. (2021). Corrigendum: Genome-wide identification and expression pattern analysis of the HAK/KUP/KT gene family of cotton in fiber development and under stresses. *Front. Genet.* 12:632854. doi: 10.3389/fgene.2021.632854

Yang, Y., Wu, Y., Ma, L., Yang, Z., Dong, Q., Li, Q., et al. (2019). The Ca^{2+} sensor SCaBP3/CBL7 modulates plasma membrane H^+-ATPase activity and promotes alkali tolerance in *Arabidopsis*. *Plant Cell* 31, 1367–1384. doi: 10.1105/tpc.18.00568

Ye, C. Y., Yang, X., Xia, X., and Yin, W. (2013). Comparative analysis of cation/proton antiporter superfamily in plants. *Gene* 521, 245–251. doi: 10.1016/j.gene.2013.03.104

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

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

Zheng, S., Pan, T., Fan, L., and Qiu, Q. S. (2013). A novel AtKEA gene family, homolog of bacterial K^+/H^+ antiporters, plays potential roles in K^+ homeostasis and osmotic adjustment in *Arabidopsis*. *PLoS One* 8:e81463. doi: 10.1371/journal.pone.0081463

Zhou, H., Qi, K., Liu, X., Yin, H., Wang, P., Chen, J., et al. (2016). Genome-wide identification and comparative analysis of the cation proton antiporters family in pear and four other Rosaceae species. *Mol. Genet. Genomics* 291, 1727–1742. doi: 10.1007/s00438-016-1215-y

Zhu, X., Pan, T., Zhang, X., Fan, L., Quintero, F. J., Zhao, H., et al. (2018). K^+ efflux antiporters 4, 5, and 6 mediate pH and K^+ homeostasis in endomembrane compartments. *Plant Physiol.* 178, 1657–1678. doi: 10.1104/pp.18.01053

Zhu, Y. X., and Li, F. G. (2013). The *Gossypium raimondii* genome, a huge leap forward in cotton genomics. *J. Integr. Plant Biol.* 55, 570–571. doi: 10.1111/jipb.12076

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

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

frontiers in Plant Science

ORIGINAL RESEARCH
published: 15 December 2021
doi: 10.3389/fpls.2021.799904

The Complete Chloroplast Genome Sequences of Eight *Fagopyrum* Species: Insights Into Genome Evolution and Phylogenetic Relationships

Yu Fan[1,2], Ya'nan Jin[2,3], Mengqi Ding[2], Yu Tang[4], Jianping Cheng[1], Kaixuan Zhang[2]* and Meiliang Zhou[2]**

[1] College of Agriculture, Guizhou University, Guiyang, China, [2] Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, China, [3] College of Life Sciences and Food Engineering, Inner Mongolia MINZU University, Tongliao, China, [4] College of Food Science and Technology, Sichuan Tourism University, Chengdu, China

OPEN ACCESS

Edited by:
Wei Hu,
Institute of Tropical Bioscience and Biotechnology, Chinese Academy of Tropical Agricultural Sciences, China

Reviewed by:
Yun-peng Du,
Beijing Academy of Agriculture and Forestry Sciences, China
Alexander Betekhtin,
University of Silesia in Katowice, Poland

***Correspondence:**
Jianping Cheng
chengjianping63@qq.com
Kaixuan Zhang
zhangkaixuan@caas.cn
Meiliang Zhou
zhoumeiliang@caas.cn

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

Received: *22 October 2021*
Accepted: *18 November 2021*
Published: *15 December 2021*

Citation:
Fan Y, Jin Y, Ding M, Tang Y, Cheng J, Zhang K and Zhou M (2021) The Complete Chloroplast Genome Sequences of Eight Fagopyrum Species: Insights Into Genome Evolution and Phylogenetic Relationships. Front. Plant Sci. 12:799904. doi: 10.3389/fpls.2021.799904

Buckwheat (*Fagopyrum* genus, Polygonaceae), is an annual or perennial, herbaceous or semi-shrub dicotyledonous plant. There are mainly three cultivated buckwheat species, common buckwheat (*Fagopyrum esculentum)* is widely cultivated in Asia, Europe, and America, while Tartary buckwheat (*F. tataricum*) and *F. cymosum* (also known as *F. dibotrys*) are mainly cultivated in China. The genus *Fagopyrum* is taxonomically confusing due to the complex phenotypes of different *Fagopyrum* species. In this study, the chloroplast (cp) genomes of three *Fagopyrum* species, *F. longistylum*, *F. leptopodum*, *F. urophyllum*, were sequenced, and five published cp genomes of *Fagopyrum* were retrieved for comparative analyses. We determined the sequence differentiation, repeated sequences of the cp genomes, and the phylogeny of *Fagopyrum* species. The eight cp genomes ranged, gene number, gene order, and GC content were presented. Most of variations of *Fagopyrum* species cp genomes existed in the LSC and SSC regions. Among eight *Fagopyrum* chloroplast genomes, six variable regions (*ndhF-rpl32*, *trnS-trnG*, *trnC*, *trnE-trnT*, *psbD,* and *trnV*) were detected as promising DNA barcodes. In addition, a total of 66 different SSR (simple sequence repeats) types were found in the eight *Fagopyrum* species, ranging from 8 to 16 bp. Interestingly, many SSRs showed significant differences especially in some photosystem genes, which provided valuable information for understanding the differences in light adaptation among different *Fagopyrum* species. Genus *Fagopyrum* has shown a typical branch that is distinguished from the *Rumex*, *Rheum*, and *Reynoutria*, which supports the unique taxonomic status in *Fagopyrum* among the Polygonaceae. In addition, phylogenetic analysis based on the cp genomes strongly supported the division of eight *Fagopyrum* species into two independent evolutionary directions, suggesting that the separation of cymosum group and urophyllum group may be earlier than the flower type differentiation in *Fagopyrum* plants. The results of the chloroplast-based phylogenetic tree were further supported by the *matK* and Internal Transcribed Spacer

(ITS) sequences of 17 *Fagopyrum* species, which may help to further anchor the taxonomic status of other members in the urophyllum group in *Fagopyrum*. This study provides valuable information and high-quality cp genomes for identifying species and evolutionary analysis for future *Fagopyrum* research.

Keywords: *Fagopyrum*, Polygonaceae, chloroplast genome, comparative analysis, phylogenetic relationship

INTRODUCTION

As the organelle specialized for carrying out photosynthesis in plants, the chloroplast is descended from cyanobacteria, and occurs in eukaryotic autotrophs such as land plants and algae (Jin and Daniell, 2015; Gao et al., 2019). Chloroplasts are involved in photosynthesis and important biochemical processes including storage of starch, and the biosynthesis of sugars, several amino acids, lipids, vitamins, and pigments within plant cells, as well as sulfate reduction and nitrogen cycle supplying for the driving force of plants growth and development (Neuhaus and Emes, 2000; Jarvis and Soll, 2001; Leister, 2003; Bausher et al., 2006). As the center of photosynthesis, chloroplast has a complete genetic system, in which the genetic material is the cp genome (Zhao et al., 2019). Like nuclear DNA, chloroplasts have the same functions of replication, transcription, and inheritance, and cp genomes in plants are generally 10–20% of total genomes with an average length of about 120–170 kb (kilo-base pair) in tetrad ring structure (Shinozaki et al., 1986; Ruhlman and Jansen, 2014). The average cp genome size of land plants is 151 kb, with most species ranging from 130–170 kb in length, as well as the average GC content is 36.3%. The circle cp genome was separated by two inverted repeats (IRs, 20–28 kb) generating the large single copy (LSC, 16–27 kb) and the small single copy (SSC) (Jansen et al., 2007), which can provide abundant information for solving plant phylogenetic relationships and trends. Gene contents and sequences of cp genomes of angiosperm are generally conserved including 4 rRNAs, 30 tRNAs, and 80 unique proteins (Chumley et al., 2006). With the characteristics of parthenogenetic inheritance (maternal inheritance), relatively small genome and slow genome mutation rate (Palmer et al., 1988), analysis of the phylogenetic relationships of multiple chloroplast DNA can help to understand plant phylogeny, population genetic analysis, and taxonomic status at the molecular level (Alwadani et al., 2019). Although cp genomes of angiosperms are generally conserved in gene numbers and sequences (Jansen and Ruhlman, 2012), levels of structural variation in the genome different from various families and genera existed, such as gene duplication and large-scale rearrangement of genes, introns, and IR domains (Cosner et al., 2004; Lee et al., 2007; Cai et al., 2008; Guisinger et al., 2010; Martin et al., 2014).

The size of the cp genome was correlated with plant habits, environments, and other functional traits (Beaulieu et al., 2008; Li et al., 2018), making it a promising tool in studies of phylogeny, evolution, and population genetics of angiosperms (Tonti-Filippini et al., 2017). For example, the phylogenetic relationships among the main branches of flowering angiosperms were analyzed by using the coding genes from 64 cp genomes in *Amborella Baill* (Jansen et al., 2007); moreover, the relationship between genome evolution and phylogeny of *Zingiberaceae* was identified using the complete genome sequences of 14 chloroplasts of *Curcuma* Species (Liang et al., 2020).

Fagopyrum genus belongs to the Polygonaceae family, which are annual or perennial herb or semi-shrub plants (Zhang et al., 2021a). Wild buckwheats are mainly distributed in the regions of southwest China, which was recognized as the center of buckwheat origin and diversity (Ohnishi, 1995, 1998; Ohsako et al., 2002; Saski et al., 2005; Tang et al., 2010; Shao et al., 2011; Zhou et al., 2018). In 1742, *Fagopyrum* was established by Tourn, and named *Fagopyrum* Tourn ex Hall (Linnaeus, 1753). In 1992,the taxonomic status of buckwheat was confirmed, and the embryo position, morphology of cotyledon and perianth segments, characteristics of the pollen grain, and the basic number of chromosomes were taken as the basis for distinguishing *Fagopyrum* from *Polygonum* (Ye and Guo, 1992). With the continuous introduction of various buckwheat species, the classification based on morphological features gradually complicated, and plants from *Fagopyrum* were classified into 22–28 different species comprising two variants and two subspecies until 2021 (Zhang et al., 2021a). Due to the long-term change of buckwheat classification status, a consistent view of buckwheat was scarce, which limited the utilization of wild buckwheat varieties in plant improvement (Sharma and Jana, 2002; Neethirajan et al., 2011; Nagatomo et al., 2014). The controversies on buckwheat classification were including but were not limited to the following: (1) the genetic relationships among *F. tataricum*, *F. esculentum, F. esculentum* subsp. *ancestrale*, and *F. cymosum*. (2) The evolutionary paths between the cymosum group and urophyllum group are intersected or separated? (3) How to define the taxonomic status and phylogenetic relationship among *Fagopyrum* species in urophyllum group?

The rapid development of molecular biology and genomics provides favorable conditions for the study of cp genome of buckwheat plants, as well as the important genetic information for taxonomic status, phylogeny, and species identification. At present, five buckwheat cp genomes had been published, including *F. tataricum*, *F. esculentum*, *F. esculentum* subsp. *ancestrale*, *F. cymosum,* and *F. luojishanense* (Liu et al., 2008; Logacheva et al., 2008; Cho et al., 2015; Hou et al., 2015; Wang et al., 2017a; Zhang and Chen, 2018). However, the in-depth and conjoint study of *Fagopyrum* cp genome data sets was lacking, as well as the researches on buckwheat phylogeny and interspecific differences.

In this study, three cp genomes of *Fagopyrum* were sequenced, assembled, and annotated, then their cp genome

data with five published ones were analyzed comprehensively, including characteristics of *Fagopyrum* cp genomes, codon usage, expansion of IR regions, SSRs analysis, and phylogenetic analysis of eight *Fagopyrum* species. Our objectives in this study were: (1) To present the complete sequence of cp genomes of three newly assembled buckwheat plants and to compare the global structure with five other previously published species (including one subspecies) within genus species comparisons; (2) SSR variations in the cp genome sequences of eight buckwheat plants were detected to develop a series of SSRs molecular markers that could be used to distinguish the relationship between different species; (3) The phylogenetic relationship and evolutionary path of buckwheat were reconstructed by combining genetic sequences based on eight cp genomes and six highly variable regions developed. (4) The taxonomic status of 17 buckwheat plants was discussed by using ITS and *matK* gene sequences.

MATERIALS AND METHODS

Plant Material, Morphological Analysis, and DNA Extraction

In previous reports, we investigated in detail the survival status of *Fagopyrum* plants in southwest China (Cheng et al., 2020; Zhang et al., 2021a). The mature seeds of these plant materials are collected in the wild, then they are grown in the greenhouse of the institute of crop science, Chinese Academy of Agricultural Sciences (CAAS) in Beijing. The morphological details of eight *Fagopyrum* species were further observed. We mainly investigated the differences in plant type, leaf, inflorescence, seed and distribution (Cheng et al., 2020).

Further, the fresh leaves from three *Fagopyrum* species, *F. longistylum*, *F. leptopodum*, *F. urophyllum* were collected in Sichuan Province in 2020 (**Supplementary Table 1**). Voucher specimens of these samples were deposited in the Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, China. Total genomic DNA was isolated from 2 g of silica-dried leaf sample using the modified CTAB method (Doyle, 1987). In addition, we downloaded the available complete cp genomes of five other *Fagopyrum* species and three Polygonaceae species from GenBank [*F. tataricum*, MT712164.1; *F. cymosum* (*F. dibotrys*), KY275181.1; *F. esculentum*, MT364821.1; *F. esculentum* subsp. *ancestrale*, EU254477.1; *F. luojishanense*, KY275182.1; *Rumex hypogaeus*, MT017652.1; *Reynoutria japonica* (also known as *Polygonum cuspidatum*) MW411186.1; *Rheum officinale* MN564925.1] for phylogeny study.

Genome Sequencing, Assembly, Annotation

The total DNA was disrupted by ultrasonic wave, and DNA libraries were read of 350 bp with purified DNA constructed by Library Prep Kit from NEBNext®. Total DNA was sequenced in Hiseq 4000 PE150. After filtering the low-quality data, raw sequencing data were checked and spliced using SPAdes 3.6.1 (Bankevich et al., 2012). Contigs were used to screen the cp genome by Blast Software, using published *F. esculentum* cp genome (MT364821) as the reference genome (Altschul et al., 1997). Selected contigs of the cp genome were assembled using Sequencher 4.10 Software (GeneCodes Corp., Ann Arbor, MI, United States), and all reads were mapped to validate cp genome using Geneious 8.1 Software (Kearse et al., 2012). Polymerase Chain Reaction (PCR) was done with specific primers of gaps, which were born after assembling genomes. The PCR products were sequenced by ABI 3730, and were involved in manually correcting annotations. The circular structure map was constructed by Organellar Genome DRAW[1] (Lohse et al., 2013).

Codon Usage Analysis

Codon Usage analysis was done by codonW 1.4.4 (Peden, 2000), and the values of relative synonymous codon usage (RSCU) were used to evaluate codon preference.

Comparative Genomic Analysis

The divergence of 11 Polygonaceae genomes was counted by mVISTA in LAGAN mode (Frazer et al., 2004), and *Rumex hypogaeus* (MT017652), *Polygonum cuspidatum* (MW411186), and *Rheum officinale* (MN564925) were considered as the reference genomes. MAFFT was used to align all *Fagopyrum* species genome (Zhang et al., 2018), and the nucleotide diversity (Pi) of all complete cp genome was calculated using Launch DnaSP6 (Rozas et al., 2017), and the results were presented through a sliding window analysis with a window length of 600 bp and step size of 200 bp. Boundaries of inverted repeat (IR) regions, contraction, and expansion of eight cp genomes were determined using IRscope (Amiryousefi et al., 2018).

Simple Sequence Repeats Analysis

To identify the microsatellites, the Perl script MISA70 and the SSRs parameter were used to analyze the SSRs detection based on the following conditions (Beier et al., 2017); thresholds were set as eight repeat units for mononucleotide SSRs, four repeat units for dinucleotide SSRs, four repeat units for trinucleotide SSRs, and three repeat units for tetranucleotide, pentanucleotide, and hexanucleotide SSRs.

Phylogenetic Analysis

We used the 11 above-mentioned cp genomes to analyze the phylogenetic relationships among *Fagopyrum* species, including eight *Fagopyrum* species, and three Polygonaceae species (*Rumex hypogaeus*, *Rheum officinale*, and *Reynoutria japonica*) were used as outgroups. These cp sequences were aligned with the default parameters set using MAFFT program (Katoh and Standley, 2013) in GENEIOUS R8, and were manually adjusted in MEGA 6.0. The nucleotide sequence (*matK* and ITS) data were obtained from NCBI (**Supplementary Table 9**). The RAxML v7.2.8 program (Stamatakis, 2006) was used to perform the phylogenetic trees based on maximum likelihood analysis with 1000 bootstrap replicates. Bayesian inference was performed using the MrBayes v3.1.27 program (Ronquist and Huelsenbeck, 2003). Markov chain Monte Carlo simulations have two parallel runs with 2000,000 generations independently,

[1]https://chlorobox.mpimp-golm.mpg.de/OGDraw.html

and sampling trees every 100 generations. The initial 25% of trees were discarded as burn-in, and the remaining dates were used to construct a majority-rule consensus tree. Convergence diagnostics were monitored by examining the average standard deviation of split frequencies below 0.01.

RESULTS AND ANALYSIS

Morphological Analysis in Eight *Fagopyrum* Species

The morphological characters of eight *Fagopyrum* species are further analyzed in this section. Buckwheat is a rare cereal crop that does not belong to Gramineae. *Fagopyrum* contains plants of both self-compatible (homostyly) and self-incompatible (heterostyly) species. Therefore, *Fagopyrum* species are good materials for studying the origin and spread of cultivated crops, as well as hot issues such as phylogenetic evolution of plants (Zhou et al., 2018). Morphological characteristics of eight typical different *Fagopyrum* species (including seven species and one subspecies) were systematically analyzed, and their differences were mainly concentrated in stems, leaves, flowers, and fruits (**Figure 1** and **Supplementary Table 1**). In general, the morphology of *Fagopyrum* plants is relatively complex and their habits and features are various. In this study, three *Fagopyrum* species which cp genomes were not revealed were fully considered based on plant characteristics. *F. leptopodum,* which was commonly found in rocks and dry-hot valley areas, was considered to be a highly drought-resistant and barren resistant species. *F. longistylum,* a self-compatible but heteromorphic species, was a very rare phenomenon in plants. In addition, *F. urophyllum,* contained semi-woody branches and perennial rhizomes, which are considered as transitional species from herbaceous to woody plants (Ohnishi and Matsuoka, 1996; Zhang et al., 2021b).

Characteristics of *Fagopyrum* Chloroplast Genomes

The cp genomes of three wild *Fagopyrum* species were sequenced in this study, including two annual species (*F. longistylum* and *F. leptopodum*) and one perennial species (*F. urophyllum*). We obtained the complete cp genome sequences of 159,325 bp for *F. longistylum*, 159,350 bp for *F. urophyllum*, and 159,376 bp for *F. leptopodum*. Other published cp genomes of *Fagopyrum* were obtained from National Center for Biotechnology Information (NCBI), and all cp genomes ranged in size from 159,265 bp (*F. luojishanense*) to 159,599 bp (*F. esculentum* ssp. *ancestrale*) with 37.78–37.99% GC contents (**Figure 2** and **Table 1**). Similar to other Polygonaceae, all cp genomes of cultivated and wild *Fagopyrum* species comprised a typical circular structure with four regions (Wu et al., 2020), and two inverted repeats (IR, IRa, and IRb) regions were separated by a LSC and a SSC (**Figure 2**). The LSC region in *Fagopyrum* accounted for 52.87–53.19% of the total cp genomes and ranged in size from 84,250 bp (*F. urophyllum*) to 84,885 bp (*F. esculentum* ssp. *ancestrale*); the SSC region in *Fagopyrum* accounted for 8.22–8.41% and ranged in size from 13,094 bp (*F. luojishanense*) to 13,406 bp (*F. urophyllum*); the *Fagopyrum* IR region accounted for 19.23–19.38% of the total size and ranged from 30,6845 bp (*F. esculentum* and *F. esculentum* ssp. *ancestrale*) to 30,870 bp (*F. luojishanense*). Moreover, the GC contents of all *Fagopyrum* cp genomes were similar, and the GC content of IR region was highest (41.26–41.48%), followed by the LSC region (36.01–36.32%) and the SSC region (31.97–32.99%).

There was little difference in coding regions in eight *Fagopyrum* species. Overall, they encode a total of 108–113 chloroplast genes, including 76–79 protein-coding genes, 28–30 tRNAs, and 4 rRNAs (**Figure 2** and **Table 2**). All the above-mentioned genes were furtherly categorized as three parts, of which 47 genes belong to photosynthesis related genes (including rubisco, photosystem I, assembly/stability of photosystem I, photosystem II, ATP synthase, cytochrome b/f complex, cytochrome c synthesis, and NADPH dehydrogenase), 60 genes belong to transcription and translation related genes (including transcription, ribosomal proteins, and translation initiation factor, ribosomal RNA, and transfer RNA), and the remaining genes belong to biomacromolecule metabolism related genes or other unknown functions (**Table 2**). Moreover, among these various 113 genes, 15 genes contained one intron comprising 9 protein-coding genes (*atpF, petB, petD, ndhA, ndhB, rpoC1, rps12, rpl2,* and *rpl16*) and 6 tRNA genes (*trnA, trnG, trnI, trnK, trnL,* and *trnV*), while 2 genes (*ycf3*, *clpP*) contained two introns. In addition, *rps12* was identified as a noticeable trans-splicing gene of all *Fagopyrum* species, because the 5′ end of *rps12* exon was located in the LSC region but the other end of that was located in the IR domain.

Codon Usage

Codon is the connection between the nucleic acids and proteins, and codon usage reflects the preference for selective use of codons encoding specific amino acids with genetic information (Wanga et al., 2021). The codon usage frequency of 79 protein-coding genes for 8 *Fagopyrum* species were calculated, and 64 codons were involved in encoding proteins containing three termination codons, such as UAA, UAG, and UGA (**Table 3**). The relative synonymous codon usage (RSCU) analysis showed that 30 codons of 8 *Fagopyrum* species were > 1, and the UUA encoding leucine had the highest RSCU with 1.85–1.87 in 8 *Fagopyrum* species. While the lowest RSCU was 0.33–0.36 with the CGC encoding arginine.

Comparative Genomic Analysis

The genome of *F. tataricum* was served as the reference to conduct the mVISTA program for discovering *Fagopyrum* genome divergence, and three other genomes from Polygonaceae were regarded as the outgroups covering *Rumex hypogaeus*, *Polygonum cuspidatum*, and *Rheum officinale*. Results revealed that 11 cp genomes were relatively conserved (**Figure 3**). The three cultivated *Fagopyrum* species, four wild *Fagopyrum* species, and three outgroup members had higher similarity and low divergence, respectively. Furthermore, the divergence of LSC and SSC regions were higher than that of IR regions,

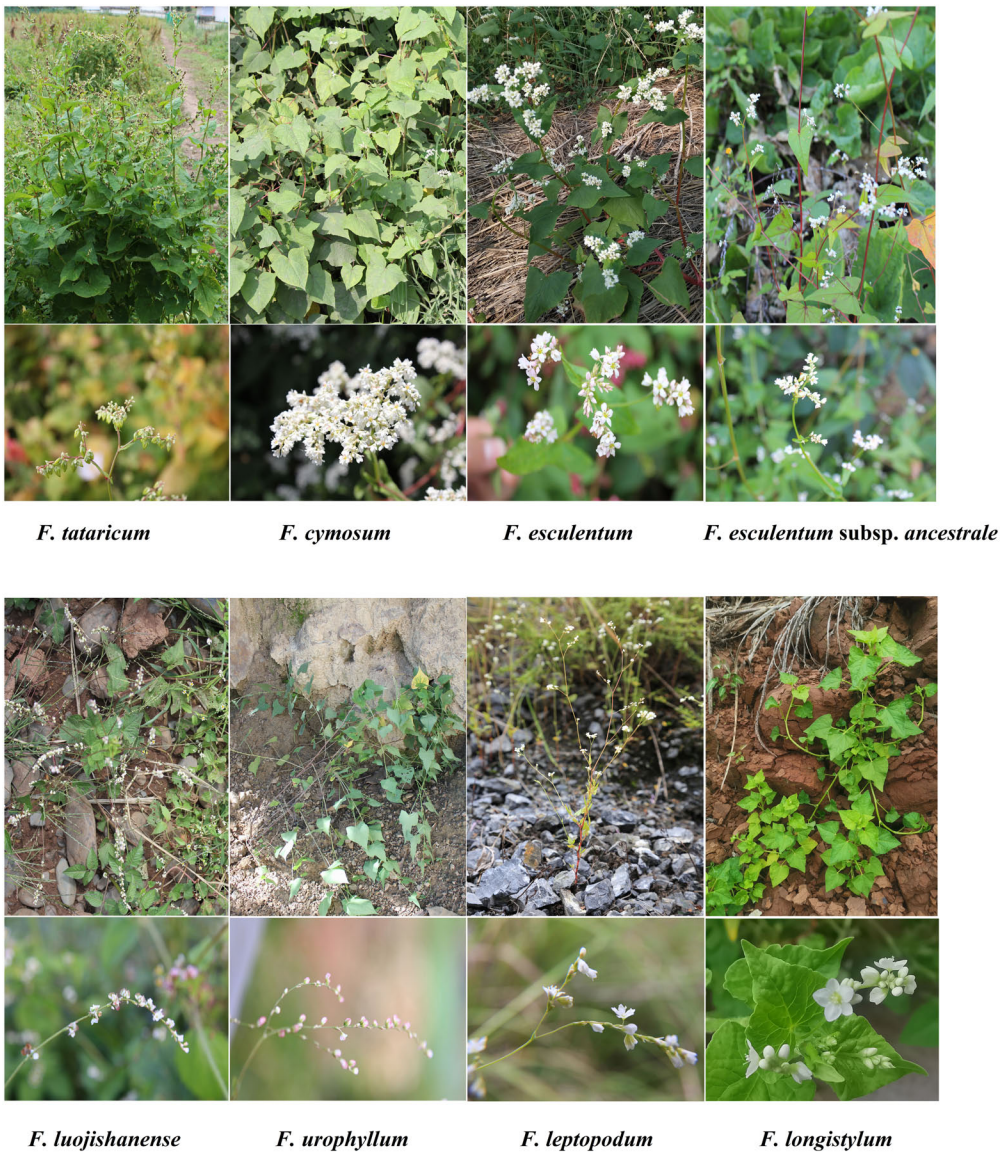

FIGURE 1 | The morphological characters of plants and flowers of eight *Fagopyrum* species. **(A)** *F. tataricum*; **(B)** *F. cymosum*; **(C)** *F. esculentum*; **(D)** *F. esculentum* subsp. *ancestrale*; **(E)** *F. luojishanense*; **(F)** *F. urophyllum*; **(G)** *F. leptopodum*; **(H)** *F. longistylum*.

and the non-coding regions exhibited greater variation than the coding regions.

To further know the genetic diversity of various *Fagopyrum* species and exploit suitable polymorphic genes for identifying novel species, we calculate the nucleotide diversity (Pi) of eight *Fagopyrum* species. The Pi values were ranged from 0 to 0.10179 in the total cp genomes. The average Pi values of LSC and SSC regions were 0.0356 and 0.0445, respectively, but that of IR regions was 0.0084 (**Supplementary Table 2**). Most of the variations of *Fagopyrum* species cp genomes existed in the LSC and SSC regions. That is to say, two IR regions were more conserved than another two regions. A sliding window analysis showed that the Pi values of six regions were > 0.08, and these most divergent regions included *ndhF-rpl32*, *trnS-trnG*, *trnC*, *trnE-trnT*, *psbD,* and *trnV* (**Figure 4** and **Supplementary Table 2**). Among them, three coding genes (*ndhF*, *rpl31,* and *psbD*) were highlighted, because coding genes were generally conserved. These polymorphic regions might be the critical loci for population genetic studies of *Fagopyrum* species.

Contraction and Expansion of Inverted Repeats Regions Among Eight *Fagopyrum* Species

As we all know, contraction and expansion of the IR regions are strongly linked to the length of cp genomes (Liang et al., 2020), therefore the IR boundaries were detected to explain the

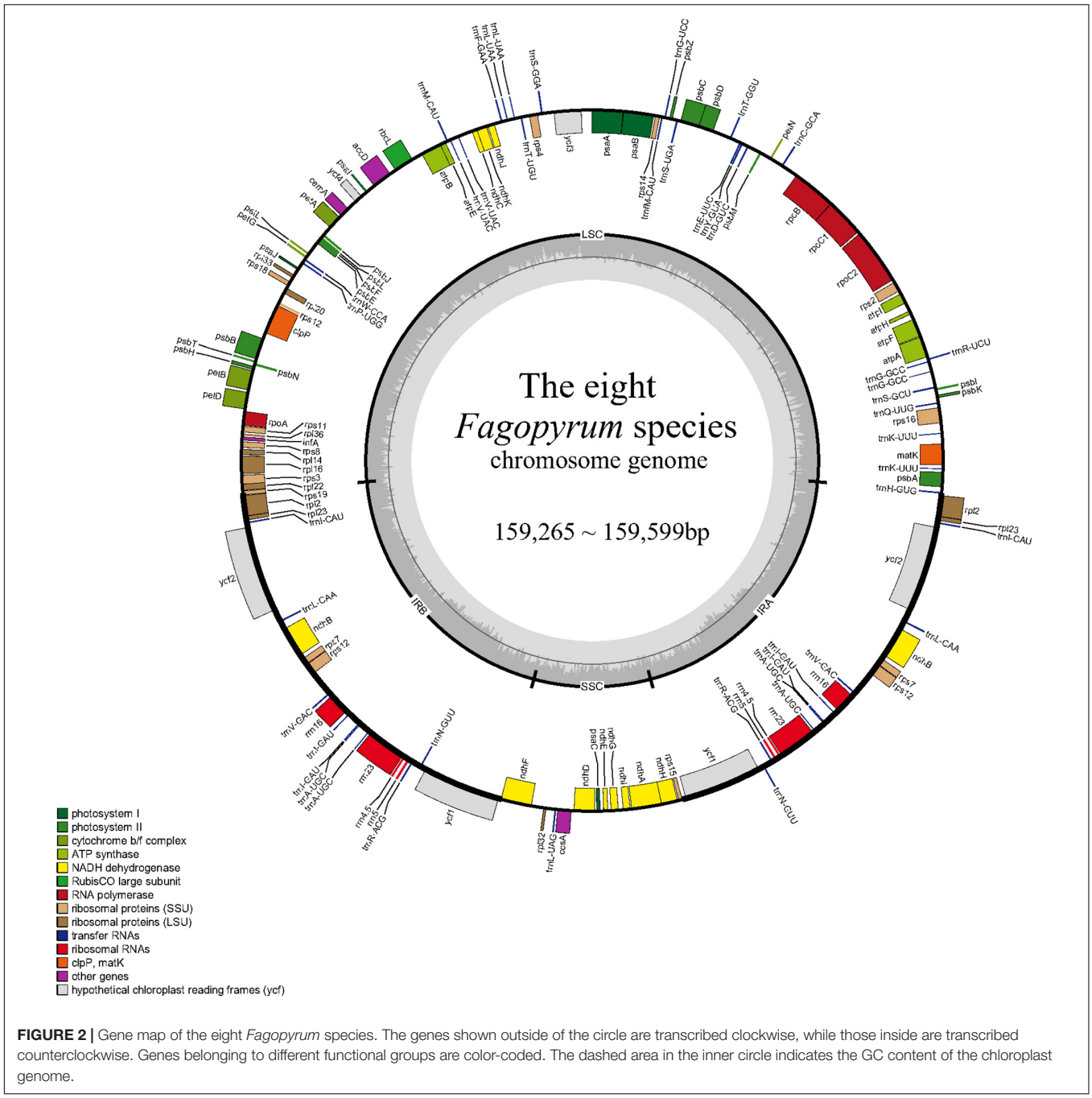

FIGURE 2 | Gene map of the eight *Fagopyrum* species. The genes shown outside of the circle are transcribed clockwise, while those inside are transcribed counterclockwise. Genes belonging to different functional groups are color-coded. The dashed area in the inner circle indicates the GC content of the chloroplast genome.

differences in *Fagopyrum* cp genome size. In general, IRs of wild *Fagopyrum* species (*F. longistylum*, *F. leptopodum*, *F. urophyllum*, and *F. luojishanense*) were longer than cultivated *Fagopyrum* species (*F. tataricum*, *F. cymosum*, and *F. esculentum*). Among them, the size of the IR regions of the two *F. esculentum* was the shortest (30,685 bp) and that of *F. luojishanense* was the longest (30,870 bp) (**Figure 5**).

Within the 8 *Fagopyrum* species, the *rps19* genes were located in the boundaries of LSC/IRb regions (JLB) consistently, except for the location of *rps19* from *F. esculentum* ssp. *ancestrale* in JLB was more forward than other members (1 bp). The SSC and IRb regions (JSB) were connected by *ndhF* genes, and the length of the *ndhF* in IRb from the JLB was 54–90 bp. In the JSA (SSC/IRa) regions, only JSA of three species were embedded in *rps15* gene, including the two *F. esculentum* and *F. luojishanense*. Specifically, the *rps15* gene was located on the right of the two *F. esculentum* with the distance of 2 bp, but that of *F. luojishanense* was 23 bp. The LSC/IRa (JLA) junctions in the cp genomes of 8 *Fagopyrum* species were identical. All in all, the IR boundaries of *F. tataricum* and *F. cymosum* were similar, as well as two *F. esculentum* species, and three wild species (*F. longistylum*, *F. leptopodum*, and *F. urophyllum*), respectively.

TABLE 1 | Comparison of the complete chloroplast genomes for eight *Fagopyrum* species.

		F. tataricum	*F. cymosum*	*F. esculentum*	*F. esculentum* ssp. *ancestrale*	*F. longistylum*	*F. leptopodum*	*F. urophyllum*	*F. luojishanense*
	Accession number	MT712164	KY275181	MT364821	EU254477	OK054489	OK054491	OK054490	KY275182
Total	Total length (bp)	159,272	159,320	159,576	159,599	159,325	159,376	159,350	159,265
	GC (%)	37.87	37.93	37.95	37.99	37.82	37.79	37.78	37.84
LSC	Length (bp)	84,397	84,422	84,875	84,885	84,417	84,282	84,250	84,431
	GC (%)	36.20	36.26	36.29	36.32	36.03	36.02	36.01	36.05
	Length (%)	52.99	52.99	53.19	53.19	52.98	52.88	52.87	53.01
SSC	Length (bp)	13,241	13,264	13,331	13,344	13,226	13,402	13,406	13,094
	GC (%)	32.78	32.90	32.99	32.96	32.17	32.10	31.97	32.36
	Length (%)	8.31	8.33	8.35	8.36	8.30	8.41	8.41	8.22
IRa/IRb	Length (bp)	30,817	30,817	30,685	30,685	30,841	30,846	30,847	30,870
	GC (%)	41.26	41.29	41.38	41.37	41.48	41.45	41.46	41.45
	Length (%)	19.35	19.34	19.23	19.23	19.36	19.35	19.36	19.38

Simple Sequence Repeats Analysis

Simple sequence repeats, also known as microsatellites, consisted of short tandem repeats of 1–6 bp in length (Li B. et al., 2020). SSRs are widely distributed in the cp genome, and play a key role in the identification of plant genetic relationships and taxonomic status (Yang et al., 2019; Li Y. et al., 2020). In the cp genome sequence of the eight *Fagopyrum* species, SSRs were mainly located in the intergene region (∼57.72%), followed by the genic region (∼42.28%), while no SSR was observed in tRNAs and rRNAs (**Figure 6A** and **Supplementary Table 3**), which is consistent with the report of Wang et al. (2017b). Of note, the SSR numbers of *F. leptopodum* (133, ∼59.38%), *F. longistylum* (138, ∼60.26%), *F. luojishanense* (131, ∼58.48%), and *F. urophyllum* (143, ∼60.59%) in the intergene region were significantly higher than that of *F. tataricum* (110, ∼53.66%), *F. cymosum* (115, ∼56.65%), *F. esculentum* (119, ∼55.61%) and *F. esculentum* subsp. *ancestrale* (120, ∼56.34%). Most SSRs were located in LSC region (∼64.63%), followed by IR region (∼26.38%) and SSC region (∼8.99%) (**Figure 6B** and **Supplementary Tables 4, 5**). *F. cymosum* (129, ∼63.55%) had the least number of SSR in LSC region, followed by *F. tataricum* (130, ∼63.41%), *F. esculentum* (139, ∼64.95%) and *F. esculentum* subsp. *ancestrale* (138, ∼64.79%), in general, their number was significantly lower than *F. leptopodum* (146, ∼65.18%), *F. luojishanense* (145, ∼64.73%), *F. longistylum* (148, ∼64.35%), and *F. urophyllum* (156, ∼66.10%). Interestingly, as two typical cultivars, *F. tataricum* (58, ∼28.29%) and *F. esculentum* (56, ∼26.17%) showed significant expansion in SSR proportion in IR region. Further, a total of 24 gene located in different regions were found, which may be the result of co-evolution of cp genomes (Zhao et al., 2021). Among them, *ndhB*, *ycf2*, and *ycf1* are in the IRb/IRa region, *atpA*, *rbcL*, *rpl20*, *rpl22*, *rpoA*, *ycf4*, *cemA*, *petB*, *ycf3*, *petA*, *rpoB*, *atpF*, *rpoC1*, *rpl16*, and *rpoC2* are located in LSC region, and *rps15*, *ndhF*, *ndhD* are located in SSC region.

The distribution range of SSRs ranged from 8 to 16 bp in eight *Fagopyrum* species, with a total of 66 different types(**Figure 6C** and **Supplementary Tables 4, 5**). There were no hexanucleotide repeats have been found in these SSR sequences, and pentanucleotide repeats were only found in the cp genomes of *F. urophyllum* (ATTAT), *F. tataricum* (TTTTA), and *F. cymosum* (TCTAT/TTTTA). Among all *Fagopyrum* species, the number of mononucleotide repeats in the cymosum group was significantly lower than that in the urophyllum group. In general, this study supports that mononucleotide repeats may play a more important role in genetic variation in buckwheat than other SSR types (Huang et al., 2017; Liang et al., 2020). Although the chloroplast evolution of *Fagopyrum* species were relatively conserved, the cymosum group may be subjected to stronger selection and evolutionary pressure, resulting in the decline of SSR genetic diversity. Meanwhile, the number and types of SSR of the eight buckwheat plants in this study were further analyzed (**Figure 6D** and **Supplementary Tables 5, 6**). Further, the proportion of mononucleotide repeats for A/T and C/G types were 71.52 and 1.86%, respectively (**Figure 6D** and **Supplementary Tables 5–7**). This is similar to Zingiberales, Salicaceae, and Ranunculaceae, etc., indicating that mononucleotide repeats of A/T type may always be the most abundant base of simple repeat sequences (Huang et al., 2017; Liang et al., 2020; Park and Park, 2021). In addition, the number of mononucleotide repeats of A/T types or C/G types in the cymosum group was significantly lower than the urophyllum group, indicating that the number of SSR may still be similar in different subgroups of *Fagopyrum* species. The dinucleotides of eight *Fagopyrum* species were divided into four categories, which showed differences in some gene regions and repeated fragments among different groups. For example, repeat sequences of AG/CT and GA/TC types do not differ significantly between the cymosum group and urophyllum group. However, the proportion of CA/TG repeats in the cymosum group (∼0.96%) was much higher than that in the urophyllum group (∼0.44%). Similarly, AT/TA type accounted for the highest proportion of all dinucleotides (∼14.16%), which further confirmed the activity of A/T base in the cp genome. In this study, *F. tataricum* (27, ∼13.17%)/*F. cymosum* (27, ∼13.30%), *F. esculentum* (32, ∼14.95%)/*F. esculentum* subsp. *ancestrale* (32,

TABLE 2 | Genes contained in the chloroplast genome of eight *Fagopyrum* species.

Category for genes	Groups of genes	Name of genes
Photosynthesis related genes	Rubisco	*rbcL*
	Photosystem I	*psaA, psaB, psaC, psaI, psaJ*
	Assembly/stability of photosystem I	*ycf3*[a], *ycf4*
	Photosystem II	*psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ*
	ATP synthase	*atpA, atpB, atpE, atpF*[a], *atpH, atpI*
	Cytochrome b/f complex	*petA, petB*[a], *petD*[a], *petG, petL, petN*
	Cytochrome c synthesis	*ccsA*
	NADPH dehydrogenase	*ndhA*[a], *ndhB*[a,b], *ndhC, ndhD, ndhE, ndhF*[b], *ndhK ndhG, ndhH, ndhI, ndhJ*
Transcription and translation related genes	Transcription	*rpoA, rpoB, rpoC1*[a], *rpoC2*
	Ribosomal proteins	*rps2, rps3, rps4, rps7*[b], *rps8, rps11, rps12*[a,b], *rps14,rps15, rps16, rps18, rps19*[b], *rpl2*[a,b], *rpl14, rpl16*[a], *rpl20, rpl22, rpl23, rpl32*[b], *rpl33, rpl36*
	Translation initiation factor	*infA*
RNA genes	Ribosomal RNA	*rrn5*[b], *rrn4.5*[b], *rrn16*[b], *rrn23*[b]
	Transfer RNA	*trnA-UGC*[a,b], *trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnG-UCC, trnG-GCC*[a], *trnH-GUG, trnI-CAU*[b], *trnI-GAU*[a,b], *trnK-UUU*[a], *trnL-CAA*[b], *trnL-UAA*[a], *trnL-UAG, trnfM-CAUI, 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,b], *trnW-CCA, trnY-GUA*
Other genes	RNA processing	*matK*
	Carbon metabolism	*cemA*
	Fatty acid synthesis	*accD*
	Proteolysis	*clpP*[a]
Genes of unknown function	Conserved reading frames	*ycf1*[b], *ycf2*[b]
Pseudogenes		*ycf15*

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

∼15.02%) had similar AT/TA types in number and proportion, which supported their genetic relationship to a certain extent. In addition, nucleotide repeats of AAT/TTA type did not exist in the four species of cymosum group (0), while *F. longistylum* (∼0.87%), *F. leptopodum* (∼0.89%), *F. luojishanense* (∼0.89%), and *F. urophyllum* (3, ∼1.27%) had a similar proportion. Therefore, there may exist two divergent evolutionary directions between the cymosum group and the urophyllum group. These results suggest that SSR can be used to identify genetic diversity, study evolution and develop molecular markers in buckwheat.

Phylogenetic Analysis of Eight *Fagopyrum* Species Based on cp Genome

Chloroplast genome sequences of eight *Fagopyrum* species and three Polygonaceae plants, which were selected as the outgroup, were used to construct phylogenetic trees to elucidate their genetic relationships (**Figure 7**). The numbers on the branches show the bootstrap value of the maximum likelihood analysis. The results showed that all *Fagopyrum* species clustered together at a very high resolution, and the three Polygonaceae plants and the eight *Fagopyrum* species were divided into two main types, which confirmed the independent differentiation status of the *Fagopyrum* from other genera of Polygonaceae. Further, eight *Fagopyrum* species were classified into two typical subclades. Among them, *F. tataricum* and *F. cymosum* formed a subgroup different from *F. esculentum*, which further supports that they may have a relatively high degree of homology and a closer genetic relationship. And then, they gradually converged with *F. esculentum* and *F. esculentum* subsp. *ancestrale* to form a subbranch. In addition, *F. longistylum* first approximates to *F. luojishanense*, and then gradually forms with *F. urophyllum* and *F. leptopodum*. These results showed that there might be two different subgroups among the eight *Fagopyrum* species, and the cymosum group and the urophyllum group evolved independently. Further, we developed six molecular marker sequences based on Pi values (**Supplementary Figures 1A–F** and **Supplementary Table 8**). And, six cluster trees were constructed based on these sequences using the neighbor-joining method (NJ). Among them, *trnS-trnG* and *trnV* trees supported the topological structure of the cp genome, which can be further applied in the identification of genetic relationships in *Fagopyrum* species.

Phylogenetic Relationship Based on the ITS and *matK*

The most widely used chloroplast gene *matK* and nuclear marker ITS were selected to further speculate the genetic relationship of eighteen *Fagopyrum* species (including one variety: *F. gracilipes* var. *odontopterum*) (**Supplementary Figures 2A,B** and **Supplementary Table 9**). In general, the two ML trees based on ITS and *matK* supported the above-mentioned cp genome tree results: *F. tataricum* and *F. cymosum* in the two phylogenetic trees are first clustered into one branch, then clustered with *F. esculentum*, and then gradually clustered into other wild species. Therefore, phylogenetic trees based on different markers in this study all supported the conclusion that *F. tataricum* and *F. cymosum* in the cymosum group has a more close relationship than *F. esculentum*, which consisted with the

TABLE 3 | Codon content of amino acids and stop codon of eight *Fagopyrum* species.

Amino acid	Codon	RSCUa																	
		F. tataricum	*F. cymosum*	*F. esculentum*	*F. esculentum ssp. ancestrale*	*F. longistylum*	*F. leptopodum*	*F. urophyllum*	*F. luojishanense*			*F. tataricum*	*F. cymosum*	*F. esculentum*	*F. esculentum ssp. ancestrale*	*F. longistylum*	*F. leptopodum*	*F. urophyllum*	*F. luojishanense*
Ala	GCA	1.12	1.13	1.14	1.11	1.16	1.16	1.16	1.15	Ile	AUA	0.97	0.97	0.97	0.96	0.97	0.97	0.98	0.98
	GCC	0.76	0.76	0.77	0.78	0.71	0.7	0.71	0.71		AUC	0.57	0.57	0.56	0.56	0.55	0.55	0.55	0.55
	GCG	0.48	0.48	0.47	0.47	0.49	0.5	0.49	0.49		AUU	1.46	1.46	1.47	1.48	1.48	1.48	1.47	1.48
	GCU	1.64	1.64	1.63	1.64	1.64	1.64	1.64	1.64	Lys	AAA	1.5	1.5	1.49	1.49	1.5	1.5	1.5	1.51
Arg	AGA	1.67	1.67	1.65	1.66	1.71	1.7	1.7	1.71		AAG	0.5	0.5	0.51	0.51	0.5	0.5	0.5	0.49
	AGG	0.75	0.75	0.77	0.76	0.73	0.74	0.73	0.73	Met	AUG	1	1	1	1	1	1	1	1
	CGA	1.45	1.45	1.48	1.48	1.46	1.47	1.45	1.45	Phe	UUC	0.65	0.66	0.67	0.67	0.66	0.65	0.65	0.66
	CGC	0.36	0.36	0.36	0.36	0.35	0.33	0.35	0.34		UUU	1.35	1.34	1.33	1.33	1.34	1.35	1.35	1.34
	CGG	0.45	0.44	0.43	0.45	0.43	0.44	0.45	0.43	Pro	CCA	1.08	1.08	1.06	1.08	1.08	1.09	1.08	1.08
	CGU	1.32	1.32	1.31	1.3	1.33	1.32	1.32	1.33		CCC	0.74	0.74	0.76	0.78	0.77	0.77	0.77	0.77
Asn	AAC	0.47	0.47	0.48	0.48	0.47	0.48	0.48	0.47		CCG	0.61	0.61	0.62	0.6	0.63	0.62	0.61	0.63
	AAU	1.53	1.53	1.52	1.52	1.53	1.52	1.52	1.53		CCU	1.57	1.57	1.56	1.54	1.53	1.52	1.54	1.53
Asp	GAC	0.42	0.42	0.42	0.42	0.41	0.41	0.41	0.42	Ser	AGC	0.43	0.43	0.44	0.44	0.42	0.42	0.43	0.43
	GAU	1.58	1.58	1.58	1.58	1.59	1.59	1.59	1.58		AGU	1.14	1.14	1.15	1.15	1.14	1.15	1.14	1.13
Cys	UGC	0.58	0.58	0.58	0.57	0.6	0.61	0.6	0.61		UCA	1.2	1.19	1.22	1.21	1.19	1.18	1.19	1.19
	UGU	1.42	1.42	1.42	1.43	1.4	1.39	1.4	1.39		UCC	0.94	0.95	0.94	0.94	0.96	0.97	0.95	0.96
Gln	CAA	1.53	1.53	1.52	1.51	1.52	1.52	1.51	1.52		UCG	0.65	0.66	0.63	0.62	0.64	0.65	0.65	0.64
	CAG	0.47	0.47	0.48	0.49	0.48	0.48	0.49	0.48		UCU	1.63	1.63	1.61	1.63	1.65	1.63	1.64	1.64
Glu	GAA	1.47	1.47	1.46	1.47	1.47	1.47	1.48	1.48	Thr	ACA	1.21	1.21	1.22	1.22	1.21	1.22	1.22	1.21
	GAG	0.53	0.53	0.54	0.53	0.53	0.53	0.52	0.52		ACC	0.74	0.75	0.74	0.73	0.71	0.71	0.71	0.71
Gly	GGA	1.53	1.52	1.52	1.52	1.57	1.57	1.57	1.57		ACG	0.54	0.53	0.53	0.53	0.58	0.56	0.57	0.58
	GGC	0.5	0.5	0.51	0.51	0.52	0.52	0.52	0.52		ACU	1.52	1.51	1.51	1.52	1.5	1.51	1.5	1.5
	GGG	0.74	0.75	0.75	0.74	0.69	0.69	0.7	0.69	Trp	UGG	1	1	1	1	1	1	1	1
	GGU	1.23	1.23	1.22	1.23	1.22	1.22	1.22	1.22	Tyr	UAC	0.44	0.44	0.42	0.41	0.42	0.42	0.42	0.42
His	CAC	0.47	0.47	0.47	0.47	0.47	0.47	0.48	0.47		UAU	1.56	1.56	1.58	1.59	1.58	1.58	1.58	1.58
	CAU	1.53	1.53	1.53	1.53	1.53	1.53	1.52	1.53	Val	GUA	1.41	1.41	1.41	1.41	1.39	1.41	1.41	1.4
Leu	CUA	0.85	0.84	0.85	0.84	0.84	0.85	0.85	0.84		GUC	0.58	0.58	0.58	0.59	0.59	0.58	0.58	0.58
	CUC	0.41	0.41	0.41	0.41	0.41	0.41	0.41	0.42		GUG	0.54	0.54	0.54	0.54	0.54	0.53	0.53	0.54
	CUG	0.38	0.38	0.41	0.4	0.42	0.41	0.41	0.42		GUU	1.46	1.47	1.46	1.46	1.48	1.48	1.48	1.48
	CUU	1.27	1.27	1.25	1.25	1.27	1.27	1.28	1.27	Stop	UAA	1.68	1.68	1.74	1.74	1.63	1.68	1.63	1.63
	UUA	1.87	1.87	1.86	1.87	1.86	1.86	1.85	1.85		UAG	0.74	0.74	0.63	0.63	0.79	0.74	0.79	0.74
	UUG	1.22	1.23	1.22	1.22	1.2	1.2	1.2	1.19		UGA	0.58	0.58	0.63	0.63	0.58	0.58	0.58	0.63

FIGURE 3 | Sequence alignment of chloroplast genome among eight *Fagopyrum* species and three Polygonaceae species (*Rumex hypogaeus*, *Reynoutria japonica*, and *Rheum officinale*) with *F. tataricum* as a reference by using mVISTA. The Y-scale represents the percentage of identity ranging from 50 to 100%. Coding and non-coding regions are marked in purple and pink, respectively.

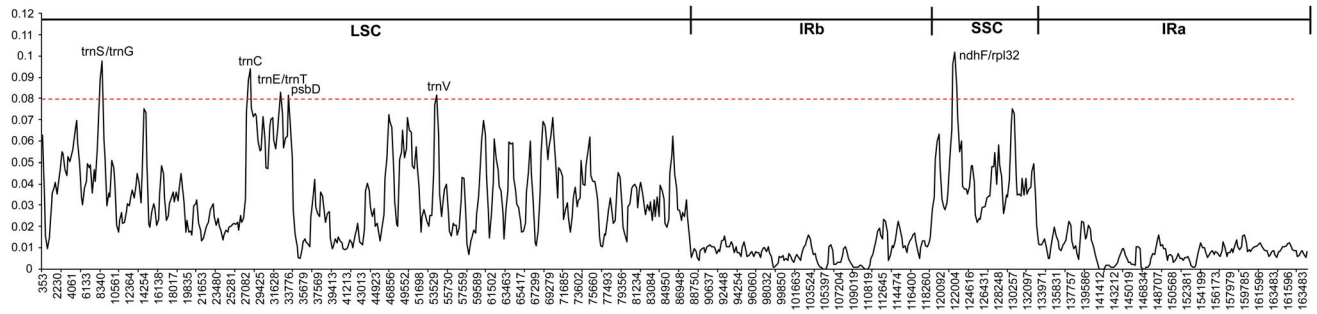

FIGURE 4 | Comparison of nucleotide diversity (Pi) values among the eight *Fagopyrum* species. *X*-axis, position of the midpoint of each window; *Y*-axis, nucleotide diversity (Pi) of each window.

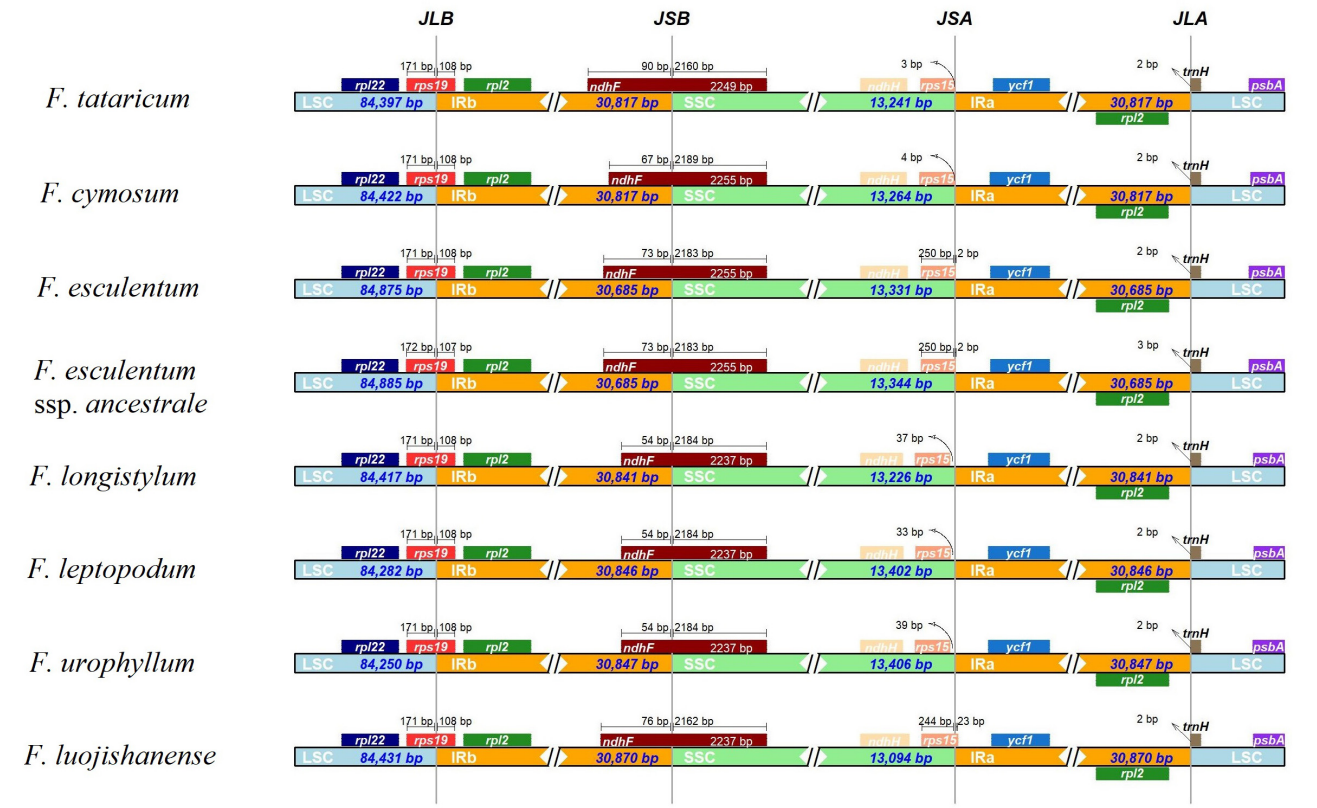

FIGURE 5 | Comparison of the junctions between the LSC, SSC, and IR regions among eight *Fagopyrum* species chloroplast genomes. The figure is not scaled LSC, SSC, and IR.

previous study (Zhang et al., 2021a). Similarly, *F. luojishanense* and *F. longistylum* of the urophyllum group may be closely related, and then cluster with *F. leptopodum* and *F. urophyllum*. These results further supported the chloroplast phylogenetic tree results. Therefore, the relationship of *Fagopyrum* plants was further inferred, *F. luojishanense*, *F. longistylum*, *F. gracilipes*, *F. gracilipes* var. *odontopterum* and other wild species may have a close relationship. According to the clustering results, *F. gracilipes* var. *odontopterum* as the division of *F. gracilipes* is considered reasonable. The *F. lineare* and *F. leptopodum* may be closely related to each other. They are both short plants, thin stem nodes, and highly adaptable in these *Fagopyrum* plants. Moreover, the two evolutionary trees supported *F. caudatum* and *F. pugense* were closely related. In general, these sequences of molecular markers with stable phylogenetic relationships of *Fagopyrum* plants will be considered as "references" to further infer taxonomic status among other species. However, it should be pointed out that the phylogenetic trees based on *matK* and ITS sequences could not completely define the relationships of some *Fagopyrum* species. For example, the genetic relationship between

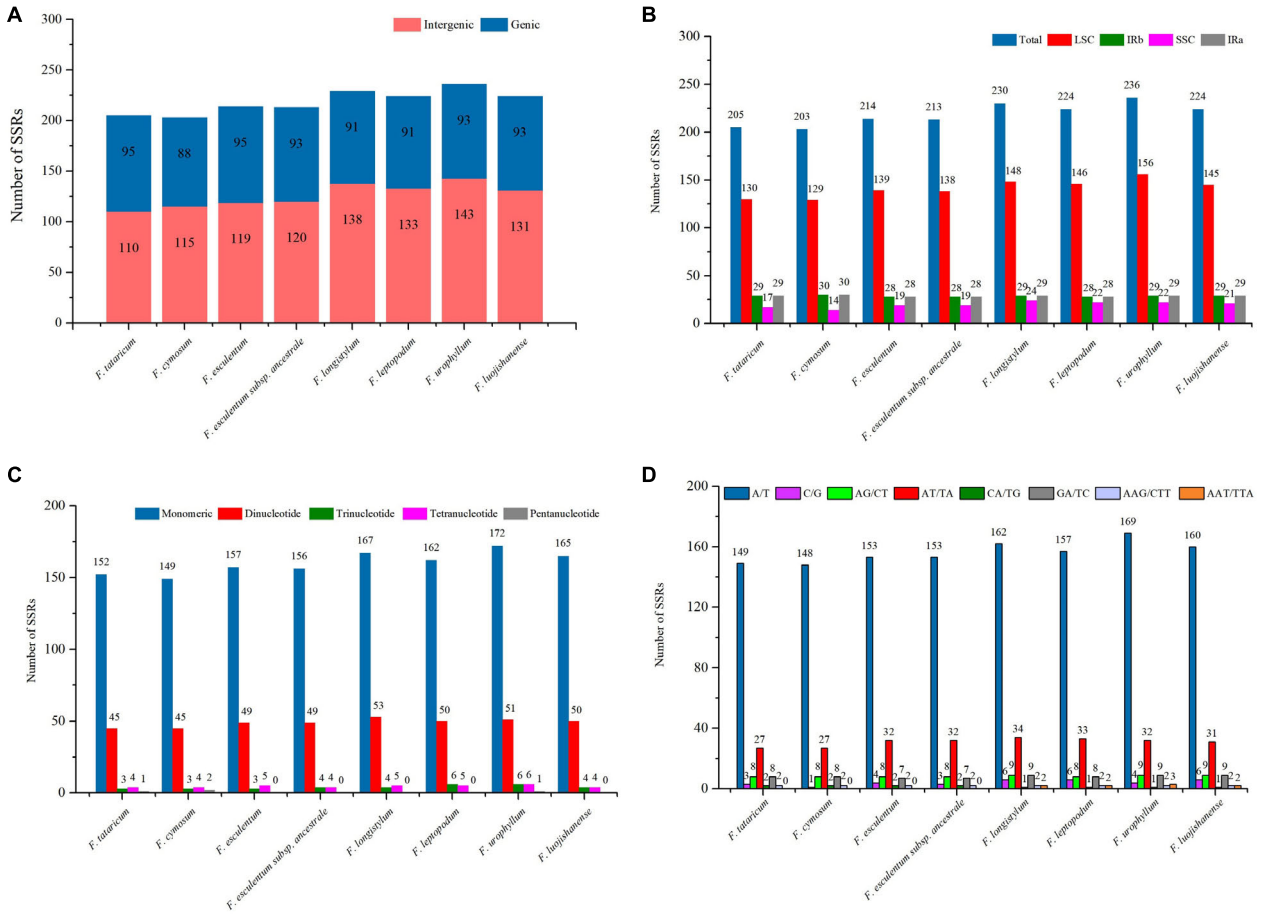

FIGURE 6 | Analysis of SSRs in the eight *Fagopyrum* species cp genomes. **(A)** Distribution of SSRs in genic and intergenic. **(B)** Distribution of SSRs in LSC, IRa/IRb, and SSC. **(C)** The number of different SSR types detected in the eight *Fagopyrum* species cp genomes. **(D)** The number of different base SSR types in the eight *Fagopyrum* species cp genomes.

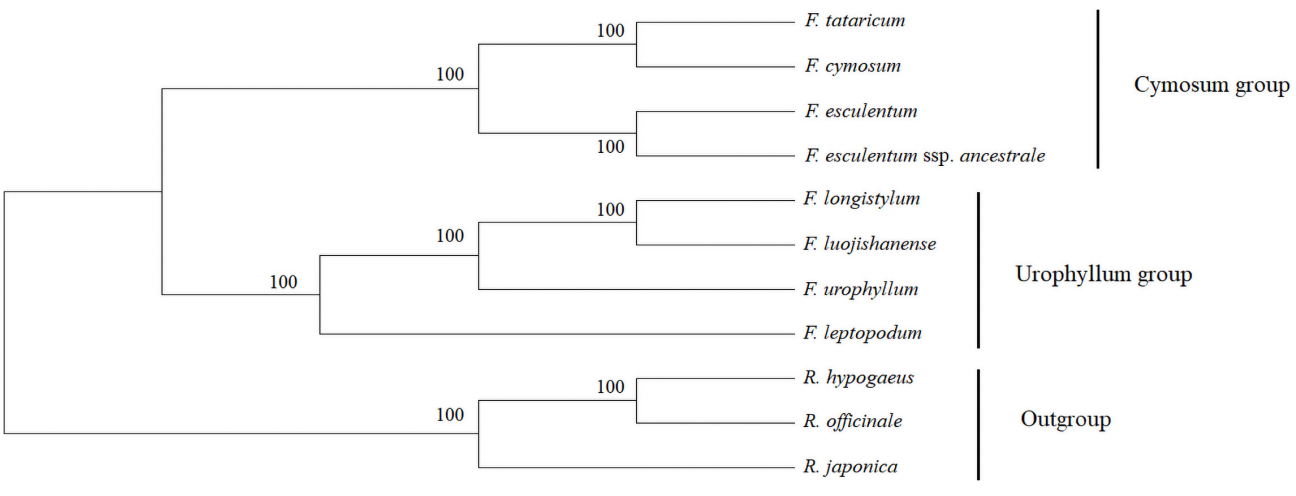

FIGURE 7 | Phylogenetic tree obtained for eight *Fagopyrum* species inferred from ML analysis constructed by the complete chloroplast genomes. The number on the branches displayed the bootstrap support values.

F. macrocarpum and *F. qiangcai* is still unstable. Therefore, it is necessary to further analyze the taxonomic status of *Fagopyrum* plants through extensive molecular marker sequences or complete genome sequencing.

DISCUSSION

Sequence Differentiation

In this study, we compared the complete cp genomes of eight *Fagopyrum* species, which showed a typical circular tetrad structure. It consisted of a LSC region (84,494.9 bp in average), a SSC region (13,288.5 bp in average), and two reverse repeats (IR) regions (30,801 bp in average). The structures, genome lengths and proportion of these cp genomes were highly conserved. Among the eight cp genomes, the gene spacer is the largest variable region, which is consistent with most angiosperms (Wicke et al., 2011). The total GC ranges from 37.78 to 37.99%, which are higher than that of *Euonymus*, and *Curcuma* (Liang et al., 2020; Li et al., 2021). The GC ratios of the cp genome of angiosperms are usually between 34 and 40%, which plays an important role in the transmission of gene information (Zhu et al., 2017). The cp genome differences of different species are obvious through changes in base composition. These GC contents of the *Fagopyrum* species are the highest in IRa/IRb region, and the uneven distribution of GC ratio and gene conversion between IR sequences, which may be the reason why the IR region is more conserved than the LSC and SSC region (Khakhlova and Bock, 2006; Fan et al., 2018).

The contraction or expansion of the IR boundary is one of the main driving forces of cp genome length and structure difference, and the change of IR/SC connection location is a typical evolutionary phenomenon in plants (He et al., 2017). Interestingly, we found significant expansion of the LSC region in *F. esculentum* and *F. esculentum* ssp. *ancestrale*, which may be direct evidence of both cp genome length expansion and IRb region contraction. In addition, a significant contraction was observed in the SSC region of *F. luojishanense* (∼13,094 bp), which had the largest IRa/IRb region (∼30,870 bp), resulting in the C terminal of *rps15* crossing into the IRb region (∼23 bp). Furthermore, we found that the loss of functional genes in cymosum members were significantly higher than that in urophyllum group. And, this phenomenon was more obvious in many transfer RNAs. Therefore, we hypothesized that this deletion may result from the apparent activity of the highly structured chloroplast genome in cymosum group. For example, *trnfM-CAU* lost in *F. esculentum* and *F. esculentum* ssp. *ancestrale*. The chloroplast genome structures of urophyllum members were more conserved, and there were little difference in the numbers and positions of encoded genes. In addition, *trnfM-CAU*/*trnM-CAU*, *trnG-UCC*/*trnG-GCC* in cymosum group were significant differences in gene location in cp genomes. tRNAs are one of the most important and versatile molecules responsible for the maintenance and maintenance of protein translation mechanisms (Mohanta et al., 2019). Differences in the number and distribution of tRNAs in the cp genome may result in significantly influences of post-translational modification processes on genes in the photosynthetic system, especially *rpoA*, *rpoB*, and *rpoC* genes (Little and Hallick, 1988; Zhang, 2020). In addition, deletion of *rpl23* gene in cp genomes of two cultivated species (*F. tataricum* and *F. esculentum*) were observed. This phenomenon illustrated a typical case of protein (gene) substitution in the evolution of chloroplast ribosomes in *Fagopyrum* plants, and nuclear genome could progressively exert stronger over the chloroplast translational system (Bubunenko et al., 1994). It is worth noting that *F. esculentum*, as a *Fagopyrum* plant which is mostly distributed in the middle and high latitude areas of the northern hemisphere with long sunshine, is observed the most loss of functional genes, such as *trnT-UGU*, *rpl23*, *trnI-CAU*, etc.

Divergence Hotspot Regions

DNA barcoding is widely used in species identification, germplasm management, genetic diversity analysis, phylogeny, and evolution (Gregory, 2005; Liu et al., 2019). In previous studies, the phylogeny of structural *Fagopyrum* plants was mainly based on SSR markers (Ma et al., 2009; Yang et al., 2020), single-copy nuclear gene (Ohnishi and Matsuoka, 1996; Ohsako and Ohnishi, 1998). The taxonomic analysis and genetic identification of *Fagopyrum* species are hampered by the lack of genomic information. Cp genome sequences are relatively conserved, which is less affected by non-parallel evolutionary in functional genes of nuclear genes in phylogenetic tree construction. Therefore, the cp genome sequences are often used in angiosperms phylogenetic prediction in recent years (Zhang et al., 2017; Zhao et al., 2020). To determine divergence packaging, the mVISTA program was used to compare the cp genome sequences of eight *Fagopyrum* species. The results showed that the cp genomes of eight *Fagopyrum* species were rich in the variable sites, and some regions with high variable frequency could be directly used as potential molecular markers for species identification (Song et al., 2017; Xu et al., 2017). In general, the proportion of variable loci in the non-coding region was higher than that in the coding region. Meanwhile, sequence differentiation in the IR region was slower and more conserved than that in LSC and SSC region. These results are consistent with most cp genome studies in plants, and we speculate that this may be due to higher gene conversion between the two IR regions (Khakhlova and Bock, 2006; Jansen and Ruhlman, 2012; Huang et al., 2014). In addition, the nucleotide diversity (Pi) of eight *Fagopyrum* species were assessed by sliding window analysis. These results of Pi values were generally consistent with mVISTA analysis, and the nucleotide diversity in the non-coding region was higher than that in the coding region. Six variable regions (*ndhF-rpl32*, *trnS-trnG*, *trnC*, *trnE-trnT*, *psbD,* and *trnV*) were identified as highly variable sites at the species level of *Fagopyrum*. These variable regions were further used to identify the genetic relationship of eight *Fagopyrum* species. And, the results showed that *trnS-trnG*, and *trnV* trees showed highly consistent results with cp genomes, so that they were further recommended as potential molecular markers in genetic development analysis and assisted breeding in *Fagopyrum* plants.

Identification of Repeated Sequences

Simple repeat sequences play important role in the combination and arrangement of cp genome structures, which are highly variable in different species of the same genus, Thus, SSRs have been widely used in population genetics and species biodiversity studies (Thiel et al., 2003; Zhou et al., 2019). In this study, it was found that the SSR polymorphism levels of the four major components of these cp genomes were inconsistent. SSRs were mainly found in the LSC region of the eight *Fagopyrum* species, which was closely related to the interval length. The distribution density of SSRs in the eight *Fagopyrum* species were uneven, and there may be more SSRs in some sections and gene locus. For example, *matK*, *rpoC2*, *clpP*, *ycf1*, *ycf2*, *ycf3*, and other gene regions showed higher SSR density, which was consistent with Zingiberales and other plants (Liang et al., 2020). Although the cp genome evolution of *Fagopyrum* plants is generally co-evolutionary, some functional gene regions may respond to important biological effects and thus be subjected to more significant evolutionary pressures (Williams et al., 2019). At present, only a few "star genes," such as *matK*, *rbcL*, *ycf1*, and *ycf2*, have been found as common positive selection sites (Liang et al., 2020; Li et al., 2021), other studies on the response evolution and biological role of chloroplast functional genes are still scarce. Nevertheless, it is desirable to select some segments or polymorphism of repeating sequence fragments from the cp genome as new tools for studying systematic differentiation.

A total of 110 (∼*F. tataricum*) ∼143 SSR markers (∼*F. urophyllum*) were found in the cp genomes of eight *Fagopyrum* plants, including mononucleotides, dinucleotides, tetranucleotides, trinucleotides, pentucleotide. Notably, there were no hexonotides found in all *Fagopyrum* species, which is inconsistent with Euonymus, Zanthoxylum, Curcuma, Wurfbainia Villosa, Amomum, Kaempferia, etc. (Liang et al., 2020; Li et al., 2021; Zhao et al., 2021). A/T and AT/TA repeats are the main SSR types, which may be because A/T bases are more easily changed than G/C bases (Li et al., 2021). However, these AT-rich regions did not contribute significantly to the expansion of cp genome size (**Figure 6D**). Compared with the gene regions, most of the SSRs were distributed in the intergene region (IGS region), which was more obvious in the members of the urophyllum group. It should be noted that there were significant differences in SSR markers in some gene regions between the urophyllum group and the cymosum group. For example, CA (4) existed only in cymosum group members, while AAT (4), AG (5), GA (5), TCAA (3), and TTA (4) were all found in urophyllum group members. These markers can be further applied to the identification of the two subgroups. In addition, many unique SSR markers were found in some *Fagopyrum* species, which can be used in the identification of different species. For example, AAAT (3) only existed in tartary buckwheat, AATT (4), A (16), TCTAT (3) only exist in *F. cymosum*, AATG (4) only existed in *F. longistylum*. Interestingly, there are still some unique SSR markers in *F. esculentum* and *F. esculentum* ssp. *ancestrale*, which will be effectively used in the identification of cultivated and wild ancestor species. For example, TTGA (3) was found in *F. esculentum*, while GTA (5), and C (12) were unique to *F. esculentum* ssp. *ancestrale*.

Interestingly, we observed significant differences in repeat sequences among some photosystem genes between members of the cymosum group and urophyllum group (**Supplementary Table 7**). For example, *ycf1* and two ribosome large subunit genes (*rpl32*, *rps15*) at the IR boundary showed significant SSR expansion in the cymosum group. This may contribute to the light adaptation of cymosum group members, which is conducive to planting (Fan et al., 2018; Liang et al., 2020). Photosystem subunit genes (*psaJ*, *psbK*, *psbZ*) showed significant SSR expansion in *F. esculentum* and *F. esculentum* subsp. *ancestrale*. They are more adapted to the long-sunshine of the northern hemisphere (Ikeuchi et al., 1991; Sugimoto and Takahashi, 2003). In addition, the urophyllum group members have a narrower distribution range, mainly growing in mountainous areas of southwest China. However, they are more adaptable to complex geographical environments, such as mountain areas and sandy areas, which are too harsh for the cultivated species (Zhou et al., 2018). In general, the process of artificial domestication or natural selection pressure leads to a significant decline in genetic diversity in the genome (Louwaars, 2018; Zhu et al., 2019). However, this was not significantly reflected in cp genomes of *F. cymosum*, *F. esculentum*, and *F. tataricum*. Therefore, we speculate that these domestication intervals may exist mainly in the nuclear genome. In conclusion, SSR markers of eight *Fagopyrum* species were systematically reported for the first time, which can provide a reference for the subsequent study of molecular evolution and phylogeny of *Fagopyrum* genus and Polygonaceae family.

Phylogenetic Relationships

For a long time, the taxonomic status of *Fagopyrum* genus has changed frequently, and no consensus has been reached among different species (Linnaeus, 1753; Miller, 1754; Meisner, 1826; Gross, 1913; Stewad, 1930; Zhang et al., 2021b). In this study, the phylogenetic trees based on cp genomes of eight *Fagopyrum* species and *Rumex*, *Rheum*, and *Reynoutria* supported the independent evolution of *Fagopyrum* plants. Therefore, it is reliable that *Fagopyrum* has a separate taxonomic status in the Polygonaceae.

Furthermore, the cymosum members (*F. tataricum*, *F. cymosum*, *F. esculentum*, *F. esculentum* subsp. *ancestrale*) had significant independent cluster branches into the urophyllum group. Therefore, we infer that the evolutionary processes of the two groups of *Fagopyrum* species may be independent rather than overlapping. Similarly, the separation of the cymosum group and the urophyllum group may be earlier than the flower type differentiation of *Fagopyrum* plants, and then two pollination modes of self-pollination (self-compatibility) and cross-pollination (self-incompatibility) are produced. In addition, this study concluded that the genetic relationship in the cymosum group is clear, the *F. cymosum* and *F. tataricum* are more closely related than *F. esculentum*, although their pollination patterns are not consistent. However, the taxonomic status of the members of the urophyllum group are more complicated, as the urophyllum group consists of 18 species. Although there were significant differences in differentiation rates between nuclear and cp genomes, ITS clearly supported the clustering results of the urophyllum group in the evolutionary

tree of cp genomes. Four urophyllum group members can further anchor the taxonomic status of other wild species members, which is further supported by the previous study (Cheng et al., 2020; Zhang et al., 2021a). It should be noted that the taxonomic status of some members of the urophyllum Group cannot be significantly anchored by a single molecular marker, which may require further molecular evidence.

DATA AVAILABILITY STATEMENT

The data presented in the study are deposited in the National Center for Biotechnology Information (NCBI) repository, accession number were: *F. longistylum* (OK054489), *F. urophyllum* (OK054490), *F. leptopodum* (OK054491).

AUTHOR CONTRIBUTIONS

KZ, MZ, JC, and YF conceived and designed the work. YT and MD collected the samples. YF, YJ, and KZ performed the experiments and analyzed the data. YF and YJ wrote the manuscript. MD, MZ, and JC revised the manuscript. All the authors have read and agreed to the published version of the manuscript.

FUNDING

This work was financially supported by the National Key R&D Program of China (2019YFD1000700 and 2019YFD1000703) and National Science Foundation of China (31560578).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.799904/full#supplementary-material

Supplementary Figure S1 | Phylogenetic tree based on the *ndhF-rpl32* **(A)**, *psbD* **(B)**, *trnC* **(C)**, *trnE-trnT* **(D)**, *trnS-trnG* **(E)**, and *trnV* **(F)** sequences of eight *Fagopyrum* species constructed from NJ analysis.

Supplementary Figure S2 | Phylogenetic tree based on the ITS **(A)** and *matK* **(B)** sequences of eighteen *Fagopyrum* species constructed from ML analysis.

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

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

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

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

Bausher, M. G., Singh, N. D., Lee, S. B., Jansen, R. K., and Daniell, H. (2006). The complete chloroplast genome sequence of *Citrus sinensis* (L.) Osbeck var 'Ridge Pineapple': organization and phylogenetic relationships to other angiosperms. *BMC Plant Biol.* 6:21. doi: 10.1186/1471-2229-6-21

Beaulieu, J. M., Leitch, I. J., Patel, S., Pendharkar, A., and Knight, C. A. (2008). Genome size is a strong predictor of cell size and stomatal density in angiosperms. *New Phytol.* 179, 975–986. doi: 10.1111/j.1469-8137.2008.02528.x

Beier, S., Thiel, T., Münch, T., Scholz, U., and Mascher, M. (2017). MISA-web: a web server for microsatellite prediction. *Bioinformatics* 33, 2583–2585. doi: 10.1093/bioinformatics/btx198

Bubunenko, M. G., Schmidt, J., and Subramanian, A. R. (1994). Protein substitution in chloroplast ribosome evolution. A eukaryotic cytosolic protein has replaced its organelle homologue (L23) in spinach. *J. Mol. Biol.* 240, 28–41. doi: 10.1006/jmbi.1994.1415

Cai, Z., Guisinger, M., Kim, H. G., Ruck, E., Blazier, J. C., McMurtry, V., et al. (2008). Extensive reorganization of the plastid genome of *Trifolium subterraneum* (Fabaceae) is associated with numerous repeated sequences and novel DNA insertions. *J. Mol. Evol.* 67, 696–704. doi: 10.1007/s00239-008-9180-7

Cheng, C., Fan, Y., Tang, Y., Zhang, K., Joshi, D. C., Jha, R., et al. (2020). Fagopyrum esculentum ssp. *ancestrale*-a hybrid species between diploid *F. cymosum* and *F. esculentum*. *Front. Plant Sci.* 11:1073. doi: 10.3389/fpls.2020.01073

Cho, K. S., Yun, B. K., Yoon, Y. H., Hong, S. Y., Mekapogu, M., Kim, K. H., et al. (2015). Complete chloroplast genome sequence of tartary buckwheat (*Fagopyrum tataricum*) and comparative analysis with common buckwheat (*F. esculentum*). *PLoS One* 10:e0125332. doi: 10.1371/journal.pone.0125332

Chumley, T. W., Palmer, J. D., Mower, J. P., Fourcade, H. M., Calie, P. J., Boore, J. L., et al. (2006). The complete chloroplast genome sequence of *Pelargonium* x *hortorum*: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. *Mol. Biol. Evol.* 23, 2175–2190. doi: 10.1093/molbev/msl089

Cosner, M. E., Raubeson, L. A., and Jansen, R. K. (2004). Chloroplast DNA rearrangements in Campanulaceae: phylogenetic utility of highly rearranged genomes. *BMC Evol. Biol.* 4:27. doi: 10.1186/1471-2148-4-27

Doyle, J. J. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* 19, 11–15.

Fan, W. B., Wu, Y., Yang, J., Shahzad, K., and Li, Z. H. (2018). Comparative chloroplast genomics of dipsacales species: insights into sequence variation, adaptive evolution, and phylogenetic relationships. *Front. Plant Sci.* 9:689. doi: 10.3389/fpls.2018.00689

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

Gao, L. Z., Liu, Y. L., Zhang, D., Li, W., Gao, J., Liu, Y., et al. (2019). Evolution of Oryza chloroplast genomes promoted adaptation to diverse ecological habitats. *Commun. Bio.* 2:278. doi: 10.1038/s42003-019-0531-2

Gregory, T. R. (2005). DNA barcoding does not compete with taxonomy. *Nature* 434:1067. doi: 10.1038/4341067b

Gross, M. H. (1913). Remarques sur les polygonees del'Asie orientale. *Bull. Torrey Bot. Club* 23, 7–32.

Guisinger, M. M., Chumley, T. W., Kuehl, J. V., Boore, J. L., and Jansen, R. K. (2010). Implications of the plastid genome sequence of Typha (Typhaceae, Poales) for understanding genome evolution in Poaceae. *J. Mol. Evol.* 70, 149–166. doi: 10.1007/s00239-009-9317-3

He, L., Qian, J., Li, X., Sun, Z., Xu, X., and Chen, S. (2017). Complete chloroplast genome of medicinal plant *Lonicera japonica*: genome rearrangement, intron gain and loss, and implications for phylogenetic studies. *Molecules* 22:249. doi: 10.3390/molecules22020249

Hou, L. L., Zhou, M. L., Zhang, Q., Qi, L. P., Yang, B. X., Tang, Y., et al. (2015). *Fagopyrum luojishanense*, a new species of polygonaceae from sichuan, China. *Novon J. Bot. Nomenclat.* 24, 22–26. doi: 10.3417/2013047

Huang, H., Shi, C., Liu, Y., Mao, S. Y., and Gao, L. Z. (2014). Thirteen *Camellia* chloroplast genome sequences determined by high-throughput sequencing:

genome structure and phylogenetic relationships. *BMC Evol. Biol.* 14:151. doi: 10.1186/1471-2148-14-151

Huang, Y., Wang, J., Yang, Y., Fan, C., and Chen, J. (2017). Phylogenomic analysis and dynamic evolution of chloroplast genomes in salicaceae. *Front. Plant Sci.* 8:1050. doi: 10.3389/fpls.2017.01050

Ikeuchi, M., Eggers, B., Shen, G. Z., Webber, A., Yu, J. J., Hirano, A., et al. (1991). Cloning of the *psb*K gene from *Synechocystis* sp. PCC 6803 and characterization of photosystem II in mutants lacking PSII-K. *J. Biol. Chem.* 266, 11111–11115.

Jansen, R. K., and Ruhlman, T. A. (2012). *Plastid Genomes of Seed Plants, Genomics of Chloroplasts, and Mitochondria*. Dordrecht: Springer. 103–126. doi: 10.1007/978-94-007-2920-9_5

Jansen, R. K., Cai, Z., Raubeson, L. A., Daniell, H., Depamphilis, C. W., Leebens-Mack, J., et al. (2007). Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proc. Natl. Acad. Sci. USA* 104, 19369–19374. doi: 10.1073/pnas.0709121104

Jarvis, P., and Soll, J. (2001). Toc, Tic, and chloroplast protein import. *Biochimica et biophysica acta.* 1541, 64–79. doi: 10.1016/s0167-4889(01)00147-1

Jin, S., and Daniell, H. (2015). The engineered chloroplast genome just got smarter. *Trends Plant Sci.* 20, 622–640. doi: 10.1016/j.tplants.2015.07.004

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

Khakhlova, O., and Bock, R. (2006). Elimination of deleterious mutations in plastid genomes by gene conversion. *Plant J. Cell Mol. Biol.* 46, 85–94. doi: 10.1111/j.1365-313X.2006.02673.x

Lee, H. L., Jansen, R. K., Chumley, T. W., and Kim, K. J. (2007). Gene relocations within chloroplast genomes of *Jasminum* and *Menodora* (Oleaceae) are due to multiple, overlapping inversions. *Mol. Biol. Evol.* 24, 1161–1180. doi: 10.1093/molbev/msm036

Leister, D. (2003). Chloroplast research in the genomic age. *Trends Genet.* 19, 47–56. doi: 10.1016/s0168-9525(02)00003-3

Li, B., Lin, F., Huang, P., Guo, W., and Zheng, Y. (2020). Development of nuclear SSR and chloroplast genome markers in diverse *Liriodendron chinense* germplasm based on low-coverage whole genome sequencing. *Biol. Res.* 53:21. doi: 10.1186/s40659-020-00289-0

Li, X., Li, Y., Zang, M., Li, M., and Fang, Y. (2018). Complete chloroplast genome sequence and phylogenetic analysis of *quercus acutissima*. *Int. J. Mol. Sci.* 19:2443. doi: 10.3390/ijms19082443

Li, Y., Chen, X., Wu, K., Pan, J., Long, H., and Yan, Y. (2020). Characterization of simple sequence repeats (SSRs) in ciliated protists inferred by comparative genomics. *Microorganisms* 8:662. doi: 10.3390/microorganisms8050662

Li, Y., Dong, Y., Liu, Y., Yu, X., Yang, M., and Huang, Y. (2021). Comparative analyses of *euonymus* chloroplast genomes: genetic structure, screening for loci with suitable polymorphism, positive selection genes, and phylogenetic relationships within celastrineae. *Front. Plant Sci.* 11:593984. doi: 10.3389/fpls.2020.593984

Liang, H., Zhang, Y., Deng, J., Gao, G., Ding, C., Zhang, L., et al. (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

Linnaeus, C. (1753). *Species plantarum I:359*. Holmiae: Laurentius Salvius.

Little, M. C., and Hallick, R. B. (1988). Chloroplast rpoA, rpoB, and rpoC genes specify at least three components of a chloroplast DNA-dependent RNA polymerase active in tRNA and mRNA transcription. *J. Biol. Chem.* 263, 14302–14307.

Liu, M., Li, X. W., Liao, B. S., Luo, L., and Ren, Y. Y. (2019). Species identification of poisonous medicinal plant using DNA barcoding. *Chin. J. Nat. Med.* 17, 585–590. doi: 10.1016/S1875-5364(19)30060-3

Liu, J. L., Tang, Y., Xia, M. Z., Shao, J. R., Cai, G. Z., Luo, Q., et al. (2008). *Fagopyrum crispatifolium* a new species of Polygonaceae from Sichuan, China. *J. Syst. Evol.* 46, 929–932.

Logacheva, M. D., Samigullin, T. H., Dhingra, A., and Penin, A. A. (2008). Comparative chloroplast genomics and phylogenetics of *Fagopyrum esculentum* ssp. *ancestrale* - A wild ancestor of cultivated buckwheat. *BMC Plant Biol.* 8:59. doi: 10.1186/1471-2229-8-59

Lohse, M., Drechsel, O., Kahlau, S., and Bock, R. (2013). Organellar Genome DRAW–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

Louwaars, N. P. (2018). Plant breeding and diversity: A troubled relationship? *Euphytica: Netherlands J. Plant Breed.* 214:114. doi: 10.1007/s10681-018-2192-5

Ma, K. H., Kim, N. S., Lee, G. A., Lee, S. Y., Lee, J. K., Yi, J. Y., et al. (2009). Development of SSR markers for studies of diversity in the genus *Fagopyrum*. TAG. Theoretical and applied genetics. *Theoretische und angewandte Genetik* 119, 1247–1254. doi: 10.1007/s00122-009-1129-8

Martin, G. E., Rousseau-Gueutin, M., Cordonnier, S., Lima, O., Michon-Coudouel, S., Naquin, D., et al. (2014). The first complete chloroplast genome of the Genistoid legume *Lupinus luteus*: evidence for a novel major lineage-specific rearrangement and new insights regarding plastome evolution in the legume family. *Ann. Bot.* 113, 1197–1210. doi: 10.1093/aob/mcu050

Meisner, C. F. (1826). *Monographiae Generis Polygoni Prodromus*. Genevae: Sumtibus Auctoris.

Miller, P. H. (1754). *The Gardeners Dictionary. Abridged(edition 4)*. London: Self-publishing.

Mohanta, T. K., Khan, A. L., Hashem, A., Allah, E., Yadav, D., and Al-Harrasi, A. (2019). Genomic and evolutionary aspects of chloroplast tRNA in monocot plants. *BMC Plant Biol.* 19:39. doi: 10.1186/s12870-018-1625-6

Nagatomo, Y., Usui, S., Ito, T., Kato, A., Shimosaka, M., and Taguchi, G. (2014). Purification, molecular cloning and functional characterization of flavonoid C-glucosyltransferases from *Fagopyrum esculentum* M. (buckwheat) cotyledon. *Plant J. Cell Mol. Biol.* 80, 437–448. doi: 10.1111/tpj.12645

Neethirajan, S., Hirose, T., Wakayama, J., Tsukamoto, K., Kanahara, H., and Sugiyama, S. (2011). Karyotype analysis of buckwheat using atomic force microscopy. *Microsc Microanal.* 17, 572–577. doi: 10.1017/S1431927611000481

Neuhaus, H. E., and Emes, M. J. (2000). Nonphotosynthetic metabolism in plastids. *Ann. Rev. Plant Physiol. Plant Mol. Biol.* 51, 111–140. doi: 10.1146/annurev.arplant.51.1.111

Ohnishi, O. (1995). Discovery of new *Fagopyrum* species and its implication for the studies of evolution of *Fagopyrum* and of the origin of cultivated buckwheat. *Proc. Intl. Symp. Buckwheat* 1995, 175–190.

Ohnishi, O. (1998). Search for the wild ancestor of buckwheat I Description of new *Fagopyrum* (Polygonaceae) species and their distribution in China. *Fagopyrum* 15, 18–28.

Ohnishi, O., and Matsuoka, Y. (1996). Search for the wild ancestor of buckwheat ii. taxonomy of *Fagopyrum* (polygonaceae) species based on morphology, isozymes and cpdna variability. *Genes Genetic Syst.* 71, 383–390.

Ohsako, T., and Ohnishi, O. (1998). New *Fagopyrum* species revealed by morphological and molecular analyses. *Genes Genetic Syst.* 73, 85–94.

Ohsako, T., Yamane, K., and Ohnishi, O. (2002). Two new *Fagopyrum* (po1ygonaceae) species *F. gracilipedoides* and *F. jinshaense* from Yunnan. *Genes Genetic Syst.* 77, 399–408.

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, K. T., and Park, S. (2021). Phylogenomic Analyses of *Hepatica* Species and Comparative Analyses Within Tribe Anemoneae (Ranunculaceae). *Front. Plant Sci.* 12:638580. doi: 10.3389/fpls.2021.638580

Peden, J. F. (2000). Analysis of codon usage. *Univ. Nottingham.* 90, 73–74. doi: 10.1006/expr.1997.4185

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.* 34, 3299–3302. doi: 10.1093/molbev/msx248

Ruhlman, T. A., and Jansen, R. K. (2014). The plastid genomes of flowering plants. *Methods Mol. Biol.* 1132, 3–38. doi: 10.1007/978-1-62703-995-6_1

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

analysis with other legume genomes. *Plant Mol. Biol.* 59, 309–322. doi: 10.1007/s11103-005-8882-0

Shao, J. R., Zhou, M. L., Zhu, X. M., Wang, D. Z., and Bai, D. Q. (2011). *Fagopyrum wenchuanense* and *Fagopyrum qiangcai*, two new species of polygonaceae from sichuan, china. *Novon* 21, 256–261.

Sharma, R., and Jana, S. (2002). Species relationships in *Fagopyrum* revealed by PCR-based DNA fingerprinting. TAG. Theoretical and applied genetics. *Theoretische und angewandte Genetik.* 105, 306–312. doi: 10.1007/s00122-002-0938-9

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.

Song, Y., Wang, S., Ding, Y., Xu, J., Li, M. F., Zhu, S., et al. (2017). Chloroplast genomic resource of Paris for species discrimination. *Sci. Rep.* 7:3427. doi: 10.1038/s41598-017-02083-7

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22, 2688–2690. doi: 10.1093/bioinformatics/btl446

Stewad, A. N. (1930). The Polygoneae of eastern Asia. *Contrib. Gray Herbarium Harvard Universit.* 88, 1–129.

Sugimoto, I., and Takahashi, Y. (2003). Evidence that the *Psb*K polypeptide is associated with the photosystem II core antenna complex CP43. *J. Biol. Chem.* 278, 45004–45010. doi: 10.1074/jbc.M307537200

Tang, Y., Zhou, M. L., Bai, D. Q., Shao, J. R., Zhu, X. M., Wang, D. Z., et al. (2010). *Fagopyrum pugense* (Polygonaceae), a new species from Sichuan, China. *Novon J. Bot. Nomenclature* 20, 239–242.

Thiel, T., Michalek, W., Varshney, R. K., 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

Tonti-Filippini, J., Nevill, P. G., Dixon, K., and Small, I. (2017). What can we do with 1000 plastid genomes?. *Plant J. Cell Mol. Biol.* 90, 808–818. doi: 10.1111/tpj.13491

Wang, C. L., Ding, M. Q., Zou, C. Y., Zhu, X. M., Tang, Y., Zhou, M. L., et al. (2017b). Comparative analysis of four buckwheat species based on morphology and complete chloroplast genome sequences. *Sci. Rep.* 7:6514. doi: 10.1038/s41598-017-06638-6

Wang, C. L., Li, Z. Q., Ding, M. Q., Tang, Y., Zhu, X., and Liu, J. (2017a). *Fagopyrum longzhoushanense*, a new species of Polygonaceae from Sichuan. *China Phytotaxa.* 291, 73–80.

Wanga, V. O., Dong, X., Oulo, M. A., Mkala, E. M., Yang, J. X., Onjalalaina, G. E., et al. (2021). Complete chloroplast genomes of *Acanthochlamys bracteata* (China) and *Xerophyta* (Africa) (Velloziaceae): comparative genomics and phylogenomic placement. *Front. Plant Sci.* 12:691833. doi: 10.3389/fpls.2021.691833

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

Wu, C. X., Zhai, C. C., and Fan, S. J. (2020). Characterization of the complete chloroplast genome of *Rumex nepalensis* (Polygonaceae). *Mitochondrial DNA B Resour.* 5, 2458–2459. doi: 10.1080/23802359.2020.1778568

Xu, C., Dong, W., Li, W., Lu, Y., Xie, X., Jin, X., et al. (2017). Comparative analysis of six *Lagerstroemia* complete chloroplast genomes. *Front. Plant Sci.* 8:15. doi: 10.3389/fpls.2017.00015

Yang, B., Li, L., Liu, J., and Zhang, L. (2020). Plastome and phylogenetic relationship of the woody buckwheat *Fagopyrum tibeticum* in the Qinghai-Tibet Plateau. *Plant Divers.* 43, 198–205. doi: 10.1016/j.pld.2020.10.001

Yang, Y., Zhang, Y., Chen, Y., Gul, J., Zhang, J., Liu, Q., et al. (2019). Complete chloroplast genome sequence of the mangrove species Kandelia obovata and comparative analyses with related species. *PeerJ* 7:e7713. doi: 10.7717/peerj.7713

Ye, N. G., and Guo, G. Q. (1992). Classification, origin and evolution of genus *Fagopyrum* in China. *Taiyuan: Agricult. Publ. House* 1992, 19–28.

Zhang, D., Gao, F., Li, W. X., Jakovlic, I., Zou, H., Zhang, J., et al. (2018). 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, K., Fan, Y., Weng, W. F., Tang, Y., and Zhou, M. L. (2021a). *Fagopyrum longistylum* (Polygonaceae), a new species from Sichuan. *China. Phytotaxa* 482, 173–182.

Zhang, K., He, M., Fan, Y., Zhao, H., Gao, B., Yang, K., et al. (2021b). Resequencing of global Tartary buckwheat accessions reveals multiple domestication events and key loci associated with agronomic traits. *Genome Biol.* 22:23. doi: 10.1186/s13059-020-02217-7

Zhang, S., Jin, 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, T. (2020). The butterfly effect: natural variation of a chloroplast tRNA-modifying enzyme leads to pleiotropic developmental defects in rice. *Plant Cell* 32, 2073–2074. doi: 10.1105/tpc.20.00342

Zhang, Y., and Chen, C. (2018). The complete chloroplast genome sequence of the medicinal plant *Fagopyrum dibotrys* (Polygonaceae). *Mitochondrial DNA Part B Res.* 3, 1087–1089. doi: 10.1080/23802359.2018.1483761

Zhao, K., Li, L., Lu, Y., Yang, J., Zhang, Z., Zhao, F., et al. (2020). Characterization and comparative analysis of two rheum complete chloroplast genomes. *Biomed. Res. Int.* 2020, 1–11. doi: 10.1155/2020/6490164

Zhao, K., Li, L., Quan, H., Yang, J., Zhang, Z., Liao, Z., et al. (2021). Comparative Analyses of Chloroplast Genomes From 14 *Zanthoxylum* Species: Identification of Variable DNA Markers and Phylogenetic Relationships Within the Genus. *Front. Plant Sci.* 11:605793. doi: 10.3389/fpls.2020.605793

Zhao, Z., Gao, A., and Huang, J. (2019). Sequencing and analysis of chloroplast genome of *Clausena lansium* (lour.) Skeels. *Anhui. Agric. Sci.* 47, 115–118. doi: 10.3969/j.issn.0517-6611.2019.11.032

Zhou, M. L., Tang, Y., Deng, X. Y., Ruan, C., Tang, Y. X., and Wu, Y. M. (2018). *Classification and Nomenclature of Buckwheat Plants, Buckwheat Germplasm in the World*. Cambridge: Academic Press. 9–20.

Zhou, T., Ruhsam, M., Wang, J., Zhu, H., Li, W., Zhang, X., et al. (2019). The complete chloroplast genome of *Euphrasia regelii*, Pseudogenization of ndh genes and the phylogenetic relationships within Orobanchaceae. *Front. Genet.* 10:444. doi: 10.3389/fgene.2019.00444

Zhu, G., Li, W., Wang, G., Li, L., Si, Q., Cai, C., et al. (2019). Genetic basis of fiber improvement and decreased stress tolerance in cultivated versus semi-domesticated upland cotton. *Front. Plant Sci.* 10:1572. doi: 10.3389/fpls.2019.01572

Zhu, T., Zhang, L., Chen, W., Yin, J., and Li, Q. (2017). Analysis of chloroplast genomes in 1342 plants. *Genom. Appl. Biol.* 36, 4323–4333. doi: 10.13417/j.gab.036.004323

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

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

frontiers in Plant Science

ORIGINAL RESEARCH
published: 17 December 2021
doi: 10.3389/fpls.2021.804568

Chloroplast Phylogenomic Analyses Reveal a Maternal Hybridization Event Leading to the Formation of Cultivated Peanuts

Xiangyu Tian[1], Luye Shi[1], Jia Guo[1], Liuyang Fu[1,2], Pei Du[2], Bingyan Huang[2], Yue Wu[1], Xinyou Zhang[2] and Zhenlong Wang[1]**

[1] School of Life Sciences, Zhengzhou University, Zhengzhou, China, [2] Key Laboratory of Oil Crops in Huang-Huai-Hai Plains, Ministry of Agriculture and Rural Affairs, Henan Provincial Key Laboratory for Oil Crops Improvement, Henan Institute of Crop Molecular Breeding, Henan Academy of Agricultural Sciences, Zhengzhou, China

Peanuts (*Arachis hypogaea* L.) offer numerous healthy benefits, and the production of peanuts has a prominent role in global food security. As a result, it is in the interest of society to improve the productivity and quality of peanuts with transgenic means. However, the lack of a robust phylogeny of cultivated and wild peanut species has limited the utilization of genetic resources in peanut molecular breeding. In this study, a total of 33 complete peanut plastomes were sequenced, analyzed and used for phylogenetic analyses. Our results suggest that sect. *Arachis* can be subdivided into two lineages. All the cultivated species are contained in Lineage I with AABB and AA are the two predominant genome types present, while species in Lineage II possess diverse genome types, including BB, KK, GG, *etc*. Phylogenetic studies also indicate that all allotetraploid cultivated peanut species have been derived from a possible maternal hybridization event with one of the diploid *Arachis duranensis* accessions being a potential AA sub-genome ancestor. In addition, *Arachis monticola,* a tetraploid wild species, is placed in the same group with all the cultivated peanuts, and it may represent a transitional species, which has been through the recent hybridization event. This research could facilitate a better understanding of the taxonomic status of various *Arachis* species/accessions and the evolutionary relationship among them, and assists in the correct and efficient use of germplasm resources in breeding efforts to improve peanuts for the benefit of human beings.

Keywords: *Arachis*, whole plastid genome, genetic structure, phylogenomics, maternal hybridization event

OPEN ACCESS

Edited by:
Hai Du,
Southwest University, China

Reviewed by:
Abdullah,
Quaid-i-Azam University, Pakistan
Xu Zhang,
Wuhan Botanical Garden, Chinese Academy of Sciences (CAS), China

****Correspondence:***
Xinyou Zhang
haasz@126.com
Zhenlong Wang
wzl@zzu.edu.cn

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

Received: *29 October 2021*
Accepted: *30 November 2021*
Published: *17 December 2021*

Citation:
Tian X, Shi L, Guo J, Fu L, Du P, Huang B, Wu Y, Zhang X and Wang Z (2021) Chloroplast Phylogenomic Analyses Reveal a Maternal Hybridization Event Leading to the Formation of Cultivated Peanuts. Front. Plant Sci. 12:804568. doi: 10.3389/fpls.2021.804568

INTRODUCTION

The genus *Arachis* consists of approximately 81 species, which represent nine sections and 16 genome types, and are mainly distributed in the tropics and subtropics of South America (Stalker, 2017). Among these, peanut or groundnut (*Arachis hypogaea* L.) is a world-famous legume crop and cultivated by more than one hundred countries in the tropical and subtropical regions (Singh and Moss, 1982; Varshney et al., 2009; Pandey et al., 2020). Peanut was domesticated about 3,500–9400 years ago in South America (Bertioli et al., 2019; Chen et al., 2019; Zhuang et al., 2019). It

is known as the "longevity fruit," "poor man's almonds" because it is an excellent source of good fats and proteins (∼80% of seed content). Peanut has also become one of the most important contributors to human health and food security (Konate et al., 2020). In addition to cultivated peanuts, some wild species including *Arachis glabrata*, *Arachis pintoi*, *Arachis stenosperma,* and *Arachis villosulicarpa, etc*. are also used as food and medicine (Stalker, 2017). More importantly, some wild *Arachis* species possess many agronomic traits, such as disease and pest resistances (Subrahmanyam et al., 2001; Tallury et al., 2014), which are important in crop improvement, but these traits are not present in cultivated species (Upadhyaya et al., 2011). Although progress has been made through conventional breeding, yet the confusing species barrier between cultivated peanuts and wild species makes the utilization of genetic resources very difficult. The lack of a robust phylogeny of the *Arachis* genus has impeded the advances in basic biological research and molecular breeding of the cultivated peanuts.

Allotetraploidy, which are evident in soybean, *Brassica*, wheat, cotton, and peanut via whole chromosomal genome (Gill et al., 2009; Feldman et al., 2012; Paterson et al., 2012; Chalhoub et al., 2014; Bertioli et al., 2019; Zhuang et al., 2019), plays a critical role in the evolving history of most domesticated crop species. However, how allotetraploids species (e.g., cultivated peanut) have evolved from their diploid parents remains largely unknown (Bertioli et al., 2020; Zhuang et al., 2020). The lack of information is caused by two possible reasons: (1) morphological and molecular phylogenetic studies are not efficient in distinguishing taxonomic species for some horticulture features may have resulted from domestication. (2) Genetic diversity introduced by multiple parental inheritance makes it difficult to detect homology among sequences. According to a few previous studies, cultivated peanuts are allotetraploid (AABB genome type) and derived from two diploids wild species by a recent hybridization event (Bertioli et al., 2019; Zhuang et al., 2019). Many studies suggest that *A. duranensis* Krapov. & W.C.Greg. (AA) and *Arachis ipaensis* Krapov. & W.C.Greg. (BB) are the progenitor species, which provide valuable genetic resources to *A. hypogaea* (Kochert et al., 1996; Koppolu et al., 2010; Bertioli et al., 2011, 2016). However, some other studies support that cultivated peanuts may have been derived from more than two progenitor species, including *Arachis diogoi* Hoehne (AA), *Arachis correntina* (Burkart) Krapov. & W.C.Greg. (AA), *Arachis cardenasii* Krapov. & W.C.Greg. (AA), *A. batizocoi* Krapov. & W.C.Greg. (KK), *A. trinitensis* Krapov. & W.C.Greg. (FF), and *A. williamsii* Krapov. & W.C.Greg. (BB) (Stalker et al., 1991; Singh et al., 1994; Leal-Bertioli et al., 2014; Wang et al., 2019; Zhuang et al., 2019). The origination and evolution of the cultivated peanut species remains elusive, and it is extremely difficult to demarcate the boundary of some peanut species due to gene introgression, ancestral polymorphism and various speciation rates in different species (Moretzsohn et al., 2013; Bertioli et al., 2019).

The previous classification has put cultivated peanuts into two groups, subsp. *hypogaea* and subsp. *fastigiata*, based on some morphological and physiological characteristics, such as the presence of flower on main stem, time of maturation, the presence of seed dormancy, *etc.* (Gibbons et al., 1972; Krapovickas et al., 2007; Belamkar et al., 2011). According to some early classification work, which studied the growth habit, leaflet surface, branching pattern and pod traits of various peanuts (Ferguson et al., 2004; Krapovickas et al., 2007), subsp. *hypogaea* contain two botanical varieties, var. *hypogaea*, var. *hirsute*, while four varieties (var. *fastigiata*, var. *peruviana*, var. *vulgaris*, and var. *aequatoriana*) are present in subsp. *fastigiata*. However, classification based on morphological and physiological characteristics is not consistently supported by works done at the molecular level when employing different methods or using different genetic markers (He and Prakash, 2001; Gimenes et al., 2002; Moretzsohn et al., 2004; Koppolu et al., 2010). A molecular analysis using the AFLP approach shows that var. *aequatoriana* and var. *peruviana* are closely related to subsp. *hypogaea* (He and Prakash, 2001). Furthermore, a study carried out with SSRs markers put var. *peruviana* into subsp. *hypogaea*. More interestingly, var. *hypogaea* and var. *hirsute*, which are originally placed in subsp. *hypogaea*, are not even closely related according to Ferguson et al. (2004). The conventional classification of cultivated peanuts is supported by one recent study, which looked at high-quality SNPs in the peanut nuclear genomes (Zheng et al., 2018). However, the taxonomic boundaries among some botanical varieties cannot be clearly delimited in this study. Var. *hypogaea* and var. *hirsute* could not be distinguished due to difficulties in putting different accessions of the same variety into one cluster. A close evolutionary relationship was inferred between var. *hirsute* and var. *vulgaris* when using the plastomics approach (Wang et al., 2018, 2019). This study also supports a close relationship between var. *hypogaea* and var. *fastigiata*, which is different from what we would expect based on the previous classification. It seems that nuclear genomic sequence data is not sufficient or reliable in interpreting evolutionary relationship among allotetraploid species. Due to the lack of consistency, a study carried out with a different type of sequence data (i.e., plastomic data) or employing various analytic methods would be appropriate when trying to reconstruct the phylogeny of cultivated peanuts.

Plastomics provide a powerful tool in phylogenetic studies involving particular evolutionary events, such as interspecific hybridization, allopolyploidization, rapid evolution, *etc.* (Moore et al., 2007). In contrast to nuclear genomes, plastomes are maternally inherited. The evolutionary rate of plastomes is low, and there is no recombination during chloroplast division (Daniell et al., 2016). Therefore, plastomes are good resources for studying maternal evolutionary dynamics (Tonti-Filippini et al., 2017). Chloroplast genomes are highly conserved in angiosperms, which share a quadripartite structure containing a large single copy (LSC; 80–90 kb) and a small single copy (SSC; 16–27 kb) separated by two inverted repeats (IR; 20–28 kb) (Daniell et al., 2016). In green plants, plastomes typically range from 120 to 218 kb in size (Wicke et al., 2011), and such a variety in size is mainly caused by IR contraction and expansion (Choi et al., 2020; Henriquez et al., 2020). To take an extreme example, the IR region is completely lost in *Erodium* L'Herit. and some papilionoid legumes (Blazier et al., 2016; Lee et al., 2021). Angiosperm plastomes generally encode 110–130 genes, which

include approximately 80 protein coding genes, 30 transfer RNA genes, and four ribosomal RNA genes (Daniell et al., 2016). Even though the loss of genes (Song et al., 2017; Alqahtani and Jansen, 2021) or introns (Jansen et al., 2007), and pseudogenization (Abdullah et al., 2021a; Li et al., 2021) have been reported in the plastomes of diverse plant species, plastomics still provide a reliable tool in phylogenetic studies, and plastid genomes have been largely used to reconstruct the phylogeny of many crop and horticulture species in recent years (Li et al., 2017; Xue et al., 2019; Guo et al., 2020; Hassoubah et al., 2020; Moner et al., 2020; Tyagi et al., 2020). However, there are only a limited number of peanut plastomes that have been sequenced and analyzed to date, including that of *A. hypogaea* and a few other related wild species (Prabhudas et al., 2016; Yin et al., 2017; Wang et al., 2018, 2019, 2021). This is insufficient in gaining a full picture of what has happened in the evolutionary history of cultivated peanuts and some wild species, and the relationship between cultivated peanuts and their potential wild maternal progenitor species is still unclear.

In this study, we assembled 33 *Arachis* plastomes including both cultivated and wild peanut species. Through comparative analysis with other peanut plastomes, which are currently available at NCBI, we aim to provide insights into species delimitation of *Arachis* and to identify the potential maternal genome progenitor species of cultivated peanuts. This work will serve as a foundation for the utilization of peanut genetic resources and the development of high-quality peanut varieties through molecular breeding.

MATERIALS AND METHODS

Plant Sampling

In this study, Fresh young leave samples of 33 peanut accessions (24 species) representing 11 different genome types were collected from Henan Academy of Agricultural Sciences, Zhengzhou, China (HNAAS) and used for further analysis (**Table 1**). These include five botanical varieties of *A. hypogaea*, var. *hypogae* (Lainongzao), var. *hirsute* (Bajisitanhuapi), var. *fastigiate* (PI493938), var. *peruviana* (NcAc17090), var. *vulgaris* (Yiya). Samples were stored immediately in a −80°C freezer prior to DNA extraction. All the voucher specimens were deposited to the Herbarium of Zhengzhou University (**Supplementary Table 1**).

Genomic DNA Extraction and Sequencing

Total genomic DNA of the 33 samples were extracted with the Tiangen Plant Genomic DNA Kit (Tiangen Inc., China) following the protocol provided by the manufacturer. DNA purity was assessed using the Qubit 2.0 (Invitrogen Inc., United States) and a NanoDrop machine (Thermo Scientific Inc., United States). DNA libraries were constructed using the Illumina Paired-End DNA library Kit and sequenced with a NovaSeq 6000 platform (Illumina Inc., United States) with a paired-end read length of 150 bp (NovoGene Inc., China). Upon completion, more than 6.0 GB raw reads were retrieved for each sample. The GetOrganelle toolkit was used for *de novo* assembling of the complete plastid genomes (Jin et al., 2020). The published plastomic sequences of *Arachis* (**Supplementary Table 1**) from GenBank were used as the seed file ("embplant_pt") for the assembling process, as well as a template to estimate the possible circular sequence pattern.

Plastome Annotation and Comparison

The Plastid Genome Annotator (PGA) software (Qu et al., 2019) was employed in the annotation of the selected peanut plastomes using *A. hypogaea* (accession no. MT712165) as a reference. The accuracy of annotation was evaluated with GeSeq (Tillich et al., 2017), HMMER (Wheeler and Eddy, 2013), and tRNAscan-SE (Lowe and Eddy, 1997) programs implemented in the CHLOROBOX web toolbox[1] with a default setting. Chloroplot was used to visualize the plastid genomes as a physical map (Zheng et al., 2020). MISA-web (Beier et al., 2017) was used to identify simple sequence repeats (SSRs) with the following criteria: 10, 5, 4, 3, 3, and 3 repeat units are for mono-, di-, tri-, tetra-, penta-, and hexa-nucleotides, respectively. In addition, forward, palindrome, reverse, and complement repeated elements were identified using REPuter (Kurtz et al., 2001) with a minimal length of 30 bp, an identity value of more than 90% and a Hamming distance of 3. The comparison among whole chloroplast genomes in genus *Arachis* species were using data from 33 new sequenced plastomes, and published plastomes of five cultivated peanuts (Prabhudas et al., 2016; Wang et al., 2018), 12 wild peanuts (Wang et al., 2019) which downloaded from the NCBI database (**Supplementary Table 1**). Nucleotide diversity (Pi) of the plastomic sequences of *Arachis* species were obtained in this study and the published sequences were calculated using a sliding window method with a window length of 600 bp and a step size of 200 bp by DnaSP (Rozas et al., 2017).

Phylogenetic Analysis

To reconstruct the phylogeny of peanut species and to identify the potential maternal progenitor species, the complete plastomes of 53 species (**Supplementary Table 1**) were retrieved from various databases and used to make a multiple sequence alignment with MAFFT under a default setting (Katoh and Standley, 2013). Among these species, *Dalbergia hupeana* Hance from the Tribe Dalbergieae was defined as outgroup. Phylogenetic trees were constructed with the 53 sequences using both the Maximum likelihood (ML) and the Bayesian inference method (BI), which are implemented in IQ-TREE (Nguyen et al., 2014) and MrBayes (Ronquist et al., 2012), respectively. The best fit nucleotide substitution models, TVM + F + R3 for ML analysis and GTR + F + I + G4 for BI analysis, were selected using the ModelFinder (Kalyaanamoorthy et al., 2017) according to the AIC criterion. In the ML analysis, 50,000 bootstrap replicates were carried out with the SH-aLRT branch test. The BI analysis was performed with two independent Markov Chain Monte Carlo chains with 2,000,000 generations, and it was considered to be stationary when the average standard deviation of split frequencies fell below 0.01. The first 25% of trees were discarded

[1]https://chlorobox.mpimp-golm.mpg.de/geseq.html

TABLE 1 | Complete plastome features of the 33 *Arachis* accessions.

Species	Strains Information	Section	Plastome Size (bp)	IR (bp)	LSC (bp)	SSC (bp)	Number of genes (PCGs/tRNA/rRNA)	GC content (%; IR/LSC/SSC)
A. batizocoi	PI 298639	Arachis	156,340	25,781	85,846	18,932	109 (76/29/4)	36.4 (42.9/33.8/30.2)
A. cardenasii	PI 475996	Arachis	156,394	25,824	85,946	18,800	109 (76/29/4)	36.4 (42.9/33.8/30.2)
A. cardenasii	PI 476014	Arachis	156,410	25,825	85,958	18,802	109 (76/29/4)	36.4 (42.9/33.8/30.2)
A. cruziana	PI 476003	Arachis	156,364	25,785	85,851	18,943	109 (76/29/4)	36.4 (42.9/33.8/30.2)
A. decora	Grif 7721	Arachis	156,247	25,757	85,739	18,994	109 (76/29/4)	36.4 (42.9/33.8/30.2)
A. diogoi	PI 276235	Arachis	156,377	25,824	85,933	18,796	109 (76/29/4)	36.4 (42.9/33.8/30.2)
A. duranensis	PI 475844	Arachis	156,392	25,824	85,948	18,796	109 (76/29/4)	36.4 (42.9/33.8/30.2)
A. duranensis	PI 468200	Arachis	156,424	25,824	85,964	18,812	109 (76/29/4)	36.4 (42.9/33.8/30.2)
A. duranensis	PI 468323	Arachis	156,433	25,824	85,968	18,817	109 (76/29/4)	36.4 (42.9/33.8/30.2)
A. duranensis	PI 219823	Arachis	156,383	25,825	85,937	18,796	109 (76/29/4)	36.4 (42.9/33.8/30.2)
A. glandulifera	PI 468336	Arachis	156,363	25,774	85,870	18,945	109 (76/29/4)	36.4 (42.9/33.8/30.2)
A. herzogii	PI 476008	Arachis	156,420	25,825	85,962	18,808	109 (76/29/4)	36.4 (42.9/33.8/30.2)
A. hoehnei	Grif 7682	Arachis	156,379	25,824	85,942	18,789	109 (76/29/4)	36.4 (42.9/33.8/30.2)
A. hypogaea var. *fastigiata*	PI493938	Arachis	156,384	25,825	85,938	18,796	109 (76/29/4)	36.4 (42.9/33.8/30.2)
A. hypogaea var. *hirsuta*	Bajisitanhuapi	Arachis	156,387	25,825	85,942	18,795	109 (76/29/4)	36.4 (42.9/33.8/30.2)
A. hypogaea var. *hypogae*	Lainongzao	Arachis	156,387	25,825	85,942	18,795	109 (76/29/4)	36.4 (42.9/33.8/30.2)
A. hypogaea var. *peruviana*	NcAc17090	Arachis	156,387	25,825	85,942	18,795	109 (76/29/4)	36.4 (42.9/33.8/30.2)
A. hypogaea var. *vulgaris*	Yiya	Arachis	156,384	25,825	85,938	18,796	109 (76/29/4)	36.4 (42.9/33.8/30.2)
A. ipaensis	–	Arachis	156,394	25,776	85,904	18,938	109 (76/29/4)	36.4 (42.9/33.8/30.2)
A. kempff-mercadoi	PI 468330	Arachis	156,429	25,824	85,965	18,816	109 (76/29/4)	36.4 (42.9/33.8/30.2)
A. microsperma	PI 674407	Arachis	156,326	25,787	85,939	18,813	109 (76/29/4)	36.4 (42.9/33.8/30.2)
A. monticola	PI 263393	Arachis	156,388	25,824	85,945	18,795	109 (76/29/4)	36.4 (42.9/33.8/30.2)
A. monticola	PI 219824	Arachis	156,388	25,824	85,945	18,795	109 (76/29/4)	36.4 (42.9/33.8/30.2)
A. palustris	PI 666093	Arachis	156,220	25,827	85,750	18,907	109 (76/29/4)	36.4 (42.9/33.8/30.2)
A. simpsonii	Grif 14534	Arachis	156,385	25,824	85,941	18,796	109 (76/29/4)	36.4 (42.9/33.8/30.2)
A. trinitensis	PI 666101	Arachis	156,354	25,776	85,860	18,941	109 (76/29/4)	36.4 (42.9/33.8/30.2)
A. valida	PI 666103	Arachis	156,369	25,774	85,877	18,944	109 (76/29/4)	36.4 (42.9/33.8/30.2)
A. villosa	PI 298636	Arachis	156,464	25,824	85,958	18,858	109 (76/29/4)	36.4 (42.9/33.8/30.2)
A. pintoi	–	Caulorrhizae	156,311	25,757	85,736	18,970	109 (76/29/4)	36.4 (42.9/33.9/30.3)
A. dardonoi	–	Heteranthae	156,630	25,857	85,990	18,926	109 (76/29/4)	36.3 (42.9/33.8/30.2)
A. pusilla	PI 497572	Heteranthae	156,476	25,862	85,889	18,863	109 (76/29/4)	36.3 (42.9/33.8/30.2)
A. rigonii	PI 262142	Procumbentes	156,476	25,862	85,889	18,863	109 (76/29/4)	36.3 (42.9/33.8/30.2)
A. glabrata	PI 468366	Rhizomatosae	156,428	25,824	85,969	18,811	109 (76/29/4)	36.4 (42.9/33.8/30.2)

as burn-ins, and the remaining trees were used to construct a consensus tree.

RESULTS

Characterization of the Peanut Plastomes

The size of the studied *Arachis* plastomes ranges from 156,220 bp (*Arachis palustris*) to 156,630 bp (*Arachis dardonoi*) in length (**Table 1** and **Figure 1**), while the *Arachis* species ranging from 156,220 bp (*A. palustris*) to 156,878 bp (*A. hypogaea* var. *hirsute* AHL) in whole chloroplast genome length. Moreover, the cultivated peanut plastomes ranges from 156,354 to 156,878 bp in length, with *A. hypogaea* var. *hirsute* AHL being the largest. There is a 10–100 bp difference in length when comparing our sequencing data with the published data of a few species including *A. batizocoi*, *A. cardenasii*, *A. duranensis*, *A. ipaensis*, *Arachis villosa,* and *A. monticola* PI 219824. All the sequenced plastomes share a G + C content of 36.4% except for *A. dardonoi*, *A. pusilla*, and *A. rigonii*, which share a G + C content of 36.3%. All peanut plastomes contain a large single-copy (LSC), a small single-copy (SSC), and two inverted repeats (IRa/IRb). The LSC regions range from 85,736 bp (*A. pintoi*) to 85,990 bp (*A. dardonoi*) in length, with the G + C contents falling between 33.8 and 33.9%. The SSC regions vary from 18,789 bp (*Arachis hoehnei*) to 18,994 bp (*Arachis decora*) in length and the G + C content falls between 30.2 and 30.3%. *A. pintoi* has the smallest IRs, which is 25,757 bp in length, while a maximum IR length of 25,862 bp was observed in both *A. pusilla* and *A. rigonii*. These regions have a G + C content of 42.9%, which is significantly higher than that of the LSCs and SSCs. All plastomes included in our studies contain 109 unique genes, encoding 76 protein genes, 29 tRNAs, and 4 rRNAs (**Table 1** and **Figure 1**), which is comparable with some well-studied *Arachis* species. Based on their annotated functions, these genes can be classified into four categories (**Table 2**), namely

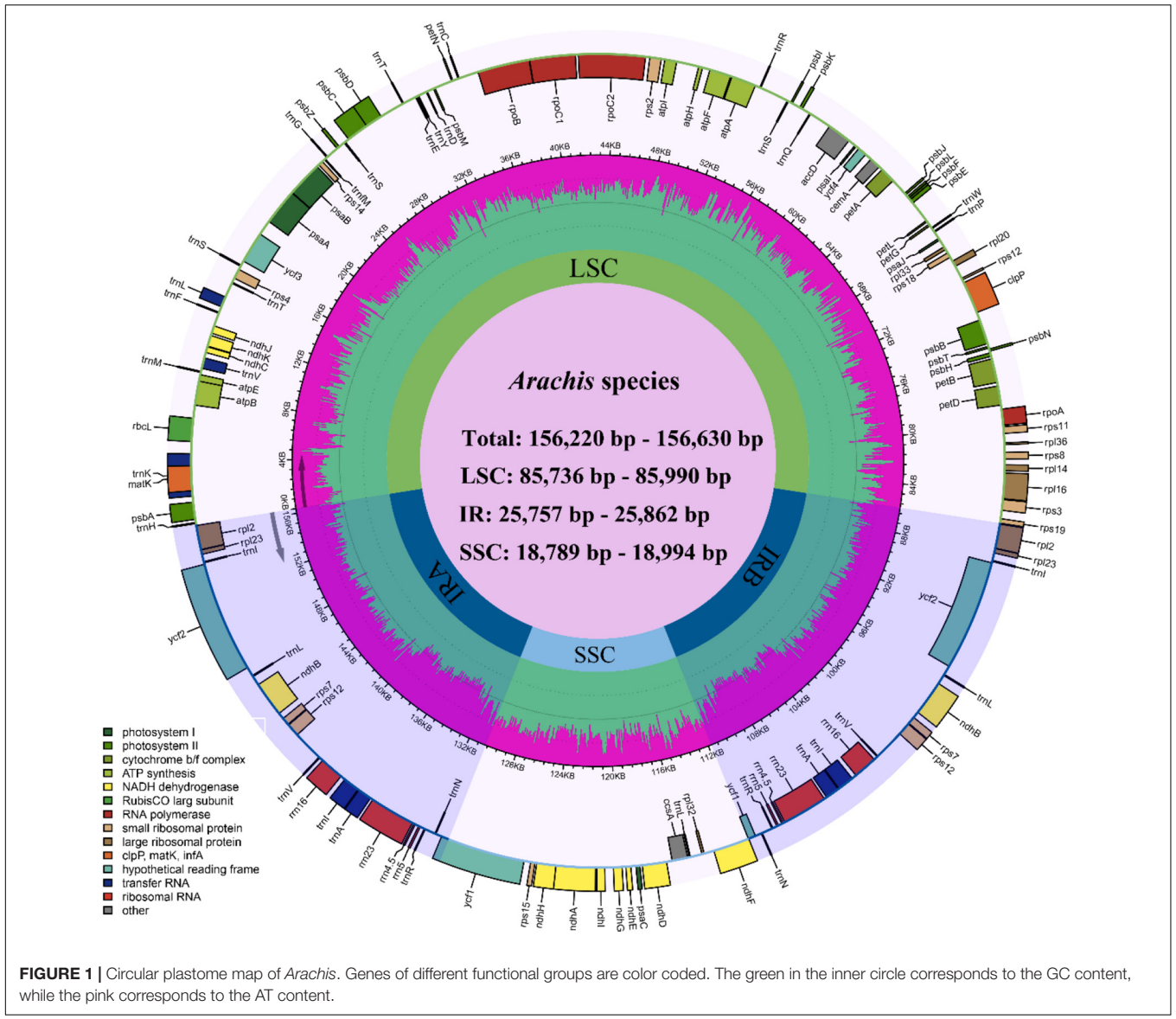

FIGURE 1 | Circular plastome map of *Arachis*. Genes of different functional groups are color coded. The green in the inner circle corresponds to the GC content, while the pink corresponds to the AT content.

self-replication genes, photosynthesis related genes, other genes, and unknown function genes.

Comparative Plastomic Analysis

Analysis of the 33 new sequenced plastomes revealed 1,593 tandem repeats with complement, forward, reverse, and palindromic elements (>30 bp). The number of repeats present in each plastome varies considerably, ranging from 38 in *A. decora* to 50 in most other species (**Figure 2A**). In average, 17 forward, 27 palindromic, 2 complement, and 3 reverse repeats were estimated in each plastome. Among the species, in which repeats were identified, *A. dardonoi* lacks complement repeats, while *A. pintoi*, *A. pusilla*, and *A. rigonii* do not have complement or reverse repeats. Most repeats among *Arachis* species plastomes are present in the intergenic spacer regions. With the MISA analysis, 60 universal SSR loci were detected in the plastomes of *A. pusilla* and *A. rigonii* while 83 was in *A. dardonoi* (**Figure 2B**). Based on the SSR analysis, 40–57 of the identified SSRs are mononucleotidic, 14–20 are dinucleotidic, 1–4 are trinucleotidic, and 5–9 are tetranucleotidic (**Supplementary Table 2**). Among these SSRs, most of the identified mononucleotidic SSRs are composed of A/T, and the dinucleotidic ones contain AT/TA. Moreover, the pentanucleotidic SSRs in *A. dardonoi* and *A. ipaensis* have a typical sequence of AATAG/CTATT or TATAA/TTATA, and the hexanucleotidic SSRs in *A. cardenasii*, *A. dardonoi*, *A. duranensis*, *A. glabrata*, *Arachis herzogii*, and *Arachis microsperma* contain either AATGGA/TCCATT or ATAGCA/TGCTAT (**Figure 2B**).

A total number of 3,416 polymorphic sites (Pi: 0.227%) were detected in the 52 cultivated and wild peanut plastomes (**Supplementary Table 1**), including 1,670 singleton variable sites and 1,746 parsimony informative sites. The alignment of seventeen peanut complex (see discussion) sharing high sequence similarity revels 54 singleton variable sites and 20 parsimony

TABLE 2 | List of the annotated genes in the plastomes of the selected *Arachis* accessions.

Category	Gene group	Gene name
Photosynthesis related genes	Rubisco	rbcL
	Photosystem I	psaA, psaB, psaC, psaI, psaJ
	Assembly/stability of photosystem I	ycf3
	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
	Cytochrome c synthesis	ccsA
	NADPH dehydrogenase	ndhA, ndhB(×2), ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK
Transcription and translation related genes	Transcription	rpoA, rpoB, rpoC1, rpoC2
	Ribosomal proteins	rpl14, rpl16, rpl2(×2), rpl20, rpl23(×2), rpl32, rpl33, rpl36, rps11, rps12(×2), rps14, rps15, rps18, rps19, rps2, rps3, rps3, rps4, rps7(×2), rps8
RNA genes	Ribosomal RNA	rrn4.5(×2), rrn5(×2), rrn16(×2), rrn23(×2)
	Transfer RNA	trnA-UGC(×2), trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnfM-CAU, trnG-UCC, trnH-GUG, trnI-CAU(×2), trnI-GAU(×2), trnK-UUU, trnL-CAA(×2), trnL-UAA, trnL-UAG, trnM-CAU, trnN-GUU(×2), trnP-UGG, trnQ-UUG, trnR-ACG(×2), trnR-UCU, trnS-GCU, trnS-GGA, trnS-UGA, trnT-GGU, trnT-UGU, trnV-GAC(×2), trnV-UAC, trnW-CCA, trnY-GUA
Other genes	RNA processing	matK
	Carbon metabolism	cemA
	Fatty acid synthesis	accD
	Proteolysis	clpP
Genes of unknown function	Conserved reading frames	ycf1(×2), ycf2(×2), ycf4

×2 denotes two gene copies in the IR region.

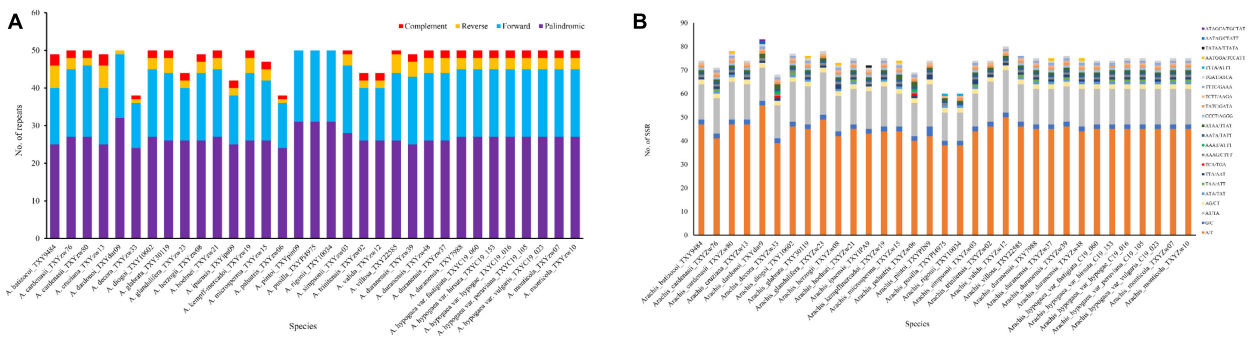

FIGURE 2 | Analyses of repeated sequences in the plastomes of the 33 *Arachis* accessions. **(A)** Numbers of repeats and their types; **(B)** number of identified SSRs motifs and their types.

informative sites, which are also highly conserved across all the analyzed plastomes. Pi values among different plastomes were computed by a sliding window method with a window length of 600 bp and a step size of 200 bp (**Figure 3**). In addition, six hotspot regions with high Pi values were identified among various cultivated peanut accessions and other wild species, which include two protein-coding genes (*rpoC2* and *ycf1*) and four intergenic spacer regions (*trnS-UGA-psbC*, *atpA-trnR-UCU*, *psbE-petL*, and *rpl32-trnL-UAG*). These regions could be potentially used as DNA markers in phylogenetic studies of different *Arachis* species.

Phylogeny of *Arachis* Based on Whole Plastomes

In our study, *Arachis* is recovered as monophyletic, which is well supported by both ML and BI analyses (**Figures 4A,B**). The basal position of *A. dardonoi* (HH) from section *Heteranthae* is strongly supported by both methods. Within section *Arachis*, two major lineages, Lineage I and II, were clearly defined (**Figure 4**). Another species *A. pusilla* (HH) from section *Heteranthae* is grouped into one clade with *A. rigonii* (PR) from section *Procumbentes*. However, ML and BI analysis did not provide consistent result in terms of the taxonomic statuses of *A. duranensis*, *A. monticola*, and cultivated peanut (**Figures 4A,B**).

Species with a genome type of AA are mainly distributed in Lineage I, which are further divided into three clades. Based on the BI analysis, the newly sequenced peanut varieties, var. *hypogaea* and var. *hirsuta*, are clustered into one clade, while var. *fastigiata*, var. *vulgaris*, and var. *peruviana* are placed in another clade with relatively low bootstrap values (**Figure 4A**). In the ML analysis, the six varieties are placed in one big clade, and it is impossible to draw a clear boundary

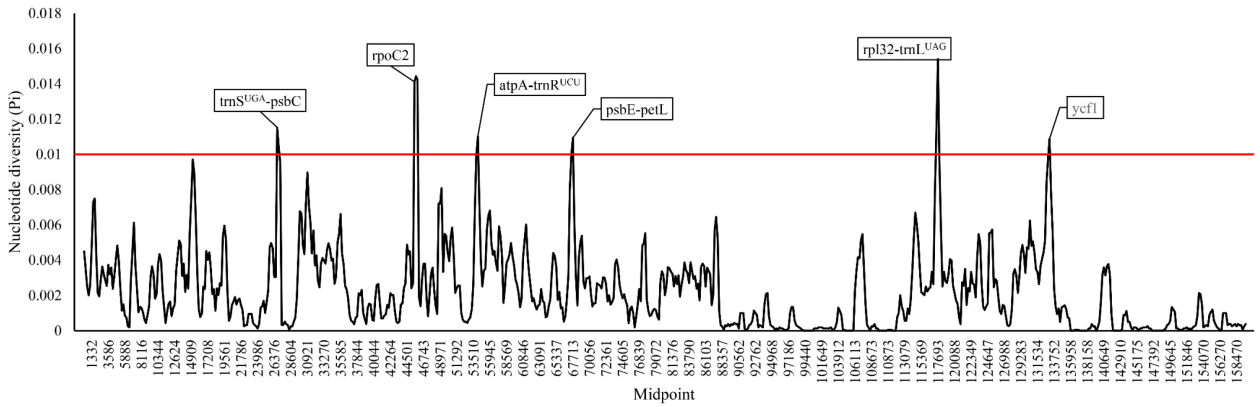

FIGURE 3 | Sliding window analysis of the whole plastomes of the 52 *Arachis* accessions.

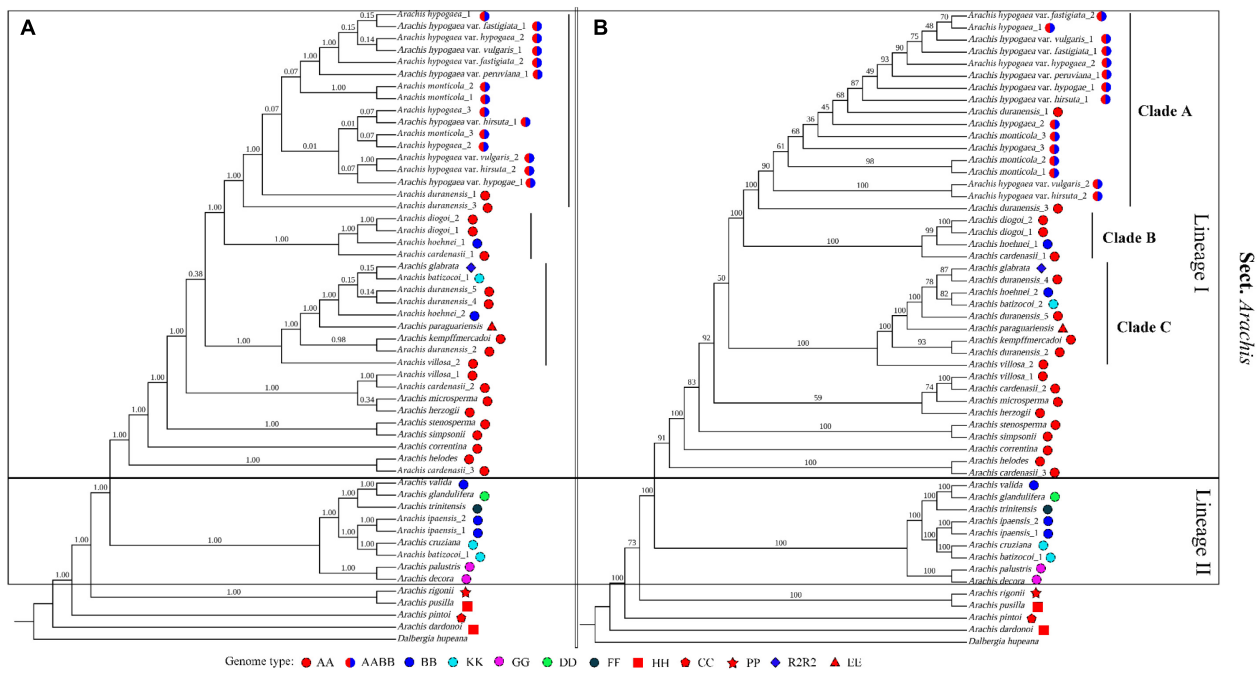

FIGURE 4 | Phylogenetic trees constructed with 52 *Arachis* plastomic sequences using the Bayesian inferred **(A)** and Maximum likelihood **(B)** methods. Numbers above the branch represent the confidence level.

between subsp. *hypogaea* and subsp. *fastigiated* (**Figure 4B**). Var. *vulgaris* (Yiya vs. AHZ) and var. *hirsute* (Bajisitanhuapi vs. AHL) are placed in two separate clades in this study. The cultivated peanuts and two wild species, *A. monticola* and *A. duranensis* (PI219823 and PI 475844), are grouped together as the "peanut complex" clade (Clade A), members of which demonstrate a diversity in morphological features, and the boundary between Clade A and other clades is not well defined or supported by the phylogenetic analyses (**Figure 5**). Both the ML and BI analyses support that *A. duranensis* (AA) is the wild diploid progenitor of all cultivated peanuts. In Lineage I, Clades B and C are not monophyletic, which contain species with various genome types, such as *A. hoehnei* (BB), *A. glabrata* (R2) from section *Rhizomatosae*, *A. batizocoi* (KK), and *Arachis paraguariensis* (EE) from section *Erectoides* (**Figure 4**).

The phylogenetic structure of Lineage II is strongly supported by both ML and BI analyses. It contains nine *Arachis* species/accessions with diverse genome types. For example, the two accessions of *A. ipaensis* and *Arachis valida* contain BB genome type. *A. decora* and *A. palustris* ($2n = 18$), share a genome type of GG and are placed together with high confidence scores. *A. valida* shows a sisterhood relationship with *Arachis trinitensis* (FF) and *Arachis glandulifera* (DD), while *A. ipaensis*, another possible diploid progenitor of cultivated peanuts, is grouped together with *A. batizocoi* (AA) and *Arachis cruziana* (KK).

FIGURE 5 | Morphological differences among selected *Arachis* species (three accessions of *A. duranensis*, two accessions of *A. cardenasii* and five cultivated peanuts).

DISCUSSION

Arachis Plastomes Are Highly Conserved

All *Arachis* plastomes share a typical quadripartite structure, consisting of one LSC region and one SSC regions separated by a pair of IRs. The same structure has also been reported in other angiosperms (Xu et al., 2015; Daniell et al., 2016; Tonti-Filippini et al., 2017). All the *Arachis* plastomes covered in this study are highly conserved in genome size and structure, G + C content, and gene number, which are also comparable to the plastomes of previously published *Arachis* species (Prabhudas et al., 2016; Yin et al., 2017; Wang et al., 2018, 2019). Plastomes of angiosperms tend to vary in size, the size of a typical *Arachis* plastome is approximately 156 kb (**Supplementary Table 1**), similar with the plastomes length of soybean (*Glycine*) in 152 kb, but more than the length of wheat (*Tribe Triticeae*), which varies from 133 to 137 kb (Middleton et al., 2014), rice (*Oryza*) of 135 kb in size (Asaf et al., 2017), and less than buckwheat (*Fagopyrum*) of 159 kb in total length (Wang et al., 2017). Genome size change was suggested to be linked variation of intergenic region, InDel events and oligonucleotide/microsatellites repeats within the related species, while gene loss, expansion/contraction of an IR region among seed plants (Xu et al., 2015; Zheng et al., 2017).

All the published *Arachis* plastomes share the same number of protein coding genes (**Table 2**) with only a few exceptions. Prabhudas et al. (2016) was not able to detect NADH dehydrogenase subunit 2 gene (*ndhB*) in *A. hypogaea* Co7, and *orf42* and *ycf68* were miss annotated in another two studies by Yin et al. (2017) and Wang et al. (2019). A closer look at the coding regions reveals that five tRNAs and 11 protein coding genes harbor at least one intron. Among these, *ycf3*, *clpP,* and *rps12* (a *trans*-splicing gene) contain two introns (Xu et al., 2015; Liu et al., 2020). The total number of tRNA genes present in our sequenced plastomes is 29, and the same conclusion was reached in two other studies by Schwarz et al. (2015) and Wang et al. (2019). However, one extra tRNA gene was annotated in one previous study carried out by Prabhudas et al. (2016). This one extra gene is *trnP*-GGG, which overlaps with another tRNA gene[2]. According to wild Roses, *trnP*-GGG gene in the region of *trnP*-UGG gene also exists (Jeon and Kim, 2019). Former studies demonstrated a widely distributed of *trnP*-GGG gene present in charophyte to gymnosperm, while *trnP*-UGG gene in plastomes from algae to higher plants (Turmel et al., 2002; Sugiura and Sugita, 2004).

Microsatellites and oligonucleotide repeats play an important role in the identification of regions with a large number of mutations, and are helpful in the study of population genetics (Ahmed et al., 2012; Abdullah et al., 2019). A consistent result was obtained when comparing SSRs and oligonucleotide repeats across different *Arachis* plastomes (Yin et al., 2017; Wang et al., 2019), with A/T and AT/TA being the most common mononucleotidic SSRs and mononucleotidic SSRs, respectively. A similar pattern is also reported in plastomes of many other angiosperms (Tian et al., 2019; Mehmood et al., 2020; Abdullah et al., 2021b). The SSRs loci identified in this work could serve as potential molecular markers for understanding the population genetic structure among various *Arachis* species. Here, we also identified some oligonucleotide repeats, which are associated with nucleotide substitution, mutation and InDel events in the genomes (Abdullah et al., 2021b,c). Most of the oligonucleotide repeats were found in the intergenic regions, and a similar pattern is observed in the plastomes of many other vascular

[2]http://www.ncbi.nlm.nih.gov/nuccore/KX257487

plants (Kuang et al., 2011; Li et al., 2017; Sigmon et al., 2017; Wang et al., 2019). Our results also showed a high abundance of complement and forward oligonucleotide repeats across different *Arachis* species. Oligonucleotide repeats could be used for the identification of regions with mutations and the reconstruction of accurate phylogeny of *Arachis* species (Mehmood et al., 2020; Abdullah et al., 2021b).

Linking Phylogeny With Genome Type

The genus *Arachis* consists of 81 species demonstrating a huge diversity in genome types (A, B, AB, C, D, E, EX, F, H, K, PR, R1, R2, T, and TE). Linking phylogenetic analysis with their genome type information could allow us to better understand the origination and evolution of cultivated peanuts. Based on our study, hybridization seems to play a major role in the evolution history of cultivated species (Garcia et al., 1995; Jarvis et al., 2003). However, problems within several clades are still unsolved. Our results show that the taxonomic relationship based on morphology should be revised (He et al., 2014; Vishwakarma et al., 2017). Two studies working with plastomics data (Wang et al., 2019) and microsatellite markers (Moretzsohn et al., 2013) also reached the same conclusion. In addition, one clade may contain species with various genome types, which is supported by this study and two other phylogenetic studies working with intron sequences and microsatellite markers (Moretzsohn et al., 2013). Again, it is very difficult to delimit the boundary of different *Arachis* species. In fact, all *Arachis* species look very similar morphologically, and leaf shape could probably be the only morphological trait, which could potentially be used in putting species into different taxonomic groups (**Supplementary Figure 1**). We speculate that recent speciation events play an important role in the evolution of *Arachis*. Both underground fruiting and clistogamy are thought to limit gene flows and seed dispersal in peanuts (Tan et al., 2010; Zhang et al., 2017), which should allow each species to keep its distinct identity (Yu et al., 2020). However, it is very interesting to see that the flowers and stems of *Arachis* plant could attract small insects, such as ants (**Supplementary Figure 2**). The movement of ants between different plants could cause the pollen of one species to be transferred to another species, and therefore promote gene flow between different *Arachis* species. In fact, genome introgression was detected among the interspecific hybrid population of peanuts (Garcia et al., 1995).

Plastomes are highly conserved and tend to have low nucleotide variations (Sigmon et al., 2017; Wang et al., 2018; Nock et al., 2019) (**Figure 4**). In this study, only 74 nucleotide polymorphisms were detected among different species of the cultivated peanut complex, indicating that the plastomes of cultivated peanuts are highly conserved (Wang et al., 2018). This observation could also be explained with a low nucleotide substitution rate. Peanut has only been domesticated for several thousand years, there is not enough time to accumulate many genetic variations (Bertioli et al., 2019). Although most botanical varieties examined in this study do demonstrate differences in their morphology (**Figure 5**), there are no distinguishable morphological features, which could be used to put different species into the two subspecies groups. For example, var. *fastigiate*, var. *vulgaris* and var. *hirsute* coming from two different groups all have three or more seeds in each shell (**Figure 4A**). The overall phylogeny obtained in this study is in agreement with the conventional classification based on studies looking at other features, including morphology (Krapovickas et al., 2007), AFLP markers (He and Prakash, 2001), simple sequence repeats (Ferguson et al., 2004), and single nucleotide polymorphisms (Zheng et al., 2018). However, violations do exist when it comes to the phylogenetic relationship of different varieties, such as, var. *peruviana* does not belong to subsp. *fastigiata* (He and Prakash, 2001; Ferguson et al., 2004). Var. *hypogaea* and var. *hirsute* should not be placed in subsp. *hypogaea* according to the conventional classification.

Maternal Hybridization Event in the History of Cultivated Peanuts

Our results strongly support the hypothesis that *A. duranensis* is the wild diploid progenitor (with a genome type of A) of cultivated peanuts (**Figure 4**). This result is compatible with the earlier view, which is based on multiple lines of evidence from comparative genomics, geographic distribution, phylogenetic reconstruction, *etc.* (Kochert et al., 1996; Seijo et al., 2004; Fávero et al., 2006; da Cunha et al., 2008; Bertioli et al., 2016; Chen et al., 2016; Wang et al., 2019). Furthermore, phylogenomic investigation using both ML and BI methods suggests that *A. duranensis* have diverged into two groups. *A. duranensis* (PI219823 and PI 475844) shows a closer relationship with *A. hypogaea*, while *A. duranensis* PI 468200, PI 468323, and PI263133 (Genbank no. MK144822) are grouped in another clade containing *A. batizocoi*, *A. glabrata*, *A. hoehnei*, *Arachis kempff-mercadoi*, and *A. paraguariensis* (**Figure 4**). This topology was generally consistent with that of the ML tree, in which the three botanical accessions of *A. duranensis* (ICG 8138, ICG 8123, and PI 262133) are distributed in different clades (Zhuang et al., 2019). Moreover, the 42 accessions of *A. duranensis* demonstrate clear variations in morphological features (Singh et al., 1996). In agreement with Bertioli's work (Bertioli et al., 2019), some accession of *A. duranensis* may have served as the AA sub-genome maternal progenitor of *A. hypogaea*. However, the status of *A. diogoi* (former known as *Arachis chacoensis*) and *A. cardenasii* as another two potential progenitors is not supported by our study.

This does not contrary to the earlier view that *A. monticola* is the direct progenitor of cultivated peanuts, and that it plays a vital role in the transition of diploid wild species to tetraploid cultivated species (Simpson et al., 2001; Yin et al., 2020). Cultivated peanut (*A. hypogaea*) and wild *A. monticola* are allotetraploids (AABB), while other 30 described wild species are diploid (Stalker, 2017). The previous phylogeographical analyses often group these two species (*A. hypogaea* and *A. monticola*) together (Gimenes et al., 2002; Seijo et al., 2004). As former documented, *A. monticola* is a weedy subspecies of cultivated peanuts, and it is placed in one group with *A. hypogaea* in earlier phylogenetic studies (Koppolu et al., 2010; Stalker, 2017; Vishwakarma et al., 2017; Wang et al., 2019). Their close relationship can be further supported with the following

evidence. Firstly, *A. hypogaea* is able to produce fertile hybrids when hybridized with *A. monticola* (Stalker and Moss, 1987). Secondly, this is in agreement with the results of previous studies focusing on somatic chromosomes, such as the virtually identical centromeric bands and *in situ* hybridization between *A. hypogaea* and *A. monticola* (Raina and Mukai, 1999). Thirdly, *A. monticola* may have been derived from a more ancient hybridization event according to the phylogenetic studies on the two *FAD2A* alleles, while the accessions of *A. hypogaea* may have evolved latter (Jung et al., 2003). During its evolution, *A. monticola* has accumulated more mutations in its plastome than most other cultivated peanuts do, which could be possibly traced back to different evolution rates or natural selection. Nevertheless, plastomics approach is very useful in inferring the maternal origin of cultivated peanuts and explaining the close phylogenetic relationship between *A. monticola* and *A. hypogaea*.

CONCLUSION

In summary, 33 *Arachis* plastomes were sequenced and analyzed in a comparative framework with the published plastomics data of cultivated and wild peanut species. These plastomes share similar structural organization with low nucleotide variations. The phylogenetic topology obtained in this study shows that plastomics could facilitate a better understanding of the phylogeny among deep lineages of *Arachis*. Based on our result, it is speculated that cultivated peanuts have experienced a multi-maternal hybridization event with a recent origin. Some wild species of the *A. duranensis* accessions might have contributed the maternal sub genomes to cultivated peanuts and *A. monticol*a, which represents a transitional species between wild diploid species and tetraploid cultivated species. Owing to interspecific gene flow and recent speciation, the relationship among different *Arachis* species inferred based on phylogeny do not always go along with their genome types. As a result, more *Arachis* species with various genome types should be included in future study to fully elucidate the origin and evolutionary history of *Arachis*.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

ZW, XZ, and BH conceived the ideas. PD and LF contributed to the sampling. XT and LS performed the experiments. XT and YW analyzed the data. The manuscript was written and improved by XT, LS, JG, ZW, XZ, and BH. All authors contributed to the article and approved the submitted version.

FUNDING

This research was financially supported by the China Postdoctoral Science Foundation (2020M672264), China Agricultural Research System (CARS-13), and Special Funds for Scientific and Technological Development from Henan Academy of Agricultural Sciences (2020CY07).

ACKNOWLEDGMENTS

We are grateful to Yongsheng Chen from Peking University for his critical review of the manuscript and Ziqi Sun from Henan Academy of Agricultural Sciences for providing the cultivated peanut accessions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.804568/full#supplementary-material

REFERENCES

Abdullah, Mehmood, F., Heidari, P., Rahim, A., Ahmed, I., and Poczai, P. (2021a). Pseudogenization of the chloroplast threonine (trnT-GGU) gene in the sunflower family (Asteraceae). *Sci. Rep.* 11:21122. doi: 10.1038/s41598-021-00510-4

Abdullah, Mehmood, F., Rahim, A., Heidari, P., Ahmed, I., and Poczai, P. (2021b). Comparative plastome analysis of Blumea, with implications for genome evolution and phylogeny of Asteroideae. *Ecol. Evol.* 11, 7810–7826. doi: 10.1002/ece3.7614

Abdullah, Mehmood, F., Shahzadi, I., Ali, Z., Islam, M., Naeem, M., et al. (2021c). Correlations among oligonucleotide repeats, nucleotide substitutions, and insertion–deletion mutations in chloroplast genomes of plant family Malvaceae. *J. Syst. Evol.* 59, 388–402. doi: 10.1111/jse.12585

Abdullah, Shahzadi, I., Mehmood, F., Ali, Z., Malik, M. S., Waseem, S., et al. (2019). Comparative analyses of chloroplast genomes among three Firmiana species: identification of mutational hotspots and phylogenetic relationship with other species of Malvaceae. *Plant Gene* 19:100199. doi: 10.1016/j.plgene.2019.100199

Ahmed, I., Biggs, P. J., Matthews, P. J., Collins, L. J., Hendy, M. D., and Lockhart, P. J. (2012). Mutational Dynamics of Aroid Chloroplast Genomes. *Genome Biol. Evol.* 4, 1316–1323. doi: 10.1093/gbe/evs110

Alqahtani, A. A., and Jansen, R. K. (2021). The evolutionary fate of rpl32 and rps16 losses in the Euphorbia schimperi (Euphorbiaceae) plastome. *Sci. Rep.* 11:7466. doi: 10.1038/s41598-021-86820-z

Asaf, S., Waqas, M., Khan, A. L., Khan, M. A., Kang, S.-M., Imran, Q. M., et al. (2017). The Complete Chloroplast Genome of Wild Rice (Oryza minuta) and Its Comparison to Related Species. *Front. Plant Sci.* 8:304. doi: 10.3389/fpls.2017.00304

Beier, S., Thiel, T., Münch, T., Scholz, U., and Mascher, M. (2017). MISA-web: a web server for microsatellite prediction. *Bioinformatics* 33, 2583–2585. doi: 10.1093/bioinformatics/btx198

Belamkar, V., Selvaraj, M. G., Ayers, J. L., Payton, P. R., Puppala, N., and Burow, M. D. (2011). A first insight into population structure and linkage

disequilibrium in the US peanut minicore collection. *Genetica* 139:411. doi: 10.1007/s10709-011-9556-2

Bertioli, D. J., Abernathy, B., Seijo, G., Clevenger, J., and Cannon, S. B. (2020). Evaluating two different models of peanut's origin. *Nat. Genet.* 52, 557–559. doi: 10.1038/s41588-020-0626-1

Bertioli, D. J., Cannon, S. B., Froenicke, L., Huang, G., Farmer, A. D., Cannon, E. K. S., et al. (2016). The genome sequences of Arachis duranensis and Arachis ipaensis, the diploid ancestors of cultivated peanut. *Nat. Genet.* 48, 438–446. doi: 10.1038/ng.3517

Bertioli, D. J., Jenkins, J., Clevenger, J., Dudchenko, O., Gao, D., Seijo, G., et al. (2019). The genome sequence of segmental allotetraploid peanut Arachis hypogaea. *Nat. Genet.* 51, 877–884. doi: 10.1038/s41588-019-0405-z

Bertioli, D. J., Seijo, G., Freitas, F. O., Valls, J. F. M., Leal-Bertioli, S. C. M., and Moretzsohn, M. C. (2011). An overview of peanut and its wild relatives. *Plant Genet. Resour.* 9, 134–149. doi: 10.1017/s1479262110000444

Blazier, J. C., Jansen, R. K., Mower, J. P., Govindu, M., Zhang, J., Weng, M.-L., et al. (2016). Variable presence of the inverted repeat and plastome stability in Erodium. *Ann. Bot.* 117, 1209–1220. doi: 10.1093/aob/mcw065

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I. A. P., Tang, H., Wang, X., et al. (2014). Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *Science* 345, 950–953. doi: 10.1126/science.1253435

Chen, X., Li, H., Pandey, M. K., Yang, Q., Wang, X., Garg, V., et al. (2016). Draft genome of the peanut A-genome progenitor (Arachis duranensis) provides insights into geocarpy, oil biosynthesis, and allergens. *Proc. Natl. Acad. Sci. U. S. A.* 113:6785. doi: 10.1073/pnas.1600899113

Chen, X., Lu, Q., Liu, H., Zhang, J., Hong, Y., Lan, H., et al. (2019). Sequencing of Cultivated Peanut, Arachis hypogaea, Yields Insights into Genome Evolution and Oil Improvement. *Mol. Plant* 12, 920–934. doi: 10.1016/j.molp.2019.03.005

Choi, I.-S., Jansen, R., and Ruhlman, T. (2020). Caught in the Act: variation in plastid genome inverted repeat expansion within and between populations of Medicago minima. *Ecol. Evol.* 10, 12129–12137. doi: 10.1002/ece3.6839

da Cunha, F. B., Nobile, P. M., Hoshino, A. A., Moretzsohn, M. D. C., Lopes, C. R., and Gimenes, M. A. (2008). Genetic relationships among Arachis hypogaea L. (AABB) and diploid Arachis species with AA and BB genomes. *Genet. Resour. Crop Evol.* 55, 15–20. doi: 10.1016/j.gene.2021.145539

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

Fávero, A. P., Simpson, C. E., Valls, J. F. M., and Vello, N. A. (2006). Study of the Evolution of Cultivated Peanut through Crossability Studies among Arachis ipaënsis, A. duranensis, and A. hypogaea. *Crop Sci.* 46, 1546–1552. doi: 10.2135/cropsci2005.09-0331

Feldman, M., Levy, A. A., Fahima, T., and Korol, A. (2012). Genomic asymmetry in allopolyploid plants: wheat as a model. *J. Exp. Bot.* 63, 5045–5059. doi: 10.1093/jxb/ers192

Ferguson, M. E., Bramel, P. J., and Chandra, S. (2004). Gene Diversity among Botanical Varieties in Peanut (Arachis hypogaea L.). *Crop Sci.* 44, 1847–1854. doi: 10.2135/cropsci2004.1847

Garcia, G. M., Stalker, H. T., and Kochert, G. (1995). Introgression analysis of an interspecific hybrid population in peanuts (Arachis hypogaea L.) using RFLP and RAPD markers. *Genome* 38, 166–176. doi: 10.1139/g95-021

Gibbons, R. W., Bunting, A. H., and Smartt, J. (1972). The classification of varieties of groundnut (Arachis hypogaea L.). *Euphytica* 21, 78–85. doi: 10.1007/bf00040550

Gill, N., Findley, S., Walling, J. G., Hans, C., Ma, J., Doyle, J., et al. (2009). Molecular and Chromosomal Evidence for Allopolyploidy in Soybean. *Plant Physiol.* 151, 1167–1174. doi: 10.1104/pp.109.137935

Gimenes, M. A., Lopes, C. R., Galgaro, M. L., Valls, J. F. M., and Kochert, G. (2002). RFLP analysis of genetic variation in species of section Arachis, genus Arachis (Leguminosae). *Euphytica* 123, 421–429. doi: 10.1007/s00122-005-0017-0

Guo, L., Guo, S., Xu, J., He, L., Carlson, J. E., and Hou, X. (2020). Phylogenetic analysis based on chloroplast genome uncover evolutionary relationship of all the nine species and six cultivars of tree peony. *Ind. Crops Prod.* 153:112567. doi: 10.1016/j.indcrop.2020.112567

Hassoubah, S., Farsi, R., Alrahimi, D., Nass, N., and Bahieldin, A. (2020). Comparison of Plastome SNPs/INDELs among different Wheat (Triticumsp.) Cultivars. *Biosci. Biotechnol. Res. Asia* 17, 27–44. doi: 10.13005/bbra/2807

He, G., Barkley, N. A., Zhao, Y., Yuan, M., and Prakash, C. S. (2014). Phylogenetic relationships of species of genus Arachis based on genic sequences. *Genome* 57, 327–334. doi: 10.1139/gen-2014-0037

He, G., and Prakash, C. (2001). Evaluation of genetic relationships among botanical varieties of cultivated peanut (Arachis hypogaea L.) using AFLP markers. *Genet. Resour. Crop Evol.* 48, 347–352.

Henriquez, C. L., Abdullah, Ahmed, I., Carlsen, M. M., Zuluaga, A., Croat, T. B., et al. (2020). Molecular evolution of chloroplast genomes in Monsteroideae (Araceae). *Planta* 251:72. doi: 10.1007/s00425-020-03365-7

Jansen, R. K., Cai, Z., Raubeson, L. A., Daniell, H., Depamphilis, C. W., Leebens-Mack, J., et al. (2007). Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proc. Natl. Acad. Sci. U. S. A.* 104:19369. doi: 10.1073/pnas.0709121104

Jarvis, A., Ferguson, M. E., Williams, D. E., Guarino, L., Jones, P. G., Stalker, H. T., et al. (2003). Biogeography of Wild Arachis. *Crop Sci.* 43, 1100–1108. doi: 10.2135/cropsci2003.1100

Jeon, J.-H., and Kim, S.-C. (2019). Comparative Analysis of the Complete Chloroplast Genome Sequences of Three Closely Related East-Asian Wild Roses (Rosa sect. Synstylae; Rosaceae). *Genes* 10:23. doi: 10.3390/genes10010023

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

Jung, S., Tate, P. L., Horn, R., Kochert, G., Moore, K., and Abbott, A. G. (2003). The Phylogenetic Relationship of Possible Progenitors of the Cultivated Peanut. *J. Hered.* 94, 334–340. doi: 10.1093/jhered/esg061

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

Kochert, G., Stalker, H. T., Gimenes, M., Galgaro, L., Lopes, C. R., and Moore, K. (1996). RFLP and Cytogenetic Evidence on the Origin and Evolution of Allotetraploid Domesticated Peanut, Arachis hypogaea (Leguminosae). *Am. J. Bot.* 83, 1282–1291.

Konate, M., Sanou, J., Miningou, A., Okello, D., Desmae, H., Janila, P., et al. (2020). Past, Present and Future Perspectives on Groundnut Breeding in Burkina Faso. *Agronomy* 10:704. doi: 10.3390/agronomy10050704

Koppolu, R., Upadhyaya, H. D., Dwivedi, S. L., Hoisington, D. A., and Varshney, R. K. (2010). Genetic relationships among seven sections of genus Arachis studied by using SSR markers. *BMC Plant Biol.* 10:15. doi: 10.1186/1471-2229-10-15

Krapovickas, A., Gregory, W. C., Williams, D. E., and Simpson, C. E. (2007). Taxonomy of the genus Aeachis (Leguminosae). *Bonplandia* 16, 7–205.

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

Leal-Bertioli, S. C. M., Santos, S. P., Dantas, K. M., Inglis, P. W., Nielen, S., Araujo, A. C. G., et al. (2014). Arachis batizocoi: a study of its relationship to cultivated peanut (A. hypogaea) and its potential for introgression of wild genes into the peanut crop using induced allotetraploids. *Ann. Bot.* 115, 237–249. doi: 10.1093/aob/mcu237

Lee, C., Choi, I.-S., Cardoso, D., De Lima, H. C., De Queiroz, L. P., Wojciechowski, M. F., et al. (2021). The chicken or the egg? Plastome evolution and an independent loss of the inverted repeat in papilionoid legumes. *Plant J.* 107, 861–875. doi: 10.1111/tpj.15351

Li, P., Zhang, S., Li, F., Zhang, S., Zhang, H., Wang, X., et al. (2017). A Phylogenetic Analysis of Chloroplast Genomes Elucidates the Relationships of the Six Economically Important Brassica Species Comprising the Triangle of U. *Front. Plant Sci.* 8:111. doi: 10.3389/fpls.2017.00111

Li, X., Yang, J.-B., Wang, H., Song, Y., Corlett, R. T., Yao, X., et al. (2021). Plastid NDH Pseudogenization and Gene Loss in a Recently Derived Lineage from the Largest Hemiparasitic Plant Genus Pedicularis (Orobanchaceae). *Plant Cell Physiol.* 62, 971–984. doi: 10.1093/pcp/pcab074

Liu, S., Wang, Z., Wang, H., Su, Y., and Wang, T. (2020). Patterns and Rates of Plastid rps12 Gene Evolution Inferred in a Phylogenetic Context using Plastomic Data of Ferns. *Sci. Rep.* 10:9394. doi: 10.1038/s41598-020-66219-y

Lowe, T. M., and Eddy, S. R. (1997). tRNAscan-SE: a Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Res.* 25, 955–964. doi: 10.1093/nar/25.5.955

Mehmood, F., Abdullah, Shahzadi, I., Ahmed, I., Waheed, M. T., and Mirza, B. (2020). Characterization of Withania somnifera chloroplast genome and its comparison with other selected species of Solanaceae. *Genomics* 112, 1522–1530. doi: 10.1016/j.ygeno.2019.08.024

Middleton, C. P., Senerchia, N., Stein, N., Akhunov, E. D., Keller, B., Wicker, T., et al. (2014). Sequencing of Chloroplast Genomes from Wheat, Barley, Rye and Their Relatives Provides a Detailed Insight into the Evolution of the Triticeae Tribe. *PLoS One* 9:e85761. doi: 10.1371/journal.pone.0085761

Moner, A. M., Furtado, A., and Henry, R. J. (2020). Two divergent chloroplast genome sequence clades captured in the domesticated rice gene pool may have significance for rice production. *BMC Plant Biol.* 20:472. doi: 10.1186/s12870-020-02689-6

Moore, M. J., Bell, C. D., Soltis, P. S., and Soltis, D. E. (2007). Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proc. Natl. Acad. Sci. U. S. A.* 104:19363. doi: 10.1073/pnas.0708072104

Moretzsohn, M. C., Gouvea, E. G., Inglis, P. W., Leal-Bertioli, S. C. M., Valls, J. F. M., and Bertioli, D. J. (2013). A study of the relationships of cultivated peanut (Arachis hypogaea) and its most closely related wild species using intron sequences and microsatellite markers. *Ann. Bot.* 111, 113–126. doi: 10.1093/aob/mcs237

Moretzsohn, M. D. C., Hopkins, M. S., Mitchell, S. E., Kresovich, S., Valls, J. F. M., and Ferreira, M. E. (2004). Genetic diversity of peanut (Arachis hypogaea L.) and its wild relatives based on the analysis of hypervariable regions of the genome. *BMC Plant Biol.* 4:11. doi: 10.1186/1471-2229-4-11

Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2014). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Nock, C. J., Hardner, C. M., Montenegro, J. D., Ahmad Termizi, A. A., Hayashi, S., Playford, J., et al. (2019). Wild Origins of Macadamia Domestication Identified Through Intraspecific Chloroplast Genome Sequencing. *Front. Plant Sci.* 10:334. doi: 10.3389/fpls.2019.00334

Pandey, M. K., Pandey, A. K., Kumar, R., Nwosu, C. V., Guo, B., Wright, G. C., et al. (2020). Translational genomics for achieving higher genetic gains in groundnut. *Theor. Appl. Genet.* 133, 1679–1702. doi: 10.1007/s00122-020-03592-2

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

Prabhudas, S. K., Prayaga, S., Madasamy, P., and Natarajan, P. (2016). Shallow Whole Genome Sequencing for the Assembly of Complete Chloroplast Genome Sequence of Arachis hypogaea L. *Front. Plant Sci.* 7:1106. doi: 10.3389/fpls.2016.01106

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

Raina, S. N., and Mukai, Y. (1999). Genomic in situ hybridization inArachis (Fabaceae) identifies the diploid wild progenitors of cultivated (A. hypogaea) and related wild (A. monticola) peanut species. *Plant Syst. Evol.* 214, 251–262.

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

Rozas, J., Ferrer-Mata, A., Sánchez-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.* 34, 3299–3302. doi: 10.1093/molbev/msx248

Schwarz, E. N., Ruhlman, T. A., Sabir, J. S. M., Hajrah, N. H., Alharbi, N. S., Al-Malki, A. L., et al. (2015). Plastid genome sequences of legumes reveal parallel inversions and multiple losses of rps16 in papilionoids. *J. Syst. Evol.* 53, 458–468. doi: 10.1111/jse.12179

Seijo, J. G., Lavia, G. I., Fernández, A., Krapovickas, A., Ducasse, D., and Moscone, E. A. (2004). Physical mapping of the 5S and 18S–25S rRNA genes by FISH as evidence that Arachis duranensis and A. ipaensis are the wild diploid progenitors of A. hypogaea (Leguminosae). *Am. J. Bot.* 91, 1294–1303. doi: 10.3732/ajb.91.9.1294

Sigmon, B. A., Adams, R. P., and Mower, J. P. (2017). Complete chloroplast genome sequencing of vetiver grass (Chrysopogon zizanioides) identifies markers that distinguish the non-fertile 'Sunshine' cultivar from other accessions. *Ind. Crops Prod.* 108, 629–635. doi: 10.1016/j.indcrop.2017.07.029

Simpson, C. E., Krapovickas, A., and Valls, J. F. M. (2001). History of Arachis Including Evidence of *A. hypogaea* L. Progenitors. *Peanut Sci.* 28, 78–80. doi: 10.3146/i0095-3679-28-2-7

Singh, A. K., Gurtu, S., and Jambunathan, R. (1994). Phylogenetic relationships in the genus Arachis based on seed protein profiles. *Euphytica* 74, 219–225.

Singh, A. K., and Moss, J. P. (1982). Utilization of wild relatives in genetic improvement of Arachis hypogaea L. *Theor. Appl. Genet.* 61, 305–314.

Singh, A. K., Subrahmanyam, P., and Gurtu, S. (1996). Variation in a wild groundnut species, Arachis duranensis Krapov. & W.C. Gregory. *Genet. Resour. Crop Evol.* 43, 135–142.

Song, Y., Yu, W.-B., Tan, Y., Liu, B., Yao, X., Jin, J., et al. (2017). Evolutionary Comparisons of the Chloroplast Genome in Lauraceae and Insights into Loss Events in the Magnoliids. *Genome Biol. Evol.* 9, 2354–2364. doi: 10.1093/gbe/evx180

Stalker, H. T. (2017). Utilizing Wild Species for Peanut Improvement. *Crop Sci.* 57, 1102–1120.

Stalker, H. T., Dhesi, J. S., Parry, D. C., and Hahn, J. H. (1991). Cytological and Interfertility Relationships of Arachis Section Arachis. *Am. J. Bot.* 78, 238–246.

Stalker, H. T., and Moss, J. P. (1987). Speciation, Cytogenetics, and Utilization of Arachis Species. *Adv. Agron.* 41, 1–40.

Subrahmanyam, P., Anaidu, R., Reddy, L. J., Kumar, P. L., and Ferguson, M. E. (2001). Resistance to groundnut rosette disease in wild Arachis species. *Ann. Appl. Biol.* 139, 45–50. doi: 10.1111/j.1744-7348.2001.tb00129.x

Sugiura, C., and Sugita, M. (2004). Plastid transformation reveals that moss tRNAArg-CCG is not essential for plastid function. *Plant J.* 40, 314–321.

Tallury, S. P., Hollowell, J. E., Isleib, T. G., and Stalker, H. T. (2014). Greenhouse Evaluation of Section Arachis Wild Species for Sclerotinia Blight and Cylindrocladium Black Rot Resistance. *Peanut Sci.* 41, 17–24. doi: 10.3146/ps13-02.1

Tan, D., Zhang, Y., and Wang, A. (2010). A review of geocarpy and amphicarpy in angiosperms, with special reference to their ecological adaptive significance. *Chin. J. Plant Ecol.* 34, 72–88.

Tian, X., Ye, J., and Song, Y. (2019). Plastome sequences help to improve the systematic position of trinerved Lindera species in the family Lauraceae. *PeerJ* 7:e7662. doi: 10.7717/peerj.7662

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

Tonti-Filippini, J., Nevill, P. G., Dixon, K., and Small, I. (2017). What can we do with 1000 plastid genomes?. *Plant J.* 90, 808–818. doi: 10.1111/tpj.13491

Turmel, M., Otis, C., and Lemieux, C. (2002). The chloroplast and mitochondrial genome sequences of the charophyte ⁢em>Chaetosphaeridium globosum⁢/em>: insights into the timing of the events that restructured organelle DNAs within the green algal lineage that led to land plants. *Proc. Natl. Acad. Sci. U. S. A.* 99:11275.

Tyagi, S., Jung, J.-A., Kim, J. S., and Won, S. Y. (2020). A comparative analysis of the complete chloroplast genomes of three Chrysanthemum boreale strains. *PeerJ* 8:e9448. doi: 10.7717/peerj.9448

Upadhyaya, H. D., Dwivedi, S. L., Nadaf, H. L., and Singh, S. (2011). Phenotypic diversity and identification of wild Arachis accessions with useful agronomic and nutritional traits. *Euphytica* 182:103.

Varshney, R. K., Mahendar, T., Aruna, R., Nigam, S. N., Neelima, K., Vadez, V., et al. (2009). High level of natural variation in a groundnut (Arachis hypogaea L.) germplasm collection assayed by selected informative SSR markers. *Plant Breed.* 128, 486–494. doi: 10.1111/j.1439-0523.2009.01638.x

Vishwakarma, M. K., Kale, S. M., Sriswathi, M., Naresh, T., Shasidhar, Y., Garg, V., et al. (2017). Genome-Wide Discovery and Deployment of Insertions and Deletions Markers Provided Greater Insights on Species, Genomes, and Sections Relationships in the Genus Arachis. *Front. Plant Sci.* 8:2064. doi: 10.3389/fpls.2017.02064

Wang, C.-L., Ding, M.-Q., Zou, C.-Y., Zhu, X.-M., Tang, Y., Zhou, M.-L., et al. (2017). Comparative Analysis of Four Buckwheat Species Based on Morphology

and Complete Chloroplast Genome Sequences. *Sci. Rep.* 7:6514. doi: 10.1038/s41598-017-06638-6

Wang, J., Li, C., Shi, D., Liu, Y., Tang, R., He, L., et al. (2021). Verifying high variation regions based on sect. Arachis chloroplast genome and revealing the interspecies genetic relationship. *Chin. J. Oil Crop Sci.* 43:495.

Wang, J., Li, C., Yan, C., Zhao, X., and Shan, S. (2018). A comparative analysis of the complete chloroplast genome sequences of four peanut botanical varieties. *PeerJ* 6:e5349. doi: 10.7717/peerj.5349

Wang, J., Li, Y., Li, C., Yan, C., Zhao, X., Yuan, C., et al. (2019). Twelve complete chloroplast genomes of wild peanuts: great genetic resources and a better understanding of Arachis phylogeny. *BMC Plant Biol.* 19:504. doi: 10.1186/s12870-019-2121-3

Wheeler, T. J., and Eddy, S. R. (2013). nhmmer: DNA homology search with profile HMMs. *Bioinformatics* 29, 2487–2489. doi: 10.1093/bioinformatics/btt403

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

Xu, J.-H., Liu, Q., Hu, W., Wang, T., Xue, Q., and Messing, J. (2015). Dynamics of chloroplast genomes in green plants. *Genomics* 106, 221–231. doi: 10.1016/j.ygeno.2015.07.004

Xue, S., Shi, T., Luo, W., Ni, X., Iqbal, S., Ni, Z., et al. (2019). Comparative analysis of the complete chloroplast genome among Prunus mume, P. armeniaca, and P. salicina. *Hortic. Res.* 6:89. doi: 10.1038/s41438-019-0171-1

Yin, D., Ji, C., Song, Q., Zhang, W., Zhang, X., Zhao, K., et al. (2020). Comparison of Arachis monticola with Diploid and Cultivated Tetraploid Genomes Reveals Asymmetric Subgenome Evolution and Improvement of Peanut. *Adv. Sci.* 7:1901672. doi: 10.1002/advs.201901672

Yin, D., Wang, Y., Zhang, X., Ma, X., He, X., and Zhang, J. (2017). Development of chloroplast genome resources for peanut (Arachis hypogaea L.) and other species of Arachis. *Sci. Rep.* 7:11649.

Yu, J., Xu, F., Wei, Z., Zhang, X., Chen, T., and Pu, L. (2020). Epigenomic landscape and epigenetic regulation in maize. *Theor. Appl. Genet.* 133, 1467–1489. doi: 10.1007/s00122-020-03549-5

Zhang, D., Luo, K., Wu, F., Wang, Y., and Zhang, J. (2017). Advances in cleistogamy of angiosperms. *Pratacultural Sci.* 34, 1215–1227. doi: 10.1111/tpj.12693

Zheng, S., Poczai, P., Hyvönen, J., Tang, J., and Amiryousefi, A. (2020). Chloroplot: an Online Program for the Versatile Plotting of Organelle Genomes. *Front. Genet.* 11:576124. doi: 10.3389/fgene.2020.576124

Zheng, X., Wang, J., Feng, L., Liu, S., Pang, H., Qi, L., et al. (2017). Inferring the evolutionary mechanism of the chloroplast genome size by comparing whole-chloroplast genome sequences in seed plants. *Sci. Rep.* 7:1555. doi: 10.1038/s41598-017-01518-5

Zheng, Z., Sun, Z., Fang, Y., Qi, F., Liu, H., Miao, L., et al. (2018). Genetic Diversity, Population Structure, and Botanical Variety of 320 Global Peanut Accessions Revealed Through Tunable Genotyping-by-Sequencing. *Sci. Rep.* 8:14500. doi: 10.1038/s41598-018-32800-9

Zhuang, W., Chen, H., Yang, M., Wang, J., Pandey, M. K., Zhang, C., et al. (2019). The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. *Nat. Genet.* 51, 865–876. doi: 10.1038/s41588-019-0402-2

Zhuang, W., Wang, X., Paterson, A. H., Chen, H., Yang, M., Zhang, C., et al. (2020). Reply to: evaluating two different models of peanut's origin. *Nat. Genet.* 52, 560–563. doi: 10.1038/s41588-020-0627-0

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

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

frontiers in Plant Science

ORIGINAL RESEARCH
published: 20 January 2022
doi: 10.3389/fpls.2021.814252

TALE Transcription Factors in Sweet Orange (*Citrus sinensis*): Genome-Wide Identification, Characterization, and Expression in Response to Biotic and Abiotic Stresses

Weiye Peng[1,2†], Yang Yang[1,2†], Jing Xu[1,2], Erping Peng[1,2], Suming Dai[3,4], Liangying Dai[1,2], Yunsheng Wang[1,2], Tuyong Yi[1,2], Bing Wang[1,2], Dazhi Li[3,4*] and Na Song[1,2*]*

[1] College of Plant Protection, Hunan Agricultural University, Changsha, China, [2] Hunan Provincial Key Laboratory for Biology and Control of Plant Diseases and Insect Pests, Hunan Agricultural University, Changsha, China, [3] Horticulture College, Hunan Agricultural University, Changsha, China, [4] National Center for Citrus Improvement Changsha, Changsha, China

Three-amino-acid-loop-extension (TALE) transcription factors comprise one of the largest gene families in plants, in which they contribute to regulation of a wide variety of biological processes, including plant growth and development, as well as governing stress responses. Although sweet orange (*Citrus sinensis*) is among the most commercially important fruit crops cultivated worldwide, there have been relatively few functional studies on *TALE* genes in this species. In this study, we investigated 18 *CsTALE* gene family members with respect to their phylogeny, physicochemical properties, conserved motif/domain sequences, gene structures, chromosomal location, *cis*-acting regulatory elements, and protein–protein interactions (PPIs). These *CsTALE* genes were classified into two subfamilies based on sequence homology and phylogenetic analyses, and the classification was equally strongly supported by the highly conserved gene structures and motif/domain compositions. *CsTALEs* were found to be unevenly distributed on the chromosomes, and duplication analysis revealed that segmental duplication and purifying selection have been major driving force in the evolution of these genes. Expression profile analysis indicated that *CsTALE* genes exhibit a discernible spatial expression pattern in different tissues and differing expression patterns in response to different biotic/abiotic stresses. Of the 18 *CsTALE* genes examined, 10 were found to be responsive to high temperature, four to low temperature, eight to salt, and four to wounding. Moreover, the expression of *CsTALE3/8/12/16* was induced in response to infection with the fungal pathogen *Diaporthe citri* and bacterial pathogen *Candidatus* Liberibacter asiaticus, whereas the expression of *CsTALE15/17* was strongly suppressed. The transcriptional activity of *Cs*TALE proteins was also verified in yeast, with yeast two-hybrid assays indicating that

OPEN ACCESS

Edited by:
Hai Du,
Southwest University, China

Reviewed by:
Jin-zhi Zhang,
Huazhong Agricultural University, China
Qiang Zhou,
Lanzhou University, China
Ling Xu,
Zhejiang Sci-Tech University, China

***Correspondence:**
Bing Wang
zhufu@hunau.edu.cn
Dazhi Li
ldazhi@163.com
Na Song
songna@hunau.edu.cn

† These authors have contributed equally to this work

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

Received: *13 November 2021*
Accepted: *13 December 2021*
Published: *20 January 2022*

Citation:
Peng W, Yang Y, Xu J, Peng E, Dai S, Dai L, Wang Y, Yi T, Wang B, Li D and Song N (2022) TALE Transcription Factors in Sweet Orange (Citrus sinensis): Genome-Wide Identification, Characterization, and Expression in Response to Biotic and Abiotic Stresses. Front. Plant Sci. 12:814252. doi: 10.3389/fpls.2021.814252

CsTALE3/CsTALE8, CsTALE3/CsTALE11, CsTALE10/CsTALE12, CsTALE14/CsTALE8, CsTALE14/CsTALE11 can form respective heterodimers. The findings of this study could lay the foundations for elucidating the biological functions of the *TALE* family genes in sweet orange and contribute to the breeding of stress-tolerant plants.

Keywords: *Citrus sinensis*, genome-wide characterization, expression analysis, *TALE* transcription factor, biotic and abiotic stresses

INTRODUCTION

In plants, numerous transcription factors (TFs) have been identified and shown to play significant roles in the regulation of developmental processes, stress responses, and genetic control (Liu X. et al., 2019). TFs in the three-amino-loop-extension (TALE) gene family have established to be relatively numerous and highly conserved in different plant species (Choe et al., 2014). These genes are classified into two subfamilies, namely, the KNOX (KNOTTED-like homeodomain) and BEL (BEL1-Like homeodomain) subfamilies, which normally function as heterodimeric TF complexes that contribute to modifying physiological and biochemical properties, particularly those associated with the metabolism and biosynthesis of lignin (Yoon et al., 2014, 2017). TALE proteins have a distinctive common characteristic in that interactions can occur either between TALE and non-TALE members or among different TALE family members (Hudry et al., 2014). In barley, for example, BKN3 (KNOX protein) has been shown to interact with JUBEL1 and JUBEL2 (BEL proteins; Müller et al., 2001), whereas SHOOT MERISTEMLESS (STM), a MEINOX domain protein, has been demonstrated to be a common interacting partner of three BEL homeodomain members (ATH1, BLH3, and BLH9; Cole et al., 2006).

The nutritional and economic value of fruits is dependent to a large extent on their developmental status, which is often determined by *TALE* genes. In tomato (*Solanum lycopersicum*), for example, the *TALE* gene *TKN2/4* has been demonstrated to specifically influence fruit chloroplast development and thereby nutrient composition and flavor (Nadakuduti et al., 2014). Similarly, *LeT6/TKn2* has been reported to be involved in tomato fruit morphological development (Avivi et al., 2000). In addition, CcBLH6 had been found to play an active role in the lignification and lignin biosynthesis pathway of *Camellia chekiangoleosa* fruit (Yan et al., 2021). Moreover, the activity of *TALE* family members is believed have a considerable influence on the size, yield, and quality of fruit in many fruit crops, including *Actinidia chinensis*, *Fragaria vesca*, and *Litchi chinensis* (Shahan et al., 2019; Zhao et al., 2020; Brian et al., 2021).

In recent years, an increasing amount of evidence has accumulated to indicate that *TALE* genes play important roles not only in growth and development but also in the adaptation of stress responses in different plant species (Butenko and Simon, 2015). For instance, *GmSBH1*, the first *TALE* gene identified in *Glycine max*, has been shown to influence leaf phenotype and enhance plant tolerance to high temperatures or humidity (Shu et al., 2015). In *Populus*, the type I *KNOX* gene *PagKNAT2/6b* has been demonstrated to directly suppress gibberellin biosynthesis, thereby promoting phenotypic alteration and enhancing plant drought stress tolerance (Song X. et al., 2021). Similarly, POTH15, a type I KNOX gene, has been characterized as a regulator of photoperiodic development, meristem maintenance, and leaf development, and is believed to be involved in responses to plant hormone signal transduction and biotic or abiotic stresses, based on RNA sequencing and quantitative real-time PCR (qRT-PCR) validation (Mahajan et al., 2016).

Citrus species are among of the most widely cultivated and economically significant fruit crops (Xu et al., 2021). Given the large planting areas, citrus producers face multiple challenges relating to the dynamic environment and myriad stresses (Yu et al., 2020). Recent research showed that agricultural producers are facing several problems due to biotic and abiotic stresses like ubiquitous phytopathogens and changeable weather (high or low temperature, and soil salinity) which seriously reduce the Citrus yield and quality. For example, melanose disease caused by the fungal pathogen *Diaporthe citri*, which harms both leaves and fruits, contributes to massive reductions in yield and loss of quality (Mondal et al., 2007; Chaisiri et al., 2020), whereas citrus greening disease (Huanglongbing, HLB) is recognized as the most serious and fatal bacterial diseases threatening the citrus industry worldwide (Qiu et al., 2020; Yao et al., 2021). Unfortunately, HLB remains incurable, with all diseased plants eventually succumbing to the disease (Thapa et al., 2020). Currently, there are no known commercial citrus varieties with effective resistance to the phloem-residing HLB-associated bacterium *Candidatus* Liberibacter asiaticus (*C*Las; Iftikhar et al., 2016). Within a plant, the phloem is the predominant passageway for the long-distance transport of solutes and signaling, at the same time, provides an effective avenue of phloem-inhabiting bacteria spread systemically throughout a host plant (Welker et al., 2021). With respect to the breeding of resistant varieties, it is anticipated that on the basis comparative pathological, transcriptomic, and anatomical investigations using HLB-tolerant and -sensitive cultivars, phloem regeneration will become one of the most important and promising research directions in citrus production (Deng et al., 2019; Curtolo et al., 2020). It has long been established that KNAT6 (KNOX subfamily) is particularly enriched in phloem and required for correct lateral root formation in Arabidopsis (Dean et al., 2004). Similar findings have been reported in potato, in which the KNOX subfamily protein POTH1 interacts with the BEL subfamily protein StBEL5, a phloem-mobile messenger that regulates phloem transport activities (Mahajan et al., 2012; Hannapel et al., 2013). Thus, it would be of interest to investigate the potential function and underlying regulatory mechanisms of TALE family genes in host plant resistance to HLB pathogens.

To date, however, there has been no relevant research on the TALE family in sweet orange. Nevertheless, recent publication of the complete genome sequence of sweet orange now makes it feasible to conduct genome-wide identification and comparative analyses of the *TALE* gene family in sweet orange. In this study, we identified *CsTALE* genes in sweet orange, using which, we performed a comprehensive analysis, examining gene phylogeny, chromosomal position, duplication events, gene/protein structures, *cis*-acting regulatory elements (CREs), PPI networks, subcellular localization, and transcriptional activation, and undertaking yeast-two-hybrid validation. Moreover, we also examined expression profiles of all *CsTALE* genes in different sweet orange tissues and in response to different abiotic and biotic stresses. By adopting this integrative approach, we provide a basis for further elucidating the functional and mechanistic characteristics of the *TALE* genes. In addition, identification of stress resistance genes will provide a basis for effective engineering strategies to improve crop stress tolerance.

MATERIALS AND METHODS

Identification and Phylogenetic Analysis

Publicly available information relating to the sweet orange genome sequences and gene annotations were downloaded from the National Center for Biotechnology Information (NCBI) and the *Citrus sinensis* Genome Annotation Project (Xu et al., 2013; Wu et al., 2018). All Hidden Markov Model (HMM) profile files of the TALE domain (Accession no. PF05920) were downloaded from the Pfam database, version 34.0[1].

Sequences of *Arabidopsis thaliana* TALEs (*At*TALE), *Oryza sativa* TALEs (*Os*TALE), and *Populus trichocarpa* TALEs (*Pt*TALE) were obtained from previous studies (Hamant and Pautot, 2010; Zhao et al., 2019). Multiple alignments of TALE member amino acid sequences were performed using ClustalX software V2.1, employing default parameters with subsequent manual adjustment. A phylogenetic tree was generated using MEGA-X v10.2.4 software based on the neighbor-joining (NJ) algorithm, with the following parameters: Poisson correction, pair-wise deletion and bootstrap sampling (1000 replicates; random seed).

Chromosomal Distribution of *CsTALE* Genes and Duplication Events

The chromosomal positions of *CsTALE* genes were extracted from the sweet orange genome annotation information in GFF3 format and visualized using Toolkit for Biologists stand-alone software v1.0986 (Chen et al., 2020). Chromosome size and gene density were determined with reference to the sweet orange genomic annotation information. *CsTALE* gene replication events were identified using a multiple collinear scan kit (MCScanX) program with default settings. For synteny analysis, the genome sequence and gene structure annotation files of sweet orange and Arabidopsis were inputted into One Step MCScanX, followed by the visualization with Dual Synteny Plot plugin embedded in TBtools software. KaKs_Calculator2.0 (MA model) was selected to calculate non-synonymous (Ka), synonymous (Ks), and Ka/Ks values.

Gene Characteristic and Structural Analyses

The theoretical isoelectric point (pI) and molecular weight (MW) of entered protein sequence were estimated using Expert Protein Analysis System 3.0[2] (Duvaud et al., 2021). The subcellular localization of CsTALE proteins was predicted using the online bioinformatics tools Plant-mPLoc[3] and WoLF PSORT[4]. On the basis of the genome and coding sequences, the gene structure of each *TALE* gene was obtained using the Gene Structure Display server[5]. The conserved motifs of *Cs*TALE proteins were identified using the online MEME Suite Programs in classic mode. Domain-based analyses were performed using the SMART server[6] in default mode (Letunic et al., 2021).

Cis-Acting Regulatory Elements and Protein Interaction Network Predictions

In order to identify CREs in the promoter sequence of sweet orange *TALE* genes, we extracted genomic DNA sequences extending 2000 bp upstream of the transcription start site, and then submitted these to the PlantCare website[7]. Potential PPIs were predicted using the STRING online portal (Version 11.0[8]).

Plant Materials and Treatments

The sweet orange materials used in tissue-specific expression pattern and stress response analyses were obtained from the National Center for Citrus Improvement, Hunan Agricultural University, Hunan Province, China. For analysis, three samples of different tissues (leaf, stem, and flower) were sampled from the same sweet orange plant at the flowering stage, and ripe fruits were subsequently obtained.

The different stress treatments performed in this study were carried out as previously described, with each experiment being conducted with three replicates (Song N. et al., 2021). For the purposes of stress analysis, we used 1-month-old sweet orange plants that had been grown in greenhouse at 25°C under an 8-h dark/16-h light photoperiod.

Salt Stress Assays

Sweet orange seedlings with good health and the same growth potential were transferred to flasks containing 100 mM NaCl, with sterile distilled water serving as a control. Samples then collected at 0, 12, 24, and 48 h after treatment.

[1]https://pfam.xfam.org/

[2]https://web.expasy.org/protparam/

[3]http://www.csbio.sjtu.edu.cn/bioinf/plant-multi/

[4]https://wolfpsort.hgc.jp/

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

[6]http://smart.embl.de/

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

[8]https://string-db.org/cgi/input.pl

Wounding Assays

The leaves of the well-growth sweet orange were gently stab a wound with a pipette tip, with non-wounded plants as a control. Samples were taken at 0, 12, 24, and 48 h after treatment.

High or Low Temperature Stress Assays

Sweet orange seedlings with good health and the same growth potential were transferred to plant growth cabinet at 40 or 4°C for high temperature and low temperature treatments, with normal growth conditions as a control. Samples were taken at 0, 12, 24, and 48 h after treatment.

For each stress type, three independent samples were harvested at 0, 12, 24, and 48 h after treatment, and then immediately snap-frozen in liquid nitrogen and thereafter maintained at –80°C until used for RNA extraction.

Diaporthe citri spores were incubated on oat agar medium at 25°C until germinating, A suspension of these spores (1×10^6 spores/mL) was subsequently used to inoculate 1-month-old sweet orange plants using the spray method as previously described (Agostini et al., 2003). *C*Las inoculation was performed using to a slightly modified version of the method described by Martins Cristina de Paula Santos et al. (de Paula Santos Martins et al., 2015). *C*las-infected sweet oranges showing typical HLB symptoms were collected from commercial citrus growing plantations in the central south region of Hunan Province, China. The seedlings were graft-inoculated with budwood from *C*Las-free and *C*Las-infected sweet orange to obtain healthy and infected plants, respectively, as described previously (Suh et al., 2021). For all new mature leaves, the presence of *C*Las was examined based on qRT-PCR analysis for 6 months post inoculation (Maheshwari et al., 2021). Three independent samples were collected at 0, 24, and 48 h post inoculation and maintained as described above.

RNA Extraction and Quantitative Real-Time PCR

Total RNA was extracted from sweet orange leaves using TransZol (TransGen Biotech, Beijing, China) in accordance with the manufacturer's instructions. One microgram of total RNA was used for first-strand complementary DNA synthesis using a Goldenstar RT6 cDNA Synthesis Kit (Tsingke Biotechnology, Beijing, China). All qRT-PCR reactions were run and analyzed using a CFX96 Touch Deep Well Real-Time PCR Detection System (Bio-Rad, Munich, Germany) with a SYBR Green PCR Mastermix (Solarbio, Beijing, China). qRT-PCR was conducted following standard procedures and conditions as previously described (Peng et al., 2021b). qRT-PCR gene-specific primers were designed using Oligo7 software (**Supplementary Table 7**).

Subcellular Localization

Amplified full-length *TALE* fragments were cloned into a linearized pCAMBIAI1132 vector between a CaMV35S promoter and green fluorescent protein tag using ClonExpress II One Step Cloning Kit (Vazyme, Nanjing, China). The resulting vectors were introduced into *Agrobacterium tumefaciens* EHA105 by electroporation, followed infiltration into *Nicotiana benthamiana* leaves using a needleless syringe. Subsequently, samples were viewed under a CarlZeiss LSM710 confocal laser scanning microscopy.

Transcriptional Activation and Yeast Two-Hybrid Assay

The transcriptional activation of TALE was analyzed according to a previously reported method (Liu et al., 2021). The full-length sequences of TALE proteins were fused in a pGBKT7 vector. Subsequently, pGBKT7-TALE recombinant vectors and a negative control pGBKT7 empty vector were separately transformed into the yeast strain AH109 in accordance with the manufacturer's protocol (Weidi Biotechnology, Shanghai, China). The resulting culture was diluted and dropped on SD/–Trp, SD/–Trp/–His/–Ade and SD/–Trp/–His/–Ade/X-α-Gal synthetic dropout medium (Clontech, Mountain View, CA, United States), followed by incubation at 28°C for 3 days.

In order to validate the interactions between members of the *CsTALE* gene family, we performed a yeast two-hybridization (Y2H) assay using the constructed bait (pGBKT7) and prey (pGADT7) vectors. Yeast transformation was performed using Yeastmaker Yeast Transformation System 2 (Takara, Tokyo, Japan) according to the manufacturer's instructions. Recombinant vectors and negative (pGBKT7/pGADT7) and positive (pGBKT7-53/pGADT7-T) control vectors were separately transformed into yeast strain AH109 as described above. Thereafter, the transformed yeasts were plated on SD/–Trp/-His, SD/–Trp/–Leu/–His, and SD/–Trp/–Leu/–His/–Ade/X-α-Gal synthetic dropout medium and incubated at 28°C for 3 or 4 days.

RESULTS

Genome-Wide Identification and Phylogenetic Analysis of *TALE* Genes in Sweet Orange

To identify TALE genes in sweet orange, initial candidates were retrieved from the NCBI and *Citrus sinensis* Genome Annotation Project databases. HMMER (Hidden Markov Model) matrices specific to TALE family were then constructed using HMMBUILD (HMMER-3.1) and scanned against the PFAM domain (PF05920). On the basis of a rigorous two-staged screening process, we identified a total of 18 TALE superfamily genes were identified in sweet orange, which account for approximately 0.06% of the entire sweet orange genome (29,445 predicted genes in sweet orange). These 18 TALE family members were named based on the order of their chromosomal location (*Cs*TALE1–*Cs*TALE18) and different transcripts were distinguished by the postscripts a/b/c/d. Lengths of the open reading frames of sweet orange TALE genes ranged from 582 to 2520 bp, and calculated theoretical MWs of the TALE proteins varied ranged 22.16–92.52 kDa. The gene id, protein sequence, physicochemical properties and subcellular localization prediction of the characterized TALE genes/proteins are presented in **Supplementary Table 1**. A pairwise identity (%) matrix revealed similarities among of the sweet orange TALE

family nucleotide and amino acid sequences, among which, the highest degree of similarity (93.44/96.89%) was obtained for *CsTALE4* and *CsTALE16* (**Supplementary Table 2**).

In order to reconstruct the evolutionary relationships among sweet orange and Arabidopsis TALE members, 39 aligned TALE protein sequences from sweet orange (18 TALE proteins), Arabidopsis (21 TALE proteins), rice (26 TALE proteins), poplar (35 TALE proteins) were used to generate a phylogenetic tree using MEGA X software and the neighbor-joining method (**Figure 1**). The phylogenetic distribution clearly indicated that the *TALE* genes clustered into two subfamilies (KNOX and BEL). The KNOX and BEL subfamilies were found to contain 10 and seven *CsTALE* genes, respectively, whereas *CsTALE7* forms a separate evolutionary branch. According to the current classification, these two clusters show obvious differences with respect to TALE sequence length, with the average length of KNOX and BEL subfamily proteins being 626 and 336 amino acids, respectively. The relevant grouping information, gene ids, and gene names are provided in **Supplementary Table 3**.

Chromosomal Position and Duplication Analysis of *CsTALE* Genes

In order to determine the chromosomal distribution of *CsTALE* genes, the positions of *CsTALEs* were mapped on the chromosomes of sweet orange based on the NCBI *Citrus sinensis* genome sequence (Assembly Csi_valencia_1.0). The assessment results revealed a sparse distribution of the 18 *CsTALE* genes across all chromosomes, with the exception of chromosome 9 (**Supplementary Figure 1**), with variable *CsTALE* gene densities on individual chromosomes. The highest *CsTALE* gene frequency (three) was detected on chromosome 7, whereas chromosomes 1 and Un each harbored only a single gene (*CsTALE1* and *CsTALE18*, respectively). For the purposes of the present study, we defined tandem duplicated pairs as a genomic region harboring two or more neighboring *CsTALE* genes residing within a 20 kb sequence. Among all *CsTALE* genes, we detected only a single tandem duplicated pair (*CsTALE4*/*CsTALE5*) located adjacent to each other in a chromosomal region.

To further examine the relationship between genetic divergence and gene duplication, we performed comparative syntenic and duplication pair analyses. On the basis of our analysis of the sweet orange genome, we identified 10 segmental duplication events involving 10 *CsTALE* genes (**Figure 2A**). With the exceptions of chromosomes 4 and 9, segmental duplicates were detected on all chromosomes. *CsTALE3/9/11* were found to be involved in three duplication events, whereas others have been involved in two events (*CsTALE8/14/18*) or one event (*CsTALE1/4/15/16*). To assess the direction and strength of natural selection pressure, we estimated the rates of *K*a to *K*s substitution. The ratios of *K*a to *K*s for the 10 pairs of *CsTALE* genes were less than 1, ranging from 0.11 to 0.38, which indicates that the *CsTALE* gene pairs in sweet orange have undergone purifying selection during the course of evolution (**Supplementary Table 4**). *K*s values are routinely used to obtain approximate estimates of the evolutionary dates of segmental duplication events. We established that the duplication of *CsTALE* genes occurred during the from 3.70 Mya to 12.72 Mya, with a mean date of 8.86 Mya. In order to clarify the evolution and collinearity of sweet orange TALE family members among species, we sought to identify members of the *Cs*TALE family that had colinear relationships with those in the model plant *A. thaliana*, and accordingly identified 15 colinear gene pairs (**Figure 2B** and **Supplementary Table 5**). Syntenic relations of the TALE members among *C. sinensis*, *A. thaliana*, *O. sativa,* and *P. trichocarpa* are visualized in **Supplementary Figure 2**.

Structural and *Cis*-Acting Regulatory Element Analysis of *CsTALE* Genes

Bioinformatics data obtained for proteins can enable us to establish correlations between structure and function, and in this regard, we determined motif/domain and exon/intron structures based on the corresponding amino acid and genome sequences (**Figure 3**). Structural analyses of these genes revealed that members of the BEL subfamily have the same number of exons, namely four, whereas KONX subfamily members are characterized by a diverse exon complement, ranging from three to six (**Figure 3B**). Moreover, all BEL subfamily members contain a 5′-UTR (except *CsTALE17*) and truncations to the 5′-UTR and/or 3′-UTR were found to be common among the *CsTALE* genes. In addition, BEL subfamily proteins all contain POX domains, whereas in the KONX subfamily, all proteins contain KONX1 and KONX2 domains, which are notably consistent with the subfamily clustering (**Figure 3C**). Some KONX subfamily proteins (*Cs*TALE2/3/9/14) are also characterized by an additional ELK domain. Conversely, with the exception of individual variants (*Cs*TALE6c/13c/13d), both BEL and KONX subfamily proteins possess a Homeobox_KN domain. In total, we identified 9 conserved motifs, designated motifs 1–9, in *Cs*TALE, with the number of conserved motifs in each *Cs*TALE ranging from 2 to 9 (**Figure 3D**). Notably, some characterized motifs were found to be present exclusively in one or the other subfamily, namely, motifs 5 and 8 in the BEL subfamily and motifs 3, 4, and 6 in the KONX subfamily.

cis-acting regulatory elements (CREs), located upstream of the promoter region, are essential sites for TFs that are associated with the initiation of transcription, and function as control centers for gene transcription. Among the CREs identified, abiotic stress responsive elements and phytohormone-related elements were selected for analysis. We detected marked differences in the number, location, and type of CREs among the promoters of different *CsTALE* genes (**Supplementary Figure 3A**), and also observed the presence of two or three-tandem CREs, some of which may overlap with others. **Supplementary Figure 3B** presents details of the analyzed CREs, including the total number of each CRE type and the corresponding CREs in each gene. Among these, most *CsTALE* genes contain all types, with CREs involved in abscisic acid response occurring at the highest frequency.

Expression Profile of *CsTALE* Genes

Spatial patterns of gene expression can often provide valuable clues regarding gene function. Accordingly, to assess the potential

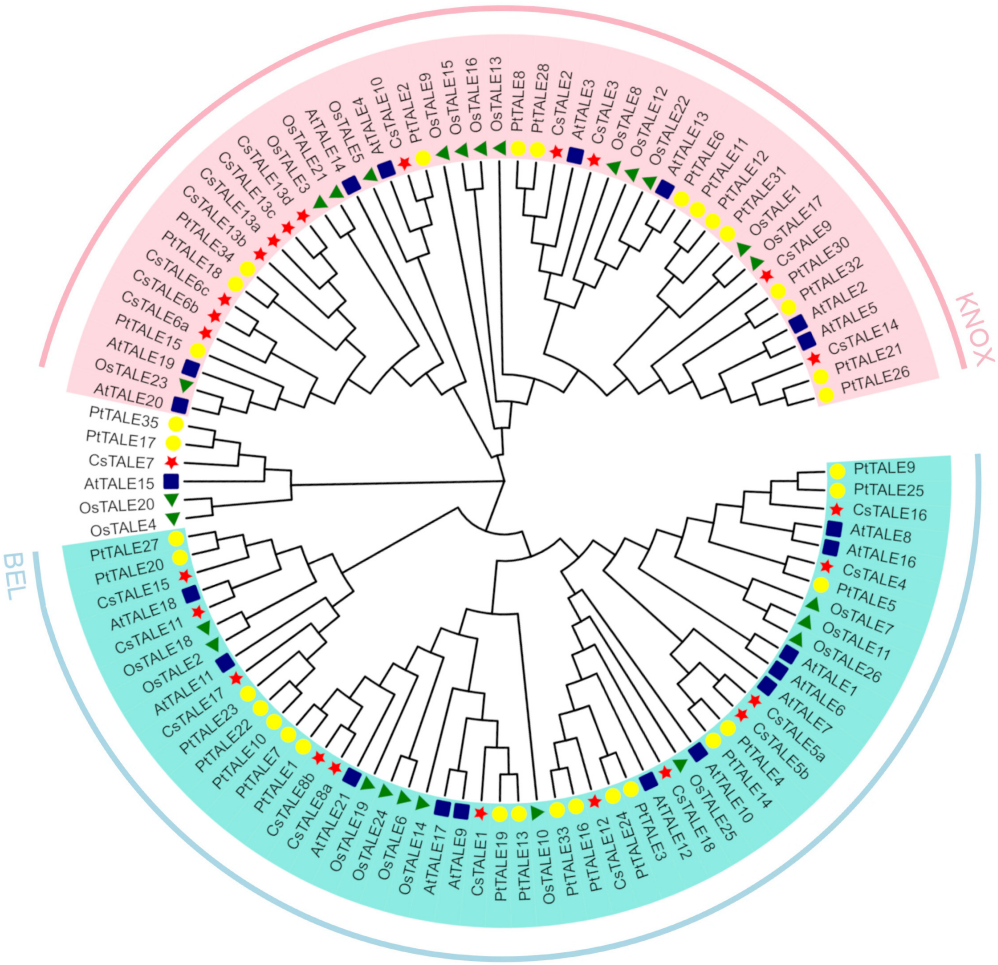

FIGURE 1 | Phylogenetic tree of *Citrus sinensis*, *Arabidopsis thaliana*, *Oryza sativa,* and *Populus trichocarpa* TALE proteins. The phylogenetic tree was constructed from amino sequences using MEGA-X v10.2.4 software by the neighbor-joining method with 1000 bootstrap replicates. The TALE proteins are clustered into 2 subgroups, marked by different colors. Red star indicates *Citrus sinensis* TALEs (*Cs*TALE), blue square indicates *Arabidopsis thaliana* TALEs (*At*TALE), green triangle indicates *Oryza sativa* TALEs (*Os*TALE), and yellow circle indicates *Populus trichocarpa* TALEs (*Pt*TALE).

functions of *CsTALE* genes in sweet orange development, we characterized the expression profiles of all 18 *CsTALE* genes in different tissues (stem, leaf, flower, and fruit) based on qRT-PCR analyses. Associated heat-maps revealed diverse patterns in the relative expression of *CsTALE* genes in different tissues (**Supplementary Figure 4**). In general, 11 *CsTALE* genes were found to be highly expressed in stems, whereas 14 show relatively low expression in fruit. *CsTALE13* is notably expressed at a high level in all examined tissues, and *CsTALE1/3/9/10* are highly expressed in stems, leaves, and flowers. In contrast, *CsTALE12/14* were observed to be weakly expressed in all tissues.

To identify those *CsTALE* genes that play a potential role in stress responses, we exposed *Citrus* seedlings to bacterial and fungal infection (CLas and *D. citri*, respectively) and abiotic stresses (high and low temperature, salt, and wounding), and examined the expression patterns of the 18 *CsTALE* genes at 0, 12, 24, and 48 h post-treatment using qRT-PCR (**Figure 4**). qRT-PCR analyses revealed that most of the *CsTALE* genes underwent changes in expression in response to different stresses over the course of the experiment. For example, *CsTALE7/8* were found to be induced by salt and high and low temperature treatments, whereas *CsTALE11/16* were induced in response to both salt and high temperature, and *CsTALE1/17* were induced by wounding or high temperature treatment. Intriguingly, some *CsTALE* genes showed significantly contrasting expression patterns in response to different stress types. For example, whereas *CsTALE1/2/10/11/16/17* were induced by a high temperature, their expression was significantly inhibited by exposure to a low temperature treatment. In contrast, *CsTALE6* was characterized by the converse pattern of expression.

We also investigated the expression of *CsTALE* in sweet orange infected with the fungal pathogen *D. citri* and bacterial pathogen CLas. **Figure 5A** shows the *CsTALE* genes differentially expressed in response to *D. citri* infection at 0, 24, and 48 h post-infection. Eight *CsTALE* genes were observed to be significantly up-regulated by *D. citri* inoculation, with greater

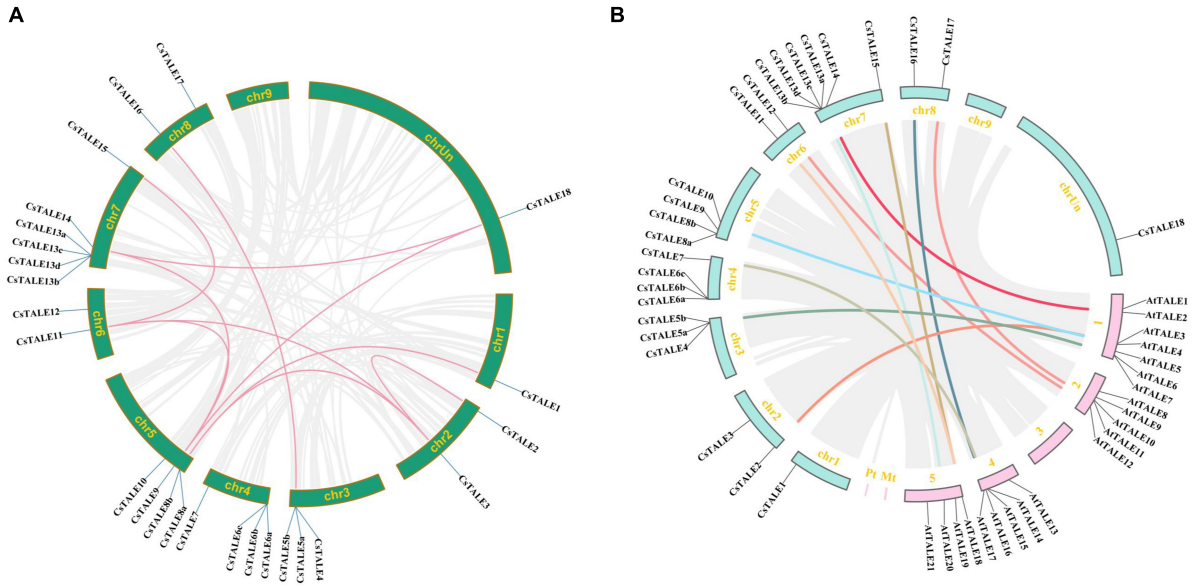

FIGURE 2 | Duplication events of CsTALE genes. **(A)** Circle plot displayed duplication events of *CsTALE* genes based on the sweet orange genome. Colored lines indicated duplication of *CsTALE* genes. **(B)** The homologous gene pairs between sweet orange and Arabidopsis. Gray lines indicate all the collinear blocks within sweet orange and Arabidopsis. Other colored lines indicate TALE homologous gene pairs.

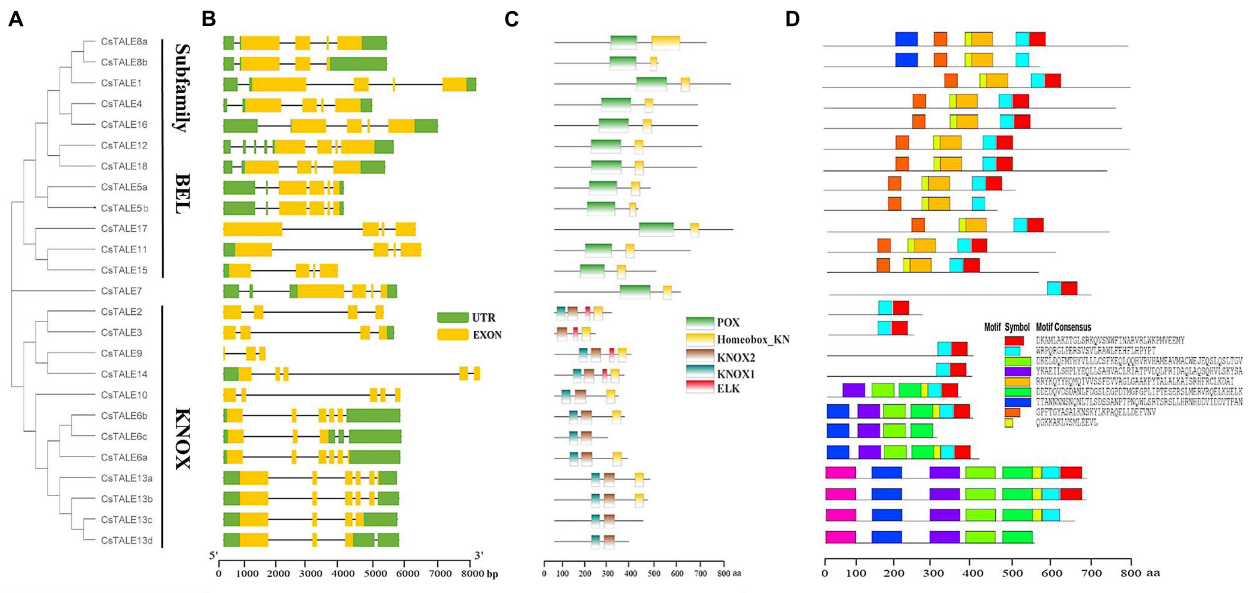

FIGURE 3 | Phylogenetic relationships and gene/protein structure of CsTALE. **(A)** NJ phylogenetic tree was created in MEGAX software according to the CsTALE amino acid sequence. **(B)** Gene structure. Exons, UTR and introns are indicated by yellow rounded rectangles, green rounded rectangles and black lines, respectively. **(C)** Conserved domains. Different domains are represented in different color patches. **(D)** Protein motif. Schematic diagram of the conserved motifs of CsTALE proteins and the motif name indicated in the bottom right corner. The scale bar at the bottom is used to estimate the sizes of protein structure and gene structure.

than threefold changes, among which, *CsTALE4/6/9/12/16* showed highest up-regulated expression at 24 h, whereas the expression of *CsTALE2/3/8* peaked at 48 h. Conversely, the expression of four BEL subfamily genes (*CsTALE11/15/17/18*) and one KNOX subfamily gene (*CsTALE10*) was markedly inhibited. The expression profiles of *CsTALE* genes in CLas-infected sweet orange revealed that most of these genes were up-regulated in CLas-infected plants compared with healthy plants, although exceptions were noted. Specifically, we detected no appreciable changes in the relative expression of *CsTALE4/5/9*,

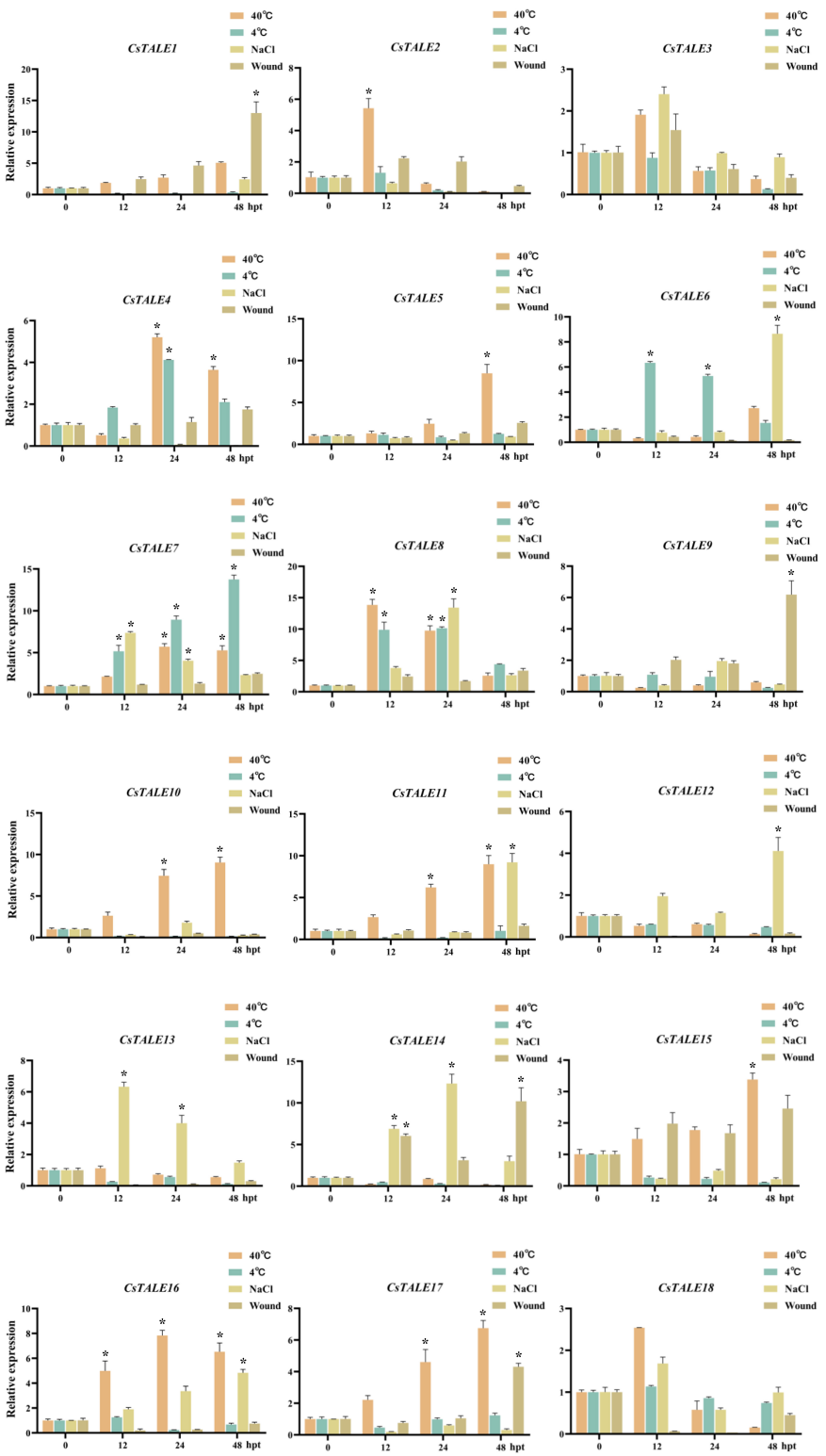

FIGURE 4 | Expression levels of *CsTALE* genes under different abiotic stress treatment. The *Y*-axis represent the relative expression level of *CsTALE* genes and the *X*-axis indicate different time points post abiotic stress treatment. Different colors represent different stress treatment. The standard errors are plotted using vertical lines. * Represents significant difference ($p < 0.05$). The experiments in all panels were repeated three times with similar results.

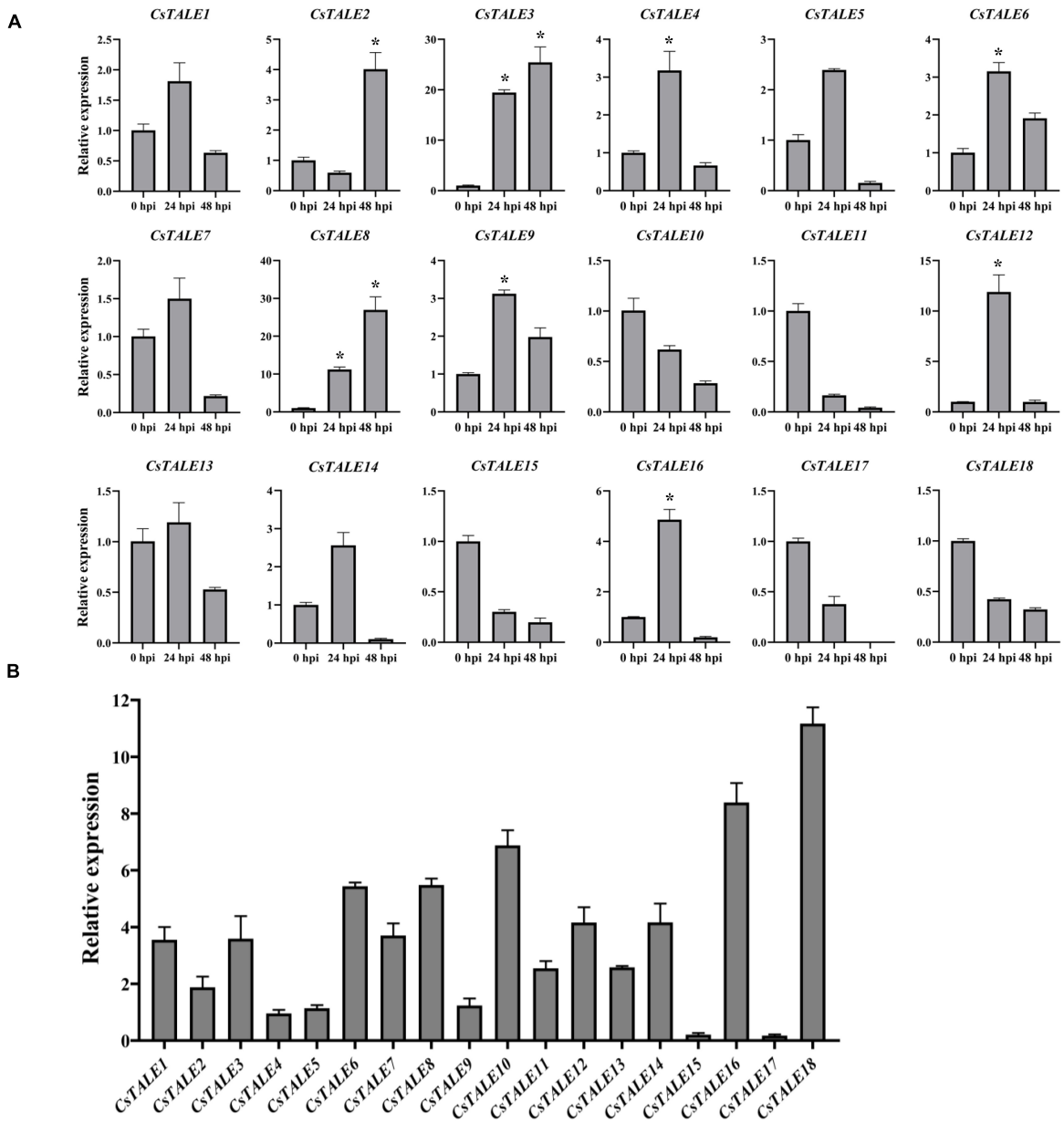

FIGURE 5 | Expression levels of *CsTALE* genes under different biotic stress. **(A)** The *Y*-axis represent the relative expression level of *CsTALE* genes and the *X*-axis indicate different time points post *Diaporthe citri* inoculation. **(B)** The *X*-axis represented the different *CsTALE* genes and the *Y*-axis represent the relative expression level after *Candidatus* Liberibacter asiaticus-infected. The gene transcription levels in CLas-free plants were normalized as 1. The standard errors are plotted using vertical lines. * Represents significant difference ($p < 0.05$). The experiments in all panels were repeated three times with similar results.

whereas the expression of *CsTALE15*/*17* appeared to be strongly suppressed (**Figure 5B**).

Subcellular Localization and Transcriptional Activation

To examine the subcellular localization of *Cs*TALE, we initially employed both the Plant-mPLoc and WoLF PSORT web-servers to predict subcellular localizations. Prediction results indicated that all these proteins are localized in the nucleus (**Supplementary Table 1**). To further substantiate these results, we generated 35S:TALE-GFP constructs and used these to observe transient expression in *N. benthamiana* leaves. This accordingly enabled us to confirm nuclear localization of the *Cs*TALE3/10 proteins (**Figure 6**).

In order to ascertain whether these TALE proteins have transcriptional activation, we constructed yeast expression vectors (pGBKT7-TALE), which were used to transform the AH109 yeast strain. We accordingly found that all transformed yeasts grew normally on the SD/–Trp medium

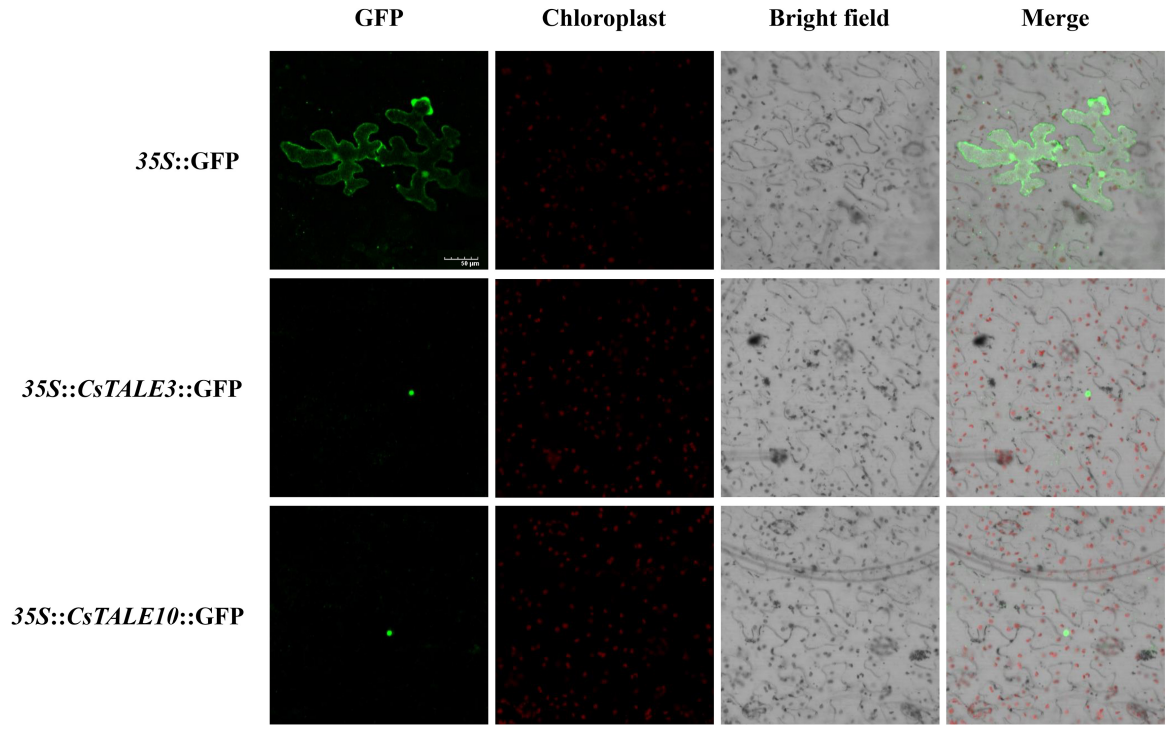

FIGURE 6 | Subcellular localization of CsTALE. The fusion constructs (35S:TALE-GFP) and GFP (control) were separately transiently expressed in tobacco leaves and visualized under a confocal laser scanning microscopy. Bar denotes 50 μm.

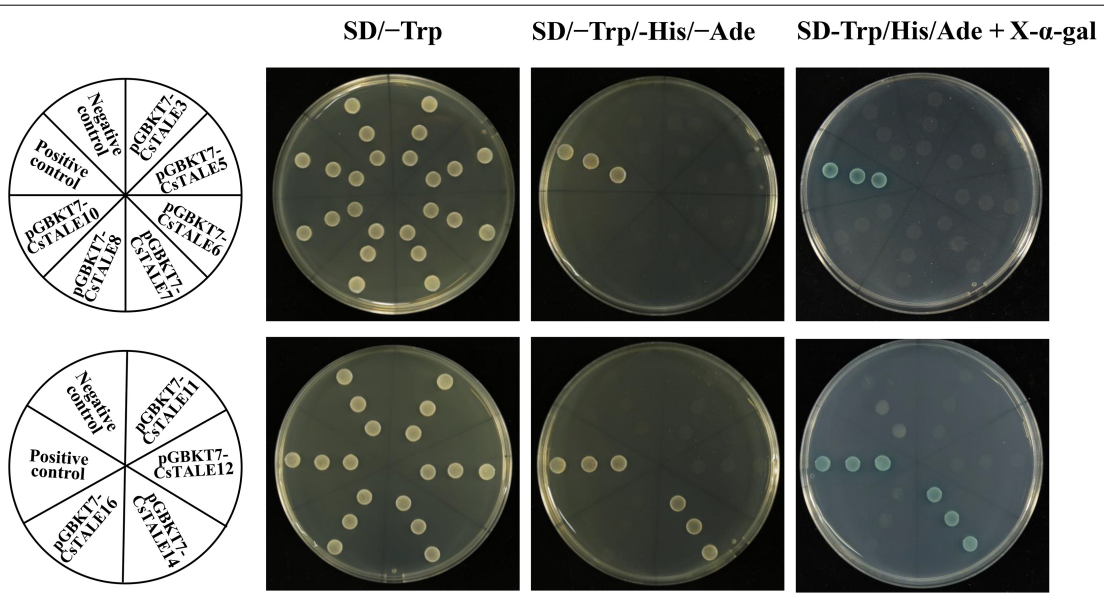

FIGURE 7 | Transcriptional activation of CsTALEs. Transactivational analyses. Fusion constructs (pGBKT7-CsTALE) and negative control (pGBKT7 empty vectors), were transformed into yeast AH109 strain and incubated in SD/-Trp, SD/-Trp/–His/-Ade and SD-Trp/His/Ade + X-α-gal medium.

(**Figure 7**). pGBKT7-*Cs*TALE14 co-transformed yeast cells grew well on SD/–Trp/–Leu medium and turned blue on X-α-gal-supplemented SD–Trp/His/Ade medium, thereby indicating that they have transcriptional activation ability. However, yeast transformed with pGBKT7-*Cs*TALE3/5/6/7/8/10/11/12/16 and the negative control pGBKT7 empty vector were only able to grow on the SD/–Trp, and were neither able grow nor turned blue on the SD–Trp/His/Ade + X-α-gal medium.

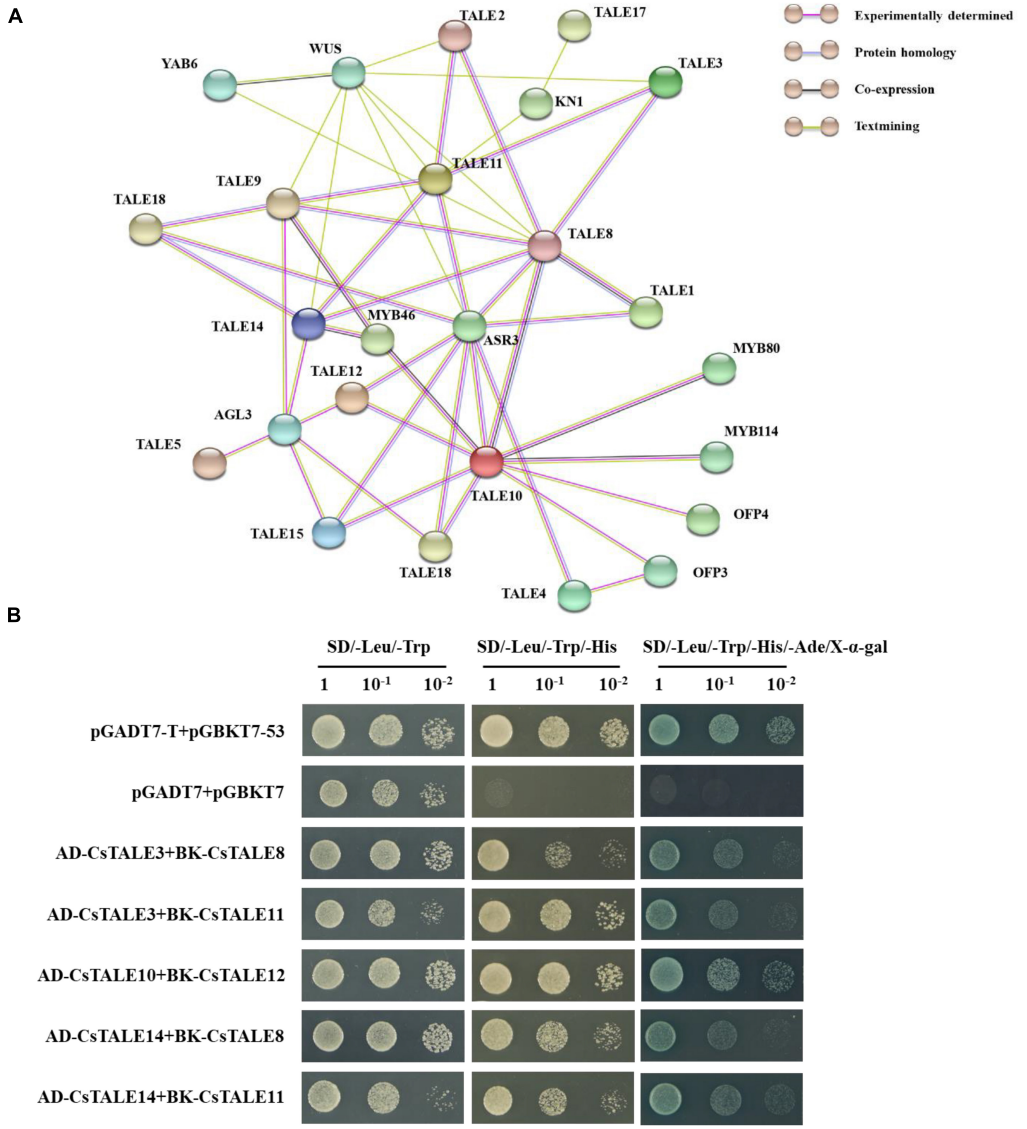

FIGURE 8 | Interaction network and protein interaction of CsTALEs. **(A)** Interaction network of TALE. Nodes represent proteins, and lines represent protein interaction pairs. Line colors represent different types evidence of protein interaction pairs. **(B)** Yeast two-hybrid assays. The co-transformed yeast cells were diluted to different concentrations (1, 10^{-1}, 10^{-2}) and cultured in SD/–Leu/Trp, SD/–Leu/Trp/His and SD/–Leu/Trp/His/Ade + X-α-gal medium.

These observations thus tended to indicate that pGBKT7-*Cs*TALE3/5/6/7/8/10/11/12/16 did not show the transcriptional activation in transformed yeast.

CsTALE Protein Interaction Network and Interaction Analysis

The web-based database for PPI networks, with predicted and known protein interactions, including direct (physical) and indirect (functional) associations, provides a valuable basis assessing the biological functions of uncharacterized proteins. The PPI network we constructed for *Cs*TALE proteins comprised 28 nodes and 49 edges, with an average node degree of 3.5 (**Figure 8A** and **Supplementary Table 6**). The network revealed that several *Cs*TALE proteins interact directly or indirectly with other *Cs*TALE members, among which *Cs*TALE8 and *Cs*TALE10 are predicted to interact with five and four *Cs*TALE proteins, respectively, and thus could represent the key connector proteins in the PPI network.

On the basis of this network, we performed Y2H assays to systematically assess the interactions between predicted pairwise *Cs*TALE proteins. These assays revealed that yeast co-transfected with *Cs*TALEs, including pGADT7-*Cs*TALE3/pGBKT7-*Cs*TALE8, pGADT7-CsTALE3/pGBKT7-CsTALE11, pGADT7-*Cs*TALE10/pGBKT7-CsTALE12, pGADT7-CsTALE14/pGBKT7-*Cs*TALE8, and pGADT7-*Cs*TALE14/pGBKT7-*Cs*TALE11 complex vectors, can grow well on SD/-Trp/–Leu, and SD/–Trp/–Leu/–His media and colonies turned blue on

SD/–Trp/–Leu/–His/–Ade medium supplemented with X-α-gal (**Figure 8B**). Moreover, the growth of all five recombinant yeast was comparable to that of the positive control and clearly distinct from negative controls (**Supplementary Figure 5**). These results unequivocally provide evidence to indicate the interaction between these pairs of *Cs*TALE proteins.

DISCUSSION

TALE family genes, which are widely distributed in both plant and animal genomes, play prominent roles in numerous cellular processes, growth, and stress responses (Wang et al., 2021). In recent years, benefiting from the notable advances in bioinformatics and genomic technologies, *TALE* family genes in *A. thaliana* (Bellaoui et al., 2001), *Solanum tuberosum* (Sharma et al., 2014), *Gossypium hirsutum* (Ma et al., 2019), and *G. max* (Wang et al., 2021) have been systematically studied and characterized. In contrast, there have been no comparable genome-wide studies and characterization of the *TALE* gene family in sweet orange. To rectify this deficiency, we performed a comprehensive integrative genomic analysis of *CsTALE* genes in sweet orange.

Different species have been found to vary considerably with respect to the number of TALE family members they harbor. In this study, we identified a total of 18 *CsTALE* family genes in sweet orange, which compares with the 35 in *Populus trichocarpa* (Zhao et al., 2019), 18 in *Ananas comosus* (Ali et al., 2019), and 14 in *Medicago truncatula* (Dolgikh et al., 2020), which could reflect differences in genome size and ploidy level. We established that the 18 *Cs*TALE proteins differ notably in terms of amino acid residues and physicochemical properties. In lines with expectations, as TFs, all identified *Cs*TALE members are predicted to be nuclear localized. Similar to the soybean and poplar TALEs, we found that the number of amino acid residues and MWs of KNOX subfamily *Cs*TALEs are considerably smaller than those in the BEL subfamily (Zhao et al., 2019; Wang et al., 2021).

In order to gain a better understanding of evolutionary relationships among the identified TALE proteins, we constructed a neighbor-joining phylogenetic tree, on the basis of which, the *Cs*TALE proteins were classified into two subfamily, BEL and KNOX, which is consistent with previously reported observations (Ruiz-Estévez et al., 2017). Notably, *Cs*TALE7 and *At*TALE15 were found to cluster in the branch of the phylogenetic tree separate from the other assessed TALES. Given that members of the same phylogenetic cluster are generally assumed have similar functions, we speculate that *Cs*TALE7 in sweet orange has a growth-related function comparable to that of *At*TALE15 (ATH1), which has been demonstrated to influence the growth of either vegetative or reproductive organs and represses stem development (Song et al., 2020). In addition, it has been established that the domains/motifs of BEL and KNOX subfamily proteins tend to show a strong subfamily specificity, which is also consistent with our classification results. Particular protein domains/motifs have been shown to contribute in defining the functionality of certain DNA binding and PPIs (Liu M. et al., 2019). For example, POX, a plant-specific domain found in BEL subfamily members, is reported to function in association with homeobox domains. In Arabidopsis, VAAMANA, a BEL1-like homeodomain protein, interacts specifically with KNAT6 and STM to promote appropriate inflorescence development (Bhatt et al., 2004). KNOX1 is known to exert potent effects in inhibiting the expression of downstream target genes, whereas KNOX2 has been found to be essential for homodimer formation, and a combination of both KNOX1 and KNOX 2 can form a MEINOX domain (Nagasaki et al., 2001). Furthermore, the ELK domain has been speculated to serve as a nuclear localization signal, as well as a PPI domain (Jia et al., 2020). Accordingly, we would anticipate the different domain/motif types of *Cs*TALE members might provide clues as to the distinct or specialized functions of these proteins, which thus should be examined by future studies. Gene structure analysis revealed that 13 and 11 of the 18 *CsTALE* genes contain 5'- and 3'-UTRs, respectively. Previous study has demonstrated that 5'-UTRs plays a role in regulating mRNA stability, whereas 3'-UTRs may function as miRNA binding sites, which would thus confer *CsTALE* genes with rich and complex properties with respect to the regulation of downstream genes (Peng et al., 2012). In summary, we detected similarities in protein/gene structures of *Cs*TALEs grouped within the same subfamily or clade, although structures typically differ between members of the different subfamilies. Hence, structural consistencies or discrepancies may also contribute to similarities or diversity in the function of *Cs*TALE members.

Common patterns of duplication events, including tandem, segmental, and genomic duplications, are among the most important factors influencing biological evolution and the expansion of different gene families in eukaryotic genomes (Peng et al., 2021a). Most *CsTALE* genes appear to have originated from segmental duplication, which has been established to be the main evolutionary driving force, followed by tandem duplication. By determining the ratios of *K*a to *K*s, we were able to characterize the evolutionary history and differentiation paths of *CsTALE* genes, which indicated that these genes have primarily evolved under the influence of purifying selection, and that KONX subfamily members appeared later than those in the BEL subfamily. These evolutionary patterns of *CsTALE* gene origin and divergence are similar to those reported for *G. max*, which thus tends to indicate that *TALE* gene families have evolutionarily conserved mechanisms and functions (Wang et al., 2021).

The correct identification of orthologous genes in extensively studied model plants may provide important clues as to the properties of newly discovered members (Kristensen et al., 2011). Accordingly, we sought to identify the biological functions of *CsTALE* genes based on comparative synteny analysis of these genes and *TALE* genes from the Arabidopsis genome. We detected a total of 15 collinear gene pairs between sweet orange and Arabidopsis, and can thus speculate that these paired genes may have originated from a single common ancestral gene and that their role may have been broadly conserved over the subsequent course of evolution. For example, *At*TALE21 (BEL1) and *At*TALE15 (ATH1) have been reported to play

roles in complex networks involved in early developmental stages of the inflorescence meristem (Smith and Hake, 2003), and AtTALE3 (KNAT7) and AtTALE15 (BLH6) have been demonstrated to influence secondary cell wall development by specifically interacting with one another (Liu et al., 2015). In addition, several reports have described certain functionally enriched homologous genes, including AtTALE3 (STM; Cole et al., 2006), AtTALE17 (BLH2; Xu et al., 2020), and AtTALE20 (KNAT3; Pagnussat et al., 2007). Although numerous genetic resources provide valuable insight into the molecular bases of different gene functions, further investigations are required to define the biological functions and associated molecular mechanisms for each candidate gene. In this regard, our CRE analysis revealed the potentially diverse roles of the identified *CsTALE* genes implicated in regulation of different biological processes in sweet orange, including responses to different phytohormones and stress.

The findings of previous studies have indicated that the transcript abundances of TALE family members differ considerably among different tissues, in which they perform different biological functions (Wuddineh et al., 2016). Thus, TALE gene expression patterns can provide important information regarding the function of candidate genes. Our subcellular localization analysis based on gene expression profiles indicated that most *CsTALE* genes are expressed in the stem at a high level of expression, thereby indicating that these genes may have certain tissue-specific properties and fulfill different functions in different tissues. For example, potato POTH1, a KNOX family protein, has been shown interact with seven BEL family proteins based on Y2H screening, and thereby regulates shoot tip cytokinin levels and tuber formation (Chen et al., 2003).

Various environmental stresses, both abiotic and biotic, can have pronounced detrimental effects that contribute to substantial reductions in citrus crop yields and productivity (Sun et al., 2019). In this regard, the findings of a recent study provide evidence to indicate that the *Gh*BLH7/*Gh*OFP complex in cotton functions as a negative regulator in regulating resistance to Verticillium wilt by inhibiting lignin biosynthesis and the JA signaling pathway (Ma et al., 2020). Furthermore, on the basis of regulatory network analyses, *TALE* family genes have been predicted to be key factors mediating resistance to bacterial spot disease in pepper (Zhu et al., 2021). However, there have been comparatively few studies that have investigated the involvement of *CsTALE* genes in biotic stress responses, although we assume that their roles in this respect have been underestimated. Thus, in the present investigation, we utilized qRT-PCR analysis and integrated the overall levels of gene expression profiles to assess the magnitude of the responses of all identified *CsTALE* genes to different biotic and abiotic stresses. We accordingly observed that in response to *D. citri* infection, 13 of the 18 *CsTALE* genes were upregulated and five were down-regulated. Interestingly, in conjunction with PPI network analysis, we found that the expression profiles of *CsTALE* genes in response to *D. citri* infection indicate that for two *Cs*TALE proteins with predicted interactions, one is up-regulated and the other is down-regulated (e.g., CsTALE14–CsTALE18–CsTALE9–CsTALE11–CsTALE3 and CsTALE8–CsTALE10–CsTALE12). Moreover, when we examined the expression profiles of *CsTALE* genes in response to both *D. citri* and CLas infection, we found that *CsTALE3*/*6*/*8*/*12*/*16* were significantly upregulated and *CsTALE15*/*17* were strongly suppressed, thus indicating that these genes might play conserved roles in sweet orange disease resistance, via either positive or negative regulation. Moreover, *CsTALE10*/*11*/*18* were significantly upregulated in response to CLas infection, although were strongly inhibited by *D. citri* infection, which indicates that these genes may play unique immunological roles in HLB resistance. Given that *CsTALE* genes respond to different abiotic and biotic stresses to varying degrees, we speculate that these genes may play a dynamic regulatory role in the stress-induced gene regulation network of sweet orange; however, the underlying mechanisms need to be further investigated.

To gain further insights into the functions of *Cs*TALE proteins, we proceeded to investigate the subcellular localization and transcriptional activation of these proteins. Consistent with the established characteristic of TFs, we observed that *Cs*TALE3/10 localize exclusively to the nucleus. Subsequently, transcriptional activity experiments indicated that *Cs*TALE14 has transcriptional activation activity and may thus regulate the coordinate expressions of downstream genes. In contrast, *Cs*TALE3/5/6/7/8/10/11/12/16 showed no comparable transcriptional activation, which may indicate that these proteins initially need to form complexes with partners to exert their transcriptional activation function.

Protein–protein interactions network analysis revealed the identity of several functional partners among *Cs*TALE members. With respect to non-TALE interacting partners, OFP and MYB family proteins have been the most frequently reported TALE-interacting proteins (Gong et al., 2014; Wang et al., 2015), which is consistent with the interactions depicted in our PPI networks. In Arabidopsis, interaction between KNOX and BEL subfamily members has repeatedly been reported and demonstrated to play a key role in growth and developmental processes. The most well-studied and representative example of this phenomenon is the formation a transcriptional activation complex among BEL1 and KNAT1, KNAT2, STM, and KNAT5 proteins (Bellaoui et al., 2001). In this context, it is worth noting that the interaction of Arabidopsis protein pairs orthologous to *Cs*TALE14/*Cs*TALE11 and *Cs*TALE14/*Cs*TALE8 have previously demonstrated, and that the former pair has clearly characterized functions in the regulation of inflorescence development (Bhatt et al., 2004; Ragni et al., 2008). Somewhat surprisingly, in the present study, we identified interactions between the pairs CsTALE10/CsTALE12, CsTALE3/CsTALE8, and CsTALE3/CsTALE11, which have not previously been reported and could thus be species-specific.

Collectively, our characterization of the interactions of *Cs*TALE proteins reveals a certain degree of conservation, as indicated by comparisons with the homologous proteins in Arabidopsis. Nevertheless, we also identified certain differences indicative of multiple novel regulatory mechanisms among the *Cs*TALE family genes in sweet orange. A determination of complete or near complete interaction networks in further studies will hopefully enable us to clarify these mechanisms.

CONCLUSION

In this study, we undertook a comprehensive and systematic analysis of the TALE family proteins in sweet orange. In total, 18 *CsTALE* genes were identified, which were unevenly distributed on nine chromosomes. We analyzed their phylogenetic relationships, duplication events, and protein/gene structures, and complemented these analyses with predictions of *cis*-acting regulatory elements and PPIs. In addition, we examined the expression of the 18 *CsTALE* genes in different tissues and in response to different abiotic and biotic stresses. Furthermore, yeast two-hybrid assays enabled us to determine the interaction between BEL and KNOX subfamily members. Taken together, the findings of this study yielded important new information that will provide a basis for further studies examining the roles of *CsTALE* genes in regulating sweet orange growth and stress tolerance, as well as contributing to future sweet orange breeding programs.

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

WP and YY performed the experiments, collected the data, and wrote the main manuscript text. JX and EP searched the literature and prepared the materials. SD and LD provided the value comments and analyzed the experimental data. TY and YW provided help in statistical and bioinformatics tools. DL and NS helped to typeset and proofread this manuscript. BW supervised the study, designed the experiments and assisted in editing the revisions of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Key Research and Development Project of China (2019YFE0104100), the Scientific Research Fund of Hunan Provincial Education Department (20B288), and the Youth Fund Project of Hunan Agricultural University (19QN30).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.814252/full#supplementary-material

Supplementary Figure 1 | Chromosomal location of *CsTALE* genes. Left bar represents chromosome length. The chromosome numbers are labeled on left of the chromosomes. Gene density was calculated based on the annotation information within a genomic region.

Supplementary Figure 2 | Syntenic relations of the TALE members among *Citrus sinensis* and three representative plant species. Light-colored lines in the background represents the collinear relationship within *Citrus sinensis* and other plant genomes, and the deep red lines represent the collinearity of CsTALEs. Cs stands for *Citrus sinensis*, At for *Arabidopsis thaliana*, Os for *Oryza sativa* and Pt for *Populus trichocarpa*.

Supplementary Figure 3 | Diagram of CREs in promoter sequences of *CsTALE* genes. **(A)** The box in different colors indicated different CREs. The description of the eight CREs were depicted on the right side. **(B)** The details of the CREs analysis statistics. Left bar chart showed the total number of each type CREs of the *CsTALE* genes. Upset plot showed the corresponding CREs of each gene. The black point indicated which sets are included in *CsTALE* genes.

Supplementary Figure 4 | Quantitative real time-PCR analysis of *CsTALE* gene expression levels in various sweet orange tissues. Heatmap showing the expression of *CsTALE* genes in different tissues. qRT-PCR analysis of *CsTALE* genes expression in different tissues. The heat map was generated on the basis of $\log_2$ normalized intensity value. The color bar from blue-to-red indicated expression levels from high to low.

Supplementary Figure 5 | The control of yeast two-hybrid analysis. pGBKT7-53/pGADT7-T were used as a positive control. TALE and the empty vector (pGADT7 or pGBKT7) were used as a negative control.

REFERENCES

Agostini, J. P., Bushong, P. M., Bhatia, A., and Timmer, L. W. (2003). Influence of Environmental Factors on Severity of Citrus Scab and Melanose. *Plant Dis.* 87, 1102–1106. doi: 10.1094/PDIS.2003.87.9.1102

Ali, H., Liu, Y., Azam, S. M., Ali, I., Ali, U., Li, W., et al. (2019). Genome Wide Identification and Expression Profiles of TALE Genes in Pineapple (*Ananas comosus* L). *Trop. Plant Biol.* 12, 304–317. doi: 10.1007/s12042-019-09232-4

Avivi, Y., Lev-Yadun, S., Morozova, N., Libs, L., Williams, L., Zhao, J., et al. (2000). Clausa, a Tomato Mutant with a Wide Range of Phenotypic Perturbations, Displays a Cell Type-Dependent Expression of the Homeobox Gene *LeT6/TKn2*. *Plant Physiol.* 124, 541–552. doi: 10.1104/pp.124.2.541

Bellaoui, M., Pidkowich, M. S., Samach, A., Kushalappa, K., Kohalmi, S. E., Modrusan, Z., et al. (2001). The Arabidopsis BELL1 and KNOX TALE Homeodomain Proteins Interact through a Domain Conserved between Plants and Animals. *Plant Cell* 13, 2455–2470. doi: 10.1105/tpc.010161

Bhatt, A. M., Etchells, J. P., Canales, C., Lagodienko, A., and Dickinson, H. (2004). VAAMANA—a BEL1-like homeodomain protein, interacts with KNOX proteins BP and STM and regulates inflorescence stem growth in Arabidopsis. *Gene* 328, 103–111. doi: 10.1016/j.gene.2003.12.033

Brian, L., Warren, B., McAtee, P., Rodrigues, J., Nieuwenhuizen, N., Pasha, A., et al. (2021). A gene expression atlas for kiwifruit (*Actinidia chinensis*) and network analysis of transcription factors. *BMC Plant Biol.* 21:121. doi: 10.1186/s12870-021-02894-x

Butenko, M. A., and Simon, R. (2015). Beyond the meristems: similarities in the CLAVATA3 and INFLORESCENCE DEFICIENT IN ABSCISSION peptide mediated signalling pathways. *J. Exp. Bot.* 66, 5195–5203. doi: 10.1093/jxb/erv310

Chaisiri, C., Liu, X.-Y., Lin, Y., Li, J.-B., Xiong, B., and Luo, C.-X. (2020). Phylogenetic Analysis and Development of Molecular Tool for Detection of *Diaporthe citri* Causing Melanose Disease of Citrus. *Plants* 9:329. doi: 10.3390/plants9030329

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, H., Rosin, F. M., Prat, S., and Hannapel, D. J. (2003). Interacting Transcription Factors from the Three-Amino Acid Loop Extension Superclass Regulate Tuber Formation. *Plant Physiol.* 132, 1391–1404. doi: 10.1104/pp.103.022434

Choe, S.-K., Ladam, F., and Sagerström, C. G. (2014). TALE Factors Poise Promoters for Activation by Hox Proteins. *Dev. Cell* 28, 203–211. doi: 10.1016/j.devcel.2013.12.011

Cole, M., Nolte, C., and Werr, W. (2006). Nuclear import of the transcription factor SHOOT MERISTEMLESS depends on heterodimerization with BLH proteins expressed in discrete sub-domains of the shoot apical meristem of *Arabidopsis thaliana*. *Nucleic Acids Res.* 34, 1281–1292. doi: 10.1093/nar/gkl016

Curtolo, M., de Souza Pacheco, I., Boava, L. P., Takita, M. A., Granato, L. M., Galdeano, D. M., et al. (2020). Wide-ranging transcriptomic analysis of *Poncirus trifoliata*, *Citrus sunki*, *Citrus sinensis* and contrasting hybrids reveals HLB tolerance mechanisms. *Sci. Rep.* 10:20865. doi: 10.1038/s41598-020-77840-2

de Paula Santos Martins, C., Pedrosa, A. M., Du, D., Gonçalves, L. P., Yu, Q., Gmitter, F. G., et al. (2015). Genome-Wide Characterization and Expression Analysis of Major Intrinsic Proteins during Abiotic and Biotic Stresses in Sweet Orange (*Citrus sinensis* L. Osb.). *PLoS One* 10:e0138786. doi: 10.1371/journal.pone.0138786

Dean, G., Casson, S., and Lindsey, K. (2004). KNAT6 gene of Arabidopsis is expressed in roots and is required for correct lateral root formation. *Plant Mol. Biol.* 54, 71–84. doi: 10.1023/B:PLAN.0000028772.22892.2d

Deng, H., Achor, D., Exteberria, E., Yu, Q., Du, D., Stanton, D., et al. (2019). Phloem Regeneration Is a Mechanism for Huanglongbing-Tolerance of "Bearss" Lemon and "LB8-9" Sugar Belle® Mandarin. *Front. Plant Sci.* 10:277. doi: 10.3389/fpls.2019.00277

Dolgikh, A. V., Rudaya, E. S., and Dolgikh, E. A. (2020). Identification of BELL Transcription Factors Involved in Nodule Initiation and Development in the Legumes *Pisum sativum* and *Medicago truncatula*. *Plants* 9:1808. doi: 10.3390/plants9121808

Duvaud, S., Gabella, C., Lisacek, F., Stockinger, H., Ioannidis, V., and Durinx, C. (2021). Expasy, the Swiss Bioinformatics Resource Portal, as designed by its users. *Nucleic Acids Res.* 49, W216–W227. doi: 10.1093/nar/gkab225

Gong, S.-Y., Huang, G.-Q., Sun, X., Qin, L.-X., Li, Y., Zhou, L., et al. (2014). Cotton KNL1, encoding a class II KNOX transcription factor, is involved in regulation of fibre development. *J. Exp. Bot.* 65, 4133–4147. doi: 10.1093/jxb/eru182

Hamant, O., and Pautot, V. (2010). Plant development: a TALE story. *C. R. Biol.* 333, 371–381. doi: 10.1016/j.crvi.2010.01.015

Hannapel, D. J., Sharma, P., and Lin, T. (2013). Phloem-mobile messenger RNAs and root development. *Front. Plant Sci.* 4:257. doi: 10.3389/fpls.2013.00257

Hudry, B., Thomas-Chollier, M., Volovik, Y., Duffraisse, M., Dard, A., Frank, D., et al. (2014). Molecular insights into the origin of the Hox-TALE patterning system. *Elife* 3:e01939. doi: 10.7554/eLife.01939

Iftikhar, Y., Rauf, S., Shahzad, U., and Zahid, M. A. (2016). Huanglongbing: pathogen detection system for integrated disease management – A review. *J. Saudi Soc. Agric. Sci.* 15, 1–11. doi: 10.1016/j.jssas.2014.04.006

Jia, P., Zhang, C., Xing, L., Li, Y., Shah, K., Zuo, X., et al. (2020). Genome-Wide Identification of the MdKNOX Gene Family and Characterization of Its Transcriptional Regulation in *Malus domestica*. *Front. Plant Sci.* 11:128. doi: 10.3389/fpls.2020.00128

Kristensen, D. M., Wolf, Y. I., Mushegian, A. R., and Koonin, E. V. (2011). Computational methods for Gene Orthology inference. *Brief. Bioinform.* 12, 379–391. doi: 10.1093/bib/bbr030

Letunic, I., Khedkar, S., and Bork, P. (2021). SMART: recent updates, new developments and status in 2020. *Nucleic Acids Res.* 49, D458–D460. doi: 10.1093/nar/gkaa937

Liu, M., Huang, L., Ma, Z., Sun, W., Wu, Q., Tang, Z., et al. (2019). Genome-wide identification, expression analysis and functional study of the GRAS gene family in Tartary buckwheat (*Fagopyrum tataricum*). *BMC Plant Biol.* 19:342. doi: 10.1186/s12870-019-1951-3

Liu, X., Li, D., Zhang, S., Xu, Y., and Zhang, Z. (2019). Genome-wide characterization of the rose (*Rosa chinensis*) WRKY family and role of RcWRKY41 in gray mold resistance. *BMC Plant Biol.* 19:522. doi: 10.1186/s12870-019-2139-6

Liu, R., Wu, M., Liu, H., Gao, Y., Chen, J., Yan, H., et al. (2021). Genome-wide identification and expression analysis of the NF-Y transcription factor family in *Populus*. *Physiol. Plant.* 171, 309–327. doi: 10.1111/ppl.13084

Liu, Y., You, S., Taylor-Teeples, M., Li, W. L., Schuetz, M., Brady, S. M., et al. (2015). BEL1-LIKE HOMEODOMAIN6 and KNOTTED ARABIDOPSIS THALIANA7 Interact and Regulate Secondary Cell Wall Formation via Repression of *REVOLUTA*. *Plant Cell* 26, 4843–4861. doi: 10.1105/tpc.114.128322

Ma, Q., Wang, N., Hao, P., Sun, H., Wang, C., Ma, L., et al. (2019). Genome-wide identification and characterization of TALE superfamily genes in cotton reveals their functions in regulating secondary cell wall biosynthesis. *BMC Plant Biol.* 19:432. doi: 10.1186/s12870-019-2026-1

Ma, Q., Wang, N., Ma, L., Lu, J., Wang, H., Wang, C., et al. (2020). The Cotton BEL1-Like Transcription Factor GhBLH7-D06 Negatively Regulates the Defense Response against *Verticillium dahliae*. *Int. J. Mol. Sci.* 21:7126. doi: 10.3390/ijms21197126

Mahajan, A., Bhogale, S., Kang, I. H., Hannapel, D. J., and Banerjee, A. K. (2012). The mRNA of a Knotted1-like transcription factor of potato is phloem mobile. *Plant Mol. Biol.* 79, 595–608. doi: 10.1007/s11103-012-9931-0

Mahajan, A. S., Kondhare, K. R., Rajabhoj, M. P., Kumar, A., Ghate, T., Ravindran, N., et al. (2016). Regulation, overexpression, and target gene identification of *Potato Homeobox 15* (*POTH15*) – a class-I *KNOX* gene in potato. *J. Exp. Bot.* 67, 4255–4272. doi: 10.1093/jxb/erw205

Maheshwari, Y., Selvaraj, V., Godfrey, K., Hajeri, S., and Yokomi, R. (2021). Multiplex detection of "*Candidatus* Liberibacter asiaticus" and *Spiroplasma citri* by qPCR and droplet digital PCR. *PLoS One* 16:e0242392. doi: 10.1371/journal.pone.0242392

Mondal, S. N., Vicent, A., Reis, R. F., and Timmer, L. W. (2007). Saprophytic Colonization of Citrus Twigs by *Diaporthe citri* and Factors Affecting Pycnidial Production and Conidial Survival. *Plant Dis.* 91, 387–392. doi: 10.1094/PDIS-91-4-0387

Müller, J., Wang, Y., Franzen, R., Santi, L., Salamini, F., and Rohde, W. (2001). *In vitro* interactions between barley TALE homeodomain proteins suggest a role for protein-protein associations in the regulation of *Knox* gene function: interactions between barley TALE homeodomain proteins. *Plant J.* 27, 13–23. doi: 10.1046/j.1365-313x.2001.01064.x

Nadakuduti, S. S., Holdsworth, W. L., Klein, C. L., and Barry, C. S. (2014). KNOX genes influence a gradient of fruit chloroplast development through regulation of GOLDEN2-LIKE expression in tomato. *Plant J.* 78, 1022–1033. doi: 10.1111/tpj.12529

Nagasaki, H., Sakamoto, T., Sato, Y., and Matsuoka, M. (2001). Functional Analysis of the Conserved Domains of a Rice KNOX Homeodomain Protein, OSH15. *Plant Cell* 13, 2085–2098. doi: 10.1105/tpc.010113

Pagnussat, G. C., Yu, H.-J., and Sundaresan, V. (2007). Cell-Fate Switch of Synergid to Egg Cell in *Arabidopsis eostre* Mutant Embryo Sacs Arises from Misexpression of the BEL1-Like Homeodomain Gene *BLH1*. *Plant Cell* 19, 3578–3592. doi: 10.1105/tpc.107.054890

Peng, W., Song, N., Li, W., Yan, M., Huang, C., Yang, Y., et al. (2021b). Integrated Analysis of MicroRNA and Target Genes in *Brachypodium distachyon* Infected by *Magnaporthe oryzae* by Small RNA and Degradome Sequencing. *Front. Plant Sci.* 12:742347. doi: 10.3389/fpls.2021.742347

Peng, W., Li, W., Song, N., Tang, Z., Liu, J., Wang, Y., et al. (2021a). Genome-Wide Characterization, Evolution, and Expression Profile Analysis of GATA Transcription Factors in *Brachypodium distachyon*. *Int. J. Mol. Sci.* 22:2026. doi: 10.3390/ijms22042026

Peng, Y., Soper, T. J., and Woodson, S. A. (2012). "RNase Footprinting of Protein Binding Sites on an mRNA Target of Small RNAs," in *Bacterial Regulatory RNA*, ed. K. C. Keiler (Totowa: Humana Press), 213–224. doi: 10.1007/978-1-61779-949-5_13

Qiu, W., Soares, J., Pang, Z., Huang, Y., Sun, Z., Wang, N., et al. (2020). Potential Mechanisms of AtNPR1 Mediated Resistance against Huanglongbing (HLB) in Citrus. *Int. J. Mol. Sci.* 21:2009. doi: 10.3390/ijms21062009

Ragni, L., Belles-Boix, E., Günl, M., and Pautot, V. (2008). Interaction of *KNAT6* and *KNAT2* with *BREVIPEDICELLUS* and *PENNYWISE* in *Arabidopsis* Inflorescences. *Plant Cell* 20, 888–900. doi: 10.1105/tpc.108.058230

Ruiz-Estévez, M., Bakkali, M., Martín-Blázquez, R., and Garrido-Ramos, M. (2017). Identification and Characterization of TALE Homeobox Genes in the Endangered Fern *Vandenboschia speciosa*. *Genes* 8:275. doi: 10.3390/genes8100275

Shahan, R., Li, D., and Liu, Z. (2019). Identification of genes preferentially expressed in wild strawberry receptacle fruit and demonstration of their promoter activities. *Hortic. Res.* 6:50. doi: 10.1038/s41438-019-0134-6

Sharma, P., Lin, T., Grandellis, C., Yu, M., and Hannapel, D. J. (2014). The BEL1-like family of transcription factors in potato. *J. Exp. Bot.* 65, 709–723. doi: 10.1093/jxb/ert432

Shu, Y., Tao, Y., Wang, S., Huang, L., Yu, X., Wang, Z., et al. (2015). GmSBH1, a homeobox transcription factor gene, relates to growth and development and involves in response to high temperature and humidity stress in soybean. *Plant Cell Rep.* 34, 1927–1937. doi: 10.1007/s00299-015-1840-7

Smith, H. M. S., and Hake, S. (2003). The Interaction of Two Homeobox Genes, *BREVIPEDICELLUS* and *PENNYWISE*, Regulates Internode Patterning in the Arabidopsis Inflorescence. *Plant Cell* 15, 1717–1727. doi: 10.1105/tpc.012856

Song, J., Chen, C., Zhang, S., Wang, J., Huang, Z., Chen, M., et al. (2020). Systematic analysis of the Capsicum ERF transcription factor family: identification of regulatory factors involved in the regulation of species-specific metabolites. *BMC Genomics* 21:573. doi: 10.1186/s12864-020-06983-3

Song, N., Cheng, Y., Peng, W., Peng, E., Zhao, Z., Liu, T., et al. (2021). Genome-Wide Characterization and Expression Analysis of the SBP-Box Gene Family in Sweet Orange (*Citrus sinensis*). *Int. J. Mol. Sci.* 22:8918. doi: 10.3390/ijms22168918

Song, X., Zhao, Y., Wang, J., and Lu, M.-Z. (2021). The transcription factor KNAT2/6b mediates changes in plant architecture in response to drought via down-regulating *GA20ox1* in *Populus alba* × *P. glandulosa*. *Int. J. Mol. Sci.* 72, 5625–5637. doi: 10.1093/jxb/erab201

Suh, J. H., Tang, X., Zhang, Y., Gmitter, F. G., and Wang, Y. (2021). Metabolomic Analysis Provides New Insight Into Tolerance of Huanglongbing in Citrus. *Front. Plant Sci.* 12:710598. doi: 10.3389/fpls.2021.710598

Sun, L., Nasrullah, Ke, F., Nie, Z., Wang, P., and Xu, J. (2019). Citrus Genetic Engineering for Disease Resistance: past, Present and Future. *Int. J. Mol. Sci.* 20:5256. doi: 10.3390/ijms20215256

Thapa, S. P., De Francesco, A., Trinh, J., Gurung, F. B., Pang, Z., Vidalakis, G., et al. (2020). Genome-wide analyses of *Liberibacter* species provides insights into evolution, phylogenetic relationships, and virulence factors. *Mol. Plant Pathol.* 21, 716–731. doi: 10.1111/mpp.12925

Wang, L., Yang, X., Gao, Y., and Yang, S. (2021). Genome-Wide Identification and Characterization of TALE Superfamily Genes in Soybean (*Glycine max* L.). *Int. J. Mol. Sci.* 22:4117. doi: 10.3390/ijms22084117

Wang, S., Li, E., Porth, I., Chen, J.-G., Mansfield, S. D., and Douglas, C. J. (2015). Regulation of secondary cell wall biosynthesis by poplar R2R3 MYB transcription factor PtrMYB152 in Arabidopsis. *Sci. Rep.* 4:5054. doi: 10.1038/srep05054

Welker, S., Pierre, M., Santiago, J. P., Dutt, M., Vincent, C., and Levy, A. (2021). A. Phloem transport limitation in Huanglongbing affected sweet orange is dependent on phloem-limited bacteria and callose. *Tree Physiol.* 38:tab134. doi: 10.1093/treephys/tpab134

Wu, G. A., Terol, J., Ibanez, V., López-García, A., Pérez-Román, E., Borredá, C., et al. (2018). Genomics of the origin and evolution of Citrus. *Nature* 554, 311–316. doi: 10.1038/nature25447

Wuddineh, W. A., Mazarei, M., Zhang, J.-Y., Turner, G. B., Sykes, R. W., Decker, S. R., et al. (2016). Identification and Overexpression of a Knotted1-Like Transcription Factor in Switchgrass (*Panicum virgatum* L.) for Lignocellulosic Feedstock Improvement. *Front. Plant Sci.* 7:520. doi: 10.3389/fpls.2016.00520

Xu, Q., Chen, L.-L., Ruan, X., Chen, D., Zhu, A., Chen, C., et al. (2013). The draft genome of sweet orange (*Citrus sinensis*). *Nat. Genet.* 45, 59–66. doi: 10.1038/ng.2472

Xu, Y., Wang, Y., Wang, X., Pei, S., Kong, Y., Hu, R., et al. (2020). Transcription Factors BLH2 and BLH4 Regulate Demethylesterification of Homogalacturonan in Seed Mucilage. *Plant Physiol.* 183, 96–111. doi: 10.1104/pp.20.00011

Xu, Y.-Y., Liu, S.-R., Gan, Z.-M., Zeng, R.-F., Zhang, J.-Z., and Hu, C.-G. (2021). High-Density Genetic Map Construction and Identification of QTLs Controlling Leaf Abscission Trait in *Poncirus trifoliata*. *Int. J. Mol. Sci.* 22:5723. doi: 10.3390/ijms22115723

Yan, C., Hu, Z., Nie, Z., Li, J., Yao, X., and Yin, H. (2021). CcBLH6, a bell-like homeodomain-containing transcription factor, regulates the fruit lignification pattern. *Planta* 253:90. doi: 10.1007/s00425-021-03610-7

Yao, T., Zhou, Y., Hu, J., Xiao, T., and Zhou, C. (2021). Genomic evolutionary relationship of SWEET genes and their responses to HLB disease and oxytetracycline treatment in Valencia sweet orange. *Biologia* 76, 1685–1689. doi: 10.1007/s11756-021-00745-6

Yoon, J., Cho, L.-H., Antt, H. W., Koh, H.-J., and An, G. (2017). KNOX Protein OSH15 Induces Grain Shattering by Repressing Lignin Biosynthesis Genes. *Plant Physiol.* 174, 312–325. doi: 10.1104/pp.17.00298

Yoon, J., Cho, L.-H., Kim, S. L., Choi, H., Koh, H.-J., and An, G. (2014). The BEL1-type homeobox gene *SH5* induces seed shattering by enhancing abscission-zone development and inhibiting lignin biosynthesis. *Plant J.* 79, 717–728. doi: 10.1111/tpj.12581

Yu, L., Liu, D., Chen, S., Dai, Y., Guo, W., Zhang, X., et al. (2020). Evolution and Expression of the Membrane Attack Complex and Perforin Gene Family in the Poaceae. *Int. J. Mol. Sci.* 21:5736. doi: 10.3390/ijms21165736

Zhao, K., Zhang, X., Cheng, Z., Yao, W., Li, R., Jiang, T., et al. (2019). Comprehensive analysis of the three-amino-acid-loop-extension gene family and its tissue-differential expression in response to salt stress in poplar. *Plant Physiol. Biochem.* 136, 1–12. doi: 10.1016/j.plaphy.2019.01.003

Zhao, M., Li, C., Ma, X., Xia, R., Chen, J., Liu, X., et al. (2020). KNOX protein KNAT1 regulates fruitlet abscission in litchi by repressing ethylene biosynthetic genes. *J. Exp. Bot.* 71, 4069–4082. doi: 10.1093/jxb/eraa162

Zhu, Q., Gao, S., and Zhang, W. (2021). Identification of Key Transcription Factors Related to Bacterial Spot Resistance in Pepper through Regulatory Network Analyses. *Genes* 12:1351. doi: 10.3390/genes12091351

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

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

frontiers in Plant Science

ORIGINAL RESEARCH
published: 21 January 2022
doi: 10.3389/fpls.2021.801061

Phylogenetic Analysis of the SQUAMOSA Promoter-Binding Protein-Like Genes in Four *Ipomoea* Species and Expression Profiling of the *IbSPLs* During Storage Root Development in Sweet Potato (*Ipomoea batatas*)

Haoyun Sun[1†], Jingzhao Mei[2†], Weiwei Zhao[1†], Wenqian Hou[1], Yang Zhang[1], Tao Xu[1], Shaoyuan Wu[1,2]* and Lei Zhang[1]**

[1] Jiangsu Key Laboratory of Phylogenomics and Comparative Genomics, School of Life Sciences, Jiangsu Normal University, Xuzhou, China, [2] Department of Biochemistry and Molecular Biology, 2011 Collaborative Innovation Center of Tianjin for Medical Epigenetics, Tianjin Key Laboratory of Medical Epigenetics, Key Laboratory of Immune Microenvironment and Disease (Ministry of Education), School of Basic Medical Sciences, Tianjin Medical University, Tianjin, China

As a major plant-specific transcription factor family, *SPL* genes play a crucial role in plant growth, development, and stress tolerance. The SPL transcription factor family has been widely studied in various plant species; however, systematic studies on *SPL* genes in the genus *Ipomoea* are lacking. Here, we identified a total of 29, 27, 26, and 23 *SPLs* in *Ipomoea batatas*, *Ipomoea trifida*, *Ipomoea triloba*, and *Ipomoea nil*, respectively. Based on the phylogenetic analysis of SPL proteins from model plants, the *Ipomoea* SPLs were classified into eight clades, which included conserved gene structures, domain organizations and motif compositions. Moreover, segmental duplication, which is derived from the *Ipomoea* lineage-specific whole-genome triplication event, was speculated to have a predominant role in *Ipomoea SPL* expansion. Particularly, tandem duplication was primarily responsible for the expansion of SPL subclades IV-b and IV-c. Furthermore, 25 interspecific orthologous groups were identified in *Ipomoea*, rice, *Arabidopsis*, and tomato. These findings support the expansion of *SPLs* in *Ipomoea* genus, with most of the *SPLs* being evolutionarily conserved. Of the 105 *Ipomoea SPLs*, 69 were predicted to be the targets of miR156, with seven *IbSPLs* being further verified as targets using degradome-seq data. Using transcriptomic data from aboveground and underground sweet potato tissues, *IbSPLs* showed diverse expression patterns, including seven highly expressed *IbSPLs* in the underground tissues. Furthermore, the expression of 11 *IbSPLs* was validated using qRT-PCR, and two (*IbSPL17/IbSPL28*) showed significantly increased expression during root development. Additionally, the qRT-PCR analysis revealed that six *IbSPLs* were strongly induced in the roots under phytohormone treatments,

OPEN ACCESS

Edited by:
Hai Du,
Southwest University, China

Reviewed by:
Moyang Liu,
Shanghai Jiao Tong University, China
Satyabrata Nanda,
Centurion University of Technology and Management, India

***Correspondence:**
Lei Zhang
leizhang@jsnu.edu.cn
Shaoyuan Wu
shaoyuanwu@outlook.com
Tao Xu
xutao_yr@126.com

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

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

Received: *24 October 2021*
Accepted: *17 December 2021*
Published: *21 January 2022*

Citation:
Sun H, Mei J, Zhao W, Hou W, Zhang Y, Xu T, Wu S and Zhang L (2022) Phylogenetic Analysis of the SQUAMOSA Promoter-Binding Protein-Like Genes in Four Ipomoea Species and Expression Profiling of the IbSPLs During Storage Root Development in Sweet Potato (Ipomoea batatas). Front. Plant Sci. 12:801061. doi: 10.3389/fpls.2021.801061

particularly zeatin and abscisic acid. Finally, the transcriptomic data of storage roots from 88 sweet potato accessions were used for weighted gene co-expression network analysis, which revealed four *IbSPLs* (*IbSPL16/IbSPL17/IbSPL21/IbSPL28*) clusters with genes involved in "regulation of root morphogenesis," "cell division," "cytoskeleton organization," and "plant-type cell wall organization or biogenesis," indicating their potential role in storage root development. This study not only provides novel insights into the evolutionary and functional divergence of the *SPLs* in the genus *Ipomoea* but also lays a foundation for further elucidation of the potential functional roles of *IbSPLs* on storage root development.

Keywords: *Ipomoea*, SPL transcription factor, evolutionary patterns, root development, expression profiles

INTRODUCTION

The SQUAMOSA promoter-binding protein-like (*SPL*) genes are of the plant-specific transcription factor families, which play fundamental roles in plant growth, development, and stress tolerance (Guo et al., 2008; Chen et al., 2010; Preston and Hileman, 2013; Chen et al., 2015; Wang and Wang, 2015). The *SPL* genes predominantly contain the SQUAMOSA promoter-binding (SBP) domain, which comprises three distinct motifs: two non-interleaved zinc-binding sites (Cys-Cys-Cys-His and Cys-Cys-His-Cys) and one nuclear localization signal (NLS) at the C-terminus (Guo et al., 2008). *SPL* genes (*AmSBP1* and *AmSBP2*) were first discovered in snapdragon (*Antirrhinum majus*), following their role in flower development (Klein et al., 1996). Since then, numerous homologs have been identified and characterized in model species, such as *Arabidopsis thaliana* (Cardon et al., 1999), rice (*Oryza sativa*) (Xie et al., 2006), and tomato (*Solanum lycopersicum*) (Salinas et al., 2012). With the increasing number of sequenced genomes, *SPL* members have been increasingly annotated and reported in non-model plants, such as apple (*Malus domestica*) (Li et al., 2013), *Jatropha curcas* (Yu et al., 2020), pepper (*Capsicum annuum*) (Zhang et al., 2016), poplar (*Populus trichocarpa*) (Li and Lu, 2014), and soybean (*Glycine max*) (Tripathi et al., 2017).

SPL genes in model plants have been well studied, showing functional divergence. In *A. thaliana*, AtSPL3, AtSPL9, and AtSPL10 have been reported to regulate root development (Yu et al., 2015; Barrera-Rojas et al., 2020). Similar roles have also been reported in rice (Shao et al., 2019) and apple (Xu et al., 2017). Various studies have further found that SPL proteins participate in vegetative and reproductive phase transitions, for example, AtSPL3/4/5 promotes flowering by directly inducing AP1, FUL and LFY expression, which are flowering integrator genes (Yamaguchi et al., 2009). Additionally, SPL proteins are involved in fruit development and grain yield. For example, LeSPL-CNR regulates cell wall disassembly and carotenoid biosynthesis during fruit ripening in tomato (Orfila et al., 2002; Manning et al., 2006); *OsSPL16* expression promotes cell division and grain filling, with positive results in grain width and yield in rice (Wang et al., 2015). *SPL* genes are also considered as miR156 targets, thus forming a functional miR156-SPL regulatory network (Wang and Wang, 2015). In the miR156-SPL network, AtSPL9 negatively affects anthocyanin accumulation (Gou et al., 2011), whereas OsSPL7 enhances disease resistance against bacterial blight (Liu et al., 2019). The functions of *SPL* genes have been comprehensively studied in *Arabidopsis* and other model plants; however, their functionality in *Ipomoea* are relatively scarce.

The genus *Ipomoea*, which includes 500–600 species, possesses the largest number of species in the family *Convolvulaceae* (Austin and Huáman, 1996). *Ipomoea* species are widely and globally distributed with great value in the fields of industry and agriculture (Austin and Huáman, 1996; Liu, 2017; Morita and Hoshino, 2018). For example, Japanese morning glory, *Ipomoea nil* (L.) Roth. (2n = 2x = 30), is cultivated as an ornamental plant due to its diverse flower color patterns (Morita and Hoshino, 2018). Sweet potato, *Ipomoea batatas* (L.) Lam. (2n = 6x = 90), is ranked as the seventh most important crop globally due to its strong adaptability, stable yields, and high nutritional value (Liu, 2017). The storage roots of sweet potato, which are mainly harvested, have significant nutrient content and yield (Zhang et al., 2020). The initiation and development of storage roots is known as a complex and genetically programmed process (Ravi et al., 2014). Although several studies have reported the formation and development of storage root at the morphological, physiological, and molecular level (Nakatani, 1991; Wang et al., 2005; Tanaka et al., 2008; Noh et al., 2010; Dong et al., 2019; Huan et al., 2020), the underlying mechanisms of storage root development have not yet been fully elucidated. Up to now, the genomes of four species (*I. batatas, Ipomoea trifida*, *Ipomoea triloba,* and *I. nil*) have been sequenced in *Ipomoea* (Hoshino et al., 2016; Yang et al., 2017; Wu et al., 2018). Among these species, *I. trifida* is the most closely related diploid to *I. batatas*, followed by *I. triloba* and *I. nil* (Wu et al., 2018). The reported haplotype-resolved genome assembly of *I. batatas* is of low-quality, making it difficult to accurately identify and characterize genes. Contrastingly, the genome assembly of the other three diploid relatives is of high-quality and can be used as robust references for *I. batatas*. Therefore, genome availability makes it possible to perform a genome-wide comparative analysis of *SPL* genes in *Ipomoea*.

In this study, genome-wide identification and characterization of *SPL* genes were performed in four publicly available *Ipomoea* species, including *I. batatas*, *I. trifida*, *I. triloba,* and *I. nil*. Following this, the phylogenetic relationships and evolutionary patterns of *SPL* genes were investigated in these

four species. *IbSPL* gene expression patterns were determined using transcriptome and qRT-PCR in different organs or under various hormone treatments. Finally, weighted gene co-expression network analysis (WGCNA) was used to construct the co-expression network and infer the putative functions for the *IbSPL* genes in the storage root of sweet potato. This work not only provides insights into the evolutionary conservation and diversification of *SPL* genes in the genus *Ipomoea* but also lays the foundation for further research on *IbSPL* genes related to storage root development in sweet potato.

MATERIALS AND METHODS

Identification of SQUAMOSA Promoter-Binding Protein-Like Genes in *I. nil, I. triloba, I. trifida, and I. batatas*

The genomes of four *Ipomoea* species (including *I. nil, I. triloba*, *I. trifida,* and *I. batatas*) were downloaded from the "*Ipomoea nil* Genome Project[1]" (Hoshino et al., 2016), "Sweetpotato Genomic Resource[2]" (Wu et al., 2018), and "Sweet potato genome browser[3]" (Yang et al., 2017), respectively. *SPL* genes in these four species were identified using the following three methods. First, the 16 *A. thaliana* SPL proteins[4] were used as queries to find *SPL* homologs using BLASTP program with a threshold of *e*-value < 1*e*-3. Second, the Hidden Markov Model (HMM) of the SBP (PF03110) domain was downloaded from the Pfam database (El-Gebali et al., 2019) and used to identify putative SPL proteins using the hmmsearch (El-Gebali et al., 2019) program. Third, all the candidate SPL proteins obtained from the BLASTP and hmmsearch analysis were submitted to the SMART (Letunic and Bork, 2018) and ScanProsite databases (de Castro et al., 2006) to confirm an SBP domain presence. Proteins lacking the SBP domain were excluded, while the remaining were considered as the SPL proteins. Additionally, a manual examination was performed on the structures of identified *IbSPLs* to correct genome assembly errors.

The BUSCA online software (Savojardo et al., 2018) was used to predict the subcellular localization of *Ipomoea* SPL proteins. An in-house Perl script was used to analyze the physical and chemical properties of *Ipomoea* SPL proteins, such as protein length, molecular weight (MW, kD) and isoelectric point (pI).

Phylogenetic Analysis of SQUAMOSA Promoter-Binding Protein-Like Proteins

Phylogenetic analysis was performed on the SPL proteins from *Chlamydomonas reinhardtii*, *A. thaliana*, *J. curcas*, *M. domestica*, *O. sativa*, *P. trichocarpa*, *S. lycopersicum,* and the four *Ipomoea* species using the following steps: first, MAFFT software (v7.45) (Katoh and Standley, 2013) was used to obtain full-length SPL proteins' multiple alignments; second, Gblocks program (v0.71b) (Castresana, 2000) was used to select conserved blocks from the multiple alignments; third, MEGA X software (Kumar et al., 2018) was used to construct a neighbor-joining phylogenetic tree with 1000 bootstrap replications, with the CRR1 protein from *C. reinhardtii* was set as an outgroup; finally, Evolview website was used to visualize the tree (Subramanian et al., 2019). Similarly for the *Ipomoea* SPL proteins, a neighbor-joining phylogenetic tree was constructed using the aforementioned.

Analysis of the Gene Structure, Protein Domain, and Motif

The exon/intron positions of all *Ipomoea SPL* genes were obtained from the downloaded GFF3 files of the genomic database. The domain organizations of *Ipomoea* SPL proteins were annotated based on the SMART database results (Letunic and Bork, 2018). The conserved sequences in each domain were shaded at four levels using GeneDoc. The motif compositions of *Ipomoea* SPL proteins were analyzed through MEME online database (Bailey et al., 2009), with the maximum number set to 10. Finally, the gene structure, domain organization, and motif composition were drawn using Tbtools (Chen et al., 2020).

Gene Duplication, Orthology, and Selection Analysis

MCScanX software (Wang et al., 2012) was used to identify collinear blocks within or between species to classify the *SPL* genes into five different types: singleton, dispersed, proximal, tandem, and segmental duplication. The synteny relationships of the collinearity blocks in each *Ipomoea* species were visualized using Circos (Krzywinski et al., 2009). OrthoMCL software (Li et al., 2003) was used to detect orthologous groups among the diverse *SPL* genes. For each orthologous gene pair, *Ks* (synonymous substitution rate), *Ka* (non-synonymous substitution rate), and *Ka/Ks* ratio (evolutionary constraint) were calculated using PAML (Yang, 2007).

Prediction of miR156-Targeted Genes

Publicly available datasets were used to identify miR156 sequences in *Ipomoea*. A total of 58 miRNA transcriptomes deposited in National Center for Biotechnology Information (NCBI) (Coordinators, 2018) (16 in PRJNA471495, 2 in PRJNA474012, 11 in PRJNA592001, 12 in PRJNA599544, 12 in PRJNA600587, and 5 in PRJNA638516) were collected (**Supplementary Table 5**; Kuo et al., 2019; Saminathan et al., 2019; Yang et al., 2020; Liu et al., 2021). Trimmomatic software (version 0.39) (Bolger et al., 2014) was used to filter the raw miRNA sequencing data, which removed low-quality reads and sequencing adaptors. Finally, using *I. batatas* as the reference genome, the miRDeep2 (version 1.1.4) pipeline (Kuang et al., 2019) was employed to identify miR156 sequences with default parameters.

For the *Ipomoea SPL* genes, miR156 target sites were predicted using the psRNATarget server (Dai et al., 2018) with default settings. The predicted miR156-SPL interactions in *I. batatas* were validated using five degradomes (one in PRJNA592001 and four in PRJNA600587) (Yang et al., 2020; Liu et al., 2021)

[1]http://viewer.shigen.info/asagao

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

[3]http://public-genomes-ngs.molgen.mpg.de/SweetPotato

[4]https://www.arabidopsis.org

downloaded from public databases (**Supplementary Table 5**). After filtering out the low-quality reads and sequencing adaptors, the CleaveLand4 pipeline (Addo-Quaye et al., 2009) was used to identify miR156 cleavage sites. The identified targets with categories 0–3 and p-values < 0.05 were considered to be reliable miR156 target genes.

Promoter Analysis of *Ipomoea* SQUAMOSA Promoter-Binding Protein-Like Genes

The 2000 bp sequence upstream of the start codon for all *Ipomoea SPL* genes was retrieved using an in-house Perl script and submitted to the PlantCARE program (Lescot et al., 2002) to predict *cis*-acting elements as previously described (Chen et al., 2019). The distribution of *cis*-acting elements in each promoter was determined using TBtools (Chen et al., 2020).

Plant Materials and Hormone Treatments

Sweet potato (*I. batatas* cv. Xuyu34) plants used in this study were provided by the Xuzhou Academy of Agricultural Sciences, Xuzhou, Jiangsu, China. According to institutional, national, and international guidelines, these samples do not require specific permissions for research purposes. The plants were grown in greenhouses on the campus of Jiangsu Normal University, Xuzhou, China. For organ-specific expression analysis, the tissues of young leaves, mature leaves, flowers, and roots (10, 20, 40, 60, 80, 90, and 100 DAT roots with 0.3, 2, 7, 25, 37, 52, and 60 mm in diameter, respectively) were collected. For hormone treatments, stems with 4–5 leaves were cut and planted in 1/8 Hoagland solution to initiate adventitious root development for 10 days. Then stem cuttings with similar growth conditions were chosen and planted in 1/8 Hoagland solution separately containing 100 μM abscisic acid (ABA), indole-3-acetic acid (IAA), zeatin (ZT), and methyl-Jasmonate (MeJA). Stem cuttings without any hormone treatment were set as a control. Adventitious roots from the stem cuttings were collected at 0, 6, 12, 24, and 48 h post the treatments. Three biological replicates were collected for each sample. All samples were frozen in liquid nitrogen and finally stored at –80°C for subsequent use.

RNA Extraction and qRT-PCR Analysis

For analyzing expression patterns of *IbSPL* genes in different tissues or under phytohormone treatment, total RNA for each sample was extracted using the RNApure Plant Kit (CWBio, Beijing, China), following the manufacturers' instructions. For investigating the miR156-SPL interactions, total RNA was extracted using TRIzol reagent (Invitrogen, CA, United States) according to the manufacturers' instructions. The first cDNA strand was synthesized from 1.0 μg total pure RNA using the HiFiScript cDNA Synthesis Kit (CWBio, Beijing, China). The reverse transcription primer and qRT-PCR primer for miR156 were designed as previous study described (Zhou et al., 2020). Gene-specific primers for each *IbSPL* gene were designed using primer3 (Untergasser et al., 2012). qRT-PCR was performed *via* the Bio-rad CFX Connect^{TM} Real-Time System (Bio-Rad, CA, United States) using 2 × Q3 SYBR qPCR Master Mix (Universal) premix (Tolo Biotechnology, Shanghai, China). *IbARF* gene was used as a reference gene for normalizing the expression levels (Park et al., 2012). The relative transcript abundance for each gene was calculated with mean ± SD of biological triplicate samples using the $2^{-\Delta\Delta CT}$ approach (Livak and Schmittgen, 2001). The primers used are listed in **Supplementary Table 10**.

Analysis of the Expression Patterns of *IbSPLs* Using Published Transcriptomic Data

To explore tissue- and developmental stage-specific expression patterns of *IbSPL* genes, publicly available transcriptome datasets from two previous studies (Ding et al., 2017; Wu et al., 2018; **Supplementary Table 8**) were used: one included eight different tissues from cultivar Xuzi3 and Yan252 under the BioProject accession number PRJCA000640 in National Genomics Data Center (NGDC) (National Genomics Data Center and Partners, 2020), and the other included eight different stages during root development from cultivar Beauregard under the BioProject accession number PRJNA491292 in NCBI (Coordinators, 2018). Transcriptome analysis was performed as described in our previous study (Zhang et al., 2020). The downloaded raw fastq files were filtered using Trimmomatic (version 0.39) (Bolger et al., 2014), and then were mapped to sweet potato genome Taizhong6 (Yang et al., 2017) using STAR (version 2.7.1a) software under the 2-pass mapping mode (Dobin et al., 2013). RSEM (Li and Dewey, 2011) was used to calculate Fragments Per Kilobase of transcript per Million mapped reads (FPKM) values for each gene. Finally, a heatmap was plotted based on the normalized expression values of 29 *IbSPL* genes using the pheatmap package in R.

Construction of Co-expression Networks Involving *IbSPL* and Other *I. batatas* Genes in Sweet Potato Storage Root

Transcriptomic datasets of mature storage roots of 88 sweet potato accessions were obtained from a previous study (**Supplementary Table 11**; Ding et al., 2017) under the BioProject accession number PRJCA000642 in NGDC (National Genomics Data Center and Partners, 2020). Weighted co-expression network construction and module detection were performed using the R package WGCNA (version 1.4.9) (Langfelder and Horvath, 2008) with the following parameters: power = 9, minModuleSize = 30, cutHeight = 0.25, and network module export weight threshold = 0.05. The sub-network was subsequently visualized using Cytoscape (Smoot et al., 2011). *eggNOG-mapper* (version 2) (Huerta-Cepas et al., 2017) was used to assign the functional annotation to sweet potato genes, and Clusterprofiler (Yu et al., 2012) was used to perform GO enrichment analysis for genes co-expressed with *IbSPLs* (adjusted P-value < 0.05).

Statistical Analysis

The qRT-PCR results were analyzed using ANOVA (one-way analysis of variance) followed by LSD test. Statistically significant differences at $p < 0.05$ are indicated using different letters.

RESULTS

Identification of SQUAMOSA Promoter-Binding Protein-Like Genes in Four *Ipomoea* Species

BLASTP and HMM were used to identify the *SPL* genes in *Ipomoea* species, while SMART and ScanProsite were used to validate the results. A total of 29, 27, 26, and 23 *SPL* genes were identified in *I. batatas* (*Ib*), *I. trifida* (*Itf*), *I. triloba* (*Itb*), and *I. nil* (*In*), respectively. The *Ipomoea SPL* genes were renamed according to their chromosomal location (**Supplementary Table 1**). The numbers of *SPL* genes and their total percentage in each species are displayed in **Figure 1A**. The results showed that the *I. nil* genome had the least number of *SPL* genes compared to the other three species.

Subcellular localization analysis showed the nuclear localization of most *Ipomoea* SPL proteins (94, 89.52%) (**Supplementary Table 1**), suggesting their critical role in regulatory functions. Furthermore, the physical and chemical properties of SPL proteins were significantly differed within species but exhibited similar patterns among the four species (**Figures 1B–D** and **Supplementary Table 1**). Moreover, amino acid numbers in the *Ipomoea* SPL proteins ranged from 103 (*InSPL9*) to 1141 (*IbSPL11*), the MWs varied between 12.01 (*InSPL9*) and 126.51 (*IbSPL11*) kDa, and the pIs ranged from 5.40 (*ItfSPL14*) to 10.55 (*InSPL15*).

Comparative Phylogenetic Analysis of *Ipomoea* SQUAMOSA Promoter-Binding Protein-Like Genes

The evolutionary relationship between the *Ipomoea SPL* genes was explored *via* a rooted neighbor-joining phylogenetic tree, which was constructed using 105 SPL proteins from four *Ipomoea* species and 124 SPL proteins from seven other plant species (*A. thaliana*, *J. curcas*, *M. domestica*, *O. sativa*, *P. trichocarpa*, *S. lycopersicum,* and *Chlamydomonas reinhardii*) (**Figure 2** and **Supplementary Table 2**). Based on the classification of SPLs from *A. thaliana, S. lycopersicum,* and *O. sativa* (Cardon et al., 1999; Xie et al., 2006; Salinas et al., 2012), *Ipomoea SPL* genes were classified into eight clades (I–VIII) (**Figures 1**, **2**). All clades had at least one *SPL* gene in each *Ipomoea* species, indicating *SPL* conservation across *Ipomoea* genomes. However,

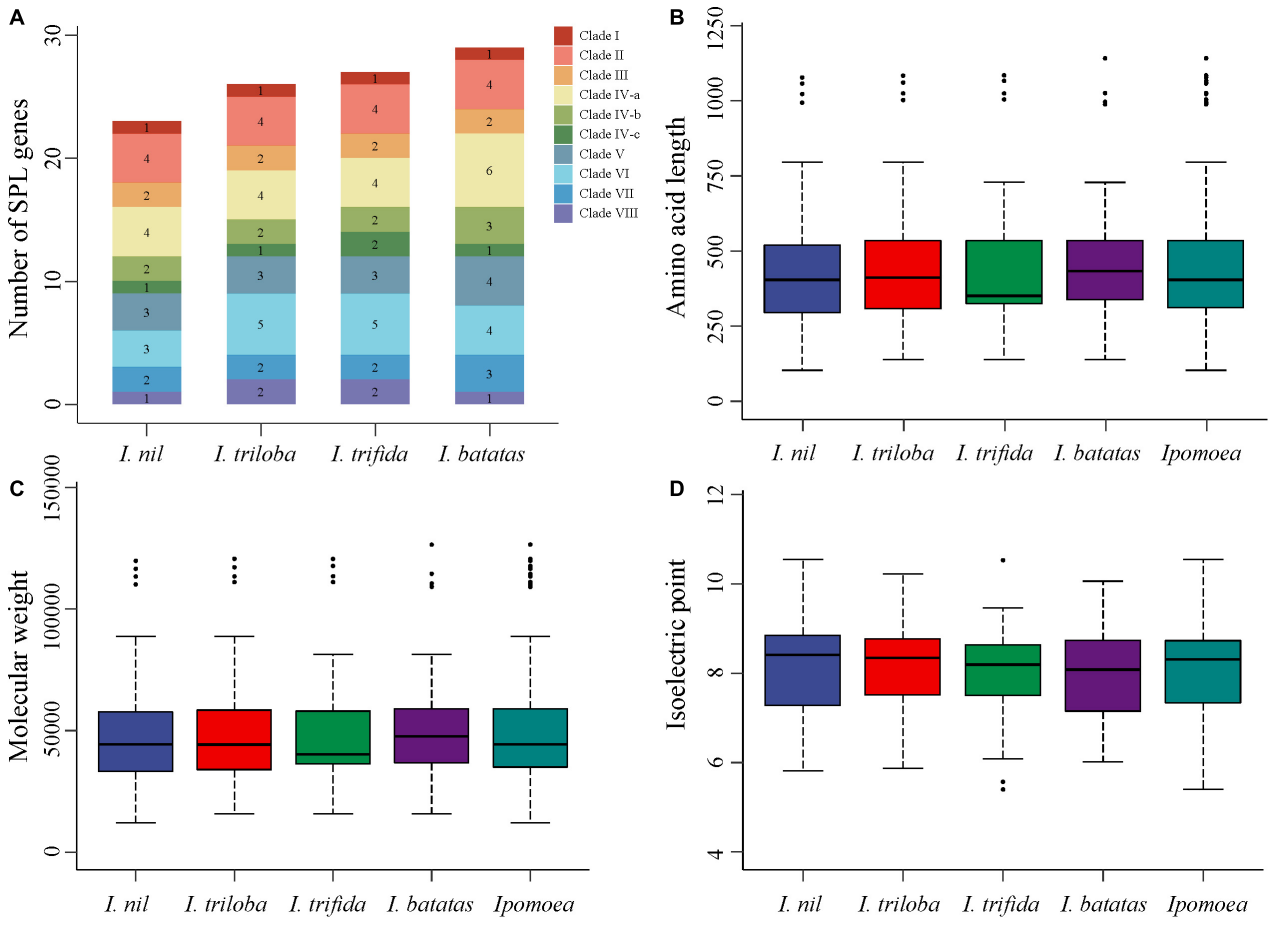

FIGURE 1 | Comparison of *SPL* genes in four *Ipomoea* species. **(A)** Comparison of the number and ratio of *SPL* genes in four *Ipomoea* species, with different colors representing each clade. **(B–D)** Comparison of SPL protein lengths, MWs, and pIs in four *Ipomoea* species, respectively.

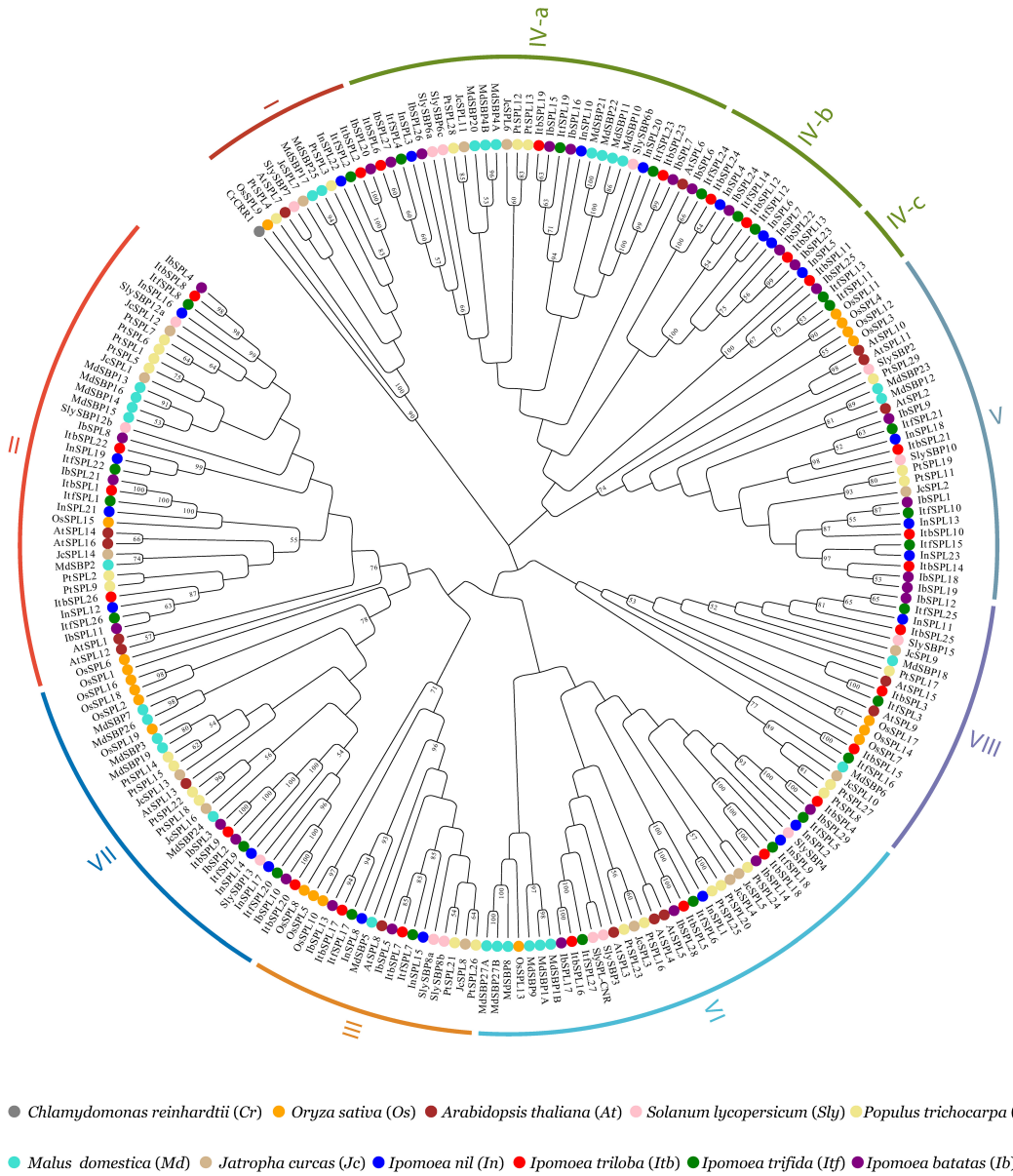

FIGURE 2 | Phylogenetic analysis of SPL proteins from *I. batatas*, *I. trifida*, *I. triloba*, *I. nil*, *C. reinhardtii*, *A. thaliana*, *J. curcas*, *M. domestica*, *P. trichocarpa*, *S. lycopersicum*, and *O. sativa*. The rooted phylogenetic tree was constructed based on the conserved domain of 229 SPLs using the neighbor-joining method with 1000 bootstrap replications. The CrCRR1 protein from *C. reinhardtii* was used as an outgroup to root the tree. Numbers on the tree indicate bootstrap support (values <50% not shown). Each colored arcs indicates the different clades of the SPLs. SPL members from the same species are marked with the same colors: Blue, *I. nil*; red, *I. triloba*; green, *I. trifida*; purple, *I. batatas*; gray, *C. reinhardtii*; orange, *O. sativa*; brown, *A. thaliana*; pink, *S. lycopersicum*; tan, *J. curcas*; turquoise, *M. domestica*; khaki, *P. trichocarpa*.

the number of *SPLs* in certain clades was highly variable among *Ipomoea* species, suggesting a diversity of *SPLs* in the genus *Ipomoea*. Clade I was the smallest subfamily, containing only one member for each *Ipomoea* species while clade IV had the highest number of *SPLs* (>26%) in genus *Ipomoea*, with further divisions into three subclades: IV-a, IV-b, and IV-c. Members of the IV-b and IV-c subclades only comprised *SPL* genes from the *Ipomoea* species and no homologs of other species, indicating that the *Ipomoea SPL* genes in these two subclades were evolutionary conserved. Additionally, the phylogenetic analysis also indicated that most *IbSPL* genes were closer to *ItfSPL* genes than either *ItbSPL* or *InSPL* genes, supporting the fact that *I. trifida* is the most closely related diploid to hexaploid sweet potato (Wu et al., 2018). Moreover, the number of *SPL* genes in *Ipomoea* species (the average number of *SPL* genes in the four *Ipomoea* species was 26) greatly increased by approximately 2 times compared to that in *S. lycopersicum* (13), respectively. These results indicate the extensive expansion of *Ipomoea SPL* genes after the speciation of *S. lycopersicum*.

Gene and Protein Structure of the *Ipomoea* SQUAMOSA Promoter-Binding Protein-Like Family

The structural diversity of the *Ipomoea SPL* genes was explored using intron/exon structure analysis (**Supplementary Figure 1**). Gene structure illustrations showed a high variation in the number of exons, ranging from 2 to 15 (**Supplementary Figure 1a** and **Supplementary Table 1**). For example, *IbSPL20* contained the highest exons (15), whereas most *Ipomoea SPL* genes in Clade VI contained the least exons (2). Moreover, most *Ipomoea SPL* genes in the same clade exhibited similar gene structures, despite belonging to different species. *SPL* gene in clades I and II contained the highest exons, ranging from 10 to 15, while *SPL* genes in the remaining clades had 2–7 exons (except *IbSPL19*). These results suggested that the gain or loss of exon/intron had occurred during the *Ipomoea SPL* gene evolution, resulting in their functional divergence.

Ipomoea SPL protein features were investigated by analyzing the conserved domains using multiple sequence alignment. The results showed that the SPL members had the SBP domain, which comprised two non-interleaved zinc finger-like structures (Zn-1/2) and one NLS motif (**Supplementary Figures 1c, 2**). Based on the alignments of the *Ipomoea* SPLs, the Zn-2 motif showed higher conservation than the Zn-1 and NLS motifs, which was consistent with the results in *Rosaceae* (Jiang et al., 2021) and *Oryza* species (Zhong et al., 2019; **Supplementary Figure 2**). The Zn-2 motif in all *Ipomoea* SPLs was a Cys-Cys-His-Cys (C2HC) type (except IbSPL23) whereas the Zn-1 motif showed varied types: Cys-Cys-Cys-Cys (C4) type in clade I and Cys-Cys-Cys-His (C3H) type in the remaining clades. Moreover, other conserved domains were identified in specific clades. For instance, SPLs in clade I and II possessed a DEXDc domain (**Supplementary Figure 3**), which is involved in ATP-dependent DNA unwinding (Caruthers and McKay, 2002); SPLs in clade II contained Ankyrin repeats (**Supplementary Figure 4**), which are considered to be significant for mediating protein-protein interactions (Li et al., 2006).

To gain a better understanding of *Ipomoea* SPL protein characteristics, the MEME software was used to explore the motif compositions (**Supplementary Figures 1d, 5**). The results showed that *Ipomoea* SPL proteins within the same clade showed similar motif compositions, while those in different clades exhibited distinct variations in motif composition. In brief, all *Ipomoea* SPL proteins had two motifs (motif 2 and 3), which were a part of the SBP domain (**Supplementary Figure 2**); clade I and II had four motifs (motif 4, 5, 6, and 7), with motif 4 being the DEXDc domain (**Supplementary Figure 3**);clade II had motif 8, which consisted of Ankyrin repeats (**Supplementary Figure 4**); clade IV had motif 10, which had unknown functions.

Gene Duplication, Orthology Relationship, and Selective Pressure of *Ipomoea* SQUAMOSA Promoter-Binding Protein-Like Genes

To investigate the gene duplication modes of *Ipomoea SPL* genes, MCScanX (Wang et al., 2012) was used to perform gene collinearity analysis in each *Ipomoea* species. All *Ipomoea SPL* genes were estimated to be duplicated (absence of singleton mode), with segmental (51, 48.57%) mode as the dominant mode compared to the other duplication modes: dispersed (39, 37.14%), tandem (13, 12.38%), and proximal (2, 1.90%) (**Figures 3A,B**; **Supplementary Table 1**; **Supplementary Figure 6**). These results indicate that segmental duplication has played a predominant role in the evolution and expansion of *Ipomoea SPL* genes. Additionally, tandem duplication was found to be the predominant model in the IV-b and IV-c subclades, suggesting the expansion of *SPL* genes in these two clades *via* tandem duplication.

The orthologous relationships among the *SPL* genes were determined using OrthoMCL (Li et al., 2003) across *O. sativa*, *A. thaliana*, *S. lycopersicum,* and the four *Ipomoea species*. A total of 25 (1, 2, 2, 8, 3, 5, 2, and 2) orthologous groups in the eight clades (clade I to VIII) were identified, respectively (**Supplementary Table 3**). Among these groups, nine groups had genes originating from *O. sativa*, suggesting that these *SPL* genes may have originated prior to the split of monocots and dicots; four groups had genes from *A. thaliana* but were absent in *O. sativa*, implying that they originated after the divergence of monocots and dicots; ten groups had genes that existed only in the genus *Ipomoea*, indicating their origination *via* a common ancestor of the *Ipomoea* lineage. Furthermore, the potential functions of certain *Ipomoea SPL* genes could be inferred from their orthologs in *O. sativa*, *A. thaliana,* and *S. lycopersicum*.

To understand the divergence of *Ipomoea SPL* genes, the *Ka*, *Ks*, and *Ka/Ks* ratios for all orthologous groups were calculated using PAML software (Yang, 2007; **Figure 3C**; **Supplementary Table 4**). As a result, the mean *Ka/Ks* values of all clades were lower than 1.0, suggesting the evolution of *Ipomoea SPL* genes under the pressure of purifying selection. Genes in clade VIII showed the lowest mean *Ka/Ks* values (0.15) compared to those in the other clades, indicating their evolution under strong positive selection. Contrastingly, genes in subclades IV-b and IV-c exhibited the highest *Ka/Ks* values, implying that these two subclades have generally diverged much more rapidly than the other clades.

miR156 Target Site of *Ipomoea* SQUAMOSA Promoter-Binding Protein-Like Genes

A total of nine IbmiR156 members were identified in *I. batatas* (**Figure 4A**) using the publicly available miRNA transcriptomes (**Supplementary Table 5**). To explore the roles of miR156-mediated post-transcriptional regulation of *SPLs* in the genus *Ipomoea*, the transcripts of all the 105 *Ipomoea SPL* genes were searched for the target site of miR156 using psRNATarget (Dai et al., 2018). As a result, a total of 69 *SPL* genes were found to be potential miR156 targets, including 10 *InSPLs*, 19 *ItbSPLs*, 18 *ItfSPLs,* and 22 *IbSPLs* (**Figure 4B** and **Supplementary Table 1**). Among the miR156 target *SPL* genes, most of which (84%) the sites recognized by miR156 were located downstream of the SBP domain in the CDS region, then followed by the 3′-UTR (**Figure 4B**).

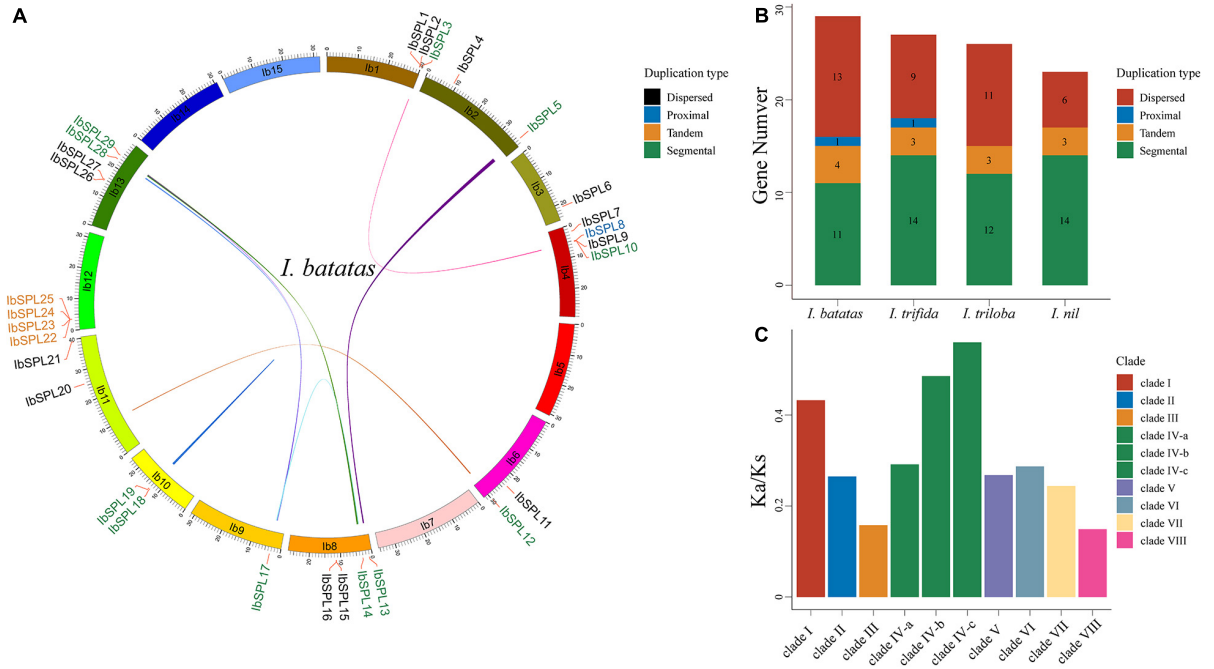

FIGURE 3 | Syntenic relationships and selective pressure of *Ipomoea* SPLs. **(A)** The collinear relationship of *SPL* genes in *I. batatas*, wherein the colored lines indicate syntenic regions. *IbSPL* genes are classified into four duplicated types, each indicated by a different color. **(B)** The distribution of *SPLs* duplication modes in four *Ipomoea* species. **(C)** The mean *Ka/Ks* values in each clade.

The predicted miR156-SPL interactions in *I. batatas* were validated using publicly available degradomes (**Supplementary Table 5**). The results showed that seven miR156-SPL interactions predicted by psRNATarget were confirmed by degradome sequencing (**Figure 4C** and **Supplementary Table 1**). Notably, the miR156 target sites of *IbSPL9*, *IbSPL10*, *IbSPL12*, and *IbSPL15* were located in the CDS region, while the target sites of *IbSPL16*, *IbSPL17*, and *IbSPL28* were located in the 3′-UTR region. The four IbmiR156-IbSPL pairs (IbmiR156b-IbSPL10, IbmiR156d-IbSPL17, IbmiR156e-IbSPL9, and IbmiR156h-IbSPL15) validated through degradomes data, were further selected for expression analysis by qRT-PCR. To understand the regulatory mechanisms of selected miR156 genes, correlation in the expression pattern of miR156 and their target *SPLs* was determined in different tissues (**Figure 4D**). The expression pattern of IbmiR156e was higher in flower, followed by mature leaf, young leaf, 10DAT root, 60DAT root, 20DAT root, and stem; conversely, the opposite trend was observed for its target IbSPL9. The expression of IbmiR156d and IbSPL17 also showed negative correlation in different tissues. While the expression of IbmiR156b-IbSPL10 and IbmiR156h-IbSPL15 were partially negatively correlated in some tissues.

Cis-Acting Elements in the Promoters of *Ipomoea* SQUAMOSA Promoter-Binding Protein-Like Genes

To understand the regulatory mechanisms and potential functions of *Ipomoea SPL* genes, *cis*-acting elements were analyzed in the 2000 bp upstream sequence from the start codon for all *SPL* genes by using the PlantCARE database (Lescot et al., 2002). A total of 4088 putative *cis*-acting elements were identified and divided into four categories: light responsiveness, plant growth, phytohormone, and abiotic/biotic stress response (**Supplementary Tables 6, 7**). As shown in **Supplementary Figure 7a**, *SPL* genes in the same clades exhibited similar *cis*-acting element compositions in the promoter, indicating their conserved biological functions. Among these categories, the abiotic/biotic stress response category covered the largest portion (43.66%), followed by the light response (25.93%), phytohormone response (20.77%), and plant growth (9.64%) categories (**Supplementary Figure 7b**). In the abiotic/biotic stress response category, MYB/MYC (responds to abiotic stress signals), STRE (metal-responsive element), WUN-motif (wound-responsive element), and LTR (low-temperature-responsive element) elements were found. In the light responsiveness category, Box 4, G-box, GT1-motif, TCT-motif, GATA-motif, and MRE elements were found, with the Box 4 motif as the most common (24%) element. As for the phytohormone response category, the ABRE element (responds to ABA), CGTCA-motif (responds to MeJA), ERE (responds to ethylene), TCA-element (responds to salicylic acid), and as-1 (responds to auxin) were commonly found, appearing in more than 60 *Ipomoea SPL* genes. In the plant growth category, ARE elements essential for the anaerobic induction, CAT-box related to meristem expression and O2-site involved in zein metabolism regulation were the three major elements. Therefore, analysis of the *cis*-acting elements

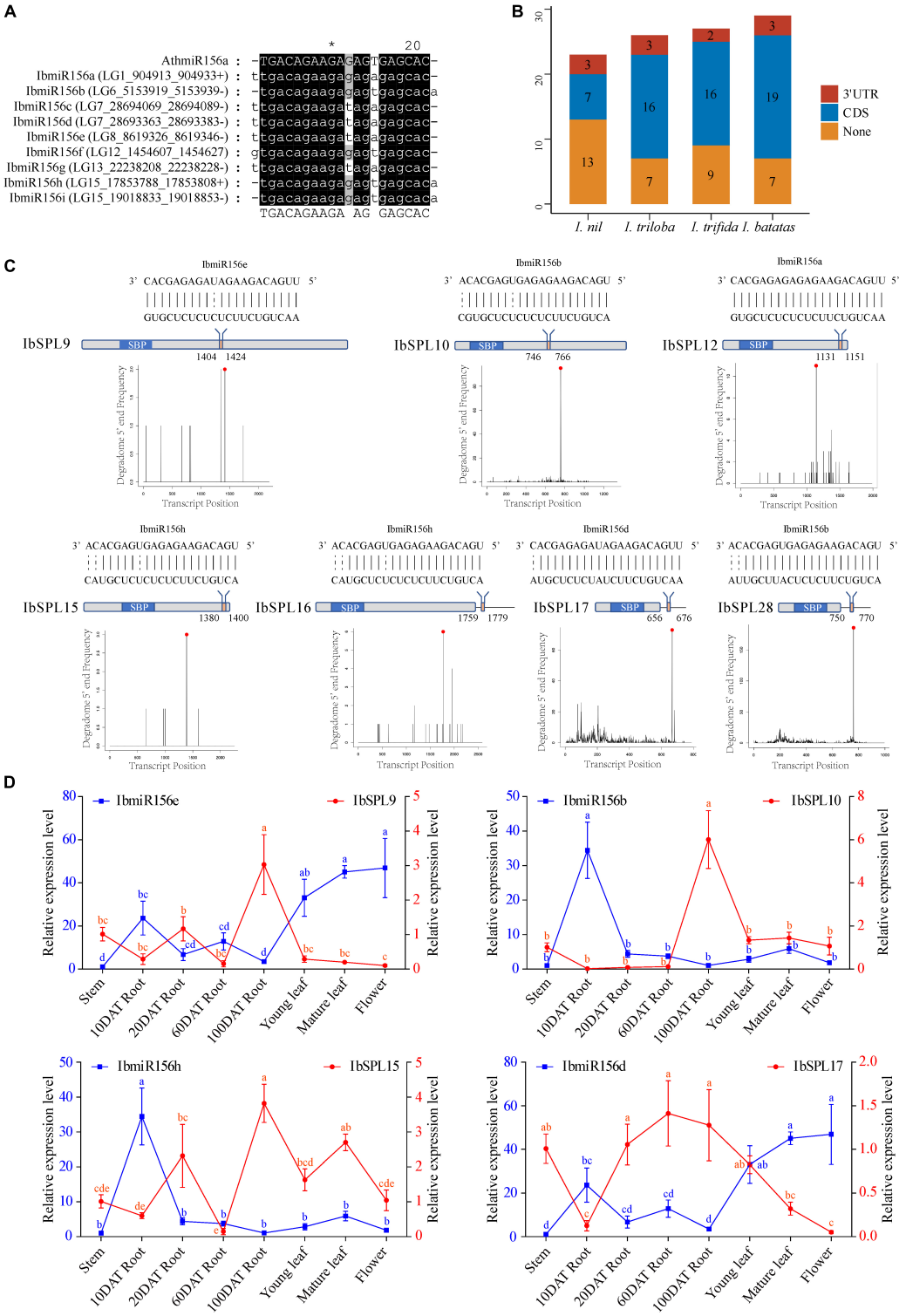

FIGURE 4 | miR156 target site of the *Ipomoea SPL* genes. **(A)** Multiple alignments of identified *IbmiR156* sequences. **(B)** The summary of miR156-targeted *SPL* genes in the *Ipomoea* genus. **(C)** Schematic diagram of *IbSPL* genes targeted by *IbmiR156*. The top diagram represents the complementary sequences between *IbmiR156* and their targets. The gray box indicates the CDS region, the blue box represents the SBP domain, and the red box indicates the *IbmiR156* target site. The diagram below is the target plots (t-plots) of *IbmiR156* targets confirmed by degradome sequencing. **(D)** Expression correlation between *IbmiR156* and *IbSPLs*. The lines represent the abundance of *IbmiR156* and *IbSPLs* in different tissues. The Y-axis on the left and right indicates the relative expression levels of the *IbmiR156* and *IbSPLs*, respectively. Relative expression was calculated using the $2^{-\Delta\Delta CT}$ method. The error bars indicate the standard deviations of the three biological replicates. Different letters indicate statistically significant differences at $p < 0.05$.

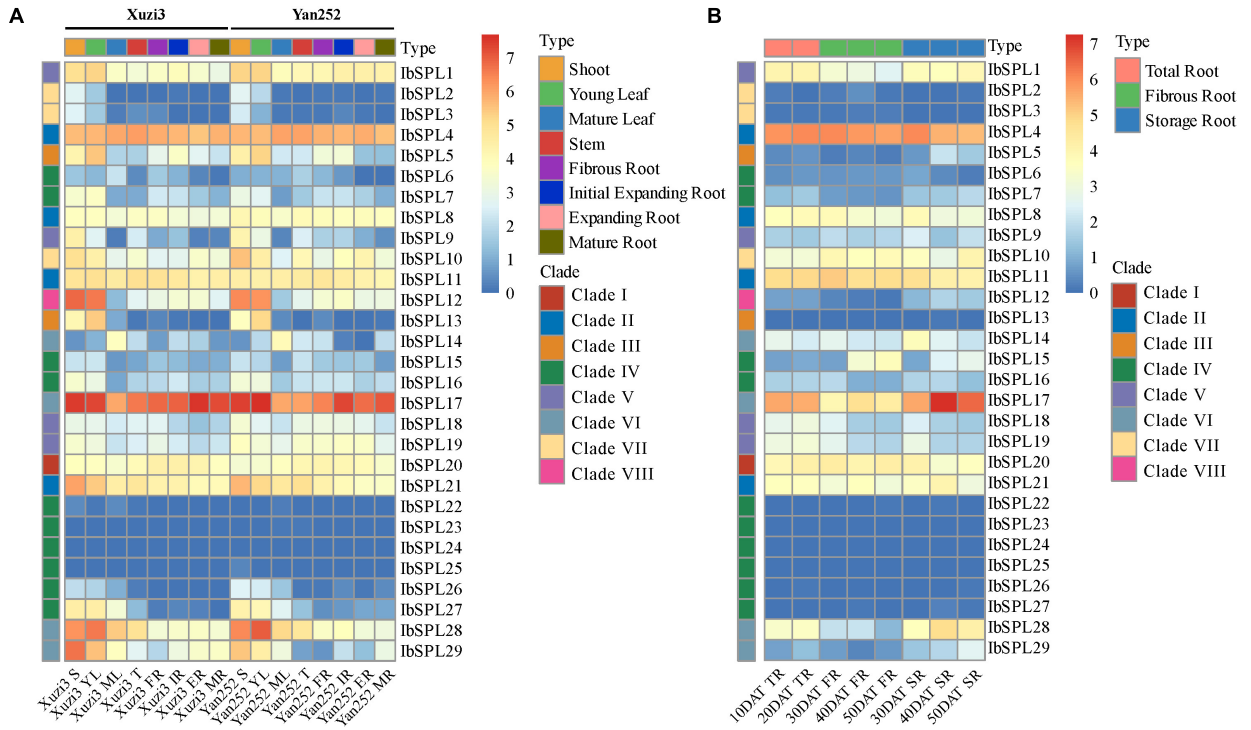

FIGURE 5 | Expression profiles of *IbSPL* genes in different tissues and developmental stages using transcriptome analysis. **(A)** Heatmap representing the expression patterns of *IbSPL* genes in eight different tissues of two sweet potato cultivars (Xuzi3 and Yan252). Xuzi3: sweet potato cultivar Xuzi3, Yan252: sweet potato cultivar Yan252, S: shoot, YL: young leaf, ML: mature leaf, T: stem, FR: fibrous root, IR: initial tuberous root, ER: expanding storage root, MR: mature storage root. **(B)** Heatmap representing the expression patterns of *IbSPL* genes in eight different stages of root development. 10DAT_TR: undifferentiated root (10 days after transplanting), 20DAT_TR: undifferentiated root (20 days after transplanting), 30DAT_FR: fibrous root (30 days after transplanting), 40DAT_FR: fibrous root (days after transplanting), 50DAT_FR: fibrous root (50 days after transplanting), 30DAT_SR: storage root (30 days after transplanting), 40DAT_SR: storage root (40 days after transplanting), 50DAT_SR: storage root (50 days after transplanting). Heatmaps were generated based on the $\log_2$(FPKM + 1) values for each *IbSPL* gene.

suggested that *Ipomoea SPL* genes participate in various biological processes.

Expression Profiles of *IbSPL* Genes in Different Tissues

Among the four *Ipomoea* species, *I. batatas* is the most important crop cultivated globally. To explore the putative roles of *IbSPL* genes, the tissue-specific expression patterns of *IbSPLs* were analyzed in eight tissues (four aboveground and four underground tissues) of two sweet potato cultivars (Xuzi3 and Yan252) using publicly available transcriptomic data (**Supplementary Table 8**; Ding et al., 2017). FPKM values were calculated to evaluate gene expression levels (**Supplementary Table 9**). As shown in **Figure 5A**, the expression patterns of *IbSPLs* were classified into three groups. The first group included seven *IbSPL* genes (*IbSPL6*/*IbSPL15*/*IbSPL22*/*IbSPL23*/*IbSPL24*/*IbSPL25*/*IbSPL26*), with lowest expression levels [$\log_2$(FPKM) $<$ 2] in all tissues. The second group included five *IbSPL* genes (*IbSPL4*/*IbSPL8*/*IbSPL11*/*IbSPL17*/*IbSPL20*), with relatively high expression levels in all tissues. The third group included the remaining 17 *IbSPL* genes, with high expression in some aboveground tissues, especially in shoots or young leaves. Additionally, the gene expression profiles in underground tissues (fibrous and tuberous root) were investigated, with seven *IbSPL* genes highly expressed in underground tissues (mean FPKM $>$ 10), such as *IbSPL1*, *IbSPL4*, *IbSPL11*, *IbSPL17*, *IbSPL20*, *IbSPL21*, and *IbSPL28*, implying their potential functionality in root development. Furthermore, these results showed that *IbSPL* genes within the same clades exhibit distinct expression patterns, such as *IbSPL14*, *IbSPL17*, *IbSPL28*, and *IbSPL29* in clade VI.

To further investigate the expression profiles of *IbSPL* genes in underground tissues, publicly available transcriptomic data of eight different stages during root development from the cultivar 'Beauregard' were used (**Figure 5B** and **Supplementary Tables 8, 9**). The results showed that the overall expression patterns of *IbSPL* genes in the roots of the cultivar 'Beauregard' were similar to those of cultivar 'Xuzi3' and 'Yan252.' Specifically, the expression levels of *IbSPL17*, *IbSPL28*, and *IbSPL29* were the highest in storage roots compared with undifferentiated and fibrous roots. *IbSPL1* showed the highest expression in undifferentiated roots whereas *IbSPL10* showed the highest expression in fibrous roots. However, the expression levels of *IbSPL4*, *IbSPL8*, *IbSPL11*, *IbSPL20*, and

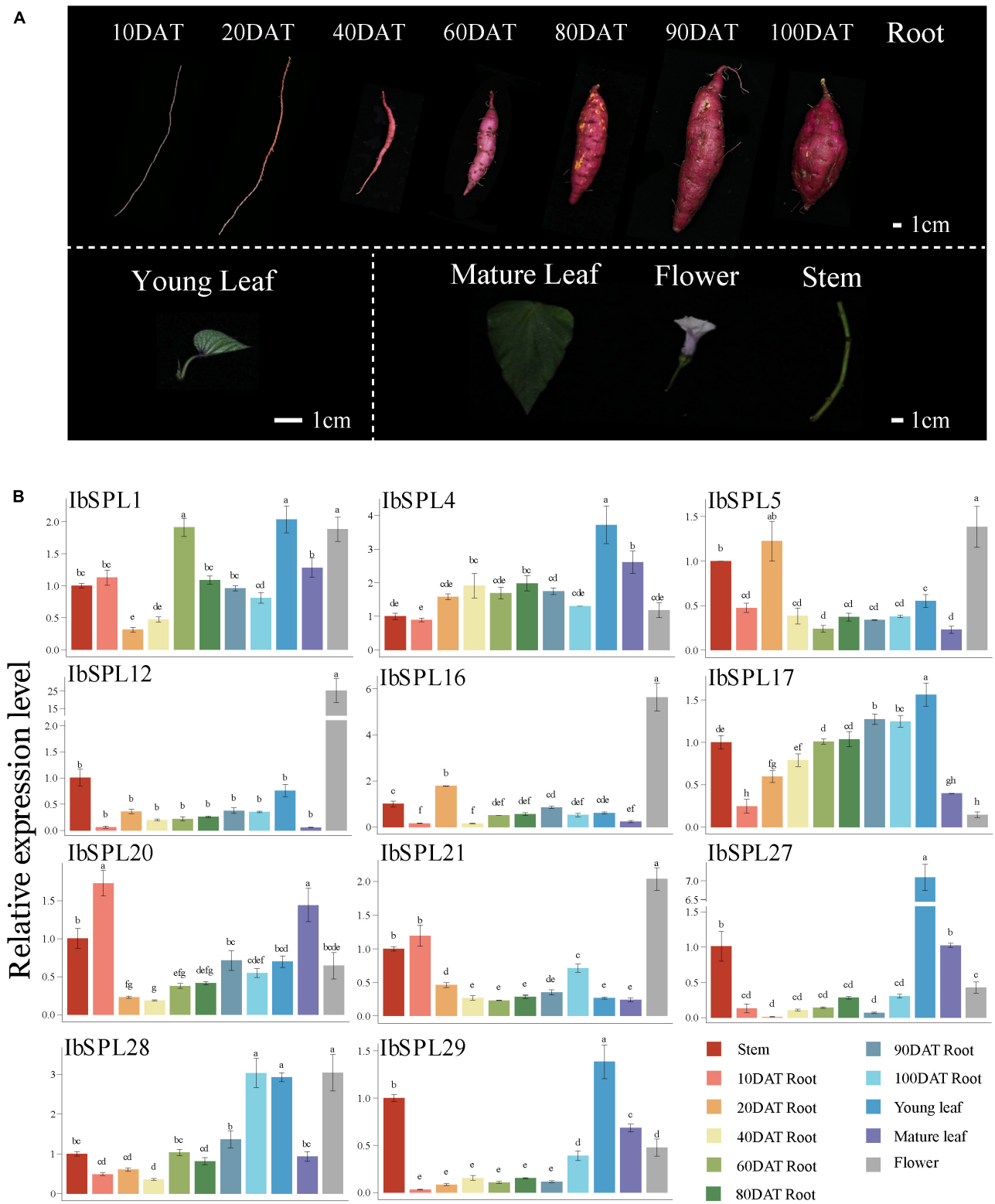

FIGURE 6 | Expression profiles of 11 *IbSPL* genes in different tissues and at different developmental stages using qRT-PCR. **(A)** Different tissues from cultivar Xuyu34 are shown, including stem, leaf, flower, and root. Black bar is scale of 1 cm. **(B)** Expression patterns of 11 *IbSPL* genes determined using qRT-PCR. The *x*-axis indicates different tissues. The *y*-axis represents the relative expression calculated using the $2^{-\Delta\Delta CT}$ method. The error bars indicate the standard deviations of the three biological replicates. Different letters indicate statistically significant differences at $p < 0.05$.

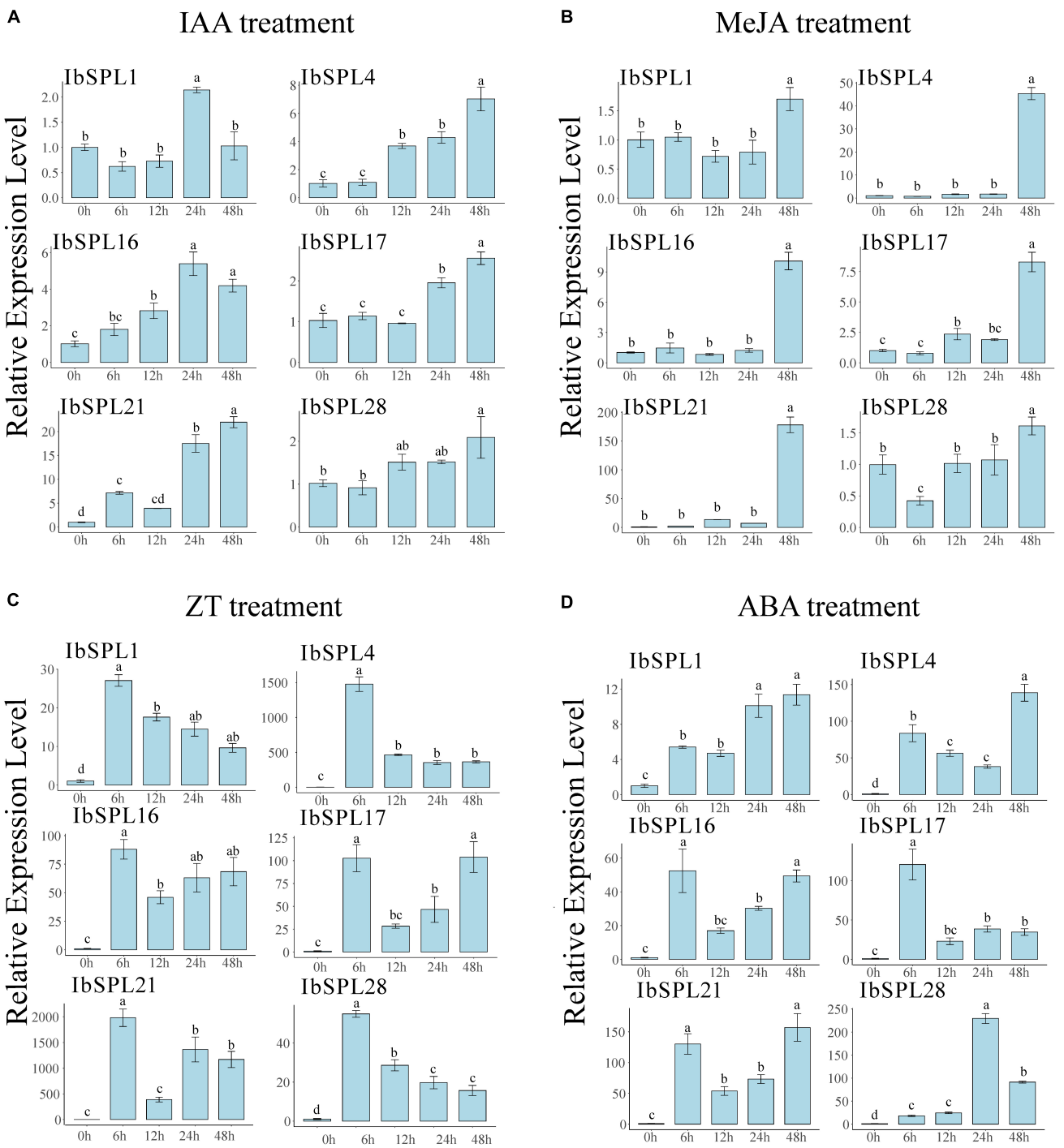

FIGURE 7 | Expression patterns of six *IbSPL* genes under phytohormone treatment. **(A)** Expression patterns of the six *IbSPL* genes under IAA treatment. **(B)** Expression patterns of the six *IbSPL* genes under MeJA treatment. **(C)** Expression patterns of the six *IbSPL* genes under ZT treatment. **(D)** Expression patterns of the six *IbSPL* genes under ABA treatment. Samples were collected at 0, 6, 12, 24, and 48 h after treatment. The *x*-axis indicates different samples, while the *y*-axis represents the relative expression calculated using the $2^{-\Delta\Delta CT}$ method. The error bars indicate the standard deviations of the three biological replicates. Different letters indicate statistically significant differences at $p < 0.05$.

IbSPL21 showed no distinct variation in all tissues. These results, therefore, imply that these genes may play important roles in root development.

To confirm the expression patterns of *IbSPLs* derived from the transcriptomic data, a total of 11 *IbSPL* genes (*IbSPL1*, *IbSPL4*, *IbSPL5*, *IbSPL12*, *IbSPL16*, *IbSPL17*, *IbSPL20*,

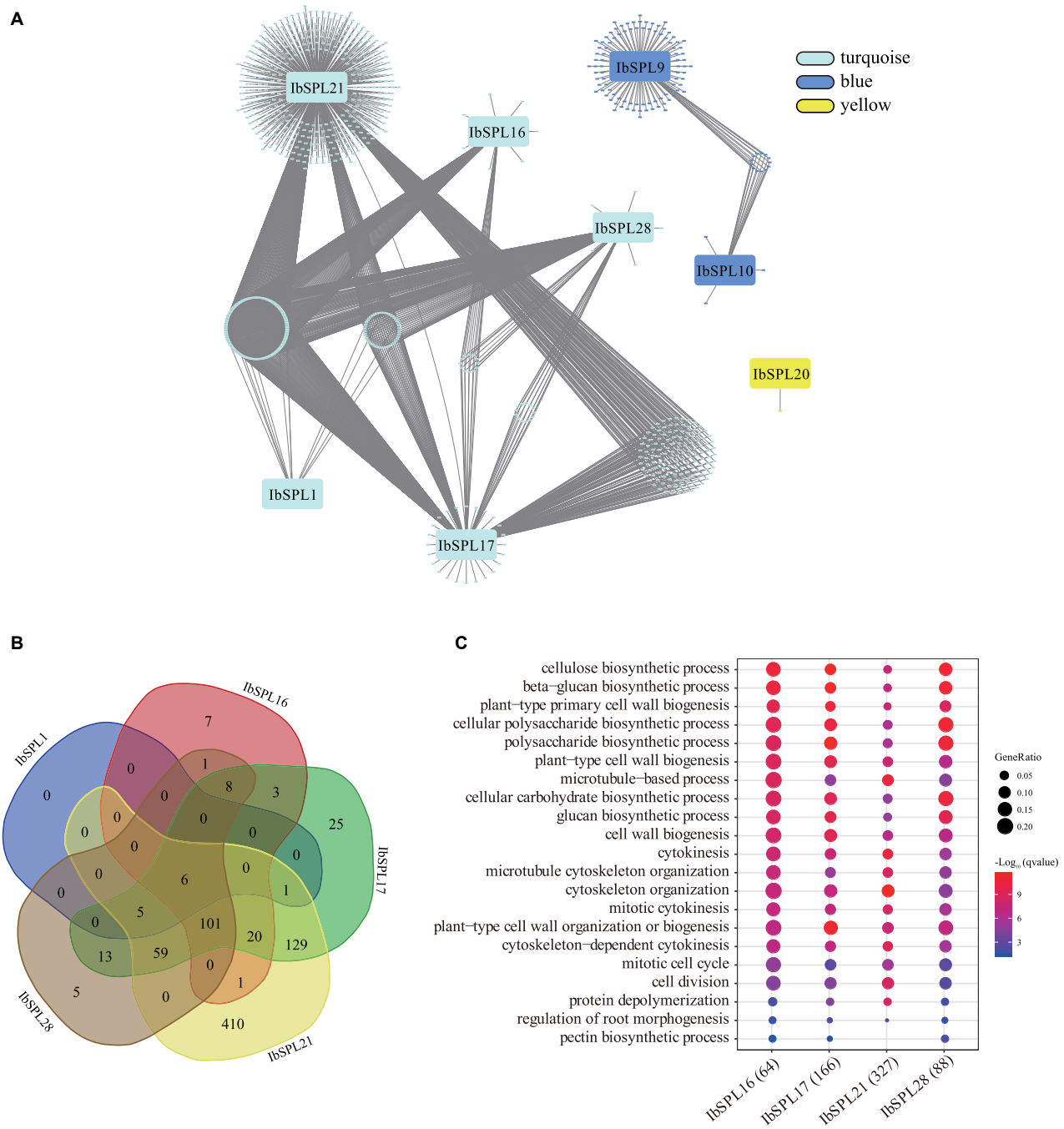

FIGURE 8 | Regulatory sub-networks involving *IbSPLs* and other *I. batatas* genes in the storage root of sweet potato. **(A)** Co-expression sub-network visualization for *IbSPLs* in the storage root of sweet potato. Only the edges with a weight above a threshold of 0.05 are displayed. **(B)** Venn diagram of the number of co-expressed genes for *IbSPL1*, *IbSPL16, IbSPL17*, *IbSPL21*, and *IbSPL28* in the turquoise module. **(C)** GO enrichment analysis of co-expressed genes for *IbSPL16*, *IbSPL17*, *IbSPL21*, and *IbSPL28*.

IbSPL21, *IbSPL27*, *IbSPL28*, and *IbSPL29*) highly expressed in aboveground or underground tissues were selected for qRT-PCR analysis in 11 tissues of cultivar 'Xuyu34' (**Figure 6**). The results showed consistent expression patterns of *IbSPL* genes between the transcriptomic data and qRT-PCR results. Moreover, the expression levels of *IbSPLs* differed in various tissues. For example, *IbSPL27* and *IbSPL29* were highly expressed in young leaves, while *IbSPL12*, *IbSPL16*, and *IbSPL21* were highly expressed in flower. Moreover, the gradual increase in *IbSPL17* and *IbSPL28* expressions indicated their differential roles in storage root development in sweet potato.

IbSPL Genes in Response to Exogenous Phytohormones

The promoter analysis revealed that *IbSPL* genes could be regulated by various phytohormones, which are known regulators of plant growth and development. To reveal the potential roles of *IbSPLs* in hormone signaling pathways, six *IbSPL* genes (*IbSPL1*, *IbSPL4*, *IbSPL16*, *IbSPL17*, *IbSPL21,* and *IbSPL28*) highly expressed in roots were selected to perform qRT-PCR analysis under exogenous phytohormone treatments, which included IAA, MeJA, ZT, and ABA. The expression analysis indicated that the six *IbSPL* genes exhibited highly divergent response patterns under phytohormone treatment in the adventitious root (**Figure 7**). Under IAA treatment, *IbSPL16* and *IbSPL21* were rapidly upregulated after 6 h of treatment, while the other four genes were upregulated after 12 or 24 h of treatment. Under MeJA treatment, all *IbSPL* genes were significantly upregulated after 48 h of treatment, with *IbSPL4* and *IbSPL21* upregulated around 45.3 and 178.1 folds, respectively. Under ZT treatment, all *IbSPL* genes were highly upregulated after 6 h of treatment, with a subsequent decline followed by approximately 10-fold increased expression than CK. Notably, *IbSPL21* showed a particularly positive response to ZT treatment. Under ABA treatment, all examined *IbSPL* genes, particularly *IbSPL1*, *IbSPL4*, *IbSPL16*, *IbSPL17*, and *IbSPL21*, were rapidly upregulated after 6 h of treatment, whereas *IbSPL28* showed significantly upregulation after 24 h of treatment.

Regulatory Sub-Networks Involving *IbSPLs* and Other *I. batatas* Genes in the Storage Root of Sweet Potato

To identify the regulatory sub-networks involving *IbSPLs* in the storage root of sweet potato, WGCNA was performed based on transcriptomic data of mature storage roots from 88 sweet potato accessions (**Supplementary Figure 8** and **Supplementary Table 11**). A total of 19 modules were obtained from this analysis, with three modules (turquoise, blue and yellow) containing totally eight *IbSPL* genes (**Figure 8A**). In the yellow module, *IbSPL20* showed co-expression with only one gene. In the blue module, *IbSPL9* and *IbSPL10* were co-expressed with 17 and 107 genes, respectively. In the turquoise module, *IbSPL1*, *IbSPL16, IbSPL17*, *IbSPL21*, and *IbSPL28* had 12, 147, 370, 732, and 198 co-expressed genes, respectively (**Supplementary Table 12**). Interestingly, *IbSPL16, IbSPL17*, *IbSPL21*, and *IbSPL28* in the turquoise module share 101 co-expressed genes (**Figure 8B**), indicating functionality in similar biological processes.

To further explore the putative functions of the SPLs in the storage root, GO enrichment analysis was performed on the co-expressed genes. For *IbSPL1, IbSPL20,* and *IbSPL9*, GO enrichment results could not be obtained due to the small number of co-expressed genes. For *IbSPL10* in the blue module, co-expressed genes related to "response to chitin" and "response to organonitrogen compound" were enriched (**Supplementary Table 13**). For *IbSPL16*, *IbSPL17*, *IbSPL21*, and *IbSPL28* in the turquoise module, the similar GO terms were enriched, such as "regulation of root morphogenesis," "cell division," "cytoskeleton organization," "plant-type cell wall organization or biogenesis," and "cellulose biosynthetic process" (**Figure 8C** and **Supplementary Table 13**). It is known that the cytoskeleton, cell division, and cell wall organization/biogenesis are important biological processes involved in storage root development and formation (Dong et al., 2019). Therefore, these results indicated that *IbSPL16*/*IbSPL17*/*IbSPL21*/*IbSPL28* may play a key role in storage root development in sweet potato.

DISCUSSION

SPL genes are important plant-specific transcription factors with a highly conserved SBP domain. Since its discovery in *A. majus*, *SPL* gene members have been increasingly identified in plants (Klein et al., 1996; Cardon et al., 1999; Xie et al., 2006; Salinas et al., 2012; Li et al., 2013; Zhang et al., 2016; Tripathi et al., 2017). However, comprehensive molecular, evolutionary and functional analysis of the *SPL* genes in the genus *Ipomoea* are lacking. The genus *Ipomoea* has significant nutritional and economic value for humans, including the seventh most important crop *I. batatas* and ornamental plant *I. nil*. Up to now, four *Ipomoea* species have been sequenced: *I. nil*, *I. triloba*, *I. trifida,* and *I. batatas*. Utilizing these genomes, this study systematically analyzed the *Ipomoea SPL* genes, including molecular characteristics, evolutionary process, post-transcriptional regulation, and physiological function.

Comparative Analysis of SQUAMOSA Promoter-Binding Protein-Like Genes in the Genus *Ipomoea*

Recently, some important transcription factor gene families have been investigated in *Ipomoea* species, such as bZIPs (Yang et al., 2019), WRKYs (Li Y. et al., 2019), GRPs (Lu et al., 2019), MADS (Zhu et al., 2020), and DEAD-box (Wan et al., 2020). However, these studies have focused only on the gene families in a single *Ipomoea* species, and comparative analysis of gene families in the genus *Ipomoea* are scarce. This study dissects the evolutionary dynamics of *SPL* genes in the genus *Ipomoea*, identifying *SPL* genes in four *Ipomoea* species: 29 *IbSPLs*, 27 *ItfSPLs*, 26 *ItbSPLs,* and 23 *IniSPLs* (**Figure 1**). Notably, the number of *SPL* genes in sweet potato is approximately equal to that of the diploid wild relatives (*I. nil*, *I. triloba*, and *I. trifida*), owing to the haplotype-resolved genome assembly of hexaploid sweet potato (Yang et al., 2017). Following the classifications of *A. thaliana, S. lycopersicum,* and *O. sativa* (Cardon et al., 1999; Xie et al., 2006; Salinas et al., 2012), the *Ipomoea SPL* genes were also divided into eight clades (**Figure 2**), with clade IV comprising the highest members (32) and clade I comprising the lowest (four). Gene and protein structure analysis revealed that most *Ipomoea* SPLs from the same phylogenetic clade share similar intron/exon structures, domain organizations, and motif compositions (**Supplementary Figure 1**), indicating that *SPLs* within the same clade may have similar functions in *Ipomoea* species. Interestingly, apart from the conserved domain and motifs present in the SPL proteins, other domains or motifs were found in clades I, II, and IV, such as DEXDc domain

(motif 4) and Ankyrin repeats (motif 8), which were also observed in papaya (Xu et al., 2020) and barley (Tripathi et al., 2018). These results suggest that the SPLs in these clades may have undergone evolutionary functional differentiation and/or neofunctionalization.

Ipomoea species were found to possess more *SPL* genes than dicotyledonous model plants, such as *A. thaliana* and *S. lycopersicum* (Cardon et al., 1999; Salinas et al., 2012), implying the genus-specific expansion of the *SPL* gene family in *Ipomoea* species. The expansion of gene families is a result of evolutionary duplication events (Moore and Purugganan, 2005). This study showed that segmental duplication plays a major role in the evolution and expansion of *Ipomoea SPL* genes (**Figure 3B**), which is consistent with findings in cotton (Cai et al., 2018), *Rosacea* species (Abdullah et al., 2018) and *Euphorbiaceae* species (Li J. et al., 2019). Previous studies have reported that a whole-genome triplication event occurred 46.1 million years ago (Mya) in the progenitor of the genus *Ipomoea* (Yang et al., 2017), and thereby the derivation of the segmental duplication of SPLs from this event was speculated. Furthermore, tandem duplication was found to be the most frequent event in the IV-b and IV-c subclades (**Supplementary Table 1**) and orthologs were found to be absent in dicots (**Supplementary Table 3**), implying that the *SPL* members in these two subclades were tandem duplicated from a recent event. Therefore, the *SPL* genes were speculated to have undergone replication expansion in the progenitor of the genus *Ipomoea* and that various *SPL* members were retained due to their important role in growth and development during the *Ipomoea* species differentiation. Additionally, the *Ka/Ks* ratios were less than 1 for all *Ipomoea SPL* ortholog gene pairs, indicating that the *Ipomoea SPL* genes were under strong purifying selection (**Supplementary Table 4**).

Many *SPL* genes are miR156 targets, thus forming a functional regulatory network of miR156-SPL, which plays an important role in plant growth and development (Gou et al., 2011; Wang and Wang, 2015; Liu et al., 2019). More than half of the *SPL* gene family members have been reported to be targeted by miR156 in various plant species, such as rice (Xie et al., 2006), tomato (Salinas et al., 2012), and apple (Li et al., 2013). In this study, two-thirds of *SPL* genes in each *Ipomoea* species were predicted to be miR156 targets (**Figure 4B**). Phylogenetic analysis showed that all *SPL* genes in clade III lacked miR156 binding sites, which is consistent with the reported results of *A. thaliana*, rice, and tomato (Preston and Hileman, 2013). Additionally, two binding site types of miR156 were identified in *SPL* genes: one located in CDS and the other located in the 3′-UTR (**Figure 4B**), which is consistent with the observation in other plants, such as rice (Xie et al., 2006), tomato (Salinas et al., 2012), apple (Li et al., 2013), and papaya (Xu et al., 2020). The degradome data of *I. batatas* further confirmed seven *IbSPL* genes as the targets of miR156 (**Figure 4C** and **Supplementary Table 1**). Among these *IbSPL* genes, the miR156 binding site for *IbSPL17* and *IbSPL28* is located in the 3′-UTR, which is consistent with the miR156 binding site of their orthologous genes (*AtSPL3* and *LeSPL-CNR*) in *A. thaliana* and tomato (Gandikota et al., 2007; Chen et al., 2015). This suggests the high conservation of the miR156-SPL regulatory module in plants. Additionally, our results showed a negative correlation in the expression pattern of miR156 and their target *SPL* genes, suggesting that *SPLs* might be regulated by miR156 at the post-transcriptional level. However, most of the miR156-SPL interactions in this study were predicted using *in silico* analysis, requiring further experimental verification for the miR156-SPL interactions in the genus *Ipomoea*.

IbSPL Genes Are Putatively Involved in Storage Root Development

Sweet potato, the seventh most important crop globally, has strong adaptability, stable yield, and high nutritional value (Liu, 2017). The storage root of sweet potato is economically useful for its nutrient content and yield, and thus, dissecting the mechanisms underlying storage root formation and development is significant to improve sweet potato nutrient content and yield. Considering the key regulatory roles of *SPL* genes in root architecture (Yu et al., 2015; Barrera-Rojas et al., 2020) and biomass enhancement (Wang et al., 2015), the expression patterns of *IbSPL* genes in different tissues or at different developmental stages were evaluated using the public transcriptome data. Most of the *IbSPLs* were found to be highly expressed in aboveground tissues, especially in shoots or young leaves; however, only some *IbSPLs* were found to be highly expressed in underground tissues (**Figure 7**). qRT-PCR analysis of the expression levels of two *IbSPL* genes (*IbSPL17*/*IbSPL28*) revealed a significant increase with storage root development (**Figure 6**). This study provides evidence that *SPL* genes have important functions during storage root development in sweet potato.

The formation and development of storage roots is a complex physiological process that includes the cessation of root elongation, genesis and development of the primary and secondary vascular cambium, increase in radial growth and accumulation of starch and storage proteins (Ravi et al., 2009). These processes are closely related to the endogenous phytohormones, such as IAA, cytokinins (CTKs), JA, and ABA (Nakatani, 1991; Tanaka et al., 2008; Ravi et al., 2009; Dong et al., 2019). For instance, IAA is involved in early stages of storage root formation and primary storage root thickening (Noh et al., 2010); ABA plays a significant role in storage root bulking by activating cell division (Huan et al., 2020) and CTKs play a key role in storage root initiation and expansion as a pre-requirement for cambial cell proliferation (Dong et al., 2019). Moreover, storage root yields are positively correlated with ABA and CTK contents (Wang et al., 2005). In the present study, different kinds of hormone-responsive elements were found by analyzing the *IbSPL* promoters (**Supplementary Figure 7**), implying that *IbSPL* genes may participate in hormone signaling pathways. qRT-PCR analysis further confirmed that the expression of the tested *IbSPLs* (*IbSPL1*, *IbSPL4*, *IbSPL16*, *IbSPL17*, *IbSPL21,* and *IbSPL28*) was strongly induced under exogenous phytohormone treatments, particularly ZT and ABA, suggesting their crucial roles in root development.

Storage root formation and development is maintained by coordinated cellular behaviors, such as cell division, expansion, and differentiation. Previous studies have revealed that cell wall biosynthesis and cytoskeleton organization are critical in these cellular behaviors (Bashline et al., 2014; Brasil et al., 2017).

The regulatory sub-networks in this study were analyzed using WGCNA, which indicated that eight *IbSPL* genes were co-expressed with at least one other *I. batatas* genes (**Figure 8A**). GO enrichment analysis of co-expressed genes speculated the role of *IbSPL* genes in stress responses, root morphogenesis, and cell division (**Supplementary Table 13**). Moreover, the genes co-expressed with *IbSPL16*/*IbSPL17*/*IbSPL21*/*IbSPL28* in the turquoise module were all significantly enriched for "regulation of root morphogenesis," "cell division," "cytoskeleton organization," "plant-type cell wall organization or biogenesis," and "cellulose biosynthetic process." These enriched processes are essential for cell morphogenesis and cell cycles, implying their key roles in storage root development. In the future, functional characterization is needed to elucidate the specific roles of *IbSPLs* in storage root development.

CONCLUSION

In summary, a genome-wide analysis of the *SPL* gene family in four *Ipomoea* species, including *I. batatas*, *I. trifida*, *I. triloba*, and *I. nil* was performed. A total of 105 *Ipomoea SPL* genes were identified and divided into eight clades. Genes in one clade were found to harbor similar gene structures, domain organizations, motif compositions, and *cis*-acting elements, suggesting potential functional similarity. Moreover, segmental duplication was predominantly responsible for the expansion of the *Ipomoea SPL* gene family. On combining the results from the expression patterns and regulatory sub-networks, *IbSPL16*/*IbSPL17*/*IbSPL21*/*IbSPL28* were found to play an important role in storage root development. Therefore, this study not only provides novel insights into the evolutionary and functional divergence of the *SPL* genes in the genus *Ipomoea* but also lays a foundation for further elucidation of the potential functional roles of *IbSPL* genes during storage root 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 in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

LZ, SW, and TX conceived and designed the research. HS, JM, WZ, LZ, WH, and YZ performed the research and analyzed the data. LZ and HS wrote the manuscript. All authors have read and approved the manuscript.

FUNDING

This work was supported by the Natural Science Foundation of Jiangsu Province (No. BK20190995), the Shanghai Municipal Afforestation and City Appearance and Environmental Sanitation Administration (No. G212402), the Natural Science Foundation of Xuzhou City (No. KC19070), Postgraduate Research and Practice Innovation Program of Jiangsu Province (No. KYCX20_2319), and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). Funders had no role in the design of the study and collection, analysis, and interpretation of data, and in writing the manuscript.

ACKNOWLEDGMENTS

We would like to thank Meng Kou (Xuzhou Academy of Agricultural Sciences) for providing the sweet potato (*I. batatas* cv. Xuyu34) plants.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.801061/full#supplementary-material

REFERENCES

Abdullah, M., Cao, Y., Cheng, X., Shakoor, A., Su, X., Gao, J., et al. (2018). Genome-Wide Analysis Characterization and Evolution of SBP Genes in Fragaria vesca, Pyrus bretschneideri, Prunus persica and Prunus mume. *Front. Genet.* 9:64. doi: 10.3389/fgene.2018.00064

Addo-Quaye, C., Miller, W., and Axtell, M. J. (2009). CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. *Bioinformatics* 25, 130–131. doi: 10.1093/bioinformatics/btn604

Austin, D. F., and Huáman, Z. (1996). A synopsis of Ipomoea (Convolvulaceae) in the Americas. *Taxon* 45, 3–38.

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

Barrera-Rojas, C. H., Rocha, G. H. B., Polverari, L., Pinheiro Brito, D. A., Batista, D. S., Notini, M. M., et al. (2020). miR156-targeted SPL10 controls Arabidopsis root meristem activity and root-derived de novo shoot regeneration via cytokinin responses. *J. Exp. Bot.* 71, 934–950. doi: 10.1093/jxb/erz475

Bashline, L., Lei, L., Li, S., and Gu, Y. (2014). Cell wall, cytoskeleton, and cell expansion in higher plants. *Mol. Plant* 7, 586–600. doi: 10.1093/mp/ssu018

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

Brasil, J. N., Costa, C. N. M., Cabral, L. M., Ferreira, P. C. G., and Hemerly, A. S. (2017). The plant cell cycle: Pre-Replication complex formation and controls. *Genet. Mol. Biol.* 40(1 Suppl. 1), 276–291. doi: 10.1590/1678-4685-GMB-2016-0118

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

Cardon, G., Hohmann, S., Klein, J., Nettesheim, K., Saedler, H., and Huijser, P. (1999). Molecular characterisation of the Arabidopsis

SBP-box genes. *Gene* 237, 91–104. doi: 10.1016/s0378-1119(99)00308-x

Caruthers, J. M., and McKay, D. B. (2002). Helicase structure and mechanism. *Curr. Opin. Struct. Biol.* 12, 123–133. doi: 10.1016/s0959-440x(02)00298-1

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mole. Biol. Evol.* 17, 540–552.

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, W., Kong, J., Lai, T., Manning, K., Wu, C., Wang, Y., et al. (2015). Tuning LeSPL-CNR expression by SlymiR157 affects tomato fruit ripening. *Sci. Rep.* 5:7852. doi: 10.1038/srep07852

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

Chen, Y., Zhu, P., Wu, S., Lu, Y., Sun, J., Cao, Q., et al. (2019). Identification and expression analysis of GRAS transcription factors in the wild relative of sweet potato Ipomoea trifida. *BMC Genom.* 20:911. doi: 10.1186/s12864-019-6316-7

Coordinators, N. R. (2018). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 46, D8–D13. doi: 10.1093/nar/gkx1095

Dai, X., Zhuang, Z., and Zhao, P. X. (2018). psRNATarget: a plant small RNA target analysis server (2017 release). *Nucleic Acids Res.* 46, W49–W54. doi: 10.1093/nar/gky316

de Castro, E., Sigrist, C. J., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P. S., Gasteiger, E., et al. (2006). ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res.* 34, W362–W365. doi: 10.1093/nar/gkl124

Ding, N., Wang, A., Zhang, X., Wu, Y., Wang, R., Cui, H., et al. (2017). Identification and analysis of glutathione S-transferase gene family in sweet potato reveal divergent GST-mediated networks in aboveground and underground tissues in response to abiotic stresses. *BMC Plant Biol.* 17:225. doi: 10.1186/s12870-017-1179-z

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

Dong, T., Zhu, M., Yu, J., Han, R., Tang, C., Xu, T., et al. (2019). RNA-Seq and iTRAQ reveal multiple pathways involved in storage root formation and development in sweet potato (Ipomoea batatas L.). *BMC Plant Biol.* 19:136. doi: 10.1186/s12870-019-1731-0

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., et al. (2019). The Pfam protein families database in 2019. *Nucleic Acids Res.* 47, D427–D432. doi: 10.1093/nar/gky995

Gandikota, M., Birkenbihl, R. P., Hohmann, S., Cardon, G. H., Saedler, H., and Huijser, P. (2007). The miRNA156/157 recognition element in the 3' UTR of the Arabidopsis SBP box gene SPL3 prevents early flowering by translational inhibition in seedlings. *Plant J.* 49, 683–693. doi: 10.1111/j.1365-313X.2006.02983.x

Gou, J. Y., Felippes, F. F., Liu, C. J., Weigel, D., and Wang, J. W. (2011). Negative regulation of anthocyanin biosynthesis in Arabidopsis by a miR156-targeted SPL transcription factor. *Plant Cell* 23, 1512–1522. doi: 10.1105/tpc.111.084525

Guo, A. Y., Zhu, Q. H., Gu, X., Ge, S., Yang, J., and Luo, J. (2008). Genome-wide identification and evolutionary analysis of the plant specific SBP-box transcription factor family. *Gene* 418, 1–8. doi: 10.1016/j.gene.2008.03.016

Hoshino, A., Jayakumar, V., Nitasaka, E., Toyoda, A., Noguchi, H., Itoh, T., et al. (2016). Genome sequence and analysis of the Japanese morning glory Ipomoea nil. *Nat. Commun.* 7:13295. doi: 10.1038/ncomms13295

Huan, L., Jin-Qiang, W., and Qing, L. (2020). Photosynthesis product allocation and yield in sweet potato with spraying exogenous hormones under drought stress. *J. Plant Physiol.* 253, 153265. doi: 10.1016/j.jplph.2020.153265

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., et al. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* 34, 2115–2122. doi: 10.1093/molbev/msx148

Jiang, X., Chen, P., Zhang, X., Liu, Q., and Li, H. (2021). Comparative analysis of the SPL gene family in five Rosaceae species: Fragaria vesca, Malus domestica, Prunus persica, Rubus occidentalis, and Pyrus pyrifolia. *Open Life Sci.* 16, 160–171. doi: 10.1515/biol-2021-0020

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

Klein, J., Saedler, H., and Huijser, P. (1996). A new family of DNA binding proteins includes putative transcriptional regulators of the Antirrhinum majus floral meristem identity gene SQUAMOSA. *Mol. Gen. Genet.* 250, 7–16. doi: 10.1007/BF02191820

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

Kuang, Z., Wang, Y., Li, L., and Yang, X. (2019). miRDeep-P2: accurate and fast analysis of the microRNA transcriptome in plants. *Bioinformatics* 35, 2521–2522. doi: 10.1093/bioinformatics/bty972

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

Kuo, Y.-W., Lin, J.-S., Li, Y.-C., Jhu, M.-Y., King, Y.-C., and Jeng, S.-T. (2019). MicroR408 regulates defense response upon wounding in sweet potato. *J. Exp. Bot.* 70, 469–483.

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

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

Letunic, I., and Bork, P. (2018). 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.* 46, D493–D496. doi: 10.1093/nar/gkx922

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, C., and Lu, S. (2014). Molecular characterization of the SPL gene family in Populus trichocarpa. *BMC Plant Biol.* 14:131. doi: 10.1186/1471-2229-14-131

Li, J., Gao, X., Sang, S., and Liu, C. (2019). Genome-wide identification, phylogeny, and expression analysis of the SBP-box gene family in Euphorbiaceae. *BMC Genomics* 20(Suppl. 9):912. doi: 10.1186/s12864-019-6319-4

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

Li, J., Mahajan, A., and Tsai, M. D. (2006). Ankyrin repeat: a unique motif mediating protein-protein interactions. *Biochemistry* 45, 15168–15178. doi: 10.1021/bi062188q

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, Y., Zhang, L., Zhu, P., Cao, Q., Sun, J., Li, Z., et al. (2019). Genome-wide identification, characterisation and functional evaluation of WRKY genes in the sweet potato wild ancestor Ipomoea trifida (H.B.K.) G. Don. under abiotic stresses. *BMC Genet.* 20:90. doi: 10.1186/s12863-019-0789-x

Liu, M., Shi, Z., Zhang, X., Wang, M., Zhang, L., Zheng, K., et al. (2019). Inducible overexpression of Ideal Plant Architecture1 improves both yield and disease resistance in rice. *Nat. Plants* 5, 389–400. doi: 10.1038/s41477-019-0383-2

Liu, Q. (2017). Improvement for agronomically important traits by gene engineering in sweetpotato. *Breed Sci.* 67, 15–26. doi: 10.1270/jsbbs.16126

Liu, Y., Su, W., Wang, L., Lei, J., Chai, S., Zhang, W., et al. (2021). Integrated transcriptome, small RNA and degradome sequencing approaches proffer insights into chlorogenic acid biosynthesis in leafy sweet potato. *PLoS One* 16:e0245266. doi: 10.1371/journal.pone.0245266

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

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

Manning, K., Tor, M., Poole, M., Hong, Y., Thompson, A. J., King, G. J., et al. (2006). A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. *Nat. Genet.* 38, 948–952. doi: 10.1038/ng1841

Moore, R. C., and Purugganan, M. D. (2005). The evolutionary dynamics of plant duplicate genes. *Curr. Opin. Plant Biol.* 8, 122–128. doi: 10.1016/j.pbi.2004.12.001

Morita, Y., and Hoshino, A. (2018). Recent advances in flower color variation and patterning of Japanese morning glory and petunia. *Breed Sci.* 68, 128–138. doi: 10.1270/jsbbs.17107

Nakatani, M. (1991). Changes in endogenous level of zeatin riboside, abscisic acid and indole acetic acid during formation and thickening of tuberous root in sweet potato. *Jpn. J. Crop* 1991:60.

National Genomics Data Center, M., and Partners (2020). Database Resources of the National Genomics Data Center in 2020. *Nucleic Acids Res.* 48, D24–D33. doi: 10.1093/nar/gkz913

Noh, S. A., Lee, H. S., Huh, E. J., Huh, G. H., Paek, K. H., Shin, J. S., et al. (2010). SRD1 is involved in the auxin-mediated initial thickening growth of storage root by enhancing proliferation of metaxylem and cambium cells in sweetpotato (Ipomoea batatas). *J. Exp. Bot.* 61, 1337–1349. doi: 10.1093/jxb/erp399

Orfila, C., Huisman, M. M., Willats, W. G., van Alebeek, G. J., Schols, H. A., Seymour, G. B., et al. (2002). Altered cell wall disassembly during ripening of Cnr tomato fruit: implications for cell adhesion and fruit softening. *Planta* 215, 440–447. doi: 10.1007/s00425-002-0753-1

Park, S. C., Kim, Y. H., Ji, C. Y., Park, S., Jeong, J. C., Lee, H. S., et al. (2012). Stable internal reference genes for the normalization of real-time PCR in different sweetpotato cultivars subjected to abiotic stress conditions. *PLoS One* 7:e51502. doi: 10.1371/journal.pone.0051502

Preston, J. C., and Hileman, L. C. (2013). Functional Evolution in the Plant SQUAMOSA-PROMOTER BINDING PROTEIN-LIKE (SPL) Gene Family. *Front. Plant Sci.* 4:80. doi: 10.3389/fpls.2013.00080

Ravi, V., Chakrabarti, S. K., Makeshkumar, T., and Saravanan, R. (2014). Molecular Regulation of Storage Root Formation and Development in Sweet Potato. *Horticult. Rev.* 42, 157–208.

Ravi, V., Naskar, S. K., Makeshkumar, T., Babu, B., and Krishnan, B. P. (2009). Molecular physiology of storage root formation and development in sweet potato (Ipomoea batatas (L.) Lam.). *J. Root Crops* 35, 1–27.

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

Saminathan, T., Alvarado, A., Lopez, C., Shinde, S., Gajanayake, B., Abburi, V. L., et al. (2019). Elevated carbon dioxide and drought modulate physiology and storage-root development in sweet potato by regulating microRNAs. *Funct. Integr. Genomics* 19, 171–190. doi: 10.1007/s10142-018-0635-7

Savojardo, C., Martelli, P. L., Fariselli, P., Profiti, G., and Casadio, R. (2018). BUSCA: an integrative web server to predict subcellular localization of proteins. *Nucleic Acids Res.* 46, W459–W466. doi: 10.1093/nar/gky320

Shao, Y., Zhou, H. Z., Wu, Y., Zhang, H., Lin, J., Jiang, X., et al. (2019). OsSPL3, an SBP-Domain Protein, Regulates Crown Root Development in Rice. *Plant Cell* 31, 1257–1275. doi: 10.1105/tpc.19.00038

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27, 431–432.

Subramanian, B., Gao, S., Lercher, M. J., Hu, S., and Chen, W. H. (2019). Evolview v3: a webserver for visualization, annotation, and management of phylogenetic trees. *Nucleic Acids Res.* 47, W270–W275. doi: 10.1093/nar/gkz357

Tanaka, M., Kato, N., Nakayama, H., Nakatani, M., and Takahata, Y. (2008). Expression of class I knotted1-like homeobox genes in the storage roots of sweetpotato (Ipomoea batatas). *J. Plant Physiol.* 165, 1726–1735. doi: 10.1016/j.jplph.2007.11.009

Tripathi, R. K., Bregitzer, P., and Singh, J. (2018). Genome-wide analysis of the SPL/miR156 module and its interaction with the AP2/miR172 unit in barley. *Sci. Rep.* 8:7085. doi: 10.1038/s41598-018-25349-0

Tripathi, R. K., Goel, R., Kumari, S., and Dahuja, A. (2017). Genomic organization, phylogenetic comparison, and expression profiles of the SPL family genes and their regulation in soybean. *Dev. Genes Evol.* 227, 101–119. doi: 10.1007/s00427-017-0574-7

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., et al. (2012). Primer3—new capabilities and interfaces. *Nucleic Acids Res.* 40, e115–e115.

Wan, R., Liu, J., Yang, Z., Zhu, P., Cao, Q., and Xu, T. (2020). Genome-wide identification, characterisation and expression profile analysis of DEAD-box family genes in sweet potato wild ancestor Ipomoea trifida under abiotic stresses. *Genes Genomics* 42, 325–335. doi: 10.1007/s13258-019-00910-x

Wang, H., and Wang, H. (2015). The miR156/SPL Module, a Regulatory Hub and Versatile Toolbox, Gears up Crops for Enhanced Agronomic Traits. *Mol. Plant* 8, 677–688. doi: 10.1016/j.molp.2015.01.008

Wang, Q. M., Zhang, L. M., and Wang, Z. L. (2005). Formation and Thickening of Tuberous Roots in Relation to the Endogenous Hormone Concentrations in Sweetpotato. *Scientia Agricult. Sinica* 38, 2414–2420.

Wang, S., Li, S., Liu, Q., Wu, K., Zhang, J., Wang, S., et al. (2015). The OsSPL16-GW7 regulatory module determines grain shape and simultaneously improves rice yield and grain quality. *Nat. Genet.* 47, 949–954. doi: 10.1038/ng.3352

Wang, 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

Wu, S., Lau, K. H., Cao, Q., Hamilton, J. P., Sun, H., Zhou, C., et al. (2018). Genome sequences of two diploid wild relatives of cultivated sweetpotato reveal targets for genetic improvement. *Nat. Commun.* 9:4580. doi: 10.1038/s41467-018-06983-8

Xie, K., Wu, C., and Xiong, L. (2006). Genomic organization, differential expression, and interaction of SQUAMOSA promoter-binding-like transcription factors and microRNA156 in rice. *Plant Physiol.* 142, 280–293. doi: 10.1104/pp.106.084475

Xu, X., Li, X., Hu, X., Wu, T., Wang, Y., Xu, X., et al. (2017). High miR156 Expression Is Required for Auxin-Induced Adventitious Root Formation via MxSPL26 Independent of PINs and ARFs in Malus xiaojinensis. *Front Plant Sci.* 8:1059. doi: 10.3389/fpls.2017.01059

Xu, Y., Xu, H., Wall, M. M., and Yang, J. (2020). Roles of transcription factor SQUAMOSA promoter binding protein-like gene family in papaya (Carica papaya) development and ripening. *Genomics* 112, 2734–2747. doi: 10.1016/j.ygeno.2020.03.009

Yamaguchi, A., Wu, M. F., Yang, L., Wu, G., Poethig, R. S., and Wagner, D. (2009). The microRNA-regulated SBP-Box transcription factor SPL3 is a direct upstream activator of LEAFY, FRUITFULL, and APETALA1. *Dev. Cell* 17, 268–278. doi: 10.1016/j.devcel.2009.06.007

Yang, J., Moeinzadeh, M. H., Kuhl, H., Helmuth, J., Xiao, P., Haas, S., et al. (2017). Haplotype-resolved sweet potato genome traces back its hexaploidization history. *Nat. Plants* 3, 696–703. doi: 10.1038/s41477-017-0002-z

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

Yang, Z., Sun, J., Chen, Y., Zhu, P., Zhang, L., Wu, S., et al. (2019). Genome-wide identification, structural and gene expression analysis of the bZIP transcription factor family in sweet potato wild relative Ipomoea trifida. *BMC Genet.* 20:41. doi: 10.1186/s12863-019-0743-y

Yang, Z., Zhu, P., Kang, H., Liu, L., Cao, Q., Sun, J., et al. (2020). High-throughput deep sequencing reveals the important role that microRNAs play in the salt response in sweet potato (Ipomoea batatas L.). *BMC Genomics* 21:164. doi: 10.1186/s12864-020-6567-3

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

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

Yu, N., Yang, J. C., Yin, G. T., Li, R. S., and Zou, W. T. (2020). Genome-wide characterization of the SPL gene family involved in the age development of Jatropha curcas. *BMC Genomics* 21:368. doi: 10.1186/s12864-020-06776-8

Zhang, H. X., Jin, J. H., He, Y. M., Lu, B. Y., Li, D. W., Chai, W. G., et al. (2016). Genome-Wide Identification and Analysis of the SBP-Box Family Genes under

Phytophthora capsici Stress in Pepper (Capsicum annuum L.). *Front. Plant Sci.* 7:504. doi: 10.3389/fpls.2016.00504

Zhang, L., Yu, Y., Shi, T., Kou, M., Sun, J., Xu, T., et al. (2020). Genome-wide analysis of expression quantitative trait loci (eQTLs) reveals the regulatory architecture of gene expression variation in the storage roots of sweet potato. *Horticult. Res.* 7:90. doi: 10.1038/s41438-020-0314-4

Zhong, H., Kong, W., Gong, Z., Fang, X., Deng, X., Liu, C., et al. (2019). Evolutionary Analyses Reveal Diverged Patterns of SQUAMOSA Promoter Binding Protein-Like (SPL) Gene Family in Oryza Genus. *Front. Plant Sci.* 10:565. doi: 10.3389/fpls.2019.00565

Zhou, R., Yu, X., Ottosen, C. O., Zhang, T., Wu, Z., and Zhao, T. (2020). Unique miRNAs and their targets in tomato leaf responding to combined drought and heat stress. *BMC Plant Biol.* 20:107. doi: 10.1186/s12870-020-2313-x

Zhu, P., Dong, T., Xu, T., and Kang, H. (2020). Identification, characterisation and expression analysis of MADS-box genes in sweetpotato wild relative Ipomoea trifida. *Acta Physiolog. Plant.* 42:163. doi: 10.1007/s11738-020-03153-6

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.

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

frontiers in Plant Science

ORIGINAL RESEARCH
published: 10 February 2022
doi: 10.3389/fpls.2022.832884

Comparative Analysis the Complete Chloroplast Genomes of Nine *Musa* Species: Genomic Features, Comparative Analysis, and Phylogenetic Implications

Weicai Song[1], Chuxuan Ji[2], Zimeng Chen[1], Haohong Cai[1], Xiaomeng Wu[1], Chao Shi[1,3] and Shuo Wang[1]**

[1] College of Marine Science and Biological Engineering, Qingdao University of Science and Technology, Qingdao, China, [2] Department of Life Sciences, Imperial College London, Silwood Park, London, United Kingdom, [3] Plant Germplasm and Genomics Center, Germplasm Bank of Wild Species in Southwest China, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, China

OPEN ACCESS

Edited by:
Wei Hu,
Institute of Tropical Bioscience and Biotechnology, Chinese Academy of Tropical Agricultural Sciences (CATAS), China

Reviewed by:
Jin Xu,
Chinese Academy of Inspection and Quarantine (CAIQ), China
Zefu Wang,
Sichuan University, China
Yong Qi Zheng,
Chinese Academy of Forestry, China
Xinyi Guo,
Central European Institute of Technology (CEITEC), Czechia

***Correspondence:**
Chao Shi
chsh1111@aliyun.com
Shuo Wang
shuowang@qust.edu.cn

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

Received: *10 December 2021*
Accepted: *07 January 2022*
Published: *10 February 2022*

Citation:
Song WC, Ji CX, Chen ZM, Cai HH, Wu XM, Shi C and Wang S (2022) Comparative Analysis the Complete Chloroplast Genomes of Nine Musa Species: Genomic Features, Comparative Analysis, and Phylogenetic Implications. Front. Plant Sci. 13:832884. doi: 10.3389/fpls.2022.832884

Musa (family Musaceae) is monocotyledonous plants in order Zingiberales, which grows in tropical and subtropical regions. It is one of the most important tropical fruit trees in the world. Herein, we used next-generation sequencing technology to assemble and perform in-depth analysis of the chloroplast genome of nine new *Musa* plants for the first time, including genome structure, GC content, repeat structure, codon usage, nucleotide diversity and etc. The entire length of the *Musa* chloroplast genome ranged from 167,975 to 172,653 bp, including 113 distinct genes comprising 79 protein-coding genes, 30 transfer RNA (tRNA) genes and four ribosomal RNA (rRNA) genes. In comparative analysis, we found that the contraction and expansion of the inverted repeat (IR) regions resulted in the doubling of the *rps19* gene. The several non-coding sites (*psbI–atpA, atpH–atpI, rpoB–petN, psbM–psbD, ndhf–rpl32,* and *ndhG–ndhI*) and three genes (*ycf1, ycf2,* and *accD*) showed significant variation, indicating that they have the potential of molecular markers. Phylogenetic analysis based on the complete chloroplast genome and coding sequences of 77 protein-coding genes confirmed that *Musa* can be mainly divided into two groups. These genomic sequences provide molecular foundation for the development and utilization of *Musa* plants resources. This result may contribute to the understanding of the evolution pattern, phylogenetic relationships as well as classification of *Musa* plants.

Keywords: *Musa*, chloroplast genome, genetic structure, comparative analysis, phylogenetic analysis, interspecific relationships

INTRODUCTION

Musaceae is a small family of Zingiberales in monocotyledonous plants, mostly distributed in tropical regions in Australia, Africa, and Asia. It is closest to Strelitziaceae, Lowiaceae, and Heliconiaceae in phylogenetic position (Kress et al., 2001). Three genera are commonly recognized within Musaceae. *Ensete* is a small genus with eight to nine species found in Madagascar, sub-Saharan Africa and Asia, *Musella* is a monotypic genus native to southwest China (Li et al., 2010).

While most species of the family, which occur mainly in Southeast Asia, are classified into the *Musa* group (Häkkinen and Väre, 2008). *Musa* grow in tropical and subtropical regions and is one of the most important tropical fruit trees in the world. According to molecular analysis, wild *Musa* species are reclassified into two groups, *Musa* L. sect. *Musa* (by merging *Eumusa* with *Rhodochlamys*) and *Musa* sect. *Callimusa*, including the previously classified *M*. sect. *Australimusa* and *M*. sect. *Ingentimusa* (Häkkinen, 2013). Banana fiber has become one of the high potential biological resources in new material field due to it' s characteristics such as sustainability, low cost and environmental friendliness (Pappu et al., 2015; Vishnuvarthanan et al., 2019). For example, the leaf fibers of abaca (*Musa textilis*) are ideal raw materials for manufacturing specialty paper (del Río and Gutiérrez, 2006). Many organs of *Musa* plants are being used in various fields. Banana peels not only have effect in purifying Cr(III), Cr(VI), Cu(II), and radioactive substances (uranium and thorium) in water (Pakshirajan et al., 2013; Oyewo et al., 2016), but also were used as a new type of bio-sorbent to adsorb aflatoxins and ochratoxin A (Shar et al., 2016). Tree trunks and leaves can be used as precursors for the production of adsorbents for the purification of various pollutants (Ahmad and Danish, 2018). The dry biomass of banana pseudo stem can remove the reactive blue 5G (RB5G) dye (Jarvis and López-Juez, 2013). At the same time, many parts of banana can be used to produce industrial raw materials, such as ethanol, polyhydroxy butyrate (PHB), etc. (Oberoi et al., 2011; Ingale et al., 2014; Naranjo et al., 2014). Banana starch also plays an important role in the food, pharmaceutical, and cosmetic industries (Ramírez-Hernández et al., 2017; Arias et al., 2021; Taweechat et al., 2021; Thanyapanich et al., 2021).

Chloroplasts are an energy converter that provides energy for higher plants and algae, which are a unique structure of plant cells. At the same time, chloroplasts play a vital role in many functions of plant growth, including starch storage, sugar synthesis, the production of several amino acids, lipids, vitamins and pigments, essential sulfur and nitrogen metabolic pathways (Jarvis and López-Juez, 2013; Martin et al., 2013; Nielsen et al., 2016). In angiosperms, chloroplast (cp) genome is mainly a circular structure with the length is between 120–180 kb (Provan et al., 2001). The chloroplast genome is a circular double-stranded structure, which is divided into four parts, two of which are called single-copy regions, including a large single-copy region (LSC) and a small single-copy region (SSC) (Kolodner and Tewari, 1979), and the other two almost identical regions separating the single-copy regions are called inverted repeat sequences A and B (IRa, IRb) (Wicke et al., 2011). Compared with the nuclear and mitochondrial genomes, the chloroplast genome is relatively conserved in gene structure and composition (Asaf et al., 2017a). With the rapid development of Next Generation Sequencing (NGS), the National Center for Biotechnology Information (NCBI) database provides more and more chloroplast genomes, enabling people to have a better understanding of the relationship between chloroplast structure and genetic evolution, which also heavily facilitated the research of chloroplast genomes (Yang et al., 2014; Li et al., 2017; Amiryousefi et al., 2018b). The polymorphic sites of the chloroplast genome can be used to develop reliable and stable molecular markers, which will help us to study population genetics and phylogeny (Ahmed et al., 2013; Sheng et al., 2021).

The relatively conservative chloroplast genome is an ideal research method for studying genetic relationship identification. It is of great significance to analyze the chloroplast genome of *Musa*, including structural characteristics, phylogenetic relationships and population genetics. As a supplementary technology, chloroplast sequencing not only provide part of the genetic diversity information about *Musa* germplasm resources, but also clarifies the genes and potential functions of *Musa* plants. So far, the complete chloroplast sequences of *Musa* plants have been obtained in *Musa acuminata* (Martin et al., 2013), *Musa balbisiana* (Shetty et al., 2016), *Musa beccarii* (Feng et al., 2020) and *Musa ornate* (Liu et al., 2018) and so forth. Here, we reported the complete chloroplast genomes of nine *Musa* species, which was the first comprehensive comparison of these nine species. We compared the structure and content patterns of nine *Musa* chloroplast genomes; explored the sequence differences in nine *Musa* cp genomes; detected simple sequence repeats (SSR) and long repeats; calculated codon usage bias and putative RNA editing site. We also studied the genetic variation between *Musa* species, including inverted repeat (IR) contraction/expansion; gene duplication and loss during evolution; the ratio of non-synonymous (K_a) to synonymous substitutions (K_s), which may help uncover the genetic relationship between *Musa* species. We also performed phylogenetic analyses using chloroplast genome sequences from other related species to further determine the taxonomy of *Musa* genus. These results perfect the existing genetic information of *Musa* species and provide a valuable reference for the DNA molecular research of *Musa* species. Application of these results will help assess the genetic variation and phylogenetic relationships between closely related species and support the development of wild germplasm resources.

MATERIALS AND METHODS

Sample Collection, DNA Extraction, and Sequencing Plants

In this study, the nine species of *Musa* were collected from Plant Germplasm and Genomics Center, Kunming Institute of Botany, the Chinese Academy of Sciences, and was approved by Kunming Institute of Botany and local policy. The voucher specimen and DNA were deposited at Qingdao University of Science and Technology (specimen code BJ210253-BJ210261). Total genomic DNA was extracted from fresh leaves using modified CTAB (Porebski et al., 1997). According to the manufacturer's protocol, the Illumina TruSeq Library Preparation Kit (Illumina, San Diego, CA, United States) was used to prepare approximately 500 bp of paired-end libraries for DNA inserts. These libraries were sequenced on the Illumina HiSeq 4000 platform in Novogene (Beijing, China), generating raw data of 150 bp paired-end reads. About 3 Gb high quality, 2 × 150 bp pair-end raw reads were obtained and were used to assemble the complete chloroplast genome of *Musa*.

Chloroplast Genome *de novo* Assembly and Annotation

Trimmomatic 0.39 software were used preprocessed the raw data (Bolger et al., 2014), including removal of adapter sequences and other sequences introduced in the sequencing, removing low-quality and over-N-base reads, etc. The quality of newly produced clean short reads was assessed using FASTQC v0.11.9 and MULTIQC software (Ewels et al., 2016), and high-quality data with Phred scores averaging above 35 were screened out. According to the reference sequence (*Musa balbisiana*), the chloroplast-like (cp) reads were isolated from clean reads by BLAST (Shetty et al., 2016). Short reads were *de novo* assembled into long contigs using SOAPdenovo 2.04 (Luo et al., 2012) by setting kmer values of 35, 44, 71, and 101. Furthermore, the long-contigs was expanded and gap-filed using Geneious ver 8.1 (Muraguri et al., 2020), which forms the whole chloroplast genome. The complete chloroplast genome was further validated and calibrated by using *de novo* splicing script NOVOplsty 4.2 (Dierckxsens et al., 2017). In addition, GeSeq (Tillich et al., 2017) was used to annotate the *de novo* assembled genomes, RNAmmer (Lagesen et al., 2007) was used to validate rRNA genes with default settings, and tRNAscanSE ver 1.21 (Lowe and Eddy, 1997) was applied to detect tRNA genes with default settings. Finally, we compared the results with the reference sequence and corrected the misannotated genes by GB2Sequin (Lehwark and Greiner, 2019) in an artificial way. The circular map of the genomes was drawn by using Organellar Genome DRAW (OGDRAW) (Lohse et al., 2007). The nine newly assembled *Musa* chloroplasts genomes were deposited in GenBank with the accession numbers NC_056826 - NC_056834.

Plastome Structural Analysis

Chloroplast Microsatellites or simple sequence repeats (SSRs) were detected in the perl script MISA (Beier et al., 2017). The basic repeat setting of SSRs was determined: ten for mononucleotide, five for dinucleotide, four for trinucleotide and three for tetranucleotide pentanucleotide hexanucleotide. The REPuter tool (Kurtz et al., 2001) was applied to analyze forward (F), reverse (R), complement (C), and palindromic (P) oligonucleotide repeats. The following parameters were used to identify repeats with: (1) hamming distance equal to 3; (2) minimal repeat size set to 30 bp; and (3) maximum computed repeats set to 300 bp. Relative synonymous codon usage (RSCU) and amino acid frequency in the protein coding gene region were determined by MEGA-X (Kumar et al., 2018). The putative RNA editing sites in 35 genes were investigated in the coding gene using PREP-cp (Predictive RNA Editors for Plants chloroplast) (Mower, 2009).

Genome Comparison

We compared and analyzed the basic features of nine chloroplast genomes using Geneious software, including calculating the length of the region sequence, GC content in different regions, and the proportions of different sequences. The junction sites of various regions of the chloroplast genome were analyzed in IRscope (Amiryousefi et al., 2018a) to visualize the expansion and contraction of reverse repeats (IR). We used KaKs_Calculator 2.0 software (Wang et al., 2010) to calculate the rate values of K_s (synonymous substitution) and K_a (non-synonymous substitution) with the YN method. Shuffle-LAGAN mode alignment program in mVISTA (Brudno et al., 2003) was used to evaluate structural similarity for the nine species, with the annotation of *M. balbisiana* as the reference.

Phylogenetic Analysis

The complete chloroplast genomic sequences from 17 species of *Musa* (nine sequences newly generated and eight species obtained from GenBank) were performed for phylogenetic analyses (**Supplementary Table 8**). *Heliconia collinsiana* (accession number NC_020362) and *Ravenala madagascariensis* (accession number NC_022927) were downloaded from the NCBI (National Center of Biotechnology Information) as an outgroup of the evolutionary tree. Multiple sequence alignment was aligned using MAFFT and GTR-GAMMA (GTR + G) model was selected using model test applying the Bayesian information criterion (BIC) (Posada and Crandall, 1998). All InDels were excluded from the alignment sequence to construct a phylogenetic tree based on only substitutions. The maximum likelihood (ML) trees were conducted by MEGA-X and 1,000 bootstrap replicates were set to evaluate the branch support values. Finally, the 79 protein-coding genes from the 19 species were also extracted to reconstructed ML trees using the same methods.

RESULTS

Assembly and Annotation of the Chloroplast Genomes of Nine *Musa* Species

Genome-skimming data were generated about 3.2–5.7 GB by the Illumina HiSeq 2500 in each of the sequenced *Musa* species. The complete chloroplast genomes of these nine species were typical circular double-stranded structures and ranged from 167,975 bp (*Musa jackeyi*) to 172,653 bp (*Musa rubinea*) (**Table 1**). All nine sequence presented the quadripartite structure, including large single copy (LSC) region, the small single copy (SSC) region and a pair of inverted repeat (IR) regions. The length of the LSC region ranged between 88,330 and 89,997 bp, with the GC content of 34.8–35.2%. The length of the SSC region was distributed between 10,773 and 11768 bp. The GC content of SSC regions was similar in nine species, ranging from 30.1% in *M. rubinea* to 31.2% in *Musa laterita*. 33,864–35,522 bp was the length range of the IR region of nine *Musa* species, which contains 39.5–40.0% GC content. The complete chloroplast genome sequences of the nine *Musa* species were provided in GenBank (under accession number NC_056826–NC_056834).

Although the length of the chloroplast genomes of the nine species was some different, the analyses of the genetic composition showed that they have some similarities. The

TABLE 1 | Chloroplast genome features of nine species of *Musa*.

Genome features	*Musa ingens*	*Musa jackeyi*	*Musa laterita*	*Musa lolodensis*	*Musa mannii*	*Musa nagensium*	*Musa rubinea*	*Musa troglodytarum*	*Musa yunnanensis*
Genome size (bp)	168,471	167,975	170,565	168,542	170,699	170,304	172,653	168,121	169,816
LSC size (bp)	89,888	88,422	88,748	88,330	88,883	88,420	89,997	88,724	90,720
SSC size (bp)	10,855	11,049	10,773	11,060	10,816	11,082	11,768	11,049	11,072
IR size (bp)	33,864	34,252	35,522	34,576	35,500	35,401	35,444	34,174	34,012
Total GC content (%)	36.8	36.9	36.8	36.8	36.8	36.6	36.4	36.8	36.8
GC content in LSC (%)	35.2	35.1	35.2	35.2	35.1	35.0	34.8	35.1	35.1
GC content in SSC (%)	31.0	31.1	31.2	31.1	31.2	30.8	30.1	31.1	31.1
GC content in IR (%)	39.6	40.0	39.5	39.9	39.5	39.5	39.5	40.0	39.9
Number of genes (unique)	135(113)	135(113)	136(113)	135(113)	136(113)	136(113)	136(113)	135(113)	136(113)
Protein-coding genes (unique)	89(79)	89(79)	90(79)	89(79)	90(79)	90(79)	90(79)	89(79)	90(79)
tRNA genes (unique)	38(30)	38(30)	38(30)	38(30)	38(30)	38(30)	38(30)	38(30)	38(30)
rRNA genes (unique)	8(4)	8(4)	8(4)	8(4)	8(4)	8(4)	8(4)	8(4)	8(4)
Accession numbers in GenBank	NC_056826	NC_056827	NC_056828	NC_056829	NC_056830	NC_056831	NC_056832	NC_056833	NC_056834

positions of the genes were visualized in **Figure 1**. A total of 135 functional genes were predicted in all nine *Musa.* sps, including 113 unique genes comprising 79 protein-coding genes, 30 transfer RNA (tRNA) genes and four ribosomal RNA (rRNA) genes. These genes, represented by *Musa nagensium*, can be roughly divided into three categories: photosynthesis-related genes, chloroplast self-replication genes, and other genes (**Table 2**). Among the genes, 18 intron-containing genes (ICG) were found, covering 12 protein-coding genes and 6 tRNA genes (**Supplementary Table 1**). Among these ICG, *ycf3*, and *clpP* possessed two introns, respectively, while the rest of ICG contained only one intron. The *rps12* gene has *trans*-splicing, and its 3′-end is duplicated in the IRs region, while its 5′-end is present in the LSC region. As a regional demarcation gene, the *ndhA* gene starts at the IRs region and ends at the SSC region.

Codon Usage Bias

In this study, we analyzed the codon usage bias and relative synonymous codon usage (RSCU) based on the protein coding gene of *Musa*'s chloroplast genome, and a total of 28,690–29,360 codons were identified (**Supplementary Table 2**). Analysis showed that codons containing A or T instead of C or G at the 3′-end of the codon have a higher encoding rate. The RSCU of codons containing A/T at the 3′-end was mainly greater than 1, and the codons containing C or G at the 3′-end mostly have RSCU $\leq$ 1. In addition, there were 29 codons with RSCU values greater than 1, 2 of them were equal to 1, and 30 of them are less than 1. Among them, AUU (4.15–4.24%, Isoleucine), AAA (4.15–4.36%, Lysine), and GAA (4.21–4.39%, Glutamic acid) were the most frequently used codons, while UGC (0.30–0.31%, Cysteine) and CGC (0.31–0.33%, Arginine) had the lowest usage rates. In addition, most amino acids possessed at least two synonymous codons, except for methionine (AUG) and tryptophan (UGG), which had no codon usage preference since they only have one coding codon. Among all codons with an RSCU value greater than 1, the vast majority of codons presented a higher A/T appreciation in the third codon. Overall, we found that the nine *Musa* species have high similarities in codon usage and amino acid frequency. This result is very common in the chloroplast genome of higher plants (Gichira et al., 2017).

Positive Selection Analysis and Putative RNA Editing Site

The ratio of non-synonymous (K_a) to synonymous substitutions (K_s), K_a/K_s, has been widely used to evaluate the natural selection pressure and evolution rates of nucleotides in genes (Li et al., 1985). The results of the statistical neutrality test indicated that 77 protein-coding genes were relatively stable during the evolution process, but two genes (*ycf1* and *ycf2*) were under positive selection (**Supplementary Table 3**). The K_a/K_s ratio of the *ycf2* gene of the nine species in *Musa* are all greater than 1 (2.66–5.22). Except for the K_a/K_s ratio of the *ycf1* gene of *Musa mannii* (0.9), that of the other eight species are also all positive selection status (1.16–2.29).

In order to gain a deeper insight into the RNA metabolism of *Musa* species, we used PREP to predict 74–77 post-transcriptional RNA editing modifications of 26 protein-coding genes (**Supplementary Table 4**). Most RNA editing sites were located in *ndhB* (11 editing sites, 14.3–14.8%), while *ndhD* (5–7 editing sites, 6.6–9.2%), *ndhF* (6–7 editing sites, 7.8–9.4%), and *rpoB* (5–6 editing sites, 6.7–8.0%) also had a great portion of editing sites. The types of RNA editing sites reported here were all C to U and all affect a single site. All changes occurred in the first or second nucleotides of the codon. Among the amino acid conversions caused by RNA editing sites, the transformation of serine to leucine accounted for one-third of the total.

Inverted Repeat Contraction and Expansion

Our research revealed that all nine *Musa* species have *ndhA* genes that spanned the SSC and IRa regions (**Figure 2**). Only the *ndhA* gene of *Musa yunnanensis* was longer in the SSC region than in the IRa region. In comparison, the length of the *ndhA* gene of the remaining eight species were not much different in the SSC and IRa regions. We speculated that this may be due to the expansion

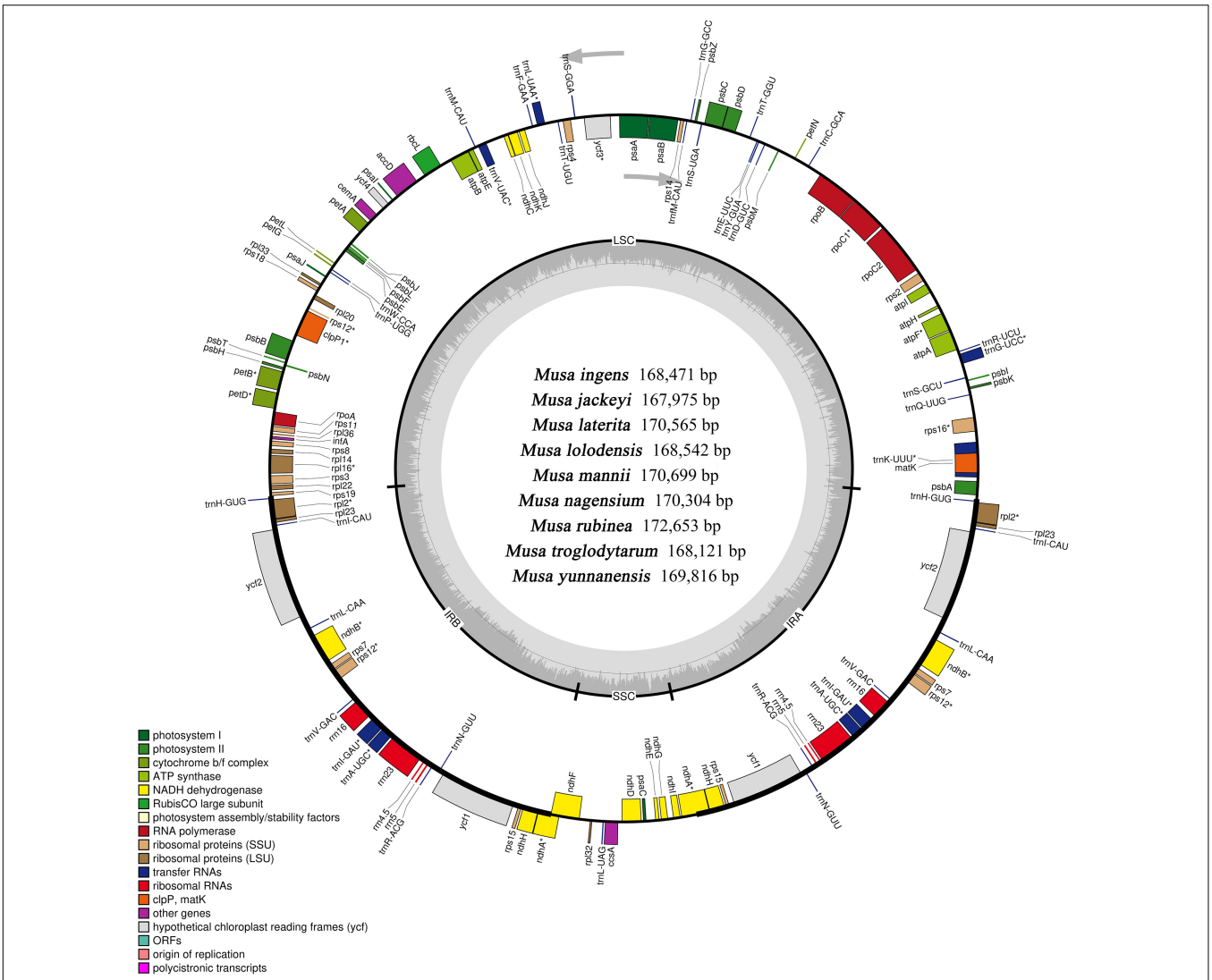

FIGURE 1 | Complete genome map of the chloroplast genome of *Musa*. The inner gray ring is divided into four areas, Clockwise, and they are: small single-copy region (SSC), inverted repeat sequences B (IRb), large single-copy region (LSC), and inverted repeat sequences A (Ira). The genes in the outer ring region are transcribed clockwise, while those in the inner ring are transcribed counterclockwise. The inner ring dark gray indicates the GC content, the light gray reaction AT content. In the lower left is a legend that classifies chloroplast genes according to their functions.

of the IRa region of *M. yunnanensis*. Primarily, at the junction of LSC/IRb (JLB), the *rpl2* gene was located in the IRb region, while the *rpl2* genes of *Musa ingens* and *M. yunnanensis* spanned the LSC and the IRb region. According to the distribution of *rps19* gene, nine *Musa* species can be roughly divided into three categories. The *rps19* gene of *M. laterita, M. mannii, M. nagensium,* and *M. rubinea* in the first category were entirely located in the IR region, 100–131 bp apart from LSC/IRb and IRa/LSC. The second type of species (*M. jackeyi, Musa lolodensis*, and *Musa troglodytarum*) were where the *rps19* gene was situated at the junction of LSC/IRb and were 18–19 bp away from the IRb region. Moreover, the *rps19* genes were entirely located in the IRb region (*M. ingens and M. yunnanensis*), suggesting that this phenomenon may occur with the contraction IRb area. However, at the junction of IRa/LSC (JLA), *M. yunnanensis* processed two *rps19* genes, so we speculated that *rps19* was deleted in *M. ingens*.

Repeat Sequence and Simple Sequence Repeats Analysis (Analysis of Microsatellites and Oligonucleotide Repeats)

This study counted all the interspersed repetitive sequences in the *Musa* chloroplast genome with a repeat unit length of more than 30 bp. At the same time, we detected four types of repeats, including forward repeats (F), inverted repeats (R), complementary repeats (C), and palindromic repeats (P) (**Supplementary Table 5**). Repeat analysis showed 50–170 forward duplications, 0–26 inverted duplications, 0–13

TABLE 2 | List of predicted genes in the *Musa* chloroplast genome.

Category for gene	Group of gene	Name of gene				
	Subunits of Photosystem I	*psaA*	*psaB*	*psaC*	*psaI*	*psaJ*
	Subunits of Photosystem II	*psbA*	*psbB*	*psbC*	*psbD*	*psbE*
		psbF	*psbH*	*psbI*	*psbJ*	*psbK*
		psbL	*psbM*	*psbN*	*psbT*	*psbZ*
Genes for photosynthesis	Subunits of NADH oxidoreductase	*ndhA*[ab]	*ndhB*[ab]	*ndhC*	*ndhD*	*ndhE*
		ndhF	*ndhG*	*ndhH*[a]	*ndhI*	*ndhJ*
		ndhK				
	Cytochrome b6/f complex	*petA*	*petB*	*petD*	*petG*	*petL*
		petN				
	Subunits of ATP synthase	*atpA*	*atpB*	*atpE*	*atpF*[b]	
		atpI				
	Large subunit of RuBisCo	*rbcL*				
	Large subunit of ribosomal proteins	*rps2*	*rps3*	*rps4*	*rps7*[a]	*rps8*
		rps11	*rps12*[ab]	*rps14*	*rps15*[a]	*rps16*[b]
		rps18	*rps19*[a]			
	Small subunit of ribosomal proteins	*rpl2*[ab]	*rpl14*	*rpl16*	*rpl20*	*rpl22*
		rpl23[a]	*rpl32*	*rpl33*	*rpl36*	
	DNA-dependent RNA polymerase rRNA	*rpoA*	*rpoB*	*rpoC1*[b]	*rpoC2*	
Self-replication	Ribosomal RNA genes	*rrn4.5*[a]	*rrn5*[a]	*rrn16*[a]	*rrn23*[a]	
	Transfer RNA genes	*trnA-UGC*[ab]	*trnC-GCA*	*trnD-GUC*	*trnE-UUC*	*trnF-GAA*
		trnfM-CAU	*trnG-GCC*[ab]	*trnH-GUG*[a]	*trnI-CAU*[a]	*trnI-GAU*[ab]
		trnK-UUU[b]	*trnL-CAA*[a]	*trnL-UA*[ab]	*trnL-UAG*	*trnM-CAU*
		trnN-GUU[a]	*trnP-UGG*	*trnQ-UUG*	*trnR-ACG*[a]	*trnR-UCU*
		trnS-GCU	*trnS-GGA*	*trnS-UGA*	*trnT-GGU*	*trnT-UGU*
		trnV-GAC[a]	*trnV-UAC*[b]	*trnW-CCA*	*trnY-GUA*	
	Translational initiation factor	*infA*				
Other genes	Maturase	*matK*				
	Protease	*clpP*[c]				
	Envelope membrane protein	*cemA*				
	Subunit acetyl-CoA-carboxylase	*accD*				
	c-Type cytochrome synthesis gene	*ccsA*				
	Conserved open reading frames	*ycf1*[a]	*ycf2*[a]	*ycf3*[c]	*ycf4*	

[a] *Two gene copies in IRs;* [b] *gene containing a single intron;* [c] *gene containing two introns.*

complementary repeats and 37–140 palindromic repeats in nine *Musa* species (**Figure 3A**). The length of the repetitive sequence varied from species to species, but most of the repetitive sequence length existed in the range of 30–50 bp (40.76–82.1%) (**Figure 3B**). Compared with the LSC and SSC regions, the IR region contained most of the repetitive sequences, and the chloroplast genome regions also shared most of the repetitive sequences. Among them, the repetitive sequences in the IR region of *M. nagensium* accounted for the highest proportion of all repetitive sequences (96.3%), and the IR region of *M. troglodytarum* had the lowest proportion of repetitive sequences (60%) (**Figure 3C**).

We analyzed the simple sequence repeats (SSRs) in the chloroplast genomes of nine *Musa* species (**Figure 3D**). A total of six types of SSR (mono-/di-/tri-/tetra-/penta-/hexa-nucleotide repeats) were detected, the first four microsatellites accounted for 86.17–94.52%, and the penta- or hexanucleotide repeats was very small (no more than 8) or even non-existent (**Figure 3E**). In the MISA analysis, the number of SSRs detected in the nine *Musa* species was 73–93. At the same time, the distribution of SSR in the LSC region (61.54–72.29%) was higher than that in the IR region (21.69–28.57%) and SSC region (6.02–11.54%) (**Figure 3F**). Analysis revealed that SSRs were mainly distributed in the non-coding areas (51.9–60.26%). The number of SSRs in the coding region of *M. rubinea* (41) were the largest, while that of *M. mannii* (31) was the lowest.

Sequence Divergence in the Nine *Musa* Chloroplast Genomes

Using *M. Balbisiana* as a reference, we used the DnaSP6 to detect single nucleotide polymorphisms (SNPs) in the chloroplast genomes of nine *Musa* species (**Table 3**). Through analysis, we divided these nine species into two groups. The first group contained four species (*M. ingens, M. jackeyi, M. lolodensis,* and *M. troglodytarum*). In comparison, the second group comprised five species (*M. laterita, M. mannii, M. nagensium, M. rubinea*, and *M. yunnanensis*). Among them, 1,419–1,459 SNPs were

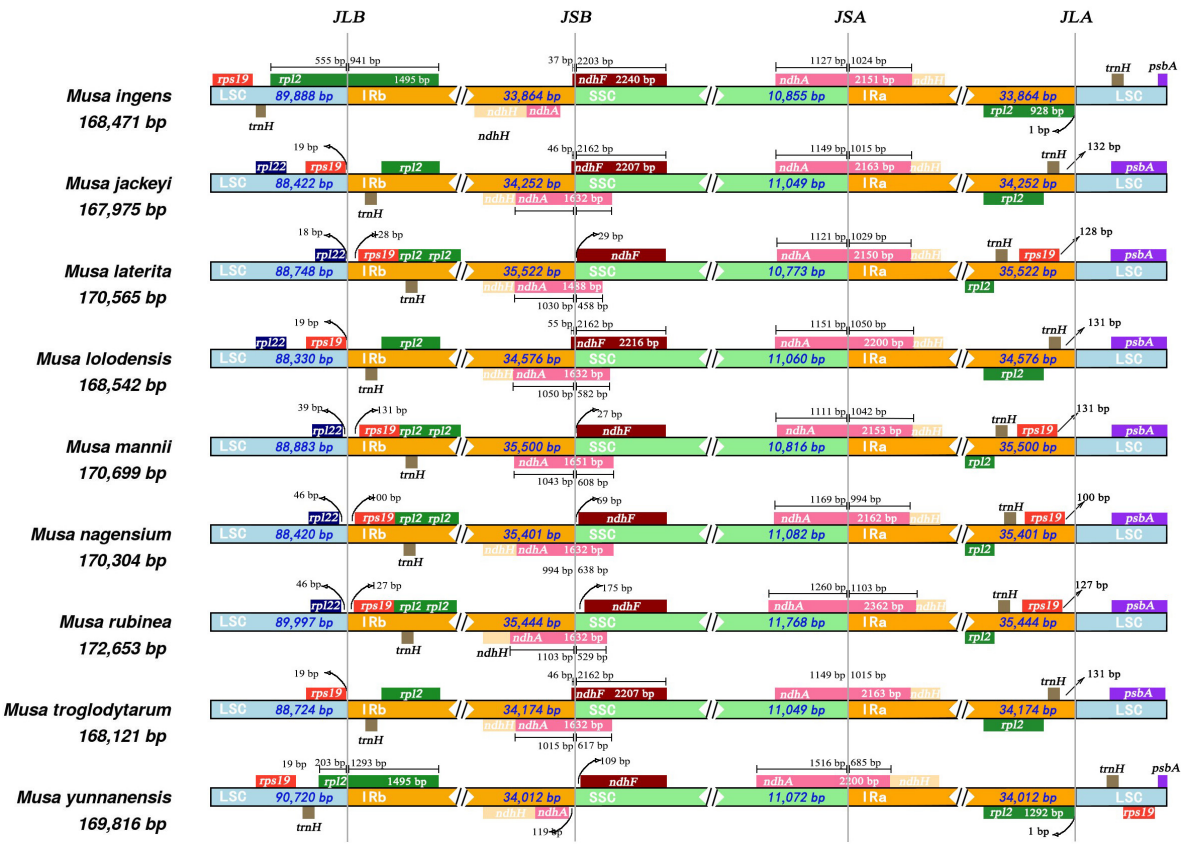

FIGURE 2 | Comparison of SSC, LSC, IRB, and IRB boundary regions in the chloroplast genomes of nine species of *Musa*. Comparative analysis of junction sites in *Musa* chloroplast genomes. The coordinate position of the start or end of each gene from the corresponding junction is shown with arrows. All the genes those integrate from one region of chloroplast genome to another region are shown with the T bar above or below. The T bars show the length of base pair for which the integration of genes has been occur. JLA (IRa/LSC), JLB (IRb/LSC), JSA (SSC/IRa), and JSB (IRb/SSC) denote the junction sites between each corresponding two regions on the genome.

detected in the first group of four species, and 628–716 SNP sites were seen in the second group. We found that the distribution of SNPs of the nine species in the LSC region is not much different, and the SNP of each species accounted for the highest proportion in the LSC region (61.17–65.29%) (**Supplementary Table 6**). However, in the statistics of SNP content in the IR region, the ratio of the first group (16.30–19.97%) was slightly lower than that of the second group (21.85–22.62%), and the proportion of SNP in the SSC region was slightly higher (the first group: 15.45–21.04%, the second group: 14.16–14.59%). The mutation frequencies of the corresponding LSC, SSC, and IR regions of the first group were 1.020–1.035, 1.846–1.953, and 0.448–0.487%, respectively, while the second group was 0.452–0.497, 0.876–1.281, and 0.155–0.202%. We also analyzed the insertions and deletions of the chloroplast genomes of nine species, which found they have similar rules in SNPs. 126–160 insertions were detected in nine species, respectively. The detection rates of LSC, SSC and IR were 61.90–67.42, 3.79–14.81, and 20.99–30.60%. The deletion mainly occurred in the LSC region (the first group: 58.66–66.64%, the second group: 59.09–66.04%), followed by the IR region (the first group: 26.06–30.60%, the second group 16.39–24.82%), and finally, the SSC region (the first group: 9.84–15.08%, the second group: 13.21–22.13%).

Comparative Genomic Analysis in the Nine *Musa* Chloroplast Genomes

To study the level of sequence polymorphism, we used DnaSP6 and mVISTA programs to calculate the genetic differences between nine *Musa* plants and compared the whole chloroplast genomes (**Figure 4**) with reference sequence of *M. balbisiana* set. In this study, the IR region variation of the *Musa* chloroplast genome was lower than that of LSC and SSC. In the coding region, the *ycf1*, *accD,* and *ycf2* of were quite different from each other. In general, non-coding regions were more susceptible to mutations than coding regions. Chloroplast genome of *Musa* is also consistent with this characteristic, and high variable regions are mainly found in IGS, such as *psbI–atpA, atpH–atpI, rpoB–petN, psbM–psbD, ndhF–rpl32, psaC–ndhE,* and *ndhG-ndhI.* These hot spots can be applied

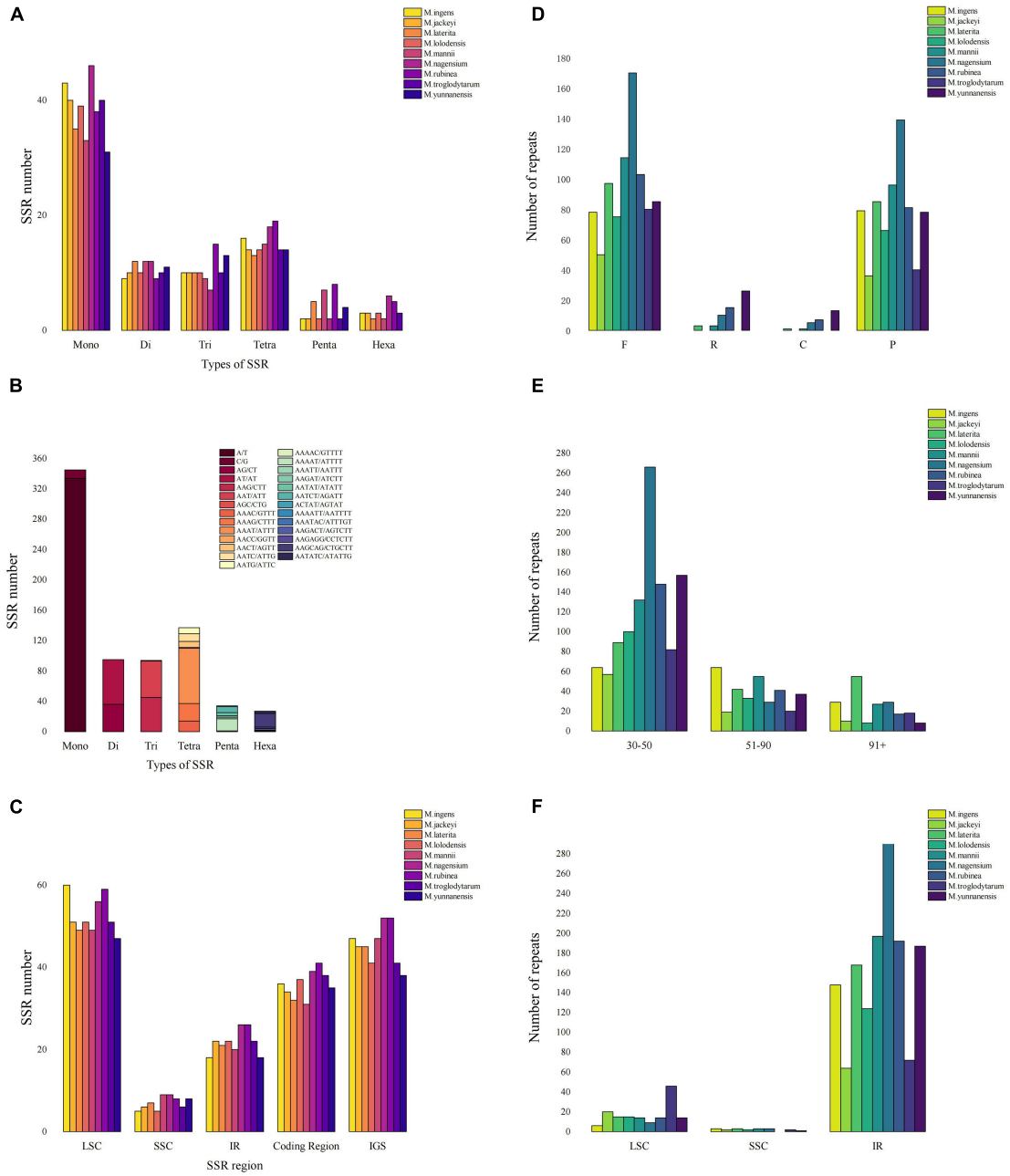

FIGURE 3 | Comparison of microsatellites and long repeats in the chloroplast genomes of *Musa* species. **(A)** The number of SSRs of different types of SSR for nine *Musa* species. **(B)** Details in SSR types among nine *Musa* species. **(C)** The number of SSR markers in the LSC/SSC/IR region along with coding region and IGS. **(D)** Number of four long repeat sequences in nine species: complement repeats. F represents forward repeats, P represents palindromic repeats, R represents reverse repeats, C represents complement repeats. **(E)** Number of long repeat sequences with different lengths in nine species. **(F)** The distribution of long repeats in LSC, SSC and IR regions.

to DNA barcode encoding and phylogenetic analysis of *Musa* genus. With the rapid development of the chloroplast genome, comparing the differences in chloroplast genome sequences of different taxa can, it not only effectively screen out information-rich DNA fragments, but also promote the development of species identification and population diversity. The nucleotide variation (Pi) of nine species ranged from 0 to 0.08264, with an average value of 0.00792 (**Supplementary Table 7**). The average Pi of the SSC area was 0.01188, the average Pi of the LSC area was 0.00862, and the average Pi of the IR area was 0.00502. It can be seen that inverted repeats were more conservative than the single copy regions, and the coding regions were more conservative than the non-coding regions (**Figure 5**).

TABLE 3 | Details of single nucleotide polymorphisms (SNP) and InDel sites in large single-copy region (LSC), small single-copy region (SSC), and inverted repeat (IR) regions in the complete chloroplast genomes of nine *Musa* species.

SNP and indel			*M. ingens*	*M. jackeyi*	*M. laterita*	*M. lolodensis*	*M. mannii*	*M. nagensium*	*M. rubinea*	*M. troglodytarum*	*M. yunnanensis*
Number	SNP	Transition	954	917	361	912	349	322	365	915	347
		Transversion	505	524	309	507	342	353	351	528	281
		Total	1459	1441	670	1419	691	675	716	1443	628
	Insertion		160	132	134	155	143	139	162	132	126
	Deletion		179	180	122	165	132	88	106	183	141
Region	SNP	LSC	917	915	423	902	442	423	438	914	410
		SSC	212	204	114	207	116	142	135	205	97
		IR	330	322	133	310	133	110	143	324	121
	Insertion	LSC	103	87	84	96	94	90	104	89	78
		SSC	14	5	9	14	12	15	24	14	12
		IR	43	40	41	45	37	34	34	29	36
	Deletion	LSC	105	109	75	105	78	56	70	109	86
		SSC	27	19	27	17	29	15	14	18	20
		IR	47	52	20	43	25	17	22	56	35

Phylogenetic Analyses

To further understand the phylogenetic status of *Musa* plants and their relationship with other closely related species, the chloroplast whole genome and the shared protein-coding genes of 22 *Zingiberales* plants (including 18 *Musa* species) were used to constructed phylogenetic tree using maximum-likelihood (ML) method and bootstrap with 1,000 times iteration (**Figure 6**). The 22 *Zingiberales* plants were clustered as a large group, including many important crops, such as abaca (*M. textilis*), an excellent raw material for making specialty paper, and the primary sources of high-quality fiber-Abacá and Chinese dwarf banana, which was regarded as an essential Chinese medicinal material and rare ornamental plant, etc. The bootstrap values for almost all relationships inferred from all chloroplast genome data were very high. The results of the evolutionary tree we constructed can be divided into approximately four parts, which are *M. mannii–M. yunnanensis*, *M. balbisiana–Musa formobisiana*, *Musa coccinea–M. troglodytarum* and outgroups. In third part of the two evolutionary trees, there is a slight divergence, the **Figure 6A** showed that *M. textilis* and the sub-branches containing *M. beccarii*, *M. lolodensis*, *M. jackeyi,* and *M. troglodytarum* had sister relationship, and **Figure 6B** indicated that *M. textilis* and *M. beccarii* were sister species. *Musa lasiocarpa* is closer to *Ensete glaucum* of outgroups than to other 17 *Musa* species. The phylogenetic tree of nine *Musa* plants showed that *Musa* L. sect. *Musa* and *Musa* sect. *Callimusa* had a sister relationship, which was further verifying the latest *Musa* species classification (Weiner et al., 2019).

DISCUSSION

Comparison of Chloroplast Genomes in the *Musa* Species

The chloroplast genome of angiosperms has made essential contributions to the study of phylogeny and the analysis of evolutionary relationships in phylogeny (Lee et al., 2019). The rich information in the chloroplast genome is very suitable as a DNA barcoding for species identification (Millen et al., 2001). However, among the 86 species belonging to *Musa* genus, there was very little analysis of complete chloroplast genomes. At this stage, only the complete chloroplast genomes of few species have been reported (Martin et al., 2013; Shetty et al., 2016; Liu et al., 2018; Feng et al., 2020), herein, we have added nine *Musa* species. The chloroplast genomes of most land plants are highly conserved, while during the evolution of angiosperms, one of the most fluid chloroplast genes, *infA*, was discovered (Millen et al., 2001). The *chlB*, *chlL*, *accD*, *ycf1*, *ycf68*, *infA*, *ycf15*, *ycf2*, *rpl22*, *rps16*, *rpl23*, *ndhF*, *chlN,* and *trnP* (GGG) genes in the plastid genome of some angiosperms were observed to be missing (Liaud et al., 1990; Liu and Xue, 2005; Jansen et al., 2007; Sheng et al., 2021). Among them, the deletion of four genes [*chlB*, *chlL*, *chlN* and *trnP* (GGG)] represents the homomorphism of flowering plants (Shahzadi et al., 2020). The deletions of the above four genes were found in the chloroplast genomes of all nine *Musa* species, including the missing of *ycf15* and *ycf68*. *M. laterita*, *M. mannii*, *M. nagensium*, *M. rubinea,* and *M. yunnanensis* all had two *rps19* genes, but only one in the chloroplast genomes of the other four species. This phenomenon is consistent with the classification of previous studies that the first five *Musa* sps. belong to *Musa* L. sect. *Musa*, and the last four species belong to *Musa* sect. *Callimusa* (Jiang et al., 2017).

Codon usage bias helped revealing the interaction between the chloroplast genome and its nuclear genome (Yang Y. et al., 2018). In many previous studies, the codons for leucine and isoleucine are the most common codons in the chloroplast, and the codons for cysteine are the least (Asaf et al., 2017b; Yang Y. et al., 2018; Shahzadi et al., 2020). The nine *Musa* species in this study also meet this feature. In the chloroplast genome of angiosperms, most codons showed higher A/T preference in the third codon. Our results followed this trend, and this phenomenon was also observed in *Forsythia suspensa* (Tian et al., 2018), *Epipremnum aureum* (Abdullah et al., 2020),

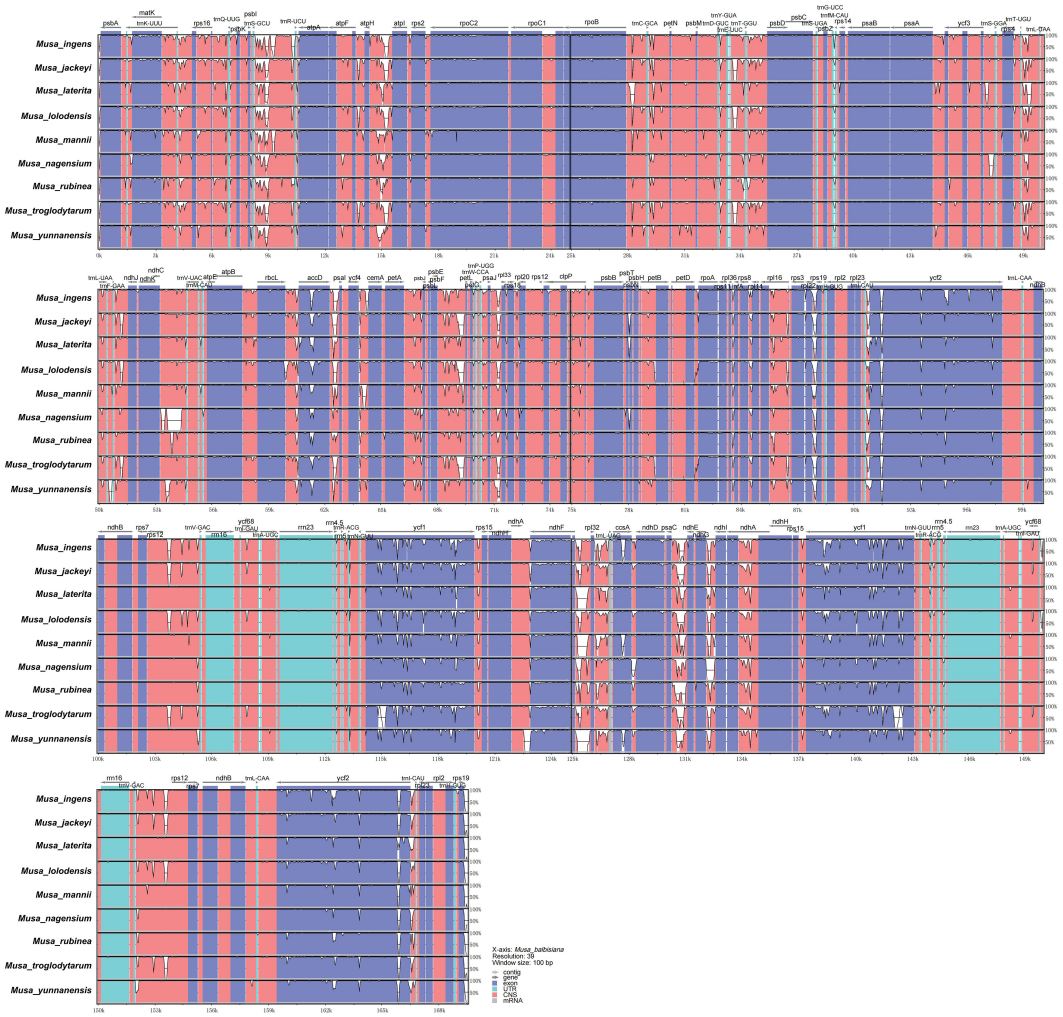

FIGURE 4 | mVISTA map of chloroplast genome of nine species of *Musa*. Sequence identity plot comparing the chloroplast genome of nine *Musa* species. The vertical scale indicates the percentage of identity, ranging from 50 to 100%. The horizontal axis indicates the coordinates within the chloroplast genome. Genome regions are color-coded as protein-coding, rRNA, tRNA, intron, and conserved non-coding sequences (CNS).

Zingibereae sp. (Saina et al., 2018a), two Artemisia species (Piot et al., 2018), and other species. The main reason for this situation may be related to the abundance of A or T in the IR region (Chen et al., 2015).

Long repeats (LR) were essential when analyzing genome reorganization, rearrangement, and phylogeny, or inducing substitutions and insertions in the chloroplast genome (Chumley et al., 2006). We detected 86–324 LRs in nine *Musa* species, most of which were located in the IR region. This phenomenon was different from some species (Tian et al., 2018; Abdullah et al., 2020; Zhu et al., 2021). The IR regions of *Musa* sp. stabilizes plastid chromosomes through a repair mechanism induced by homologous recombination (Maréchal and Brisson, 2010). At the same time, our analysis shows that the proportion of LRs of *M. laterita*, *M. mannii*, *M. nagensium*, *M. rubinea,* and *M. yunnanensis* in the IR regions were greater than that of the other four species, which also may play a role in the genetic diversity and evolution of different *Musa* branches.

In the chloroplast genome, SSR was considered an important role in population genetics and phylogenetic analysis (Terrab et al., 2006). The number of SSRs were detected in the nine *Musa* species ranged from 73 to 93. The distribution of SSRs in the LSC region was higher than that in the IR and SSC region. At the same time, analysis shows that SSRs were mainly distributed in non-coding regions. These results were supported by previous studies on the chloroplast genome of angiosperms (Kim et al., 2009; Xu et al., 2012; Cheng et al., 2017). The SSRs analysis in this study showed that single nucleotide SSRs (A/T) had the highest content among the nine *Musa* plants, reaching 334, and mono-/di-/tri-/tetra-nucleotide repeats accounted for 86.17–94.52%, the penta- or hexanucleotide repeats were very few. The AT content in the chloroplast genome of nine *Musa* plants were higher than the GC content, and SSRs shows a strong AT bias, which was a common phenomenon in the chloroplast genome of higher plants (Kuang et al., 2011; Lei et al., 2016). Repetitive sequences played a vital role in generating insertion mutations and substitution

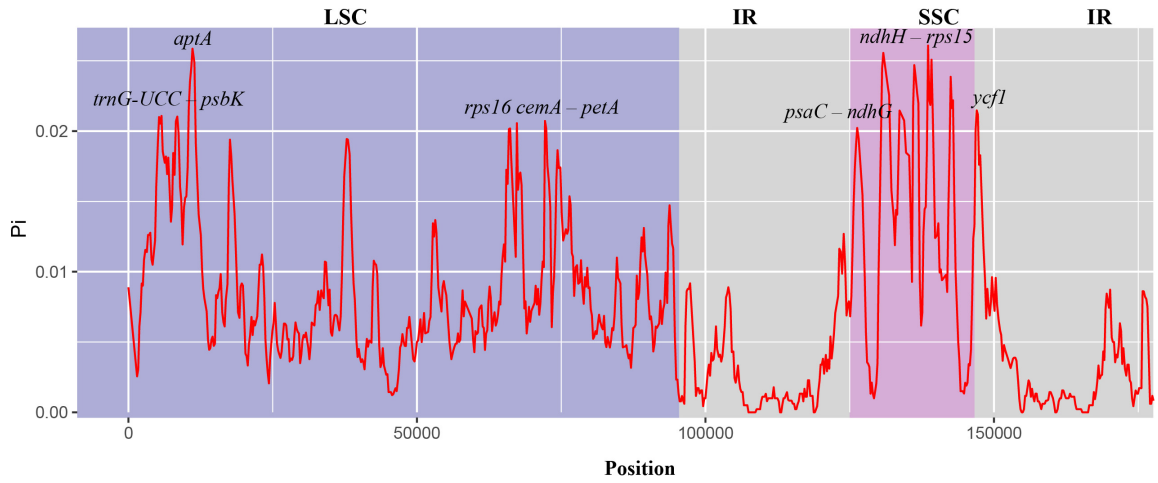

FIGURE 5 | Nucleotide diversity in chloroplast genomes of nine species of *Musa*. The abscissa represents the position, and the red line represents the average of the nucleotide variations of the nine species (Pi).

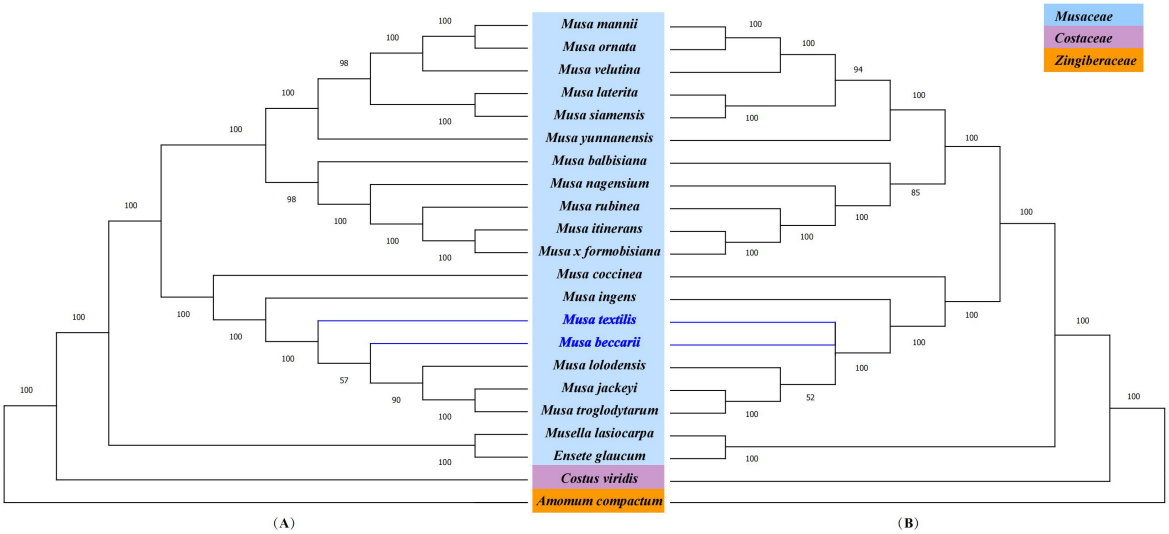

FIGURE 6 | Phylogenetic analysis. **(A)** Phylogenetic tree based on the complete chloroplast genome. **(B)** Phylogenetic tree based on shared protein-coding genes. *Costus viridis* and *Amomum compactum* were selected as out groups. Numbers at branch nodes are bootstrap values.

mutations (McDonald et al., 2011). Previous studies have shown widespread substitutions and deletions in the LSC and SSC regions of the chloroplast genome (Ahmed et al., 2012).

Comparison of the Sequences Within *Musa* Complex Species and Phylogenetic Relationships

The IR/LSC boundary position was not fixed during the evolution of angiosperms but can expand and contract moderately (Ahmed et al., 2012). The large inverted repeat sequence may be directly related to the structural conservation of the chloroplast genome (Palmer and Thompson, 1982). In some angiosperms, the expansion or contraction of IR is usually accompanied by the change of gene position. For example, the *ycf1* gene often is pseudogene because it crosses the boundary between LSC–IR and SSC–IR (Saina et al., 2018b; Yang Z. et al., 2018; Shahzadi et al., 2020). In our research, we divided these nine species into three categories based on the location of *rps19* gene. In contrast, the *M. yunnanensis* in the third category and the four species in the first category belong to *Musa* L. sect. *Musa*, the three species in the second category and the *M. ingens* in the third category all belong to *Musa* sect. *Callimusa*. In the chloroplast genome of *M. yunnanensis*, we speculated that two *rps19* genes appear in the LSC region due to the contraction of the IR region. In contrast, the *ndhB* gene remained in the IR region, thus evolving into a part of *Musa* sect. *Callimusa*. The shrinkage or expansion of the IR region was one of the essential features

for understanding the evolution and structure of the chloroplast genome (Jiang et al., 2017).

The SNP distributions of the nine species were very similar. The SNPs of each species account for the highest proportion in the LSC region. Except for *M. nagensium*, the SNPs of the other species in the IR region were more than the SSC region. We also analyzed the insertions and deletions of the chloroplast genomes of nine species, and the results found that they have similar rules as SNPs. In that case, it is possible to predict mutation hot spots and better study population genetics and analyze the phylogenetic relationship of species (Du et al., 2017; Keller et al., 2017).

K_a/K_s is used to assess nucleotides' natural selection pressure and evolution rate, which is a meaningful marker in species evolution (Li et al., 1985). In our study, K_s was much higher than K_a, which means that the evolution of *Musa* species was relatively slow. Only two genes (*ycf1* and *ycf2*) were under positive selection, and this was also somewhat different from the species of the Zingiberaceae (Liang and Chen, 2021). Consistent with many previous studies, the evolution of photosynthesis genes was slower than other types of protein-coding genes (Wicke et al., 2011; Saina et al., 2018a; Tian et al., 2018). Genes under positive selection often inserted many repetitive amino acid sequences to varying degrees, which may be evidence of adaptation to new ecological conditions or the result of co-evolution (Piot et al., 2018).

The chloroplast genome sequence contains highly variable regions. Finding more regions with a higher evolution rate is helpful to distinguish closely related species or genus, which is of great significance to the study of DNA barcodes (Dong et al., 2012). The chloroplast genomes contained two huge genes, *ycf1* and *ycf2*, which were indispensable chloroplast genes in higher plants (Drescher et al., 2000). The proteins that control transcription play an important role in cell survival. In the chloroplast genomes of most flowering plants, the *accD* gene encodes the β-carboxyl transferase subunit of acetyl-CoA carboxylase, which is essential for plant leaf development (Kode et al., 2005). Since they are all protein-coding genes, they may provide information about the evolution of *Musa* plants. Our comparative analysis identified several non-coding sites (*psbI–atpA, atpH–atpI, rpoB–petN, psbM–psbD, ndhf-rpl32, psaC-ndhE*, and *ndhG-ndhI*) and three genes (*ycf1, ycf2,* and *accD*). These mutation hotpots with high nucleotide diversity were particularly suitable for *Musa* genus' further molecular phylogeny and population genetics research.

In recent years, many studies have used protein-coding regions or chloroplast whole-genome sequence for phylogenetic analysis (Henriquez et al., 2014). The results of this study revealed the genetic relationship between *Musa* plants. It is generally believed that *Musa* genus includes *Musa* sect. *Rhodochlamys*, *Musa* sect. *Eumusa*, *Musa* sect. *Australimusa*, and *Musa* sect. *Ingentimusa* (Cheesman, 1947). Currently, the *Musa* genus is divided into two sections, *Musa* L. sect. *Musa* and *Musa* sect. *Callimusa* (Häkkinen, 2013). At the same time, the 19 unlinked nuclear genes confirmed the close relationship of *Australimusa* and *Callimusa* sections and showed that *Eumusa* and *Rhodochlamys* sections are not reciprocally monophyletic (Christelov et al., 2011). Our analysis revealed that *Musa* sect. *Rhodochlamys* and *Musa* sect. *Eumusa* were sisterly related to *Musa* sect. *Australimusa* and *Musa* sect. *Ingentimusa*. This result further verified that *Musa* L. sect. *Musa* included *Musa* sect. *Rhodochlamys* and *Musa* sect. *Eumusa*, and *Musa* sect. *Callimusa* comprised *Musa* sect. *Australimusa* and *Musa* sect. *Ingentimusa*. Based on the evolutionary tree, we also found that *M. lasiocarpa* is a basal species in the genus of *Musa* (Novák et al., 2014), which will help to deduce the time of origin of *Musa*. In addition, the results we obtained were different from previous studies (Liu et al., 2018; Feng et al., 2020). For example, the findings of their results concluded that *M. textilis* was the sister group of *M. balbisiana* and *M. beccarii* was closer to the roots of the evolutionary tree than *Musa itinerans*, which may be related to the other genomic regions and species collected. At present, the phylogenetic analysis of *Musa* species we have done was based on the complete chloroplast genome and protein-coding genes were the most comprehensive, which provided a theoretical foundation and technical support for the development and utilization of *Musa* plants resources.

CONCLUSION

In this study, we reported and compared the complete chloroplast genomes of nine *Musa* species in the first time, greatly increasing the available molecular sequences for this genus. The complete chloroplast genomes of these nine species were typical circular double-stranded quadripartite structure and ranged from 167,975 to 172,653 bp in the length. We analyzed the sequences of the chloroplast genomes of nine *Musa* species, such as the sequence length of each region, the number of different types of genes, and the types of intron genes. Codon bias analysis presented an extensively preferences for codons containing A/T at the 3′ end, especially for those who showed RSCU greater than one. We detected most of repetitive sequence existed in range of 30–50 bp. As shown in K_a/K_s evaluation, 77 of protein-coding genes was relatively stable during evolution process, while two genes (*ycf1* and *ycf2*) were under positive selection. Our research also revealed that all nine *Musa* species have *ndhA* genes that spanned the SSC and IRa regions, and notably, the *rps19* gene was entirely located in the IRb regions, suggesting that this phenomenon may occur with the contraction IRb area. SNP and InDels analysis divided nine *Musa* species into two groups in terms of the abundance and distribution of nucleotide polymorphic phenomenon, which was further confirmed by the phylogenetic tree. In summary, comparing the chloroplast genomes of *Musa* can deepen our understanding of the evolution of the Musaceae and may be suitable for the phylogenetic analysis and classification of *Mus*a genus.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available. The data that support the findings of this study are

openly available in the Genbank database at https://www.ncbi.nlm.nih.gov/ under accession number NC_056826 - NC_056834.

AUTHOR CONTRIBUTIONS

CJ and CS: conceptualization. HC, XW, and SW: data curation. SW: formal analysis. CS and SW: funding acquisition. CJ, HC, and XW: investigation. WS: methodology. WS, CS, and SW: project administration. WS and ZC: software, visualization, and writing – review and editing. CS: supervision. HC and XW: validation. WS and CJ: writing – original draft. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (No. 31801022) and Shandong Province Natural Science Foundation of China (No. ZR2019BC094). We are thankful to Beijing-based Novogene for their NGS service that was instrumental to the execution of the project.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.832884/full#supplementary-material

Supplementary Table 1 | Intron details.

Supplementary Table 2 | Codon usage details all nine *Musa* species.

Supplementary Table 3 | K_a/K_s raw data all nine *Musa* species.

Supplementary Table 4 | RNA editing sites raw data of all nine *Musa* species.

Supplementary Table 5 | Detail of SSR and long repeats.

Supplementary Table 6 | SNP and InDels statistics.

Supplementary Table 7 | Raw data of Pi values.

Supplementary Table 8 | Information of species in phylogenetic tree.

REFERENCES

Abdullah, Mehmood, F., Shahzadi, I., Waseem, S., Mirza, B., Ahmed, I., et al. (2020). Chloroplast genome of Hibiscus rosa-sinensis (Malvaceae): Comparative analyses and identification of mutational hotspots. *Genomics* 112, 581–591. doi: 10.1016/j.ygeno.2019.04.010

Ahmad, T., and Danish, M. (2018). Prospects of banana waste utilization in wastewater treatment: a review. *J. Environ. Manag.* 206, 330–348. doi: 10.1016/j.jenvman.2017.10.061

Ahmed, I., Biggs, P. J., Matthews, P. J., Collins, L. J., Hendy, M. D., and Lockhart, P. J. (2012). Mutational dynamics of aroid chloroplast genomes. *Genome Biol. Evol.* 4, 1316–1323. doi: 10.1093/gbe/evs110

Ahmed, I., Matthews, P. J., Biggs, P. J., Naeem, M., Mclenachan, P. A., and Lockhart, P. J. (2013). Identification of chloroplast genome loci suitable for high-resolution phylogeographic studies of *Colocasia esculenta* (L.) Schott (Araceae) and closely related taxa. *Mol. Ecol. Res.* 13, 929–937. doi: 10.1111/1755-0998.12128

Amiryousefi, A., Hyvönen, J., and Poczai, P. (2018a). IRscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics* 34, 3030–3031. doi: 10.1093/bioinformatics/bty220

Amiryousefi, A., Hyvönen, J., and Poczai, P. (2018b). The chloroplast genome sequence of bittersweet (*Solanum dulcamara*): Plastid genome structure evolution in Solanaceae. *PLoS One* 13:69. doi: 10.1371/journal.pone.0196069

Arias, D., Rodríguez, J., López, B., and Méndez, P. (2021). Evaluation of the physicochemical properties of pectin extracted from *Musa paradisiaca* banana peels at different pH conditions in the formation of nanoparticles. *Heliyon* 7:59. doi: 10.1016/j.heliyon.2021.e06059

Asaf, S., Khan, A. L., Khan, M. A., Imran, Q. M., Kang, S. M., Al-Hosni, K., et al. (2017a). Comparative analysis of complete plastid genomes from wild soybean (*Glycine soja*) and nine other Glycine species. *PLoS One* 12:182281. doi: 10.1371/journal.pone.0182281

Asaf, S., Waqas, M., Khan, A. L., Khan, M. A., Kang, S. M., Imran, Q. M., et al. (2017b). The complete chloroplast genome of wild rice (*Oryza minuta*) and its comparison to related species. *Front. Plant Sci.* 8:304. doi: 10.3389/fpls.2017.00304

Beier, S., Thiel, T., Münch, T., Scholz, U., and Mascher, M. (2017). MISA-web: A web server for microsatellite prediction. *Bioinformatics* 33, 2583–2585. doi: 10.1093/bioinformatics/btx198

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

Brudno, M., Malde, S., Poliakov, A., Do, C. B., Couronne, O., Dubchak, I., et al. (2003). Glocal alignment: Finding rearrangements during alignment. *Bioinformatics* 19:5. doi: 10.1093/bioinformatics/btg1005

Cheesman, E. E. (1947). Classification of the Bananas: The Genus *Musa* L. *Kew Bullet.* 2:106. doi: 10.2307/4109207

Chen, J., Hao, Z., Xu, H., Yang, L., Liu, G., Sheng, Y., et al. (2015). The complete chloroplast genome sequence of the relict woody plant metasequoia glyptostroboides. *Front. Plant Sci.* 6:1–11. doi: 10.3389/fpls.2015.00447

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* 2017:e3919. doi: 10.7717/peerj.3919

Christelov, P., Valrik, M., Hibov, E., De Langhe, E., and Doleel, J. (2011). A multi gene sequence-based phylogeny of the Musaceae (banana) family. *BMC Evol. Biol.* 11:103. doi: 10.1186/1471-2148-11-103

Chumley, T. W., Palmer, J. D., Mower, J. P., Fourcade, H. M., Calie, P. J., Boore, J. L., et al. (2006). The complete chloroplast genome sequence of *Pelargonium* × *hortorum*: Organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. *Mol. Biol. Evol.* 23, 2175–2190. doi: 10.1093/molbev/msl089

del Río, J. C., and Gutiérrez, A. (2006). Chemical composition of abaca (*Musa textilis*) leaf fibers used for manufacturing of high quality paper pulps. *J. Agricult. Food Chem.* 54, 4600–4610. doi: 10.1021/JF053016N

Dierckxsens, N., Mardulyn, P., and Smits, G. (2017). NOVOPlasty: De novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* 45:e18. doi: 10.1093/nar/gkw955

Dong, W., Liu, J., Yu, J., Wang, L., and Zhou, S. (2012). Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. *PloS One* 7:1–9. doi: 10.1371/journal.pone.0035071

Drescher, A., Stephanie, R., Calsa, T., Carrer, H., and Bock, R. (2000). The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. *Plant J.* 22, 97–104. doi: 10.1046/j.1365-313X.2000.00722.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. doi: 10.1038/s41598-017-06210-2

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32, 3047–3048. doi: 10.1093/bioinformatics/btw354

Feng, H., Chen, Y., Xu, X., Luo, H., Wu, Y., and He, C. (2020). The complete chloroplast genome of *Musa beccarii*. *Mitochon. DNA Part B: Res.* 5, 2384–2385. doi: 10.1080/23802359.2020.1775513

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* 2017:2846. doi: 10.7717/peerj.2846

Häkkinen, M. (2013). Reappraisal of sectional taxonomy in *Musa* (Musaceae). *Taxon* 62, 809–813. doi: 10.12705/624.3

Häkkinen, M., and Väre, H. (2008). Typification and check-list of *Musa* L. names (Musaceae) with nomenclatural notes. *Adansonia* 30, 63–112. doi: 10.5281/zenodo.5190398

Henriquez, C. L., Arias, T., Pires, J. C., Croat, T. B., and Schaal, B. A. (2014). Phylogenomics of the plant family Araceae. *Mol. Phylogenet. Evol.* 75, 91–102. doi: 10.1016/j.ympev.2014.02.017

Ingale, S., Joshi, S. J., and Gupte, A. (2014). Production of bioethanol using agricultural waste: Banana pseudo stem. *Braz. J. Microbiol.* 45, 885–892. doi: 10.1590/s1517-83822014000300018

Jansen, R. K., Cai, Z., Raubeson, L. A., Daniell, H., Depamphilis, C. W., Leebens-Mack, J., et al. (2007). Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proc. Nat. Acad. Sci. USA* 104, 19369–19374. doi: 10.1073/pnas.0709121104

Jarvis, P., and López-Juez, E. (2013). Biogenesis and homeostasis of chloroplasts and other plastids. *Nat. Rev. Mol. Cell Biol.* 14, 787–802. doi: 10.1038/nrm3702

Jiang, D., Zhao, Z., Zhang, T., Zhong, W., Liu, C., Yuan, Q. J., et al. (2017). The chloroplast genome sequence of *Scutellaria baicalensis* provides insight into intraspecific and interspecific chloroplast genome diversity in Scutellaria. *Genes* 8, 1–13. doi: 10.3390/genes8090227

Keller, J., Rousseau-Gueutin, M., Martin, G. E., Morice, J., Boutte, J., Coissac, E., et al. (2017). The evolutionary fate of the chloroplast and nuclear *rps16* genes as revealed through the sequencing and comparative analyses of four novel legume chloroplast genomes from Lupinus. *DNA Res.* 24, 343–358. doi: 10.1093/dnares/dsx006

Kim, Y. K., Park, C. W., and Kim, K. J. (2009). Complete chloroplast DNA sequence from a Korean endemic genus, *Megaleranthis saniculifolia*, and its evolutionary implications. *Mol. Cells* 27, 365–381. doi: 10.1007/s10059-009-0047-6

Kode, V., Mudd, E. A., Iamtham, S., and Day, A. (2005). The tobacco plastid *accD* gene is essential and is required for leaf development. *Plant J.* 44, 237–244. doi: 10.1111/j.1365-313X.2005.02533.x

Kolodner, R., and Tewari, K. K. (1979). Inverted repeats in chloroplast DNA from higher plants. *Proc. Nat. Acad. Sci.* 76, 41–45. doi: 10.1073/PNAS.76.1.41

Kress, W. J., Prince, L. M., Hahn, W. J., and Zimmer, E. A. (2001). Unraveling the evolutionary radiation of the families of the Zingiberales using morphological and molecular evidence. *Syst. Biol.* 50, 926–944. doi: 10.1080/106351501753462885

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

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

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

Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H. H., Rognes, T., and Ussery, D. W. (2007). RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* 35, 3100–3108. doi: 10.1093/nar/gkm160

Lee, H. O., Joh, H. J., Kim, K., Lee, S. C., Kim, N. H., Park, J. Y., et al. (2019). Dynamic chloroplast genome rearrangement and DNA barcoding for three Apiaceae species known as the medicinal herb "bang-poong.". *Int. J. Mol. Sci.* 20:2196. doi: 10.3390/ijms20092196

Lehwark, P., and Greiner, S. (2019). GB2sequin - A file converter preparing custom GenBank files for database submission. *Genomics* 111, 759–761. doi: 10.1016/j.ygeno.2018.05.003

Lei, W., Ni, D., Wang, Y., Shao, J., Wang, X., Yang, D., et al. (2016). Intraspecific and heteroplasmic variations, gene losses and inversions in the chloroplast genome of Astragalus membranaceus. *Sci. Rep.* 6, 1–13. doi: 10.1038/srep21669

Li, L. F., Häkkinen, M., Yuan, Y. M., Hao, G., and Ge, X. J. (2010). Molecular phylogeny and systematics of the banana family (Musaceae) inferred from multiple nuclear and chloroplast DNA fragments, with a special reference to the genus. *Musa*. *Mol. Phylogenet. Evol.* 57, 1–10. doi: 10.1016/j.ympev.2010.06.021

Li, P., Zhang, S., Li, F., Zhang, S., Zhang, H., Wang, X., et al. (2017). A phylogenetic analysis of chloroplast genomes elucidates the relationships of the six economically important *Brassica* species comprising the triangle of U. *Front. Plant Sci.* 8:1–13. doi: 10.3389/fpls.2017.00111

Li, W. H., Wu, C. I., and Luo, C. C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* 2, 150–174. doi: 10.1093/OXFORDJOURNALS.MOLBEV.A040343

Liang, H., and Chen, J. (2021). Comparison and phylogenetic analyses of nine complete chloroplast genomes of Zingibereae. *Forests* 12:710. doi: 10.3390/f12060710

Liaud, M. F., Zhang, D. X., and Cerff, R. (1990). Differential intron loss and endosymbiotic transfer of chloroplast glyceraldehyde-3-phosphate dehydrogenase genes to the nucleus. *Proc. Nat. Acad. Sci. USA* 87, 8918–8922. doi: 10.1073/PNAS.87.22.8918

Liu, J., Gao, C. W., and Niu, Y. F. (2018). The complete chloroplast genome sequence of flowering banana, *Musa ornata*. *Mitochon. DNA Part B: Res.* 3, 962–963. doi: 10.1080/23802359.2018.1507647

Liu, Q., and Xue, Q. (2005). Comparative studies on codon usage pattern of chloroplasts and their host nuclear genes in four plant species. *J. Genet.* 84, 55–62. doi: 10.1007/BF02715890

Lohse, M., Drechsel, O., and Bock, R. (2007). OrganellarGenomeDRAW (OGDRAW): A tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Curr. Genet.* 52, 267–274. doi: 10.1007/s00294-007-0161-y

Lowe, T. M., and Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA Genes in Genomic Sequence. *Nucleic Acids Res.* 25, 955–964. doi: 10.1093/nar/25.5.955

Luo, R., Lam, T.-W., Liu, B., Xie, Y., Li, Z., Huang, W., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* 1, 1–6. doi: 10.1186/2047-217X-1-18

Maréchal, A., and Brisson, N. (2010). Recombination and the maintenance of plant organelle genome stability. *New Phytol.* 186, 299–317. doi: 10.1111/j.1469-8137.2010.03195.x

Martin, G., Baurens, F. C., Cardi, C., D'Hont, A., and Aury, J. M. (2013). The complete chloroplast genome of banana (*musa acuminata*, zingiberales): insight into plastid monocotyledon evolution. *PLoS One* 8:67350. doi: 10.1371/journal.pone.0067350

McDonald, M. J., Wang, W. C., Huang, H., and Leu, J. Y. (2011). Clusters of Nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences. *PLoS Biol.* 9:e1000622. doi: 10.1371/journal.pbio.1000622

Millen, R. S., Olmstead, R. G., Adams, K. L., Palmer, J. D., Lao, N. T., Heggie, L., et al. (2001). Many parallel losses of *infa* from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. *Plant Cell* 13, 645–658. doi: 10.1105/tpc.13.3.645

Mower, J. P. (2009). The PREP suite: Predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. *Nucleic Acids Res.* 37, 12–14. doi: 10.1093/nar/gkp337

Muraguri, S., Xu, W., Chapman, M., Muchugi, A., Oluwaniyi, A., Oyebanji, O., et al. (2020). Intraspecific variation within Castor bean (*Ricinus communis* L.) based on chloroplast genomes. *Indus. Crops Prod.* 155:112779. doi: 10.1016/J.INDCROP.2020.112779

Naranjo, J. M., Cardona, C. A., and Higuita, J. C. (2014). Use of residual banana for polyhydroxybutyrate (PHB) production: Case of study in an integrated biorefinery. *Waste Manag.* 34, 2634–2640. doi: 10.1016/j.wasman.2014.09.007

Nielsen, A. Z., Mellor, S. B., Vavitsas, K., Wlodarczyk, A. J., Gnanasekaran, T., Perestrello Ramos, H., et al. (2016). Extending the biosynthetic repertoires of cyanobacteria and chloroplasts. *Plant J.* 87, 87–102. doi: 10.1111/tpj.13173

Novák, P., Høibová, E., Neumann, P., Koblížková, A., Doležel, J., and Macas, J. (2014). Genome-wide analysis of repeat diversity across the family Musaceae. *PLoS One* 9:98918. doi: 10.1371/journal.pone.0098918

Oberoi, H. S., Vadlani, P. V., Saida, L., Bansal, S., and Hughes, J. D. (2011). Ethanol production from banana peels using statistically optimized simultaneous saccharification and fermentation process. *Waste Manag.* 31, 1576–1584. doi: 10.1016/j.wasman.2011.02.007

Oyewo, O. A., Onyango, M. S., and Wolkersdorfer, C. (2016). Application of banana peels nanosorbent for the removal of radioactive minerals from real mine water. *J. Environ. Radioact.* 164, 369–376. doi: 10.1016/j.jenvrad.2016.08.014

Pakshirajan, K., Worku, A. N., Acheampong, M. A., Lubberding, H. J., and Lens, P. N. L. (2013). Cr (III) and Cr(VI) removal from aqueous solutions by cheaply available fruit waste and algal biomass. *Appl. Biochem. Biotechnol.* 3, 498–513. doi: 10.1007/S12010-013-0202-6

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

Pappu, A., Patil, V., Jain, S., Mahindrakar, A., Haque, R., and Thakur, V. K. (2015). Advances in industrial prospective of cellulosic macromolecules enriched banana biofibre resources: A review. *Int. J. Biol. Macromol.* 79, 449–458. doi: 10.1016/j.ijbiomac.2015.05.013

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

Porebski, S., Bailey, L. G., and Baum, B. R. (1997). Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Mol. Biol. Rep.* 1, 8–15. doi: 10.1007/BF02772108

Posada, D., and Crandall, K. A. (1998). MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14, 817–818. doi: 10.1093/BIOINFORMATICS/14.9.817

Provan, J., Powell, W., and Hollingsworth, P. M. (2001). Chloroplast microsatellites: New tools for studies in plant ecology and evolution. *Trends Ecol. Evol.* 16:2097. doi: 10.1016/S0169-5347(00)02097-8

Ramírez-Hernández, A., Aparicio-Saguilán, A., Reynoso-Meza, G., and Carrillo-Ahumada, J. (2017). Multi-objective optimization of process conditions in the manufacturing of banana (*Musa paradisiaca* L.) starch/natural rubber films. *Carbohyd. Polym.* 157, 1125–1133. doi: 10.1016/j.carbpol.2016.10.083

Saina, J. K., Gichira, A. W., Li, Z. Z., Hu, G. W., Wang, Q. F., and Liao, K. (2018a). The complete chloroplast genome sequence of *Dodonaea viscosa*: comparative and phylogenetic analyses. *Genetica* 146, 101–113. doi: 10.1007/s10709-017-0003-x

Saina, J. K., Li, Z. Z., Gichira, A. W., and Liao, Y. Y. (2018b). The complete chloroplast genome sequence of tree of heaven (*Ailanthus altissima* (mill.) (Sapindales: Simaroubaceae), an important pantropical tree. *Int. J. Mol. Sci.* 19:929. doi: 10.3390/ijms19040929

Shahzadi, I., Abdullah, Mehmood, F., Ali, Z., Ahmed, I., and Mirza, B. (2020). Chloroplast genome sequences of *Artemisia maritima* and *Artemisia absinthium*: Comparative analyses, mutational hotspots in genus Artemisia and phylogeny in family Asteraceae. *Genomics* 112, 1454–1463. doi: 10.1016/j.ygeno.2019.08.016

Shar, Z. H., Fletcher, M. T., Sumbal, G. A., Sherazi, S. T. H., Giles, C., Bhanger, M. I., et al. (2016). Banana peel: an effective biosorbent for aflatoxins. *Food Addit. Contam. Part A Chem. Anal. Control Expo. Risk Assess.* 33, 849–860. doi: 10.1080/19440049.2016.1175155

Sheng, J., Yan, M., Wang, J., Zhao, L., Zhou, F., Hu, Z., et al. (2021). The complete chloroplast genome sequences of five *Miscanthus* species, and comparative analyses with other grass plastomes. *Indus. Crops Prod.* 162:113248. doi: 10.1016/j.indcrop.2021.113248

Shetty, S. M., MdShah, M. U., Makale, K., Mohd-Yusuf, Y., Khalid, N., and Othman, R. Y. (2016). Complete chloroplast genome sequence of *musa balbisiana* corroborates structural heterogeneity of inverted repeats in wild progenitors of cultivated bananas and plantains. *Plant Genome* 9:89. doi: 10.3835/PLANTGENOME2015.09.0089

Taweechat, C., Wongsooka, T., and Rawdkuen, S. (2021). Properties of banana (*Cavendish* spp.) starch film incorporated with banana peel extract and its application. *Molecules* 26:51406. doi: 10.3390/molecules26051406

Terrab, A., Paun, O., Talavera, S., Tremetsberger, K., Arista, M., and Stuessy, T. F. (2006). Genetic diversity and population structure in natural populations of Moroccan Atlas cedar (*Cedrus atlantica*; Pinaceae) determined with cpSSR markers. *Am. J. Bot.* 93, 1274–1280. doi: 10.3732/ajb.93.9.1274

Thanyapanich, N., Jimtaisong, A., and Rawdkuen, S. (2021). Functional properties of Banana starch (*Musa* spp.) and its utilization in cosmetics. *Molecules* 26:26123637. doi: 10.3390/molecules26123637

Tian, N., Han, L., Chen, C., and Wang, Z. (2018). The complete chloroplast genome sequence of *Epipremnum aureum* and its comparative analysis among eight Araceae species. *PLoS One* 13:e0192956. doi: 10.1371/journal.pone.0192956

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

Vishnuvarthanan, M., Dharunya, R., Jayashree, S., Karpagam, B., and Sowndharya, R. (2019). Environment-friendly packaging material: banana fiber/cowdung composite paperboard. *Environ. Chem. Lett.* 17, 1143–1429. doi: 10.1007/s10311-019-00879-9

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. Proteom. Bioinform.* 8, 77–80. doi: 10.1016/S1672-0229(10)60008-3

Weiner, I., Shahar, N., Marco, P., Yacoby, I., and Tuller, T. (2019). Solving the riddle of the evolution of shine-dalgarno based translation in chloroplasts. *Mol. Biol. Evol.* 36, 2854–2860. doi: 10.1093/molbev/msz210

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

Xu, Q., Xiong, G., Li, P., He, F., Huang, Y., Wang, K., et al. (2012). Analysis of complete nucleotide sequences of 12 *Gossypium* chloroplast genomes: Origin and evolution of Allotetraploids. *PLoS One* 7:e37128. doi: 10.1371/journal.pone.0037128

Yang, J. B., Li, D. Z., and Li, H. T. (2014). Highly effective sequencing whole chloroplast genomes of angiosperms by nine novel universal primer pairs. *Mol. Ecol. Res.* 14, 1024–1031. doi: 10.1111/1755-0998.12251

Yang, Y., Zhu, J., Feng, L., Zhou, T., Bai, G., Yang, J., et al. (2018). Plastid genome comparative and phylogenetic analyses of the key genera in fagaceae: Highlighting the effect of codon composition bias in phylogenetic inference. *Front. Plant Sci.* 9:1–13. doi: 10.3389/fpls.2018.00082

Yang, Z., Zhao, T., Ma, Q., Liang, L., and Wang, G. (2018). Comparative genomics and phylogenetic analysis revealed the chloroplast genome variation and interspecific relationships of *Corylus* (Betulaceae) species. *Front. Plant Sci.* 9:927. doi: 10.3389/fpls.2018.00927

Zhu, B., Qian, F., Hou, Y., Yang, W., Cai, M., and Wu, X. (2021). Complete chloroplast genome features and phylogenetic analysis of *Eruca sativa* (Brassicaceae). *PLoS One* 16:248556. doi: 10.1371/journal.pone.0248556

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

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

frontiers in Plant Science

ORIGINAL RESEARCH
published: 16 February 2022
doi: 10.3389/fpls.2022.829668

Genome-Wide Characterization of Serine/Arginine-Rich Gene Family and Its Genetic Effects on Agronomic Traits of *Brassica napus*

Meili Xie, Rong Zuo, Zetao Bai, Lingli Yang, Chuanji Zhao, Feng Gao, Xiaohui Cheng, Junyan Huang, Yueying Liu, Yang Li, Chaobo Tong and Shengyi Liu**

The Key Laboratory of Biology and Genetic Improvement of Oil Crops, The Ministry of Agriculture and Rural Affairs of the PRC, Oil Crops Research Institute of the Chinese Academy of Agricultural Sciences, Wuhan, China

OPEN ACCESS

Edited by:
Hai Du,
Southwest University, China

Reviewed by:
Kun Lu,
Southwest University, China
Xiaoming Song,
North China University of Science and Technology, China
Haifeng Li,
Northwest A&F University, China

***Correspondence:**
Chaobo Tong
tongchaobo@126.com
Shengyi Liu
liusy@oilcrops.cn

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

Received: *06 December 2021*
Accepted: *10 January 2022*
Published: *16 February 2022*

Citation:
Xie M, Zuo R, Bai Z, Yang L, Zhao C, Gao F, Cheng X, Huang J, Liu Y, Li Y, Tong C and Liu S (2022) Genome-Wide Characterization of Serine/Arginine-Rich Gene Family and Its Genetic Effects on Agronomic Traits of Brassica napus. Front. Plant Sci. 13:829668. doi: 10.3389/fpls.2022.829668

Serine/arginine-rich (SR) proteins are indispensable factors for RNA splicing, and they play important roles in development and abiotic stress responses. However, little information on *SR* genes in *Brassica napus* is available. In this study, 59 *SR* genes were identified and classified into seven subfamilies: SR, SCL, RS2Z, RSZ, RS, SR45, and SC. In each subfamily, the genes showed relatively conserved structures and motifs, but displayed distinct expression patterns in different tissues and under abiotic stress, which might be caused by the varied *cis*-acting regulatory elements among them. Transcriptome datasets from Pacbio/Illumina platforms showed that alternative splicing of *SR* genes was widespread in *B. napus* and the majority of paralogous gene pairs displayed different splicing patterns. Protein-protein interaction analysis indicated that SR proteins were involved in the regulation of the whole lifecycle of mRNA, from synthesis to decay. Moreover, the association mapping analysis suggested that 12 *SR* genes were candidate genes for regulating specific agronomic traits, which indicated that *SR* genes could affect the development and hence influence the important agronomic traits of *B. napus*. In summary, this study provided elaborate information on *SR* genes in *B. napus*, which will aid further functional studies and genetic improvement of agronomic traits in *B. napus*.

Keywords: serine/arginine-rich gene family, *Brassica napus*, expression pattern, alternative splicing, association mapping analysis, agronomic traits

INTRODUCTION

RNA splicing is an important process in eukaryotes that could produce one or multiple mature mRNAs via different splicing sites, which significantly increases the flexibility of gene expression regulation and the diversity of transcriptome and proteome (Black, 2003). The process is mediated by the spliceosome, a large macromolecule complex composed of five small nuclear ribonucleoproteins (snRNPs) and a mass of proteins (Will and Luhrmann, 2010). Among these proteins, the serine/arginine-rich (SR) proteins are vital splicing factors to regulate the selections of splicing

sites by binding splicing enhancers on the pre-mRNA (Zahler et al., 1992). The structure of SR proteins is conserved, containing one or two RNA binding domains (RBDs) at the N-terminus and an arginine/serine-rich (RS) domain at the C-terminus (Shepard and Hertel, 2009). The RBDs are responsible to recognize and bind to specific RNA regions, while the RS domain contributes to the protein-protein interactions. The subcellular localization of SR proteins is directly related to their molecular functions, and it has been reported that they are localized in the nuclear speckles (Caceres et al., 1997), a subset of them could shuttle between the nucleus and cytoplasm (Sapra et al., 2009).

In plants, SR proteins were initially identified in Arabidopsis (Kalyna and Barta, 2004), then in rice, maize, wheat, tomato, cassava, and so on (Isshiki et al., 2006; Richardson et al., 2011; Yoon et al., 2018; Chen et al., 2019, 2020b; Gu et al., 2020; Rosenkranz et al., 2021). According to sequence similarity, SR proteins could be divided into seven subfamilies: SR, SCL, RS2Z, RSZ, RS, SR45, SC, and three of them (SCL, RS2Z, RS) are plant-specific (Richardson et al., 2011). Subfamily SCL is the largest plant-specific one containing members from dicots, monocots, moss, and green algae. Subfamily RS2Z was mainly composed of dicots and monocots, whereas most members of subfamily RS came from photosynthetic eukaryotes. Many studies have shown that the *SR* genes play important roles in plant developmental processes and respond to hormonal signaling or environmental stress (Isshiki et al., 2006; Palusa et al., 2007; Reddy and Shad Ali, 2011; Melo et al., 2020). For example, the life cycle of *Atsr45-1* was significantly shorter, the leaves of *Atsr45-1* were elongated and curly, and the number of petals and stamens was also significantly different from the wild type (Ali et al., 2007). Overexpression of *RSZ33* in Arabidopsis can result in developmental abnormalities in embryos and root apical meristem (Kalyna et al., 2003). And knockout SC and SCL in Arabidopsis could affect the transcriptions of many genes, resulting in serrated leaves, late flowering, shorter roots and abnormal silique phyllotaxy (Yan et al., 2017). Most members of the plant-specific SCL are involved in stress responses mediated by exogenous abscisic acid (ABA) (Cruz et al., 2014). In terms of environmental stress, *Atsr34B* reduces plant tolerance to calcium by regulating the expression of *IRT1* (Zhang et al., 2014), while *AtRS40* and *AtRS41* act as critical modulators under salt stress (Chen et al., 2013).

In addition to regulating the splicing of other genes, *SR* genes also could be alternatively spliced. A total of 19 *SR* genes were identified in Arabidopsis (Kalyna and Barta, 2004). Among them, 15 genes could produce 95 transcripts under hormone induction or abiotic stress, which greatly increased the complexity of the *SR* genes by sixfold (Palusa et al., 2007). There were 21 and 18 *SR* genes in maize and sorghum, respectively, whereas 92 and 62 transcripts were detected in each of them, and the majority of SR transcripts were not conserved between maize and sorghum (Rauch et al., 2014). *SR* genes in tomato showed different splicing profiles in various organs as well as in response to heat stress (Rosenkranz et al., 2021). And a variety of AS events occurred in *SR* genes of *Brassica rapa* under abiotic stresses (Yoon et al., 2018). Recently, an increasing number of studies focused on the detailed functional and regulatory mechanisms of the varied SR transcripts. Numerous SR transcripts contained premature termination codons (PTCs) which might elicit nonsense-mediated mRNA decay (NMD) to regulate the gene expression (McGlincy and Smith, 2008; Palusa and Reddy, 2010). And other SR transcripts showed distinct biological functions, like salt-responsive gene *SR45a* could generate two transcripts SR45a-1a and SR45a-1b, the first of which directly interacted with the cap-binding protein 20 (CBP20), whereas the latter promoted the association of SR45a-1a with CBP20, through the fine-tune regulatory mechanism, it was conducive for the plants to response to salt stress (Li et al., 2021).

Brassica napus is an important global oil crop (Chalhoub et al., 2014), which is an allotetraploid species derived from hybridization between *B. rapa* and *Brassica oleracea*. To date, it is unclear how many *SR* genes/transcripts are present in *B. napus* and how they perform their function to affect the oil crop. Now the genome sequences and various transcriptome datasets of *B. napus* are available (Chalhoub et al., 2014; Zhang et al., 2019; Yao et al., 2020), which provide an ample resource to investigate the specific genes at the genome-wide level. In this study, *SR* genes were identified in *B. napus*, the phylogenetic relationship, gene structures, conserved motifs, gene duplications and protein interactions were also analyzed. The transcriptome data from various tissues and environmental stresses were used for the expression patterns and alternative splicing analysis of *SR* genes in *B. napus*. Moreover, genetic variations of *SR* genes in a worldwide core collection germplasm (Tang, 2019) were also investigated, and the association mapping analysis revealed that 12 *SR* genes were candidate genes for agronomic traits in *B. napus*. This study expanded our understanding of *SR* genes in *B. napus* and provided a foundation for further functional studies.

MATERIALS AND METHODS

Identification of *SR* Genes in *Brassica napus*

The genome and annotation information of the *B. napus* cultivar "*Darmor*-bzh" were obtained from the Brassicaceae Database (BRAD)[1] (Chalhoub et al., 2014). The amino acid sequences of the SR family in *Arabidopsis thaliana* (Kalyna and Barta, 2004) were obtained to build a Hidden Markov Model (HMM), and HMMER3.0 (Mistry et al., 2013) was used to search for *SR* genes in *B. napus* (E value was set to 1e-5). The NCBI Conserved Domain Database[2] (Lu et al., 2020)and the SMART databases[3] (Letunic et al., 2020)were used for verification of candidate genes, preserving the ones containing RRM and RS domains. Moreover, ProtParam,[4] an online software of SWISS-PROT, was used to predict the molecular weights (MW) and isoelectric point (pI) of SR proteins, and CELLO v2.5 (Yu et al., 2006) was used to predict the subcellular location of these proteins.

[1]http://brassicadb.cn/

[2]https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

[3]http://smart.embl.de/

[4]https://web.expasy.org/protparam/

Chromosomal Location and Gene Duplication Analysis

The locations of *SR* genes were obtained from the annotation of *B. napus* genome. To identify gene duplication events, BLASTP with the e-value of 1e–10 was used to align the sequence, and MCScanX (Wang et al., 2012) was used to detect the duplication patterns including segmental and tandem duplication. Chromosomal locations and duplication events were visualized using the Circos software (Krzywinski et al., 2009). The ratios of non-synonymous to synonymous substitutions (Ka/Ks) of duplicate gene pairs were counted by ParaAT2.0 (Zhang et al., 2012), which aligned the protein sequence by Muscle (Edgar, 2004) and calculated the Ka/Ks ratio by KaKs_Calculator (Wang et al., 2010).

Gene Structure, Conserved Motifs, and *cis*-Acting Regulatory Elements Analysis

TBtools (Chen et al., 2020a) and Multiple Expectation Maximization for Motif Elicitation (MEME) (Bailey et al., 2015) were used to display the gene structures and to analyze the conserved motifs in SR proteins. To identify the *cis*-acting regulatory elements of *SR* genes, promoters (2 kb upstream sequences from initiation codon) were extracted and predicted by PlantCARE[5] (Magali, 2002). The location was displayed by Gene Structure Display Server (GSDS 2.0) (Hu et al., 2015), the amount heatmap was visualized by R.[6]

Phylogenetic Analysis of SR Family Members

To gain insights into the evolutionary relationships of SR family members, multiple sequence alignments of SR amino acids of *A. thaliana* and *B. napus* were performed using the ClustalW (Larkin, 2007). Phylogenetic trees were generated with the MEGA 11 program using the Neighbor-Joining (NJ) method with 1,000 bootstrap replications (Tamura et al., 2021). The tree was visualized using Evolview[7] (He et al., 2016).

Prediction of Protein-Protein Interactions

The Protein-Protein Interactions of *A. thaliana* were downloaded from STRING[8] (Szklarczyk et al., 2021), the interaction networks of SR proteins in *B. napus* were predicted based on the homologs in *A. thaliana*, and Cytoscape (Shannon et al., 2003) was used to display the interaction. To investigate the involved biological process, genes that interacted with SR proteins were taken out for Gene Ontology and KEGG enrichment analysis by clusterProfiler in R (Yu et al., 2012).

Expression Analysis of *SR* Genes in *Brassica napus*

Transcriptome data from five tissues (leaf, callus, bud, root, and young silique) and different stress conditions (dehydration, salt, cold and ABA) of *B. napus* cultivar "ZS11" were used in this study (Zhang et al., 2019; Yao et al., 2020), the expression levels of *SR* genes were calculated with Stringtie (Pertea et al., 2015) after alignment with Hisat2 (Kim et al., 2015), and displayed by Pheatmap and UpSet in R. And expression patterns of four genes were showed by TBtools-eFP (Chen et al., 2020a).[9]

Alternative Splicing Analysis of *SR* Genes in *Brassica napus*

Based on the two sets of transcriptome data, alternative splicing of *SR* genes were also investigated. For the transcript isoform catalog of *B. napus* obtained from Iso-seq (Yao et al., 2020), the AS events were identified by Astalavista (Sylvain and Michael, 2007) and the expression of alternative splicing transcripts of *SR* genes in various tissues were calculated with Stringtie. For the RNA-seq of different stress conditions (Zhang et al., 2019), transcripts were assembled by Stringtie firstly, then the AS events and the expression of alternative splicing transcripts were counted. In order to verify the AS events between paralogous gene pairs, transcriptome data based on EST sequencing of *B. napus* were downloaded and analyzed (Troncoso-Ponce et al., 2011).

RNA Isolation and qRT-PCR Analysis of *SR* Genes

The seeds of *B. napus* cultivar "ZS11" were germinated and grown in a growth room at 24°C, with a 16/8 h light/dark photoperiod. The leaves and roots were collected from 20-day-old seedlings, while buds were collected from 70-day-old seedings, siliques were harvested 90 days after germination. Samples were immediately stored in liquid nitrogen, and total RNA was extracted from samples using Invitrogen trizol reagent (TRIzol^{TM}15596026, United States) according to the manufacturer's instructions. Total RNA was then reverse-transcribed into complementary DNAs by using the PrimeScript RT reagent Kit With gDNA Eraser (Takara, Japan). The complementary DNAs were used as templates in quantitative reverse-transcription polymerase chain reaction (qRT-PCR) with the gene-specific primers (**Supplementary Table 1**). qRT-PCR was performed by using SYBR Green Real-time PCR Master Mix (Bio-Rad, United States) in 20 μl reaction mixture and run on CFX96 Real-time PCR system (Bio-Rad, United States). *B. napus* β-actin gene was used as internal standard. All assays were conducted with three biological repeats, and each with three technical repeats. The relative expression level was obtained using the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001).

Association Mapping of *SR* Genes in a Natural Population of *Brassica napus*

To understand the natural variations of *SR* genes in *B. napus*, a natural population with 324 worldwide accessions was used in this study (Tang, 2019). SNPs in the gene regions of *SR* genes were extracted and annotated by SnpEff (Cingolani et al., 2012). The agronomic traits including primary flowering

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

[6]https://cran.r-project.org/

[7]https://www.evolgenius.info/evolview/

[8]https://www.string-db.org/

[9]http://yanglab.hzau.edu.cn/BnTIR/eFP

time (PFT), full flowering time (FFT1), final flowering time (FFT2), early flowering stage (EFS), late-flowering stage (LFS), flowering period (FP), plant height (PH), branch number (BN), branch height (BH), main inflorescence length (MIL), main inflorescence silique number (MISN), main inflorescence silique density (MISD) were selected (Tang, 2019). With the mixed linear model, a family-based association mapping analysis considering population structure and relative kinship was performed by EMMAX (Kang et al., 2010). The linkage disequilibrium and haplotype blocks were made by LDBlockShow (Dong et al., 2020) and the enriched Gene Ontology terms of interacted proteins were drawn by Cytoscape (Shannon et al., 2003).

RESULTS

SR Genes Form Seven Subfamilies in *Brassica napus*

After performing HMM search and domain verification, a total of 59 *SR* genes were identified in *B. napus*. The detailed information of each SR was listed in **Table 1**, including gene ID, genomic location, amino acids (AA) length, isoelectric point (pI), and molecular weight (MW) and so on. The lengths of SR proteins ranged from 130 to 412 AA, with an average length of 293 AA. The pI values were varied from 7.31 to 12.41 and the MW values were varied from 14.92 to 47.02 kDa. According to the prediction of CELLO, it showed that all the SR proteins were located in nuclear.

To understand the evolutionary relationships of *SR* genes between *B. napus* and *A. thaliana*, a phylogenetic tree was constructed based on their protein sequences. Finally, 19 *AtSRs* and 59 *BnSRs* were clustered into seven subfamilies (**Figure 1** and **Table 1**). According to the previous nomenclature system, subfamily SR, SCL, RS2Z, RSZ, RS, SR45, and SC were also used in this study. Subfamily SCL and SR were the largest, each of which included 12 *SR* genes, while subfamily SC was the smallest, with only 3 *SR* genes, and the other subfamily RS2Z, RS, RSZ, and SR45 contained 10, 9, 7, and 6 *SR* genes, respectively.

Chromosomal Distribution and Gene Duplication of *SR* Genes in *Brassica napus*

In *B. napus*, 46 of 59 *SR* genes were unevenly distributed over 19 chromosomes, while the other 13 *SR* genes were assigned to unanchored scaffolds (**Figure 2**). In total, 26 and 33 *SR* genes were located on the A subgenome and C subgenome, respectively. There were 5 SR, 5 SCL, 5 RS2Z, 4 RSZ, 3 SR45, 3 RS, and 1 SC subfamily genes on A subgenome, with compared to 7 SR, 7 SCL, 5 RS2Z, 3 RSZ, 3 SR45, 6 RS, and 2 SC on C subgenome. Chromosomes C03, C05, and C08 had the most *SR* genes (5 genes per chromosome), while chromosomes A01, A02, and C02 contained only one *SR* gene, respectively, and no *SR* gene was located on chromosomes A07, A10, and C09.

According to BLAST and MCScanX, gene duplication events of the *SR* genes were detected in *B. napus*. In short, all 59 *SR* genes were derived from duplication (**Table 1**), 89.83% of them (53 *SR* genes) were originated from whole-genome duplication (WGD) or segmental duplications, while the other 6 *SR* genes resulted from dispersed duplications. Moreover, there were 91 paralogous gene pairs in *B. napus* (**Figure 2** and **Supplementary Table 2**), 15 of them occurred in the A subgenome, 21 of them took place in the C subgenome, and the other 55 duplication events occurred between the A and C subgenome. To estimate the selection mode of *SR* genes in *B. napus*, the ratios of non-synonymous to synonymous substitutions (Ka/Ks) for paralogous gene pairs were calculated. Generally, Ka/Ks $>$ 1 means positive selection, Ka/Ks = 1 means neutral selection, and Ka/Ks $<$ 1 represents purifying selection. In this work, Ka/Ks ratios of all the paralogous gene pairs were less than 1, suggesting that *SR* genes were under purifying selection (**Supplementary Table 2**).

Gene Structure, Conserved Motifs, and *cis*-Acting Regulatory Elements Analysis of *SR* Genes in *Brassica napus*

The exon-intron structure of 59 *SR* genes in seven subfamilies was displayed (**Figures 3A,B**), On average, each gene included 7 exons, but the exon numbers differed widely, ranging from 3 to 13, and different subfamilies exhibited different exon numbers. While the genes in the same subfamily tended to possess similar gene structures, for example, in subfamily SC, all the *SR* genes had 9 exons, and in subfamily SR45, all the *SR* genes contained 12 exons except *BnaA08g23570D*.

In total, 9 conserved motifs were identified in 59 *SR* genes (**Figure 3C**). All the *SR* genes contained motif 1 and motif 2 except *BnaC08g31720D*, which lacked motif 1. All the *SR* genes possessed motif 9, except those in subfamily SC. Apparently, the motif structures of distinct subfamilies varied. For example, the pattern of subfamily SC was motif 2-1-2-4-8, while subfamily RS2Z was motif 2-1-3-4-6-9. And some subfamilies had a few specific motifs, like motif 7 was unique to subfamily SR, motif 8 only existed in subfamily RS2Z.

Promoter regions were found to be critical for gene expression (Oudelaar and Higgs, 2021), so *cis*-acting regulatory elements in these regions were investigated for *SR* genes. *Cis*-acting regulatory elements related to stress, hormone and development (ranging from 5 to 23) were detected in promoters of *SR* genes (**Figure 4**, **Supplementary Figure 1**, and **Supplementary Table 3**). The majority of *SR* genes (56/59, 94.92%) contained ARE elements, which is essential for anaerobic induction. Moreover, stress-responsive elements such as TC-rich repeats (involved in defense and stress responsiveness, 33/59, 55.93%), LTR (involved in low-temperature responsiveness, 33/59, 55.93%) and MBS (involved in drought-inducibility, 29/59, 49.15%) were also common in promoters of *SR* genes. Hormone-responsive elements like ABRE (involved in the abscisic acid responsiveness), CGTCA-motif (involved in the MeJA-responsiveness) and ERE (involved in the ethylene

TABLE 1 | Characteristics of the *SR* genes in *B. napus* (pI, isoelectric point; MW, molecular weight).

Gene ID	Subfamily	Chromosome	Start	End	Amino acids	pI	MW(kDa)	Exon number	Duplication type	Subcellular Location
BnaA01g14750D	RS	A01	7452346	7455318	338	10.29	39.08	5	WGD or segmental	Nuclear
BnaA03g12870D	RS	A03	5857719	5860659	402	9.97	47.02	6	WGD or segmental	Nuclear
BnaA08g30960D	RS	A08_random	1785553	1788161	348	10.15	40.26	5	WGD or segmental	Nuclear
BnaC01g41640D	RS	C01_random	812746	815639	339	10.29	39.32	5	Dispersed	Nuclear
BnaC03g15710D	RS	C03	7919987	7922439	348	9.87	41.1	5	WGD or segmental	Nuclear
BnaC04g00810D	RS	C04	681412	683420	246	9.87	29.23	5	WGD or segmental	Nuclear
BnaC07g39690D	RS	C07	40437893	40439448	308	9.9	35.26	4	WGD or segmental	Nuclear
BnaC08g31720D	RS	C08	30974005	30975333	276	9.62	31.75	7	WGD or segmental	Nuclear
BnaC08g47240D	RS	C08_random	2078021	2080623	348	10.12	40.27	5	WGD or segmental	Nuclear
BnaA03g00590D	RS2Z	A03	271562	273654	263	10.18	29.47	5	WGD or segmental	Nuclear
BnaA03g17170D	RS2Z	A03	8046335	8048874	316	10.01	35.85	7	WGD or segmental	Nuclear
BnaA05g28890D	RS2Z	A05	20374400	20376670	295	10.03	33.88	6	WGD or segmental	Nuclear
BnaA07g37700D	RS2Z	A07_random	1239429	1241695	291	10.13	32.62	6	WGD or segmental	Nuclear
BnaA09g33780D	RS2Z	A09	24828942	24830905	283	10.13	31.92	6	WGD or segmental	Nuclear
BnaC03g00890D	RS2Z	C03	425179	427495	265	10.29	29.9	5	WGD or segmental	Nuclear
BnaC03g20680D	RS2Z	C03	10976632	10979194	293	10.1	33.37	6	WGD or segmental	Nuclear
BnaC05g43360D	RS2Z	C05	40270977	40273246	287	9.88	33	6	WGD or segmental	Nuclear
BnaC06g14780D	RS2Z	C06	17526686	17529186	288	10.13	32.39	6	WGD or segmental	Nuclear
BnaC08g24530D	RS2Z	C08	26559105	26561152	284	10.07	31.91	6	WGD or segmental	Nuclear
BnaA03g51620D	RSZ	A03	26830836	26832115	199	11.25	22.85	5	WGD or segmental	Nuclear
BnaA04g14520D	RSZ	A04	12206509	12208772	196	11.06	22.03	5	WGD or segmental	Nuclear
BnaA09g54590D	RSZ	A09_random	2700552	2702123	130	9.86	14.92	3	WGD or segmental	Nuclear
BnaAnng28560D	RSZ	Ann_random	32673950	32675820	194	11.22	21.96	6	Dispersed	Nuclear
BnaC04g36280D	RSZ	C04	37798706	37800989	196	11.06	22.03	5	WGD or segmental	Nuclear
BnaC05g19020D	RSZ	C05	12576277	12578087	185	11.28	21.32	4	WGD or segmental	Nuclear
BnaC07g43350D	RSZ	C07	42484671	42485905	197	11.36	22.76	5	WGD or segmental	Nuclear
BnaA09g52820D	SC	A09_random	666454	668617	381	7.38	41.67	9	Dispersed	Nuclear
BnaCnng35170D	SC	Cnn_random	33346654	33348807	366	7.31	40.1	9	Dispersed	Nuclear
BnaCnng52140D	SC	Cnn_random	51593258	51595460	381	7.38	41.66	9	Dispersed	Nuclear
BnaA04g03560D	SCL	A04	2430038	2432468	282	11.37	31.86	7	WGD or segmental	Nuclear
BnaA05g13830D	SCL	A05	8419061	8421417	318	11.77	36.72	5	WGD or segmental	Nuclear
BnaA05g27090D	SCL	A05	19602609	19604747	205	10.75	24.04	5	WGD or segmental	Nuclear
BnaA06g00410D	SCL	A06	214562	217292	284	11.78	32.55	5	WGD or segmental	Nuclear

(Continued)

TABLE 1 | (Continued)

Gene ID	Subfamily	Chromosome	Start	End	Amino acids	pI	MW(kDa)	Exon number	Duplication type	Subcellular Location
BnaA08g00560D	SCL	A08	347635	349821	232	11.29	26.56	8	WGD or segmental	Nuclear
BnaC01g37560D	SCL	C01	36838369	36841052	261	11.51	30.38	4	WGD or segmental	Nuclear
BnaC03g70430D	SCL	C03	60111052	60114048	337	11.91	39.05	7	WGD or segmental	Nuclear
BnaC04g25450D	SCL	C04	26579193	26581824	278	11.41	31.53	7	WGD or segmental	Nuclear
BnaC05g41220D	SCL	C05	39032165	39034551	238	10.78	28.34	7	WGD or segmental	Nuclear
BnaC06g07190D	SCL	C06	7792680	7794731	287	11.77	32.59	5	WGD or segmental	Nuclear
BnaC06g10860D	SCL	C06	12913198	12916033	297	11.67	34.02	7	WGD or segmental	Nuclear
BnaCnng00990D	SCL	Cnn_random	1223600	1225715	263	11.54	30.55	4	WGD or segmental	Nuclear
BnaA02g20550D	SR	A02	12929940	12932521	305	11.18	34.06	11	WGD or segmental	Nuclear
BnaA06g15930D	SR	A06	8756159	8759128	289	10.21	32.17	11	WGD or segmental	Nuclear
BnaA06g21030D	SR	A06	14628976	14631673	275	10.55	31.27	12	WGD or segmental	Nuclear
BnaA06g37780D	SR	A06_random	174729	177826	253	9.96	28.94	11	WGD or segmental	Nuclear
BnaA09g00790D	SR	A09	496380	499254	309	11.11	34.38	11	WGD or segmental	Nuclear
BnaC01g26160D	SR	C01	22702333	22705203	282	10.35	31.78	10	WGD or segmental	Nuclear
BnaC02g27300D	SR	C02	25201876	25204492	299	11.01	33.73	11	WGD or segmental	Nuclear
BnaC03g52440D	SR	C03	37261995	37265320	269	10.47	30.55	12	WGD or segmental	Nuclear
BnaC05g06630D	SR	C05	3275956	3279260	317	9.68	36.59	13	WGD or segmental	Nuclear
BnaC07g38480D	SR	C07	39812212	39814540	200	7.67	22.91	9	Dispersed	Nuclear
BnaC08g21130D	SR	C08	23678348	23681381	295	10.32	33.04	11	WGD or segmental	Nuclear
BnaCnng19170D	SR	Cnn_random	17870537	17873473	308	11.05	34.33	11	WGD or segmental	Nuclear
BnaA06g11140D	SR45	A06	5847134	5850099	396	12.41	43.55	12	WGD or segmental	Nuclear
BnaA08g23570D	SR45	A08	16731975	16735088	366	12.34	40.69	11	WGD or segmental	Nuclear
BnaA09g56240D	SR45	A09_random	3650300	3653282	411	12.41	45.25	12	WGD or segmental	Nuclear
BnaC05g12680D	SR45	C05	7414534	7417852	399	12.38	44.09	12	WGD or segmental	Nuclear
BnaC08g16960D	SR45	C08	20742488	20745684	387	12.36	42.48	12	WGD or segmental	Nuclear
BnaC08g38300D	SR45	C08	34660215	34663377	412	12.41	45.29	12	WGD or segmental	Nuclear

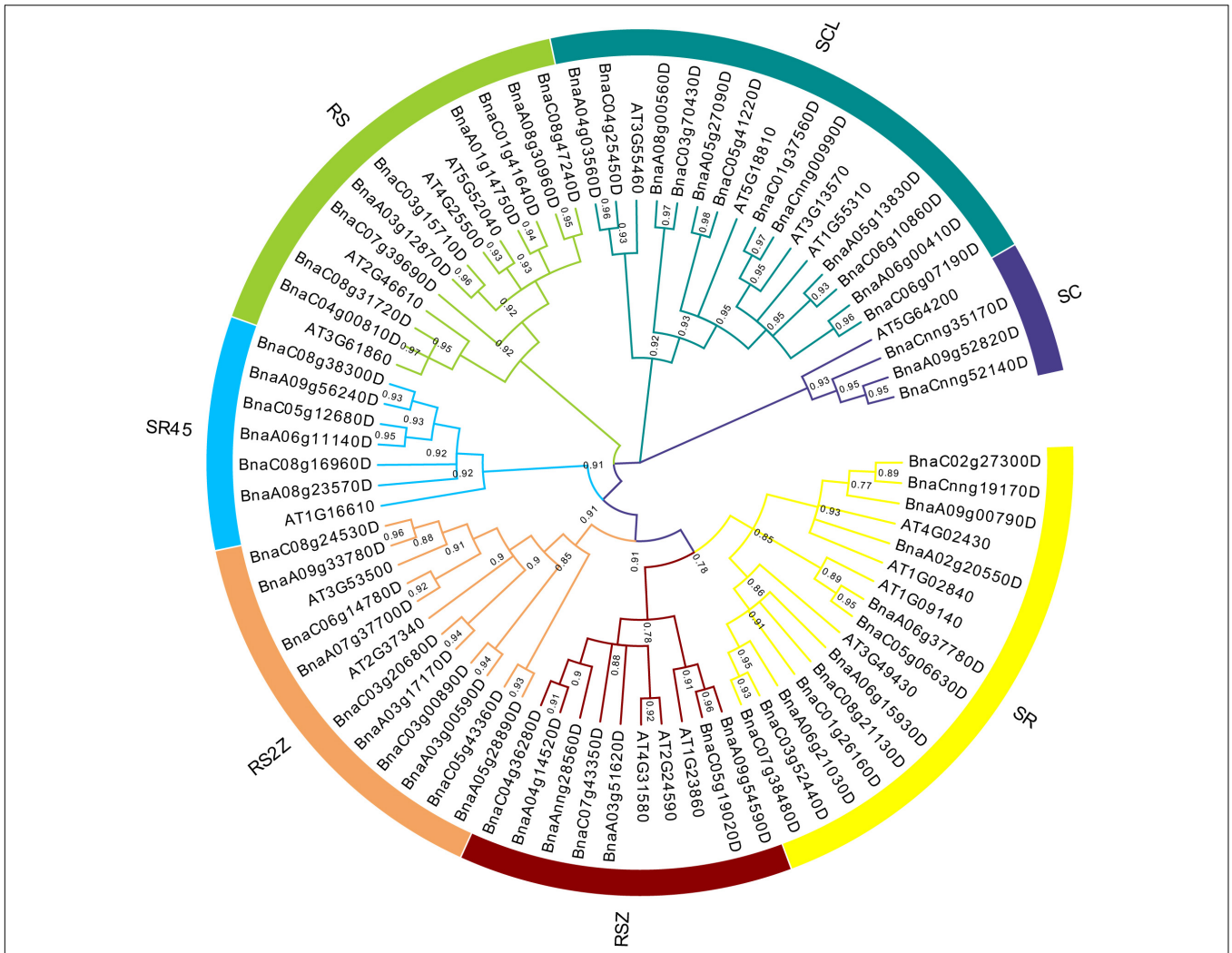

FIGURE 1 | Phylogenetic analysis of SR proteins in *B. napus* and *A. thaliana*. All SR proteins were clustered into seven subfamilies, and each subfamily was represented by a different color.

responsiveness) existed in most promoters of *SR* genes. In terms of development-related elements, CAT-box (24/59, 40.68%), which is related to meristem expression, was most frequently observed in the promoters of *SR* genes. The results indicated that many *SR* genes in *B. napus* were responsible for plant growth and stress response.

Predicted Protein Interactions of *SR* Proteins in *Brassica napus*

SR proteins were the key components of the spliceosome and they always interacted with other proteins to perform their functions (Shepard and Hertel, 2009; Will and Luhrmann, 2010). To understand the biological processes involved by SR proteins in *B. napus*, interaction networks were predicted according to known protein interactions in Arabidopsis. The homologous proteins of 59 BnSR proteins interacted with 3,528 proteins in Arabidopsis, which were homologous to 13,591 proteins in *B. napus* (**Figure 5A**). It demonstrated that SR proteins were the core nodes in the network, most SR proteins interacted with each other, meanwhile, they also interacted with other proteins to participate in different biological processes. KEGG enrichment analysis showed these interacted proteins were involved in a variety of processes including RNA degradation, ribosome biogenesis, RNA polymerase, proteasome, circadian rhythm, and so on (**Figure 5B** and **Supplementary Table 4**). Gene Ontology enrichment analysis (**Figure 5C** and **Supplementary Table 5**) showed that ribosome biogenesis, mRNA splicing and protein import into the nucleus were the significantly enriched terms in the biological process category. While the terms including cytosolic small ribosomal subunit and ribosome in the cellular component category were highly enriched, and in the molecular function category, translation initiation factor activity, proton symporter activity and RNA binding were significantly enriched. Protein-protein interactions analysis showed that SR proteins played important

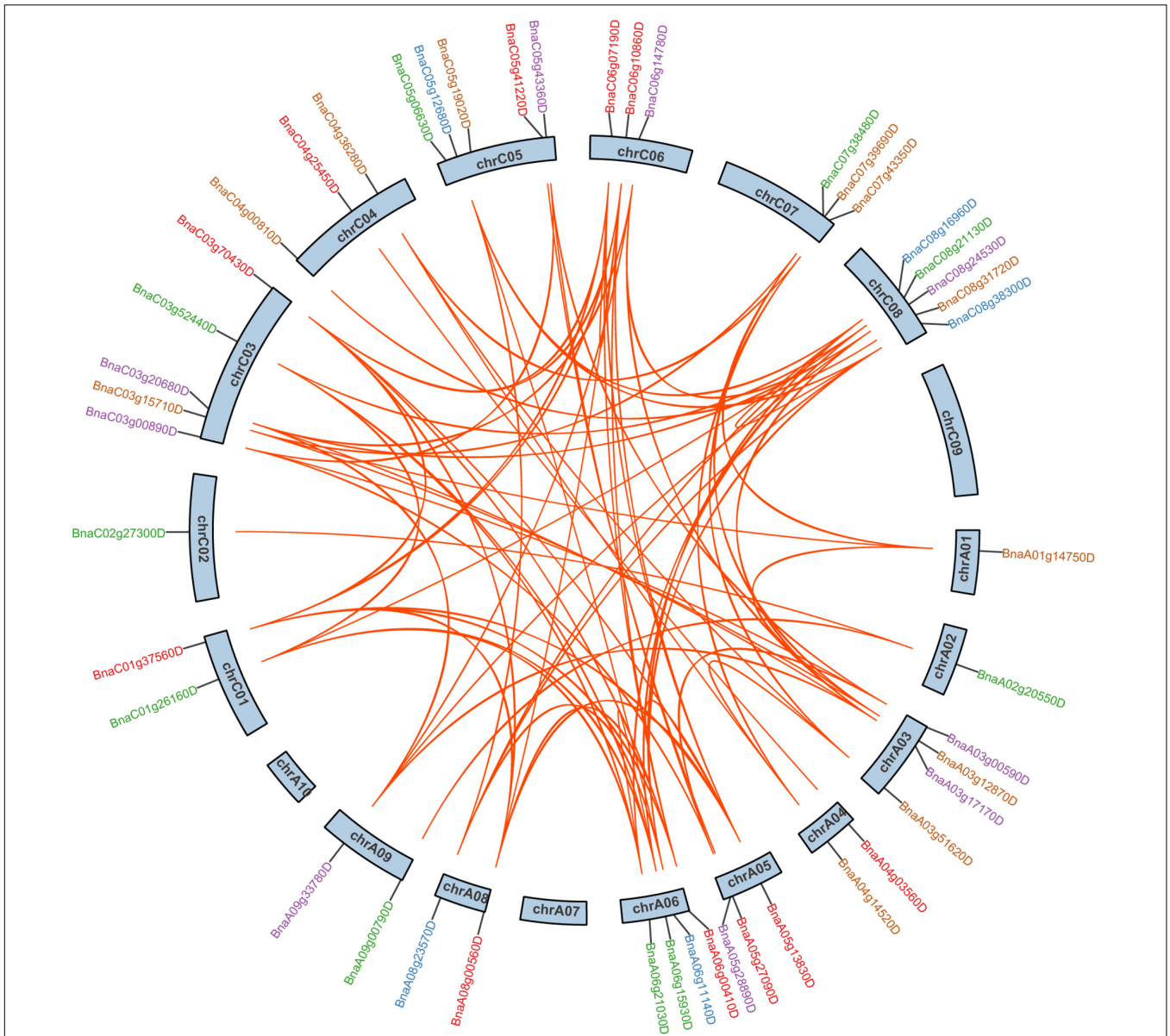

FIGURE 2 | The chromosomal distribution and duplication analysis of *SR* genes in *B. napus*. The locations of all the chromosomal *SR* genes were shown in different chromosomes. The different colors indicated different subfamilies of the *SR* genes. The orange lines highlighted the duplicated *SR* gene pairs.

roles in the regulation of the whole lifecycle of mRNA, from synthesis to decay.

Various Expression Patterns of *SR* Genes in Different Tissues and Under Abiotic Stresses in *Brassica napus*

To predict the potential functions of *SR* genes, expression patterns based on RNA-Seq of five tissues in *B. napus* cultivar "ZS11" (Yao et al., 2020) were displayed in **Figure 6**. *SR* genes showed different expression patterns among the five tissues. The expression profiles of *SR* genes in the silique and root displayed similar patterns. Almost all the *SR* genes were expressed highly in bud, root, silique and callus, but lowly in leaf (**Figure 6A**). There was 34 *SR* genes expressed in all of the five tissues based on the threshold value (FPKM > 5), and some of the *SR* genes were tissue-specific or preferential expression (**Figure 6B**). Like *BnaA01g14750D* showed the highest expression in callus (**Figure 6C**), and *BnaC06g14780D* expressed at a high level in silique and bud (**Figure 6D**), nevertheless, both of them expressed lowly in leaf. Meanwhile, a few *SR* genes expressed highly in callus and lowly in silique. And two *SR* genes (*BnaC08g31720D* and *BnaC07g39690D*) barely expressed in these five tissues.

SR genes in subfamily RS2Z, SR45, and SC showed similar expression patterns, paralogous gene pairs in these subfamilies also owned similar expression

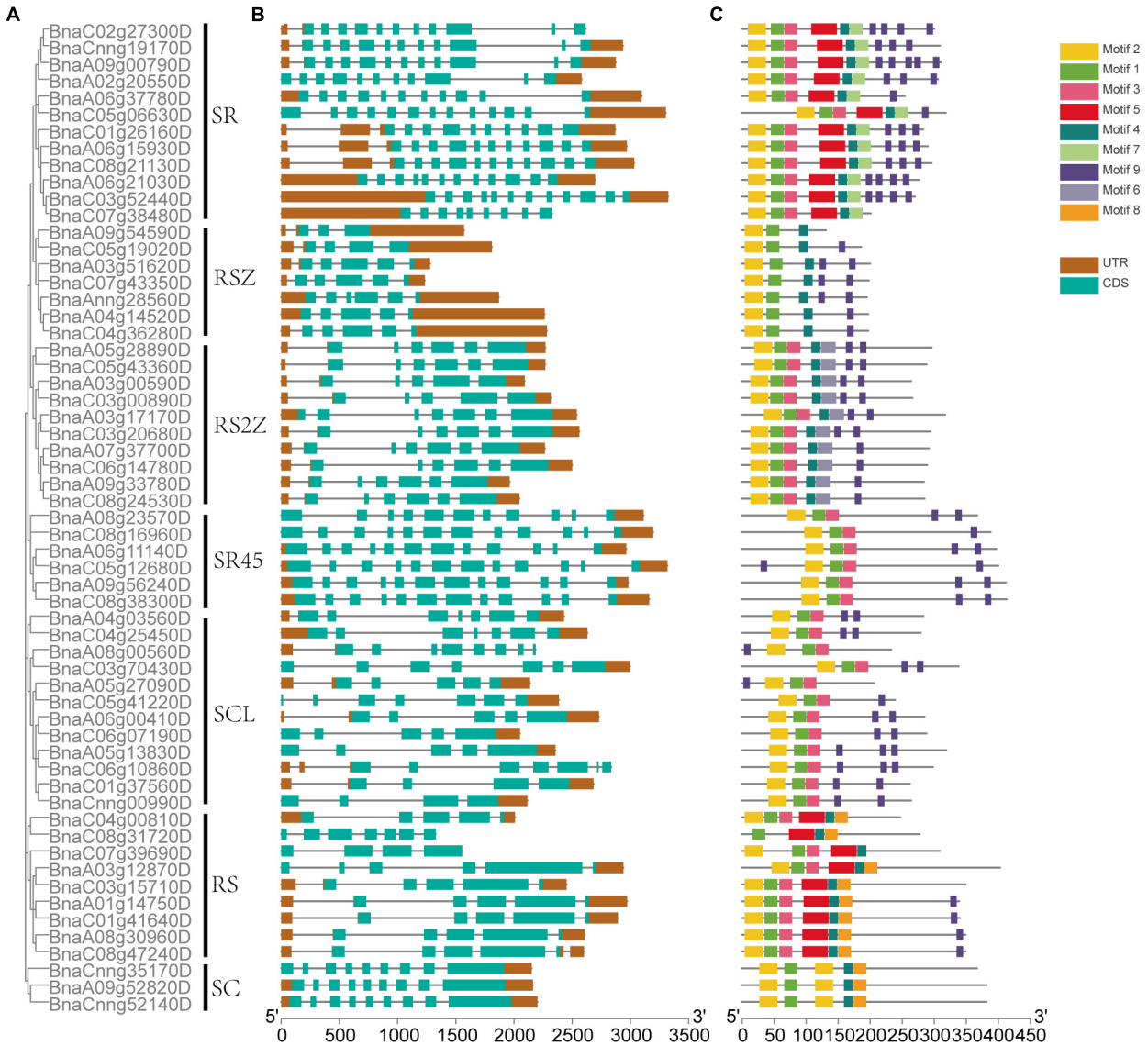

FIGURE 3 | The phylogenetic relationship, gene structure, and conserved motifs of *SR* genes in *B. napus*. **(A)** The phylogenetic relationships of 59 SR proteins based on the NJ method. **(B)** Gene structures of the *SR* genes. Dark brown boxes represented UTR, and indigo boxes represented CDS. **(C)** The motif composition of SR proteins. Numbers 1–9 were displayed in different colored boxes.

patterns, like *BnaA09g33780D*/*BnaC06g14780D* in RS2Z, *BnaA06g11140D*/*BnaC05g12680D* in SR45. Nevertheless, in other subfamilies, different patterns were observed, for example, paralogous gene pairs (*BnaA04g03560D*/*BnaC04g25450D*) in subfamily SCL expressed at the same pattern, while in subfamily RS *BnaC08g31720D* barely expressed in five tissues, its paralogous gene *BnaC04g00810D* expressed at a high level in callus, bud, root and silique, and in subfamily RSZ, *BnaC04g36280D* and its paralogous gene *BnaA04g14520D* expressed at a high level in each tissue (**Figure 6E**), but their paralogous gene *BnaA03g51620D* and *BnaC07g43350D* weakly expressed (**Figure 6F**). Moreover, 14 *SR* genes from different subfamilies were selected for qRT-PCR analysis (**Figure 7** and **Supplementary Table 1**), similarly, most of these genes expressed higher in bud, and the expression patterns of two genes (*BnaA09g52820D* and *BnaCnng52140D*) from subfamily SC were almost the same, while in subfamily SCL, *BnaCnng00990D* showed different expression patterns with *BnaA05g27090D* and *BnaC05g41220D*.

In spite of expression patterns in various tissues were investigated, the expression profiles of *SR* genes under different abiotic stresses were also analyzed. In this study, RNA-Seq data of samples from different abiotic treatments including cold, drought, salinity, ABA induction (Zhang et al., 2019) were utilized to analyze the expression pattern of *SR* genes in *B. napus* (**Figure 8**). Obviously, all the *SR* genes expressed higher

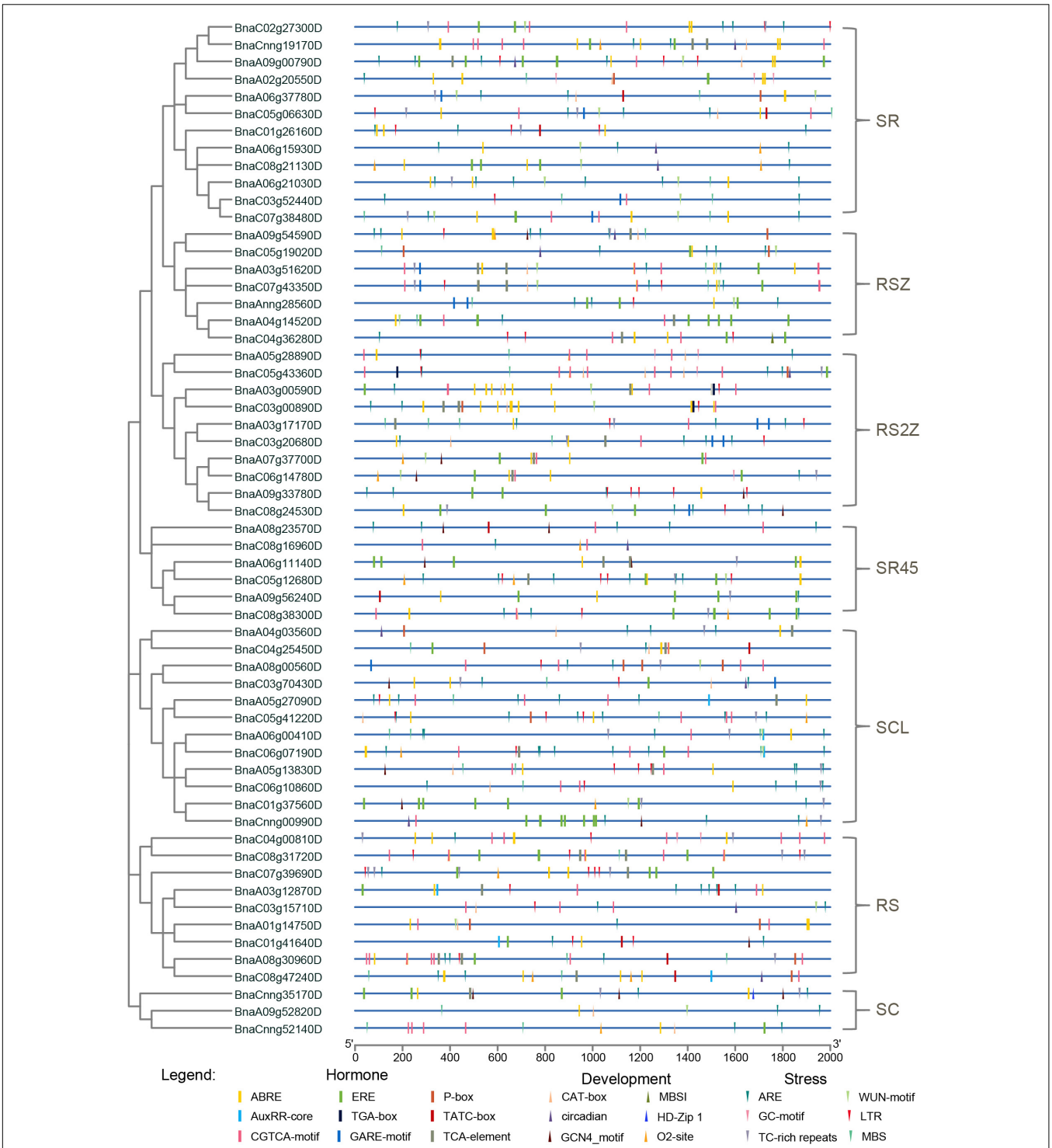

FIGURE 4 | *Cis*-acting regulatory elements identified in promoters of *SR* genes in *B. napus*. Boxes indicated hormone-related elements, up-triangles indicated development-related elements, and down-triangles indicated stress-related elements. Different colors indicated different elements.

after the treatment of abiotic stresses except those unexpressed or low-expressed genes. The expression of *BnaC07g39690D* was apparently up-regulated under dehydration stress. The expression of *BnaC05g06630D* dramatically increased under ABA induction as well as cold and salt stress, and it was noticed that elements response to these stresses (ABRE, LTR, and TC-rich repeats) were enriched in its promoter. All the *SR* genes expressed at a higher level in both subfamily

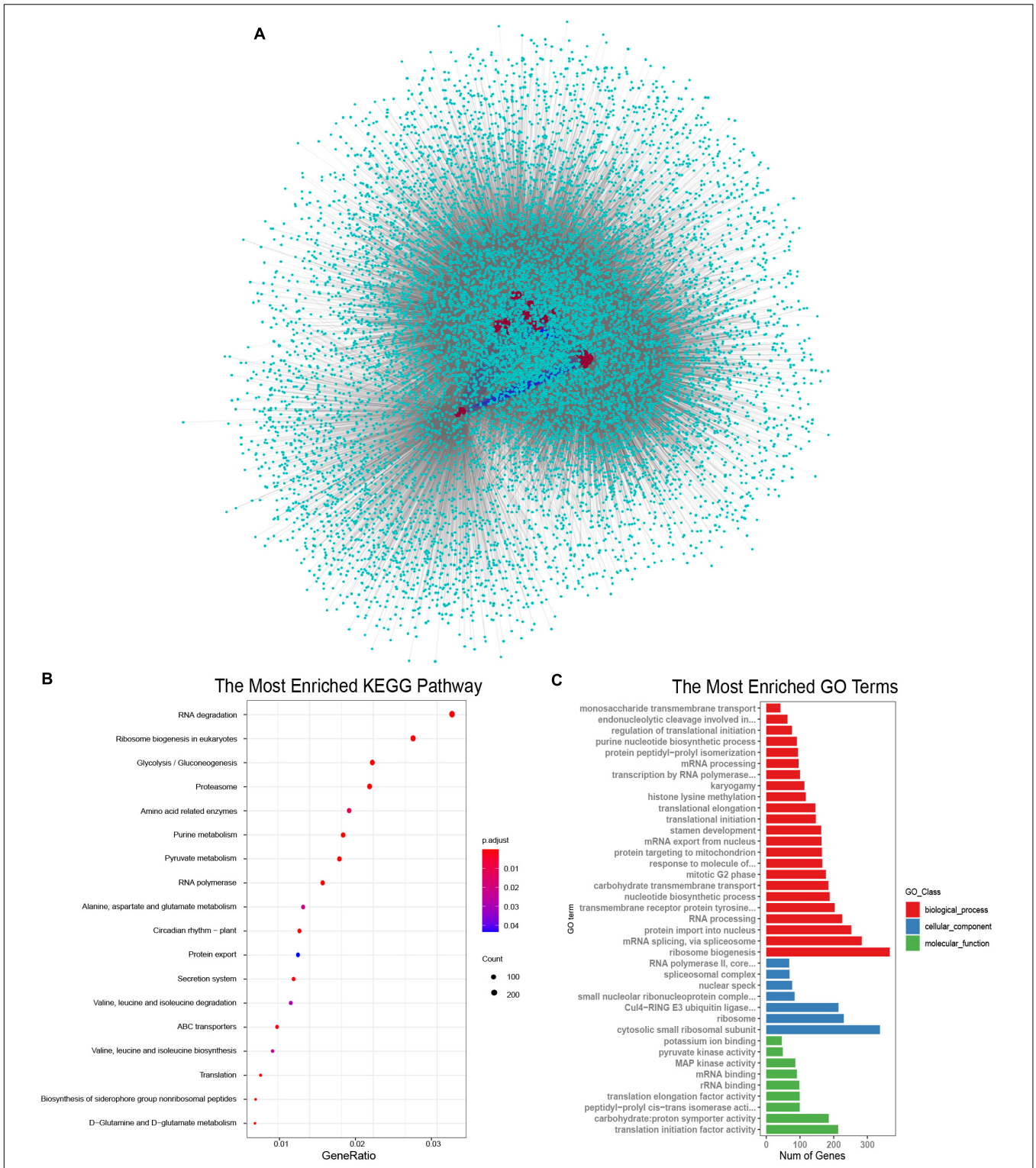

FIGURE 5 | Proteins interacted with SR proteins in *B. napus*. **(A)** Protein-protein interaction network of SR proteins in *B. napus*. The red circles represented the SR proteins, the indigo circles represented proteins that interacted with SR proteins. The blue lines represented the interaction between SR proteins, and the gray lines represented the interaction between SR proteins and other proteins. **(B)** KEGG pathway enrichment analysis of proteins interacted with SR proteins. **(C)** Gene Ontology enrichment analysis of proteins interacted with SR proteins.

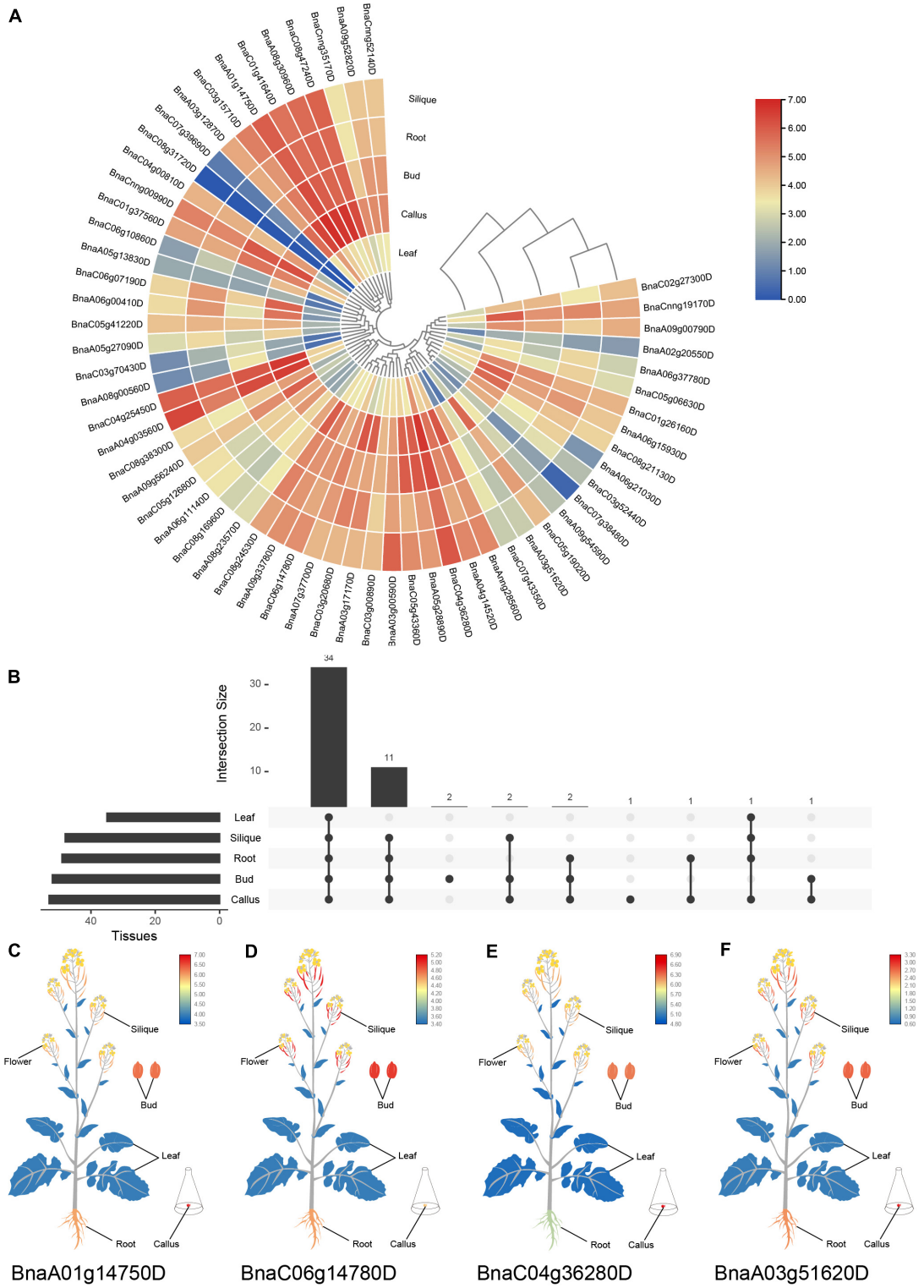

FIGURE 6 | Expression profiles of *SR* genes in different tissues of *B. napus*. **(A)** Heatmap representation of 59 *SR* genes in different tissues. **(B)** Number of *SR* genes that were expressed in various tissues. **(C–F)** The expression patterns of four selected *SR* genes in *B. napus* plants. Expression data were processed with $\log_2$ normalization. The color scale represented relative expression levels from low (blue color) to high (red color).

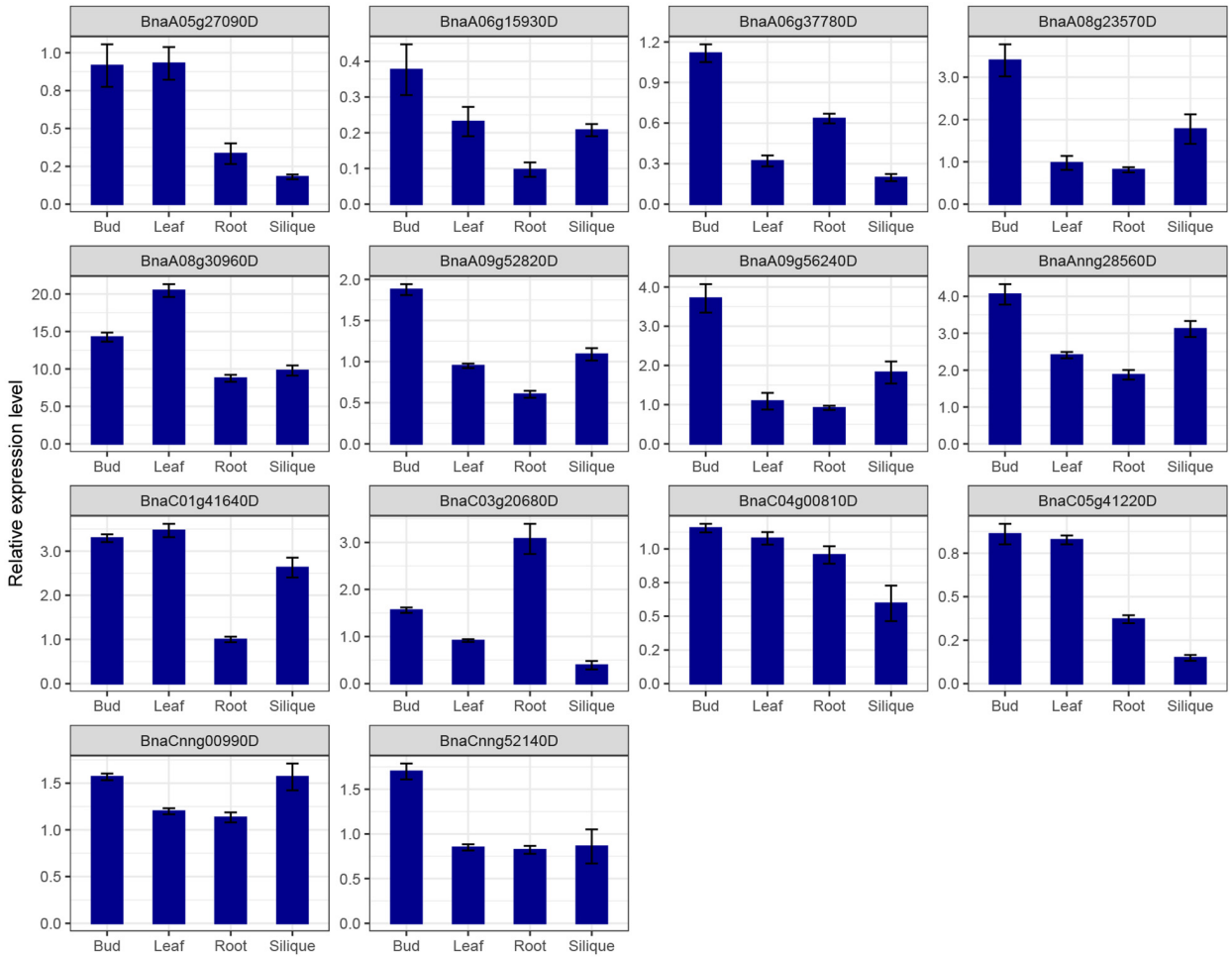

FIGURE 7 | qRT-PCR expression analysis of 14 *SR* genes in different tissues of *B. napus*. The error bars represented the standard error of the means of three replicates.

RS2Z and subfamily SC, but in other subfamilies, different expression patterns were observed, especially for some paralogous gene pairs, like *BnaC03g15710D*/*BnaC07g39690D*, *BnaC04g00810D*/*BnaC08g31720D*, and *BnaA02g20550D*/*BnaA09g00790D*, coincidentally, these gene pairs also showed different patterns in various tissues, which suggested they were differentiated into different directions, and the low-expressed genes like *BnaC07g39690D*, *BnaC04g00810D* and *BnaA02g20550D* may become pseudogenes.

Alternative Splicing of *SR* Genes Is Widespread in *Brassica napus*

In Arabidopsis, maize and sorghum, most of the *SR* genes could be alternatively spliced, in order to investigate the alternative splicing (AS) of *SR* genes in *B. napus*, we used the dataset from Pacbio Iso-Seq technique, which could directly detect the existed mRNA and provide full-length transcripts. Based on Iso-Seq of *B. napus* cultivar "ZS11" (Yao et al., 2020), 51 of 59 *SR* genes were detected in this dataset, and 41 *SR* genes were alternative spliced, yielding 206 transcripts, an average of 5 transcripts for each gene (**Figure 9A** and **Supplementary Table 6**). As to each subfamily, *SR* genes in subfamily RS owned the most transcripts per gene (average 6.4 transcripts), whereas *SR* genes in subfamily SC contained the least transcripts, only 1.7 transcripts per gene, and the other subfamily RS, SR45, RS2Z, SCL, and RSZ contained 6.2, 4.3, 4.3, 2, and 1.8 transcripts, respectively. In the multi-exon *SR* genes, a total of 163 AS events were discovered, intron retention (IR) was the most one (87), followed by alternative 3′ splice site (A3SS, 38), alternative 5′ splice site (A5SS, 21) and exon skipping (ES, 17) (**Figure 9B**). Subfamily RS had 51 AS events (IR-29, A3SS-8, A5SS-9, ES-5), which was the most and consistent with its most transcripts. While the fewer transcripts in subfamily RSZ and SC contained fewer AS events. Most of the paralogous gene pairs displayed distinct splicing patterns, the first one was the transcripts number varied between paralogous gene pairs, like 2 transcripts of *BnaA06g11140D* vs. 4 transcripts of *BnaC05g12680D*, and 8 transcripts of *BnaA03g17170D* vs. 3 transcripts of *BnaA07g37700D*, the second one was the AS events varied between paralogous gene pairs, both *BnaA04g03560D* and

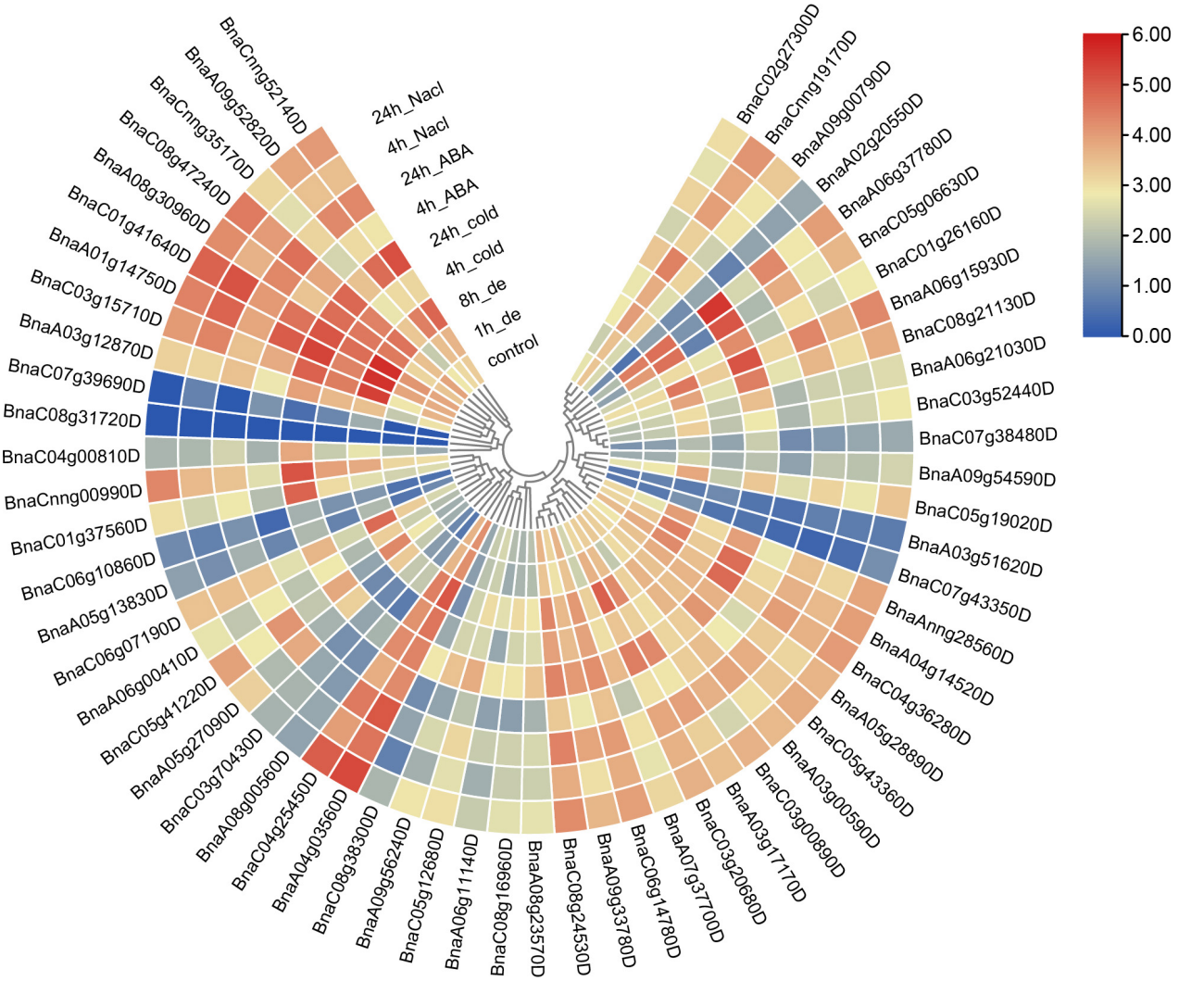

FIGURE 8 | Expression profiles of *SR* genes under abiotic stress conditions. Expression data were processed with $\log_2$ normalization. The color scale represented relative expression levels from low (blue color) to high (red color).

BnaC04g25450D had 2 transcripts, but the identified AS events were different (**Figure 9C**). To verify the AS events, the detailed alignment information was displayed, and it showed that a small number of reads could span the splice sites (**Supplementary Figure 2**). Moreover, EST dataset was also used to blast against the alternative splicing transcripts, and the results revealed that the different AS events really existed (**Supplementary Table 7**). To find out the expression patterns of transcripts in various tissues, the expression levels of all the transcripts of *SR* genes were also counted (**Supplementary Figure 3**), and it showed that only a fraction of them expressed higher in these tissues, for paralogous gene pair *BnaA04g03560D*/*BnaC04g25450D*, the expression patterns of their transcripts were also different.

Moreover, in the RNA-seq of abiotic stresses, the short reads were assemblied to predict the splicing profiles (**Supplementary Figure 4**), finally 124 transcripts were detected in 46 genes, and 61 AS events were identified. In this dataset, IR was not the most prevalent AS type, instead, A3SS was more prevalent. Five transcripts of *BnaA06g37780D*, *BnaC05g06630D*, and *BnaA01g14750D* were obviously induced by all four stresses, and the increment was obvious as the treatment time increased (**Supplementary Figure 4**), indicating that they were the responsible splicing factors responding to abiotic stress in *B. napus*.

Genetic Effects of *SR* Genes on Agronomic Traits of *Brassica napus*

To investigate the genetic variations of *SR* genes, SNPs were identified in a natural population containing 324 accessions collected from worldwide countries (**Supplementary Table 8**; Tang, 2019). Averagely, each *SR* gene contained 43 SNPs, lower than the whole genome level (94 SNPs in each gene). In consideration of genome size, we calculated the

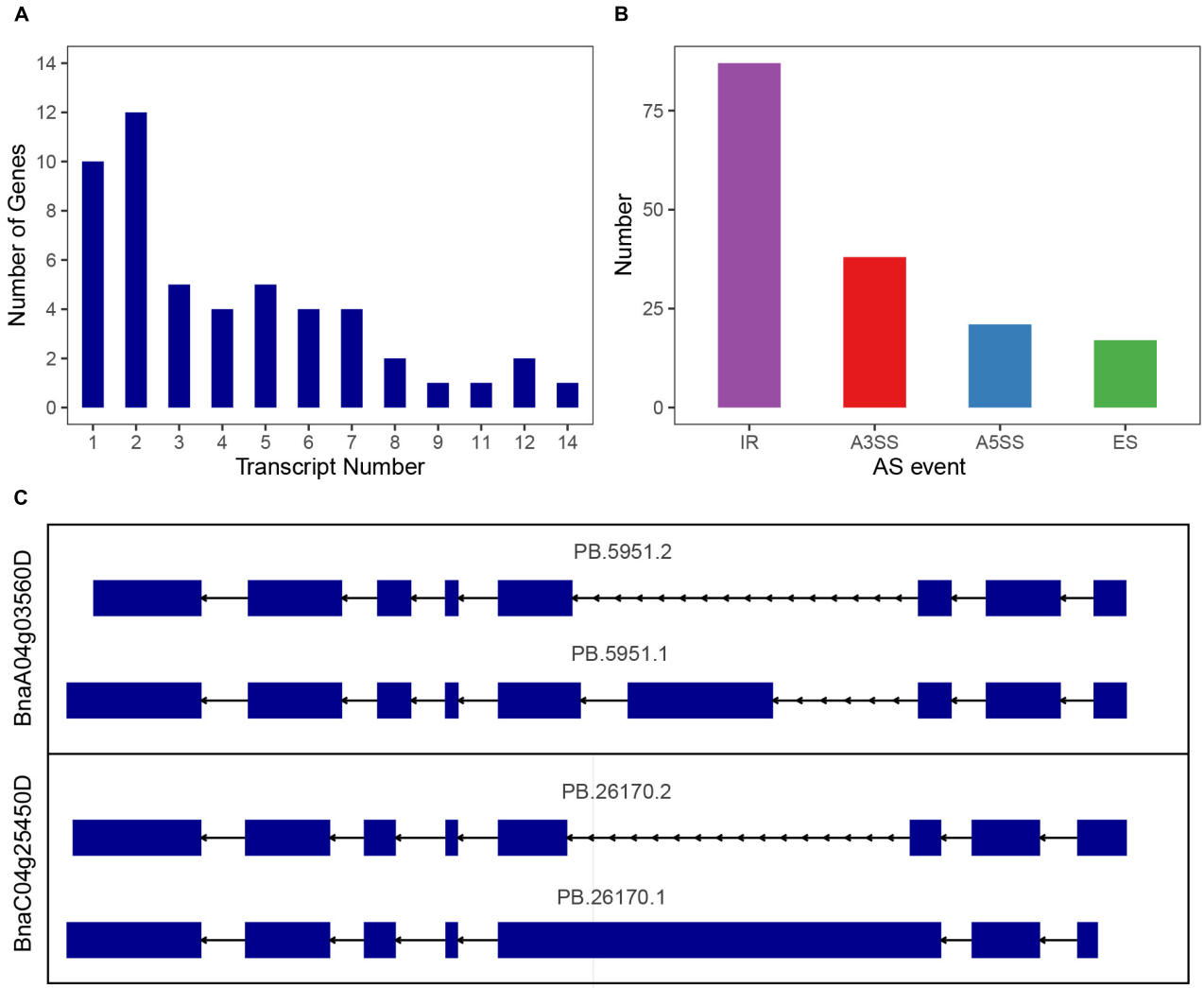

FIGURE 9 | Splicing profiles of *SR* genes in *B. napus* based on Iso-Seq data. **(A)** The numbers of *SR* genes that produced one or more transcripts from Iso-Seq data. **(B)** Classification of AS events. IR, intron retention; A3SS, alternative 3′ splice site; A5SS, alternative 5′ splice site; ES, exon skipping. **(C)** Different splicing patterns of a paralogous gene pair.

average SNP number of each kilobase (kb), all the *SR* genes were 17 SNPs/kb, while the whole genome level was 11 SNPs/kb. The SNP density of *SR* genes in the A subgenome (22 SNPs/kb) was slightly higher than the C subgenome (13 SNPs/kb). Moreover, the SNP density varied in different subfamilies, like subfamily SR45 had the most, with an average of 90 SNPs, followed by RSZ (41 SNPs) and SCL (39 SNPs), while SC had the fewest (only 29 SNPs). We also examined the genetic variations of paralogous gene pairs, there were 97 SNPs in *BnaA09g00790D*, but none in its paralogous gene *BnaCnng19170D*, while paralogous gene pairs *BnaC04g00810D*/*BnaC08g31720D*, had 49 and 5 SNPs, respectively. On the whole, most paralogous gene pairs exhibited unequal variations. Finally, SNP annotation showed that 658 SNPs occurred in exon regions and 194 SNPs in 39 *SR* genes resulted in missense mutations.

For *SR* genes were the fundamental regulators in pre-mRNA processing, it could affect various physiological processes, and finally result in diverse phenotype (Shepard and Hertel, 2009; Reddy and Shad Ali, 2011). In order to study the impact of *SR* genes on agronomic traits in *B. napus*, the association mapping analysis was conducted for 12 agronomic traits. In total, 49 SNPs (corresponding to 12 *SR* genes, **Supplementary Table 8**) located on A03, A05, A09, C03, C04, C05, C06, C07 and unanchored scaffolds were significantly associated with one or more agronomic traits ($p < 0.001$). *BnaC04g00810D* was significantly associated with main inflorescence silique density (**Figures 10A,B**), and the missense mutation in the coding sequence changed the arginine to histidine (305G > A). According to the genotype, two groups were divided and the main inflorescence silique density was significantly different based on the *t*-test ($p < 3.2e{-}10$) (**Figure 10C**).

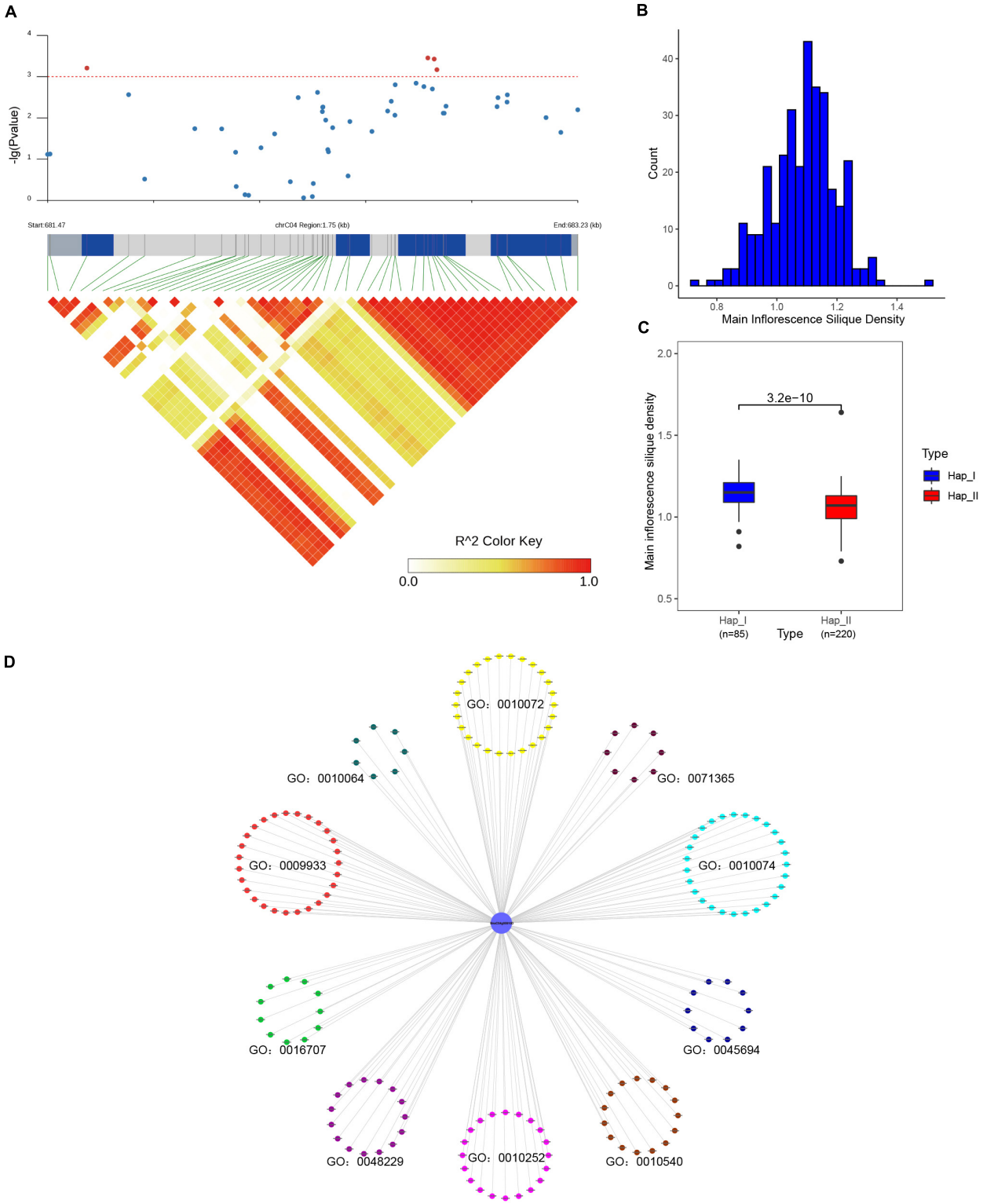

FIGURE 10 | Association mapping analysis of *BnaC04g00810D* in 324 core collections of *B. napus* germplasm. **(A)** *BnaC04g00810D* was significantly associated with main inflorescence silique density. **(B)** The distribution of main inflorescence silique density. **(C)** Comparison of main inflorescence silique density between the two haplotypes based on the most significantly associated SNP of *BnaC04g00810D*. **(D)** The enriched Gene Ontology terms of interacted proteins of *BnaC04g00810D*.

The interacted proteins of *BnaC04g00810D* were analyzed, they were not only enriched in mRNA splicing and spliceosome, but also enriched in the maintenance of meristem identity (GO:0010074), regulation of embryo sac egg cell differentiation (GO:0045694), meristem structural organization (GO:0009933), primary shoot apical meristem specification (GO:0010072), embryonic shoot morphogenesis (GO:0010064), gibberellin 3-beta-dioxygenase activity (GO:0016707), auxin homeostasis (GO:0010252), basipetal auxin transport (GO:0010540), cellular response to auxin stimulus (GO:0071365) (**Figure 10D**). As we knew, gibberellins (GAs) could promote stem elongation and floral development during bolting (Olszewski et al., 2002), auxin biosynthesis and transport played an important role in floral meristem initiation and inflorescence organization (Teo et al., 2014). All these processes were related with the regulation of endogenous hormone and the development of meristem/gametophyte, which could affect the silique density (Ren et al., 2018). The interacted proteins of *BnaC04g00810D* took part in these processes, like GA3OX1/2/4 in GO:0016707 were responsible for the last step of the biosynthetic of active GAs (Williams et al., 1998), ABCB19 in GO:0010540 mediated polar auxin transport (Wu et al., 2016), and GAF1 was involved in female gametophyte development (Zhu et al., 2016). Therefore, it was speculated that *BnaC04g00810D* also participated in the above processes through interacting with related proteins and might be an important candidate gene for silique density in *B. napus*. Moreover, *BnaA03g12870D* was significantly associated with flowering time and branch number, whereas *BnaC03g20680D* was significantly associated with the flowering period (**Supplementary Figure 5**), and the involved processes of their interacted proteins were also enriched in meristem structural organization, regulation of flower development and so on. Overall, the results suggested that sequence variations of *SR* genes could affect the development of *B. napus* and, ultimately influence the important agronomic traits.

DISCUSSION

Alternative splicing plays important role in the plant growth and development process, especially enhancing the adaptability of plants under stress conditions (Black, 2003; Palusa et al., 2007). Splicing factors are essential for the execution and regulation of splicing. Among them, SR proteins are the prominent factors involved in the assembly of spliceosomes, recognition and splicing of pre-mRNAs (Zahler et al., 1992). Recently, SR proteins in many plants have been studied at the genome-wide level to understand their evolution and function (Kalyna and Barta, 2004; Isshiki et al., 2006; Richardson et al., 2011; Chen et al., 2019, 2020b; Gu et al., 2020). In this study, 59 *SR* genes were identified and characterized in *B. napus*. A systematical analysis of *SR* genes including chromosomal locations, gene structures, conserved motifs, phylogenetic relationships, and protein-protein interactions was performed. Moreover, the expression patterns and AS types of *SR* genes in various tissues and stresses were analyzed. Variations in *SR* gene sequences and the association mapping analysis based on various agronomic traits were also performed to detect the relationship between *SR* genes and the final phenotype in *B. napus*.

After divergence from Arabidopsis lineage, the genus *Brassica* underwent a genome triplication event that occurred 13 million years ago, then interspecific hybridization between *B. rapa* and *B. oleracea* formed the allotetraploid *B. napus* (Allender and King, 2010). All the genes in *B. napus* expanded during its evolution and formation (Chalhoub et al., 2014). Many studies had shown that whole-genome duplication (WGD) and segmental duplications were the key factors to produce duplicated genes and result in the expansion of gene families (Ma et al., 2017; Wu et al., 2018; Zhu et al., 2020), as well as observed in *SR* genes in this study. Based on the effect of two recent duplication events, six homologs for each Arabidopsis gene were expected to present in *B. napus*, but we only found 59 *SR* genes in *B. napus* (about threefold of AtSRs), which indicated that gene loss happened (Albalat and Canestro, 2016). And the distribution of *SR* genes in the A and C subgenome implied the gene loss is asymmetrical, which is consistent with the genome level (Chalhoub et al., 2014). According to the Ka/Ks ratios of paralogous gene pairs, it is suggested that purifying selection played an important role in the evolution of *SR* genes in *B. napus*.

In plants, *SR* gene family had been investigated in Arabidopsis, rice, maize, wheat, tomato, cassava, and so on (Kalyna and Barta, 2004; Isshiki et al., 2006; Richardson et al., 2011; Yoon et al., 2018; Chen et al., 2019, 2020b; Gu et al., 2020; Rosenkranz et al., 2021). Most of the *SR* genes were divided into five to seven subfamilies according to the domain sequence or the whole sequence, likewise, 59 *SR* genes in *B. napus* were also classified into seven subfamilies. The proportion of plant-specific subfamily members in *B. napus* (31/59, 52.54%) was similar to that of other plants (Chen et al., 2019). Most genes in the same subfamily shared similar gene structures, conserved motifs, but the *cis*-acting regulatory elements in promoters emerged a big difference, which would affect the expression patterns (Zou et al., 2011; Oudelaar and Higgs, 2021). In the RNA-seq of various tissues, *SR* genes expressed obviously lower in leaf in comparison with bud, root, silique and callus, which was probably due to more complex splicing events in differentiated organs than mature organs, similarly, it had been proved that many *SR* genes expressed highly in early stages of fruit growth and development in tomato, which indicated a higher demand for factors to regulate pre-mRNA processing during cell expansion in immature green fruits (Rosenkranz et al., 2021). Various expression patterns of duplicated genes were also observed in this study, and it had been proved as one common way to lead to pseudogenization, neofunctionalization, or subfunctionalization in polyploids (Chaudhary et al., 2009). The lifestyle of plants is sessile, which is different from animals, environmental factors such as light, temperature, water or soil characteristics strongly influence their growth and development. As a result, plants have intelligently evolved various strategies for fleetly responding to changes (Meena et al., 2017). The diverse *cis*-acting regulatory elements in

the promoter regions of different *SR* genes indicated their expression could be induced by hormones or abiotic stress. The different types, copy numbers and combinations of *cis*-acting regulatory elements predicted the diversity of *SR* genes expression patterns and flexibility in response to different stresses. Under environmental stress or hormone induction, the expression patterns of most *SR* genes changed. Expression of *BnaA06g37780D* and *BnaC05g06630D* increased with the treatment of cold, drought, salinity and ABA, and it had been verified that its orthologous gene *AtSR30* was up-regulated by salinity stress (Tanabe et al., 2007).

Transcription is a flexible mechanism, which not only alters the gene expression but also could create diverse transcripts (Herbert and Rich, 1999). With the development of sequencing technology, it is possible to provide full-length transcripts by Iso-Seq directly (Abdel-Ghany et al., 2016; Wang et al., 2017), avoiding sequence assembly by short reads from RNA-seq. In the Iso-Seq data of the five tissues (Yao et al., 2020), 41 *SR* genes were alternatively spliced to produce 206 transcripts, which increased the transcriptome complexity greatly. If datasets from other various tissues and treatments were obtained, it was speculated that the amounts of SR transcripts were astounding in *B. napus*. AS not only regulated the gene expression, but also could cause neofunctionalization or subfunctionalization between paralogous genes (Zhang et al., 2009). Here we found diverse AS patterns that occurred in the paralogous gene pairs, this result supplied a clue for further functional study which would focus on the different transcripts of *SR* genes. Furthermore, *SR* genes generated a variety of transcripts by alternative splicing in response to abiotic stress. In Arabidopsis, it had been proved that the alternatively spliced transcripts of several *SR* genes were directly associated with plants' ability to adapt to different environmental stresses (Palusa et al., 2007; Rauch et al., 2014). Similarly, 21 SR transcripts were detected under salt stress in cassava, which indicated these transcripts might participate in the biological process induced by salt (Gu et al., 2020). In this study, five transcripts from three *SR* genes obviously increased their expression after prolonged treatments of four different stresses. However, further research is required to determine the precise function and regulatory mechanisms of these SR transcripts in response to abiotic stress.

Sequence variations of *SR* genes were investigated in a natural population of *B. napus* (Tang, 2019), the SNP density in *SR* genes was higher than the average level of the genome, implying that abundant variations have accumulated in the evolution of *SR* gene family. The greater SNP prevalence of *SR* genes in the A subgenome was consistent with other gene families such as *GATAs* in a core collection of *B. napus* (Zhu et al., 2020). For genes in polyploids, after predicting function through their orthologs, to distinguish the one which performs function among several paralogous genes is another question. One way is to verify the function of paralogous genes one by one through traditional transgenic analysis, another way is with the aid of association mapping analysis. Typically, changes between paralogous gene pairs were distinct, leading to pseudogenization, neofunctionalization or subfunctionalization (Schiessl et al., 2017). For example, in contrast to *Bn-CLG1C*, a dominant point mutation in *Bn-CLG1A* led to cleistogamy in *B. napus*, which was regarded as a gain-of-function semi-dominant mutation (Lu et al., 2012). A single "C-T" mutation in the coding sequence of *BnaA03.CHL*H hindered chloroplast development, resulting in yellow-virescent leaf, while *BnaC03.CHLH* maintained the virescent color of the leaf (Zhao et al., 2020). In this study, 194 missense mutations could introduce various divergences of *SR* genes in *B. napus*. For paralogous gene pairs *BnaC04g00810D*/*BnaC08g31720D*, the expressions of *BnaC04g00810D* in tissues were higher than *BnaC08g31720D*, the missense mutation in the coding sequence of *BnaC04g00810D* changed the arginine to histidine, the association analysis and enriched processes of interacted proteins indicated that it was candidate gene for regulating silique density in *B. napus*. In previous studies, over-expression or transgenic analysis had proved that *SR* genes could affect the development and morphology in Arabidopsis (Kalyna et al., 2003; Ali et al., 2007), although none of the *SR* genes were studied by experimental analysis in *B. napus*, the association mapping analysis performed in this study could provide a useful clue for understanding the effect of *SR* genes on final phenotype and supply candidate genes for further improving agronomic traits in *B. napus*.

CONCLUSION

In this study, a comprehensive genome-wide identification and characterization of *SR* genes in *B. napus* were conducted. In total, 59 *SR* genes were identified and classified into seven subfamilies. Genes belonging to the same subfamily shared similar gene structures and motifs. *Cis*-acting regulatory elements in the promoters of *SR* genes and expression patterns in various tissues and environmental stresses revealed that they played important roles in development and stress responses. Transcriptome datasets from Pacbio/Illumina platforms showed that alternative splicing of *SR* genes was widespread in *B. napus* and the majority of paralogous gene pairs displayed different splicing patterns. Protein-protein interaction analysis showed that *SR* genes were involved in the whole lifecycle of mRNA, from synthesis to decay. Furthermore, genetic variations in *SR* genes were also investigated, and the association mapping results indicated that 12 *SR* genes were candidate genes for regulating specific agronomic traits. In summary, these findings provide elaborate information about *SR* genes in *B. napus* and may serve as a platform for further functional studies and genetic improvement of agronomic traits in *B. napus*.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

MX, CT, and SL designed the research. MX, RZ, ZB, CZ, and LY performed the experiments. MX, RZ, FG, XC, JH, and YuL analyzed the data. MX, CT, and YaL wrote and revised the manuscript. All authors have read and approved the current version of the manuscript.

FUNDING

This research was funded by the National Key Research and Development Program of China (2018YFE0108000), the Agricultural Science and Technology Innovation Program of Chinese Academy of Agricultural Sciences (CAAS-ASTIP-2013-OCRI), China Agriculture Research System of MOF and MARA (CARS-12), and the Young Top-notch Talent Cultivation Program of Hubei Province.

ACKNOWLEDGMENTS

We thank Isobel Parkin and Gary Peng from Saskatoon Research Centre of Agriculture and Agri-Food Canada for meaningful discussion and constructive comments. We thank Zhixian Qiao of the Analysis and Testing Center at IHB for technical supports in RNA-seq analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.829668/full#supplementary-material

Supplementary Figure 1 | Amount of *cis*-acting regulatory elements in promoters of *SR* genes in *B. napus*. Elements numbers were processed with $\log_2$ normalization. The color scale represented amounts from low (blue color) to high (red color).

Supplementary Figure 2 | The alignment information of *BnaA04g03560D* and *BnaC04g25450D.*

Supplementary Figure 3 | Heatmap representation of transcripts of *SR* genes in different tissues. Expression data were processed with $\log_2$ normalization. The color scale represented relative expression levels from low (blue color) to high (red color).

Supplementary Figure 4 | Splicing profiles of *SR* genes in *B. napus* under abiotic stress condition. **(A)** Distribution of genes that produced one or more transcripts from RNA-Seq data. **(B)** Classification of AS events from RNA-Seq data. IR, intron retention; A3SS, alternative 3′ splice site; A5SS, alternative 5′ splice site; ES, exon skipping. **(C)** Five transcripts were obviously induced by all four stresses. Expression data were processed with $\log_2$ normalization. The color scale represented relative expression levels from low (blue color) to high (red color).

Supplementary Figure 5 | Association mapping analysis of *SR* genes in 324 core collections of *B. napus* germplasm. **(A,B)** *BnaA03g12870D* was significantly associated with primary flowering time. **(C,D)** *BnaA03g12870D* was significantly associated with branch number. **(E,F)** *BnaC03g20680D* was significantly associated with the flowering period.

REFERENCES

Abdel-Ghany, S. E., Hamilton, M., Jacobi, J. L., Ngam, P., Devitt, N., Schilkey, F., et al. (2016). A survey of the sorghum transcriptome using single-molecule long reads. *Nat. Commun.* 7:11706. doi: 10.1038/ncomms11706

Albalat, R., and Canestro, C. (2016). Evolution by gene loss. *Nat. Rev. Genet.* 17, 379–391. doi: 10.1038/nrg.2016.39

Ali, G. S., Palusa, S. G., Golovkin, M., Prasad, J., Manley, J. L., and Reddy, A. S. N. (2007). Regulation of Plant Developmental Processes by a Novel Splicing Factor. *PLoS One* 2:e471. doi: 10.1371/journal.pone.0000471

Allender, C., and King, G. (2010). Origins of the amphiploid species Brassica napus L. investigated by chloroplast and nuclear molecular markers. *BMC Plant Biol.* 10:54. doi: 10.1186/1471-2229-10-54

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

Black, D. L. (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* 72, 291–336. doi: 10.1146/annurev.biochem.72.121801.161720

Caceres, J. F., Misteli, T., Screaton, G. R., Spector, D. L., and Krainer, A. R. (1997). Role of the Modular Domains of SR Proteins in Subnuclear Localization and Alternative Splicing Specificity. *J. Cell Biol.* 138, 225–238. doi: 10.1083/jcb.138.2.225

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I. A., Tang, H., Wang, X., et al. (2014). Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *Science* 345, 950–953. doi: 10.1126/science.1253435

Chaudhary, B., Flagel, L., Stupar, R. M., Udall, J. A., Verma, N., Springer, N. M., et al. (2009). Reciprocal Silencing, Transcriptional Bias and Functional Divergence of Homeologs in Polyploid Cotton (Gossypium). *Genetics* 182, 503–517. doi: 10.1534/genetics.109.102608

Chen, C., Chen, H., Zhang, Y., Thomas, H., Frank, M., He, Y., et al. (2020a). TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp.2020.06.009

Chen, S., Li, J., Liu, Y., and Li, H. (2019). Genome-Wide Analysis of Serine/Arginine-Rich Protein Family in Wheat and Brachypodium distachyon. *Plants* 8:188. doi: 10.3390/plants8070188

Chen, T., Cui, P., Chen, H., Ali, S., Zhang, S., Xiong, L., et al. (2013). A KH-Domain RNA-Binding Protein Interacts with FIERY2/CTD Phosphatase-Like 1 and Splicing Factors and Is Important for Pre-mRNA Splicing in Arabidopsis. *PLoS Genet.* 9:e1003875. doi: 10.1371/journal.pgen.1003875

Chen, X., Huang, S., Jiang, M., Chen, Y., XuHan, X., Zhang, Z., et al. (2020b). Genome-wide identification and expression analysis of the SR gene family in longan (Dimocarpus longan Lour.). *PLoS One* 15:e0238032. doi: 10.1371/journal.pone.0238032

Cingolani, P., Platts, A., Wang, I. L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 6, 80–92. doi: 10.4161/fly.19695

Cruz, T. M., Carvalho, R. F., Richardson, D. N., and Duque, P. (2014). Abscisic acid (ABA) regulation of Arabidopsis SR protein gene expression. *Int. J. Mol. Sci.* 15, 17541–17564. doi: 10.3390/ijms151017541

Dong, S. S., He, W. M., Ji, J. J., Zhang, C., and Yang, T. L. (2020). LDBlockShow: a fast and convenient tool for visualizing linkage disequilibrium and haplotype blocks based on variant call format files. *Brief. Bioinformatics* 22:bbaa227. doi: 10.1093/bib/bbaa227

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

Gu, J., Ma, S., Zhang, Y., Wang, D., Cao, S., and Wang, Z. Y. (2020). Genome-Wide Identification of Cassava Serine/Arginine-Rich Proteins: Insights into Alternative Splicing of Pre-mRNAs and Response to Abiotic Stress. *Plant Cell Physiol.* 61, 178–191. doi: 10.1093/pcp/pcz190

He, Z., Zhang, H., Gao, S., Lercher, M. J., Chen, W. H., and Hu, S. (2016). Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. *Nucleic Acids Res.* 44, W236–W241. doi: 10.1093/nar/gkw370

Herbert, A., and Rich, A. (1999). RNA processing and the evolution of eukaryotes. *Nat. Genet.* 21, 265–269. doi: 10.1038/6780

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

Isshiki, M., Tsumoto, A., and Shimamoto, K. (2006). The Serine/Arginine-Rich Protein Family in Rice Plays Important Roles in Constitutive and Alternative Splicing of Pre-mRNA. *Plant Cell* 18, 146–158. doi: 10.1105/tpc.105.037069

Kalyna, M., and Barta, A. (2004). A plethora of plant serine/arginine-rich proteins: redundancy or evolution of novel gene functions? *Biochem. Soc. Trans.* 32, 561–564. doi: 10.1042/BST0320561

Kalyna, M., Lopato, S., and Barta, A. (2003). Ectopic Expression of atRSZ33 Reveals Its Function in Splicing and Causes Pleiotropic Changes in Development. *Mol. Biol. Cell* 14, 3565–3577. doi: 10.1091/mbc.e03-02-0109

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42, 348–354. doi: 10.1038/ng.548

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

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. (2007). Clustal W and Clustal X v. 2.0. *Bioinformatics* 23, 2947–2948. doi: 10.1093/bioinformatics/btm404

Letunic, I., Khedkar, S., and Bork, P. (2020). SMART: recent updates, new developments and status in 2020. *Nucleic Acids Res.* 49, D458–D460. doi: 10.1093/nar/gkaa937

Li, Y., Guo, Q., Liu, P., Huang, J., Zhang, S., Yang, G., et al. (2021). Dual roles of the serine/arginine-rich splicing factor SR45a in promoting and interacting with nuclear cap-binding complex to modulate the salt-stress response in Arabidopsis. *New Phytol.* 230, 641–655. doi: 10.1111/nph.17175

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2 Δ Δ C T Method. *Methods* 25, 402–408. doi: 10.1006/meth.2001.1262

Lu, S., Wang, J., Chitsaz, F., Derbyshire, M., Geer, R., Gonzales, N., et al. (2020). CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.* 48, D265–D268. doi: 10.1093/nar/gkz991

Lu, Y. H., Arnaud, D., Belcram, H., Falentin, C., Rouault, P., Piel, N., et al. (2012). A dominant point mutation in a RINGv E3 ubiquitin ligase homoeologous gene leads to cleistogamy in Brassica napus. *Plant Cell* 24, 4875–4891. doi: 10.1105/tpc.112.104315

Ma, J. Q., Jian, H. J., Yang, B., Lu, K., Zhang, A. X., Liu, P., et al. (2017). Genome-wide analysis and expression profiling of the GRF gene family in oilseed rape (Brassica napus L.). *Gene* 620, 36–45. doi: 10.1016/j.gene.2017.03.030

Magali, L. (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

McGlincy, N. J., and Smith, C. W. J. (2008). Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? *Trends Biochem. Sci.* 33, 385–393. doi: 10.1016/j.tibs.2008.06.001

Meena, K. K., Sorty, A. M., Bitla, U. M., Choudhary, K., Gupta, P., Pareek, A., et al. (2017). Abiotic Stress Responses and Microbe-Mediated Mitigation in Plants: The Omics Strategies. *Front. Plant Sci.* 8:172. doi: 10.3389/fpls.2017.00172

Melo, J. P., Kalyna, M., and Duque, P. (2020). Current Challenges in Studying Alternative Splicing in Plants: The Case of Physcomitrella patens SR Proteins. *Front. Plant Sci.* 11:286. doi: 10.3389/fpls.2020.00286

Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., and Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* 41:e121. doi: 10.1093/nar/gkt263

Olszewski, N., Sun, T. P., and Gubler, F. (2002). Gibberellin signaling: biosynthesis, catabolism, and response pathways. *Plant Cell* 14, S61–S80. doi: 10.1105/tpc.010476

Oudelaar, A., and Higgs, D. (2021). The relationship between genome structure and function. *Nat. Rev. Genet.* 22, 154–168. doi: 10.1038/s41576-020-00303-x

Palusa, S. G., Ali, G. S., and Reddy, A. S. (2007). Alternative splicing of pre-mRNAs of Arabidopsis serine/arginine-rich proteins: regulation by hormones and stresses. *Plant J.* 49, 1091–1107. doi: 10.1111/j.1365-313X.2006.03020.x

Palusa, S. G., and Reddy, A. (2010). Extensive coupling of alternative splicing of pre-mRNAs of serine/arginine (SR) genes with nonsense-mediated decay. *New Phytol.* 185, 83–89. doi: 10.1111/j.1469-8137.2009.03065.x

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295. doi: 10.1038/nbt.3122

Rauch, H. B., Patrick, T. L., Klusman, K. M., Battistuzzi, F. U., Mei, W., Brendel, V. P., et al. (2014). Discovery and expression analysis of alternative splicing events conserved among plant SR proteins. *Mol. Biol. Evol.* 31, 605–613. doi: 10.1093/molbev/mst238

Reddy, A. S., and Shad Ali, G. (2011). Plant serine/arginine-rich proteins: roles in precursor messenger RNA splicing, plant development, and stress responses. *Wiley Interdiscip. Rev. RNA* 2, 875–889. doi: 10.1002/wrna.98

Ren, Y., Cui, C., Wang, Q., Tang, Z., and Zhou, Q. (2018). Genome-wide association analysis of silique density on racemes and its component traits in Brassica napus L. *Sci. Agricult. Sin.* 54, 1020–1033. doi: 10.3864/j.issn.0578-1752.2018.06.002

Richardson, D. N., Rogers, M. F., Labadorf, A., Ben-Hur, A., Guo, H., Paterson, A. H., et al. (2011). Comparative analysis of serine/arginine-rich proteins across 27 eukaryotes: insights into sub-family classification and extent of alternative splicing. *PLoS One* 6:e24542. doi: 10.1371/journal.pone.0024542

Rosenkranz, R. R. E., Bachiri, S., Vraggalas, S., Keller, M., Simm, S., Schleiff, E., et al. (2021). Identification and Regulation of Tomato Serine/Arginine-Rich Proteins Under High Temperatures. *Front. Plant Sci.* 12:645689. doi: 10.3389/fpls.2021.645689

Sapra, A. K., Ank, M. L., Grishina, I., Lorenz, M., and Neugebauer, K. M. (2009). SR Protein Family Members Display Diverse Activities in the Formation of Nascent and Mature mRNPs In Vivo. *Mol. Cell* 34, 179–190. doi: 10.1016/j.molcel.2009.02.031

Schiessl, S., Huettel, B., Kuehn, D., Reinhardt, R., and Snowdon, R. (2017). Post-polyploidisation morphotype diversification associates with gene copy number variation. *Sci. Rep.* 7:41845. doi: 10.1038/srep41845

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

Shepard, P. J., and Hertel, K. J. (2009). The SR protein family. *Genome Biol.* 10:242. doi: 10.1186/gb-2009-10-10-242

Sylvain, F., and Michael, S. (2007). ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. *Nucleic Acids Res.* 35, W297–W299. doi: 10.1093/nar/gkm311

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., et al. (2021). The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* 49, D605–D612. doi: 10.1093/nar/gkaa1074

Tamura, K., Stecher, G., and Kumar, S. (2021). MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Mol. Biol. Evol.* 38, 3022–3027. doi: 10.1093/molbev/msab120

Tanabe, N., Yoshimura, K., Kimura, A., Yabuta, Y., and Shigeoka, S. (2007). Differential expression of alternatively spliced mRNAs of Arabidopsis SR protein homologs, atSR30 and atSR45a, in response to environmental stress. *Plant Cell Physiol.* 48, 1036–1049. doi: 10.1093/pcp/pcm069

Tang, M. (2019). *Population genome variations and subgenome asymmetry in Brassica napus L.* Huazhong: Huazhong Agricultural University.

Teo, Z. W., Song, S., Wang, Y. Q., Liu, J., and Yu, H. (2014). New insights into the regulation of inflorescence architecture. *Trends Plant Sci.* 19, 158–165. doi: 10.1016/j.tplants.2013.11.001

Troncoso-Ponce, M. A., Kilaru, A., Cao, X., Durrett, T. P., Fan, J., Jensen, J. K., et al. (2011). Comparative deep transcriptional profiling of four developing oilseeds. *Plant J.* 68, 1014–1027. doi: 10.1111/j.1365-313X.2011.04751.x

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. Protem. Bioinf.* 8, 77–80. doi: 10.1016/S1672-0229(10)60008-3

Wang, M., Wang, P., Liang, F., Ye, Z., Li, J., Shen, C., et al. (2017). A global survey of alternative splicing in allopolyploid cotton: landscape, complexity and regulation. *New Phytol.* 217, 163–178. doi: 10.1111/nph.14762

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

Will, C. L., and Luhrmann, R. (2010). Spliceosome Structure and Function. *Cold Spring Harb. Perspect. Biol.* 3:a003707. doi: 10.1101/cshperspect.a003707

Williams, J., Phillips, A. L., Gaskin, P., and Hedden, P. (1998). Function and Substrate Specificity of the Gibberellin 3β-Hydroxylase Encoded by the Arabidopsis GA4 Gene. *Plant Physiol.* 117, 559–563. doi: 10.1104/pp.117.2.559

Wu, G., Carville, J. S., and Spalding, E. P. (2016). ABCB19-mediated polar auxin transport modulates Arabidopsis hypocotyl elongation and the endoreplication variant of the cell cycle. *Plant J.* 85, 209–218. doi: 10.1111/tpj.13095

Wu, Y., Ke, Y., Wen, J., Guo, P., Ran, F., Wang, M., et al. (2018). Evolution and expression analyses of the MADS-box gene family in Brassica napus. *PLoS One* 13:e0200762. doi: 10.1371/journal.pone.0200762

Yan, Q., Xia, X., Sun, Z., and Fang, Y. (2017). Depletion of Arabidopsis SC35 and SC35-like serine/arginine-rich proteins affects the transcription and splicing of a subset of genes. *PLoS Genet.* 13:e1006663. doi: 10.1371/journal.pgen.1006663

Yao, S., Liang, F., Gill, R. A., Huang, J., Cheng, X., Liu, Y., et al. (2020). A global survey of the transcriptome of allopolyploid Brassica napus based on single-molecule long-read isoform sequencing and Illumina-based RNA sequencing data. *Plant J.* 103, 843–857. doi: 10.1111/tpj.14754

Yoon, E. K., Krishnamurthy, P., Kim, J. A., Jeong, M.-J., and Lee, S. I. (2018). Genome-wide Characterization of Brassica rapa Genes Encoding Serine/arginine-rich Proteins: Expression and Alternative Splicing Events by Abiotic Stresses. *J. Plant Biol.* 61, 198–209. doi: 10.1007/s12374-017-0391-6

Yu, C. S., Chen, Y. C., Lu, C. H., and Hwang, J. K. (2006). Prediction of protein subcellular localization. *Proteins Struct. Funct. Bioinform.* 64, 643–651. doi: 10.1002/prot.21018

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

Zahler, A. M., Lane, W. S., Stolk, J. A., and Roth, M. B. (1992). SR proteins: a conserved family of pre-mRNA splicing factors. *Genes Dev.* 6, 837–847. doi: 10.1101/gad.6.5.837

Zhang, W., Du, B., Di, L., and Qi, X. (2014). Splicing factor SR34b mutation reduces cadmium tolerance in Arabidopsis by regulating iron-regulated transporter 1 gene. *Biochem. Biophys. Res. Commun.* 455, 312–317. doi: 10.1016/j.bbrc.2014.11.017

Zhang, Y., Ali, U., Zhang, G., Yu, L., Fang, S., Iqbal, S., et al. (2019). Transcriptome analysis reveals genes commonly responding to multiple abiotic stresses in rapeseed. *Mol. Breed.* 39:158. doi: 10.1007/s11032-019-1052-x

Zhang, Z., Li, Z., Ping, W., Yang, L., Chen, X., and Hu, L. (2009). Divergence of exonic splicing elements after gene duplication and the impact on gene structures. *Genome Biol.* 10:R120. doi: 10.1186/gb-2009-10-11-r120

Zhang, Z., Xiao, J., Wu, J., Zhang, H., Liu, G., Wang, X., et al. (2012). ParaAT: A parallel tool for constructing multiple protein-coding DNA alignments. *Biochem. Biophys. Res. Commun.* 419, 779–781. doi: 10.1016/j.bbrc.2012.02.101

Zhao, C., Liu, L., Safdar, L. B., Xie, M., Xiaohui Cheng, Liu, Y., et al. (2020). Characterization and Fine Mapping of a Yellow-Virescent Gene Regulating Chlorophyll Biosynthesis and Early Stage Chloroplast Development in Brassica napus. *G3 Genes Genom. Genet.* 10, 3201–3211. doi: 10.1534/g3.120.401460

Zhu, D. Z., Zhao, X. F., Liu, C. Z., Ma, F. F., Wang, F., Gao, X. Q., et al. (2016). Interaction between RNA helicase ROOT INITIATION DEFECTIVE 1 and GAMETOPHYTIC FACTOR 1 is involved in female gametophyte development in Arabidopsis. *J. Exp. Bot.* 67, 5757–5768. doi: 10.1093/jxb/erw341

Zhu, W., Guo, Y., Chen, Y., Wu, D., and Jiang, L. (2020). Genome-wide identification, phylogenetic and expression pattern analysis of GATA family genes in Brassica napus. *BMC Plant Biol.* 20:543. doi: 10.1186/s12870-020-02752-2

Zou, C., Sun, K., Mackaluso, J. D., Seddon, A. E., Jin, R., Thomashow, M. F., et al. (2011). Cis-regulatory code of stress-responsive transcription in Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U S A.* 108, 14992–14997. doi: 10.1073/pnas.1103202108

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

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

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 23 June 2022
doi: 10.3389/fpls.2022.924892

Genome-Wide Association Mapping of Hulless Barely Phenotypes in Drought Environment

Jie Li[1,2,3,4], *Xiaohua Yao*[2,3,4], *Youhua Yao*[2,3,4], *Likun An*[2,3,4], *Zongyun Feng*[1,5]* and *Kunlun Wu*[2,3,4]*

[1] College of Agronomy Sichuan Agricultural University, Chengdu, China, [2] Academy of Agricultural and Forestry Sciences, Qinghai University, Xining, China, [3] Qinghai Key Laboratory of Hulless Barley Genetics and Breeding, Xining, China, [4] Qinghai Subcenter of National Hulless Barley Improvement, Xining, China, [5] State Key Laboratory of Crop Gene Exploration and Utilization in Southwest China, Chengdu, China

Drought stress is one of the main factors restricting hulless barley (*Hordeum vulgare* L. var. *nudum* Hook. f.) yield. Genome-wide association study was performed using 269 lines of hulless barley to identify single-nucleotide polymorphism (SNP) markers associated with drought-resistance traits. The plants were cultured under either normal or drought conditions, and various quantitative traits including shoot fresh weight, shoot dry weight, root fresh weight, root dry weight, leaf fresh weight, leaf saturated fresh weight, leaf dry weight, ratio of root and shoot fresh weight, ratio of root and shoot dry weight, shoot water loss rate, root water loss rate, leaf water content and leaf relative water content, and field phenotypes including main spike length, grain number per plant, grain weight per plant, thousand grain weight (TGW), main spike number, plant height, and effective spike number of plants were collected. After genotyping the plants, a total of 8,936,130 highly consistent population SNP markers were obtained with integrity > 0.5 and minor allele frequency > 0.05. Eight candidate genes potentially contributed to the hulless barley drought resistance were obtained at loci near significant SNPs. For example, *EMB506*, *DCR*, and *APD2* genes for effective spike number of plants, *ABCG11* gene for main spike number (MEN), *CLPR2* gene for main spike length, *YIP4B* gene for root and shoot dry weight (RSWD), and *GLYK* and *BTS* genes for TGW. The SNPs and candidate genes identified in this study will be useful in hulless barley breeding under drought resistance.

Keywords: hulless barley, GWAS, drought resistance, high throughput sequencing, quantitative traits, SNP

OPEN ACCESS

Edited by:
Zhe Liang,
Biotechnology Research Institute (CAAS), China

Reviewed by:
Maolin Wang,
Sichuan University, China
Mingxun Chen,
Northwest A&F University, China

***Correspondence:**
Zongyun Feng
zyfeng49@126.com
Kunlun Wu
wklqaaf@163.com

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

Received: *20 April 2022*
Accepted: *18 May 2022*
Published: *23 June 2022*

Citation:
Li J, Yao X, Yao Y, An L, Feng Z and Wu K (2022) Genome-Wide Association Mapping of Hulless Barely Phenotypes in Drought Environment. Front. Plant Sci. 13:924892. doi: 10.3389/fpls.2022.924892

INTRODUCTION

Plants live in complex and changeable environmental conditions, often bring huge misfortune on plant growth (Zhu, 2016). As the global climate becomes drier and warmer, more than 15% of the world's population faces severe water shortages (Schewe et al., 2014; Gong et al., 2020). Drylands cover 40% of the global land surface and drought has caused losses in agriculture up to $30 billion over the past decade (Dai, 2013; Gupta et al., 2020). Drought has brought a great strain on the growth of plants, at the meantime, plants also have corresponding effective measures to prevent water loss, maintain cell water content, and help plants to survive the difficult drought period. Understanding drought resistance and water use efficiency of plants will provide guarantee for

maintaining normal plant growth and improving agricultural yield under drought (Gupta et al., 2020; Yu et al., 2021).

Hulless barley (*Hordeum vulgare* L. var. *nudum* Hook. f.) is an important economic crop (He and Jia, 2008). As the only crop growing at high altitude, the planting area of hulless barley accounts for 43% of the grain crop area on the Qinghai Tibet Plateau (Dai et al., 2012; Zhong et al., 2016). Hulless barley has made great contribution as the main food, fuel, and livestock feed of the Tibetan people, and also is the raw material for beer, medicine, and health care products (Yang et al., 2013; Zhu et al., 2015; Liu et al., 2018). Hulless barley is rich in β-glucan, phenolic acid, and anthocyanins, which has high nutritional and medicinal value and is of great significance to human health (Bonoli et al., 2004; Siebenhandl et al., 2007; Kohyama et al., 2008; Zhao et al., 2015). The climate inside the Qinghai–Tibet Plateau is gradually drying out, and some scientists predict that only plants that can tolerate drought conditions will be able to settle on the plateau's platforms (Meng et al., 2017). Therefore, it is very important to study the drought tolerance of hulless barley.

To predict the important agronomic traits such as drought tolerance, it is necessary to understand the specific loci based on phenotype and the genetic structure of the traits. Genome-wide association study (GWAS) is just such a powerful tool for connecting genotypes–phenotypes (Korte and Farlow, 2013). Genome-wide association study refers to the association analysis of traits through the sequence and the SNP marker information on the whole genome so as to detect the loci significantly associated with the target trait (Li, 2013; Tam et al., 2019). Genome-wide association study provides higher resolution and finer scale association, and has been widely used in the identification of markers associated with desirable traits in crops (Nordborg and Weigel, 2008; Xu et al., 2017).

This study based on the identification results of hulless barley drought tolerance traits in 269 lines, SNP markers were developed by simplified genome sequencing (SLAF) to genotype natural populations. Using linear mixed model (LMM) and EmMax, the association between the quantitative traits of drought tolerance and genotype was analyzed, and the SNP loci and chromosome segments significantly associated with the target traits were screened.

MATERIALS AND METHODS

Genetic Materials

The 269 hulless barley lines with different drought resistance assessment were used as the GWAS panel in this study (**Supplementary Table 1**). Phenotypic observation was performed on each line, both in the laboratory and in the field. The laboratory experiment was conducted in two growth condition with three biological replicates. The normal culture group was used as control, and the treatment group was applied with PEG-6000 to simulate drought stress. The associated phenotypes including shoot fresh weight SFW (g),shoot dry weight SDW (g), root fresh weight RFW (g), root dry weight RDW (g), leaf fresh weight LFW (g), leaf saturated fresh weight SFW (g), leaf dry weight LDW (g), ratio of root and shoot fresh weight RSFW (%), ratio of root and shoot dry weight RSWD (%), shoot water loss rate SWLR (%), root water loss rate RWLR (%), leaf water content WC (%), and leaf relative water content RWC (%) were measured. Field planting data were collected in 2019–2020 from three different growing environments at two sites, including drought treatment and natural irrigation at two different habitats. The associated phenotypes of different habitats consisted of main spike length MSL (cm), grain number per plant GNPP, grain weight per plant GWPP (g), thousand grain weight TGW(g), main spike number MEN(g), plant height (cm) and effective spike number of plants ESNP.

Single-Nucleotide Polymorphism-Based Genotyping for 269 Hulless Barley Lines

In 2021, 269 pieces of hulless barley lines were planted in germinating boxes and cultured in greenhouse to two leaves stage. Whole-genome DNA of each germplasm resource leaves was extracted by CTAB method (Allen et al., 2006). The DNA quality and concentration were detected by 0.1% agarose gel electrophoresis, and whole-genome SNP genotyping was produced by Biomarker technologies company. The SLAF tags were developed by enzyme digestion (RsaI) of the genomic DNA, followed by adaptor ligation, amplification and purification. Then, the SLAF library were sequenced by Illumina Novaseq 6000. The sequencing reads were mapped to the reference genome by BWA software (Li and Durbin, 2009). GATK (McKenna et al., 2010) and samtools (Li et al., 2009) were used to identify SNPs. The intersection of SNP markers obtained by the two methods was used as the final reliable SNP marker dataset, and a total of 5,949,446 SNPs were obtained. The genotypic data obtained were screened as integrity > 0.8 and minor allele frequency (MAF) > 0.05.

Structure of Hulless Barley Population

Based on the SNPs obtained from the above genotypes, 269 phylogenetic trees of hulless barley was constructed by neighbor-joining (NJ) method (1,000 replicates) with Kimura 2-parameter (K2-P) model using MEGA X software (Kumar et al., 2018). The phylogenetic tree was colored based on the analysis results of STRUCTURE.

Genome-Wide Association Study and Candidate Gene Screening

Based on the developed high-density SNP molecular markers, GEMMA, FaST-LMM, and EMMAX were used for association analysis. Correlation analysis between phenotypic value of drought-tolerant-resistant traits and genotypes was carried out to obtain the p-value of each SNP. Screened with $p < 5 \times 10^{-6}$, the genetic variation loci most likely to affect the trait was selected. The quantile–quantile (Q–Q) scatter plot and Manhattan plot were made by the qqman package in R software.

To screen the drought-tolerant-resistant genes near the significant associative loci, the genetic information of specific association regions was queried from barley genome in plant whole-genome information database (http://plants.ensembl.org/index.html). All genes with coding regions in the 100–500-kb window were used for subsequent analysis.

TABLE 1 | Distribution of SLAF markers on chromosomes.

Chromosome ID	Chromosome length	SNP number	SNP number	Polymorphic SLAF
chr1H	558,535,432	914,610	103,229	55,541
chr2H	768,075,024	1,289,234	139,589	77,829
chr3H	699,711,114	1,532,190	130,401	75,513
chr4H	647,060,158	1,186,220	122,917	67,774
chr5H	670,030,160	1,451,138	117,223	66,198
chr6H	583,380,513	1,216,744	108,416	63,442
chr7H	657,224,000	1,345,994	118,728	68,117
chrUn	249,774,706	123,320	22,256	6,334

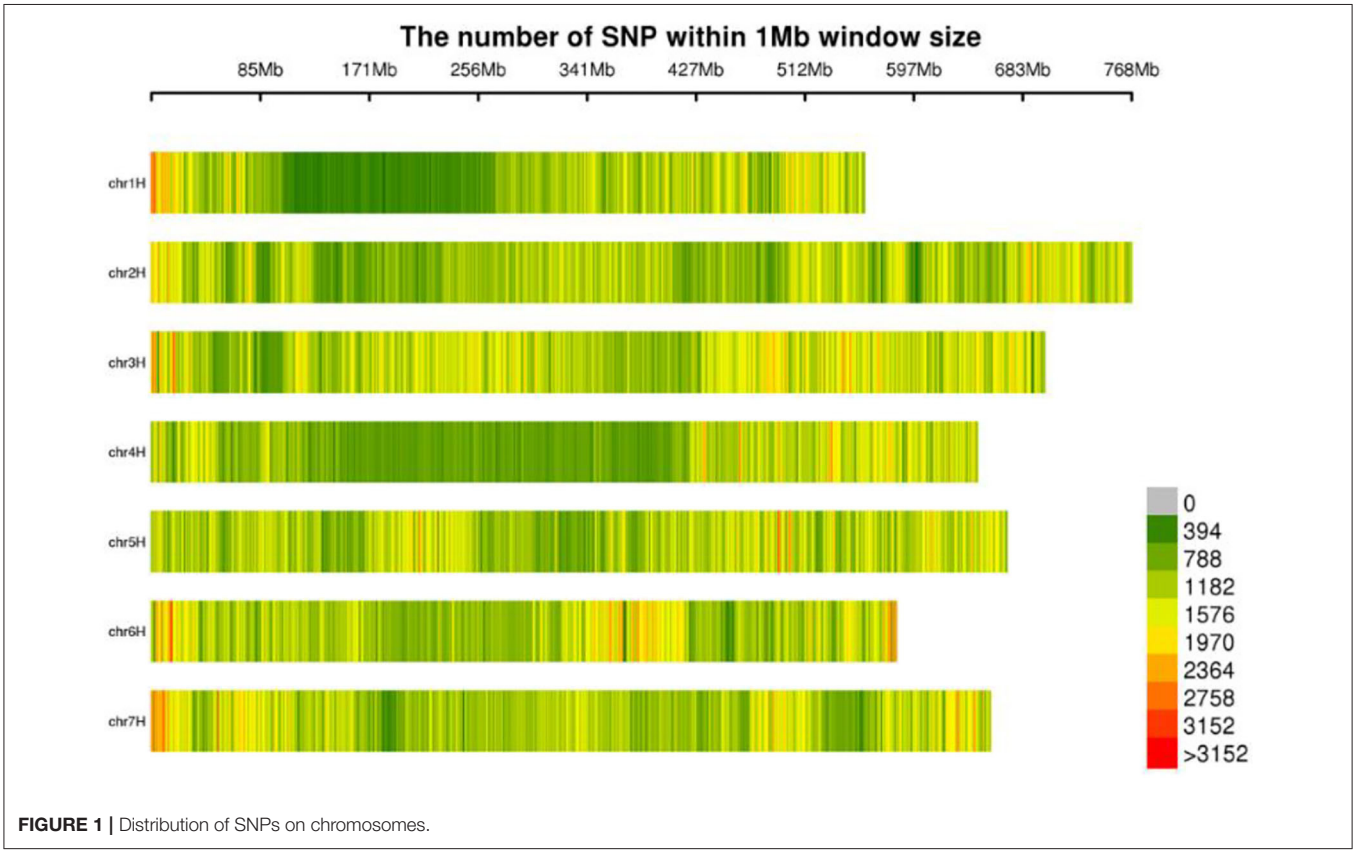

FIGURE 1 | Distribution of SNPs on chromosomes.

RESULTS

Genomic Library Construction and SNP Markers Development

The molecular markers of 269 hulless barley lines were developed by Specific-Locus Amplified Fragment Sequencing (LAF-SEQ) to obtain molecular markers in the whole genome. An average of 311,695 SLAF tags were developed per sample for a total of 862,999, including 480,790 polymorphic SLAF tags and 5,532,468 SNP markers. The average sequencing depth of SLAF tags was 10.43×, and 1,067.96 Mb reads data were generated. These markers were evenly distributed on the chromosomes of hulless barley (**Table 1**, **Figure 1**). The average Q30 of sequences was 94.78%, and the average GC content was 44.23%. A total of 8,936,130 SNP markers with high consistency were obtained from 269 hulless barley lines filtered by integrity > 0.5 and MAF > 0.05. Chromosome 3 had the largest number of SNP markers (1,532,190), with an average label distance of 456 bp. On the contrary, chromosome 1 had the lowest number of SNP markers (914,610), with an average label distance of 610 bp.

Genetic Structure of Hulless Barley Population

The 269 lines were divided into 6 groups according to the geographic location information of the samples, and the group information was used for linkage disequilibrium (LD) and evolutionary tree analysis. The phylogenetic tree was colored by the clustering result of STRUCTURE, basically, each cluster was

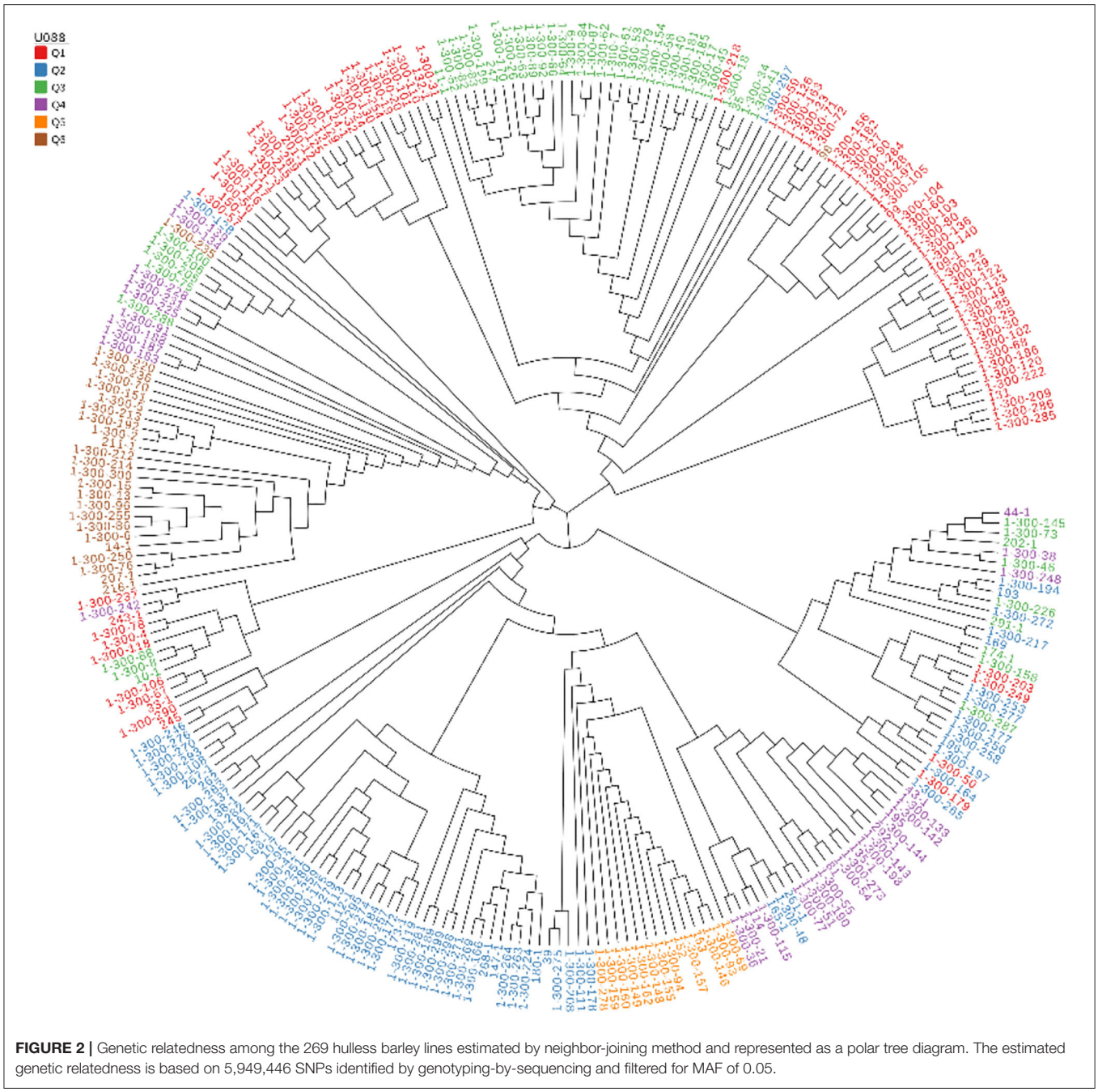

FIGURE 2 | Genetic relatedness among the 269 hulless barley lines estimated by neighbor-joining method and represented as a polar tree diagram. The estimated genetic relatedness is based on 5,949,446 SNPs identified by genotyping-by-sequencing and filtered for MAF of 0.05.

gathered into one block in the phylogenetic tree, especially Q5 and Q6 (**Figure 2**).

Using the SNP information mentioned above, the principal component analysis (PCA) was conducted. The top three principal components could explain 34.23% of the genomic variations, and principal component 1 could explain 18.3%. Consistent with the phylogenetic tree, Q5 and Q6 were separate from other populations (**Figure 3**).

Plink2 software (Chen et al., 2019) was used to calculate the linkage disequilibrium (LD) between two SNP pairs within a certain distance (1,000 kb) on the same chromosome, and the linkage disequilibrium intensity was represented by r^2. The closer r^2 is to 1, the stronger the linkage disequilibrium intensity. The distance between SNPs and r^2 was fitted, and the curve of r^2 changes with distance was presented. Generally speaking, the closer the distance between SNPs is, the larger r^2 is and *vice versa*. The LD decay (LDD) distance was used as the distance traveled when the maximum r^2 value dropped to half. The longer LDD, the lower the probability of recombination within the same physical distance. It should be noted that some regions of Q5 and Q6 had very strong linkage, but the

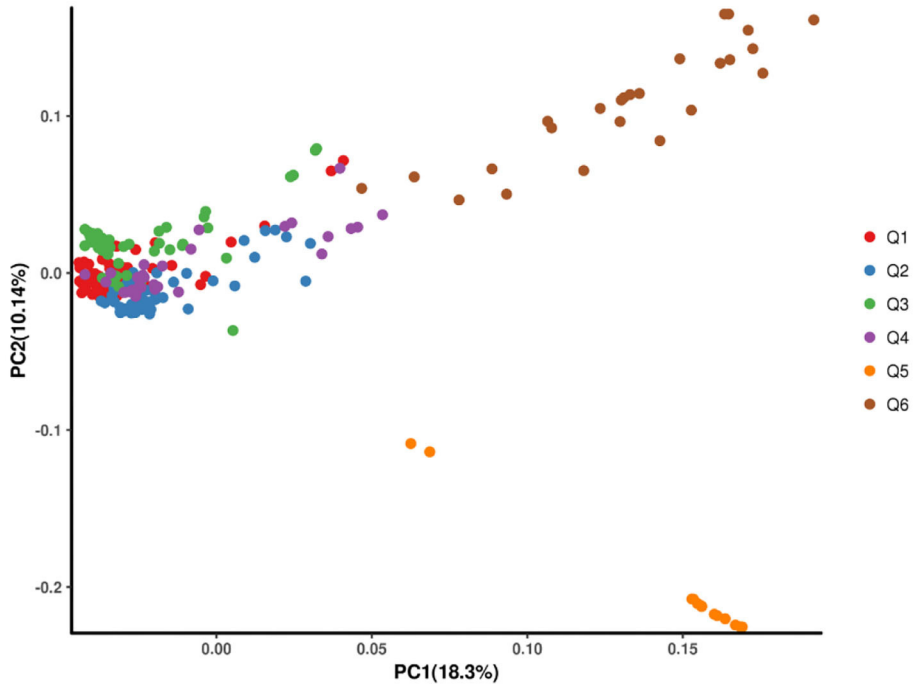

FIGURE 3 | The scatter plots of the first two principal components (PCs) showing the distribution of the 269 hulless barley lines in PC1 vs. PC2.

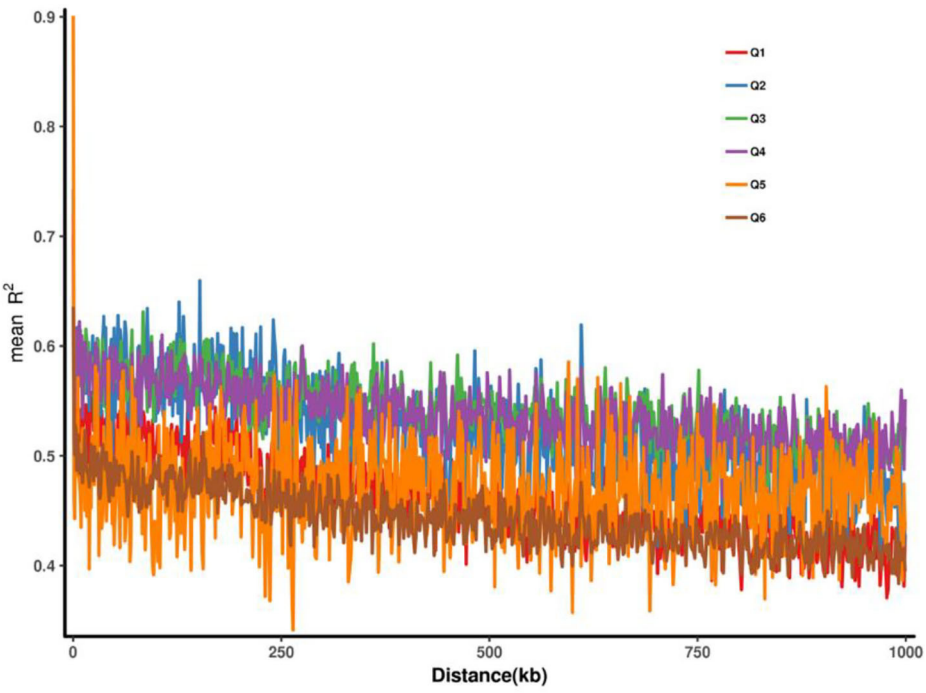

FIGURE 4 | Linkage disequilibrium decay based on six groups.

length of the strong linkage was short, indicating that these two groups were subjected to some artificial selection pressure and some loci were selected, leading to linkage in some regions (**Figure 4**).

TABLE 2 | SNP markers for each phenotype based on different GWAS analysis models.

Phenotype	Year	Station	Conditions	LMM	EMMAX	FASTLMM	Shared SNP markers
The field data							
ESN	2019	Menyuan	Field	18	6	10	6
	2019	Xining	Field	84	22	54	21
	2019	Xining	Greenhouse	10	3	7	3
	2020	Menyuan	Field	44	10	18	10
	2020	Xining	Field	117	6	74	6
	2020	Xining	Greenhouse	5	1	2	1
GW	2019	Menyuan	Field	10	3	11	3
	2019	Xining	Field	11	0	14	0
	2019	Xining	Greenhouse	6	1	6	1
	2020	Menyuan	Field	4	1	1	1
	2020	Xining	Field	11	1	6	1
	2020	Xining	Greenhouse	0	0	0	0
MSL	2019	Menyuan	Field	2	0	2	0
	2019	Xining	Field	7	2	5	2
	2019	Xining	Greenhouse	1	0	1	0
	2020	Menyuan	Field	191	1	41	1
	2020	Xining	Field	6	0	1	0
	2020	Xining	Greenhouse	1	0	2	0
PH	2019	Menyuan	Field	10	2	39	2
	2019	Xining	Field	7	1	6	1
	2019	Xining	Greenhouse	28	10	23	8
	2020	Menyuan	Field	14	2	6	2
	2020	Xining	Field	2	0	3	0
	2020	Xining	Greenhouse	0	0	0	0
SN	2019	Menyuan	Field	37	2	108	1
	2019	Xining	Field	32	26	34	23
	2019	Xining	Greenhouse	2	1	2	1
	2020	Menyuan	Field	24	6	79	6
	2020	Xining	Field	3	3	3	3
	2020	Xining	Greenhouse	5	3	5	3
SPP	2019	Menyuan	Field	0	0	0	0
	2019	Xining	Field	28	0	13	0
	2019	Xining	Greenhouse	12	1	7	1
	2020	Menyuan	Field	0	0	4	0
	2020	Xining	Field	7	0	3	0
	2020	Xining	Greenhouse	1	1	1	1
TGW	2019	Menyuan	Field	16	17	24	16
	2019	Xining	Field	14	2	8	2
	2019	Xining	Greenhouse	15	1	2	1
	2020	Menyuan	Field	47	33	55	31
	2020	Xining	Field	25	1	19	1
	2020	Xining	Greenhouse	4	0	4	0
Laboratory data							
SFW			Control	3	3	3	3
			PEG	2	0	1	0
LDMC			Control	0	0	0	0
			PEG-6000	0	0	0	0
LDW			Control	4	0	5	0
			PEG-6000	5	1	1	1

(Continued)

TABLE 2 | Continued

Phenotype	Year	Station	Conditions	LMM	EMMAX	FASTLMM	Shared SNP markers
LFW			Control	3	0	0	0
			PEG-6000	2	0	2	0
RDW			Control	1	0	0	0
			PEG-6000	2	1	2	1
RFW			Control	0	0	0	0
			PEG-6000	13	3	10	3
RSDW			Control	3	0	3	0
			PEG-6000	3	4	5	2
RSFW			Control	0	0	0	0
			PEG-6000	3	2	2	2
RWC			Control	1	1	1	1
			PEG-6000	23	1	48	1
RWLR			Control	3	3	4	2
			PEG-6000	1	2	2	1
SDW			Control	5	0	5	0
			PEG-6000	2	1	2	1
SWLR			Control	1,148	0	696	0
			PEG-6000	1	2	2	1
WC			Control	1	1	1	1
			PEG-6000	10	11	14	7

Genome-Wide Association Study Analysis of Traits in Hulless Barley

This analysis was based on SNP data from mutation detection, filtered by secondary allele frequency (MAF > 0.05) and locus integrity (integrity > 0.8) to obtain highly consistent SNP loci for GWAS analysis. Genome-wide association analysis was performed using LMM, EMMAX, and FaST-LMM models, respectively. The following table showed the number of significant SNP markers obtained for each phenotype corresponding to each mode and the number of common SNP markers in each model (**Table 2**).

Association Analysis

In the following part, we made a detailed explanation of some phenotypes which have shared SNP markers in the three relational models. The Manhattan plots showed significant correlation between SNP markers on multiple chromosomes and traits, while the Q–Q plots showed the relationship between observed *p*-values and expected *p*-values for each SNP marker (**Figure 5**, **Supplementary Table 2**).

Figure 5 showed the association between SNP markers and effective spike number of plants (ESNP) phenotype with Manhattan map and corresponding Q–Q plot. For plants cultured in greenhouse in Xining city during 2019, two SNP markers chr7H_102929775 and chr7H_102929728 with known functions were obtained. Both of them locate on the gene Ankyrin repeat domain-containing protein (*EMB506*). Gene *EMB506* is expressed at flowering and heading stage and closely associated with the character (spike number) (Despres et al., 2001), so it is likely to be the effector gene. There were another two SNP markers detected in plants cultured in filed in Menyuan city during 2019. One was chr1H_12690373, located on the Protein Dicer (*DCR*) gene, which is required for cutin polyester formation (Panikashvili et al., 2009). The other was chr1H_512690373, located on *APD2* gene involved in male gametophyte development (Luo et al., 2012). For main spike number (MEN) phenotype, 23 SNP markers were detected on chromosomes 1H, 2H, and 7H in plants which were cultured in filed in Xining city during 2019. On chromosome 2H, 18 SNP markers were found, including one aldehyde reductase gene Neuroplastin (*SDR1*), one gene interrelated with chloroplast development and plant growth named Probable GTP-binding protein (*OBGC1*), and three bacterial infection related genes Peroxidase 2 (*PRX112*), Probable acyl-CoA dehydrogenase (*IBR3*) and Ethylene-responsive transcription factor (*RAP2*-3). The ABC transporter G family member 11 (*ABCG11*) gene located on chromosome 2H was highly expressed in flowers and young seeds and was closely related to spike number (Panikashvili et al., 2010) so that *ABCG11* was likely to be the effector gene of the trait. Our results showed that two SNP markers (chr3H_152206655 and chr5H_250095923) connected with main spike length (MSL) phenotype. Thereinto, ATP-dependent Clp protease proteolytic subunit-related protein 1 (*CLPR*) is considered to regulate chloroplast and plant development. Deletion of *CLPR* alleles resulted in embryonic development delay and leaf albinism (Kim et al., 2009). Root and shoot dry weight (RSWD) phenotype was represented by chr1H_64014764 on Ypt Interacting Protein 4b (*YIP4B*) gene. The *YIP4B* regulates cell wall composition and participate in root and hypocotyl elongation (Gendre et al., 2013). As for thousand grain weight (TGW) phenotype, the first SNP peak was found

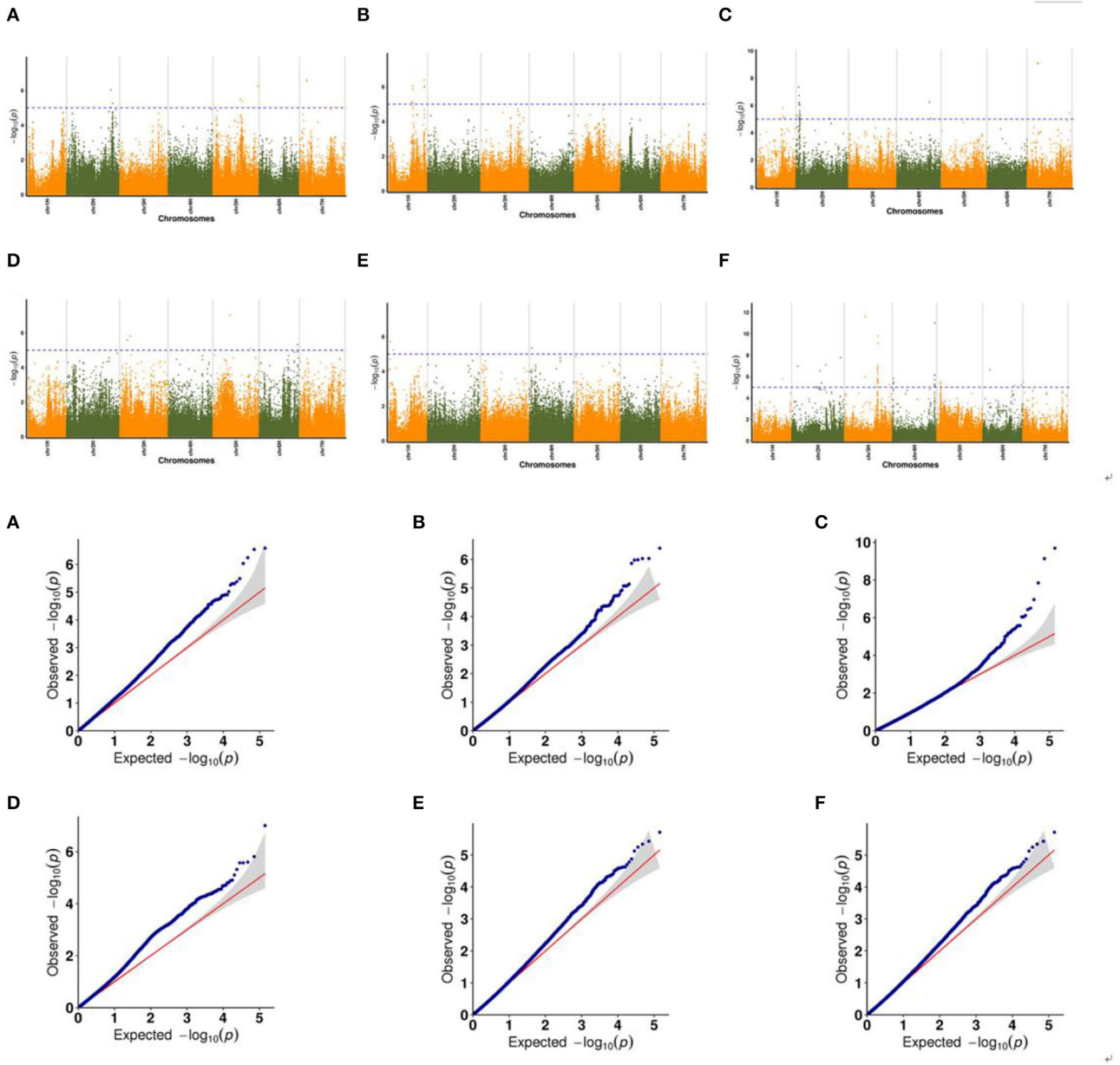

FIGURE 5 | Manhattan maps and Q–Q plots representing the SNP markers associated with drought resistant in multiple trials of GWAS associated with spike development. **(A,B)** ESNP, **(C)** MEN, **(D)** MSL, **(E)** RSWD, and **(F)** TGW.

at chr3H_482958549, and mapped to photosynthesis related gene D-glycerate 3-kinase, chloroplastic (*GLYK*). The second SNP peak was found at chr3H_489630701, and mapped to iron accumulation associated gene Geranylgeranyl pyrophosphate synthase (*BTS*).

DISCUSSION

Hulless barley is rich in nutrients and is the main food source for Tibetan people (Bonoli et al., 2004; Siebenhandl et al., 2007; Kohyama et al., 2008; Zhao et al., 2015). It grows on the Qinghai–Tibet Plateau and is the only crop that can grow at high altitude of 4,200–4,500 m (Dai et al., 2012; Zhong et al., 2016). However, at present, with the aggravation of drought on the Qinghai–Tibet Plateau (Meng et al., 2017), the selection of drought-tolerant hulless barley strains has become an urgent affair. However, the genetic resource that could be used to assist hulless barley molecular breeding was scarce. In this study, GWAS was used to map SNP markers related to drought tolerance in hulless barley. The SNP markers identified in this study will be used to analyze drought tolerance of hulless barley and facilitate the selection of drought-tolerant strains.

In this study, 269 lines of hulless barley were selected for drought treatment under laboratory and field conditions. Significant phenotypic variation in effective spike number of plants (ESNP), main spike number (MEN), main spike length (MSL), root and shoot dry weight (RSWD), and thousand grain weight (TGW) have been identified under drought conditions. These results indicated that the selected lines could play an important role in exploring the drought-tolerance genes of hulless barley. Hulless barley has great plasticity in adapting to drought stress, which will provide reference for the breeding process of superior hulless barley strains and improve the drought tolerance of hulless barley.

Using the hulless barley GWAS panel, 29 SNPs loci and five candidate genes connected with all spike traits (including ESNP MEN and MSL) were identified. As for ENSP, markers distributed on chromosomes 7H and chromosomes 1H were correlated, a total of 4 SNPs loci on three genes (*EMB506*, *DCR*, and *APD2*) were identified. *ABCG11* and *CLPR2* are two effector genes for MEN and MSL traits, respectively. Among the genes related to spike traits, *DCR* and *ABCG11* plays a key role in cuticle formation (Panikashvili et al., 2010; Rani et al., 2010). As the contact zone between the plant and the environment, cuticle has been well-characterized for its multiple roles in the regulation of gas exchange, epidermal permeability, and non-stomatal water loss (Sieber et al., 2000). So, it is not surprising that *dcr* mutants show increased water loss and increased sensitivity to drought conditions (Panikashvili et al., 2009). Besides, *EMB506* and *CLPR2* genes are associated with chloroplast and plant growth (Despres et al., 2001; Rudella et al., 2006). The lack of *CLPR2* gene causes leaf albinism and undoubtedly affects photosynthetic efficiency and crop yield (Kim et al., 2009). For RSWD at the dehydrated growth condition, *YIP4B* gene represented by chr1H_64014764 SNP was identified. In *Arabidopsis thaliana*, *YIP4B* affects root and hypocotyl growth through elongation rather than cell division (Gendre et al., 2013). Two genes located on Chr3H were identified for TGW trait. Among them, *GLYK* catalyzes the termination of the C2 cycle in photosynthesis, which is an indispensable auxiliary metabolic pathway for the C3 cycle of photosynthesis. The presence of this gene ensures the normal growth of terrestrial plants in an oxygen-containing atmosphere and avoids photoinhibition (Boldt et al., 2005). As iron sensors, *BTS* gene plays a vital role in modulating iron homeostasis (Zhang et al., 2015).

The mutation of these loci under drought conditions and the resulting phenotypic changes undisputedly gives us huge inspiration. Further development of these SNPs and genes will provide new insights into improving crop phenotypic traits and make plants develop in an environment-adapted direction. For instance, drought-tolerant, higher-yielding plants could be created by genetically modifying these loci. These findings will simplify the tedious process of hybridization and culture, and turn to use molecular methods for seedling breeding, which will reduce our experimental time greatly. What is more exciting is that it also provides direction for drought-tolerance selection of other economic crops besides hulless barley.

In summary, we used 5,532,468 SNP markers from 269 hulless barley lines to analyze the association between phenotypic values and genotypes of drought-tolerance traits in this study. The SNP markers association with spike traits (chr7H_102929775, chr7H_102929728, chr1H_512690373, and chr1H_512690373, chr1H_349621827, chr1H_349622062, chr2H_39562071, chr2H_47246481, chr2H_47623192, chr2H_47623303, chr2H_48148956, chr2H_48534579, chr2H_48534763, chr2H_48796138, chr2H_496935399, chr2H_496935424, chr2H_54436239, chr2H_54436346, chr2H_55411310, chr2H_55768675, chr2H_57397060, chr2H_57539586, chr2H_62893696, chr2H_64170225, chr7H_149587366, chr7H_158720411, chr7H_621337028, chr3H_152206655, and chr5H_250095923), RSWD trait (chr1H_64014764), and TGW trait (chr3H_482958549, chr3H_489630701) were identified. Under drought conditions, the mutation of these SNPs loci possibly lead to phenotypic changes and improve the adaptation of hulless barley to drought environment. In conclusion, the SNPs identified in this study can be used in drought-tolerance gene analysis, and can provide valuable information for further improvement of crop yield, quality, and adaptability.

DATA AVAILABILITY STATEMENT

All the raw genome data can be found in the National Genomics Data Center (NGDC) BioProject database with the accession number PRJCA009869.

AUTHOR CONTRIBUTIONS

ZF and KW contributed to the conceptualization and methodology. JL contributed to the experimentation, data analysis, and first draft preparation. XY, YY, JL, and LA contributed to the experimentation and data analysis. JL, ZF, and KW contributed to the supervision and manuscript editing. All authors have read and agreed to the published version of the manuscript.

FUNDING

This work was supported by funds from the National Key R & D Program of China (2018YFD1000705 and 2018YFD1000700), National Natural science Foundation of China (32060480), the China Agriculture Research System (CARS-05), the International Science and Technology Cooperation in Sichuan Province of China (Grant No. 2021YFH0113), Qinghai Provincial Academy of Agriculture and Forestry Innovation Fund (2018-NKY-12), and the Double Support Program for Discipline Construction of Sichuan Agricultural University in China.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.924892/full#supplementary-material

REFERENCES

Allen, G. C., Flores-Vergara, M., Krasynanski, S., Kumar, S., and Thompson, W. (2006). A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. *Nat. Protoc.* 1, 2320–2325. doi: 10.1038/nprot.2006.384

Boldt, R., Edner, C., Kolukisaoglu, U., Hagemann, M., Weckwerth, W., Wienkoop, S., et al. (2005). D-GLYCERATE 3-KINASE, the last unknown enzyme in the photorespiratory cycle in Arabidopsis, belongs to a novel kinase family. *Plant Cell.* 17, 2413–2420. doi: 10.1105/tpc.105.033993

Bonoli, M., Verardo, V., Marconi, E., and Caboni, M. F. (2004). Antioxidant phenols in barley (*Hordeum vulgare* L.) flour: comparative spectrophotometric study among extraction methods of free and bound phenolic compounds. *J. Agric. Food Chem.* 52, 5195–5200. doi: 10.1021/jf040075c

Chen, Z. L., Meng, J. M., Cao, Y., Yin, J. L., Fang, R. Q., Fan, S. B., et al. (2019). A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides. *Nat. Commun.* 10:3404. doi: 10.1038/s41467-019-11337-z

Dai, A. (2013). Increasing drought under global warming in observations and models. *Nat. Clim. Chang.* 3, 52–58. doi: 10.1038/nclimate1633

Dai, F., Nevo, E., Wu, D., Comadran, J., Zhou, M., Qiu, L., et al. (2012). Tibet is one of the centers of domestication of cultivated barley. *Proc. Natl. Acad. Sci. U.S.A.* 109, 16969–16973. doi: 10.1073/pnas.1215265109

Despres, B., Delseny, M., and Devic, M. (2001). Partial complementation of embryo defective mutations: a general strategy to elucidate gene function. *Plant J.* 27, 149–159. doi: 10.1046/j.1365-313x.2001.01078.x

Gendre, D., McFarlane, H. E., Johnson, E., Mouille, G., Sjödin, A., Oh, J., et al. (2013). Trans-Golgi network localized ECHIDNA/Ypt interacting protein complex is required for the secretion of cell wall polysaccharides in Arabidopsis. *Plant Cell.* 25, 2633–2646. doi: 10.1105/tpc.113.112482

Gong, Z., Xiong, L., Shi, H., Yang, S., Herrera-Estrella, L., Xu, G., et al. (2020). Plant abiotic stress response and nutrient use efficiency. *Sci. China Life Sci.* 63, 635–674. doi: 10.1007/s11427-020-1683-x

Gupta, A., Rico-Medina, A., and Caño-Delgado, A. I. (2020). The physiology of plant responses to drought. *Science* 368, 266–269. doi: 10.1126/science.aaz7614

He, T., and Jia, J. F. (2008). High frequency plant regeneration from mature embryo explants of highland barley (*Hordeum vulgare* L. var. nudum Hk. f.) under endosperm-supported culture. *Plant Cell Tissue Organ Cult.* 95, 251–254. doi: 10.1007/s11240-008-9437-2

Kim, J., Rudella, A., Ramirez Rodriguez, V., Zybailov, B., Olinares, P. D., and van Wijk, K. J. (2009). Subunits of the plastid ClpPR protease complex have differential contributions to embryogenesis, plastid biogenesis, and plant development in Arabidopsis. *Plant Cell* 21, 1669–1692. doi: 10.1105/tpc.108.063784

Kohyama, N., Ono, H., and Yanagisawa, T. (2008). Changes in anthocyanins in the grains of purple waxy hull-less barley during seed maturation and after harvest. *J. Agric. Food Chem.* 56, 5770–5774. doi: 10.1021/jf800626b

Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 9:29. doi: 10.1186/1746-4811-9-29

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, 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, 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, W. (2013). "Genome-wide association study," in *Encyclopedia of Systems Biology*, eds W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota (New York, NY: Springer New York), 834.

Liu, H., Chen, X., Zhang, D., Wang, J., Wang, S., and Sun, B. (2018). Effects of highland barley bran extract rich in phenolic acids on the formation of N(ε)-carboxymethyllysine in a biscuit model. *J. Agric. Food Chem.* 66, 1916–1922. doi: 10.1021/acs.jafc.7b04957

Luo, G., Gu, H., Liu, J., and Qu, L. J. (2012). Four closely-related RING-type E3 ligases, APD1-4, are involved in pollen mitosis II regulation in Arabidopsis. *J. Integr. Plant Biol.* 54, 814–827. doi: 10.1111/j.1744-7909.2012.01152.x

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

Meng, L. H., Yang, J., Guo, W., Tian, B., Chen, G. J., Yang, Y. P., et al. (2017). Differentiation in drought tolerance mirrors the geographic distributions of alpine plants on the Qinghai-Tibet Plateau and adjacent highlands. *Sci. Rep.* 7:42466. doi: 10.1038/srep42466

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

Panikashvili, D., Shi, J. X., Bocobza, S., Franke, R. B., Schreiber, L., and Aharoni, A. (2010). The Arabidopsis DSO/ABCG11 transporter affects cutin metabolism in reproductive organs and suberin in roots. *Mol. Plant.* 3, 563–575. doi: 10.1093/mp/ssp103

Panikashvili, D., Shi, J. X., Schreiber, L., and Aharoni, A. (2009). The Arabidopsis DCR encoding a soluble BAHD acyltransferase is required for cutin polyester formation and seed hydration properties. *Plant Physiol.* 151, 1773–1789. doi: 10.1104/pp.109.143388

Rani, S. H., Krishna, T. H., Saha, S., Negi, A. S., and Rajasekharan, R. (2010). Defective in cuticular ridges (DCR) of Arabidopsis thaliana, a gene associated with surface cutin formation, encodes a soluble diacylglycerol acyltransferase. *J. Biol. Chem.* 285, 38337–38347. doi: 10.1074/jbc.M110.133116

Rudella, A., Friso, G., Alonso, J. M., Ecker, J. R., and van Wijk, K. J. (2006). Downregulation of ClpR2 leads to reduced accumulation of the ClpPRS protease complex and defects in chloroplast biogenesis in Arabidopsis. *Plant Cell.* 18, 1704–1721. doi: 10.1105/tpc.106.042861

Schewe, J., Heinke, J., Gerten, D., Haddeland, I., Arnell, N. W., Clark, D. B., et al. (2014). Multimodel assessment of water scarcity under climate change. *Proc. Natl. Acad. Sci. U.S.A.* 111, 3245–3250. doi: 10.1073/pnas.1222460110

Siebenhandl, S., Grausgruber, H., Pellegrini, N., Del Rio, D., Fogliano, V., Pernice, R., et al. (2007). Phytochemical profile of main antioxidants in different fractions of purple and blue wheat, black barley. *J. Agric. Food Chem.* 55, 8541–8547. doi: 10.1021/jf072021j

Sieber, P., Schorderet, M., Ryser, U., Buchala, A., Kolattukudy, P., Métraux, J. P., et al. (2000). Transgenic Arabidopsis plants expressing a fungal cutinase show alterations in the structure and properties of the cuticle and postgenital organ fusions. *Plant Cell.* 12, 721–738. doi: 10.1105/tpc.12.5.721

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* 20, 467–484. doi: 10.1038/s41576-019-0127-1

Xu, Y., Li, P., Yang, Z., and Xu, C. (2017). Genetic mapping of quantitative trait loci in crops. *Crop J.* 5, 175–184. doi: 10.1016/j.cj.2016.06.003

Yang, L., Christensen, D., McKinnon, J., Beattie, A., and Yu, P. (2013). Effect of altered carbohydrate traits in hulless barley (*Hordeum vulgare* L.) on nutrient profiles and availability and nitrogen to energy synchronization. *J. Cereal Sci.* 58, 182–190. doi: 10.1016/j.jcs.2013.05.005

Yu, L., Zhao, X., Gao, X., Jia, R., and Siddique, K. (2021). Effect of natural factors and management practices on agricultural water use efficiency under drought: a meta-analysis of global drylands. *J. Hydrol.* 594:125977. doi: 10.1016/j.jhydrol.2021.125977

Zhang, J., Liu, B., Li, M., Feng, D., Jin, H., Wang, P., et al. (2015). The bHLH transcription factor bHLH104 interacts with IAA-LEUCINE RESISTANT3 and modulates iron homeostasis in Arabidopsis. *Plant Cell* 27, 787–805. doi: 10.1105/tpc.114.132704

Zhao, C., Wang, X., Wang, X., Wu, K., Li, P., Chang, N., et al. (2015). Glucose-6-phosphate dehydrogenase and alternative oxidase are involved in the cross tolerance of highland barley to salt stress and UV-B radiation. *J. Plant Physiol.* 181, 83–95. doi: 10.1016/j.jplph.2015.03.016

Zhong, Z.-M., Shen, Z.-X., and Fu, G. (2016). Response of soil respiration to experimental warming in a highland barley of the Tibet. *Springerplus* 5:137. doi: 10.1186/s40064-016-1761-0

Zhu, F., Du, B., and Xu, B. (2015). Superfine grinding improves functional properties and antioxidant capacities of bran dietary fibre from Qingke (hull-less barley) grown in Qinghai-Tibet Plateau, China. *J. Cereal Sci.* 65, 43–47. doi: 10.1016/j.jcs.2015.06.006

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

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

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

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 27 June 2022
doi: 10.3389/fpls.2022.932947

Cassava (*Manihot esculenta*) Slow Anion Channel (*MeSLAH4*) Gene Overexpression Enhances Nitrogen Assimilation, Growth, and Yield in Rice

Linhu Song[1,2†], Xingmei Wang[1†], Liangping Zou[3†], Zakaria Prodhan[2†], Jiaheng Yang[1], Jianping Yang[1], Li Ji[1,4], Guanhui Li[1], Runcong Zhang[1], Changyu Wang[1], Shi Li[1], Yan Zhang[1], Xiang Ji[1], Xu Zheng[1], Wanchen Li[4]* and Zhiyong Zhang[2]**

[1]State Key Laboratory of Wheat and Maize Crop Science and Center for Crop Genome Engineering, College of Agronomy, Henan Agricultural University, Zhengzhou, China, [2]College of Life Sciences, Neijiang Normal University, Neijiang, China, [3]Institute of Tropical Bioscience and Biotechnology, Chinese Academy of Tropical Agricultural Sciences, Haikou, China, [4]Key Laboratory of Biology and Genetic Improvement of Maize in Southwest Region, Ministry of Agriculture, Maize Research Institute, Sichuan Agricultural University, Chengdu, China

Nitrogen is one of the most important nutrient elements required for plant growth and development, which is also immensely related to the efficient use of nitrogen by crop plants. Therefore, plants evolved sophisticated mechanisms and anion channels to extract inorganic nitrogen (nitrate) from the soil or nutrient solutions, assimilate, and recycle the organic nitrogen. Hence, developing crop plants with a greater capability of using nitrogen efficiently is the fundamental research objective for attaining better agricultural productivity and environmental sustainability. In this context, an in-depth investigation has been conducted into the cassava slow type anion channels (*SLAHs*) gene family, including genome-wide expression analysis, phylogenetic relationships with other related organisms, chromosome localization, and functional analysis. A potential and nitrogen-responsive gene of cassava (*MeSLAH4*) was identified and selected for overexpression (OE) analysis in rice, which increased the grain yield and root growth related performance. The morpho-physiological response of OE lines was better under low nitrogen (0.01 mm NH_4NO_3) conditions compared to the wild type (WT) and OE lines under normal nitrogen (0.5 mm NH_4NO_3) conditions. The relative expression of the *MeSLAH4* gene was higher (about 80-fold) in the OE line than in the wild type. The accumulation and flux assay showed higher accumulation of NO_3^- and more expansion of root cells and grain dimension of OE lines compared to the wild type plants. The results of this experiment demonstrated that the *MeSLAH4* gene may play a vital role in enhancing the efficient use of nitrogen in rice, which could be utilized for high-yielding crop production.

Keywords: cassava, slow anion channel, transgenic rice, nitrogen use efficiency, root phenotype

OPEN ACCESS

Edited by:
Hai Du,
Southwest University, China

Reviewed by:
Chaowen Xiao,
Sichuan University, China
Zhiguo Zhang,
Biotechnology Research Institute (CAAS), China

****Correspondence:***
Xu Zheng
zhengxu@henau.edu.cn
Wanchen Li
aumdyms@sicau.edu.cn
Zhiyong Zhang
10001080@njtc.edu.cn

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

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

Received: *30 April 2022*
Accepted: *08 June 2022*
Published: *27 June 2022*

Citation:
Song L, Wang X, Zou L, Prodhan Z, Yang J, Yang J, Ji L, Li G, Zhang R, Wang C, Li S, Zhang Y, Ji X, Zheng X, Li W and Zhang Z (2022) Cassava (Manihot esculenta) Slow Anion Channel (MeSLAH4) Gene Overexpression Enhances Nitrogen Assimilation, Growth, and Yield in Rice. Front. Plant Sci. 13:932947. doi: 10.3389/fpls.2022.932947

INTRODUCTION

Nitrogen (N) is one of the most important macronutrients and plays a significant role in the photosynthesis, growth, development, and reproduction of plants. Nitrogen is an essential component element of many enzymes that control and direct many biochemical reactions in plants (Kraiser et al., 2011; Li et al., 2020). Therefore, N availability and utilization are the key factors for adequate biomass accumulation, proper crop growth, yield, and productivity. Plants can absorb N from various sources in the form of organic nitrogen compounds, nitrate (NO_3^-) and ammonium (NH_4^+; Vidal et al., 2020). The efficient use of N by crop plants involves several steps, including uptake, assimilation, translocation, and, when the plant ages, recycling and remobilization (Masclaux-Daubresse et al., 2010). Furthermore, N plays an important signaling role in plant growth and metabolism, such as breaking seed dormancy, controlling lateral root and leaf development, regulating blooming time, and activating associated genes (Ho and Tsay, 2010; Hachiya and Sakakibara, 2017).

Plants have evolved a sophisticated mechanism for up-taking, allocation, and storage of inorganic and organic N, including low-affinity transport systems (LATS), which operate at high nutrient concentrations (> 1 mm), and high-affinity transport systems (HATS) that predominate in the micromolar range (Wang et al., 1993; Kraiser et al., 2011). There are four gene families, including *Nitrate Transporter 1*/*Peptide Transporter* (*NRT1*/*PTR*, *NPF*), *NRT2*, *Chloride Channel* (*CLC*), and *Slow Anion Channel* (*SLAC1*/*SLAH*), involved in the nitrate transport system (Krapp et al., 2014; O'Brien et al., 2016). Among these subfamilies, the *SLAC*/*SLAH* members play an important role in anion transport, stress signaling, growth and development, and hormonal response in plants (Vahisalu et al., 2008; Nan et al., 2021).

Slow anion channel proteins, particularly those implicated in nitrate absorption and transport, are the subject of an increasing amount of research. Members of the *SLAC*/*SLAH* family have been found and researched in a variety of plants, including *Arabidopsis* (Zheng et al., 2015), rice (Sun et al., 2016), maize (Qi et al., 2018), barley (Liu et al., 2014), tobacco (Kurusu et al., 2013), poplar (Jaborsky et al., 2016), pear (Chen et al., 2019a), and *Brassica napus* (Nan et al., 2021). A total of five *SLAC*/*SLAH* genes were identified in *Arabidopsis* (Vahisalu et al., 2008), 23 genes in *B. napus* (Nan et al., 2021), and 9 genes in rice (Sun et al., 2016). Differential expression and assembly of *SLAH1*/*SLAH3* anion channel subunits are used by plants to regulate the transport of NO_3^- and Cl^- between the root and shoot, where the *AtSLAH1* gene is co-localized with *AtSLAH3* (Cubero-Font et al., 2016). The *SLAC1* anion channel, as well as its homologs, *SLAH3* and *SLAH2*, have been functionally characterized in *Arabidopsis* and *Xenopus* oocytes (Negi et al., 2008; Maierhofer et al., 2014a). The *SLAC1* gene is mostly found in guard cells and is phosphorylated by the *Open stomata 1* (*OST1*) kinase, causing anion efflux from guard cells, which mediates stomatal closure and increases drought tolerance (Vahisalu et al., 2010; Geiger et al., 2011). The interaction of *AtSLAC1* and *AtSLAH3* with several kinase phosphatases is linked to water stress signals (Brandt et al., 2012). The SLAH3 protein is phosphorylated by calcium-dependent protein kinases such as CPK2 and CPK20, which regulate pollen tube formation *via* regulating *SLAC*/*SLAH* expression in *Arabidopsis* (Gutermuth et al., 2013). The *SLAH2* gene, which is the most similar protein to *SLAH3*, absorbs only nitrate, unlike other *SLAC*/*SLAH* members, which absorb both nitrate and chloride (Maierhofer et al., 2014b). The *PbrSLAH3* gene is localized in the plasma membrane without expression in flowers, and has a strong selective absorption for nitrate and no permeability to chlorine (Chen et al., 2019b). Nevertheless, the *PttSLAH3* gene of poplar is not activated by protein kinase phosphorylation to absorb nitrate and chloride ions (Jaborsky et al., 2016). The *AtSLAH4* gene, which is phylogenetically linked to *AtSLAH1*, has a similar expression pattern as *AtSLAH3*, but is greater toward the root tip (Zheng et al., 2015). However, research on the *SLAH4* gene is very limited and molecular, physiological, and functional studies have not been carried out completely, which makes *SLAH4* a promising candidate gene for plant genetic engineering and biotechnological investigations.

Cassava is a short-day dicot plant in the Euphorbiaceae family that is used as a food crop as well as a possible biofuel crop (Drunkler et al., 2012). Cassava is a durable and easy-to-plant tropical commercial crop with a high degree of adaptability, and it may produce a huge yield in dry and barren mountainous and hilly areas (Wang, 2002). Cassava can obtain sufficient nitrogen from the soil to fulfill its own growth and development requirements without requiring excessive nitrogen fertilizer throughout the growing phase (Jiang et al., 2016). Therefore, it is crucial to identify the key genes involved in cassava's nitrogen-efficient utilization for the improvement and production of nitrogen-efficient germplasm resources through genetic modification in other crops, particularly in rice.

In this study, six *SLAH* genes were identified in the cassava genome and their phylogenetic relationships, chromosomal localization, and morpho-physiological characteristics have been analyzed. A potential candidate gene, *MeSLAH4*, which was localized in the plasma membrane and in the nucleus, highly expressed in the roots under low nitrate conditions, was selected for overexpression analysis. Furthermore, several parameters related to plant growth, development, and yield were evaluated to demonstrate the role of this gene in improving nitrogen use efficiently in rice. The results of these experiments suggested that overexpression of *MeSLAH4* could increase plant growth, grain dimension, root systems indices (root morphology), and yield in rice. The findings of this research would shed new light on the possibility of a genetic engineering approach of key candidate genes in nitrogen uptake and utilization.

MATERIALS AND METHODS

Plant Materials and Growth Conditions

The cassava (*Manihot esculenta*) variety ("Huanan5") was chosen as the wild type during this experiment. Cassava seedings were grown on half-strength Murashige and Skoog

(MS) medium at 26°C with a 16 h light and 8 h dark cycle and 70% relative humidity conditions in a growth chamber. The tobacco (*Nicotiana benthamiana*) plants were also grown in a greenhouse under cycles of 16 h light and 8 h dark at 25°C. The rice (*Oryza sativa*) seedlings were grown in controlled conditions of 16 h light (30°C) and 8 h dark (28°C) photoperiods with ~70% relative humidity. For the overexpression experiment, the coding sequences of the *MeSLAH4* gene were amplified and inserted into pCAMBIA2300-35S-eGFPvector. These constructs were subsequently transferred into an *Agrobacterium* strain (EHA105) for rice transformation. For the hydroponic culture, the plants were grown in a nutrient solution containing 1.5 mm NH_4NO_3, 0.3 mm NaH_2PO_4, 0.3 mm K_2SO_4, 1.0 mm $CaCl_2$, 1.6 mm $MgSO_4$, 0.5 mm Na_2SiO_3, 20 µm Fe-EDTA, 18.9 µm H_3BO_3, 9.5 µm $MnCl_2$, 0.1 µm $CuSO_4$, 0.2 µm $ZnSO_4$, and 0.39 mm Na_2MoO_4, but supplied with different N concentrations, termed as normal nitrogen NN (0.5 mm NH_4NO_3) and low nitrogen LN (0.01 mm NH_4NO_3), pH 5.5. The nutrient solution was changed every 3 days.

Phylogenetic Analysis and Expression Pattern of Cassava *SLAH* Genes

The protein sequences of *SLAH* genes (gene name; locus identifiers) of *Arabidopsis*, including *AtSLAH1* (*AT1G12480*), *AtSLAH1* (*AT1G62280*), *AtSLAH2* (*AT4G27970*), *AtSLAH3* (*AT5G24030*) and *AtSLAH4* (*AT1G62262*), were downloaded from the TAIR database[1] (Swarbreck et al., 2007). The SLAH protein sequences of genes in rice, such as *Os01g0623200* (*LOC_Os01g43460*), *Os01g0385400* (*LOC_Os01g28840*), *Os05g0219900* (*LOC_Os05g13320*), *Os07g0181100* (*LOC_Os07g08350*), *Os01g0226600* (*LOC_Os01g12680*), *Os04g0574700* (*LOC_Os04g48530*), *Os01g0247700* (*LOC_Os01g14520*), *Os05g0269200* (*LOC_Os05g18670*), and *Os05g0584900* (*LOC_Os05g50770*) were downloaded from the RAP_DB database.[2] The SLAH protein sequence of *M. esculenta* (Cassava), including *MANES_05G153100* (SLAH4), *MANES_11G124900* (SLAC1 homolog 1), *MANES_14G020300* (SLAC1 homolog 3), *MANES_06G154500* (SLAC1 homolog 3), *MANES_06G154600* (SLAC1 homolog 3), and *MANES_S089100*, along with their homologues in other species such as *Zea mays* (Corn), *Triticum aestivum* (Wheat), *B. napus* (Rapeseed), *Selaginella moellendorffii* (Spikemoss) etc., were downloaded from the Phytozome database[3] (Nan et al., 2021). The phylogenetic trees were constructed based on the SLAC/SLAH protein sequences by IQ-TREE using the Maximum Likelihood (ML) method with 1,000 replicates of bootstrap alignments. RNA-seq data and differential gene expression information were obtained from a published database.[4]

Chromosomal Localization Analysis

The chromosomal localization information of the *SLAC*/*SLAH* genes was obtained from sequences of the cassava genome, and the MG2C[5] was used to draw the chromosomal distribution of *MeSLAH* genes.

Gene Structure and Conserved Motifs Analysis

The structures of the *SLAH* genes were analyzed using the Gene Structure Display Server (GSDS 2.0)[6] by aligning the cDNA sequences with their corresponding genomic DNA sequences. Conserved motifs of the SLAH proteins were identified using the online Multiple Expectation Maximization for Motif Elicitation (MEME[7]; Bailey et al., 2006). All obtained SLAH protein sequences were analyzed against the Pfam database to verify the presence of SLAC1 domains (**Supplementary Figures 2**, **3**). The SLAC1 domain was also detected by the SMART program (SMART).[8] Protein sequences lacking the SLAC1 domain or having E-values of more than 1 e-6 were removed.

Subcellular Localization of MeSLAH4 Protein

Rice protoplasts were isolated for transient transformation of the *MeSLAH4* gene (Zhang et al., 2011; Burman et al., 2020). The open reading frame (ORF) of the *MeSLAH4* gene was amplified by PCR (95°C for 5 min, then 35 cycles of 95°C for 30 s, 58°C for 30 s, and 72°C for 70 s, with a final extension at 72°C for 5 min). The PCR products (**Supplementary Figure 4**) were cloned into the 35S-eGFP vector in between the XhoI and HindIII sites and under the control of the cauliflower mosaic virus 35S promoter. The competent cells of *Escherichia coli* (DH5α) and *Agrobacterium* (LBA4404) were used for the transformation of recombinants. Listed primers (**Supplementary Table 2**) for gene cloning and vector construction and plasma membrane marker (35 s-ZmCDPK7-MCHERRY and 35S-HY5-Mcherry; Zhao et al., 2021) were used in this study. The MeSLAH4-eGFP and 35 s-ZmCDPK7-MCHERRY fusion constructs were transiently expressed in rice protoplasts using a polyethylene glycol calcium-mediated method. An empty 35S-GFP vector was served as a negative control. Transfected protoplasts were observed after 16 h of incubation by a confocal laser scanning microscope (Olympus FV3000, Tokyo, Japan).

Measurements of Morpho-Physiological Traits in Rice

The morphological, physiological, and agronomic traits of each transgenic rice line (OE) and wild type (WT) was measured at the 2-week-seedling stage and the maturity stage. The morphological characters, including seedlings height (cm) and root length (cm), were measured using a ruler or meter stick. The grains were lined up lengthwise and widthwise along a ruler to measure grain length (mm) and grain breadth (mm), respectively, and the measurements were confirmed using an

[1]http://www.arabidopsis.org/
[2]https://rapdb.dna.affrc.go.jp/
[3]http://phytozome.jgi.doe.gov/pz/portal.html
[4]https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc
[5]http://mg2c.iask.in/mg2c_v2.1/
[6]http://gsds.cbi.pku.edu.cn/
[7]http://meme.nbcr.net/meme/cgibin/meme.cgi
[8]http://smart.emblheidelberg.de/

MRS-9600TFU2L (MICROTEK) grain observation instrument. The shoot weight (g. plant^{-1} FW), root weight (g. plant^{-1} FW), and grain yield of a single spikelet (g) were measured using a weighing balance. The root fork numbers (number of branches), root tip numbers, and grain numbers of a single spikelet were calculated with the eye and confirmed by capturing an image with an Epson Expression 10000XL (Epson, Japan) and counting with winRHIZO software (Li et al., 2016a). The root average diagram (mm), root surface area (cm^2), root volume (cm^3), and grain diameter (mm) measurements were performed using a microscope (MVX10, Olympus), and the winRHIZO scanner-based image analysis system (Regent Instruments, Montreal, QC, Canada; Sun et al., 2014). Total grain protein content (%), and grain moisture content (%) were detected using an XDS Near-Infrared Rapid Content Analyzer (Foss® Analytical, Hilleroed, Denmark; Li et al., 2014).

The chlorophyll was extracted from 0.15 g of fresh leaves at the booting stage using 95.0% ethanol. Briefly, leaves were cut into 3 mm pieces and immersed in 95.0% ethanol for 24 h in the dark at 26°C. The absorbance of the extract was measured using a spectrophotometer at A665 and A649. The chlorophyll-a content (mg. g^{-1} FW), chlorophyll-b content (mg. g^{-1} FW), and total chlorophyll content (mg. g^{-1} FW) contents were determined by the method reported by Arnon (1949). A total of 10 individuals of each transgenic line and wild type plants were assayed (Li et al., 2016b).

The activities of catalase (CAT), peroxidase (POD), and superoxide dismutase (SOD) were measured by employing 0.5 g of seedlings in 5 ml of extraction buffer containing 0.05 M phosphate buffer (Li et al., 2018). The CAT activity was determined spectrophotometrically based on the decrease in absorbance of H_2O_2 (extinction coefficient of 43.6 M^{-1} cm^{-1}) at 240 nm for 1 min (Aebi, 1984). The POD was measured as the absorbance at 470 nm. The SOD activity was assayed by measuring the ability of the enzyme extract to inhibit the photochemical reduction of nitroblue tetrazolium (NBT; Yoshimura et al., 2000). Phenotypic measurements of the positive transgenic plants were undertaken using three independent lines at least (Li et al., 2014).

Gene Expression Analysis Using Quantitative Real-Time PCR

Total RNA was isolated from different tissues (leaves, stems, and roots) of cassava and rice using an RNA kit (TRNzol universal reagent, TIANGEN Biotech, Beijing, China) according to the manufacturer's instructions. The RNA was then reverse transcribed into cDNA using the oligo (dT) primers and ImProm-II reverse transcriptase (Promega). The specific primers for the *MeSLAH4* genes along with the housekeeping Actin genes were designed using the Primer Premier 5.0 software (**Supplementary Table 2**). The real-time PCR reactions were conducted with 2 µl of diluted cDNA, 200 nM of each primer, 2× SYBER GREEN Master Mix (Green I Master Mix, Roche) in a final volume of 20 µl of double sterile water. The thermal cycle conditions were pre-incubation at 95°C for 5 min, then 40 cycles of 95°C for 3 s, 60°C for 10 s, and 72°C for 30 s, with a final extension at 72°C for 3 min in the Light Cycler 480 (Roche, United States). The gene expression levels were calculated with the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001), and the qRT-PCR assays were performed with three biological and three technical replicates.

Nitrogen Accumulation Analysis

Fresh samples (whole plant, grain, and glume) of WT or transgenic lines were harvested at the rice mature stage ($n = 4$) and heated at 105°C for 30 min. The samples were then dried for 3 days at 75°C. Dry weights were recorded as biomass values. Total N accumulation was assessed using the Kjeldahl method in the different plant samples by multiplying the N concentration with the corresponding biomass.

Statistical Analysis

The data from the experiments were analyzed by one-way ANOVA, Duncan's multiple range test, and Tukey's test at $p < 0.05$ to determine the statistically significant differences among different treatments. All the statistical evaluations were performed using SPSS version 20.0 statistical software (SPSS Inc., Chicago, IL, United States) and MS-Office 2019 software.

RESULTS

Phylogenetic and Expression Pattern Analysis of SLAH Genes

In the present study, a close phylogenetic relationship of entire *SLAH* genes in cassava with previously reported *SLAH* genes in another species was detected (**Figure 1A**) by analyzing with a Hidden Markov Model (HMM) profile search along with a conserved model (SLAC1, PF03595) for the SLAH proteins. The phylogenetic tree displayed seven clades (I to VII), and a very close relationship among the six *SLAH* genes in the cassava genome, five *SLAH* genes in *Arabidopsis*, and nine *SLAH* genes in rice were identified. However, two *SLAC1* gene homologues (*Os05g0269200* and *Os01g0247700*) in rice and two *SLAC1* homologues (*AT1G62280*, *SLAH1*; and *AT1G62262*, *SLAH4*) in *Arabidopsis* were detected in the same clade (Clade II) as the *MeSLAH4* (*MANES05G153100*) gene. Moreover, the MeSLAH4 protein demonstrated about 49.40% similarity with the Os05g0269200 amino acid sequence (**Supplementary Table 1**; **Supplementary Figure 1**). Besides, the phylogenetic tree using IQ-TREE of SLAH protein also revealed an evolutionary relationship with other species, including *Chondrus crispus* (Carrageen Irish moss), *Physcornitrella patens* (Bryophyta), *Marchantia polymorpha* (Liverwort), *Amborella trichopoda*, *Z. mays* (Corn), *T. aestivum* (Wheat), *B. napus* (Rapeseed), and *Selaginella moellendorffii* (Spikemoss).

RNA-seq data which are available on the database represents diverse expression pattern of *SLAH* genes in cassava, and the *SLAH* genes are expressed differentially in different tissues of cassava under different nitrogen conditions (**Figure 1B**). In particular, the *MeSLAH4* (*MANES_05G153100*) gene is highly expressed in the root under free nitrate concentration

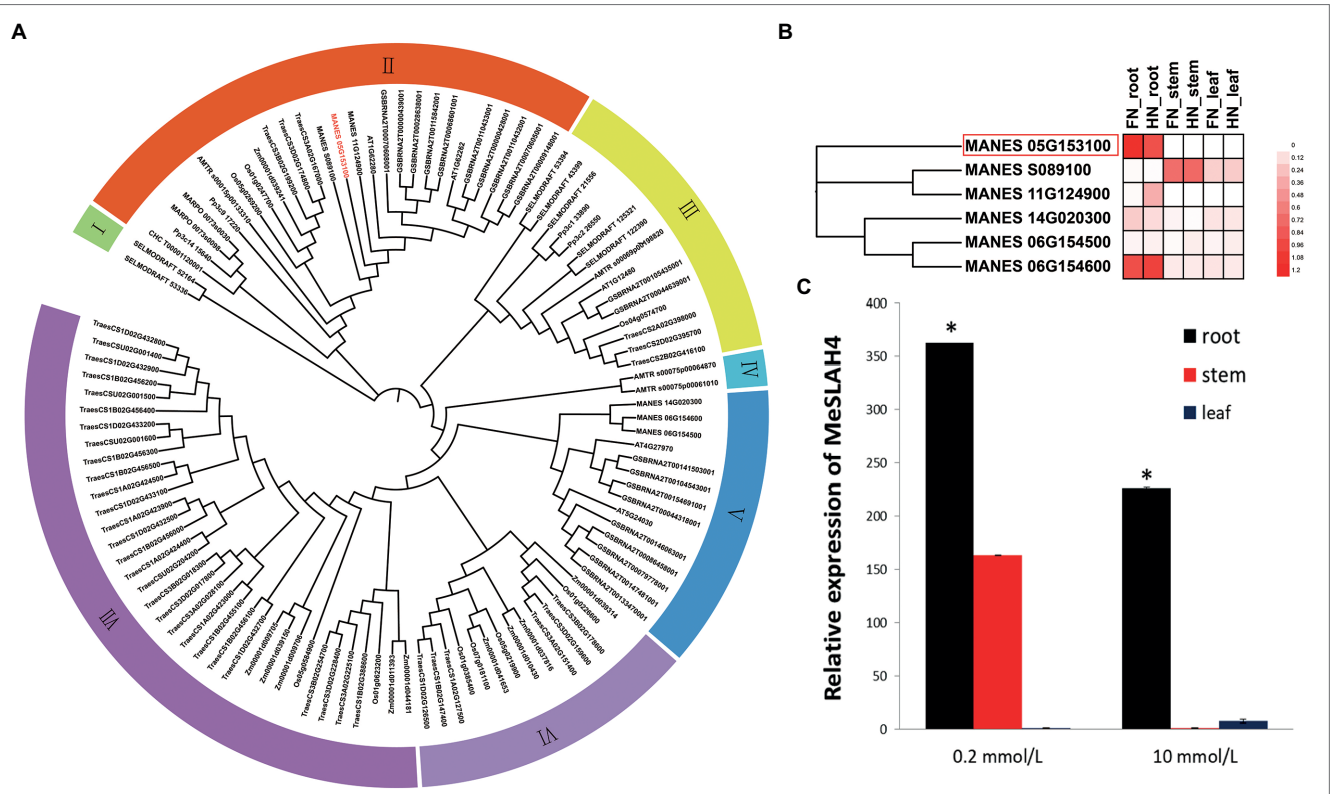

FIGURE 1 | Phylogenetic relationship and expression analysis of *SLAH* genes. **(A)** Phylogenetic tree of SLAH proteins constructed using ML method with 1,000 bootstrap replications in different species. The prefixes CHC, Pp3c, MARPO, AMTR, Os, Zm, Traes, MANES, AT, GSBRNA, and SELMODRAFT represent *Chondrus crispus*, *Physcornitrella patens*, *Marchantia polymorpha*, *Amborella trichopoda*, *Oryza sativa*, *Zea mays*, *Triticum aestivum*, *Manihot esculenta*, *Arabidopsis thaliana*, *Brassica napus*, and *Selaginella moellendorffii*. **(B)** Expression profiles of *SLAH* genes in the root, stem, and leaf of cassava under different nitrogen conditions. FN (0 mmol/l of nitrate nitrogen), and HN (10 mmol/l of nitrate nitrogen). **(C)** Relative expression of the *MeSLAH4* gene in cassava. The relative expression levels were obtained by normalization with the *MeActin* gene. The error bars indicate standard deviations, and the data are shown as mean values ± SD, while * represents significant differences at $p < 0.05$ on one-way ANOVA analysis.

(FN, 0 mmol/l, around 1.2-fold) as well as at high nitrate concentration (HN, 10 mm/l, approximately 1.08-fold). Another *MeSLAH1* homologue 3 (*MANES_06G154600*) is also expressed in the roots of cassava but a little bit lower (FN, about 1.08-fold, and HN, roughly 0.96-fold) than the *MeSLAH4* gene.

The relative expression pattern of the *MeSLAH4* gene in different tissues of cassava revealed variations in expression levels at various concentrations of nitrate levels (**Figure 1C**). The transcript accumulation patterns that were analyzed in roots, stems, and leaves indicated that the *MeSLAH4* gene was mainly expressed in the root. A significantly higher expression (about 370-fold) was observed in the root under a lower concentration of nitrate (0.2 mmol/l) treatment (**Figure 1C**). Thus, the *MeSLAH4* gene exhibited significantly higher relative gene expression levels at lower nitrate concentrations in the root, indicating a potential role in enhancing nitrogen use in plants.

As the root is the principal organ for nutrient uptake in plants, the up-regulation of the *MeSLAH4* gene could correlate the relationships among different nitrate concentrations with plant growth and development.

Chromosome Localization of *SLAH* Genes and Analysis of the Promoter Regions of *SLAH* Genes With *cis*-Acting Elements

The *SLAH* genes are distributed on four chromosomes in cassava (chromosomes 5, 6, 11, and chromosome 14). In cassava, one *SLAH* gene, which has been identified as *MeSLAH4* genes, was present on chromosome 5, two *SLAH* genes were located on chromosomes 6, one gene on chromosome 11, and one gene on chromosome 14 (**Figure 2A**). The *SLAH* genes in *Arabidopsis* are detected on three chromosomes (three genes on chromosome 1, one gene on chromosome 4, and one gene on chromosome 5). In rice, four *SLAH* genes were found on chromosome 1, one gene on chromosome 4, two genes on chromosome 5, and one gene on chromosome 7 (**Figure 2A**).

Analysis of the upstream promoter region of *SLAH* genes represented transcriptional regulation mechanisms. About 2 kb upstream of the initiation codon of *SLAH* genes of *Arabidopsis*, cassava, and rice were obtained and submitted to the Plant CARE database for investigating *cis*-regulatory elements. A total of 9 different *cis*-elements associated with light responsiveness, stress responsiveness, phytohormone responsiveness and growth regulation have been identified in upstream regions of *SLAH* genes (**Figure 2B**).

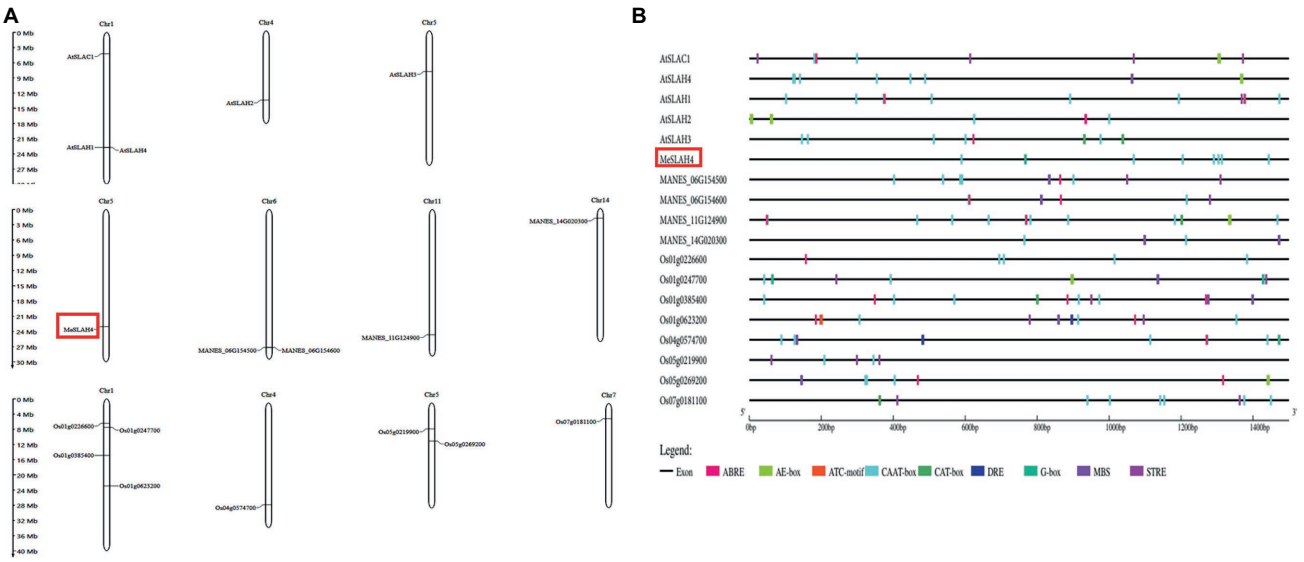

FIGURE 2 | Chromosomal localization and *cis*-elements analysis. **(A)** Chromosome localization of *SLAH* genes in *Arabidopsis thaliana* (At), *Manihot esculenta* (MANES), and *Oryza sativa* (Os). The approximate positions of the *SLAH* genes are presented on the respective chromosome size. **(B)** *Cis*-elements in the promoter regions (2.0-kb upstream regions) of the *SLAH* genes in *Arabidopsis*, cassava, and rice. Different colors represent various types of *cis*-elements, including core promoter elements, light responsive, phytohormone responsive, abiotic stress responsive and others related to growth. ABRE (ABA-responsive elements); E-box (enhancer box); CCAAT box/CAAT box/CAT box (GGCCAATCT consensus sequence); DRE (DNA Replication-related Element); G-box (G - box sequence, CACGTG); MBS (multichain binding site); STRE (stress-responsive elements).

A linear line has been constructed to present regulatory elements in each corresponding gene (**Figure 2B**). These results indicate that complex regulatory networks may be implicated in the transcriptional regulation of *SLAH* genes in different plants. *Cis*-regulatory elements, CAAT-box was commonly shared by all *SLAH* genes. G-box elements responding to light existed in the 2-kb upstream region of *SLAH* genes. Most *SLAH* genes contain ABRE elements (ABA responsive), but they are absent in the cassava *MeSLAH4* gene which suggested that this gene might not involve in regulation and physiological responses of various processes, including stomatal closure, seed and bud dormancy. Moreover, *SLAH* genes harbored drought responsive *cis*-elements (DRE) that were not present in the *MeSLAH4* gene. The *MeSLAH4* gene contains several copies of the CAAT-box and a copy of the G-box, indicating a higher transcription rate with sufficient quantities of suitable binding sites for several transcription factors as well as a highly conserved sequence for evolutionary process and epigenetic regulation.

Subcellular Localization of MeSLAH4 Protein

The transiently expressed fusion protein driven by the 35S promoter through protoplast transformation of the *MeSLAH4* gene represented a clear subcellular localization in rice (**Figure 3**). The green fluorescent signal of eGFP, which represented a negative control, was observed in the cytoplasm. However, the signal of the MeSLAH4-eGFP fusion protein, which coincides with the red fluorescent signal of plasma membrane-localized protein, was detected in the plasma membrane and in the nucleus. The protein localization was further confirmed by the protoplast transformation, which indicated that MeSLAH4 proteins are localized in the plasma membrane and in the nucleus. The microscopic visualization exhibited that the green fluorescence was distributed throughout the whole cell when the control (empty) vector was used. The green fluorescence was exclusively detected on the plasma membrane and nucleus by confocal microscopy when the vectors contained MeSLAH4 (**Figure 3**). These results indicate that the *MeSLAH4* gene may be involved in other functions in the plants.

Influence of *MeSLAH4* Overexpression on Morpho-Physiological Traits in Transgenic Rice

In the current experiment, the plant phenotype exhibited higher overall growth in OE lines under both nitrated concentrations (0.5 mm NH_4NO_3, NN, and 0.01 mm NH_4NO_3, LN) compared to the WT (**Figure 4A**). Plant height was significantly different ($p < 0.01$) under both nitrated concentrations (NN and LN) in both WT and OE lines, but they demonstrated higher plant height at LN compared to NN (**Figure 4B**). The shoot and root weight exhibited non-significant differences in both WT and OE lines under both nitrate concentrations (LN and NN; **Figures 4C,D**). However, shoot weight was higher (0.75 g. plant^{-1} FW) in OE lines at LN compared to both lines (WT and OE) in NN condition. Conversely, the root weight was higher in the WT (4.8 g. plant^{-1} FW) under NN than under LN. These results point out that the overexpression of the *MeSLAH4* gene

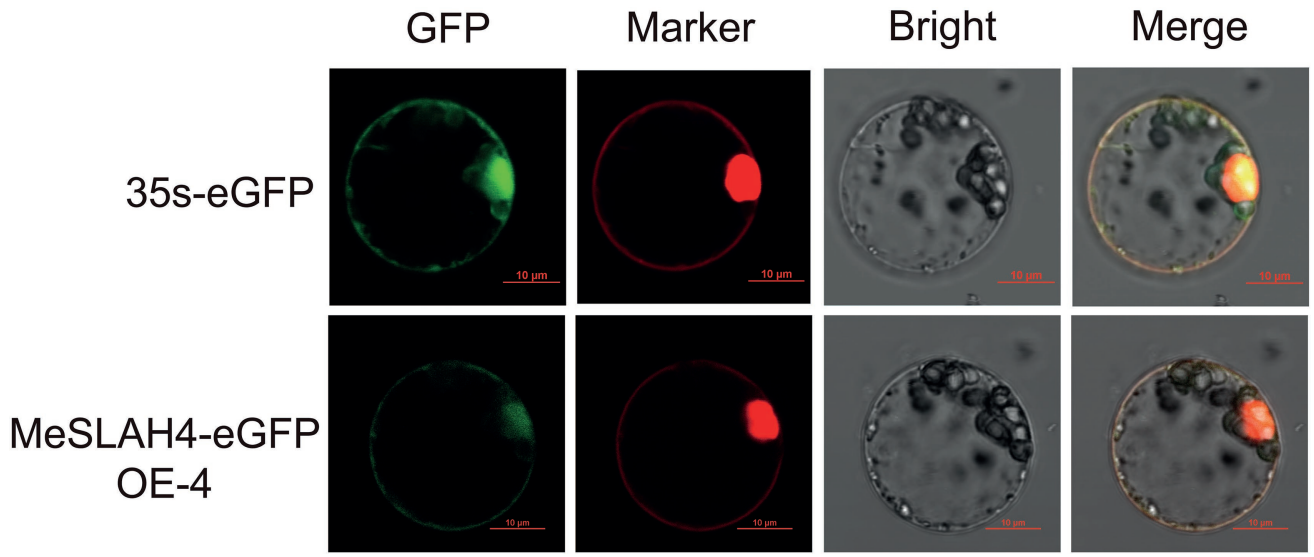

FIGURE 3 | Subcellular location of MeSLAH4 proteins. MeSLAH4-eGFP or 35 s-eGFP driven by the 35S promoter were transiently expressed in rice mesophyll protoplasts. Green signals indicate eGFP and red signals represent mCherry fluorescence. The merged images include the green fluorescence channel (first panels) and the chloroplast autofluorescence channel (second panels). The corresponding bright field images are shown on the right. Bar = 10 μm.

enhances aboveground biomass (plant height and shoot weight) but decreases the lower ground parts (root weights) under low nitrate conditions.

The chlorophyll-a content was non-significantly different at LN but significantly different ($p < 0.05$) under NN while it was higher at NN condition (**Figure 4E**). Conversely, the chlorophyll-b content was significantly different ($p < 0.05$) under LN but a non-significant difference was observed under NN, while the OE lines showed higher chlorophyll-b content compared with WT and OE lines under LN conditions (**Figure 4F**). The OE lines showed higher total chlorophyll content under LN conditions, which also demonstrated significant differences ($p < 0.05$) compared to the WT (**Figure 4G**). Higher chlorophyll content in OE lines under LN indicates the increasing nitrogen use efficiency and higher conversion of photosynthesis by the plants under low nitrate concentration.

The CAT activity was significantly different under the LN ($p < 0.01$) and NN ($p < 0.001$) conditions, but both lines (WT and OE) demonstrated higher activity under the NN condition, where WT had more activity than OE lines (**Figure 4H**). The POD activity of the OE line was higher than wild type (WT) under both (LN and NN) conditions, but it was highly significant ($p < 0.0001$) under the NN condition (**Figure 4I**). The SOD activity was complicated because the WT demonstrated higher SOD activity (650 $U.g^{-1}min^{-1}$) under the LN condition while the OE lines exhibited higher activity (700 $U.g^{-1}min^{-1}$) under the NN condition (**Figure 4J**). Lower CAT and SOD activity of OE lines in LN conditions indicates higher photosynthetic and stress-responsive activities, while the activity of POD in OE lines under both (LN and NN) conditions demonstrates a higher ability to scavenge hydrogen peroxide under prolonged nitrated conditions.

Effects of *MeSLAH4* Overexpression on Grain Morpho-Physiological Traits in Transgenic Rice

In the field trial, the grain length (**Figures 5A,C**) and breadth (**Figures 5B,D**) of transgenic rice were increased significantly relative to the wild type, and the highest increment was observed in the 35S:MeSLAH4 OE-4 line. The relative gene expression of this line (35S,MeSLAH4 OE-4) was higher (about 80-fold, **Figure 5E**) and the panicle morphology (**Figure 5F**) was better than the wild type, hence the OE-4 lines have been selected for all other experiments. The grain numbers of a single spike (around 60) in the OE lines have been increased (**Figure 5G**). These results indicate that the OE lines have up-regulated *MeSLAH4* gene expression, which has facilitated higher assimilation rates and more storage in the grain compared to the wild types.

As shown in **Figure 5**, all other traits of the OE line, including grain protein content (12%), grain moisture content (14%), grain diameter (4.8 mm), and grain yield of a single spike (1.48 g), were higher compared to the WT. Thus, the results of this experiment demonstrated that overexpression of the *MeSLAH4* gene increases protein content with higher grain expansion and yield in rice.

Effects of *MeSLAH4* Overexpression on Root Morphological Traits in Transgenic Rice

The root system indices (phenotype of roots) demonstrated changes in terms of size and shapes in the wild type (WT) and overexpression (OE) lines under different nitrate conditions (**Figure 6A**). The root fork numbers (number of branches) and root tip numbers were increased (580 and 490, respectively)

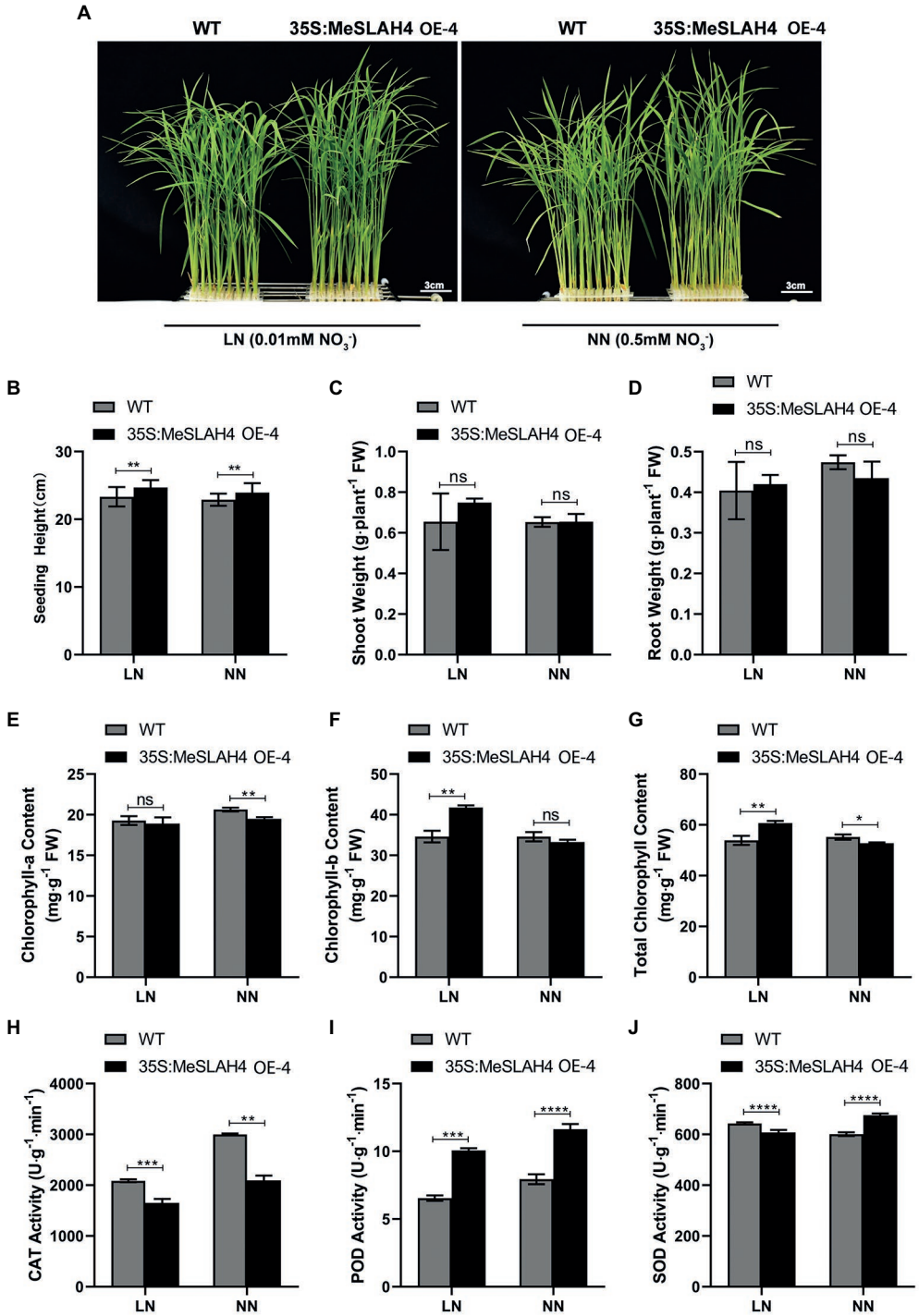

FIGURE 4 | Morpho-physiological traits of wild type and OE lines under different nitrogen conditions. **(A)** Phenotypes of wild type (WT) and MeSLAH4-OElines (35S:MeSLAH4) grown in hydroponic medium with different nitrate concentrations for 14 days. Bar = 3.0 cm. **(B)** Plant height (cm). **(C)** Shoot weight (g plant^{-1} FW). **(D)** Root weight (g plant^{-1} FW). **(E)** Chlorophyll-a content (mg. g^{-1} FW). **(F)** Chlorophyll-b content (mg. g^{-1} FW). **(G)** Total chlorophyll content (mg. g^{-1} FW). **(H)** CAT activity (U.g^{-1}.min^{-1}). **(I)** POD activity (U.g^{-1}.min^{-1}). **(J)** SOD activity of seedings (U.g^{-1}.min^{-1}). The letters "ns" indicate non-significant differences while *, **, ***, and **** represent significant differences at $p < 0.05$, $p < 0.01$, $p < 0.001$, $p < 0.0001$, respectively on one-way ANOVA analysis.

under the LN condition, and the increment was higher in the OE line compared to the WT line (**Figures 6B,C**). These results indicate that a lower nitrate concentration facilitated the formation of a higher root number. The root average diagram of WT was increased in NN, but it was increased higher in OE under the LN condition compared to WT, as well as both (WT and

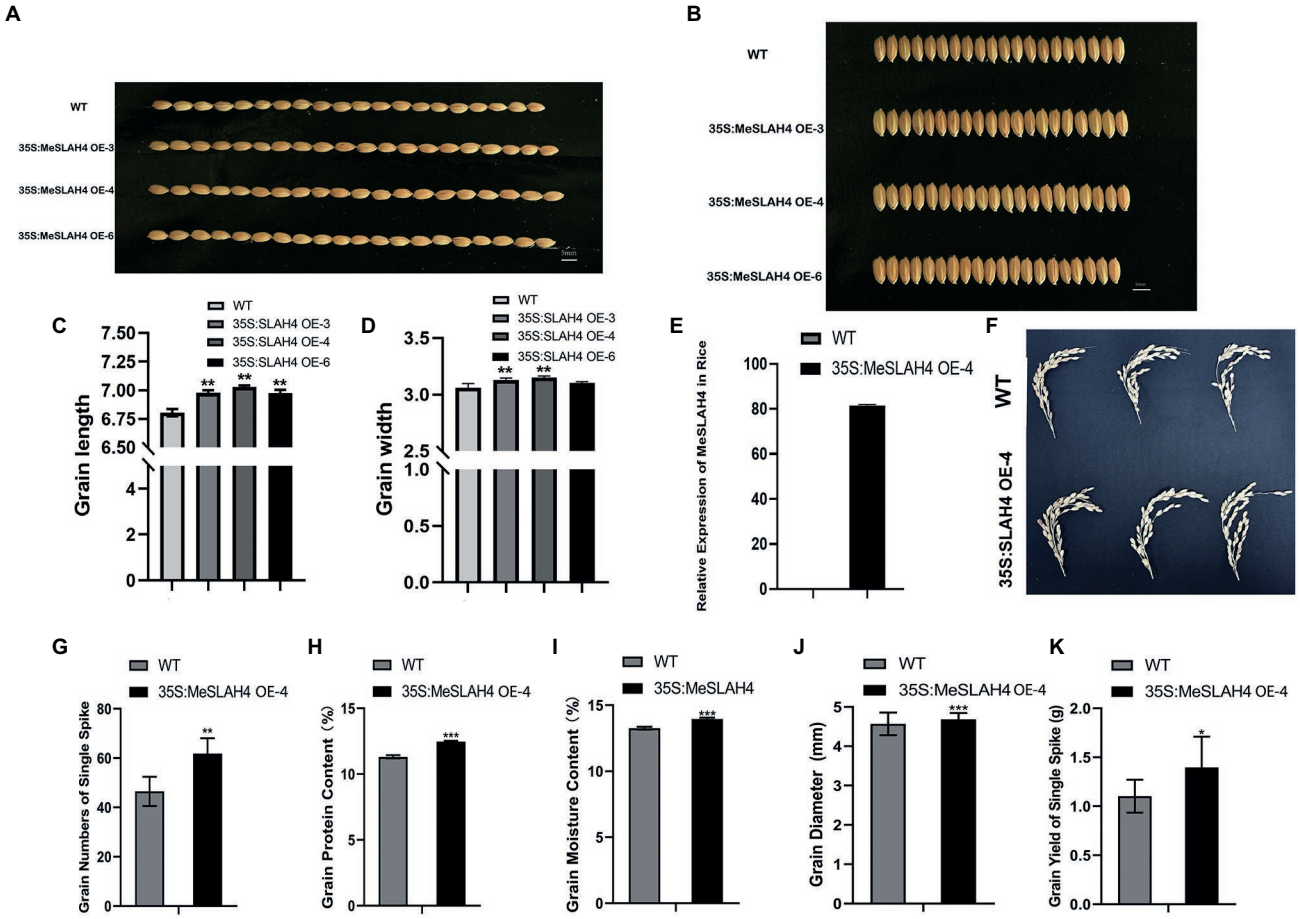

FIGURE 5 | Morpho-physiological traits of grains and relative expression of 35S:MeSLAH4 gene in the experimental lines. **(A,C)** grain length (mm). **(B,D)** grain breadth (mm). Bar = 5.0 mm. **(E)** Relative expression of the 35S:MeSLAH4OE4 line related to *OsActin* gene. **(F)** Panicle morphology. **(G)** grain numbers on a spike. **(H)** Grain protein content (%). **(I)** Grain moisture content (%). **(J)** Grain diameter (mm). **(K)** Grain yield of a spike (g). The symbols *, **, and *** represent significant differences at $p < 0.05$, $p < 0.01$, $p < 0.001$, respectively, on one-way ANOVA analysis.

OE) under the NN condition (**Figure 6D**). Similar types of expansions were observed in the OE line for root surface area (22.0 cm^2, **Figure 6E**) and root volume (22.0 cm^3, **Figure 6F**). However, root length was higher in the OE line (160 cm) under NN conditions, and it demonstrated non-significant changes compared to the WT (**Figure 6G**). Enlargement and expansion of root average diagram, root surface area, and root volume indicate that limited nitrate concentration does not inhibit root growth but rather allows it to optimize for absorption of more nutrient resources.

Nitrogen Accumulation in Rice Lines

The nitrate accumulation in the transgenic whole plants (35S:MeSLAH4) was higher under both the low nitrate (LN) and normal nitrate (NN) conditions compared to their wild type. However, there was a significant difference in the nitrate accumulation in whole plants at the low nitrate concentration (**Figure 7A**). During organ or tissue-specific nitrate accumulation analysis, rice grain exhibited significantly higher nitrate accumulation in the transgenic lines compared to the wild type lines (**Figure 7B**). Conversely, transgenic rice lines demonstrated lower nitrate accumulation in the glume (**Figure 7C**).

These results showed that overexpression of the *MeSLAH4* gene led to more nitrate accumulation by whole plants, more storage in the grain, but lower translocations to the glume in rice.

DISCUSSION

Nitrogen (N) is one of the most important micronutrients required for plant growth and development, and hence, plants have evolved different strategies, sophisticated mechanisms, and adaption processes depending on soil N availability and distribution. Among the four gene families involved in the nitrate transport system, the *SLAC/SLAH* members play an important role in anion transport, stress signaling, growth and development, and hormonal response (Vahisalu et al., 2008; Nan et al., 2021). A total of five *SLAC/SLAH* genes were

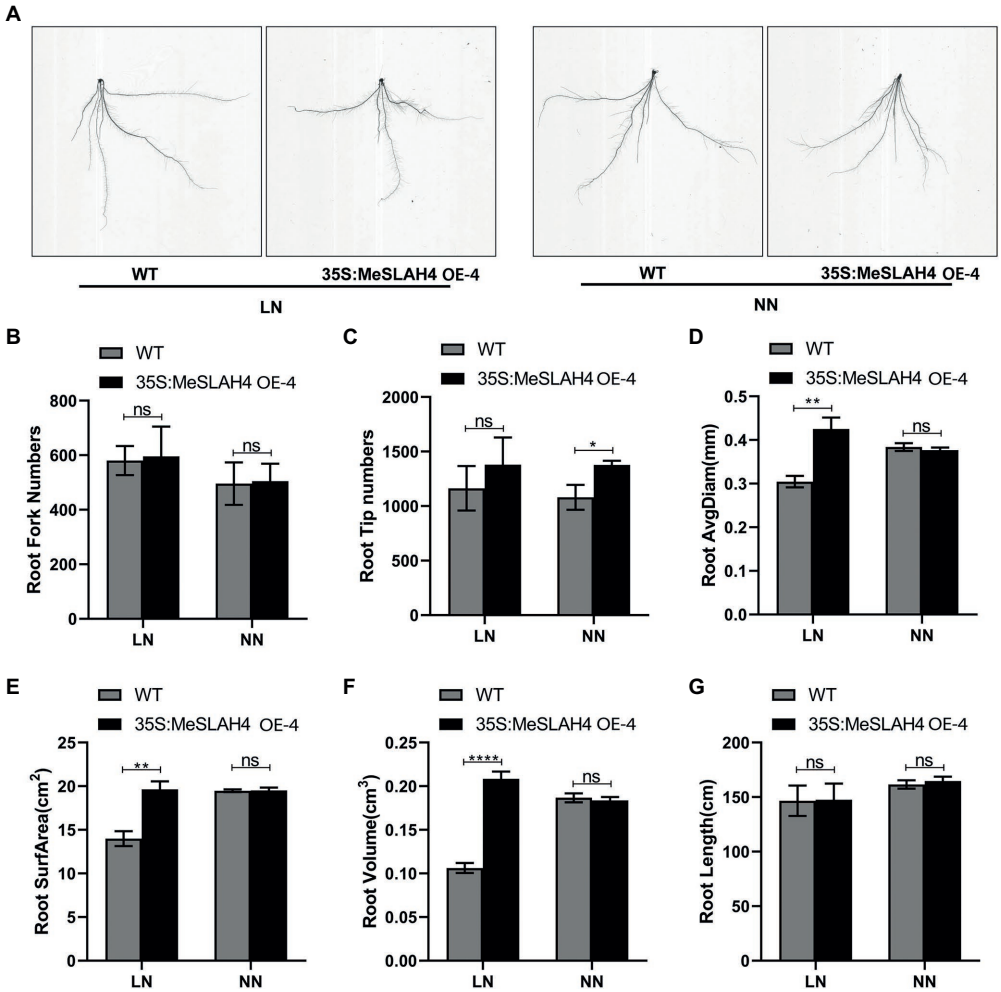

FIGURE 6 | Root phenotypes of the wild type and OE line under different nitrate conditions. **(A)** Phenotypes of wild type (WT), and MeSLAH4-OE line (35S:MeSLAH4) grown in hydroponic medium treated with LN (0.01 mm NO_3^-) and NN (0.50 mm NO_3^-) for 14 days. **(B)** Root fork numbers (number of branches). **(C)** Root tip numbers. **(D)** Root average diagram (mm). **(E)** Root surface area (cm^2). **(F)** Root volume (cm^3). **(G)** Root length (cm) of seedings. The letters "ns" indicate non-significant differences while *, **, ***, and **** represent significant differences at $p < 0.05$, $p < 0.01$, $p < 0.001$, $p < 0.0001$, respectively on one-way ANOVA analysis.

identified in *Arabidopsis* (Vahisalu et al., 2008), 23 genes in *B. napus* (Nan et al., 2021), and 9 genes in rice (Kusumi et al., 2012; Sun et al., 2016). In this study, six *SLAH* genes were identified in cassava, and these genes showed close phylogenetic relationships with other organisms (**Figure 1A**). These *SLAH* genes are expressed differentially in different tissues of cassava under varying nitrogen concentrations. Predominantly, the *MeSLAH4* (*MANES_05G153100*) gene was highly expressed in the root under free nitrate concentration (FN) as well as at high nitrate concentration (HN; **Figure 1B**), indicating a potential role in enhancing nitrogen use in plants. Since the root is the principal part for nutrient uptake in plants, the overexpression of the *MeSLAH4* gene has been tested in rice, which could correlate the relationships between different nitrate concentrations with plant growth and development parameters. In *Arabidopsis*, the *SLAH3* (*SLAC1* homologue 3) gene closely related to the *SLAC1* gene showed an overlapping function with *SLAC1* in guard cells (Negi et al., 2008; Geiger et al., 2011). The expression of the *SLAH3* gene was also detected in guard cells, albeit at much lower levels than the expression of *SLAC1* (Geiger et al., 2011; Zheng et al., 2015). High expression levels of the *SLAH3* gene were observed in roots and exhibited stronger selectivity for nitrate over chloride compared to *SLAC1* (Geiger et al., 2009; Lee et al., 2009), and therefore, it was considered a nitrate efflux channel. Although the *SLAH2* gene, which is the closest homolog of the *SLAH3* gene, is also expressed in root vascular tissues, it did not show any related phenotype under the same conditions as the *SLAH3* gene, indicating non-overlapping function (Zheng et al., 2015). In *Arabidopsis*, *SLAH1* and *SLAH4* genes share similar duplicates, and both are members of a clade that predates seed plants. However, similar to *SLAH3*, the *SLAH4* gene is also expressed in roots, and shows relatively stronger expression near the root tip (Zheng et al., 2015).

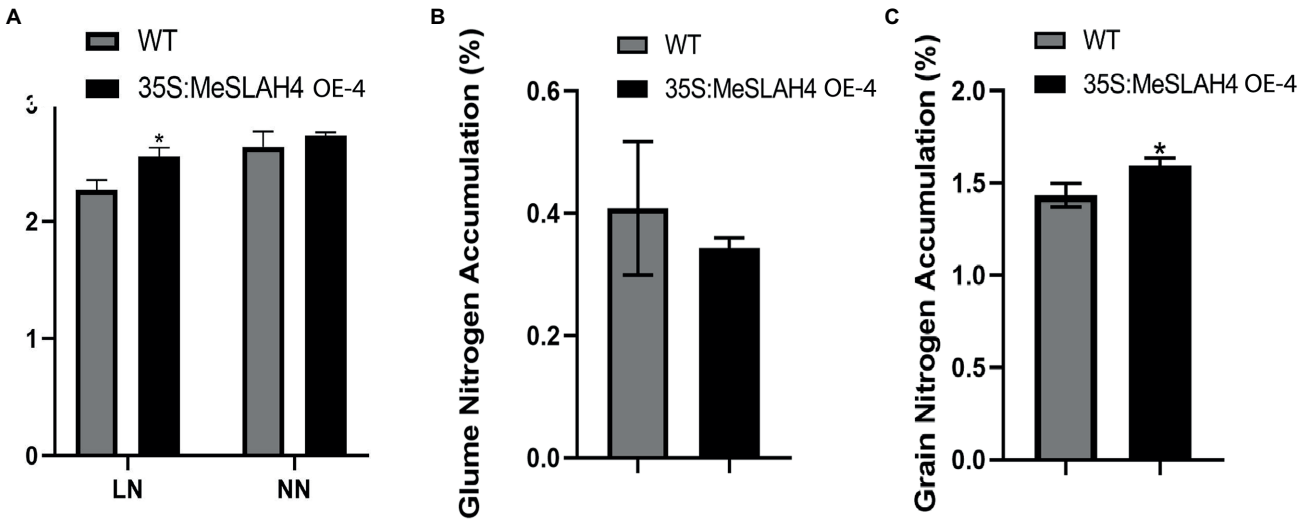

FIGURE 7 | Nitrogen accumulation in whole plants, grain, and glume in rice. **(A)** Nitrogen accumulation in whole rice plants. LN, 0.01 mm NH_4NO_3; NN, 0.5 mm NH_4NO_3. **(B)** Nitrogen accumulation (%) in glume. **(C)** Nitrogen accumulation (%) in grain. The error bars indicate standard deviations and the data are shown as mean values ± SD, while * represents significant differences at $p < 0.05$ on one-way ANOVA analysis.

In this study, the cassava *MeSLAH4* gene was identified on chromosome 5 (**Figure 2A**) and localized in the plasma membrane and nucleus (**Figure 3**). The localization of the *MeSLAH4* protein in the plasma membrane and nucleus indicated that MeSLAH4 protein may be involved in other cellular functions. Previously, confocal microscopy observations pointed out that BnSLAH1-1, BnSLAH3-2, and BnSLAH3-3 were localized on the plasma membrane the same as in *Arabidopsis* and pear (Chen et al., 2019a; Nan et al., 2021).

Analysis of the upstream promoter region of *SLAH* genes (**Figure 2B**) showed that the *cis*-regulatory elements, CAAT-box, were commonly shared by all *SLAH* genes, and most *SLAH* genes contained ABRE elements (ABA-responsive) and drought-responsive *cis*-elements (DRE), which are absent in the cassava *MeSLAH4* gene suggested that this gene might not be involved in regulation and physiological responses of various processes, including stomatal closure, seed, bud dormancy, and stresses (Gómez-Porras et al., 2007). Besides, the *MeSLAH4* gene was observed to contain several copies of the CAAT-box and a copy of the G-box, indicating a higher transcription rate with sufficient quantities of suitable binding sites for several transcription factors. The CAAT box is generally located approximately 80 bp upstream of the transcription start site (TSS) and significantly influences gene expression efficiency (Biłas et al., 2016). In addition, the presence of the highly conserved G-box motif (CACGTG) indicated frequent binding with the basic helix–loop–helix (bHLH) and basic Leu zipper (bZIP) TF families (Ezer et al., 2017). In *B. napus*, promoter analysis showed the presence of different kinds of *cis*-elements involved in the light response, phytohormone response, drought response, low temperature response, and growth regulation. It was assumed that the *BnSLAC*/*SLAH* may function in the abiotic stress tolerance, and growth regulation (Nan et al., 2021).

A total of 10 motifs (motifs 1 to 10) were identified in the Arabidopsis, rice, and cassava *SLAH* genes (**Supplementary Figures 2**, **3**), while the *MeSLAH4* gene in the cassava contains 5 motifs (motifs 1, 2, 4, 5, and motif 8), which might represent the conserved functional motif of this gene. In previous experiments on conserved motif analysis, it was suggested that the presence of motifs 1, 3, 4, 8, and 10 indicated a conserved functional motif in the *SLAC*/*SLAH* gene family of Rosaceae (Chen et al., 2019a). However, the *BnSLAH3* subfamily was found to contain motifs 1 to 10, while *BnSLAH2* contained motifs 1, 5, and 7. In the same experiment, motifs 1 to 7 were found to be widely distributed in the *BnSLAH1*, *BnSLAH4*, and *BnSLAC1* subfamilies (Nan et al., 2021). These conserved motifs were considered to have functional or structural roles in active proteins, indicating functional diversity during growth and development in plants (Nan et al., 2021).

In the current experiment, plant height was significantly different ($p < 0.01$) under both nitrated concentrations (LN and NN) in both WT and OE lines, but they demonstrated higher plant height at LN compared to NN (**Figure 4B**). The shoot weight was higher in OE lines at LN compared to both lines (WT and OE) in NN condition. Conversely, the root weight was higher in the WT under NN than under LN. Thus, overexpression of the *MeSLAH4* gene enhances aboveground biomass (plant height and shoot weight) but decreases the lower ground parts (root weights) under low nitrate conditions. In an earlier experiment, overexpression of the *OsNLP4* gene significantly increased N uptake and assimilation in rice, thus enhancing plant growth, grain yield and NUE compared with the wild type under all N conditions (Wu et al., 2021). The OE lines showed higher total chlorophyll content compared to the WT (**Figure 4G**) under LN conditions with significant differences ($p < 0.05$), indicating higher nitrogen use efficiency with higher conversion of photosynthesis under low nitrate

concentration. Previously, transgenic plants (overexpression of *OsGS1;1* and *OsGS2* genes) exhibited higher chlorophyll fluorescence under stress (drought and salinity) compared to control rice plants, which indicated that the transgenic lines had enhanced protection of the photosynthetic machinery, leading to improved post-stress recovery (James et al., 2018). Lower CAT (**Figure 4H**) activity and less SOD (**Figure 4J**) activity of OE lines in the LN condition indicated higher stress-responsive activities, while higher activity of POD (**Figure 4I**) in OE lines under both (LN and NN) conditions demonstrated a higher ability to scavenge hydrogen peroxide under prolonged nitrated conditions. The plants that were deficient in CAT indicated an association with photorespiratory H_2O_2 accumulation and downstream oxidative signaling (Vandenabeele et al., 2004). The SOD enzyme catalyzes the dismutation of the superoxide anion ($O_2^{\bullet-}$) into hydrogen peroxide and molecular oxygen, which play the most important roles in protecting against oxidative stress as well as in the survival of plants under stressful conditions (Gill and Tuteja, 2010). The activity of POD is increased under decreased CAT activity to compensate for the lack of H_2O_2 scavenging capacity in rice under stress conditions (Wang et al., 2019).

The root fork numbers (number of branches) and root tip numbers were increased under the LN condition, and the increment was higher in the OE line compared to the WT line (**Figures 6B,C**). Enlargement and expansion of root average diagram (**Figure 6D**), root surface area (**Figure 6E**), and root volume (**Figure 6F**) in OE lines indicating optimize condition for higher absorption of nutrients. Former researchers discovered that *BnSLAH3-2*, *BnSLAH3-3*, and *BnSLAH3-4* were up-regulated in roots 12 h after low nitrate treatment (0.19 mm), indicating that the *BnSLAH3* genes could respond quickly to low nitrate stress and may promote nitrate uptake and transport in rapeseed roots. Conversely, a high concentration (64 mm) of nitrate was detected to induce expression of *SLAC*/*SLAH* genes in pear, which indicated that gene expression varies depending on species, nitrate concentration, and treatment time (Chen et al., 2019a; Nan et al., 2021).

The nitrate accumulation in the transgenic plants (35S,MeSLAH4) was higher under both nitrate (LN and NN) conditions compared to their wild type, but it was significantly different at the low nitrate concentration (**Figure 7A**). However, higher nitrate accumulation led to more storage in the grain but lower translocations to the glume in rice. In the present experiments, other traits, including grain numbers of a single spike, grain protein content, grain moisture content, grain diameter, and grain yield of a single spike of the OE line, were higher compared to the WT (**Figure 5**). It is well known that crop yield is closely related to N utilization, and it mainly depends on nitrogen absorption by plants before flowering and nitrogen remobilization during seed maturation (Kichey et al., 2007; Masclaux-Daubresse et al., 2008). Current research reveals that overexpression of the *MeSLAH4* gene significantly enhances grain size as well as nitrate influx in OE-lines compared to the wild type. Hence, the *MeSLAH4* gene might play an important role in the process of nitrogen transport and nitrogen utilization efficiency, which could be useful in developing high-yielding crop varieties. In addition, this study found that *MeSLAH4* has great impacts on the biological function, regulatory mechanism of nitrate absorption and utilization, and enhanced performance of yield-related traits in rice.

CONCLUSION

Cassava is a short-day, durable, and easy-to-plant dicot plant with high adaptability and a huge yield that can obtain sufficient nitrogen from the soil without requiring excessive nitrogen fertilizer. For the improvement and production of nitrogen-efficient germplasm resources, it is crucial to identify the key genes involved in nitrogen-efficient utilization. However, it is well evident that the *SLAC*/*SLAH* genes play important roles in responses to nitrate transport, stress signaling, and growth and development in plants. Till date, detailed bioinformatic analyses of the *SLAC*/*SLAH* gene family in the cassava genome have not been reported completely. Only some identified gene information is available in the Phytozome and NCBI databases. The functional characterization and expression analysis of these genes remain to be elucidated. Hence, in this study, six *SLAC*/*SLAH* genes were identified in the cassava genomes, which demonstrated a close phylogenetic relationship with other organisms. The structural characteristics of the promoter region, gene expression analyses, motif and sequence logo comparisons, and chromosomal localizations with *Arabidopsis* and rice homologs have provided a suitable framework for analyzing the *SLAC*/*SLAH* genes in the cassava genome. Cassava *SLAH* genes, particularly the *MeSLAH4* gene, respond significantly to different concentrations of nitrate ions and are expressed highly in the roots and enhance grain dimension while increasing yield in rice. The *MeSLAH4* gene is identified on chromosome 5 and is localized in the plasma membrane and nucleus. The overexpression (OE) rice lines showed higher total chlorophyll content, increased root fork numbers (number of branches), and root tip numbers compared to the WT under low nitrate (LN) conditions. The findings of these studies revealed the potential of the *MeSLAH4* gene for use in high-yielding crop production, as well as laid the groundwork for future research into the other *SLAC*/*SLAH* genes found in the cassava genome.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

LS, XW, ZP, LJ, and LZ accomplished and finalized the experiment, performed data analysis, and prepared a draft of the manuscript. LS, LJ, JhY, JnY, and GL conducted experimental trails and collected data. LZ, XW, CW, and RZ constructed the transformation vector and produced transgenic plants. LS, SL, JhY, and YZ participated in morpho-physiological

data measurements and prepared figures. XZ and WL provided guidance for the experimental design, analysis, and writing. ZP, XJ, XZ, WL, and ZZ designed, monitored, and validated the experimental procedures and corrected the final manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Hainan Provincial Natural Science Foundation of China (grant no. 2019RC303), the Major Science and Technology Plan of Hainan Province (grant no. ZDKJ2021012), the Advanced Scientific Program for the Returned Overseas Chinese Scholars, Henan Province (grant no. 30602724), the Henan Province Science and Technology Attack Project (grant no. 222102110465), the Special Fund for High-level Talent Research Team of Neijiang Normal University (grant no. RSC202102), and the State Key Laboratory for Managing Biotic and Chemical Treats to the Quality and Safety of Agro-products (grant nos. KF20200107 and KF202218).

ACKNOWLEDGMENTS

The authors are thankful to the Henan Agricultural University, Henan, China, Neijiang Normal University, Sichuan, China, Chinses Academy of Tropical Agricultural Sciences, Hainan, China, and Sichuan Agricultural University, Sichuan, China for providing laboratory facilities and logistic support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.932947/full#supplementary-material

Supplementary Figure 1 | Amino acid alignment of MeSLAH4 protein in cassava, and Os05g0269200 protein in rice. Here, MANES (*Manihot esculenta*) MeSLAH4 amino acid sequence was the query sequence and it was aligned with Os05g0269200 amino acid sequence.

Supplementary Figure 2 | Conserved motifs of *SLAH* genes. Here, MANES (*Manihot esculenta*), At (*Arabidopsis thaliana*), and Os (*Oryza sativa*) genes and their motifs are depicted. Various color represents different motifs. The lengths and positions of the colored blocks correspond to the lengths and positions of motifs in the individual protein sequences. The scale indicates the lengths of the proteins as well as the motifs.

Supplementary Figure 3 | Sequence logos of the conserved motifs of *SLAH* genes. Over-represented motifs were identified using the MEME tool. The stack's height indicates the level of sequence conservation. The heights of the residues within the stack indicate the relative frequencies of each residue at that position.

Supplementary Figure 4 | Identification of transgenic rice lines. Lane 1–7 is a single transgenic strain of rice (35S:MeSLAH4OE-4), – WT control, + is MeSLAH4 plasmid.

REFERENCES

Aebi, H. (1984). Catalase *in vitro*. *Meth. Enzymol* 105, 121–126. doi: 10.1016/S0076-6879(84)05016-3

Arnon, D. I. (1949). Copper enzymes in isolated chloroplasts. Polyphenoloxidase in Beta vulgaris. *Plant Physiol.* 24, 1–15.

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

Biłas, R., Szafran, K., Hnatuszko-Konka, K., and Kononowicz, A. K. (2016). *Cis*-regulatory elements used to control gene expression in plants. *Plant Cell Tissue Organ Cult.* 127, 269–287. doi: 10.1007/s11240-016-1057-7

Brandt, B., Brodsky, D. E., Xue, S., Negi, J., Iba, K., Kangasjärvi, J., et al. (2012). Reconstitution of abscisic acid activation of SLAC1 anion channel by CPK6 and OST1 kinases and branched ABI1 PP2C phosphatase action. *Proc. Natl. Acad. Sci.* 109, 10593–10598. doi: 10.1073/pnas.1116590109

Burman, N., Chandran, D., and Khurana, J. P. (2020). A rapid and highly efficient method for transient gene expression in rice plants. *Front. Plant Sci.* 11:584011. doi: 10.3389/fpls.2020.584011

Chen, G., Li, X., Qiao, X., Li, J., Wang, L., Kou, X., et al. (2019a). Genome-wide survey and expression analysis of the *SLAC/SLAH* gene family in pear (*Pyrus bretschneideri*) and other members of the Rosaceae. *Genomics* 111, 1097–1107. doi: 10.1016/j.ygeno.2018.07.004

Chen, G., Wang, L., Chen, Q., Qi, K., Yin, H., Cao, P., et al. (2019b). *PbrSLAH3* is a nitrate-selective anion channel which is modulated by calcium-dependent protein kinase 32 in pear. *BMC Plant Biol.* 19, 1–12. doi: 10.1186/s12870-019-1813-z

Cubero-Font, P., Maierhofer, T., Jaslan, J., Rosales, M. A., Espartero, J., Díaz-Rueda, P., et al. (2016). Silent S-type anion channel subunit *SLAH1* gates *SLAH3* open for chloride root-to-shoot translocation. *Curr. Biol.* 26, 2213–2220. doi: 10.1016/j.cub.2016.06.045

Drunkler, N. L., Leite, R. S., Mandarino, J. M. G., Ida, E. I., and Demiate, I. M. (2012). Cassava starch as a stabilizer of soy-based beverages. *Food Sci. Technol. Int.* 18, 489–499. doi: 10.1177/1082013211433072

Ezer, D., Shepherd, S. J., Brestovitsky, A., Dickinson, P., Cortijo, S., Charoensawan, V., et al. (2017). The G-box transcriptional regulatory code in Arabidopsis. *Plant Physiol.* 175, 628–640. doi: 10.1104/pp.17.01086

Geiger, D., Maierhofer, T., Al-Rasheid, K. A., Scherzer, S., Mumm, P., Liese, A., et al. (2011). Stomatal closure by fast abscisic acid signaling is mediated by the guard cell anion channel SLAH3 and the receptor RCAR1. *Sci. Signal.* 4:ra32. doi: 10.1126/scisignal.2001346

Geiger, D., Scherzer, S., Mumm, P., Stange, A., Marten, I., Bauer, H., et al. (2009). Activity of guard cell anion channel *SLAC1* is controlled by drought-stress signaling kinase-phosphatase pair. *Proc. Natl. Acad. Sci.* 106, 21425–21430. doi: 10.1073/pnas.0912021106

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

Gómez-Porras, J. L., Riaño-Pachón, D. M., Dreyer, I., Mayer, J. E., and Mueller-Roeber, B. (2007). Genome-wide analysis of ABA-responsive elements ABRE and CE3 reveals divergent patterns in *Arabidopsis* and rice. *BMC Genomics* 8, 1–13. doi: 10.1186/1471-2164-8-260

Gutermuth, T., Lassig, R., Portes, M. T., Maierhofer, T., Romeis, T., Borst, J. W., et al. (2013). Pollen tube growth regulation by free anions depends on the interaction between the anion channel *SLAH3* and calcium-dependent protein kinases CPK2 and CPK20. *Plant Cell* 25, 4525–4543. doi: 10.1105/tpc.113.118463

Hachiya, T., and Sakakibara, H. (2017). Interactions between nitrate and ammonium in their uptake, allocation, assimilation, and signaling in plants. *J. Exp. Bot.* 68, 2501–2512. doi: 10.1093/jxb/erw449

Ho, C. H., and Tsay, Y. F. (2010). Nitrate, ammonium, and potassium sensing and signaling. *Curr. Opin. Plant Biol.* 13, 604–610. doi: 10.1016/j.pbi.2010.08.005

Jaborsky, M., Maierhofer, T., Olbrich, A., Escalante-Pérez, M., Müller, H. M., Simon, J., et al. (2016). SLAH3-type anion channel expressed in poplar secretory epithelia operates in calcium kinase CPK-autonomous manner. *New Phytol.* 210, 922–933. doi: 10.1111/nph.13841

James, D., Borphukan, B., Fartyal, D., Ram, B., Singh, J., Manna, M., et al. (2018). Concurrent overexpression of *OsGS1; 1* and *OsGS2* genes in transgenic

rice (*Oryza sativa* L.): impact on tolerance to abiotic stresses. *Front. Plant Sci.* 9:786. doi: 10.3389/fpls.2018.00786

Jiang, Q., Kang, L., Zhang, X., Yao, Y., Liang, Q., Gu, M., et al. (2016). Effects of nitrogen level on source-sink relationship of cassava. *Southwest China J. Agric. Sci.* 29, 2162–2166. doi: 10.16213/j.cnki.scjas.2016.09.026

Kichey, T., Hirel, B., Heumez, E., Dubois, F., and Le Gouis, J. (2007). In winter wheat (*Triticum aestivum* L.), post-anthesis nitrogen uptake and remobilisation to the grain correlates with agronomic traits and nitrogen physiological markers. *Field Crops Res.* 102, 22–32. doi: 10.1016/j.fcr.2007.01.002

Kraiser, T., Gras, D. E., Gutiérrez, A. G., González, B., and Gutiérrez, R. A. (2011). A holistic view of nitrogen acquisition in plants. *J. Exp. Bot.* 62, 1455–1466. doi: 10.1093/jxb/erq425

Krapp, A., David, L. C., Chardin, C., Girin, T., Marmagne, A., Leprince, A. S., et al. (2014). Nitrate transport and signalling in *Arabidopsis*. *J. Exp. Bot.* 65, 789–798. doi: 10.1093/jxb/eru001

Kurusu, T., Saito, K., Horikoshi, S., Hanamata, S., Negi, J., Yagi, C., et al. (2013). An S-type anion channel *SLAC1* is involved in cryptogein-induced ion fluxes and modulates hypersensitive responses in tobacco BY-2 cells. *PLoS One* 8:e70623. doi: 10.1371/journal.pone.0070623

Kusumi, K., Hirotsuka, S., Kumamaru, T., and Iba, K. (2012). Increased leaf photosynthesis caused by elevated stomatal conductance in a rice mutant deficient in *SLAC1*, a guard cell anion channel protein. *J. Exp. Bot.* 63, 5635–5644. doi: 10.1093/jxb/ers216

Lee, S. C., Lan, W., Buchanan, B. B., and Luan, S. (2009). A protein kinase-phosphatase pair interacts with an Ion channel to regulate ABA signaling in plant guard cells. *Proc. Natl. Acad. Sci.* 106, 21419–21424. doi: 10.1073/pnas.0910601106

Li, Y., Fan, C., Xing, Y., Yun, P., Luo, L., Yan, B., et al. (2014). *Chalk5* encodes a vacuolar H^+-translocating pyrophosphatase influencing grain chalkiness in rice. *Nat. Genet.* 46, 398–404. doi: 10.1038/ng.2923

Li, B., Liu, D., Li, Q., Mao, X., Li, A., Wang, J., et al. (2016a). Overexpression of wheat gene *TaMOR* improves root system architecture and grain yield in *Oryza sativa*. *J. Exp. Bot.* 67, 4155–4167. doi: 10.1093/jxb/erw193

Li, Y. Y., Shen, A., Xiong, W., Sun, Q. L., Luo, Q., Song, T., et al. (2016b). Overexpression of *OsHox32* results in pleiotropic effects on plant type architecture and leaf development in rice. *Rice* 9:46. doi: 10.1186/s12284-016-0118-1

Li, D., Zhang, P., Chen, T., and Qin, W. (2020). Recent development and challenges in spectroscopy and machine vision Technologies for Crop Nitrogen Diagnosis: a review. *Remote Sens.* 12:2578. doi: 10.3390/rs12162578

Li, F., Zhang, H., Zhao, H., Gao, T., Song, A., Jiang, J., et al. (2018). Chrysanthemum *CmHSFA4* gene positively regulates salt stress tolerance in transgenic chrysanthemum. *Plant Biotechnol. J.* 16, 1311–1321. doi: 10.1111/pbi.12871

Liu, X., Mak, M., Babla, M., Wang, F., Chen, G., Veljanoski, F., et al. (2014). Linking stomatal traits and expression of slow anion channel genes *HvSLAH1* and *HvSLAC1* with grain yield for increasing salinity tolerance in barley. *Front. Plant Sci.* 5:634. doi: 10.3389/fpls.2014.00634

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

Maierhofer, T., Diekmann, M., Offenborn, J. N., Lind, C., Bauer, H., Hashimoto, K., et al. (2014a). Site-and kinase-specific phosphorylation-mediated activation of *SLAC1*, a guard cell anion channel stimulated by abscisic acid. *Sci. Signal.* 7:ra86. doi: 10.1126/scisignal.2005703

Maierhofer, T., Lind, C., Hüttl, S., Scherzer, S., Papenfuß, M., Simon, J., et al. (2014b). A single-pore residue renders the *Arabidopsis* root anion channel SLAH2 highly nitrate selective. *Plant Cell* 26, 2554–2567. doi: 10.1105/tpc.114.125849

Masclaux-Daubresse, C., Daniel-Vedele, F., Dechorgnat, J., Chardon, F., Gaufichon, L., and Suzuki, A. (2010). Nitrogen uptake, assimilation and remobilization in plants: challenges for sustainable and productive agriculture. *Ann. Bot.* 105, 1141–1157. doi: 10.1093/aob/mcq028

Masclaux-Daubresse, C., Reisdorf-Cren, M., and Orsel, M. (2008). Leaf nitrogen remobilisation for plant development and grain filling. *Plant Biol.* 10, 23–36. doi: 10.1111/j.1438-8677.2008.00097.x

Nan, Y., Xie, Y., Atif, A., Wang, X., Zhang, Y., Tian, H., et al. (2021). Identification and expression analysis of *SLAC/SLAH* gene family in *Brassica napus* L. *Int. J. Mol. Sci.* 22:4671. doi: 10.3390/ijms22094671

Negi, J., Matsuda, O., Nagasawa, T., Oba, Y., Takahashi, H., Kawai-Yamada, M., et al. (2008). CO_2 regulator SLAC1 and its homologues are essential for anion homeostasis in plant cells. *Nature* 452, 483–486. doi: 10.1038/nature06720

O'Brien, J. A., Vega, A., Bouguyon, E., Krouk, G., Gojon, A., Coruzzi, G., et al. (2016). Nitrate transport, sensing, and responses in plants. *Mol. Plant* 9, 837–856. doi: 10.1016/j.molp.2016.05.004

Qi, G. N., Yao, F. Y., Ren, H. M., Sun, S. J., Tan, Y. Q., Zhang, Z. C., et al. (2018). The S-type anion channel *ZmSLAC1* plays essential roles in stomatal closure by mediating nitrate efflux in maize. *Plant Cell Physiol.* 59, 614–623. doi: 10.1093/pcp/pcy015

Sun, S. J., Qi, G. N., Gao, Q. F., Wang, H. Q., Yao, F. Y., Hussain, J., et al. (2016). Protein kinase *OsSAPK8* functions as an essential activator of S-type anion channel *OsSLAC1*, which is nitrate-selective in rice. *Planta* 243, 489–500. doi: 10.1007/s00425-015-2418-x

Sun, H., Tao, J., Liu, S., Huang, S., Chen, S., Xie, X., et al. (2014). Strigolactones are involved in phosphate-and nitrate-deficiency-induced root development and auxin transport in rice. *J. Exp. Bot.* 65, 6735–6746. doi: 10.1093/jxb/eru029

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., et al. (2007). The *Arabidopsis* information resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.* 36, D1009–D1014. doi: 10.1093/nar/gkm965

Vahisalu, T., Kollist, H., Wang, Y. F., Nishimura, N., Chan, W. Y., Valerio, G., et al. (2008). *SLAC1* is required for plant guard cell S-type anion channel function in stomatal signalling. *Nature* 452, 487–491. doi: 10.1038/nature06608

Vahisalu, T., Puzõrjova, I., Brosché, M., Valk, E., Lepiku, M., Moldau, H., et al. (2010). Ozone-triggered rapid stomatal response involves the production of reactive oxygen species, and is controlled by SLAC1 and OST1. *Plant J.* 62, 442–453. doi: 10.1111/j.1365-313X.2010.04159.x

Vandenabeele, S., Vanderauwera, S., Vuylsteke, M., Rombauts, S., Langebartels, C., Seidlitz, H. K., et al. (2004). Catalase deficiency drastically affects gene expression induced by high light in *Arabidopsis thaliana*. *Plant J.* 39, 45–58. doi: 10.1111/j.1365-313X.2004.02105.x

Vidal, E. A., Alvarez, J. M., Araus, V., Riveras, E., Brooks, M. D., Krouk, G., et al. (2020). Nitrate in 2020: thirty years from transport to signaling networks. *Plant Cell* 32, 2094–2119. doi: 10.1105/tpc.19.00748

Wang, W. (2002). "Cassava production for industrial utilization in China–present and future perspective." in *7th Cassava Regional Conference Proceeding*; October 28, 2002; Bangkok, Thailand, 33.

Wang, X., Liu, H., Yu, F., Hu, B., Jia, Y., Sha, H., et al. (2019). Differential activity of the antioxidant defence system and alterations in the accumulation of osmolyte and reactive oxygen species under drought stress and recovery in rice (*Oryza sativa* L.) tillering. *Sci. Rep.* 9, 1–11. doi: 10.1038/s41598-019-44958-x

Wang, M. Y., Siddiqi, M. Y., Ruth, T. J., and Glass, A. D. (1993). Ammonium uptake by rice roots (II. Kinetics of $13NH_4^+$ influx across the plasmalemma). *Plant Physiol.* 103, 1259–1267. doi: 10.1104/pp.103.4.1259

Wu, J., Zhang, Z. S., Xia, J. Q., Alfatih, A., Song, Y., Huang, Y. J., et al. (2021). Rice NIN-LIKE PROTEIN 4 plays a pivotal role in nitrogen use efficiency. *Plant Biotechnol. J.* 19, 448–461. doi: 10.1111/pbi.13475

Yoshimura, K., Yabuta, Y., Ishikawa, T., and Shigeoka, S. (2000). Expression of spinach ascorbate peroxidase isoenzymes in response to oxidative stresses. *Plant Physiol.* 123, 223–234. doi: 10.1104/pp.123.1.223

Zhang, Y., Su, J., Duan, S., Ao, Y., Dai, J., Liu, J., et al. (2011). A highly efficient rice green tissue protoplast system for transient gene expression and studying light/chloroplast-related processes. *Plant Methods* 7, 1–14. doi: 10.1186/1746-4811-7-30

Zhao, Y., Du, H., Wang, Y., Wang, H., Yang, S., Li, C., et al. (2021). The calcium-dependent protein kinase *ZmCDPK7* functions in heat-stress tolerance in maize. *J. Integr. Plant Biol.* 63, 510–527. doi: 10.1111/jipb.13056

Zheng, X., He, K., Kleist, T., Chen, F., and Luan, S. (2015). Anion channel *SLAH3* functions in nitrate-dependent alleviation of ammonium toxicity in Arabidopsis. *Plant Cell Environ.* 38, 474–486. doi: 10.1111/pce.12389

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations,

or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 19 July 2022
doi: 10.3389/fpls.2022.960302

Transcriptome Analysis of Moso Bamboo (*Phyllostachys edulis*) Reveals Candidate Genes Involved in Response to Dehydration and Cold Stresses

Zhuo Huang, Peilei Zhu, Xiaojuan Zhong, Jiarui Qiu, Wenxin Xu and Li Song*

College of Landscape Architecture, Sichuan Agricultural University, Chengdu, China

Bamboo (Bambusoideae) belongs to the grass family (Poaceae) and has been utilized as one of the most important nontimber forest resources in the world. Moso bamboo (*Phyllostachys edulis*) is a large woody bamboo with high ecological and economic values. Global climate change brings potential challenges to the normal growth of moso bamboo, and hence its production. Despite the release of moso bamboo genome sequence, the knowledge on genome-wide responses to abiotic stress is still limited. In this study, we generated a transcriptome data set with respect to dehydration and cold responses of moso bamboo using RNA-seq technology. The differentially expressed genes (DEGs) under treatments of dehydration and cold stresses were identified. By combining comprehensive gene ontology (GO) analysis, time-series analysis, and co-expression analysis, candidate genes involved in dehydration and cold responses were identified, which encode abscisic acid (ABA)/water deficit stress (WDS)-induced protein, late embryogenesis abundant (LEA) protein, 9-cis-epoxycarotenoid dioxygenase (NCED), anti-oxidation enzymes, transcription factors, etc. Additionally, we used *PeLEA14,* a dehydration-induced gene encoding an "atypical" LEA protein, as an example to validate the function of the identified stress-related gene in tolerance to abiotic stresses, such as drought and salt. In this study, we provided a valuable genomic resource for future excavation of key genes involved in abiotic stress responses and genetic improvement of moso bamboo to meet the requirement for environmental resilience and sustainable production.

Keywords: moso bamboo (*Phyllostachys edulis*), abiotic stress response, transcriptome, dehydration, late embryogenesis abundant protein

OPEN ACCESS

Edited by:
Hai Du,
Southwest University, China

Reviewed by:
Maoqun Yu,
Chengdu Institute of Biology (CAS), China
Chao Ma,
China Agricultural University, China

****Correspondence:***
Zhuo Huang
huangzhuo@sicau.edu.cn

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

Received: *02 June 2022*
Accepted: *21 June 2022*
Published: *19 July 2022*

Citation:
Huang Z, Zhu P, Zhong X, Qiu J, Xu W and Song L (2022) Transcriptome Analysis of Moso Bamboo (Phyllostachys edulis) Reveals Candidate Genes Involved in Response to Dehydration and Cold Stresses. Front. Plant Sci. 13:960302. doi: 10.3389/fpls.2022.960302

INTRODUCTION

Bambusoideae, also called bamboo, belongs to the grass family (Poaceae) and is comprised of more than 1,400 species. Compared with other herbaceous species of Poaceae, the species of Bambusoideae are predominantly arborescent and perennial woody species. They can grow large woody culms up to 30 cm in diameter and 12 m in height (Barker et al., 2001) and are utilized as one of the most important nontimber forest resources in the world. According to the statistics,

bamboo covers over 30 million hectares (ha) worldwide, and approximately 2.5 billion people depend economically on bamboo (Lobovikov, 2005), accounting for 68.8 billion US dollars in international trade in 2018 (International Bamboo and Rattan Organization).

Moso bamboo (*Phyllostachys edulis*) is a large woody bamboo with high ecological and economic values. It serves as a promising bio-resource for renewable forestry products and accounts for over two-thirds of the total bamboo growing area (4.43 million ha) in China (Peng et al., 2013). As sessile organisms, plants have evolved a wide spectrum of adaptations to cope with the inevitable challenges from environmental stress, such as drought, high salinity, and cold. Many aspects of these adaptation processes, including developmental, physiological, and biochemical changes, are regulated or achieved by stress-responsive gene expression (Huang et al., 2016a). Therefore, the identification of key genes involved in abiotic stress response is essential for the dissection of the complex mechanism underlying the stress tolerance, which will provide guidance for plant genetic improvement to meet continuous economic requirements and environmental resilience.

The draft genome sequence of moso bamboo was released previously (Peng et al., 2013), and an updated chromosomal level reference genome was also reported recently (Zhao et al., 2018). These genomic resources provide an opportunity for the excavation of stress-related genes at a genome-wide level. Huang et al. (2016a,b) analyzed genes encoding TIFY transcription factor and late embryogenesis abundant protein families in the moso bamboo genome and identified some stress-responsive genes. Jin et al. (2020) identified the expansin (EX) gene family in the moso bamboo genome and found that the expression of some *PeEXs* was induced by abscisic acid (ABA) and polyethylene glycol (PEG) treatments. Studies on stress-related functional genes have gradually increased. A moso bamboo WRKY gene *PeWRKY83* confers salinity tolerance in transgenic *Arabidopsis* plants (Wu et al., 2017); overexpression of *PeVQ28* in *Arabidopsis* increased resistance to salt stress and enhanced sensitivity to ABA (Cheng et al., 2020); a moso bamboo homeodomain-leucine zipper (HD-Zip) transcription factor Phehdz1 positively regulates the drought stress response of transgenic rice (Gao et al., 2021). *Arabidopsis* overexpressing *PheWRKY50-1* showed higher resistance to stress than the wild type (WT) (Huang et al., 2022). These studies made insights into moso bamboo's responses to abiotic stress. However, the genome-wide data for abiotic stress responses is still lacking. Recently, transcriptome profiling was conducted and revealed the crucial biological pathways involved in the cold response in moso bamboo (Liu et al., 2020).

To generate more genomic resources for stress-responsive gene mining, in this study, RNA-seq was employed to analyze transcriptomal responses of moso bamboo to dehydration and cold stresses. Differentially expressed genes (DEGs) at different time points of treatment were identified, and comprehensive stress-responsive gene exploration was also performed. A dehydration-responsive gene *PeLEA14*, encoding an "atypical" late embryogenesis abundant (LEA) protein, was selected as a tester to preliminarily investigate its function in drought and salt tolerance.

MATERIALS AND METHODS

Plant Materials and Growth Condition

The *P. edulis* plants used in this study were approximately 2-year-old plants of *P. edulis*, which were manually planted and grown under the natural condition at Linyanshan Experimental Base (N31°00 33.2000, E103°360 51.9500) of Sichuan Agricultural University, Dujiangyan, Sichuan, China. The approximately 20 cM branch containing young unexpanded leaves of 4–5 centimeters long was cut from five plants in similar growth status. All these samples were collected in the morning (at approximately 10 o'clock and it was cloudy with a temperature of ∼22°C). *Nicotiana tabacum* L. was used for *Agrobacterium*-mediated genetic transformation using the leaf disk method.

Stress Treatments, Sampling, and RNA-seq

For dehydration treatment, the branches were placed on the dry filter paper and treated under room temperature (20°C and ∼50% humidity). For cold treatment, the branches were put into a dark chamber set to 0°C. At different time points of treatments (2, 4, and 8 h for dehydration treatment, and 2 h and 4 h for cold treatment), ten unexpanded leaves were detached from the base and immediately frozen in liquid nitrogen and stored in the refrigerator at −80°C. The same amount of untreated leaves was also sampled and processed, which were used as control. One biological repeat for all samples was collected.

The total RNA was extracted according to the manual of the TRIZOL RNA Kit (TIANGEN, Beijing, China). The qualities and quantities of extracted nucleotide were measured using the NanoDrop 2000 Spectrophotometer (Thermo Fisher, USA) and the Agilent 2100 RNA 6000 Nano kit. The threshold of the quality of extracted RNA was RIN $\geq$ 7 with a concentration of $\geq$ 150 ng/μl and an amount of $\geq$ 5 μg. The cDNA library construction and pair-end sequencing on Illumina HiSeq^{TM} 4000 platform were performed by Onmath Co. (Chengdu, China), following the manufacturer's standard protocol.

Quantification of Gene Expression Levels

The eXpress was used for transcripts quantification. An eXpress is a streaming tool for quantifying the abundances of a set of target sequences from sampled subsequences. A probabilistic model developed for RNA-seq is the underlying model of eXpress. In general, clean reads were aligned to reference transcripts of moso bamboo genome, and quantification was conducted based on alignments with the probabilistic model. As a result, estimated read count and transcript per million (TPM) ($\mathbf{TPM_i} = \frac{\mathbf{q_i}/\mathbf{l_i}}{\sum_j (\mathbf{q_j}/\mathbf{l_j})} * \mathbf{10^6}$, in which $\mathbf{q_i}$ indicates reads mapped to the transcript, $\mathbf{l_i}$indicates the transcript length, and $\sum_j (\mathbf{q_j}/\mathbf{l_j})$ denotes the sum of mapped reads to transcript normalized by transcript length), were obtained for every single transcript in every sample, even for the multi-isoform from the same gene.

Differential Expression Test

A differential expression test was conducted using DESeq R packages according to the packages manual. Raw count data were prepared using the custom perl script based on results of eXpress software and were imported into the DESeq framework.

Information on experiment design was also imported into the DESeq framework to form a Count Data Set. Filtering was performed to remove transcripts in the lowest 40% quantile of the overall sum of counts (irrespective of biological condition) to increase the differential expression transcript detection rate. The estimated SizeFactors function was used to estimate the effective library size in order to normalize the transcripts counts. The estimate Dispersions function was used to estimate dispersion. The nbinomTest function was used to see whether there is a differential expression between two conditions. FDRs were controlled using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995).

Gene Ontology Enrichment Analysis

Gene ontology (GO) enrichment analysis of the DEGs was implemented using the GOseq R packages based on Wallenius noncentral hypergeometric distribution (Young et al., 2010), which can adjust for gene length bias in DEGs.

Plasmid Construction and Transformation in *N. tabacum*

Based on the coding sequence of *PeLEA14* (PH01001932G0350), a pair of primers, LEA-F: 5′-CCAAGCTTATGGCGCAGCTGATGGACAA-3′ (*Hind*III) and LEA-R: 5′-CGGGATCCTTAGAAGATGGTGGAGAGCGT-3′ (*Bam*HI), containing restriction enzyme sites (underlined), were designed to amplify the coding sequence of *PeLEA14* from cDNA using Phanta Max Super-Fidelity DNA Polymerase (Vazyme Biotech Co., Nanjing, China). The amplified fragment was double-digested and ligated onto the corresponding sites of the pGSA-1403 vector by T4 DNA ligase. The resulting construct 35S::pGSA1403-PeLEA14 was introduced into the *Agrobacterium tumefaciens* strain GV3101 and then transformed into *N. tabacum* using the leaf disk method.

The T_0 seedlings were screened on 1/2 MS medium supplied with kanamycin (50 μg/ml). The seedlings resistant to kanamycin were transplanted into pots with soil, and the positive transgenic plants were further verified by PCR. The T3 homozygous positive lines were used for further investigation.

Evaluation of Abiotic Stress Tolerance

For stress tolerance assays at the seedling stage, the WT and transgenic seedlings were placed in cultivation bottles with 1/2 MS solid medium containing 200 mM mannitol and 100 mM NaCl, respectively. The cultivation bottles were incubated with a cycle of 16 h/8 h of light (24°C)/dark (22°C).

For stress tolerance assays at the mature stage, 3-day acclimatization was performed for the WT and transgenic seedlings (with 4-5 leaves) grown on a 1/2 MS solid medium. Then, the plants were transplanted into pots (one plant per pot) containing an equal amount of sterilized soil and grown in an incubator with a cycle of 16 h/8 h of light (24°C)/dark (22°C). After 7 days, the plants with a similar growth status were selected for stress treatment. For natural drought treatment, the soil was fully watered, and then the watering was stopped for several days. For salt treatment, the soil was fully infiltrated with water, and then the plants were irrigated by applying enough 300 mM NaCl solution into the tray of cultivation pots and keeping the soil moist during the processing. The morphological changes of the plants were constantly observed and photographed.

Measurements of Physiological Parameters Related to Stress Responses

Chlorophyll was extracted from leaf tissue in 95% ethanol as previously described (Palta, 1990). Proline was measured following the modified method of acidic ninhydrin reaction as reported previously (Bates et al., 1973). The enzyme liquid was extracted for the determination of superoxide dismutase (SOD), peroxidase (POD), and catalase (CAT) activities as well as malondialdehyde (MDA) content. Detailed descriptions of these assays were elaborated by Du and Bramlage (1992) and Zheng et al. (2007). Three replicates were executed for these experiments. Samples used for physiological index measurements were obtained through drought treatment (withhold watering) for 10 days and salt treatment (300 mM NaCl) for 7 days, respectively.

RESULTS AND DISCUSSION

RNA Sequencing, Mapping, and Transcript Quantification

To understand transcriptomic responses of moso bamboo to abiotic stresses, RNA-seq analysis of bamboo leaves under dehydration and cold stresses was performed. We generated 33,498,161 to 58,049,004 clean reads for each sample by pair-end sequencing, corresponding to approximately 5.02 to 8.71 giga base pairs (Gb), with an average of 6.48 Gb (**Supplementary Table S1**). All clean reads were mapped to moso bamboo genome (Peng et al., 2013). Notably, ∼66.6 to 71.9% of clean reads could be uniquely mapped and used for further analysis.

Subsequently, the transcripts per million (TPM) value was calculated to quantify the transcript abundance in each sample. To evaluate the repeatability of biological repeats, the correlation coefficient among samples was calculated based on the TPM values. The results showed that the correlation coefficient among the two biological repeats was more than 0.9, which was significantly higher than between other samples (**Figure 1A**). The cluster analysis also showed that the two biological repeats of each treatment were clustered together and separated from other samples (**Figure 1A**). These results suggested that the obtained RNA-seq data have good repeatability, providing a guarantee for subsequent data mining.

Identification of DEGs Related to Dehydration and Cold Responses

We used false-discovery rate (FDR)-corrected *p*-value (adjusted *p*-value, padj), which was defined by Benjamini and Hochberg (1995), to identify DEGs between treatments. When set padj$<$0.01 as the threshold, 661 to 10,054 DEGs between CK and each dehydration and cold treatment were obtained (**Figures 1B,C**; **Supplementary Figure S1**). For three dehydration time points (2 h, 4 h, and 8 h), 10,054,

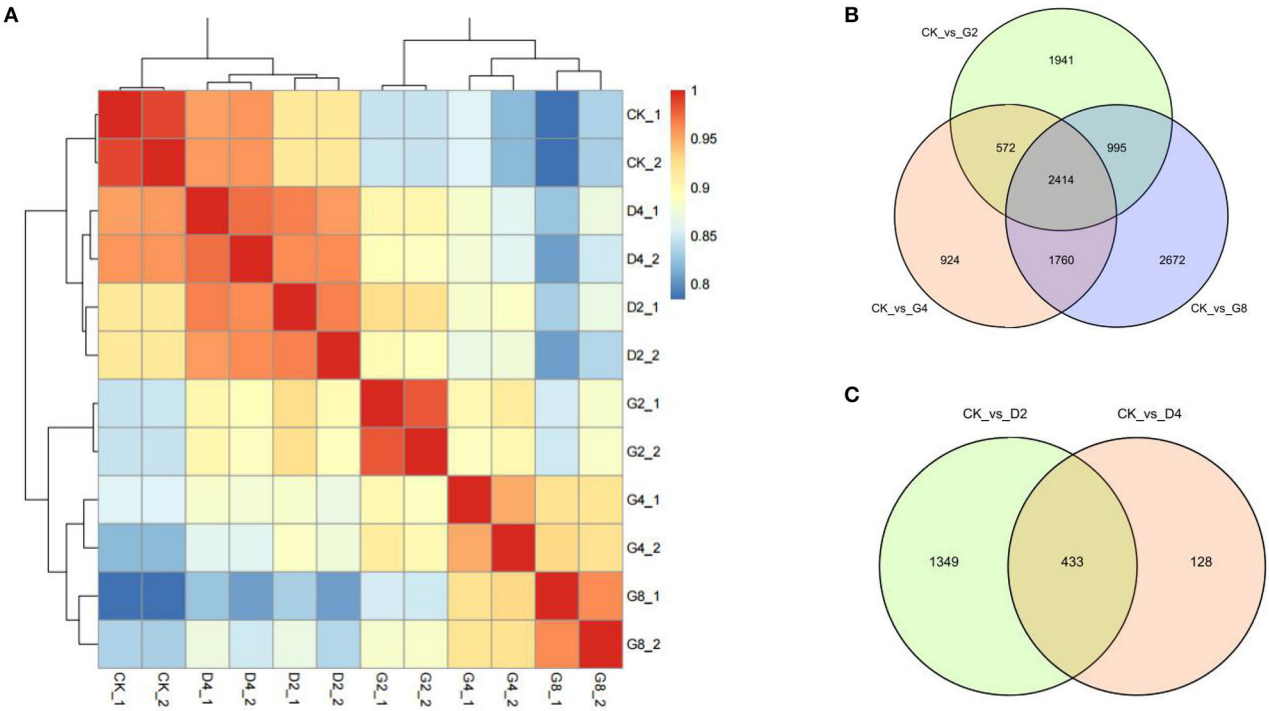

FIGURE 1 | RNA-seq of moso bamboo leaves under dehydration and cold treatments. **(A)** Correlation and clustering of each sample. **(B,C)** Venn diagram of differentially expressed genes (DEGs) under two treatments. CK, control; D2 and D4 indicated 2 h and 4 h of cold treatment, and G2, G4, and G8 indicated 2 h, 4 h, and 8 h of dehydration treatment, respectively.

5,967, and 8,398 DEGs were identified, in which 5,388, 2,784, and 4,109 genes were upregulated, and 4,666, 3,183, and 4,289 genes were downregulated, respectively (**Figure 1B**, **Supplementary Figures S2A,B**). Notably, 1,364 and 1,027 genes were upregulated or downregulated at all time points of dehydration (**Supplementary Figures S2A,B**). For cold treatments, 2,322 and 661 DEGs were found at 2 h and 4 h of cold treatment, in which 1,418 and 565 DEGs were upregulated, and 904 and 96 DEGs were downregulated, respectively. A total of 433 DEGs showed differential expression at all two time points (**Figure 1C**, **Supplementary Figures S2C,D**). We noticed that the DEG number under cold treatment was much less than that obtained under dehydration. This phenomenon was also found in the previous study on transcriptomic responses to cold of moso bamboo (Liu et al., 2020). They only found <100 DEGs at early stages (0.5 and 1 h) of cold treatment (0.5 and 1 h).

Functional Annotation of DEGs by GO Enrichment

We analyzed GO classification to categorize the functions of DEGs during dehydration and cold stresses (**Supplementary Table S2**). GO includes three main ontologies, namely, molecular function (MF), biological process (BP), and cellular component (CC). Under dehydration stress, the GO terms of BP in which the DEGs enriched are related to some basic BPs, such as carbohydrate metabolic process, proteolysis, translation, and transport, as well as those closely related to stress responses, such as response to oxidative stress, signal transduction, cell redox homeostasis, and response to stress (**Figures 2A,B**). For the MF, the predominant GO terms are nucleotide-binding, sequence-specific DNA binding, heme binding, binding, structural constituent of ribosome, RNA binding, hydrolase activity, hydrolyzing O-glycosyl compounds, iron ion binding, transporter activity, calcium ion binding, etc. (**Figure 2C**). For the CC, intracellular, ribosome, cytoplasm, extracellular region, nucleosome, cell wall, and photosystem II are among the most DEG-enriched GO terms (**Figure 2D**).

Under cold treatment, GO terms of BP containing a predominant number of DEGs are carbohydrate metabolic process, proteolysis, response to oxidative stress, protein dephosphorylation, lipid metabolic process, and transport (**Figure 3A**). There are five GO terms apparently related to environmental responses, i.e., response to oxidative stress, response to stress, response to auxin, response to desiccation, and response to water (**Figure 3B**). For the MF, DEGs were enriched in GO terms of sequence-specific DNA binding, heme binding, ADP binding, calcium ion binding, iron ion binding, hydrolase activity, oxidoreductase activity, protein dimerization activity, and binding (**Figure 3C**). For the CC, DEGs were enriched in GO terms of intracellular, ribosome, cytoplasm, extracellular region, cell wall, photosystem II, and apoplast (**Figure 3D**).

In previous cold-responsive transcriptomic studies, 89 and 79 DEGs were identified at 0.5 h and 1 h cold treatment, respectively, and 125 DEGs were annotated with GO terms.

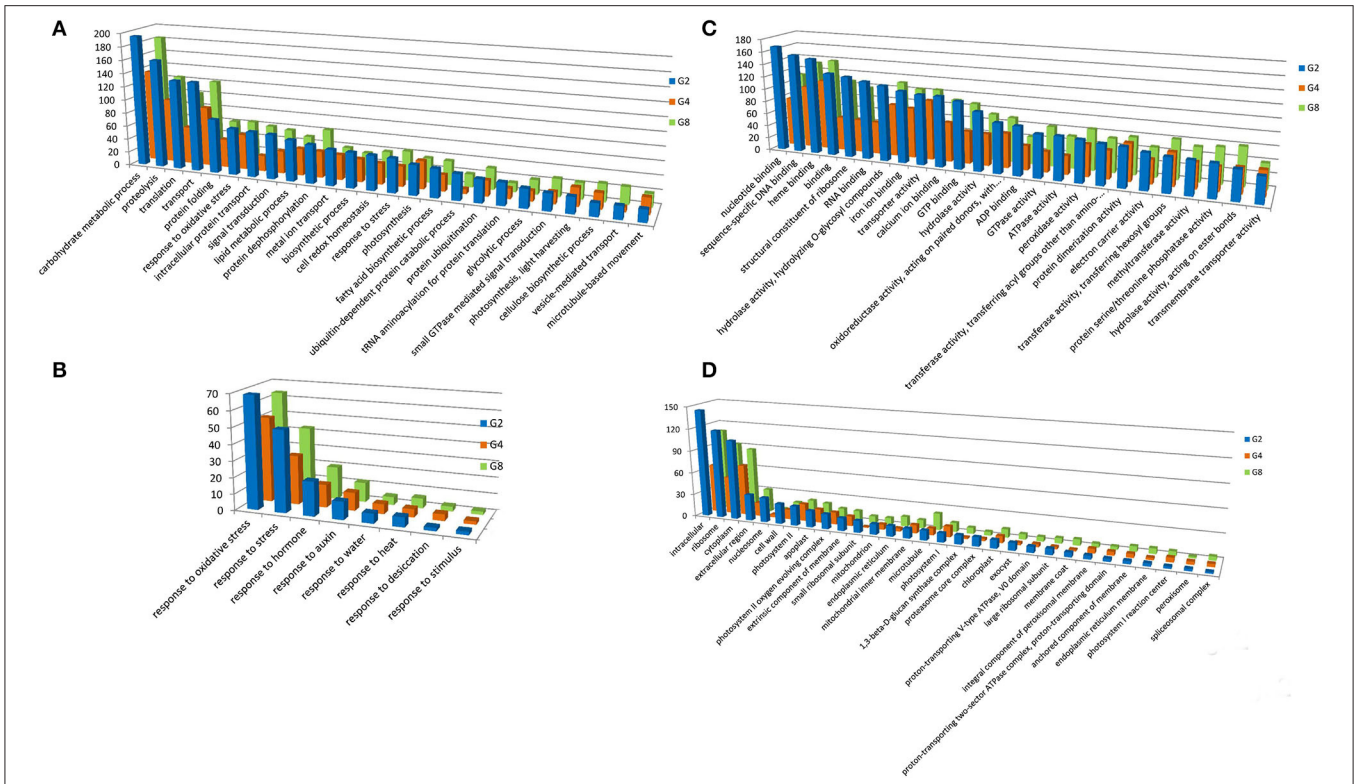

FIGURE 2 | Gene ontology (GO) terms with predominant DEG number under dehydration treatment. GO terms of BP **(A)**, MF **(C)**, and CC **(D)** were shown. **(B)** Indicates GO terms of MF, which are closely related to stress responses. Y-axis indicated the DEG number assigned in a GO term. CK, control; G2, G4, and G8 indicated 2 h, 4 h, and 8 h of dehydration treatment, respectively.

Among them, 21 DEGs were assigned GO terms of BP, including metal ion transport, response to UV-B, flavonoid biosynthetic process, response to light stimulus, response to salt stress, calcium ion transport, transcription, and response to salt stress. A total of 70 DEGs were assigned with MF GO terms. The GO terms with predominant DEG numbers included calcium ion binding (6 DEGs), DNA binding (11 DEGs), iron ion binding (5 DEGs), sequence-specific DNA binding (12 DEGs), protein serine/threonine kinase activity (3 DEGs), and protein serine/threonine phosphatase activity (DEGs). For the CC, the major GO terms were plastid (8 DEGs), integral component of membrane (5 DEGs), plasma membrane (4 DEGs), and cytoplasmic vesicle (3 DEGs). These results showed significantly different responses to cold at different time points. Interestingly, sequence-specific DNA binding, calcium ion binding, and iron ion binding are predominant GO terms in our and previous studies, indicating their general and important roles in cold responses of moso bamboo.

Differentially expressed genes enriched in these GO terms reflected transcriptomal responses of *P. edulis* to dehydration and cold stresses. For example, response to oxidative stress is one of the predominant GO terms under both stresses, indicating that oxidation is one of the most serious stresses under the two treatments (**Supplementary Tables S2**, **S3**). DEGs involved in this GO term mainly encode catalase, glutathione peroxidase, and a group of plant peroxidase (**Supplementary Table S4**). These enzymes were well known to scavenge excess reactive oxygen species (ROS) generated by stress (Zhang et al., 2019).

Response to stress is another common GO term. Under dehydration treatment, 24 DEGs enriched in this GO term were simultaneously found in three time points (**Supplementary Table S5**). Genes encoding ABA/water deficit stress (WDS)-induced protein, LEA protein (including dehydrin), and universal stress protein A (UspA) are among the most abundant (**Supplementary Table S5**). ABA/WDS-induced protein, also known as ASR, is a family of plant proteins induced by water deficit stress (Padmanabhan et al., 1997), or ABA stress and ripening (Canel et al., 1995) and has been extensively studied. The maize ZmASR1 acts as both a transcriptional regulator and a chaperone-like protein (Virlouvet et al., 2011). The rice *ASR5* was reported to be involved in response to drought stress by regulating ABA biosynthesis, promoting stomatal closure, and also acts as a chaperone-like protein possibly preventing drought stress-related proteins from inactivation (Li et al., 2017). LEA proteins are well known to play a protective role during exposure to different abiotic stresses. The Universal Stress Protein (USP) contains a Universal Stress Protein A domain comprising 140–160 highly conserved residues and is significantly overexpressed under multiple unfavorable environmental stresses (Kvint et al., 2002; Ndimba et al.,

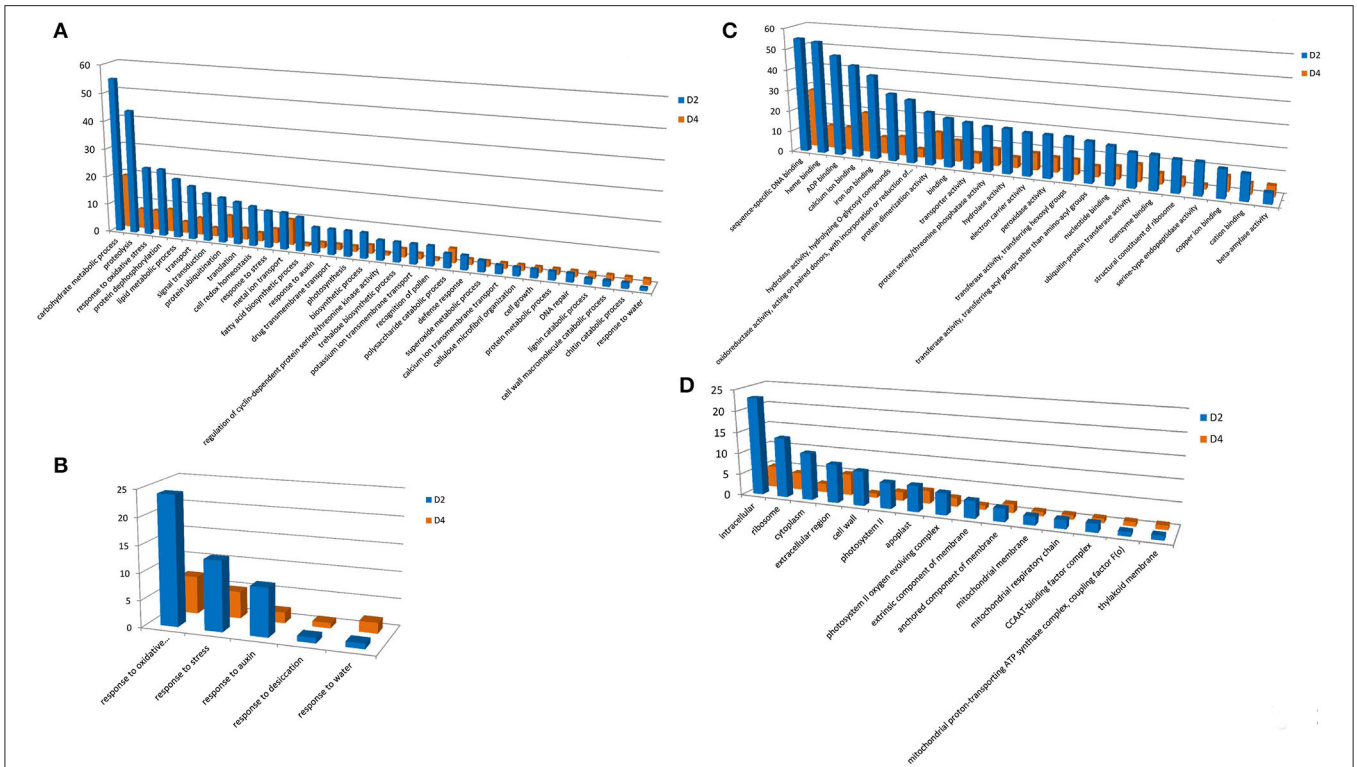

FIGURE 3 | Gene ontology terms with predominant DEG number under cold treatment. GO terms of BP **(A)**, MF **(C)**, and CC **(D)** were shown. **(B)** Indicates GO terms of MF, which are closely related to stress responses. Y-axis indicated the DEG number assigned in a GO term. CK, control; D2 and D4 indicated 2 h and 4 h of cold treatment, respectively.

2005; Persson et al., 2007). Under cold treatment, four DEGs, PH01000696G0320, PH01001317G0160, PH01003240G0100, and PH01004855G0070, with GO term of response to stress were found in both time points (**Supplementary Table S5**). They encode ABA/WDS-induced protein, LEA protein 3, dehydrin, and calmodulin-binding protein-like (SARD1) (**Supplementary Table S5**) and also showed differential expression levels under dehydration treatment, indicating that these genes may play roles in both dehydration and cold stresses.

We also analyzed the DEGs enriched in sequence-specific DNA binding of GO category MF, as they usually include transcription factors. Crossing the three time points of dehydration treatment, 63 DEGs were assigned in this GO term. They all encode putative transcription factors, including 17 basic-leucine zipper (bZIP), 18 WRKY, six heat shock factor (HSF) type, 15 homeodomain, and seven GATA type (**Supplementary Table S6**). Under cold treatment, 25 DEGs enriched in sequence-specific DNA binding were detected in the two time points, including six bZIP, 18 WRKY, and one HSF (**Supplementary Table S5**). Eleven DEGs were simultaneously found in all time points of both treatments (**Supplementary Tables S4**, **S5**), indicating that they may play dual roles in responding to dehydration and cold stresses. Under unfavorable environmental conditions, plants have evolved diverse stress-response mechanisms, such as induction of defense gene expression. Therefore, activating the overall defense reaction ultimately contribute to stress response ad tolerance (Fraire-Velázquez et al., 2011). Transcription factors (TFs) are important regulators for the control of gene expression in all living organisms and play crucial roles in plant development, cell cycling, cell signaling, and stress response (Gonzalez, 2016). Extensive studies proved that TF families, such as AP2/ERF, MYB, NAC, and WRKY, are crucial regulators of various stress-responsive genes (Wang et al., 2016). Therefore, The TFs encoded DEGs identified in this study may help to obtain a better understanding of the mechanisms of abiotic stress response of moso bamboo and could be considered an ideal choice for genetic engineering in order to enhance stress tolerance.

Time-Series Analysis of DEGs

To characterize dynamic expression patterns of DEGs following the time points of dehydration and cold treatments, an R package Mfuzz (Kumar and Futschik, 2007) was employed to perform time-series "soft clustering" based on TPM values (**Figures 4A,B**). DEGs with padj <0.01 between at least two time points were used as input for the clustering. The number of clusters was set to 16, and the fuzzifier coefficient was set to 1.55.

We focused on clusters in which the two biological repeats are at similar levels and showed stress-responsive patterns. For dehydration treatment, clusters 11, 13, and 15 showed a quick upregulation pattern (type I) at 2 h, whereas DEGs

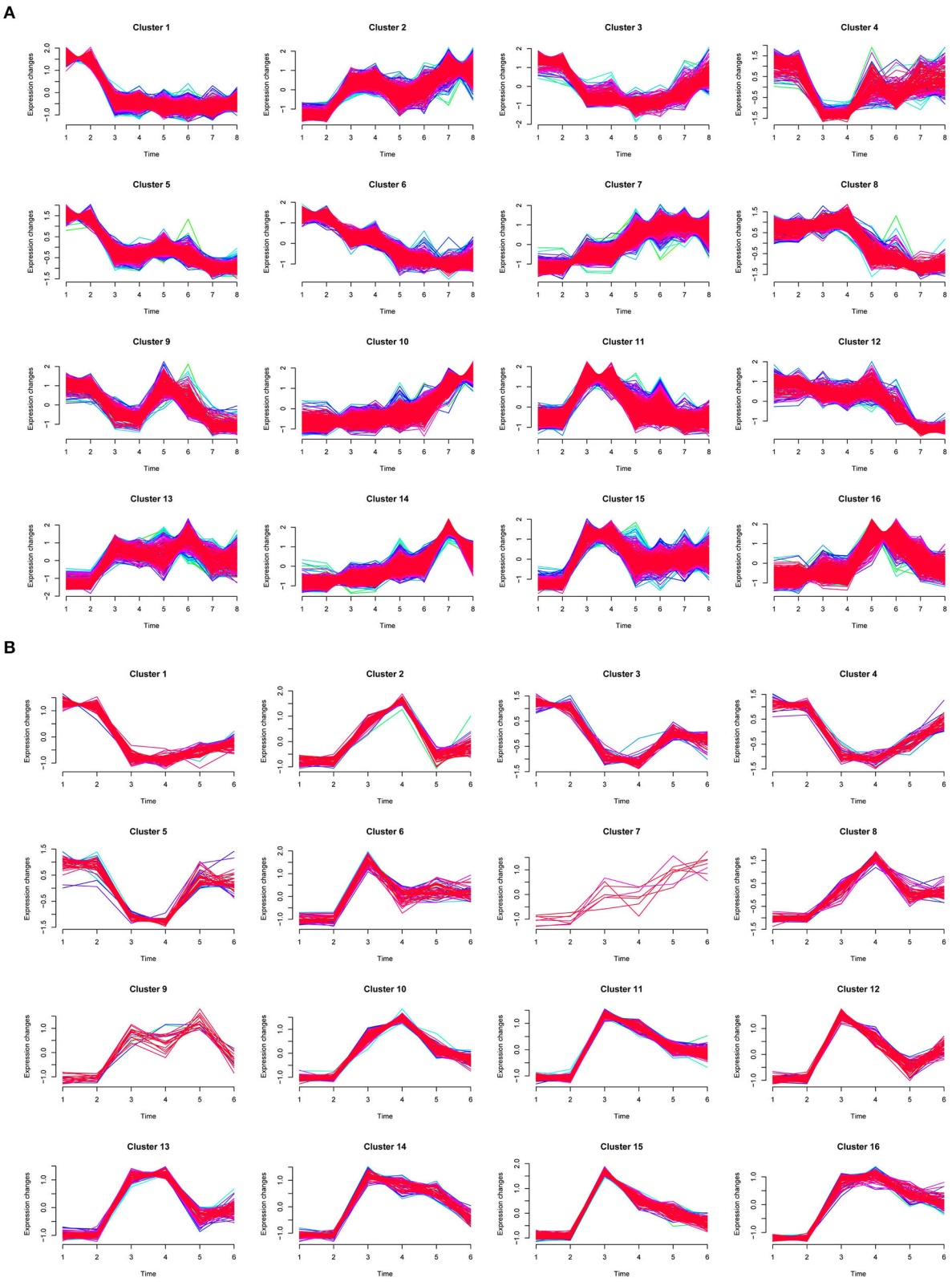

FIGURE 4 | Time-series clustering of DEGs for dehydration **(A)** and cold **(B)** treatments. In **(A)**, 1 and 2, 3 and 4, 5 and 6, and 7 and 8 on *X*-axis indicated two biological repeats of CK, G2, G4, and G8, respectively. In **(B)**, 1 and 2, 3 and 4, and 5 and 6 on *X*-axis indicated two biological repeats of CK, D2, and D4, respectively.

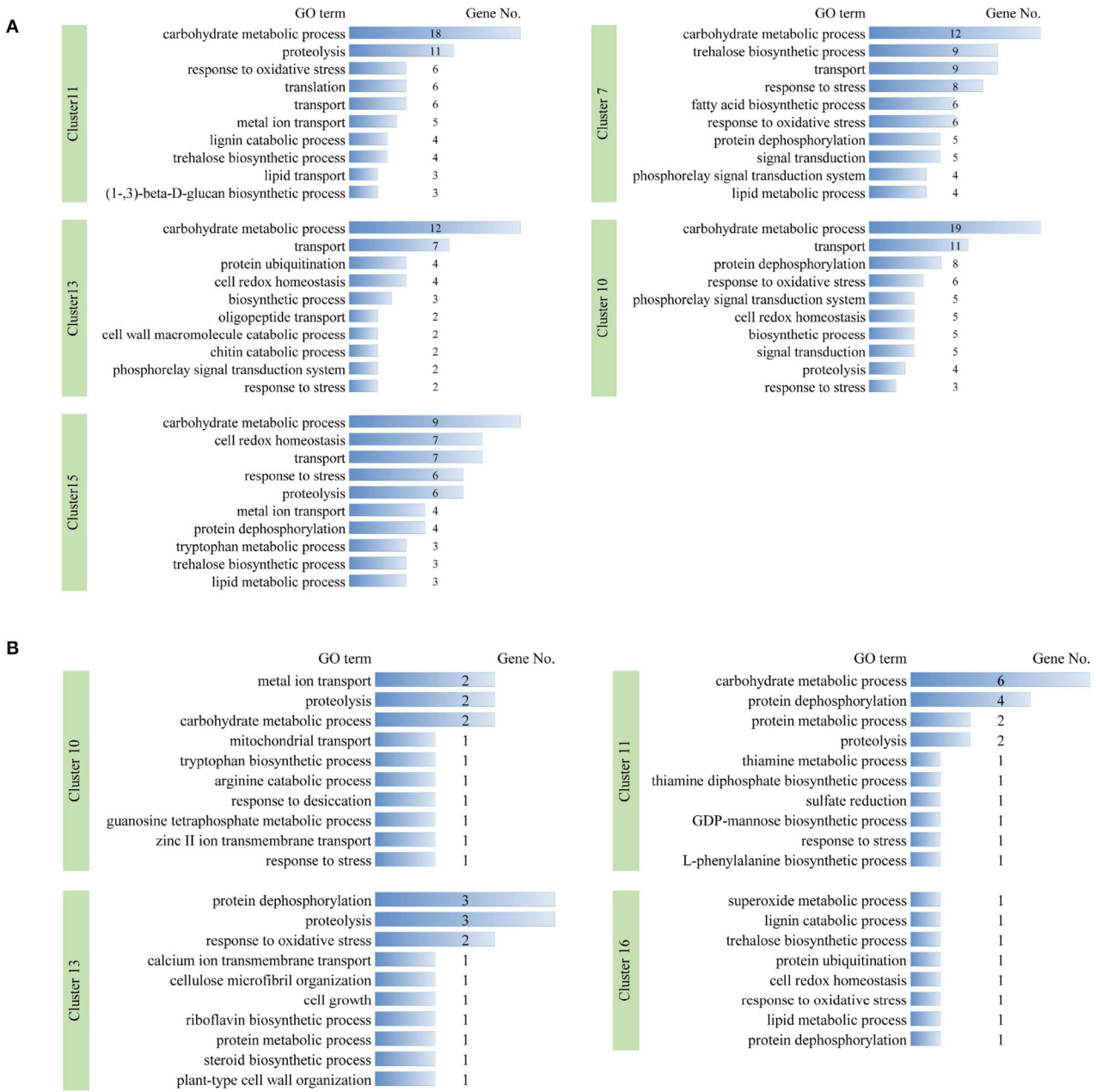

FIGURE 5 | Gene ontology terms with predominant numbers of DEGs in stress-responsive timer-series clusters under dehydration **(A)** and cold **(B)** stresses, respectively. DEG number in each GO term is shown.

in clusters 7 and 10 were gradually upregulated (type II); clusters 1, 3, 4, and 5 exhibited early quick downregulation patterns (type III), and clusters 6 and 8 were gradually downregulated (type IV), respectively (**Figure 4A**). We further examined the functions of DEGs in upregulation clusters. GO terms with predominant gene numbers for type I and II clusters included carbohydrate metabolic process, proteolysis, response to oxidative stress, transport, protein ubiquitination, cell redox homeostasis, response to stress, trehalose biosynthetic process, and fatty acid biosynthetic process (**Figure 5A**). For cold treatment, clusters 10, 11, 13, and 16 showed type I upregulated patterns, and clusters 1 and 3 exhibited type III downregulated patterns (**Figure 4B**). GO terms for genes in these clusters included carbohydrate metabolic process, metal ion transport, protein dephosphorylation, proteolysis, and response to oxidative stress (**Figure 5B**). These data provided a resource to characterize gene sets showing similar patterns responsive to dehydration and cold stress.

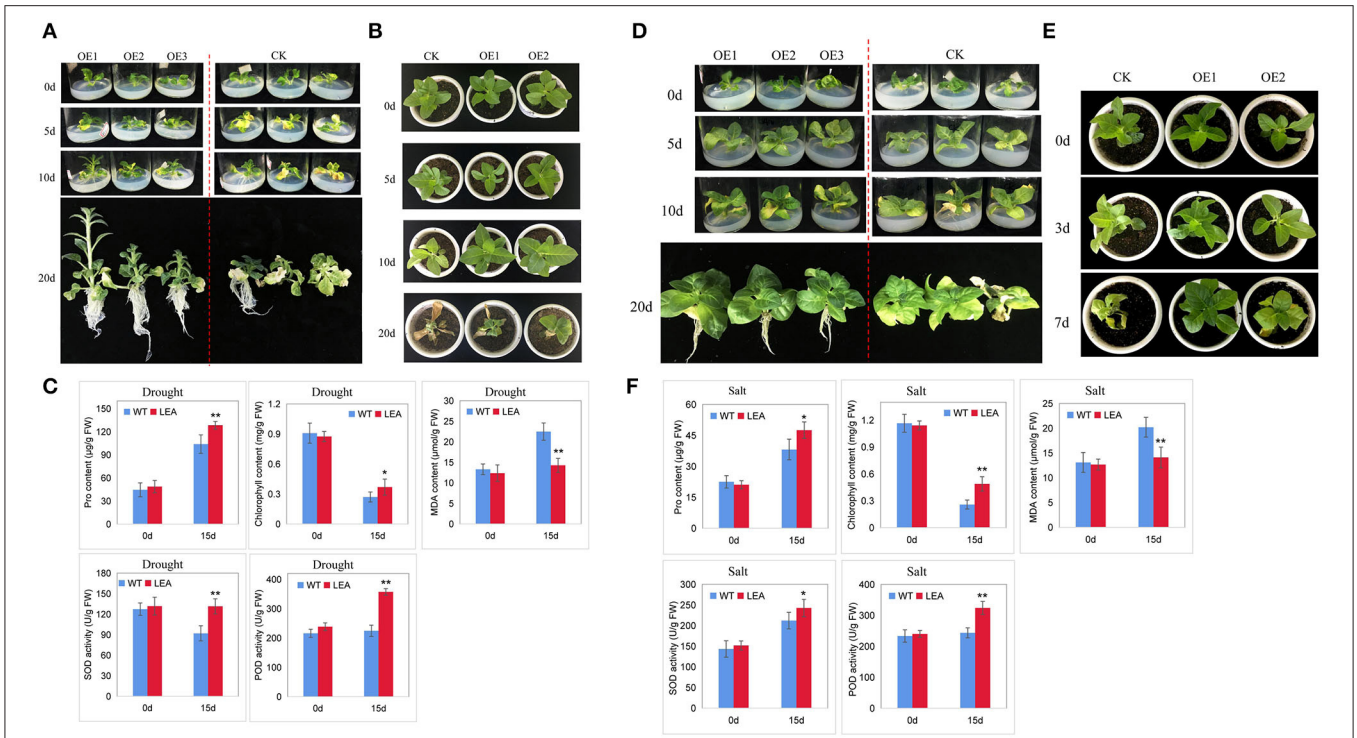

FIGURE 6 | Overexpression of *PeLEA14* in tobacco. **(A–C)** indicate evaluation of tolerance of wild type (WT) and transgenic lines (OE) to drought stress at seedling (1/2 MS solid medium containing 200 mM mannitol) and adult (natural drought by withholding watering) stages and stress-related physiological parameters, respectively. **(D–F)** indicate evaluation of tolerance of WT and transgenic lines (OE) to salt stress at seedling (1/2 MS solid medium containing 100 mM NaCl) and adult (applying 300 mM NaCl solution into the tray) stages and stress-related physiological parameters, respectively.

Identify Co-Expressed Hub Genes Based on Pearson'S Correlations

To identify potential key genes involved in dehydration and cold responses, we calculated Pearson's correlations for all gene pairs crossing all time points of two treatments. The gene filtering was performed with a cutoff value of 0.85, and the degree (number of associated genes fitting the cutoff) for each qualified gene was calculated. The top 50 genes were identified as potential hub genes, including those encoding 9-cis-epoxycarotenoid dioxygenase (NCED), LEA protein, probable indole-3-acetic acid-amido synthetase, protein phosphatase 2C, ubiquitin-activating enzyme, as well as bZIP, NAC, F-box, and SBP-box transcription factors (**Supplementary Table S8**). Some of these are well known to function in stress responses. For example, NCED is the key enzyme for the biosynthesis of ABAs (Hauser et al., 2011), which plays a critical role in abiotic stress response (Nambara and Marion-Poll, 2005). These results, therefore, provide candidate core genes, especially those with unknown functions, involved in the abiotic stress response of moso bamboo.

Function Validation of Stress-Responsive Gene

To preliminarily test the effectiveness of the identified genes in abiotic stress tolerance, we selected a dehydration-responsive gene *PeLEA14* (PH01001932G0350) for functional validation. *PeLEA14* was slightly upregulated by cold but significantly induced by dehydration (**Supplementary Figure S3**). We overexpressed *PeLEA14* in tobacco. The positive transgenic lines were acquired from kanamycin resistance screening and PCR (**Supplementary Figure S4**). Artificially simulated drought treatment at the seedling stage was performed by growing plants on 1/2 MS solid medium containing 200 mM mannitol. The WT and transgenic lines were in similar status before treatment. After 5 days of treatment, the leaves of WT were slightly withered, and such symptoms became more serious by 10 days. At 20 days of treatment, almost all leaves of WT were withered, but the transgenic lines showed significantly better growth status and root system (**Figure 6A**). The natural drought treatment was also applied. Ten days after withholding water, the dehydration symptoms (chlorosis) could be observed on leaves of both WT and transgenic lines. However, the vegetative growth of the latter is better than that of the former. Twenty days after withholding water, the whole plant of WT was severely withered, whereas the growth status of transgenic lines was significantly better than WT (**Figure 6B**). We further evaluated physiological parameters related to stress responses. Under natural drought stress, the transgenic line exhibited significantly higher proline and chlorophyll content, lower MDA content, and higher SOD and POD activities (**Figure 6C**). We also evaluated tolerance to salt stress. Similar results to those under drought stresses were

obtained and indicated that the tobacco lines overexpressing *PeLEA14* showed better growth performance under salt stress at both seedlings and adult stages, as well as higher antioxidant capacity (**Figures 6D–F**).

Late embryogenesis abundant proteins constitute a family of hydrophilic proteins that are presumed to play a protective role during exposure to different abiotic stresses. They were initially classified into six subgroups on the basis of specific domains (Dure et al., 1989), and different researchers have also tried to use different classification methods (Tunnacliffe and Wise, 2007; Battaglia et al., 2008; Bies-Ethève et al., 2008; Hundertmark and Hincha, 2008; Shih et al., 2008; Battaglia and Covarrubias, 2013). *PeLEA14* encodes 151 amino acids and is a homolog of *AtLEA14* (AT1G01470, *E*-value = 5e-63) of *Arabidopsis* and belongs to the LEA_2 group (also be classified as group 5 by a different nomenclature). The typical LEA proteins, such as those of groups 1, 2, 3, and 4, are genuine hydrophilic and share specific sequence motifs within each group. However, group 5 LEAs lack significant signature motifs or consensus sequences, contain a significantly higher proportion of hydrophobic residues than typical LEA proteins, and therefore were considered "atypical" LEA proteins (Baker et al., 1988; Galau et al., 1993; Battaglia et al., 2008).

Although this group has not been extensively investigated, some reports indicated that they will accumulate in response to diverse stresses in plants (Baker et al., 1988; Piatkowski et al., 1990; Maitra and Cushman, 1994; Zegzouti et al., 1997; Kimura et al., 2003). In this study, we found that *PeLEA14* was significantly induced by dehydration (**Supplementary Figure S3**). Additionally, a transgenic sweet potato that overexpressed *IbLEA14* showed increased tolerance to osmotic and salt stress by enhancing lignification (Park et al., 2011). Overexpression of *SiLEA14* of foxtail millet improved tolerance to salt and drought stresses (Wang et al., 2014). Overexpression of *OsLea14-A* in rice improved tolerance to dehydration, high salinity, $CuSO_4$, and $HgCl_2$ (Hu et al., 2019). Our study showed that the overexpression of *PeLEA14* enhanced the tolerance of tobacco to drought and salt stresses, possibly through, at least in part, increasing antioxidant capacity. These results indicated that the "atypical" LEA_2 group proteins may play important roles in plant stress protection. The data set provided in this study will lay a foundation for future discovery of stress-tolerant genes in moso bamboo.

CONCLUSION

In this study, we generated an expression data set with respect to dehydration and cold responses of moso bamboo. With repeatable RNA-seq data and a strict screening cutoff, we identify a lot of DEGs. Combining comprehensive GO enrichment analysis, time-series analysis, and co-expression analysis, we identified DEGs closely related to dehydration and cold responses, stress-responsive DEGs with similar expression patterns, as well as potential core genes, which may play an important role in dehydration and/or cold stress. We used *PeLEA14* as an example to validate the function of the identified stress-related gene in tolerance to abiotic stresses, such as drought and salt. These data may be valuable for future excavation of key genes involved in abiotic stress responses and genetic improvement of moso bamboo.

DATA AVAILABILITY STATEMENT

All data present in this study can be found in the article or **Supplementary Materials**.

AUTHOR CONTRIBUTIONS

ZH: conceptualization, writing-original draft preparation, project administration, and funding acquisition. PZ and XZ: methodology. PZ and JQ: validation. WX and LS: data curation. All authors have read and agreed to the published version of the manuscript.

FUNDING

This study was supported by the International Cooperation Project (2022YFH0066) funded by the Science and Technology Department of Sichuan Province, China, and the Shuangzhi Plan funded by the Sichuan Agricultural University. The funders had no role in the design of this study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.960302/full#supplementary-material

Supplementary Table S1 | Statistics of RNA sequencing and mapping.

Supplementary Table S2 | Gene ontology (GO) enrichment of differentially expressed genes (DEGs) under dehydration treatment.

Supplementary Table S3 | Gene ontology enrichment of DEGs under cold treatment.

Supplementary Table S4 | DEGs under dehydration treatment and enriched in response to oxidative stress.

Supplementary Table S5 | DEGs under dehydration treatment and enriched in response to stress.

Supplementary Table S6 | Consensus DEGs at three time points of dehydration and enriched in GO term Sequence-specific DNA binding.

Supplementary Table S7 | Consensus DEGs at two time points of cold treatment and enriched in GO term sequence-specific DNA binding.

Supplementary Table S8 | List of hub genes.

Supplementary Figure S1 | Identification of differentially expressed genes (DEGs) at each of the time points.

Supplementary Figure S2 | Venn diagram of upregulated and downregulated DEGs under dehydration and cold treatment.

Supplementary Figure S3 | Expression patterns of *PeLEA14* under dehydration and cold treatments.

Supplementary Figure S4 | Identification of positive transgenic lines by PCR. M, DNA marker. The amplified fragment is ~500 bp, which fits the length of *PeLEA14* (456 bp).

REFERENCES

Baker, J., Steele, C., and Dure, L. (1988). Sequence and characterization of 6 LEA proteins and their genes from cotton. *Plant Mol. Bio*. 11, 277–291. doi: 10.1007/BF00027385

Barker, N. P., Clark, L. G., Davis, J. I., Duvall, M. R., Guala, G.F., Hsiao, C., et al. (2001). Phylogeny and subfamilial classification of the grasses (Poaceae). *Ann. Mo. Bot. Gard*. 88, 373–457. doi: 10.2307/3298585

Bates, L. S., Waldren, R. P., and Teare, I. D. (1973). Rapid determination of free proline for water-stress studies. *Plant Soil*. 39, 205–207. doi: 10.1007/BF00018060

Battaglia, M., and Covarrubias, A. A. (2013). Late Embryogenesis Abundant (LEA) Proteins in Legumes. *Front Plant Sci*. 25, 190. doi: 10.3389/fpls.2013.00190

Battaglia, M., Olvera-Carrillo, Y., Garciarrubio, A., Campos, F., and Covarrubias, A. A. (2008). The enigmatic LEA proteins and other hydrophilins. Plant Physiol. 148, 6–24. doi: 10.1104/pp.108.120725

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple hypothesis testing. *J. R. Stat. Soc. B*. 57, 289–300 doi: 10.1111/j.2517-6161.1995.tb02031.x

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, 107–24. doi: 10.1007/s11103-008-9304-x

Canel, C., Bailey-Serres, J. N., and Roose, M. L. (1995). Pummelo fruit transcript homologous to ripening-induced genes. *Plant Physiol*. 108, 1323–1324. doi: 10.1104/pp.108.3.1323

Cheng, X. R., Wang, Y. J., Xiong, R., Gao, Y. M., Yan, H. W., and Xiang, Y. (2020). A Moso bamboo gene *VQ28* confers salt tolerance to transgenic Arabidopsis plants. *Planta,*. 251, 99. doi: 10.1007/s00425-020-03391-5

Du, Z. Y., and Bramlage, W. J. (1992). Modified thiobarbituric acid assay for measuring lipid oxidation in sugar-rich plant tissue extracts. *J. Agric. Food Chem*. 40, 1566–1570. doi: 10.1021/jf00021a018

Dure, L., Crouch, M., Harada, J., Ho, T. H., Mundy, J., Quatrano, R., et al. (1989). Common amino acid sequence domains among the LEA proteins of higher plants. *Plant Mol. Biol.*12, 475–486. doi: 10.1007/BF00036962

Fraire-Velázquez, S., Rodríguez-Guerra, R., and Sánchez-Calderón, L. (2011). *Abiotic Stress Response in Plants-Physiological, Biochemical and Genetic Perspectives*. London, UK: IntechOpen. Abiotic and biotic stress response crosstalk in plants. 3–26.

Galau, G. A., Wang, H. Y C., and Hughes, D. W. (1993). Cotton Lea5 and Lea14 encode a typical late embryogenesis-abundant proteins. *Plant Physiol*. 101, 695–696. doi: 10.1104/pp.101.2.695

Gao, Y. M., Liu, H. L., Zhang, K. M., Li, F., Wu, M., and Xiang, Y. (2021). A moso bamboo transcription factor, Phehdz1, positively regulates the drought stress response of transgenic rice. *Plant Cell Rep*. 40, 187–204. doi: 10.1007/s00299-020-02625-w

Gonzalez, D. H. (2016). "Introduction to transcription factor structure and function", in *Plant Transcription Factors*, ed. D. H. Gonzalez (Boston, MA: Academic Press), 3–11. doi: 10.1016/B978-0-12-800854-6.00001-4

Hauser, F., Waadtl, R., and Schroeder, J. I. (2011). Evolution of abscisic acid synthesis and signaling mechanisms. *Curr. Biol*. 21, R346–R355. doi: 10.1016/j.cub.2011.03.015

Hu, T.Z., Liu, Y.L., Zhu, S.S., Qin, J., Li, W.P., and Zhou, N. (2019). Overexpression of *OsLea14*-A improves the tolerance of rice and increases Hg accumulation under diverse stresses. *Environ. Sci. Pollut. Res. Int*. 26, 10537–10551. doi: 10.1007/s11356-019-04464-z

Huang, R., Gao, H.Y., Liu, J., and Li, X.P. (2022). WRKY transcription factors in moso bamboo that are responsive to abiotic stresses. *J. Plant Biochem. Biotechnol*. 31, 107–114. doi: 10.1007/s13562-021-00661-5

Huang, Z., Jin, S. H., Guo, H. D., Zhong, X. J., He, Jiao., Li, Xi., et al. (2016b). Genome-wide identification and characterization of TIFY family genes in Moso Bamboo (*Phyllostachys edulis*) and expression profiling analysis under dehydration and cold stresses. *PeerJ*. 4, e2620. doi: 10.7717/peerj.2620

Huang, Z., Zhong, X. J., He, J., Jin, S. H., Guo, H. D., Yu, X. F., et al. (2016a). Genome-Wide Identification, Characterization, and Stress Responsive Expression Profiling of Genes Encoding LEA (Late Embryogenesis Abundant) Proteins in Moso Bamboo (*Phyllostachys edulis*). *PLoS ONE*. 11, e0165953. doi: 10.1371/journal.pone.0165953

Hundertmark, M., and Hincha, D. K. (2008). LEA (Late Embryogenesis Abundant) proteins and their encoding genes in *Arabidopsis thaliana*. *BMC Genomics*. 9, 118. doi: 10.1186/1471-2164-9-118

Jin, K. M., Zhuo, R. Y., Xu, D., Wang, Y. J., Fan, H. J., Huang, B. Y., et al. (2020). Genome-wide identification of the expansin gene family and its potential association with drought stress in Moso Bamboo. *Int. J. Mol. Sci*. 21, 9491. doi: 10.3390/ijms21249491

Kimura, M., Yamamoto, Y. Y., Seki, M., Sakurai, T., Abe, T., Yoshida, S., et al. (2003). Identification of Arabidopsis genes regulated by high light-stress using cDNA microarray. *Photochem. Photobiol*. 77, 226–233. doi: 10.1562/0031-8655(2003)0770226IOAGRB2.0.CO2

Kumar, L., and Futschik, M. E. (2007). Mfuzz: a software package for soft clustering of microarray data. *Bioinformation*. 2, 5–7. doi: 10.6026/97320630002005

Kvint, K., Nachin, L., Diez, A., and Nyström, T. (2002). The bacterial universal stress protein: function and regulation. *Curr. Opin. Microbiol*. 6, 140–145. doi: 10.1016/S1369-5274(03)00025-0

Li, J., Li, Y., Yin, Z., Jiang, J., Zhang, M., Guo, X., et al. (2017). OsASR5 enhances drought tolerance through a stomatal closure pathway associated with ABA and H_2O_2 signalling in rice. *Plant Biotechnol. J*. 15, 183–196. doi: 10.1111/pbi.12601

Liu, Y. Y., Wu, C., Hu, X., Gao, H. Y., Wang, Y., Hong, Luo., et al. (2020). Transcriptome profiling reveals the crucial biological pathways involved in cold response in Moso bamboo (*Phyllostachys edulis*). *Tree Physiol*. 40, 538–556. doi: 10.1093/treephys/tpz133

Lobovikov, M, Paudel, S, Piazza, M, Ren, H, and Wu, J. (2005). *World Bamboo Resources: A Thematic Study Prepared in the Framework of the Global Forest Resources Assessment*. Rome: Food and Agriculture Organization of the United Nations (FAO).

Maitra, N., and Cushman, J. (1994). Isolation and Characterization of a Drought-Induced Soybean cDNA Encoding a D95 Family Late-Embryogenesis-Abundant Protein. *Plant Physiol*. 106, 805–806. doi: 10.1104/pp.106.2.805

Nambara, E., and Marion-Poll, A. (2005). Abscisic acid biosynthesis and catabolism. *Annu. Rev. Plant Biol*. 56, 165–185. doi: 10.1146/annurev.arplant.56.032604.144046

Ndimba, B. K., Chivasa, S., Simon, W. J., and Slabas, A. R. (2005). Identification of Arabidopsis salt and osmotic stress responsive proteins using two-dimensional difference gel electrophoresis and mass spectrometry. *Proteomics*. 5, 4185–4196. doi: 10.1002/pmic.200401282

Padmanabhan, V., Dias, D. M., and Newton, R. J. (1997). Expression analysis of a gene family in loblolly pine (*Pinus taeda* L.) induced by water deficit stress. *Plant Mol. Biol*. 35, 801–807. doi: 10.1023/A:1005897921567

Palta, J.P. (1990). Leaf chlorophyll content. *Remote Sens. Rev*. 5, 207–213. doi: 10.1080/02757259009532129

Park, S. C., Kim, Y. H., Jeong, J. C., Kim, C. Y., Lee, H. S., Bang, J. W., et al. (2011). Sweetpotato *late embryogenesis abundant 14* (*IbLEA14*) gene influences lignifications and increases osmotic- and salt stress-tolerance of transgenic calli. *Planta*. 233, 621–634. doi: 10.1007/s00425-010-1326-3

Peng, Z. H., Lu, Y, Li, L. B., Zhao, Q, Feng, Q, Gao, Z., et al. (2013). The draft genome of the fast-growing nontimber forest species moso bamboo (*Phyllostachys heterocycla*). *Nat. Genet*. 45, 456–461. doi: 10.1038/ng.2569

Persson, O., Valadi, A., Nyström, T., and Farewell, A. (2007). Metabolic control of the *Escherichia coli* universal stress protein response through fructose-6-phosphate. *Mol. Microbiol*. 65, 968–978. doi: 10.1111/j.1365-2958.2007.05838.x

Piatkowski, D., Schneider, K., Salamini, F., and Bartels, D. (1990). Characterization of Five Abscisic Acid-Responsive cDNA Clones Isolated from the Desiccation-Tolerant Plant *Craterostigma plantagineum* and Their Relationship to Other Water-Stress Genes. *Plant Physiol*. 94, 1682–1688. doi: 10.1104/pp.94.4.1682

Shih, M. D., Hoekstra, F. A., and Hsing, Y. I. C. (2008). Late embryogenesis abundant proteins. *Adv. Bot. Res*. 48, 211–255. doi: 10.1016/S0065-2296(08)00404-7

Tunnacliffe, A., and Wise, M. (2007). The continuing conundrum of the LEA proteins. *Naturwissenschaften*. 94, 791-812. doi: 10.1007/s00114-007-0254-y

Virlouvet, L., Jacquemot, M. P., Gerentes, D., Corti, H., Bouton, S., Gilard, F., et al. (2011). The *ZmASR1* protein influences branched-chain amino acid biosynthesis and maintains kernel yield in maize under water-limited conditions. *Plant Physiol*. 157, 917–936. doi: 10.1104/pp.111.176818

Wang, H., Wang, H., Shao, H., and Tang, X. (2016). Recent advances in utilizing transcription factors to improve plant abiotic stress tolerance by transgenic technology. *Front. Plant. Sci*. 7, 67. doi: 10.3389/fpls.2016.00067

Wang, M. Z., Li, P., Li, C., Pan, Y., Jiang, X., Zhu, D., et al. (2014). SiLEA14, a novel atypical LEA protein, confers abiotic stress resistance in foxtail millet. *BMC Plant Biol*. 14, 290. doi: 10.1186/s12870-014-0290-7

Wu, M., Liu, H. L., Han, G. M., Cai, R. H., Pan, F., and Yan, X. (2017). A moso bamboo WRKY gene PeWRKY83 confers salinity tolerance in transgenic Arabidopsis plants. *Sci. Rep*. 7, 11721. doi: 10.1038/s41598-017-10795-z

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

Zegzouti, H., Jones, B., Marty, C., Lelievre, J.M., Latche, A., and Pech, J.C. (1997). ER5, a tomato cDNA encoding an ethylene-responsive LEA-like protein: characterization and expression in response to drought, ABA and wounding. *Plant Mol. Biol.* 35, 847–854. doi: 10.1023/A:1005860302313

Zhang, L. P., Wu, M., Teng, Y. J., Jia, S. H., Yu, D. S., Wei, T., et al. (2019). Overexpression of the *Glutathione Peroxidase 5* (*RcGPX5*) Gene From *Rhodiola crenulata* Increases Drought Tolerance in *Salvia miltiorrhiza*. *Front. Plant Sci.* 9, 1950. doi: 10.3389/fpls.2018.01950

Zhao, H. S., Gao, Z. M., Wang, L., Wang, J. L., Wang, S. B., Fei, B. H., et al. (2018). Chromosome-level reference genome and alternative splicing atlas of moso bamboo (*Phyllostachys edulis*). *Gigascience*. 7, giy115. doi: 10.1093/gigascience/giy115

Zheng, X., Tian, S., Meng, X., and Li, B. (2007). Physiological and biochemical responses in peach fruit to oxalic acid treatment during storage at room temperature. *Food Chem.* 104, 156–162. doi: 10.1016/j.foodchem.2006.11.015

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

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

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