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: researchtopics@frontiersin.org
EVOLUTION OF GENE REGULATORY NETWORKS IN PLANT DEVELOPMENT

Topic Editors:

Federico Valverde, Institute for Plant Biochemistry and Photosynthesis, CSIC, Universidad de Sevilla, Spain
Andrew Groover, United States Forest Service (USDA), University of California, Davis, United States
José M. Romero, Institute for Plant Biochemistry and Photosynthesis, CSIC, Universidad de Sevilla, Spain

During their life cycle plants undergo a wide variety of morphological and developmental changes. Impinging these developmental processes there is a layer of gene, protein and metabolic networks that are responsible for the initiation of the correct developmental transitions at the right time of the year to ensure plant life success. New omic technologies are allowing the acquisition of massive amount of data to develop holistic and integrative analysis to understand complex processes. Among them, Microarray, Next-generation Sequencing (NGS) and Proteomics are providing enormous amount of data from different plant species and developmental stages, thus allowing the analysis of gene networks globally. Besides, the comparison of molecular networks from different species is providing information on their evolutionary history, shedding light on the origin of many key genes/proteins. Moreover, developmental processes are not only genetically programed but are also affected by internal and external signals. Metabolism, light, hormone action, temperature, biotic and abiotic stresses, etc. have a deep effect on developmental programs. The interface and interplay between these internal and external circuits with developmental programs can be unraveled through the integration of systematic experimentation with the computational analysis of the generated omics data (Molecular Systems Biology).

This Research Topic intends to deepen in the different plant developmental pathways and how the corresponding gene networks evolved from a Molecular Systems Biology perspective. Global approaches for photoperiod, circadian clock and hormone regulated processes; pattern forma-
tion, phase-transitions, organ development, etc. will provide new insights on how plant complexity was built during evolution. Understanding the interface and interplay between different regulatory networks will also provide fundamental information on plant biology and focus on those traits that may be important for next-generation agriculture.

# Table of Contents

## Evolution of Gene Regulatory Networks in Plant Development

### 06 Editorial: Evolution of Gene Regulatory Networks in Plant Development
Federico Valverde, Andrew Groover and José M. Romero

## Systems Biology Approaches in Plant Reproduction

### 10 Evolution of Daily Gene Co-expression Patterns from Algae to Plants
Pedro de los Reyes, Francisco J. Romero-Campero, M. Teresa Ruiz, José M. Romero and Federico Valverde

### 32 The Importance of Being on Time: Regulatory Networks Controlling Photoperiodic Flowering in Cereals
Vittoria Brambilla, Jorge Gomez-Ariza, Martina Cerise and Fabio Fornara

### 40 Expansion and Functional Divergence of AP2 Group Genes in Spermatophytes Determined by Molecular Evolution and Arabidopsis Mutant Analysis
Pengkai Wang, Tielong Cheng, Mengzhu Lu, Guangxin Liu, Meiping Li, Jisen Shi, Ye Lu, Thomas Laux and Jinhui Chen

### 55 The Role of SHI/STY/SRS Genes in Organ Growth and Carpel Development Is Conserved in the Distant Eudicot Species Arabidopsis thaliana and Nicotiana benthamiana
Africa Gomariz-Fernández, Verónica Sánchez-Gerschon, Chloé Fourquin and Cristina Ferrándiz

### 72 Overview of OVATE FAMILY PROTEINS, A Novel Class of Plant-Specific Growth Regulators
Shucai Wang, Ying Chang and Brian Ellis

### 80 Perspectives for a Framework to Understand Aril Initiation and Development
Sylvia R. Silveira, Marcelo C. Dornelas and Adriana P. Martinelli

### 87 Genome-Wide Characterization of the MADS-Box Gene Family in Radish (Raphanus sativus L.) and Assessment of Its Roles in Flowering and Floral Organogenesis
Chao Li, Yan Wang, Liang Xu, Shanshan Nie, Yinglong Chen, Dongyi Liang, Xiaochuan Sun, Benard K. Karanja, Xiaobo Luo and Liwang Liu

### 101 Evolution and Expression Patterns of CYC/TB1 Genes in Anacyclus: Phylogenetic Insights for Floral Symmetry Genes in Asteraceae
María A. Bello, Pilar Cubas, Inés Álvarez, Guillermo Sanjuanbenito and Javier Fuertes-Aguilar

### 113 Evolution and Expression Patterns of TCP Genes in Asparagales
Yesenia Madrigal, Juan F. Alzate and Natalia Pabón-Mora
TCP Transcription Factors at the Interface between Environmental Challenges and the Plant’s Growth Responses
Selahattin Danisman

A Conserved Carbon Starvation Response Underlies Bud Dormancy in Woody and Herbaceous Species
Carlos Tarancón, Eduardo González-Grandío, Juan C. Oliveros, Michael Nicolas and Pilar Cubas

Evolution of Hormonal Signals
Evolutionary Analysis of DELLA-Associated Transcriptional Networks
Asier Briones-Moreno, Jorge Hernández-García, Carlos Vargas-Chávez, Francisco J. Romero-Campero, José M. Romero, Federico Valverde and Miguel A. Blázquez

Transcriptomic Analysis Implies That GA Regulates Sex Expression via Ethylene-Dependent and Ethylene-Independent Pathways in Cucumber (Cucumis sativus L.)
Yan Zhang, Guiye Zhao, Yushun Li, Ning Mo, Jie Zhang and Yan Liang

Genome Wide Association Approaches
Genome-Wide Analysis of Soybean JmjC Domain-Containing Proteins Suggests Evolutionary Conservation Following Whole-Genome Duplication
Yapeng Han, Xiangyong Li, Lin Cheng, Yanchun Liu, Hui Wang, Danxia Ke, Hongyu Yuan, Liangsheng Zhang and Lei Wang

Comprehensive Analysis of the CDPK-SnRK Superfamily Genes in Chinese Cabbage and Its Evolutionary Implications in Plants
Peng Wu, Wenli Wang, Weike Duan, Ying Li and Xilin Hou

Expression of the KNOTTED HOMEOBOX Genes in the Cactaceae Cambial Zone Suggests Their Involvement in Wood Development
Jorge Reyes-Rivera, Gustavo Rodríguez-Alonso, Emilio Petrone, Alejandra Vasco, Francisco Vergara-Silva, Svetlana Shishkova and Teresa Terrazas

CRP1 Protein: (dis)similarities between Arabidopsis thaliana and Zea mays
Roberto Ferrari, Luca Tadini, Fabio Moratti, Marie-Kristin Lehninger, Alex Costa, Fabio Rossi, Monica Colombo, Simona Masiero, Christian Schmitz-Linneweber and Paolo Pesaresi
Editorial: Evolution of Gene Regulatory Networks in Plant Development

Federico Valverde¹, Andrew Groover²,³ and José M. Romero¹,⁴*

¹ Plant Development Unit, Institute for Plant Biochemistry and Photosynthesis, Consejo Superior de Investigaciones Científicas-Universidad de Sevilla, Seville, Spain, ² US Forest Service, Pacific Southwest Research Station, Davis, CA, United States, ³ Department of Plant Biology, University of California, Davis, Davis, CA, United States, ⁴ Departamento de Bioquímica Vegetal y Biología Molecular, Universidad de Sevilla, Seville, Spain

Keywords: gene regulatory networks, plant development, evolution, omics, molecular system biology

Editorial on the Research Topic

Evolution of Gene Regulatory Networks in Plant Development

The mechanisms regulating developmental processes in plants are very sophisticated as a result of a continuous increase in complexity during evolution. Plants undergo a wide variety of morphological and developmental changes in their lifetime. The regulation of gene expression is a primary mechanism of regulating development. Gene expression is controlled by regulators that, together with their regulatory interactions, integrate environmental cues and coordinate different developmental programmes (Kaufmann and Chen, 2017) through gene regulatory networks (GRNs). GRNs can be defined as a series of regulatory factors that interact with each other and with other regulators to control the levels of mRNA and proteins to specify temporal and spatial patterns (Davidson and Levin, 2005; Levine and Davidson, 2005), and often involve interactions with metabolic networks. New technological advances use the computational analysis of massive “omics” data to generate holistic views of complex biological processes. This Research Topic focuses on the evolution of genes, gene families and GRNs involved in plant development.

Plant growth and development is widely regulated by the circadian clock, which allows plants to anticipate light transitions. The GRN of the circadian clock generates 24-h rhythms influencing various aspects of plant biology including gene expression, metabolism, developmental programmes and transitions, such as flowering (Millar, 2016; Nohales and Kay, 2016). A significant number of genes are regulated by the circadian clock in photosynthetic organisms. It is estimated that in algae about 80–90% and in plants between 30 and 50% of their genes show daily rhythmic patterns (Covington et al., 2008; Zones et al., 2015). That means that despite the evolutionary distance between algae and plants, the expression patterns of many gene clusters showing daily rhythms are conserved (de los Reyes et al.; Serrano-Bueno et al., 2017). Analysis of gene–co-expression networks in plants and microalgae and the use of a new algorithm MBBH (Multiple Bidirectional Best Hits) has allowed the identification of orthologous genes from species evolutionarily very distant, and reveal a wide conservation of genes under daily rhythms in the green lineage (de los Reyes et al.). Photoperiodic sensing (the detection of day length) is crucial for plants to make important decisions during their life, such as the best time to flower. Photoperiod sensing is closely associated with photoreceptor signaling and the circadian clock, processes that are conserved during the evolution of the green lineage (Serrano-Bueno et al., 2017). Brambilla et al. describe the GRNs controlling photoperiodic flowering in long-day and short-day cereals showing their flexibility and how the photoperiodic gene networks have evolved in crops. They also discuss the fact that some regulators are not conserved in all lineages, and that several conserved elements of these GRNs are dedicated to novel functions.
Among many other developmental processes in plants, floral transition and the subsequent floral organ development are triggered by both internal and external cues and are coordinated by gene networks highly conserved in evolution (Romero-Campero et al., 2013). Floral organ identity is determined by a GRNs defined by MADS-box transcription factors, according to the ABCE model for floral development (Lucas-Reina et al., 2016). The APETAL2 (AP2) gene participates in the determination of sepals and petals, belonging to the AP2/ERF family and being the only gene of the ABCE model that is not a MADS-box. A comprehensive analysis of the AP2 genes in a series of spermatophytes indicates that these genes suffered from both negative and positive selection through evolution, first appearing in Gymnosperms and evolved by gene duplication in Angiosperms (Wang et al.). The formation of carpels is directed by the combined action of the C and E functions of the flower organogenesis model. However, carpel development requires a plethora of regulatory genes some of them related to auxin signaling. The SHORT INTERNODES (SHI), STYLISH (STY), and SHI RELATED SEQUENCE (SRS) gene family of zinc-finger transcription factors (SHI/STY/SRS) is involved in carpel formation in distantly related flowering plants as Arabidopsis thaliana and Nicotiana benthamiana, suggesting a common evolutionary origin of the GRN directing carpel development, and possibly the conservation of their targets (Gomariz-Fernández et al.). Ovule development, fruit shape, size and ripening are controlled by multiple factors. The OVATE family of proteins (OPF) is specific to plants and was first identified because of their influence on fruit shape and size in tomato, and constitute a large gene family in all plants studied. OFPs participate in many aspects of plant development and growth as ovule development, fruit shape, fruit ripening, vascular development or even DNA repair (Wang et al.), however further evidence is needed to understand their function. In the developmental programme of seeds, many plants generate a variety of structures related to seed protection and dispersion. Arils are structures, often fleshy, that can accumulate sugars and other substances that help dispersion. Due to the fact that plant model systems do not develop arils, little is known about their origin and the molecular GRN involved in its formation. The origin and the state of the art of aril development and evolution, as well as the molecular pathways known to date is covered by Silveira et al.

The MADS-box family of transcription factors is implicated in flower development, as indicated above, but also in many other developmental and growth programmes. A comprehensive analysis of the MADS-box family in radish (Raphanus sativus) genome indicates that 144 genes code for MADS-box proteins. These data in combination with the characterization of their differential expression pattern in diverse developmental stages and tissues by Next Generation Sequencing (NGS), provides a basic tool to study flowering and floral development in radish, an economically important root crop (Li et al.).

The form and symmetry of the flower in plants is extremely variable. Flower symmetry is controlled by a set of transcription factors including CYCLOIDEA (CYC) and DICHOTOMA (DICH), originally characterized in the eudicot Antirrhinum majus (Luo et al., 1999), that belong to the TCP gene family (TEOSINTE BRANCHED (TB), CYCLOIDEA (CYC), and PROLIFERATION CELL FACTOR (PCF). It is considered that the success of Asteraceae, one of the largest family of vascular plants, is related to their head-like inflorescences, or capitulum (Broholm et al., 2014). By analysing the phylogeny of CYC/TB1 genes together with their expression patterns in Anacyclus flowers, it has been shown that CYC paralogs play a largely conserved role in determining floral symmetry (Bello et al.). Although the flower structure of monocots differs significantly from dicots, the role of the genes involved in flower symmetry are evolutionarily conserved. Madrigh et al. performed a large-scale study of the TCP gene family and their pattern of expression in Asparagales and showed that in this monocot group the number of CYC-like genes is reduced in relation to eudicots, while the converse is observed for PCF-like and CINCINNATA-like genes. This analysis provides basic information for further functional studies on flower symmetry in monocots. Although TCP transcription factors were first characterized as floral symmetry regulators, they are also involved in different developmental processes affecting growth and plant architecture. TCPs connect growth responses (including interactions with plant hormones) to the changing environment as light quality, abiotic and biotic stress or the availability of nutrients (Danisman). TCP transcription factors are also involved in bud dormancy. The Arabidopsis BRANCHEDI (BRC1), a TCP transcription factor, is implicated within the bud in the control of bud dormancy (Aguilar-Martinez et al., 2007) and regulates different GRNs. The analysis of transcriptomic data has allowed the identification of four interconnected GRNs associated with bud dormancy in A. thaliana (Tarancón et al.). They share a significant proportion of genes, indicating that they act coordinately, possibly through hormonal control. Several genes related to carbon starvation response are present in these GRNs, and it has been shown that this syndrome precedes the transition to dormancy both in herbaceous and woody plants, suggesting that bud dormancy could be an ancient response to unfavorable environmental conditions (Tarancón et al.).

Plant hormones play a crucial role in interpreting environmental signals (detected by receptors) and induce developmental programmes through different interconnected pathways. Gibberellins (GAs) induce the degradation of DELLA proteins (named by the amino acid sequence D-E-L-L-A in their N-terminal region) allowing the integration of environmental signals by controlling over 100 transcription factors. By the generation of gene co-expression networks in different organisms, it has been proposed that DELLA proteins had a basic role in coordinating diverse developmental programmes during evolution of the green lineage. The ancient DELLA proteins present in bryophytes do not respond to GAs as in vascular plants. Thus, the recruitment of DELLAs in the gibberellin signaling pathway might have increased their importance in coordinating different processes (Briones-Moreno et al.). Sex determination in monoecious plants is controlled, among other factors, by plant hormones like ethylene and GAs. In cucumber (Cucumis sativus L.), a monoecious plant,
GAs are involved in the formation of the male flower. Transcriptomic analysis of shoot apices in GA-treated and control cucumber plants have provided evidences that GA regulation of sex determination takes place through both, ethylene–dependent and ethylene-independent pathways (Zhang et al.).

The function and components of many GRNs are conserved in evolution based in the innovation, amplification and divergence theory (Bergthorsson et al., 2007; Romero-Campero et al., 2013). Genome-wide analysis (GWA) is a potent tool to identify genes involved in complex traits (Bergthorsson et al., 2007; Romero-Campero et al., 2013).

Dependent Protein Kinases/Snf-1 (CDPK-SnRK) superfamily members are involved in the establishment of organelle RNA levels and participate in organelle biogenesis (Barkan and Small, 2014). Comparative analysis of CRP1 (Chloroplast RNA Processing 1), belonging to the PPR (AtCRP1) and maize (ZmCRP1) showed the conservation of the functionality of these genes between monocots and dicots (Ferrari et al.).

Evaluation of sex determination takes place through both, ethylene–dependent and ethylene-independent pathways (Zhang et al.). The function and components of many GRNs are conserved in evolution based in the innovation, amplification and divergence theory (Bergthorsson et al., 2007; Romero-Campero et al., 2013). Genome-wide analysis (GWA) is a potent tool to identify genes involved in complex traits (Bergthorsson et al., 2007; Romero-Campero et al., 2013).

Comparative analysis of CRP1 (Chloroplast RNA Processing 1), belonging to the PPR (AtCRP1) and maize (ZmCRP1) showed the conservation of the functionality of these genes between monocots and dicots (Ferrari et al.).

This Research Topic is intended to deepen our understanding of the evolution of GRNs underlying different plant developmental pathways and processes. The papers herein describing and comparing GRNs associated with diverse processes including photoperiod, circadian clock, hormone regulation, pattern formation, phase-transitions, organ development, etc. illustrate how the regulatory networks of plants are organized and evolved. We hope that the work presented in this Research Topic will support the growing international efforts in molecular systems biology that are providing exciting new insights into developmental processes in plants.

**AUTHOR CONTRIBUTIONS**

All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

**FUNDING**

The authors would like to thank funding from projects BIO2011-28847-C02-00 and BIO2014-52425-P (Spanish Ministry of Economy and Competitiveness, MINECO) partially supported by FEDER funding.

---

**REFERENCES**


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

Copyright © 2017 Valverde, Groover and Romero. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.
Evolution of Daily Gene Co-expression Patterns from Algae to Plants

Pedro de los Reyes 1, Francisco J. Romero-Campero 1, 2, M. Teresa Ruiz 1, José M. Romero 1 and Federico Valverde 1*

1 Plant Development Unit, Institute for Plant Biochemistry and Photosynthesis, Consejo Superior de Investigaciones Científicas, Universidad de Sevilla, Seville, Spain, 2 Department of Computer Science and Artificial Intelligence, Universidad de Sevilla, Seville, Spain

Daily rhythms play a key role in transcriptome regulation in plants and microalgae orchestrating responses that, among other processes, anticipate light transitions that are essential for their metabolism and development. The recent accumulation of genome-wide transcriptomic data generated under alternating light:dark periods from plants and microalgae has made possible integrative and comparative analysis that could contribute to shed light on the evolution of daily rhythms in the green lineage. In this work, RNA-seq and microarray data generated over 24 h periods in different light regimes from the eudicot Arabidopsis thaliana and the microalgae Chlamydomonas reinhardtii and Ostreococcus tauri have been integrated and analyzed using gene co-expression networks. This analysis revealed a reduction in the size of the daily rhythmic transcriptome from around 90% in Ostreococcus, being heavily influenced by light transitions, to around 40% in Arabidopsis, where a certain independence from light transitions can be observed. A novel Multiple Bidirectional Best Hit (MBBH) algorithm was applied to associate single genes with a family of potential orthologues from evolutionary distant species. Gene duplication, amplification and divergence of rhythmic expression profiles seems to have played a central role in the evolution of gene families in the green lineage such as Pseudo Response Regulators (PRRs), CONSTANS-Likes (COLs), and DNA-binding with One Finger (DOFs). Gene clustering and functional enrichment have been used to identify groups of genes with similar rhythmic gene expression patterns. The comparison of gene clusters between species based on potential orthologous relationships has unveiled a low to moderate level of conservation of daily rhythmic expression patterns. However, a strikingly high conservation was found for the gene clusters exhibiting their highest and/or lowest expression value during the light transitions.

Keywords: daily rhythmic genes, evolution, co-expression networks, systems biology, circadian, Arabidopsis, Chlamydomonas, Ostreococcus

INTRODUCTION

The evolution of the green lineage of photosynthetic organisms from unicellular green algae to land plants is subject to intense research. Recently, genomic analysis have been applied to study key events in the evolution of the green lineage such as terrestrialization (Delwiche and Cooper, 2015; de Vries et al., 2016). Nevertheless, the integration of genomics and transcriptomics to study the
evolution of gene expression patterns in the green lineage has only recently been explored (Romero-Campero et al., 2013; Ruprecht et al., 2017). Photosynthetic organisms are particularly influenced by daily light/dark transitions as their main energy income is a very demanding light-dependent process. Therefore, those plant ancestors (whose extant representative species are unicellular green algae) that were able to anticipate daily fluctuations and schedule their physiological processes accordingly, evolved into the present plant species (Mora-Garcia et al., 2017). Transcriptomics have become a powerful tool to study the global influence of daily light/dark cycles in photosynthetic organisms. Thus, in higher plants, it has been described that between a third and a half of their gene expression is regulated by the circadian clock (Covington et al., 2008; Michael et al., 2008), whereas in algae between 80 and 90% of the transcriptome follows light-dependent daily rhythmic patterns (Monnier et al., 2010; Zones et al., 2015). It seems then that daily rhythmic regulation of gene expression reached a maximum in algae and has substantially decreased in Spermatophytes. Key biological processes such as starch and sugar metabolism exhibit daily rhythmic patterns in both plants and algae (Bläsiing et al., 2005; Sorokina et al., 2011), but these analysis remain fragmented and focused on individual species. In order to better understand the evolution of daily rhythmic gene expression between algae and plants, confirm the tendency to decrease daily regulation and determine the evolution of daily regulated biological processes from algae to plants, a series of comparative Systems Biology analysis integrating genomics and transcriptomics has been used in this study.

The availability of massive amounts of transcriptomic data obtained from different species under equivalent environmental conditions has enabled the use of comparative transcriptomics methodologies to study the evolution of key biological processes (Trachana et al., 2010). In this work, database available RNA-seq and microarray data generated over 24 h periods in neutral conditions (ND: 12 h of light/12 h of dark) and long days (LD: 16 h of light/8 h of dark) conditions from three different model photosynthetic species (Arabidopsis thaliana, Chlamydomonas reinhardtii, and Ostreococcus tauri) have been integrated and analyzed. These photosynthetic eukaryotes represent distant phylogenetic groups. Among the Chlorophyta algae division, the marine Prasinophyceae microalga Ostreococcus is considered a representative of ancient microalgae, one of the smallest eukaryote (picoeukaryote) and constitutes an important part of sea phytoplankton (Derelle et al., 2006; Palenik et al., 2007). Due to its small genome (13 Mb), whose sequence has recently been improved (Blanc-Mathieu et al., 2014), its non-flagellar small body (0.8 µm), its single copy organella (Henderson et al., 2007) and its planktonic life style, it has been considered as the smallest possible photosynthetic eukaryote (Raven et al., 2013) and promising organism for Systems Biology (Thommen et al., 2015). Chlamydomonas is a Chlorophyceae microalg and has been used as a model for photosynthetic organisms’ studies for many years (Harris, 2001). Chlamydomonas has two polar flagella (10 µm body), a much bigger genome than Ostreococcus (120 Mb; Merchant et al., 2007) and lives in sweet water environments. Omics are currently intensely used to explore Chlamydomonas potential in biotechnological applications (Aucoin et al., 2016).

As a representative Spermatophyte or seed plant, Arabidopsis has several characteristics that make it especially useful for Systems Biology studies in general and circadian experiments in particular, as well as a bigger genome and a complex physiology in comparison to microalgae (Van Norman and Benfey, 2009; Koornneef and Meinke, 2010).

Arabidopsis has been extensively used as a model to describe the basic aspects of plant circadian clock regulation over daily rhythms (Millar, 2016; Nohales and Kay, 2016). In a few words, it is formed by three interlocked positive/negative feedback loops. CIRCADIAN CLOCK-ASSOCIATED 1 (CCA1) and LATE ELONGATED HYPOCHOTYL (LHY) myb transcription factors constitute the positive/negative “morning loop” together with PRR9, PRR7 and PRR5 (Nakamichi et al., 2012; Liu et al., 2013, 2016; Kamioka et al., 2016). The negative “central loop” is constituted by CCA1/LHY repression over the gene TIME OF CAB EXPRESSION 1 (TOC1) that, in turn, represses the “morning loop” genes and activates evening-regulated genes in a 24 h-long feed-back loop (Huang et al., 2012). Over this basic core, the circadian genes: GIGANTEA (GI), ZEITLUPE (ZTL), EARLY LIGHT FLOWERING 3 (ELF3), ELF4, LUX ARRHYTHMO (LUX) among others, constitute the “evening loop” that refines the clock and allows for the complex response to the changing daily conditions, including light and temperature inputs (Miyazaki et al., 2015). Outputs of this clock are cell wall synthesis, photosynthetic and starch metabolism genes, among many others (Adams and Carré, 2011). Taking Arabidopsis as model, a much simpler clock has been described in Chlamydomonas and Ostreococcus, where only some genes of the higher plant circadian clock have been identified so far (Mittag et al., 2005; Corellou et al., 2009). In this way, most evolutionary studies have focused on phylogenetic analyses of the key genes regulating the circadian clock (Serrano-Bueno et al., 2017) whereas the analysis of global rhythmic patterns conservation and evolution among different photosynthetic species still remains to be explored.

In this work, phylogenomic and transcriptomic data integration and analysis have been performed by gene co-expression networks construction (Romero-Campero et al., 2013; Gehan et al., 2015; Ruprecht et al., 2017) and a novel algorithm for the identification of potential orthologues called Multiple Best Bi-directional Hit (MBBH). Using clustering techniques, specific gene clusters or modules that showed a rhythmic daily regulation have been identified. Interestingly, these clusters consist of groups of highly co-expressed genes involved in particular biological processes such as cell cycle progression, photosynthesis and ribogenesis, revealing a significant temporal organization in their specific gene co-expression patterns. By comparing the gene modules identified in the different gene co-expression networks obtained for each species, it was possible to determine which biological processes have conserved a daily rhythmic co-expression pattern over the green lineage and which ones have evolved into different patterns. Additionally, a web based software tool, CircadiaNET (http://viridiplantae.ibvf.csic.es/circadiaNet/) has been developed that will allow researchers to independently
analyze their circadian genes of interest studying the biological processes they are potentially involved in, the conservation or evolution of the gene co-expression patterns they follow, as well as the transcription factor (TF) binding sites that are significantly present in their promoters.

**MATERIALS AND METHODS**

**Identification of Putative Orthologous Proteins**

The protein sequences of the three photosynthetic species analyzed in this study were downloaded from publicly available databases. *Chlamydomonas* v5.5 and *Arabidopsis* TAIR10 proteins were downloaded from Phytozone (http://www.phytozone.net/) (Goodstein et al., 2012), while *Ostreococcus* proteins v2 were downloaded from ORCAE (http://bioinformatics.psb.ugent.be/orcae/) (Sterck et al., 2012; Blanc-Mathieu et al., 2014). Using the tools available from the Pfam database (http://pfam.xfam.org/) (Punta et al., 2012), with default parameters (E-value 1.0), the protein domains in all protein sequences were identified. Potential homologous proteins between species were identified based on sequence similarity. We developed a variant of the Bidirectional Best Hit (BBH) that takes into account several candidates. We called this variant Multiple Bidirectional Best Hit (MBBH) and happen to be specially suited for duplication-enriched species such as *Arabidopsis*. The algorithm receives as input two fasta files containing the protein sequences of the two species to compare and the number of multiple bidirectional best hits to consider (Figure 1). In this study *N* was fixed to 20. For each protein Prot$_{1i}$ (encoded by Gene$_{1i}$, deep green) from the first species, the *N* proteins (encoded by Gene$_{2i}$, red) from the second species, Prot$_{2j}$, Prot$_{2j+2}$,...,Prot$_{2j+N}$, exhibiting the highest sequence similarity are selected using the Needleman-Wunsch algorithm implemented in the R Bioconductor package Biostrings (Pagès et al., 2016). These proteins will be called “initial best hits.”

The same process is carried out for species 2. For each initial hit from species 2, the Protein$_{i}$ encoding the Genes encoded by the Protein$_{i}$ are filtered. Only those proteins that exhibit the highest sequence similarity are selected. Next, for each Protein$_{i}$ from the first species its initial best hits, Prot$_{i1}$, Prot$_{i2}$,...,Prot$_{iN}$, are filtered. Only those Prot$_{i}$ from species 2 present Protein$_{i}$ in their *N* initial best hits are kept. The same process is carried out for each protein Prot$_{i}$ from species 2 keeping only the initial best hits that present Prot$_{i}$ among their *N* initial best hits. An additional filtering process is performed by removing from the best bidirectional hits those that do not share at least one domain with the corresponding protein. Finally, we also assign putative orthology to proteins showing no MBBH target, but sharing exactly identical number and order of one or more, non-overlapping pfam domains. This means that due to the specificity of pfam domains, the presence of a single domain does not immediately imply the assignation of a putative orthologue as would be the case with specific higher plant domains such as NAM (Figure 3C) or specific microalgal domains. In this way, our approach combines global sequence similarity with protein domain structure and, additionally, allows for multiple best hits in order to capture the evolution of a single gene into a multi-gene family. In general, MBBH could be a useful tool for those researchers trying to identify orthologues from any given gene in different species, even when they belong to distantly related ones. The tools developed in this article will be operative in the web site: (http://viridiplantae.ibvf.csic.es/circadiaNet/).

**Transcriptomic Data Acquisition and Processing**

For the three photosynthetic species analyzed in this study, transcriptomic data comprising time series of 48 or 72 h were collected from the GEO database (https://www.ncbi.nlm.nih.gov/geo/) (Barrett et al., 2013). In the case of *Arabidopsis*, two different data sets were used. The first one identified with the accession number GSE3416 (Bläsiing et al., 2005) consists of microarray data taken in ND conditions (12:12). The second one identified with the accession number GSE43865 (Rugnone et al., 2013) consists of RNA-seq data taken in LD conditions (16:8). For *Chlamydomonas*, a single RNA-seq data set was collected identified with the accession number GSE71469 (Zones et al., 2015) taken in ND conditions. For *Ostreococcus*, a single microarray data set was used, identified with the accession number GSE16422 (Monnier et al., 2010) taken in ND conditions. The microarray data were processed using the Robust Multi-array Average (RMA) implemented in the Bioconductor R package affy (Gautier et al., 2004). The RNA-seq data was processed following the Tuxedo protocol (Trapnell et al., 2012). Reads were mapped to the reference genomes downloaded from Phytozone using Tophat. Transcripts were
assembled using Cufflinks. Default parameters were used for the different software tools. Gene expression levels were measured as FPKM (Fragments Per Kilobase of exon per Million fragments mapped). Since RNA-seq and microarray gene expression levels have different measurement units these data were normalized by subtracting the mean and dividing by the standard deviation using the R function scale from the base package. This allows representing both types of data on a common scale. Samples were collected every 3 or 4 h in the data analyzed in this study. In order to produce a smooth representation of the gene expression profiles in heatmap graphs these data were linearly interpolated to produce time series consisting of 24 h using the R function approx from the stats package. In heatmaps, genes were sorted according to their expression profile similarity using hierarchical clustering.

**Identification and Clustering of Rhythmic Daily Gene Expression Patterns**

The detection of significant periodic patterns in the transcriptomic data analyzed was performed using RAIN (Rhythmicity Analysis Incorporating Nonparametric methods). This Bioconductor R package consists of functions that implement robust nonparametric methods for the detection of rhythms with arbitrary wave forms and pre-specified periods (Thaben and Westermark, 2014). The main function `rain` was used with the following parameters: A numeric matrix comprising the gene expression levels from the different biological replicates of the 24 h time series; a sampling interval of 4 h for *Arabidopsis* and 3 h for *Chlamydomonas* and *Ostreococcus*; a period of 24 h; and a number of three replicates for *Arabidopsis* and *Ostreococcus* and two for *Chlamydomonas*. A significance level \( \alpha = 1\% \) was chosen to assume that a certain gene exhibits a rhythmic daily expression pattern. The wave form of a gene exhibiting rhythmic daily expression was characterized using its peak (time point where the highest expression value is reached) and its trough (time point where the lowest expression value is reached). Daily rhythmic genes were thus classified into 16 different clusters according to their peaks and troughs at a particular time of the day, that is, all genes in the same cluster present their peaks and troughs at the same time point.

**Gene Ontology Term and Pathway Enrichment Analysis**

Gene Ontology (GO) terms associated to each *Arabidopsis* and *Chlamydomonas* gene were downloaded from Phytozone. For *Ostreococcus*, GO terms were downloaded from ORCAE. The Bioconductor R package topGO (Alexa and Rahnenfuhrer, 2016) was used to determine GO terms significantly enriched in different gene sets. The entire genome of the corresponding species was used as gene background. The statistical significance test selected was Fisher's exact test with a significance level \( \alpha = 5\% \). The web based tool REVIGO (Supek et al., 2011) was used to remove redundancy from the enriched GO terms and produced a summary. The identification of enriched pathways in *Arabidopsis* gene sets was performed with the Bioconductor R package clusterProfiler (Yu et al., 2012). The enrichKEGG function was used with Bejamin-Hochberg as \( p \)-value correction method for the multiple testing and a \( q \)-value cutoff of 0.05.

**Gene Co-expression Network Construction, Visualization, and Analysis**

Gene co-expression was measured using the correlation between gene expression profiles over the 24 h time series. Two daily rhythmic genes were assumed to be co-expressed when the Pearson correlation coefficient between their expression profiles was greater than 0.95. A gene co-expression network was constructed for each species where the nodes represent daily rhythmic genes and an edge is drawn between nodes when the corresponding genes are co-expressed according to the previous criterion. Cytoscape, a software tool for the representation and analysis of complex networks (Shannon et al., 2003), was used to visualize the gene co-expression networks applying the Prefuse Force-Directed Layout. The analysis of the networks was performed using the R package igraph (Csardi and Nepusz, 2006). The scale-free property was tested using linear regression over the logarithmic transform of the degree distribution. The small-world property was tested by generating \( 10^4 \) random scale-free networks with the same number of nodes and edges as the corresponding network, using the `barabasi.game` function from the igraph package.
Module Conservation Analysis
Two daily rhythmic genes from two different species determined as potential orthologues using MBHB were defined as exhibiting a conserved daily pattern when both belonged to the same cluster (they presented their peak and trough in the same time interval) or when the Pearson correlation coefficient between their expression profiles were higher than 0.98. The conservation among the co-expression patterns of two different sets of genes from two different species were computed according to the summary composite conservation statistic “Zsummary” as defined in Langfelder et al. (2011). A Zsummary value lower than 2 indicates no conservation, a Zsummary value 2–10 implies moderate conservation, while a Zsummary greater than 10 constitutes evidence of a great level of conservation. The R package WGCNA (Langfelder and Horvath, 2012) and the function implemented therein were used.

Plant, Algal Material, and Growth Conditions
Three independent biological replicates for plants and algae were grown in a model SG-1400 phytotron (Radibex SA, Spain) under LD conditions with temperatures ranging from 22°C (day) to 18°C (night) and 75 μmol m−2 s−1 light intensity. Arabidopsis thaliana Col-0 wild type seeds were incubated 4 days at 4°C in the dark before sowing in MS plates. 12-day-old seedlings were collected every 4 h during a 24 h period. Time points were denoted as Zeitgeber Time N (ZTN) indicating the time point N hours after the lights are switched on in the phytotron (ZT0), mimicking dawn. Chlamydomonas reinhardtii wild type CW15 was grown in flasks with minimal Sueoka medium for 12 days. Similarly, Ostreococcus tauri wild type RCC 745 was grown in flasks with Keller medium (Keller et al., 1987) for 12 days. Algae were harvested every 4 h during a 24 h period.

RNA Isolation and QPCR
RNA was isolated from Arabidopsis seedlings (0.1 g leaf tissue), Chlamydomonas and Ostreococcus (pellet of 25 ml culture at exponential phase) using a modified TRIZOL (Invitrogen) protocol as described by the manufacturer. Briefly, the sample was mixed with 1 ml of TRIZOL and 0.2 ml of chloroform and the mixture was centrifuged at 16,000 g for 10 min at 4°C. The supernatant was treated with 1 volume of 100% (v/v) 2-propanol, incubated 15 min at room temperature and centrifuged at 16,000 g for 10 min at 4°C. The pellet was dissolved in 0.75 ml 3 M LiCl, incubated for t > 10 min at room temperature and centrifuged at 16,000 g for 10 min at 4°C. This pellet was washed with 80% (v/v) ethanol and centrifuged at 16,000 g for 10 min at 4°C. The final RNA sample was suspended in 21 µl of DEPC treated water and 1 µl quantified employing a ND-1000 Spectrophotometer (Nanodrop).

One micro gram of TRIZOL-isolated RNA was used to synthesize cDNA employing the Quantitect® Reverse kit (Qiagen) following the instructions recommended by the manufacturer. cDNA samples were diluted to a final concentration of 10 ng/µL and stored at −20°C until QPCR was performed. Primers to amplify the 3’ translated region of AtHY5, CrHY5, OtHY5, including AtUBQ10, CrTUB, and OtEF1α as housekeeping genes (Table 1) were designed using the Oligo analyzer program from Integrated DNA Technologies (http://eu.idtdna.com/analyzer/Applications/OligoAnalyzer/).

QPCR was performed in a Multicolor Real-Time PCR Detection System iQTM5 (Bio-Rad) in a 10 µL reaction: primer concentration 0.2 µM, 10 ng cDNA and 5 µL SensiFAST TM SYBR & Fluorescein Kit (Bioline). The QPCR program consisted in (i) 1 cycle (95°C, 2 min); (ii) 40 cycle (95°C, 5 s; 60°C, 10 s and 72°C, 6 s) (iii) 1 cycle (72°C, 6 s). Fluorescence was measured at the end of each extension step and the melting curve was performed between 55 and 95°C. Three biological replicates with three technical replicates from each species were used for every time point. The QPCR results were estimated using the ddCt R Bioconductor package (Zhang et al., 2015).

RESULTS AND DISCUSSIONS
Most Ostreococcus Proteins Present Potential Orthologues in Arabidopsis and Chlamydomonas, but This Is Not the Case with the Other Two Species
The identification of potential orthologues between two species constitutes one of the most important bottlenecks in comparative genomics and transcriptomics (Dessimoz et al., 2012). The Bidirectional Best Hit (BBH) algorithm is the most commonly used method for the automated identification of putative orthologues. In spite of its simplicity, BBH is highly accurate when dealing with bacterial and archaeal genomes (Wolf and Koonin, 2012). Nevertheless, BBH does not perform optimally, missing as much as 60% of orthologues, in gene-duplication-enriched genomes such as those of plants and animals (Dalquen and Dessimoz, 2013). In this work, in order to identify orthology among distantly related species, we have developed a variant of the BBH algorithm termed Multiple Bidirectional Best Hit (MBBH). MBBH assigns to each gene g_i from the first species k ≤ N potential orthologues from the second species, g_i→g_j, based on sequence similarity of the proteins they encode if, and only if, the query gene g_i is among the N most similar genes from the first species (Section Materials and Methods, Figure 1).

Using MBBH tool, more than 97% of Ostreococcus proteins could be ascribed to a potential orthologue either in Arabidopsis or Chlamydomonas (Supplementary Table 1). This is in agreement with the reduced and compact Ostreococcus genome (Palenik et al., 2007) and suggests, as has been proposed (Raven et al., 2013), that this marine picoeukaryote contains the minimal genome for a functional photosynthetic eukaryote and that it is almost entirely regulated by the circadian clock (see below). Confirming this idea, a GO term enrichment analysis over the set of genes without potential Arabidopsis or Chlamydomonas orthologues did not produce any significant result, indicating that these genes are not involved in any specific biological process. In fact, the highest number of genes without potential orthologues was located on chromosomes 2 and 19. Some of these genes codified for a number of unnamed products (ostta02g03690, ostta02g05500, ostta02g00800) and a group of Tcl-like/mariner transposases (ostta02g01245, ostta02g01247, ostta02g02355).
FIGURE 2 | Functional annotation of genes exclusively identified in *Chlamydomonas*. No potential *Ostreococcus* or *Arabidopsis* orthologues were identified for 15% of *Chlamydomonas* proteins. (A) Non-redundant GO term enrichment analysis over the set of *Chlamydomonas* genes with no potential orthologue in the other species, suggesting that they are mainly involved in “movement of cell or subcellular component,” “organic cyclic compound metabolism,” and “DNA integration.” Each rectangle area in the treemap represents the $-\log_{10}(p$-value) for the corresponding GO-term. (B) Domain structure of the proteins encoded by Flagellar inner arm dynein 1 (Cre14.g624950) and axonemal dynein heavy chain 6 (Cre14.g627576), two *Chlamydomonas*-specific flagellar proteins absent in *Ostreococcus* and *Arabidopsis* involved in “movement of cell” or “subcellular component.” (C) Domain structure of class III guanylyl and adenylyl cyclase encoded by CYG11 (Cre07.g320700), a *Chlamydomonas*-specific protein involved in “organic cyclic compound metabolism.” This protein shows a higher homology to animal cyclases than to plant ones. (D) Domain structure of the protein encoded by gag-pol-related retrotransposon (Cre01.g045850), an example of *Chlamydomonas* specific proteins associated to the “DNA integration” GO term. Color boxes represent domains identified in pfam database including their identification codes.

This supports previous analysis on the heterogeneity of *Ostreococcus* genome that identified chromosomes 2 and 19 as different from the other chromosomes in terms of GC content, codon usage and number of transposable elements (Derelle et al., 2006).

Following the same comparative analysis, approximately 85% of *Chlamydomonas* proteins present potential orthologues in *Ostreococcus* and *Arabidopsis* (Supplementary Table 2). Therefore, a set of approximately 2,600 *Chlamydomonas* genes that codify for specific proteins could not be traced to any orthologue in the other two species. An ontology enrichment analysis performed over this set of genes using the R Bioconductor package topGO and summarized with REVIGO revealed that the top three most significant and non-redundant GO terms were “Movement of cell or subcellular component,” “Organic cyclic compound metabolism,” and “DNA integration” (Figure 2A). The same set of genes were associated with the GO terms “Movement of cell or subcellular component,” “Anatomical structure homeostasis” and “single-organism process.” Therefore, the last two GO terms were considered redundant and only “Movement of cell or subcellular component” is discussed. Similarly, the GO terms “Organic cyclic compound metabolism” and “cellular aromatic compound metabolism” included the same list of associated genes than “Organic cyclic compound metabolism,” so only the latter is discussed. Genes codifying flagellar proteins such as flagellar inner arm dynein 1 (Cre14.g624950) and axonemal dynein heavy chain 6 (Cre14g.627576) (Figure 2B) are representative of the group “Movement of cell or subcellular component.” *Chlamydomonas* cells exhibit two polar flagella whereas no flagellar-like structures are present in *Ostreococcus* or *Arabidopsis* cells, explaining the lack of orthologues in the non-flagellated organisms. Besides, the *Chlamydomonas* specific genes annotated with the GO term “Organic cyclic compound metabolism” include the class III guanylyl and adenylyl cyclase family including genes such as adelynate/guanylate cyclase CYG11 (Cre07.g320700) (Figure 2C). These enzymes that catalyze the synthesis of cGMP and cAMP in animals are absent in plants, hence the lack of potential orthologues in *Ostreococcus* and *Arabidopsis*. In fact, this concurs with the idea that some *Chlamydomonas* genes are closer to animal than to plant ones (Merchant et al., 2007). On the other hand, the unique gag-pol-related
retrotransposon (Cre01g.045850) (Figure 2D) is an example of the Chlamydomonas specific genes associated with “DNA integration” GO term, suggesting that this set of viral pathogens infecting Chlamydomonas is specific for the alga.

The analysis also identified potential Ostreococcus and Chlamydomonas orthologues for approximately 75% of Arabidopsis proteins (Supplementary Table 3), suggesting that one fourth of the Arabidopsis proteins are not present in microalgae and have potentially been acquired during the course of higher plant evolution. In accordance with this idea, a GO term enrichment analysis showed that these Arabidopsis-specific proteins are involved in the following top three most significant non-redundant terms: “Cell communication,” “Cell wall organization,” and “Multiorganism reproductive process” (Figure 3A). Similar to the previous result, the GO term “Signaling” was found to share the same gene list with “Cell communication”; “Cellular glucan metabolism” presented the same associated genes as “Cell wall organization” and the set of genes assigned to “Biological regulation” and “Multiorganism reproductive process” were identical. Therefore, the former GO terms were considered redundant and only the latter GO terms are discussed. These GO terms correspond to complex multicellular plants features that are absent in microalgae. This way, key proteins involved in cell wall biogenesis and cellular glucan metabolism in Arabidopsis such as Pectin Methyl esterase 1 (PME1, At1g53840) lack Ostreococcus and Chlamydomonas potential orthologues (Figure 3B). Indeed, the evolutionary history of PME genes has been previously studied, establishing their appearance in multicellular Charophyte algae and supporting their absence in more primitive algae (Wang et al., 2013). A case of proteins annotated as involved in “Multi-cellular processes” without potential orthologues in Ostreococcus and Chlamydomonas is Arabidopsis NAC domain containing protein 98 (ANAC098, At1g53950) (Figure 3C). The family of NAC TFs is one of the largest in plants and is involved in multiple key developmental processes such as floral development. Again, this TF family first originated in Charophytes (Zhu et al., 2012) and is absent in Chlamydomonas and Ostreococcus. Finally, the Recognition of Peronospora Parasitica 1 (RPP1, At3g44880) is an example of a specific Arabidopsis protein involved in “Cell communication/signaling” and is part of the set of specific Arabidopsis genes related to pathogen resistance and programmed cell death (Schreiber et al., 2016; Figure 3D).

Large TF Families in Arabidopsis Have Single Orthologues in Microalgae Suggesting Gene Amplification and Functional Diversification Processes

The approach to define potential orthologues by MBBH can detect several candidates for any given gene, identifying multiple orthologous genes that could have appeared from processes of gene duplication. In Arabidopsis, on average, 5.19 and 7.83 genes could be ascribed to a homolog in Chlamydomonas and Ostreococcus, indicating that Arabidopsis gene families are, on average, five to eight times larger than in those organisms. This concurs with the idea that whole genome duplication and gene duplication events were crucial in the evolution of plant gene families (Romero-Campero et al., 2013; Rensing, 2014). Interestingly, this process is particularly frequent in TF families and MBBH tool has efficiently detected functional domains in the protein sequences of the three species under study and identified the TF family they belong to. In general, their sizes coincided with the data available in the Plant Transcription Factor Database, PlantTFDB (Jin et al., 2017), confirming the accuracy of this approach. On average, less than 4 and 7 protein members formed each Ostreococcus and Chlamydomonas TF family respectively, whereas a media of 30 members constituted Arabidopsis TF families, further supporting the idea of multiple duplication events.

Two clear examples of amplification in TFs are the DOF and COL protein families (Figures 4A,B) that present one single member in Chlamydomonas and two members in Ostreococcus.
FIGURE 4 | Amplification of TF families through evolution. Arabidopsis TF families are respectively on average five and eight times larger than Chlamydomonas or Ostreococcus ones, indicating an evolutionary history with multiple gene duplication events. (A) Domain structure of DOF proteins. Ostreococcus has two members; one of them exhibiting a Response Regulator domain in the N-terminal (OtDOF2), while it is constituted by a single member in Chlamydomonas. Of the 47 members of the family in Arabidopsis, the closest homologs to algal DOFs are the CDFs, involved in photoperiodic flowering. (B) Domain structure of the COL TF family. It has two members in Ostreococcus, a single member in Chlamydomonas and 22 in Arabidopsis. Some variation in the number of B-boxes in the N-terminal domain is observed. CO and closest homologs are shown. (C) Domain structure of the MADS-box TF family. This family has gone through an intense process of amplification from the single copy gene present in Ostreococcus and Chlamydomonas to the multi-gene family, constituted by 146 members, in Arabidopsis. The acquisition of novel protein domains in Arabidopsis would have given rise to different subfamilies. (D) Domain structure of PRR protein family. Although relatively small in Arabidopsis, with around 10 members, it has also been submitted to a process of gene amplification starting from the single copy gene in Ostreococcus and two genes in Chlamydomonas. Color boxes represent domains identified in pfam database including their identification codes.

respectively, whereas in Arabidopsis 47 DOF and 22 COL proteins are present. A subset of plant DOF genes are regulated by the clock (CYCLING DOF FACTORS, CDFs) and, in time, control the daily expression of CONSTANS (CO) in Arabidopsis and potato during photoperiodic flowering (Imaizumi et al., 2005; Formara et al., 2009; Kloosterman et al., 2013). This constitutes a conserved DOF-CO module in Spermatophytes that is also conserved in Chlamydomonas (Lucas-Reina et al., 2015). In Ostreococcus it might be different. While the Chlamydomonas and Arabidopsis orthologues presented a single DOF domain, one of the two Ostreococcus DOF proteins (OtDOF2, ostta04g02850) presented an additional N-terminal Response Regulator domain with a potential phosphoacceptor aspartic acid-aspartic acid-lysine (DDK) motif (Figure 4A). This suggests that OtDOF2 could be part of a phosphorelay system (Djouani-Tahri et al., 2011). The COL gene family (Figure 4B) seemed to have appeared in microalgae, with a single representative in Chlamydomonas (CrCO, Cre06.g278159) (Valverde, 2011) and two putative orthologues in Ostreococcus (OtCOL1, ostta04g3620 and OtCOL2, ostta09g01510). In Chlamydomonas CrCO is involved in the control of the cell cycle, starch synthesis and oil content (Serrano et al., 2009; Deng et al., 2015) and some of these functions have been conserved in some Arabidopsis orthologues (Romero-Campero et al., 2013; Ortiz-Marchena et al., 2014).

The MADS-box and PRR families (Figures 4C,D) constitute other interesting cases. MADS-box TF family is an example of frequent amplification over the course of plant evolution, since only a single copy is present in Chlamydomonas and Ostreococcus, whereas in Arabidopsis this family contains 146 members. MADS-boxes have been classified into different subfamilies depending on the recruitment of new protein domains besides the conserved MADS DNA-binding domain, such as the K-boxes (Figure 4C), and that may explain the ample functional diversity of these TF family in higher plants. On the other hand, modern plant PRRs (Figure 4D) contain an N-terminal Pseudo Response Regulator domain and a C-terminal CCT domain, but algal proteins are different. Spermatophytic PRRs are similar in size to COLs and, like them, seem to have experienced a similar amplification process from a single gene copy in Ostreococcus (Otprr, ostta13g01820) and two genes in Chlamydomonas (CrPPR1, Cre02g094150; CrPPR2, Cre16.g567421) to a multi-gene family in Arabidopsis. Nevertheless, while in Ostreococcus and Chlamydomonas...
proteins, a potential true phosphoacceptor DDK motif is present, suggesting that the algal proteins still retain part of the ancestral phosphorelay signaling mechanism, this motif is missing in the higher plant domain, constituting a "Pseudo" Response Regulator (Mizuno and Nakamichi, 2005; Satbhai et al., 2011).

To study to what extent the conservation in protein sequence and domain structure is accompanied by conservation in the expression profiles, transcriptomic data consisting of 24 h time series for the three species under study was analyzed. Not surprisingly, differences in the expression profiles of Arabidopsis genes were observed, some genes retaining the same one as in Ostreococcus and Chlamydomonas, while others acquired completely new expression profiles (Figure 5). This way, in the DOF family (Figure 5A) the expression profile of the Arabidopsis genes CDF1 and CDF2 have retained similar expression profiles as CrDOF and OtDOF1 with a peak around ZT0 and a trough around ZT12. However, CDF4 and OtDOF2 peak around ZT21, while CDF3 exhibits an expression profile with a peak at ZT3 and trough at ZT18, not observed in Chlamydomonas or Ostreococcus genes. A similar situation is observed for the COL (Figure 5B), MADS-box (Figure 5C) and PRR (Figure 5D) TF families. The Arabidopsis COL genes COL1, COL2 and COL3 exhibit very similar expression profiles to OtCOL1 and CrCO, showing a trough at ZT12 (Figure 5B). On the other hand, OtCO2 and CO present distinctly different expression profiles with their minimum expression levels at ZT0 and ZT6, respectively. The MADS-box family presents a great diversification in the expression profile of their members, with a substantial difference between the expression patterns of Arabidopsis, Chlamydomonas and Ostreococcus genes (Figure 5C). In the PRR family, OtPRR presents the same expression pattern as Arabidopsis PRR1 (TOC1) and PRR3, peaking at ZT12, whereas PRR5 and PRR7 peak earlier, at around ZT9. However, CrPRR1 and CrPRR2 present completely different profiles, peaking at ZT21 and ZT0, respectively (Figure 5D). This could imply that, in general, gene amplification is accompanied by expression profile diversification and hence, functional diversification (Romero-Campero et al., 2013).

**Most Genes in Chlamydomonas and Ostreococcus Exhibit Light-Dependent Daily Rhythmic Expression Patterns, but Only 40% of the Arabidopsis Transcriptome Shows a Periodic Expression**

Recently, massive amounts of transcriptomic data for photosynthetic organisms under diverse environmental and physiological conditions have been produced. One of the most studied environmental signals is the alternation of light and dark cycles. Nevertheless, the analysis of the diurnal changes in the transcriptome has been independently performed in different species, making necessary the application of Molecular Systems Biology techniques to integrate and compare them. In this study we have analyzed microarray data collected over 24 h periods in ND conditions for Ostreococcus (Monnier et al., 2010) and Arabidopsis (Bläsing et al., 2005), as well as RNA-seq data for Chlamydomonas (Zones et al., 2015) in ND conditions and Arabidopsis in LD conditions (Rugnone et al., 2013). Details on transcriptomic data processing are described in the Materials and Methods section. In order to determine the effect of the alternation of light and dark periods over the transcriptome we identified genes exhibiting rhythmic or oscillating expression patterns with a period of approximately 24 h in the three species under study. Nonparametric methods implemented in the Bioconductor R package RAIN (Rhythmicity Analysis Incorporating Nonparametric methods) were used to detect significant gene expression patterns with arbitrary wave forms and a pre-specified period of 24 h (Thaben and Westermark, 2014), see Materials and Methods section for details.

The analysis revealed that more than 90% of the Ostreococcus genes targeted in the microarray exhibited daily rhythmic patterns (Supplementary Table 1). No GO term significantly enriched in the non-periodic genes was identified. This suggests that practically the entire Ostreococcus transcriptome is periodic and that most Ostreococcus biological processes are strongly affected by the alternation of light and dark cycles. In Chlamydomonas approximately 70% of the genes were found to follow significant daily rhythmic expression patterns (Supplementary Table 2). The GO term enrichment analysis over the non-periodic Chlamydomonas genes identified only a
few significant biological processes. The two most significant processes were “DNA integration” and “Defense response to virus” including genes such as the reverse transcriptase Cre65.g235102, the DNA ligase Cre60.g277801 and the 2′-5′ oligoadenylate synthetase Cre15.g641050. These have been identified as biological processes induced by biotic stimuli, and are thus independent from the abiotic inputs from light/dark cycles. Finally, only 43.18% of the Arabidopsis genes showed significant circadian patterns according to our analysis (Supplementary Table 3), suggesting that during the evolution of higher plants many biological processes were uncoupled from the external influence of alternating light and dark cycles. These percentages of rhythmic genes are largely in agreement with previous results (Bläsing et al., 2005; Monnier et al., 2010; Zones et al., 2015).

The pathway enrichment analysis performed over the Arabidopsis daily rhythmic genes revealed several key significant pathways influenced by rhythmic changes (Table 2). As expected, the most significant pathway represented in the enrichment was the one termed “Circadian rhythms in plants.” Also expected were the pathways “Porphyran and chlorophyll metabolism” and “Pentose phosphate pathway”. These pathways are involved in photosynthesis and hence are expected to be highly regulated by light/dark cycles exhibiting rhythmic patterns. Key metabolic pathways such as “Fatty acid degradation” and “Starch and sucrose metabolism” were also detected as significantly enriched in daily rhythmic genes indicating that Arabidopsis metabolism is still highly affected by alternating periods of light and dark following rhythmic patterns. Somewhat surprising was the identification of the “Alpha-Linolenic acid metabolism” and “Plant hormone signal transduction” among the enriched GO groups. This result suggests either that these pathways are ancient or that during the course of plant speciation new or exclusive plant pathways involved in hormone synthesis and sensing were recruited and acquired a daily rhythmic regulatory pattern. Curiously, although the course of evolution seems to have made the clock more complex in Arabidopsis, a lower percentage of the transcriptome seems to be regulated in a light-dependent periodic manner. It could be that a precise, fine tuning of the clock was recruited to make new processes independent from light/dark transitions in higher plants. More circadian experiments in different tissues or developmental stages may be needed to explain this apparent paradox.

With the aim of representing the daily rhythmic genes and their complex co-expression relationships, a gene co-expression network for each photosynthetic species analyzed in this study was constructed (Figures 6A–C). Nodes represent daily rhythmic genes and edges co-expression relationships, so that two circadian genes are assumed to be co-expressed when the Pearson correlation index between their 24 h expression profiles is greater than 0.95 and an edge is drawn between them. The three gene co-expression networks were visualized using the Prefuse Force Directed layout implemented in the software tool Cytoscape (Shannon et al., 2003). Interestingly, the three different gene co-expression networks acquired the same ring-like structure, capturing the chronological relationship between the co-expression patterns of the periodic genes. The basic topological parameters in network analysis, namely node degree distribution and clustering coefficient, were computed for each network (Figure 6D). The degree distribution of the three networks follows an exponential negative distribution with p-values below 2.2 · 10−16, which implies that they are scale-free networks (Barabási, 2015), suggesting that the global daily rhythmic co-expression patterns in the three species are robust to random changes or mutations but fragile to direct changes or mutations in hub genes, those co-expressed with a high number of genes (Aoki et al., 2007). Additionally, the clustering coefficient of the three networks was significantly high when compared to the clustering coefficient of random scale-free networks with the same number of nodes and edges. Therefore, all three networks constitute small-world networks (Barabási, 2015) with expected short paths connecting any two nodes of the network. This is assumed to facilitate the quick propagation of information between genes with different periodic expression patterns.

**Daily Rhythmic Genes Clusters Constitute Transcriptional Programs That Confer Temporal Separation to Different Biological Processes with Different Levels of Conservation between Ostreococcus, Chlamydomonas, and Arabidopsis**

The daily rhythmic expression pattern of individual genes can be described by using its peak, the time point when the expression

<table>
<thead>
<tr>
<th>Pathway ID</th>
<th>Description</th>
<th>q-value</th>
<th>Ratio cercadian/total</th>
</tr>
</thead>
<tbody>
<tr>
<td>ath04712</td>
<td>Circadian rhythm-plant</td>
<td>4.20 · 10−6</td>
<td>29/36</td>
</tr>
<tr>
<td>ath00030</td>
<td>Pentose phosphate pathway</td>
<td>5.25 · 10−3</td>
<td>35/58</td>
</tr>
<tr>
<td>ath00052</td>
<td>Galactose metabolism</td>
<td>1.54 · 10−2</td>
<td>32/55</td>
</tr>
<tr>
<td>ath00590</td>
<td>Arachidonic acid metabolism</td>
<td>1.54 · 10−2</td>
<td>13/17</td>
</tr>
<tr>
<td>ath00966</td>
<td>Glucosinolate biosynthesis</td>
<td>1.54 · 10−2</td>
<td>14/19</td>
</tr>
<tr>
<td>ath00860</td>
<td>Porphyrin and chlorophyll</td>
<td>1.76 · 10−2</td>
<td>28/48</td>
</tr>
<tr>
<td>ath00360</td>
<td>Phenylalanine metabolism</td>
<td>1.84 · 10−2</td>
<td>25/42</td>
</tr>
<tr>
<td>ath00071</td>
<td>Fatty acid degradation</td>
<td>2.66 · 10−2</td>
<td>24/41</td>
</tr>
<tr>
<td>ath00500</td>
<td>Starch and sucrose metabolism</td>
<td>2.66 · 10−2</td>
<td>66/139</td>
</tr>
<tr>
<td>ath0130</td>
<td>Ubiqinone and other</td>
<td>2.66 · 10−2</td>
<td>21/35</td>
</tr>
<tr>
<td></td>
<td>terpenoid-quinone biosynthesis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ath00260</td>
<td>Glycine, serine and</td>
<td>3.54 · 10−2</td>
<td>37/72</td>
</tr>
<tr>
<td></td>
<td>threonine metabolism</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ath00592</td>
<td>alpha-Linolenic acid metabolism</td>
<td>3.54 · 10−2</td>
<td>21/36</td>
</tr>
<tr>
<td>ath01200</td>
<td>Carbon metabolism</td>
<td>3.62 · 10−2</td>
<td>114/262</td>
</tr>
<tr>
<td>ath01230</td>
<td>Biosynthesis of amino acids</td>
<td>3.62 · 10−2</td>
<td>111/255</td>
</tr>
<tr>
<td>ath04075</td>
<td>Plant hormone signal transduction</td>
<td>3.62 · 10−2</td>
<td>118/273</td>
</tr>
<tr>
<td>ath01210</td>
<td>2-Oxocarboxylic acid metabolism</td>
<td>4.66 · 10−2</td>
<td>37/74</td>
</tr>
</tbody>
</table>
FIGURE 6 | Circadian gene co-expression networks. Genes are represented by nodes and edges represent co-expression between them (Pearson correlation index between their 24 h-expression profiles greater than 0.95). The three networks acquired a ring-like structure mirroring the periodic expression pattern of daily rhythmic genes. (A) Arabidopsis daily rhythmic gene co-expression network comprised 7639 circadian genes with 537027 co-expression relationships. (B) Chlamydomonas daily rhythmic gene co-expression network consisted of 10338 circadian genes and 573980 co-expression interactions. (C) Ostreococcus daily rhythmic gene co-expression network comprised 5782 nodes and 222493 co-expression relationships. (D) The topological analysis of the three gene co-expression network shows that they are scale-free and small-world networks: (1) Their degree distributions (above, red) follow an exponential negative distribution. (2) Their average clustering coefficient (below, blue) is significantly greater than random networks with equal number of nodes and edges. Gene cluster with a peak at Dawn is represented in blue, that with a peak during the Day in yellow, the one showing a peak at Dusk in red and that with a peak at Night in black.
level is highest and its trough, the time point when the expression level is lowest (Figure 7A, Supplementary Tables 1–3). The transcriptomic data used in the co-expression networks was collected in ND conditions, but using different time sets. While Ostreococcus and Chlamydomonas data were collected every 3 h, Arabidopsis data were collected every 4 h. In order to compare the expression profiles from the different species, a 24 h day was divided into four temporal intervals that contained all points from the different time series. Therefore, Dawn (blue) was defined as the time interval from ZT21 to ZT3; Day (yellow) was considered from ZT3 to ZT9; Dusk (red) consisted of the time period from ZT9 to ZT15 and Night (black) was assumed

**FIGURE 7 |** Daily rhythmic gene clustering. The wave form of daily periodic genes expression profile can be characterized by its peak (maximum) and trough (minimum). (A) Example of gene expression for two daily rhythmic genes in a 72 h course. Time is shown in hours (h) as Zeitgeber Time (ZT). The first one exhibits its peak around the light/dark transition and its trough around the dark/light, while the second one presents a symmetric profile. (B) The 24 h day was divided into four different time intervals: “Dawn” was defined as the time interval from ZT21 to ZT3 (blue); “Day” was considered from ZT3 to ZT9 (yellow); “Dusk” consisted of the time period from ZT9 to ZT15 (red) and “Night” was assumed to be from ZT15 to ZT21 (black). (C) Four different clusters were defined for visualization purposes following the same color code: genes peaking at Dawn, Day, Dusk, and Night. A 48 h course normalized expression profile of every gene in each cluster was represented, as well as the average gene expression profile for each cluster. Note that the wave form of the average expression profile for each gene cluster is similar in the three different species.
to be from ZT15 to ZT21 (Figure 7B). Thus, daily rhythmic genes could be classified into 16 different clusters according to the time interval where their peak was located (Dawn, Day, Dusk, and Night) and the time interval containing their trough (again Dawn, Day, Dusk, and Night). Due to the amount of data engaged, the graphic representation of these 16 clusters can be confusing, so, in order to facilitate their visualization, we decided to merge all clusters peaking at the same interval into a single one. This way, for each generalized gene cluster we represented the expression profile of every gene and the average expression profile of the entire cluster (Figure 7C). Interestingly, no major differences were apparent between the average expression profiles from the three different species under analysis. Furthermore, when we colored the location of these clusters in the corresponding gene co-expression networks (Figures 6A–C), we observed that these clusters presented the same localization in the three different networks. This is a strong support for the veracity of our approach and shows, at a single view, that daily rhythmic genes associate in true temporal sets of genes performing synchronous actions and roughly reflecting the natural course of a 24 h clock. In this analysis the boundaries separating clusters were more sharply defined in Arabidopsis and Chlamydomonas than Ostreococcus that showed a higher degree of genes mixing between temporal clusters (Figures 6A–C). This supports the already established idea that the primitive Ostreococcus clock is not as efficient as those of the other more complex organisms.

The two largest gene clusters in Ostreococcus and Chlamydomonas contain genes peaking at Dusk with their troughs at Dawn (1620 and 3751 genes, respectively) and genes peaking at Dawn with their troughs at Dusk (1744 and 2869 genes, respectively; Table 3). The genes of these two clusters comprise 56.85 and 49.98% of the entire set of daily rhythmic genes in Ostreococcus and Chlamydomonas, respectively and concur with the common notion that the light/dark and dark/light transitions play central roles in daily rhythmic gene expression regulation in both algae. Surprisingly, in Arabidopsis these two gene clusters are not the largest ones, comprising only 22.51% of all periodic genes. This suggests a degree of uncoupling between regulation of circadian gene expression and light/dark and dark/light transitions in higher plants. Nevertheless, the largest gene cluster in Arabidopsis is made up of 1209 genes peaking at Dusk with their troughs in the Day, indicating that the dark/light transition still plays a key role in the regulation of daily rhythmic genes in Arabidopsis. The second largest cluster in Arabidopsis contains 1014 genes whose expression peaks take place at Night and troughs during the Day. This suggests certain independence in the regulation of daily rhythmic gene expression in Arabidopsis from the light/dark transitions since a large number of genes present their peaks or troughs in the periods of light and dark.

We performed GO term enrichment analysis over the different gene clusters (Table 3) in order to determine the biological processes that were carried out by the different rhythmic gene clusters. Different clusters were enriched in distinct biological processes indicating a temporal separation among them. Some biological processes show enrichment in the same clusters for the three different species such as “DNA metabolic process” (gene clusters with peak at Dusk and trough at Dawn), “photosynthesis” (gene clusters with peak at Day and trough at Night) and “carbohydrate catabolic process” (gene clusters that have the peak during the Dawn and trough at Dusk) indicating a conservation in their daily rhythmic expression patterns (Table 2). On the contrary, an anticipation, delay or uncoupling from periodic expression of certain biological processes was apparent when the three species were compared.

Cell cycle-related GO terms such as “DNA metabolic process” were enriched in the gene cluster with peak at Dusk and trough at Dawn in the three species, including genes such as REPLICATION PROTEIN A (Cre16.g65100, ostta18g01440 and At4g19130), ORIGIN RECOGNITION COMPLEX 1 (Cre10.g455600, ostta04g05220, and At1g26840) and CELL DIVISION CYCLE 45 (Cre06.g270250, ostta04g04640, and At3g25100) that showed very similar expression profiles in the three species (Figure 8A). Interestingly, the GO terms “DNA replication” and “DNA metabolic process” are also enriched in the gene cluster with peak at Day and trough at Dawn only in Ostreococcus and Arabidopsis indicating an apparent anticipation or broader gene expression peaks in these two species with respect to Chlamydomonas, where narrow peaks are observed at the light/dark transition (Figure 8A). For instance, DNA POLYMERASE A4 (Cre07.g312350, ostta13g02040, and At5g41880), MINI CHROMOSOME MAINTENANCE 2 (Cre07.g338000, ostta01g02580, and At1g44900), CYCLIN A1 (Cre03.g207900, ostta02g00150, and At1g44110) present a narrow peak centered on ZT12 in Chlamydomonas, whereas in Ostreococcus and Arabidopsis they present a broader peak centered on ZT6. This could also indicate a better culture synchronization in Chlamydomonas than Ostreococcus and the lack of synchronicity between the different tissues in Arabidopsis. The GO term “Photosynthesis” appeared significantly enriched in the gene clusters with peak at the Day and trough at Night in the three different species (Table 3). Specifically, PHOTOSYSTEM I SUBUNIT D (Cre05.g238332, ostta10g03280, and At1g03130) and PHOTOSYSTEM II SUBUNIT O (Cre09.g396213, ostta14g00150, and At3g50820) exhibit very similar expression patterns with broader peaks in Ostreococcus and Arabidopsis than in Chlamydomonas (Figure 8B). Only small variations were detected, for instance, in the expression profiles of genes codifying for components of plastid ATP synthase and b6f complex. ATP CHLOROPLAST SYNTHASE 1 (ATPC, Cre06.g255900, ostta09g01080, and At4g04640) and PHOTOSYNTHETIC ELECTRON TRANSFER C (PETC, Cre11g467689, ostta07g02450, and At4g03280) present a peak at dawn and trough at dusk in Chlamydomonas and Ostreococcus, while in Arabidopsis ATPC1 is delayed, peaking during the day and PETC anticipates dawn, peaking at night (Figure 8B).

Ribogenesis, the process of ribosome making, is another interesting example of conservation and evolution of the periodic pattern of a biological process. In fact, “Ribosome biogenesis” seems to be enriched mainly in the gene cluster with peak at Dawn and a trough during the Day in Ostreococcus and Chlamydomonas, whereas in Arabidopsis is
<table>
<thead>
<tr>
<th></th>
<th>Arabidopsis thaliana</th>
<th></th>
<th>Ostreococcus tauri</th>
<th></th>
<th>Chlamydomonas reinhardtii</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Size</td>
<td>Enriched GO terms</td>
<td>Genes</td>
<td>Size</td>
<td>Enriched GO terms</td>
<td>Genes</td>
</tr>
<tr>
<td>Peak dawn</td>
<td>0</td>
<td>Protein catabolic process</td>
<td>osta11g00020</td>
<td>422</td>
<td>Transcription from RNA polymerase II promoter</td>
<td>osta14g01300</td>
</tr>
<tr>
<td>Trough dawn</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RNA metabolic process</td>
<td>osta10g00100</td>
</tr>
<tr>
<td>Peak dawn</td>
<td>1,209</td>
<td>Carboxylic acid metabolic process</td>
<td>AT4G17090</td>
<td>1,384</td>
<td>Ribosome biogenesis</td>
<td>osta06g01560</td>
</tr>
<tr>
<td>Trough day</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RNA metabolic process</td>
<td></td>
</tr>
<tr>
<td>Peak dawn</td>
<td>811</td>
<td>Response to auxin</td>
<td>AT1G299440</td>
<td>1,744</td>
<td>Carbohydrate biosynthetic process</td>
<td></td>
</tr>
<tr>
<td>Trough dusk</td>
<td></td>
<td>Carbohydrate biosynthetic process</td>
<td>AT1G70290</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peak dawn</td>
<td>132</td>
<td>Regulation of transcription</td>
<td>AT5G62430</td>
<td>739</td>
<td>Protein localization to membrane</td>
<td>osta02g05390</td>
</tr>
<tr>
<td>Trough night</td>
<td></td>
<td>Organic cyclic compound metabolic process</td>
<td>AT1G20330</td>
<td></td>
<td>Nucleotide metabolic process</td>
<td>osta15g00150</td>
</tr>
<tr>
<td>Peak day</td>
<td>980</td>
<td>DNA replication</td>
<td>AT4G02060</td>
<td>997</td>
<td>DNA metabolic process</td>
<td>osta02g01890</td>
</tr>
<tr>
<td>Trough dusk</td>
<td></td>
<td>Actin cytoskeleton organization</td>
<td>AT1G13180</td>
<td>141</td>
<td>Protein localization to membrane</td>
<td>osta01g00410</td>
</tr>
<tr>
<td>Peak day</td>
<td>127</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trough day</td>
<td></td>
<td>Tricarboxylic acid metabolic process</td>
<td>AT5G04950</td>
<td>514</td>
<td>Tetrapyrrole biosynthetic process</td>
<td>osta02g02380</td>
</tr>
<tr>
<td>Peak day</td>
<td>222</td>
<td></td>
<td></td>
<td></td>
<td>Response to stress</td>
<td>osta08g03450</td>
</tr>
<tr>
<td>Trough dusk</td>
<td></td>
<td>Isoprenoid metabolic process</td>
<td>AT4G15560</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peak day</td>
<td>935</td>
<td>Photosynthesis</td>
<td>AT4G10340</td>
<td>1,060</td>
<td>Signal transduction</td>
<td>osta04g01810</td>
</tr>
<tr>
<td>Trough night</td>
<td></td>
<td>Ribosome biogenesis</td>
<td>AT5G16750</td>
<td></td>
<td>DNA replication</td>
<td>osta10g02580</td>
</tr>
<tr>
<td>Peak day</td>
<td>907</td>
<td>DNA metabolic process</td>
<td>AT2G42120</td>
<td>1,620</td>
<td>DNA metabolic process</td>
<td>osta02g00690</td>
</tr>
<tr>
<td>Trough dusk</td>
<td></td>
<td>Carbohydrate catabolic process</td>
<td>AT2G32230</td>
<td></td>
<td>Carbohydrate metabolic process</td>
<td>osta05g04330</td>
</tr>
<tr>
<td>Peak dusk</td>
<td>432</td>
<td>Vitamin metabolic process</td>
<td>AT2G29630</td>
<td>437</td>
<td>Microtubule-based process</td>
<td>osta10g00680</td>
</tr>
<tr>
<td>Trough day</td>
<td></td>
<td>Peroxisome organization</td>
<td>AT1G47750</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peak dusk</td>
<td>196</td>
<td>Peptide biosynthetic process</td>
<td>AT4G29430</td>
<td>696</td>
<td>Vesicle-mediated transport</td>
<td>osta09g03760</td>
</tr>
<tr>
<td>Trough night</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>DNA replication</td>
<td>osta04g04640</td>
</tr>
<tr>
<td>Peak nightt</td>
<td>122</td>
<td>Hexose metabolic process</td>
<td>AT1G24280</td>
<td>540</td>
<td>RNA processing</td>
<td>osta11g03270</td>
</tr>
<tr>
<td>Trough dusk</td>
<td></td>
<td>Protein modification process</td>
<td>AT1G12830</td>
<td></td>
<td>Protein modification process</td>
<td>osta11g02830</td>
</tr>
<tr>
<td>Peak night</td>
<td>1,014</td>
<td>Deyphosphorylation</td>
<td>AT1G71860</td>
<td>1,234</td>
<td>Ribosome biogenesis</td>
<td>osta09g00820</td>
</tr>
</tbody>
</table>

(Continued)
A closer inspection into ribogenesis in Chlamydomonas reveals an uncoupling between the genesis of cytosolic, plastid and mitochondrial ribosomes (Zones et al., 2015). Cytosolic ribosome components peak at Dusk (i.e., CrRPL5, Cre14.g621450), plastid ribosome components peak at Dawn (i.e., CrPRPL1, Cre02.g088900) and mitochondrial ribosome component present a broad peak at Dawn (i.e., CrMRPL21, Cre09.g388550) (Figure 8C). This uncoupling is less evident in Ostreococcus where genes codifying for components of cytosolic (i.e., OtRPL5, ostta15g01160), plastid (i.e., OtPRPL1, ostta12g02760) and mitochondrial (i.e., OtMRPL21, ostta01g04030) ribosomes exhibit a peak at Dawn and a trough at Day (Figure 8C). No apparent synchronization is observed between genes codifying for ribosome components in Arabidopsis (Figure 8C). Therefore, the analysis shows how the light/dark cycles in ancient algae synchronized all ribosome synthesis at the same time point. In Chlamydomonas ribogenesis was divided into three different stages depending on the nuclear, chloroplast or mitochondrial genome control, while Arabidopsis showed a ribosome biogenesis almost independent from daily rhythms.

The conservation and evolution of the central transcriptional regulators in circadian/photoperiod response in Arabidopsis was analyzed. A strong conservation was observed between the expression profiles of the key circadian/photoperiod regulators in Ostreococcus and Arabidopsis with some variations in Chlamydomonas (Figure 8D). For instance, the three different species showed similar expression profiles of two central photoperiodic genes; DOF (Cre12.g51440, ostta04g02850, At5g62430) and COL (Cre06.g278159, ostta04g03620, At5g15840) peaking at Dawn, with their troughs at Dusk (Figure 8D). Remarkably, two of the central genes in the circadian clock; CCA1 (At2g46830, ostta06g01220) and TOC1 (At5g61380, ostta13g01820) exhibit the same symmetric expression profiles in Ostreococcus and Arabidopsis, while in Chlamydomonas the putative orthologues CrCCA1 (Cre12.g514400) and CrTOC1 (Cre02.g094150) detected in our analysis present the same expression profile, peaking at Dawn with their troughs at Dusk (Figure 8D). This suggests an independent and divergent evolution in Chlamydomonas of the core regulatory circuit of circadian rhythms as previously suggested (Mittag et al., 2005). Nevertheless, the full validation of these results would require more extensive analysis and experimental work.

Other relevant genes involved in photomorphogenesis such as the bZIP TF LONGER HYPOCOTYL 5 (HY5) and HY5-HOMOLOG (HYH) exhibit the same expression pattern in Chlamydomonas (Cre07.g318050, Cre10.g438850), Ostreococcus (ostta03g00340, ostta01g03880), and Arabidopsis (At5g11260, At3g17609), peaking at Dawn with their troughs at Day (Figures 8D, 11C,D).

It was also interesting to observe how the clustering approach worked when the photoperiod changed from ND to LD (Figure 9). This way, when the expression profile by RNAseq of a 24 h Col-0 plant grown in LD (red) was compared with the profile in ND (blue), more than 41% of the total genes only significantly detected in the gene cluster with peak at Day and trough at Night (Table 3).
FIGURE 8 | Comparison of gene expression patterns in key biological processes in *Chlamydomonas*, *Ostreococcus* and *Arabidopsis*. Heatmaps are used to represent gene expression with genes sorted according to their expression profile similarity using hierarchical clustering as in Figure 5. (A) Cell cycle genes such as *REPLICATION PROTEIN A* (*RPA*) and *ORIGIN RECOGNITION COMPLEX 1* (*ORC1*) exhibit a peak at Dusk and trough at Dawn in the three species. In *Ostreococcus* and *Arabidopsis* genes such as *DNA POLYMERASE A4* (*POLA4*) and *MINI CHROMOSOME MAINTENANCE 2* (*MCM2*) are expressed previously to those in *Chlamydomonas*, presenting broader peaks expanding from Day to Dusk with respect to the narrow peaks at Dusk observed in *Chlamydomonas*. (B) Photosystem components such as *PHOTOSYSTEM I SUBUNIT D* (*PSAD*) and *PHOTOSYSTEM II SUBUNIT O* (*PSBO*) follow similar expression patterns in the three species with peaks at Day and troughs at Night. Only small variations were detected in photosynthesis-involved genes such as the expression profiles of *ATP CHLOROPLAST SYNTHASE 1* (*ATPC1*) and *PHOTOSYNTHETIC ELECTRON TRANSFER C* (*PETC*). These genes present a peak at Dawn and trough at Dusk in both algae, while in *Arabidopsis*, *ATPC1* is delayed, peaking during the day and *PETC* anticipates dawn peaking at night. (C) Ribogenesis-involved genes present their peaks at Dawn and troughs at Day in the algae, such as *PLASTID RIBOSOMAL PROTEIN L1* (*PRPL1*) and *MITOCHONDRIAL RIBOSOMAL PROTEIN L21* (*MRPL21*). Cytosolic ribogenesis is uncoupled from organella ribogenesis in *Chlamydomonas* with genes such as *RIBOSOMAL PROTEIN L5, L21* (*RPL5, RPL21*), peaking at Dusk. This uncoupling is not evident in *Ostreococcus* and no apparent synchronization is observed between genes codifying for ribosome components in *Arabidopsis*. (D) Photoperiod/circadian central transcriptional regulators exhibit highly similar expression profiles in the three species. *DOF*, *CCA1*, and *COL* present the same expression pattern; peaking at Dawn with their troughs at Dusk. Some genes, such as *OtPRR* and *TOC1*, present very similar profiles (peaking at Dusk with their troughs at Dawn) in *Ostreococcus* and *Arabidopsis*, whereas in *Chlamydomonas* *CrPRR* follows a symmetric pattern peaking at Dawn and a trough at Dusk.

expressed in LD presented a daily rhythmic profile (11157), while the percentage was reduced to 37% of genes in ND (7640). Out of the 7640 genes showing a periodic regulation in ND, 5143 (deep purple, 67%) could be also identified in LD, while approximately half of the 11157 genes in LD (6014, 54%) did not follow a daily rhythmic pattern in ND. This could mean that increasing the photoperiod also augments the daily rhythmic regulation of the transcriptome in higher plants, although never reaching the numbers shown in microalgae. When we organized the periodic-expressed genes in clusters, based on their peak and trough during the day as in Figure 8, a clear clock-wise distribution was observed (Figure 9B). This seems to indicate that the basic organization of the daily rhythms is independent of the photoperiod. However, when a temporal-clustered gene co-expression network was constructed with the intersection genes between daily rhythmic genes in LD and ND (5143 genes, deep purple) a diffused circadian pattern of expression was observed, suggesting a displacement or alteration in the daily rhythmic expression of a substantial number of genes due to the extended photoperiod (Figure 9C).
applied to determine functional orthologues (Romero-Campero and co-expression patterns have been recently successfully on the integration of sequence similarity, gene expression profiles previously using the MBBH method. Similar approaches based expression patterns was integrated with the results obtained between the three different species, information regarding co-

To study the conservation of daily rhythmic patterns in the different clusters among the three species, beyond the comparison between individual gene expression profiles, the Summary Composite Conservation Statistic (Zsummary) was computed as defined in Langfelder et al. (2011) (Figure 10A). A Zsummary value lower than 2 indicates no conservation, a Zsummary value 2–10 implies a moderate conservation, while Zsummary greater than 10 constitutes evidence of a great level of conservation. For each daily rhythmic gene cluster, the Zsummary for the six different possible comparisons (Arabidopsis vs. Ostreococcus, Arabidopsis vs. Chlamydomonas, Ostreococcus vs. Arabidopsis, Ostreococcus vs. Chlamydomonas, Chlamydomonas vs. Arabidopsis and Chlamydomonas vs. Ostreococcus) was computed and the corresponding average and standard deviation was plotted. In Figure 10A, it can be observed that clusters with a peak at Dawn present a moderate level of daily rhythmic pattern co-expression conservation. In fact, the highest level of conservation was obtained for the cluster with peak at Dawn and trough at Dusk and the cluster peak at Dusk and trough at Dawn (Zsummary > 10). This suggests that evolution has mostly conserved periodic genes with a strong influence from the light/dark and dark/light transitions in their expression profiles, again hinting the importance of these transitory states in plant physiology and their conservation across the entire green lineage.

To corroborate the conservation of co-expression patterns in the three different species, two central pathways identified as daily-regulated in Arabidopsis were chosen (Table 2). For these gene sub-clusters, smaller gene co-expression networks were constructed, namely key regulators in circadian/photoperiod system (Figure 10B) and central enzymes in starch/sucrose metabolism (Figure 10C). In these networks, where red edges indicate positive correlations and blue edges indicate negative correlations, a general conservation between the plots can be observed (Figures 10B,C). The circadian/photoperiod plot revealed a high conservation among the three sub-clusters. Nevertheless, a general higher conservation between Arabidopsis and Ostreococcus display, when compared to Chlamydomonas could be observed, indicating again that this microalgae has slightly differentiated from the core Arabidopsis model clock (Figure 10B). On the contrary, daily-regulated starch metabolic

**Module and Pathway Conservation Reveals High Level of Conservation between Ostreococcus and Chlamydomonas and a Moderate Level of Conservation between Microalgae and Arabidopsis**

With the aim of determining highly likely functional orthologues between the three different species, information regarding co-expression patterns was integrated with the results obtained previously using the MBBH method. Similar approaches based on the integration of sequence similarity, gene expression profiles and co-expression patterns have been recently successfully applied to determine functional orthologues (Romero-Campero et al., 2013; Das et al., 2016). Thus, it was assumed that two MBBH potential orthologous periodic genes from two different species exhibited a conserved daily pattern when both presented their peak and trough in the same time interval (i.e., same circadian cluster) or when the Pearson correlation coefficient between their expression profiles was higher than 0.98. Therefore, such two genes could be named as expresologues (Das et al., 2016).

According to this criterion approximately 34% of Arabidopsis daily rhythmic genes presented an expresologue in Ostreococcus or Chlamydomonas. This suggests a high level of conservation in spite of the large evolutionary distance between flowering plants and microalgae and presumably, along the plant evolutionary lineage. Interestingly, only the Arabidopsis merged cluster of daily rhythmic genes peaking at Dusk was significantly enriched in Ostreococcus and Chlamydomonas expresologues (p-value of 4.84 × 10−5 by Fisher's exact test), suggesting that most of these conserved genes maintain a strong influence of light/dark transitions on their expression control. In Ostreococcus, 83.62% of the daily rhythmic genes present an expresologue in Chlamydomonas but just 36.16% have an expresologue in Arabidopsis. In Chlamydomonas, 52.71% of the daily rhythmic genes present an expresologue in Ostreococcus whereas only 19.05% have an Arabidopsis expresologue, which supports a more divergent evolution in this species when compared to Ostreococcus.

To study the conservation of daily rhythmic patterns in the different clusters among the three species, beyond the comparison between individual gene expression profiles, the Summary Composite Conservation Statistic (Zsummary) was computed as defined in Langfelder et al. (2011) (Figure 10A). A Zsummary value lower than 2 indicates no conservation, a Zsummary value 2–10 implies a moderate conservation, while Zsummary greater than 10 constitutes evidence of a great level of conservation. For each daily rhythmic gene cluster, the Zsummary for the six different possible comparisons (Arabidopsis vs. Ostreococcus, Arabidopsis vs. Chlamydomonas, Ostreococcus vs. Arabidopsis, Ostreococcus vs. Chlamydomonas, Chlamydomonas vs. Arabidopsis and Chlamydomonas vs. Ostreococcus) was computed and the corresponding average and standard deviation was plotted. In Figure 10A, it can be observed that clusters with a peak at Dawn present a moderate level of daily rhythmic pattern co-expression conservation. In fact, the highest level of conservation was obtained for the cluster with peak at Dawn and trough at Dusk and the cluster peak at Dusk and trough at Dawn (Zsummary > 10). This suggests that evolution has mostly conserved periodic genes with a strong influence from the light/dark and dark/light transitions in their expression profiles, again hinting the importance of these transitory states in plant physiology and their conservation across the entire green lineage.

To corroborate the conservation of co-expression patterns in the three different species, two central pathways identified as daily-regulated in Arabidopsis were chosen (Table 2). For these gene sub-clusters, smaller gene co-expression networks were constructed, namely key regulators in circadian/photoperiod system (Figure 10B) and central enzymes in starch/sucrose metabolism (Figure 10C). In these networks, where red edges indicate positive correlations and blue edges indicate negative correlations, a general conservation between the plots can be observed (Figures 10B,C). The circadian/photoperiod plot revealed a high conservation among the three sub-clusters. Nevertheless, a general higher conservation between Arabidopsis and Ostreococcus display, when compared to Chlamydomonas could be observed, indicating again that this microalgae has slightly differentiated from the core Arabidopsis model clock (Figure 10B). On the contrary, daily-regulated starch metabolic

---

**Figure 9** The clustering of the clock-regulated genes does not change substantially between ND and LD conditions in Arabidopsis. (A) Venn diagrams representing genes in LD (red) and ND (blue) clusters. In LD conditions a higher percentage of genes (41%) follow a circadian expression than in ND (33%). Most genes identified as circadian in ND conditions (7640) exhibit circadian patterns also in LD (5143, 76%). Nevertheless, new genes are identified as circadian in LD (6014, 83.62%) indicating a higher circadian dependence in LD. (B) In the gene co-expression network composed of circadian genes in LD (11157), the same ring-like structure as in ND conditions is observed. This is constituted by the sequential distribution of the clusters constituted by genes peaking at Dawn (red), Day (yellow), Dusk (blue), and Night (black). (C) In the representation of the clusters obtained in ND conditions intersecting with the gene co-expression network in LD conditions (deep purple, 5143 genes), an apparent conservation of the circadian pattern is observed.

---

**TABLE 2** Summary Composite Conservation Statistic (Zsummary) for the Arabidopsis vs. Ostreococcus, Arabidopsis vs. Chlamydomonas, Ostreococcus vs. Arabidopsis and Ostreococcus vs. Chlamydomonas comparisons. The conservation level is described as follows: less than 2 indicates no conservation, 2–10 implies a moderate conservation, and greater than 10 constitutes evidence of a great level of conservation.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Zsummary</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis vs. Ostreococcus</td>
<td>8.95</td>
<td>4.84×10−5</td>
</tr>
<tr>
<td>Arabidopsis vs. Chlamydomonas</td>
<td>2.71</td>
<td>4.84×10−10</td>
</tr>
<tr>
<td>Ostreococcus vs. Arabidopsis</td>
<td>1.73</td>
<td>4.84×10−7</td>
</tr>
<tr>
<td>Ostreococcus vs. Chlamydomonas</td>
<td>1.09</td>
<td>4.84×10−4</td>
</tr>
</tbody>
</table>

---

**FIGURE 10A** The clustering of the clock-regulated genes does not change substantially between ND and LD conditions in Arabidopsis. (A) Venn diagrams representing genes in LD (red) and ND (blue) clusters. In LD conditions a higher percentage of genes (41%) follow a circadian expression than in ND (33%). Most genes identified as circadian in ND conditions (7640) exhibit circadian patterns also in LD (5143, 76%). Nevertheless, new genes are identified as circadian in LD (6014, 83.62%) indicating a higher circadian dependence in LD. (B) In the gene co-expression network composed of circadian genes in LD (11157), the same ring-like structure as in ND conditions is observed. This is constituted by the sequential distribution of the clusters constituted by genes peaking at Dawn (red), Day (yellow), Dusk (blue), and Night (black). (C) In the representation of the clusters obtained in ND conditions intersecting with the gene co-expression network in LD conditions (deep purple, 5143 genes), an apparent conservation of the circadian pattern is observed.

---

**FIGURE 10B** The clustering of the clock-regulated genes does not change substantially between ND and LD conditions in Arabidopsis. (A) Venn diagrams representing genes in LD (red) and ND (blue) clusters. In LD conditions a higher percentage of genes (41%) follow a circadian expression than in ND (33%). Most genes identified as circadian in ND conditions (7640) exhibit circadian patterns also in LD (5143, 76%). Nevertheless, new genes are identified as circadian in LD (6014, 83.62%) indicating a higher circadian dependence in LD. (B) In the gene co-expression network composed of circadian genes in LD (11157), the same ring-like structure as in ND conditions is observed. This is constituted by the sequential distribution of the clusters constituted by genes peaking at Dawn (red), Day (yellow), Dusk (blue), and Night (black). (C) In the representation of the clusters obtained in ND conditions intersecting with the gene co-expression network in LD conditions (deep purple, 5143 genes), an apparent conservation of the circadian pattern is observed.
FIGURE 10 | Conservation of circadian patterns in Arabidopsis, Chlamydomonas, and Ostreococcus. (A) The average and standard deviation of the Zsummary conservation score between the three different species was computed for all clusters capturing different daily rhythmic patterns. Only clusters that exhibit a periodic pattern with peak or trough at Dusk or Dawn are significantly conserved over the three different species. (B) The co-expression pattern of the key circadian/photoperiod regulators exhibits a high level of conservation in the three species, but it is more evident for Arabidopsis and Ostreococcus. Red edges represent positive correlation between the corresponding gene expression profiles whereas blue edges indicate negative correlation. Edge width is proportional to the absolute value of the correlation between the corresponding expression profiles. (C) The co-expression patterns between key enzymes in starch/sucrose metabolism are highly conserved among the three different species with slight differences in Chlamydomonas.

Genes showed a higher conservation degree between Ostreococcus and Chlamydomonas, maybe revealing that through the course of evolution, additional regulatory steps were incorporated into the higher plant starch synthesis control (Figure 10C). On the whole, this sub-cluster visualization allowed us to compare different subset of genes that had showed a correlation in our general analysis (Supplementary Tables 1–3) further indicating levels of conservation and divergence among the species and constituting efficient tools to initiate the study on the evolution of particular processes.

To further test the veracity of our approach, the promoters of genes showing a high level of conservation both in sequence similarity (Figures 11A,B, leftmost) and co-expression patterns (Figures 11A,B, middle) in the three species were analyzed. As observed in Figures 11A,B rightmost, the promoters of COLs (Figure 11A) and GBSSs (Figure 11B) orthologues showed also conservation in the TF binding sites identified in the corresponding gene promoters. This result suggests that through evolution, in order to maintain the regulation of these functional networks, whole set of genes have to evolve synchronously and reveal why these networks are so resilient to change even in long evolutionary time scales. Finally, in order to provide an independent validation of the conservation of these daily rhythmic patterns and the capacity to predict orthology in modern plants and algae, the normalized expression of the Arabidopsis bZIP gene HY5 and the putative orthologues detected in this study in Chlamydomonas (CrHY5) and Ostreococcus (OtHY5) from the ND microarray analysis (Figure 11C) were plotted and compared with the expression detected in a 24 h LD course by QPCR experiments (Figure 11D). Both plots showed a remarkable similarity, except that the LD expression profiles of HY5 and orthologues prolonged the daily minimal expression levels compared to the ND profile until the last hours of light, when the mRNA levels started to rise again. Therefore, the expression profiles of HY5 and orthologues showed a very clear conserved pattern, as had been shown in our previous analysis (Figure 8) indicating that the three genes have a high chance of expressing functional orthologues, a line of research that is now being followed in the laboratory.
FIGURE 11 | Putative TF orthologues show a conserved protein domain structure, promoter distribution and circadian expression profile in algae and higher plants. (A) Gene structure (leftmost), expression in 24h ND course (middle) and TF binding site in the promoter of Arabidopsis COL2 gene and putative orthologues from the two microalgae. (B) As above, showing Arabidopsis GBSS gene and putative orthologues. (C) Normalized expression level in 24h ND photoperiod from the microarray and RNAseq experiments of HY5 gene from Arabidopsis (blue), ChlHY5 from Chlamydomonas (red) and OtHY5 from Ostreococcus (green). Notice the high similarity in the expression profile of the three species. (D) Expression levels (Arbitrary Units) of HY5 (blue) from Arabidopsis, ChlHY5 from Chlamydomonas (red), and OtHY5 from Ostreococcus (green) in a 24h course in LD by QPCR. Notice how HY5 circadian expression pattern is also conserved in the three species, although the daily expression lengthens, strongly supporting the idea that they are true orthologues. Time is shown in hours (h) as Zeitgeber Time (ZT). For each time point of the analysis, three biological replicates and three technical replicates were analyzed. Each point shows standard error bars ± s.e.m.

CONCLUSIONS

As the amount of transcriptomics and phylogenomics data accumulates, the complexity of the evolution of the mechanisms governing the control of gene expression in eukaryotes becomes more evident. A paradigm of a specific pattern that controls the transcriptome occurs under alternating light/dark cycles (Millar, 2016). In photosynthetic organisms this control is particularly important due to their dependence on sunlight for most of their main physiological processes and thus, a complex periodic gene expression mechanism can be found as early as in microalgae (Mittag et al., 2005; Corellou et al., 2009). Here, a Systems Biology approach has been used to dissect this dependence and find out mechanisms anciently controlled by daily rhythms and those that have acquired a new regulation. To achieve this, MBBH tool was developed to help discern orthology from species evolutionarily very distant as algae and modern plants. When used in parallel with gene co-expression networks, a very strong toolset to understand how different processes have conserved a periodic regulation through long evolutionary distances has been built. Furthermore, a web-based tool has been implemented to allow the study of the daily rhythmic evolution of any set of genes from algae to modern plants in a similar way to previous web-based tools that allow researchers to compare diurnal expression profiling between monocots and dicots (Mockler et al., 2007). Together with the available pipelines and databases presented in former papers (Romero-Campero et al., 2013, 2016) they constitute strong resources for the research community.

Using these tools, we confirmed that primitive picoeukaryote microalgae, such as Ostreococcus, govern most of its transcriptome (a practical 100%) by a daily rhythmic mechanism under alternating light/dark cycles (Monnier et al., 2010). Other
microalgae have reduced this light dependence, but the number of genes regulated by light is still significantly high. This is the case of *Chlamydomonas* (Harris, 2001), which can perform some complex physiological processes independently of light/dark cycles, now controlled by other factors such as external biotic stimuli. Following a similar tendency, *Arabidopsis* has reduced the temporal dependence of its gene expression control under light/dark cycles to just over a third of its transcriptome (Michael et al., 2008) but is still significantly higher than the 10–15% of genes showing a periodic control in mammals (Lowrey and Takahashi, 2011). Nevertheless, it has been reported that up to 89% of the *Arabidopsis* transcriptome could follow a daily rhythmic pattern when different environmental signals, such as temperature, are combined with light/dark cycles (Michael et al., 2008). Therefore, plants seem to possess several molecular mechanisms to integrate diverse environmental signals beside light/dark cycles to exert daily rhythmic regulation over their transcriptome that are missing in microalgae. Temperature seems to be a key factor among these signals. The lack of potential microalgal orthologues for key genes in the evening loop such as GI and ZTL, that exert the temperature compensation to the clock in *Arabidopsis*, could explain these differences.

Finally, the clustering analysis performed attending to gene rhythmic expression has shown that the co-expression networks in the three organisms are distributed in the same order as the clock, giving the circadian timer a real associative and temporal dimension that participates in the coordination of different physiological processes such as ribosome synthesis or starch metabolism. In algae, those clusters that showed maximum expression during the light/dark transitions are the most abundant, while in *Arabidopsis*, these clusters containing orthologous genes, seem to be expressed during the dark or light periods. It seems then that *Arabidopsis* has been able to predict dawn and dusk in a better way than algae and has advanced gene expression in order to anticipate these transitions. Moreover, when a global conservation analysis is performed, only genes with their peaks and troughs at dawn and dusk are strongly conserved between microalgae and plants, again suggesting a diversification in the temporal control of gene expression during their evolution. Finally, we have also shown the powerful predictable capacity of the approach by picking up putative orthologues that may result in good candidates for further studies in the future. This approach could help to better understand daily rhythmic regulation in algal systems, which may be important to understand the more complex mechanisms of higher plants and shade some light on how these mechanisms have evolved in the green lineage.

**AUTHOR CONTRIBUTIONS**

Pd and FR performed bioinformatics analysis. Pd and MR carried out experimental work. FV, JR, and FR conceived the study, interpreted the results and wrote the paper. All authors read and approved the manuscript.

**FUNDING**

The authors would like to thank funding from projects BIO2011-28847-C02-00 and BIO2014-52425-P (Spanish Ministry of Economy and Competitiveness, MINECO) partially supported by FEDER funding.

**ACKNOWLEDGMENTS**

The authors would like to acknowledge Rafael Álvarez-Romo for the maintenance and assistance with the use of the high performance computing facilities of the CIC-Cartuja Data Processing Center that have been instrumental to the development of this study. We acknowledge support of the publication fee by the CSIC Open Access Publication Support Initiative through its Unit of Information Resources for Research (URICI).

**SUPPLEMENTARY MATERIAL**

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2017.01217/full#supplementary-material

---

**REFERENCES**


Kloosterman, B., Abelenda, J. A., Gomez, M. D. M. C., Oortwijn, M., de los Reyes et al. Plant Circadian Clock Evolution


The Importance of Being on Time: Regulatory Networks Controlling Photoperiodic Flowering in Cereals

Vittoria Brambilla, Jorge Gomez-Ariza, Martina Cerise and Fabio Fornara*

Department of Biosciences, University of Milan, Milan, Italy

Flowering is the result of the coordination between genetic information and environmental cues. Gene regulatory networks have evolved in plants in order to measure diurnal and seasonal variation of day length (or photoperiod), thus aligning the reproductive phase with the most favorable season of the year. The capacity of plants to discriminate distinct photoperiods classifies them into long and short day species, depending on the conditions that induce flowering. Plants of tropical origin and adapted to short day lengths include rice, maize, and sorghum, whereas wheat and barley were originally domesticated in the Fertile Crescent and are considered long day species. In these and other crops, day length measurement mechanisms have been artificially modified during domestication and breeding to adapt plants to novel areas, to the extent that a wide diversity of responses exists within any given species. Notwithstanding the ample natural and artificial variation of day length responses, some of the basic molecular elements governing photoperiodic flowering are widely conserved. However, as our understanding of the underlying mechanisms improves, it becomes evident that specific regulators exist in many lineages that are not shared by others, while apparently conserved components can be recruited to novel functions during evolution.

Keywords: photoperiod, florigen, cereals, flowering, gene regulatory network

INTRODUCTION

Several plant species measure day length to start specific developmental switches, e.g., the transition to reproductive growth, during the most appropriate times of the year. Seasonal variation of day length provides a fundamental parameter to synchronize developmental changes, because it is not subject to fluctuations like other environmental cues such as temperature.

Plants can be categorized as long day (LD) or short day (SD) species, depending on the photoperiod most effective at triggering reproductive growth. When day length exceeds a specific critical threshold, flowering is promoted in LD plants, whereas SD plants flower in response to reduction of day length below a critical threshold. Such thresholds are characteristic of each species and largely determined by the region where the species originated and first adapted. Plants growing at low tropical latitudes tend to flower in response to exposure to long nights, whereas species adapted to higher latitudes promote flowering during seasons characterized by LD, indicative of the warm days of spring and summer. Plants adapted to temperate regions that germinate before winter, often also need to satisfy a vernalization requirement (exposure to low non-freezing temperatures for several weeks) to become competent to respond to photoperiodic induction. Additionally, many plants can promote flowering even after long exposures to non-inductive
photoperiodic conditions, indicating a facultative response to day length and the existence of floral promoting stimuli that can bypass the requirement for specific conditions. Therefore, plant interactions with its growth environment can be complex, and gene networks have evolved that respond to changing seasonal parameters.

In crop species, responses to day length have been extensively manipulated, creating varieties that can grow, flower and set seeds at latitudes outside of the range occupied by the wild progenitor. Artificial adaptation to broad latitudinal ranges has been a key step during domestication of several species, allowing cultivation and diversification in many regions of the globe. Natural genetic variation has offered the substrate for human selection and remarkably, many domestication loci encode orthologous genes in distantly related species providing a molecular perspective to look at conservation and evolution of pathways regulating flowering.

Here, we will summarize recent advances in understanding photoperiodic flowering regulation in crop species, focusing on cereals. Starting with the tenets established using Arabidopsis as model system, we will discuss how conserved and unique elements have been deployed to evolve flowering networks of LD and SD plants and how they control production of a florigenic systemic signal in leaves.

**ARABIDOPSIS CONTRIBUTED TO DEVELOP THE BASIC TENETS OF PHOTOPERIODIC FLOWERING**

Photoperiodic flowering has been mostly studied using the dicot Arabidopsis, through which core genetic and molecular mechanisms at the base of the process have been characterized (Song et al., 2015). Arabidopsis might not be fully representative of all plant species but it provides a conceptual framework that can be implemented in other species and also used to discuss evolution of distinct mechanisms typical of distantly related plants (Table 1).

Flowering of Arabidopsis is promoted under LD. The circadian clock is responsible for the rhythmic expression of several factors implicated in environmental responses. Among them, the GIGANTEA (GI) and FLAVIN BINDING KELCH REPEAT F-BOX PROTEIN 1 (FKF1) proteins are expressed at the end of the light phase and interact in a light-dependent fashion (Sawa et al., 2007). The resulting complex targets a group of CYCLING DOF FACTORS (CDFs) for proteasome-mediated degradation (Fornara et al., 2009). The CDFs encode transcriptional repressors that limit expression of the CONSTANS (CO) zinc finger transcription factor, a central regulator within the photoperiodic flowering pathway (Putterill et al., 1995). Besides the major GI-FKF1-CDFs module, several additional mechanisms contribute to CO expression at the transcriptional and post-transcriptional level, including regulation by transcription factors (Ito et al., 2012), alternative splicing (Gil et al., 2016), photoreceptors (Valverde et al., 2004; Song et al., 2014), as well as ambient temperature signals (Fernández et al., 2016), hormonal signals (Wang et al., 2016) and post-translational modifications (Sarid-Krebs et al., 2015). However, central to the current model for photoperiodic flowering, the most prominent feature of CO is its light-dependent stability (Valverde et al., 2004; Song et al., 2012). During the night and the morning, CO protein is unstable and quickly degraded (Jang et al., 2008; Song et al., 2014; Lazaro et al., 2015). Consequently, its expression is shaped to be highest under LD, during the light phase. At this time of the diurnal cycle, CO protein, acting in the companion cells of the phloem, can directly promote expression of FLOWERING LOCUS T (FT), component of the systemic florigenic signal (An et al., 2004; Corbesier et al., 2007; Mathieu et al., 2007).

The effects of CO protein on the levels and rhythmicity of FT mRNA abundance are mediated by several classes of protein interactors that include transcription factors and transcriptional co-regulators, photoreceptors, histone-like proteins, and ubiquitin ligases (see Brambilla and Fornara, 2016 and references therein). Therefore, the photoperiodic flowering pathway, despite being largely interconnected with other regulatory pathways, can be simplified into a linear molecular cascade, whose major output is the FT protein (Figure 1).

**REWIRING PHOTOPERIODIC NETWORKS IN RICE MODIFIES DAY LENGTH RESPONSES**

Rice flowering is accelerated by exposure to SD. Seasonal and diurnal time measurements are mediated by a circadian clock that shares components with that of Arabidopsis, and when mutated results in altered sensitivity to the length of the day (Izawa et al., 2011; Matsubara et al., 2012). Homologs of GI, FKF1, the CDFs, CO, and FT exist in rice and have been partly linked in a cascade that resembles the photoperiodic pathway of Arabidopsis (Shrestha et al., 2014) (Table 1). The OsGI and OsFKF1 proteins can interact with each other and with a CDF protein, OsDOF12, similarly to their Arabidopsis homologs (Li et al., 2009; Han et al., 2015). However, mutations in OsFKF1 delay flowering under any photoperiod tested, whereas osgi mutants are late flowering under SD, while having only mild effects under LD (Hayama et al., 2003; Izawa et al., 2011). The phenotypic effects of the two mutations are therefore different. Overexpression of OsDOF12 increases transcription of Heading Date 3a (Hd3a), a homolog of FT, under LD while having no impact on transcription of Heading date 1 (Hd1), a homolog of CO. Thus, the function of OsDOF12 is opposite to that of Arabidopsis CDFs, effectively promoting flowering (Li et al., 2009). It is still unclear whether the interaction between OsGI and OsFKF1 is dependent upon the photoperiod, or if it is necessary for the degradation of OsDOF12 or other DOF proteins. These data indicate that a similar arrangement of regulators exists upstream of Hd3a, but that their molecular function or day length-dependency is very different from Arabidopsis. Both the DOF-CO and the GI-FKF1 modules are evolutionarily ancient as indicated by data from the unicellular alga Chlamydomonas reinhardtii and the liverwort Marchantia polymorpha, where they control phase transition (Kubota et al., 2014; Lucas-Reina et al., 2015). However, evolution
has likely re-shaped the function of the dimer several times, readjusting it depending on the species.

Cloning of *Hd1* indicated that it encodes a homolog of *CO* (Yano et al., 2000). However, the Hd1 protein not only promotes flowering under SD but also represses it under LD. Mutations in *Hd1* result in accelerated flowering under LD and have been extensively introgressed in varieties cultivated at high latitudes (Izawa et al., 2002; Hayama et al., 2003; Gao et al., 2014; Gómez-Ariza et al., 2015; Goretti et al., 2017). A second important flowering QTL, *Early Heading Date 1* (*Ehd1*) was later cloned and shown to encode a B-type response regulator (Doi et al., 2004). *Ehd1* integrates circadian and light inputs and is required to promote flowering under both LD and SD (Itoh et al., 2010), and to modulate it also in response to abiotic stress, including water deficit (Galbiati et al., 2016; Zhang et al., 2016). Under SD, *Ehd1* induces flowering mainly by promoting *Hd3a* expression, and this function is not shared with dicot species (Zhao et al., 2015). Under LD, expression of *Ehd1* is limited by several repressors that delay flowering, including *Grain Number Plant Height and Heading Date 7* (*Ghd7*), *Hd1*, and *Pseudo Response Regulator 37* (*PRR37*) (Gao et al., 2014; Gómez-Ariza et al., 2015). The *Hd1* and *Ghd7* proteins interact forming a repressor dimer and at least the *Ghd7* protein can directly bind the promoter of *Ehd1* (Nemoto et al., 2016). Thus, genetic and molecular evidences indicate how a conserved inductive cascade has been repurposed and integrated with unique components to create a novel network topology (Figure 1).

As with all photoperiodic response networks, the major outputs of the regulatory cascade include *Hd3a* and its paralog.
MECHANISMS OF PHOTOPERIODIC FLOWERING IN OTHER SHORT DAY MONOCOTS INCLUDING SORGHUM AND MAIZE

Sorghum (Sorghum bicolor) is a SD plant evolved in Africa, in the Sudan region. Six major QTLs controlling flowering time and termed Maturation loci (Ma1–Ma6) have been detected in sorghum. Almost all QTLs have been identified as photoperiodic flowering regulators and their study is demonstrating the strong homology occurring between the sorghum and rice pathways (Wolabu and Tadege, 2016) (Table 1).

Cloning of the Ma3 locus showed that it encodes SbPhyB, a light receptor which can mediate light signaling and flowering repression (Childs et al., 1997). When SbPhyB is mutated, sorghum becomes insensitive to the photoperiod and flowers early compared to the wild type both under LD and SD (Yang et al., 2014a). One of the functions of SbPhyB is to promote the transcription of SbPRR37 (possibly Ma1) and SbGhd7 (Ma6). These genes encode flowering repressors that limit mRNA expression of downstream targets under LD, including Ehd1, SbFT, and SbZCN8 (collinear orthologs of Ha3a and maize ZCN8, respectively) (Murphy et al., 2011). The flowering suppressor role of these sorghum genes reflects the function of rice OsPRR37 and Ghd7, indicating that these components are shared among SD cereals. Recent data suggested that the Ma1 QTL does not correspond to PRR37, but rather to an FT-like gene, SbFT12, that could act as floral suppressor (Cuevas et al., 2016; Wolabu and Tadege, 2016). Additional data will be required to confirm the true identity of the Ma1 gene.

The regulation of SbCO transcription mediated by SbPRR37 has also been investigated. The data suggest that SbPRR37 modulates SbCO expression at dawn, promoting its transcription under LD, whereas under SD SbCO expression seems not to depend upon SbPRR37 (Murphy et al., 2011). SbCO can activate florigen production under both SD and LD conditions through the activation of SbEhd1, SbCN8, and SbCN12 (Yang et al., 2014b). The role of sorghum SbCO as constitutive floral activator is therefore different from that of rice Hd1, implicating a different regulatory mechanism.

Thirteen different FT-like genes have been identified in the sorghum genome, three of which (SbFT11/SbCN15, SbFT8/SbCN12, and SbFT10/SbCN8) could promote flowering when constitutively expressed in Arabidopsis (Yang et al., 2014b; Wolabu et al., 2016). The transcripts of SbCN8, SbCN12, and SbCN15 peak at dawn but show distinct sensitivities to SbCO mutations. Whereas the transcripts of SbCN8 and SbCN12 are
strongly reduced in the Sbco mutant background under LD, SbCN15 shows only a phase shift, suggesting different regulation by SbCO (Yang et al., 2014b). The transcriptional patterns of SbCN8, SbCN12, and SbCN15 under different photoperiods and mutant backgrounds could provide in the future valuable data to understand similarities and differences with the dual florigen system of rice.

Maize (Zea mays) was domesticated in central Mexico from Teosinte, which is a SD plant. The first flowering gene cloned in maize was INDETERMINATE 1 (ID1): plants with mutations in this gene delay the floral transition and produce aberrant inflorescences (Colasanti et al., 1998). ID1 encodes a zinc-finger transcription factor expressed in immature leaves which can activate the floral transition and is not under the control of the circadian clock (Wong and Colasanti, 2007). Although the precise function of ID1 in the photoperiodic pathway is still unclear, recent analyses demonstrated that ID1 controls chromatin modifications of loci encoding maize florigens, and that it can regulate flowering through histone methylations (Mascheretti et al., 2015). A rice homolog of ID1, OsEhd2, is required to induce OsEhd1 expression (Matsubara et al., 2008) (Table 1). Although a maize Ehd1 homolog has not yet been found, the high homology between ID1 and OsEhd2 could suggest a similar regulatory mechanism, possibly indicating the existence of a ZmEhd1-like protein subject to similar regulation. Indirect evidence supporting this view is that the CCT-domain transcription factor ZmCCT shows sequence homology with OsGhd7, and encodes a strong LD flowering repressor (Figure 1). Mutations in ZmCCT cause early flowering and have been artificially selected to expand maize cultivation to higher latitudes (Hung et al., 2012).

Two GI homologs are present in maize, GIGANTEA1 (GI1) and GIGANTEA2 (GI2) (Miller et al., 2008). In Arabidopsis and rice, GI is under circadian clock control and regulates the expression of several genes important for the floral induction. In maize, gi1 mutations cause early flowering under LD conditions. Transcriptional analysis of these mutants demonstrated that GI1 is necessary to repress transcription of CONZ1 (homolog of OsHd1) and ZCN8 (homolog of Hd3a), both of which displayed increased expression in the gi1 background (Bendix et al., 2013). These data demonstrate that ZmGI function is similar to OsGI which can repress flowering under LD conditions, a function opposite to that of AtGI (Hayama et al., 2003). Whether mutations in CONZ1 influence flowering is unknown, but the data suggest it to be downstream of GI1, and possibly upstream of ZCN8 as positive regulator of flowering (Miller et al., 2008).

From the analysis of 15 maize FT-like genes, ZCN8 was identified as the strongest candidate for the maize florigen (Meng et al., 2011). ZCN8 encodes a homolog of FT that delays flowering if silenced, and can complement ft mutants when expressed in Arabidopsis (Lazakis et al., 2011). The regulation of ZCN8 is similar to that of another putative maize florigen, ZCN7, and is under the control of chromatin modifications governed by ID1 (Mascheretti et al., 2015). However, whether ZCN7 satisfies the criteria of a florigenic protein is still to be clarified.

FLOWERING MECHANISMS IN LONG DAY TEMPERATE CEREALS

Differently from rice, sorghum, and maize, the temperate cereals wheat (Triticum spp.) and barley (Hordeum vulgare) were domesticated in the Eastern Mediterranean region, in areas characterized by the alternation of cold and warm seasons. These cereals have evolved mechanisms to prevent flowering when temperatures are low, to protect the meristem from cold damage. Flowering is promoted after exposure to vernalizing conditions, when plants resume growth in the spring. During domestication, some cultivars of these species have lost sensitivity to vernalization and, depending on the response to cold, they could be classified as winter or spring types. Winter-types have an obligate vernalization requirement. Such response is controlled by the VERNALIZATION (VRN) loci (Ream et al., 2012). VRN1 is a MADS-box floral promoter homologous to FRUITFULL (FUL) and APETALA1 (API) of Arabidopsis, whereas VRN2 is a floral repressor sharing sequence similarity to Ghd7 of rice (Table 1).

Under low temperatures, the expression of VRN1 is induced and the protein directly binds to the promoter of VRN2 to reduce its expression during vernalization (Treviskis, 2006; Deng et al., 2015). Dominant mutations in VRN1 or recessive mutations in VRN2 confer a spring growth habit, and have been exploited by breeders to expand cultivation areas (Yan et al., 2004; Fu et al., 2005; Loukoianov, 2005).

Downregulation of VRN2 is required to induce VRN3 expression during the floral transition. VRN3 proteins (designated as TaFT and HvFT in wheat and barley, respectively) are homologs of the Arabidopsis and rice florigens, and move to the apical meristem to promote flowering upon exposure to warm temperatures and LD (Yan et al., 2006; Li and Dubcovsky, 2008). Thus, cold signals coordinate VRN expression to activate flowering and long-distance florigenic signaling only when a vernalization requirement has been satisfied.

As soon as VRN2 levels decrease, exposure to LDs is required to promote flowering. Temperate cereals flower earlier under LDs, whereas exposure to SDs delays flowering. The PHOTOPERIOD 1 (Ppd1) gene has been described as the major factor controlling sensitivity to day length in wheat and barley (Turner et al., 2005; Beales et al., 2007). Mutations in PPD1 delay flowering under LD and reduce VRN3/FT expression. PPD1 proteins are homologous to PRR37 proteins of rice and sorghum, both of which repress flowering under LD. The functional divergence of PRR37 proteins observed among LD temperate and SD tropical cereals deserves further attention, as it might be at the base of their distinct photoperiodic requirements.

Homologs of CO and Hd1 have been identified in wheat and barley (Campoli and Von Korff, 2014). The TaHd1-1 gene could complement a rice hd1 mutant, suggesting functional conservation of protein function in a heterologous system (Nemoto et al., 2003). In barley, studies based on overexpression have provided important clues to the position of Hd1 homologs in flowering regulatory networks. Overexpression of HvCO1 and HvCO2 promoted flowering under both LD and SD, but plants retained sensitivity to the photoperiod, because of independent control of HvFT1 by PPD1 (Campoli et al., 2012). Thus, barley
flowering depends on two parallel pathways controlling FT expression (Figure 1). Interestingly, overexpression of HvCO2 was recently shown to increase expression of VRN2 under LD and SD in a winter variety (Mulki and von Korff, 2016). Despite such increase of the VRN2 repressor, overexpression of HvCO2 could still promote flowering, likely through a VRN2-independent pathway. The data might suggest that HvCO2 mediates a floral repressive function through VRN2, to limit FT expression. Whether barley orthologs of Hdl display dual functions similarly in rice Hdl awaits further testing. The use of mutant resources and possibly of edited alleles might help to address this issue.

CONCLUDING REMARKS

The examples discussed above illustrate the flexibility of photoperiodic flowering networks and how adaptation to distinct environments modifies their topology. Major changes include the integration of vernalization modules in some networks and the recruitment of non-shared regulators, such as Ehd1 and Ghd7, in others. A common theme appears to be the requirement for upstream master regulators to control expression of FT-like genes, but their number and relative contributions to heading time broadly varies between species. While in Arabidopsis, CO acts as central and primary regulator of FT, CO homologs in crops are coupled to parallel pathways largely sharing the workload, and FT expression often strongly depends on additional regulators.

Efforts will be needed in the future to isolate all components of the networks in crop species, many of which are still to be cloned. Quantification of transcripts offers a rapid way of determining relationships between genes, but provides only limited information on protein expression or biochemical function. Finally, molecular networks are starting to be built, based on protein–protein or protein–RNA interactions especially in rice. Expanding these efforts toward other crops will prove necessary.

REFERENCES


Conflict of Interest Statement: 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.
Expansion and Functional Divergence of AP2 Group Genes in Spermatophytes Determined by Molecular Evolution and Arabidopsis Mutant Analysis

Pengkai Wang1,2, Tielong Cheng1, Mengzhu Lu3, Guangxin Liu1, Meiping Li1, Jisen Shi1, Ye Lu1, Thomas Laux4* and Jinhui Chen1*

1 Ministry of Education, Key Laboratory of Forest Genetics and Biotechnology, Nanjing Forestry University, Nanjing, China, 2 Suzhou Polytechnic Institute of Agriculture, Suzhou, China, 3 Laboratory of Biotechnology, Chinese Academy of Forestry, Beijing, China, 4 Institute of Biology III, University of Freiburg, Freiburg, Germany

The APETAL2 (AP2) genes represent the AP2 group within a large group of DNA-binding proteins called AP2/EREBP. The AP2 gene is functional and necessary for flower development, stem cell maintenance, and seed development, whereas the other members of AP2 group redundantly affect flowering time. Here we study the phylogeny of AP2 group genes in spermatophytes. Spermatophyte AP2 group genes can be classified into AP2 and TOE types, six clades, and we found that the AP2 group homologs in gymnosperms belong to the AP2 type, whereas TOE types are absent, which indicates the AP2 type gene are more ancient and TOE type was split out of AP2 type and losing the major function. In Brassicaceae, the expansion of AP2 and TOE type lead to the gene number of AP2 group were up to six. Purifying selection appears to have been the primary driving force of spermatophyte AP2 group evolution, although positive selection occurred in the AP2 clade. The transition from exon to intron of AtAP2 in Arabidopsis mutant leads to the loss of gene function and the same situation was found in AtTOE2. Combining this evolutionary analysis and published research, the results suggest that typical AP2 group genes may first appear in gymnosperms and diverged in angiosperms, following expansion of group members and functional differentiation. In angiosperms, AP2 genes (AP2 clade) inherited key functions from ancestors and other genes of AP2 group lost most function but just remained flowering time controlling in gene formation. In this study, the phylogenies of AP2 group genes in spermatophytes was analyzed, which supported the evidence for the research of gene functional evolution of AP2 group.

Keywords: AP2 group gene, spermatophyte, phylogeny, selective pressures, arabidopsis mutants, functional divergence

INTRODUCTION

The genes in AP2/ERF family can be divided into subfamilies according to the number of AP2/ERF domains. There is a single AP2/ERF domain in each member of the EREBP subfamily, which function in the signal transduction pathways of stress responses and cambial tissue development (Mizoi et al., 2012; Licausi et al., 2013). The members of the AP2 subfamily contain two
AP2/ERF domains is further classified into two monophyletic groups: the AP2 group and the AINTEGUMENTA (ANT group, Shigyo et al., 2006). We selected the AP2 group genes for our study because they played key roles in the reproductive and vegetative organs development (Ohto et al., 2005; Huijser and Schmid, 2011).

There are seven conservative domains in a typical AP2 group gene, including one ethylene-responsive element binding factors (ERF) - associated amphiphilic repression motif (EAR) (Kagale et al., 2010) or EAR-like domain, a nuclear localization signal (NLS) domain, two AP2 (AP2-R1 and AP2-R2) domain (Kim et al., 2006), a linkage domain (connecting the AP2-R1 with R2), another EAR domain, and a MIR172 target site (Image 1). However, not all AP2 group genes contain two typical complete AP2 domains. For example, there are six members in the AP2 gene group in Arabidopsis, TARGET OF EARLY ACTIVATION TAGGED 1-3 (TOE1-3), AP2, SCHLAFMUTZE (SMZ), and SCHNARCHZAPFEN (SNZ). Among these six Arabidopsis genes, AP2, TOE3, and TOE1 contain both complete AP2 domains (AP2-R1 and R2 domains) but there is only one typical AP2 domain (AP2-R1 domain) in TOE2, SMZ, and SNZ (Image 1). The AP2-R2 domain in these three genes are not the same as AP2, TOE3, and TOE1.

In Arabidopsis, AP2 regulates floral development, (Jofuku et al., 1994), stem cell maintenance, (Wurschum et al., 2005), seed developmenta, whereas the remaining five genes (TOE1-3, SMZ, and SNZ) act redundantly as flowering repressors (Zhu and Helliwell, 2011). mRNA abundance and translation of TOE1-2, AP2, SMZ, and SNZ in Arabidopsis are regulated by miR172, which is also important for regulating phase transition and determining floral organ identity in monocotyledons (Nair et al., 2010; Zhu and Helliwell, 2011). TOE3 most likely acts redundantly with TOE1 and TOE2 to repress flowering. A good candidate for such a repressor is SMZ, which was originally identified in an activation-tagging screen because of its dominant late-flowering phenotype. Additionally, SNZ, a paralog of SMZ, represses flowering when expressed at high levels. Several other known regulators of flowering time have been identified as SMZ targets. Among them are SMZ itself, SNZ, AP2, and TOE3, suggesting complex feedback regulation among AP2 group members.

Several studies have examined the origin, phylogeny and evolution of the AP2/ERF family and the AP2 subfamily (Magnani et al., 2004; Kim et al., 2006; Shigyo et al., 2006). With the development of high-throughput DNA sequencing techniques, increasing information has become available in recent years for the AP2/ERF family based on genomic data from species such as rice, apple, peach, Hevea brasiliensis, Prunus mume, Vitis vinifera, Populus trichocarpa, Chinese cabbage, and others (Zhuang et al., 2008; Licausi et al., 2010; Sharoni et al., 2011; Zhang et al., 2012; Du et al., 2013; Duan et al., 2013; Song et al., 2013). However, these studies have mainly focused on the identification, classification and expression of AP2/ERF family genes. To date, there have been no studies describing the molecular evolutionary history and the structural characterization of the AP2 group in spermatophytes. As such, an evolutionary and structural analysis of this group may provide both a reference for further functional studies and evidence for gene functional diversification.

Here, we examined 105 spermatophyte AP2 group genes from 56 spermatophytes by phylogenetic analyses, comparing whole gene sequences, homeodomains, and other motifs. The results revealed that, in general, the spermatophyte AP2 group experienced background purifying selection throughout evolution. However, analyses also showed that the AP2 group genes have also undergone positive selection, despite little evidence for positive pressure on these genes. In particular, the evolutionary relationships among members of the AP2 group were apparent from the divergence between the TOE and AP2 type. By analyzing the expression patterns, functional data and phylogenetic relationships among AP2 genes, we reveal rules concerning the formation of new genes in the AP2 group and identified the pathway of functional evolution. We also find evidence that the AP2 function in maintaining the stem-cell niche is to be conserved in spermatophytes.

RESULTS

The Orthologs of AP2 Group Genes from Spermatophytes Differ and Can Be Classified into Two Types and Six Clades

The composition of AP2 group orthologs differs among spermatophyte species. It is well known that the Arabidopsis AP2 group has six members, namely AtAP2, AtTOE1-3, AtSMZ, and AtSNZ. In fact, blast analysis of spermatophyte AP2 group gene sequences revealed that only some species in Brassicaceae contain five or six AP2 group members. Orthologs of AtTOE2, AtTOE3, AtSMZ, and AtSNZ were not found in the other species included in our study. Although the AP2-R2 domain in the AtTOEI orthologs of some species is not complete, these genes still clustered together to form the TOE1 clade in the prephylogenetic analysis. Therefore, the orthologs of AtTOE2, AtTOE3, AtSMZ, and AtSNZ are only been identified from Brassicaceae (Data sheet 1).

All predicted spermatophyte AP2 group protein sequences (105, Data sheet 1) were retrieved from the plant genome (Phytozome and NCBI) and protein databases (NCBI) and used to construct a maximum-likelihood phylogenetic tree (Figure 1 and Image 2). According to the simplified phylogenetic tree (Figure 1) of spermatophyte AP2 group, all genes were categorized as two types: the AP2 type, which included the three clades TOE3, AP2-like and AP2, and the TOE1 type, which included the three clades TOE1, TOE2, SMZ/SNZ. The results of the phylogenetic analysis were consistent with those of the sequence search. For each ortholog, most of the spermatophyte sequences clustered together to form an independent clade, except in gymnosperms. The genes AP2, TOE1, TOE2, SMZ/SNZ, and AP2L from gymnosperms Cucurbitaceae (CyrAP2L), Ginkgo biloba (GibAP2L), Picea sitchensis (PiaAP2L), Larix × marschallii (LamAP2La, b), Picea abies (PiaAP2La, b), and Pinus thunbergii (PitAP2La, b) are well clustered into form six independent clades (Figure 1, bootstrap...
value > 80%). AP2L sequences from gymnosperms were obtained from the NCBI database and clustered together with the AP2 and TOE3 clades to form a larger group, which implied that AP2 genes might be relatively ancient in the AP2 group. The two sub-branches of Pinaceae in the AP2L cluster reflected the duplication of AP2 genes in gymnosperms. In each clade, most of the sequences from species within a single family or order clustered together well to form an independent group (Image 2, bootstrap value > 69%). These results indicated that most of these sequences are specific at the family level. Intriguingly, in the branches of AP2 and TOE1, the phylogenetic trees are very similar to the structure of the Angiosperm Phylogeny Group system, which classifies basal angiosperms, monocots, and dicots into three independent groups. In the AP2 clade, the dicots such as Ricinus communis, Manihot esculenta, Jatropha curcas, Populus trichocarpa, Carica papaya, Vitis vinifera, and Betula platyphylla with unisexual flowers formed a branch with lower bootstrap value (bootstrap value = 25). However, the branch of unisexual flower didn’t appear in the TOE1 clade.

The Distribution of Homeodomains Varies Significantly in Different Clades and Types

There are seven common homeodomains in the typical AP2 group genes according the analytical results of MEME and Pfam (Bailey et al., 2009; Punta et al., 2011): the first EAR domain (DLNxxP or LxLxL), NLS domain, AP2-R1 domain, linkage domain, AP2-R2 domain, the second EAR domain (LxLxL), and miRNA172 target site. These homeodomains except miRNA172 target site are conserved in amino acid sequences. NLS domain, AP2-R1 domain, linkage domain, the second EAR domain and miRNA172 target site have greater sequence similarity than the first EAR domain and AP2-R2 domain.

Notably, the EAR domains differed among the clades. In the TOE3 clade (Image 2), there is no first EAR domain. Likewise, PtAP2, AdAP2, EgAP2, StAP2, MdTOE1, AtTOE1, GpTOE1, NhTOE1, and orthologs from the Rutaceae family also do not contain the first EAR domain. Interestingly, both EAR domains are missing in TOE1 orthologs from the Poaceae family. LamAP2La, PiaAP2La, and PisAP2L, three complete protein-coding genes in the AP2L clade, contain two EAR...
motifs. Because some N-terminal amino acid sequences of the AP2L clade were incomplete, it was not clear whether the first EAR domain exists in all AP2L clade members from gymnosperms. The EAR domains are of two types, namely DLNxxP and LxLxL. The amino acid sequence of the first EAR domain in the AP2 type (including LamAP2La, PiaAP2La, and PisAP2L) was DLNxxP, but the sequence in the TOE1 type was LxLxL, which is the same as the second EAR domain (LxLxL). Outside of the EAR domains, differences in the AP2-R2 domain were identified among AP2 group proteins. There was an incomplete AP2-R2 domain in the amino acid sequence of the TOE2 clade lacking a 15-residue insertion (Image 1). There are significant amino-acid sequence differences between the typical AP2-R2 domain and the SMZ/SNZ clade. The TOE2 clade has a closer phylogenetic relationship to the SMZ/SNZ clade, forming a TOE type with the TOE1 clade, indicating a common ancestry separate from the AP2 type.

The Distribution of Motifs Reflects Differences and Phylogenetic Relationships across Clades and Species

Motifs of the AP2 group were investigated, and 25 motifs including the AP2-R2 domain of the SMZ/SNZ clade, i.e., Motif 8, were then characterized relative to the homeodomains (Figure 2). Most of these motifs were located upstream of the NLS (7/25) or downstream of the second EAR domain (14/25). Only one motif was identified at the C-terminus (i.e., the downstream of the miR172 target site) of most AP2- and TOE-type proteins, and no C-terminal motif was seen in the proteins of clades TOE3 and SMZ/SNZ and TOE1 proteins of
the Brassicaceae and Poaceae families (Figure 2). We also noted that most AP2 clade proteins contained more motifs (15/25), but there was only one motif (Motif 4) in TOE3 clade proteins. The differences in numbers of motifs were seen mostly in one region, i.e., between the second EAR motif and the miR172 target site. Most AP2 clade proteins had three or four motifs, whereas other proteins had only one or two. The different clades and species could be characterized based on the distribution of motifs. Every clade contained its own unique motifs: AP2 clade, Motifs 1 and 9–17; AP2L clade, Motif 18; TOE1 clade, Motifs 21 and 24; TOE2 clade, Motif 25; SMZ/SNZ clade, Motifs 7, 8, and 22). This indicated that the motifs might be related to the functional divergence of the AP2 group (Figure 2). Motif 4 was shared by almost all AP2 group proteins (Figure 2).

Interestingly, the distribution of motifs revealed a phylogenetic relationship among AP2 group proteins that agreed with the results of the phylogenetic analysis by MEGA (Figure 1). In the AP2L clade, all sequences were from gymnosperms and contained five or six motifs (Motifs 2, 4, 5, 18, 19, and 23), suggesting that these motifs are more primitive. Notably, Motifs 5 and 23 are specific to clades AP2 and AP2L, and Motif 19 is also specific to clades TOE1 and TOE2, which may reflect the phylogenetic relationship between AP2 group members from clade AP2L and clades AP2 and TOE1. The proteins in the AP2 clade from Brassicaceae and Poaceae contain unique motifs, i.e., Motifs 13–15 in Brassicaceae and Motifs 16 and 17 in Poaceae, forming two separate branches. By contrast, Brassicaceae and Poaceae TOE1 proteins lacked unique motifs, suggesting that AP2 evolved faster than TOE1. Another observation supporting this view is that Motif 19 is present in clades AP2L and TOE1 but not in clade AP2.

Purifying Selection was the Main Driving Force in the Evolution of Spermatophyte AP2 Group Genes, but Positive Selection Still Occurred, Mostly in Clade AP2

One criterion for assessing the type of selective pressure at the protein level is to calculate $\omega$, i.e., $dN/dS$, for protein-coding genes (Seo et al., 2004; Kryazhimskiy and Plotkin, 2008). The $dN/dS$ ratio $\omega$ provides a criterion for assessing selective pressure at the protein level (Zhang et al., 2006). The $\omega$-values of $>1$, $1$, and $<1$ imply positive selection, neutral evolution and purifying selection, respectively. In the graph of the AP2 group sequences, most of the points fell between the $dS$ axis and the diagonal, indicating $dN < dS$ (Figure 3) and suggesting that purifying selection dominated the selection process during evolution. Similar $\omega$-values were obtained for each family (order) (Figure 3B), which also contained the different clades and the comparison between them ($p < 0.01$, Z-test, Figure 3B). Calculation of the ratio of nucleotide substitutions in a one-by-one comparison of $dN$ to $dS$ for individual AP2 group genes within families (orders) (Figure 3B) and clades (Figure 3B) provided further evidence for purifying selection.

The $dN$ of the TOE1 clade was higher than that of the AP2 clade (Figure 3B), implying that more amino acids changes accumulated in TOE1 during evolution, similar to the results of the sliding-window analyses of $\omega$ among clades AP2, AP2L, and TOE1 (Figure 4). The sliding-window $\omega$-tests on clades AP2, AP2L, and TOE1 (Figure 4) showed similar $\omega$ curves in the different clade lineages. The $\omega$-values of the seven common homeodomains (the two EARs, NLS, AP2-R1, linkage, and AP2-R2 domains and the miR172 target site) in the three

![FIGURE 3](#) | Pairwise Comparison Plots of $dN$ and $dS$ Values for each family (order) AP2 group genes (A) and different clades (B).
lineages were almost all much less than 1 except for the end of the AP2-R2 domain in AP2L, which further suggested the functional importance of the common homeodomains in AP2 group proteins. The $\omega$-values are not shown for the miR172 target site because $dN$ was zero. The peaks above the line that marks $dN = dS$ in Figure 4 suggest the existence of positive selection, primarily on either side of the two AP2 domains, especially downstream of domain AP2-R2. There were more $dN > dS$ peaks in the graph of the AP2L clade because fewer AP2 ortholog sequences are available for gymnosperms (only eight sequences) for analysis of evolutionary pressure, but the distribution trend of positive-selection peaks agreed with that of clades AP2 and TOE1. Compared with clade AP2, TOE1 contained more positive-selection peaks, implying that the purifying selective pressure on the TOE1 clade was relatively weak and enabled more nonsynonymous substitutions to be retained. Potentially, the $\omega$-value differences of different regions are related to functional divergence in AP2 group genes.

To examine if $\omega$ varied among branches of each clade (Figure 1), the free-ratio and one-ratio models in the Codeml program of PAML 4.2 were chosen and used to detect selective pressure acting on some branches (Data sheet 2). The values of $\omega$ for these AP2 group genes were similar (0.2083–0.2897) and substantially less than 1. However, the free-ratio model fit the data better than the one-ratio model for the protein-coding sequences from clades AP2, AP2L, and TOE1, suggesting that the genes from these three clades possibly experienced different selective pressures. Conversely, when coding sites in genes of clades TOE3, TOE2, and SMZ/SNZ were analyzed, the codon-substitution free-ratio model, which allows for different $\omega$-values among the branches, did not fit the data any better than the one-ratio model, which assumes a single mean $\omega$-value for the branches. The primary reason for this result was that the genes in clades TOE3, TOE2, and SMZ/SNZ were all from Brassicaceae. All AP2 group genes were analyzed by the free-ratio and one-ratio models, and the results also suggested that $\omega$ varied among branches. Therefore, genes of different clades experienced different selective pressures.

Six codon substitution models, namely M0 (one-ratio), M1a (nearly neutral), M2a (positive selection), M3 (discrete), M7 (beta), and M8 (beta and $\omega$), were implemented in PAML 4.2 to analyze the positive selection and identify positively selected sites in all AP2 group genes. The likelihood values and parameter estimates of all AP2 group gene sequences from the six models applied in the Codeml program are listed in Table 1. The average $\omega$-values in the six models ranged from 0.2794 to 0.5952, providing evidence for purifying selection. Although the average $\omega$ was 0.2794 for all sites of the AP2 group genes by the M0 model, this model was rejected as a result of the low-likelihood value ($-87829.48516$) and the LRT statistic (2 delta lambda statistic, $2\Delta\lambda$, Table 1). No positively selected sites were identified by the M3 model because $\omega < 1$, but two models (M2a and M8) allowed for positive selection, indicated by 38 and 17 positively selected sites with $\omega > 1$, respectively. Because of the overestimate of the number of actual positively selected sites (Anisimova et al., 2001, 2002), the results under model M3 were not used to identify positively selected sites. To reduce or avoid possible false-positive results, positively selected sites identified simultaneously by models M2a and M8 in Codeml were defined as positively selected. The LRT statistic demonstrated that the
two selection models fitted the data significantly better than the null models without positive selection, supporting the view that certain amino acids in AP2 group proteins experienced strong positive selection. At the level of posterior probability >0.95, 22, and 9 sites in the AP2 group genes were identified as being under positive selection (ω > 1) by the selection models M2a and M8, respectively (Table 1). There were 15 sites with posterior probability >0.99 among the 22 positively selected sites in the M2a model and 7 among the 9 positively selected sites in the M8 model. All 9 positively selected sites detected by M8 were also identified by M2a at the level of posterior probability >0.99. The positively selected sites were mainly concentrated in two regions—upstream of the NLS and downstream of the second EAR domain. Both regions showed the corresponding positive-selection peaks in Figure 4. All of this evidence supports the existence of positive selection and positively selected sites in AP2 group genes during spermatophyte evolution.

The likelihood values and parameters of the three main clade branches (AP2, AP2L, and TOE1) of AP2 group proteins were estimated by the six models to detect whether there was positive selection in clades AP2, AP2L, and TOE1 (Table 2). Positive selection was only detected in clade AP2 by the two positive selection models (M2a and M8), and no such positively selected sites were discovered in AP2L and TOE1. The positively selected sites in clade AP2 were not identified by the M3 (discrete) model because ω was <1, which was similar to the results for all AP2 genes. Of the sites with posterior probability >0.95 in AP2 clade proteins, six and four sites were identified to be under positive selection (ω > 1) by selection models M2a and M8, respectively, and the number of positively selective sites was four and two in M2a and M8, respectively, at posterior probability >0.99. The positively selected sites of clade AP2 were mainly concentrated upstream of the miR172 target site, and there were also corresponding positive-selection peaks in Figure 4. The parameter estimates from the six models were quite similar between clades AP2L and TOE1. The parameters of the M2a model for AP2L and TOE1 revealed a lack of positive selection in the two clades because ω = 1. The LRT statistic demonstrated that the M8 model did not fit the data significantly better than the M7 model without positive selection, suggesting that no amino acid sites in AP2L and TOE1 underwent positive selection. Therefore, the analysis of selective pressure of the three main clade branches of spermatophyte AP2 group genes indicated that clades AP2L and TOE1 experienced similar adaptive evolutionary mechanisms and only the AP2 clade underwent positive selection. Because TOE2, TOE3, SMZ, and SNZ are only found in Brassicaceae, the analysis of positive selection of Brassicaceae AP2 group genes and five branches (AP2, TOE1, TOE2, TOE3, and SMZ/SNZ) was performed (Data sheet 3), with the results revealing a lack of positive selection in the AP2 group genes of Brassicaceae, which indicates that the expansion of the number of AP2 group genes in Brassicaceae was not caused by positive selection.

Certain Amino Acid Residues in Common Homeodomains Reflect the Evolutionary Relationship from AP2L to AP2 and TOE1

Both the distribution of motifs and the analysis of selective pressure suggested that AP2L may have diverged to yield the two structurally and functionally distinct genes AP2 and TOE1. Specific motifs (Motifs 5, 19, and 23) of clades AP2 and TOE1 were also found in AP2L proteins, and AP2L and TOE1 experienced similar adaptive evolutionary processes. By comparing all AP2, AP2L, and TOE1 proteins, 10 amino acid sites that may reflect the evolutionary relationship in common homeodomains were identified (Figure 5), one in the AP2-R1 domain, three in the linkage domain, five in the AP2-R2 domain, and one in the second EAR domain. These sites could be divided into three categories: AP2L having the same amino acids as TOE1 (three sites), AP2L having the same amino acids as AP2 (one site) and AP2L having two amino acids from AP2 and two from TOE1 (six sites). Only one site belonging to the second category suggested that AP2 evolved faster than TOE1,
TABLE 2 | Likelihood Values and Parameter Estimates for AP2, AP2L, and TOE1 Clades Genes Sequences.

<table>
<thead>
<tr>
<th>Clade</th>
<th>Model code</th>
<th>InL</th>
<th>dN/dS</th>
<th>Estimates of parameters</th>
<th>2Δ</th>
<th>Positive selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP2</td>
<td>M0 (one-ratio)</td>
<td>−32964.4428</td>
<td>0.2083</td>
<td>ω = 0.2083</td>
<td></td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>M3 (discrete)</td>
<td>−30570.7018</td>
<td>0.2613</td>
<td>ρ0 = 0.5033, p1 = 0.3059 (p2 = 0.1908), a0 = 0.0129, a1 = 0.2785 a2 = 0.8891</td>
<td>4787.48207</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>M1a (NearlyNeutral)</td>
<td>−31071.0231</td>
<td>0.4177</td>
<td>ρ0 = 0.6109 (p1 = 0.3891)</td>
<td>70.79546</td>
<td>Not allowed</td>
</tr>
<tr>
<td></td>
<td>M2a (PositiveSelection)</td>
<td>−31036.8254</td>
<td>0.4636</td>
<td>ρ0 = 0.6053, p1 = 0.3574 (p2 = 0.0373), a0 = 0.0486, a1 = 1.0000 a2 = 2.0876</td>
<td>68.95064</td>
<td>Not allowed</td>
</tr>
<tr>
<td></td>
<td>M7 (beta)</td>
<td>−30488.3157</td>
<td>0.2556</td>
<td>p = 0.2052, q = 0.5973</td>
<td>24.89248</td>
<td>Not allowed</td>
</tr>
<tr>
<td></td>
<td>M8 (betaω)</td>
<td>−30475.8965</td>
<td>0.2654</td>
<td>ρ0 = 0.9713 (p1 = 0.0287)</td>
<td></td>
<td>2986.303D 306A 307G</td>
</tr>
</tbody>
</table>

| AP2L        | M0 (one-ratio) | −11930.6163 | 0.2897 | ω = 0.2897              |    | None |
|             | M3 (discrete)   | −11582.3172 | 0.3653 | ρ0 = 0.2811, p1 = 0.5226 (p2 = 0.1963), a0 = 0.0063, a1 = 0.3192 a2 = 1.1041 | 696.59821 | None |
|             | M1a (NearlyNeutral) | −11675.8826 | 0.4763 | ρ0 = 0.6021 (p1 = 0.3979) | NA | None |
|             | M2a (PositiveSelection) | −11675.8826 | 0.4763 | ρ0 = 0.6021, p1 = 0.3028 (p2 = 0.0951), a0 = 0.1302, a1 = 1.0000 a2 = 1.0000 | NA | None |
|             | M7 (beta)       | −11599.3964 | 0.3641 | p = 0.4323, q = 0.7545 | 924.9825 | Not allowed |
|             | M8 (betaω)      | −11597.7023 | 0.3996 | ρ0 = 0.9625 (p1 = 0.0375) p = 0.4627, q = 0.9010, ω = 1.8976 | 3.388152 5 | None |

| TOE1        | M0 (one-ratio) | −31052.6234 | 0.2710 | ω = 0.2710              |    | None |
|             | M3 (discrete)   | −29048.0536 | 0.3427 | ρ0 = 0.3819, p1 = 0.3179 (p2 = 0.3002), a0 = 0.0129, a1 = 0.2758 a2 = 0.8330 | 4009.19619 | None |
|             | M1a (NearlyNeutral) | −29473.6938 | 0.5461 | ρ0 = 0.4793 (p1 = 0.5207) | NA | None |
|             | M2a (PositiveSelection) | −29473.6938 | 0.5461 | ρ0 = 0.4793, p1 = 0.0583 (p2 = 0.4824), a0 = 0.0531, a1 = 1.0000 a2 = 1.0000 | NA | None |
|             | M7 (beta)       | −29004.4021 | 0.3346 | p = 0.2721, q = 0.5409 | 198.9409 | Not allowed |
|             | M8 (betaω)      | −29001.2895 | 0.3491 | ρ0 = 0.9613 (p1 = 0.0387) |    | None |

InL: the log-likelihood difference between the two models; 2Δ: twice the log-likelihood difference between the two models. The values in parentheses represent the significant level of the χ2 distribution at d.f. = 4 (M0 vs. M3) or 2 (M1a vs. M2a and M7 vs. M8). The amino acid sequences of AtAP2, PiaAP2La, and AtTOE1 were respectively used as the sequences reference. Positive selected sites in AP2 and AP2L clades were identified with posterior probability p > 0.95, in boldface, p > 0.99.

in agreement with the results of the adaptive evolution analysis. The sites having two amino acids from AP2 and TOE1 also provided evidence that AP2L diverged into two structurally and functionally distinct genes, AP2 and TOE1 OR genes, AP2 and TOE1, through structural and functional changes.

Alignment of AP2 Group Gene Homolog Sequences from A. thaliana Demonstrated the Expanded Mode of Spermatophyte AP2 Group

Like certain other Brassicaceae plants, Arabidopsis has six AP2 group genes, which is unique because the orthologs of TOE2, TOE3, SMZ, and SNZ were not found in the other spermatophytes analyzed in this study. Accordingly, the detailed genomic information available for Arabidopsis was very helpful for exploring the expansion of the AP2 group in spermatophytes. Alignment of AP2 group genes of Arabidopsis revealed conserved regions (Figure 6). Ten exons were identified in the ATAP2 genomic DNA sequence, nine in AtTOE1–3 and seven in AtSMZ and AtSNZ. Unsurprisingly, the exons corresponding to the first EAR, NLS, AP2-R1, and linkage domains exhibited stronger conservation in these six genes, and the conserved sequences of the AP2-R2 domain in ATAP2, AtAP2L, and AtTOE1 were also shown in Figure 6. Most notably, a region in intron 5 of AtTOE2 was very similar to exon 6 of AtAP2, AtTOE1, and AtTOE3, strongly suggesting that exon 6 was lost in the course of ATAP2 evolution. There was also some sequence conservation in introns 1, 3, and 4. Overall, AtTOE1 and AtTOE2 were the most closely related of these six genes, and the similarity between AtSMZ and AtSNZ was highest. The conservation among other genes was mainly in the introns, such as between AtTOE3 and AtSMZ.
and between AtTOE2 and AtSNZ. In the phylogenetic tree of all AP2 group genes in Brassicaceae (Figure 1 and Image 3), the relationship between clades SMZ and SNZ was paralogous and so were the relationships between the AP2 clade and the TOE3, TOE2, and SMZ/SNZ clades and the AP2 and TOE1 types. These results indicated that gene duplication was an important cause of the expansion of the Brassicaceae AP2 group. Structural changes and rearrangements after gene duplication could have resulted in functional divergence.

**Arabidopsis** Mutant Analysis Supports Evidence for Functional Divergence after Gene Expansion of the AP2 Group

Mutant studies indicate that AP2 is involved in the regulation of the stem-cell niche in the shoot meristem (Wurschum et al., 2005), floral development (Jofuku et al., 1994), and seed mass (Ohto et al., 2005), whereas AtSMZ, AtSNZ, AtTOE1, AtTOE2, and AtTOE3 redundantly affect flowering time (Jung et al., 2007, 2014; Yant et al., 2010). We isolated a novel ap2 allele from an ethyl methanesulfonate mutagenesis screen for abnormal expression patterns of the shoot meristem stem cell marker pCLV3:YFPer transgenic line. In contrast to wild-type plants, where pCLV3:YFPer is expressed in the three stem-cell layers of torpedo-stage embryos (Image 4A), pCLV3:YFPer signal was only observed in the epidermal layer of the mutant ap2 (2-132) mutant at levels comparable to wild-type, whereas expression in the subjacent layers was strongly reduced (Image 4B). At the seedling stage, ap2 (2-132) mutants failed to develop a wild-type like shoot meristem. ap2 (2-132) plants also were late flowering with abnormal flower phenotype and abnormal CLV3 expression in the stem-cell niche of shoots (Image 4A–D,H).

Comparing the genomic and coding sequences of AtAP2 in the wild type and 2-132, a single-base exchange in intron 6 of AtAP2 was identified (Image 4E), which affected pre-mRNA splicing and led to loss of exon 6 of AtAP2 in 2-132 (Image 4F), and consequently 15 amino acids were lost from the AP2-R2 domain (Image 4G). Remarkably, exon loss also occurred in AtTOE2, and this lost exon is homologous with exon 6 of AtAP2 (Figure 6 and Image 4G). Thus, fluctuation in the number of exons may facilitate divergence in gene function and also may be one of the ways new genes are formed.

**DISCUSSION**

**Typical AP2 Group Genes First Appeared in Gymnosperms and Evolved into AP2 and TOE Types through Whole-Genome and Gene Duplication in Angiosperms**

The AP2 domain was previously considered a plant specific core construct of the AP2 family, but it has more recently also been found in cyanobacterium, ciliates, and viruses (Magnani et al., 2004). These newly identified non-plant proteins with an AP2 domain are predicted HNH endonucleases, a kind of homing endonuclease (Chevalier and Stoddard, 2001). An HNH-AP2 homing endonuclease may have been transported into plants via endosymbiosis, horizontal transfer or other lateral gene transfer events. In the process of formation of the AP2 subfamily containing two AP2 domains, tandem duplication is likely to have played a major role (Magnani et al., 2004). A protein containing two AP2 domains has been identified in *Chlamydomonas reinhardtii*, but it does not clustered with the AP2 subfamily, though the amino acid composition of its AP2
domain is very similar to the AP2 group (Shigyo et al., 2006). It is regarded as a sister to the AP2 and ANT groups in terms of phylogenetic relationships (Shigyo et al., 2006). AP2/ERF proteins with two AP2 domains from *Physcomitrella* cluster with the ANT group (Kim et al., 2006). *C. reinhardtii* belongs to the Chlorophyta lineage, which is sister to the Streptophyta lineage (Charophyceae and land plants, Karol et al., 2001). This suggests that the AP2 and EREBP subfamilies diverged before the Chlorophyta lineage diverged from the Streptophyta lineage (Shigyo et al., 2006). In addition, we found no orthologs of AP2 group genes in our database searches of alga, moss and fern. AP2L from gymnosperms were identified in searching the orthologs of AP2 group genes, demonstrating the ancestral polyploidy event during the formation of gymnosperms (Jiao et al., 2011). AP2 group genes were also detected in basal angiosperms (*Amborella trichopoda*, *AmtAP2*; *Gnetum parvifolium*, *GpTOE1*; *Nymphaea hybrid cultivar*, *NhTOE1*) and respectively clustered into the AP2 and TOE1 clades. The presence of angiosperm genes in both the AP2- and TOE-type lineages suggests the duplication that gave rise to these two lineages followed the divergence of gymnosperms and angiosperms. The whole-genome duplication in ancestral angiosperms (Jiao et al., 2011) may have led to the AP2 group genes falling into two broad categories: the AP2 and TOE types.

**Gene Duplication and Motif Changes Produced New Genes in the Angiosperm AP2 Group**

The distribution diagram (Figure 2) of motifs and homeodomains exhibited the evolutionary relationships in the AP2 group. There were ANT orthologs but no AP2 orthologs in alga, moss and fern, which supports the argument that the AP2 group first appeared in gymnosperms. There was no differentiation between the AP2 and TOE types in gymnosperms, though all AP2Ls from Pinaceae were divided into two sub-branches by the whole-genome duplication in ancestral gymnosperms, which indicates that the duplication did not lead to the formation of TOE-type genes. The whole-genome duplication in the ancestral angiosperm was the likely basis of AP2 group gene differentiation. In this process, there were more new motifs in AP2-type than TOE-type genes (Figure 2). The homeodomains have evolved little in the AP2 group gene differentiation, especially the NLS and AP2-R1 domains. Most obviously, the first EAR domain transformed from DLNxXP to LxLxL in the TOE type and disappeared in the TOE3 clade. These changes caused AP2 group genes to diverge into the AP2 and TOE types. Our analysis demonstrated there are AP2 and TOE1 orthologs in most angiosperms except Brassicaceae, which contains six AP2 group genes, suggesting that other gene or whole-genome duplication events occurred in the course of evolution. In fact, the extensive complete genome analyses in *Arabidopsis* supports the model that two recent whole-genome duplication events occurred in Brassicaceae and one triplication event occurred in eudicots (Bowers et al., 2003; Tuskan et al., 2006; Lyons et al., 2008; Barker et al., 2009). Interestingly, only in Brassicaceae, new functional AP2 group orthologs appeared. In most angiosperms, polyploidy could simply cause increased gene copy numbers of AP2 and TOE1. Some changes of motifs and homeodomains have taken place in the new AP2 group genes. The AP2 group was different in AP2 domain and motifs than the ANT group (Kim et al., 2006). Like the isolation of AP2 and ANT groups, the new AP2 group genes in Brassicaceae may have been formed by a similar mechanism. For instance, the deletion and amino acid changes mainly occurred in the TOE2 and SMZ/SNZ AP2-R2 domains but TOE3 was formed by the removal of motifs. Compared with TOE2s, SMZs/SNZs evolved two new specific motifs (Figure 2). The analysis of mutant and genomic sequence alignment supports the view that new genes and functional differentiation could be produced by exon changes and genomic sequence rearrangements.

**Different Selective Pressures Drove the Evolution of Different Clades in Spermatophyte AP2 Group**

The selective pressures analysis of all and each clade of the AP2 group genes in spermatophyte suggested that AP2 group genes experienced different evolutionary patterns and each clade encountered various selective pressures, demonstrating that complex selective pressures drove the evolution of the AP2 group. As DNA-binding proteins containing two AP2 domains, AP2
group genes needed to maintain high conservation in the AP2 domains and NLS i.e., to facilitate nuclear translocation and DNA binding. All ω-values were <1 in the one-by-one comparisons of AP2 group genes, showing a background of purifying selection during evolution and coinciding with the conservation of sequence. However, likelihood values and parameter estimates in PAML demonstrated the presence of positive selection in the evolution of all AP2 group genes, although the positively selected sites were few (<3% of sites) and the distribution was relatively concentrated (upstream of the NLS and downstream of the second EAR domain). The analysis further showed that the positive selection occurred exclusively in the AP2 clade. Related to the fact that AP2 was the main functional gene in the AP2 group, this positive selection might have lead to the protein functional changes. There was no positive selection in other clades of the AP2 group other than the AP2 clade. Accordingly, there may have been positive selection at the time of divergence of the AP2 and TOE types. In the subsequent evolution, every clade experienced differential selective pressures, particularly the AP2 and TOE1 clades. Notably, more amino acid changes accumulated in the TOE1 clade during evolution, which was supported by the pairwise and sliding-window analysis of dN and dS in all AP2 group genes. As suggested above, different clades of the AP2 group experienced different evolutionary patterns, which might be associated with gene function. AP2 clade genes, as the main functional gene of the AP2 group, required high conservation and probably changed accordingly with angiosperm diversification.

**AP2 Clade Genes Retained Ancestor Gene Function of AP2 Group**

Phylogenetic analysis demonstrated that all orthologs of AP2 group genes (AP2L) in gymnosperms belong to the AP2 type and are most closely related to the AP2 clade, which suggests that gymnosperms only contain AP2 homologous genes, and that the ancestor of seed plants had AP2-like genes but no TOE1-like genes. In spermatophyte evolution, the AP2 group genes diversified but some functions might be conserved in the most recent common ancestor of extant spermatophytes. In the model plant *Arabidopsis*, the functions of the six AP2 group members (AtAP2, AtTOE1–3, AtSMZ, and AtSNZ) have been fully studied, and all are repressors of flowering, but only AtAP2 exhibits multiple functions in the development of flowers, fruit, seeds and stem cells (Wurschum et al., 2005; Mathieu et al., 2009; Huijser and Schmid, 2011; Ripoll et al., 2011). The expression patterns of AP2Ls in gymnosperms (Larix × martschinsii, P. abies, P. thunbergii) have been examined in previous studies. AP2Ls are expressed in female and male cones, leaves, stems and roots (Vahala et al., 2001; Shigyo and Ito, 2004). AP2L from P. abies shows functional similarities to AtAP2 in floral patterning when overexpressed in *Arabidopsis* (Nilsson et al., 2007) and the overexpression of AtAP2 also affects floral patterning in *Nicotiana benthamiana* (Mlotshwa et al., 2006). AP2Ls are also expressed during somatic embryogenesis in gymnosperms (Guillaumot et al., 2008). Remarkably, there are multiple homologous genes of the AP2 group in gymnosperms and their expression patterns differ from each other, though they belong to the AP2 type together with the AP2 clade in the phylogenetic tree, which implies that there has been a certain degree of functional differentiation in gymnosperm AP2Ls. Having gone through whole-genome duplication in the ancestral angiosperm, AP2 clade proteins in angiosperms likely had similar functions to gymnosperm AP2Ls inheriting from the common ancestor. This model is widely supported by the fact that AP2 clade proteins from *Oryza sativa* (Lee et al., 2007), *Zea mays* (Chuck et al., 2007), *Hordeum vulgare* (Nair et al., 2010), *Solanum lycopersicon* (Karlova et al., 2011), *Solanum tuberosum* (Martin et al., 2009), *Actinidia delicosa* (Varkonyi-Gasic et al., 2012), *Petunia hybrida* (Maes et al., 1999, 2001), and *Crocus sativus* (Tsafaris et al., 2012) have been linked to AP2Ls and their function in floral organ identity and development, fruit development, lodicules development, branching and tuberization. To meet the complex requirements of biological and functional diversity, AP2 clades genes have been constantly evolving by structural changes and adaptive evolution, which is in good agreement with our analysis of structural (motifs and homeodomains) alignment and selective pressure.

**Gene Numeric Expansion of AP2 Group**

With gene or whole-genome duplications in spermatophyte evolution, there was expansion of the number of AP2 group genes, especially in Brassicaceae. The phylogenetic tree of the Brassicaceae AP2 group reflects the evolutionary relationship: The TOE3 clade is a sister to the AP2 clade belonging to the AP2 type; the TOE2 and SMZ/SNZ clades are paralogous and cluster together with the TOE1 clade in the TOE type. The analysis results have gained support from the functional research of AP2 group proteins in *Arabidopsis*. AP2 group proteins from *Arabidopsis* are control factors of flowering. AP2-type proteins (AtAP2 and AtTOE3) participate in floral organ identity and development (Aukerman and Sakai, 2003; Chen, 2004; Jung et al., 2014), but TOE-type proteins (AtTOE1 and AtTOE2) are involved in flowering control and developmental phases of plant (Huijser and Schmid, 2011). The single mutant and overexpression of ATTO3 showed no visible phenotypic effects, while the expression pattern was different from other AP2 group genes in *Arabidopsis* (Jung et al., 2007; Yant et al., 2010). Recent research has shown that overexpression of an miR172-resistant ATTO3 can control floral organ identity and flowering time when the miR172 target site is mutated (Jung et al., 2014). ATTO3 binds to the second intron of AGAMOUS (AtAG) and represses its expression like AtAP2 (Yant et al., 2010; Jung et al., 2014), which means that the function of ATTO3 is similar to AtAP2, but is strongly constrained by miR172. Overexpression of AtSMZ, AtSNZ, AtTOE1, or AtTOE2 causes late flowering (Aukerman and Sakai, 2003; Chen, 2004; Jung et al., 2007) and quadruple (smz snz toe1 toe2) and sextuple (ap2 toe3 smz snz toe1 toe2) mutants flower earlier than any single or double mutant (Jung et al., 2007; Mathieu et al., 2009; Yant et al., 2010), which also indicates that the function of TOE-type genes is similar in
Arabidopsis. The functional studies of the AP2 group proteins in Arabidopsis supports the model that orthologs formed by gene or whole-genome duplications could be transformed into new genes by changes in motifs and homeodomains.

AP2 Function to Maintain the Stem Cell Niche was Conservative Function in Spermatophyte

In Arabidopsis, the marker genes WUS and CLV3 of shoot meristem stem-cell niche are regulated by AP2 and TOE3, and TOE1 does not act redundantly with AP2 in stem-cell maintenance (Wurschum et al., 2005). Likewise, the expression of WUS and CLV3 is changed in Arabidopsis by overexpressing AP2L of P. abies (Nilsson et al., 2007). Overexpression of AtAP2 in N. benthamiana also causes expression changes of NbWUS (Mlotshwa et al., 2006). Accordingly, the function of maintaining the stem-cell niche was conserved in the spermatophyte AP2 and AP2L clades. AP2Ls from Larix × marschilinsii are expressed during somatic embryogenesis and germination (Guillaumot et al., 2008), which also provides support for this view. Mutant analysis showed that changes in AP2-R1 and R2 lead to loss of stem-cell maintenance. AtTOE3 and AtTOE1 do not function in stem-cell maintenance, which indicates that the changes in motifs and homeodomains could also cause the loss of function in stem-cell maintenance.

The spermatophyte AP2 group contains many orthologs because of gene or whole-genome duplications. These orthologs may have evolved different functions, which means that the AP2 group may be a very valuable research area for examining new gene formation. The problems in AP2 group research focus on two aspects: evolution and functional differentiation. In gymnosperm, there was no differentiation into AP2 type and TOE type. In angiosperm, AP2 group genes were divided into two types and finished the functional differentiation. But more gene and genome data was needed to support the conclusion, especially gymnosperm. There are only two members (AP2s and TOE1s) in angiosperm AP2 group expect Brassicaceae. Current data suggested TOE2s, TOE3s, SMZs, and SNZs only exist in Brassicaceae, which means gene expansion and functional differentiation occurred during the formation of Brassicaceae but the detailed process was still unknown. AP2s are the the main function genes in AP2 group. After the formation of TOE1s, the function of flower time was only preserved. In Brassicaceae, SMZs and SNZs that belong to the same type with TOE1s own some new function. But so far, that still needs to be further studied. The correspondence among expression patterns, function and phylogenetic relatedness requires further study. The functionally similar genes in this study suggest functional diversification within the AP2 group. Comprehensive studies of expression and function and intensive phylogenetic characterization of the AP2 group genes will give us a clearer indication of the roles these genes play in the developmental processes of different species as well as the function of AP2 group genes in the evolutionary history of spermatophytes.

MATERIALS AND METHODS

Sequence Data
We retrieved the nucleotide and amino acid sequences for the Arabidopsis AP2 group from the Arabidopsis Information Resource database (www.arabidopsis.org). A BLASTP search was then performed using the AtAP2, AtTOE1–3, AtSMZ, and AtSNZ sequences as the query to retrieve AP2 group gene sequences from the NCBI (www.ncbi.nlm.nih.gov) and Phytozome databases (www.phytozome.org). The identified sequences were from the 56 species of spermatophytes (Data sheet 1). All selected AP2 group amino acid sequences contain six or seven conserved domains (Image 1). Prior to the phylogenetic analysis of these AP2 group gene protein sequences with Arabidopsis AP2 group, we ensured that all the AP2 group sequences clustered together as well as with the AtAP2 group. For our analysis, we selected the most typical AP2 group genes from a large number of paralogs but retained all query results of gymnosperms (Data sheet 1). The genes that fell outside the AP2 group were not analyzed.

Phylogenetic Analysis and Tree Construction
After deletion of identical sequences, only 105 sequences were used for phylogenetic analysis (Data sheet 1). They were aligned together using CLUSTAL 1.83. The phylogenetic tree of AP2 group genes was obtained by using ML (maximum likelihood) (MEGA 6.0) methods, and the reliability of the trees was evaluated by the bootstrap method with 1000 replications. The dN/dS value was used to detect positive selection.

Identification of Sequence Motifs
To identify motifs shared among related proteins within the AP2 group gene, we used the MEME motif search tool was used with its default settings. The maximum number of possible motifs was set to 35, and the maximum width was 300. Identified motifs were annotated using SMART (http://smart.embl-heidelberg.de/) and Pfam (http://pfam.sanger.ac.uk/).

Analysis of Adaptive Evolution and Identification of Selective Pressures
The program Codeml implemented in the PAML 4.0 software package was used to investigate the adaptive evolution of AP2 group protein-coding sequences. A total of 105 aligned AP2 group genes sequences, isolated from the different clades, were selected to test whether they were under purifying selection. Six models of codon substitution, M0 (one-ratio), M1a (Nearly Neutral), M2a (Positive Selection), M3 (discrete), M7 (beta), and M8 (beta & ω) were used in the analysis. M0 assumes that all sites have the same ω ratio. M1a assumes two classes of sites in proteins in proportions p0 and p1 (1–p0) with $0 < \omega_0 < 1$ (purifying selection) and $\omega_1 = 1$ (neutral sites). M2a adds a proportion (p2) to account for a class of sites where $\omega_2$ is estimated from the data and can be $> 1$. M3 uses a general discrete distribution with three site classes, with the proportions (p0, p1, and p2) and the ω ratios ($\omega_0$, $\omega_1$, and $\omega_2$) estimated from the data. M7 assumes a beta distribution (p, q) for 10 different ω
ratios in the interval (0, 1). M8 adds an extra class of sites with positive selection \((\omega > 1)\) to the beta (M7) model. Therefore, the null models M0, M1a and M7 fix the \(\omega\) ratios between 0 and 1, and do not allow the presence of positively selected sites. The alternative models M2a, M3, and M8 account for positive selection by using parameters, which estimate \(\omega\) greater than 1, and allow for the variable \(\omega\) along codon sequence.

The likelihood ratio test (LRT) was performed to detect the presence of positively selected sites by comparing the models that do not allow for positive selection with the models that allow for positive selection. The LRT was performed by taking twice the difference in log likelihood between nested models and testing for significance using the \(\chi^2\) distribution with the degrees of freedom equivalent to the difference in the number of parameters between models. If the LRT yields a statistically significant result, then positive selection is inferred. In the present study, three LRTs (M0 vs. M3, M1a vs. M2a, and M7 vs. M8) were used to detect positive selection. The Bayes empirical Bayes (BEB) approach implemented in M2a and M8 was used to determine the positively selected sites by calculating the posterior probabilities \((p)\) of \(\omega\) classes for each site. The sites with high posterior probabilities \((p > 0.95)\) coming from the class with \(\omega > 1\) were believed to be under positive selection.

**Plant Growth, Mutant Lines, and Mapping**

All plants were in the Landsberg erecta (Ler) ecotype, which was also used as the wild-type control. The 2-132 mutant was identified in the M2 generation of ethylmethanesulfonate-mutagenized pCLV3:YFPer plants. \(ap2-1\) and \(ap2-2\) mutants have been described.

The 2-132 mutation was mapped using plants with a wild-type-like phenotype in the F2 generation from a cross of an 2-132/+ plant with the Columbia ecotype. The initial mapping was done using CAPS markers from The Arabidopsis Information Resource (http://www.arabidopsis.org). The dCAPS markers used for fine mapping of the 2-132 mutation were generated based on the information available from CEREON.

**AUTHOR CONTRIBUTIONS**

PW, JS, and JC planned and designed the research. PW, TC, and Meiping Li performed experiments. Mengzhu Lu, TL conducted experiments. PW, GL, YL, and JC collected and analyzed data. PW wrote the manuscript.

**FUNDING**

This work was supported by National High Technology Research and Development Program of China (863 Program, 2013AA102705), Specialized National Basic Research Program of China (973 Program, 2012CB114504), Natural Science Foundation of Jiangsu University grant 13KJA220001, Talent project by the Ministry of Science and Technology, Co-Innovation Center for Sustainable Forestry in Southern China, Nanjing Forestry University, and Priority Academic Program Development of Jiangsu Higher Education Institutions.

**SUPPLEMENTARY MATERIAL**

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2016.01383

**Image 1** | **Conservative Domains Distributions of AP2 Group and the Consensus Amino Acid Sequence of AP2-R2 Domain in Brassicaceae.** Different colors represent different conservative domains (EAR domain, yellow; purple, NL5 domain; aqua, AP2-R1 domain; gold, linkage domain; red and green, AP2-R2 domain). The amino acid composition of AP2-R2 domain in different AP2 group genes is poor conservative, especially in TOE2s, SMZ2s, and SNZs.

**Image 2** | **Detailed Phylogenetic Trees of five clade in Figure 1.** The different color represent different clades (red, AP2 clades; green, TOE1 clade; olive, NLS domain; aqua, AP2-R1 domain; gold, linkage domain; red and green, AP2-R2 domain). The point mutations were generated in 100 self-crossed (simplified phylogenetic tree).

**Image 3** | **Phylogenetic Analysis of AP2 Group Protein in Brassicaceae (simplified phylogenetic tree).** The ML tree was constructed based on the whole protein sequences of spermatophyte AP2 Group gene using MEGA6.0 with 1000 bootstrap replications and Jones-Taylor-Thornton (JTT) + Gamma Distributed model (Discrete Gamma Categories = 5).

**Image 4** | **Phenotype, Site of Mutation and Mechanism of Mutation of 2-132. (A,B) Fluorescence microscope of pCLV3:YFPer in wild-type (A) and 2-132 (B) homozygous mutant torpedo stage embryos. The yellow fluorescent protein indicate the location of the embryonic stem cell niche. In 100 self-crossed progeny embryos of 2-132 heterozygote, the number of abnormal and normal yellow fluorescent in embryonic stem cell niche was 29: 71, which was no significant difference with 3: 1 by \(\chi^2\)-test. (C,D) The phenotype of flower in wild-type (C) and 2-132 (D) homozygous mutant. The sepals of 2-132 (D) homozygous mutant transform into leaves morphologically and the petals are like sepals. E Genomic organization of AP2. The mutant sites of 2-132, l28, ap2-1, ap2-2, and ap2-7 mutations are shown. The exon sequences of the two AP2 domains are marked (aqua, AP2-R1 domain; red, AP2-R2 domain). The point mutation in the genomic sequence of 2-132 is highlighted. (F,G) The sequencing results of AP2 (genomic DNA and miRNA) from wild-type and 2-132 homozygous mutant. The sequences of miRNA show there are 45 bases deletion in 2-132 homozygous mutant which happens to be the sixth exon of wild-type AP2. In the ATTOE2 of wild-type, this exon also does not exist. (H) The phenotype of flowering and height growth (centimeter) of wild-type and 2-132 homozygous mutant. The number of rosette leaves in 2-132 (B) homozygous mutant during flowering time is more than wild-type but the height growth is less.

**Data Sheet 1** | **The List of AP2 Group Genes in this Article.**

**Data Sheet 2** | **Branch Model Test for Each Clade Genes.**

**Data Sheet 3** | **Likelihood Values and Parameter Estimates for AP2s, TOE1s, TOE3s, TOE2s, and SMZSNZs in Brassicaceae.**
REFERENCES


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

Copyright © 2016 Wang, Cheng, Lu, Liu, Li, Shi, Lu, Laux and Chen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.
The Role of SHI/STY/SRS Genes in Organ Growth and Carpel Development Is Conserved in the Distant Eudicot Species Arabidopsis thaliana and Nicotiana benthamiana

Africa Gomariz-Fernández, Verónica Sánchez-Gerschon, Chloé Fourquin and Cristina Ferrándiz

Instituto de Biología Molecular y Celular de Plantas, Consejo Superior de Investigaciones Científicas–Universidad Politécnica de Valencia, Valencia, Spain

Carpels are a distinctive feature of angiosperms, the ovule-bearing female reproductive organs that endow them with multiple selective advantages likely linked to the evolutionary success of flowering plants. Gene regulatory networks directing the development of carpel specialized tissues and patterning have been proposed based on genetic and molecular studies carried out in Arabidopsis thaliana. However, studies on the conservation/diversification of the elements and the topology of this network are still scarce. In this work, we have studied the functional conservation of transcription factors belonging to the SHI/STY/SRS family in two distant species within the eudicots, Eschscholzia californica and Nicotiana benthamiana. We have found that the expression patterns of EcSRS-L and NbSRS-L genes during flower development are similar to each other and to those reported for Arabidopsis SHI/STY/SRS genes. We have also characterized the phenotypic effects of NbSRS-L gene inactivation and overexpression in Nicotiana. Our results support the widely conserved role of SHI/STY/SRS genes at the top of the regulatory network directing style and stigma development, specialized tissues specific to the angiosperm carpels, at least within core eudicots, providing new insights on the possible evolutionary origin of the carpels.

Keywords: carpel evolution, Eschscholzia californica, gynoecium, Nicotiana benthamiana, style and stigma, virus-induced gene silencing (VIGS), SHI STY SRS factors

INTRODUCTION

Organ development is directed by gene regulatory networks (GRNs) that control the temporal and spatial expression of downstream effectors responsible for creating morphogenetic outputs. GRNs are composed mainly of transcription factors and other regulators of gene expression such as microRNAs or chromatin-modifying elements, and signaling molecules such as mobile peptides or hormones, that interact extensively at different levels to generate morphogenetic patterns (Davidson and Levine, 2008). Over the last few decades, a wealth of genetic and molecular studies has uncovered major players that participate in organ formation and patterning in plants, and more recently, system biology approaches have complemented these studies to generate global and more
detailed pictures of the regulatory networks that underlie organ development. Most of these studies have been carried out in the model species Arabidopsis thaliana, where, for example, meristem formation, root development, or floral organ specification are increasingly understood from this global perspective (Alvarez-Buylla et al., 2007; De Lucas and Brady, 2013; O’Maoileidigh et al., 2014; Tian and Jiao, 2015; Davila-Velderrain et al., 2016). In addition, the rapidly increasing toolkit of genetic, molecular and functional resources across the plant kingdom is also expanding this knowledge to many other species and boosting evo-devo approaches to understand the evolution of plant form (Viallette-Guiraud et al., 2016).

Angiosperms are the largest and most diverse group of land plants. Carpels are the ovule-bearing structures of angiosperms, representing a major evolutionary innovation for this group that was probably key for their success. The carpels provide a confined casing to protect the ovules in their development. In the flower they may occur as single carpels, multiple unfused carpels or a syncarpic structure resulting from multiple fused carpels, where each individual structure is termed pistil. While morphological diversity of pistils is huge across angiosperms, they mostly share a basic organization plan. Apically, specialized cells form the stigma, which receives, discriminates and helps to germinate the pollen grains. The stigma is connected to the ovary through the style, generally a tube-like structure containing transmitting tissues that serve to grow and direct the pollen tubes toward the ovules. The ovary occupies a basal position, housing the ovules and, upon fertilization, it becomes a fruit, which may or may not incorporate additional parts of the flower, and serves to protect the developing seeds and later to facilitate seed dispersal (Ferrandiz et al., 2010). GRNs directing pistil patterning have been studied in Arabidopsis. Several genetic and hormonal factors required for the specification of carpel identity or the development of the specialized pistil tissues have been identified in the last few years, as well as some of their interactions and regulatory hierarchies. From these studies, GRNs directing the different functional modules in the Arabidopsis carpels have been proposed, and, although we are still far from completing an integrated network that provides a comprehensive view of spatial–temporal pistil morphogenesis, we increasingly understand how the basic blocks that compose a functional pistil are formed (Ferrandiz et al., 2010; Reyes-Olalde et al., 2013; Chávez Montes et al., 2015; Ballester and Ferrandiz, 2016; Marsch-Martinez and de Folter, 2016).

The evolutionary importance of carpels has also prompted many different labs to explore questions mainly focused on the evolutionary origin of carpels and the conservation of the genetic functions that specify carpel identity (Bowman et al., 1989; Bradley et al., 1993; Davies et al., 1999; Pan et al., 2010; Yellina et al., 2010; Dreni et al., 2011; Fourquin and Ferrandiz, 2012). However, only more recently these studies are being extended to other components of the emerging GRN proposed to direct pistil patterning. The increasing availability of plant genome sequences has allowed to reconstruct phylogenies for many of the gene families involved in carpel development with good taxonomic sampling, and thus to propose hypotheses on the possible evolution of the pistil GRNs (Pabon-Mora et al., 2014; Pfannebecker et al., 2017a,b). However, it is necessary to complement these works by carrying out functional studies in different taxa, which are still scarce. In this context, it is especially interesting to explore the functional conservation of the elements that drive style and stigma specification, since these tissues are only found across angiosperms and intimately linked to the evolutionary origin of carpels.

In Arabidopsis, it has been shown that correct auxin signaling is essential to establish apical–basal polarity in the Arabidopsis pistil, for correct development of the style and stigma and to ensure apical closure (Sundberg and Ostergaard, 2009; Larsson et al., 2013; Larsson et al., 2014; Moubayidin and Ostergaard, 2014). Two families of transcription factors are essential for style and stigma formation, and, at least in part, they exert these functions by regulating auxin synthesis, transport and response. The four NGA (NGA) genes belong to the RAV clade of the B3-domain transcription factor family and act redundantly to specify style and stigma identity. The Arabidopsis nga quadruple mutants completely fail to form these apical tissues and are female sterile, but develop fairly normal ovaries (Alvarez et al., 2009; Trigueros et al., 2009). The SHI/STY/SRS family of zinc-finger transcription factors, named after the members of this family in Arabidopsis SHORT INTERNODES (SHI), STYLISH (STY), and SHI RELATED SEQUENCE (SRS), play similar roles in pistil development, and multiple combinations of mutants in four or more members of the family cause almost identical phenotypes to those of quadruple nga mutants (Kuusk et al., 2002, 2006). NGA and SHI/STY/SRS factors share similar patterns of expression and common targets, and appear to work as master regulators of the GRN directing style and stigma development. In fact, simultaneous overexpression of NGA3 and STY1 in Arabidopsis is sufficient to direct ectopic style tissue formation to the whole surface of the ovary (Kuusk et al., 2002, 2006; Sohlig et al., 2006; Trigueros et al., 2009; Staldal et al., 2012; Martinez-Fernandez et al., 2014). The role of NGA orthologs in pistil development is strongly conserved in distant species such as the basal eudicot Eschscholzia californica or the asterid core eudicot Nicotiana benthamiana, where NGA inactivation leads to the absence of style and stigma differentiation (Fourquin and Ferrandiz, 2014). Several reports show that the connection of SHI/STY/SRS genes to auxin signaling is also conserved across land plants, as well as a general role in controlling plant architecture and possibly other hormone pathways (Eklund et al., 2010a,b; Zawaski et al., 2011; Islam et al., 2013; Youssef et al., 2017). However, the role of SHI/STY/SRS genes in pistil development has not been explored in detail outside Brassicaceae, although it has been described that barley mutants in the Lks2 gene, a member of the SHI/STY/SRS family, have short awns and defective stigmas (Yuo et al., 2012).

In this work, we have studied whether members of the SHI/STY/SRS family have conserved roles in style and stigma development in E. californica and N. benthamiana. This study supports that, as shown for NGA genes, SHI/STY/SRS factors are essential for the development of the apical tissues of pistils at least
in the core eudicots and that their downstream effectors leading to apical-basal patterning in the carpels are also likely conserved.

RESULTS

Identification of SHI/STY/SRS Genes in *E. californica* and *N. benthamiana*

The SHI/STY/SRS gene family in Arabidopsis comprises 10 members. Published phylogenies show that they belong to a plant-specific family, where homologs can be found already in the moss *Physcomitrella patens* or the lycophyte *Selaginella moellendorffii* (Eklund et al., 2010b; Pfannebecker et al., 2017a). SHI/STY/SRS factors share two highly conserved domains: a C3HC3H RING zinc finger domain and a IGGH motif that appears to be specific to this family (Fridborg et al., 2001), but outside these conserved regions the protein sequences are highly divergent (Supplementary Table S1).

To search for SHI/STY/SRS homologs in *E. californica*, we designed degenerate primers based on the sequence the RING-like and IGGH conserved domains of SHI/STY/SRS genes from other species. Only one putative SHI/STY/SRS gene, named EcSRS-L, was amplified from cDNA of *E. californica* flowers. The complete coding sequence of EcSRS-L was subsequently amplified by TAIL PCR and by the use of an adapted oligo-dT primer. The predicted EcSRS-L protein sequence possessed the typical RING domain and IGGH motif, but was different from the EscaSTY-L protein sequence recently published (Pfannebecker et al., 2017a), that we were not able to amplify with this strategy, probably because it contains a variant of the IGGH domain (Fridborg et al., 2001), but outside these conserved regions the protein sequences are highly divergent (Supplementary Table S1).

To search for SHI/STY/SRS homologs in *E. californica*, we designed degenerate primers based on the sequence the RING-like and IGGH conserved domains of SHI/STY/SRS genes from other species. Only one putative SHI/STY/SRS gene, named EcSRS-L, was amplified from cDNA of *E. californica* flowers. The complete coding sequence of EcSRS-L was subsequently amplified by TAIL PCR and by the use of an adapted oligo-dT primer. The predicted EcSRS-L protein sequence possessed the typical RING domain and IGGH motif, but was different from the EscaSTY-L protein sequence recently published (Pfannebecker et al., 2017a), that we were not able to amplify with this strategy, probably because it contains a variant of the IGGH domain that does not align with our degenerate primers (Supplementary Figure S1). Eleven SRS-related sequences were identified by searching the most recent draft of *N. benthamiana* genome (Figure 1) (Bombarely et al., 2012), all of them coding for predicted proteins that contained the RING and IGGH domains.

Using the identified *E. californica* and *N. benthamiana* predicted protein homologs and the Arabidopsis protein sequences of the SHI/STY/SRS family members we performed comparative sequence analyses, using the Neighbor joining algorithm built into the Mega7 software (Figure 1). The resulting tree had an overall topology similar to that of other previously published, although some differences were found mainly affecting less-supported clades and probably due to the use of different tree reconstruction methods and datasets. The clade comprising AtSTY1, AtSHI, and AtSRS8 was related to one formed by five Nicotiana predicted proteins. Four additional Nicotiana homologs were identified in a well-supported clade related to AtSRS5 and AtSRS7, while both EcSRS-L and EscaSTY-L grouped in the same clade as AtSRS3. Finally, two Nicotiana predicted proteins clustered as an outgroup with AtLRP. Low support of some branches did not allow to unequivocally establish direct relationships of some of the *N. benthamiana* and *E. californica* factors with the Arabidopsis SHI/STY/SRS proteins, but the tree strongly suggested that the duplication events that resulted in the high number of homologs found in Arabidopsis and Nicotiana were independent, and also that in Nicotiana, a recent allotetraploid, two copies of each gene were usually present.

Functional studies in Arabidopsis have shown that members of the family have redundant functions and that the degree of this redundancy does not depend strongly on how similar are their sequences. Thus, *sty1-1* is the only single mutant that shows an abnormal phenotype in gynoecium development, but combinations with mutant alleles in either *LRP,* *SRS5,* *STY2,* or *SHI* genes, which belong to different subclades in the family and have, respectively, a similarity score of 30, 34, 42, and 54% with *STY1,* similarly enhance the *sty1-1* defects in carpel development (Kuusk et al., 2002, 2006). For this reason, and since none of the *NbSTY/SHI/SRS* genes showed a clear orthology relationship with *AtSTY1,* we decided to focus for this study on three *N. benthamiana* genes: Niben101Scf06228g01001.1, a member of a sister clade to the one formed by the Arabidopsis *AtSRS5* and *AtSRS7* genes, and a pair of closely related genes that grouped together with the *AtSTY1/SHI/SRS* clade (Niben101Scf00158g00608.1 and Niben101Scf08273g05002.1, Figure 1). For simplicity, we renamed these as *NbSRS-L1,* *NbSRS-L2,* and *NbSRS-L3* (Figure 1). In addition, we also chose for functional characterization the *EcSRS-L* gene that we were able to amplify from *E. californica* flowers. An alignment of all the sequences of the predicted proteins used in this study, also including the Arabidopsis *AtSTY1,* *AtSRS3,* *AtSRS5,* and *EscaSTY-I* factors is shown in Supplementary Figure S1.

**SHI/STY/SRS Gene Expression Patterns Are Similar in Arabidopsis, *E. californica,* and *N. benthamiana***

The expression pattern of the SRS-L genes identified in *E. californica* and *N. benthamiana* was characterized by RNA *in situ* hybridization on young flower buds.

*EcSRS-L* transcripts were weakly detected in young flowers from early stages of development. In very young *E. californica* buds, *EcSRS-L* was detected in the emerging stamen and carpel primordia (Figure 2A). At later stages, *EcSRS-L* mRNA accumulated weakly in anthers and in the placenta and ovules (Figures 2B,C). *EcSRS-L* expression could also be detected in the most distal cells of the developing gynoecium, in the presumptive domain that would further develop into the style and the stigma (Figures 2B–D). This expression pattern generally resembled those described for Arabidopsis *SHI/STY/SRS* genes, which showed distal accumulation in the developing gynoecium and, at least in the case of *AtSTY1,* expression in ovules (Kuusk et al., 2002). In *N. benthamiana,* the three *SHI/STY/SRS* genes included in this study showed similar expression patterns, although not identical. *NbSRS-L1* was detected in very young floral buds, before floral organs in inner whorls were initiated (Figure 2E). At later stages, *NbSRS-L1* expression concentrated in the distal end of growing petals and of the gynoecium, as well as in the placenta and the anthers (Figure 2F). After style fusion, *NbSRS-L1* was mainly detected in ovules and pollen grains, while only weakly present in the transmitting tissues at the central domain of the style and in the stigma (Figures 2G,H). *NbSRS-L2* showed a more expanded expression in floral organs. In young buds, *NbSRS-L2*
expression was associated to all incipient floral organ primordia (Figure 2I). At later stages, expression was detected in growing petals, stamens and gynoecia, mainly at the distal end (Figure 2J). After style fusion, NbSRS-L2 was expressed mainly in ovules and it could be also found in the central domain of the style, the ovary wall and the distal end of growing petals (Figures 2K,L). NbSRS-L3 expression pattern was most similar to that of NbSRS-L1 (Figures 2M–O), although at later stages expression in the inner style was stronger than for NbSRS-L1 (Figure 2P).

In summary, NbSRS-L1, NbSRS-L2, and NbSRS-L3 genes showed remarkably similar expression patterns during flower development, being detected in all floral organ primordia from early stages of development and then mainly associated to the distal end of petals and carpels, anthers, placentae, and ovule primordia, as well as the inner style in more advanced stages. Again, these expression patterns resembled those described for several Arabidopsis SHI/STY/SRS genes (Kuusk et al., 2002, 2006).

**Overexpression of NbSHI/STY/SRS Genes in Arabidopsis thaliana Mimics the Effect of the Ectopic Expression of the Endogenous Arabidopsis STY Genes**

*Eschscholzia californica*, *N. benthamiana*, and Arabidopsis SHI/STY/SRS genes showed similar expression patterns, suggesting that they might have conserved functions. However, outside the RING and IGGH domains, low similarity could be found among *E. californica*, *N. benthamiana*, and Arabidopsis SHI/STY/SRS factors (Supplementary Figure S1 and Table S1). Nevertheless, extensive functional redundancy of Arabidopsis genes has been shown in spite of sequence divergence (Kuusk et al., 2006), suggesting that the RING and IGGH domains might be sufficient for their shared function. To explore whether the NbSRS-L genes under study had similar functional properties among them and also to their Arabidopsis homologs, we transformed Arabidopsis wild type and sty1 sty2 mutants with constructs for the overexpression of the three NbSRS-L genes and observed the corresponding phenotypes of at least 30 individual T1 plants for each construct and background. Overexpression of the three different NbSRS-L genes caused similar phenotypic alterations in carpel and fruit development when transformed into wild type Arabidopsis plants, resembling those observed for previously reported lines constitutively expressing *AtSTY1* or *AtSTY2* (Kuusk et al., 2002) (Figure 3A).

35S::NbSRS-L1 and 35S::NbSRS-L3 caused milder defects, mainly affecting the shape and length of the style, which was reduced and did not elongate as in wild type fruits, and the overall shape of the fruit, which appeared wider and blunt at the distal end in the lines with weaker phenotypic defects (Figures 3A,B, category 1), or shorter than wild type, wider specially at the apical end, and with an irregular ovary surface (Figures 3A,B,G,H, category 2) in the lines with stronger defects. NbSRS-L2 overexpression caused related but more dramatic phenotypic alterations in fruit development. About one third of the lines showed fruit phenotypes resembling those of the most
affected NbSRS-L1 or NbSRS-L3 overexpressors (Figures 3A,B), but most of the transgenic lines showed further enhanced defects that were subdivided in two additional categories (Figures 3A,B, categories 3 and 4). Fruits in category 3 had short and inflated ovaries (Figure 3A). The cell in the valves had similar shapes to wild type but were not oriented in parallel to the apical-basal axis of fruit growth, but formed an angle with the replum (Figures 3C,I). In addition, the valve margins at the apical end were not clearly defined, the style was short and expanded laterally, and the demarcation between the style and the valves was not discernible (Figures 3C,D,J). Finally, about 40% of the 35S::NbSRS-L2 lines had short fruits of tapered shape, with short ectopic style cells developing extensively at the lateral domains of the apical half of the valves (Figures 3E,F), and cells closer to the replum that retained valve identity but elongated forming a wider angle with the apical-basal axis (Figure 3E). When the same constructs were introduced into Arabidopsis sty1 sty2 mutants, all of them were able to partially or fully rescue the phenotypic defects of sty1 sty2 stigma and style (Figures 4A–F), although at different degrees.
FIGURE 3 | Phenotypes of Arabidopsis transgenic lines overexpressing the NbSRS-L genes. (A) Three to five fruits from a single plant representative of the different phenotypic categories and genotypes are shown. The AtSTYox fruits belong to a transgenic line obtained from E. Sundberg’s lab (Kuusk et al., 2002). (B) Proportions of the different phenotypic categories among the T1 transgenic lines obtained for each construct, where 1 (light blue) represents the weakest phenotype and 4 (dark blue) the strongest phenotype. White color corresponds to plants showing no conspicuous differences respect to wild type. (C–J) Scanning electron micrographs of fruits from selected genotypes and categories. (C,D) Fruits from 35S::NbSRS-L2 lines in category 3. (C) Apical part of the fruit. (D) Close up of the style/valve area. While cells retain their identity, there is no clear demarcation between the two zones. (E) Fruit from a 35S::NbSRS-L3 line in category 4. (F) Close up of the apical part of the fruit shown in (E). All the cells have a crenelated surface with the typical morphology of style cells and no valve cells can be distinguished. (G) Lateral view of a 35S::NbSRS-L3 fruit in category 2. The valves grow laterally to the same level as the stigma and style and have a square morphology. (H) Close up of the apical valve cells of the fruit shown in panel (G). All cells have valve identity and no ectopic formation of style cells can be observed. (I) Wild type fruit. (J) Close up of the apical domain, where the demarcation of stigma, style, and valves is clearly seen. Bars in (C–G): 50 µm.

Altogether, these results indicated that the three N. benthamiana SHI/STY/SRS factors under study had similar molecular properties among them and also, at least, to the Arabidopsis STY1 or STY2 genes in spite of high sequence divergence outside the RING and IGGH domains.

A similar approach was undertaken with the EsSRS-L gene identified in this study and 30 independent 35S::EcSRS-L T1 lines were generated in wild type and sty1 sty2 backgrounds. However, EcSRS-L overexpression was not able to produce neither any phenotypic effect in the wild type background, nor complementation of the sty1 sty2 mutant phenotype (Supplementary Figure S3).

Silencing of NbSHI/STY/SRS Genes in Nicotiana benthamiana Using VIGS Alters Floral Organ Growth and Style and Stigma Development

The results of overexpressing the three NbSRS-L genes in Arabidopsis wild type or sty1 sty2 plants suggested that the Nicotiana NbSRS-L factors were functionally equivalent to AtSTY1 or AtSTY2 to a large extent, but did not clarify whether these genes had similar roles in Nicotiana development to those described for the SHI/STY/SRS genes in Arabidopsis. To investigate the function of the three NbSRS-L genes in this study, we used Virus Induced Gene Silencing (VIGS) to reduce their transcript levels (Ratcliff et al., 2001; Wege et al., 2007; Fourquin and Ferrandiz, 2012). We generated three different TRV constructs designed to downregulate NbSRS-L1, NbSRS-L2, or NbSRS-L3. Twelve plants were inoculated with each construct independently and 5–20 flowers from each plant were chosen for further characterization (Table 1).

Nicotiana benthamiana has pentamerous flowers, where the perianth is composed of a calyx of five sepals, and five fused petals forming a tubular corolla. The five stamens are epipetalous and the gynoecium is bicarpellate and formed by a short bilocular ovary with central placentation and an elongated style, measuring more than 3 cm at anthesis, capped by a round flat stigma (Figures 5A–F). Plants inoculated with any of the three TRV2-NbSRS-L vectors showed similar phenotypic
alterations, although TRV2-NbSRS-L1 had a greater efficiency (64% of the flowers showed abnormal development) when compared with TRV2-NbSRS-L2 (15%) or TRV2-NbSRS-L3 (22%) (Table 1). These defects affected upper corolla development, and most frequently consisted of alterations in the shape, the color or the symmetry of the petal lobes (Figures 5G,L,P,M). Style and stigma development were also strongly affected. Style length was reduced, and very frequently the style was unfused or even split at the apical end, where the stigma adopted irregular shapes (Figures 5H–K,N,O,Q–S). Aiming to enhance the phenotypic defects caused by NbSRS-L inactivation, 12 plants were also inoculated simultaneously with TRV2-NbSRS-L1/TRV2-NbSRS-L3 vectors or with TRV2-NbSRS-L1/TRV2-NbSRS-L2/TRV2-NbSRS-L3 vectors. These combined treatments produced phenotypic effects in 70 and 64% of the observed flowers, respectively, therefore at levels comparable to those resulting of TRV2-NbSRS-L1 inoculation alone. However, in addition to the phenotypes already described for the individual constructs, these combinations induced novel alterations mainly related to floral organ number. In plants treated with TRV2-NbSRS-L1/TRV2-NbSRS-L3, the calix was frequently formed by six sepals (Figure 5T). Gynoecia formed by four fused carpels were also observed, sometimes even capped by multiple styles derived from each carpel (Figures 5U,W), while in other flowers two bicarpellated fused pistils were formed (Figure 5V). Simultaneous inoculation with the three vectors

<table>
<thead>
<tr>
<th>VIGS construct</th>
<th>N° plants inoculated (dead)</th>
<th>N° plants with phenotype</th>
<th>N° flowers/plant observed</th>
<th>N° flowers/plant with phenotype</th>
<th>N° total flowers with phenotype</th>
<th>% Flowers with phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>N. benthamiana</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRV2-NbSRS-L1</td>
<td>12 (5)</td>
<td>7</td>
<td>10–20</td>
<td>4–14</td>
<td>82/128</td>
<td>64%</td>
</tr>
<tr>
<td>TRV2-NbSRS-L2</td>
<td>12 (2)</td>
<td>8</td>
<td>7–20</td>
<td>1–6</td>
<td>22/143</td>
<td>15%</td>
</tr>
<tr>
<td>TRV2-NbSRS-L3</td>
<td>12 (2)</td>
<td>9</td>
<td>5–20</td>
<td>2–8</td>
<td>30/137</td>
<td>22%</td>
</tr>
<tr>
<td>TRV-NbSRS-L1/2</td>
<td>12 (3)</td>
<td>9</td>
<td>15–20</td>
<td>8–18</td>
<td>120/172</td>
<td>70%</td>
</tr>
<tr>
<td>TRV-NbSRS-L1/2/3</td>
<td>12 (3)</td>
<td>10</td>
<td>3–20</td>
<td>2–14</td>
<td>63/99</td>
<td>64%</td>
</tr>
<tr>
<td>Empty vector</td>
<td>6 (1)</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>0/100</td>
<td>0%</td>
</tr>
<tr>
<td>E. californica</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRV2-EcSRS-L</td>
<td>120 (11)</td>
<td>0</td>
<td>1–5</td>
<td>0</td>
<td>0/0</td>
<td>0%</td>
</tr>
<tr>
<td>TRV2-EcPDS</td>
<td>70 (22)</td>
<td>48</td>
<td>n.d.</td>
<td>n.d.</td>
<td>48/70*</td>
<td>68%*</td>
</tr>
<tr>
<td>Empty vector</td>
<td>60 (16)</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0/88</td>
<td>0%</td>
</tr>
</tbody>
</table>

* Bleaching caused by PDS inactivation.
did not further enhance the defects observed in TRV2-NbSRS-L1/TRV2-NbSRS-L3 treated plants (Figures 5X–Z), except for occasional development of abnormal anthers (Figure 5Z) and stronger defects in corolla development (Figures 5X–Z).

The similar effect of all the VIGS treatments tested suggested that the three NbSRS-L genes under study had related functions, an idea already supported by the results of their heterologous constitutive expression in Arabidopsis. However, since it was likely that these related functions were redundant, it was surprising that the inoculation with combinations of different constructs did not dramatically enhance the associated phenotypic defects, suggesting that the different TRV2-NbSRS-L vectors could be targeting more than one gene in the family and therefore inducing transitory inactivation of several NbSRS-L genes at once. The efficiency and specificity of each VIGS treatment was assessed by measuring the level of expression of NbSRS-L1, NbSRS-L2, and NbSRS-L3 on flowers from two treated plants for each treatment that showed inactivation-related phenotypes. The results from these experiments showed that the VIGS treatments were effective in reducing the expression levels of the NbSRS-L genes, but that they did not achieve a high degree of specificity and therefore the observed phenotypes were likely caused by simultaneous inactivation of several NbSRS-L genes (Supplementary Figure S4).

The effect of transient inactivation of EcSRS-L in E. californica was also tested by inoculation of plants with a TRV2-EcSRS-L construct. One hundred and twenty plants were inoculated but none of them showed evident phenotypic defects (Table 1), in spite of the effective inactivation caused by the VIGS treatment as quantified by qRT-PCR (Supplementary Figure S5), indicating that either the EcSRS-L gene was fully redundant with the EscaSTY gene described in (Plannebecker et al., 2017a), that the residual expression of EcSRS-L in VIGS treated plants was sufficient to provide its function, or that the phenotype associated with EcSRS-L inactivation was not affecting morphogenesis in a conspicuous way.

Overexpression of NbSHI/STY/SRS Genes in Nicotiana benthamiana Cause Ectopic Style and Stigma Development

To study the effect of ectopic expression of the NbSRS-L genes under study in N. benthamiana we generated transgenic plants in which we introduced the 35S::NbSRS-L1 and 35S::NbSRS-L2 constructs previously used for heterologous expression in Arabidopsis. Twenty-three independent transgenic lines were obtained for 35S::NbSRS-L1 and 9 for 35S::NbSRS-L2, and around half of them showed conspicuous phenotypic changes in gynoecium morphology (Figures 6A,E,J). Both 35S::NbSRS-L1 and 35S::NbSRS-L2 plants showed similar phenotypes, although, as already observed in Arabidopsis, the constitutive expression of NbSRS-L2 had more dramatic effects.

The pistils of plants overexpressing NbSRS-L1 had stigma and style of normal morphology, only shorter (Figure 6E). The demarcation between style and ovary was not defined as in wild type (Figure 6B), but showed a gradual transition between the elongated zone with columnar cells typical of the style and the nearly isodiametric small cells of the ovary (Figures 6E–G). The valve margins in the ovary, which in the wild type are poorly defined and adjacent (Figure 6C), in the 35S::NbSRS-L1 pistils appeared more pronounced and separated by several cell files, that in the upper part of the ovary bulged forming ridges capped by stigmatic cells (Figures 6E–H). Occasionally, small cylindrical protrusions, similar to styles and terminated by a rounded stigma, grew out of these ridges (Figure 6F). Finally, the gynophore was elongated and showed a clear demarcation from the ovary (Figures 6D,L). 35S::NbSRS-L2 pistils showed similar phenotypes, only enhanced. The junction between style and ovary showed stronger alterations and more frequently gave rise to what appeared to be ectopic styles (Figures 6J–L). The surface of the ovary was irregular and grew unevenly, with intermixed domains of different style/ovary cell identity (Figures 6J,L,M).

The gynophore was also more pronounced and separated from the ovary by a bulging ridge of tissue (Figure 6N). Anatomical sections confirmed that the style and stigma of the transgenic lines were similar to wild type (Figures 7A–F). The ovaries of the lines overexpressing NbSRS-L1 had thickened walls with between 2 and 4 more cell layers in the mesocarp than the wild type, and the ridges bulging at the valve margins clearly showed mixed identity of ovary and style/stigma tissues (Figures 7G,H,J,K). In the 35S::NbSRS-L2 lines, the upper part of the ovary showed thickened and irregular walls that did not resemble clearly the ovary wall morphology of the wild type (Figure 7L); the basal part of the ovary was less affected, but still showed irregularities in shape (Figure 7L).

In addition to changes in gynoecium morphology, overexpression of NbSRS-L genes caused strong alterations in leaf development, especially in 35S::NbSRS-L2 lines. The leaves of the transgenic plants were darker than wild type and growth was constricted in the margins, causing the blade to adopt a hood-like appearance (Figure 8).

DISCUSSION

In this work, we have studied the functional conservation of members of the SHI/STY/SRS gene family in the asterid core eudicot N. benthamiana and the basal eudicot E. californica. We have determined the expression patterns of the genes under study during flower development, characterized the floral phenotypes caused by their downregulation, and the effects of their overexpression both in the heterologous system A. thaliana and in N. benthamiana. Our work supports the conserved function of STY/SHI/SRS genes in apical gynoecium development and their position at the top of the GRN directing style and stigma development in core eudicots.

SHI/STY/SRS Function in Apical Gynoecium Development Is Conserved in Core Eudicots

The SHI/STY/SRS genes of Arabidopsis act redundantly to specify style and stigma development, as revealed by the phenotype of multiple combinations of mutations in members of the family (Kuusk et al., 2006). The expression of different
SHI/STY/SRS genes in Arabidopsis has been analyzed by mRNA in situ hybridization or the use of promoter::GUS reporters. These experiments show largely overlapping expression patterns, generally associated with domains of auxin accumulation in all lateral organs and in the apical tissues of the developing gynoecium, consistent with the high level of redundancy found in the family (Fridborg et al., 2001; Kuusk et al., 2002, 2006; Eklund et al., 2011). In this work, we have characterized the expression of a SHI/STY/SRS gene from the basal eudicot *E. californica* and of three members of the family from the core eudicot *N. benthamiana*, observing a significant conservation in the pattern of mRNA accumulation in the flowers of these two species and of Arabidopsis. In all cases, we have detected expression in the distal domain of growing floral organs in...
young floral buds, the ovule primordia and the apical domain of the pistil, indicating that the regulatory regions of these genes contain conserved elements, despite the evolutionary distance of the corresponding species and the differences among paralogs within the same genome. Interestingly, it has been shown that a regulatory element present in the promoters of most Arabidopsis SHI/STY/SRS genes, a GCC-box bound by transcription factors of the AP2/ERF family, is required for the expression of SHI/STY/SRS genes in the distal domain of lateral organs including the apical tissues of the developing pistils (Eklund et al., 2011). A recent work which includes a comprehensive phylogeny of the SHI/STY/SRS family also reports the search for GCC-boxes in the promoters of selected genes from this study, finding that it is present in almost all of the angiosperm sequences but not in more basal taxa such as those including mosses, lycophytes or conifers (Pfannebecker et al., 2017a). Not surprisingly, we could also detect conserved GCC-boxes within the 1 kb upstream promoter sequences of the three Nicotiana genes, for which genomic sequences were available, supporting their importance in mediating apical gynoecium expression.

Heterologous constitutive expression of *N. benthamiana* SHI/STY/SRS genes in Arabidopsis caused phenotypic effects similar to those of overexpressing the endogenous genes (Kuusk et al., 2002; Kim et al., 2010), including ectopic style tissue formation, loss of definition between the style and the ovary, and an overgrowth of the valves at the apical end of the pistil, a morphology typically associated to auxin accumulation and that mimics exogenous treatment with auxin (Ståldal et al., 2008). Despite the sequence divergence between the sequences of Arabidopsis and Nicotiana proteins, which is high outside the conserved RING and IGGH domains, and the occurrence of independent duplication events in each species, the overexpression phenotypes associated to the three *NbSRS-L* genes were similar, as well as the ability of the different *NbSRS-L* factors to complement the *sty1 sty2* mutant phenotype. These experiments support that many of the SHI/STY/SRS proteins are basically equivalent in molecular function, even if outside the conserved RING and IGGH domains they show low sequence similarity, suggesting that most of their interactions with DNA targets and other proteins would be mediated by these conserved
domains. In this context, it was even more surprising that the EcSRS-L gene did not complement the Arabidopsis sty1 sty2 mutant phenotype, neither caused phenotypic defects when overexpressed in Arabidopsis wild type plants. A close inspection of the RING and IGGH domains showed high sequence similarity with that of other members of the family, although the last Cys residue in the zinc finger domain is preceded by a Pro residue in the E. californica proteins while in most members of the family this position is most frequently a charged or polar residue like Asp, Glu, His, or Gln. Given the structural properties of Pro residues, which have an exceptional conformational rigidity that usually impacts protein secondary structure and protein–protein interactions (Morris et al., 1992), it is possible that this difference may affect function. This Pro residue is not exclusive of the E. californica homologs, but it was also found in 5 of the 91 predicted proteins included in the phylogenetic study by Pfannebecker et al. (2017a). However, since none of these homologs has been functionally characterized, it would be necessary to perform equivalent analyses with some of them to test this hypothesis.

Silencing of the NbSRS-L homologs by VIGS also supported their conserved role in style and stigma development at least in N. benthamiana. The pistils of VIGS-NbSRS-L plants displayed a range of phenotypic defects that affected strongly these tissues, mostly shortening of the style and split and abnormal stigma formation. In addition, other phenotypic defects associated

**FIGURE 7** | Phenotypic characterization of *N. benthamiana* pistils from 35S::NbSRS-L1 and 35S::NbSRS-L2 lines. Histological sections of anthesis pistils. (A,D,G,J) Wild type. (B,E,H,K) 35S::NbSRS-L1. (C,F,I,L) 35S::NbSRS-L2. (A–C) Transversal sections of stigmas. Note the similar morphology in all cases. (D–F) Transversal sections of the style. Again, style morphology appears similar in all three genotypes. (G–I) Transversal sections at the distal portion of the ovary. Note the medial protrusions capped with stigmatic tissue in 35S::NbSRS-L1 (H) and the engrossed and irregular ovary walls in 35S::NbSRS-L2 (I). (J–L) Transversal sections of the basal domain of the ovary. The morphology of 35S::NbSRS-L pistils is more similar to wild type, but still medial protrusions and additional cell layers in the ovary wall can be observed in 35S::NbSRS-L1 (K) and the ovary walls are irregular in shape in 35S::NbSRS-L2 (L). Bars: 200 µm.
previously with loss of SHI/STY/SRS function in Arabidopsis, like general defects in floral organ growth and anther development, were also observed (Kuusk et al., 2006). We found the VIGS treatments not to be specific among the NbSRS-L genes under study, so it is not possible to estimate the level of genetic redundancy or the specific functions of each individual gene. However, since the three of them show similar patterns of expression and cause similar effects when overexpressed in Arabidopsis or in Nicotiana, it seems most likely that they work redundantly as described for SHI/STY/SRS family members in Arabidopsis. Unfortunately, the EcSRS-L-VIGS treatments were not sufficient to induce phenotypic changes in *E. californica*, so we cannot infer any significant conclusion. The lack of abnormal phenotypes in VIGS-treated plants and in the Arabidopsis plants expressing EcSRS-L may suggest that the EcSRS-L protein is not functional and that other members of this or other families provide the style-stigma specification function. Only one additional SHI/STY/SRS gene has been described in *E. californica*, which presents high level of homology with the EcSRS-L gene in this study, including the distinctive Pro residue in the RING domain and thus unlikely to be functionally divergent (Pfannebecker et al., 2017a). However, the *E. californica* genome has not been sequenced yet, and it is possible that it encodes other members of the family that may have a role in style and stigma development. It would be interesting to extend this type of studies to other basal dicots and even more basal taxa within angiosperms to conclusively support the conservation of the role of SHI/STY/SRS factors in style and stigma differentiation across angiosperms. However, the reports of abnormal style and stigma development in the monocot *Hordeum vulgare* short awn mutants, affected in the SHI/STY/SRS gene *Lks2*, strongly support this idea (Yuo et al., 2012).

NbSRS-L VIGS-treated plants in Nicotiana occasionally showed increased number of floral organs or multicarpellate pistils. These phenotypes had not been previously reported for *shi/sty* mutants in Arabidopsis, and could reflect a specific function for SHI/STY/SRS genes in Nicotiana. However, it has been recently reported that the *vrs2* mutants in barley, affected in a SHI/STY/SRS gene, form supernumerary spikelet meristems in the inflorescence (Youssef et al., 2017), an indeterminate behavior somehow reminiscent of the supernumerary floral organs found in VIGS-NbSRS-L plants and that could reveal an additional conserved role of the members of the SHI/STY/SRS family, that might be related to their described conserved function in hormone homeostasis (Fridborg et al., 2001; Eklund et al., 2010b, 2011; Kim et al., 2010; Zawaski et al., 2011; Youssef et al., 2017).

**A Conserved GRN for Carpel Development**

The phenotypes of Nicotiana 35S::NbSRS-L lines parallel those described for Arabidopsis 35S::STY1 plants. In addition to support functional conservation among SHI/STY/SRS genes at the top of the GRN involved in style and stigma specification, these results also suggest that the set of targets for these factors in both species is also very conserved. Thus, the downstream pathways leading to style and stigma development are probably...
similar, despite the dramatic differences of style and stigma morphologies in both species. Moreover, in addition to directing ectopic development of these tissues, the overall changes in pistil morphology, such as the less conspicuous demarcation between ovary and style, or the irregular proliferation of cells in the ovary walls causing crinkled ovary surface, suggest that the process of patterning the pistil into different territories and functional domains responds to similar cues in species with highly diverse pistil morphologies.

Our study also reveals that, as in Arabidopsis, NGA and SHI/STY/SRS genes also share their functions in pistil development in Nicotiana. NbNGA and NbSRS-L genes show very similar expression patterns in flower development and lead to highly related phenotypes when downregulated by VIGS (Fourquin and Ferrandiz, 2014). However, while NbNGA inactivation caused a complete lack of stigma and style tissues in Nicotiana pistils, NbSRS-L downregulation only produced milder defects in these same tissues, which could suggest that the contribution of NbSRS-L factors to this function is less important. However, VIGS treatments were not as efficient for NbSRS-L inactivation as for the reported VIGS-NbNGA studies, and the number of SHI/STY/SRS genes in the N. benthamiana genome is also higher than the number of NbNGA homologs, and therefore it is likely that the observed phenotypes of the NbSRS-L-VIGS plants do not reflect the effect of a full loss of SHI/STY/SRS function (Fourquin and Ferrandiz, 2014). Interestingly, the phenotypes of N. benthamiana 35S:NbSRS-L lines also resemble some aspects of the effect of overexpressing the AtNGA genes in Arabidopsis, such as an enlarged and bulging replum (the region between the valve margins) or the elongation of the gynophore, also supporting the conserved position of both classes of factors at the top of the GRN directing style and stigma development where they would converge in common targets and functions (Trigueros et al., 2009).

NGA genes belong to the RAV subfamily of B3-domain transcription factors present in all land plants, but they possess three characteristic conserved domains only found in angiosperm NGA proteins (Fourquin and Ferrandiz, 2014; Pfannebecker et al., 2017b). SHI/STY/SRS homologs are found already in bryophytes, although the GCC-box conferring carpel expression has only been found in the promoters of angiosperm SHI/STY/SRS genes (Pfannebecker et al., 2017a). This suggests that the NGA and the SHI/STY/SRS genes may have acquired angiosperm-specific functions linked to the evolutionary origin of the style and the stigma, angiosperm-specific tissues themselves.

In addition to NGA and SHI/STY/SRS genes, other factors involved in style and stigma development have been shown to have conserved functions in several angiosperm species. Members of the PLE-subclade of MADS-box genes have been related to style and stigma development in different dicot species (Colombo et al., 2010; Fourquin and Ferrandiz, 2012; Heijmans et al., 2012). Likewise, orthologs of CRABS CLAW (CRC) gene, a member of the YABBY family of transcription factors required for correct style development and apical gynoecium closure in Arabidopsis, also have a conserved role in these functions in a wide range of angiosperm species (Yamaguchi et al., 2004; Fourquin et al., 2005, 2007, 2014; Lee et al., 2005; Ishikawa et al., 2009; Orashakova et al., 2009; Yamada et al., 2011). Taking all this evidence together, we could propose an emerging evolutionary conserved GRN directing carpel patterning which would include NGA, SHI/STY/SRS, CRC, and PLE-like genes as a core of factors linked to style and stigma development, although we still need to investigate in more depth the conservation of their regulatory relationships. In addition, it would be interesting to expand these studies to other genes with a described role in apical pistil development in Arabidopsis, such as HECATE, NO TRANSMITTING TRACT or HALF FILLED (Crawford et al., 2007; Gremski et al., 2007; Crawford and Yanofsky, 2011), which have not been functionally characterized in other taxa, and to connect this apical gynoecium network with those directing valve margin, dehiscence or fruit patterning, which also have been addressed lately (Ferrandiz and Fourquin, 2014; Pabon-Mora et al., 2014).

**MATERIALS AND METHODS**

**Plant Material and Growth Conditions**

*Eschscholzia californica* Cham. and *Nicotiana benthamiana* L. plants were grown in the greenhouse, at 22°C: 18°C (day : night), with a 16 : 8 h, light : dark photoperiod, in soil irrigated with Hoagland no. 1 solution supplemented with oligoelements (Hewitt, 1966). *E. californica* germplasm used in this study (accession no. PI 599252) was obtained from the National Genetic Resources Program (United States).

For Arabidopsis transformation, *EcSRS-L* and *NbSRS-L* coding sequences were amplified with primers EcSTYF/EcSTYR (EcSRS-L), NbSTY1For/NbSTY1Rev (NbSRS-L1), NbSTY2For/ NbSTY2Rev (NbSRS-L2), NbSTY3For/NbSTY3Rev (NbSRS-L3), cloned into PCR8/GW/TOPO (Invitrogen) and then transferred by Gateway reactions into the pMCD32 destination vector (Curtis and Grossniklaus, 2003). Each vector was introduced into *Agrobacterium tumefaciens* PMP90 for Arabidopsis transformation using the floral dip protocol (Clough and Bent, 1998) into the wild type Columbia background or the sty1-1 sty2-1 double mutants (Kuusk et al., 2002). T1 plants were selected based on kanamycin selection. *N. benthamiana* transformation was done using the same vectors and following previously described standard procedures (Clemente, 2006). Three to five T1 individuals from each transformation were tested to confirm that the transgene was being expressed by qRT-PCR. See Supplementary Table S2 for primer sequences.

**Cloning and Sequence Analysis**

The partial coding sequence of EcSRS-L gene was isolated by RT-PCR on cDNA of flowers of *E. californica* using the degenerate primers EcSTYdF2/EcSTYdR2 designed from the conserved motifs of SRS homologs from other species. The 3’ end of EcSRS-L was then isolated by reverse transcription polymerase chain reaction (RT-PCR) using the primers PoppyRT and RT (sequence added to the oligoT primer used for retrotranscription). To amplify the *NbSRS-L* genes, we designed primers from the annotated genomic sequences retrieved from solgenomics.net
(same pairs as described for cloning the CDS in the previous subheading). The deduced amino acid sequences alignments were analyzed using the MUSCLE tool built in the Macvector 15.1 software (MacVector Inc., Cary, NC, United States). The phylogenetic tree was inferred using the Neighbor-Joining method (Saitou and Nei, 1987). The optimal tree with the sum of branch length = 4.39003061 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (2000 replicates) are shown next to the branches. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the p-distance method and are in the units of the number of amino acid differences per site. The analysis involved 24 amino acid sequences. All ambiguous positions were removed for each sequence pair. There was a total of 723 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Kumar et al., 2016).

**In Situ Hybridization**

RNA *in situ* hybridization with digoxigenin-labeled probes was performed on 8 μm paraffin sections of *E. californica* and *N. benthamiana* buds, as described by Ferrándiz et al. (2000). The RNA antisense and sense probes were generated from a 294 bp fragment of the EcSRS-L cDNA (positions 1-294), from a 438 bp fragment of the NbSRS-L2 cDNA (positions 1-407) and a 438 bp fragment of the NbSRS-L3 cDNA (positions 607-1045), from a 438 bp fragment of the NbSRS-L2 cDNA (positions 1-407) and from a 536 bp fragment of the NbSRS-L3 cDNA (positions 1-536). Each fragment was cloned into the pGemT-Easy vector (Promega), and sense and antisense probes were synthesized using the corresponding SP6 or T7 polymerases.

**Virus-Induced Gene Silencing (VIGS)**

The same regions of EcSRS-L, NbSRS-L1, and NbSRS-L3 coding sequences used for *in situ* hybridization were used for the VIGS experiments, while for NbSRS-L2 a fragment of 537 bp (positions 1-537) was used. A XbaI restriction site was added to the 5′ end of the PCR fragment and a BamH1 restriction site was added to the 3′ end. The amplicon was digested by Xba1 and Bam H1 and cloned into a similarly digested pTRV2 vector. The resulting plasmids were confirmed by digestion and sequencing, before being introduced into the *Agrobacterium tumefaciens* strain GV3101. The agroinoculation of *E. californica* seedlings was performed as described (Pabon-Mora et al., 2012). The agroinoculation of *N. benthamiana* leaves was performed as previously described (Dinesh-Kumar et al., 2003).

**Quantitative RT-PCR**

Total RNA was extracted from flowers in anthesis with the RNeasy Plant Mini kit (Qiagen). Four micrograms of total RNA were used for cDNA synthesis performed with the First-Strand cDNA Synthesis kit (Invitrogen) and the qPCR master mix was prepared using the iQTM SYBR Green Supermix (Bio-Rad). The primers used to amplify EcSRS-L1 (qEcSTYFor and qEcSTYRev), NbSRS-L1 (qNbSTY1For and qNbSTY1Rev), NbSRS-L2 (qNbSTY2For and qNbSTY2Rev), and NbSRS-L3 (qNbSTY3For and qNbSTY3Rev) did not show any cross-amplification. Results were normalized to the expression of the ACTIN gene of *E. californica*, amplified by EcACTFor and EcCTRev, and to the Elongation Factor 1 (*EF1*) gene of *N. benthamiana* (accession no. AY206004), amplified by qNeEF1For and qNeEF1Rev (Fourquin and Ferrandiz, 2014). The efficiencies in the amplification of the genes of interest and the corresponding reference gene were similar. Three technical and two biological replicates were performed for each sample. The PCR reactions were run and analyzed using the ABI PRISM 7700 Sequence detection system (Applied Biosystems Inc., Life Technologies Corp., Carlsbad, CA, United States). See Supplementary Table S2 for primer sequences.

**Scanning Electron Microscopy (SEM) and Histology**

Plants treated with VIGS were analyzed by cryoSEM on fresh tissue under a JEOL 5410 microscope equipped with a CRIOSEM instrument CT 15000-C (Oxford Instruments®). Samples were collected for histological analyses, fixed in FAA (3.7% formaldehyde, 5% acetic acid, 50% ethanol) under vacuum and embedded into paraffin. Sections 10 μm thick were stained in 0.2% toluidine blue solution, and observed under a Nikon Eclipse E-600 microscope².

**AUTHOR CONTRIBUTIONS**

CFe conceived the project and designed the experiments together with CFo. AG-F, VS-G, CFo, and CFe performed the experiments. CFe wrote the paper.

**FUNDING**

This work was supported by the Spanish MINECO/FEDER grants no BIO2012-32902 and BIO2015-64531-R to CFe. AG-F was supported by a predoctoral contract of the Generalitat Valenciana (ACIF/2013/044).

**ACKNOWLEDGMENTS**

We thank David Parejo and Victoria Palau (IBMCP) for greenhouse support, Marisol Gascón (IBMCP) for technical advice in microscopy and personnel at Servicio de Microscopia Electrónica (UPV) for help and technical assistance. *E. californica* germplasm used in this study was obtained from the National Genetic Resources Program (United States) and Arabidopsis *sty1-1 sty2-1* and 35S::STY1 seeds from Eva Sundberg. We acknowledge support of the publication fee by the CSIC Open Access Publication Support Initiative through its Unit of Information Resources for Research (URICI).

---

1http://www.oxford-instruments.com

2http://www.nikoninstruments.com
SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2017.00814/full#supplementary-material

FIGURE S1 | Amino acid identity of the SHI/STY/SRS homologs included in this study. The conserved C3HC3RING Zinc finger and IGGH domains are highlighted in blue. In red, the atypical Pro residues in the E. californica proteins.


REFERENCES


Heijmans, K., Ament, K., Rijpkema, A. S., Zethof, J., Wolters-Arts, M., Gerats, T.,


Frontiers in Plant Science | www.frontiersin.org May 2017 | Volume 8 | Article 814

Martinez-Fernandez, I., Sanchis, S., Marini, N., Balanza, V., Ballester, P., Navarrete-


doi: 10.1093/molbev/msw054


STY1 and STY2 promote the formation of apical tissues during Arabidopsis


redundant SHI family genes regulate Arabidopsis gynoecium development in a


Polar auxin transport is essential for medial versus lateral tissue specification and


doi: 10.1104/pp.111.179501

Lee, J.-Y., Baum, S. F., Oh, S.-H., Jiang, C.-Z., Chen, J.-C., and Bowman, J. L. (2005). Recruitment of CRABS CLAW to promote nectary development within the


 Trigueros, M., Navarrete-Gómez, M., Sato, S., Christensen, S. K., Pelaz, S.,
Weigel, D., et al. (2009). The NGATHA genes direct style development in the

Bot. 67, 2549–2563. doi: 10.1093/jxb/erw119

mcm118


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

Copyright © 2017 Gomariz-Fernández, Sánchez-Gerschon, Fourquin and Ferrándiz. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.
Overview of OVATE FAMILY PROTEINS, A Novel Class of Plant-Specific Growth Regulators

Shucai Wang1*, Ying Chang2 and Brian Ellis3

1 Key Laboratory of Molecular Epigenetics of MOE and Institute of Genetics and Cytology, Northeast Normal University, Changchun, China, 2 College of Life Science, Northeast Agricultural University, Harbin, China, 3 Michael Smith Laboratories, The University of British Columbia, Vancouver, BC, Canada

OVATE FAMILY PROTEINS (OFPs) are a class of proteins with a conserved OVATE domain. OVATE was first identified in tomato as a key regulator of fruit shape. OFPs are plant-specific proteins that are widely distributed in the plant kingdom including mosses and lycophytes. Transcriptional activity analysis of Arabidopsis OFPs (AtOFPs) in protoplasts suggests that they act as transcription repressors. Functional characterization of OFPs from different plant species including Arabidopsis, rice, tomato, pepper, and banana suggests that OFPs regulate multiple aspects of plant growth and development, which is likely achieved by interacting with different types of transcription factors including the KNOX and BELL classes, and/or directly regulating the expression of target genes such as Gibberellin 20 oxidase (GA20ox). Here, we examine how OVATE was originally identified, summarize recent progress in elucidation of the roles of OFPs in regulating plant growth and development, and describe possible mechanisms underpinning this regulation. Finally, we review potential new research directions that could shed additional light on the functional biology of OFPs in plants.

Keywords: OVATE, OVATE FAMILY PROTEINS, fruit shape, transcription factor, plant growth and development, Arabidopsis, rice, pepper

INTRODUCTION

More than 100 years ago, it was proposed that the pear-shaped fruit phenotype in tomato (Solanum lycopersicum) might be genetically controlled by a single recessive locus, namely ovate (Hedrick and Booth, 1907; Price and Drinkard, 1908). However, it was only at the end of the last century that Ku et al. (1999) established, by construction and analysis of near-isogenic lines (NILs), that the ovate locus could account for both the pear shape and the elongated fruit shape in tomato. The ovate locus was later mapped to chromosome 2 (Ku et al., 2001), and the ovate gene was finally cloned in 2002 (Liu et al., 2002). Amino acid sequence analysis showed that OVATE is different from all the previously characterized plant genetic regulators, indicating that OVATE FAMILY PROTEINS (OFPs) represent a novel class of plant regulators (Liu et al., 2002).

Subsequent studies revealed that OFPs are widely distributed in the plant kingdom, and that they regulate multiple aspects of plant growth and development (Gui and Wang, 2007; Rodriguez et al., 2011; Tsaballa et al., 2011; Wang et al., 2011; Huang et al., 2013).
THE 100-YEAR PATHWAY TO THE IDENTIFICATION OF THE OVATE GENE

About a century ago, a series of genetic studies in tomato suggested that fruit size and shape in this economically important species are quantitatively inherited (Hedrick and Booth, 1907; Price and Drinkard, 1908; Grane, 1915; Lindstrom, 1926, 1927, 1929, 1932; MacArthur, 1926; Tanksley, 2004). More specifically, pear-shaped fruit form in tomato was proposed to be controlled by a single recessive quantitative trait locus (QTL), which was named pyriform (pr) (Hedrick and Booth, 1907; Price and Drinkard, 1908). Later, pr was found to co-segregate with the locus causing oblate- to oval-shaped fruit, and was therefore renamed ovate (o). In the late 1920s, based on its linkage to the dwarf locus, ovate was placed on chromosome 2 (Lindstrom, 1926, 1927, 1929; MacArthur, 1926).

Nearly 70 years later, Ku et al. (1999) conducted a more detailed molecular marker-based analysis of fruit shape in Lycopersicon. By crossing Lycopersicon esculentum cv. Yellow Pear (TA503), a variety of tomato bearing small, pear-shaped fruit, with Lycopersicon esculentum pimpinellifolium (LA1589), a wild tomato species bearing round-shaped fruit, and examining the F2 population, they found that the pear-shaped fruit phenotype is largely controlled by a major QTL on chromosome 2. This observation was confirmed by analyzing F2 populations from crosses between TA503 and a round-fruited introgression tomato line IL2-5, which carried the distal ovate region on chromosome 2 from the L. pennellii genome. Based on these results, they proposed that the QTL detected on chromosome 2 corresponds to the ovate locus (Ku et al., 1999).

High-resolution mapping of the ovate region on chromosome 2 using a total of 3000 near-isogenic lines (NILs) derived from TA503 and L. pennellii, allowed Ku et al. (2001) to place ovate adjacent to a known marker. By screening tomato BAC (bacterial artificial chromosome) and binary BAC libraries with the known marker-derived probe, and mapping the ends of the selected BAC clones, they were able to identify two ovate-containing BAC clones (Ku et al., 2001). A combination of sequencing and fine mapping analysis of one of these BAC clones obtained further narrowed the ovate locus to a 55 kb fragment that contained eight open reading frames (ORFs) (Liu et al., 2002). To identify the OVATE gene, Liu et al. (2002) amplified and compared the corresponding 55 kb fragments from TA496, a round-fruited wild type L. esculentum genotype, and TA493 (L. esculentum cv. Heinz 1706), an ovate genotype. They identified a G\textsubscript{T}A496→T\textsubscript{TA493} nucleotide polymorphism (SNP) in one of ORFs. This SNP created an early stop codon in the ovate genotype, leading to a 75-aa truncation in the C-terminus of the predicted protein. Sequence comparison of the corresponding ORF in several pear-shaped tomato varieties including TA503, LA791 (L. esculentum cv. Longjohn), and LA0025, as well as complementation of the pear-shaped fruit phenotype in TA503 by over-expression of the genomic DNA fragment containing the ORF and its 5′ and 3′ untranslated regions from TA496, confirmed the identity of the tomato OVATE gene.

Amino acid sequence analysis showed that the OVATE protein contains an ~70-aa carboxyl-terminal domain, referred to as the OVATE domain, that is conserved in both Arabidopsis and rice. The premature termination occurring in the ovate genotype eliminates most of this conserved OVATE domain (Liu et al., 2002). The OVATE protein also contains a putative bipartite nuclear localization signal (NLS), and two putative Von Willebrand factor type C (VWFC) domains required for protein–protein interaction, features that distinguish OVATE from any of the previously identified plant genetic regulators (Liu et al., 2002).

Thanks to fast development of new sequencing technologies over the recent years, the genomes of many more plant species have now been fully sequenced, which has greatly facilitated the identification of protein homologs and phylogenetic studies in plants. Based on amino acid sequence similarity analysis, OFPs are found exclusively in plants (Hackbusch et al., 2005; Gui and Wang, 2007; Wang et al., 2007, 2011; Rodriguez et al., 2011; Tsabella et al., 2011; Huang et al., 2013). By using the amino acid sequences of OFPs in Arabidopsis and the OVATE protein in tomato to search genomes of 13 land plants including Solanum lycopersicum, Solanum tuberosum, Mimalius guttatus, Arabidopsis, Vitis vinifera, Populus trichocarpa, Prunus persica, Carica papaya, Aquilegia coerulea, rice (Oryza sativa), Zea mays, Selaginella moellendorffii, and Physcomitrella patens, Liu et al. (2014) found that OFPs are distributed in all the plants examined, including the seedless vascular plant Selaginella moellendorffii, and the non-vascular plant Physcomitrella patens. In contrast, no OFPs were identified in chlorophytes, a division of the green algae (Liu et al., 2014). Interestingly, monocot species appeared to have more OFP family members than eudicots; for example, Zea mays had 45 OFPs, and rice had 33 OFPs, whereas tomato had 26 OFPs and Vitis vinifera had only 9 (Liu et al., 2014). It was subsequently reported that 31 genes in rice encode full-length OFPs (Yu et al., 2015). It should also be noted that Liu et al. (2014) found that the Arabidopsis genome contains 19 OFP-encoding genes, rather than 18 as reported previously (Hackbusch et al., 2005; Wang et al., 2011).

ROLES OF OFPs IN PLANT GROWTH AND DEVELOPMENT

Although OFPs are widely distributed in the plant kingdoms, their biological functions in plants remain largely unknown. However, limited studies in several different plant species including tomato, Arabidopsis, pepper (Capsicum annuum) and banana (Musa acuminate) have revealed that OFPs proteins are involved in regulation of several aspects of plant growth and development.

Ovule Development

Characterization of a T-DNA insertion mutant for AtOFP5 suggests that this member of the Arabidopsis OFP family may be involved in the regulation of ovule development. When Pagnussat et al. (2007) failed to isolate homozygous plants from a T-DNA insertion line of AtOFP5 (SALK_010386), they examined female gametophyte development in pistils of the
ofp5−/− heterozygotes. This revealed that, out of the 256 embryo
sacs studied, ~38% of them seemed to collapse at the FG2
female gametophyte development stage, when two-nucleate cells
are normally produced, and ~8% of the embryo sacs showed
abnormal micropylar cells with two egg cells, indicating that
AtOFP5 is required for proper female gametophyte development.
Because some embryo sacs in the ofp5- genetic background
had two egg cells, AtOFP5 may be specifically involved in
the regulation of a cell fate switch from synergid to egg cell
development (Pagnussat et al., 2007).

Vascular Development
OsOFP2, a rice OFP gene, was found to be expressed mainly
within the vasculature in all growth stages examined in rice,
as shown by GUS staining in OsOFP2pro:GUS transgenic rice
plants. Rice plants over-expressing OsOFP2 under the control
of the 35S promoter showed reduced height and exhibited
altered leaf morphology, seed shape, and positioning of vascular
bundles in the stems. Transcriptome analysis of the OsOFP2
over-expressing rice plants showed that the expression of genes
associated with vascular development, lignin biosynthesis, and
hormone homeostasis was altered, consistent with a role for
OsOFP2 in the regulation of vascular development (Schmitz et al.,
2015).

Fruit Shape
Consistent with OVATE’s role as a key regulator of fruit shape in
tomato (Liu et al., 2002; Azzi et al., 2015; Wang et al., 2015; Wu
et al., 2015), the OVATE gene is expressed mainly in reproductive
organs. Its transcripts can be detected in tomato flowers 10 days
before anthesis and begin to decrease in developing fruit 8 days
after anthesis (Liu et al., 2002). CaOVate, an OVATE family
member in pepper that shares 63% identity with the tomato
OVATE protein, was also found to be involved in the regulation
of fruit shape (Tsaballa et al., 2011). However, unlike in tomato,
differences in fruit shape in two pepper cultivars, cv. “Mytilini
Round” and cv. “Piperaki Long,” are associated with different
defaults of CaOVate expression, rather than with an ORF
mutation; i.e., the expression of CaOVate is higher in cv. “Mytilini
Round” than in cv. “Piperaki Long.” Down-regulation of CaOVate
in cv. “Mytilini Round” produced through virus-induced gene
silencing (VIGS) thus changed its fruit to a more obovate form
(Tsaballa et al., 2011). Beyond the Solanaceae, several QTLs
controlling fruit shape in melon (Cucumis melo) have also been
identified, and several melon OFP homologs (GmOFP) were
found to co-map with seven fruit shape QTLs, indicating that
OFPs in melon may also control fruit shape (Monforte et al.,
2014).

It should be noted that the G-to-T mutation in OVATE is
not associated with a single fruit shape phenotype in tomato.
In some genetic backgrounds, the mutation leads to pear-shape
fruits, but in other backgrounds, fruit shape remains largely
unchanged, indicating that the OVATE locus may interact with
other regulators to control a specific fruit shape (Tanksley, 2004;
Gonzalo and van der Knaap, 2008). Indeed, two suppressor loci
for OVATE, suppressor of OVATE1 (sov1) and sov2, have been
identified in tomato (Rodriguez et al., 2013).

Grafting of Capsicum cv. “Mytilini Round” scion on the
long-shaped cultivar, cv. “Piperaki Long” rootstock, also resulted
in inheritable fruit shape changes in the scion, and these fruit
phenotypic changes were retained through two generations of
seed-derived progeny (Tsaballa et al., 2013). However, only a
slight difference in CaOVate gene expression accompanied this
change in the fruit shape, and simple sequence repeat (ISSR)
analysis of the progenies of the scion fruits showed that their
genetic profile closely resembled the parental scion genetic
profile. This suggests either that only slight changes in the
expression of CaOVate are enough to induce fruit shape changes,
or that CaOVate interacts with other genes in an epistatic manner
to regulate fruit shape in pepper (Tsaballa et al., 2013). It is also
possible that epigenetic modification plays a role in the regulation
of fruit shape in pepper.

DNA Repair
When Hackbusch et al. (2005) tried to characterize Arabidopsis
ofp1 mutants, they failed to recover any homozygous plants
from three independent lines with T-DNA inserted either in
the single exon or in the 5′ upstream region of AtOFP1. They
concluded that AtOFP1 must be required for essential
processes in gametophyte or sporophyte development. However,
they did not observe any apparent morphological aberrations
in either the pollen or ovules in plants heterozygous for the
T-DNA insertion. When they further investigated male and
female transmission of the T-DNA insertion by reciprocally
crossing the mutants with wild type plants, they found that
when the heterozygous mutant lines were used as pollen donors,
no T-DNA insertion in the AtOFP1 gene was detected in the
F1 plants. The authors therefore concluded that AtOFP1 is
essential for male gamete and pollen function (Hackbusch et al.,
2005).

In contrast to this report, Wang et al. (2007) successfully
obtained homozygous mutant plants from ofp1-2 and ofp1-3,
two mutant lines derived from the same T-DNA insertion lines
used by Hackbusch et al. (2005). These authors reported that,
in the ofp1-2 mutant, the T-DNA was actually inserted in
the 5′-UTR of AtOFP1 rather than in the exon as indicated by
the T-DNA Express database (http://signal.salk.edu/cgi-in/tdnaexpress). The expression of AtOFP1 in the ofp1-2 mutant
was largely unaffected. Wang et al. (2007) also identified an
ofp1-1 mutant, a true loss-of-function mutant line for AtOFP1
derived from a transposon insertion event, and found that this
mutant is morphologically similar to the wild type plant. These
results are inconsistent with the earlier proposal that OFP1
plays a critical role in male gamete transmission or pollen
function.

Although the ofp1-1 mutant is largely similar to wild type
plant (Wang et al., 2007), ofp1-1 mutant seedlings were reported
to be more sensitive to treatment with methyl methanesulfonylate
(MMS), a DNA-damaging reagent. The ofp1-1 mutants also
showed relative lower non-homologous end-joining (NHEJ)
activity in vivo. Because the NHEJ pathway is thought to be
involved in the repair of DNA double-strand breaks (DSBs), these
results suggest that OFP1 may play a role in this DNA repair
pathway (Wang et al., 2010).
Secondary Cell Wall Formation

Although all the single and double AtOFP gene mutants identified in Arabidopsis are morphologically similar to wild type plants (Wang et al., 2011), careful examination of the anatomy of stem inflorescence cross-sections showed that ofp4 mutants exhibited an irregular xylem (irx) phenotype, marked by increased thickness of interfascicular fiber cell walls, and decreased wall thickness of vessel and xylary fiber cells (Li et al., 2011). This phenotype is similar to that of knat7, a loss-of-function mutant of the KNOTTED ARABIDOPSIS THALIANA7 transcription factor gene (Brown et al., 2005). Further analysis showed that the phenotype of the ofp4 knat7 double mutant was similar to that of the ofp4 and knat7 single mutants, and that the AtOFP4 over-expression phenotype (kidney-shaped cotyledons, and round and curled leaves) (Wang et al., 2011), was suppressed in a knat7 mutant background. Taken together, these results suggest that AtOFP4 is involved in the regulation of secondary cell wall formation, and that its function is at least partially dependent on KNA7.

Unlike ofp4, xylem and interfascicular fiber morphology in the ofp1-1 mutant is indistinguishable from that in wild type plants (Li et al., 2011). However, the AtOFP1 over-expression phenotype is similar to that of AtOFP4 over-expression plants (Wang et al., 2011), and is also suppressed in a knat7 mutant background, suggesting that AtOFP1 may also be involved in the regulation of secondary cell wall formation (Li et al., 2011). However, the possibility cannot be ruled out that this behavior may simply reflect the close homologous relationship between AtOFP1 and AtOFP4.

Fruit Ripening

A role for OFPs in regulating fruit ripening has so far only been reported for banana. The banana OFP1 (MaOFP1) was identified in a yeast two-hybrid (Y2H) study when using the banana MADS-box protein MuMADS1 as bait to screen a 2-day-postharvest (DPH) banana fruit cDNA library (Liu et al., 2015). Quantitative RT-PCR (qRT-PCR) analysis showed that both MuMADS1 and MaOFP1 were highly expressed in 0 DPH fruit, but had relative low levels of expression in the stem. However, different expression patterns of MuMADS1 and MaOFP1 were observed in different tissues and developing fruits. qRT-PCR analysis also showed that expression of MuMADS1 and MaOFP1 was differentially regulated by ethylene, i.e., expression of MuMADS1 was induced by exogenously applied ethylene, and suppressed by the ethylene competitor 1-methylcyclopropene (1-MCP), whereas expression of MaOFP1 was suppressed by ethylene, and induced by 1MCP. These results indicate that MuMADS1 and MaOFP1 may play antagonistic roles in ethylene-induced postharvest fruit ripening in banana (Liu et al., 2015).

Pleiotropic Effects

Although some evidence suggests that OFPs may have specific effects on different aspects of plant growth and development, as described above, other evidence suggests that OFPs can also have complex pleiotropic effects on plant growth and development. For example, in addition to resulting in round fruit, over-expression of Ovate in tomato also produced a number of abnormal phenotypes, including reduced size of floral organs, dwarf plants with smaller compound-leaf size and rounder leaflets (Liu et al., 2002). Similarly, over-expression of some Arabidopsis OFP genes also resulted in inhibited plant growth and development (Hackbusch et al., 2005; Wang et al., 2011).

Analysis of transgenic plants over-expressing AtOFPs in Arabidopsis also reveals that they regulate multiple aspects of plant growth and development in this species. Based on the similarity of the phenotypes observed, AtOFPs could be grouped into different groups. Over-expression of AtOFP1, AtOFP2, AtOFP4, AtOFP5 and AtOFP7 resulted in similar phenotypes including kidney-shaped cotyledons, as well as round and curled leaves, and these AtOFPs were designated as Class I AtOFPs. Over-expression of AtOFP6 and AtOFP8, on the other hand, resulted in a different phenotype including flat, thick and cyan leaves, and these were therefore designated as Class II AtOFPs. Over-expression of AtOFP13, AtOFP15, AtOFP16 and AtOFP18 led to another distinct phenotype including blunt-end siliques, and these AtOFPs were designated as Class III AtOFPs. Plants over-expressing all other AtOFPs examined were largely indistinguishable from wild type plants (Wang et al., 2011). Interestingly, this functional-based classification is largely consistent with the subgroups of the AtOFPs identified in phylogenetic analysis (Figure 1). Because AtOFP19 and AtOFP20 are newly identified OFPs in Arabidopsis (Liu et al., 2014), their functions have not yet been examined. Consistent with their functions in regulating multiple aspects of plant growth and development, expression of most of the OFPs in tomato, rice and Arabidopsis was detectable in all the tissues and organs examined (Wang et al., 2011; Liu et al., 2014; Schmitz et al., 2015). Nevertheless, differential expression patterns are also observed; for example, nearly half of the OsOFPs were found to be more highly expressed during the stages of rice seed development (Yu et al., 2015).

It is interesting that all the AtOFPs shown through genetic manipulations, as described above, to have a specific role in regulating plant growth and development are from the Class I subfamily. All knockout mutants identified so far for other AtOFP genes, including AtOFP8, AtOFP10, AtOFP15, and AtOFP16, are morphologically similar to wild type plants (Wang et al., 2011). A double mutant between two class III OFP genes, ofp15 ofp16 is also indistinguishable from wild type plants (Wang et al., 2011).

MECHANISMS UNDERLYING THE REGULATION OF PLANT GROWTH AND DEVELOPMENT BY OFPs

Regulation of Target Gene Expression

The evidence available so far supports a model in which OFPs regulate plant growth and development by directly influencing expression of their target genes, and/or through interaction with other transcription factors. Both AtOFP1 and AtOFP5 were found to associate with the cytoskeleton in transient expression...
Assays in tobacco leaves (Hackbusch et al., 2005). However, similar to the tomato OVATE protein, AtOFP1 also possesses a putative NLS (Liu et al., 2002; Wang et al., 2007), and in both transient transfection assays in Arabidopsis protoplasts, and stable transformed Arabidopsis plants, AtOFP1 was found to localize in the nucleus (Wang et al., 2007). AtOFP1 also contains an LxLxL motif (Wang et al., 2007), which is also found in Aux/IAA proteins and ERF transcription factors and is required for their transcriptional repression functions (Ohta et al., 2001; Hiratsu et al., 2003; Tiwari et al., 2004). Transfection assays in Arabidopsis protoplasts showed that AtOFP1 repressed reporter gene expression when recruited to the promoter region of the Gal4:GUS reporter gene by a fused Gal4 DNA binding domain (GD), suggesting that AtOFP1 may function in vivo as a transcriptional repressor (Wang et al., 2007). Although some of the AtOFPs do not contain a LxLxL motif, all the AtOFPs examined repressed Gal4:GUS reporter gene expression in transfected protoplasts, indicating that AtOFPs in Arabidopsis are a novel transcription repressor family (Wang et al., 2011).

**Targets of OFP Proteins**

A gain-of-function mutant of AtOFP1, *opf1-1D*, as well as *AtOFP1* over-expressing plants, all showed reduced lengths in their aerial organs, including hypocotyl, rosette leaves, cauline leaves, inflorescence stem, floral organs and siliques (Hackbusch et al., 2005; Wang et al., 2007). Detailed analysis showed that this growth phenotype was the result of a reduction in cell elongation, rather than in cell division (Wang et al., 2007), suggesting that AtOFP1 represses cell elongation. Hackbusch et al. (2005) had earlier shown that expression of *AtGA20ox1*, a gene encoding a key enzyme in gibberellic acid (GA) biosynthesis, was reduced in plants over-expressing AtOFP1, and Wang et al. (2007) showed, by use of chromatin immunoprecipitation assays, that the *AtGA20ox1* promoter is a direct target of AtOFP1.

*GA20ox* is likely also a target of OFPs in pepper and rice. In pepper, two lines of evidence suggest that *GA20ox* is likely a target gene of the OFP homolog, CaOvate. First, RT-PCR analysis showed that although *GA20ox* has a similar expression level at most growth stages in both cv. “Mytilini Round” and cv. “Piperaki Long,” its expression was significantly higher in cv. “Piperaki Long” fruits 10 days after anthesis. In addition, expression of *GA20ox* was elevated after VIGS suppression of *CaOvate* expression (Tsaballa et al., 2011, 2012). Similarly, rice *GA20ox7* is likely a target gene of OsOFP2, since expression of *GA20ox7* was reduced in transgenic rice plants over-expressing OsOFP2 (Schmitz et al., 2015).

Although *GA20ox1* is probably a direct target of AtOFP1 (Wang et al., 2007), and reduction in the expression of rice *GA20ox7* is correlated with lower gibberellin content in transgenic rice plants (Schmitz et al., 2015), exogenous GA can only partially restore the defects in cell elongation in Arabidopsis transgenic plants over-expressing AtOFP1 (Wang et al., 2007). This indicates that AtOFP1 may have other target genes, and indeed, microarray-based gene expression assays showed that a total of 129 genes were down-regulated at least twofold when AtOFP1-GR (glucocorticoid receptor) transgenic plants were treated with DEX (dexamethasone), which will allow the AtOFP1-GR protein to relocate into the nucleus (Wang et al., 2011). However, it remains unknown if any of these genes is directly targeted by AtOFP1.

**FIGURE 1** | Phylogenetic analysis of Arabidopsis OVATE FAMILY PROTEINS (AtOFPs). The entire amino acid sequences of AtOFPs were obtained from Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html) and used to generate the phylogenetic tree by using “One Click” mode with default settings on Phylogeny (www.phylogeny.fr). Branch support values are indicated above the branches, and Bar indicates branch length.
Interaction of OFPs with Other Transcription Factors

So far nearly all OFPs with known functions were found to regulate plant growth and development via interaction with homeodomain proteins (Table 1). In a yeast two-hybridization screen, nine AtOFPs were found to interact with the 3-amino acid loop extension homeodomain (TALE) transcription factors KNOX and BELL (BEL1-like homeodomain) (Hackbusch et al., 2005). Consistent with the observation that the ofp4 knat7 double mutant phenotype was similar to that of the ofp4 or knat7 single mutants, and that the AtOFP1 and AtOFP4 over-expression phenotype was suppressed in a knat7 mutant background, transfection assays in Arabidopsis protoplasts showed that both AtOFP1 and AtOFP4 physically interacted with KNAT7 (Li et al., 2011). Transfection assays in Arabidopsis protoplasts also showed both OFP1 and OFP4 interacted with BLH6 (Liu and Douglas, 2013). Considering that MYB75/PAP1 (PRODUCTION OF ANTHOCYANIN PIGMENT 1), a R2R3 MYB transcription factor that has been shown to regulate phenylpropanoid biosynthesis in Arabidopsis (Borevitz et al., 2000), interacts with KNAT7 to modulate secondary cell wall deposition in stems and seed coat in Arabidopsis (Bhargava et al., 2013), it is likely that AtOFPs, MYB transcription factors, and TALE homeodomain proteins form one or more multi-protein complexes to regulate secondary cell wall formation. Recently, AtOFP1 was also found to interact with the BELL transcription factor BLH3 to regulate the vegetative to reproductive phase transition in Arabidopsis (Zhang et al., 2016).

Homologs of Arabidopsis KNOX and BELL proteins have also been found to interact with OFPs and to regulate secondary cell wall formation in other plant species. For example, GhKNL1, a homeodomain protein in cotton (Gossypium hirsutum) is found to be preferentially expressed in developing fibers at the stage of secondary cell wall biosynthesis, and ectopic expression of GhKNL1 can partially rescue the cell wall defective phenotype of the Arabidopsis knat7 mutant. Yeast two-hybrid assays showed that GhKNL1 interacts with GhOFP4, as well as with AtOFP1, AtOFP4, and AtMYB75 (Gong et al., 2014). In rice, OsOFP2 was found to interact with putative vascular development KNOX and BELL proteins, so it is also likely that OsOFP2 can modulate KNOX-BELL function in this species to regulate diverse aspects of development, including vascular development (Schmitz et al., 2015).

AtOFP5 has been shown to be involved in the regulation of female gametophyte development. The “two egg cells” phenotype in eostre-1, a female gametophyte mutant, was caused by elevated expression of the BELL transcription factor gene BLH1 in the embryo sac, and this phenotype is dependent upon the function of the class II knox gene, KNAT3, (Pagnussat et al., 2007). Disruption of AtOFP5, a known interactor of KNAT3 and BLH1, also partially phenocopies the eostre mutation (Pagnussat et al., 2007), suggesting that the roles of AtOFP5 in female gametophyte development may also involve interactions of AtOFP5 with TALE homeodomain proteins.

On the other hand, Y2H screening using tomato OVATE protein as bait did not lead to the identification of any KNOX or BELL transcription factors. Instead, the screen revealed interactions of OVATE with 11 out of 26 members of the TONNEAU1 Recruiting Motif (TRM) superfamily, including the putative tomato ortholog of AtTRM17/20 (van der Knaap et al., 2014). Because some of the TRMs can bind microtubules and are likely centrosomal components in plant cells (Drevensek et al., 2012), this result suggests that, in addition to interacting with KNOX or BELL transcription factors, and acting as transcriptional repressors to repress target gene expression, OFPs may also interact with TRMs and microtubules to regulate plant growth and development (van der Knaap et al., 2014).

CONCLUDING REMARKS AND FUTURE PERSPECTIVES

Studies in recent years reveal that OFPs are widely distributed in the plant kingdom, and that they help regulate multiple aspects of plant growth and development in various species (Liu et al., 2002; Hackbusch et al., 2005; Gui and Wang, 2007; Wang et al., 2007, 2011; Rodriguez et al., 2011; Tsaballa et al., 2011; Huang et al., 2013). Beyond this general observation, however, it will be important to establish the extent to which OFP family members are functionally conserved in different taxa. As described above, it would appear that some OFP homologs may have conserved functions; for example, OVATE protein was identified as a key regulator of fruit shape in tomato (Liu et al., 2002), while fruit shape in pepper is also at least partially controlled by an OFP homologue, CaOVate (Tsaballa et al., 2011). However, tomato and Capsicum are closely related genera, so it remains to be established whether fruit shape in other plants is also controlled by homologous OFPs. In addition, a major QTL, f10.1, has been shown to control fruit elongation in Capsicum, but it is unclear if it represents an OFP gene (Borovsky and Paran, 2011). Similarly,

<table>
<thead>
<tr>
<th>OFPs</th>
<th>Interactors</th>
<th>Functions in plants</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>AtOFP1</td>
<td>KNAT7, BLH6</td>
<td>Secondary cell wall formation</td>
<td>Li et al., 2011; Liu and Douglas, 2015</td>
</tr>
<tr>
<td>AtOFP1</td>
<td>BLH3</td>
<td>Phase transition</td>
<td>Zhang et al., 2016</td>
</tr>
<tr>
<td>AtOFP4</td>
<td>KNAT7, BLH6</td>
<td>Secondary cell wall formation</td>
<td>Li et al., 2011; Liu and Douglas, 2015</td>
</tr>
<tr>
<td>AtOFP5</td>
<td>KANT3, BLH1</td>
<td>Female gametophyte development</td>
<td>Pagnussat et al., 2007</td>
</tr>
<tr>
<td>GhOFP4</td>
<td>GhKNL1</td>
<td>Secondary cell wall formation</td>
<td>Gong et al., 2014</td>
</tr>
<tr>
<td>OsOFP2</td>
<td>OsKNNAT7, BLH6-like1, bHLH6-like2</td>
<td>Vascular development</td>
<td>Schmitz et al., 2015</td>
</tr>
<tr>
<td>MaOFP1</td>
<td>MuMADS1</td>
<td>Fruit ripening</td>
<td>Liu et al., 2015</td>
</tr>
</tbody>
</table>
several QTLs controlling fruit shape in apple have been identified, but so far it is unclear whether any of them is associated with an OFP gene (Cao et al., 2015). Once these genomes have been fully sequenced, the candidate regions can be explored for the possible presence of OFP homologs.

Further evidence is also required to establish the extent to which the function of OFPs in controlling secondary cell formation might be evolutionarily conserved. In Arabidopsis, interaction of specific OFPs with homeodomain proteins strongly influences secondary cell wall formation (Li et al., 2011; Liu and Douglas, 2015). In cotton (Gossypium hirsutum), GhKNL1 has been reported to interact with GhOFP4, but it remains to be demonstrated that GhOFP4 is also involved in the regulation of secondary cell wall formation in cotton fibers.

So far only AtGA20ox1 has been shown to be a direct target of AtOFP1. However, the observation that expression of >100 genes was down-regulated in DEX-treated AtOFP1-GR transgenic plants (Wang et al., 2011), raises the possibility that some of these genes may also serve as direct targets of AtOFP1.

The possibility cannot be excluded that OFPs may regulate plant growth and development in ways other than directly targeting expression of particular genes or interaction with other transcription factors. Phenotype similarity-based characterization may make it possible to detect the operation of a previously unsuspected mechanism. For example, transgenic Arabidopsis over-expressing Class III AtOFP genes display blunt siliques, a phenotype also observed in the mutants erecta, a loss-of-function mutant of ERECTA (ER), a putative receptor protein kinase, and agb1-1, a loss-of-function mutant of the heterotrimeric G-protein β subunit gene AGB1 (Torii et al., 1996; Lease et al., 2001). The agb1-1 mutant was originally identified as an erecta-like mutant, elk4, during the characterization of an ER signaling pathway, and was renamed as agb1-1 after ELK4 was found to encode AGB1. Their results indicate that AGB1 can influence silique morphology via an ER-type Leucine-rich repeat (LRR) receptor-like kinase signaling pathway. Considering that an ER LRR receptor-like kinase signaling pathway has been shown to regulate multiple aspects of plant growth and development, including organ shape (Shpak et al., 2003), and that the transcription factor SPEECHLESS has been shown to be phosphorylated by a MAP kinase signaling pathway (Lampard et al., 2008), it will be worthwhile to investigate whether Class III AtOFPs might be phosphorylated by kinases operating downstream of an ER LRR receptor-like kinase signaling pathway, and whether such post-translational modification of OFPs plays a role in regulating silique morphology.

**AUTHOR CONTRIBUTIONS**

SW conceived the review topic, BE and YC participated in the discussion of topic, SW drafted the manuscript, BE reversed the manuscript, and all the authors approved the final version of manuscript.

**ACKNOWLEDGMENTS**

This work was supported by the Natural National Science Foundation of China (31470297, 31370221), and a Discovery Grant from the Natural Science and Engineering Research Council of Canada (BEE). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**REFERENCES**


REFERENCES
Lindstrom, E. W. (1929). Fruit size and shape genes on the first chromosome of the tomato.


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

Copyright © 2016 Wang, Chang and Ellis. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.
Perspectives for a Framework to Understand Aril Initiation and Development

Sylvia R. Silveira¹, Marcelo C. Dornelas² and Adriana P. Martinelli¹*

¹ Laboratório de Biotecnologia Vegetal, Centro de Energia Nuclear na Agricultura, Universidade de São Paulo, Piracicaba, Brazil, ² Departamento de Biologia Vegetal, Instituto de Biologia, Universidade Estadual de Campinas, Campinas, Brazil

A differentiated structure called “aril” has been described in seeds of several plant species during the course of evolution and might be considered as a supernumerary integument. Besides its ecological function in seed dispersal, the structure also represents a relevant character for systematic classification and exhibits important properties that impart agronomic value in certain species. Little is known about the molecular pathways underlying this morphological innovation because it is absent in currently used model species. A remarkable feature of the seeds of *Passiflora* species is the presence of a conspicuous aril. This genus is known for the ornamental, medicinal, and food values of its species. In view of the molecular resources and tools available for some *Passiflora* species, we highlight the potential of these species as models for developmental studies of the aril.

Keywords: aril development, integument, model species, ovule, *Passiflora*, seed

INTRODUCTION

The morphological diversity among plant species results from differential gene expression controlling the development of novel features that ensure the adaptation and reproductive success of a species. An important question in plant biology is when and how these features emerged during evolution. One of such novel features is the aril. The aril is a differentiated structure present in seeds of several gymnosperm and angiosperm species, forming seed dispersal units. In many species, the aril accumulates several nutritional compounds attracting and rewarding frugivorous animals. There is a great amount of information available about morphological and molecular development of plant ovules and seeds and they can be used as initial clues to investigate aril development. These appendages are often used in systematics classification, since its presence, absence, form, and function vary among taxa. The well known model species do not exhibit this feature evidencing the need for novel models to study this specific structure. A better understanding of the processes involved in aril origin and development is interesting and necessary due to its economical, ecological and phylogenetic importance.

ARIL ORIGIN AND IMPORTANCE

Several plant species develop differentiated structures associated with their seeds, often constituting diasporas, which are plant dispersal units mostly related to their dispersion syndrome (Corner, 1976). Some authors also believe that these structures originated as a protection mechanism...
for seeds and embryos, regardless of their role in dispersion (Mack, 2000). Also associated to the ovule/seed, either one or more integuments are found. The current theory of the evolution of integuments states that there are different evolutionary origins for the outer and inner integuments in flowering plants (Endress, 2011). In angiosperms, the inner integument is considered homologous to the single integument of extant and fossil gymnosperms (Reinheimer and Kellogg, 2009), and the outer integument may have been derived from a cupule/leaf-like structure found in several gymnosperms (Gasser et al., 1998). The integuments may or may not originate appendages that perform a defined role in seed dispersion. Such seed appendages may be wings, spines, hairs, plumes, fibers, or fleshy tissues, receiving different denominations in the literature.

Both gymnosperms and angiosperms evolved the habit of enveloping the seeds with a fleshy tissue (Lovisetto et al., 2012). Such tissue, called “aril,” generally accumulate sugars and other substances that will confer biological roles similar to those of fruits (Herrera, 1989).

The use of the term “aril” is quite controversial in the literature. It has been used both in a broader sense, referring to any fleshy structure associated with the seed, but also to designate structures with a specific anatomical origin. According to Corner (1976), the term defines a structure varying from a fleshy to a more-or-less hard consistency, which develops from part of the ovule after fertilization and envelopes the seed partially or completely. Van der Pijl (1972) preferred to distinguish these structures according to their anatomical origin, the aril being originated from the funiculus. Therefore, a structure developing from other parts of the ovule are usually called arillode, false aril, or aril-like structure (Figure 1). Both, aril and arillode, are somehow associated with integuments. In fact, some authors consider the “true” aril as a supernumerary integument (Maheshwari, 1950; Kapil and Vasil, 1963; Endress, 2011).

As arils are generally fleshy structures, they are of extreme importance because during their development and ripening they accumulate substances that confer properties that not only attract dispersion agents, but also arouse interest for human consumption. Arils are very common in tropical and subtropical species and might accumulate oils (e.g., Ricinus communis), flavor- and aroma-rich compounds (Myristica fragrans), nutrients, and sugars (Passiflora edulis), among other substances.

**ARIL ONTOGENY**

Few studies describe the ontogeny and/or morphological aspects of aril formation and associate these with ovule development; this lack of information probably led to its controversial nomenclature. Additionally, the current model plant species do not exhibit this unique structure, making it difficult to characterize its development, especially at the molecular level.

Aril developmental stages were observed in some species of Passiflora (Raju, 1956; Singh, 1962; Dathan and Singh, 1973), and described in greater detail in P. suberosa and Turnera ulmifolia (Kloos and Bouman, 1980). It has also been described in Leguminosae, such as Eriosoma glaziovi (Grear and Dengler, 1976), Cytisus striatus, and C. multiflorus (Rodriguez-Riaño et al., 2006). More recently, the development of an aril was described in Celastraceae, however, the authors showed that the origin of the aril-like structure was not from the funiculus, calling it “caruncula” (Zhang et al., 2011).

Aril development has been divided into stages by some authors, and ontogenetic descriptions suggest that it is a pre-anthesis event originating during megagametogenesis from periclinical divisions of epidermal cells of the funiculus, followed by anticlinical divisions, forming a ring or collar-like structure surrounding the ovule (Kloos and Bouman, 1980; Rodriguez-Riaño et al., 2006). The specific stage of ovule development in which the aril initiates is not very clear in most of the reports. The first divisions might be observed between the tetrad formation stage, when integuments are elongating toward the nucellus, and the beginning of megagametogenesis, when the outer integument has already enveloped the inner integument and the nucellus, forming the micropyle (Raju, 1956; Singh, 1962; Dathan and Singh, 1973; Grear and Dengler, 1976; Kloos and Bouman, 1980; Rodriguez-Riaño et al., 2006).

**MOLECULAR MECHANISMS CONTROLLING INTEGUMENT INITIATION AND GROWTH**

As mentioned, the aril initiates during ovule development after the emergence and growth of integuments, resembling its development and exhibiting similar patterns of polarity. Thus, to speculate on whether the aril is an extra integument, and which molecular mechanisms might be involved in its identity and development, one should look closely to the molecular basis at the integument initiation and growth.

The development of the ovule in plants has been well characterized in model species, such as Arabidopsis and Petunia, through molecular genetic studies. Several genes involved in different events of ovule development where identified through mutant screening, as reviewed by Angenent and Colombo (1996), Gasser et al. (1998), and Schneitz (1999). The results obtained from mutant characterization, patterns of gene expression, and transcriptomic analyses in the last two decades allowed for the elucidation of regulatory networks controlling the initiation and development of integuments. Most of the genes characterized encode transcription factors, and molecular studies have been performed to better understand the means by which these factors act, and how they interact regulating integument morphogenesis.

Integument formation marks the transition from the earlier established proximal-distal axis of the ovule primordia to an additional adaxial/abaxial polarity axis. Integument initiation is characterized by epidermal cell proliferation in a region between the nucellus and the funiculus. The putative transcriptional regulator NOZZLE/SPOROCYTELESS (NZZ) is required for maintaining the homeobox gene WUSCHEL (WUS) expression limited to the nucellus (Figure 2) (Sieber et al., 2004). Another factor restraining WUS in the nucellus is the interaction of BEL1 (a homeodomain protein) with an integument identity
protein complex that represses WUS in the chalaza, and activates INNER NO OUTER (INO) for outer integument development (Figure 2) (Brambilla et al., 2008). WUS, in turn, is sufficient to induce integument formation from the underlying chalazal tissue, since it generates downstream signals inducing meristematic activity even where it is not expressed (Gross-Hardt et al., 2002). An evidence for this is the induction of ectopic structures resembling integuments at the flanks of the funiculus, when WUS is ectopically expressed in the chalaza, under the control of the AINTEGUMENTA (ANT) promoter (Gross-Hardt et al., 2002). Thus, ectopic WUS expression caused by natural gain-of-function mutation(s) might be involved in the evolutionary origin of supernumerary integuments and, therefore, in structures resembling arils.

Additionally, NZZ is known to restrict both the homeodomain-leucine zipper gene PHABULOSA (PHB) in the abaxial domain of the chalazal region where the inner integument initiates (Sieber et al., 2004), and INO, which is responsible for outer integument differentiation (Figure 2) (Schneitz et al., 1997; Villanueva et al., 1999). INO expression, in turn, is restricted to the outer integument by WUS and, more specifically, to the abaxial side, where it is repressed by SUPERMAN (SUP). Thus, INO and SUP are responsible for the asymmetric growth of the outer integument (Figure 2) (Meister et al., 2002). BEL1, ANT, and HUELLENLOS (HLL) also participate directly or indirectly in INO negative spatial regulation (Villanueva et al., 1999). These antagonistic relations control integument polarity.

An additional mutant in which both integuments are present, but exhibits aberrant features is worth mentioning. The unicorn (unc) mutation results in excrescences emerging from the outer integument (Schneitz et al., 1997). Later on, UNC was found to encode an AGC VIII kinase that directly interacts with and represses the activity of ABERRANT TESTA SHAPE (ATS), a transcriptional regulator belonging to the KANADI family (Figure 2) (Enugutti et al., 2012; Enugutti and Schneitz, 2013). Thus, ectopic expression of ATS would provide another mechanism by which additional initiation and growth of integument-derived tissue may occur, therefore indicating an alternative possible molecular mechanism underlying the evolutionary origin of aril or aril-like structures.

Considering the amount of information on regulatory networks for integument initiation and growth, along with the fact that most of these mechanisms are conserved among
different taxa, and the known morphoanatomy of arils, it becomes possible to identify the initial cues on the molecular basis of aril origin and development.

**MOLECULAR ASPECTS OF THE “RIPENING” OF FLESHY SEED STRUCTURES**

The development of fleshy seed structures such as the aril can be divided in three main stages: (1) initiation, which includes cell proliferation; (2) growth, with cell expansion, mainly; and (3) accumulation of storage products, which would be equivalent to a “ripening” stage. As we are assuming a similarity between integument and aril development, we considered the first two stages in the previous section, and we will now consider the third stage.

Since gymnosperms do not form ovaries that will develop into fruits after fertilization, many species developed fruit-like fleshy structures around their seeds to attract frugivorous animals that act as seed dispersers (Herrera, 1989). Because of its importance in the formation of reproductive structures in both gymnosperms and angiosperms, the involvement of MADS-Box genes in the development of fleshy structures was investigated in *Ginkgo biloba* and *Taxus baccata*, both gymnosperms (Lovisetto et al., 2012, 2013, 2015a), and in *Magnolia grandiflora*, a basal angiosperm (Lovisetto et al., 2015b). Gene expression analyses showed that AGL6, AGL16 (a gene phylogenetically close to the SEPALLATA clade), and TM8-like genes, are involved in the development of fleshy structures in both the sarcotesta of *Ginkgo* and the aril of *Taxus*, regardless of their anatomical origin (Lovisetto et al., 2012). Moreover, activated forms of AGL6 (AGL6::VP16) triggered ectopic outgrowths on the surface of reproductive structures in *Arabidopsis* (Koo et al., 2010). A subfamily of MADS-Box, the B-sister genes, is believed to be required for the correct development of ovule and seed, with their expression analyzed in the gymnosperms mentioned above (Lovisetto et al., 2013). The pattern of gene expression differed between these two species, being weaker throughout aril development in *Taxus*, indicating that the involvement of B-sisters in the formation of fleshy fruit-like structures might be dependent of their origin. In *Magnolia*, with a fleshy tissue also originating from the seed tegument, AGAMOUS, AGL6, SEPALLATA, and B-sister were also detected during the sarcotesta formation and growth (Lovisetto et al., 2015b).

There is evidence that a common set of genes was recruited independently in distantly related taxa, regulating the development of all fleshy structures, regardless of their anatomic origin in, both, gymnosperms and angiosperms. Accordingly, a group of tomato MADS-box genes have been implicated in fruit ripening, including members of the SEPALLATA and B-sister clades (Vrebalov et al., 2002; Yasuhiryo, 2016). Altogether, these observations suggest that fleshy tissues that undergo physiological changes that involve tissue softening, pigmentation, and accumulation of sugars, aroma, and flavor (or “ripening syndrome,” in general), appeared independently in fruits and seeds but are likely to be regulated, at the molecular level, by conserved pathways.

**Passiflora AS A SUGGESTED MODEL SYSTEM TO STUDY ARIL DEVELOPMENT**

Among the angiosperms, species belonging to *Passiflora* are noteworthy regarding their aril, which are often cited in anatomical and morphological literature as an example of a true aril. *Passiflora* is the largest genus of the family Passifloraceae with over 500 species, mostly originated in neotropical regions, with hundreds of species throughout Latin America (Kugler and King, 2004; Ulmer and MacDougal, 2004). *Passiflora* also include commercial species, such as *P. edulis*, *P. alata*, and *P. incarnata*, which are important for their ornamental, medicinal, and food values, the latter given specifically by the aril in which juice is produced and accumulated. Typically, the fruits of *Passiflora* are indehiscent berries, rarely a dehiscent capsule, very variable in shape, size, and color, and in general produces a mucilaginous or aqueous acid pulp, forming a cupuliform or saccate aril, covering each of numerous seeds (Figure 3) (Cervi, 1997; Dhawan et al., 2004). Passionfruit propagation is mainly carried out by seeds (Pereira and Dias, 2000), and the aril works as a reward for its dispersing agents (Ulmer and MacDougal, 2004), therefore, being directly related to the reproductive success of the wild species (Fenster et al., 2004), which highlights the ecological importance of this structure.

Studies about the aril ontogeny in *Passiflora* are scarce, although it has been addressed in descriptions of the embryology or seed coat structure of *Passifloraceae* (Raju, 1956; Singh, 1962; Dathan and Singh, 1973; Kloos and Bouman, 1980). The first mention of aril initiation in *Passiflora* is from *P. suberosa* and describes it as a ring around the distal area of the funiculus (Kratzer, 1918 in Kloos and Bouman, 1980). Later studies also refer to the aril primordium as a rim, collar or ring around the funiculus in several species of the family, *P. calcicara* (Raju, 1956; Singh, 1962), *P. foetida* (Singh, 1962), *P. caerulea*, *P. mollissima* (Dathan and Singh, 1973), and *P. edulis* (Dathan and Singh, 1973; Corner, 1976). These authors describe the origin of the aril as dermal, epidermal or hypodermal, and are not precise whether it develops from the funiculus, exostome, hilum, micropyle or raphe. A more detailed case study of aril development was performed using *P. suberosa* and *T. ulmifolia*, (Kloos and Bouman, 1980). According to this description, the aril is initiated during megagametogenesis by periclinal and anticlinial divisions of dermal cells, forming a rim around the funiculus from the raphe to the outer integument region at the micropyle (Kloos and Bouman, 1980). Differences among species occur mainly after fertilization. The aril of *P. suberosa* continues to grow, covers the micropyle, and by division of its apical cells, equally envelopes the developing seed, while the aril of *T. ulmifolia* grows unilaterally leaving the exostome exposed.

In spite of these descriptions of the initiation and development of the aril in *Passiflora* species, the molecular mechanisms implicated in these processes have not been described yet. Few
studies addressed gene expression in *Passiflora* arils, such as the analysis of differential expression among *PeETR1*, *PeERS1*, and *PeERS*. These genes encode proteins involved in ethylene perception in *Passiflora* fruit tissues, with higher levels of mRNA in arils than in seeds during fruit ripening (Mita et al., 1998; Mita et al., 2002). Nevertheless, these focused mainly on fruit ripening and, therefore, in genes involved in later aril developmental stages, and not in the identity and differentiation of this specialized structure.

Although in recent decades there has been a breakthrough in genome sequencing and genomic data analysis from crop species, efforts for entire genome sequencing were not done in *Passiflora* species, and very little is known about the genomics of this genus. The currently available sequence data in public databases are molecular markers used in phylogenetic and genetic diversity studies, such as microsatellites (Oliveira et al., 2005, 2008; Pádua et al., 2005; Cazé et al., 2012; Cerqueira-Silva et al., 2012, 2014), and internal transcribed spacers (Muschner et al., 2003; Yockteng and Nadot, 2004). On the other hand, specific transcript and genomic libraries for *Passiflora* have been constructed: a database of expressed sequence tags (ESTs) from libraries derived from *P. edulis* and *P. suberosa* reproductive tissues (Cutri and Dornelas, 2012), and a large-insert bacterial artificial chromosome (BAC) library of *P. edulis* (Santos et al., 2014). These are very resourceful for genomic studies allowing a greater understanding of gene structure and function, and the process of differentiation of complex morphological characters, which provide the diversity found among plants, such as the aril. Another useful resource that should aid these functional and developmental studies is the availability of genetic transformation and *in vitro* regeneration protocols for *Passiflora* species. Such protocols where generated by the large number of studies aiming at the genetic improvement of passion fruit, that have been carried out since the 1990s (Cerqueira-Silva et al., 2014), mainly to obtain transgenic plants resistant to the woodiness virus in *P. edulis* (Manders et al., 1994; Alfenas et al., 2005; Trevisan et al., 2006; Monteiro-Hara et al., 2011), and *P. alata* (Correa et al., 2015). Several protocols for *in vitro* regeneration via organogenesis or somatic embryogenesis for a large number of *Passiflora* species where established aiming at germplasm preservation, and recovery of transgenic plants, as reviewed by Vieira and Carneiro (2004) and Otoni et al. (2013). Although designed for breeding purposes, these methodologies are important tools to study the molecular basis of aril development. Novel genomic editing tools, such as the CRISPR/Cas9 technology, will also help in the genetic and molecular analysis of aril development.

**CONCLUSION**

Aris are accessory seed structures present in both gymnosperms and angiosperms, being important for seed dispersal, and might possess economic importance. Nonetheless, aril evolutionary origin and ontogenesis are largely unknown, with, both, structural and molecular information lacking and needed. Here we established parallels between ovule integuments...
and arils that might help the design of further studies. Our testable statements need a novel model species, since the traditional plant models do not develop arils. We postulate that *Passiflora* species are good candidates for such needed model.

**AUTHOR CONTRIBUTIONS**

SRS, MD, and AM designed the initial manuscript. SRS wrote the initial draft of the manuscript and conceived the figures.

**REFERENCES**


SRS, MD, and AM contributed reviewing and discussing the manuscript to produce its final version.

**ACKNOWLEDGMENTS**

The authors acknowledge financial support from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Brazil), Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, São Paulo, Brazil), and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brazil).


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

Copyright © 2016 Silveira, Dornelas and Martinelli. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.
Genome-Wide Characterization of the MADS-Box Gene Family in Radish (Raphanus sativus L.) and Assessment of Its Roles in Flowering and Floral Organogenesis

Chao Li††, Yan Wang††, Liang Xu†, Shanshan Nie†, Yinglong Chen‡, Dongyi Liang†, Xiaochuan Sun†, Benard K. Karanja†, Xiaobo Luo† and Liwang Liu*†

1 National Key Laboratory of Crop Genetics and Germplasm Enhancement, College of Horticulture, Nanjing Agricultural University, Nanjing, China, 2 School of Earth and Environment, The UWA Institute of Agriculture, The University of Western Australia, Perth, WA, Australia

The MADS-box gene family is an important transcription factor (TF) family that is involved in various aspects of plant growth and development, especially flowering time and floral organogenesis. Although it has been reported in many plant species, the systematic identification and characterization of MADS-box TF family is still limited in radish (Raphanus sativus L.). In the present study, a comprehensive analysis of MADS-box genes was performed, and a total of 144 MADS-box family members were identified from the whole radish genome. Meanwhile, a detailed list of MADS-box genes from other 28 plant species was also investigated. Through the phylogenetic analysis between radish and Arabidopsis thaliana, all the RsMADS genes were classified into two groups including 68 type I (31 Mα, 12 Mβ and 25Mγ) and 76 type II (70 MIKC and 6 MIKC∗). Among them, 41 (28.47%) RsMADS genes were located in nine linkage groups of radish from R1 to R9. Moreover, the homologous MADS-box gene pairs were identified among radish, A. thaliana, Chinese cabbage and rice. Additionally, the expression profiles of RsMADS genes were systematically investigated in different tissues and growth stages. Furthermore, quantitative real-time PCR analysis was employed to validate expression patterns of some crucial RsMADS genes. These results could provide a valuable resource to explore the potential functions of RsMADS genes in radish, and facilitate dissecting MADS-box gene-mediated molecular mechanisms underlying flowering and floral organogenesis in root vegetable crops.

Keywords: radish, MADS-box genes, flowering, floral organogenesis, RT-qPCR

INTRODUCTION

MADS-box genes, which were primarily identified as floral homeotic genes, encode a family of transcription factors (TFs) containing a highly conserved MADS domain of approximately 60-amino-acid sequences in the N-terminal region (Norman et al., 1988), which bind to (CC[A/T]6GG) that is also known as CArG boxes (Pellegrini et al., 1995; Shore and Sharrocks, 1995; Sasaki et al., 2010). Based on phylogenetic relationships, MADS-box genes have been
classified into two broad groups, type I and type II genes, which were generated by single gene duplication (Alvarez-Buylla et al., 2000; Liu et al., 2013). Among them, type I proteins are further divided into three subgroups including Ma, Mb and My, while type II can be classified into the subgroups MIKC\(^\ast\) and MIKC\(^\ast\) according to the sequence divergence at I domain (De Bodt et al., 2003; Wells et al., 2015). The MIKC\(^\ast\) type proteins have a longer I domain and a less conserved K domain than the MIKC\(^\ast\) type (Henschel et al., 2002; Gramzow and Theißen, 2013). Previous reports revealed the type I MADS-box genes encode SRF-like domain proteins, while type II genes encode MEF2-like proteins and MIKC-type proteins (De Bodt et al., 2003; Wells et al., 2015). Intriguingly, the most well-known MADS-box proteins belong to MIKC-type proteins which contains four common domains including MADS (M), Intervening (I), Keratin (K) and the C-terminal (C) domain (Kaufmann et al., 2005). Compared with type II, the type I proteins lack the K domain and show a relatively simple gene structure that usually only have one or two exons (Smaczniak et al., 2012; Kaufmann et al., 2005). At present, 62 Type I and 46 Type II genes have been identified and characterized in \(A.\) thaliana (Parenicová et al., 2003). Among the 46 Type II genes, 39 MIKC\(^\ast\) type genes were further classified into 12 groups based on their phylogenetic relationships, nevertheless, there were only seven genes belonging to the MIKC\(^\ast\) type (Duan et al., 2015).

In plants, increasing evidences from genetic and molecular analyses have revealed that MADS-box genes could play critical roles in regulating diverse developmental processes, such as flower organogenesis (Zahn et al., 2006), determination of flowering time (Moon et al., 2003; Adamczyk et al., 2007; Lee et al., 2007; Liu et al., 2008; Hu et al., 2014), regulation of fruit ripening (Liljegren et al., 2000), development of vegetative organs (Tapia-López et al., 2008), seed pigmentation and embryo development (Nesi et al., 2002). MIKC\(^\ast\)-type MADS-box genes play fundamental roles especially in flowering time control and floral organ identity. Based on the proposed ABC model (Haughn and Somerville, 1988), the ABCDE model that determines the identity of floral organs has been presented. Different floral organs identities are controlled by various combinations of types of genes, sepalas (A+E), petals (A+B+E), stamens (B+C+E), carpels (C+E) and ovules (D+E) (Zahn et al., 2006). A series of correlation functional genes were found to be involved in this process, such as Class A, APETALA1 (AP1); Class B, PISTILATA (PI) and AP3; Class C, AGAMOUS (AG); Class D, SEEDSTICK (STK); and Class E, SEPALLATA (SEP1, SEP2, SEP3 and SEPF4) (Parenicová et al., 2003).

In recent decades, several crucial MIKC\(^\ast\)-type genes have been suggested to modulate flowering time in \(A.\) thaliana. For instance, FLOWERING LOCUS C (FLC) gene has been found to inhibit flowering by encoding a specific MADS domain protein (Michaels and Amasino, 1999). SUPPRESSOR OF OVEREXPRESSION OF CO1 (SOC1) gene plays a critical role in vernalization and gibberellin signal integration for flowering (Moon et al., 2003). SHORT VEGETATIVE PHASE (SVP) is considered as an important control factor of flowering time by ambient temperature (Lee et al., 2007). Moreover, AGAMOUS-LIKE16 (AGL16) gene targeted by miR824 contributes to the repression of plant flowering time (Hu et al., 2014). In addition, several other MIKC\(^\ast\)-type genes have also been proven to be involved in flowering time, such as AGAMOUS-LIKE 24 (AGL24) (Liu et al., 2008), MADS AFFECTING FLOWERING (MAFI/FLM) (Ratcliffe et al., 2003) and AGAMOUS-LIKE 15/18 (AGL15/AGL18) (Adamczyk et al., 2007). More intriguingly, compared to MIKC\(^\ast\)-type genes, relatively less study has been conducted on the functions of MIKC\(^\ast\)-type and Type I genes. To date, MIKC\(^\ast\)-type and Type I genes only have been shown to participate in the \(A.\) thaliana male and female gametophyte, respectively (Zobell et al., 2010; Masiere et al., 2011). Furthermore, recent studies have revealed that Type I genes are primarily expressed in developing seed of \(A.\) thaliana (Barker and Ashton, 2013).

Radish (\textit{Raphanus sativus} L., 2\(n = 2x = 18\)) is an important root vegetable crop of Brassicaceae family worldwide (Xu et al., 2013). In the complete life cycle of radish, bolting and flowering are some of the critical factors which affect the yield and quality. Premature bolting seriously decreases the production of vegetable crops which ultimately lead to the reduction of economic benefits (Nie et al., 2015). Consequently, it is extremely essential to explore the MADS-box gene family whose primary function is to regulate flowering time and floral organ development. Recently, genome-wide identification and characterization of MADS-box genes were reported in some plant species including \(A.\) thaliana (Parenicová et al., 2003), rice (Arora et al., 2007), Chinese cabbage (Duan et al., 2015), cucumber (Hu and Liu, 2012), soybean (Fan et al., 2013) and maize (Zhao et al., 2011). However, the genome-wide analysis and characterization of MADS-box genes in radish remain lacking. Especially, it is ambiguous how MADS-box genes control flowering time and floral organ development in radish. Fortunately, the completion of the radish genome sequencing makes it possible to analyze MADS-box genes (Mitsui et al., 2015). In the present study, MADS-box members from radish genome were firstly identified and divided into different classes, and the gene structures, conserved motifs and phylogenetic relationships between these members were systematically analyzed. Additionally, linkage group locations and primary prediction of gene functions were also investigated, and the expression patterns of all MIKC\(^\ast\) genes in radish were carried out with RT-qPCR. These results would greatly contribute to gain insight into functional analysis of MADS-box genes and facilitate dissecting MADS-box gene-mediated molecular mechanisms underlying flowering and floral organogenesis in radish and other root vegetable crops.

**MATERIALS AND METHODS**

**Identification of MADS-Box Genes**

All radish genome sequences used to identify the MADS-box genes were available from the NODAI Radish genome database\(^1\)

\(^1\)http://www.nodai-genome-d.org/
(Mitsui et al., 2015). To confirm the candidates of radish MADS-box genes, the proteins with SRF-TF domain (Pfam accession number:PF00319) were searched against the genome protein sequences using HMM search tool with an E-value cut-off 1.0 (Finn et al., 2011; Finn et al., 2015). Each sequence predicted was subsequently verified through the public databases including NCBI1, Pfam and SMART4 to confirm its reliability (Letunic et al., 2012).

**Sequence Collection from Various Plant Species**

The MADS-box protein sequences of *A. thaliana* and *Brassica* annuum and *A. thaliana* were downloaded from TAIR database5 (Finn et al., 2011; Finn et al., 2015). Each sequence predicted number:PF00319) were searched against the genome protein sequences were retrieved from pepper genome platform6 (Kim et al., 2014) and Brassica database, respectively. The genome data of Beta vulgaris, Fragaria vesca, Phaseolus vulgaris, Ricinus communis, Brachypodium distachyon, Setaria italica, Amborella trichopoda and Chlamydomonas reinhardtii were downloaded from the genome browser phytozome5. All these collected genome sequences were used to screen MADS-box genes from various plant species through the Pfam database. All the sequences of the other species used in this study were collected from previous reports (Parenicova et al., 2003; Leseberg et al., 2006; Arora et al., 2007; Diaz-Riquelme et al., 2009; Zhao et al., 2011; Gramzow et al., 2012; Barker and Ashton, 2013; Fan et al., 2013; Duan et al., 2015).

**Linkage Group Localization and Identification of Orthologous and Paralogous Genes**

The sequences of RsMADS genes were searched against the genomic sequences of the scaffolds that were anchored to the integrated high-density linkage map (Kitashiba et al., 2014). The gene sequences with identity ≥98% and length difference ≤5 bp were considered to be the same genes between the two genomes, and localized to the linkage groups according to the corresponding location parameters using MapInspect Software4.

To gain insight into the homology relationship between MADS-box genes of radish and other species, we investigated the orthologous and paralogous MADS-box genes in radish, *A. thaliana*, Chinese cabbage and rice using OrthoMCL program19 (Li et al., 2003). Subsequently, the relationship networks of homologous genes in radish and *A. thaliana* was visualized using Cytoscape software (Shannon et al., 2003).

**Identification of Protein Properties, Gene Structure and Conserved Motifs and Phylogenetic Analysis**

Protparam tool of ExPasy11 was employed to analyze series of RsMADS protein properties like molecular weight, theoretical pI and instability index. The Pfam database and SMART were employed to determine conserved domains of proteins. After that, the GSDS12 was adopted to reveal intron-exon structure of RsMADS genes. Conserved motifs were identified using Motif Elicitation (MEME) software13, and the parameters settings as follows: (1) 10 ≤ optimum motif width ≤100; and (2) maximum number of motifs = 15. In addition, multiple alignments of MADS-box gene sequences were performed using ClustalX 2.0 with default parameters. MEGA 5.1 ( Tamura et al., 2011) was then used to construct the phylogenetic tree based on neighbor-joining (NJ) method, and bootstrap values were set to 1,000 replications.

**Prediction of miRNAs Targeting the RsMADS Genes**

To identify potential miRNAs targeting the RsMADS genes, all RsMADS genes were searched against a comprehensive miRNA library on psRNATarget Server14 with default parameters (Dai and Zhao, 2011), which was constructed according to the previously established five miRNA libraries (Xu et al., 2013; Nie et al., 2015; Sun X. et al., 2015; Sun Y. et al., 2015; Yu et al., 2015). After that, Cytoscape software was utilized to visualize the targeted relationship between predicted miRNA and corresponding RsMADS genes.

**Expression Analysis Using Radish RNA-seq Data**

Illumina RNA sequencing showed gene expression of radish varied in the different tissues and developmental stages (Mitsui et al., 2015). In this study, the Illumina RNA-Seq data, which were downloaded from NODAI radish genome database, were used for the transcriptional profiling of RsMADS genes in five tissues (cortical, cambium, xylem, root tip and leaf) and six stages of leaf [7, 14, 20, 40, 60 and 90 days after sowing(DAS)]. The expression level for each RsMADS gene was presented by the RPKM (Reads Per kb per Million reads) method (Mitsui et al., 2015). Lastly, heat maps were generated by Cluster 3.015 (de Hoon et al., 2004) and Tree View16 (Saldanha, 2004).

**Plant Material and Treatments**

The radish advanced inbred line, 'NAU-DY13,' was used in the current study. Germinated seeds were vernalized and sown in plastic pots and cultivated in controlled-environment growth chamber with day/night temperature of 28/18°C. For

---

1http://pfam.xfam.org/
3http://smart.embl-heidelberg.de/
4http://www.arabidopsis.org/
5http://brassicadb.org/brad/
6http://peppergenome.snu.ac.kr/
7http://phytozome.net/
8http://mapinspect.software.informer.com/
9http://www.orthomcl.org/cgi-bin/OrthoMcDWeb.cgi
10http://meme-suite.org/tools/meme
11http://expasy.org/protparam/
12http://gsds.cbi.pku.edu.cn/
13http://plantgrn.noble.org/psRNATarget/
14http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm
15http://jtreeview.sourceforge.net/
16http://treeview.sourceforge.net/
vernalization treatments, germinated seeds were vernalized at 2–4°C for 0, 10 and 30 days, respectively, and grew under the middle-day (12 h light/12 h dark). For photoperiodic treatments, unvernalized seedlings were cultured under long-day (16 h light/8 h dark) and short-day (8 h light/16 h dark) treatments, respectively. Furthermore, the rest of unvernalized seedlings were treated with 200 mg/L and 800 mg/L GA3 every other day for a week from 2-week-old seedlings under the middle-day condition. Unvernalized seedlings without any treatment grown under the middle-day were set as control (CK). Leaf samples were collected when treated seedlings were grown to three weeks old. Different flower tissues from control plants, including sepal, petal, stamen and carpel, were collected separately at reproductive stage. All the samples were collected from three randomly selected individuals and immediately frozen in liquid nitrogen and stored at −80°C for further use.

RNA Isolation and RT-qPCR
Total RNA of each sample was isolated using Trizol reagent according to the manufacturer’s instructions (Invitrogen, Carlsbad, CA, USA). Then, the first-strand cDNA was synthesized using the Superscript III First-Strand Synthesis System (Invitrogen). The specific primers of RsMADS genes for RT-qPCR were designed using Beacon Designer 7.7 (Premier Biosoft International, Palo Alto, CA, USA). To confirm results reliability, three biological and three technical replicates were adopted. RT-qPCR reaction system and cycling profile were carried out on Bio-Rad iQ5 Real-Time PCR System. RsActin gene was selected as the reference gene (Xu et al., 2013). The primers used for RT-qPCR were shown in Supplementary Table S1. Finally, formula $-\Delta\Delta C_t$ and $2^{-\Delta C_t}$ were applied to calculate the relative expression ratio. The data were statistically analyzed with Duncan’s multiple range test at the $P < 0.05$ level of significance using SPSS 20 software (SPSS Inc., USA).

RESULTS
Identification and Analysis of MADS-Box Proteins in Radish
To define the candidate MADS-box proteins in radish, a profile hidden Markov model (HMM) search against NODAI radish genome protein sequences was carried out using the SRF-TF domain (PF00319), and totally 157 putative MADS-box protein genes were obtained. The low quality sequences without start and/or stop codons were removed to ensure the reliability of these sequences, and finally a total of 146 sequences were retained. Subsequently, all remaining sequences were verified through the public databases including NCBI, Pfam and SMART. All these radish MADS-box protein genes were named as RsMADS001 to RsMADS146, respectively (Supplementary Table S2). After searching these protein sequences against A. thaliana on TAIR database by BLASTp, RsMADS040 and RsMADS091 were removed, because they contained other functional domains and their homologous proteins were non-MADS-box proteins (Supplementary Figure S1). To study the comparative evolution among various plant species, MADS-box genes from 28 other plant species were also collected by searching for SRF-TF domain (PF00319) in their genomes (Figure 1A; Supplementary Tables S3 and S4). Compared with other species, radish had a relatively large MADS-box gene family of 144 members, and the members of MADS-box gene family subgroups were also identified (Figure 1A).

Comparative Phylogenetic Analysis of RsMADS Genes
To better understand the phylogenetic relationships of the MADS-box genes between radish and A. thaliana, the classification of 108 MADS-box genes from A. thaliana was performed (Supplementary Figure S2A). An unrooted phylogenetic tree of MADS-box genes between radish and A. thaliana was constructed by the NJ method (Supplementary Figure S2B). It is quite obvious that RsMADS genes were divided into five clades according to the classification of the A. thaliana, namely subfamilies MIKCc (70), MIKC* (6), Mα (31), Mβ (12) and Mγ (25) (Supplementary Figures S2B,D). Additionally, an unrooted phylogenetic tree was produced using MADS-box proteins from radish, A. thaliana and Chinese cabbage to further confirm the phylogenetic relationships and classification of RsMADS proteins (Supplementary Figure S2C).

Phylogenetic trees for type I and type II MADS-box genes were separately generated using A. thaliana and radish proteins (Figures 2A,B). Totally 49% (70) of the 144 RsMADS genes belongs to the MIKCc-type genes, which could be further divided into 12 subfamilies (Figures 1B and 2A). Subgroup SOCl and subgroup Bs, respectively, showed the largest (~21%) and smallest (~3%) number of RsMADS genes. However, in A. thaliana the largest and smallest proportion of subgroup is ~15% (subgroup SOCl or FLC) and ~3% (subgroup AGL12), respectively (Parenicová et al., 2003). The number of RsMADS genes from other subgroups ranged from four to 13, interestingly, subgroup C/D and E had seven members, while subgroup B, FLC and AGL15 consisted of five members (Figures 1B and 2A).

Linkage Group Localization and Orthologous Relationship Analysis
In total, 41 RsMADS genes (20 type I and 21 type II) accounting for 28.5% of the total MADS-box gene number were separately anchored onto the approximate location of linkage group (LG) R1-R9 (Figure 3; Supplementary Table S5). On the whole, the distribution of 41 RsMADS genes was relatively dispersed, but there were also some RsMADS gene cluster, for example, six genes clustered in the front of LG R2. Among the nine LGs, LG R4 contained the most RsMADS genes (9 members, ~22%), while LG R7 and R8 presented the least member (1 member, ~2%) (Figure 3). Moreover, there were two MIKC* -type genes that were successfully anchored on the LG R9.
In the present study, orthologous and paralogous MADS-box genes between radish and other three plant species (Chinese cabbage, A. thaliana and rice) were comparatively analyzed by OrthoMCL Software. Among all the MADS-box genes, 60 orthologous and 38 co-orthologous gene pairs were found between radish and A. thaliana. Nevertheless, only 19 orthologous and 22 co-orthologous gene pairs were detected between radish and rice, and 16 orthologous and 18 co-orthologous gene pairs were found between A. thaliana and rice (Figure 4; Supplementary Table S6). Furthermore, a relational graph was used to visualize all the relationships among the orthologous, co-orthologous and paralogous MADS-box genes between radish and A. thaliana (Supplementary Figure S3). 29 and 50 AtMADS genes were determined to have no and only one orthologous gene in radish, respectively. While, 68 and 70 paralogous MADS-box gene pairs were detected in radish and A. thaliana, respectively (Supplementary Figure S3).

Characterization of RsMADS Proteins, Conserved Motif Distribution and Intron–Exon Structure

To gain insight into the molecular characterization of 144 RsMADS proteins, their physical and chemical properties including molecular weights (MWs), theoretical isoelectric points (pI), instability index and aliphatic index were analyzed, and all RsMADS proteins were hydrophilic (Supplementary Table S2, Supplementary Figure S4). To analyze the features of RsMADS protein sequences, the 15 conserved motifs of 144 RsMADS proteins within the different groups were predicted by performing MEME motif search tool, and the LOGO of 15 amino acid motifs were generated (Supplementary Figure S5). Moreover, all of the RsMADS proteins contained motif 1 and motif 3, which indicated that this highly conserved domain was MADS domain. Nevertheless, motif 4 and motif 6 were present...
FIGURE 2 | Phylogenetic relationship of radish and A. thaliana type II (A) and type I (B) MADS-box proteins. Subgroups are marked in different colors.
Li et al. Characterization of Radish MADS-Box Gene Family

FIGURE 3 | Linkage group (LG) localization of RsMADS genes. (A) Distribution of the RsMADS genes on nine radish LGs. (B) The percentage of RsMADS gene anchored onto each LG.

in most of the MIKC-type genes, and thus were predicted to be K-box domain. In addition, 15 motifs were submitted to the Pfam and SMART website for further identification, and provided strong evidences supporting our predictions (Supplementary Figure S6). Moreover, according to previous reports (Saha et al., 2015; Rameneni et al., 2014; Shu et al., 2013) and the conservative characteristics of motifs, motif 8 was predicted to represent the I domain, and motif 10 and motif 14 specified the C-terminal domain. It should be emphasized that type I group had more distinct motifs at their C-terminal regions except the MADS domain, which were more divergent than those in the type II group, and these motifs were identified as unknown by Pfam and SMART (Supplementary Figure S6D). Motif analysis showed that the majority of RsMADS proteins in the same subgroup shared similar motif distribution, suggesting that the proteins from the same subgroup probably had similar functions (Parenicová et al., 2003; Song et al., 2015).

Additionally, the intron–exon patterns were analyzed to investigate the structural diversity of RsMADS genes. Comparison of genomic DNA and cDNA showed that type I RsMADS genes had no or only one intron except RsMADS133 containing five introns (Supplementary Figure S6C, Supplementary Table S2). Compared with type I, type II RsMADS genes had more complex structures. The intron numbers of type II RsMADS varied from 0 to 16 with an average of 5.6, and 60 (78.9%) members were consisted of at least five introns (Supplementary Figure S6C, Supplementary Table S2).

Analysis of miRNAs Targeting RsMADS Genes

To have a better understanding of the function of MADS-box gene family in radish, a comprehensive miRNA library consisting of five miRNA libraries reported from our previous studies was used to determine miRNAs targeting RsMADS genes by psRNATarget program. Totally, 19 known miRNAs and six potential novel miRNAs (named Rsa-miR1-Rsa-miR6) belonging to 25 miRNA families were identified as putative miRNAs which could target 25 RsMADS target transcripts (Supplementary Table S7). The regulatory relationship between putative miRNAs and their targets were presented in Supplementary Figure S7. RsMADS027 was the target transcript of miR8154, miR5293 and miR831-5p, RsMADS084 was targeted by miR5174e-5p, Rsa-miR4 and Rsa-miR3, while four transcripts (RsMADS087, RsMADS125, RsMADS138 and RsMADS140) were targeted by miR5021 (Supplementary Figure S7). It is worth noting that miR824 was predicted to target RsMADS020 and RsMADS044, whose sequences showed high similarity with AGL16 in A. thaliana (Supplementary Table S2).

Differential Expression Analysis of RsMADS Genes

To estimate the expression levels of RsMADS genes, RPKM of 144 RsMADS genes in leaves from seven different development
stages and in five different tissues was obtained (Mitsui et al., 2015). The results showed that the transcript abundances of different RsMADS genes were extremely diverse in radish (Figure 5; Supplementary Table S8). On the whole, almost all Type I RsMADS genes either maintained a relatively low transcriptional level or had no expression in RNA-Seq libraries except RsMADS093, RsMADS097, RsMADS106 and RsMADS111 (Figure 5B). The expression of RsMADS097 and RsMADS106 were downregulated in leaves with the development of radish. RsMADS093 and RsMADS111 have high expression levels in roots but were hardly expressed in the leaves, indicating that they perhaps were root-specific and play a vital role in root development (Figure 5B).

Compared with Type I RsMADS genes, Type II genes showed a higher expression level both in the roots and leaves except subgroup B and C/D. With the growth of radish, expression levels of some genes increased gradually, including RsMADS042, RsMADS050 (subgroup A); RsMADS001, RsMADS002, RsMADS010, RsMADS029 (subgroup SOC1); RsMADS135 (subgroup AGL15/18); RsMADS032, RsMADS065 (subgroup SVP), while others decreased such as RsMADS33 (subgroup E), RsMADS13 (subgroup AGL15/18), RsMADS44 (subgroup ANR1) and RsMADS36 (subgroup C/D) (Figure 5A). In leaves and roots, the ABCDE model genes showed low transcript levels, while SOC1, AGL15/18, ANR1, SVP and FLC categories had high expression levels (Figure 5A). Interestingly, some genes exhibited tissue-specific expression (Figure 5A).

For example, RsMADS074 (subgroup AGL13), RsMADS050 (subgroup A) and RsMADS089 (subgroup MIKC∗) were specifically expressed in leaves, whereas some genes such as RsMADS004, RsMADS008 (subgroup SOC1) and RsMADS043, RsMADS044 (subgroup ANR1), were specifically expressed in roots. In addition, RsMADS052, RsMADS053 and RsMADS077 (subgroup AGL12) displayed a high expression level in root tips (Figure 5A).

Expression Analysis of MIKC C Genes by RT-qPCR
To further reveal the function of 12 subgroups of radish MIKC C genes, the relative expression levels of genes in A, B,
genes were more diverse. The expression levels of E subgroup (AP1) exhibited high expression level in sepal and petal, as compared with that in stamen and carpel, whereas RsMADS057 (AP3) and RsMADS078 (PI) were significantly expressed in petal and stamen (Figure 6). More interestingly, RsMADS036 (AG) tended to be expressed in stamen and carpel, while RsMADS047 (STK) was significantly up-regulated only in carpel (Figure 6). Moreover, the expression patterns of E subgroup genes were more diverse. The expression levels of RsMADS017 (SEP1) and RsMADS023 (SEP2) were relatively steady in the four tissues, whereas RsMADS033 (SEP3) and RsMADS026 (SEP4) maintained relatively high expression levels in one (petal) and two (petal and carpel) specific tissues, respectively (Figure 6).

Additionally, it is apparent that three different treatments (GA, light and vernalization) resulted in a wide variety of expression profiles among RsMADS genes (Supplementary Figure S8). The orthologous RsMADS genes with A. thaliana ABCDE model genes were selected for further analysis. RsMADS68 (AP1) exhibited high expression level in sepal and petal, as compared with that in stamen and carpel, whereas RsMADS057 (AP3) and RsMADS078 (PI) were significantly expressed in petal and stamen (Figure 6). More interestingly, RsMADS036 (AG) tended to be expressed in stamen and carpel, while RsMADS047 (STK) was significantly up-regulated only in carpel (Figure 6). Moreover, the expression patterns of E subgroup genes were more diverse. The expression levels of RsMADS017 (SEP1) and RsMADS023 (SEP2) were relatively steady in the four tissues, whereas RsMADS033 (SEP3) and RsMADS026 (SEP4) maintained relatively high expression levels in one (petal) and two (petal and carpel) specific tissues, respectively (Figure 6).

Additionally, it is apparent that three different treatments (GA, light and vernalization) resulted in a wide variety of expression profiles among RsMADS genes (Supplementary Figure S8). The orthologous seven representative genes including RsMADS002, RsMADS012, RsMADS020, RsMADS021, RsMADS015, RsMADS064 and RsMADS065 were reported to be crucial in flowering control in A. thaliana. The results showed that RsMADS002 (SOCI) and RsMADS065 (AGL24) were up-regulated under different concentrations of GA treatments, but RsMADS015 (SVP), RsMADS020 (AGL16), RsMADS021 (AGL15) and RsMADS064 (FLC) were obviously down-regulated (Figure 7A). RsMADS015 (SVP) were down-regulated following the decrease of light lengths, whereas the transcript accumulation of RsMADS020 (AGL16) and RsMADS065 (AGL24) were relative lower at short-day (SD) and peaked at long-day (LD) (Figure 7B). Intriguingly, most members showed strong sensitivity toward vernalization treatment. Along with prolonging vernalization time, RsMADS002 (SOCI) and RsMADS065 (AGL24) were evidently induced, by contrast, the other five genes were inhibited inordinately (Figure 7C).

**The MADS-Box Gene-Mediated Regulation Associated with Flowering and Floral Organogenesis**

In the present study, 16 critical RsMADS genes including RsMADS002 (SOCI), RsMADS012 (AGL18), RsMADS015 (SVP), RsMADS020 (AGL16), RsMADS021 (AGL15), RsMADS064 (FLC), RsMADS065 (AGL24), RsMADS68 (AP1), RsMADS057 (AP3), RsMADS078 (PI), RsMADS036 (AG), RsMADS047 (STK), RsMADS017 (SEP1), RsMADS033 (SEP3), RsMADS023 (SEP2) and RsMADS026 (SEP4) were identified to be involved radish flowering and floral organ formation. According to the reported A. thaliana and radish flowering and floral organogenesis regulatory network (Posé et al., 2012; Sun et al., 2013; Zhou et al., 2012; Khan et al., 2014; Nie et al., 2016), some critical floral pathway integrator genes, such as FLC, SVP, AGL16, SOCI, AGL24 and AGL19, were considered to respond to environmental and endogenous factors directly or indirectly through interacting with other genes, and then these genes further regulate the expression of downstream floral organ identity genes including AP1, AG, AP3 and SEP3.

**DISCUSSION**

Higher plants routinely go through various phase transitions from germination to death mainly including juvenile phase, vegetative growth and reproductive development. The vegetative
phase change is essential for plants in response to environmental and endogenous factors, so as to complete their life cycle and achieve reproduction successfully (Khan et al., 2014). Flowering, as a symbol of plants into the reproductive growth phase, is determined by a complex gene interaction which is composed of a crowd of flowering and organogenesis related genes including most of MADS-box family genes (Smaczniak et al., 2012). In recent years, the bioinformatics analysis of gene families in different species facilitated the identification of various gene families with the completion of the genome sequencing. MADS-box gene families were identified and analyzed at a genome-wide scale in a series of plant species such as A. thaliana (Parenicová et al., 2003), rice (Arora et al., 2007), Chinese cabbage (Duan et al., 2015), cucumber (Hu and Liu, 2012), soybean (Fan et al., 2013) and maize (Zhao et al., 2011). However, it is still deficient for the genome-wide identification and analysis of MADS-box genes in radish. In the present study, a comprehensive analysis of MADS-box genes was performed and a total of 144 MADS-box family members were identified from the whole radish genome.

**Overview of MADS-Box Gene Family in Radish**

In the current study, apart from 18 species reported previously, MADS-box gene families from other 10 species were firstly identified (Supplementary Tables S3 and S4). As previously observed, it could be suggested that the number of MADS-box genes in Angiospermae was obviously larger than that in other species belonging to Algae, Bryophyta and Lycophytes (Figure 1A), indicating that a great expansion of MADS-box gene family members occurred after the angiosperm evolution (Gramzow et al., 2014; Duan et al., 2015). Simultaneously, the analysis of phylogenetic relationships between radish and other plant species, especially A. thaliana, provided a solid foundation for better understanding the function of RsMADS genes (Wang et al., 2015; Duan et al., 2015). In addition, the intron–exon structure feature has a potential influence on alternative splicing of gene to a certain extent, and the function of the protein will be affected (Tian et al., 2015). For type II RsMADS genes with more complex structures, it could be inferred that type II genes had

---

**FIGURE 7** The expression levels of representative RsMADS genes under GA (A), photoperiod (B) and vernalization (C) treatments. The subgroup is marked in different colors under the gene name. Each bar shows the mean ± SE of the triplicate assay. The value with different letter indicates significant difference at $P < 0.05$ according to Duncan's multiple range tests.
more variable and intricate function than type I genes, which was in accordance with previous results in *A. thaliana* (Parenicová et al., 2003), Chinese cabbage (Duan et al., 2015) and soybean (Fan et al., 2013).

Previous evidence has shown that known miR5227 and novel Rsa-miR4 played a role in the bolting and flowering process of radish by high-throughput sequencing technology (Nie et al., 2015). Therefore, their target genes RsMADS115 (AGL103) and RsMADS084 (AGL30) might be associated with regulation of bolting and flowering time in radish. Furthermore, previous observations confirmed that miR824-regulated AGL16 inhibited flowering in *A. thaliana* (Hu et al., 2014). In this study, miR824 was identified to target RsMADS020 and RsMADS044 (AGL16), revealing that these target genes may contribute to flowering time repression in radish.

**Characterization of Critical RsMADS Genes in Flowering and Floral Organ Development**

Biochemical and genetic studies have indicated that flowering and floral organogenesis can be modulated by MADS-box genes especially MIKC type in higher plants (Lee et al., 2013; O’Maoléidigh et al., 2014). Meanwhile, a growing number of key MADS-box genes including FLC, SOC1, SVP, AGL24, AGL16, AGL15 and AGL18, and ABCDE model genes involved in this process have been widely recognized (Ferrario et al., 2004; Khan et al., 2014).

Control of flowering time is an intricate genetic circuitry in response to various endogenous and exogenous cues (Wullschleger and Weston, 2012). In *A. thaliana*, molecular genetics and physiological studies revealed that five main pathways of vernalization, photoperiod gibberellin, autonomy and age controlled flowering time (Kim et al., 2009; Mutasa-Göttgens and Hedden, 2009; Srikanth and Schmid, 2011; Wang, 2014). In this study, expression profiles of seven representative genes were investigated and the results suggested that RsMADS015 (SVP), RsMADS020 (AGL16), RsMADS021 (AGL15) and RsMADS064 (FLC) might act as flowering repressor, while RsMADS002 (SOC1) and 065 (AGL24) contributed to the flowering promotion in radish (Figure 7).

Flower meristem and floral identity had been explained perfectly by five kinds of genetic function genes (A-B-C-D-E), which were important in regulating different flower whors from sepals to carpels (Ferrario et al., 2004; Li et al., 2015). In the present study, the regulatory relationships between ABCDE genes and floral organ development were analyzed in radish, and a schematic ABCDE model was proposed (Supplementary Figure S9). RNA-Seq and RT-qPCR analysis revealed the expression patterns of ABCDE model orthologous genes in different tissues and at different stages of flower development. These genes exhibited relatively low abundant transcripts in the leaf and root (Figure 5A) and a regular expression patterns in different flower whors (Figure 6; Supplementary Figure S8A), which were consistent with previous studies (Su et al., 2013; O’Maoléidigh et al., 2014; Xie et al., 2015), suggesting that ABCDE model genes worked in a combinatorial manner to regulate the floral morphogenesis in radish.

**The Roles of MADS-Box Genes in Flowering and Flower Formation in Radish**

Flowering is a coherent and sophisticated development process (Nie et al., 2015). Flowering-related genes were affected by multiple flowering signals converging on the regulation of floral organ identity genes including SEP3, AP1, AG and AP3, leading to flower formation eventually (Posé et al., 2012; Sun et al., 2013; Zhou et al., 2013; Khan et al., 2014). Considerable reports have indicated that FLC and SOC1 as floral pathway integrators which were regulated by numerous genes and flowering pathways, played important roles in the flowering process (Lee et al., 2007; Franks et al., 2015). Genetic studies showed that FLC could block the transcriptional activation of SOC1 and required SVP and FRI to delay flowering strongly (Helliwell et al., 2006; Lee et al., 2007; Geraldo et al., 2009; Franks et al., 2015). AGL16, a target gene of miR824, can help to repress flowering time by interacting indirectly with FLC and directly with SVP in *A. thaliana* (Hu et al., 2014). In this study, RT-qPCR validation showed that FLC, SVP and AGL16 orthologous genes were down-expressed with the increase of GA concentration, light length and vernalization time (Figure 7), indicating that they may be repressors of flowering in radish.

Moreover, two other critical MADS-box genes, SOC1 and AGL24, could promote flowering by responding to GA signaling (Moon et al., 2003). Additionally, AGL15 and AGL18 acted as the floral repressors via controlling the regulation of SOC1 and FT, and aGL15 aGL18 mutations presented a quick increase in SOC1 and FT levels, leading to early flowering (Adamczyk et al., 2007; Fernandez et al., 2015). In the current study, AGL24, SOC1, AGL15, and AGL18 orthologous genes were identified in radish, and RT-qPCR profiling showed that AGL24 and SOC1 orthologous genes were up-regulated, while AGL15 and AGL18 orthologous genes were obviously down-regulated when treated with different flowering-induced factors (Figure 7). These results suggested that AGL24 and SOC1 promoted flowering, whereas AGL15 and AGL18 inhibited flowering in radish. Therefore, it could be suggested that MADS-box gene family play a major role in regulating flowering time and floral meristem identity in radish.

**CONCLUSION**

In conclusion, a total of 144 genes encoding MADS-box TF including 68 type I and 76 type II genes were identified in the whole radish genome. Among them, 41 genes were localized on the nine linkage groups of radish. A comparative phylogenetic analysis of the MADS-box genes was carried out between radish and *A. thaliana* to classify the MADS-box proteins. Furthermore, identification of miRNAs targeting RsMADS transcripts shed a novel insight into the functions of
RsMADS genes at transcriptional and post-transcriptional level. In addition, RT-qPCR analysis provided a better understanding of critical functions of candidate RsMADS genes involved in flowering and floral organ identity in radish. Taken together, in this study, radish MADS-box gene family was comprehensively characterized, which facilitated dissecting RsMADS gene-mediated molecular mechanism underlying flowering and floral organogenesis in radish.

AUTHOR CONTRIBUTIONS
All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS
This work was in part supported by grants from the NSFC (31372064, 31501759, 31601766), National Key Technologies R&D Program of P. R China (2016YFD0100204-25), Key Technology R&D Program of Jiangsu Province (BE2016379), JASTIF(CX(16)2012) and the PAPD.

SUPPLEMENTARY MATERIAL
The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2016.01390

REFERENCES

FIGURE S4 | The physical and chemical properties of RsMADS proteins. (A) The distribution of putative isoelectric points and molecular weights. (B) The distribution of Grand Average of hydropathicity (GRAVY), instability index and aliphatic index.

FIGURE S5 | Sequence logos of MADS domains in radish. The overall height of the stack indicates the level of sequence conservation. The height of residues within the stack indicates the relative frequency of each residue at that position.

FIGURE S6 | The analysis of RsMADS proteins and RsMADS genes structure. (A) The classification of RsMADS genes. (B) Phylogenetic tree of RsMADS proteins. (C) Intron-exon structure distribution of 144 RsMADS genes. (D) Conserved motif distribution of 144 RsMADS proteins.

FIGURE S7 | Predicted targeted regulatory network between RsMADS genes and miRNAs.

FIGURE S8 | Expression analysis of RsMADS at different flower whors and different treatments. Heat map representation and hierarchical clustering of RsMADS genes during sepalas, petals, stamens, carpels and ovules (A), and under vernalization, photoperiod and GA treatments (B). The scale represents relative expression value. The subgroup is marked in different color on the right side of the gene list.

FIGURE S9 | Putative schematic ABCDE model of floral organ development in radish.

TABLE S1 | The primer sequences used for RT-qPCR of actin and MIKC genes.

TABLE S2 | The information of 144 RsMADS genes in radish.

TABLE S3 | The Information of MADS-box genes of various species from the previous reports.

TABLE S4 | The Information of MADS-box genes from different species.

TABLE S5 | The information of linkage group localization of RsMADS genes.

TABLE S6 | The orthologous gene pairs and co-orthologous gene pairs in MADS-box proteins of radish, Arabidopsis, Chinese cabbage and rice, and the paralogous gene pairs among these species.

TABLE S7 | The information of miRNAs targeting RsMADS genes identified from previous five libraries.

TABLE S8 | The RPMK values of RsMADS genes.
Hu, J. Y., Zhou, Y., He, F., Coupland, G., et al. (2014). miR824


Evolution and Expression Patterns of CYC/TB1 Genes in Anacyclus: Phylogenetic Insights for Floral Symmetry Genes in Asteraceae

Maria A. Bello¹, Pilar Cubas², Inés Álvarez¹, Guillermo Sanjuanbenito¹ and Javier Fuertes-Aguilar¹*  
¹ Plant Evolutionary Biology Group, Real Jardín Botánico (CSIC), Madrid, Spain, ² Department of Plant Molecular Genetics, Centro Nacional de Biotecnología, CSIC-Universidad Autónoma de Madrid, Madrid, Spain

Homologs of the CYC/TB1 gene family have been independently recruited many times across the eudicots to control aspects of floral symmetry. The family Asteraceae exhibits the largest known diversification in this gene paralog family accompanied by a parallel morphological floral richness in its specialized head-like inflorescence. In Asteraceae, whether or not CYC/TB1 gene floral symmetry function is preserved along organismic and gene lineages is unknown. In this study, we used phylogenetic, structural and expression analyses focused on the highly derived genus Anacyclus (tribe Anthemidae) to address this question. Phylogenetic reconstruction recovered eight main gene lineages present in Asteraceae: two from CYC1, four from CYC2 and two from CYC3-like genes. The species phylogeny was recovered in most of the gene lineages, allowing the delimitation of orthologous sets of CYC/TB1 genes in Asteraceae. Quantitative real-time PCR analysis indicated that in Anacyclus three of the four isolated ACYC2 genes are more highly expressed in ray flowers. The expression of the four AcCYC2 genes overlaps in several organs including the ligule of ray flowers, as well as in anthers and ovules throughout development.

Keywords: Anacyclus, Asteraceae, CYC2 diversification, CYCLOIDEA, CYC/TB1, floral symmetry

INTRODUCTION

Asteraceae is the largest family of vascular plants with more than 23,600 species distributed among 13 different lineages (Panero et al., 2014). By contrast, its closest relatives, Calyceraceae and Goodeniaceae, have around 54 and 404 species, respectively (Carolin, 2007; Pozner et al., 2012). The evolutionary success of Asteraceae is strongly associated with its head-like inflorescence or capitulum (Broholm et al., 2014). Variations in perianth morphology, symmetry and sexuality of the flowers along the radial axis of this inflorescence have been used to discriminate different types of capitula (i.e., bilabiate, ligulate, radiate, discoid, and disciform; Jeffrey, 1977; Bremer, 1994). This diversity has important consequences, such as differential attractiveness to pollinators, increased rate of outcrossing and fitness by the presence of peripheral ray flowers (e.g., Senecio; Chapman and Abbott, 2009), or differential rates of seed germination across the capitulum (e.g., Anacyclus; Torices et al., 2013).
Floral symmetry is one of the most striking morphological variations within the capitulum. The evolution of floral symmetry in angiosperms is controlled by a restricted set of gene families (Luo et al., 1996; Costa et al., 2005; Broholm et al., 2008; Hileman, 2014). Among these, the TCP factor CYCLOIDEA (CYC) (Luo et al., 1996) has been recruited for independent transitions but then further modified or lost in conjunction with reversals to radial symmetry. The TCP gene family (Cubas et al., 1999) has expanded 600–800 million years ago, forming the class I and class II TCP subfamilies (Navaud et al., 2007). The TCP class II subfamily is formed by the CYCLOIDEA/TEOSINTE BRANCHED1 (CYC/TB1) and CINCINNATA (CIN) clades (Martin-Trillo and Cubas, 2010). The CYC/TB1 or ECE clade is specific to angiosperms and underwent duplication before the core eudicot diversification, producing three main groups: CYC1, CYC2, and CYC3 (Howarth and Donoghue, 2006; Citerne, et al., 2013). CYC3, and particularly CYC1 genes, have a major role in the regulation of axillary bud outgrowth (Aguilar-Martinez et al., 2007; Finlayson, 2007), whereas CYC2 genes control the growth patterns of flower meristems and regulate the establishment of bilateral symmetry of flowers in asterids and rosids (Busch and Zachgo, 2009; Preston et al., 2011; Nicolas and Cubas, 2015).

Investigation of Asteraceae CYC/TB1 gene family evolution performed in Gerbera (Mutisieae), Helianthus (Heliantheae), and Senecio (Senecioneae) revealed that CYC2 genes play a critical role in controlling ray flower identity (Broholm et al., 2008, 2014; Chapman et al., 2008; Kim et al., 2008; Juntheikli-Palovaara et al., 2014), and changes in temporal and spatial expression of CYC2 genes are associated with modifications of the floral symmetry pattern (Busch and Zachgo, 2007; Gao et al., 2008; Zhou et al., 2008). The acquisition of bilateral flowers in Asteraceae involved parallel recruitment of the diversified CYC2 genes, according to recent phylogenetic reconstruction of CYC/TB1 genes (Chapman et al., 2008, 2012; Tähtiharju et al., 2012).

We characterized the CYC/TB1 genes of the Mediterranean Anacyclus (Anthemideae), inferred their phylogenetic relationships and expression patterns, and compared them with the identified models in Gerbera, Helianthus and Senecio in order to test if the bilateral symmetry evolved early in the family due to cooption of CYC2 clade genes. Anacyclus L. with homogamous (i.e., all flowers bisexual, tubular and pentameros flowers) and heterogamous capitula (with modified female, bilateral and trimerous ray flowers surrounding the bisexual, tubular and pentameros disc flowers) (Bello et al., 2013) is a suitable model with which to carry out this comparative analysis because it is mostly annual, diploid (2n = 18, x = 9; Ehrendorfer et al., 1977; Humphries, 1981) and is suitable for ex situ cultivation. Also, the floral development of Anacyclus, Gerbera and Helianthus is similar, excepting the delayed development of the ray flowers relative to the disc flowers in Anacyclus, the exclusive presence of transitional flowers in Gerbera, the sterile condition of the ray flowers in Helianthus and the late zygomorphy of the disc flowers in Anacyclus (Laitinen et al., 2006; Tähtiharju et al., 2012; Bello et al., 2013). Moreover, unusual heterogamous capitula with peripheral "trumpet" flowers in natural populations of A. clavatus (Desf.) Pers. and A. valentinus L. documented from southern Spain (Bello et al., 2013) are very similar to tub (Berti et al., 2005) and turf (Chapman et al., 2012) mutant individuals of Helianthus. These "trumpet" flowers differ from the typical ray flowers in their tubular five-lobed perianth with radial symmetry and the labile presence of stamens (Bello et al., 2013).

As Anacyclus represents the derived and highly diversified tribe Anthemideae, where at least 20 genera have species presenting inflorescences with and without ray flowers, we extended the phylogenetic range of the CYC/TB1 studies. Although there are previous partial Asteraceae CYC/TB1 phylogenetic reconstructions (Chapman et al., 2008, 2012; Kim et al., 2008; Tähtiharju et al., 2012), a framework to visualize the entire diversification scenario in the family based on nucleotide variation is lacking. With the inclusion of several eudicot CYC/TB1 sequences in our analysis, together with those available from Asteraceae and the isolated CYC-like genes from Anacyclus, we have reconstructed a wider lineage profile and propose it as a model system for the classification and identification of paralogous and orthologous groups of CYC/TB1 genes in Asteraceae. Having this phylogenetic framework, we have explored if CYC/TB1 diversification involved positive selection, differential rates of evolution or differential expression patterns of the paralogs in Anacyclus.

MATERIALS AND METHODS

Plant Materials

Seeds and entire plants of wild Anacyclus clavatus (IA 2006, Soto del Real, Madrid), A. valentinus (LM 4435, Altea, Valencia), Matricaria aurea (IA 1996, El Retiro, Madrid) and Matricaria chamomilla (IA 1996, living collection Real Jardín Botánico, Madrid) were collected in 2008–2009 and treated as indicated in Bello et al. (2013). A. clavatus with trumpet phenotypes were obtained from seeds collected in May 2012 (one population from Carchuna, Granada), sowed in November 2010 and harvested in May 2013.

Cycloidea Gene Analysis

CYC-like genes of Anacyclus and Matricaria species were amplified from genomic DNA and cDNA with previously published (Chapman et al., 2008) and own-designed primers (Table S1). The DNeasy® and RNeasy® Plant Mini kits from Qiagen® were used for DNA and RNA extraction, respectively. RNA extraction of individual plant tissues was performed after their dissection, fixation and disruption in liquid nitrogen. RNA concentration was measured by spectrophotometry (NanoDrop 1000 v3.7, Thermo Fisher Scientific Inc.) and adjusted among tissues. cDNA synthesis was performed with the InVitrogen™ ThermoScript™ RT-PCR system kit. Semi quantitative RT-PCR of CYC-like amplicons from young roots (10-cm-long), leaves (2-cm-long), peduncules (1-cm-long), capitula (ca. 1 cm diameter), inflorescence bracts (ca. 0.5-cm-long), young/full expanded ray flowers (1–2.5-cm-long) and young/mature disc flowers (0.5–1.5-cm-long) of A. clavatus was carried out three times using different CYC and actin primer sets (Table S1). PCR amplification was performed with Ready-To-Go PCR beads (Illustra™) using a general program as follows: 95°C/5 min followed by 35 cycles
of 95°C/30 s, annealing temperature for 30 s and 72°C/45 s, and a final extension of 72°C/7 min (annealing temperatures are listed in Table S1). Amplified sequences were cloned using the Promega pGEM®-T Easy vector system (JM109 competent cells) and sequenced on a 3730 DNA Analyzer (Center for Research Support CAI, Universidad Complutense, Madrid) and an ABI 3700 instrument (STAB VIDA DNA sequencing service, Oeiras, Portugal). Inverse PCR and RACE techniques were used to amplify longer sequences from Anacyclus. For inverse PCR, DNA from A. clavatus and A. valentinus was extracted (Doyle and Doyle, 1987) and digested with 1.5 µL of restriction enzyme (BamHI, PstI, EcoRI, HindIII, XhoI, NcoI; New England Biolabs® Inc.) in a reaction containing 1 µL of DNA, 2 µL of buffer, and 15.5 µL of water. After 3 h of incubation (37°C) and 10 min of enzyme deactivation (65°C), the reactions were diluted (280 µL water) and ligation was performed using 26 µL of DNA, 3 µL of ligase buffer and 1 µL of ligation. Ligation products were amplified with specific primer pairs (Table S2) and cloned for sequencing. For 3′ RACE amplification of CYC genes from young capitula of A. clavatus, the SMARTer® RACE cDNA Amplification Kit and Advantage® 2 PCR Taq polymerase (Clontech Laboratories, Inc.) were used with the specific primer Ha2c_11 (Table S1).

Phylogenetic Analysis
Selected clones isolated from Anacyclus (79) and Matricaria (16) were aligned with other CYC-like genes from other species of Asteraceae (44) and other eudicots (54) including Calyceraceae and Goodeniaceae (Table S3). Initial alignments were performed with Geneious Pro 5.5.5 (Biomatters, http://www.geneious.com/; Kearse et al., 2012) using the default options of Geneious, MUSCLE and ClustalW. Nucleotides were aligned considering the codon arrangement in the amino acid alignment. Nucleotide (4) and amino acid (1) matrices were assembled (Table S4). The models of evolution for CYC amino acid (JTT + I + G, −lnL = 11,683.44) and nucleotide matrices (GTT + I + G, −lnL = 22,128.4089) were estimated using ProtTest 3 (Darriba et al., 2011), jModelTest 2 (Darriba et al., 2012) and Modeltest V 3.8 (Posada and Crandall, 1998). For Bayesian analyses (Ronquist et al., 2012), the matrices were analyzed with Mr Bayes 3.2.2 on XSEDE as implemented in the CIPRES Science Gateway (http://www.phylo.org/sub_sections/portal; Miller et al., 2010). For the amino acid matrix, the model was set to fixed (Jones), the rates to gamma distribution, the number of generations to 5,000,000, the number of chains to 1, and the sample frequency to 2,000. The nucleotide analyses were performed with and without removal of the third codon. In two other data sets, ambiguously aligned positions were removed and CYC2 genes from Asterales were analyzed independently (Table S4). For these analyses, the number of chains was set to four with 15,000,000 generations, a sample frequency of 2,000 and a diagnostic frequency of 5,000. The selected outgroup for all analyses was ActBBl (Acorus calamus) except the CYC2 genes from Asterales dataset that used SLCYC1 (Solanum lycopersicum) as outgroup. In all cases, the post-burn in trees were selected after discarding 25% of the trees. Final topologies were visualized with figTree v1.1.2 (http://tree.bio.ed.ac.uk/software/figtree). Mapping of the CYC amino acids along the trees was done with Mesquite 3.0 (Maddison and Maddison, 2011) using parsimony optimization. The amino acids were tracked on a post burn-in tree (tree 4,800,000) resulting from the amino acid Bayesian analysis. Maximum likelihood (ML) analysis was performed with GARLI 2.1 (Bazinet et al., 2014) and the bootstrap support was estimated with a 1,000 replicate search in Bootstrap RAxML (Stamatakis et al., 2008) in the CIPRES portal.

Diversifying/Purifying Selection of CYC Genes
Recombination Detection Program RDP v4.36 (Heath et al., 2006) was used to identify potential cases of recombination that could affect the estimate for selection, implementing the RDP (Martin and Rybicki, 2000) and MaxChi (Smith, 1992) methods. To detect individual sites subject to episodic diversifying selection, the CYC nucleotide matrix was analyzed under the mixed effects model of evolution (MEME) and the fixed effects likelihood approaches (FEL) in Datamonkey (Delport et al., 2010). In MEME, the distribution of the rate ω varies from site to site and also from branch to branch in a site, capturing the footprints of episodic and pervasive positive selection, whereas in FEL the synonymous and non-synonymous rates are fitted at each site with no variation along branches (Kosakovsky Pond et al., 2005; Murrell et al., 2012).

Diversification Rates Analysis
We used BAMM (Rabosky et al., 2014) to estimate rates of diversification across different gene lineages across the CYC phylogeny. The general model in this Bayesian method assumes that phylogenetic trees may have been shaped by a heterogeneous mixture of different evolutionary rates of gene diversification and extinction. Our working hypothesis was that, given a balanced sampling across paralogs of CYC2 in Asteraeae, which exhibit a number of paralogs larger than in other eudicot families, we could detect a significant heterogeneity across branches of arising paralogs in Anthemideae. We allowed each regime to be characterized by a distinct time-varying speciation process, where the diversification rate varies exponentially through time. The model of exponential change has been used in taxon diversification studies and is also expected as an approximation to diversity-dependent changes in gene diversification rates through time (Rabosky, 2014). We accounted for incomplete taxon sampling using the analytical correction implemented in BAMM, assuming that our sampling included 95% of extant Anacyclus CYC diversity. Visualization was performed using R scripts available through the R package BAMMTOOLS (Rabosky, 2014; Rabosky et al., 2014).

Expression Analysis by Quantitative PCR
Expression of the CYC genes was compared in wild rayed and “trumpet” inflorescences of A. clavatus using young plant tissues: roots (10-cm-long), leaves (2-cm-long), capitula stage 1 (<0.5 cm diameter), capitula stage 2 (>1 cm diameter), mature ligules (>3-cm-long, full expanded) and closed disc flowers (0.5-mm-long). Three different individuals of wild and "trumpet" A. clavatus were included in the analysis, as well as three
technical replicates of each tissue. RNA extraction was performed as described above. cDNA synthesis was carried out with the Transcriptor Universal cDNA Master kit® (Roche) using the following conditions: 25°C/5, 55°C/10, and 85°C/5 min. RNA concentration was adjusted among tissues, adding up to 15 ng/μL per reaction. The qPCR was run on a LightCycler 2.0 using 4 μL of Sensimix Capillary Kit, 0.2 μL of Sybr green and 0.75 μL of MyFiTM DNA Polymerase (Bioline) together with designed lineage-specific CYC primers (final concentration 0.2 μM, Table S5). Positive (genomic DNA) and negative (without nucleic acids) controls were included in each qPCR run to test the resultant crossing-point (Cp) values. Discarded Cp values included those higher than the Cp of the negative control, values above 35, and dissimilar melting temperatures compared against the positive control. Actin was used as the reference gene and was included those higher than the Cp of the negative control, values included those higher than the Cp of the negative control, values above 35, and dissimilar melting temperatures compared against the positive control. Actin was used as the reference gene and was amplified in all tissues. The primer efficiency (E) was calculated from the amplification of three replicates of the “capitula stage 1” tissue (Cp control) and the E target ΔCt/E control ΔCt ratio calculated (Pfaffl, 2001). For a graphical format of the CYC gene expression, the average of three tissue replicates of this ratio was calculated together with the standard deviation. The Kolmogorov-Smirnov test was used to evaluate the probability distribution of the expression average ration in trumpet and wild individuals (α = 0.05, Table S7). A t-test paired two sample for means was conducted to test if there is significant difference between the mean expression of AcCYC2 genes in trumpet and wild individuals (α = 0.05, Tables S8, S9).

**RNA In situ Hybridization**

A non-radioactive in situ protocol using RNA probes was followed using wild inflorescences of A. clavatus at different developmental stages. Capitula of different stages with a maximum of 1 cm of diameter were dissected under a Leica M165FC stereo microscope and fixed in 4% formaldehyde with 0.1% Tween-20 and 0.1% Triton X-100 (Jackson, 1991). Gene-specific probes for the CYC2 clade genes 2A, 2B, 2C, and 2D were amplified using the following general program: 95°C/5 min, 35 cycles with 95°C/30 s, annealing temperatures for 1 min, 72°C/2 min, and a final extension of 72°C/12 min (Table S10). A sense probe was amplified with the 2b set of primers and used in further analyses. Anti-sense and sense probes were tested with M13 primers in combination with the CYC2 primers. Probes were cloned into pBluescript II SK and linearized with BamHI before digoxigenin labeling (Coen et al., 1990), which was performed with anti-digoxigenin-AP, Fab fragments, T7 RNA polymerase and deoxynucleoside triphosphates (Roche). Tissue pretreatment, hybridization, washing and antibody staining steps followed Coen et al. (1990) and Fobert et al. (1994). The reaction to visualize the hybridized probes was incubated for 24-48 h at room temperature (~23°C). Sections were mounted with DePeX mounting medium (Serva) and observed and photographed using a Leica DMR microscope with an Olympus DP70 camera. Images were edited and organized in Adobe Photoshop CS4.

**RESULTS**

**Phylogenetic Reconstruction Reveals Eight CYC/TB1 Gene Lineages in Asteraceae**

The CYC/TB1 phylogeny reconstruction reveals new and previously identified gene lineages in Asteraceae (Figure 1A). We define a lineage as a group formed by genes from different Asteraceae species representing an orthologous gene set. Although the identified CYC/TB1 lineages are unevenly sampled for all the Asteraceae species, each of them is congruent with the species phylogeny (Figure 1). For tribe Anthemideae, we identified cDNAs from 10 putative CYC/TB1 genes (Figure 1) from A. clavatus (AcCYC1a, AcCYC1b, AcCYC2a, AcCYC2a1, AcCYC2b, AcCYC2c, AcCYC2c1, AcCYC2d, AcCYC3a, and AcCYC3b), three from A. valentinus (AvCYC2b, AcCYC2c1, and AvCYC2d), four from Matricaria aurea (MaCYC2a1, MaCYC2c, MaCYC2c1, and MACYC2d), and three from M. chamomilla (McCYC2b, McCYC2c1, and McCYC2d). We also identified several allelic variants for each gene (Table S1, Figure 1). The coding sequences (CDS) range between 783 and 900 bps. In the Bayesian and ML analyses reconstructed from the amino acid and nucleotide data, eight main lineages were recovered (Figure 1A): two formed by “CYC1-type” genes (CYC1a, CYC1b), four by CYC2 genes (CYC2a, CYC2b, CYC2c, CYC2d) and two by CYC3 genes (CYC3a, CYC3b). CYC2a1 and CYC2c1 are nested lineages in CYC2a and CYC2c, respectively. CYC2a1 represents a gene diversification (Figure 1) congruent with the genera phylogeny. The CYC2c1 genes display diversification of Anacyclus and Matricaria genes only (Figure 1). The CYC2a1 lineage involves Gerbera and Anacyclus genes, whereas in CYC2c, genes from different species display a diversification congruent with the species phylogeny (Figure 1A). CYC2b and CYC2d represent relatively well-sampled gene lineages congruent with the species phylogeny.

CYC2 and CYC3 genes form monophyletic groups, but their relationship with other clades is unresolved or poorly supported (Figure 1A, Table S11). By contrast, genes formerly identified as “CYC1” do not form a single clade. Some “CYC1-like” genes, including Solanaceae and the Asteraceae 1A, 1B lineages, are more closely related to CYC2 genes than to other “CYC1-like” eudicot loci (Figure S1, Tree V). In all topologies, the earliest divergent CYC2 genes are those from Adoxaceae, Brassicaceae, Caprifoliaceae, Gesneriaceae, Goodeniaceae, Plantaginaceae, Solanaceae, and Vitaceae (Figure 1). The remaining CYC2 members form a well-supported clade with genes from Asteraceae and Calyceraceae (Table S11). Hereafter, this clade will be called the Ast/Cal clade (Figure 1A). A close relationship between CYC2b and CYC2c is seen in all resultant trees excepting tree V from the GARLI analysis (Figure S1D). All CYC2 genes from Asteraceae are placed in the CYC2a–CYC2d lineages except some sequences from Dasyphyllum, Gerbera and Helianthus, which display unstable locations when different topologies are compared (Figure 1, Figure S1, Table S3).
FIGURE 1 | (A) Majority rule consensus tree from Bayesian analysis of the nucleotide dataset with excluded third codon positions (tree II) with the main CYC/TB1 gene lineages identified by colored boxes. Genes from Anthemideae are in bold. Branches affected by episodes of positive selection are indicated by black thick lines. Numbers in red represent posterior probabilities above 0.50. (B) Summary tree of the phylogeny of Asteraceae (modified from Panero et al., 2014) illustrating the relationships among the genera included in this study and their corresponding tribes.
**CYC/TB1 Genes Display a Pervasive Purifying Selection with Bursts of Episodic Positive Selection**

We tested whether the CYC/TB1 genes included in the phylogeny have been under positive selection. After rejection of recombination in the CYC/TB1 complete nucleotide matrix by RDP analysis \((P = 0.05)\), the dataset was evaluated by MEME and FEL. The analysis with MEME, with higher log-likelihood values than FEL, suggested episodic positive selection for 42 codons (Table S12). The output from FEL indicated negative and neutral evolution for 100 and 146 codons, respectively. Neither MEME nor FEL suggested evidence of pervasive positive selection.

Although most amino acid positions with episodic positive selection within the TCP domain lie in the basic domain and in the loop, additional sites were detected within helix 1 and adjacent to helix 2. The maximum likelihood estimate (MLE) of the synonymous rate \(\alpha\) is always higher than the non-synonymous rate \(\beta\) in a few conservative sites \((\alpha = \beta)\) between TCP and ECE domains (Table S12). On the other hand, identified sites with high selective constrain \((\beta = 0)\) occur in the TCP domain and outside the ECE and R domains. The proportion of the branches evolving under unconstrained non-synonymous rate \(\beta^{+}\) is always small \((q^{+} < 38\%)\) in comparison with branches where the synonymous substitutions prevail \((q^{-} > 63\%)\). Examination of the magnitude of the Empirical Bayes Factor (EBF) and the single nucleotide substitutions on different branches of the MEME output trees (not shown) revealed that several clades were affected by episodic positive selection (Figure 1). From the identified CYC/TB1 lineages, the CYC3b clade is the only one affected by episodic positive selection just before its ortholog diversification. Although there is a pervasive purifying selection trend in the CYC/TB1 genes here analyzed, there are episodes of positive selection not associated with the diversification of the main orthologous gene sets here identified, excepting CYC3b.

**Inference of Diversification Rate Shifts**

The phylogenetic data coupled with the BAMM model is designed to automatically detect changes of speciation rates. The BAMM analyses converged well as indicated by high ESS values (ESS log-likelihood = 1244.05, ESS number of shifts = 1333.27). BAMM failed to detect any significant rate-shift configuration associated with CYC lineage diversification, and the best-fit model to the phylogeny was one involving a homogeneous process of near constant per-lineage diversification rates except for lineage 2c1 restricted to Anacyclus (Figure S2). Running the analyses with the different prior settings did not change the overall pattern.

**AcCYC2b, AcCYC2c, and AcCYC2d Genes Are Highly Expressed in Wild but Not in Trumpet Ray Flowers**

There was little or no expression of CYC2 genes in vegetative tissues compared with flowers or inflorescences (Figure 2), except for AcCYC2a gene expression in roots and leaves of trumpet individuals. In ray and disc flowers, the expression of AcCYC2a was not different between wild and trumpet individuals (Figures 2A–C). AcCYC2a was expressed at slightly higher levels in young capitula (Cap1) and ray flowers than in mature capitula (Cap2) and disc flowers (Figure 2A). Expression of AcCYC2a in wild individuals (not shown) was high (97.45%) in young ray flowers with unexpanded ligules (ca. 0.5-cm-long). In addition, the relative expression of AcCYC2b, AcCYC2c, and AcCYC2d genes was much higher in ray flowers of the wild individuals than in other tissues (Figures 2D–F). In trumpet individuals, the expression was below 20% in young capitula (Figures 2D–E), mature capitula (Figure 2E), rays (Figure 2D), and disc flowers (Figures 2E–F). The average target Eexp\(\Delta Ct\) control Eexp\(\Delta Ct\) results for each gene follow a normal distribution according to the Kolmogorov-Smirnov test \((\alpha = 0.05\), Table S11\). The t-test indicates that the expression of the AcCYC2 genes in wild and trumpet is significantly different for all genes \((\alpha = 0.05\), Table S12\).

**Pattern of AcCYC2 Gene Expression in Early Flower Development**

To investigate the mRNA distribution of AcCYC2 genes during capitulum and early flower development we carried out in situ hybridizations with digoxigenin labeled RNA probes complementary to these genes. AcCYC2b and AcCYC2c transcripts could not be detected in our experiments. In contrast, AcCYC2d was strongly expressed in young disc flower meristems and during floral organ initiation (Figures 3A, B). AcCYC2d mRNA also accumulated in young developing stamens and ovules (Figures 3A, C). Likewise, AcCYC2a transcripts were detectable in developing disc flowers, both in the developing stamens and the ovules (Figures 3D, E). At this floral stage AcCYC2d and AcCYC2a signals were clearly excluded from developing corolla lobes (Figures 3A, D). Sense probes of these genes gave no detectable signal in sections of similar tissues (Figure S3).

**DISCUSSION**

**CYC/TB1 Diversification in Asteraceae Reflects Species Phylogeny**

The phylogenetic reconstruction shown in this study (Figure 1) is congruent with previous partial analyses of CYC/TB1-like genes of Asteraceae (Chapman et al., 2008, 2012; Kim et al., 2008; Tähtiharju et al., 2012). The CYC2 orthologous sets proposed by Tähtiharju et al. (2012) correspond to our 2a–2d lineages, with the difference that we found two lineages within 2a (2a1 and 2a2) and a new diversification within Anthemideae (2c1 and 2c2). Within CYC3, we also identified the two orthologous groups 3a and 3b reported by Tähtiharju et al. (2012), and in CYC1 we inferred the lineages 1a and 1b. Although congruence between CYC-like genes and species trees is not easy to interpret (e.g., in Dipsacaceae; Carlson et al., 2011), our gene tree reconstruction for each of the CYC lineages is broadly consistent with the species tree phylogeny (Panero and Funk, 2008; Figure 1B). For example, in the lineage 2b, the topology of the genes from [(Berkheya
(Helianthus (Senecio+Callistephus)) (Anacyclus+Matricaria)] is congruent with the species phylogeny of Asteraceae (Figure 1B). The same consistency is maintained in lineages 1b, 2a1, 2a2, 2b, 2c, and most of 2d, where the only incongruence with the species tree is the closer relationship of HaCYC2a from Helianthus to Anthemideae (AcCYC2d, AvCYC2d, MaCYC2d, and McCYC2d) rather than to Callistephus (CcCYC2a) genes (Figure 1).

Despite this general congruence, the distribution of the CYC2 genes indicates specific gene gain/loss events in relatively well-sampled genera, such as Gerbera, Helianthus and Senecio. Tracking the CYC2 paralogous genes of Anacyclus and Matricaria, we found that they are distributed in six gene clades, whereas Gerbera genes are absent from 2a1, 2b, and 2c clades (however, there is a phylogenetically unstable clade composed of GhCYC0 and GhCYC3 with an unresolved position with the CYC2 lineages, which could be potentially included in any of the CYC2 lineages; Figure 1, Figure S1). Helianthus genes are absent from the 2a1 clade and HaCYC2c relationships are unresolved, while Senecio RAY1 and RAY2 are present only in the 2b and 2c clades. Aside from the obvious lack of gene and species sampling that can alter the CYC/TB1 diversification pattern, it would be important to test whether the differential gene diversification of the species is linked to morphological or
FIGURE 3 | Morphology of Anacyclus capitulum and florets, and tissue-specific expression of AcCYC2 genes during capitulum development. (A) Section of a Anacyclus capitulum hybridized with a digoxigenin-labeled probe complementary to AcCYC2d. Youngest flower meristems (fm) can be seen at the top, older flowers to the left and right. AcCYC2d transcripts accumulate in young flower meristems and in meristems that are initiating corolla lobe primordia (*). In older flowers mRNA is detectable in developing stamens (st) and ovules (ov). No signal is detectable in developing corolla lobes (lb). (B) Scanning electron microscopy (SEM) image of a capitulum of an age comparable to that in (A). (C) Close up of an ovule displaying AcCYC2d expression. (D) Young disc floret showing AcCYC2a mRNA signal in stamens and ovule but not in corolla lobes. (E) SEM image of disc flowers comparable to those in (D).

functional evolutionary transitions. Moreover, the orthologous CYC2 genes in Helianthus and Gerbera do not necessarily share the same function (Broholm et al., 2014), which might allow similar gene repertoires shaping different inflorescence morphologies. The maintenance of each CYC2 paralog in different species of Asteraceae suggests that they existed before the species diversification had taken place in this family. This recruitment and maintenance of CYC2 function along the history of Asteraceae results in independent evolution of an adaptive trait, the heterogamous capitula (Chapman et al., 2012; Hileman, 2014). The convergent headed inflorescence in Asteraceae and Dipsacaceae could be the consequence of similar diversification patterns of CYC1- and CYC2-like genes in both families (Howarth and Donoghue, 2005; Carlson et al., 2011; Specht and Howarth, 2014). Although the CYC/TB1 phylogenetic pattern is not a consequence of differential rate shifts between the inferred CYC lineages according to the BAMM analyses performed here, whole genome duplication events in Asteraceae occurring in the last 40 million years (Barker et al., 2008, 2015) could be responsible for the observed diversity of CYC/TB1 genes.

Now that several CYC/TB1 genes isolated from separate studies consistently fit into specific CYC lineages in a comprehensive phylogeny (Figure 1A), it will be convenient from here onto add and assign newly isolated genes to this framework. It will encourage an appropriate association of new CYC/TB1 isolated genes with its phylogenetic origin and will allow a more consistent classification.

CYC/TB1 Proteins Evolve under Purifying and Episodic Positive Selection

Analysis of our CYC/TB1 gene dataset suggests a pervasive purifying selection with bursts of episodic positive selection, where a very small proportion of sites evolve at unconstrained non-synonymous rate ($q^+ < 38\%$; Table S12). Chapman et al. (2008) found that the per-site frequency of synonymous substitutions was saturated on many internal branches of Helianthus CYC/TB1 genes and that TCP and R domain were evolving under strong purifying selection. Similarly, CYC/TB1 genes from Antirrhineae were also subjected to strong purifying selection (Hileman and Baum, 2003). Even in recent duplicates of RAY2 in Senecio vulgaris, there is no evidence of positive selection that justifies their divergence (Chapman and Abbott, 2009). In terms of events of episodic codon selection associated with the stem/crown clades of the main CYC lineages identified in this study, only one codon has episodic positive selection at the stem node of the lineage CYC3b (Figure 1). For the remaining main CYC lineages, there were no identified episodes of positive selection associated with their origin or diversification, suggesting that selection changes are not important performers of the CYC/TB1 main phylogenetic patterns. Although some minor clades seem to be diversified after the positive selection (Figure 1), it might be affecting the quaternary rather than the primary/secondary protein structure used for phylogenetic reconstruction. With the lack of a significant difference between the diversification rates in the CYC/TB1 gene lineages suggested by our BAMM analysis, it seems that it is a more uniform and recent pattern of evolution of these genes.

Frontiers in Plant Science | www.frontiersin.org April 2017 | Volume 8 | Article 589 | 108
In cases, such as the bird toll-like receptors (Grueber et al., 2014) and the eudicot X-intrinsic proteins (Venkathe et al., 2015), there is a similar pattern of predominant purifying selection with rounds of episodic positive selection as inferred here for the CYC/TB1 genes. The high constraints imposed by the purifying selection of CYC/TB1 proteins could be maintaining their general patterns/functions conserved along the eudicots, whereas the episodic positive selection might allow a subtle modulation of protein-protein interactions, such as binding regulation or protein differential heteromeric combinations (see e.g., differential capacity of dimerization of CYC/TB1 proteins in Gerbera and Helianthus in Tähtiharju et al., 2012).

Expression Patterns of CYC/TB1 Are Similar in Asteraceae Orthologs

Comparing the CYC/TB1 genes of Anacyclus clavatus with their identified orthologs, the expression patterns are similar. For example, AcYC2b, detected in young capitate and highly expressed in ray flowers (Figure 2D), lies in the orthologous set of Senecio RAY1 and HaCYC2d (lineage CYC 2b; Figure 1). RAY1 is expressed in young inflorescences in the peripheral area of the ray floral meristem in radiate and non-radiate capitulo of Senecio (see Figure 2 in Kim et al., 2008) and HaCYC2d is one of the strongest candidates for conferring ray flower identity in Helianthus (Tähtiharju et al., 2012). AcYC2c2d has a similar expression profile to AcYC2b in qPCR analyses (Figures 2F), and is orthologous to HaCYC2a and GhCYC7 (lineage CYC 2d; Figure 1). GhCYC7 and HaCYC2a are expressed in different tissues, but GhCYC7 appears in earlier stages of ray and trans flowers, similar to HaCYC2a in ray flowers (Chapman et al., 2008, 2012; Juntheikki-Palovaara et al., 2014). Also, AcYC2d (Figure 3A) and GhCYC7 are expressed in early stamen primordia of disc and ray flowers (Juntheikki-Palovaara et al., 2014).

Nevertheless, in the lineages CYC2c and CYC2a (Figure 1A, Figure S4), the expression pattern of orthologous genes is not as similar as in CYC 2b and CYC 2d. In CYC2c, the orthologous AcYC2c, RAY2 and HaCYC2c genes are highly expressed in young and mature ray flowers (Kim et al., 2008; Tähtiharju et al., 2012) but HaCYC2e appears widely expressed in several tissues in PCR assays (Chapman et al., 2008). In the lineage CYC2a, the genes AcYC2a, GhCYC4, GhCYC9, and HaCYC2b are expressed in different tissues, but expression is nonetheless higher in ray flowers of Gerbera and Helianthus (Figure 2C; Chapman et al., 2008, 2012; Tähtiharju et al., 2012). The expression of AcYC2a and HaCYC2b seems not affected when actinomorphic tubular ray flowers are formed in the trumpet individual (Figure 2C) and the tubular mutants of Helianthus, respectively (see Figure 2B in Chapman et al., 2012). Therefore, despite the fact that in Asteraceae the CYC2b and CYC2d orthologous genes display similar expression patterns, it is not always possible to predict a particular CYC/TB1 gene expression pattern from the phylogenetic framework.

In the case of genes with unstable phylogenetic positions, such as HaCYC2c, GhCYC2, GhCYC3, and GhCYC5 lying outside the CYC2 lineages 2a–2d (Figure 1, Figure S1), there are redundant expression patterns and multiple functions (Figure S4). Whereas, GhCYC5 seems to be involved in the control of the flower density and in floral organ fusion, GhCYC2 is expressed in the dorsal part of the ray flowers, reproductive whorls, the ligule and the perianth throat of ray flowers (Broholm et al., 2008; Tähtiharju et al., 2012; Juntheikki-Palovaara et al., 2014). GhCYC3 and HaCYC2c are crucial for the ray flower identity and are expressed in meristem, perianth and ovules of ray flowers (Tähtiharju et al., 2012; Chapman et al., 2008). Juntheikki-Palovaara et al. (2014) suggest that redundancy of the CYC2 genes in Gerbera reflect a functional specificity for the CYC2 proteins obtained by the formation of specific protein complexes. In Asteraceae, the maintenance of the inflorescence unit may require a cross regulation between the CYC2 genes from different lineages, analogous to the interactions of CYC2 genes identified in Primulina heterotricha from Gesneriaceae (Gao et al., 2008; Yang et al., 2012).

A gradient of expression of the CYC/TB1 genes occurs in pseudanthial structures bearing different morphologies (e.g., in the Myrtaceae Actinodium cunninghamii where a cluster of fertile actinomorphic flowers are surrounded by ray-shaped branched shoots; Claßen-Bockhoff et al., 2013). This pattern is coincident with the centripetal gradient of floral morphology of the Asteracean inflorescence (Harris, 1999; Giterne et al., 2010). Aside from HaCYC2b in wild type Helianthus (Chapman et al., 2012), the CYC2 genes of wild Anacyclus (AcYC2c), Gerbera (GhCYC), and Helianthus (HaCYC) are usually highly expressed in the zygomorphic ray flowers relative to the disc flowers (Figure 3; Chapman et al., 2012; Tähtiharju et al., 2012). Expression patterns in Helianthus mutants agree with this general Asteraceae profile (Berti et al., 2005; Fambrini et al., 2006, 2011; Chapman et al., 2012). The double floved mutant (dbl), with disc flowers displaying bilateral ray-like corollas, expresses Ha/CYC2c ectopically, an important loci for the establishment of ray flower identity. On the other hand, in the tubular-rayed (tub) mutants with ray flowers displaying tubular actinomorphic corollas (similar to the trumpet phenotype in Anacyclus; Figure 2B), HaCYC2c is expressed at lower levels due to the presence of transposable elements. Although we cannot suggest a direct ortholog gene of HaCYC2c in Anthemideae due its phylogenetic unstable position (Figure 1, Figure S1), qPCR analysis in Anacyclus suggests a lower expression of the CYC2 genes AcYC2b, AcYC2c and AcYC2d in the actinomorphic ray flowers of the trumpets (Figures 2B–D).

Our results support the role of CYC 2 genes in the evolution of Asteraceae flower morphological diversity and illustrate their evolution, diversification, and expression patterns in Anacyclus. From an evolutionary perspective, the phylogenetic analyses show that CYC2 gene family has diversified in Asteraceae into four main paralogs, which has been accompanied by an increased structural and functional complexity in inflorescences across the different lineages (Chapman et al., 2008; Tähtiharju et al., 2012). The comparison of gene expression analyses in CYC paralogs and their phylogenetic relationship suggests that different Asteraceae lineages have mostly conserved their roles in determining floral symmetry (Garcés et al., 2016). However, this work also confirms previous evidence proposed by Fambrini...
and Pugliesi (2017) for a consistent functional recruitment of CYC2 genes in the development of microspores (pollen) and macrospores (ovule) in female and bisexual flowers of the capitulum. This observation opens a new field for the study of the involvement of CYC2 genes in the evolution of new sexual systems in Asteraceae.

**AUTHOR CONTRIBUTIONS**

JF, MB, IA, and PC conceived the study. IA did the fieldwork, MB and IA maintained Anacyclus living collections, MB performed phylogenetic analyses, MB and PC conducted the in situ hybridization, MB and GS carried out the qPCR analysis, JF performed the speciation rate shifts analysis. MB, JF, IA, and PC discussed the results and wrote the manuscript.

**ACKNOWLEDGMENTS**

We are grateful to F. Durán, A. Gallego, A. Herrero, B. Ríos, Y. Ruiz (RJB-CSIC) and R. Torices (Université de Lausanne) for their contribution during the collection of plant material and CYC sequences of Anacyclus and Matricaria. R. Riina photographed the trumpet phenotypes. L. Barrios (CTI-CSIC) provided helpful advice on qPCR statistical analysis. We particularly acknowledge support from F. Chevalier, I. Domínguez, E. Gonzalez-Grandío and M. Nicolas (CNB-CSIC) for help with in situ hybridization and their valuable comments. We thank K. McCreath for revision of the English language. This work was supported by grants from the Spanish Ministry of Economy and Competitiveness, Plan Estatal de I+D+I and European Social Fund to JF, IA, and PC (CGL2007-66516, CGL2013-49097-C2-2-P and BIO2014-57011-R), “Juan de la Cierva” program co-financed by the European Social Fund to MB (JCI-2010-07374) and Proyecto Intramural CSIC 201430E023 to JF.

**REFERENCES**


SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2017.00589/full#supplementary-material

**Figure S1** | Comparison of the CYC/TB1 summary trees based on nucleotide (A,C,D) and amino acid (B) data sets.

**Figure S2** | Maximum clade credibility (MCC) tree from the Bayesian inference analysis of CYC/TB1 genes.

**Figure S3** | Longitudinal sections of floral tissues of *A. clavatus* hybridized with the sense probe.

**Figure S4** | Summary of expression patterns of selected CYC2 genes in the Asteraceae/Calyceraceae clade.

**Table S1** | CYC-like genes/clones isolated from Anacyclus and Matricaria (Asteraceae, Anthemideae).

**Table S2** | Primers used for inverse PCR.

**Table S3** | Eudicot CYC-like genes included in the phylogenetic analyses.

**Table S4** | Datasets, trees and main statistics of the Bayesian analyses.

**Table S5** | Annealing temperature, product size and efficiency of the CYC2 gene specific primers used for qPCR in Anacyclus clavatus tissues.

**Table S6** | Main results of the qPCR essays performed on wild and trumpet individuals of *A. clavatus* following the Pfaffl method.

**Table S7** | Kolmogorov-Smirnoff test for probability distribution. α = 0.05.

**Table S8** | Ratio of expression (target Eexp/ΔCt) of *AcCYC2* genes in trumpet and wild individuals.

**Table S9** | t-test paired two sample for means of trumpet and wild expression of *AcCYC2* genes. α = 0.05.

**Table S10** | Primers used to amplify the CYC2 gene specific probes for in situ hybridization in *Anacyclus clavatus*.

**Table S11** | Main lineages reconstructed from the different Bayesian and ML analyses performed and their supports.

**Table S12** | Forty two CYC/TB1 codons with significant evidence of positive selection according to the Mixed Effects Model of Episodic Selection (MEME) method.


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

Copyright © 2017 Bello, Cubas, Álvarez, Sanjuanbenito and Fuertes-Aguilar. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.
Evolution and Expression Patterns of TCP Genes in Asparagales

Yesenia Madrigal 1, Juan F. Alzate 2 and Natalia Pabón-Mora 1*

1 Facultad de Ciencias Exactas y Naturales, Instituto de Biología, Universidad de Antioquia, Medellín, Colombia, 2 Centro Nacional de Secuenciación Genómica, Sede de Investigación Universitaria, Facultad de Medicina, Universidad de Antioquia, Medellín, Colombia

CYCLOIDEA-like genes are involved in the symmetry gene network, limiting cell proliferation in the dorsal regions of bilateral flowers in core eudicots. CYC-like and closely related TCP genes (acronym for TEOSINTE BRANCHED1, CYCLOIDEA, and PROLIFERATION CELL FACTOR) have been poorly studied in Asparagales, the largest order of monocots that includes both bilateral flowers in Orchidaceae (ca. 25,000 spp) and radially symmetrical flowers in Hypoxidaceae (ca. 200 spp). With the aim of assessing TCP gene evolution in the Asparagales, we isolated TCP-like genes from publicly available databases and our own transcriptomes of Cattleya trianae (Orchidaceae) and Hypoxis decumbens (Hypoxidaceae). Our matrix contains 452 sequences representing the three major clades of TCP genes. Besides the previously identified CYC specific core eudicot duplications, our ML phylogenetic analyses recovered an early CIN-like duplication predating all angiosperms, two CIN-like Asparagales-specific duplications and a duplication prior to the diversification of Orchidoideae and Epidendroideae. In addition, we provide evidence of at least three duplications of PCF-like genes in Asparagales. While CIN-like and PCF-like genes have multiplied in Asparagales, likely enhancing the genetic network for cell proliferation, CYC-like genes remain as single, shorter copies with low expression. Homogeneous expression of CYC-like genes in the labellum as well as the lateral petals suggests little contribution to the bilateral perianth in C. trianae. CIN-like and PCF-like gene expression suggests conserved roles in cell proliferation in leaves, sepals and petals, carpels, ovules and fruits in Asparagales by comparison with previously reported functions in core eudicots and monocots. This is the first large scale analysis of TCP-like genes in Asparagales that will serve as a platform for in-depth functional studies in emerging model monocots.

Keywords: Cattleya trianae, CINcINNA, CYCLOIDEA, Hypoxidaceae, Hypoxis decumbens, floral symmetry, Orchidaceae, PROLIFERATION CELL FACTOR

INTRODUCTION

As currently circumscribed the order Asparagales is a species-rich group comprising ca. 50% of all monocots, corresponding to 10–15% of flowering plants (Chase et al., 2009, 2016; Chen et al., 2013; Givnish et al., 2016). The most recent phylogenetic analyses in the monocots place Orchidaceae as sister to all other Asparagales (Chen et al., 2013). The family is divided into five subfamilies: Apostasioideae, Vanilloideae, Cypripedioideae, Orchidoideae, and Epidendroideae (Chase et al., 2015; Endress, 2016). The floral groundplan in Asparagales varies primarily in the floral symmetry and the number of stamens (Simpson, 2006). The floral morphology of...
Asparagales outside orchids consists of radially symmetrical, trimerous flowers with tepaloid perianth and free floral organs, although a few exceptions have been documented in Aspidistra (Asparagaceae), Gethyllis (Amaryllidaceae), Neocatelia (Asteliacae) and Pauridia (Hypoxidaceae) (Rudall, 2002; Rudall and Bateman, 2002, 2004; Kocyan, 2007). Conversely, orchid flowers are variously bilateral and undergo extreme elaboration of some organs, including differentiation of perianth parts, stamen abortion, and fusion of floral parts from the same whorl or from different whorls (Rudall, 2002). In the bilateral resupinated orchid flowers the two dorsal petals are very similar to each other, whereas the ventral one (the lip or labellum) often undergoes extreme elaboration in shape, color, size and epidermal specializations (Rudall and Bateman, 2004; Pabón-Mora and González, 2008; Mondragón-Palomino and Theißen, 2009; Rudall et al., 2013; Endress, 2016). In the inner floral whorls bilateral symmetry is evident by the formation of a gynostemium that results from the congenital fusion between the single fertile stamen (sometimes two fertile stamens) and stigmas (Rudall and Bateman, 2002; Pabón-Mora and González, 2008; Endress, 2016).

Such floral elaboration has been linked to extremely specialized pollination mechanisms and the exceedingly high diversification rates in Orchidaceae (Gong and Huang, 2009; Mondragón-Palomino and Theißen, 2009; Mondragón-Palomino, 2013).

The genetic network underlying bilateral floral symmetry has been assessed using Antirrhinum majus floral symmetry mutants (Luo et al., 1996). This network includes the differential dorsiventral expression of four transcription factors in the two-lipped flowers of this species. Three transcription factors, CYCLOIDEA (CYC), DICHTOMOTA (DICH) and RADIALIS (RAD) regulate cell division on the dorsal portion of the flower primordium and during dorsal petal and stamen primordia initiation. Additionally, RAD outcompetes DIVARICATA (DIV) for binding proteins in the dorsal side of the flower, restricting DIV function to the ventral and lateral petals (Almeida et al., 1997; Galego and Almeida, 2002; Raimundo et al., 2013). Thus, cyc/dich mutants show radialy symmetrical ventralized flowers (Luo et al., 1996, 1999). Both, CYC and DICH genes belong to the TCP gene family ( acronym for TEOSINTE BRANCHED 1 -TB1- from Z. mays, CYCLOIDEA -CYC- from A. majus and PROLIFERATION CELL FACTOR 1 and 2 -PCF1 and PCF2- from Oryza sativa) (Doebley et al., 1997; Kosugi and Ohashi, 1997; Luo et al., 1999). RAD and DIV belong to the MYB (Myeloblastosis) gene family (Luo et al., 1999; Galego and Almeida, 2002; Corley et al., 2005; Costa et al., 2005).

The TCP genes encode putative basic-Helix-Loop-Helix (bHLH) transcription factors (Cubas et al., 1999). The bHLH domain recognizes a consensus sequence GGNCCTAC/GTCCNTCCC required for DNA binding and activation or repression of transcription (Kosugi and Ohashi, 2002; Martin-Trillo and Cubas, 2010). Gene evolution analyses have identified two large groups of TCP genes, namely Class I (which include PCF homologs) and Class II (containing the CIN/CYC/TB1-like genes) (Cubas et al., 1999; Damerval and Manuel, 2003; Reeves and Oilmstead, 2003; Broholm, 2009; Mondragón-Palomino and Trontin, 2011). Additional large scale duplications (i.e., those occurring prior to the diversification of major inclusive hierarchical groupings) have been found within the CYC genes. Two rounds of duplication occurred specifically in core eudicots, resulting in CYC1, CYC2, and CYC3 clades, and one duplication specific to monocots resulting in the RETARDED PALEA 1 (REPI) and TEOSINTE BRANCHED 1 (TB1) clades (Vieira et al., 1999; Damerval and Manuel, 2003; Howarth and Donoghue, 2006; Navaud et al., 2007; Yao et al., 2007; Mondragón-Palomino and Trontin, 2011). In addition, species specific duplications (i.e., those occurring in a single species) have also been reported, often linked to polyploidy (Ma et al., 2016). Non-core eudicot homologs are known as the CYC-like genes (Damerval et al., 2007; Preston and Hileman, 2012; Horn et al., 2015). Functional characterization has concentrated in CYC2 orthologs in eudicots, including Asterales, Brassicales, Dipsacales, Fabales, Lamiales and Malpighiales, among others (Busch and Zachgo, 2007; Gao et al., 2008; Preston et al., 2009; Wang et al., 2010; Zhang et al., 2010, 2013; Howarth et al., 2011; Tähtiharju et al., 2012; Yang et al., 2012). These studies have found CYC2 expression restricted to the same dorsal floral domain and a conserved role as cell proliferation repressors resulting in bilateral symmetry (reviewed in Hileman, 2014). Fewer studies have been made in basal eudicots, but dissymmetric Fumarioids (Papaveraceae) do have asymmetric expression of CYC-like genes, suggesting that CYC-like recruitment to form bilateral flowers has occurred independently several times in eudicots (Damerval et al., 2007, 2013).

Less is known about the role of pre-duplication CYC-like genes in monocots (Bartlett and Specht, 2011; Mondragón-Palomino and Trontin, 2011; Preston and Hileman, 2012). Expression analyses of CYC-like genes in Costus (Costaceae; Zingiberales) and Commelina (Commelinaceae; Commelinales) suggest that they play a role in bilateral symmetry (Bartlett and Specht, 2011; Preston and Hileman, 2012). Functional studies in O. sativa (Poaceae, Poales) confirm that these genes contribute to the asymmetric growth of the dorsal versus the ventral portions of the flower, as shown by the rep1 mutants which exhibit a smaller palea due to cell division arrest (Yuan et al., 2009). The only two studies available in Orchidaceae are particularly intriguing as they show very different expression patterns of CYC/TB1-like orthologs. Whereas, the only copy of CYC/TB1-like in Orchis italicca (OitaTB1) is expressed exclusively in leaves (De Paolo et al., 2015), two of three CYC/TB1-like copies in Phalaenopsis equestris, PeCYC1 and PeCYC2, seem to be expressed in higher levels (2–10 times more) in the dorsal sepal and the labellum compared to the ventral sepal and the lateral petals (Lin et al., 2016). Furthermore, some authors have hypothesized that the expression gradient of TCP genes is largely controlled by upstream expression of the AP3/DEF petal-stamen identity genes, resulting in higher concentrations of CYC/TB1-like genes in the dorsal floral regions; however, more experimental data is needed to support this (Mondragón-Palomino and Theißen, 2009).

It is unclear whether closely related TCP-like CINNINNATA (CIN) and PROLIFERATION CELL FACTOR (PCF) genes play any role in floral symmetry. CIN was originally characterized in A. majus and more recently in Arabidopsis thaliana (Crawford...
et al., 2004; Nag et al., 2009; Sarvepalli and Nath, 2011; Danisman et al., 2013). In both species CIN controls cellular proliferation in petals and cellular arrest in leaves (Crawford et al., 2004; Nag et al., 2009). On the other hand, O. sativa PCF1 and PCF2 are involved in auxillary meristem repression, likely via the activation of PROLIFERATING CELL NUCLEAR ANTIGEN (PCNA), which encodes a protein involved in DNA replication and repair, maintenance of chromatin structure, chromosome segregation, and cell-cycle progression (Kosugi and Ohashi, 1997). Other studies in A. thaliana suggest that PCF-like genes are also involved in gametophyte development, transduction of hormonal signals, mitochondrial biogenesis, leaf and flower morphogenesis, seed germination, branching, and even circadian clock regulation (Koyama et al., 2007; Pruneda-Paz et al., 2009; Giraud et al., 2010; Kieffer et al., 2011; Resentini et al., 2015). Recent studies in the model orchid P. equestris have found that PcPeCIN8 (CIN-like) and PcPePCF10 (PCF-like) control cell proliferation and cell shape in petals, ovules and leaves (Lin et al., 2016).

In order to study the contribution of CYC/TB1-like and the closely related CIN-like and PCF-like genes to floral patterning in Asparagales, we first determined copy number and characteristic protein motifs and then assessed gene lineage evolution including a vast sampling of TCP-like genes across angiosperms and particularly of Asparagales monocots. Next, we evaluated the expression patterns of all TCP-like genes in dissected floral organs, young leaves, and fruits of Hypoxis decumbens (Hypoxidaceae) which has typical asparagalean radial trimerous flowers with free parts, and Cattleya trianae (Orchidaceae), that has bilateral flowers, and a single fertile stamen fused with the tree stigmas (i.e., gynostemium). Finally, we propose hypotheses on functional evolution based on previous literature reports and comparisons with our results that suggest different trends among gene clades when comparing Asparagales to model core eudicots.

MATERIALS AND METHODS

Gene Isolation and Phylogenetic Analyses

In order to isolate putative TCP-like homologs in Asparagales, searches were performed using previously reported TCP genes from eudicots, monocots and in particular Orchidaceae as queries (Mondragón-Palomino and Trontin, 2011; Preston and Hileman, 2012; De Paolo et al., 2015; Horn et al., 2015). Searches included homologs from all the three main clades of TCP genes: CYC-like, CIN-like and PCF-like. Searches were done using BLAST tools (Altschul et al., 1990) in the orchid specific available databases including Orchidbase 2.0 (http://orchidbase.itsp.ncku.edu.tw/). (Tsai et al., 2013), Orchidrsta (http://orchidrsta2.abrc.sinica.edu.tw/orchidrsta2/index.php) (Su et al., 2013) as well as the more inclusive OneKP database (http://www.bioinfodata.org/Blast4OneKP/). All core eudicot sequences, were isolated from Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html) and genbank (https://www.ncbi.nlm.nih.gov/genbank/).

In addition to the available databases we generated two transcriptomes from C. trianae (Orchidaceae) and H. decumbens (Hypoxidaceae). The transcriptome for each species was generated from mixed material from 3 biological replicates and included vegetative and reproductive meristems, floral buds, young leaves and fruits in as many developmental stages as possible. Total RNA was purified and used for the preparation of one mRNA (polyA) HiSeq library for each species. RNA-seq experiments were conducted using truseq mRNA library construction kit (Illumina, San Diego, California, USA) and sequenced in a HiSeq 2000 instrument reading 100 base paired end reads.

The transcriptome was assembled de novo with Trinity v2 following default settings. Read cleaning was performed with prinsesq-lite v0.20.4 with a quality threshold of Q35 and a minimum read length of 50 bases. Contig metrics are as follows: (1) H. decumbens total assembled bases: 73,787,751; total number of contigs (>101 bases): 157,153; average contig length: 469 bp; largest contig: 15,554 bp; contig N50: 1075 bp; contig GC%: 46.42. (2) C. trianae total assembled bases: 63,287,862; total number of contigs (>101 bases): 109,708; average contig length: 576 bp; largest contig: 9321 bp; contig N50: 1401 bp; contig GC%: 42.73. Homologous gene sequence was performed using BLASTN with the query sequences downloaded from GenBank and other databases (see above). In order to estimate the relative abundance of the assembled contigs, cleaned reads were mapped against the de novo assembled dataset and counted with two different strategies. The first one involved the mapping algorithm of the software Newbler v2.9 where only the unique matching read pairs were accepted as positive counts. The second one involved the mapping algorithm BOWTIE2 and raw reads counts as well as RPKM were calculated to each assembled contig (Table 1).

All sequences isolated were compiled with Bioedit (http://www.ncbi.nlm.nih.gov/Bioedit.html). Sequences shorter than 200 bp lacking similarity with a region of the putative bHLH motif were discarded. Nucleotide sequences were subsequently aligned using the online version of MAFFT (http://mafft.cbrc.jp/alignment/software/) (Katoh et al., 2002) with a gap open penalty of 3.0, offset value of 1.0 and all other default settings. The alignment was then refined by hand using Bioedit considering as a reference the 60–70 aa reported as conserved in the TCP protein domain (Cubas et al., 1999). To better understand the evolution of the TCP gene lineage, and to integrate previous eudicot and monocot specific phylogenetic analyses (Damerval and Manuel, 2003; Hileman and Baum, 2003; Reeves and Olmstead, 2003; Howarth and Donoghue, 2006; Damerval et al., 2007; Bartlett and Specht, 2011; Mondragón-Palomino and Trontin, 2011; Preston and Hileman, 2012; Horn et al., 2015), we performed Maximum likelihood (ML) phylogenetic analyses using the nucleotide sequences with RaxML-HPC2 BlackBox through the CIPRES Science Gateway (https://www.phylo.org/) (Miller et al., 2010). Bootstrapping (BS) was performed according to the default criteria in RAXML (200–600 replicates). The PCF-like gene from Amborella trichopoda (AtrTCP4) as well as all other PCF-like sequences were used as the outgroup. To find the molecular evolution model that best fit our data, we used the jModelTest package implemented in MEGA6 (Posada and Crandall, 1998). Trees were viewed and edited using FigTree v1.4.0 (http://tree.bio.ed.ac.uk/software/figtree/) (Rambaut, 2014). Newly isolated sequences from our own generated transcriptomes from C. trianae (Orchidaceae) and H. decumbens (Hypoxidaceae) can
TABLE 1 | Selected features of TCP and AP3-like gene contigs taken from sequences and the transcriptomes of Cattleya trianae and Hypoxis decumbens.

<table>
<thead>
<tr>
<th>Gene</th>
<th>CDS</th>
<th>Contig</th>
<th>CDS size (in bp)</th>
<th>Protein size (in AA)</th>
<th>URP* Newbler</th>
<th>RRC** RPKM</th>
<th>RPKM rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>CtrAP3a</td>
<td>c19830_g1_i1</td>
<td>684</td>
<td>227</td>
<td>4047</td>
<td>3877</td>
<td>91,023,385,79</td>
<td>16,899</td>
</tr>
<tr>
<td>CtrAP3b</td>
<td>c16430_g1_i2</td>
<td>696</td>
<td>231</td>
<td>15</td>
<td>719</td>
<td>16,641,985,41</td>
<td>7,551</td>
</tr>
<tr>
<td>CtrAP3c</td>
<td>c16430_g1_i1</td>
<td>669</td>
<td>222</td>
<td>184</td>
<td>1002</td>
<td>23,724,784,36</td>
<td>5,775</td>
</tr>
<tr>
<td>CtrAP3d</td>
<td>c14728_g2_i1</td>
<td>628</td>
<td>208</td>
<td>1218</td>
<td>1228</td>
<td>38,723,939,68</td>
<td>3,786</td>
</tr>
<tr>
<td>CtrAP3e</td>
<td>c13983_g1_i1</td>
<td>601</td>
<td>199</td>
<td>321</td>
<td>269</td>
<td>9,445,131,34</td>
<td>11,262</td>
</tr>
<tr>
<td>CtrCIN1</td>
<td>c10815_g1_i1</td>
<td>945</td>
<td>314</td>
<td>565</td>
<td>540</td>
<td>7,473,218,106</td>
<td>13,078</td>
</tr>
<tr>
<td>CtrCIN2</td>
<td>c20571_g1_i1</td>
<td>975</td>
<td>324</td>
<td>5413</td>
<td>5455</td>
<td>58,713,801,49</td>
<td>2,597</td>
</tr>
<tr>
<td>CtrCIN3</td>
<td>c20515_g1_i2</td>
<td>1158</td>
<td>385</td>
<td>2065</td>
<td>2091</td>
<td>28,277,405,53</td>
<td>4,987</td>
</tr>
<tr>
<td>CtrCIN4</td>
<td>c20515_g1_i1</td>
<td>1116</td>
<td>371</td>
<td>371</td>
<td>491</td>
<td>6,141,833,864</td>
<td>14,794</td>
</tr>
<tr>
<td>CtrCIN5</td>
<td>c19791_g1_i1</td>
<td>957</td>
<td>319</td>
<td>453</td>
<td>411</td>
<td>8,006,529,24</td>
<td>12,506</td>
</tr>
<tr>
<td>CtrCIN6</td>
<td>c10815_g1_i1</td>
<td>930</td>
<td>309</td>
<td>260</td>
<td>258</td>
<td>8,318,287,47</td>
<td>12,198</td>
</tr>
<tr>
<td>CtrPCF1</td>
<td>c16466_g1_i1</td>
<td>799</td>
<td>252</td>
<td>360</td>
<td>308</td>
<td>8,606,589,31</td>
<td>14,228</td>
</tr>
<tr>
<td>CtrPCF2</td>
<td>c10284_g1_i1</td>
<td>882</td>
<td>293</td>
<td>1109</td>
<td>1118</td>
<td>23,042,432,74</td>
<td>5,910</td>
</tr>
<tr>
<td>CtrPCF3</td>
<td>c10751_g1_i1</td>
<td>561</td>
<td>187</td>
<td>315</td>
<td>230</td>
<td>11,375,259,56</td>
<td>9,984</td>
</tr>
<tr>
<td>CtrPCF4</td>
<td>c13983_g1_i1</td>
<td>900</td>
<td>299</td>
<td>493</td>
<td>401</td>
<td>6,065,203,82</td>
<td>14,927</td>
</tr>
<tr>
<td>CtrPCF5</td>
<td>c20571_g1_i1</td>
<td>828</td>
<td>275</td>
<td>5821</td>
<td>5849</td>
<td>12,805,290,89</td>
<td>5,397</td>
</tr>
<tr>
<td>CtrPCF6</td>
<td>c16850_g1_i1</td>
<td>642</td>
<td>213</td>
<td>1109</td>
<td>1118</td>
<td>23,042,432,74</td>
<td>5,910</td>
</tr>
</tbody>
</table>

*Unique read pairs - number of specific/uniform read pairs supporting each contig.
** Raw read counts.

be found under Genbank numbers KY296315–KY296347. All sequences included in the phylogenetic analyses can be found in the Supplementary Table 1.

Identification of New Protein Motifs
In order to detect previously reported, as well as to identify new, conserved motifs, 77 TCP-like genes were selected representing major model eudicot and monocot groups of this study (i.e., A. majus, O. sativa, Aloe vera, P. equestris, C. trianae and H. decumbens). Sequences were permanently translated and uploaded as amino acids to the online MEME server (http://meme-suite.org/tools/meme) and run with all the default options (Bailey et al., 2006). Specific analyses for each of the TCP clades (i.e., CYC-like, CIN-like, PCF-like) were also performed.
Expression Analyses by RT-PCR

To examine and compare the expression patterns of TCP-like genes we used floral buds, dissected floral organs, leaves, and fruits of C. trianae and H. decumbens. Preanthetic floral buds of H. decumbens were dissected into sepals, petals, stamens and carpels. In addition whole floral buds, immature fruits (F1, right after tepals shed off), mature fruits (F2, before lignification), and young leaves were also collected. Preanthetic floral buds of C. trianae were dissected into sepals, lateral petals, labellum (or lip), gynostemium, and ovary. Young leaves were also collected. Total RNA was isolated from each organ collected using the SV Total RNA Isolation System Kit (Promega, Madison, WI, USA), and resuspended in 20 µl of DEPC water. RNA was treated with DNaseI (Roche, Basel, Switzerland) and quantified with a NanoDrop 2000 (Thermo Scientific, Waltham, MA) (Wilfinger et al., 1997). Three Micrograms of RNA were used as a template for cDNA synthesis (SuperScriptIII RT, Invitrogen) using Oligo-dT primers. The cDNA was diluted 1:4 for amplification reactions by RT-PCR. Primers were designed in specific regions like portions flanking the conserved domains for each copy found in C. trianae and H. decumbens (Supplementary Table 2). Each amplification reaction incorporated 9 µl of EconoTaq (Lucigen, Middleton, WI), 6 µl of nuclease free water, 1 µl of BSA (5 µg/ml), 1 µl of Q solution (5 µg/ml) 1 µl f/w primer (10 mM), 1 µl rev primer (10 mM), and 1 µl of template cDNA for a total of 20 µl. Thermal cycling profiles followed an initial denaturation step (94°C for 30 s), an annealing step (50–59°C for 30 s) and an extension step with polymerase (72°C for 10 min), all by 30–38 amplification cycles. ACTIN2 was used as a load control. PCR products were run on a 1.0% agarose gel stained with ethidium bromide and digitally photographed using a Whatman Biometra® BioDoc Analyzer.

RESULTS

Exhaustive search from available databases retrieved 452 TCP Class I and Class II sequences from flowering plants. From these 138 belong to the Orchidaceae, including 18 homologs from C. trianae and 110 sequences belong to non-Orchidaceae Asparagales, including 15 homologs from H. decumbens (Supplementary Table 1). ML analyses were performed using the complete nucleotide sequences of all TCP-like genes isolated. The A. trichopoda AtrTCP4 together with all other isolated PCF-like genes were used as the outgroup. The analysis recovered two clades previously reported in TCP genes, namely the CYC/TB1-like clade (with a Bootstrap Support, BS = 98) and the CIN-like clade (BS = 67) (Figure 1). We will discuss our results for each clade separately.

CYC/TB1-like Gene Evolution

We were able to isolate 168 sequences belonging to the CYC/TB1-like clade (Supplementary Table 1). Our sampling includes 15 sequences from four species of Poales, 10 sequences from seven species of Asparagales (incl. Orchidaceae), eight from five species of Commelinaceae (Commelina, Alocroemia, Tradescantia), 14 from nine species of basal angiosperms and 121 from 46 species of eudicots. Only one homolog from Curculigo spp. (CurcTB1) and one homolog from H. decumbens (HydTB1) were recovered from our blast searches. Moreover, an exhaustive search in Orchidaceae specific databases resulted in eight additional copies: three from P. equestris (PETCP06750, PETCP06749, PETCP11715), two from P. aphrodite (PaTB1, PaTCP06749) and one from C. trianae (CtrTB1) (Epipendroideae); one homolog from O. italica (OitaTB1) (Orchidoideae); and one homolog from Vanilla shenzhenica (VaTCP06749) (Vanilloideae). CYC/TB1-like homologs seem to have undergone size reduction in Asparagales, and only the searches made with monocot TB1 genes yielded positive hits. Interestingly, in the two transcriptomes newly obtained in the present research, the CYC/TB1-like contigs were supported by fewer reads (when compared to MADS-box APETALA3 floral organ identity genes and other TCP-like genes; Table 1) suggesting low expression of these transcripts.

The resulting ML topology recovered the three previously established core eudicot subclades (Howarth and Donoghue, 2006) with very low support (BS < 50), namely CYC1/TCP18, CYC2/TCP1, and CYC3/TCP12 (Figure 1). Our analysis also recovers the previously identified duplication of CYC-like genes in basal eudicots (Citerne et al., 2013), and another in Poales, the latter resulting in the REP1/TB1 clades (Mondragón-Palomino and Trontin, 2011). Most basal angiosperms and many monocots outside Poales have single copy CYC-like genes that predate the independent duplications in eudicots and Poales (Figure 1). Intraspecific duplications in monocots have occurred in Zea mays, as well as in the orchids P. equestris and P. aphrodite (Figure 1; BS = 100). Outside of the monocots, local specific duplications have also occurred in the basal angiosperms Aristolochia ringens and Persea americana, in the basal eudicots Circaeaeret and Nelumbo, and in core eudicots such as Gerbera, Antirrhinum, Citrus, Glycine, Populus, and Gossypium (Figure 1).

Members of the CYC/TB1-like clade show very little variation in the ca. 60 amino acid TCP domain (sensu Cubas et al., 1999) consisting of a putative basic-Helix-Loop-Helix (bHLH) domain (Figure 2). We were able to identify the highly conserved residues previously reported for the putative bipartite Nuclear Localization Signal (NLS) at the N-terminus flanking the bHLH, which provide hydrophobicity in the α-helices and in the loop region between the two. The loop itself is highly conserved in all CYC/TB1-like sequences except for AmDICH and AmCYC that have an α > P change at position 42. The second helix contains the LxxLL motif in all CYC2 proteins; this motif is modified into a VxWLx motif in other CYC-like proteins.

Our MEME analysis identified motifs 1 and 2 corresponding to the TCP domain (Supplementary Figure 1). At the start of Helix I, between positions 24–29 we found specific amino acids exclusive to CYC protein homologs. Outside the TCP domain, motifs 7 and 10 (reported also by Bartlett and Specht, 2011) and motifs 36–40, 42, and 43 (reported also by De Paolo et al., 2015) were recovered in our analysis as conserved in all CYC proteins. In addition, the protein interaction R domain (motif 11) putatively involved in hydrophilic α-helix formation in TCP Class II genes (shared between CYC and CIN proteins) was also identified (Supplementary Figure 1; Cubas et al., 1999). However, this motif is absent from CtrTB1 and REP-1 (Supplementary Figure 1; Yuan et al., 2009).
FIGURE 1 | ML analysis of CYC/TB1-like genes. Overview (upper left) summary Maximum Likelihood (ML) analysis of all TCP-like genes, where (1) corresponds to PCF-like homologs, (2) to the CIN-like clade and (3) to the CYC/TB1-like clade. Detailed (right) gene tree expanded in the CYC/TB1-like clade (3); yellow stars indicate large-scale duplication events in core, basal eudicots and monocots; resulting in the core eudicot clades CYC3/TCP12, CYC2/TCP1 and CYC1/TCP18, the basal eudicot CYC-like clades and the TB1 and REP1 clades in Poales; red stars indicate species-specific duplication events; blue and pink arrows indicate H. decumbens and C. trianae homologs, respectively; branch and taxa colors correspond to those in the overview tree to the upper left. BS values $\geq$ 50 are shown.
FIGURE 2 | TCP protein domain alignment with Asparagales representative sequences. Oryza sativa (Poales) was used for reference. Names to the left indicate the clade to which sequences belong according to Figures 1, 4, 5 and Supplementary Table 1. The upper bars point to the putative structure bHLH (basic-Helix-Loop-Helix) at the TCP domain. Circles indicate residues forming part of the putative bipartite NLS; asterisks indicate conserved hydrophobic residues in the helices; black arrowheads point to residues (glycine or proline) that disrupt α-helix formation (modified following Cubas et al., 1999). The green box indicates changes in residues between CYC-like and CIN-like proteins. The pink box indicates the LxxLL motif with significant variations outside the CYC2 clade.
unidentified motifs include motif 4, exclusive to Epidendroideae and motif 5, exclusive to Phalaenopsis species. Whereas, most orchid CYC-like proteins (including *O. italica* OitaTB1 and *C. trianae* CtrTB1) do not share any common motifs with the canonical *A. majus* paralogs outside the TCP domain, the *Phalaenopsis* CYC-like homolog (PETTCP11715) shares motifs 6, 9, and 12 with AmDICH or AmCYC.

**CIN-like Gene Evolution**

A total of 155 CIN-like homologs were recovered and unlike the CYC-like sampling, most CIN-like sequences belong to the Asparagales (*Supplementary Table 1*). Our sampling contains 57 sequences from 35 non-Orchidaceae Asparagales species, including four paralogs from *H. decumbens* labeled HydCIN1-HydCIN4. A total of 78 CIN-like homologs were isolated, including four homologs from two Apostasioideae species, nine homologs from three Vanilloideae species, five homologs from three Cyprinidiodaeae species, 22 homologs from seven Orchidoideae species and 38 homologs from 11 Epidendroideae species. Furthermore, six paralogs were identified in *C. trianae* labeled CtrCIN1-CtrCIN6. Searches outside Asparagales were restricted to six homologs from *O. sativa*, two from *A. trichopoda* and 12 homologs from five eudicots species, including the canonical *A. majus* CINCNNATA (*AmCIN*).

The CIN-like ML analysis shows a duplication (BS = 75) that predates the diversification of angiosperms resulting in the CIN1 (BS = 79) and CIN2 clades (BS = 99) (*Figure 3*). This is confirmed by the position of the two *A. trichopoda* CIN paralogs, *AtTCPF* and *AtTCP2*, each in its own clade (*Figure 3*). Additional support for this early duplication is found in the topology yielded by a second complementary analysis that includes 11 Solanaceae CIN homologs, where CIN1 and CIN2 clades have monocot and core eudicot representatives and, at least CIN2 is well supported (BS = 93) (*Supplementary Figure 2*). The CIN1 clade has undergone at least two additional duplications resulting in the CIN1a-c clades. It is likely that the duplication resulting in CIN1a and CIN1b/c occurred exclusively in monocots, although the exact timing is unclear. The other duplication resulting in CIN1b and CIN1c appears to be Orchidaceae-specific, prior to the diversification of Orchidoideae and Epidendroideae (*Figure 3*).

On the other hand, the CIN2 clade underwent an independent duplication predating the diversification of Asparagales, resulting in CIN2a and CIN2b subclades. Intrasppecies duplications were identified in *Hesperaloe*, *Disporopsis*, *Maianthemum*, *Rhodophiaia*, *Sanseveria* and *Yucca*. (Figure 3). Poales CIN homologs form a clade, with a low BS, in the first analysis, with the exception of *OsPCF5* clustered with *AmCIN* (*Figure 3*). However, our second analysis shows two Poales clades nested in each angiosperm paralogous CIN1 and CIN2 clades (with low BS), suggesting that the two Poales clades likely resulted from the angiosperm CIN1/2 duplication (*Supplementary Figure 2*).

CIN-like sequences show high degree of conservation at the N-flank of the TCP domain (*Figure 2*). The only changes with respect to the key aminoacids in the bHLH domain in CYC proteins are at the second helix where the LxxLL motif shifts to V/IxxLL (*Figure 2*). Toward the 3’ end of the TCP domain proteins are highly variable, except for motifs 13, 14, 16–18 and 21, reported also by De Paolo et al. (2015) (*Supplementary Figure 1*). The R domain (motif 11) in CIN proteins is only present in the CIN1a clade and *O. sativa* homologs OsTCP21, OSTCP8 and OsTCP10. Motif 14, which corresponds to the miR319 binding site is present in most CIN-like sequences. The miR319 binding motif is lacking in HydCIN2, HydCIN3, OsTCP21, OsTCP27, and OsTCP10. All CIN1 subclade sequences share motifs 13, 17, 19, 20, 21, and 25; the CIN1a subclade shares motifs 11, 28, 29, 34 and 35, while the CIN1b subclade shares motifs 15, 23, 26, and 30. Synapomorphies for the CIN2 subclade include motifs 16, 18, 22, 24 and only motif 27 is exclusive to Orchidaceae. Finally, CIN2b homologs share motifs 32 and 33. The most divergent *C. trianae* sequence is CtrCIN6, which only has the motifs 1, 2, 3, 11, 13, 14, 17, and 21.

**PCF-like Gene Evolution**

Our analysis recovered 129 PCF-like homologs (*Supplementary Table 1*). Similarly to CIN-like genes most sampling is concentrated in Asparagales, thus 53 homologs belong to 35 species of non-Orchidaceae Asparagales and 52 sequences correspond to Orchidaceae. *H. decumbens* has 10 PCF-like copies (HydPCF1-HydPCF10). Sampling in Orchidaceae includes one homolog from one Vanilloideae species, 14 homologs from seven Orchidoideae species and 37 homologs from 11 Epidendroideae species. We recovered 11 PCF-like homologs from *C. trianae* (CtrPCF1-CtrPCF11). Sampling in monocots outside Asparagales include 10 homologs from *O. sativa* (Poales) and sequences that are not monocots are restricted to one homolog from *A. trichopoda* and 13 homologs from *A. thaliana*.

Our analysis detected at least five duplication events of PCF-like prior to the diversification of Asparagales, however support is low for all clades (BS = < 50) (*Figure 4*). In addition, our complementary analysis including 18 Solanaceae PCF-like genes also shows support for at least two rounds of core eudicot specific PCF-like duplications (*Supplementary Figure 2*).

PCF-like sequences exhibit the shortest basic motif in TCP proteins, with a deletion in the bipartite NLS between positions 10 and 13 (*Figure 2*; Cubas et al., 1999). The TCP domain shows little conservation in comparison to the TCP II class (CYC and CIN) proteins. For instance, there is a four amino acid deletion in the middle of the basic motif, and only 12 out of the 23 amino acids characterized in both helices and the loop are conserved (*Figure 2*). Additionally our MEME analysis show that PCF-like proteins do not have an R domain (motif 11), nor a target sequence for miR319 (motif 14) (*Supplementary Figure 1*).

**Expression of TCP-like Homologs from Hypoxis decumbens and Cattleya trianae**

In order to hypothesize functional roles for the Asparagales TCP-like homologs, the expression patterns of all homologs isolated from transcriptomic analysis in *H. decumbens* and *C. trianae* were evaluated (*Figure 5*). Although in both species we were able to dissect floral organs in preanthesis and young leaves, we were not able to find fruits of *C. trianae*, thus only young and old fruits of *H. decumbens* were included in the expression study.
FIGURE 3 | ML analysis of CIN-like genes. Overview (upper left), summary tree as in Figure 1. To the right, ML phylogenetic analysis of TCP genes expanded to show the CIN-like clade (2). Yellow stars indicate large scale duplication events: one prior to the diversification of angiosperms yielding clades CIN1 and CIN2, two before the radiation of Asparagales and one prior to the origin of Epidendroideae and Orchidoideae; red stars indicate species-specific duplication events; blue and pink arrows indicate H. decumbens and C. trianae homologs, respectively. Branch and taxa colors correspond to those in the conventions to the left. BS values $\geq 50$ are shown.
FIGURE 4 | ML analysis of PCF-like genes. Overview (upper left) summary tree as in Figure 1. To the right, ML phylogenetic analysis of TCP genes expanded to show the PCF-like clade (1); yellow stars indicate large scale duplication events at least three before the diversification of Asparagales; red stars indicate species-specific duplication events; blue and pink arrows indicate H. decumbens and C. trianae homologs, respectively. Branch and taxa colors correspond to those in the conventions to the left. BS values ≥ 50 are shown.
The CYC/TB1-like homologs have very different expression patterns in C. trianae and H. decumbens. HydTB1 is expressed in the floral bud, sepals, carpels, young and mature fruits and leaves, whereas CtrTB1 is expressed in all floral whorls and is not expressed in leaves (Figure 5). CIN-like homologs also exhibit different expression patterns in C. trianae and H. decumbens. HydCIN1, HydCIN2 and HydCIN3 are expressed in the floral bud, carpels, fruits and leaves. Only HydCIN2 expression is extended to sepals, petals and stamens at very low levels. Interestingly both HydCIN2 and HydCIN3 lack the miR319 binding site. HydCIN4 expression is restricted to fruits and leaves. CtrCIN1, CtrCIN2, CtrCIN5, and CtrCIN6 are expressed in sepals, petals, lip, gynostemium, and ovary, although, CtrCIN2 expression in sepals and gynostemium occurs at low levels. CtrCIN3 and CtrCIN4 have similar expression patterns to CtrCIN1, however, they are poorly expressed in the ovary and are the only CIN-like homologs that extend their expression to leaves.

Similar to CYC/TB1-like and CIN-like genes, the expression of PCF-like homologs varies dramatically between C. trianae and H. decumbens (Figure 5). CtrPCF1, CtrPCF2, CtrPCF3, CtrPCF4, CtrPCF5, CtrPCF6, CtrPCF9, and CtrPCF11 are expressed in all floral organs and the only copy with expression in leaves is CtrPCF4. CtrPCF3 has low expression in sepals and petals. CtrPCF7 is only expressed in petals and ovary. Finally, CtrPCF10 has expression in all floral whorls except in petals. In H. decumbens only HydPCF3, HydPCF4, HydPCF5, and HydPCF6 are expressed in the floral buds, carpels, fruits and leaves, although HydPCF3 and HydPCF4 have low expression in carpels. HydPCF5 is also expressed in the perianth with higher expression in sepals than in petals. HydPCF7 has restricted expression to the floral buds and the young fruits. HydPCF9 and HydPCF10 are only expressed in young fruits and the remaining copies (HydPCF1, HydPCF2, and HydPCF8) are only expressed at very low levels in the floral buds.

The CYC/TB1-like homologs have very different expression patterns in C. trianae and H. decumbens. HydTB1 is expressed in the floral bud, sepals, carpels, young and mature fruits and leaves, whereas CtrTB1 is expressed in all floral whorls and is not expressed in leaves (Figure 5). CIN-like homologs also exhibit different expression patterns in C. trianae and H. decumbens. HydCIN1, HydCIN2 and HydCIN3 are expressed in the floral bud, carpels, fruits and leaves. Only HydCIN2 expression is extended to sepals, petals and stamens at very low levels. Interestingly both HydCIN2 and HydCIN3 lack the miR319 binding site. HydCIN4 expression is restricted to fruits and leaves. CtrCIN1, CtrCIN2, CtrCIN5, and CtrCIN6 are expressed in sepals, petals, lip, gynostemium, and ovary, although, CtrCIN2 expression in sepals and gynostemium occurs at low levels. CtrCIN3 and CtrCIN4 have similar expression patterns to CtrCIN1, however, they are poorly expressed in the ovary and are the only CIN-like homologs that extend their expression to leaves.

The CYC/TB1-like homologs have very different expression patterns in C. trianae and H. decumbens. HydTB1 is expressed in the floral bud, sepals, carpels, young and mature fruits and leaves, whereas CtrTB1 is expressed in all floral whorls and is not expressed in leaves (Figure 5). CIN-like homologs also exhibit different expression patterns in C. trianae and H. decumbens. HydCIN1, HydCIN2 and HydCIN3 are expressed in the floral bud, carpels, fruits and leaves. Only HydCIN2 expression is extended to sepals, petals and stamens at very low levels. Interestingly both HydCIN2 and HydCIN3 lack the miR319 binding site. HydCIN4 expression is restricted to fruits and leaves. CtrCIN1, CtrCIN2, CtrCIN5, and CtrCIN6 are expressed in sepals, petals, lip, gynostemium, and ovary, although, CtrCIN2 expression in sepals and gynostemium occurs at low levels. CtrCIN3 and CtrCIN4 have similar expression patterns to CtrCIN1, however, they are poorly expressed in the ovary and are the only CIN-like homologs that extend their expression to leaves.

The CYC/TB1-like homologs have very different expression patterns in C. trianae and H. decumbens. HydTB1 is expressed in the floral bud, sepals, carpels, young and mature fruits and leaves, whereas CtrTB1 is expressed in all floral whorls and is not expressed in leaves (Figure 5). CIN-like homologs also exhibit different expression patterns in C. trianae and H. decumbens. HydCIN1, HydCIN2 and HydCIN3 are expressed in the floral bud, carpels, fruits and leaves. Only HydCIN2 expression is extended to sepals, petals and stamens at very low levels. Interestingly both HydCIN2 and HydCIN3 lack the miR319 binding site. HydCIN4 expression is restricted to fruits and leaves. CtrCIN1, CtrCIN2, CtrCIN5, and CtrCIN6 are expressed in sepals, petals, lip, gynostemium, and ovary, although, CtrCIN2 expression in sepals and gynostemium occurs at low levels. CtrCIN3 and CtrCIN4 have similar expression patterns to CtrCIN1, however, they are poorly expressed in the ovary and are the only CIN-like homologs that extend their expression to leaves.
DISCUSSION

Most functional studies on TCP genes have focused on identifying their contribution to floral symmetry and plant architecture, as expected by the functions of the canonical CYC and DICH from A. majus and AtTCP1 from A. thaliana respectively (Luo et al., 1996, 1999; Costa et al., 2005). Studies on the evolution of TCP transcription factors have concentrated on core eudicots and particularly on CYC2 homologs (Damerel and Manuel, 2003; Howarth and Donoghue, 2006; Preston and Hileman, 2009; Martin-Trillo and Cubas, 2010; Mondragón-Palomino and Trontin, 2011; Sarvepalli and Nath, 2011; Danisman et al., 2012, 2013; Uberti-Manassero et al., 2012; Aguilar-Martínez and Sinha, 2013; Das Gupta et al., 2014; Lin et al., 2016). Our matrix includes sampling from the Phalaenopsis genome as well as all transcriptomes available for Asparagales. The phylogenetic analysis made with the full-length coding sequences, allowed us to identify a number of large scale as well as local TCP gene duplications and changes in protein sequences linked to these duplications. Moreover, this is the first large scale analysis of CIN-like and PCF-like genes. We are able to report a comparative expression pattern in two Asparagales species, representing the two floral groundplans in the order, and present hypotheses on the putative role of these genes in floral patterning in representative Asparagales.

Asparagales CYC/TB1 Homologs Are Found Predominantly as Single Copies and Have Divergent Expression Patterns in Hypoxis decumbens and Cattleya trianae

Our study detected a single copy CYC/TB1 gene in each of the species investigated in the Asparagales. The Asparagales CYC/TB1 homologs fall outside of the Poales (TB1/REP clades) or Commelinales identified local duplications, described before and recovered here (Doebley et al., 1995, 1997; Yuan et al., 2009; Mondragón-Palomino and Trontin, 2011). The tree topology suggests that independent duplications have occurred in CYC-like genes in the monocots (Yuan et al., 2009; Mondragón-Palomino and Trontin, 2011; Hileman, 2014). Species-specific duplications in Asparagales were found only in Phalaenopsis (Orchidaceae). Moreover, expression data of CYC/TB1 genes in Asparagales show remarkable differences between H. decumbens and C. trianae, even though sampled organs correspond to fairly well developed tissues. CtrTB1 is expressed homogeneously in all floral whors while HydTB1 is expressed predominantly in carpels, fruits and leaves (Figure 5). Homogeneous expression of CYC-like genes in dorsal and ventral floral organs in C. trianae suggest that CtrTB1 is likely not playing important roles in maintenance of bilateral symmetry in orchid flowers (see also Horn et al., 2015). However, only pre-anthethic floral buds were sampled and earlier stages are needed to test whether CtrTB1 can be contributing to the establishment of bilateral symmetry in the flower primordia. Comparative expression studies of CYC/TB1-like genes in other Orchidaceae point to significant variations. For instance, while the O. italica homolog OitaTB1 is only expressed in leaves (De Paolo et al., 2015), two CYC/TB1 genes in P. equestris do exhibit differential dorsiventral expression in the floral bud suggesting species specific roles in bilateral symmetry establishment (Lin et al., 2016).

Expression of CYC/TB1 genes in Orchidaceae contrasts with studies in Commeliniales, Zingiberales, and Alstroemeriae that show differential dorsiventral expression of CYC/TB1 genes and hence support convergent recruitment of CYC homologs in the acquisition of bilateral symmetry in different monocots (Bartlett and Specht, 2011; Preston and Hileman, 2012; Hoshino et al., 2014). Within core eudicots, only the CYC2/TCP1 clade members have been linked to shifts toward bilateral floral symmetry. This has been extensively documented for CYC and DICH in A. majus (recent paralogs within the CYC2/TCP1 clade) which are expressed in the dorsal regions of the floral meristems and negatively regulate cell proliferation (Luo et al., 1996, 1999). However, many other CYC2 orthologs have been shown to control bilateral symmetry in Asterales, Brassicales, Dipsacales, Fabales, Lamiales, and Malpighiales (Busch and Zachgo, 2007; Gao et al., 2008; Preston et al., 2009; Wang et al., 2010; Zhang et al., 2010, 2013; Howarth et al., 2011; Tähtiharju et al., 2012; Yang et al., 2012; Ma et al., 2016).

Recruitment of CYC2 homologs in bilateral symmetry is likely facilitated by conserved protein-protein interactions mediated by the LxxLL motif (Heery et al., 1997; Damerel and Manuel, 2003; Reeves and Olmstead, 2003; Howarth and Donoghue, 2006; Li et al., 2009; Preston et al., 2009; Tähtiharju et al., 2012; Parapunova et al., 2014; Ma et al., 2016). If so, it is possible that different CYC-like proteins can form specific homo- and heterodimers, or even have unique partners, and thus protein motifs can provide clues to protein affinity and functional specificity (Kosugi and Ohashi, 2002). Our MEME analysis rescues motifs 6, 9 and 12, shared only between P. equestris PETCP11715 and the A. majus AmCYC and AmDICH, which are not present in other orchid CYC/TB1 proteins (i.e., C. trianae and O. italica) and hence putatively involved in establishing early bilateral floral symmetry in some Orchidaceae species (Supplementary Figure 1). When full length CtrTB1 is compared to the three P. equestris CYC/TB1 proteins very little similarity is detected (PETCP11715–0.38; PETCP06750–0.22; PETCP06750–0.24); for instance, CtrTB1 lacks motif 11 (R-domain). Nevertheless, the TCP domain is extremely conserved (0.91, 0.80, and 0.84 respectively). These results suggest that besides the key bHLH amino acids that target conserved genes, there are likely important motifs in the flanking regions allowing unique interactions and downstream partners in each species.

Conversely, the expression of HydTB1 in H. decumbens is indicative of exclusive roles in carpels, fruits and leaves. This data suggests that HydTB1 may have similar roles to other CYC genes that do not participate in establishing floral bilateral symmetry. Such is the case of the A. thaliana, AtTCP1, which promotes shoot growth and regulates leaf lamina size (Costa et al., 2005; Guo et al., 2010; Koyama et al., 2010b). To date less attention has been given to the putative role of CYC genes in the development of leaves, carpels and fruits, despite the fact that expression has been detected in these organs in other core eudicots. This is the case for CYC homologs in Citrullus, Gerbera, Gossypium, Lotus, Solanum, and some Papaveraceae that have been detected in seedlings,
young leaves, and immature fruits (Damerval et al., 2007; Wang et al., 2010; Parapunova et al., 2014; Ma et al., 2016; Shi et al., 2016).

**CIN-like Genes Have Undergone Numerous Duplications in Angiosperms, Monocots and Orchidaceae and Show Broad Expression Patterns in Cattleya trianae When Compared to Hypoxis decumbens**

Here we show the first comprehensive phylogenetic analysis of CIN-like genes. Our results point to the occurrence of at least one duplication event predating angiosperm diversification, at least one duplication occurring prior to the origin of the Asparagales and one specific duplication prior to the diversification of Orchidoideae and Epidendroideae (Figure 3, Supplementary Figure 2; Floyd and Bowman, 2007; Martin-Trillo and Cubas, 2010). By comparison to CYC genes, CIN genes functional characterization is restricted to model species only. The canonical CINCINNATA in A. majus has dual roles in limiting the growth of leaf margins while promoting epidermal cell differentiation in petals (Nath et al., 2003; Crawford et al., 2004). The cin mutant in A. majus, as well as the tcp4 mutant in A. thaliana, exhibit curly leaves as a result of excessive growth in leaf margins (Crawford et al., 2004; Koyama et al., 2010a). CIN regulates leaf shape through direct or indirect negative regulation of the boundary CUP-SHAPED COTYLEDON (CUC) genes, likely via the activation of ASSYMMETRIC LEAVES 1 (AS1), miR164, INDOLE-3-ACETIC ACID/SHORT HYPOCOTYL2 (IAA3/SHY2), and SMALL AUXIN UP RNA (SAUR) (Koyama et al., 2007, 2010a). CIN also controls leaf development through the regulation of cell proliferation by activating miR396, CYCLIN-DEPENDENT KINASE INHIBITOR/KIP RELATED PROTEIN 1 (ICK1/KRP1) and jasmonate biosynthesis (Schommer et al., 2014). Similarly, AtTCP4 regulates the transition between cell proliferation and differentiation, controlling cytokinin and auxin receptors (Efroni et al., 2013; Das Gupta et al., 2014). AtTCP4 also controls leaf senescence, maintains petal growth, and regulates early embryo development and seed viability, and finally it regulates jasmonic acid (JA) biosynthesis by the activation of LIPOXIGENASE2 (LOX2) (Schommer et al., 2008; Nag et al., 2009; Sarvepalli and Nath, 2011; Danisman et al., 2012). Leaf patterning is also controlled by CIN-like homologs in tomato and rice (Ori et al., 2007; Yang et al., 2013; Zhou et al., 2013; Ballester et al., 2015).

In addition to the roles of CIN genes in leaf patterning, other functions in carpel and fruit development have been identified. AAtTCP2 and AtTCP3 are known to activate NGATHA genes that regulate carpel apical patterning. NGA orthologs from A. thaliana, rice, tomato and bean have conserved putative TCP binding site suggesting that this regulation is conserved in monocots and dicots (Ballester et al., 2015). Moreover, the tcp3 mutant has shorter silique with a wrinkled surface (Koyama et al., 2007; Ballester et al., 2015). Additionally, expression analyses in different Solanaceae species suggest an important contribution of these genes to fruit development and maturation, as they are downstream targets of key ripening regulators including RIPENING INHIBITOR (RIN), COLORLESS NON-RIPENING (CNR) and SIAP2a (Supplementary Figure 3; Crawford et al., 2004; Parapunova et al., 2014). More recently PeCIN8, a P. equestris CIN-like homolog, was shown to be broadly expressed and to have roles in leaf cell proliferation, determining the fruit final size and controlling proper embryo and ovule development (Lin et al., 2016). In summary, CIN-like genes are pleiotropic regulators of cell division and differentiation in leaves, petals, carpels, ovules, fruits and seeds across angiosperms.

This study identified significant changes in the helix I residues and the loop between Orchidaceae sequences and other Asparagales CIN-like proteins (Figure 2). For instance, a number of motifs involved in protein interaction, including the R domain characteristic of class II TCP proteins were only identified in the CIN1a clade, but are absent in all other paralogous clades (Supplementary Figure 1, Cubas et al., 1999; Damerval and Manuel, 2003). Here we have also identified conserved motifs like the mir319 binding site, previously reported in CIN-like homologs from O. italica (Nag et al., 2009; De Paolo et al., 2015), for all CIN-like sequences in Asparagales. This indicates, that Asparagales homologs are also regulated by mir319, similar to AtTCP4, and other CIN-like Arabidopsis paralogs, including AtTCP2/3/10 and 24 (Palatnik et al., 2003; Koyama et al., 2007; Schommer et al., 2008, 2012; Koyama et al., 2010a; Danisman et al., 2012).

In comparison with reported functional data from model eudicots and monocots, the CIN-like differential expression observed in C. trianae and H. decumbens homologs points to three testable hypotheses. (1) While in H. decumbens all CIN-like genes are expressed in young leaves and may play roles in leaf development, only two of the six paralogs identified in C. trianae (CtCIN3/4) are likely involved in cell division and differentiation during leaf development, similar to the P. equestris PeCIN8 (Lin et al., 2016). (2) While CIN-like genes are likely playing key roles in both perianth and fertile organs development and growth in C. trianae, their contribution to perianth development and growth is less clear in H. decumbens, perhaps only with HydCIN2 involved in cell proliferation in the perianth. (3) In both species CIN-like genes are strongly expressed in carpels and fruits, suggesting that their role in carpel patterning, ovule development as well as fruit maturation is likely conserved in Asparagales. Nevertheless, mRNA expression data for all CIN-like genes must be interpreted with caution given the putative conserved mir139 posttranscriptional regulation.

**PCF-like Genes Are Extensively Duplicated and Have Overlapping Expression Patterns with CIN-like Genes in Asparagales**

Our results on the evolution of PCF-like genes points to numerous duplication events within Asparagales in comparison to the CYC-like and the CIN-like clades. In addition, the topology recovered suggests that PCF-like gene duplications in monocots are independent from the one that occurred in core eudicots (Figure 4, Supplementary Figure 2). Functional data available for PCF-like genes suggest redundancy with CIN-like
genes (Aguilar-Martinez and Sinha, 2013; Danisman et al., 2013). For instance, both PCF-like and CIN-like genes control leaf development through regulation of LOX2. However, while AtTCP20 (PCF-like) inhibits, AtTCP4 (CIN-like) induces the expression of LOX2 (Danisman et al., 2012). PCF-like genes are also involved in the regulation of circadian clock genes, that is the case of CCA1 Hiking Expedition CHE (AtTCP21) (Pruneda-Paz et al., 2009; Giraud et al., 2010). In addition dimers formed between AtTCP15 and other TCP-like proteins (AtTCP2, AtTCP3, AtTCP11) are known to regulate circadian cycles, cell proliferation in floral organs and leaves, and to promote seed germination (Koyama et al., 2007; Kieffer et al., 2011; Resentini et al., 2015). PCF-like genes are also expressed in ovule, seed and fruit development in Solanum lycopersicum, Solanum tuberosum, O. sativa and in P. equestris, suggesting that they mediate cell proliferation in the carpel to fruit transition in both eudicots and monocots (Supplementary Figure 3; Kosugi and Ohashi, 1997; Yao et al., 2007; Parapunova et al., 2014; Lin et al., 2016).

Expression detected here of PCF-like homologs in C. trianae and H. decumbens show broad expression of most paralogs in C. trianae (except CtrPCF7) contrasting with a restricted expression of most paralogs in H. decumbens (except HydPCF5; Figure 5). Such expression patterns are in accordance with all putative functions identified in other monocots and in core eudicots including, but not restricted to, cell proliferation control in leaves, carpels, ovules, fruits and seeds (Lin et al., 2016). All data available point to pleiotropic roles of TCP-like genes with a high degree of redundancy among paralogs. Interestingly, when compared to CYC-like and CIN-like genes, the PCF-like gene HydPCF5 in H. decumbens is more likely to be playing cell proliferation roles in the perianth.

In conclusion, the Asparagales, unlike core eudicot model plants, have a reduction in the number of CYC-like homologs and an increase of CIN-like and PCF-like copies. Characteristic key amino acids in the bHLH domain, flanking motifs and binding sites are for the most part conserved in Asparagales CIN-like and PCF like proteins suggesting similar conserved mechanisms of post-transcriptional regulation and interacting partners. The most notorious exception to this is the lack of a miR319 binding site in HydCIN2/3. Nevertheless, the CYC-like proteins in Asparagales seem to be poorly expressed and have undergone important shifts in protein domains, suggesting changes in regulation as well as in protein-protein interactions. The orchid included in this study has similar numbers of TCP copies when compared to Asparagales with radially symmetrical flowers, and only the Epidendrioideae and Orchidoideae seem to have an additional CIN paralog lacking in other Orchidaceae and Asparagales. Expression data suggests different roles of TCP-like genes in C. trianae and H. decumbens, pointing to: (1) a possible decoupling of TB1 homologs from bilateral symmetry in C. trianae; (2) conserved roles of CIN-like and PCF-like genes in the control of cell proliferation in carpels, ovules and fruits in both species; and (3) preferential leaf expression of CIN-like and PCF-like genes in H. decumbens and perianth expression in C. trianae. Here we have performed the first large scale analysis of TCP genes in Asparagales which will provide a platform for in depth comparative expression analyses as well as much needed functional studies of these genes in emerging model orchids.

AUTHOR CONTRIBUTIONS

YM and NP planned and designed the research. YM and NP conducted fieldwork. YM, JA, and NP performed experiments. YM, JA, and NP analyzed the data and wrote and approved the final version of the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

We thank the recognition obtained by the Premio Fomento a la Investigación, Alcaldía de Medellín, 2016. We thank the OneKP repository staff (University of Alberta), the Orchidstra staff (Academia Sinica) and Orchidbase staff (National Cheng Kung University) for facilitating access to the online databases. We thank Favio González (Universidad Nacional de Colombia) and Mariana Mondragón-Palomino (Universität Regensburg) for their comments to the manuscript and to preliminary results presented in meetings, respectively. We thank Ricardo Callejas (Universidad de Antioquia) for allowing us to use office and laboratory space and Cecilia Zunajo-Cardona (The New York Botanical Garden) for laboratory assistance. We thank Victor Acosta, Francisco Villegas and the staff at Vivero Sol Rojo for maintaining cultivated plant material of C. trianae. Finally we thank two reviewers for their insightful comments on the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2017.00009/full#supplementary-material

Supplementary Figure 1 | Conserved motifs in and Orchidaceae and non-Orchidaceae Asparagales TCP-like proteins. Model core eudicots and monocots used as reference include Arabidopsis thaliana, Antirrhinum majus, and Oryza sativa. Motifs 1, 2, and 3 correspond to the conserved TCP domain. Motif 11 indicates the characteristic R domain in Class II TCP-like genes. Motif 14 corresponds to the miR319 binding site in CIN-like genes.

Supplementary Figure 2 | ML analysis of TCP-like genes with extended Solanaceae sampling. ML phylogenetic analysis of TCP-like genes with reduced sampling including only model organisms like Arabidopsis thaliana, Oryza sativa, Solanum lycopersicum, Solanum tuberosum, two Orchidaceae species, Cattleya trianae, Orchis italica, one non-Orchidaceae Asparagales, Hypoxis decumbens and the early diverging angiosperm Amborella trichopoda. Yellow stars indicate large scale duplication events. TCP major clades are labeled to the right. Branch and taxa colors correspond to those in the conventions. BS values ≥ 50 are shown.

Supplementary Figure 3 | Expression patterns of TCP-like homologs in Arabidopsis thaliana (Brassicaceae), Solanum lycopersicum and Solanum tuberosum (Solanaceae). (A) Arabidopsis thaliana AtTCP24 (CIN-like). (B) Arabidopsis thaliana AtTCP20 (PCF-like). (C) Solanum lycopersicum StTCP24 (CIN-like). (D) Solanum lycopersicum StTCP11 (PCF-like). (E) Solanum tuberosum StTCP2 (CIN-like). (F) Solanum tuberosum StTCP11 (PCF-like). Expression patterns show here were selected for the genes that showed the broader, most comprehensive expression patterns after comparing expression for all paralogs, and are used to summarize putative expression patterns for each
gene clade. Expression levels vary in intensity between different copies of CIN-like and PCF-like. Taken from the eFP browser, last accessed Sep 5/2016 (http://bar.utoronto.ca/efp_tomato/cgi-bin/efpWeb.cgi and http://bar.utoronto.ca/efp_potato/cgi-bin/efpWeb.cgi).

**REFERENCES**


Broholm, S. (2009). The Role of MADS and TCP Transcription Factors in Gerbera Hybrida Flower Development. Available online at: https://helda.helsinki.fi/handle/10138/22335


**Supplementary Table 1 | List of sequences used in this study.**

**Supplementary Table 2 | Primers used for TCP-like gene expression analyses.**


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

Copyright © 2017 Madrigal, Alzate and Pabón-Mora. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.
TCP Transcription Factors at the Interface between Environmental Challenges and the Plant’s Growth Responses

Selahattin Danisman*
Molecular Cell Physiology, Faculty of Biology, Bielefeld University, Bielefeld, Germany

Plants are sessile and as such their reactions to environmental challenges differ from those of mobile organisms. Many adaptations involve growth responses and hence, growth regulation is one of the most crucial biological processes for plant survival and fitness. The plant-specific TEOSINTE BRANCHED 1, CYCLOIDEA, PCF1 (TCP) transcription factor family is involved in plant development from cradle to grave, i.e., from seed germination throughout vegetative development until the formation of flowers and fruits. TCP transcription factors have an evolutionary conserved role as regulators in a variety of plant species, including orchids, tomatoes, peas, poplar, cotton, rice and the model plant *Arabidopsis*. Early TCP research focused on the regulatory functions of TCPs in the development of diverse organs via the cell cycle. Later research uncovered that TCP transcription factors are not static developmental regulators but crucial growth regulators that translate diverse endogenous and environmental signals into growth responses best fitted to ensure plant fitness and health. I will recapitulate the research on TCPs in this review focusing on two topics: the discovery of TCPs and the elucidation of their evolutionarily conserved roles across the plant kingdom, and the variety of signals, both endogenous (circadian clock, plant hormones) and environmental (pathogens, light, nutrients), TCPs respond to in the course of their developmental roles.

**Keywords:** transcription factor, TCP, development, evolution, plant hormones, signaling

**DISCOVERY OF TCPs – OF PELORIA AND OTHER MUTANTS**

Developmental plasticity is important for plant survival because plants are sessile organisms that have to adapt to suboptimal environmental conditions. It is crucial that these developmental adaptations are balanced, which means that multiple environmental stimuli have to be perceived and weighed against each other before a plant adjusts its growth. Hence, a plethora of regulatory proteins is involved in governing developmental responses to the environment. One family of

**Abbreviations:** bHLH, basic helix-loop-helix; BRC, BRANCHED; CCA1, CIRCADIAN CLOCK ASSOCIATED 1; CIN, CINCINNATA; CUC, CUP-SHAPED COTYLEDON; CYC, CYCLOIDEA; FT, FLOWERING LOCUS T; GA, gibberellic acid; IC51, ISOCHORISMATE SYNTHASE 1; IDR, intrinsically disordered region; jaw-D, JAGGED AND WAVY-D; LA, LANCEOLATE; LHY, LATE ELONGATED HYPOCOTYLI; LOX, LIPOXYGENASE; NLS, nuclear localization signal; PRR, PSEUDO RESPONSE REGULATOR; TB1, TEOSINTE BRANCHED 1; TCP, TEOSINTE BRANCHED 1, CYCLOIDEA, PCF1; TOC1, TIMING OF CAB EXPRESSION1.
transcription factors that is involved in multiple developmental processes are the TEOSINTE BRANCHED 1, CYCLOIDEA, PCF1 (TCP) proteins.

The common toadflax (*Linaria vulgaris*) is a perennial plant with bilateral, zygomorphic flowers that is native to Europe and large parts of northern Asia. When Carl Linnaeus was presented with a common toadflax that did not exhibit zygomorphic but radially symmetric flowers, he called it *peloria* from the Old Greek πελορία (peÍor), which means monster. Linnaeus speculated that this monster was a hybrid between the common toadflax and a thitherto unknown plant and he was surprised to see that this hybrid was nevertheless able to propagate through seeds (Linnaeus and Rudberg, 1744). Whereas his hybrid hypothesis proved to be wrong, he used this case as evidence against immutability, the belief that all species are created at the beginning of the world and are unchanging (Smith, 1821). *Peloria* is a natural variation that occurs in toadflax, snapdragons (*Antirrhinum majus*) (Darwin, 1868) and in foxgloves (*Digitalis purpurea*) (Keeble et al., 1910), amongst other species.

About 250 years later, Luo et al. (1996) isolated the CYCLOIDEA (*CYC*) gene which is only expressed in the dorsal parts of the snapdragon flower and which is responsible for the regulation of zygomorphic flowers. A double mutant of *CYC* and its close homolog DICHTOMA leads to radially symmetric snapdragon flowers (Luo et al., 1996). Cubas et al. (1999b) found that a homolog of the *CYC* gene was also responsible for floral symmetry in the common toadflax. Here, they could show that the *CYC* gene in peloric mutants was extensively methylated and silenced (Cubas et al., 1999b).

At about the same time, Doebley et al. (1995) analyzed two quantitative trait loci that control morphological differences between domesticated maize (*Zea mays*) and its wild progenitor teosinte. They found the teosinte branched 1 (*tb1*) mutation, which leads to increased side shoot outgrowth, and showed that the difference between the maize and the teosinte variant of *TB1* lies mainly in the regulatory regions of the gene, i.e., whereas the function remains the same, the expression pattern is different between domesticated maize and teosinte (Wang et al., 1999).

Kosugi et al. (1995) found that two promoter motifs that are important for the transcriptional regulation of the proliferating cell nuclear antigen (PCNA) gene in rice (*Oryza sativa*) were bound by two transcription factors that were designated PCF1 and PCF2 (Kosugi and Ohashi, 1997). Finally, Cubas et al. (1999a) determined that the above described proteins *TB1, CYC* and *PCF1* and *PCF2* share a conserved non-canonical bHLH region, the eponymous TCP domain (Kosugi and Ohashi, 1997).

**FORM AND FUNCTION OF TCP TRANSCRIPTION FACTORS**

Whereas, this review will mainly focus on the evolutionarily conserved roles of TCPs in the regulation of plant development and their interactions with endogenous and environmental signals, it is crucial to understand how they function. TCP transcription factors are divided into two classes, class I and class II TCPs. These classes differ in the composition of their respective NLSs, the length of the second helix in the bHLH domain, and the presence of an arginine-rich domain of unknown functionality outside the bHLH domain (Cubas et al., 1999a). This so-called R domain is not found in class I TCPs and was predicted to form a hydrophilic α-helix or a coiled-coil structure that mediates protein–protein interactions (Lupas et al., 1991; Cubas et al., 1999a).

The basic region of the TCP domain is essential for DNA binding. Replacement of a conserved glycine–proline pair in the basic region by two lysines completely abolished DNA binding activity of TCP4 in electrophoretic mobility shift studies (Aggarwal et al., 2010). Addition of the major groove binding dye methyl green reduced TCP4 binding to DNA, indicating that TCP4 binds to the major groove in double stranded DNA (Aggarwal et al., 2010).

In various experimental approaches, class I and class II TCP proteins have been shown to recognize GC-rich sequences in target gene promoters (Kosugi and Ohashi, 1997; Li et al., 2005; Viola et al., 2011; Danisman et al., 2012). The differences between class I and class II binding preferences are dependent on the presence of glycine or aspartic acid at positions 11 or 15, respectively (Viola et al., 2012). Interestingly, the class I and class II consensus binding site sequences are not mutually exclusive, indicating that at least a subset of potential target genes are targeted by both class I and class II TCP proteins. This led to speculations about a possible antagonistic relation between class I and class II TCPs, where these proteins compete for common target genes and inhibit or activate gene expression depending on which class dominates the target gene promoter (Li et al., 2005). So far, this was shown in one case only, where the *Arabidopsis* class I TCP transcription factor TCP20 binds to the same promoter as the class II TCP4 and regulates the target gene *LIPOXYGENASE2 (LOX2)* in the opposite direction to TCP4 (Danisman et al., 2012). It is likely though that more cases of class I-class II TCP antagonisms will be discovered in the future, as the two classes are frequently discovered to be involved in the same biological processes.

Similar to many transcription factor families, TCPs require dimerization to bind to DNA, as addition of deoxycholate, an inhibitor of protein–protein interactions, to electrophoretic mobility shift assays leads to a reduction of TCP binding to target sequences (Trémousaygue et al., 2003). Dimerization between TCP transcription factors first has been described between PCF1 and PCF2 in rice, which form homo- and heterodimers (Kosugi and Ohashi, 1997). Whereas the homodimer of TCP20 for example does not bind to the promoter of the iron homeostasis regulator *BHLLHE9* in yeast one-hybrid experiments, the TCP20 heterodimer with TCP8 or TCP21 does (Andrianakaja et al., 2014). A systematic yeast two-hybrid approach between *Arabidopsis* TCPs found that many protein–protein interactions are possible between TCPs and that there is a preference to bind to TCPs of the own class, i.e., class I TCPs preferably interact with class I TCPs and class II TCPs preferably interact with class II TCPs (Danisman et al., 2013). Dimerization of TCPs are facilitated by IDR (Valsechi et al., 2013). These are characterized by low compactness, low globularity and higher structural flexibility and are typically present extensively in eukaryotic transcription
factors (Liu et al., 2006). The C-terminal IDR of TCP8 is needed for self-assembly of TCP8 in dimers and higher order complexes. These IDRs potentially facilitate the flexibility of TCPs in the choice of interacting partners and thus increases the number of potential functions TCP transcription factors can be involved in Thieulin-Pardo et al. (2015). TCPs not only interact with TCPs: protein–protein interactions with a plethora of other proteins has been described, including negative regulators of effector-triggered immunity (Kim et al., 2014), components of the circadian clock (Pruneda-Paz et al., 2009, 2014; Giraud et al., 2010), and others (Trémosaygue et al., 2003; Tao et al., 2013).

**EVOLUTIONARY CONSERVED ROLES OF TCPs**

The three eponymous TCP proteins were characterized as regulators of branching, floral symmetry, and the cell cycle (Doebley et al., 1995; Luo et al., 1996). Later, both CYC-like and the PCF-like TCPs were shown to be involved in leaf development (Kosugi and Ohashi, 1997; Palatnik et al., 2003). TCP research since then has focused on these three developmental processes, mainly identifying evolutionarily conserved processes in a wide array of plant species and the role of cell cycle regulation in the observed phenotypes. Recently it became clear however that TCPs are not limited to branching, floral symmetry and leaf development, and neither are they limited to cell cycle mediated regulation of growth. Both will be discussed further below.

TEOSINTE BRANCHED 1, CYCLOIDEA, PCF1 transcription factors belong to an evolutionary conserved family that first appears in fresh water algae of the Charophyta family (Navaud et al., 2007). In the bryophyte Physcomitrella patens, knockout of the TCP transcription factor PpTCP5 leads to increased numbers of sporangia that are attached to a single seta, reminiscent of branching phenotypes of tcp mutants in higher land plants (Ortiz-Ramírez et al., 2016). Hence, control of meristematic activity of axillary meristems with a subsequent effect on branching patterns seems to be an ancient role of TCP transcription factors (Ortiz-Ramírez et al., 2016). Consistent with this finding, branching phenotypes are apparent both in monocot and dicot plant species. Overexpression of the rice OsTB1, an ortholog of maize TB1, led to a strong decrease in tiller number. The number of axillary buds was not affected in these plants but their outgrowth was Takeda et al. (2003). This fits to the observation that it is not the formation of axillary meristems but the outgrowth of these that is affected by TCPs (Braun et al., 2012). This has been shown in peas (Braun et al., 2012), poplar (Muhr et al., 2016), Arabidopsis (Aguilar-Martinez et al., 2007; Poza-Carrión et al., 2007) and potato (Nicolas et al., 2015).

TCP effect on floral development was shown in a wide range of plant species, including Arabidopsis, Antirrhinum, annual candytuft (Iberis amara) (Busch and Zachgo, 2007; Busch et al., 2012), angiosperms like Aristolochia arborea and Saruma henryi (Horn et al., 2015), Gerbera species (Broholm et al., 2008), rice (Yuan et al., 2009), sunflowers (Fambirini et al., 2012), peas (Wang et al., 2008), ragworts (Kim et al., 2008), Morrow’s honeysuckle (Lonicera morrowii) (Howarth and Donoghue, 2006), Knautia macedonica (Berger et al., 2016), and orchids (De Paolo et al., 2015).

Phylogenetic analysis revealed that the CYCLOIDEA-like TCPs underwent two major duplication events that both predate the formation of core eudicots (Howarth and Donoghue, 2006). In Arabidopsis, all three CYC clades are represented by TCP12, TCP1 and TCP18, respectively (Howarth and Donoghue, 2006). Especially the CYC2 clade, represented by TCP1 in Arabidopsis, underwent multiple additional duplications and has been studied for its effect on floral symmetry, as it contains the original CYC gene of Antirrhinum (Howarth and Donoghue, 2006). An interesting side note is that the duplication of the CYCLOIDEA-like TCPs nearly coincides with the major duplication events of the homeotic MADS-box transcription factors APETALA3, AGAMOUS and SEPALLATA, all three of them important factors for the definition of organ identity in flowering plants (Howarth and Donoghue, 2006). This suggests that the genetic components that are important for the definition of floral organs diversified at a similar time as the components that are important for the growth regulation of these. TCP transcription factors have been identified as targets of Arabidopsis APETALA1 and SEPALLATA3 (Kaufmann et al., 2009, 2010), highlighting a possible link between organ identity formation and growth regulation between MADS-box transcription factors and TCPs (Dornelas et al., 2011).

In Antirrhinum, CYC regulates symmetry via the Myb-domain transcription factor RADIALIS (Corley et al., 2005). Overexpression of CYC in Arabidopsis leads to larger petals containing enlarged petal cells (Costa et al., 2005). Regulation of floral growth is not restricted to the CYC-like class II TCPs. In the jaw-D mutant, petal development is different from wild type Arabidopsis (Palatnik et al., 2003). Nag et al. (2009) showed that this depends on miR319 regulation of TCP4. A microRNA-resistant form of TCP4 under the control of an APETALA3 promoter is expressed in floral organs only and leads to dramatically smaller flowers that only consist of carpels and sepals, missing any petals or stamens, whereas the seedlings of these plants look normal (Nag et al., 2009).

The zinc-finger transcriptional repressor RABBIT EARS controls the expression of the TCPs TCP5, TCP13, and TCP17 and misexpression of both RABBIT EARS and these TCPs leads to aberrant petal development in Arabidopsis (Huang and Irish, 2015). Repression of these TCPs leads to an early step of mitotic activity during petal development (Huang and Irish, 2015). Interestingly the opposite occurs upon downregulation of TCP5, TCP13, and TCP17 in leaves, where leaf cells continue with mitotic divisions for a longer time than in wild type plants (Efroni et al., 2008). Here, the effect of TCP transcription factors on organ development is dependent on the organ-context. This underlines the importance of the regulatory interplay between TCPs and organ identity regulators. While there are hints at this interplay between TCPs and MADS box transcription factors during flower development, such an interplay remains to be shown during the development of other organs (Dornelas et al., 2011).

First indications for a role of TCPs in leaf development comes from work in Antirrhinum (Nath et al., 2003). The Antirrhinum...
class II TCP mutant cin displays crinkly leaves, which are the result of a change in the regulation of the cell cycle during leaf development (Nath et al., 2003). Essentially, mitotic divisions of developing leaf cells in the leaf tip are arrested first and those at the leaf base are arrested last. The result of this successive arresting behavior is a so called arrest front that moves from the leaf tip to the leaf base. The form of this arrest front is different in cin leaves than in wild type leaves, leading to a modified leaf curvature (Nath et al., 2003). In Arabidopsis, similar behavior is observed in the jaw-D mutant (Palatnik et al., 2003). Jaw-D is an overexpressor of the microRNA miR319a in which the CIN-like class II TCPs TCP2, TCP3, TCP4, TCP10, and TCP24 are downregulated (Palatnik et al., 2003). Jaw-D mutants display serrated leaves, abnormal petals and delayed leaf development and senescence (Palatnik et al., 2003). This phenotype derives from delayed leaf development, in which the mitotic arrest front starts later than in wild type plants (Efroni et al., 2008). Recently, it was shown that miR319a-regulated TCP transcription factors act redundantly with NGATHA transcription factors to limit meristematic activity of leaf meristems during leaf development (Alvarez et al., 2016). This phenotype was also apparent in plants expressing an artificial microRNA against the class II TCPs TCP5, TCP13, and TCP17 and the phenotype was extremely strong when these plants were crossed with jaw-D plants (Efroni et al., 2008).

Class II TCPs also regulate leaf development in tomato compound leaves. An ortholog of the Arabidopsis miR319-sensitive TCPs in tomato is LA and it is under the control of the tomato miR319 (Ori et al., 2007). La mutants exhibit simple leaves, whereas overexpression of miR319 without LA insensitivity to the microRNA leads to increased partitioning of the compound leaves. Also, miR319 overexpressing tomato leaves grow 3 months longer than wild type leaves and show the marks of late differentiation, which is a behavior that is identical to Arabidopsis jaw-D plants (Ori et al., 2007; Efroni et al., 2008). Overexpression of miR319 in the monocot Agrostis stolonifera (creeping bentgrass) leads to downregulation of class II TCPs and to the formation of wider and thicker leaves that are different from the wild type (Zhou et al., 2013). This phenotype stems from an increased number of cells in the transgenic bentgrass, similar to jaw-D in Arabidopsis (Efroni et al., 2008; Zhou et al., 2013). In general, expression of CIN-like genes is closely correlated with leaf shapes both in Solanaceae species and in the desert poplar (Populus euphratica) (Shleizer-Burko et al., 2011; Ma et al., 2016).

Expression of TCP3 with a dominant repressor domain led to severe disturbance of Arabidopsis development in all organs (Koyama et al., 2007), involving ectopic shoot formation, serrated leaves, modified sepals and petals, and wavy silique formation. This was due to misexpression of boundary specific genes, i.e., CUC and LATERAL ORGAN BOUNDARIES (Koyama et al., 2007). Also in Antirrhinum, an ortholog of Arabidopsis TCP15 was found to interact with CUPULIFORMIS, a protein that is related to Arabidopsis CUC proteins (Weir et al., 2004). Furthermore, the two Arabidopsis class I TCPs TCP14 and TCP15 were shown to be redundant in affecting cell proliferation during leaf development and in other tissues in Arabidopsis. The most obvious effect though was seen in internode length, which is reduced in tcp14 tcp15 mutants and leads to shorter plants (Kieffer et al., 2011).

Whereas TCP functions have thus been very well-characterized in these branching, flower and leaf development over a wide array of plant species (Figure 1), there are hints that this is just a subset of TCP roles in development. TCPs were shown to be upregulated upon imbibition of dry seeds and germination of tcp14 transposon insertion lines seemed to be lower than in wild type seeds (Tatematsu et al., 2008). Although here, expression of TCP14 in the transposon lines was not necessarily lower than in the wild type, indicating that TCP14 may not be the only cause of the reduced germination rate (Tatematsu et al., 2008). Downregulation of TCP expression in cotton led to reduced cotton hair fiber length as well as a higher of lateral shoots and a stunted growth indicative of a reduced apical dominance (Hao et al., 2012). Overexpression of miR319 in Chinese cabbage not only led to altered leaf development, also the cabbage heads were rounder than in cabbage with low miR319 expression and higher expression of its target gene BrpTCP4-1 (Mao et al., 2014). Heterologous expression of the rice OsTCP19 in Arabidopsis led to a lower number of lateral roots (Mukhopadhyay et al., 2015). In cucumber, mutation of a TCP gene led to a unique plant phenotype. The affected cucumber plants did not develop tendrils but shoots instead. The authors of this study hypothesize that here TCPs not only affect growth of an organ but also determine organ identity (Wang S. et al., 2015). A similar phenotype was found in melons where a single-nucleotide mutation in CmtTCP1 led to the Chiba tendril-less mutation. Also here, the tendrils were converted to shoot and leaf-like structures (Mizuno et al., 2015). This would be the first indication that TCPs can act as organ identity regulators. Further research has yet to uncover whether the function of TCPs in organ identity regulation of tendrils is a unique and novel role or whether other plant organs also need TCPs to define their identity.

TCP FUNCTIONS EFFECT ON THE CELL CYCLE – DIRECT OR INDIRECT?

Early, TCP research focused on the cell cycle as main target of TCP regulation (Kosugi and Ohashi, 1997; Li et al., 2005). Whereas, binding to cell cycle genes has been shown in certain cases (Li et al., 2005; Davière et al., 2014), close analysis of cell division patterns and transcript changes during jaw-D leaf development indicated that the class II-TCP dependent regulation of the cell cycle may be indirect (Efroni et al., 2008). Also binding of the class I TCP TCP20 to cell cycle genes has been shown only once and in vitro (Li et al., 2005), whereas direct target gene analysis indicate that hormone synthesis, especially jasmonate synthesis, is rather directly targeted by TCP20 (Danisman et al., 2012). Both TCP4 and TCP20 affect leaf development via the synthesis of methyl jasmonate, a hormone that has multiple functions in plant development and response to wounding and pathogens (Schommer et al., 2008). Jasmonate, usually known for its role in wounding and pathogen response, does also affect the cell cycle (Świątek et al., 2002).
Jasmonate is not the only plant hormone that may mediate TCP regulation to the cell cycle. The evidence for hormone involvement in TCP-mediated growth regulation accumulated in the recent years (Nicolas and Cubas, 2016). TCP functions have been associated with abscisic acid (Tatematsu et al., 2008; González-Grandio et al., 2013; Mukhopadhyay et al., 2015), auxin (Kosugi et al., 1995; Ben-Gera and Ori, 2012; Uberti-Manassero et al., 2012; Das Gupta et al., 2014), brassinosteroid (Guo et al., 2010), cytokinin (Steiner et al., 2012; Efroni et al., 2013), GA (Yanai et al., 2011; Das Gupta et al., 2014; Davière et al., 2014), jasmonic acid (Schommer et al., 2008; Danisman et al., 2012), salicylic acid (Wang X.et al., 2015), and strigolactone signaling pathways (Dun et al., 2012; Hu et al., 2014) (Figure 2).

Apart from hormonal control of growth, TCP transcription factors are also involved in other biological processes that in turn affect growth. For example, binding sites of TCP transcription factors have been identified in the promoters of CYTOCHROME C1 and 103 genes that are encoding components of the mitochondrial oxidative phosphorylation machinery and protein biogenesis (Welchen and Gonzalez, 2006). The authors of this study proposed that the TCP transcription factors binding these sites coordinate mitochondria genesis and function with growth in new organs (Welchen and Gonzalez, 2006). Another
study showed these genes contain a GGGC(C/T) element in their promoters which is important for diurnal regulation of their gene expression (Giraud et al., 2010). These promoters are bound by TCP transcription factors, implying a role in diurnal regulation of transcripts of the mitochondrial oxidative phosphorylation machinery (Giraud et al., 2010). Earlier TCP21 was found to bind to the promoter of the core clock gene CCA1 and regulate its expression (Pruneda-Paz et al., 2009). TCP21 serves as an inhibitor of CCA1 during the day and dimerization of TOC1 with TCP21 abolishes its binding to the CCA1 promoter. In a double mutant with the clock gene LHY, tcp21/lhy greatly reduces the period of CCA1 expression (Pruneda-Paz et al., 2009). Not only TCP21, other TCPs have also been found to bind to CCA1 in yeast based studies and co-immunoprecipitation experiments (Giraud et al., 2010; Pruneda-Paz et al., 2014). A recent study also showed that TCP20 and TCP22 act as activators of CCA1 in the morning, fulfilling an important role in the circuity of the circadian clock (Wu et al., 2016) (Figure 3). Altogether, it becomes clear that TCPs not only affect growth via the cell cycle. Instead, they act in different biological processes that directly or indirectly affect growth.

**MEDIATING ENVIRONMENTAL SIGNALS INTO GROWTH RESPONSES**

This picture becomes even more complex, as TCPs also mediate environmental signals into growth responses. TCPs were found to be involved in pathogen defense. First, an extensive study showed that both Pseudomonas syringae and Hyaloperonospora arabidopsidis infection led to reduction of TCP14 protein (Mukhtar et al., 2011). Secreted proteins from pathogenic bacteria transferred by the Aster leafhopper (Macrosteles quadrilineatus) to Arabidopsis were able to dimerize with and destabilize TCP2, TCP4, and TCP7 proteins, comprising both classes of TCP transcription factors (Sugio et al., 2011, 2014). Overexpression of the responsible phytoplasma protein SECRETED ASTER YELLOWS-WITCHES BROOM PROTEIN 11 in Arabidopsis destabilizes TCP2, TCP3, TCP4, TCP5, TCP10,
TCP13, TCP17, and TCP24 and leads to jaw-D-like phenotypes (Sugio et al., 2011). Additionally, jasmonic acid levels in infected Arabidopsis leaves are significantly reduced in comparison with untreated leaves, indicating that the plant’s defense mechanisms are reduced upon infection by the pathogen. A similar effect has been found in apples, where the plant pathogen Candidatus Phytoplasma mali binds to two TCP transcription factors and induces morphogenetic changes that co-occur with reduction of jasmonic acid, salicylic acid, and abscisic acid levels (Janik et al., 2016). Further studies identified the class I TCPs TCP8 and TCP9 as important factors for the expression of ICS1, which encodes for a key enzyme in salicylic acid synthesis (Wang X.et al., 2015). In another study, TCP21 has been identified to bind to the promoter of ICS1 and induction of ICS1 expression by salicylic acid is blocked in tcp21 mutants (Zheng et al., 2015). Class I TCPs also interact with proteins known to regulate ICS1 expression, i.e., the transcription factors WRKY28, NAC019 and ETHYLENE INSENSITIVE 1 and the calmodulin binding protein SYSTEMIC ACQUIRED RESISTANCE DEFICIENT 1. Consequently, the tcp8 tcp9 double mutant shows increased sensitivity to infection with Pseudomonas syringae pv. maculicola ES4326 (Wang X.et al., 2015). TCP transcription factors partially control pathogen defense via a second pathway, i.e., by antagonizing the effect of SUPPRESSOR OF rps4-RLD1, a protein that negatively regulates effector-triggered immunity in Arabidopsis (Kim et al., 2014). Lack of TCPs in the triple mutant tcp8 tcp14 tcp15 leads to increased growth of Pseudomonas syringae DC3000 when compared to wild type plants (Kim et al., 2014).

Recent studies showed that TCP transcription factors regulate flowering time. A knockout of the class I TCP transcription factors...
factor TCP23 led to earlier flowering than the wild type, whereas TCP23 overexpressing lines showed delayed flowering behavior (Balsemão-Pires et al., 2013). The floral transition of axillary meristems in *Arabidopsis* is controlled by an interaction between the flowering time proteins FT and TWIN SISTER OF FT and BRC1 (Niwa et al., 2013). The protein–protein interactions between these transcription factors have been shown in yeast two-hybrid, bimolecular fluorescence complementation, and *in vitro* pull-down assays (Niwa et al., 2013). As brc1 mutants exhibit accelerated flowering and ft and twin sister of ft mutants exhibit slower flowering of axillary meristems, respectively, it seems that there is an antagonistic relationship between BRC1 and the flowering time proteins (Niwa et al., 2013). It is likely that dimerization of BRC1 with FT and TWIN SISTER OF FT represses their function in axillary meristems (Niwa et al., 2013). The apple FT orthologs MdFT1 and 2 were also found to interact with TCP transcription factors (Mimida et al., 2011). Overexpression of the tomato miR319 led to flowering with fewer leaves than in wild type tomato and it was shown that LA binds to the promoters of the tomato *APETALA1* and *FRUITFUL* orthologs (Burko et al., 2013).

Perception of the red to far-red light ratio (R:FR) informs a plant of shading by neighboring vegetation and a lower R:FR ratio leads to suppressed axillary meristem outgrowth, allowing the plant to invest in a longer hypocotyl and eventually avoid the shading. In *Arabidopsis*, hypocotyl elongation is regulated via the bHLH transcription factor PHYTOCHROME INTERACTING FACTOR 4, which among others activates *YUCCA8* expression to promote cell elongation (Sun et al., 2012). *YUCCA2, 5, and 8* are also direct target genes of TCP4. In fact, induced overexpression of TCP4 leads to elongated hypocotyls and this effect is dependent on both auxin and brassinosteroid signaling (Challa et al., 2016). In potato, *BRC1* regulation is dependent on the R:FR. BRC1 comes in two forms: the short form (BRC1aS) and the alternatively spliced long version (BRC1aL). Both result in proteins but the shorter
form is cytoplasmatic and does not bind to target genes in the nucleus. The ratio between these two forms changes upon decapitation of potato shoots, exposure to darkness, and under low R:FR conditions (Nicolas et al., 2015). Whereas decapitation leads to a relative increase in BRC1a5, darkness and low R:FR treatments lead to a relative increase in BRC1aL content. The longer BRC1aL protein subsequently inhibits axillary branch elongation in potato shoots and stolons (Nicolas et al., 2015). Arabidopsis brc1 and brc2 (tcp12 and tcp18) show a reduced response to R:FR and the response is abolished in the brc1 brc2 double mutant (González-Grandio et al., 2013). TCP transcription factors are also involved in axillary bud outgrowth of Petunia. Here, GtTCP3 acts in conjunction with DECREASED APICAL DOMINANCE 2, a receptor protein that normally inactivates strigolactones in response to decreased R:FR (Drummond et al., 2015). Rice OsTCP15 is involved in the mesocotyl elongation in response to darkness and responds to strigolactone and cytokinin treatments, outlining the interplay between TCPs and different plant hormones in developmental regulation that is responsive to the environment (Hu et al., 2014).

Viola et al. (2013) showed that class I TCPs contain a conserved cysteine-20 which is sensitive to treatments by oxidants in a dose-dependent manner. This redox-dependent behavior of TCP15 is important for its effect in anthocyanin biosynthesis. A mutant in which the cysteine-20 of TCP15 was replaced by a serine accumulates less anthocyanin under high light stress than wild type plants (Viola et al., 2016). Plant extracts from TCP15 overexpressing plants showed that exposure to prolonged high light conditions leads to an abolishment of TCP15 DNA-binding activity in vivo, mirroring the in vitro phenotype (Viola et al., 2013, 2016). Thus, TCP15 function is reactive to high light input. While the anthocyanin response is not a direct developmental response, further analysis may show that there is a developmental effect.

Not only light affects TCPs, also other signals are perceived and lead to TCP-mediated growth regulation. For example, Guan et al. (2014) showed that TCP20 is involved in nutrient foraging of Arabidopsis roots. In split-root experiments wild type Arabidopsis develops an increased number of lateral roots in medium containing high nitrte concentrations (i.e., 5 mM NO$_3^-$) and close to no lateral roots in medium containing low nitrate concentrations (i.e., 0 mM NO$_3^-$). Tcp20 plants do not exhibit this behavior, indicating that the regulation of root foraging is under the control of TCP20 (Guan et al., 2014). Interestingly, TCP20 transcript levels are not under the control of nitrate levels, indicating that TCP20 is regulated on protein level, either by forming specific protein–protein dimers in the case of nitrate deficiency or via another regulatory mechanism. In rice, the transcript of the class I TCP OsTCP19 is upregulated during salt stress and water-deficit treatments (Mukhopadhyay et al., 2015). Heterologous OsTCP19 overexpression in Arabidopsis leads to reduced numbers of lateral roots but increased abiotic stress tolerance, i.e., plants grew better on Mannitol-containing medium and recovered better after water-deficit treatments. Here, LOX2 expression was reduced in the OsTCP19 overexpressers and ABA signaling genes were upregulated (Mukhopadhyay et al., 2015). Recent experiments revealed up- and down-regulation of several TCPs in Arabidopsis under osmotic stress, although a functional analysis of their role in response to osmotic stress has not been done yet (Kumar et al., 2016). In summary, these few results are first indications that TCPs are no mere static regulators of development, but that they do directly translate environmental signals into growth regulation (Figure 4).

CONCLUSION AND OUTLOOK

TCP transcription factors play a role in a multitude of growth processes over a wide range of plant species (Figure 1). They affect growth directly via the cell cycle and indirectly via influencing plant hormonal signaling and the circadian clock (Figures 2 and 3). Additionally, recent discoveries link TCP-controlled growth responses with environmental signals such as R:FR, high light stress, salt stress or the presence or absence of nutrients.

TCP transcription factors are involved in so many important developmental processes and interact with so many plant hormones that it is likely that future plant research will also uncover a lot more signals that TCPs react to. This will also mean that future TCP research will have to more closely elucidate how the interaction of TCPs with different signaling networks is regulated to ensure a measured response to environmental challenges. This research will have to uncover the roles of dimerization, transcriptional and post-transcriptional regulation as well as post-translational modifications in controlling and ensuring specific TCP functions in plant development.

Plant pathogens are targeting TCP transcription factors to manipulate plant architecture in their favor. If plant pathogens use TCPs in their best interests, maybe so should we. TCP transcription factors will be valuable tools in optimizing plant architecture and hardening plants in response to environmental challenges.

AUTHOR CONTRIBUTIONS

SD drafted, wrote and critically revised the article.

FUNDING

Work in the author’s lab is supported by the Bielefeld Young Researcher’s Fund and the core grant of Bielefeld University to D. Staiger.

ACKNOWLEDGMENTS

We acknowledge support for the Article Processing Charge by the Deutsche Forschungsgemeinschaft and the Open Access Publication Fund of Bielefeld University.
REFERENCES


Danisman TCPs between the Environment and Plant Growth


**Conflicts of Interest Statement:** The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Copyright © 2016 Danisman. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.**
A Conserved Carbon Starvation Response Underlies Bud Dormancy in Woody and Herbaceous Species

Carlos Tarancón, Eduardo González-Grandío, Juan C. Oliveros, Michael Nicolas and Pilar Cubas

Plant shoot systems give rise to characteristic above-ground plant architectures. Shoots are formed from axillary meristems and buds, whose growth and development is modulated by systemic and local signals. These cues convey information about nutrient and water availability, light quality, sink/source organ activity and other variables that determine the timeliness and competence to maintain development of new shoots. This information is translated into a local response, in meristems and buds, of growth or quiescence. Although some key genes involved in the onset of bud latency have been identified, the gene regulatory networks (GRNs) controlled by these genes are not well defined. Moreover, it has not been determined whether bud dormancy induced by environmental cues, such as a low red-to-far-red light ratio, shares genetic mechanisms with bud latency induced by other causes, such as apical dominance or a short-day photoperiod. Furthermore, the evolution and conservation of these GRNs throughout angiosperms is not well established. We have reanalyzed public transcriptomic datasets that compare quiescent and active axillary buds of Arabidopsis, with datasets of axillary buds of the woody species Vitis vinifera (grapevine) and apical buds of Populus tremula x Populus alba (poplar) during the bud growth-to-dormancy transition. Our aim was to identify potentially common GRNs induced during the process that leads to bud para-, eco- and endodormancy. In Arabidopsis buds that are entering eco- or paradormancy, we have identified four induced interrelated GRNs that correspond to a carbon (C) starvation syndrome, typical of tissues undergoing low C supply. This response is also detectable in poplar and grapevine buds before and during the transition to dormancy. In all eukaryotes, C-limiting conditions are coupled to growth arrest and latency like that observed in dormant axillary buds. Bud dormancy might thus be partly a consequence of the underlying C starvation syndrome triggered by environmental and endogenous cues that anticipate or signal conditions unfavorable for sustained shoot growth.

Keywords: bud dormancy, carbon starvation, gene regulatory networks, shoot architecture, plant evolution and development
INTRODUCTION

Shoot branching patterns define overall above-ground plant architecture. In angiosperms, shoots are formed from auxiliary meristems initiated at the base of leaves. These meristems grow and develop into axillary buds that contain, preformed, most of the elements of adult branches (shoot meristems, leaf primordia, reproductive meristems). Axillary buds can enter a quiescent state, rather than growing out immediately to give a branch. In this latent or dormant state their metabolic activity and cell division are very limited (Shimizu and Mori, 1998; Rutting et al., 2007). Bud dormancy and bud activation are influenced by environmental signals such as nutrient and water availability, light quality, day-length and temperature, and by endogenous signals such as sink/source organ activity and hormone signaling. Once dormant, buds require changes in specific developmental and/or environmental cues to resume growth and generate an elongated branch. These cues are monitored in different organs, and inform the plant as to when develop new shoots. This information is transduced to the bud and translated into a gene response that leads to quiescence or growth activation (Rameau et al., 2015).

Bud dormancy is therefore an adaptive trait that allows plants to endure adverse situations until conditions are favorable for development of new shoots. It has great impact on plant reproductive success and productivity, and on survival in temperate woody species. Evolution of this trait might have allowed plants to colonize habitats with fluctuating conditions not always suitable for sustained, uninterrupted growth. Depending on the type of stimulus that promotes growth arrest, Lang et al. (1985, 1987) distinguished three types of bud dormancy. When dormancy is induced by environmental factors, it is termed bud dormancy. When dormancy is induced by environmental arrest, Lang et al. (1985, 1987) distinguished three types of growth. Depending on the type of stimulus that promotes growth conditions not always suitable for sustained, uninterrupted transitions between different dormant states throughout the year. Paradoxical dormancy in response to changes in daylength and temperature. Chilling promotes transition from endo- to ecodormancy, after which the buds are susceptible to growth in response to mild temperatures (Rohde and Bhalerao, 2007).

Transcriptomic studies have been carried out in several herbaceous and woody species to define expression changes in buds during the transitions into and out of different types of dormancy, in response to changes in daylength, light quality, and apical dominance, and in mutant genotypes in which bud growth is affected (e.g., Tatematsu et al., 2005; Rutting et al., 2007; González-Grandio et al., 2013; Reddy et al., 2013; Ueno et al., 2013; Porto et al., 2015). The GRNs that act inside the bud to control the stages leading to dormancy nonetheless remain little known. It is also largely unknown whether different types of dormancy share common underlying genetic mechanisms. Even less is known about the degree of conservation and evolution of the genetic control of this process in different plant species. Comparative analyses to identify common themes among different types of dormancy, or across species, are scarce (González-Grandio and Cubas, 2014; Fennell et al., 2015; Howe et al., 2015; Hao et al., 2017). Such comparisons could help us determine whether eco-, para- and endodormancy are variations of a single ancestral genetic program or whether each type is controlled by unrelated GRNs. It also will help elucidate whether GRNs that cause bud growth arrest are conserved in different herbaceous and woody plant species.

The master regulators that locally control the dormancy onset are also largely unknown. The best characterized are the genes that encode the TCP transcription factors (TF) teosinte branched1 (Tb1, Doebley et al., 1997), BRANCHED1 (BRC1, Aguilar-Martínez et al., 2007; Finlayson, 2007) and their orthologs in mono- and dicotyledonous species, respectively. These widely conserved factors play a very important role in the regulation of para- and ecodormancy in herbaceous plants. These genes are expressed in axillary buds and promote bud dormancy in response to fluctuating environmental cues such as light quality and quantity, and endogenous signals such as apical dominance, sugar availability and hormone signaling (reviewed in Nicolas and Cubas, 2015). In Arabidopsis thaliana, BRC1 controls transcription of several GRNs in buds; one, positively controlled by BRC1, leads to abscisic acid (ABA) accumulation and signaling, whereas another two that are downregulated by BRC1 are enriched in ribosomal protein genes in one case, and in cell division and DNA replication genes in the other (González-Grandio et al., 2013, 2017). Additional GRNs controlled by BRC1 remain to be characterized.

In this study our aim was to identify potentially common GRNs induced during the process that leads to bud para-, eco- and endodormancy. For that we compared publicly available transcriptomic data from active para- and ecodormant axillary buds of Arabidopsis, and found, induced in dormant buds, a shared transcriptomic response typical of tissue undergoing C starvation. We then detected this response also in Populus tremula × Populus alba (poplar) apical buds undergoing endodormancy and in Vitis vinifera (grapevine) axillary buds entering para-, endo- and ecodormancy. This C starvation transcriptional response, activated shortly after exposure to conditions leading to bud dormancy, anticipates and underlies the growth-to-dormancy transition in the three species. The C starvation syndrome entails a suite of interconnected transcriptional responses that include sugar signaling, sugar metabolism reprogramming, senescence, autophagy, catabolism, and ABA and ethylene signaling. It also involves downregulation of cytokinin (CK) signaling, inhibition of anabolism, and repression of protein/DNA synthesis and cell division, conditions typical of cells in dormant buds. This conserved starvation response, genetically connected to cell growth arrest, may be one of the underlying forces driving the growth-to-dormancy transition of axillary buds in response to suboptimal conditions in herbaceous and woody species.
RESULTS

Arabidopsis Bud Dormancy Is Associated With the Induction of Four GRN

Three independent transcriptomic analyses have compared active and dormant buds in Arabidopsis. One study compared the transcriptional profiling of (dormant) buds of intact plants and of (active) buds of decapitated plants at 24 h post-treatment (Tatematsu et al., 2005). Two additional experiments compared the transcripts of active vs. dormant buds of plants exposed to high red-to-far-red light ratio (R:FR, active buds) or low R:FR (dormant buds) (González-Grandio et al., 2013; Reddy et al., 2013). Here we define dormancy as a state in which bud growth is reversibly interrupted, regardless of the requirements to resume development. A search for genes upregulated in dormant buds relative to active buds in the three experiments identified 78 genes (Supplementary Figure S1 and Dataset S1; González-Grandio and Cubas, 2014). These genes correspond to the least common denominator of the three studies and were induced in para- and ecodormant buds by either correlative inhibition or low R:FR, respectively. They were also differentially expressed at 3 h (Reddy et al., 2013), 8 h (González-Grandio et al., 2013) and 24 h (Tatematsu et al., 2005) after treatment onset.

We evaluated the degree of coregulation of these genes using the most updated co-expression database of ATTED-II (15,275 microarray experiments; Obayashi et al., 2007). Hierarchical clustering analysis revealed four clusters of coregulated genes (14, 20, 13, and 31 genes; Figure 1A and Supplementary Dataset S1). We then searched for additional genes coregulated with each cluster using CoExSearch (ATTED-II, Obayashi et al., 2007) and obtained four lists of highly coregulated genes (Supplementary Dataset S1). Analysis of their fold change (FC) induction in the three active-vs.-dormant bud experiments confirmed that a significant proportion of the genes in each list were induced (FC ≥ 1.2) in dormant buds in at least one experiment (Figure 1B and Supplementary Figure S2). We termed the gene lists that comprised the bud dormancy genes of the original clusters plus their coregulated genes (induced in dormant buds in at least one experiment, red dots in Figure 1B) bud dormancy GRNI-IV, with 297, 283, 271, and 295 genes respectively (Supplementary Dataset S1).

Bud Dormancy GRNs Are Related to Hormone Signaling, Stress, Catabolism and Starvation Response

To elucidate the biological processes in which these GRNs were involved, we searched for enrichment in gene ontology (GO) terms using the Panther Classification System (Mi et al., 2017; Supplementary Dataset S2), complemented with a MapMan bin analysis (Thimm et al., 2004; Supplementary Dataset S1). GRNI was significantly enriched in terms related to ethylene, auxin and gibberellin signaling and response; GRNII in terms related to ABA, catabolism and response to abiotic stress; GRNIII in terms related to lipid and amino acid catabolism, senescence, response to starvation and biotic stress; and GRNIV in terms related to protein ubiquitination and response to sucrose starvation.

We evaluated the degree of overlap between these GRNs by seeking common genes. GRNIII and GRNIV shared one-third of their genes; GRNII and GRNIII shared 30%. and GRNI and GRNIV had 26% genes in common (Figure 1C, Supplementary Figure S3 and Dataset S3). This suggested that these GRNs are not strictly independent, but correspond to related aspects of the same syndrome, probably coordinated or maintained by ethylene, auxin and ABA signaling (Figure 1D).

Bud Dormancy GRNs Are Enriched in Genes Typical of a C Starvation Response

We observed that three robust sugar starvation gene markers, GIBBERELLIN-STIMULATED ARABIDOPSIS 6 (GASA6), DORMANCY-ASSOCIATED PROTEIN-LIKE 1 (DRM1/DYL1) and DARK INDUCIBLE 6 (DIN6) (Contento et al., 2004; Price et al., 2004; Gonzali et al., 2006; Zhong et al., 2015) were members of one or several GRNs (GASA6, GRNI; DRM1, GRNII, III and IV; DIN6, GRNIII and IV; Supplementary Dataset S3). As sugar has a prominent role in the control of shoot outgrowth (Mason et al., 2014; Barbier F. et al., 2015; reviewed in Barbier F.F. et al., 2015) and GRNIV was significantly enriched in terms related to sucrose starvation, we studied this response further. The C starvation syndrome, triggered under C-limiting conditions (e.g., an extended night), helps to obtain an alternative energy source and C skeletons. In Arabidopsis, it comprises a suite of interconnected events that result in changes in C balance and growth. They include reprogramming of sugar sensing, transport, signaling and metabolism, increased protein ubiquitination and degradation, amino acid and lipid catabolism, induction of ABA and ethylene signaling, and recycling of cell components via autophagy and senescence. In addition, CK signaling, ribosomal gene expression, DNA synthesis and cell division are inhibited (Contento et al., 2004; Lin and Wu, 2004; Thimm et al., 2004; Gonzali et al., 2006; Rolland et al., 2006; Rose et al., 2006). Remarkably, many of the GO terms and/or MapMan bins enriched in the four GRNs matched categories induced by C-limiting conditions (Table 1 and Supplementary Datasets S1, S2).

To test the possibility that these GRNs correspond to a C starvation response, we compared the GRN genes with four lists of genes induced in C-limiting conditions: (i) 26 genes of a robust core of C-signaling response shared by 21 Arabidopsis accessions (Supplementary Dataset S4; Sulpice et al., 2009), (ii) 57 sugar-responsive genes, proposed upstream components of the transcriptional response to sucrose (Supplementary Dataset S4; Osuna et al., 2007), (iii) 429 dark-induced, sugar-repressed genes (Supplementary Dataset S4; Gonzali et al., 2006) and (iv) 507 genes responsive to AKIN10, a catalytic subunit of the SUCROSE-NON-FERMENTING-1-RELATED PROTEIN KINASE (SnRK1), which integrates stress and C signals to coordinate energy balance, metabolism and growth (Supplementary Dataset S4; Baena-González et al., 2007).
FIGURE 1 | Bud dormancy genes and GRN. (A) Hierarchical clustering representation of bud dormancy genes (González-Grandio and Cubas, 2014) based on their degree of coregulation in 15,275 microarray experiments (ATTED-II; Obayashi et al., 2007). The number of coregulated genes and GO terms enriched are indicated. (B) Volcano plots representing pval (–Log10 pval, vertical axis) and relative expression (Log2 fold change, horizontal axis) of all genes in each microarray. Normalized gene intensities in dormant buds vs. normalized gene intensities in active buds were compared in all experiments [3 h low R:FR (N-2 bud) vs. high R:FR (N-2 bud); 8 h low R:FR vs. High R:FR, intact plants vs. 24 h post-decapitation]. Bud dormancy genes and their coregulated genes are highlighted. In red and green, genes induced and repressed in dormant buds, respectively. Genes highlighted in red were attributed to Bud dormancy GRN-IV (see Supplementary Dataset S1) and were used for subsequent analyses. (C) Venn diagram showing overlap between bud dormancy GRN. Number of common genes is indicated. (D) Model of the relationships between bud dormancy GRN. Line thickness indicates degree of overlap between GRN.
TABLE 1 | Bud dormancy genes from categories related to sugar sensing, transport, signaling and metabolism, protein ubiquitination and degradation, as well as amino acid and lipid catabolism, autophagy and senescence.

<table>
<thead>
<tr>
<th>AGI Name GRN</th>
<th>Name</th>
<th>AGI Name GRN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sugar</td>
<td>Senescence</td>
<td></td>
</tr>
<tr>
<td>Transport</td>
<td>ABA INSENSITIVE GROWTH 1 (ABIG1/HAT22) II-III</td>
<td></td>
</tr>
<tr>
<td>At3g48740</td>
<td>SUCROSE EFLUX TRANSPORTER SWEET1 I</td>
<td>At4g37790</td>
</tr>
<tr>
<td>At5g23660</td>
<td>SUCROSE EFLUX TRANSPORTER SWEET12 I</td>
<td>A1g01720</td>
</tr>
<tr>
<td>At2g43330</td>
<td>INOSITOL TRANSPORTER 1 I-III</td>
<td>A1g69490</td>
</tr>
<tr>
<td>At1g22710</td>
<td>SUCROSE TRANSPORTER 1 (SUT1/SUC2) I-III</td>
<td>A1g39610</td>
</tr>
<tr>
<td>At1g11260</td>
<td>STP1 II-III-I</td>
<td>A1g51070</td>
</tr>
<tr>
<td>At5g61520</td>
<td>STP3 I-II-II-II-III</td>
<td>A1g02380</td>
</tr>
<tr>
<td>At1g77210</td>
<td>STP14 I-II-II-II-III</td>
<td>A1g35770</td>
</tr>
<tr>
<td>At1g08920</td>
<td>SUGAR TRANSPORTER ERD6-LIKE 3 II-III-IV</td>
<td>A1g62300</td>
</tr>
<tr>
<td>Sensing</td>
<td>EXL2 II-III-IV</td>
<td>A1g62180</td>
</tr>
<tr>
<td>At5g64260</td>
<td>EXL4 II-III-IV</td>
<td>A1g04620</td>
</tr>
<tr>
<td>Signaling</td>
<td>ATG6 IV</td>
<td>A1g62040</td>
</tr>
<tr>
<td>At3g21170</td>
<td>Snrkl submit β1 (AKINBETA1) IV</td>
<td>A1g24170</td>
</tr>
<tr>
<td>At3g48530</td>
<td>Snrkl subunit γ1 (KING1) II-III-IV</td>
<td>A1g15650</td>
</tr>
<tr>
<td>At1g68020</td>
<td>TPS6 II</td>
<td>A1g16520</td>
</tr>
<tr>
<td>At1g70290</td>
<td>TPS8 II-III-IV</td>
<td>A1g36620</td>
</tr>
<tr>
<td>At1g23870</td>
<td>TPS9 II-III-IV</td>
<td>A1g54170</td>
</tr>
<tr>
<td>At1g60140</td>
<td>TPS10 II-III-IV</td>
<td>A1g48320</td>
</tr>
<tr>
<td>At2g18700</td>
<td>TPS11 II-III-IV</td>
<td>A1g54620</td>
</tr>
<tr>
<td>At4g24040</td>
<td>TREHALASE III-IV</td>
<td>A1g54730</td>
</tr>
<tr>
<td>Metabolism</td>
<td>6-PHOSPHOGLUCONOLACTONASE 1 IV</td>
<td>A1g17620</td>
</tr>
<tr>
<td>At1g54100</td>
<td>ALDEHYDE DEHYDROGENASE H7B4 II-III</td>
<td>A1g65110</td>
</tr>
<tr>
<td>At3g23920</td>
<td>β-AMYLASE 1 (BAM 1) II</td>
<td>A1g51840</td>
</tr>
<tr>
<td>At2g47180</td>
<td>CITRATE SYNTHASE 3 II</td>
<td>A1g6620</td>
</tr>
<tr>
<td>At2g17800</td>
<td>GALACTINOL SYNTHASE 1 (GOLS 1) II</td>
<td>A1g48320</td>
</tr>
<tr>
<td>At1g56600</td>
<td>GOLS 2 II</td>
<td>A1g18640</td>
</tr>
<tr>
<td>At1g08940</td>
<td>PHOSPHOGLYCERATE MUTASE AT74H III</td>
<td>A1g39400</td>
</tr>
<tr>
<td>At4g15530</td>
<td>PYRUVATE, PHOSPHATE DIKINASE 1 III</td>
<td>A1g73920</td>
</tr>
<tr>
<td>At5g1970</td>
<td>SORBITOL DEHYDROGENASE IV</td>
<td>A1g18460</td>
</tr>
<tr>
<td>At4g02280</td>
<td>SUCROSE SYNTHASE 3 IV</td>
<td>A1g06340</td>
</tr>
<tr>
<td>Lipid degradation</td>
<td>ACYL-COENZYM A OXIDASE 1 (ACX1) III</td>
<td></td>
</tr>
<tr>
<td>At4g16760</td>
<td>ACX2 III</td>
<td>A1g51840</td>
</tr>
<tr>
<td>At2g39400</td>
<td>ACX4 IV</td>
<td>A1g6620</td>
</tr>
<tr>
<td>At4g16760</td>
<td>1,4-DIHYDROXY-2-NAPHTHOYL-COA III</td>
<td>A1g48320</td>
</tr>
<tr>
<td>At1g54620</td>
<td>THIOESTERASE 1 II</td>
<td>A1g18640</td>
</tr>
<tr>
<td>Aminoacid degradation</td>
<td>α/β-Hydrolases superfamily protein I-IV</td>
<td></td>
</tr>
<tr>
<td>At1g55510</td>
<td>2-OXOISOVALERATE DEHYDROGENASE β1 IV</td>
<td>A1g33150</td>
</tr>
<tr>
<td>At4g33150</td>
<td>α-AMINOACIDIC SEMIALDEHYDE SYNTHASE II-III</td>
<td>A1g53970</td>
</tr>
<tr>
<td>At5g54080</td>
<td>HOMOGENTISATE 1,2-DIOXYGENASE II-III</td>
<td>A1g45300</td>
</tr>
<tr>
<td>At2g45300</td>
<td>ISOVALeryl-COA DEHYDROGENASE II-III-IV</td>
<td>A1g64860</td>
</tr>
<tr>
<td>At1g63400</td>
<td>METHIONINE GAMMA-LYASE II</td>
<td>A1g34030</td>
</tr>
<tr>
<td>At1g06340</td>
<td>METHYLcrotonoyL-COA CARBOXYLASE β IV</td>
<td>A1g06340</td>
</tr>
<tr>
<td>Autophagy</td>
<td>ORE9/MAK2 II-IV</td>
<td>A1g06340</td>
</tr>
<tr>
<td>Sensing</td>
<td>ATG18 A IV</td>
<td>A1g06340</td>
</tr>
<tr>
<td>Metabolism</td>
<td>ATG18 F II</td>
<td>A1g06340</td>
</tr>
<tr>
<td>Lipid degradation</td>
<td>ACYL-COENZYM A OXIDASE 1 (ACX1) III</td>
<td></td>
</tr>
<tr>
<td>At4g16760</td>
<td>ACX2 III</td>
<td>A1g51840</td>
</tr>
<tr>
<td>At2g39400</td>
<td>ACX4 IV</td>
<td>A1g6620</td>
</tr>
<tr>
<td>At4g16760</td>
<td>1,4-DIHYDROXY-2-NAPHTHOYL-COA III</td>
<td>A1g48320</td>
</tr>
<tr>
<td>At1g54620</td>
<td>THIOESTERASE 1 II</td>
<td>A1g18640</td>
</tr>
<tr>
<td>Aminoacid degradation</td>
<td>α/β-Hydrolases superfamily protein I-IV</td>
<td></td>
</tr>
<tr>
<td>At1g55510</td>
<td>2-OXOISOVALERATE DEHYDROGENASE β1 IV</td>
<td>A1g33150</td>
</tr>
<tr>
<td>At4g33150</td>
<td>α-AMINOACIDIC SEMIALDEHYDE SYNTHASE II-III</td>
<td>A1g53970</td>
</tr>
<tr>
<td>At5g54080</td>
<td>HOMOGENTISATE 1,2-DIOXYGENASE II-III</td>
<td>A1g45300</td>
</tr>
<tr>
<td>At2g45300</td>
<td>ISOVALeryl-COA DEHYDROGENASE II-III-IV</td>
<td>A1g64860</td>
</tr>
<tr>
<td>At1g63400</td>
<td>METHIONINE GAMMA-LYASE II</td>
<td>A1g34030</td>
</tr>
<tr>
<td>At1g06340</td>
<td>METHYLcrotonoyL-COA CARBOXYLASE β IV</td>
<td>A1g06340</td>
</tr>
<tr>
<td>Autophagy</td>
<td>ORE9/MAK2 II-IV</td>
<td>A1g06340</td>
</tr>
<tr>
<td>Sensing</td>
<td>ATG18 A IV</td>
<td>A1g06340</td>
</tr>
<tr>
<td>Metabolism</td>
<td>ATG18 F II</td>
<td>A1g06340</td>
</tr>
<tr>
<td>Lipid degradation</td>
<td>ACYL-COENZYM A OXIDASE 1 (ACX1) III</td>
<td></td>
</tr>
<tr>
<td>At4g16760</td>
<td>ACX2 III</td>
<td>A1g51840</td>
</tr>
<tr>
<td>At2g39400</td>
<td>ACX4 IV</td>
<td>A1g6620</td>
</tr>
<tr>
<td>At4g16760</td>
<td>1,4-DIHYDROXY-2-NAPHTHOYL-COA III</td>
<td>A1g48320</td>
</tr>
<tr>
<td>At1g54620</td>
<td>THIOESTERASE 1 II</td>
<td>A1g18640</td>
</tr>
<tr>
<td>Aminoacid degradation</td>
<td>α/β-Hydrolases superfamily protein I-IV</td>
<td></td>
</tr>
<tr>
<td>At1g55510</td>
<td>2-OXOISOVALERATE DEHYDROGENASE β1 IV</td>
<td>A1g33150</td>
</tr>
<tr>
<td>At4g33150</td>
<td>α-AMINOACIDIC SEMIALDEHYDE SYNTHASE II-III</td>
<td>A1g53970</td>
</tr>
<tr>
<td>At5g54080</td>
<td>HOMOGENTISATE 1,2-DIOXYGENASE II-III</td>
<td>A1g45300</td>
</tr>
<tr>
<td>At2g45300</td>
<td>ISOVALeryl-COA DEHYDROGENASE II-III-IV</td>
<td>A1g64860</td>
</tr>
<tr>
<td>At1g63400</td>
<td>METHIONINE GAMMA-LYASE II</td>
<td>A1g34030</td>
</tr>
<tr>
<td>At1g06340</td>
<td>METHYLcrotonoyL-COA CARBOXYLASE β IV</td>
<td>A1g06340</td>
</tr>
<tr>
<td>Autophagy</td>
<td>ORE9/MAK2 II-IV</td>
<td>A1g06340</td>
</tr>
<tr>
<td>Sensing</td>
<td>ATG18 A IV</td>
<td>A1g06340</td>
</tr>
<tr>
<td>Metabolism</td>
<td>ATG18 F II</td>
<td>A1g06340</td>
</tr>
</tbody>
</table>
TABLE 1 | Continued

<table>
<thead>
<tr>
<th>AGI Name</th>
<th>GRN</th>
<th>AGI Name</th>
<th>GRN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein degradation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At5g57360</td>
<td>ADAGIO 1</td>
<td>I-II</td>
<td>A1g80440</td>
</tr>
<tr>
<td>At2g19815</td>
<td>ADAGIO 2</td>
<td>III</td>
<td>A1g15670</td>
</tr>
<tr>
<td>At1g05840</td>
<td>ASPARTYL PROTEASE</td>
<td>I-III-IV</td>
<td>A2g44120</td>
</tr>
<tr>
<td>At1g21780</td>
<td>BTB/POZ domain-containing protein</td>
<td>II-IV</td>
<td>A3g59940</td>
</tr>
<tr>
<td>At5g18650</td>
<td>CHY-/CTCHY-/RING-type Zinc finger protein</td>
<td>IV</td>
<td>A4g24990</td>
</tr>
<tr>
<td>At5g22920</td>
<td>CHY-/CTCHY-/RING-type Zinc finger protein</td>
<td>IV</td>
<td>A1g23440</td>
</tr>
<tr>
<td>At3g13550</td>
<td>COP10</td>
<td>IV</td>
<td>A4g02440</td>
</tr>
<tr>
<td>At2g40880</td>
<td>Cysteine proteinase inhibitor 3</td>
<td>I</td>
<td>A1g60190</td>
</tr>
<tr>
<td>At5g05110</td>
<td>Cysteine proteinase inhibitor 7</td>
<td>II</td>
<td>A2g22690</td>
</tr>
<tr>
<td>At4g38090</td>
<td>Cysteine proteinase RD19a</td>
<td>IV</td>
<td>A1g13195</td>
</tr>
<tr>
<td>At5g04250</td>
<td>Cysteine proteinases superfamily protein</td>
<td>II</td>
<td>A1g14200</td>
</tr>
<tr>
<td>At4g23450</td>
<td>E3 ubiquitin ligase</td>
<td>I-III-IV</td>
<td>A1g20440</td>
</tr>
<tr>
<td>At5g42200</td>
<td>E3 ubiquitin ligase ATL23</td>
<td>IV</td>
<td>A1g26800</td>
</tr>
<tr>
<td>At1g74410</td>
<td>E3 ubiquitin ligase ATL24</td>
<td>II</td>
<td>A1g49860</td>
</tr>
<tr>
<td>At3g05200</td>
<td>E3 ubiquitin ligase ATL6</td>
<td>III</td>
<td>A1g55530</td>
</tr>
<tr>
<td>At1g49210</td>
<td>E3 ubiquitin ligase ATL76</td>
<td>I</td>
<td>A1g75400</td>
</tr>
<tr>
<td>At1g76410</td>
<td>E3 ubiquitin ligase ATL8</td>
<td>I-III-IV</td>
<td>A2g15880</td>
</tr>
<tr>
<td>At2g09770</td>
<td>E3 ubiquitin ligase LOG2</td>
<td>II</td>
<td>A2g37150</td>
</tr>
<tr>
<td>At4g11360</td>
<td>E3 ubiquitin ligase RH41B</td>
<td>I-III-IV</td>
<td>A3g02340</td>
</tr>
<tr>
<td>At5g22000</td>
<td>E3 ubiquitin-protein ligase RHF2A</td>
<td>II-III</td>
<td>A3g05250</td>
</tr>
<tr>
<td>At4g03510</td>
<td>E3 ubiquitin-protein ligase RMA1</td>
<td>I-III-IV</td>
<td>A3g13430</td>
</tr>
<tr>
<td>At4g28270</td>
<td>E3 ubiquitin-protein ligase RMA2</td>
<td>IV</td>
<td>A3g15070</td>
</tr>
<tr>
<td>At2g27470</td>
<td>E3 ubiquitin-protein ligase RMA3</td>
<td>IV</td>
<td>A3g17160</td>
</tr>
<tr>
<td>At5g02880</td>
<td>E3 ubiquitin-protein ligase UBC4</td>
<td>II</td>
<td>A3g61180</td>
</tr>
<tr>
<td>At2g04240</td>
<td>E3 ubiquitin-protein ligase XEROICO</td>
<td>II-IV</td>
<td>A4g39340</td>
</tr>
<tr>
<td>At5g23530</td>
<td>EIN3-binding F-box protein 2</td>
<td>I-III</td>
<td>A5g01520</td>
</tr>
<tr>
<td>At1g23390</td>
<td>F-box protein</td>
<td>IV</td>
<td>A5g03180</td>
</tr>
<tr>
<td>At4g26930</td>
<td>F-box protein</td>
<td>III</td>
<td>A5g10660</td>
</tr>
<tr>
<td>At1g30200</td>
<td>F-box protein</td>
<td>IV</td>
<td>A5g19400</td>
</tr>
<tr>
<td>At1g51550</td>
<td>F-box protein</td>
<td>I-IV</td>
<td>A5g24870</td>
</tr>
<tr>
<td>At1g55000</td>
<td>F-box protein</td>
<td>IV</td>
<td>A5g55970</td>
</tr>
<tr>
<td>At1g70590</td>
<td>F-box protein</td>
<td>I-III-IV</td>
<td>A4g00325</td>
</tr>
<tr>
<td>At2g12350</td>
<td>F-box protein</td>
<td>II</td>
<td>A5g01880</td>
</tr>
<tr>
<td>At5g27920</td>
<td>F-box protein</td>
<td>I-III-IV</td>
<td>A2g17450</td>
</tr>
<tr>
<td>At5g43190</td>
<td>F-box protein</td>
<td>III</td>
<td>A2g18670</td>
</tr>
<tr>
<td>At1g21170</td>
<td>F-box protein</td>
<td>I-II</td>
<td>A5g61610</td>
</tr>
<tr>
<td>At2g42620</td>
<td>F-box protein MAX2</td>
<td>I-II</td>
<td>A3g03000</td>
</tr>
<tr>
<td>At4g03030</td>
<td>F-box protein OR23</td>
<td>II-III-IV</td>
<td>A2g22980</td>
</tr>
<tr>
<td>At1g80110</td>
<td>F-box protein PP2-B11</td>
<td>II</td>
<td>A1g16560</td>
</tr>
<tr>
<td>At5g57900</td>
<td>F-box protein SKP1</td>
<td>I-II-IV</td>
<td>A1g45976</td>
</tr>
<tr>
<td>At1g21410</td>
<td>F-box protein SKP2A</td>
<td>II</td>
<td>A3g06380</td>
</tr>
<tr>
<td>At1g77000</td>
<td>F-box protein SKP2B</td>
<td>II</td>
<td>A1g63800</td>
</tr>
<tr>
<td>At4g10925</td>
<td>F-box protein SKP8</td>
<td>IV</td>
<td>A5g14700</td>
</tr>
<tr>
<td>At3g26000</td>
<td>F-box protein SKP14</td>
<td>II</td>
<td>A4g36410</td>
</tr>
<tr>
<td>At4g21510</td>
<td>F-box protein SKP27</td>
<td>III</td>
<td>A1g64230</td>
</tr>
<tr>
<td>At5g45360</td>
<td>F-box protein SKP31</td>
<td>IV</td>
<td>A3g17000</td>
</tr>
</tbody>
</table>

The GRN to which each gene belongs to is indicated.

Genes from these sets appeared in the GRNs at a much higher frequency than expected in a random list (pval 4.5E-11 to 7.5E-215; Table 2 and Supplementary Figure S4), indicating that the bud dormancy GRNs were very highly enriched in genes typical of a C starvation response.

GSEA Analyses Confirm a C Starvation Response in Dormant Buds

We assessed this potential C starvation response by performing a Gene Set Enrichment Analysis (GSEA) using all transcribed
genes from each experiment, rather than focusing on the bud dormancy GRNs. GSEA is a statistical approach that allows identification of overrepresented gene sets among differentially up- or downregulated genes of a transcriptomic experiment (Subramanian et al., 2005). For each “active-vs.-dormant bud” experiment, we generated a ranked gene list using relative gene expression levels and False Discovery Rate (FDR) values. We then tested whether gene sets related to a potential C starvation syndrome (sugar-, darkness- and AKIN10-responsive genes, ABA, ethylene and CK markers, ribosomal genes, cell cycle and cell division genes; Supplementary Dataset S4) were found toward the top (upregulated) or the bottom (downregulated) of the ranked gene lists. Analyses confirmed significant overrepresentation of C-signaling, sugar-repressed, AKIN10-induced and ABA and ethylene marker genes among those upregulated in the three experiments (Figures 2A, 3). In contrast, CK markers, ribosomal genes and S-phase genes were overrepresented among the downregulated genes (Figures 2B, 3). Other cell division markers such as M-phase genes, histones and kinesins were overrepresented only in the 8 and 24 h experiments, which suggests they are downregulated at later stages of the process (Figures 2B, 3).

All these results suggest that a C starvation syndrome is induced early in the growth-to-dormancy transition in para- and ecodormant axillary buds in Arabidopsis.

Regulation of the Bud Dormancy GRNs

To find potential master regulators of the C starvation-related bud dormancy GRNs, we searched for overrepresented motifs in the gene promoters (1 kilobase upstream of the transcription start site) of each GRN using Oligo-analysis and Pattern Assembly (Rsat; Medina-Rivera et al., 2015). In GRN I and GRN IV, tccTATCCAC was the most-overrepresented motif (Supplementary Figure S5 and Dataset S5); it contains the sucrose-repressible element TATCCA, bound in rice by the MYB factors OsMYBS1, 2, and 3, which mediate sugar-regulated gene expression (Lu et al., 1998, 2002). We looked for TFs within GRN I and IV that could bind this motif, based on DNA affinity purification sequencing (DAP-Seq) data (O’Malley et al., 2016) or chromatin immunoprecipitation sequencing (ChIP-Seq) data (Song et al., 2016), and that could act as master regulators of the GRNs. The Arabidopsis OsMYBS2 ortholog MYBS2 (At5g08520), which has a role in sugar and ABA signaling (Chen et al., 2016), pertains to GRN I and might bind this motif (Supplementary Figure S6A and Dataset S5). In GRN IV, MYBS2 and three other MYB-related proteins, MYBH/KUA1 (At5g47390), At1g19000 and At1g74840 could bind this motif (Supplementary Figure S6A and Dataset S5). MYBH/KUA1 has a critical role in dark-induced leaf senescence (Huang et al., 2015). These sugar-regulated genes could be instrumental in coordinating gene expression in GRN I and GRN IV.

The most overrepresented motifs in GRN II and III, gaCAGTGtc, tgaCAGTGt and gACAGT, overlap with the G-box (CAGTG), which is bound by bZIP, bHLH and NAC proteins (Supplementary Figure S5 and Dataset S5). These motifs are overrepresented in the promoters of C starvation response genes (Cookson et al., 2016). In GRN II, the master regulators of ABA signaling GBF2, GBF3, AFB3 and AFB4, and the senescence-inducing NAC factors NAP, NAC6/ORE1 and ATAF1 as well as NAC047 and NAC3 could bind these motifs (Supplementary Figure S6A and Dataset S5; Guo and Gan, 2006; Balazadeh et al., 2010; Garapati et al., 2015; O’Malley et al., 2016; Song et al., 2016). In GRN III, NAP, NAC6/ORE1, NAC047, NAC3, NAC102, RD26 (Supplementary Figure S6A and Dataset S5).
FIGURE 2 | Gene Set Enrichment Analysis (GSEA) analyses of dormant vs. active bud experiments. (A,B) Enrichment Scores (ES; green line) of selected gene sets that illustrate significant overrepresentation among up- (A) or down-regulated genes (B). Barcode-like vertical black lines represent logRatios of genes of each gene set in the ranked ordered data sets. Left (positive logRatios), genes induced in dormant buds; right (negative logRatios), genes repressed in dormant buds.
and Dataset S5) and possibly NAC19, for which there is no available binding information, might regulate these motifs and promote gene expression.

We confirmed significant enrichment of the GRNs in the target genes of these TFs by using DAP-Seq and ChIP-Seq data (O’Malley et al., 2016; Song et al., 2016); their numbers in the GRNs were significantly higher than expected in a random gene list (pval < 0.01). For instance, the number of gene targets for NAC102, RD26 and ABF4 was 6, 3.2, and 3.4 times higher, respectively; for the remainder, this value was between 1.7 and 2.7 times higher than predicted (Supplementary Figure S6B).

All these results indicate that four interrelated GRNs associated to a C starvation response are induced in para- and ecodormant Arabidopsis buds. MYB-related, bZIP and NAC TFs could have a key role in the regulation of these GRNs. A large proportion of the genes in the GRNs are rapidly repressed by sugar and upregulated by AKIN10. They are tightly coregulated with or directly involved in sugar signaling and metabolism, autophagy, senescence, catabolism of lipids and proteins, and ABA and ethylene signaling. This response is also associated with downregulation of CK signaling, protein synthesis and cell division, all conditions that lead to the cell and tissue growth arrest typical of dormant buds.

**Conservation of Bud Dormancy GRNs in Arabidopsis, Poplar and Grapevine**

We investigated whether the GRNs related to a C starvation syndrome identified in Arabidopsis were also induced during the growth-to-dormancy transition in buds of the woody plant species, poplar and grapevine. We studied two public transcriptomic experiments in which apical buds of poplar (Ruttink et al., 2007) or axillary buds of grapevine (Díaz-Riquelme et al., 2012) underwent dormancy. To induce dormancy, shoot apices of poplar plants grown in long days (LD, 16 h light-8 h darkness) were exposed to 1–6 weeks of short days (w SD, 8 h light-16 h darkness) (Ruttink et al., 2007). During treatment, the shoot apices developed into buds (1–3 w SD), grew adapted to dehydration and cold (3–6 w SD), and became dormant (5–6 w SD). Samples were collected weekly. Díaz-Riquelme et al. (2012) collected monthly samples of axillary buds.
of grapevine plants grown in natural conditions in the northern hemisphere. Grapevine axillary buds are formed between April and May; in July and August they grow, undergo flowering and develop inflorescence meristems, enter endodormancy at the end of September, and exit dormancy by the end of November. They remain ecodormant throughout December, until environmental conditions become benign around March, when they sprout (Martínez de Toda Fernández, 1991; Díaz-Riquelme et al., 2012).

We analyzed the expression patterns of the poplar and grapevine orthologs of the GRNI-IV genes. Of 838 Arabidopsis genes in these GRNs, we identified 390 poplar and 421 grapevine orthologs (Supplementary Dataset S6). In both species, we studied gene expression relative to levels in the "active bud" sample (LD in poplar, April in grapevine). In general, a large proportion of the bud dormancy genes orthologs were significantly induced at most time points in poplar and grapevine buds (Supplementary Figure S7), which supports a conservation, during the growth-to-dormancy transition in these woody species, of the responses found in Arabidopsis. In poplar, the global induction appeared to increase over the weeks in SD, especially for GRNII and III genes. In contrast, grapevine gene induction was detectable throughout the year (Supplementary Figure S7).

A group of genes showed high expression levels from the earliest stages (1–3 w SD in poplar and July in grapevine), weeks/months before endodormancy onset, and throughout the experiment (Figure 4 and Supplementary Dataset S6). In poplar, these were DRM1, HISI-3, GID1C, COR413IM1, ALANINE:GLYOXYLATE AMINOTRANSFERASE 3 (AGT3), SUCROSE SYNTHASE3 (SUS3), TEMPRANILLO1 (TEMI) and SEIPIN (Figure 4A). In grapevine, they were DRM1, HISI-3, DIN6, PIF4, CBP1/MEE14, HSPRO2, EXL4, BACAT2 and ATY13/MB31, ERF2, ABA receptor PYL9/ABI1, the protein phosphatases 2C HA11/SAG113 and AIP1/HA12, involved in ABA signaling and sucrose sensitivity (Lim et al., 2012), DOF5.4, HSF1C, PLANT U-BOX 19, GOLS1, senescence factor NAP, RD26, ALUMINUM SENSITIVE 3 (ALS3), and oxidative stress-related At3g10020 (Figure 4B).

Other bud dormancy genes were induced exclusively between 1 and 3 w SD in poplar, and in July in grapevine. Early poplar genes were GAS6, the sugar-responsive gene CBP1/MEE14 (Bi et al., 2005), AKBETA1 and LSD ONE LIKE 1 (LOL1), a positive regulator of cell death (Epple et al., 2003) (Figure 4A and Supplementary Dataset S6). July-induced grapevine genes were sugar transporters STP14 and STP1, KISS ME DEADLY 4/SKP20, which encodes an F-box protein that negatively regulates the CK response, autophagy factor ATG8I, chaperone DNAJ11, PLASMA-MEMBRANE ASSOCIATED CATION-BINDING PROTEIN 1 (PCaP1), ASD1, involved in cell wall remodeling (Chávez Montes et al., 2008), and ABA-responsive MYB3 (Figure 4B and Supplementary Dataset S6).

In summary, a large proportion of the genes orthologous to Arabidopsis bud dormancy genes are also induced, either early and transiently or early and constantly during the growth-to-dormancy transition in poplar and grapevine, which supports their functional conservation in these woody species.

The C Starvation Response Is Conserved in Poplar and Grapevine Buds
To obtain a general view of the transcriptomic responses in these experiments, we performed GSEA similar to that for Arabidopsis, using all genes with proposed Arabidopsis orthologs (8023 genes in poplar and 8390 in grapevine) (Ruttink et al., 2007; Díaz-Riquelme et al., 2012).

The sugar- and AKBET10-responsive gene sets were overrepresented among upregulated genes from 1 w SD in poplar, and July in grapevine, and were also induced throughout the treatment/year (Figure 3). This finding confirms that the C starvation response begins early, long before endodormancy onset, and underlies the entire process. The ribosomal gene set was constitutively overrepresented among downregulated genes in all three species, which confirmed that inhibition of protein synthesis is an early and sustained response in buds entering dormancy. General downregulation of cell cycle and cell division genes was also observed in grapevine, whereas in poplar, cell division gene sets were repressed more gradually and reached maximum repression at 5 w SD. In contrast to Arabidopsis, histones were not significantly downregulated in the woody species. Nevertheless, C starvation response gene sets (upregulated) and cell growth-related gene sets (downregulated) clustered together in the three species.

Unlike the gene sets discussed above, hormone responses did not appear to be strongly conserved among species, suggesting more relaxed evolution of these pathways. Whereas ABA-related genes were induced constitutively in grapevine, ABA and ethylene responses were induced from 2 w SD onward in poplar, in accordance with previous observations (Ruttink et al., 2007). CK signaling is repressed in Arabidopsis, but not notably in poplar or grapevine. An early, extended response to the senescence-associated hormone jasmonate (MJ) was repressed in two Arabidopsis experiments, and was induced in most poplar and grapevine samples (Figure 3).

This results indicate that an early and sustained sugar-starvation response associated with downregulation of ribosomal and cell cycle proteins is conserved in buds of Arabidopsis, poplar and grapevine, and might constitute a core response of buds entering dormancy in the angiosperms.

Cell Type-Specific Gene Expression of Bud Dormancy Genes in the Shoot Apex
To further analyze the function of the genes induced during the C starvation response in buds, we selected those most highly expressed in Arabidopsis, poplar and grapevine, to determine the cell types in which they are expressed. We used a high-resolution gene expression database of the Arabidopsis shoot apex, which contains the same tissues as axillary buds: meristem and leaf primordia. This database comprises gene expression profiles of different cell populations obtained by fluorescence-activated cell sorting (L1, L2 and L3 layers, central (CZ) and peripheral zone (PZ), leaf primordia, xylem and phloem) (Yadav et al., 2014). As we cannot rule out that the expression levels of these genes change in dormant axillary buds, we used this database for qualitative rather than quantitative analysis, to
FIGURE 4 | Expression profiles of bud dormancy genes in poplar and grapevine. Heatmap of gene expression for poplar (A, Po_GRN) and grapevine (B, Vi_GRN) orthologs of the Arabidopsis GRN genes. Log2 ratios of normalized gene intensities in each time point vs. normalized gene intensities on the active bud sample are indicated. For poplar and grapevine the “active bud” sample are LD and April, respectively. In red and green, genes up- and downregulated in dormant buds respectively. Genes mentioned in the text are indicated. Schematic representations based on information from Ruttink et al. (2007) (A) and Díaz-Riquelme et al. (2012) (B), below indicate the proposed developmental stage of buds in each time point.
identify the cell types in which these genes were expressed most abundantly.

Many of the most highly induced genes in buds were expressed preferentially in the vasculature (Figures 5A–D); sugar-related genes Suc2, STP3, TPS11, GOLS1, and TF HB-40, HB-12, PAT1 and CDF2 were expressed exclusively in phloem (Figure 5A). Many ABA-related genes (PYL9, HA11, NAP, NAC055, NAC002, NMYA1, HAT22, RVE6, LEA4-5, DOF5.4, HIS1-3 and TSP0), SCR, the F box-encoding genes MAX2 and KMD2, CBP1/MEE14, AGT3, and CBSX5 were expressed almost exclusively in xylem (Figure 5B). Other ABA-related genes (AFB4, HAB1, HA12, GFB3, RD26, MYB31, RAP2.3) as well as KING1, TPS10, TRE, XERICO, EXL2, KMD3, VOZ1, SIS and PUB19 were expressed in both xylem and phloem (Figures 5C,D). In the CLV3/WUS expression domain, we found TEM1, protein kinase CIPK14 that interacts with SnRK1 (Yan et al., 2014), SUS3, UDP-GLUCOSYL TRANSFERASE 87A2 (UGT87A2) and 6 PHOSPHOGLUCONOLACTONASE 1 (PGL1) (Figure 5E). The genes PIF1, PIF4, EXL4, ADAGIO (ADO1), ETR2 and COR413 IM1 accumulated preferentially in leaf primordia (Figure 5F). Other strongly expressed genes such as DRM1, PACP1, KMD1, and AFP3 were found in both vascular tissue and leaf primordia, and GID1C in xylem and the peripheral zone of the meristem. The autophagy genes ATG8E, ATG8C, ATG18A, ATG18E, ATG18H, the senescence gene SAP3, as well as ER2, GID1B, BYPASS and HISTONE DEACETYLASE 8 were widely expressed throughout the meristem.

In summary, whereas ABA signaling occurs mostly in the xylem, sugar signaling in the phloem, and ethylene in the meristem proper (Yadav et al., 2014), autophagy and arrest of cell growth take place throughout the meristem. This suggests that cell-to-cell communication and movement of signaling molecules, hormones and proteins must take place across different cell types in buds entering dormancy.

**Bud Dormancy Early Markers**

It is of great interest to identify robust, universal markers that allow diagnosis of axillary bud status. These markers should be induced early, and their expression be sustained in para-dormant buds throughout the period of bud dormancy. Based on our analysis, several genes met these criteria. We tested further four of them: DRM1, HIS1-3, GID1C, and NAP. DRM1 and HIS1-3 were expressed at high levels in the Arabidopsis, poplar and grapevine experiments. DRM1 is a well-known dormancy marker in herbaceous and woody plants species (Stafstrom et al., 1998; Park and Han, 2003; Tatematsu et al., 2005; Aguilar-Martínez et al., 2007; Kebrom et al., 2010; Wood et al., 2013). It is repressed by sugar, which supports its strong association to low sugar levels and dormancy. ABA-responsive HIS1-3 is upregulated before ABA signaling in poplar and grapevine buds. GID1C, which encodes a gibberellin receptor, is expressed at high levels in the three Arabidopsis experiments and in poplar (Figure 4A). The senescence-promoting gene NAP is expressed at very high levels in Arabidopsis and throughout the year in grapevine (Figure 4B). Both GID1C and NAP belong to the four bud dormancy GRNs (Supplementary Dataset S1).

We tested whether expression of these genes also correlated with bud dormancy in axillary buds of potato (Solanum tuberosum, Solanaceae, Asteridae), a species only distantly related to Arabidopsis, poplar (both Rosidae) and grapevine (basal angiosperms). We identified the potato ortholog genes for the four candidates. In the case of NAP, we found two potato paralogs (NAPa and NAPb). We studied their mRNA levels in buds of plants treated for 10 h with white light (W) or with W supplemented with far red light (W+FR), a treatment that promotes axillary bud dormancy in potato. We also compared mRNA levels in buds of intact and decapitated plants. Whereas DRM1 and HIS1-3 were confirmed as reliable markers of bud dormancy in potato, GID1C, NAPa and NAPb did not respond as anticipated in decapitated plants; GID1C was not upregulated in low R:FR and NAPa/NAPb were highly induced after decapitation (Figure 6).

**DISCUSSION**

**Carbon Availability, a Key Signal for Growth**

In all eukaryotes, cell proliferation and growth demand high carbohydrate levels for energy generation and macromolecule synthesis. Correspondingly, low C availability promotes a reduction in growth rate in order to retain sufficient C to support essential maintenance functions (Rolland et al., 2006). In Arabidopsis, sugar availability has a great influence on growth and development, both in seedlings and adult plants. C-limiting conditions (e.g., sucrose depletion, night extensions, short-day photoperiods, starchless mutants) trigger a suite of transcriptional responses that lead to growth cessation, and which include repression of genes involved in abscisic acid, protein synthesis, cell division, cell cycle, and DNA synthesis and repair (Moore et al., 2003; Smith and Stitt, 2007; Wiese et al., 2007).

Regarding shoot branching, it has been shown that sugar availability to buds plays a major role in its control in pea and rose (Mason et al., 2014; Barbier F. et al., 2015). In agreement, we have found that induction of GRNs typical of tissues undergoing C starvation precedes and underlies the bud growth-to-dormancy transition in Arabidopsis, poplar and grapevine. This is concomitant with transcriptional repression of ribosomal and cell-cycle genes, responses typical of buds entering dormancy as well as of tissues undergoing C limitations (e.g., Thimm et al., 2004; Smith and Stitt, 2007; González-Grandio et al., 2013). Indeed it is possible that bud dormancy is a manifestation and a consequence of the observed C starvation syndrome.

**Dormancy-Promoting Stimuli and the C Starvation Response**

How is this C starvation response induced? In apical dominance it has been proposed that the growing shoot apex acting as a sugar sink might limit sugar availability to axillary buds so this can be the direct trigger of the response (Mason et al., 2014). Tre6P, a metabolite that acts as a proxy for C status, may also promote signaling in addition to, or instead of direct sugar sensing (Paul
FIGURE 5 | Cell type-specific expression of bud dormancy genes in the Arabidopsis shoot apical meristem (SAM). Heatmap of bud dormancy gene expression in SAM cell types, normalized for each gene relative to the cell type with the highest expression levels (1.0, red). No expression (0.0), green. Different cell types are indicated on top. Horizontal color lines on the right indicate gene sets to which each gene belongs. (A) Genes expressed almost exclusively in phloem. (B) Genes expressed almost exclusively in xylem. (C,D) Genes expressed preferentially in phloem and xylem. (E) Genes expressed preferentially in layers 1, 2, 3 (L1, L2, L3), meristem (mer) and central zone (CZ). (F) Genes expressed preferentially in peripheral zone (PZ).
et al., 2008; Lunn et al., 2014). This would be in agreement with the observation that plants that express microbial trehalose-phosphate synthase (TPS) genes show increased shoot branching (Goddijn et al., 1997) and maize plants with a mutation in the trehalose-phosphate phosphatase gene RAMOSA3 have altered inflorescence branching (Satoh-Nagasawa et al., 2006).

Nevertheless, it is likely that the syndrome is not only induced by an actual sugar shortfall, but also by cues that inform of current or future suboptimal conditions which may affect energy availability and/or interfere with respiration and C assimilation (Baena-González et al., 2007; Baena-González and Sheen, 2008). Seasonal environmental changes that perturb these processes (e.g., daylength shortening, light levels, temperature, water availability) may trigger acclimatory signaling pathways that anticipate C limitations (Smith and Stitt, 2007). Those and other stimuli could feed into regulatory networks that economize resources locally, to result in a moderation of growth rate in axillary meristems and buds.

In two of the Arabidopsis experiments examined, the dormancy-inducing stimuli was an exposure to low R:FR light ratio. Low R:FR light is interpreted by plants as a situation with limited light available for photosynthesis. It severely reduces the expression of photosynthesis-related genes (Cagnola et al., 2012) and induces cell-wall remodeling in stem and petioles, which may divert carbohydrates away from axillary buds (Sasidharan et al., 2010). Furthermore, low R:FR promotes ethylene and ABA signaling and CK degradation (Carabelli et al., 2007; Cagnola et al., 2012), hormonal responses tightly linked to the C starvation response (see below).

In the poplar and grapevine studies, the sugar-repressed networks are induced in buds soon after beginning of daylength shortening: in poplar at 1 w SD; in grapevine, in July, when daylength shortening has just begun (June 21), even though buds are still growing. Short-day photoperiod leads to localized flower and seed abortion associated with low levels of C in Arabidopsis (Lauxmann et al., 2016). Likewise, in poplar a measurable shortage of sugar availability is detectable after 1 w SD (Ruttink et al., 2007). Under the natural conditions in which grapevine plants are grown, daylength shortening and C limitations are progressive, but relatively small changes in C balance may trigger the response. Indeed in Arabidopsis even minor alterations in C status, well before C starvation, lead to notable changes in C-related signaling and response (Usadel et al., 2008). In addition, genetic pathways that sense photoperiod might help anticipate and adapt to impending C-limiting conditions in short days. These pathways, controlled by phytochromes, circadian clock, and genes controlling flowering time (Horvath, 2009), may regulate and establish crosstalk with the C starvation response. Indeed, sugars affect the expression of clock genes (Haydon et al., 2013) and conversely, the clock regulates carbohydrate metabolism (Smith and Stitt, 2007). Phytochromes, which monitor changes in R:FR and in day-length, also regulate SD-induced endodormancy in woody species (Johnson et al., 1994; Reed et al., 1994; Olsen et al., 1997; Neff and Chory, 1998; Monte et al., 2003; Ruonala et al., 2008; Franklin and Quail, 2010). Changes in low R:FR light ratio or photoperiod might therefore trigger partially overlapping responses, including potential anticipation of a C-limiting situation.

Although it has not been analyzed in this work, coordination between C and N metabolic pathways probably affect this process as well, as sugar responses depend significantly on the N status of the plant.

**Figure 6** | Expression levels of candidate bud dormancy markers in potato. Transcript levels of DRM1, HST-3, GID1C and NAP potato ortholog genes in aerial axillary buds were analyzed by quantitative PCR. (A) intact and decapitated plants, 10 h after treatment. (B) Plants exposed for 10 h to white light or to a low red:far-red light ratio. Values are mean ± SEM (N = 5, N = 4). Each biological replicate is a pool of 16 axillary buds. **pval < 0.01; ***pval < 0.001; two-tailed Student’s t-test.

**The C Starvation Syndrome in Axillary Buds: Sugar Signaling**

The C starvation syndrome comprises a cascade of transcriptomic events that culminate in changes in growth and C balance (Figure 7). These events include induction of genes involved in transcriptional regulation, sugar sensing, transport and signaling, catabolism (i.e., amino acid and lipid degradation), protein ubiquitination and degradation, hormone...
signaling, autophagy and senescence. Genes required for growth, such as ribosomal, cell cycle and anabolism-related genes become downregulated (Thimm et al., 2004). Buds entering dormancy in Arabidopsis, poplar and grapevine show induction of genes of the former categories and repression of genes of the latter categories.

EXORDIUM-like (EXL2) and EXL4 are bud dormancy genes potentially involved in sugar sensing. They are induced in extended night treatments in Arabidopsis seedlings, in accordance with a role under C-limiting conditions (Schröder et al., 2012). Their close paralogs, EXORDIUM (EXO) and EXL1, are proposed to integrate apoplastic C status with intracellular responses (EXO) (Lisso et al., 2013) and to control primary and long-term adaptation to C starvation (EXL1) (Schröder et al., 2011, 2012).

Several sugar transporters are also induced in dormant buds. These are STP1, one of the most rapidly and prominently downregulated genes in response to sugars (Price et al., 2004; Cordoba et al., 2015), STP14, which is strongly repressed by sugars (Büttner, 2010) and the sucrose efflux transporters SWEET11 and SWEET12, which act with the sucrose/proton symporter SUT1/SUC2 for phloem loading and long-distance transport (Chen et al., 2012).

Sugar signaling involves Tre6P (Paul et al., 2008; Lunn et al., 2014) and four class-II TPS, TPS8, TPS9, TPS10, and TPS11, are induced in buds entering dormancy. They belong to a core C-signaling response, are usually strongly upregulated in C starvation, and are AKIN10-responsive (Contento et al., 2004; Price et al., 2004; Thimm et al., 2004; Baena-González et al., 2007). Although their proteins may be catalytically inactive, they might modulate other TPSs or act as Tre6-P sensors (Lunn, 2007). In addition, the SnRK1 protein-kinase, a central regulator of growth in response to C availability (Baena-González et al., 2007), is likely to have a key role in the induction of bud dormancy (see below).

**Ethylene, ABA, CK, Senescence and Autophagy during Bud Dormancy**

Ethylene and ABA signaling are induced during the bud dormancy transition in Arabidopsis, which agrees with studies that associated these hormones with bud dormancy in many other herbaceous and woody species (Suttle, 1998; Ruonala et al., 2006; Rohde and Bhalerao, 2007; Ruttkink et al., 2007; Horvath et al., 2008; Díaz-Riquelme et al., 2012; González-Grandío et al., 2013, 2017; González-Grandío and Cubas, 2014; Yao and Finlayson, 2015). We propose that ethylene and ABA responses are closely connected to the C starvation syndrome. Indeed, many mutants with altered responses to sugars have impaired ethylene or ABA signaling (Zhou et al., 1998; Arenas-Huertero et al., 2000; Finkelstein and Lynch, 2000; Huijser et al., 2008; Diaz-Riquelme et al., 2012; González-Grandio et al., 2013, 2017; Reddy et al., 2013; González-Grandio and Cubas, 2014; Yao and Finlayson, 2015). We propose that ethylene and ABA responses are closely connected to the C starvation syndrome. Indeed, many mutants with altered responses to sugars have impaired ethylene or ABA signaling (Zhou et al., 1998; Arenas-Huertero et al., 2000; Finkelstein and Lynch, 2000; Huijser et al., 2008; Diaz-Riquelme et al., 2012; González-Grandio et al., 2013, 2017; Reddy et al., 2013; González-Grandio and Cubas, 2014; Yao and Finlayson, 2015). We propose that ethylene and ABA responses are closely connected to the C starvation syndrome. Indeed, many mutants with altered responses to sugars have impaired ethylene or ABA signaling (Zhou et al., 1998; Arenas-Huertero et al., 2000; Finkelstein and Lynch, 2000; Huijser et al., 2008; Diaz-Riquelme et al., 2012; González-Grandio et al., 2013, 2017; Reddy et al., 2013; González-Grandio and Cubas, 2014; Yao and Finlayson, 2015). We propose that ethylene and ABA responses are closely connected to the C starvation syndrome. Indeed, many mutants with altered responses to sugars have impaired ethylene or ABA signaling (Zhou et al., 1998; Arenas-Huertero et al., 2000; Finkelstein and Lynch, 2000; Huijser et al., 2008; Diaz-Riquelme et al., 2012; González-Grandio et al., 2013, 2017; Reddy et al., 2013; González-Grandio and Cubas, 2014; Yao and Finlayson, 2015). We propose that ethylene and ABA responses are closely connected to the C starvation syndrome. Indeed, many mutants with altered responses to sugars have impaired ethylene or ABA signaling (Zhou et al., 1998; Arenas-Huertero et al., 2000; Finkelstein and Lynch, 2000; Huijser et al., 2008; Diaz-Riquelme et al., 2012; González-Grandio et al., 2013, 2017; Reddy et al., 2013; González-Grandio and Cubas, 2014; Yao and Finlayson, 2015). We propose that ethylene and ABA responses are closely connected to the C starvation syndrome. Indeed, many mutants with altered responses to sugars have impaired ethylene or ABA signaling (Zhou et al., 1998; Arenas-Huertero et al., 2000; Finkelstein and Lynch, 2000; Huijser et al., 2008; Diaz-Riquelme et al., 2012; González-Grandio et al., 2013, 2017; Reddy et al., 2013; González-Grandio and Cubas, 2014; Yao and Finlayson, 2015). We propose that ethylene and ABA responses are closely connected to the C starvation syndrome. Indeed, many mutants with altered responses to sugars have impaired ethylene or ABA signaling (Zhou et al., 1998; Arenas-Huertero et al., 2000; Finkelstein and Lynch, 2000; Huijser et al., 2008; Diaz-Riquelme et al., 2012; González-Grandio et al., 2013, 2017; Reddy et al., 2013; González-Grandio and Cubas, 2014; Yao and Finlayson, 2015). We propose that ethylene and ABA responses are closely connected to the C starvation syndrome. Indeed, many mutants with altered responses to sugars have impaired ethylene or ABA signaling (Zhou et al., 1998; Arenas-Huertero et al., 2000; Finkelstein and Lynch, 2000; Huijser et al., 2008; Diaz-Riquelme et al., 2012; González-Grandio et al., 2013, 2017; Reddy et al., 2013; González-Grandio and Cubas, 2014; Yao and Finlayson, 2015). We propose that ethylene and ABA responses are closely connected to the C starvation syndrome. Indeed, many mutants with altered responses to sugars have impaired ethylene or ABA signaling (Zhou et al., 1998; Arenas-Huertero et al., 2000; Finkelstein and Lynch, 2000; Huijser et al., 2008; Diaz-Riquelme et al., 2012; González-Grandio et al., 2013, 2017; Reddy et al., 2013; González-Grandio and Cubas, 2014; Yao and Finlayson, 2015). We propose that ethylene and ABA responses are closely connected to the C starvation syndrome. Indeed, many mutants with altered responses to sugars have impaired ethylene or ABA signaling (Zhou et al., 1998; Arenas-Huertero et al., 2000; Finkelstein and Lynch, 2000; Huijser et al., 2008; Diaz-Riquelme et al., 2012; González-Grandio et al., 2013, 2017; Reddy et al., 2013; González-Grandio and Cubas, 2014; Yao and Finlayson, 2015). We propose that ethylene and ABA responses are closely connected to the C starvation syndrome. Indeed, many mutants with altered responses to sugars have impaired ethylene or ABA signaling (Zhou et al., 1998; Arenas-Huertero et al., 2000; Finkelstein and Lynch, 2000; Huijser et al., 2008; Diaz-Riquelme et al., 2012; González-Grandio et al., 2013, 2017; Reddy et al., 2013; González-Grandio and Cubas, 2014; Yao and Finlayson, 2015).
ethylene and ABA response. C starvation leads to induction of ethylene and ABA-related genes, whereas sugar treatment has the opposite effect (Labbé et al., 2000; Rook et al., 2001; Brocard et al., 2002; León and Sheen, 2003; Yanagisawa et al., 2003; Thimm et al., 2004; Buchanan-Wollaston et al., 2005; Rolland et al., 2006). Thus C starvation signaling could trigger ethylene and ABA responses in buds. Indeed, the reduction in sugar levels detected in poplar buds during the 1 w SD is suggested to induce ethylene signaling, followed by ABA signaling (Ruttink et al., 2007).

One role of ABA and ethylene in the C starvation syndrome is induction of senescence, a genetically programmed process that promotes degradation of cell components and macromolecules, remobilizes nutrients, and optimizes resources to supply energy and C skeletons. Ethylene and ABA activate senescence-related genes and senescence induces ABA signaling (Abeles et al., 1988; Zeevaart and Creelman, 1988; Zacarias and Reid, 1990; Reid and Wu, 1992; Weaver et al., 1998; Seo et al., 2000; Yang et al., 2003; Buchanan-Wollaston et al., 2005; Lim et al., 2007). It is noteworthy that the potential master regulators of GRNII and III, AtAF1, ORE1/NAC6 and NAP, are ABA-induced factors that control senescence. AtAF1 induces a C starvation transcriptome and ABA biosynthesis (Jensen et al., 2013; Garapati et al., 2015). NAP activates SAG113/HAI1 and controls expression of ABBC3IC ALDEHYDE OXIDASE3 (AAO3), encoding an enzyme that catalyzes the final steps of ABA synthesis (Guo and Gan, 2006; Yang et al., 2014). ORE1 controls the expression of at least 78 SAGs and might also promote DNA degradation (Balazadeh et al., 2010; Matallana-Ramirez et al., 2013; Kim et al., 2014). HAT22, GFB2, GFB3, ABF3 and ABF4 are additional bud dormancy genes related both to senescence and ABA (Lin and Wu, 2004; Rivero et al., 2007; Song et al., 2016). Remarkably, MAX2/ORE9, which encodes an F-box involved directly in strigolactone perception and signaling and has a critical role in the control of shoot branching, also promotes senescence (Woo et al., 2001; Strinberg et al., 2002). Finally, the MYB genes A11g19000 and A11g74840, proposed to be master regulators of GRNII and IV, are responsive to dark-induced senescence (Lin and Wu, 2004).

In contrast, CK signaling is antagonistic to senescence, and a reduction in CK levels is a key signal for senescence initiation in Arabidopsis (Gan and Asamino, 1995; Kim et al., 2006). Consistent with this, in Arabidopsis dormant buds CK signaling is reduced, and in other species CK levels have also been reported to be reduced relative to active buds (Turnbull et al., 1997; Dun et al., 2012; Roman et al., 2016). Consistently, four genes encoding F-box proteins that promote the ubiquitination and degradation of ARR factors [KISS ME DEADLY (KMD)1-4] are bud dormancy genes.

Autophagy is another process induced by C starvation (Izumi et al., 2013) and whose markers (ATG genes) are upregulated in buds entering dormancy. This is a process by which cytoplasmic components and organelles are transported to the vacuole, where they are broken down and recycled. Under C-limiting conditions it contributes to plant energy availability (Aubert et al., 1996; Rose et al., 2006; Izumi et al., 2013). Autophagy is associated with induction of lipid degradation and upregulation of E2- and E3-ubiquitin ligase components, which promote proteasomal-dependent protein degradation (Thompson and Vierstra, 2005). We have found a remarkable number of bud dormancy genes related to autophagy, ubiquitination, protein degradation and lipid catabolism, many of them controlled by SnRK1 (see below).

**SnRK1 Could Have a Pivotal Role during the Bud Growth-to-Dormancy Transition**

SnRK1, a protein-kinase active in low energy conditions, promotes catabolism and represses anabolism, cell division and growth. Our transcriptomic data indicates that it may play an important role during the bud growth-to-dormancy transition. SnRK1 affects expression of robust dormancy markers such HIS1.3 and DRM1, and the potential master regulator of GRNI and GRNIV, MYBH/KUA1. In buds entering dormancy, the SnRK1 β subunit AKINBETA1, whose mRNA levels correlate directly with night duration (Pokhilko et al., 2014), is induced. Most importantly, our GSEA analysis indicates that the transcriptional network downstream of the catalytic SnRK1 α subunit, AKIN10, is significantly induced from the earliest stages of growth-to-dormancy transition in Arabidopsis, poplar and grapevine buds, and is maintained in para-, eco- and endormant buds. Many of the abovementioned genes involved in sugar sensing, signaling, autophagy and repression of CK signaling are AKIN10-dependent, including EXL4, STP1/14, SWEET11/12, TPS8/9/19/11, AKINBETA, ATG8β/F/G/H, ATG18F/G, and F-box genes KMD1, 3, 4. SnRK1 also causes downregulation of a large number of ribosomal genes, another conserved significant effect detected by our GSEA analysis. SnRK1 could also be responsible for at least part of the observed induction of the ubiquitination machinery and lipid degradation.

**A Conserved Core C Starvation Response Underlies Bud Dormancy in Angiosperms**

Bud dormancy is an adaptive response present in all angiosperms. It prevents shoot development when endogenous or environmental conditions are unfavorable for sustained growth. It has great impact on reproductive success, productivity and survival, and must have been influential in the colonization of habitats with fluctuating conditions.

We have found induction of a conserved C starvation syndrome that precedes and underlies the growth-to-dormancy transition in buds of three distantly-related species, one herbaceous (Arabidopsis) and two woody (poplar and grapevine). This transcriptional response, composed by several interconnected GRNs, comprises ortholog genes in Arabidopsis, poplar and grapevine, as gene sets generated in Arabidopsis were used to detect the response in the woody species. Furthermore, this syndrome has been observed in several unrelated experiments, regardless the stimulus that promoted dormancy, either environmental (low R:FR, short-day photoperiods) or endogenous (apical dominance). This remarkable conservation suggests that a syndrome aimed at
adapting to C-limiting situations is deeply rooted in the control of shoot meristem and bud development across angiosperms. Bud dormancy might thus be an ancestral response directly resulting from this C starvation syndrome, coordinated by different pathways that sense and/or anticipate situations on low C availability and feed into this core response to prevent untimely growth and development.

MATERIALS AND METHODS

Identification of Coregulated Genes in Bud Dormancy GRNs

Bud dormancy genes (Supplementary Figure S1) were obtained from González-Grandío and Cubas (2014). Coregulation of the 78 bud dormancy genes was analyzed by hierarchical clustering (Hclust, ATTED-II, Obayashi et al., 2007). Additional coregulated genes were obtained using CoEx-Search (ATTED-II, Obayashi et al., 2007). The 300 genes most coregulated with each cluster were selected. These genes were validated for induction in dormant buds in the original arrays (ATTED-II, Obayashi et al., 2007). The 300 genes most coregulated with each cluster were ranked by their log2 ratios calculated as the difference between normalized log intensity in the “dormant bud” condition minus normalized log intensity in the “active bud” condition. Intensity expression values were obtained from the references above. Gene sets for hormone markers were obtained from Nemhauser et al. (2006) and Mashiguchi et al. (2009). Gene sets related to sugar and AKIN10 responses were obtained from Sulipce et al. (2009) (Core C-signaling), Osuna et al. (2007) (Sugar-responsive), Gonzali et al. (2006) (Dark-induced, sugar-repressed), and Baena-González et al. (2007) (AKIN10-responsive). The other gene sets are from González-Grandío et al. (2013). The GSEA Normalized Enrichment Score for all gene sets in all comparisons were clustered with TM4 Multi Experiment Viewer (MeV, Saeed et al., 2003). Tree was generated by the hierarchical clustering method (HCL) using Euclidean distance and average linkage options. Complete results are in http://bioinfogp.cnb.csic.es/files/projects/tarancon_et_al_2017_supp/.

Promoter Motif Analysis

Sequences (1 kb) 5′ of the transcription start site of the bud dormancy GRN genes were retrieved with Sequence Bulk Download2. Overrepresented 6-8mer motifs were identified with Motif discovery (RSAT, Medina-Rivera et al., 2015). The oligo-analysis tool was used to find significantly overrepresented motifs, which were assembled into frequency matrices with pattern-assembly and default parameters. Matrices were converted into consensus motifs with convert-matrix and represented using WebLogo (Crooks et al., 2004).

Generation and Visualization of Poplar and Grapevine Expression Datasets

For each time point we calculated the log2 ratios of normalized gene intensities vs. normalized gene intensities on LD (active buds). Expression data was visualized and clustered with MeV. Tree was generated by HCL, using Euclidean distance and average linkage options.

Cell-Type Specific Shoot Apex Expression of Bud Dormancy GRN Genes

For each sample, we calculated the log2 ratios of normalized gene intensities vs. normalized gene intensities of the “active bud” sample: LD for poplar, April for grapevine. Expression data for selected bud dormancy genes obtained from Yadav et al. (2014) was visualized and clustered with MeV. Trees were generated by HCL using Euclidean distance and average linkage options.

Identification of Solanum tuberosum Orthologs

The putative orthologs of Arabidopsis genes were identified by a tblastn search with protein sequences as query in the Spud DB Potato Genomics Resource website3, cDNAs showing a high similarity e-value with the query were selected. Proteins were aligned with those of Arabidopsis and phylogenetic trees (BioNeighbor joining method, 500 replicates; Gascuel, 1997) were built to identify the most likely orthologs, which were selected for expression studies (Supplementary Figure S9).

---

2 www.arabidopsis.org
3 solanaceae.plantbiology.msu.edu/index.shtml
Quantitative-PCR Expression Analyses in Solanum tuberosum

Plant growth conditions, experimental design, light treatments, techniques and expression level normalization were as described in Nicolas et al. (2015). For each biological replicate, 8 axillary buds from node 2, and 8 from node 3 were dissected (node 1 = lowest plant node); 4–5 biological replicates were collected for each condition. Primers used are listed in Supplementary Table S1.

AUTHOR CONTRIBUTIONS

CT, EG-G, JO, and MN, performed experiments. PC, performed experiments, and wrote the manuscript.

REFERENCES


ACKNOWLEDGMENTS

We thank Desmond Bradley and Elena Baena for constructive criticisms of the manuscript and Catherine Mark for editorial assistance. PC is supported by a MINECO grant (BIO2014-57011-R). CT is a La Caixa predoctoral fellow. EG-G was a predoctoral fellow of Fundación Ramón Areces and a CSIC JAE-Predoc fellow. MN is a Excelence Severo Ochoa (MINECO) postdoctoral researcher.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2017.00788/full#supplementary-material
Tarancón et al.

control of oxidative stress-induced cell death in Arabidopsis by two related,
doi: 10.1073/pnas.1130421100
Fennell, A. Y., Schlauch, K. A., Gouthu, S., Deluc, L. G., Khadka, V., Sreekantan, L.,
et al. (2015). Short day transcriptomic programming during induction of
gene ABI5 encodes a basic leucine zipper transcription factor. Plant Cell 12,
bud outgrowth and is homologous to monocot Teosinte Branched1. Plant Cell
Physiol. 48, 667–677. doi: 10.1093/pcp/pcm044
1986
carbon starvation responses with trehalose metabolism. Plant Physiol. 169,
379–390. doi: 10.1104/pp.15.00917
Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on
oxfordjournals.molbev.a025808
Goddijn, O. J., Verwoerd, T. C., Voogd, E., Krutwagen, R. W., de Graaf, P. T.,
van Dun, K., et al. (1997). Inhibition of trehalase activity enhances trehalose
pp.113.1.181
associated to active and dormant buds in Arabidopsis. Plant Signal. Behav. 9,
19–21. doi: 10.4161/psb.27994
González-Grandío, E., Pajoro, A., Franco-Zorrilla, J. M., Tarancón, C., Immink,
R. G. H., and Cubas, P. (2017). Abscisic acid signaling is controlled by a
BRANCHED1 promotes axillary bud dormancy in response to shade in
Identification of sugar-modulated genes and evidence for in vivo sugar sensing
Guo, Y., and Gan, S. (2006). AtNAP, a NAC family transcription factor, has an
Comprehensive transcriptome analyses reveal differential gene expression
profiles of Camellia sinensis axillary buds at Para-, Endo-, ecodormancy, and
Haydon, M. J., Mielczarek, O., Robertson, F. C., Hubbard, K. E., and Webb, A. A. R.
Nature 502, 689–692. doi: 10.1038/nature12603
Horvath, D. P., Chao, W. S., Suttle, J. C., Thimmapuram, J., and Anderson, J. V.
(2008). Transcriptome analysis identifies novel responses and potential
regulatory genes involved in seasonal dormancy transitions of leafy
Howe, G. T., Horvath, D. P., Dharmawardhana, P., Priest, H. D., Mockler, T. C.,
doi: 10.3389/fpls.2015.00989
A single-repeat MYB transcription repressor, MYBH, participates in regulation
s11103-015-0321-2

Frontiers in Plant Science | www.frontiersin.org

Carbon Starvation Response Underlies Bud Dormancy

(2000). The Arabidopsis SUCROSE UNCOUPLED-6 gene is identical to
ABSCISIC ACID INSENSITIVE-4: involvement of abscisic acid in sugar
to nighttime energy availability for growth in Arabidopsis. Plant Physiol. 161,
1682–1693. doi: 10.1104/pp.113.215632
(2013). ATAF1 transcription factor directly regulates abscisic acid biosynthetic
j.fob.2013.07.006
Photoresponses of light-grown phyA mutants of Arabidopsis (phytochrome
a is required for the perception of daylength extensions). Plant Physiol. 105,
141–149. doi: 10.1104/pp.105.1.141
sorghum axillary bud outgrowth by shade, phyB and defoliation signalling
02050.x
Gene regulatory cascade of senescence-associated NAC transcription factors
activated by ETHYLENE-INSENSITIVE2-mediated leaf senescence signalling
(2006). Cytokinin-mediated control of leaf longevity by AHK3 through
814–819. doi: 10.1073/pnas.0505150103
sugar-insensitive mutants sis4 and sis5 are defective in abscisic acid synthesis
Lang, G. A., Early, J. D., Arroyave, N. J., Darnell, R. L., Martin, G. C., and Stutte,
20, 809–811.
and ecodormancy: physiological terminology and classification for dormancy
Lauxmann, M. A., Annunziata, M. G., Brunoud, G., Wahl, V., Koczut, A.,
transient carbohydrate limitation: flowers and very young siliques are jettisoned
and the meristem is maintained to allow successful resumption of reproductive
8, 110–116. doi: 10.1016/S1360-1385(03)00011-6
of the protein phosphatase 2C, AtAIP1, in abscisic acid signaling and sugar
01.013
responses and connects the extracellular carbon status to growth. Front. Plant
with one DNA binding repeat mediate sugar and hormone regulation of
001735
273, 10120–10131. doi: 10.1074/jbc.273.17.10120
de la Viticultura. Madrid: Mundi-Prensa.

May 2017 | Volume 8 | Article 788 | 161




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

Copyright © 2017 Tarancón, González-Grandío, Oliveros, Nicolas and Cubas. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.
Evolutionary Analysis of DELLA-Associated Transcriptional Networks

Asier Briones-Moreno¹, Jorge Hernández-García¹, Carlos Vargas-Chávez², Francisco J. Romero-Campero³⁴, José M. Romero⁴, Federico Valverde⁴ and Miguel A. Blázquez¹*¹

¹ Instituto de Biología Molecular y Celular de Plantas, Consejo Superior de Investigaciones Científicas – Universidad Politécnica de Valencia, Valencia, Spain, ² Institute for Integrative Systems Biology (I2SysBio), University of Valencia, Valencia, Spain, ³ Department of Computer Science and Artificial Intelligence, Universidad de Sevilla, Sevilla, Spain, ⁴ Instituto de Bioquímica Vegetal y Fotosíntesis, Consejo Superior de Investigaciones Científicas – Universidad de Sevilla, Sevilla, Spain

DELLA proteins are transcriptional regulators present in all land plants which have been shown to modulate the activity of over 100 transcription factors in Arabidopsis, involved in multiple physiological and developmental processes. It has been proposed that DELLAs transduce environmental information to pre-wired transcriptional circuits because their stability is regulated by gibberellins (GAs), whose homeostasis largely depends on environmental signals. The ability of GAs to promote DELLA degradation coincides with the origin of vascular plants, but the presence of DELLAs in other land plants poses at least two questions: what regulatory properties have DELLAs provided to the behavior of transcriptional networks in land plants, and how has the recruitment of DELLAs by GA signaling affected this regulation. To address these issues, we have constructed gene co-expression networks of four different organisms within the green lineage with different properties regarding DELLAs: Arabidopsis thaliana and Solanum lycopersicum (both with GA-regulated DELLA proteins), Physcomitrella patens (with GA-independent DELLA proteins) and Chlamydomonas reinhardtii (a green alga without DELLA), and we have examined the relative evolution of the subnetworks containing the potential DELLA-dependent transcriptomes. Network analysis indicates a relative increase in parameters associated with the degree of interconnectivity in the DELLA-associated subnetworks of land plants, with a stronger effect in species with GA-regulated DELLA proteins. These results suggest that DELLAs may have played a role in the coordination of multiple transcriptional programs along evolution, and the function of DELLAs as regulatory ‘hubs’ became further consolidated after their recruitment by GA signaling in higher plants.

Keywords: gene co-expression networks, integrative molecular systems biology, evo–devo, transcriptional regulation, plant signaling

INTRODUCTION

Higher plants are characterized by a particularly flexible capacity to adapt to multiple environmental conditions. In other words, environmental signals are very efficient modulators of plant developmental decisions. This ability is generally assumed to be based on at least two mechanistic features: the presence of an extensive and sensitive repertoire of elements that perceive
environmental signals (such as light photoreceptors covering a wide range of wavelengths), and the high degree of interconnectivity between the different signaling pathways to allow cellular integration of variable information (Casal et al., 2004).

Evidence has accumulated in recent years about the important role that plant hormones play in the translation of environmental signals into developmental decisions. On one hand, it has become evident that hormone pathways share common components with the pathways that transduce light and other environmental signals (Jailais and Chory, 2010); and, on the other hand, hormones have been shown to participate in the regulation of developmental processes all throughout a plant's life cycle (Alabadi et al., 2009). In this context, gibberellins (GAs) and DELLA proteins are a paradigmatic example of the mechanisms that allow environmental signal integration. DELLA proteins constitute a small clade within the GRAS family of loosely defined plant specific nuclear proteins (Vera-Sirera et al., 2015). Their name was coined on the basis of a short stretch of amino acids (D-E-L-L-A) in their N-terminal region, which is tightly conserved among all higher plant species. They also present additional conserved motifs, such as the VHYNP domain, two leucine heptad repeats which may mediate protein–protein interactions, a putative nuclear localization signal, and a putative SH2 phosphotyrosine-binding domain, among others (Vera-Sirera et al., 2015). It has been shown in Arabidopsis thaliana and rice that recognition of GAs by their GID1 receptor allows physical interaction with DELLA proteins and promotes their degradation via the proteasome. In A. thaliana, loss of DELLA function mimics the phenotype of plants treated with an excess of GAs, both anatomically and also at the transcriptional level (Schwechheimer, 2011; Locascio et al., 2013b). Work in the past few years has established that DELLAs regulate transcription through the interaction with more than 100 transcription factors (TFs) in A. thaliana (de Lucas et al., 2008; Feng et al., 2008; Crocco et al., 2010; Hou et al., 2010; Gallego-Bartolomé et al., 2012; Daviere et al., 2014; Marin-de la Rosa et al., 2014, 2015; Resentini et al., 2015). In some cases, interaction with the TF inhibits its ability to bind DNA, while in other cases DELLAs seem to act as co-activators (Locascio et al., 2013b; Daviere and Achaard, 2016). For all the cases examined in detail, the DELLA region responsible for the interaction with the TFs is the C-terminal region of the protein, the GRAS domain. Given that GA levels are strongly regulated by environmental signals such as light, temperature and photoperiod (Hedden and Thomas, 2012; Colebrook et al., 2014), cellular DELLA levels seem to be a proxy for the environmental status faced by plants (Claeys et al., 2014). Changes in DELLA levels could in turn differentially modulate distinct sets of TFs and their target genes in various developmental contexts. The promiscuous interaction with TFs, and the observation that A. thaliana dellaKO mutants display constitutive growth even under stress, and suffer from increased sensitivity to several types of stress factors such as salinity, cold, or fungal attacks (Alabadi et al., 2004; Achard et al., 2006, 2007, 2008a,b; Cheminant et al., 2011) suggests that DELLAs are potentially important ‘hubs’ in the transcriptional network that regulates the balance between growth and stress tolerance in higher plants.

Previous interest in the evolution of DELLA proteins is restricted to the question on how they were recruited to mediate cellular signaling by GAs. Based on phylogenetic analyses and shallow molecular analysis with fern and moss orthologs, it seems that the GA/GID1/DELLA module originated with early diverging tracheophytes (Wang and Deng, 2014). For instance, the Selaginella genus possesses the ability to synthesize GAs, a GID1 GA receptor, and a DELLA protein (Wang and Deng, 2014), which is sensitive to GA-induced degradation, even when introduced in an angiosperm, such as A. thaliana (Hirano et al., 2007; Yasumura et al., 2007). On the other hand, the DELLA proteins that existed in other land plants before the emergence of vascular plants were not involved in GA signaling. First, there are no bona-fide DELLA genes in algae and, second, the genomes of bryophytes like Physcomitrella patens encode DELLA proteins that lack the canonical ‘DELLA motif’ (Wang and Deng, 2014), and PpDELLAs are not sensitive to GAs when introduced in A. thaliana (Yasumura et al., 2007). However, the ability of DELLA proteins to modulate transcriptional programs relies on the GRAS domain through which interactions with TFs occur, and the evolution of this activity has not been addressed before.

In an attempt to identify the possible function of ancestral DELLAs and to delineate how evolution has shaped the functions of the GA/DELLA module in higher plants, we have addressed the analysis of the transcriptional networks potentially regulated by DELLAs in several species. For this reason, we have used gene co-expression networks, in which genes are represented as nodes, and if two genes exhibit a significant correlation value for co-expression, the corresponding nodes are joined by an edge. Importantly, if a node is a TF, first neighbors can be confidently taken as targets for that particular TF (Franco-Zorrilla et al., 2014). Therefore, the analysis of topological parameters of a gene co-expression network is an interesting tool that may reveal information about the function and evolutionary history of transcriptional programs (Aoki et al., 2007; Usadel et al., 2009).

Here we have investigated the properties of networks formed by DELLA-interacting TFs and their co-expressing genes in A. thaliana, and compared them with the orthologous networks in three other plant species: (i) Solanum lycopersicum (possessing a fully operative GA/DELLA module); (ii) P. patens (possessing GA-independent DELLA functions); and (iii) Chlamydomonas reinhardtii (without GA perception or DELLAs) (Figure 1A). All the parameters examined suggest that the functions regulated by DELLA-interacting TFs (and thus DELLAs themselves) have increased their level of coordination along evolution.

RESULTS AND DISCUSSION

Construction of Networks and Subnetworks

Gene expression data from RNA sequencing (RNA-seq) experiments in A. thaliana, S. lycopersicum, P. patens, and C. reinhardtii were obtained from the Gene Expression Omnibus, and gene co-expression networks were inferred for
each species from transcriptomic data as described in section “Materials and Methods.” All four networks are scale-free networks (Supplementary Figure S1) (Romero-Campero et al., 2013, 2016) and have comparable sizes in terms of number of nodes, but there are remarkable differences in the way they are connected (Table 1). The A. thaliana network contains more than twice as many edges than the others, the average degree of its nodes (average number of connections) is one order of magnitude higher and its average shortest path length (average number of nodes between two random nodes) is lower. Even though the number of genes of each species represented in the networks is similar, in some species they are more connected, possibly due to differences in their endogenous regulation and the availability of experimental data. For that reason, we decided to do every comparative analysis between the different species in relative terms.

To be able to compare the co-expression networks of the different species, we first identified the orthologous nodes in each of them using the OrthoMCL method (Li et al., 2003). Up to 17,053 groups of genes were obtained. Genes in the same group were considered orthologs or paralogs if they belonged to different or the same species, respectively. The four species were represented unequally, as both A. thaliana and S. lycopersicum genes were present in ca. 70% of the groups, while P. patens genes were found in little more than 50% of them, and only ca. 30% of the groups contained genes from C. reinhardtii.

Table 1 | General parameters in co-expression networks.

<table>
<thead>
<tr>
<th></th>
<th>C. reinhardtii</th>
<th>P. patens</th>
<th>S. lycopersicum</th>
<th>A. thaliana</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Neigh Ortho</td>
<td>8652 48 658</td>
<td>8564 448 1503</td>
<td>7851 1314 2885</td>
<td>5663 2070 2949</td>
</tr>
<tr>
<td>Edges</td>
<td>145903 78 1173</td>
<td>295317 15078 19828</td>
<td>287409 153396 169171</td>
<td>593730 460951 510242</td>
</tr>
<tr>
<td>Average degree</td>
<td>33.73 3.25 3.57</td>
<td>68.97 67.31 26.38</td>
<td>73.22 233.48 1.39</td>
<td>209.69 445.36 347.26</td>
</tr>
<tr>
<td>Average shortest path length</td>
<td>7.37 1.91 8.71</td>
<td>13.11 1.39 12.01</td>
<td>13.78 1.67 5.63</td>
<td>4.28 2.15 3.09</td>
</tr>
<tr>
<td>Diameter</td>
<td>23 4 24</td>
<td>46 4 41</td>
<td>44 6 25</td>
<td>20 9 12</td>
</tr>
</tbody>
</table>

Parameters of networks and subnetworks used in this study. Full, full gene co-expression network; Neigh, first neighbors subnetwork; Ortho, orthologs subnetwork; C. reinhardtii, Chlamydomonas reinhardtii; P. patens, Physcomitrella patens; S. lycopersicum, Solanum lycopersicum; A. thaliana, Arabidopsis thaliana.
been described for in their respective networks. Since most of the interactions have obtained their orthologs for the four species, and localized them in their respective networks. Since most of the interactions have been described for A. thaliana, the corresponding orthologs in the other species are only “putative interactors of the DELLA proteins” (PIDs), and the first neighbors of AtDELLA interactors and PIDs are their putative expression targets. Second, we built two different subnetworks using this information. The first one, called “Neighbors” subnetwork (abbreviated as AtNeigh, SlNeigh, PpNeigh, and CrNeigh), is composed of the DELLA interactors (or the corresponding PIDs) and their first neighbors (Figure 1C and Supplementary Table S2). The second one, called “Orthologs” subnetwork (abbreviated as AtOrtho, SlOrtho, PpOrtho, and CrOrtho), contains the orthologs of all the first neighbors of PIDs in all the species (Figure 1C and Supplementary Table S3). For a given species, the “Neighbors” subnetwork provides a good approximation to its actual DELLA-dependent transcriptome, while the “Orthologs” subnetwork represents the full landscape of potential transcriptional targets for DELLAs, since it includes orthologs of genes that are DELLA transcriptional targets in other species (Figure 2).

**DELLA-Associated Subnetworks Reflect Increased Relevance of DELLAs after Being Recruited by GA Signaling**

It is important to take into account a circumstance that affects the construction of subnetworks: OrthoMCL does not always retrieve orthologs for some of the genes, because either they do not exist in the other species, or the method does not provide high-confidence results. This results in a particular bias toward smaller subnetwork sizes with increasing phylogenetic distance (Table 1). However, the impact of this bias can be disregarded when analyzing relative parameters. Hence, regardless of the absolute sizes, we observed that the average degree in the Neighbor subnetworks increased dramatically in SlNeigh and AtNeigh with respect to their full networks (more than threefold and twofold, respectively), while this parameter did not change in PpNeigh, and it actually decreased in CrNeigh (Table 1). Similarly, the Orthologs subnetworks displayed an equivalent behavior as the Neighbors subnetworks: their diameter and average shortest path length decreased considerably more in SlOrtho and AtOrtho with respect to the full networks; and the same happened with the increase of the average degree. In summary, both subnetworks showed a higher compaction and interconnection of nodes in relative terms in the case of S. lycopersicum and A. thaliana compared with P. patens and C. reinhardtii, indicating that the putative interactors and targets of the DELLAs become more connected in those species presenting GA-regulated DELLAs.

A confirmation of the impact of GA regulation on the relevance of DELLA function is found in the analysis of neighborhood conservation. Figure 3A shows the percentage of genes with a significantly overlapping neighborhood in each comparison (see Materials and Methods). When comparing P. patens with the other species, there are no substantial differences between the full network and the Orthologs subnetwork. On the contrary, SlOrtho and AtOrtho contain a considerably higher proportion of genes with conserved neighborhood than their corresponding full networks (15% vs. 10%). Between S. lycopersicum and A. thaliana, the regulation of the putative DELLA targets is more conserved than for the network in general, so this group of genes seems to have a cohesive element in the two species.

Furthermore, we examined gene–gene co-expression values, as a measure of the conservation of individual edges. For every pair of linked genes in one species, if the corresponding orthologs are also linked in a second species, it is considered that gene–gene co-expression is conserved. Therefore, the calculation of conserved links between two subnetworks is a measure of functional conservation of a regulatory module. Interestingly, we observed that gene links between PpOrtho and SlOrtho were less conserved than in the full networks, and almost unaltered between PpOrtho and AtOrtho (Figure 3B). However, the gene–gene co-expression was three times more conserved between SlOrtho and AtOrtho than between their full networks (11% vs. 3.5%). In other words, these data are compatible with the proposition that the presence of GA-regulated DELLAs (in S. lycopersicum and A. thaliana) provides stronger links between transcriptional programs, not detected in an organism with GA-independent DELLAs (P. patens).

**Efficiency of Transcriptional Regulation Is a DELLA-Associated Parameter**

The efficiency of a transcriptional regulatory mechanism can be evaluated through two additional parameters in gene co-expression networks: shortest path length distribution and motif frequency. In network theory, average shortest-path length is defined as the average number of steps along the shortest paths for all possible pairs of network nodes. It is a measure of the efficiency of information propagation on a network, with a shorter average path length being more efficient (Vragovic et al., 2005). When we compared the distribution of shortest path lengths in full and Orthologs subnetworks, we observed a clear tendency toward shorter path lengths in the Orthologs subnetworks of organisms possessing DELLAs (S. lycopersicum, A. thaliana, and P. patens) compared with the situation in an organism without DELLAs (C. reinhardtii) (Figure 4).

Network motifs are small recurring patterns involving a few nodes that appear more frequently in biological networks than in randomized ones. They consist of a certain level of regulation which connects small sets of nodes with a particular topology. Motifs characterize a network, as some of them are useful for the regulation of determined functions, and thus conserved along evolution (Kashan and Alon, 2005). After measuring the frequency of the eight common motifs
FIGURE 2 | Gene co-expression networks. Full *Chlamydomonas reinhardtii* (A,E), *Physcomitrella patens* (B,F), *Solanum lycopersicum* (C,G), and *Arabidopsis thaliana* (D,H) gene co-expression networks. Neighbors subnetworks are comprised of yellow-marked nodes in A-D. Orthologs subnetworks are comprised of yellow-marked nodes in (E–H).

FIGURE 3 | Gene connections are more conserved in species with GA-regulated DELLAs. Pairwise comparisons of *P. patens*, *S. lycopersicum*, and *A. thaliana* Full networks and Ortho subnetworks regarding: (A) Percentage of genes with significantly overlapping neighborhoods; (B) Percentage of conserved gene–gene links.

composed of three and four nodes in the full networks, we found that there was no relative enrichment of any particular motif between species when comparing the full networks or the Orthologs subnetworks (Figure 5A). However, the AtOrtho, SlOrtho, and PpOrtho subnetworks displayed a clear enrichment in virtually every motif, compared with their respective full networks (Figure 5B). Given that the function of this sort of motifs is to allow coordinated expression of a group of genes with shared function (Alon, 2007), the increase in the proportion of small regulatory patterns among all the putative DELLA targets in species that do contain DELLAs indicates an increase in the complexity of gene regulation, in which DELLAs might mediate the coordination of transcriptional programs.
FIGURE 4 | Paths are shorter in DELLA-associated subnetworks. Shortest path length distribution in Full networks and Orthologs subnetworks from the four species. The graphs represent the relative number of nodes (y-axis) joined by a given number of intermediate nodes (x-axis).

FIGURE 5 | Network motifs are enriched in DELLA related networks. (A) Percentage of motifs found in each network compared to possible combinations of three and four nodes. (B) Ratio of motif enrichment comparing Orthologs subnetworks to Full networks per species (upper panel), and per motif (lower panel). Dashed lines in (B) mark a ratio of 1. Motifs are as depicted in X-axis. Letters indicate significant differences between groups, p < 0.01 (One way ANOVA, Tukey HSD Post Hoc test). Box-plot whiskers are Tukey-defined (extended 1.5 times the IQR from the box edges).
The Regulation of the Stress Response: A Likely Role of Ancestral DELLAs Proteins

The results shown above suggest that the origin of DELLAs in land plants would be associated to an increase in the co-expression between genes that are putative targets of DELLA-interacting TFs, both in terms of size of the gene set and degree of the co-expression value. Therefore, DELLAs would have helped in the coordination of certain transcriptional circuits, and their recruitment to mediate GA signaling later in development would have further expanded their coordination capacity. To reveal the most likely functions ultimately regulated by DELLAs in the common ancestor of land plants, we carried out Gene Ontology (GO) analyses on each of the Neighbor subnetworks, with the idea that the terms shared by those in *S. lycopersicum*, *A. thaliana*, and *P. patens* could represent likely functions regulated by the ancestral DELLAs proteins.

Not surprisingly, given the larger size of AtNeigh (Table 1), GO analysis rendered a much larger number of terms significantly enriched in this subnetwork, compared to those from the other three organisms (Supplementary Table S4). Terms referring to chloroplast function, such as plastid organization, photosynthesis, or pigment biosynthesis (including chlorophyll) were specifically enriched among the putative DELLA targets in *A. thaliana* only (Figure 6). This result might reflect functions whose regulation by DELLAs has been acquired more recently, or it could simply be a bias of the analysis, caused by the big difference in size of the analyzed sets in the different species. On the contrary, the finding that terms comprised under general ‘response to stress’ were significantly over-represented in the subnetworks of the three land plants, but not *C. reinhardtii*, suggests that this function might have been the primary target of the regulation by ancestral DELLAs through their interaction with specific TFs.
CONCLUSION

Our analysis suggests that DELLAs may have contributed to the acquisition of an increasing degree of coordination between transcriptional programs during plant evolution. Although these results are consistent with the current view of DELLAs as ‘hubs’ in transcriptional programs in higher plants, and provide a plausible evolutionary scenario, it is important to remark that further experimental work is required to validate most of the conclusions from in silico network analysis. In fact, several reasonable assumptions have been made that would be relatively easy to confirm. For instance, actual transcriptomic data of dellaKO mutants in the different species, coupled to comparative analysis would help establish the role of ancestral DELLAs. Moreover, our current analysis would be strengthened by the experimentally obtained information of which PIDs are in fact bona-fide DELLA interactors in the different species. Finally, the conclusion that DELLAs have probably contributed to the establishment of new co-regulatory circuits during land-plant evolution does not explain the molecular mechanism that supports this progressive acquisition, and it can be generated by changes in DELLA proteins, in their interactors, or in both.

MATERIALS AND METHODS

Gene Co-expression Network Inference

The C. reinhardtii and A. thaliana networks were downloaded from the web resources of previous work (Romero-Campero et al., 2013, 2016). For the new networks, RNA-seq data were selected from equivalent experiments involving comparable tissues and environmental situations (Supplementary Table S5). The P. patens gene co-expression network was inferred from the RNA-seq data freely available from the Gene Expression Omnibus identified with accession numbers GSE19824, GSE33279, GSE36274, and GSE25237. The S. lycopersicum network was constructed based on the RNA-seq data identified with the accession numbers GSE45774, GSE64665, GSE64981, GSE68018, and GSE77340 in the Gene Expression Omnibus. In both cases, RNA-seq data was processed using the Tuxedo protocol (Trapnell et al., 2012) to obtain gene expression levels measured as FPKM. Briefly, short reads were mapped to the corresponding reference genome using TopHat, transcripts were assembled using Cufflinks and expression levels were computed using Cuffdiff. The Bioconductor R package cummeRbund (Goff et al., 2013) was used for subsequent analysis of the results generated by the Tuxedo protocol. In order to reduce noise in our analysis only genes that were detected as differentially expressed in at least one of the studies integrated in this work were considered. Differentially expressed genes were determined comparing each condition with the corresponding control within each study using a fold-change threshold of two. For each species, a matrix containing the expression levels of the selected genes was extracted. The Pearson correlation coefficient between every pair of gene expression profiles was computed using the cor function from the stats R package to generate a correlation matrix. Two genes were assumed to be co-expressed when the Pearson correlation coefficient between their expression profiles over the analyzed conditions was greater than 0.95. Following this criterion, the corresponding adjacency matrix was generated from the correlation matrix. Using the R package igraph1 (Csardi and Nepusz, 2006), each network was constructed from its adjacency matrix and exported in gml format for subsequent analysis.

Data Compilation and Processing

The reference proteomes from A. thaliana TAIR10, S. lycopersicum iTAGv2.3, C. reinhardtii v5.5, and P. patens v3.3 were downloaded from Phytozome (Goodstein et al., 2012). From all the possible proteins from each locus tag only the longest protein was kept and assigned to its locus tag. These files were used to identify the orthologs among the four species with OrthoMCL (Li et al., 2003).

The networks were converted to SIF format and processed using the package igraph2 (Csardi and Nepusz, 2006) made with R2 (R Core Team, 2016). Only the edges between two non-identical nodes were conserved. If a given node was not identified in the proteome files, it was removed from the network. Afterward, components with fewer than seven elements were removed from the network to generate the complete network for each species. The orthologs for the set of manually curated DELLA interactors from A. thaliana were identified, and these nodes were selected from the complete networks. The first neighbors for all the selected nodes were identified and used to build a subnetwork. Finally, the orthologs on each species for all the genes in the previous subnetworks were identified and used to generate a new subnetwork for each species.

Network Analysis and Visualization

All networks were imported into the software package Cytoscape (Smoot et al., 2011) for their visualization using the Prefuse Force Directed layout.

The measures of network topology were calculated using both predefined and custom made functions. The gene–gene co-expression and neighborhood conservation were determined following the approach described by Netotea et al. (2014), using Fisher exact tests to check for statistical significance.

Gene Ontology analysis on Neigh subnetworks was made with AgriGO (Du et al., 2010), and represented with ReviGO (Supek et al., 2011).

AUTHOR CONTRIBUTIONS

AB-M, JH-G and MB conceived and designed the work. FR-C, JR, and FV constructed the co-expression networks. AB-M, CV-C, and JH-G performed network analyses. AB-M and MB wrote the first draft of the manuscript, to which all authors contributed.

1http://igraph.com
2http://www.Rproject.org
ACKNOWLEDGMENTS

We thank the members of the Hormone Signaling and Plasticity Lab at IBMCP (http://www.ibmcp.upv.es/BlazquezAlabadiLab/) for useful discussions and suggestions.

REFERENCES


SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2017.00626/full#supplementary-material

TABLE S1 | Compilation of DELLA interactors used in this study.

TABLE S2 | Genes included in the ‘Neighbors’ subnetworks.

TABLE S3 | Genes included in the ‘Orthologs’ subnetworks.

TABLE S4 | Gene Ontology categories enriched in the ‘Neighbors’ subnetworks.

TABLE S5 | RNA-seq datasets used for the construction of the new gene co-expression networks in Physcomitrella patens and Solanum lycopersicum.


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

Copyright © 2017 Briones-Moreno, Hernández-García, Vargas-Chávez, Romero-Campero, Romero, Valverde and Blázquez. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.
Transcriptomic Analysis Implies That GA Regulates Sex Expression via Ethylene-Dependent and Ethylene-Independent Pathways in Cucumber (Cucumis sativus L.)

Yan Zhang1,2, Guiye Zhao1,2, Yushun Li1,2, Ning Mo1,2, Jie Zhang1,2 and Yan Liang1,2*

1 College of Horticulture, Northwest A&F University, Yangling, China, 2 State Key Laboratory of Crop Stress Biology in Arid Region, Northwest A&F University, Yangling, China

Sex differentiation of flower buds is an important developmental process that directly affects fruit yield of cucumber (Cucumis sativus L.). Plant hormones, such as gibberellins (GAs) and ethylene can promote development of male and female flowers, respectively, however, the regulatory mechanisms of GA-induced male flower formation and potential involvement of ethylene in this process still remain unknown. In this study, to unravel the genes and gene networks involved in GA-regulated cucumber sexual development, we performed high throughout RNA-Seq analyses that compared the transcriptomes of shoot tips between GA3 treated and untreated gynoecious cucumber plants. Results showed that GA3 application markedly induced male flowers but decreased ethylene production in shoot tips. Furthermore, the transcript levels of M (CsACS2) gene, ethylene receptor CsETR1 and some ethylene-responsive transcription factors were dramatically changed after GA3 treatment, suggesting a potential involvement of ethylene in GA-regulated sex expression of cucumber. Interestingly, GA3 down-regulated transcript of a C-class floral homeotic gene, CAG2, indicating that GA may also influence cucumber sex determination through an ethylene-independent process. These results suggest a novel model for hormone-mediated sex differentiation and provide a theoretical basis for further dissection of the regulatory mechanism of male flower formation in cucumber.

Statement: We reveal that GA can regulate sex expression of cucumber via an ethylene-dependent manner, and the M (CsACS2), CsETR1, and ERFs are probably involved in this process. Moreover, CAG2, a C-class floral homeotic gene, may also participate in GA-modulated cucumber sex determination, but this pathway is ethylene-independent.

Keywords: cucumber, ethylene, gibberellin, sex expression, transcriptome

Abbreviations: ACC, 1-aminocyclopropane-1-carboxylate; DEG, differentially expressed gene; DGE, digital gene expression; FDR, false discovery rate; FID, flame ionization detector; FPKM, fragments per kilobase of transcript sequence per millions base pairs sequenced; GA, gibberellin; GO, gene ontology; qRT-PCR, quantitative real-time PCR.
INTRODUCTION

Cucumber (*Cucumis sativus* L.) is a typical monoecious plant with distinct male and female flowers, and has been served as a model system for studying physiological and molecular aspects of sex determination in plants (Malepszy and Niemirowicz-Szczytt, 1991; Bai and Xu, 2013). During the early stages of cucumber flower development, both stamen primordia and carpel primordia are initiated, however, sex differentiation occurs just after the hermaphroditic stage, subsequently, female or male flower is formed and developed through the selective developmental arrest of stamen or carpel, respectively (Bai et al., 2004).

Sex differentiation in cucumber is mainly determined by *F*, *M*, and *A* genes. Among them, *F* (*CsACS1G* and *M* (*CsACS2*) genes encoding two ACC synthases (key enzymes in ethylene biosynthetic pathway) govern female sex expression in cucumber, and the *F* gene promotes female flower development (Trebitsh et al., 1997; Mibus and Tatlıoğlu, 2004; Knopf and Trebitsh, 2006), while the *M* gene inhibits stamen development in flower buds (Yamasaki et al., 2001, 2003; Saito et al., 2007; Li et al., 2009, 2012). In contrast, the *A* gene inhibits female flower development and facilitates male flower formation (Pierce and Wehner, 1990). The interaction of *F*, *M*, and *A* genes eventually determines various sexual phenotypes of cucumber.

In addition to genetic control, sex expression of cucumber can be affected by phytohormones, such as ethylene and GAs. Particularly, ethylene is considered as a potent sex hormone in cucumber that can induce formation of female flowers (Malepszy and Niemirowicz-Szczytt, 1991). Ethylene content in shoot tip of gynoecious cucumber is higher than that of monoecious plant (Rudich et al., 1972; Fujita and Fujieda, 1981; Trebitsh et al., 1987). Treatment with exogenous ethylene or ethylene-releasing reagent can increase the numbers of female and bisexual flowers in monoecious and andromonoecious lines, respectively (MacMurray and Miller, 1968; Iwahori et al., 1969). Until now, the molecular mechanism of ethylene-regulated sex determination of cucumber has been well understood. Except for the *F* and *M* genes, other ethylene biosynthetic genes, such as *CSACO2* and *CSACO3*, which encode ACC oxidases, are also involved in sex expression of cucumber, but the transcript levels of *CSACO2* and *CSACO3* in the shoot tips show a negative correlation with femaleness, indicating an existence of a feedback inhibition mechanism underlying such correlation (Kahana et al., 1999). Overexpression of *CsACO2*, driven by the AP3 promoter, can arrest the stamen development by inducing chromatin condensation in *Arabidopsis* (Hao et al., 2003; Duan et al., 2008). Moreover, an ethylene receptor, *CsETR1*, has been demonstrated to play a key role in stamen arrest in female cucumber flowers through induction of DNA damage (Wang et al., 2010).

Gibberellins, one class of tetracyclic diterpenoid phytohormones, can promote the male tendency in cucumber. GA production in andromonoecious cucumber in higher than that in gynoecious and monoecious plants (Hemphill et al., 1972). Exogenous GA3 application can increase the ratio of maleness to femaleness in monoecious cucumber and induce the formation of male flowers in gynoecious plants (Wittwer and Bukovac, 1962; Pike and Peterson, 1969). In addition, GA signaling pathway is involved in stamen and anther development in hermaphroditic plants, such as *Arabidopsis* and rice (*Oryza sativa*) (Cheng et al., 2004; Fleet and Sun, 2005; Aya et al., 2009; Sun, 2010, 2011; Plackett et al., 2011; Song et al., 2013). In this pathway, GA first binds with GID1 receptor and promotes the interaction between GID1 and DELLA proteins (repressors of GA signaling), leading to a rapid degradation of DELLA proteins by an ubiquitin-proteasome pathway, and the proteolysis of DELLA proteins releases their inhibitory effect on GA action and allows plant growth and development (Fleet and Sun, 2005; Murase et al., 2008; Harberd et al., 2009; Sun, 2010, 2011; Plackett et al., 2014). *GAMYB* is a positive regulator in GA signaling pathway and acts as an important downstream gene of DELLA proteins (Olszewski et al., 2002; Achard et al., 2004; Fleet and Sun, 2005). GA can induce *GAMYB* transcript through degradation of DELLA proteins, resulting in an enhanced flowering and anther development (Achard et al., 2004). In our previous studies, we identified two GA signaling genes, *CsGAIP* and *CsGAMYB1*, which belong to DELLA and GAMYB family, respectively. Both of them were predominantly expressed in the male specific organs during cucumber flower development. *CsGAIP* can inhibit stamen development through transcriptional repression of B class floral homeotic genes *APETALA3* (AP3) and *PISTILLATA* (PI) in *Arabidopsis* (Zhang et al., 2014a). However, whether *CsGAIP* is involved in GA-regulated sex determination in cucumber flowers is still unknown. Notably, *CsGAMYB1* can also mediate sex expression of cucumber. Knockdown of *CsGAMYB1* in cucumber results in decreased ratio of nodes with male to female flowers (Zhang et al., 2014b). Despite the current knowledge of GA-regulated sex expression of cucumber, the precise regulatory pathway in this complex process remains elusive.

Although both ethylene and GA can mediate sex expression of cucumber, their regulatory functions appear to be opposite. Atsmon and Tabbak (1979) interpreted that the GA-regulated sex differentiation has no effect on ethylene production, and there is a balance in the content of ethylene and GA in controlling the sex expression of cucumber. However, Yin and Quinn (1995) proposed a “one-hormone hypothesis” which posited that ethylene plays a dominant role in cucumber sex determination and GA may regulate the maleness through inhibiting ethylene production. However, our previous studies demonstrated that GA-*CsGAMYB1* signaling could regulate sex differentiation in cucumber through an ethylene-independent process (Zhang et al., 2014b). Therefore, a potential crosstalk between GA and ethylene pathways that determine sex expression in cucumber still remains unclear.

Besides, members of the MADS-box gene family can also regulate the sexual development in cucumber. *AGAMOUS (AG)*, the C-class floral homeotic gene, specifies stamen and carpel identity (Lohmann and Weigel, 2002). There are three *AG* homologs in cucumber, *CAG1*, *CAG2*, and *CAG3*, in which *CAG1* and *CAG3* are expressed in both stamen and carpel, while *CAG2* is particularly restricted to the carpel. However, the expression levels of these three genes do not appear to be mediated by ethylene and GA (Kater et al., 1998; Perl-Trevé et al., 1998).
Moreover, *ERAF17*, an another MADS-box gene, can be induced by ethylene and may be involved in female flowers formation in cucumber (Ando et al., 2001).

In our study, in order to understand the genes and gene networks that may be involved in GA-modulated cucumber sex determination, we performed RNA-Seq analyses to compare the transcriptomes of shoot apices between GA$_3$-treated and control gynoecious cucumber plants. GA$_3$ application induced male flowers but reduced ethylene production in the shoot apices. Notably, GA-regulated sex differentiation was associated with the changes in transcript levels of *M* (CsACS2) gene, ethylene receptor *ETR1*, ethylene-responsive transcription factors, and *CAG2* (a C-class floral homeotic gene), suggesting a potential involvement of both ethylene-dependent and -independent processes in GA-mediated cucumber sexual development. Thus, our results built a foundation for dissecting the molecular mechanism of male flower development in cucumber.

**MATERIALS AND METHODS**

**Plant Materials and Growth Conditions**

A gynoecious cucumber (*C. sativus* L.) line 13-3B was used in this study. The seeds were germinated on wet filter paper in a Petri dish at 28°C in dark overnight. Then the resulting seedlings were grown in a growth chamber under 16 h/8 h with 25°C/18°C in day/night, respectively. Upon two true-leaf stage, plants were transferred to a greenhouse in the experimental field of the Northwest A&F University. Pest control and water management were carried out according to standard practices.

**Exogenous GA$_3$ Treatment**

For male flowers induction in the gynoecious cucumber 13-3B, 1000 ppm GA$_3$ (dissolved in 0.1% ethanol) or deionized water with 0.1% ethanol (Control) were applied by foliar spray for three times at 7 day intervals, starting when the first true leaf was approximately 2.5 cm in diameter. The sex of the flowers on each node of the main stem was recorded until anthesis of flowers on node 25.

In addition, ethylene production in shoot apices was measured after 7 days of the third GA$_3$ treatment. And the RNA-Seq analyses were performed in shoot apices from the cucumber plants firstly treated with GA$_3$ for 6 h, 12 h and the Control, respectively. GA$_3$ was acquired from Sigma-Aldrich Chemical Co. (Shanghai, China).

**Quantification of Ethylene**

The ethylene production was measured by gas chromatography as described previously with some modifications (Zhang et al., 2014b). In brief, the excised shoot apices from cucumber plants treated with exogenous GA$_3$ and the Control were enclosed in 10 mL vessels after weighing and sealed with rubber stoppers. After incubation at 25°C for 16 h, 1 mL of head gas was withdrawn from each vessel using a syringe and injected into a gas chromatograph (GC-9A, Shimadzu, Japan) equipped with a hydrogen FID and an activated alumina column for the measurement of ethylene production. Standard ethylene gas was used for calibrating the instrument. Amount of ethylene was calculated per 1 g fresh weight and per hour.

**RNA Extraction and Quality Test**

The shoot apices from the gynoecious cucumber plants were collected at 6 and 12 h after the first treatment with GA$_3$ and the Control. Samples were immediately frozen in liquid nitrogen and stored at −80°C for RNA-Seq analyses. Total RNA was isolated using the RNA extraction kit (Promega, USA). RNA was checked by RNase-free agarose gel electrophoresis to avoid possible degradation and contamination, and then examined using the NanoPhotometer spectrophotometer (IMPLEN, Westlake Village, CA, USA) for RNA purity. RNA concentration and integrity were measured and assessed using the Qubit RNA Assay Kit in Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA) and RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies, Santa Clara, CA, USA), respectively.

**Digital Gene Expression (DGE) Library Construction and Sequencing**

Digital gene expression libraries were constructed using the NEBNext Ultra Directional RNA Library Prep Kit for Illumina (NEB, Ipsawich, USA) following instructions of manufacturer and six index codes were added to attribute sequences to various samples (Wang et al., 2009). Briefly, poly (A) mRNA was isolated from 3 µg total RNA using oligo-dT magnetic beads (Life Technologies, Carlsbad, CA, USA), and then broken into short fragments by adding fragmentation buffer. First-strand cDNA was synthesized using random hexamer-primed reverse transcription, followed by the synthesis of the second-strand cDNA using RNase H and DNA polymerase I. After adenylation of the 3′ ends of cDNA fragments, NEBNext adapter oligonucleotides were ligated to prepare for hybridization, and then the cDNA fragments were purified using AMPure XP system (Beckman Coulter, Beverly, MA, USA) to select the fragments of preferentially 150–200 bp in length. The size-selected, adaptor-ligated cDNA fragments were enriched using Phusion High-Fidelity DNA polymerase, Universal PCR primers and Index Primer in the PCR reaction. PCR products were purified with AMPure XP system and library quality was assessed using the Agilent Bioanalyzer 2100 system. At last, the cDNA libraries were sequenced on an Illumina HiSeq 4000 platform using the paired-end technology by Novogene Co. (Beijing, China).

**Bioinformatics Analysis of DGE Data**

Raw reads were pre-processed to remove low quality sequences (there were more than 50% bases with quality lower than 20 in one sequence), reads with more than 5% N bases (bases unknown) and reads containing adaptor sequences. Then the clean reads were mapped to the cucumber genome (Chinese long) v2 using TopHat (Huang et al., 2009; Trapnell et al., 2009), allowing up to one mismatch. Unigenes mapped by at

1[^1]

[^1]: http://www.icugi.org/
least one read, in at least one sample, were identified for further analysis.

In this study, samples from three treatments (GA 6 h, GA 12 h, and Control) were prepared for genome-wide expression analyses. Two biological replicates were performed for each treatment, and thus six DGE libraries were sequenced. 44.28–60.29 million raw reads from each library were generated. After removal of low-quality tags and adapter sequences, 42.31–57.63 million high-quality clean reads with a total of 6.35–8.64 G bases were obtained. Among these clean reads, the percentage of Q20 (base quality more than 20) and GC was 97.28–97.52% and 43.46–43.63%, respectively (Table 1). Furthermore, we clustered these clean reads into unique tags, which were mapped to the cucumber genome using TopHat (Huang et al., 2009; Trapnell et al., 2009). About 36.28–49.30 million clean reads (85.21–85.74% of total clean reads) from RNA-Seq data in the six libraries were mapped uniquely to the cucumber genome (Table 1).

The R package edgeR used to identify the DGEs (Robinson et al., 2010). The expression level of each gene was calculated and normalized to FPKM. The FDR was used to determine the threshold of the P-value in multiple tests. In our study, the FDR < 0.05 and fold change > 2 were used as significance cut-offs of the expression differences.

Furthermore, GO enrichment analysis of DGEs was performed using the GOSeq R package (Young et al., 2010). GO terms with corrected P-value < 0.05 were considered significantly enriched by differential expressed genes.

Quantitative Real-Time PCR (qRT-PCR) Validation
Quantitative Real-Time PCR analyses were performed using the independent cucumber shoot apices in the same time point of GA3 application as those used for RNA-Seq. Total RNA was isolated using the RNA extraction kit (Promega, USA), and cDNA was synthesized using the PrimeScript RT reagent (TaKaRa, China). qRT-PCR was carried out using SYBR Premix Ex Taq (TaKaRa, China) on an ABI 7500 Real-Time PCR System (Applied Biosystems, USA). The cucumber α-TUBULIN (TUA) gene was used as an internal control in analyzing gene expression. Three biological replicates were performed for each experiment. The gene specific primers for qRT-PCR are listed in Supplementary Table S5.

RESULTS

Exogenous GA3 Induces Male Flowers Formation and Inhibits Ethylene Production in Gynoecious Cucumber
To verify the effect of GA on sex expression of cucumber, a gynoecious cucumber line 13-3B was treated with GA3 and the sex of flowers on each node was recorded until anthesis of the flowers on node 25 of the main stems. As shown in Figure 1, there were no male flower nodes in the control plants, however, GA3 treatment induced male flowers in the gynoecious cucumber line (Figure 1A), accounting for 51.2% of nodes with male flowers (Figure 1B). Interestingly, formation of male flowers occurred mainly at the lower node positions as compared with the location of female flowers on the main stems (Figure 1A). These observations suggested that exogenous GA can promote male flowers formation in cucumber.

Identification of Differentially Expressed Genes (DEGs) in Shoot Apices from the Gynoecious Cucumber Plants Treated with GA3 and the Control
To identify the genes and gene networks that are involved in GA-regulated sex expression, the genome-wide expression analyses were performed to compare the transcriptome profiles of the shoot apices between the gynoecious cucumber plants treated with GA3 for different time points (6 and 12 h) and the Control through the digital gene expression (DGE) approach (Eveland et al., 2010). Based on deep sequencing, 23,911 unigenes were collected in six libraries. Using fold change > 2 and FDR < 0.05 as the significance cut-offs, we identified 1073 DEGs including 727 up-regulated genes and 346 down-regulated genes after GA3 treatment for 6 h compared with the Control. And we also found that 1590 genes were differentially expressed, in which 765 genes were up-regulated and 825 genes were down-regulated after GA3 treatment for 12 h (Figure 2, Supplementary Tables S1 and S2). Moreover, 594 DEGs containing 303 up-regulated genes (Figure 2A) and 291 down-regulated genes (Figure 2B) were shared in the two sets of transcriptome comparisons.

Verification of RNA-Seq Data by Quantitative Real Time RT-PCR Analyses
To validate the DEGs identified by RNA-Seq, we performed quantitative real time RT-PCR (qRT-PCR) assays using the independent cucumber shoot apices in the same time point after GA3 treatment as those used for RNA-Seq analysis. Twenty DEGs were randomly chosen for qRT-PCR analyses, in which 10 genes including five up-regulated genes and five down-regulated genes were from the set of GA 6 h vs. Control, and the other 10 genes containing five up-regulated genes and five down-regulated genes were from the set of GA 12 h vs. Control. As shown in Figure 3, all the 20 genes showed the similar expression patterns in the qRT-PCR analyses as those in the RNA-Seq data, although the particular values of fold-change were different. The Pearson correlation coefficient between qRT-PCR and RNA-Seq data was 0.975 ($P = 3.5E−13$), indicating that the RNA-Seq results were highly reliable.
The M Gene Is Involved in GA-Regulated Sex Differentiation of Cucumber

Given that ethylene production was significantly decreased in the cucumber plants treated with GA$_3$ (Figure 1C), we screened the ethylene biosynthetic genes in the DGEs. We found that the transcript of $M$ (CsACS2) gene encoding an ACC synthase was inhibited in both sets of transcriptome comparisons (GA 6 h vs. Control and GA 12 h vs. Control) (Table 2), and the qRT-PCR verification displayed the same expression pattern (Figure 3). Since the $M$ gene is believed to inhibit stamen development in flower buds (Saito et al., 2007; Li et al., 2009, 2012), we speculated that GA might release the inhibitory effect of the $M$ gene and allow male flowers to develop in cucumber.

In addition, another two ethylene biosynthetic genes, CsACO1 and CsACO3 which encode two ACC oxidases, were also differently expressed in the shoot apices after GA$_3$ treatment. CsACO3 expression was dramatically down-regulated in the set of GA 12 h vs. Control, but not changed in GA 6 h vs. Control. However, there was an increase in the transcript level of CsACO1 in both GA 6 h vs. Control and GA 12 h vs. Control, but the

---

**TABLE 1 | Summary of the transcriptome assembly.**

<table>
<thead>
<tr>
<th>Samples</th>
<th>Control_rep1</th>
<th>Control_rep2</th>
<th>GA 6 h_rep1</th>
<th>GA 6 h_rep2</th>
<th>GA 12 h_rep1</th>
<th>GA 12 h_rep2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw reads</td>
<td>50,991,998</td>
<td>56,788,784</td>
<td>60,288,014</td>
<td>58,393,488</td>
<td>44,280,996</td>
<td>52,072,848</td>
</tr>
<tr>
<td>Clean reads (%)</td>
<td>(95.52)</td>
<td>(95.57)</td>
<td>(95.59)</td>
<td>(95.42)</td>
<td>(95.55)</td>
<td>(96.01)</td>
</tr>
<tr>
<td>Clean bases</td>
<td>7.31G</td>
<td>8.14G</td>
<td>8.64G</td>
<td>8.36G</td>
<td>6.35G</td>
<td>7.50G</td>
</tr>
<tr>
<td>Q20 (%)</td>
<td>97.52</td>
<td>97.42</td>
<td>97.38</td>
<td>97.43</td>
<td>97.28</td>
<td>97.36</td>
</tr>
<tr>
<td>GC (%)</td>
<td>43.63</td>
<td>43.60</td>
<td>43.51</td>
<td>43.53</td>
<td>43.46</td>
<td>43.54</td>
</tr>
<tr>
<td>Mapped clean reads (%)</td>
<td>42,119,358</td>
<td>46,783,558</td>
<td>49,881,555</td>
<td>48,281,412</td>
<td>36,728,637</td>
<td>43,283,227</td>
</tr>
<tr>
<td>Unique mapped clean reads (%)</td>
<td>(86.48)</td>
<td>(96.20)</td>
<td>(96.55)</td>
<td>(96.65)</td>
<td>(96.80)</td>
<td>(96.58)</td>
</tr>
</tbody>
</table>
Zhang et al. Characterization of GA-Regulated Cucumber Sex Expression

FIGURE 2 | Venn diagrams of DEGs that were significantly upregulated (A) or downregulated (B) after 6 or 12 h of GA3 treatment.

FIGURE 3 | qRT-PCR validation of DEGs identified by RNA-Seq. Twenty DEGs including ten up-regulated genes and ten down-regulated genes from the two sets of transcriptome comparisons (GA 6 h vs. Control and GA 12 h vs. Control) were randomly selected for qRT-PCR confirmation. The blue and red bars represent RNA-Seq and qRT-PCR data, respectively. The cucumber TUA gene was used as an internal control, and these experiments were repeated with three biological samples. Error bars indicate the standard errors.
fold change was lower than that of M gene (Table 2). These results suggested that the decreased transcript levels of M and CsACO3 might inhibit ethylene biosynthesis in the cucumber plants treated with GA3. Nonetheless, an increased expression of CsACO1 following GA3 treatment was insufficient to rescue the effect of M and CsACO3 on ethylene production.

### The Ethylene-Responsive Transcription Factors and Ethylene Receptor CsETR1 Participate in GA-Modulated Cucumber Sex Expression

To further understand the potential functions of DEGs identified by DGE, GO term enrichment analyses (Corrected P-value < 0.05) were carried out in both sets of RNA-Seq data. We found that the DEGs were markedly enriched in biological process and molecular function (MF) groups. For the biological process category, the most significantly enriched GO terms were “cellular carbohydrate biosynthetic process” ($P = 1.1E−02$) and “regulation of cellular macromolecule biosynthetic process” ($P = 5.8E−04$) in GA 6 h vs. Control and GA 12 h vs. Control groups, respectively (Figures 4 and 5). While five GO terms including “oxidoreductase activity, acting on paired donors” ($P = 1.1E−02$), heme binding ($P = 1.1E−02$), tetrapyrrole binding ($P = 1.5E−02$), iron ion binding ($P = 2.5E−02$), and “sequence-specific DNA binding transcription factor activity” ($P = 4.6E−02$) in the set of GA 6 h vs. Control (Figure 4) and two terms containing “sequence-specific DNA binding transcription factor activity” ($P = 2.2E−03$) and “DNA binding” ($P = 1.1E−02$) in the GA 12 h vs. Control group (Figure 5) were detected in the MF category. Furthermore, the transcription factors were dramatically enriched in the DGEs in both sets of data. Accordingly, many ethylene-responsive transcription factors (ERFs) including four in GA 6 h vs. Control (Table 3 and Supplementary Table S3) and nine in GA 12 h vs. Control (Table 4 and Supplementary Table S4) were identified to be down-regulated, consistent with the reduced ethylene production (Figure 1C). Among them, three genes, ERF43 (Csa3M895680) and CRF2s (Csa5M139630 and Csa4M051360), were shared in the two sets. These observations indicated that ERFs may be implicated in cucumber sex expression. Because the ERFs act as positive regulators in ethylene signal transduction pathway (Wang et al., 2002; Guo and Ecker, 2004; Prescott et al., 2016; Zhang et al., 2016), we speculated that the decreased ethylene production inhibited the expression of ERFs, followed by modulated sexual development in cucumber plants treated with GA3.

Interestingly, we also noticed that an ethylene receptor, CsETR1, was enriched in the GO term of “DNA binding” (Figure 5), and its expression was increased by 2.12-fold in shoot apices after GA3 treatment for 12 h (Table 4 and Supplementary Table S4), but not changed in GA 6 h vs. Control.
FIGURE 5 | Gene ontology terms that were significantly enriched in the DEGs after 12 h of GA$_3$ treatment. GO terms belong to BP and MFs were shown in blue and green, respectively. GO terms were sorted based on corrected $P$-value, and the corrected $P$-value < 0.05 was used as the significance cut-off.

TABLE 3 | List of selected ethylene signaling factors in the DEGs with enriched GO terms after GA$_3$ treatment for 6 h.

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Gene annotation</th>
<th>Fold change (GA 6 h/Control)</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Csa3M995860</td>
<td>ERF43 (Ethylene-responsive transcription factor ERF043)</td>
<td>−2.93</td>
<td>1.08E−16</td>
</tr>
<tr>
<td>Csa5M139630</td>
<td>CRF2 (Ethylene-responsive transcription factor CRF2)</td>
<td>−2.51</td>
<td>1.89E−02</td>
</tr>
<tr>
<td>Csa4M051360</td>
<td>CRF2 (Ethylene-responsive transcription factor CRF2)</td>
<td>−2.23</td>
<td>2.66E−02</td>
</tr>
<tr>
<td>Csa5M151530</td>
<td>CRF4 (Ethylene-responsive transcription factor CRF4)</td>
<td>−2.12</td>
<td>1.14E−03</td>
</tr>
</tbody>
</table>

group. While the qRT-PCR verification revealed the same expression pattern (Figure 3). ETR1 is an important member of ethylene receptors family that acts as negative regulator in the ethylene signaling pathway (Wang et al., 2002; Guo and Ecker, 2004; Light et al., 2016; Prescott et al., 2016). Previous study have confirmed that CsETR1 plays a negative role in stamen arrest during development of flower buds in cucumber (Wang et al., 2010). In accordance with these findings, we speculated that up-regulated CsETR1 may promote male flowers formation by alleviating stamen arrest in cucumber after GA$_3$ treatment.

GA May Restrain the Female Tendency via Transcriptional Inhibition on CAG2 in Cucumber

Through GO term enrichment analyses, we further identified an AG (C-class floral homeotic gene) homolog CAG2, which was enriched in the “sequence-specific DNA binding transcription factor activity” group. CAG2 is one of the three AG genes in cucumber that controls pistil development due to its specific expression in the carpel (Kater et al., 1998; Perl-Treves et al., 1998). We found that the transcript level of CAG2 was
TABLE 4 | List of selected ethylene signaling factors and AGAMOUS (AG) homolog in the DEGs with enriched GO terms after GA$_3$ treatment for 12 h.

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Gene annotation</th>
<th>Fold change (GA 12 h/Control)</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Csa2M092800</td>
<td>BBM2 (AP2-like ethylene-responsive transcription factor BBM2)</td>
<td>−∞</td>
<td>1.86E−02</td>
</tr>
<tr>
<td>Csa2M382540</td>
<td>ERF03 (Ethylene-responsive transcription factor ERF003)</td>
<td>−3.44</td>
<td>3.98E−02</td>
</tr>
<tr>
<td>Csa2M382550</td>
<td>ERF03 (Ethylene-responsive transcription factor ERF003)</td>
<td>−2.03</td>
<td>8.55E−05</td>
</tr>
<tr>
<td>Csa3M8956680</td>
<td>ERF43 (Ethylene-responsive transcription factor ERF043)</td>
<td>−3.06</td>
<td>1.93E−13</td>
</tr>
<tr>
<td>Csa4M001970</td>
<td>RAP23 (Ethylene-responsive transcription factor RAP2-3)</td>
<td>−4.08</td>
<td>5.42E−10</td>
</tr>
<tr>
<td>Csa4M051360</td>
<td>CRF2 (Ethylene-responsive transcription factor CRF2)</td>
<td>−2.44</td>
<td>3.40E−02</td>
</tr>
<tr>
<td>Csa4M192030</td>
<td>RA211 (Ethylene-responsive transcription factor RAP2-11)</td>
<td>−3.87</td>
<td>4.03E−02</td>
</tr>
<tr>
<td>Csa4M652640</td>
<td>ERF03 (Ethylene-responsive transcription factor ERF003)</td>
<td>−2.42</td>
<td>5.66E−03</td>
</tr>
<tr>
<td>Csa5M139630</td>
<td>CRF2 (Ethylene-responsive transcription factor CRF2)</td>
<td>−3.02</td>
<td>9.74E−03</td>
</tr>
<tr>
<td>Csa1M467100</td>
<td>CAG2 (Floral homeotic protein AGAMOUS 2)</td>
<td>−23.48</td>
<td>7.84E−03</td>
</tr>
</tbody>
</table>

Our data implied that GA may restrain femaleness via inhibiting the CAG2 expression in cucumber.

DISCUSSION

Sex differentiation of flower buds is an important developmental process that directly affects the product yield in cucumber. In addition to genetic control, sex expression can be modified by plant hormones and environmental conditions (Małepsz and Niemirowicz-Szczytt, 1991). Among various plant hormones, ethylene can induce female flowers formation (MacMurray and Miller, 1968; Iwahori et al., 1969), and the underlying molecular mechanism has been widely documented (Yamasaki et al., 2001, 2003; Mibus and Tatlıoğlu, 2004; Knopf and Trebitsh, 2006; Saito et al., 2007; Li et al., 2009, 2012; Wang et al., 2010). GA can promote male flowers development (Wittwer and Bukovac, 1962; Pike and Peterson, 1969), but the regulatory pathway remains elusive. In addition, a potential crosstalk between GA and ethylene in controlling sex determination of cucumber still remains disputed. In this study, through genome-wide expression analyses, we showed that GA may promote cucumber maleness via an ethylene-dependent pathway by altering expression of the M (CsACS2) gene, ethylene receptor CsETR1 and ethylene-responsive transcription factors. Nevertheless, we also found that GA may also restrain femaleness through an ethylene-independent pathway regulating CAG2, a C-class floral homeotic gene (Figure 6).
GA May Promote Male Tendency via an Ethylene-Dependent Pathway in Cucumber

Upon exogenous GA$_3$ treatment in the gynoecious cucumber line 13-3B, the male flowers were markedly induced, meanwhile, ethylene production in the shoot apices was significantly decreased (Figure 1) due to collaborative regulation by three ethylene biosynthetic genes such as $M$ (CsACS2), CsACO1, and CsACO3. Notably, $M$ and CsACO1 were significantly down-regulated and up-regulated in both sets of transcriptome comparisons, respectively, however, CsACO3 transcript was only decreased in GA 12 h vs. Control group (Figure 3; Table 2). Since both CsACO1 and CsACO3 encode ACC oxidases, and they showed opposite expression patterns and similar fold changes in RNA-Seq data (Table 2), we speculated that interplay between up-regulation of CsACO1 and down-regulation of CsACO3 offset the ACC oxidase activity in shoot apices after 12 h of GA$_3$ treatment. This implied that the reduced $M$ transcript might play major role in inhibiting ethylene biosynthesis. Given that the $M$ gene can directly inhibit stamen development in cucumber flower buds (Saito et al., 2007; Li et al., 2009, 2012), our data revealed that GA might release the inhibitory effect of the $M$ gene on stamen arrest and restrain ethylene production, by down-regulating the $M$ gene expression (Figure 6, right panel).

In ethylene signal transduction pathway, the receptors such as ETRs function as negative regulators, while ERFs, downstream components of receptors, act as positive transcription factors (Wang et al., 2002; Guo and Ecker, 2004; Light et al., 2016; Prescott et al., 2016; Zhang et al., 2016). RNA-Seq data displayed that the ethylene receptor CsETR1 in cucumber, which is thought to be a negative regulatory factor in stamen arrest of flower buds (Wang et al., 2010), was markedly up-regulated after 12 h of GA$_3$ treatment (Figure 3; Table 4 and Supplementary Table S4). Up-regulation of CsETR1 occurred at relatively later time point than down-regulation of the $M$ gene (Table 2), suggesting that the increased CsETR1 transcript was not directly caused by GA$_3$ rather by reduced ethylene production. Moreover, the expression levels of some ERFs were dramatically decreased after GA$_3$ treatment (Tables 3 and 4, Supplementary Tables S3 and S4), and they were probably involved in cucumber sex expression through inhibiting maleness or promoting femaleness. However, this understanding was based on bioinformatics analysis, the precise roles of ERFs on cucumber flower development remained unclear and should be verified in future studies using advanced physiological and molecular techniques. These observations indicated that increased CsETR1 expression may stimulate male tendency through direct inhibition in stamen arrest or down-regulation on the ERFs transcript in cucumber after GA$_3$ treatment (Figure 6, right panel).

Furthermore, despite we have shown that GA can regulate cucumber sex expression in cooperation with ethylene, how GA modulated ethylene and which genes participated in this process were unknown. Given that the DELLA proteins are central repressors of GA responses (Sun, 2010, 2011; Plackett et al., 2014), and accumulating evidence suggested that DELLA proteins play important roles in ethylene-mediated plant growth and development processes through interactions with some regulatory factors in ethylene signaling pathway, such as CTR1 (CONSTITUTIVE TRIPLE RESPONSE1), EIN3/EIL1 (ETHYLENE INSENSITIVE 3/EIN3-LIKE 1), RAP2.3 (RELATED TO APETALA 2.3), and ERF11 (ETHYLENE RESPONSE FACTOR 11) (Achard et al., 2003, 2007; Pierik et al., 2009; An et al., 2012; Luo et al., 2013; Marin-de la Rosa et al., 2014; Zhou et al., 2016). Therefore, we proposed that DELLA proteins may be involved in the sex differentiation of cucumber coupled with GA and ethylene in a collaborative regulation at the protein level, since the expressions of four DELLA homologs in cucumber, CsGAIP, CsGAI1, CsGAI2, and CsGAI3 (Zhang et al., 2014a), had no change after GA$_3$ treatment (Supplementary Tables S1 and S2).

In addition, GA and ethylene can also cooperatively regulate other aspects of plant growth and development. For example, GA promoted apical hook development of Arabidopsis, in part through transcriptional regulation of several genes in ethylene biosynthetic pathway mediated by DELLA proteins (Gallego-Bartolome et al., 2011). In this process, GA induced ethylene production, that was opposite to our results in cucumber sex expression. We speculated that this distinction may be due to different roles of hormones in various plant developmental processes. In fact, GA and ethylene showed similar functions in hook development, but they may play opposite roles in sex determination, thus, the regulatory mechanisms were different.

GA May Inhibit Femaleness or Induce Maleness of Cucumber via an Ethylene-Independent Pathway

AGAMOUS, the C-class floral homeotic gene, belongs to the MADS-box family. In Arabidopsis, GA can induce the expression of AG gene (Yu et al., 2004). However, our data provided a novel point that an AG homolog CAG2 in cucumber was down-regulated upon GA$_3$ treatment (Figure 3; Table 4 and Supplementary Table S4). This distinction may be due to different abilities of these two genes to induce reproductive organ fate, in which, AG in Arabidopsis controls stamen and carpel development (Lohmann and Weigel, 2002), but CAG2 is particularly restricted to the carpel (Kater et al., 1998; Perl-Treves et al., 1998). Previous studies showed that CAG2 transcripts were not mediated by ethylene (Perl-Treves et al., 1998), hence, we speculated that GA probably suppressed pistil development through inhibiting the CAG2 expression, thereby allowing male flowers to develop, and ethylene was not involved in this process. Moreover, in our previous study, we demonstrated that transcript of CsGAMYB1 was upregulated by GA$_3$ treatment in male flower buds and silencing of CsGAMYB1 could suppress masculinization of cucumber, but the ethylene production and expression of $F$ and $M$ genes were not changed in the CsGAMYB1-RNAi lines (Zhang et al., 2014b). These observations suggested that GA can also regulate sex expression of cucumber via an ethylene-independent pathway (Figure 6, left panel). However, the relationship between CAG2 and CsGAMYB1 regulations remains obscure and that should be verified in further studies.
Besides, DELLAs can down-regulate the expression of floral homeotic genes, AP3, PI, and AG, subsequently inhibit flower development in Arabidopsis (Yu et al., 2004). CsGAIP, a DELLA homolog in cucumber, may restrain staminate development through transcriptional repression of AP3 and PI in Arabidopsis (Zhang et al., 2014a). However, the expression levels of AP3 and PI were not changed after GA3 treatment (Supplementary Tables S1 and S2), indicating that they did not participate in GA-regulated male tendency. Accordingly, there may be a regulatory relationship between DELLAs and floral homeotic gene CAG2 during GA-modulated sexual development in cucumber, however, the mechanism would be possibly different from that of Arabidopsis.

In summary, our data revealed a novel viewpoint that GA might control sex differentiation of cucumber via both ethylene-dependent and ethylene-independent pathways, and DELLAs were likely to be involved in both processes. However, this model was proposed by bioinformatics data, therefore, elucidation of the critical roles of DELLAs in flower development by cucumber transformation, and identification of the relationships among DELLAs and ethylene regulatory factors, GA-DELLA-GASYB1 signaling and CAG2 gene, will shed new light on the molecular details of GA-regulated sex expression in cucumber.

Evolution of Unisexual Flower in Cucumber and Potential Involvement of Hormones

Generally, typical unisexual flowers have two morphological types. The type I is unisexual by abortion. Initiation of stamen and pistil occurs in all flowers, followed by the developmental arrest in one or another organ. The type II is unisexual from inception. Only stamen or pistil is initiated and it does not go through a hermaphroditic stage (Lebel-Hardenack and Grant, 1997; Ainsworth, 2000; Mitchell and Diggle, 2005). Now, it is believed that the morphology of cucumber flowers belongs to the type I (Bai et al., 2004), but its evolutionary mechanism is largely unknown. Bai and Xu proposed a “miR initiative” hypothesis (Bai and Xu, 2013), where they speculated that unisexual cucumber flowers are evolved from a hermaphroditic ancestor. The first step in the evolutionary process might be the miRNA-regulated arrest of ovary development, and this predication is based on the altered expression of miRNAs, such as miR396a, 156b, 159a, 171b, and 166a, in male flowers (Bai et al., 2004; Sun et al., 2010). And this event leads to environment-dependent andromonoecy which has no progeny. Then, the M gene is recruited. On the one hand, the M gene promotes ethylene biosynthesis, resulting in the rescue of ovary development for seed set, because the ethylene might regulate the miRNA production. On the other hand, the M gene inhibits stamen development to avoid self-pollination and maintain cross-pollination. So, the monoeocious genotype is generated through the cooption of the M gene. The andromonoecious genotype is produced by the loss-of-function m gene, which is regarded as a reverted point mutation. Further, the F gene is coopted and generate the gynoeocious genotype (Sun et al., 2010).

Until now, a potential role of GA in unisexual flower evolution of cucumber has not been reported. But based on the possible function of M gene on evolutionary development of cucumber flower and the effect of GA on the transcript of M gene (Table 2), we speculated that GA might be involved in the process of cucumber flower evolution through interaction with ethylene. In addition, previous studies showed that GA signaling system can regulate anther development by modulation of miR159/GAMYB (Achard et al., 2004). In this pathway, miR159 acts as a post-transcriptional regulator of GAMYB transcript levels. GA relieves the DELLAs repression of GAMYB, which is mediated by the GA activation of miR159. As mentioned above, miR159 is likely to participate in the arrest of ovary development in the evolutionary process of cucumber flower. And the GAMYB homolog CsGAMYB1 can regulate cucumber sex expression via an ethylene-independent pathway (Zhang et al., 2014b). These observations further revealed the possible involvement of GA in unisexual flower evolution of cucumber, but this process might be dependent on miR159 and GAMYB, and have no relationship with ethylene. Finally, it is worth noting that these viewpoints are built on the basis of the “miR initiative” hypothesis and needed to be tested in further work.

AUTHOR CONTRIBUTIONS

YZ and YaL designed the experiments. YZ, GZ, and NM performed the experiments. YZ, YuL, and JZ analyzed the data. YZ and YaL wrote the paper along with YaL. All authors reviewed the manuscript.

FUNDING

This work was supported by National Natural Science Foundation of China (31601770), Science and Technology Research and Development Program of Shaanxi Province (2015NY098), Fundamental Research Funds for Northwest A&F University (2452015024), and Doctoral Scientific Research Foundation of Northwest A&F University (Z109021504) to YZ.

ACKNOWLEDGMENTS

We thank Dr. Huazhong Ren (China Agricultural University) for providing the cucumber 13-3B seeds, and members of the Liang Laboratory for helpful discussions and technical assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2017.00010/full#supplementary-material
REFERENCES


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

Copyright © 2017 Zhang, Zhao, Li, Ma, Zhang and Liang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.
Genome-Wide Analysis of Soybean JmjC Domain-Containing Proteins Suggests Evolutionary Conservation Following Whole-Genome Duplication

Yapeng Han¹,²†, Xiangyong Li¹,²†, Lin Cheng¹,², Yanchun Liu¹,², Hui Wang¹,², Danxia Ke¹,², Hongyu Yuan¹,², Liangsheng Zhang³,⁴* and Lei Wang¹,²*

¹ College of Life Sciences, Xinyang Normal University, Xinyang, China, ² Institute for Conservation and Utilization of Agro-bioresources in Dabie Mountains, Xinyang Normal University, Xinyang, China, ³ Center for Genomics and Biotechnology, Haixia Institute of Science and Technology, Fujian Agriculture and Forestry University, Fuzhou, China, ⁴ Key Laboratory of Genetics, Breeding and Multiple Utilization of Corps (Fujian Agriculture and Forestry University), Ministry of Education, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, Fujian Agriculture and Forestry University, Fuzhou, China

Histone modifications, such as methylation and demethylation, play an important role in regulating chromatin structure and gene expression. The JmjC domain-containing proteins, an important family of histone lysine demethylases (KDMs), play a key role in maintaining homeostasis of histone methylation in vivo. In this study, we performed a comprehensive analysis of the jumonji C (JmjC) gene family in the soybean genome and identified 48 JmjC genes (GmJMJs) distributed unevenly across 18 chromosomes. Phylogenetic analysis showed that these JmjC domain-containing genes can be divided into eight groups. GmJMJs within the same phylogenetic group share similar exon/intron organization and domain composition. In addition, 16 duplicated gene pairs were formed by a Glycine-specific whole-genome duplication (WGD) event approximately 13 million years ago (Mya). By investigating the expression profiles of these gene pairs in various tissues, we showed that the expression pattern is conserved in the polyploidy-derived JmjC duplicates, demonstrating that the majority of GmJMJs were preferentially retained after the most recent WGD event and suggesting important roles for demethylase duplications in soybean evolution. These results shed light on the evolutionary history of this family in soybean and provide insights into the JmjCs which will be helpful to reveal their functions in controlling soybean development.

Keywords: soybean (Glycine max L.), JmjC gene family, genome-wide analysis, phylogeny, gene structure, expression pattern

INTRODUCTION

Histone methylation and demethylation have important roles in regulating transcription, genome integrity, and epigenetic inheritance (Klose et al., 2006; Klose and Zhang, 2007; Liu et al., 2010). Histone methylation can occur at various lysine and arginine residues, including K4, K9, K27, K36, and K79 in histone H3 and K20 in histone H4 (Allis et al., 2007). Histone methylation, which is mainly catalyzed by protein families that contain PRMT and SET domains, can have both activating and repressive effects on chromatin function (Ahmad and Cao, 2012; Zhang and Ma, 2012). Two kinds of demethylase are involved in the homeostasis of methylation in organisms. Lysine Specific Demethylase 1 (LSD1) was the first histone demethylase identified and is a member...
of the flavin-dependent amine oxidase family (Lee et al., 2005; Metzger et al., 2005; Chen et al., 2011). Genes in the second class of histone demethylases have a JmjC domain with which they catalyze histone lysine demethylation through oxidative reactions dependent on ferrous ion (Fe(II)) and α-ketoglutarate (α-KG) (Elkins et al., 2003; Trewick et al., 2005).

Plant JmjC proteins are known to play important roles in regulating epigenetic processes and in growth and development (Klose et al., 2006; Kouzarides, 2007). Many members of the JmjC gene family from different plant species have been characterized. In Arabidopsis, AtMJ11/ELF6 (EARLY FLOWERING 6) is a repressor in the photoperiodic flowering pathway, and its loss-of-function mutation causes early flowering (Noh et al., 2004; Yu et al., 2008). Its relative, AtJMJ14/REF6 (RELATED OF EARLY FLOWERING 6) has an opposite effect in the regulation of flowering time (Noh et al., 2004; Yu et al., 2008; Lu et al., 2011a).

Loss-of-function mutation of REF6 leads to increased expression of the flowering repressor FLC (FLOWERING LOCUS C) and hence late flowering (Lu et al., 2011a). In addition, AtMJ14, an active histone H3K4 demethylase (Lu et al., 2010; Yang et al., 2010), was also implicated in preventing early flowering by repressing the expression of FLOWERING LOCUS T (FT) and its homologs. Recently, Ning reported that AtMJ14 associates with two NAC transcription factors, NAC050 and NAC052, and co-occupies hundreds of common target genes, resulting in H3K4 demethylation and transcriptional repression (Ning et al., 2015). Apart from controlling flowering time, there is also evidence that AtMJ14 functions in RNA silencing and cell-to-cell movement of an RNA silencing signal (Lu et al., 2010). The histone H3K9 demethylase AtMJ25/IBM1 (INCREASE IN BONSAI METHYLATION 1) (Wang et al., 2013; Shen et al., 2014a) protects genes from CHG (H represents A, T, or G) hypermethylation by CMT3 (CHROMOMETHYLASE 3). Gain-of-function mutants of AtMJ15 showed enhanced salt tolerance, in contrast with increased salt sensitivity in the loss-of-function mutant (Shen et al., 2014a). In addition, AtMJ14 and AtMJ15 have also been shown to be involved in the control of flowering time (Yang et al., 2012). AtMJ30/JMJD5, an evening-expressed gene, is the sole AtMJ protein to show a robust circadian rhythm of expression (Mockler et al., 2007; Michael et al., 2008; Jones and Harmer, 2011; Lu et al., 2011b). The role of AtMJ30 as a genetic regulator of period length in the Arabidopsis circadian clock was confirmed by analysis of loss- and gain-of-function mutants (Lu et al., 2011b). In tomato, a similar role is played by JMJD24, which standardizes the circadian clock and also alters GA response to regulate stem elongation (Li et al., 2015). In Medicago truncatula, MtJMJC5 (Medtr4g066020), an ortholog of AtMJ30/JMJD5, may play a role in epigenetic regulation of the link between the circadian clock and cold signaling (Shen et al., 2016). In rice, OsMJ705 is a bistress-sensitive H3K27me2/3 demethylase that may remove H3K27me3 from marked defense-related genes and increase their basal and induced expression during pathogen infection (Li et al., 2013). OsMJ706, encodes a heterochromatin-associated H3K9 demethylase, is reported to involve in the regulation of flower development in rice (Sun and Zhou, 2008).

Soybean [Glycine max (L.) Merr.] is one of the most economically important crop species in the world. Its genome has undergone two rounds of whole-genome duplication (WGD; Schlueuter et al., 2004; Schmutz et al., 2010; Vanneste et al., 2014; Liu et al., 2015); one occurred approximately 59 Mya and was shared by other legumes such as Medicago and Lotus, while the other was specific to Glycine and occurred around 13 Mya. Thus, about 75% of the genes in the soybean genome have multiple paralogs (Schmutz et al., 2010; Severin et al., 2011; Singh and Jain, 2015), making it an excellent model for studying the evolution of duplicate genes following polyploidy. Here, we systematically identify the JmjC gene family members in soybean (G. max), Medicago (M. truncatula), and Lotus (L. japonicus) and subsequently analyze the evolutionary relationships between these genes among the three legumes, Oryza sativa, and Arabidopsis. In addition, we study the GmJMs in further detail, including subfamily classification, gene structures, chromosomal distribution, duplication patterns, conserved residues, and expression profiling. We propose that demethylases exhibit conservative functions through duplication events. Our data will facilitate future studies to elucidate the exact biological functions of the GmJMs.

MATERIALS AND METHODS

Identification of JmjC Domain-Containing Proteins in Soybean and Other Legumes

The G. max 2.0 genome database (https://phytozome.jgi.doe.gov/pz/#) was searched to identify JmjC domain-containing proteins using Basic Local Alignment Search Tool algorithms (BLASTP) with a threshold of e-value < 1e-10, using the published Arabidopsis (21) and O. sativa (20) JmjC domain-containing protein sequences (Table S1) as queries (Huang et al., 2016). All obtained protein sequences were examined for the presence of the JmjC domain (PF03207, SM00558) domain using the Hidden Markov Model (HMM) of Pfam (Finn et al., 2016) (http://pfam.sanger.ac.uk/search), and SMART (Letunic et al., 2015) (http://smart.embl-heidelberg.de/). Sequences with obvious errors and/or JmjC domain length of <90 amino acids were removed manually. Following the same approach, putative M. truncatula and L. japonicus JmjC domain-containing proteins were identified from Phytozome v10 (https://phytozome.jgi.doe.gov/pz/portal.html) and the L. japonicus genome assembly build 3.0 (http://www.kazusa.or.jp/lotus/), respectively.

Phylogenetic Analysis

Multi-species phylogenetic tree was constructed using MEGA 6.0 (Tamura et al., 2013) with the Neighbor-Joining (NJ) method, and bootstrap analysis was conducted using 1000 replicates with the p-distance model. The JmjC domain alone was used to set up the phylogenetic tree and define the groups (Figure S1). And then with the aim to obtain a better phylogeny within each group, we added additional conserved domains to JmjC domain (Table S2) to construct the phylogenetic tree (Figures 1A, 4A). Multiple sequences alignments were performed using ClustalW with default parameters in MEGA 6.0.
Chromosomal Locations and Gene Structure of JmjC Genes

The locations of the JmjC domain-containing genes on the soybean chromosomes were plotted using the MapChart software. The location information of each JmjC domain-containing gene on each chromosome was determined from the soybean genome annotation file (Gmax_275_Wm82.a2.v1.gene.gff3). The blocks regarded as recent duplications were obtained from SoyBase (Grant et al., 2010) (http://www.soybase.org/). The presence of introns and exons was also annotated according to the soybean genome annotation file. Schematic diagrams were
picted by using GSDFS2.0 (Gene Structure Display Server http://gsds.cbi.pku.edu.cn/).

Conserved Domains and Conserved Residues in the JmjC Domain-Containing Proteins
To explore the full-length sequences of JmjC domain-containing proteins, NCBI CDD (Marchler-Bauer et al., 2015) (http://www.ncbi.nlm.nih.gov/cdd/), SMART (http://smart.embl-heidelberg.de/), and Pfam (http://pfam.xfam.org/) were performed with default parameters to search for conserved domains. To identify conserved amino acid residues for interaction with co-factors, the sequences of JmjC domain were aligned using the DNAMAN software.

Expression Analysis of Soybean JmjC Genes
To determine the expression patterns of the JmjC genes in soybean tissues, transcriptome data was downloaded from the NCBI Short Read Archive database under the following accession numbers: SRX474427, SRX474441, SRX474445, SRX474430, SRX474431, SRX474433, SRX474432, SRX474439, SRX474442, SRX474419, SRX474428, SRX474440, SRX474443, SRX474424, SRX474423, SRX474422, SRX474434, SRX474436, SRX474437, SRX474416, SRX474435, SRX474438, SRX474421, SRX474420, SRX474446, SRX474444, SRX474426, and SRX474429 (Shen et al., 2014b). Transcriptome analysis was performed to identify expression patterns in representative tissues, including roots, cotyledons, stems, shoot meristems, leaf buds, leaves, flowers, pods, pod and seeds, and seeds (Table S3). Finally, heatmaps of GmJMJ expression were produced using the pheatmap packages in R.

Calculation of Ka/Ks-Values and Evaluation Divergence Time
To investigate whether positive Darwinian selection was involved in GmJMJ divergence following duplication and to estimate the date of the duplication pairs, the non-synonymous (Ka) and synonymous substitution (Ks) rate ratios of the paralog pairs were calculated using the YN00 method of the PAML program (Yang, 2007). Based on a rate of 6.1 × 10^{-9} substitutions per site per year, we calculated the divergence time (T) as $T = Ks/(2 \times 6.1 \times 10^{-9}) \times 10^{-6}$ Mya (Lynch and Conery, 2000).

RESULTS
Identification of JmjC Gene Family in Soybean
Using the combined methods, we identified a total of 48 GmJMJ, which is more than twice the number found in Arabidopsis (21) or rice (20) (Lu et al., 2008). To better understand the expansion and evolutionary history of GmJMJ, the same methods were used to search for JmjC genes in two other legumes, Medicago and Lotus. We identified 33 Medicago and 27 Lotus JmjC genes, which is still less than the number found in soybean.

A variety of information about GmJMJ, such as different version of gene codes, gene length, isoelectric point (pI), and molecular weight (Mw) and so on, were listed in Table S4. For example, the identified GmJMJ encode proteins ranging from 284 (GmJMJ2) to 1831 (GmJMJ35) amino acids, with the isoelectric point (pI) varying from 4.91 (GmJMJ23) to 9.25 (GmJMJ37) and the molecular weight (Mw) varying from 32.2 kD (GmJMJ2) to 209.2 kD (GmJMJ35). GmJMJ1 was excluded from further analyses, as there is no annotation data available for it.

Phylogenetic Analysis of JmjC Genes in Soybean
According to the phylogenetic analysis (Figure 1A), the JmjC genes can be divided into eight groups: PKDM9, PKDM8, KDM5, JMJ6, PKDM11, PKDM13, PKDM12, and KDM3. KDM3 has the most members, with 57 homologous JmjC genes, and KDM5 is the second largest group, containing 26 JmjC genes. The smallest clades are PKDM11 and PKDM13, which both consist of only five JmjC genes, one from each species. In general, most of the clades include genes from all five species, although the clades are also enriched in particular species. For example, doubled GmJMJ pairs are sister genes to AtJMJ28 in a clade which also includes doubled genes from the other legumes, forming a cluster of several legume JMJ genes with a single AtJMJ gene. Likewise, there is a larger of percentage of soybean (40%) than Arabidopsis (7%) genes in PKDM8 (Figure 1B). These findings indicate that different levels of gene duplication or lose may have been occurred among the five species after the divergence of eudicot and monocot.

The phylogenetic relationships of AtJMJ30, OsJMJ17, and MtJMJ5 (Mtr4g066020) in PKDM12 are consistent with a eudicot and monocot.

Chromosome Location and Duplication of GmJMJ Genes
Compared to other species, soybean has an extensively expanded GmJMJ family with more than twice as many JmjC as rice and Arabidopsis (Table 1). We carried out a comprehensive analysis of the GmJMJ with the aim of understanding their duplication status and identifying duplicated gene pairs. First, GmJMJ pairs located in a pair of paralogous blocks formed by Glycine-specific WGD were considered as candidate duplicate gene pairs. As shown in Figure 2, all 47 GmJMJ (except GmJMJ1) were randomly located on 18 of the 20 soybean chromosomes. For example, chromosome 10 possesses six GmJMJ, chromosomes 4, 6, 7, 8, and 15 each contain three GmJMJ, and chromosome 2, 3, 5, 12, 13, 14, and 17 each have only one GmJMJ. In total, we found that a large proportion (41 of 47) of the GmJMJ (linked by purple
TABLE 1 | JmjC gene distribution among species that were used in this study.

<table>
<thead>
<tr>
<th>Species name</th>
<th>Abbr.</th>
<th>PKDM9</th>
<th>PKDM8</th>
<th>KDM5</th>
<th>JMJD6</th>
<th>PKDM11</th>
<th>PKDM13</th>
<th>PKDM12</th>
<th>KDM3</th>
<th>All</th>
<th>Genome size (mb)</th>
<th>Chromosomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycine max</td>
<td>Gma</td>
<td>6</td>
<td>6</td>
<td>9</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>18</td>
<td>48</td>
<td>1100</td>
<td>40</td>
</tr>
<tr>
<td>Lotus japonicus</td>
<td>Lj</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>12</td>
<td>27</td>
<td>472</td>
<td>12</td>
</tr>
<tr>
<td>Medicago truncatula</td>
<td>Mtr</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>16</td>
<td>33</td>
<td>500</td>
<td>16</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>At</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>21</td>
<td>125</td>
<td>10</td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>Os</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>20</td>
<td>466</td>
<td>24</td>
</tr>
</tbody>
</table>

FIGURE 2 | Chromosomal locations, region duplication, and predicted cluster for soybean JmjC genes. The chromosome number is indicated above each chromosome. Throat on the chromosomes indicate the position of centromere. Only those chromosomes bearing JmjC genes (18) were represented. Segmental duplicated genes are indicated by purple lines. Colored boxes indicate the gene blocks generated by the most recently WGD event. The figure was produced and modified using the MapChart and Adobe Illustrator, respectively.

lines in Figure 2) was distributed preferentially in duplicated blocks. These 41 genes be considered as candidates of the most recent Glycine-specific WGD and used in the next analysis. Second, close phylogenetic relationships have been shown among all candidate GmJMJs (Figure S2). Third, duplication types of all GmJMJs have been obtained by MCscanX programs (Wang et al., 2012) (Table S5). Fourth, colinearity analysis (Figure 3) was carried out among each candidate duplicated gene pairs. The candidate pairs were considered as created by the recent Glycine-specific WGD duplications with at least three paralogous gene pairs along the flanking regions. In conclusion, we identified that 16 GmJMJs pairs were formed by the most recent Glycine-specific WGD (Table 2).

Based on the divergence rate of $6.1 \times 10^{-9}$ synonymous mutations per synonymous site per year which has been proposed for soybean (Lynch and Conery, 2000), among the 48 JmjCs in soybean, 73% (35 of 48) represented WGD/segmental duplication genes. Ks-value was calculated for estimating the separation time of each paralogous gene pair. All Ks-values ranged from 0.068 to 0.18, which was consist with whole genome duplication events at round 13 Mya. In addition, our divergence time analyses showed that duplications among 16 paralogous pairs occurred between 5.6 and 15.5 Mya, with an average of 9.7 Mya (Table 2).

The history of selection acting on coding sequences can also be measured based on the ratio of non-synonymous to synonymous substitutions (Ka/Ks) (Li et al., 1981). Ka and Ks can be estimated using a number of substitution models and methods, and the estimates are sensitive to these choices and other complications such as the GC content of the sequences and their genomic...
TABLE 2 | Divergence between JmjC gene pairs in soybean.

<table>
<thead>
<tr>
<th>Gene1</th>
<th>Gene2</th>
<th>Ka</th>
<th>Ks</th>
<th>Ka/Ks</th>
<th>Duplicated date (Mya)</th>
<th>Duplicated type</th>
</tr>
</thead>
<tbody>
<tr>
<td>GmJMJ4</td>
<td>GmJMJ5</td>
<td>0.101</td>
<td>0.186</td>
<td>0.541</td>
<td>15.3</td>
<td>WGD or segmental</td>
</tr>
<tr>
<td>GmJMJ7</td>
<td>GmJMJ8</td>
<td>0.072</td>
<td>0.189</td>
<td>0.381</td>
<td>15.5</td>
<td>WGD or segmental</td>
</tr>
<tr>
<td>GmJMJ9</td>
<td>GmJMJ10</td>
<td>0.046</td>
<td>0.076</td>
<td>0.069</td>
<td>6.2</td>
<td>WGD or segmental</td>
</tr>
<tr>
<td>GmJMJ13</td>
<td>GmJMJ14</td>
<td>0.034</td>
<td>0.050</td>
<td>0.380</td>
<td>7.4</td>
<td>WGD or segmental</td>
</tr>
<tr>
<td>GmJMJ15</td>
<td>GmJMJ16</td>
<td>0.026</td>
<td>0.068</td>
<td>0.381</td>
<td>5.6</td>
<td>WGD or segmental</td>
</tr>
<tr>
<td>GmJMJ17</td>
<td>GmJMJ18</td>
<td>0.018</td>
<td>0.169</td>
<td>0.104</td>
<td>13.8</td>
<td>WGD or segmental</td>
</tr>
<tr>
<td>GmJMJ19</td>
<td>GmJMJ20</td>
<td>0.028</td>
<td>0.147</td>
<td>0.191</td>
<td>12.1</td>
<td>WGD or segmental</td>
</tr>
<tr>
<td>GmJMJ21</td>
<td>GmJMJ22</td>
<td>0.072</td>
<td>0.126</td>
<td>0.574</td>
<td>10.3</td>
<td>WGD or segmental</td>
</tr>
<tr>
<td>GmJMJ26</td>
<td>GmJMJ27</td>
<td>0.043</td>
<td>0.171</td>
<td>0.250</td>
<td>14.0</td>
<td>WGD or segmental</td>
</tr>
<tr>
<td>GmJMJ28</td>
<td>GmJMJ29</td>
<td>0.036</td>
<td>0.103</td>
<td>0.346</td>
<td>8.4</td>
<td>WGD or segmental</td>
</tr>
<tr>
<td>GmJMJ30</td>
<td>GmJMJ31</td>
<td>0.035</td>
<td>0.077</td>
<td>0.458</td>
<td>6.3</td>
<td>WGD or segmental</td>
</tr>
<tr>
<td>GmJMJ32</td>
<td>GmJMJ33</td>
<td>0.026</td>
<td>0.074</td>
<td>0.359</td>
<td>6.0</td>
<td>WGD or segmental</td>
</tr>
<tr>
<td>GmJMJ34</td>
<td>GmJMJ35</td>
<td>0.024</td>
<td>0.101</td>
<td>0.241</td>
<td>8.3</td>
<td>WGD or segmental</td>
</tr>
<tr>
<td>GmJMJ41</td>
<td>GmJMJ42</td>
<td>0.026</td>
<td>0.104</td>
<td>0.250</td>
<td>8.5</td>
<td>WGD or segmental</td>
</tr>
<tr>
<td>GmJMJ43</td>
<td>GmJMJ44</td>
<td>0.030</td>
<td>0.092</td>
<td>0.333</td>
<td>7.5</td>
<td>WGD or segmental</td>
</tr>
<tr>
<td>GmJMJ47</td>
<td>GmJMJ48</td>
<td>0.045</td>
<td>0.130</td>
<td>0.349</td>
<td>10.7</td>
<td>WGD or segmental</td>
</tr>
</tbody>
</table>

FIGURE 3 | Collinearity analysis of duplicated gene pairs formed by Glycine-specific WGD event. Green and purple arrows represented duplicated JmjC gene pairs. The direction of the arrow indicates the position of the gene in the positive (to left) and negative (to right) chain of DNA.

context (Bustamante et al., 2002). Ka is usually much smaller than Ks, so a pair of sequences will have Ka/Ks << 1 if both sequences have been under purifying selection, Ka/Ks < 1 if one sequence has been under purifying selection but the other drifting naturally, and in rare, Ka/Ks > 1 when both sites are under positive selection (Juretic et al., 2005). As shown in Table 2,
the average Ka/Ks-values of the GmJMj gene pairs were 0.36. Five paralog pairs have small Ka/Ks ratios (< 0.3), most Ka/Ks ratios in the range from 0.3 to 0.7, and none of them > 1.

Exon–Intron Structure and Domain Architecture of GmJMj Genes

Structural divergence has been very prevalent in duplicate genes and, in many cases, has led to the generation of functionally distinct paralogs (Lynch and Conery, 2000). To better understand the structural diversity of the GmJMjs following duplication events, the exon/intron structures (Figure 4B) were compared using Gene Structure Display Server 2.0 (http://gsds.cbi.pku.edu.cn/). Our analysis clearly revealed that most of the paralogs share a similar gene structure. For example, 12 gene pairs (GmJMj9/-10, GmJMj13/-14, GmJMj15/-16, GmJMj19/-20, GmJMj21/-22, GmJMj28/-29, GmJMj30/-31, GmJMj32/-33, GmJMj34/-35, GmJMj41/-42, GmJMj43/-44, and GmJMj47/-48) were found to have highly consistent gene structures, including the numbers of exons/introns and the length of exons. However, there were some differences in intron lengths and in the 5' UTR region, which is related to the regulation of expression. For example, GmJMj9/-10 and GmJMj30/-31 both present a large divergence in the length of their 5' UTR, implying that a subtle distinction in function in the development and growth of soybean may have appeared between the two paralogs. In addition, four gene pairs (GmJMj4/-5, GmJMj7/-8, GmJMj17/-18, and GmJMj26/-27) had greater changes in their structural organization, especially in the numbers of exons.

We also studied the proteins encoded by the GmJMjs, using the full-length protein sequences of JmJC proteins as queries in CDD (Marchler-Bauer et al., 2015), SMART (Letunic et al., 2015), and Pfam (Finn et al., 2016) with default parameters in order to gain more insights into the diversity of the domain architecture, as shown in Figure 4C. These proteins all share a JmJC domain. The JmJN domain was the second most widespread domain, appearing in the majority of members of three groups, KDM5, PKDM8, and PKDM9. This domain, which is not adjacent with JmJC, was identified in the jumonji family (Balicunas and Ronne, 2000), and its interaction with the JmJC catalytic domain was found to be important for Jhd27 (also known as KDM5), a H3K4-specific demethylase in budding yeast (Huang et al., 2010; Quan et al., 2011). In PKDM9, the zf-C2H2 domain, which contains two cysteines and two histidines that coordinate a zinc atom to create a compact nucleic acid-binding domain (Chrispeels et al., 2000), was found in four tandem repeats. Furthermore, another zinc-finger domain, zf-C5HC2, was identified in PKDM8 and KDM5. Three groups, PKDM11, PKDM12 and PKDM13, all have only one domain (JmJC) in their full-length sequence, and can be grouped together as "JmJC domain-only proteins," in keeping with previous studies (Klose et al., 2006; Lu et al., 2008; Huang et al., 2016). Two-thirds of the members of KDM5 have FYRN and FYRC domains, which may harbor chromatin-binding activity (Lu et al., 2008) or contribute to JmJC function by interacting with other proteins. For example, it has been reported that the functional specificity of AtJMj14 in flowering time control is based on the specificity of its interaction with transcription factors through the FYRC domain (Ning et al., 2015). Two GmJMj proteins, GmJMj34 and GmJMj35, have the ARID domain (AT-rich interaction domain), which has been implicated in sequence-specific DNA binding (Gregory et al., 1996).

Strangely, we found that not only the gene pairs which share a similar gene structure but also the four pairs which had greater differences in gene structural organization all had consistent domain architectures. For example, GmJMj26/-27 share a relatively consistent functional domain in the full-length protein sequences. This implies that although the gene structure of JmJC family may change through evolution, their protein structures and functions were conserved.

Conserved Amino Acid Residues in Active Sites of GmJMj

Fe(II) iron and α-KG are needed as cofactors by JmJC-domain proteins to carry out their demethylase activity (Chen et al., 2006; Huang et al., 2016). A total of five amino acid residues are needed to bind these cofactors; three residues (His188, Glu/Asp190, and His276) bind to the Fe (II) cofactor and two other residues (Thr/Phe185 and Lys206) bind to α-KG. With the aim of clarifying whether the conserved residues interacting with cofactors had diverged among GmJMjs, we aligned the domain sequences of JmJC proteins from soybean and Arabidopsis.

Based on the alignments, we grouped these proteins into two groups according to amino acids at the conserved sites. The first group, which includes PKDM8, PKDM9, and KDM5, has the conserved amino acids His (H), Glu (E), and His (H) for Fe(II) binding, and Phe (F) and Lys (K) for α-KG binding (Figure 5A), while the second group, which includes JMD6, PKDM13, PKDM11, PKDM12, and KDM3, has conserved the residues His (H), Asp (D), and His (H) for Fe(II) and Thr (T) and Lys (K) for α-KG (Figure 5B). Both forms are compatible with histone demethylation activity (Lu et al., 2008). In general, most members have conserved residues for interacting with cofactors, although there are some exceptions. For example, a substitution can be seen in the first sites in PKDM12, where Thr (T) was changed into Ser (S) in GmJMj21/22 and AtJMj31. However, Thr and Ser have similar physical and chemical properties, so the ability to bind to cofactors may not have changed despite the presence of a different amino acid. Furthermore, the first site to interact with α-KG is absent in JMD6, which is consistent with findings in rice (Lu et al., 2008). The detection of this change in all three plants, soybean, rice, and Arabidopsis, suggests that it may have occurred in the ancestor of these plants and is necessary for their common function. Overall, the high conservation in the interaction sites implies a significant role for these sites in the demethylase activity of the JmJC gene family.

Expression Profiles of JmJC Genes in Soybean

To investigate the tissue-specific expression profiles of GmJMJs, transcriptome data (Shen et al., 2014b) were studied in 10 tissues at different developmental stages including roots, cotyledons, stems, shoot meristems, leaf buds, leaves, flowers, pods, pod
FIGURE 4 | Phylogenetic analysis, gene structure, and domain architecture of GmJMJs. (A) Phylogenetic tree construction of GmJMJs based on the JmjC domain amino acid sequences. Name of genes marked in same color are a pair of paralogs. (B) Exon/intron structures of GmJMJs genes. The black line refers introns, the green box represents exons, and the orange box refers UTR. Over-longed introns were represented with slash–slash. The sizes of exons and introns can be estimated using the scale at the bottom. (C) The domain architecture of the full-length JmjC-domain containing proteins. JmjC, Jumonji C domain; JmjN, Jumonji N domain; PHD, plant homeobox domain; ARID, AT-rich interaction domain; zf-C2H2, Zinc finger of C2H2-type; zf-C5HC2, Zinc finger of C5HC2-type; FYRN, “FY-rich” domain N-terminal; FYRC, “FY-rich” domain C-terminal; WRC, Trp, Arg, and Cys domain; RING, [Really interesting new gene] finger domain.
and seeds, and seeds. As indicated in Figure 6, the expression patterns of the GmJMJs can be divided into three clusters, C1–C3. C1 genes show hardly any expression in almost all of the tissues except certain expression in flower, implying that these genes may possess certain specific function in flower after WGD events. C2 can be divided into two subgroups, C2-Sub1 and C2-Sub2, according to their expression levels, with lower expression levels in C2-Sub1 than C2-Sub2. Genes in C2 show a low expression level in all tissues investigated. We found that genes, which belong to the same clade in the phylogenetic tree (Figure 1A), were sometimes dispersed in different clusters based on expression (Figure 6). For example, the four genes in C1 (GmJMJ37, GmJMJ38, GmJMJ39, and GmJMJ40) grouped together with GmJMJ41 and GmJMJ42 in a clade, indicating that they may have been produced through evolution after WGD events, but clustered into two clusters and show different expression patterns, implying they have acquired different functions after the duplication event.

In addition, we determined the expression profiles of the recently duplicated JmjC gene pairs in 10 tissues. Most of the paralog pairs generally have the same expression pattern. For further analysis, we divided the duplicated genes into three types based on their detailed expression patterns, shown with a blue, green, and red box in Figure 7. The 12 paralogs in the blue box all have a relatively low expression level in four tissues (pods, podseeds, roots, and seeds) compared with other tissues. These paralogs also show a complex expression pattern in another six tissues examined, but almost all have high expression in the flower, leaf, and shoot meristem. For example, the two copies GmJMJ4 and GmJMJ5 both have high expression in flowers, leaf buds and shoot meristems. The expression pattern of these genes can therefore be pictured as similar to the shape of the letter “M.” The three paralogs in the green box only show high expression in one organ, such as the flower, seed, and root. For instance, GmJMJ7/8 and GmJMJ41-42 both have a high expression level in flower and roots, respectively. The third box contains only one paralog pair (GmJMJ26 and GmJMJ27), and the expression level of one copy (GmJMJ27) is higher than the other copy (GmJMJ26) in all tissues. Overall, the paralogs showed conserved expression profiles, demonstrating that the JmjC gene family has conserved its functions through duplication events.

**DISCUSSION**

As histone demethylases, JmjC domain-containing proteins play essential roles in histone modification, which is a significant part of epigenetics (Klose et al., 2006; Chen et al., 2011). To date, many efforts on the JmjC gene family have been undertaken to elucidate their evolutionary history in a wide variety of plant species, such as *Arabidopsis* (Lu et al., 2008; Zhao et al., 2015), rice (Lu et al., 2008; Zong et al., 2013), and *Fragaria vesca* (Gu et al., 2016). However, little is known about the JmjC gene family in soybean. In this study, we performed a comprehensive analysis of GmJMJs, including their phylogenetic relationships, gene structure, domain architecture, chromosome location, duplication patterns, and expression profiles.

**Phylogeny and Domain Architectures of JmjCs in Soybean**

In total, 48 JmjC genes were identified in the soybean genome, which is larger than other model plants or the other two legumes examined. The number of JmjCs from each species in each group is summarized in Table 1. In most groups, there is still a larger number of JmjCs from soybean than any other species, indicating that these groups may have different evolutionary history among the five species. PKDM11 and PKDM13, two exception groups, both of which including a single gene from each species, may have no duplication or loss after divergence from *Arabidopsis*. The phylogenetic analysis of JmjC genes among five plant species showed that each group contains JmjCs from all species investigated, four eudicots and one monocot, revealing that the JmjC family may have already existed before the divergence of these two lineages. And combining the phylogeny with the time
FIGURE 6 | Heatmaps representing the expression profiles of GmJMJs in several tissues. The numbers following the tissues indicated different developmental stages. The color scale on the right indicates expression values, blue indicating low transcript abundance, and red indicating high levels of transcript abundance.
FIGURE 7 | Three trends of expression patterns of duplicated GmJMJ gene pairs. X-axis represents representative tissues and Y-axis represents scale. The sample used are Cotyledon1 (CO), Flower1 (FL), Leaf1 (LE), LeafBud1 (LB), Pod1 (PO), Pod and Seed1 (PS), Root1 (RO), Seed1 (SE), ShootMeristem (SM), Stem1 (ST).

estimation, 16 GmJMJs pairs, formed by the most recent WGD of soybean, were identified.

Although many GmJMJs were produced by WGD events, there has been little differentiation in their gene structure, domain architecture and the conserved residues acting with cofactors. Another interesting phenomenon is that even outside the conserved coactive-sites we found GmJMJs to have higher amino acid similarity with AtJMJs from the same sub-cluster rather than with GmJMJs from other sub-clusters even though these genes were all grouped in one clade. For example, in PKDM8, GmJMJs1 and GmJMJs2 have the same residues as AtJMJs13 around the Lys (K) site, a Cys (C) and Ser (S), whereas GmJMJs37-38-39-40 have different residues. The unconserved sites around the coactive-sites indicate that the JmjCs may use different ways to bind to the co-factors. However, further evidence is needed to demonstrate and understand this mechanism.

WGDs Contributed to JmjC Gene Expansion in Glycine Max

WGD/segmental duplication and tandem duplication might lead to duplicated gene pairs on the DNA level. Previous studies have shown that the soybean genome has undergone two rounds of WGD (Schmutz et al., 2010). In this study, we have demonstrated that 16 soybean paralogous pairs derived from the second WGD, which suggests that the WGD duplication might be the main mechanism of JmjC gene family expansion and functional diversity during the evolution of soybean. This result is consistent with some other gene families not only AAT gene family (Cheng et al., 2016), GST supergene family (Liu...
et al., 2015), and receptor-like kinase genes (Zhou et al., 2016) in soybean, but also SET domain family in Populus trichocar (Lei et al., 2012; Zhang and Ma, 2012), and 14-3-3 family genes and AP2/ERF superfamily in M. truncatula (Qin et al., 2016; Shu et al., 2016). And in Arabidopsis, previous studies have proposed that more than 90% regulatory genes increased due to WGD (Maere et al., 2005). However, the dispersed duplications and retro-transpositions played the most important role in the evolution of JmjC genes in F. vesca (Gu et al., 2016). Furthermore, KDM3 group is preferentially expanded in the soybean genome compared to other groups, consistent with F. vesca (Gu et al., 2016), indicating that KDM3 group genes may have evolved to meet some unique regulatory needs.

Expression Profiles of JmjC Genes and Functional Diversity of Duplicated Pairs in Soybean

In angiosperms and vertebrates, both of the JmjC and SET genes, maintaining homeostasis of the histones methylation, are the key regulators of chromatin structure, suggesting that the epigenetic modulation playing an important role in regulation of gene expression in developmental stages and responses to abiotic stresses (Lei et al., 2012; Zhang and Ma, 2012; Qian et al., 2015). We investigated the expression profiles of JmjC genes using public expression data and found that most JmjC genes are widely expressed (Figure 6), indicating that these genes, remained after WGD/segmental events, are likely functional. To further elucidate whether functional differentiation has occurred after the WGD event, we analyzed the expression patterns between the duplicated pairs (Figure 7). All the expression patterns of 16 duplicated gene pairs can be classified into three types according to their tendency in each tissues detected. The first type is that two copies have the complicated expression patterns in different tissues. We could hypothesize possible functions of GmJMJs by coupled their expression patterns with the functions of their Arabidopsis orthologs. For example, AtJMJ24, the ortholog of GmJMJ17 and GmJMJ18 in Arabidopsis, is a histone H3K9 demethylase (Lu et al., 2008). AtJMJ24 has been proved to promote basal level transcription of endogenous silenced loci by countering H3K9me (Deng et al., 2015). Therefore, GmJMJ17 and GmJMJ18 might have the functions similar to AtJMJ24 in reinforcing the silence. The second type showed that both copies have high expression in one organ, such as the flower, seed, and root. The third type is that one duplicate was expressed at higher levels than the other one nearly in all tissues, implying that the former one has stronger function than the latter, and implying that it may play important roles in regulating broad developmental or reproductive stages. Above all, the expression patterns among the duplicated pairs are relatively conserved, suggesting little functional differentiation has occurred following the WGD event. However, some JmjC genes were specific to soybean, for example, GmJMJ4 and GmJMJ5. All of these genes have abundant transcripts in soybean and are expressed at different levels in different tissues. These results indicate that in Arabidopsis their counterparts may be lost and the functions might have been performed by other genes.

CONCLUSIONS

Here, we performed comprehensive and evolutionary analyses of JmjC gene family in soybean, and provided detailed information on its members. A total of 48 putative JmjC genes were identified in the soybean genome, which represented non-random across all soybean genome chromosomes and majority of them expanded from WGD/segmental duplication rather than the dispersed duplications. The exon/intron compositions and domain arrangements were considerably conserved among members in the same groups or subgroups. Many duplicated genes present similar expression patterns in soybean tissues detected implying functional conservation. The close phylogenetic relationship between GmJMJs and AtJMJs in the same subgroup provided insights into their putative functions. Taken together, all of these results provided valuable clues in future efforts to identify specific gene functions of this gene family and gene diversity among different genotype of soybean and other plants in Leguminosae.

AUTHOR CONTRIBUTIONS

LW and LZ designed the research; YH, XL, and LW performed phylogenetic analysis and wrote the manuscript; LC, YL, and HW annotated the JmjC genes on chromosomes and calculated the duplication date; LC, DK, and HY analyzed the expression data.

ACKNOWLEDGMENTS

This work was supported by the foundation and frontier technology research of Henan Province (No. 162300410257), the industry-university-research cooperation of Henan Province (No. 162107000032), Funding scheme for young core teachers of Xinyang Normal University (2015), Nanhu Scholars Program for Young Scholars of XYNU, and Funding scheme for young backbone teachers of Xinyang Normal University (2016GJS-13). This work was also supported by the National Natural Science Funds of China (No. 31400213). Fujian-Taiwan Joint Innovative Center for Germplasm Resources and Cultivation of Crop [Fujian 2011 Program, (2015)75] to LZ.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2016.01800/full#supplementary-material

Figure S1 | Phylogenetic relationship of JmjC-domain containing proteins from five plant species by using the JmjC domain alone.

Figure S2 | Phylogenetic relationships of GmJMJs.

Figure S3 | Alignment of JmjC domain sequences. The conserved residues compatible with the demethylase activity within the Fe(II) binding site are highlighted in red and those in the U-KG binding site are indicated in yellow. The sequences with black, gray, light gray background indicated identical 100%, conservative (75–99%), and block (50–74%) similarity of amino acid residues, respectively.

Table S1 | Protein sequences of Arabidopsis, Oryza sativa JmjC domain-containing protein used as blast queries.


Quan, Z., Oliver, S. G., and Zhang, N. (2011). JmjN interacts with JmjC to ensure selective proteolysis of Gis1 by the proteasome. *Microbiology* 157, 2694–2701. doi: 10.1099/mic.0.048199-0


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

Copyright © 2016 Han, Li, Cheng, Liu, Wang, Ke, Yuan, Zhang and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.
Comprehensive Analysis of the CDPK-SnRK Superfamily Genes in Chinese Cabbage and Its Evolutionary Implications in Plants

Peng Wu¹, Wenli Wang¹, Weike Duan¹,², Ying Li¹ and Xilin Hou¹*¹

¹ State Key Laboratory of Crop Genetics and Germplasm Enhancement, Key Laboratory of Biology and Germplasm Enhancement of Horticultural Crops in East China, Ministry of Agriculture, Nanjing Agricultural University, Nanjing, China, ² School of Life Science and Food Engineering, Huaiyin Institute of Technology, Huaiian, China

The CDPK-SnRK (calcium-dependent protein kinase/Snf1-related protein kinase) gene superfamily plays important roles in signaling pathways for disease resistance and various stress responses, as indicated by emerging evidence. In this study, we constructed comparative analyses of gene structure, retention, expansion, whole-genome duplication (WGD) and expression patterns of CDPK-SnRK genes in Brassica rapa and their evolution in plants. A total of 49 BrCPKs, 14 BrCRKs, 3 BrPPCKs, 5 BrPEPRKs, and 56 BrSnRKs were identified in B. rapa. All BrCDPK-SnRK proteins had highly conserved kinase domains. By statistical analysis of the number of CDPK-SnRK genes in each species, we found that the expansion of the CDPK-SnRK gene family started from angiosperms. Segmental duplication played a predominant role in CDPK-SnRK gene expansion. The analysis showed that PEPRK was more preferentially retained than other subfamilies and that CPK was retained similarly to SnRK. Among the CPKs and SnRKs, CPKIII and SnRK1 genes were more preferentially retained than other groups. CRK was closest to CPK, which may share a common evolutionary origin. In addition, we identified 196 CPK genes and 252 SnRK genes in 6 species, and their different expansion and evolution types were discovered. Furthermore, the expression of BrCDPK-SnRK genes is dynamic in different tissues as well as in response to abiotic stresses, demonstrating their important roles in development in B. rapa. In summary, this study provides genome-wide insight into the evolutionary history and mechanisms of CDPK-SnRK genes following whole-genome triplication in B. rapa.

Keywords: CDPK-SnRK genes, Brassica rapa, evolutionary conservation, synteny analysis, evolutionary pattern, expression pattern

INTRODUCTION

Plants are remarkably responsive to a variety of environmental stimuli, including pathogen attack, wounding, cold, drought reception, and fluctuations in incident light (Kudla et al., 2010). Meanwhile, a variety of internal substances also affect plants growth. These external and internal signals compose a complex regulatory network that allows plants to develop in balance. Following the detection of a stress stimulus, various signal transduction pathways are switched on, resulting...
in physiological changes in the plant cell. As second messengers, calcium ions play an essential role in many important cellular processes, especially under stress conditions (Trewavas and Malhó, 1998; Sanders et al., 1999; Berridge et al., 2000). In plants, transient changes in calcium content in the cytosol (calcium signatures) have been observed during growth, development and stress conditions (Evans et al., 2001; Harper, 2001; Knight and Knight, 2001; Sanders et al., 2002). Intracellular Ca²⁺ signals are produced in plant cells by a variety of stimuli, such as changes in environmental conditions, interaction with microbes, and developmental programs (Bush, 1995; Ehrhardt et al., 1996; Hammond-Kosack and Jones, 1996; Knight et al., 1996; Taylor and Hepler, 1997; Pei et al., 2000; Assmann and Wang, 2001; Murata et al., 2001; McInish et al., 2002; Plith and Trewavas, 2002; Ritchie et al., 2002). Plants have multiple calcium stores, including the apoplast, vacuole, nuclear envelope, endoplasmic reticulum (ER), mitochondria and chloroplasts. Therefore, each stimulus can elicit a characteristic Ca²⁺ wave by specifically altering the activity of various differentially localized Ca²⁺ channels, H/Ca²⁺ antiporters, and Ca²⁺ and H⁺ ATPases (Thuleau et al., 1998; Allen et al., 2000; Harmon et al., 2000; Hwang et al., 2000). Different calcium sensors recognize specific calcium signatures and transduce them into downstream effects, including altered protein phosphorylation and gene expression patterns (Sanders et al., 1999; Rudd and Franklin-Tong, 2001).

In eukaryotes, calcium-dependent protein kinases (CDPKs) and most sucrose non-fermenting-1-related kinases (SnRKs) are involved in regulating and decoding Ca²⁺ signals (Assmann and Wang, 2001; Evans et al., 2001; Harmon et al., 2001; Cheng et al., 2002; Fasano et al., 2002; Hrabak et al., 2003; Cho et al., 2009; Kulik et al., 2011). The protein kinases also involved in stress signal transduction in plants are common to all eukaryotic organisms and include mitogen-activated protein kinases (MAPKs), glycogen synthase kinase 3 (GSK3), and S6 kinase (S6K). The CDPK-SnRK superfamily consists of seven types of protein kinases, which differ in the regulatory domains and include miotogen-activated protein kinases (MAPKs), glycogen synthase kinase 3 (GSK3), and S6 kinase (S6K). The CDPK-SnRK superfamily consists of seven types of protein kinases, which differ in the regulatory domains they contain (Harmon et al., 2001). CDPKs (also named CPKs) are activated by the binding of calcium to their calmodulin-like regulatory domains. The carboxyl terminal domains of CRKs (CDPK-related kinases) have sequence similarity to the regulatory domains of CPKs but do not bind calcium. PEPRKs (PEP carboxylase kinases) contain only one catalytic domain (Harmon et al., 2001). PPCKs (PEPC kinases) have a carboxyl-terminal domain that has no similarity to that of any other member of the superfamily (Harman et al., 2003). CCAKMs (calcium- and calmodulin-dependent protein kinases) bind both calcium ions and the calcium/calmodulin complex, whereas CaMKs (calmodulin-dependent protein kinases) bind the calcium/calmodulin complex but not calcium (Hrabak et al., 2003). In addition, there are the classical SNF1-type kinases from yeast; Halford and Hardie (1998) proposed the name SNF1-related kinase (SnRK) for this group and recognized three subgroups: SnRK1, SnRK2, and SnRK3 (Harmon et al., 2001). However, CaMK and CCAmK are absent from Arabidopsis (Hrabak et al., 2003). All members of the CDPK-SnRK gene superfamily have kinase domains of similar length and sequence and a similar general organization, with the kinase domains at or near the N-terminus, then the junction domains, followed by the regulatory domains (Harmon, 2003; Hrabak et al., 2003).

The plant CPKs characterized to date play substantive roles in diverse physiological processes. These processes include tolerance to salt, cold, and drought stress in rice (Asano et al., 2002), the regulation of development of nodule number in Medicago truncatula (Gargantini et al., 2006), and the response to ABA in Arabidopsis (Choi et al., 2005) (38). The original, systematic report on the CPG genes family in Arabidopsis thaliana identified 34 CPK genes family members (Choi et al., 2005) and was followed by research in rice (Oryza sativa) (Ray et al., 2007) and wheat (Triticum aestivum) (Li et al., 2008). Recently, genome-wide analyses of the CPK gene family have been reported in maize (Ze a mays) (Kong et al., 2013) and poplar (Populus trichocarpa) (Zuo et al., 2013). Meanwhile, more and more investigations of CPK genes have also involved horticultural plants, such as alfalfa (Davletova et al., 2001), potato (Raïces et al., 2003), strawberry (Llop-Tous et al., 2002), and tomato (Chico et al., 2002). Furthermore, research using transgenic plants has revealed the biological functions of a few CPK genes in higher plants. Transgenic rice constitutively overexpressing OsCPK7 or OsCPK13 showed enhanced tolerance to cold, salt, and drought stress (Asano et al., 2000; Komatsu et al., 2007). In tobacco, CPK-silenced plants displayed a reduced and delayed hypersensitive response to the fungal Avr9 elicitor (Romeis et al., 2001). GhCPK1 was the first cotton CPK gene to be identified and was considered to play a role in the calcium signaling events associated with fiber elongation (Huang et al., 2008). Arabidopsis thaliana CDPK23 (AtCPK23) is a positive regulator of the response to drought and salt stress (Ma and Wu, 2007), whereas AtCPK6 may be crucial in positively regulating methyl-jasmonate signaling in guard cells (Munemasa et al., 2011). In addition, the overexpression of rice (Oryza sativa) CPK7 (OsCPK7) significantly improves resistance to cold (Komatsu et al., 2007). Phytohormones are involved in the responses to abiotic stresses; therefore, the expression levels of members of the CPK gene family has also been shown to be regulated after treatment with various phytohormones, such as ABA, auxin and jasmonic acid. Recently, Ze a mays CPK11 was reported as a component of the jasmonic acid signaling pathway, and its concentration in cells was observed to increase in response to wounding and touch (Szczegielniak et al., 2012).

Sucrose non-fermenting-1-related protein kinase (SnRK) is homologous to SNF1 and AMP-activated protein kinase (AMPK), which is widely distributed in plants and is involved in a variety of signaling pathways. SnRK is the key switch in plant sugar signaling, stress, seed germination and seedling growth. SNF1 of yeast, AMPK of mammals and SnRK1 of plants are homologous, belonging to the SNF1 protein kinase superfamily. SNF1 was found in yeast (Saccharomyces cerevisiae) originally (Alderson et al., 1991). In yeast, glucose regulates the protein-protein interaction, substrate specificity and subcellular localization of the SNF1 subunit that modulates SNF1 kinase activity, resulting in the phosphorylation of activators and repressors that control transcription of multiple genes in
metabolic pathways required for the utilization of alternative energy sources. In the eukaryote, SNF1 protein kinase is very strongly conserved. Many SNF1 analogs have been identified in plants. SnRK1 was discovered initially in rye (Secale cereale L.) (Alderson et al., 1991). At present, some members of the SnRK1 subfamily have been found in variety of model plants and some important crops, such as Arabidopsis thaliana, rye, barley (Hordeum vulgare), potatoes (Solanum tuberosum), tobacco, beets, etc. It may exist in all plants (Halder and Hardie, 1998; Halford et al., 2003). Studies have shown that SnRK1 is the key switch in plant sugar signaling. In addition, the regulation of glucose metabolism, hormonal regulation and sugar signaling is directly related to signal transduction (Kleinow et al., 2000; Jossier et al., 2009; Mathieu et al., 2009). SnRK2s are a plant-specific Ser/Thr protein kinase family. All of the members have a conserved N-terminal catalytic domain similar to that of SNF1/AMPK-type kinases and a short C-terminal regulatory domain that is not highly conserved. Prior to 2000, there were only a small number of studies indicating that ABA and abiotic stresses induced the expression of some SnRK2 genes (Anderberg and Walker-Simmons, 1992; Holappa and Walker-Simmons, 1995). In 2000, SnRK2s began to be recognized as enzymes involved in abiotic stress signal transduction in plants (Li et al., 2000). By 2003, 10 SnRK2 genes had been identified and were renamed SnRK2.1 through SnRK2.10 (Hrabak et al., 2003). In 2009, independently, two laboratories obtained a triple SnRK2.2/2.3/2.6 mutant. SnRK2.2/2.3/2.6 triple-mutant plants are nearly completely insensitive to ABA, which was used to establish the role of ABA-dependent SnRK2s in the plant response to water deficit, seed maturation, and germination. These reports indicate that SnRK2.2/2.3/2.6 function as primary positive regulators and suggest that ABA signaling is controlled by the dual modulation of SnRK2.2/2.3/6 and group A PP2Cs (Fujii and Zhu, 2009; Fujii et al., 2009; Nakashima et al., 2009). SnRK3 is a protein kinase in plants, called calcineurin B-like calcium sensor-interacting protein kinase (CIPK) (Kim et al., 2000). CIPK interacts with the calcium-binding protein SOS3, ScaBPS and CBL (calcineurin B-like calcium sensor). Studies have shown that CIPK and an upstream complex of CSL interactions are involved in salt stress, sucrose and ABA signal transduction (Imamura et al., 2008). In Arabidopsis, PKS3, PKS18 and CIPK3 of the SnRK3 family can regulate plant growth, stomatal opening and closing and seed germination under ABA treatment (Kim et al., 2003). Arabidopsis AtCIPK1 forms complexes with AtCBL1 and AtCBL9, regulating ABA-independent and ABA-dependent pathways, respectively (D’Angelo et al., 2006). AtCIPK3 regulates ABA and cold signal transduction pathways (Kim et al., 2003). Girdhar’s study showed that CBL9 interacted with CIPK3 to regulate the ABA pathway, and this finding was validated in a yeast two-hybrid experiment (Pandey et al., 2008).

During their evolution, plants have substantially altered their phenotypes to adapt to environmental changes by transforming the form and function of genes. Gene duplication, even a whole-genome duplication (WGD), offers the chance for genes to change (Rensing, 2014). Angiosperm genome evolution is characterized by polyploidization through WGD followed by diploidization, which is typically accompanied by considerable homoeologous gene loss (Stebbins, 1950). After duplication, one copy of the gene might either becomes non-functional (pseudogenized or silenced, also called gene death) or acquire a novel function (neofunctionalization). Alternatively, the two duplicates might divide the original function of the gene (Innan and Kondrashov, 2010). Preliminary analyses revealed that gene duplication and subsequent divergence are the main contributors to evolutionary momentum (Ohno et al., 1968; Chothia et al., 2003). The genome of A. thaliana has experienced a paleohexaploidy (β) duplication shared with most dicots and two subsequent genome duplications (α and γ) since its divergence from Carica papaya, along with rapid DNA sequence divergence and extensive gene loss (Bowers et al., 2003). In A. thaliana, some duplicated regions found in CDPK-SnRK protein kinases indicated that CDPK-SnRK protein kinases are paralogs that arose by divergence after genome duplication events (Hrabak et al., 2003). The CIPK genes of Arabidopsis and maize have undergone both segmental and tandem duplication, contributing to the expansion of the CIPK family. In Populus, however, segmental duplication played a predominant role in the expansion of CIPK genes (Zuo et al., 2013). In addition, tandem duplication of CIPK genes has not occurred in the rice genome (Asano et al., 2005).

In this study, we constructed a comprehensive comparative analysis of CDPK-SnRK genes, including phylogenetic relationships, gene structures, chromosome distribution, gene retentions, gene expansions, gene duplication and gene expression patterns, in different tissues to characterize the divergences in composition, expansion, and expression. First, we identified 555 CPKs, 120 CRKs, 5 PPCKs, 14 PEPKs, and 697 SnRKs in 16 plant species. Second, we conducted a comparative genomic analysis of these genes with 16 other plant and species found that the expansion of the CDPK-SnRK family from angiosperms mainly relied on WGDs. Third, PEPK genes were more preferentially retained than other subfamilies and CPK genes were retained similarly to SnRK genes during divergence and extensive gene loss (Bowers et al., 2003). In Arabidopsis, some duplicated regions found in CDPK-SnRK protein kinases indicated that CDPK-SnRK protein kinases are paralogs that arose by divergence after genome duplication events (Hrabak et al., 2003). In addition, tandem duplication of CIPK genes has not occurred in the rice genome (Asano et al., 2005).

RESULTS

Identification and Classification of the CDPK-SnRK Superfamily of Protein Kinases in Brassica rapa and Comparative Analyses

In this study, genome-wide analysis of CDPK-SnRK gene family has been performed on the basis of the completed B. rapa genome sequence (Wang et al., 2011). Based on previously reported methods (Harmon et al., 2001; Hrabak et al., 2003), the homogeneous candidate CDPK-SnRK genes between
Brassica rapa and other species were identified by BLASTP (Supplementary Table 1). Subsequently, all candidate protein sequences were subjected to Pfam and SMART analyses. Finally, we identified 49 BrCPKs, 14 BrCRKs, 3 BrPPCKs, 5 BrPEPRKs, and 56 BrSnRKs named according to nomenclature proposed for CDPK-SnRK genes (Supplementary Table 2).

To better understand the expansion and evolutionary history of CDPK-SnRK genes in B. rapa, genes were also identified in 16 other species representing the major clades of plants. The evolutionary relationships of the species and the number of CDPK-SnRK genes are shown in Figure 1A. The data show that Glycine max contained the highest number of CDPK-SnRK genes (200), followed by Z. mays (193) and M. truncatula (188) (Figure 1A). However, A. trichopoda, a basal angiosperm species that was the single living representative of the sister lineage to all other extant flowering plants, contained the lowest number of CDPK-SnRK genes (28) in Angiospermae. The reason is that it originated prior to the split of eudicots and monocots and has not experienced any whole genome duplication (WGD), while the other 12 angiosperms had several rounds of WGDs/triplications after their split from A. trichopoda. Furthermore, the number of CDPK-SnRK genes in algae, Bryophyta and Pteridophyta was less than that in Angiospermae. This phenomenon was also caused by several WGD events that occurred during angiosperm evolution (Figure 1B). These results indicated that the expansion of the CDPK-SnRK family from angiosperms mainly relied on large-scale DNA rearrangements, namely, WGDs. The elevated duplication frequency and increased retention of CDPK-SnRK genes also contributed to neofunctionalization and caused them to gain important functions in angiosperm development.

Characteristics of Structure, and Expansion Analysis of BrCDPK-SnRK Proteins

To investigate the extent of lineage-specific expansion of the CDPK-SnRK genes in B. rapa, phylogenetic trees were constructed using the maximum likelihood method (Figure 2A). The phylogenetic tree showed that all the CDPK-SnRK genes were clustered into five distinct gene classes (CPK, CRK, PEPRK, PPCK, SnRK) (Figure 2A), while the CPK family was divided into four groups (I, II, III, and IV) and the SnRK family was classified into three groups (SnRK1, SnRK2, and SnRK3), consistent with the reports in A. thaliana (Hrabak et al., 2003). In B. rapa, the CPK, CRK, PEPRK, PPCK, and SnRK gene families contained 49 members, 14 members, 3 members, 5 members, and 56 members, respectively, whereas in A. thaliana, the CPK, CRK, PEPRK, PPCK, and SnRK families contained 34 members, 8 members, 2 members, 2 members, and 39 members, respectively. Next, the synteny of CDPK-SnRK genes between A. thaliana and three subgenomes in B. rapa was analyzed. There were 34 CPK, 8 CRK, 2 PEPRK, 2 PPCK, and 38 SnRK genes on the conserved collinear block (Supplementary Table 5). Meanwhile, 2 CPKs, 2 CRKs, 1 PEPRK, 1 PPCK, and 3 SnRKs were retained completely; conversely, 4 CPKs and 3 SnRKs from B. rapa were lost. Due to the Brassica-specific WGT event, the gene numbers of these genes in B. rapa were greater than those in A. thaliana.

Furthermore, the different domain architectures, motif compositions and gene structures of CDPK-SnRK were analyzed (Figure 2B, Figure S1). All members of the CDPK-SnRK superfamily have a kinase domain of similar length and sequence, with the kinase domains at or near the N-terminus, then the junction domains, followed by the regulatory domains (Figure 2B). Although CPK proteins have a functional kinase domain coupled with regulatory calcium-binding EF-hands, the C-terminal domains (EF-hands) of CRK proteins contain apparently degenerate calcium-binding sites with no function. Meanwhile, 10 conserved motifs were detected in BrCPK, BrCRK, BrPEPRK, BrPPCK, and BrSnRK, respectively (Figure S3). All BrCDPK-SnRK proteins had highly conserved kinase domains, which corresponded to motifs 1-4, 7, 9 in BrCPK and motifs 1-4, 6, 7, 9 in BrSnRK, whereas motif 8 was found in BrSnRK3.1, corresponding to the NAF/FISL domains (Figures S1A,B). The amino acid sequence of BrCDPK-SnRK was aligned with A1CDPK-SnRK protein sequences from five gene classes. In CPK, CRK, PEPRK, PPCK, and SnRK, higher sequence similarities were identified in the N-terminus, which corresponded to the conserved kinase domain (Figure S2). In addition, variable gene structures of BrCDPK-SnRK were observed. As shown in Figures 3A-C, the intron numbers of the BrCPK genes ranged from 5 to 9 with a median of 6, while the BrSnRK genes ranged from 0 to 15 with a median of 7. Interestingly, we found that 22 BrSnRK3 genes have no introns. The theoretical pI of the BrCPK gene family ranged from 4 to 9 with a median of 6, but BrSnRK proteins showed a pI range from 2 to 11 and with a median of 8 (Figure 3B). Other classes had complex theoretical pI ranging in value from 4 to 9 (Figure 3B).

Different Retention of CDPK-SnRK Genes Following WGT in Brassica rapa

To investigate different retention in CPK, CRK, PEPRK, PPCK, and SnRK during B. rapa WGT events, 44/49, 14/14, 3/3, 5/5, and 50/56 were located in the syntenic regions, respectively (Figures 4A,B and Supplementary Table 3). The results demonstrated that 43% (44/102) of the CPK genes were retained in the syntenic regions, relative to 44% (50/114) of the SnRK genes. The retention rates of CPK, PEPRK and PPCK are 58% (14/24), 50% (3/6), and 83% (5/6), respectively (Figure 4F). Additionally, we counted gene copies and analyzed the distribution of the three subgenomes by comparing the retention of CPK, CRK, PEPRK, PPCK, and SnRK (Figure 4E). The result showed that all PEPRK genes had more than two copies retained, which is more than the retention of the other subfamilies (42%). However, 3% of the CPK and SnRK genes were completely lost. Next, the proportions of CPK and SnRK genes retained were higher in the least fractionated (LF) subgenome than in the medium fractionated (MF1) and most fractionated (MF2) subgenomes, consistent with a previous report showing that the degree of retained genes in these three subgenomes (LF, MF1, and MF2) was decreased (Wang et al., 2011). In contrast, the PEPRK and PPCK families were retained more in the MF1 subgenome.
than in the LF subgenome (Figure 4F). In summary, the results confirmed that PEPRK genes were more preferentially retained than other subfamilies and that CPK genes were retained similarly to SnRK genes during diploidization following WGT in B. rapa.

Furthermore, the retention rates of four CPK (CPKI, II, III, IV) and three SnRK (SnRK1, 2, 3) groups were observed. As shown in Figure 4H, 67% of CPKIIIs and SnRK1s had more than two copies retained, which is greater than the retention of the other groups (Figures 4H,D). In addition, the proportion of CPKIIIs and SnRK1s retained was higher in the LF subgenome than in the other subgenomes, which once again confirmed that CPKIII and SnRK1 genes were more preferentially retained than other groups (Figures 4C,G).

**Chromosome Distribution, Ks and Duplication Analysis of the CDPK-SnRK Genes in B. rapa**

All BrCDPK-SnRK genes could be mapped onto 10 chromosomes of Chinese cabbage with a non-random distribution, except BrPPCK3, which is located in Scaffold000191 (Figure 5). On every chromosome, the proportion of BrCPK genes was similar to that of BrSnRK genes. However, Chromosome 09 contained more BrCPK genes (11 genes) than BrSnRK genes, whereas chromosomes 01 had the opposite. B. rapa shares two WGDs (WGD: α and β) and one whole-genome triplication event (WGT: γ) in its evolutionary history with Arabidopsis but has undergone an additional WGT event. Therefore, the B. rapa genome was further divided into three
FIGURE 3 | Characteristics of isoelectric points and introns among BrCDPK-SnRK. (A) I indicates CDPK-SnRK genes within the *B. rapa* genome. II indicates the *B. rapa* chromosome karyotype. III represents the number of introns (red) of CDPK-SnRK genes. IV indicates the Pi value (blue) of BrCDPK-SnRK genes. (B) The Pi value among gene classes BrCPK, BrCRK, BrPEPRK, BrPPCK, and BrSnRK. (C) The number of introns among gene classes BrCPK, BrCRK, BrPEPRK, BrPPCK, and BrSnRK.

FIGURE 4 | CDPK-SnRK homologous genes in the segmental syntenic regions of the genomes of *Brassica rapa* and *Arabidopsis thaliana* and their different retention. (A) CPK, CRK, PEPRK, and PPCK syntenic gene lines are shown between the 10 *B. rapa* chromosomes (Br01-Br10) and the five *A. thaliana* chromosomes (At01-At05). The pink lines represent the syntenic gene pairs between Chinese cabbage and Arabidopsis; the blue lines represent the syntenic genes in Chinese cabbage. (B) SnRK syntenic gene lines are shown between the 10 *B. rapa* chromosomes and the five *A. thaliana* chromosomes. The green lines represent the syntenic genes pairs between Chinese cabbage and Arabidopsis; the red lines represent the syntenic genes in Chinese cabbage. (C) Retention of homoeologs of SnRK genes in the three subgenomes (LF, MF1, and MF2) in *B. rapa*. LF: least fractionized subgenome; MF1: moderately fractionized subgenome; MF2: most fractionized subgenome. (D) Copy numbers of SnRK genes after genome triplication and fractionation in *B. rapa*. (E) Copy numbers of CDPK-SnRK genes after genome triplication and fractionation in *B. rapa*. (F) Retention of homoeologs of CDPK-SnRK genes in the three subgenomes (LF, MF1, and MF2) in *B. rapa*. (G) Retention of homoeologs of CPK genes in the three subgenomes (LF, MF1, and MF2) in *B. rapa*. (H) Copy numbers of CPK genes after genome triplication and fractionation in *B. rapa*. 
FIGURE 5 | The features of the CDPK-SnRK genes in B. rapa. The distribution of BrCDPK-SnRK genes on 10 chromosomes. The 8 ancestral blocks and three subgenomes, including the least fractionated (LF), medium fractionated (MF1), and most fractionated (MF2) subgenomes, were plotted as described by Cheng et al. (2013). AK represents the ancestral karyotype. The red lines connect the duplicated BrCDPK-SnRK genes.

differentially fractionated subgenomes (LF, MF1, MF2), of which LF contained more BrCPK genes and BrSnRK genes than either of the other two subgenomes. In addition, the 24 conserved ancestral genomic blocks (labeled A–X) in the B. rapa genome were reconstructed according to previous reports (Cheng et al., 2013). The color coding of these blocks was based on their positions in a proposed ancestral karyotype (AK1-8). We also found that most of the BrCPK genes cluster together in a region of AK1 (20%), whereas BrSnRK genes belonged to AK1 (18%) and AK3 (18%).

Furthermore, the types were identified by the MCScanX program, and the divergence times of the duplicated genes were estimated by calculating the synonymous substitution rates (Ks) and non-synonymous substitution rates (Ka). In total, 70 BrCDPK-SnRK duplicated gene pairs were analyzed (Supplementary Table 4). BrCPK (74%), BrCRK (86%), BrPEPRK (100%), BrPPCK (100%), and BrSnRK (70%) duplicated gene pairs were segmental duplications (Figure 6B), and all the duplicated BrCDPK-SnRK gene pairs had Ka/Ks ratios less than 1, indicating the purifying selection of these genes (Figure 6A, Supplementary Table 4). To understand the divergence time, the Ks values of the BrCPK genes ranged from 0.3 to 0.5 and had a mean of ~0.34 (~11 Myr), while the BrSnRK genes ranged from 0.2 to 0.55 and focused on ~0.25 (~8.5 MYr; Figures 2A,B, 6A). The divergence time of BrSnRK duplicated gene pairs was 8.49 MYA, which indicates that their divergence occurred during the Brassica triplication events (5 ~ 9 MYA). The divergence times obtained for the BrCPK duplicated gene pairs ranged from 10 to 16.6 MYA, indicating that these duplications occurred during the divergence of Chinese cabbage and Arabidopsis (9.6–16.1 MYA).
Evolution Pattern of CDPK-SnRK Genes in Plants

To investigate the evolution of the CDPK-SnRK family in the plant kingdom, we selected 13 Angiospermae (8 eudicots, 4 monocots and one basal angiosperm), 3 Gymnospermae, 1 Pteridophyta, 1 Bryophyta, and 1 Chlorophyta species for comparative analysis (Figure S3). We constructed a phylogenetic tree of the CDPK-SnRK genes to analyze the evolutionary relationships of these species. The phylogenetic tree showed that the CDPK-SnRK genes formed five gene classes (CPK, CRK, PEPRK, PPCK, and SnRK), which is consistent with the result for B. rapa and A. thaliana. Meanwhile, we found that no CRK, PEPRK, or PPCK genes were detected in Volvox carteri. Therefore, the CRK, PEPRK, and PPCK gene families may only exist in land plants. To further determine the relationship among the five groups, the genetic distance was analyzed (Figures S4A,B). The box plot shows the genetic distance of CPK vs. CRK, PEPRK, and PPCK, which was smaller than SnRK vs. these groups (Figure S4B). Notably, the genetic distance between CPK and CRK was smaller than that between CPK and PEPRK or CPK and PPCK. These results indicate that CPK has a closer relationship with CRK, PEPRK, and PPCK, especially the CRK closest to CPK, which is consistent with previous reports that plant CPK and CRK may share a common evolutionary origin (Hrabak et al., 2003).

To further investigated the footprint of the CPK CRK, PEPRK, and SnRK families, we selected four angiosperms (C. papaya, P. trichocarpa, A. trichopoda, and V. vinifera). The reason is that Vitis vinifera, P. trichocarpa, and C. papaya did not undergo α and β duplications and A. trichopoda, a basal angiosperm, did not undergo the γ duplication event (Jiao et al., 2011b; Albert et al., 2013; Lee et al., 2013). Phylogenetic trees in each species were constructed (Figures S5, S6). In each species, the CPK family was divided into four clades (CPKI, CPKII, CPKIII, and CPKIV), and SnRK was divided into three clades (SnRK1, SnRK2, and SnRK3). CPK, CRK, and SnRK were found to exist in Amborella trichopoda, which indicates that these three groups originated from duplication events prior to the γ event. However, PEPRK appeared between the salicoid duplication and the γ event. The PPCK family exists in P. trichocarpa, which indicates that it originated from duplication events prior to the salicoid duplication. Furthermore, due to the salicoid duplication and Brassica-specific WGT events, there were more CDPK-SnRK gene family members in P. trichocarpa and B. rapa than in other species (Tuskan et al., 2006; Wang et al., 2011). In general, during the course of evolution, CPK appeared most recently and expanded most rapidly. Above all, we inferred a possible evolutionary footprint of the CPK family (Figure S7).

The family size and the percentage of CPKs in five plants suggested that CPKs expanded rapidly during evolution and further expanded in the Brassicaceae (Figure 7). WGD is known to have important impacts on the expansion and evolution of gene families in plant genomes. However, along with the gradual increase in the CPK percentage, the genes of group III were completely lost in V. vinifera (Figure 7C). Compared with other groups, the expansion of group III was more unstable. To further investigate the relationship among the four groups,
the genetic distance was analyzed. The results indicated that the genetic distance between CPKI and CPKIII was shorter than that between CPKI and CPKIV or that between CPKI and CPKII, and the genetic distance between CPKII and CPKIV was shorter than that between CPKII and CPKIII (Figure 7B). In summary, we inferred a possible evolutionary footprint of the CPK family. Before the $\gamma$ event, all groups in the family (CPKI, CPKII, CPKII, CPKIV) had already appeared. The gene family further expanded within Brassicaceae. Thus, the CPK family almost doubled in size in the *B. rapa* genome compared with that of *A. trichopoda* through three duplications, one triplication, and fractionation. The expansion of groups I, II, and IV played a major role in the expansion of the CPK gene family.

**Comparative Expression Pattern Analysis of the CDPK-SnRK Genes in Different Tissues of *Brassica rapa* and *Arabidopsis thaliana***

To investigate the divergence expression profiles of CDPK-SnRK genes in different organs between *A. thaliana* and *B. rapa*, including roots, stems, leaves, flowers and siliques, the expression patterns of all genes were investigated (Figure S8, Supplementary Tables 5, 6). BrCDPK-SnRK genes were found to be expressed in roots (99 BrCDPK-SnRKs; 77.95%), stems (104; 81.89%), siliques (106; 83.46%), leaves (101; 79.53%), and flowers (119; 93.7%) (Figures 8A–D). A total of 75 (88%) AtCDPK-SnRKs showed high expression (mean-normalized value $> 1$) in at least one of the five tissues (Figure S8), including roots (34 AtCDPK-SnRKs; 40.00%), stems (36; 42.35%), siliques (22; 25.88%), leaves (19; 22.35%), and flowers (16; 18.23%). Among the 116 CPKs (including 46 AtCPKs and 70 BrCDPKs), 4 (CPK28, CPK29, CPK30, and CPK33) were not expressed and 2 (BrCPK4 and BrCPK25) were only expressed in flower tissue (Figures 8A,B). In addition, a total of 34 BrCPKs, 8 BrCRKs, 4 BrPEPRKs, and 2 BrPPCKs had high expression levels (FPKM value $> 10$) in at least one tissue; 13 BrCPKs, 2 BrCRKs, and 1 BrPEPRK were highly expressed in all 5 tissues. However, only 2 SnRK genes were expressed in one tissue; the remains 93 SnRK genes were expressed in five tissues. Subsequently, we selected the expression patterns of genes in five tissues on the phylogenetic tree of all CDPK-SnRK genes to investigate whether the functions of homologous genes were divergent (Figure S8). All CPKI, PEPRK,
Wu et al. Comprehensive Analysis of the CDPK-SnRK FIGURE 8 | Analysis of the CDPK-SnRK Genes in different tissues of Brassica rapa. (A,B) Heat map representation and hierarchical clustering of BrCPK, BrCRK, BrPEPRK, and BrPPCK genes in root, stem, leaf, flower, silique, and callus; Venn diagram depicting the distribution of shared expression of the BrCPK, BrCRK, BrPEPRK, and BrPPCK genes among five Brassica rapa tissues. (C,D) Heat map representation and hierarchical clustering of BrSnRK genes in root, stem, leaf, flower, silique, and callus; Venn diagram depicting the distribution of shared expression of the BrSnRK genes among five Brassica rapa tissues.

and PPCK and most CPKIII and CRK genes had high expression levels, suggesting significant roles of these genes in plant development. Most BrSnRK3s exhibited little or no expression. However, the AtSnRK3s were all expressed in five tissues, indicating that BrSnRK3 genes may have lost some functions after the duplication events (Figure S8).

Expression Divergence and Coregulatory Networks of CDPK-SnRK Genes under Multiple Treatments in Brassica rapa

CDPK-SnRKs are suggested to play an important role in the regulation of gene expression in response to abiotic stresses (Figure 9, Supplementary Table 7). To investigate the divergence information of the BrCDPK-SnRK gene family, the expression patterns following different treatments, including ABA, GA, NaCl, heat, cold, and PEG treatments, were analyzed (Figures 9, 10). The qRT-PCR results demonstrate that those BrCDPK-SnRK genes respond differentially to particular abiotic stresses. Sixteen percent of investigated BrCPK genes show increased expression levels upon GA at 6 h, while the other 84% of genes display downregulation or no significant changes. Meanwhile, 20% of genes were upregulated at 1 h and 6 h under GA treatment (Figures 9A, C). Two genes (BrCPK4 and BrCPK10) were induced under both ABA and GA treatment (Figure 9B). During four abiotic stress treatments, excluding PEG, 80% of BrCPK genes were highly responsive to cold, NaCl, and heat. We found that, with the exceptions of BrCPK2, 29, 23 and 26, all BrCPK genes were significantly induced in response to NaCl treatment (Figures 9D–F, H). Under heat treatment, we found that only two genes (BrCPK4 and BrCPK10) had
FIGURE 9 | Expression analysis of the BrCPK genes under six stress treatments in Brassica rapa. (A,C–F,H) Heat map representation the BrCPK genes under six stress treatments, namely, GA, ABA, NaCl, heat, PEG, and cold. (B,G) Venn diagram depicting the distribution of shared expression of the BrCPK genes among hormone and abiotic treatments.
no expression; other BrCPK genes were highly expressed in NaCl treatment. SnRK2 genes recognized as coding for enzymes involved in abiotic stress signal transduction in plants (Kulik et al., 2011). Therefore, we selected six BrSnRK2 genes; all of these genes had the highest expression under heat treatment (Figure 10). Furthermore, BrSnRK2.12 was induced under all stress treatments (Figure 10).

To investigate the connections between these genes, coregulatory networks were established based on the PCCs of stress-inducible BrCDPK-SnRK gene pairs (Supplementary Table 8, Figure 11). All the BrCDPK-SnRKs appeared to have different degrees of positive correlation. Next, 25 BrCPKs with PCC values that were significant at the 0.05 significance level and were greater than 0.8 were collected and visualized to construct hormone and abiotic stress coregulatory networks (Figure 11A). Six BrSnRKs and seven BrCPKs had positive significant correlations (Figure 11C). Most correlations occurred among members belonging to the same group, suggesting that the gene duplication not only led to functional divergence but also enhanced the cooperative interaction of homologs to help plants to adapt to their complex environment. In addition, based on predicted Chinese cabbage BrCDPK-SnRK superfamily protein interactions, their interactions form a complex network in both Arabidopsis thaliana and B. rapa. Moreover, to study the protein interactions of stress response genes between B. rapa and Arabidopsis, STRING 9.1 was utilized (Figure S9). Figure S9 shows four complex interaction networks of CDPK-SnRKS, providing an overall view of the relationship between B. rapa and Arabidopsis.

**DISCUSSION**

In eukaryotes, protein kinases are involved in regulating key aspects of cellular function, including cell division, metabolism, and responses to external signals. CDPK-SnRK plays an important role in stress signal transduction in plants, such as wounding, salt or drought stress (Botella et al., 1996; Patharkar and Cushman, 2000; Saijo et al., 2000), cold (Monroy and Dhindsa, 1995; Saijo et al., 2000), hormone treatment (Abo-El-Saad and Wu, 1995; Botella et al., 1996; Davletova et al., 2001), light (Frattini et al., 1999), and pathogens (Romeis et al., 2001; Murillo et al., 2002).

The gene balance hypothesis predicts that genes whose products participate in signaling networks or macromolecular complexes or are transcription factors are more likely to be retained (Birchler and Veitia, 2007; Aad et al., 2012). In this study, we identified 49 BrCPKs, 14 BrCRKs, 3 BrPPCKs, 5 BrPEPRKs, and 56 BrSnRKs in the B. rapa genome, and they contained a higher rate of copies than the B. rapa whole-genome level. This result suggests that these genes had a high degree of retention following WGD. By comparing the number of different duplicated types, gene copies, and distribution of the three subgenomes, we found that all the AtCRK, AtPEPRK, and...
AtPPCK orthologs were retained in B. rapa. In contrast, four AtCPK and two AtSnRK orthologs were completely lost. CRK, PEPRK, and PPCK were more preferentially retained than CPK and SnRK. CPK (66%) and SnRK (43%) had more than two copies retained in B. rapa, and more BrCPK (74%) and BrSnRK (70%) genes experienced segmental duplication. These preferentially retained CDPK-SnRK genes may have more important functions. At the same time, the important functions have been proved in previous researches.

For example, CDPK-SnRKs have been shown to have important roles in various physiological processes, including plant growth and development and abiotic and biotic stress responses in plants. The large number and variety of protein kinases that either have an EF-hand calcium-binding domain in their structure (CPK and CCaMK) or interact with EF-hand proteins (CCaMK, CaMK, CRK, and SnRK3) provides plants with a huge potential for interpreting specific calcium signals and for eliciting specific physiological responses. The SnRK1 kinase plays an important role in carbon-nitrogen interactions (Halford and Paul, 2004; Li et al., 2010) and in the transcription regulation of gene expression. In the developing tuber of potato, the expression level of SnRK1 was higher, lower in stem and lowest in leaf. The experiment on potato provided evidence for SnRK1 to regulate the transcription (Man et al., 1997). SnRK1 influence starch biosynthesis via regulating the expression of sucrose synthase and the activity of AGPase. SnRK1 activity can respond to sucrose appropriately (Halford and Paul, 2004). Antisense gene expressions in different plant indicate that SnRK1 has very important roles in plant growth and development processes (Halford et al., 2003). The SnRK2s are about 140–160 amino acids shorter than the SnRK1s, averaging about 40 kD in size, and have a characteristic patch of acidic amino acids in their C-terminal domains (Halford et al., 2000). SnRK2 genes, which are a significant part of the ABA signal pathway, are involved in many processes that help plants resistant environmental pressures (Wang et al., 2015). The variety is further amplified in the SnRK3 subfamily, since these kinases may interact with more than one CBL/SCaBP (Guo et al., 2001; Manns et al., 2001; Shi and Eberhart, 2001). Gene expression patterns can provide important clues for gene function. Therefore, the expression
patterns of CDPK-SnRK genes under stresses were identified based on qRT-PCR analysis. Overall, expression patterns of CDPK-SnRK genes are dynamic in different tissues during different developmental stages and in response to abiotic stresses, demonstrating their regulatory roles in various cellular processes in *B. rapa*. This finding is consistent with the gene dosage hypothesis.

During evolutionary history, all extant angiosperms genome have undergone at least one and often multiple polyploidization events (Edger and Pires, 2009; Jiao et al., 2011a; Ohno, 2013). *Brassica rapa* experienced a complex WGD history, including γ, α, and β events, and an additional WGT event, providing an excellent chance to study the relationship between gene family fractionation and changes in plant morphotypes (Wang et al., 2011; Cheng et al., 2013). An extensive phylogenetic analysis revealed the evolutionary history of those CDPK-SnRK genes in *B. rapa* and in other plants. In total, 127 CDPK-SnRK genes were grouped in seven classes in *B. rapa*. In addition, we identified CDPK-SnRK genes in 15 other plant species representing the major clades of terrestrial plants. These phylogenetic studies suggest that those identified CPK and SnRK genes were highly conserved in each class across a wide spectrum of plants, indicating their essential regulatory roles in the plant kingdom. For the CPK gene family, the number of genes grouped into each class, gene number and genome size are summarized in Figure 1. Furthermore, there is a difference among the five species in class I, II, III and IV in terms of gene numbers. It suggests that those four classes underwent extensive expansion in angiosperms. Class III was entirely lost during evolution in *V. vinifera*. For the SnRK gene family, two SnRK1 genes were found in *V. carteri*, suggesting that expansion of the SnRK gene family occurred after the divergence of green algae. Meanwhile, SnRK1, SnRK2, and SnRK3 were detected in *Physcomitrella patens*, indicating that the SnRK2 and SnRK3 gene subfamilies appear to be unique to plants, which consist with previous research (Hrabak et al., 2003). Overall, for the six species investigated, the number of CPK and SnRK genes remains different, indicating different evolutionary histories accompanied by rounds of WGD and subsequent gene losses/gains by natural selection constraints.

Most land plants have undergone polyploidization that led to WGD and provided opportunities for duplicated genes to diverge in different evolutionary ways. Each of these genes subsequently experienced one of three fates: subfunctionalization, neofunctionalization, or non-functionalization (deletion or pseudogenization). These fates provided opportunities for duplicated genes to gain functional diversification, resulting in more complex organisms. The Ka/Ks substitution rate ratio is an indicator of the selection history on genes or gene regions. Commonly, if the value of Ka/Ks is lower than 1, the duplicated gene pairs may have evolved from purifying selection (also called as negative selection); Ka/Ks = 1 means neutral selection, while Ka/Ks > 1 means positive selection. In this study, the Ka/Ks values for the duplicated gene pairs were small (<0.01); thus, it is likely that these pairs have been under purifying selection. Neofunctionalization is an adaptive process during which one copy of a duplicated gene mutates to adopt a novel function that cannot be performed by the ancestral sequence. This mechanism can lead to the retention of both copies over long periods of time. In addition, by comparing the tissue expression pattern among the CDPK-SnRK genes in *B. rapa* and *At. thaliana*, we found that most of the duplicated genes maintained similar expression patterns, but some of the duplicated CDPK-SnRK genes demonstrated higher tissue specialization and diversification, also suggesting that CDPK-SnRKS underwent more neofunctionalization and subfunctionalization. In previous reports, even the most recently diverged paralogs differed in their catalytic efficiency, expression, and/or substrate spectrum, suggesting that duplicates have a relatively high rate of divergence in function.

In this study, we analyzed the evolutionary pattern, gene duplication, and expression divergence of the CDPK-SnRK genes in plants. We conducted a comparative evolutionary analysis with the genome information of selected plants, and our research provides new insight into the evolutionary history of the CDPK-SnRK gene family in plants. For example, CPK had a closer relationship with CRK, indicating that these two may share a common evolutionary origin; and PEPRK appeared between the salicoid duplication and the γ event. By compared expression patterns among tissues, we found that the expansion of CDPK-SnRK genes seems to be associated with increasingly complex organs in the plants’ evolution. Under different stress treatments, their coregulatory networks demonstrated that the CDPK-SnRK genes enhanced the cooperative interaction of homologs to adapt to the environment. In conclusion, this study provides useful resources for functional divergence and conservation in the CDPK-SnRK gene superfamily and facilitates our understanding of the effect of polyploidy during the evolution of CDPK-SnRK genes.

**METHODS**

**Retrieval of Sequences**

The *B. rapa* sequences were downloaded from BRAD (http://brassicadb.org/brad/) (Wang et al., 2011). The *Arabidopsis* sequences were retrieved from TAIR (http://www.arabidopsis.org/), and the sequences of rice were retrieved from RGAP (http://rice.plantbiology.msu.edu/; Kawahara et al., 2013). The gene information on *A. trichopoda* genes was retrieved from the Amborella Genome Database (http://www.amborella.org/). The others 14 species’ gene information was downloaded from Phytozome v9.1 (http://www.phytozome.net/; Goodstein et al., 2012). The homologous CDPK-SnRK genes in other species were identified through comparison with *Arabidopsis*. First, BLASTP searches were performed against the rice protein sequences using an E-value threshold of $1 \times 10^{-10}$. The top-ranked rice hit was used for BLASTP searches of the *Arabidopsis* proteins to confirm homologies. Starting with both *Arabidopsis* and rice homologs, BLASTP searches were performed against the proteins of other species ($e < 1 \times 10^{-10}$, identity >70%). These potential sequences were analyzed using the tool SMART (http://smart.

Identification of Gene Synteny and Duplicated CDPK-SnRK Genes

BLAST and the Multiple Collinearity Scan toolkit (MCScanX) were used for gene synteny analysis according to previous reports (Wang et al., 2012). An all-against-all BLASTP comparison provided the pairwise gene information and the P-value for a primary clustering. Paired segments were extended by identifying the clustered genes using dynamic programming. The potentially duplicated genes were also identified using MCScanX. The positions of B. rapa CDPK-SnRK genes on the blocks were verified by searching for homologous genes between A. thaliana and three subgenomes of B. rapa (LF, MF1, and MF2) in BRAD (http://brassicadb.org/brad/searchSynteny.php; Wang et al., 2011). The syntenic diagram was drawn using Circos software (Krzywinski et al., 2009).

Phylogenetic Analysis and Characterization of the CDPK-SnRK Gene Family

Phylogenetic analyses were conducted using MEGA v5.073; maximum-likelihood (ML) trees were constructed with a bootstrap value of 1,000 replications to assess the reliability of the resulting trees. The genetic distance used in this study was also calculated with MEGA v5.0. To identify conserved motifs in the complete amino acid sequences of CDPK-SnRK proteins, we used MEME software (http://meme.sdsc.edu/meme/; Bailey et al., 2009). We researched the gene structure using Gene Structure Display (GSDS, http://gsds.cbi.pku.edu.cn/). The interaction network of CDPK-SnRK proteins in Chinese cabbage was constructed with STRING software (Search Tool for the Retrieval of Interacting Genes/Proteins, http://string-db.org/; Yamada et al., 2011).

Calculation of the Ka/Ks and Dating of the Duplication Events

The duplicated CDPK-SnRK genes from B. rapa were aligned using MUSCLE (Edgar, 2004). The protein alignments were translated into coding sequence alignments using an in-house Perl script. Ks (synonymous substitution rate) and Ka (non-synonymous substitution rate) values were calculated based on the coding sequence alignments using the method of Nei and Gojobori as implemented in KaKs_calculator (Zhang et al., 2006). The Ks values of all duplicated BrCDPK-SnRK genes were then plotted as the density and a boxplot using the R program (Ihaka and Gentleman, 1996). The divergence time was calculated with the formula $T = Ks/2r$, with Ks being the synonymous substitutions per site and r being the rate of divergence for nuclear genes from plants. The value of r was taken to be $1.5 \times 10^{-8}$ synonymous substitutions per site per year for dicotyledonous plants (Koch et al., 2000).

Expression Pattern Analysis for CDPK-SnRK Genes

For expression profiling of the CDPK-SnRK genes in B. rapa, we utilized the Illumina RNA-seq data that were previously generated and analyzed by Tong et al. (2013). Six tissues of B. rapa accession Chiifu-401-42, including callus, root, stem, leaf, flower, and silique, were analyzed. The transcript abundance was expressed as fragments per kilobase of exon model per million mapped reads (FPKM). The gene expression patterns of each tissue were analyzed using Cluster 3.0, and the expression values were log2 transformed. Finally, heat maps of hierarchical clustering were visualized using Tree View (http://jtreeview.sourceforge.net/). A. thaliana development expression profiling was analyzed using the AtGenExpress Visualization Tool (AVT; http://jsp.weigelworld.org/expviz/expviz.jsp) with mean-normalized values (Schmid et al., 2005). Venn diagrams were drawn using the R program.

Plant Material and Treatments

Chinese cabbage (Chiifu-401–42) was used for the experiments. The germinated seeds were grown in plastic pots in a 3:1 soil-vermiculite mixture, and the artificial growth conditions were set at 24/16°C, with a photoperiod of 16/8 h for day/night and a relative humidity of 65–70%. Four weeks later, seedlings were subjected to various treatments. For heat and cold treatment, the pots were exposed to 38°C or 4°C; the other growth conditions were same as described earlier. Meanwhile, plants were cultured with the following four treatments: (1) control; (2) 100 µM ABA; (3) 100 µM GA; (4) (w/v) polyethylene glycol (PEG) 6,000, and (5) 250 mM NaCl. All these treatments were performed over a continuous time course (0, 1, 6, 12 h). Each treatment consisted of three biological replicates. All of the samples were frozen in liquid nitrogen and stored at −70°C for RNA preparation.

RNA Isolation and qRT-PCR Analyses

Total RNA was isolated from treated frozen leaves using Trizol (Invitrogen, San Diego, CA, USA) according to the manufacturer’s instructions. Specific primers used for qRT-PCR were designed by Beacon Designer 7 and are shown in Supplementary Table 9. To check the specificity of the primers, we used BLAST against the Brassica genome. The qRT-PCR assays were performed with three biological and technical replicates. The reactions were performed using a StepOnePlus Real-Time PCR System (Applied Biosystems, Carlsbad, CA, USA). qRT-PCR was performed according to a previous report (Song et al., 2013). Relative fold expression changes were calculated using the comparative Ct-value method.

Pearson Correlation Analyses

Pearson correlation coefficients (PCCs) of stress-inducible CDPK-SnRK gene pairs were calculated using an in-house Perl script based on log2-transformed quantitative real-time (qRT)-PCR data (Tang et al., 2013). All gene pairs whose PCC was more than 0.8 and was significant at
the 0.05 significance level (P-value) were collected for a gene coregulatory network analysis. The coexpression networks were graphically visualized using Cytoscape based on the PCCs of these gene pairs (Shannon et al., 2003).

**AUTHOR CONTRIBUTIONS**

PW was responsible for the experimental design, data analysis, and manuscript writing. WD, WW, YL contributed to the experimental work. XH contributed to the interpretation of results and coordinated the study. All authors read and approved the final manuscript.

**REFERENCES**


Asano, T., Kunieda, N., Omura, Y., Ibe, H., Kawasaki, T., Takano, M., et al. (2002). Arabidopsis CDPK-SnRK coregulatory network analysis. The coexpression results and coordinated the study. All authors read and approved the final manuscript.

**ACKNOWLEDGMENTS**

This work was supported by the National Natural Science Foundation of China (No. 31330067, 31301782), the Science and Technology Pillar Program of Jiangsu Province (No. BE2013429), and the Agricultural Science and Technology Independent Innovation Funds of Jiangsu Province [CX(13)2006].

**SUPPLEMENTARY MATERIAL**

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2017.00162/full#supplementary-material


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

Copyright © 2017 Wu, Wang, Duan, Li and Hou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.
Expression of the \textit{KNOTTED HOMEOBOX} Genes in the Cactaceae Cambial Zone Suggests Their Involvement in Wood Development

Jorge Reyes-Rivera\textsuperscript{1}, Gustavo Rodríguez-Alonso\textsuperscript{2}, Emilio Petrone\textsuperscript{1}, Alejandra Vasco\textsuperscript{1}, Francisco Vergara-Silva\textsuperscript{3}, Svetlana Shishkova\textsuperscript{2} and Teresa Terrazas\textsuperscript{1}\textsuperscript{*}

\textsuperscript{1} Departamento de Botánica, Instituto de Biología, Universidad Nacional Autónoma de México, Mexico City, Mexico
\textsuperscript{2} Departamento de Biología Molecular de Plantas, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Mexico
\textsuperscript{3} Jardín Botánico, Instituto de Biología, Universidad Nacional Autónoma de México, Mexico City, Mexico

The vascular cambium is a lateral meristem that produces secondary xylem (i.e., wood) and phloem. Different Cactaceae species develop different types of secondary xylem; however, little is known about the mechanisms underlying wood formation in the Cactaceae. The \textit{KNOTTED HOMEOBOX} (\textit{KNOX}) gene family encodes transcription factors that regulate plant development. The role of class I \textit{KNOX} genes in the regulation of the shoot apical meristem, inflorescence architecture, and secondary growth is established in a few model species, while the functions of class II \textit{KNOX} genes are less well understood, although the \textit{Arabidopsis thaliana} class II \textit{KNOX} protein KNAT7 is known to regulate secondary cell wall biosynthesis. To explore the involvement of the \textit{KNOX} genes in the enormous variability of wood in Cactaceae, we identified orthologous genes expressed in species with fibrous (\textit{Pereskia lychnidiflora} and \textit{Pilosocereus alensis}), non-fibrous (\textit{Ariocarpus retusus}), and dimorphic (\textit{Ferocactus pilosus}) wood. Both class I and class II \textit{KNOX} genes were expressed in the cactus cambial zone, including one or two class I paralogs of \textit{KNAT1}, as well as one or two class II paralogs of \textit{KNAT3-KNAT4-KNAT5}. While the \textit{KNOX} gene \textit{SHOOTMERISTEMLESS} (\textit{STM}) and its ortholog \textit{ARK1} are expressed during secondary growth in the \textit{Arabidopsis} and \textit{Populus} stem, respectively, we did not find \textit{STM} orthologs in the Cactaceae cambial zone, which suggests possible differences in the vascular cambium genetic regulatory network in these species. Importantly, while two class II \textit{KNOX} paralogs from the \textit{KNAT7} clade were expressed in the cambial zone of \textit{A. retusus} and \textit{F. pilosus}, we did not detect \textit{KNAT7} ortholog expression in the cambial zone of \textit{P. lychnidiflora}. Differences in the transcriptional repressor activity of secondary cell wall biosynthesis by the \textit{KNAT7} orthologs could therefore explain the differences in wood development in the cactus species.

\textbf{Keywords}: Cactaceae, dimorphic wood, fibrous wood, \textit{KNAT7} ortholog, \textit{KNOX} transcription factors, non-fibrous wood, vascular cambium, wood lignification
INTRODUCTION

As reservoirs of pluripotent cells, meristems have played a leading role in the diversification of angiosperm growth forms. The shoot and root apical meristems maintain primary growth in plants, while the lateral meristems, comprised of the vascular cambium and cork cambium in Eudicotyledons and gymnosperms, are involved in secondary growth (Évert, 2006). The vascular cambium maintains a population of initial (stem) cells, which divide asymmetrically to generate two daughter cells; one maintains the cambial initial identity, while the other divides again and the daughters differentiate to generate secondary phloem or xylem (Groover et al., 2010; Miyashima et al., 2013). Vascular cambium derivatives are thought to have influenced speciation and diversification events (Spicer and Groover, 2010; Lucas et al., 2013). In the Cactaceae, the traits of the secondary xylem (i.e., wood) suggest that it has evolved by heterochronous processes where a change in the timing of developmental processes leads to morphological differences between species (Altesor et al., 1994). The larger species (≥1.5 m in height) in this family of succulent plants have fibrous wood with vessel elements in the xylem similar to those typically derived from vascular cambium, with a similar wood chemical composition to that of other woody dicot species (Reyes-Rivera et al., 2015). On the other hand, the wood found in smaller species is generally scarce and non-fibrous, with abundant wide-band tracheids and vessel elements similar to those typical of proto- and metaxylem. In these species, the level of wood lignification is insignificant and the lignin has a heterogeneous chemical composition (Reyes-Rivera et al., 2015). Many species of Cactaceae have dimorphic wood, where one type of wood is produced in the juvenile stages and a different structure is formed in the adult stages of development. To the best of our knowledge, this phenomenon is unique among Eudicotyledons and is related to the globose and globose-depressed growth forms. The shoot and root apical meristems maintain primary growth in plants, while the lateral meristems, comprised of the vascular cambium and its derivatives in the Cactaceae are well established. Later in development, genes regulating cell expansion, secondary xylem differentiation, lignification, and secondary wall deposition contribute to wood formation (reviewed in Ruizička et al., 2015; Ye and Zhong, 2015; Zhong and Ye, 2015). Transcription profiling of the vascular cambium in aspen (Populus tremula) uncovered similarities between the gene regulatory networks operating in the shoot apical meristem and the vascular cambium. In particular, it was reported that four members of the KNOX gene family, namely PttKNOX1, PttSHOOT MERISTEMLESS (PttSTM), PttKNOX2, and PttKNOX6, are highly expressed in both tissues (Schrader et al., 2004). Functional analysis suggests that the Populus orthologs of Arabidopsis thaliana (Arabidopsis) class I KNOX genes STM and KNOTTED-LIKE HOMEBOX OF ARABIDOPSIS THALIANA I (KNAT1)/BREVIPEDICELLUS (BP), PttSTM/ARBORKNOX 1 (ARK1)/ARK1a and PttKNOX1/ARK2, respectively, regulate secondary growth (Groover et al., 2006; Du et al., 2009; Liu et al., 2015). While the roles of class I KNOX genes in the regulation of shoot apical meristem, inflorescence architecture, and compound leaf development are well established, the functions of class II KNOX genes are less well understood. In general, class I transcripts are less abundant than class II KNOX transcripts, and are expressed in specific regions of meristems, particularly in the shoot apical meristem and leaf meristems. By contrast, class II transcripts are found in differentiating cells and all mature plant organs (Serikawa et al., 1997; reviewed by Hay and Tsiantis, 2010; Arnaud and Pautot, 2014), but not in the shoot apical meristem (Furumizu et al., 2015). Moreover, the dark-green serrated leaf phenotype of the class II knat3 knat4 knat5 triple loss-of-function mutant in Arabidopsis is similar to that of the class I gain-of-function mutants (Furumizu et al., 2015), suggesting opposing functions for genes of the two classes. Here, we report that class I and class II KNOX genes are expressed in the cambial zone, consisting of vascular cambium and derived cells, of Cactaceae species with fibrous (Pereskia lychnidiflora and Pilosocereus alensis), non-fibrous (Ariocarpus retusus), and dimorphic (Ferocactus pilosus) wood. We also present the phylogeny of class I and class II KNOX proteins encoded in the sequenced plant genomes retrieved from the Phytozome database, confirming monophyly of class I and class II KNOX proteins, assigning the Cactaceae KNOX proteins into clades, and exploring the number of paralogs of the plant species in every KNOX clade.

MATERIALS AND METHODS

Plant Material and Tissue Sampling

Four Cactaceae species with different wood types were used in this study: Pereskia lychnidiflora DC. (subfamily: Pereskioideae); Pilosocereus alensis (F. A. C. Weber) Byles & G. D. Rowley (subfamily: Cactoideae, tribe: Cereae); Ariocarpus retusus Scheidw (Cactoideae, tribe: Cactaeae); and Ferocactus pilosus (Galeotti ex Salm-Dyck) Werderm. (Cactoideae, Cactaeae). Samples were collected from the cambial zone of adult individuals growing in Mexico. For the tall species (F. pilosus, P. lychnidiflora), the cambial zone was harvested from one individual of both species in the field (F. pilosus at the Mexican Plateau, San Luis Potosí state, location 23°19’31”N; 100°33’30”W, voucher TT890 MEXU; and P. lychnidiflora at the Isthmus of Tehuantepec, Oaxaca state, location 16°22’52”N; 95°19’03”W, voucher TT966 MEXU), with samples collected during the rainy season to ensure that the vascular cambium
was active. All surrounding tissues including the cortex were removed as described by Liu et al. (2014). The cambial zone was peeled off with a disposable microtome knife and immediately frozen in liquid nitrogen. One whole individual of both *A. retusus* and *P. alensis* were placed in pots containing their native soil and transported from their natural habitat to the laboratory (*A. retusus* from the Mexican Plateau, Nuevo León state, location 23°23′39″N; 100°21′57″W, voucher SA1976 MEXU; and *P. alensis* from the Pacific coast, Michoacán state, location 18°46′41″N; 103°08′05″W, voucher JRR3 MEXU). Species identification was confirmed by S. Arias [Instituto de Biología, Universidad Nacional Autónoma de México (IB, UNAM)], the leading taxonomy specialist of the Cactaceae family in Mexico. The cambial zone was harvested as described above. The cambial zone samples were stored at −80°C until RNA extraction. For expression analysis by RT-PCR, one adult plant of *A. retusus*, *F. pilosus*, and *P. lychnidiflora*, one 2-year-old juvenile plant of the first two species, and two 5-year-old young individuals of *A. retusus*, previously collected from the same locations, were placed in pots containing their native soil and donated by the Botanical Garden IB, UNAM. For the root tip samples and the fibrous cambium (mature wood), and each region was processed separately. The non-fibrous cambium (mature wood), and each region was processed separately. The non-fibrous *A. retusus* samples were placed in 2-mL microcentrifuge tubes filled with Franklin solution (5:1:4 acetic acid:hydrogen peroxide:water; Ruzin, 1999), while the dimorphous *F. pilosus* samples and the fibrous *P. alensis* and *P. lychnidiflora* were placed into tubes containing Jeffrey solution (equal volumes of 10% chromic acid and 10% nitric acid; Berlyn and Miksche, 1976). For each species, 0.2 g of the tissue was used. The samples were then incubated at 56°C for 4 h (non-fibrous wood) and 24 h (dimorphic and fibrous wood). Additionally, all samples were sonicated in a Branson 200 ultrasonic cleaner (Branson Ultrasonics, Danbury, CT, USA) until completely macerated, washed with water, and dehydrated using a series of ethanol washes at 50, 70, 90%, and absolute concentrations. The *A. retusus* macerations were stained with a 0.1% aqueous solution of toluidine blue for 12 h (Ruzin, 1999) and mounted onto slides. The macerations of the wood samples from another three species were stained with Safranin for 2 h and mounted onto slides using synthetic resin (Entellan, Merck Millipore, Darmstadt, Germany). All wood elements were photographed using a BX51 optical microscope (Olympus Corporation, Tokyo, Japan) and the images were analyzed using Image-Pro Plus v. 6.1 software (Media Cybernetics, Inc., Bethesda, MD, USA).

**RNA Extraction, cDNA Synthesis, and PCR Amplification**

Total RNA was extracted using TRIzol Reagent (Invitrogen, Carlsbad, CA, USA), according to the manufacturer’s protocol, including the optional step of centrifugation before the separation of phases, or using the Spectrum Plant Total RNA Isolation Kit (Sigma–Aldrich, St. Louis, MO, USA). The cDNA was synthesized using SuperScript II Reverse Transcriptase (Invitrogen), according to the manufacturer’s instructions. Degenerate PCR primers were used to amplify the KNOX genes. Primers for amplifying putative cacti orthologs of KNOT3 were designed (KNAT3_Cact_F: 5′-GAGAGRAATAATGGCWTATCATC-3′ and KNAT3_Cact_R: 5′-CTTCTGGGTCTACTTCCCTC-3′) based on the alignments of nucleotide sequences encoding the class II KNOX proteins KNAT3 (*Arabidopsis*) and its closest homologs from *Beta vulgaris* and *Pachycereus pringlei*. For the Cactaceae orthologs of KNAT1, primers designed by Du et al. (2009) were used. PCR products were purified with Sepahex Centri-sep columns (Thermo Fisher Scientific, Waltham, MA, USA) as instructed by the manufacturer. The amplified and purified products were sequenced in a 3500xL Genetic Analyzer sequencer (Applied Biosystems, Foster City, CA, USA) using the PCR primers. Platinum Taq polymerase (Thermo Fisher Scientific, Waltham, MA, USA) was used for PCR reactions. Primers used for RT-PCR are listed in **Supplementary Table S1**. RNA-seq was performed at the Beijing Genome Institute, Hong Kong; the vascular cambium and root tip transciptomes were *de novo* assembled using Trinity v. 2.21 and CLC Genomic Workbench v. 7.5 (Qiagen3).

**Sequence Alignment and Phylogenetic Analysis**

KNOX-like protein sequences were retrieved from the Phytozome database4 (v. 11; last accessed on May 9, 2016) using the PhytoMine tools. All proteins with KNOX (IPR005540 [InterProscan definition]/PF03790 [ Pfam definition] and IPR005541/ PF03791), ELK (IPR005539/PF03789), and HD domains (IPR009057/PF05920) were retrieved. *B. vulgaris* sequences were retrieved using tBLASTn from the ReFBeet v. 1.2 genome assembly (The *Beta vulgaris* resource4, Dohm et al., 2013) using a BLOSUM80 substitution matrix (Henikoff and Henikoff, 1992; at the time, *B. vulgaris* was the only species from the order Caryophyllales with a sequenced genome). Chimeric sequences and those from the early release genomes were discarded. After that, KNOX protein sequences from the tree species *Betula luminifera* and *Juglans nigra* were added. For Cactaceae species, the KNOX protein sequences used were deduced from the amplified and sequenced PCR fragments (see previous section), from our RNA-seq data and the *de novo* assembly of the cambial zone transcriptome of the four species reported in this work (the same RNA samples were used as starting material for the cambial zone transcriptome assembly), from the root tip transcriptome of the Cactoideae subfamily species *Pachycereus pringlei* (the analysis of the transcriptomes will be reported elsewhere), and from the recently published *Lophophora williamsii* transcriptome (Ibarra-Laclette et al., 2015). The resulting sequences were

---

1. [https://github.com/trinityrnaseq/trinityrnaseq/releases](https://github.com/trinityrnaseq/trinityrnaseq/releases)
3. [http://phytozome.jgi.doe.gov](http://phytozome.jgi.doe.gov)
To explore the involvement of the Cactaceae Cambial Zone KNOX Family Genes Are Expressed in the Cactaceae Cambial Zone, which comprises the vascular cambium and recently derived cells. RT-PCR using degenerate primers for the class I KNOX genes resulted in a major amplification product of the predicted molecular weight in the four Cactaceae species studied. The amino acid sequences inferred from the amplified and sequenced fragments of these genes (ArKNOX1a from Ariocarpus retusus, FpKNOX1a from Ferocactus pilosus, PaKNOX1a from Pilosocereus alensis, and PlKNOX1a from Pereskia lychnidiflora) correspond to nearly 70% of the ARK2 and KNAT1 protein sequences of Populus and Arabidopsis, respectively, including part of the KNOX1 domain and the entirety of the KNOX2, ELK, and HD domains [Figure 2 (The last letters in the gene names refer to: “a,” amplification as a method of gene isolation; “e,” the inferred coding sequence extended after alignment with sequences resulting from the RNA-seq; and “r,” RNA-seq as a part of the KNOX1 domain and the entirety of the KNOX2, ELK, and HD domains)].

Results

Morphological Analysis of Xylem Cells
To examine the detailed cellular features of the secondary xylem in the four species studied, we performed a morphological analysis of xylem cells using wood macerations. The young xylem of A. retusus with non-fibrous wood and F. pilosus with dimorphic wood forms wide-band tracheids and vessel elements with annular and helical secondary wall patterns formed during early development. The mature fibers and vessel elements with pseudocalariform and alternate pitting were formed in adult plants during late development. No differences were observed between young and mature wood of P. alensis and P. lychnidiflora, with both having fibrous wood. The tracheary elements of A. retusus and young F. pilosus had a low proportion of lignified cell wall, similar to those of proto- and metaxylem (i.e., primary xylem), whereas vessel elements and fibers of mature F. pilosus, P. alensis, and P. lychnidiflora, had a higher proportion of lignified cell walls. This analysis allowed the unequivocal identification of each secondary xylem cell type and its cell wall features (Figure 1) and thus confirmed the wood types we previously identified by RT-PCR with degenerate primers from the three species of the Cactoideae subfamily, but were not isolated from P. lychnidiflora. The amino acid sequences deduced from the amplified and sequenced fragments of these three genes, ArKNOX3a, FpKNOX3a, and PaKNOX3a, cover nearly 95% of the KNAT3 protein sequence, including the KNOX1, KNOX2, and ELK domains, as well as most of the HD domain (Figure 3).

In addition to the seven transcripts mentioned above, eight more KNOX transcripts were de novo assembled from our RNA-seq data of the cambial zone samples of three Cactaceae species, A. retusus, F. pilosus, and P. lychnidiflora. The amino acid sequences deduced from the amplified transcripts of F. pilosus were later extended by alignment with the assembled transcripts, and were therefore renamed FpKNOX1ae and FpKNOX3ae (Figures 2–5). Expression of all KNOX transcripts in the cambial zone of A. retusus, F. pilosus, and P. lychnidiflora was confirmed by RT-PCR (Supplementary Figure S1). Remarkably, in the cambial zone of 2-year-old juvenile plants of both A. retusus and F. pilosus, one of the two orthologs of KNOX3, namely, ArKNOX3a and FpKNOX3ae, and of KNOX7, namely, ArKNOX7r1 and FpKNOX7r2, were expressed. No expression of KNOX1 orthologs was detected in young plants. At this stage, only very scarce accumulation of wood, represented by wide-band tracheids and vessel elements, was detected in these species. From the six KNOX genes expressed in the vascular cambium of adult A. retusus plant, expression of only ArKNOX7r1 was detected in the tubercle of 5-year-old plant (data not shown).

Phylogenetic Analysis of the KNOX Family
Initially, 524 KNOX-like protein sequences were retrieved from the Phytozome database. Chimeric sequences and those from draft genome releases were filtered out, and the inferred protein sequences of Cactaceae, the trees B. luminifera and J. nigra, as well as those retrieved from the B. vulgaris genome, were included. The resulting matrix contained 478 aligned protein sequences belonging to 45 species with sequenced genomes, six Cactaceae species, and two species of tree. The identifiers of the sequences retrieved from the Phytozome database are listed in Supplementary Table S2. The maximum likelihood phylogeny
grouped the KNOX proteins into five main clades (Figures 4, 5), which were named according to their putative Arabidopsis ortholog: i.e., the KNAT1 clade (gray), KNAT2-KNAT6 clade (cyan), STM clade (pink), KNAT3-KNAT4-KNAT5 clade (blue), and KNAT7 clade (green). The first three clades belong to the class I KNOX and the last two belong to the class II KNOX proteins. Figure 5 depicts the phylogenetic relationships of the subset of KNOX proteins from Arabidopsis, the Cactaceae, and B. vulgaris. Arabidopsis was used as the reference species, with B. vulgaris from order Caryophyllales, to which the Cactaceae family belongs, included as the closest sister taxon with a sequenced genome.

The molecular phylogenetic analysis confirmed that the four Cactaceae class 1 KNOX genes amplified from the cambial zone by RT-PCR with degenerate primers (ArKNOX1a, FpKNOX1ae, PaKNOX1a, and PlKNOX1a) are part of the KNAT1 clade (Figures 4, 5); moreover, two of the de novo assembled transcripts, ArKNOX1r and PlKNOX1r, also belonged to the KNAT1 clade. The class II ArKNOX3a, FpKNOX3ae, and PaKNOX3a genes expressed in the vascular zone of the three species fall into the KNAT3-KNAT4-KNAT5 clade, along with two other class II KNOX genes, ArKNOX3r and PlKNOX3r (the only Pereskia class II gene identified in this study). Four de novo assembled transcripts, ArKNOX7r1, ArKNOX7r2, FpKNOX7r1, and FpKNOX7r2, belong to the KNAT7 clade (Figures 4, 5 and Table 1). The identity and similarity percentages for the proteins encoded by the identified Cactaceae genes are shown in Supplementary Figure S2.

Of the five KNOX transcripts detected in the P. pringlei root tip, both of the class I KNOX genes were found in the KNAT2-KNAT6 clade, one class II sequence was attributed to the KNAT3-KNAT4-KNAT5 clade, and two more belonged to the KNAT7 clade. Among the five KNOX genes identified in the recently published L. williamsii transcriptome
(Ibarra-Laclette et al., 2015), one was found in each of the five clades (Figures 4, 5 and Table 1). In the global KNOX phylogeny (Figure 4), B. vulgaris sequences were always resolved as sister to the Cactaceae sequences. In the phylogeny for the subset of KNOX proteins (Figure 5), B. vulgaris class I KNOX proteins were resolved as sister to the Cactaceae sequences in all three clades, while this was not the case for two clades of class II KNOX. The only B. vulgaris paralog of KNAT3, KNAT4, and KNAT5 fell in the subclade of the A. thaliana sequences. Within the KNAT7 clade, the B. vulgaris sequence represented the sister sequence for the Cactaceae KNOX7r2 subclade, whereas the KNOX7r1 sequences grouped as a separate subclade. The possible reasons of this small inconsistency could be the incomplete sequences of many Cactaceae proteins, as well as the very restricted number of sequences in this phylogeny (Figure 5).

DISCUSSION

Although, it is well established that some KNOX transcription factors are important for the maintenance of the shoot apical meristem, their role in vascular cambium activity and secondary growth is less well understood. Four class I KNOX genes were reported to be expressed in the vascular cambium of P. tremula (Schrader et al., 2004; Du et al., 2009), while the expression of a class II KNOX gene was detected in the vascular cambium of J. nigra (Huang et al., 2009). It was therefore of particular interest to determine whether genes of both classes are expressed in the vascular zones of cacti, and whether different paralogs are expressed in species producing different types of wood. By performing RT-PCR with degenerate primers in the four Cactaceae species (A. retusus, F. pilosus, P. lychnidiflora, and P. alensis) and RNA-seq with subsequent de novo transcriptome assembly for three of these species, we have found, and consequently confirmed by RT-PCR (Supplementary Figure S1), that the transcripts of both class I and class II KNOX genes are present in the cambial zone of the adult individuals in all four species (Table 1). The possibility exists that other KNOX genes could be expressed at very low levels in the Cactaceae vascular zone, impeding successful de novo assembly. Genome sequencing of species from the Cactaceae family will facilitate a more detailed analysis of the KNOX expression patterns. To annotate the Cactaceae KNOX sequences, a phylogenetic tree was constructed.
that included KNOX sequences from sequenced angiosperm genomes from the Phytozome database and from the B. vulgaris genome as a species from the order Caryophyllales, to which the Cactaceae family belongs. Expression of all KNOX genes in the cambial zone of adult A. retusus, F. pilosus, and P. lychnidiflora plants was confirmed by RT-PCR (Supplementary Figure S1). Meanwhile, in the tubercle of 5-year-old A. retusus plant, we detected expression of only one of the six ArKNOX genes reported in this work, ArKNOX7r1. Expression of two KNOX genes, the orthologous ArKNOX7r1/FpKNOX7r1 and ArKNOX3a/FpKNOX3ae, was detected in the vascular cambium of juvenile A. retusus and F. pilosus plants, when they were just starting to accumulate wood. This finding further suggests that the KNOX genes, reported in this work, are involved in the formation of mature wood, and that the KNOX7r1 paralog has broader functions in Cactaceae development.
FIGURE 4 | Maximum likelihood phylogenetic tree of KNOX proteins. Phylogeny of the complete dataset, incorporating the deduced proteins of the five Cactaceae species analyzed in this study; the deduced proteins from the 44 sequenced plant genomes retrieved from the Phytozome database and the Beta vulgaris genome; sequences from Betula luminifera and Juglans nigra trees; and proteins deduced from the Lophophora williamsii transcriptome (Ibarra-Laclette et al., 2015). The gene identifiers for each sequence are listed in Supplementary Table S2. The branches for the Cactaceae and Beta vulgaris sequences are represented in green and indicated with green ribbons. The sequences from monocots (Poaceae and Musaceae) are indicated by brown ribbons. The BEL1 protein from Arabidopsis thaliana was used to root the tree and is indicated in red. The KNOX protein from the unicellular green alga Ostreococcus lucimarinus at the base of the tree is also indicated in red. The five main clades are named according to their putative Arabidopsis ortholog, indicated by a black arrowhead. The class I KNOX proteins belong to the KNAT1 (gray), KNAT2-KNAT6 (cyan), and STM (pink) clades. The class II KNOX proteins belong to the KNAT3-KNAT4-KNAT5 (blue) and KNAT7 (green) clades. The scale bar depicts the number of substitutions per sequence.

The six class I KNOX genes found to be expressed in the cambial zone of the four cactus species are putative orthologs of the Arabidopsis BP/KNAT1 gene, as they belong to the KNAT1 clade. The Populus gene ARK2/PtKNOX1, which is expressed in the vascular cambium and developing xylem (Schrader et al., 2004; Du et al., 2009), also belongs to this clade (white arrowhead in Figure 4). Previous work has suggested that ARK2 is involved in vascular cambium activity and secondary growth; Populus plants constitutively overexpressing ARK2 had a wider cambial zone and decreased differentiation of the lignified secondary xylem, while the downregulation of ARK2 resulted in the early differentiation of lignified secondary xylem cells (Du et al., 2009).

Importantly, we did not find putative orthologs of the Arabidopsis STM and KNAT2-KNAT6 genes expressed in the Cactaceae cambial zone. By contrast, both the ARK1a and ARK1b paralogs in Populus (orthologous to STM) are shown to be expressed in the vascular cambium (Popgenie database8). ARK1 overexpression in a hybrid Populus resulted in pleiotropic phenotypes, including the slower differentiation of cambium-derived cells (Groover et al., 2006). While the KNAT1 ortholog ARK2 is expressed in both the vascular cambium and the developing xylem of Populus, ARK1a was shown to be

8http://popgenie.org/
FIGURE 5 | Phylogenetic tree of the selected dataset of KNOX proteins from Arabidopsis, the Cactaceae, and Beta vulgaris. The class I KNOX proteins belong to the KNAT1 (gray), KNAT2-KNAT6 (cyan), and STM (pink) clades. The class II KNOX proteins belong to the KNAT3-KNAT4-KNAT5 (blue) and KNAT7 (green) clades. The domain positions for Cactaceae sequences are displayed on the right alongside the KNAT ortholog. The ELK domain in KNAT7 proteins is recognized by InterProscan but not by Pfam tools; therefore, the InterProscan database was used for the domain definition. For FpKNOX1ae and FpKNOX3ae, the portion of the original amplified sequence is shown below the assembled sequence. Note that it was not possible to obtain the full KNOX1 domain sequence for the Cactaceae sequences ending in “a”; therefore, their phylogenetic position might be slightly changed when the complete protein sequence is available. Bootstrap values from 1000 trees are displayed at the nodes. The scale depicts the number of substitutions per sequence.

downregulated in non-meristematic secondary vascular tissues (Groover et al., 2006), thus showing more similarity to the STM expression pattern in the shoot apex. STM expression was previously also detected in the Arabidopsis inflorescence stems during induced secondary growth (Ko and Han, 2004). As expected, we did not find a STM ortholog in the P. pringlei root tip transcriptome, while two P. pringlei paralogs orthologous to KNAT2-KNAT6 were expressed in the root tip (Figures 4, 5 and Table 1). The STM putative ortholog was present in the L. williamsii shoot and root transcriptome, suggesting that STM might be expressed in some Cactaceae tissues, most probably in the shoot apical meristem. The absence of putative STM orthologs in the cambial zone transcriptome suggests possible differences in the vascular cambium genetic regulatory network in the Cactaceae versus Arabidopsis and Populus. Interestingly, monocots do not have orthologs of Arabidopsis STM; the grass genes involved in shoot apical meristem maintenance (maize: KNOTTED1 and ROUGH SHEATH1; rice: ORYZA SATIVA HOMEobox1 (OSH1) and OSH15; Tsuda et al., 2011; Bolduc et al., 2014) belong to the KNAT1 clade (Figure 4). Moreover, monocot KNOX genes clustered as a single subclade within the KNAT1, KNAT2-KNAT6, and KNAT7 clades, while within the KNAT3-KNAT4-KNAT5 clade two subclades, one from Poaceae and another from Musaceae, were present (brown ribbons on Figure 4).

For the three species used for RNA-seq and the de novo transcriptome assembly, two class I paralogs were identified in the vascular zone of A. retusus and P. lychnidiflora, while only one was identified in F. pilosus (Figures 2, 4, 5 and Table 1). The missing class I KNOX paralog could have been lost from F. pilosus as an evolutionary developmental process, enabling speciation after a gene duplication in the Cactaceae, a phenomenon that has been well documented in other families. Using data from sequenced genomes (Supplementary Table S2), we showed that the number of class I and class II KNOX genes in a moss, a lycophyte, and 43 angiosperm species varies significantly, particularly between Eudicotyledonous species (Figure 6). As there is still no Cactaceae species with a sequenced
TABLE 1 | Number of KNOX genes from this study expressed in the cambial zone, root tip, and shoot/root of Cactaceae species.

<table>
<thead>
<tr>
<th>Species</th>
<th>Genes expressed in</th>
<th>Class I clades</th>
<th>Class II clades</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pereskia lycnidiiflora</td>
<td>Cambial zone</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Ariocarpus retusus</td>
<td>Cambial zone</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Ferocactus pilosus</td>
<td>Cambial zone</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Pilosocereus alensis a</td>
<td>Cambial zone</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Pachycereus pringlei</td>
<td>Root tip</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Lophophora williamsi b</td>
<td>Shoot and root</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Species</th>
<th>Genes expressed in</th>
<th>Class I clades</th>
<th>Class II clades</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pereskia lycnidiiflora</td>
<td>Cambial zone</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Ariocarpus retusus</td>
<td>Cambial zone</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Ferocactus pilosus</td>
<td>Cambial zone</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Pilosocereus alensis a</td>
<td>Cambial zone</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Pachycereus pringlei</td>
<td>Root tip</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Lophophora williamsi b</td>
<td>Shoot and root</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

a No RNA-seq data available. b For comparison, the data from Ibarra-Laclette et al. (2015) were used.

FIGURE 6 | Number of KNOX genes in species from different plant taxa. The number of both class I and class II KNOX genes are shown for the angiosperms, while for Physcomitrella patens and Selaginella moellendorffii, only the total numbers of KNOX genes are indicated. When all the species from an order belonged to the same family, the family is indicated in parentheses below the order. Where different families of an order are depicted, they are indicated with lowercase letters below the species name. For the order Malpighiales, the included families are: (a) Euphorbiaceae, (b) Linaceae, and (c) Salicaceae. For the order Brassicales, one species belongs to the (d) Caricaceae family, while the rest belong to the Brassicaceae family. The order Caryophyllales, to which the Cactaceae family belongs, is depicted in bold and indicated by a black triangle. Cactaceae species are not included, because there are no Cactaceae species with sequenced genomes.

* Species from the same genus with the same number of KNOX genes. ** Species from the same genus with different numbers of KNOX genes. Note that the number of KNOX genes belonging to class I, class II, or both, can be different.

genome, Cactaceae were not included in this analysis. All of the analyzed angiosperm species have several class II genes, and all but Carica papaya have more than one class I gene. There are 29 KNOX genes in Glycine max (17 class I and 12 class II genes) but only four in Carica (one class I and three class II genes). Furthermore, different species can have a different number of paralogs in every clade (Supplementary Table S2 and Figure 5). The number of KNOX genes can
also vary within species of the same family; for instance, within
diploid species of the Brassicaceae family (which includes the
largest number of species with sequenced genomes), the number
of KNOX genes varies from five in *Eutrema salsugineum* to 12 in
*Brassica rapa*. Even species within a single genus can have
variable numbers of KNOX genes; while the *Arabidopsis* species
*A. thaliana* and *A. lyrata* both possess eight KNOX genes, there
are six and nine genes in the *Capsella grandiflora* and *C. rubella*
genomes, respectively (Figure 6 and Supplementary Table S2).

We found that two putative orthologs of the *Arabidopsis* class
II KNOX paralogs KNAT3, KNAT4, and KNAT5 were expressed
in the cambial zone of *A. retusus*, but just one was expressed in the
cambial zone of *F. pilosus* and *P. lycnidiiflora* (Figures 3–5
and Table 1). Only one putative paralog expressed in the root
tip of *P. pringlei* belonged to this class. One paralog from the
published shoot and root transcriptome of *L. williamsii* (Ibarra-
Laclette et al., 2015) also belongs to this class. Remarkably,
among the three species used for cambial zone transcriptome
assembly, two paralogs in *A. retusus* and *F. pilosus* belonged
to the KNAT7 clade, whereas no putative KNAT7 ortholog was
found in the cambium of *P. lycnidiiflora* (*P. alesis* was not
considered as the transcriptome assembly was not performed for
this species). KNAT7 is a transcriptional repressor of secondary
cell wall biosynthesis (Liu et al., 2014). *P. lycnidiiflora* has
fibrous wood with significant cell wall lignification, *A. retusus*
has non-fibrous wood with abundant wide-band tracheids and
less lignification, while in the dimorphic wood of *F. pilosus*,
the wide-band tracheids develop before more lignified cells.
Our findings therefore suggest that the repressor activity of
the KNAT7 orthologs during secondary cell wall biosynthesis
could explain the differences in wood type among these species,
including the lower lignification level observed in *A. retusus* and
*F. pilosus*.

We found that the transcripts of both KNOX gene classes
are expressed in the cambial zone of the four Cactaceae species
studied, and therefore could be expressed either in different
cell types or in the same cell type within the vascular zone.
Several KNOX proteins have been shown to selectively move
from one cell layer to another (reviewed by Han et al., 2014),
suggesting that even if the class I and II genes are transcribed
in different cell types, the encoded proteins could coexist in
the same cell as a result of protein trafficking. In any case, as
no evidence of mutual repression was found when class I and
class II KNOX loss-of-function mutations of *Arabidopsis*
were combined (Furumizu et al., 2015), KNOX proteins belonging
to different classes could exert their functions within the same
cell. KNOX proteins are known to interact with a sister group
of proteins from the BELI-LIKE-HOMEODOMAIN (BHL, or
BELL from the founding *BELL* 1 gene) family; both KNOX
and BELL proteins belong to the TALE (three amino acid
loop extension) superclass of homeobox proteins. The KNOX-
BELL heterodimer formation affects its cellular localization
and the KNOX target selection (reviewed by Di Giacomo
et al., 2013; Arnaud and Pautot, 2014). From heterologous
expression experiments it was proposed that particular KNOX
proteins could interact with different BELL partners leading
to numerous combinations with distinct activities, and thus
regulating different sets of targets, including transcription factors
and hormonal pathways, and ultimately influence multiple plant
developmental processes. Recently, however, a more specific
selectivity was suggested for the in vivo KNOX-BELL interactions
(Furumizu et al., 2015), which could also enhance the properties
of each heterodimerization partner; thus, the interaction between
the transcriptional repressor KNAT7 and its partner BHL6 could
enhance their activity in repressing genes involved in secondary
cell wall biosynthesis (Liu et al., 2014).

Knowledge of the processes of duplication and
subfunctionalization of the regulatory genes helps us understand
the evolution of the different aspects of plant development
(Pires and Dolan, 2012), including the evolution of vascular
development. Our work contributes to the elucidation of the
mechanisms by which different wood types are formed in the
Cactaceae, and provides insight into the evolutionary history of
KNOX genes in different angiosperm species and their role in the
speciation of land plants.

AUTHOR CONTRIBUTIONS
TT, JR-R, AV, FV-S, GR-A, and SS designed the work; TT, JR-R,
EP, AV, GR-A, and SS performed the experiments and analyzed
the data; GR-A and JR-R prepared the figures; and TT and SS
wrote the manuscript. All the authors have read and approved
the manuscript.

ACKNOWLEDGMENTS
Funding was provided by the DGAPA-PAPIIT, UNAM (grants
IN209012 and 210115 to TT and IN207115 to SS) and by Consejo
Nacional de Ciencia y Tecnología (CONACyT) [CB2014-240055
grant to SS; Ph.D. scholarships 220343 to JR-R; 367139 to GR-
A and 700638 to EP, and a postdoc grant (220343) to JR-R].
We thank Jorge Nieto Sotelo for the valuable comments, Paul
Gaytan, Eugenio Lopez and personnel of the DNA synthesis and
sequencing unit of the IBT-UNAM for oligonucleotide synthesis,
and Laura Márquez Valdelamar for DNA sequencing.

SUPPLEMENTARY MATERIAL
The Supplementary Material for this article can be found online at:
full#supplementary-material

FIGURE S1 | Class I and II KNOX transcript expression detected in
cambial zone samples by end point RT-PCR.

FIGURE S2 | The identity and similarity matrix of the deduced KNOX
proteins identified in this study. (A) Class I proteins. (B) Concatenated KNOX1,
KNOX2, ELK, and HD domains of class I proteins. (C) Class II proteins. (D)
Concatenated KNOX1, KNOX2, ELK, and HD domains of class II proteins.

TABLE S1 | PCR primers used in this study.

TABLE S2 | The identifiers of the KNOX sequences used for molecular
phylogenetic analysis.
REFERENCES


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

Copyright © 2017 Reyes-Rivera, Rodríguez-Alonso, Petrone, Vasco, Vergara-Silva, Shishkova and Ruzin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.
CRP1 Protein: (dis)similarities between *Arabidopsis thaliana* and *Zea mays*

Roberto Ferrari1†, Luca Tadini1†, Fabio Moratti1†, Marie-Kristin Lehniger3, Alex Costa1, Fabio Rossi4, Monica Colombo5, Simona Masiero1, Christian Schmitz-Linneweber3 and Paolo Pesaresi6*

1 Dipartimento di Bioscienze, Università degli studi di Milano, Milano, Italy, 2 Max-Planck-Institut für Molekulare Pflanzenphysiologie, Potsdam-Golm, Germany, 3 Molecular Genetics, Institute of Biology, Humboldt University of Berlin, Berlin, Germany, 4 Dipartimento di Biotecnologie Mediche e Medicina Traslazionale, Università degli studi di Milano, Milano, Italy, 5 Centro Ricerca e Innovazione, Fondazione Edmund Mach, San Michele all’Adige, Italy, 6 Dipartimento di Scienze Agrarie e Ambientali - Produzione, Territorio, Agroenergia, Università degli studi di Milano, Milano, Italy

Biogenesis of chloroplasts in higher plants is initiated from proplastids, and involves a series of processes by which a plastid able to perform photosynthesis, to synthesize amino acids, lipids, and phytohormones is formed. All plastid protein complexes are composed of subunits encoded by the nucleus and chloroplast genomes, which require a coordinated gene expression to produce the correct concentrations of organelar proteins and to maintain organelle function. To achieve this, hundreds of nucleus-encoded factors are imported into the chloroplast to control plastid gene expression. Among these factors, members of the Pentatricopeptide Repeat (PPR) containing protein family have emerged as key regulators of the organelar post–transcriptional processing. PPR proteins represent a large family in plants, and the extent to which PPR functions are conserved between dicots and monocots deserves evaluation, in light of differences in photosynthetic metabolism (C3 vs. C4) and localization of chloroplast biogenesis (mesophyll vs. bundle sheath cells). In this work we investigated the role played in the process of chloroplast biogenesis by *At*5g42310, a member of the Arabidopsis PPR family which we here refer to as *At*CRP1 (Chloroplast RNA Processing 1), providing a comparison with the orthologous *Zm*CRP1 protein from *Zea mays*. Loss-of-function *atcrp1* mutants are characterized by yellow-albinotic cotyledons and leaves owing to defects in the accumulation of subunits of the thylakoid protein complexes. As in the case of *Zm*CRP1, *At*CRP1 associates with the 5′ UTRs of both *psaC* and, albeit very weakly, *petA* transcripts, indicating that the role of CRP1 as regulator of chloroplast protein synthesis has been conserved between maize and Arabidopsis. *At*CRP1 also interacts with the *petB-petD* intergenic region and is required for the generation of *petB* and *petD* monocistronic RNAs. A similar role has been also attributed to *Zm*CRP1, although the direct interaction of *Zm*CRP1 with the *petB-petD* intergenic region has never been reported, which could indicate that *At*CRP1 and *Zm*CRP1 differ, in part, in their plastid RNA targets.

**Keywords:** PPR, anterograde signaling, chloroplast, biogenesis, RNA metabolism
INTRODUCTION

In land-plants, nuclear-encoded pentatricopeptide repeat (PPR) containing proteins constitute a large family, which regulates organelle gene expression at the RNA level (Lurin et al., 2004; O’Toole et al., 2008; Barkan and Small, 2014). They are, indeed, a major constituent of the genome-coordinating anterograde signaling pathway that evolved to adapt the expression of the organellar genomes in response to endogenous and environmental stimuli that are perceived by the nucleus (Woodson and Chory, 2008).

A typical PPR motif is characterized by a degenerate 35-mer amino acid repeat that folds into two antiparallel alpha helices (Small and Peeters, 2000). PPR proteins contain a tandem array of 2–30 PPR motifs, which stack together to form a superhelic with a central groove that allows the protein to bind RNA (Lurin et al., 2004; Rivals et al., 2006). According to the characteristics of their repeats, PPR proteins are generally classified into P and PLS sub-families. The P-type proteins are implicated in the determination and stabilization of 5′ and/or 3′ RNA termini, RNA splicing and translation of specific RNAs in chloroplasts and mitochondria, while PLS-type proteins are generally involved in RNA editing (Barkan and Small, 2014). Higher plants harbor several hundreds of PPR proteins, which generally have distinct, non-redundant functions in organelle biogenesis, plant growth and development and adaptation to environmental cues (Barkan and Small, 2014; Manna, 2015), as revealed by the high number of ppr mutants with distinct phenotypes. This is due to their ability to recognize primary RNA sequences, with each protein having different target sites, thus implying that the elucidation of the primary role of each PPR protein is greatly facilitated by the identification of its RNA targets.

The detection of few native PPR-RNA interactions through RNA immunoprecipitation on microarray (RIP-Chip) analyses and in vitro binding assays using PPR recombinant proteins, together with PPR crystal structures indicate that PPR proteins bind their cognate RNA targets in a sequence specific manner (Meierhoff et al., 2003; Schmitz-Linneweber et al., 2005, 2006; Williams-Carrier et al., 2008; Yin et al., 2013; Okuda et al., 2014; Shen et al., 2016). The code describing how PPR proteins recognize specific nucleotides of their RNA targets relies primarily on two amino acids that are within a single PPR motif, specifically the fifth residue in the first helix and the last residue on the loop interconnecting adjacent motifs (Barkan et al., 2012; Yin et al., 2013; Cheng et al., 2016). However, the current understanding of the code does not allow accurate large-scale computational predictions of PPR targets (Takenaka et al., 2013; Kindgren et al., 2015; Hall, 2016; Harrison et al., 2016). Predictive power is constrained by the fact that the code is degenerate and by the low accuracy of current methods used for the identification of PPR domains, which in turn leads to mismatches in the amino acid/nucleotide alignments. However, a more robust annotation of PPR domains has recently been conducted and made available at the PlantPPR database1 (Cheng et al., 2016). Furthermore, more PPR-RNA interactions as well as crystal structures of PPR-RNA complexes need to be characterized in different species in order to improve the understanding of the code. This would also help to determine if the amino acid sequences of the PPR domains coevolved with the nucleotide sequences of their RNA targets and ultimately to determine whether there is functional conservation of PPR proteins among land plants.

The function of PPR proteins, and more generally the function of the nuclear gene complement involved in organelar RNA metabolism, have been primarily studied in maize, since the large seed reserves of maize support rapid heterotrophic growth of non-photosynthetic mutants and provide ready access to non-photosynthetic tissues for molecular biology and biochemical studies (Belcher et al., 2015). However, the degree of functional conservation of PPR proteins between maize and other species, including Arabidopsis thaliana, has yet to be investigated. The question is of particular interest since the elaboration of the thylakoid membrane system and the biogenesis of the multi-subunit photosynthetic complexes appear to have major differences between monocotyledonous and dicotyledonous plants (Pogson et al., 2015). Indeed in maize, and more generally in monocots, the process of chloroplast development from the proplastid to functional chloroplasts can be observed as a gradient along the leaf blade, whereas in dicots, such as Arabidopsis thaliana, the development of chloroplasts differs between developmental stages, plant organs – i.e., chloroplast development is different in cotyledons and leaves – and plant tissues (Pogson and Albrecht, 2011; Jarvis and Lopez-Juez, 2013).

The majority of PPR proteins are conserved at sequence level between dicots (Arabidopsis) and monocots (rice) (O’Toole et al., 2008). Orthologous pairs can readily be identified and in a number of cases, primary sequence conservation can be traced back to the roots of all embryophytes (O’Toole et al., 2008). As a matter of fact, functional differences between orthologous PPR proteins of maize and Arabidopsis have been observed. For example, the molecular phenotypes resulting from loss of the orthologous PPR proteins ATP4 (maize) and SVR7 (Arabidopsis) differ substantially (Liu et al., 2010; Zoschke et al., 2012, 2013a,b), as do the molecular defects in maize and Arabidopsis mutants lacking the PGR3 protein (Yamazaki et al., 2004; Cai et al., 2011; Belcher et al., 2015). Thus, the extent to which lessons on PPR proteins learnt from maize can be extrapolated to dicots, such as Arabidopsis, and more broadly to other organisms, needs further investigation.

In this context, we investigated here the function of and identified the RNA targets of the PPR protein At5g42310 from Arabidopsis thaliana, that shares high similarity with the well-characterized CRP1 (Chloroplast RNA Processing 1) protein from maize (ZmCRP1), and which we here refer to as AtCRP1. Our findings indicate that AtCRP1, like the orthologous ZmCRP1 (Barkan et al., 1994; Fisk et al., 1999; Schmitz-Linneweber et al., 2005), is essential for plant autotrophy since it plays a direct role in the accumulation of the cytochrome b6/f (Cyt b6/f) complex and of the PsaC subunit.

1http://www.plantppr.com
of photosystem I (PSI). Furthermore AtCRP1, similarly to ZmCRP1, is required for the accumulation of petB and petD monocistronic RNAs, indicating that the functional roles of CRP1 proteins are highly conserved between monocots and dicots.

MATERIALS AND METHODS

Plant Material and Growth Conditions

Arabidopsis thaliana atcrp1-1 (SALK_035048) (Alonso et al., 2003) and atcrp1-2 (SAIL_916A02) (Sessions et al., 2002) T-DNA insertion lines were identified by searching the T-DNA Express database. For promoter analyses, the putative AtCRP1 promoter region (AtCRP1p, −1062 to −2 upstream the translation starting codon) was cloned into pBGWFS7 destination vector and introduced into Arabidopsis wild type background, ecotype Columbia-0 (Col-0), by Agrobacterium tumefaciens-mediated transformation. AtCRP1-GFP transgenic lines were obtained by transformation of AtCRP1/atcrp1-1 heterozygous plants with either the AtCRP1 coding sequence fused to GFP under the control of 35S-CaMV promoter, cloned into pB7FWG2 vector, or the genomic locus fused to GFP under the control of the native promoter, cloned into a modified pGreenII vector (Gregis et al., 2009). The GUN1 coding sequence, devoid of the stop codon, was cloned into pB7RWG2 vector, carrying an RFP reporter gene. pB7FWG2, pBGWFS7, and pB7RWG2 plasmids were obtained from Flanders Interuniversity Institute for Biotechnology of Gent (Karimi et al., 2002). Primers used for amplification of the DNA fragments cloned into the vectors, reported above, are listed in Supplementary Table S2. Arabidopsis Col-0 and mutant plants were grown on soil under controlled growth chamber conditions with a 16 h light/8 h dark cycle at 22°C/18°C. In the case of mesophyll protoplast preparation, Arabidopsis plants were also grown on soil in a growth chamber under the above reported conditions. Moreover, phenotypic characterization and molecular biology analyses were also conducted on plants grown on Murashige and Skoog (MS) medium (Duchefa), supplemented with or without 1% (w/v) sucrose. Tobacco plants, employed for transient gene expression, were cultivated for 5–6 weeks in a greenhouse under a 12 h light/12 h dark cycle at 22°C/18°C.

Protoplast Transformation

Mesophyll protoplasts of Arabidopsis thaliana (Col-0) were isolated and transiently transformed according to Yoo et al. (2007) and Costa et al. (2012). Briefly, well-expanded rosette leaves from 3-to-5 week-old plants were cut into strips of 0.5–1 mm with a fresh razor blade. Leaf tissue was digested using an enzyme solution containing 1.25% cellulase Onozuka R-10 (Duchefa) and 0.3% Macerozyme R-10 (Duchefa) for 3 h at 23°C in the dark. The protoplast suspension was filtered through a 50 µm nylon mesh washed three times with W5 solution (154 mM NaCl, 125 mM CaCl2, 5 mM KCl, 2 mM MES, pH 5.7 adjusted with KOH) and used for PEG-mediated transformation. For each protoplast transformation 10 µg of a MidiPrep purified DNA (QIAGEN) plasmid harboring the 35S-CaMV::AtCRP1-GFP cassette was used. Protoplasts were maintained for 16–24 h at 23°C in the dark, before performing epifluorescent microscopy.

Transient Expression in Nicotiana benthamiana Leaves

Tobacco leaf infiltration was performed using A. tumefaciens strain GV3101/pMP90 carrying the specified constructs (see results for details) together with the p19-enhanced expression system (Voinnet et al., 2003), according to the method described by Waadt and Kudla (2008). The final OD600 for A. tumefaciens strains harboring 35S-CaMV::AtCRP1-GFP and 35S-CaMV::GUN1-RFP was 0.2 and 0.3, respectively. After infiltration, plants were incubated for 3–5 days under the conditions described above.

Confocal Microscopy Analysis

Confocal Scanning Laser Microscopy analyses were performed using an inverted microscope, Leica DMIRE2, equipped with a Leica TCS SP2 laser scanning device (Leica). For the simultaneous detection of GFP and chlorophyll autofluorescence the cells were excited (Arabidopsis mesophyll protoplasts or tobacco leaf cells) with the 488 nm line of the Argon laser and the emissions were collected between 515/535 and 650/750 nm, respectively. For RFP detection the cells were excited at 561 nm from a He/Ne laser and the emission was collected between 575/625 nm. Image analyses were performed with Fiji, an open-source platform for biological-image analysis (Schindelin et al., 2012).

Nucleic Acid Analyses

Arabidopsis DNA was isolated according to Ihnatowicz et al. (2004). Isolation of total RNA from homozygous atcrp1-1 plants at four-leaf rosette stage and RNA gel blot analyses were performed as described by Meurer et al. (2002), using 10 µg of total RNA for each sample. For the RNA slot blot hybridization experiments, one-fourth of the RNA purified from each immunoprecipitation pellet and one-tenth of the RNA purified from the corresponding supernatant were applied to a nylon membrane with a slot-blot manifold and hybridized to specific radio labeled probes (see Supplementary Table S2). 32P-labeled DNA probes, complementary to chloroplast genes, were amplified using the primer pairs listed in Supplementary Table S2. Four micrograms of total RNA, treated with TURBO DNA-free (Ambion by Life Technologies), were employed for first-strand cDNA synthesis using GoScript Reverse Transcription System (Promega) according to the supplier’s instructions. Quantitative Real-Time PCR (qRT-PCR) was carried out on an CFX96 Real-Time system (Bio-Rad), using the primer pairs

http://signal.salk.edu/cgi-bin/tdnaexpress
http://www.duchefa.com

http://www.fiji.sc/
reported in Supplementary Table S2. The SAND (Remans et al., 2008) and ubiquitin transcripts were used as internal references. Data from three biological and three technical replicates were analyzed with Bio-Rad CFX Manager software (V3.1).

**Immunoblot Analyses**

For immunoblot analyses, total proteins were prepared as described by Martinez-Garcia et al. (1999). Total proteins, corresponding to 5 mg of leaf fresh-weight (100% of WT and atcrp1-1 samples) and isolated from plants at four-leaf rosette stage, were fractionated by SDS–PAGE (12% acrylamide [w/v]; Schagger and von Jagow, 1987). Proteins were then transferred to polyvinylidene difluoride (PVDF) membranes (Ihnatowicz et al., 2004) and replicate filters were immunodecorated with antibodies specific for PSI (PsaA, PsaC, and PsaD), PSII (D1, PsbO) Cyt b6/f (PetA, PetB, and PetC), ATPase (ATPase-β) subunits, PSI (Lhca1, Lhca2) and PSII (Lhcb2, Lhcb3) antenna proteins, all obtained from Agrisera®. The GFP antibody was purchased from Life Technologies®.

**Chloroplast Stromal Preparation and Protein Immunoprecipitation**

Intact chloroplasts were isolated from 11 days old Arabidopsis plants, according to Kunst (1998), and Kupsch et al. (2012) with some modifications. Chloroplasts were directly resuspended in 300–400 µl of extraction buffer [2 mM DTT; 30 mM HEPES-KOH, pH 8.0, 60 mM KOAc, 10 mM MgOAc and proteinase inhibitor cocktail (Sigma–Aldrich-P9599)]. Two independent stromal preparations were carried out and one of them was performed in the presence of 2% sodium deoxycholate in order to solubilize the membrane-attached AtCRP1 protein fraction. Chloroplasts were then disrupted by pulling them through a syringe (0.55 mm × 40 mm) 30–40 times. The solution was centrifuged at 21,000 × g at 4°C to separate the stromal from the membrane fraction.

The isolated stromal fraction was diluted with one volume of coimmunoprecipitation (CoIP) buffer (150 mM NaCl, 20 mM Tris-HCl pH 7.5, 2 mM MgCl2, 0.5% Nonidet P-40 and 0.5 µg/mL Aprotinin). Five microliters of mouse anti-GFP antibody (Roche, No. 1181446001) were added to the stromal fraction and incubated for 1 h at 4°C and 13 rpm on an overhead shaker. Thereafter the coimmunoprecipitation was performed as described by Kupsch et al. (2012). Successful precipitation of AtCRP1-GFP was confirmed by immunoblot analyses, using the same GFP antibody.

**RNA Extraction and Labeling for RIP-Chip Assay**

RNA immunoprecipitation-chip analyses were performed using a tiling microarray covering the complete Arabidopsis chloroplast genome (Kupsch et al., 2012). The coimmunoprecipitated RNA was isolated from pellet and supernatant fractions either by phenol-chloroform extraction or using the Direct-zol™ RNA MiniPrep kit (Zymo Research). For the phenol-chloroform extraction, RNA samples were incubated in 1% SDS and 5 mM EDTA at room temperature for 5 min to dissociate RNA-protein complexes. RNA was phenol-chloroform extracted, ethanol precipitated with the addition of Glycoblot™ Coprecipitant (Thermo Fisher Scientific), washed twice with 75% ethanol, air-dried and resuspended in 20 µl RNase-free water. For the replicate, RNA was extracted using the Direct-zol™ RNA MiniPrep kit (Zymo Research) according to the manufacturer’s instructions. Before the extraction 2 µg yeast RNA was added to the coimmunoprecipitated RNA pellet. The entire RNA of the pellet fraction and 2 µg RNA of the supernatant fraction were used for labeling. The pellet and supernatant RNA were labeled with 0.5 µl Cy5 and 1 µl Cy3 dye, respectively (aRNA labeling kit, Kreatech Diagnostics). Labeling reaction, microarray hybridization, scanning, and evaluation were performed as described in Kupsch et al. (2012). Only PCR products for which more than half of all replicate spots (24 per PCR product spanning two experiments) passed our quality assessment (Kupsch et al., 2012) and were used in this analysis (Supplementary Table S1).

**In silico Prediction of AtCRP1 Binding Sites**

The putative AtCRP1 binding motif, i.e., the nucleotide preference for each of the amino acid pairs at the fifth and last position of PPR domains, was predicted in silico using the reported weighting schemes (Barkan et al., 2012; Barkan and Small, 2014; Harrison et al., 2016). The software FIMO®, which analyzes sequence databases for occurrences of known motifs (Grant et al., 2011), was employed to identify the potential binding sites of AtCRP1 within the regions enriched in our RIP-Chip experiment. Furthermore, the same regions were searched for the presence of sRNA native footprints, by consulting the JBrowse sRNA database® (Ruwe et al., 2016). Numbers that delimit the native footprints refer to the chloroplast genome of Arabidopsis thaliana (NC_000932.1).

**β-Glucuronidase (GUS) Assay**

For GUS histochemical detection, plant material was fixed in 90% acetone at −20°C for 1 h. Samples were then washed three times with NaPi buffer (NaH2PO4 50 mM, Na2HPO4 50 mM; pH 7.0) and stained overnight at 37°C with X-gluc solution [1 mM 5-bromo-4-chloro-3-indolyl-β-D-glucorionate, 2 mM K3/K4Fe(CN)6, 0.1% Triton (v/v), 10 mM EDTA, 50 mM NaPi pH 7.0]. 70% EtOH (v/v) was used as washing solution. Stained samples were then stored at 4°C and observed using a Zeiss Axioshot D1 microscope equipped with differential interference contrast (DIC) optics. Images were recorded with an AxioCam MRC5 camera (Zeiss) using the Axiosvision program (v.4.1).

---

2http://www.thermofisher.com
3http://meme-suite.org/tools/fimo
4https://www.molgen.hhu-berlin.de/projects-jbrowse-athaliana.php
Primary amino acid sequence alignment of AtCRP1 and ZmCRP1 proteins. The amino acid sequence of the Arabidopsis CRP1 (AtCRP1, At5g42310) was compared with CRP1 from Zea mays (ZmCRP1), using ClustalW2. Black boxes indicate strictly conserved amino acids, and gray boxes closely related ones. The predicted chloroplast transit peptides (ChloroP, http://www.cbs.dtu.dk/services/ChloroP/) are indicated in italics, and the PPR motives (P0-to-P14), identified using the PlantPPR database (http://www.plantppr.com), are marked with gray bars. The specificity determining amino acids in each PPR motif at position 5 and 35 are indicated by black and gray stars, respectively. Note that P0 motif was not considered to contribute to the identification of RNA targets, as previously reported by Barkan et al. (2012). P0 is composed of 30 aa, whereas all other P motifs are of 35 aa, with the exception of P2, which contains 37 aa in Arabidopsis and 38 in maize.

RESULTS

AtCRP1 Is a PPR Protein Imported into the Chloroplast

The Maize Genetics and Genomics Database (Lawrence et al., 2004) was used to identify the At5g42310 gene as the Arabidopsis ortholog of ZmCRP1 (see also Belcher et al., 2015). At5g42310 encodes a polypeptide of 709 amino acids with a calculated molecular mass of 80 kDa. Intron number (three) and position are conserved between the two genes, and BLASTP query of public Arabidopsis sequence database with ZmCRP1 amino acid sequence detected At5g42310 protein as the top hit with 55% sequence identity and 72% sequence similarity (Figure 1).

AtCRP1 is annotated as a PPR protein and shares with ZmCRP1 15 PPR tandem repeats, which were predicted by using the PlantPPR database (Cheng et al., 2016). All PPR motifs are of 35 aa, with the exception of P0 which consists of 30 aa and P2 and P3 of 37 aa in Arabidopsis and 38 aa in maize. The fifth and the last residue of each PPR domain form the amino acid pairs that specify the RNA target molecules (Cheng et al., 2016), and are labeled with gray and black stars in Figure 1. The ChloroP server (Emanuelsson et al., 1999) predicted the presence of a cTP of 54 residues (see amino acid residues in italics in Figure 1), indicating that AtCRP1, like ZmCRP1, could be imported into the chloroplast. To corroborate the in silico prediction, the AtCRP1-GFP fusion protein was expressed in transiently transformed Arabidopsis protoplasts (Figure 2). In agreement with the ChloroP prediction, the chimeric protein (GFP fluorescence) accumulated within the chloroplast in distinct fluorescent foci (CHL, autofluorescence of chloroplast chlorophylls, Figure 2A), resembling the nucleoid complexes. Indeed, AtCRP1-GFP chimera co-localized perfectly with the GUN1-RFP fusion protein, used as a nucleoid marker in this assay (RFP fluorescence, Figure 2B), (Koussevitzky et al., 2007; Colombo et al., 2016; Tadini et al., 2016), in tobacco leaf cells. To further localize AtCRP1, chloroplasts were fractionated to separate the stroma and thylakoid compartments. Immunoblot analysis, using a GFP specific antibody, allowed detection of AtCRP1-GFP specific signal in total chloroplasts, as well as in thylakoids and in the stromal fraction, indicating that the nucleoid AtCRP1 protein is both associated to membranes and soluble in the stroma (Figure 2C). These findings are in agreement with the identification of AtCRP1 as part of Megadalton complexes in the chloroplast stroma (Olinares et al., 2010), as well as in the grana of thylakoid membranes (Tomizioli et al., 2014).

AtCRP1 Is Essential for Plant Autotrophy

To investigate the role that AtCRP1 plays in Arabidopsis, two lines carrying T-DNA insertions into the coding sequence of At5g42310, renamed atcrp1-1 (Salk_035048) and atcrp1-2
Both T-DNA insertions completely suppressed the accumulation of the corresponding transcripts in homozygous mutant seedlings (Figure 3B), which were characterized by a paler pigmentation of cotyledons, visible even at the fully mature embryo stage (Figure 4A), and leaves (Figures 4B,C), and found to be seedling lethal under autotrophic growth conditions on soil and MS medium without sucrose, but able to develop yellow-albinotic rosette leaves and sterile inflorescence when sucrose was provided in the medium (Figure 4C). The mutant phenotype could be rescued by Agrobacterium tumefaciens-mediated transformation of heterozygous plants with either the appropriate coding sequence fused to the 35S promoter of cauliflower mosaic virus (35S-CaMV::AtCRP1-GFP), or the genomic sequence including a 1-Kbp fragment of the promoter region (AtCRP1p::AtCRP1-GFP), corroborating a direct correspondence between genotype and phenotype, and indicating that the AtCRP1-GFP chimera was fully functional, in both cases (Figure 4D). Interestingly, complemented plants carrying the AtCRP1-GFP construct under the control of the native promoter showed a fivefold increase in AtCRP1 gene expression (Figure 4E), most probably as consequence of the T-DNA insertion in a highly expressed

---

**Figure 2** | Subcellular localization of AtCRP1 in Arabidopsis mesophyll protoplasts and leaf cells. (A) Series of Lasers Scanning Confocal images (CLSM) of the subcellular localization of the AtCRP1-GFP fusion protein (indicated as GFP) expressed in transiently transformed Arabidopsis (ecotype Col-0) leaf mesophyll protoplasts. The GFP signal accumulates in distinct spots within the chloroplasts, visualized by the red chlorophyll autofluorescence (CHL), resembling the pattern of chloroplast nucleoids. BF, Bright Field. (B) Series of CLSM images of the subcellular localization of AtCRP1-GFP and GUN1-RFP [indicated as RFP and used as a marker of chloroplast nucleoids (Koussevitzky et al., 2007)] fusion proteins upon transient co-expression in tobacco leaf cells. The green fluorescence (GFP) co-localizes perfectly with the purple fluorescence (RFP) inside the chloroplasts (violet autofluorescence of chlorophylls, CHL), indicating that AtCRP1 protein is part of the chloroplast nucleoids. Images are representative of three independent experiments. Bar = 10 µm; p = chloroplast; n = nucleoid. (C) Immunoblot analyses of proteins from Col-0 and Arabidopsis transgenic lines containing the AtCRP1-GFP construct under the control of the AtCRP1 native promoter (approximately 1 kb upstream of the translation start codon, see also Materials and Methods). Equal protein amounts isolated from total chloroplasts, thylakoids and stroma were loaded. Filters were immunolabeled with a GFP specific antibody to detect the localization of the AtCRP1-GFP chimera. An antibody specific for the large subunit of RUBISCO (RbcL) was used as a marker of chloroplast stroma, whilst an Lhcb2 specific antibody was used as a marker of thylakoid membranes. Asterisks indicate the position of the AtCRP1-GFP fusion protein. One out of three immunoblots for each antibody is shown. Note that the AtCRP1-GFP chimera is fully functional, since it was able to rescue the atcrp1-1 mutant phenotype (see also Figure 4).

---

**Figure 3** | T-DNA tagging and expression levels of AtCRP1 gene. (A) Schematic representation of AtCRP1 gene, where exons are indicated as numbered white boxes, while introns are shown as black lines. Arrowheads indicate the positions of translation initiation and stop codons. The locations, designations and orientations of T-DNA insertions are indicated (RB, right border; LB, left border). Note that the T-DNA insertions are not drawn to scale. (B) Levels of AtCRP1 gene expression was ascertained by real-time PCR of cDNA obtained from leaves of WT (Col-0) and atcrp1-1, atcrp1-2 mutant plants. Gene expression was normalized to the level of AtCRP1 transcripts in Col-0 plants, and SAND and ubiquitin were used as internal references. The bars indicate standard deviations.

---
FIGURE 4 | Effects of loss of AtCRP1 on plant development. (A) Images of isolated fully mature embryos (bent cotyledon stage) from WT (Col-0) and atcrp1-1 seeds. The lack of AtCRP1 protein did not alter embryo development, although mutant embryos were slight larger and paler than Col-0. (B) atcrp1-1 seeds were able to germinate on soil, giving rise to yellow seedlings that accounted for about one quarter of all seedlings, indicative of a monogenic recessive trait. Mutant seedlings did not survive past the cotyledon stage. (C) Mutant seedlings showed albino cotyledons when grown on MS medium without sucrose and arrested at the cotyledon stage as in (B). However, when atcrp1-1 seedlings were grown on MS medium supplement with 1% sucrose, they showed yellow-albino cotyledons at 8 days after sowing and were able to develop up to 8–10 true leaves after 35 days. (D) The atcrp1-1 seedling lethal phenotype could be fully rescued by Agrobacterium tumefaciens-mediated transformation of AtCRP1/atcrp1-1 heterozygous plants with either the AtCRP1 coding sequence fused to GFP under the control of 3SS-CaMV promoter (3SS::AtCRP1-GFP#12), or the genomic sequence fused to GFP under the control of native promoter (AtCRP1p::AtCRP1-GFP#5). (E) Real-time PCR to monitor the expression of AtCRP1 gene in WT and complemented plants. Gene expression was normalized with respect to the level of AtCRP1 transcripts in Col-0, and SAND and ubiquitin were used as internal references. The bars indicate standard deviations. (F) Col-0 and 3SS::AtCRP1-GFP#16 transgenic line with about 15-folds more AtCRP1 transcripts than WT. In this case the transgenic line shows WT-like rosette, but it is characterized by shorter and paler stems, bleached cauline leaves, and sterile flowers. A detail of the stem and inflorescence is shown in the inset. Note that the detailed molecular characterization of AtCRP1 function was conducted on atcrp1-1 plants, since the atcrp1-2 seedlings showed an identical phenotype.

euchromatin region of the nuclear genome. Furthermore, a complete rescue of mutant plant phenotype could only be observed in 3SS::AtCRP1-GFP transgenic lines with a limited accumulation of AtCRP1 transcripts (Figures 4D,E). Higher AtCRP1 expression levels (around 15-folds in comparison to WT) led to transgenic plants with WT-like rosette but shorter and paler stems, bleached cauline leaves, and sterile flowers (Figures 4E,F).

Temporal and spatial expression patterns of AtCRP1, monitored by fusing the promoter region of the gene upstream of the GUS reporter gene (see also Materials and Methods), support further the key role played by AtCRP1 during early stages of seedling and leaf development (Figure 5). The GUS staining could, indeed, be detected in young cotyledons and in the upper portion of the hypocotyl (Figure 5A). Furthermore, intense GUS signals were observable in young developing leaves.
in cotyledons and leaves using quantitative Real-Time PCR were also obtained by monitoring the expression of AtCRP1 with respect to the level of transcripts in cotyledons at stage A, and on the first pair of true leaves at stages C–D (st. C and st. D) to monitor the accumulation of AtCRP1 transcripts. Gene expression was normalized with respect to the level of AtCRP1 transcripts in cotyledons at stage A, and SAND and ubiquitin were used as internal references. The bars indicate standard deviations.

(Figures 5C,D), whereas the GUS coloration tended to decrease in old cotyledons and leaves (Figures 5B–D). Similar results were also obtained by monitoring the expression of AtCRP1 in cotyledons and leaves using quantitative Real-Time PCR (qRT-PCR). In general, a high level of expression of AtCRP1 was observed in green developing tissues, such as young cotyledons and leaves, whereas the expression decreased in older tissues (Figure 5E).

atcrp1 Mutant Chloroplasts Fail to Accumulate Cytochrome $b_6/f$ Protein Complex and the PsA Subunit of PSI

The albino pigmentation of atcrp1 seedlings, together with their inability to grow under autotrophic conditions, indicated a defect in the thylakoid-associated photosynthetic apparatus. To verify this assumption, immunoblot analyses with antibodies specific for single subunits of the four major thylakoid protein complexes were performed on total leaf proteins. Leaf samples were harvested from atcrp1 plants at the four-leaf rosette stage and grown on MS-medium supplemented with 1% sucrose (Figure 6; see also Materials and Methods). Under standard light conditions (50 µmol photons m$^{-2}$ s$^{-1}$), subunits of Photosystem I (PsAa, PsAc, and PsD), Photosystem II (D1, PsbO), Light harvesting complexes (Lhca1, Lhca2, Lhcb2, and Lhcb3) and ATPase (ATPase-$\beta$) accumulated to levels lower than 10% with respect to wild type plants. Furthermore, subunits of the Cyt $b_6/f$ (PetA, PetB, and PetC) and PSI (PsAc) were below the limits of immunoblot detection.

In summary, these results indicate a general reduction of thylakoid protein complex subunits in atcrp1 leaves, with a particularly severe effect on the accumulation of the Cyt $b_6/f$ complex and PsA.

AtCRP1 Is Associated In vivo with psaC and petB-petD Transcripts

ZmCRP1 has been previously demonstrated to associate with the psaC and petA mRNAs in vivo by RIP-Chip analyses (Schmitz-Linneweber et al., 2005). To investigate whether AtCRP1 shares with ZmCRP1 the RNA targets, the same RIP-Chip approach employed in maize was used here. Stroma from plants expressing AtCRP1-GFP, under the control of the native promoter (AtCRP1p:AtCRP1-GFP), was isolated and the fusion protein was immunoprecipitated using an anti-GFP serum. As a control, we performed mock precipitations with stroma extracted from WT plants, using the same GFP antibody. RNA was purified from the immunoprecipitation pellets and supernatants and was labeled with Cy5 (red) and Cy3 (green) fluorescent dyes, respectively. The two RNA fractions from AtCRP1-GFP immunoprecipitations (IPs) and from mock IPs were competitively hybridized to a chloroplast genome tiling microarray (Kupsch et al., 2012). Enrichment of RNA is reflected in the ratio of red to green fluorescence for each spot on the microarray. Two biological replicate experiments were performed with stroma from AtCRP1-GFP expressing plants and two with WT stroma. Data from the four assays were normalized and used to calculate median enrichment ratios of the red and green fluorescence signals for each PCR product among the 24 replicate spots on two arrays (Supplementary Table S1). To identify enrichment of RNA species specifically in the AtCRP1-GFP immunoprecipitation,
we plotted the difference in median enrichment ratio for each DNA fragment between the AtCRP1-GFP and mock experiment against the position of the product on the plastid chromosome (Figures 7A,B). Four prominent peaks of differential enrichment were observed. One of them corresponds to the 5′UTR of psaC transcript, a target already recognized as a ligand of ZmCRP1 in RIP-Chip assays (Schmitz-Linneweber et al., 2005). A second RNA target is represented by the petB-petD intergenic region. This RNA was not identified to interact with ZmCRP1 by RIP-Chip analysis, however, ZmCRP1 is known to aid in maturation of this particular intergenic region (Barkan et al., 1994). Interestingly, the observed enrichment of rps15 transcripts might uncover a further, novel target of AtCRP1, whereas the enrichment of psbM/trnD transcripts is often observed in RIP-Chip experiments, thus this peak was considered an artifact.

To corroborate the RIP-Chip data, the AtCRP1-associated RNAs were analyzed by slot blots (Figure 7C). RNA purified from immunoprecipitation pellets and supernatants were probed with the PCR fragments that detected the most highly enriched sequences in the RIP-Chip assay. The data confirmed that the psaC and petB-petD transcripts are highly enriched in the AtCRP1-GFP immunoprecipitates, but not the rps15 RNA. ZmCRP1 was also reported to be associated with RNAs of the petA region (Schmitz-Linneweber et al., 2005; Williams-Carrier et al., 2008), however, no enrichment of petA transcripts could be observed in the AtCRP1-GFP RIP-Chip assay (Figure 7A) and a low enrichment was detected in the slot blot assay (Figure 7C), possibly indicating that the interaction of AtCRP1 with petA transcripts is not very stable. In general, our analysis cannot exclude the possibility that CRP1 binds to additional target RNAs, for example when interactions take place at chloroplast membranes. Since we are not using cross-linked material, weak RNA-protein interactions might be lost during our assay.

To support further the RIP-Chip findings, AtCRP1 target RNAs were interrogated for the presence of native footprints at the JBrowse database.1 The JBrowse database provides annotations of Arabidopsis thaliana organellar short RNA (sRNA), thought to be generated from protein-mediated temporary protection of target RNAs against exonucleolytic degradation (Ruwe et al., 2016; see also Figure 8). sRNAs were found within the 5′UTR of psaC (corresponding to the 117633–117597 region of chloroplast genome) and the petB-petD intergenic region (region 76318–76358), and an sRNA was also annotated in the 5′UTR of petA (region 61615–61643). Furthermore, AtCRP1 predicted RNA binding motifs were shown to co-map with the native footprints, when the corresponding sequences were searched for the occurrence of the consensus binding motif with the FIMO program in the MEME suite.12 A short RNA has been also mapped upstream of rps15, but this region was not enriched in the RIP-Chip assay and the match with the predicted binding site of AtCRP1 is weaker than for the psaC, petB-petD, and petA sRNAs.

---

1https://www.molgen.hu-berlin.de/projects-jbrowse-athaliana.php
2http://meme-suite.org/tools/fimo
FIGURE 7 | AtCRP1 RIP-Chip data plotted according to gene order within the plastid genome. (A) Differential enrichment ratios obtained by RNA immunoprecipitation (RIP)-Chip analysis. The enrichment ratios \( \frac{F_{635}}{F_{532}} \) obtained from an assay of \( \text{AtCRP1p::AtCRP1-GFP} \) chloroplast stroma extract were normalized with respect to a control assay that used WT (Col-0) chloroplast stroma extract (both assays were performed in duplicate). The median-normalized values for replicate spots were plotted according to gene order within the plastid genome. Fragments for which fewer than 13 spots per experiment \( \frac{\text{AtCRP1p::AtCRP1-GFP}}{\text{WT}} \) passed our manual quality control and/or yielded an \( F_{532} \) signal below background were excluded and appear as gap in the curve. The enrichment of \( psaC+5' \) UTR is in agreement with previous findings obtained by RIP-Chip analysis on CRP1 from maize (Schmitz-Linneweber et al., 2005).

(B) Immunoblot analysis of protein fractions obtained from immunoprecipitation experiments using the anti-GFP mouse antibody and stroma material from Col-0 and \( \text{AtCRP1-GFP} \) plants. Equal volumes of supernatant and pellet preparations were loaded onto the gel. Note that the pellet from AtCRP1-GFP immunoprecipitation gave a stronger signal than the corresponding supernatant, implying quantitative precipitation of AtCRP1-GFP. The fact that no signal was obtained with Col-0 extracts demonstrates the specificity of the antibody. The RbcL migration region of the Ponceau S stained nylon membrane, after transfer from SDS-PAGE, was used to verify equal loading.

(C) Verification of AtCRP1 RNA targets. Coimmunoprecipitations and RNA extractions from AtCRP1-GFP and Col-0 samples were performed as for RIP-Chip assays. The RNAs were then analyzed by slot-blot hybridization with the indicated probes (see also Materials and Methods and Supplementary Table S2). The ATPase-\( \alpha \) probe hybridization was included as a control. SUP, supernatant.
In summary, the RIP-Chip and slot blot data together with the colocalization of native footprints and AtCRP1 RNA binding motifs indicate that AtCRP1 likely binds directly to the 5′UTR of psaC and the petB-petD intergenic region and possibly to the 5′UTR of petA. On the contrary, the absence of an AtCRP1-specific footprint within the rps15 RNA, together with the failure of slot blot enrichment, makes any AtCRP1-rps15 interaction unlikely.

**AtCRP1 Is Required for the Correct Processing of petB-petD Transcripts**

To assess whether the lack of Cyt b6/f complex and PsA subunit, together with the marked reduction of all protein complex subunits observed in *atcrp1-1* thylakoids, was caused by deficiencies in transcript accumulation and AtCRP1-dependent transcript processing, we probed the identified AtCRP1 RNA targets and other plastid transcripts by gel blot hybridization (Figure 9).

We investigated the transcripts encoding the subunits CP47 (psbB), T (psbT), and H (psbH) of photosystem II (PSII), subunits A (psaA) and C (psaC) of PSII, Cyt f (petA), Cyt b6 (petB) and subunit IV (petD) of cytochrome b6/f and the alpha subunit of ATPase (ATPase-α). All these transcripts accumulated in *atcrp1-1* plastids to levels lower than WT, indicating that global plastid gene expression is affected by the *atcrp1-1* mutation, and explaining the marked reduction of thylakoid protein accumulation observed in *atcrp1-1* leaves.

Furthermore, the plastid polycistronic transcription unit *psbB-psbT-psbH-petB-petD* showed some striking alteration of transcript pattern in *atcrp1* samples (Figure 9). In particular, the monocistronic *petB* (band #4, 0.8 Kb), the dicistronic *psbH-petB* (band #3; 1.2 Kb) and the unspliced *petB* (band #2, 1.6 Kb) transcripts were barely detectable in the mutant, whereas the *petB*-unspliced *petD*-spliced dicistronic transcript (band #1, 2.2 Kb), detected with probes D, E, F, and H, accumulated to even higher levels in *atcrp1* plastids, presumably due to the failure of AtCRP1-dependent processing between the *petB* and *petD* coding regions, as also shown in *zmcrlp* mutant plants (Barkan et al.,...
FIGURE 9 | Transcript patterns of chloroplast genes in Col-0 and atcrp1-1 mutant leaves. (A) The structure of the psbB gene cluster and probes A to H used in RNA gel blots analysis in (B) are shown. Furthermore, processed and spliced transcripts that accumulate differentially between Col-0 and mutant chloroplasts are drawn to scale and numbered from 1 to 4. Upward arrow indicates transcripts that accumulate to higher levels in atcrp1-1 than Col-0 chloroplasts, whilst the downward arrow is used for transcripts less abundant or absent in mutant samples. The putative binding site of AtCRP1 within the petB-petD intergenic region is also indicated. (B) RNA gel blot analysis of the psbB gene cluster were performed using probes indicated as A to H, whilst petA, ATPase-A, psaC, and psaA specific probes are described in section “Materials and Methods.” The identity of labeled transcripts (1–4), shown in (A) together with their size, was established based on the hybridization pattern, transcript size and on data reported in Meierhoff et al. (2003) and Stoppel et al. (2011). Asterisks indicate the mature transcript forms. A portion of the ethidium bromide stained Agarose gels, containing the cytosolic 25S rRNA, is included, as loading control, below each filter. One out of three Northern-blots for each transcript-specific probe is shown.

1994; Fisk et al., 1999). In contrast with maize, monocistronic and spliced petD transcripts of ~600 nucleotides do not accumulate to significant levels in Arabidopsis, and thus its absence was not observed in atcrp1 plastids (Barkan et al., 1994; Barkan, 2011).

Moreover, the lack of the PsaC and PetA subunits could be the consequence of the simultaneous decrease of transcript accumulation and a possible defect in AtCRP1-dependent activation of psaC and petA translation, as shown in Zea mays (Barkan et al., 1994; Schmitz-Linneweber et al., 2005). However, the specific regulatory role of AtCRP1 in plastid protein translation is difficult to verify, owing to the general and pleiotropic decrease of mature plastid rRNA in atcrp1-1 leaves, in spite of WT-like accumulation of rrn23 and rrn4.5 precursor forms (Figure 10). This rRNA accumulation pattern
is very similar to the ones of mutants with impaired chloroplast translation and has been interpreted as a secondary consequence of reduced plastid protein synthesis (Tiller et al., 2012; Tadini et al., 2016).

**DISCUSSION**

In this study we have investigated the role of AtCRP1 in the biogenesis of dicotyledonous-C3 chloroplasts and compared its function to the already characterized monocotyledonous-C4 chloroplast counterpart, ZmCRP1. Both proteins are essential for chloroplast biogenesis and photosynthetic activity, since they are required for the processing and translation of specific plastid transcripts encoding subunits of the thylakoid protein complexes. Our results indicate that AtCRP1 and ZmCRP1 have very similar RNA targets and the main functional divergences are most likely due to the distinct localization of the two proteins inside the chloroplast and the partially different affinity for the RNA targets (see Table 1).

**CRP1 Proteins Are Part of Chloroplast Nucleoids**

We detected AtCRP1 in the stroma and associated with thylakoid membranes (see Figure 2; Table 1), whereas ZmCRP1 is a stromal protein with no detectable association with chloroplast membranes (Fisk et al., 1999). The dual localization of AtCRP1 within the chloroplast is supported by proteomic studies that detected AtCRP1 in the grana-fraction of Arabidopsis thylakoids (Tomizioi et al., 2014) and in the stroma proteome, as part of Megadalton complexes (Olianas et al., 2010). In particular, AtCRP1 appeared to be highly enriched in fractions that contained ribosomal proteins, translation factors, RNA helicases and other PPR proteins, suggesting a major role of AtCRP1 in chloroplast gene expression. These data, together with the co-localization with GUN1 protein (see Figure 2), indicate that AtCRP1 is integral to chloroplast nucleoids (Koussevitzky et al., 2007; Colombo et al., 2016; Tadini et al., 2016), i.e., the DNA-containing structures without defined boundaries that harbor the plastid gene expression machinery (Pfalz and Pfannschmidt, 2013; Melonek et al., 2016). Similarly, ZmCRP1 was found to be highly enriched in the nucleoid fractions of maize plastids, together with proteins involved in DNA replication, organization and repair as well as transcription, mRNA processing, splicing and editing (Majeran et al., 2012), further supporting the involvement of CRP1 proteins in plastid gene expression.

**CRP1 Proteins Are Required for the Biogenesis of the Photosynthetic Apparatus**

The yellow-albinotic and seedling lethal phenotype exhibited by atcrp1 is very similar to the chlorophyll deficient and lethal phenotype of zmcrcp1 plants (Barkan et al., 1994; Fisk et al.,...
TABLE 1 | Overview of the phenotypes of Arabidopsis and maize crp1 mutants and comparison of their molecular roles in chloroplast biogenesis.

<table>
<thead>
<tr>
<th>Plant phenotype</th>
<th>atcrp1a</th>
<th>zmcrp1b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seeding lethal with yellow-albinotic cotyledons and leaves. Plants are able to develop mature leaves and sterile flowers when grown on MS medium supplemented with sucrose</td>
<td>Seeding lethal with pale-green cotyledon and leaves. Plants are able to develop mature non-photosynthetic leaves thanks to the large reserves of maize seeds</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CRP1 protein localization</th>
<th>atcrp1</th>
<th>ZmCRP1</th>
</tr>
</thead>
<tbody>
<tr>
<td>AtCRP1 is a component of plastid nucleoids and it is found associated to thylakoid membranes and in the stroma</td>
<td>ZmCRP1 has been reported to be highly enriched in plastid nucleoids and to localize exclusively in the chloroplast stroma</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Thylakoid protein content</th>
<th>PSI</th>
<th>PSII</th>
<th>Cyt b6/f</th>
<th>ATPase</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSI</td>
<td>PSII</td>
<td>Cyt b6/f</td>
<td>ATPase</td>
<td></td>
</tr>
<tr>
<td>− (/PsaC)</td>
<td>−</td>
<td>/</td>
<td>−</td>
<td>−</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RNA targets</th>
<th>RIP-Chip</th>
<th>Slot-Blot</th>
<th>In vivo footprint</th>
<th>RIP-Chip</th>
<th>Slot-Blot</th>
<th>In vivo footprint</th>
</tr>
</thead>
<tbody>
<tr>
<td>psaC</td>
<td>psaC</td>
<td>psaC</td>
<td>n.r.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>petB-petD</td>
<td>petB-petD</td>
<td>petB-petD</td>
<td>n.r.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/</td>
<td>petA</td>
<td>petA</td>
<td>n.r.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rps15</td>
<td>/</td>
<td>/</td>
<td>n.r.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metabolism of chloroplast RNAs</th>
<th>Accumulation</th>
<th>Processing defects</th>
<th>Accumulation</th>
<th>Processing defects</th>
</tr>
</thead>
<tbody>
<tr>
<td>psaC</td>
<td>−</td>
<td>No</td>
<td>=</td>
<td>No</td>
</tr>
<tr>
<td>petB-petD</td>
<td>/</td>
<td>Yes</td>
<td>/</td>
<td>Yes</td>
</tr>
<tr>
<td>petA</td>
<td>−</td>
<td>No</td>
<td>=</td>
<td>No</td>
</tr>
</tbody>
</table>

aData are obtained from the present manuscript bData are obtained from Barkan et al. (1994, 2012), Fisk et al. (1999), Schmitz-Linneweber et al. (2005). −, marked reduction; /, complete absence; =, no changes; +, increase; n.r., not reported.

1999). Arabidopsis mutants die at the two-cotyledon stage after germination on soil, but can overcome seedling lethality on sucrose-containing media, where they develop mature leaves and sterile flowers (see Figure 4; Table 1). Similarly, non-photosynthetic zmcrp1 plants die at about 3 weeks after germination when seed reserves are exhausted. Furthermore, the atcrp1 phenotype appears to be typical of Arabidopsis mutants lacking components of the photosynthetic apparatus and not of the gene expression machinery or of the protein import apparatus, since the latter usually result in the premature arrest at the globular-to-heart stage of embryo development, when chloroplast biogenesis begins (Ruppel and Hangarter, 2007; Romani et al., 2012; Beeler et al., 2014). Nevertheless, the pale-green pigmentation of the mutant embryo at bent-cotyledon stage (see Figure 4) and the β-glucuronidase (GUS) activity observed in young developing cotyledons and rosette leaves, but not in older tissues (see Figure 5), indicate that AtCRP1 gene expression and protein accumulation is required during the very early stages of the photosynthetic apparatus assembly. Immunoblot data indicate, indeed, that AtCRP1, like ZmCRP1, might act as a nuclear-encoded anterograde regulatory component responsible for coordination of the accumulation of Cyt b6/f and PSI protein complexes (see Figure 6). Besides their role in linear electron transport (LET), Cyt b6/f and PSI indeed play a key role in Cyclic Electron Transport (CET), which has been reported to be enhanced in Arabidopsis green seeds and to be required for optimal seed vigor and seed germination rate (Allorent et al., 2015).

In contrast to zmcrp1 plants (Barkan et al., 1994), the absence of AtCRP1 destabilized the entire photosynthetic apparatus, as shown by the marked reduction of PSII core, ATPase and LHC protein levels. The general down-regulation of thylakoid complexes owing to defects in the intersystem electron transport chain appears to be a common feature of Arabidopsis photosynthetic mutants and provides clear evidence of a different adaptive response between monocot and dicot plants (Meurer et al., 1996; Varotto et al., 2000, 2002; Maiwald et al., 2003; Weigel et al., 2003; Ihnatowicz et al., 2004, 2007; Belcher et al., 2015). Furthermore, the atcrp1-1 phenotype, both in terms of plastid transcript and plastid protein accumulation, appears to be much more drastic than the one of other ppr mutants required for the processing and expression of psbB-psbT-psbH-petB-petD operon, such as hcf152 (Meierhoff et al., 2003), suggesting that the absence of AtCRP1 protein might affect the activity of other factors essential for plastid gene expression. As a matter of fact, rRNA abundance is markedly reduced in atcrp1-1 plastids, indicating a general reduction of protein synthesis, as consequence of pleiotropic effects.

RNA Targets: Commonalities and Divergences between AtCRP1 and ZmCRP1 Proteins

RNA immunoprecipitation-Chip and slot blot data suggest a physical interaction between AtCRP1 and the transcripts of
psaC, petB-petD and possibly petA, even though it is not known whether these interactions are direct or mediated by other factors (see Figure 7). However, all of these RNAs harbor a region where a native footprint is annotated, raising the tempting hypothesis that AtCRP1 is in fact the RNA-binding factor responsible for that footprint (see Figure 8). Furthermore, when these enriched fragments were searched for occurrences of the predicted binding motif of AtCRP1, each of them proved to contain a hit inside the footprint region, strongly suggesting that AtCRP1 could be the factor leaving those footprints. Nevertheless, the observation that the footprints identified in Arabidopsis psaC, petA, and petB-petD transcripts are larger than the 14 nucleotide size of the predicted AtCRP1 footprint (37, 29, and 41 nucleotides in psaC, petA, and petB-petD, respectively) supports the view that the binding of AtCRP1 to its targets in vivo could be stabilized by other protein partners. For instance, the peptide chain release factor B3 (PrfB3) has been also shown to be required for Arabidopsis autotrophic growth and for the stability of 3′-processed petB transcripts to adjust cytochrome b6/f levels (Stoppel et al., 2011), thus possibly being an AtCRP1 specific protein partner. Similarly, PPR proteins involved in RNA stabilization and editing have been shown to interact with RNA Recognition Motif (RRM) proteins and other factors, indicating that larger protein complexes assembled around a PPR protein are likely to occur (Kupsch et al., 2012; Takenaka et al., 2014; Shi et al., 2015).

The interactions with the 5′ UTR of psaC and petA have also been reported in the case of ZmCRP1 (Schmitz-Linneweber et al., 2005; Williams-Carrier et al., 2008), indicating that this feature of CRP1 function is conserved between Arabidopsis and maize. ZmCRP1 was also shown to bind directly to the 5′-UTR of petA transcripts by electrophoresis mobility shift assay (Williams-Carrier et al., 2008), favoring the possibility of a direct binding of CRP1 proteins to the corresponding RNA targets (see also Figure 8). Furthermore, ZmCRP1 has been proposed to directly control the translation of petA and psaC transcripts (Barkan et al., 1994), as shown through pulse labeling and polysome loading (in the case of petA), or deduced from the reduced association of psaC RNAs with ribosomes. Interestingly, the PsAC subunit of PSI and the PetA subunit of Cyt b6/f could not be detected in atcrp1 thylakoids, despite the accumulation of the corresponding transcripts with no processing defects (see also Figure 9), suggesting that AtCRP1 plays a major role in translation regulation also in Arabidopsis. Unfortunately, the specific requirement of AtCRP1 in plastid protein translation cannot be verified by comparing Col-0 and atcrp1-1 leaves, due to the marked reduction of rRNA accumulation in atcrp1-1 plastids.

In addition to the defects in petA translation, the complete absence of Cyt b6/f protein complex observed in atcrp1 thylakoids can also be attributed to processing alterations of the psbB-psbT-psbH-petB-petD polycistronic transcription unit. The lack of the monocistronic petB, the dicistronic psbH-petB, and the unspliced petB transcripts, together with the direct binding of AtCRP1 to the petB-petD intergenic region, strongly support the role of AtCRP1 in the metabolism of petB and petD transcripts. PPR protein-derived RNA-footprints are considered to arise due to exonucleolytic activity (Ruwe et al., 2016). Since sRNAs corresponding to predicted binding sites of AtCRP1 are identified here, the most likely role for AtCRP1 is to block exonucleases from degrading the petB and petD transcripts. A similar defect in petB-petD maturation has been reported in zmcrp1 mutant plants (Barkan et al., 1994; Fisk et al., 1999; Schmitz-Linneweber et al., 2005), although no association was detected between ZmCRP1 and the petB-petD intergenic region (Schmitz-Linneweber et al., 2005), so it is still uncertain whether the role of ZmCRP1 is direct or indirect.

**CONCLUSION**

Taken together, the characterization of the functional role of AtCRP1 in chloroplast biogenesis has highlighted several features in common with the ZmCRP1. Both proteins appear to control, directly or indirectly, the expression of plastid genes encoding subunits of Cyt b6/f and PSI protein complexes. The coordination of the accumulation of these two protein complexes is fundamental to guarantee optimal photosynthesis in mature plants, but appears also to be important during seed germination, when cyclic electron transport is highly enhanced relative to LET.

Differences in RNA targets observed by immunoprecipitation and hybridization assays between AtCRP1 and ZmCRP1 might be explained by a broad affinity for RNA targets, but may also have technical reasons (GFP antibody for Arabidopsis versus direct anti-ZmCRP1 antibody in maize). Evidence in favor of conservation of PPR protein activity between different species has been reported for the PLS and P subfamilies (Choury et al., 2005; Bolle and Kempken, 2006; Choury and Araya, 2006); for instance, the maize MPPR6 protein can complement loss-of-function Arabidopsis mutants lacking the orthologous protein (Manavski et al., 2012). However, functional divergence has been also observed, as in the case of orthologous PPR proteins ATP4 (maize) and SVR7 (Arabidopsis) (Liu et al., 2010; Zoschke et al., 2012, 2013a,b). Further studies aimed to verify the degree of protein activity conservation between monocots and dicots are needed to extend our knowledge of PPR protein functions and the degree of protein function conservation. The parallel characterization of PPR orthologs, including the relationship between their protein structures and the corresponding target RNA species, may represent an underestimated and powerful strategy to precisely determine the PPR code, essential for a fast and accurate large scale prediction of PPR targets, and for the functional characterization of the PPR-mediated nucleus-to-chloroplast anterograde signaling pathway.

**AUTHOR CONTRIBUTIONS**

RF, LT, FM, FR, SM, M-KL, CS-L, and PP participated to the organization of the manuscript. RF, LT, FM, FR, SM, MC, AC,
and PP designed and carried out the experiments related to the molecular biology and biochemical characterization of atcrp1 mutants. RF, FM, LT, M-KL, and CS-L were involved in RIP-Chip and slot blot assays, as well as in the in silico identification of native footprints and prediction of AtCRP1 binding motif. PP wrote the manuscript.

FUNDING
This work was supported by ERA-NET Cofund FACCE SURPLUS (BarPLUS grant id. 93) to PP. Work in the lab of CS-L was supported by DFG grant SCHM 1698/5-1.

REFERENCES

ACKNOWLEDGMENT
We thank Hannes Ruwe and Gongwei Wang for helpful comments on the sRNA analysis.

SUPPLEMENTARY MATERIAL
The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2017.00163/full#supplementary-material

TABLE S1 | RNA immunoprecipitation (RIP-Chip) data.

TABLE S2 | Oligonucleotides employed for cloning, genotyping, northern blot, slot blot and qRT-PCR assays.


**Conflict of Interest Statement:** 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.
## Advantages of publishing in Frontiers

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>OPEN ACCESS</td>
<td>Articles are free to read, for greatest visibility</td>
</tr>
<tr>
<td>COLLABORATIVE PEER-REVIEW</td>
<td>Designed to be rigorous – yet also collaborative, fair and constructive</td>
</tr>
<tr>
<td>FAST PUBLICATION</td>
<td>Average 85 days from submission to publication (across all journals)</td>
</tr>
<tr>
<td>COPYRIGHT TO AUTHORS</td>
<td>No limit to article distribution and re-use</td>
</tr>
<tr>
<td>TRANSPARENT</td>
<td>Editors and reviewers acknowledged by name on published articles</td>
</tr>
<tr>
<td>SUPPORT</td>
<td>By our Swiss-based editorial team</td>
</tr>
<tr>
<td>IMPACT METRICS</td>
<td>Advanced metrics track your article’s impact</td>
</tr>
<tr>
<td>GLOBAL SPREAD</td>
<td>5’100’000+ monthly article views and downloads</td>
</tr>
</tbody>
</table>

---

**Frontiers**  
EPFL Innovation Park Building 1 • 1015 Lausanne • Switzerland  
Tel: +41 21 510 17 00 • Fax: +41 21 510 17 01 • info@frontiersin.org  
www.frontiersin.org

Find us on:  
![Twitter](https://twitter.com)  
![LinkedIn](https://www.linkedin.com)  
![Google+](https://plus.google.com)  
![Facebook](https://www.facebook.com)