Journal of Integrative Agriculture 2017, 16(7): 1576–1591

Available online at www.sciencedirect.com

ScienceDirect

RESEARCH ARTICLE

In silico genome-wide identification, phylogeny and expression analysis of the *R2R3-MYB* gene family in *Medicago truncatula*

ZHENG Xing-wei, YI Deng-xia, SHAO Lin-hui, LI Cong

Institute of Animal Sciences, Chinese Academy of Agricultural Sciences, Beijing 100193, P.R.China

Abstract

The *R2R3-MYB* genes make up one of the largest transcription factor families in plants, and play regulatory roles in various biological processes such as development, metabolism and defense response. Although genome-wide analyses of this gene family have been conducted in several species, *R2R3-MYB* genes have not been systematically analyzed in *Medicago truncatula*, a sequenced model legume plant. Here, we performed a comprehensive, genome-wide computational analysis of the structural characteristics, phylogeny, functions and expression patterns of *M. truncatula R2R3-MYB* genes. DNA binding domains are highly conserved among the 155 putative MtR2R3-MYB proteins that we identified. Chromosomal location analysis revealed that these genes were distributed across all eight chromosomes. Results showed that the expansion of the *MtR2R3-MYB* family was mainly attributable to segmental duplication and tandem duplication. A comprehensive classification was performed based on phylogenetic analysis of the *R2R3-MYB* gene families in *M. truncatula*, *Arabidopsis thaliana* and other plant species. Evolutionary relationships within clades were supported by clade-specific conserved motifs outside the MYB domain. Species-specific clades have been gained or lost during evolution, resulting in functional divergence. Also, tissue-specific expression patterns were investigated. The functions of stress response-related clades were further verified by the changes in transcript levels of representative *R2R3-MYB* genes upon treatment with abiotic and biotic stresses. This study is the first report on identification and characterization of *R2R3-MYB* gene family based on the genome of *M. truncatula*, and will facilitate functional analysis of this gene family in the future.

Keywords: *R2R3-MYB*, *Medicago truncatula*, gene family, stress response, function prediction

1. Introduction

Transcription factors (TFs) play significant roles in the regulation of gene expression by activating or inhibiting the transcription of target genes (Mizoi *et al.* 2012). Most known TFs are grouped into different families based on their DNA-binding domains and structures (Riechmann *et al.* 2000).

MYB proteins constitute one of the best-characterized TF families and they are present in all eukaryotes (Dubos *et al.* 2010; Feng *et al.* 2015; Li *et al.* 2016). The MYB superfamily is defined by the presence of a highly conserved DNA-binding domain near the N-terminus (Stracke *et al.* 2014). This domain typically contains one to four imperfect repeats (R1, R2, and R3), each of which is approximately 50–53 amino acid (aa) residues in length and encodes three α-helices (Li *et al.* 2015). Within each repeat, the second and third helices form a helix-turn-helix (HTH) structure that extends

Received 3 June, 2016 Accepted 9 November, 2016
ZHENG Xing-wei, E-mail: smilezxw@126.com; Correspondence LI Cong, Tel: +86-10-62815998, E-mail: licong0520@sina.com

doi: 10.1016/S2095-3119(16)61521-6

into the major groove of DNA and mediates recognition of the target sequence (Martin and Paz-Ares 1997; Li *et al.* 2016). In contrast, the highly divergent C-terminus of MYB proteins contains the transcription activation/repression domain, which confers the regulatory role for each family member (Kranz *et al.* 1998; Franco-Zorrilla *et al.* 2014).

MYB proteins can be divided into different groups based on the number of adjacent repeats: 2R (R2R3-MYB), 3R (R1R2R3-MYB), 4R (R1R2R2R1/2-MYB) and 1R-MYB (MYB-related proteins) (Hou *et al.* 2014; Du *et al.* 2015; Liu *et al.* 2015). In plants, R2R3-MYBs are the most abundant type, making up one of the largest transcription factor families and are the most similar to their vertebrate homologs, the c-MYBs (Geethalakshmi *et al.* 2015). The *R2R3-MYB* genes probably evolved from an *R1R2R3-MYB* gene ancestor through loss of the R1 repeat (Lipsick 1996; Braun and Grotewold 1999) or from an *R1-MYB* gene through an intragenic domain duplication of the R1 repeat (Jiang *et al.* 2004).

The first cloned plant *MYB* gene *COLORED1* (*C1*) was isolated from maize (*Zea mays*), and encodes a R2R3-MYB transcription factor involved in anthocyanin biosynthesis (Paz-Ares *et al.* 1987). Subsequently, an increasing number of R2R3-MYB proteins were characterized by genetic approaches and shown to regulate a number of biological processes, including cell shape and identity (Perez-Rodriguez *et al.* 2005; Baumann *et al.* 2007; Lau *et al.* 2015; Kim *et al.* 2016), developmental control (Muller *et al.* 2006; Song *et al.* 2011; Ambawat *et al.* 2013; Xu *et al.* 2016; Zheng *et al.* 2016), hormone signal transduction (Gocal *et al.* 1999; Abe *et al.* 2003; Yu *et al.* 2016), response to abiotic stresses and pathogen defense (Zhang *et al.* 2012; Bonthala *et al.* 2016; Liu *et al.* 2016; Wang H Y *et al.* 2016). In recent years, accumulating evidence has confirmed that numerous R2R3-MYBs are key factors in regulatory networks controlling primary and secondary metabolism. *R2R3-MYB* subfamily members in combination with bHLH TFs and WD40 proteins form a ternary MYB-bHLH-WD40 complex to regulate the expression of structural genes controlling the late steps of the anthocyanin and proanthocyanidins (condensed tannins) biosynthetic pathway (Albert 2015; Passeri *et al.* 2016; Perez-Diaz *et al.* 2016). Additionally, R2R3-MYB TFs are involved in the regulation of secondary cell wall formation, by modulating the transcription of lignin, xylose and cellulose biosynthetic genes (Chai *et al.* 2014; Wang W *et al.* 2016). Moreover, a newly identified R2R3-MYB protein, RCP1, positively regulates carotenoid biosynthesis during flower development (Sagawa *et al.* 2016).

Due to its diploidy and considerably small genome (haploid size 500–550 Mb), ability of nitrogen fixation, high genetic transformation efficiency, and rapid life cycle with high seed yield, *Medicago truncatula* is considered a model species to analyze biological processes that are unique or pertinent to legumes (Young *et al.* 2011). Since the release of several whole genome sequences, *R2R3-MYB* genes have been well annotated in *Arabidopsis thaliana* (126 members) (Stracke *et al.* 2001), *Oryza sativa* (88 members) (Katiyar *et al.* 2012), *Z. mays* (157 members) (Du *et al.* 2012a), *Glycine max* (soybean, 244 members) (Du *et al.* 2012b), *Populus trichocarpa* (192 members) (Wilkins *et al.* 2009) and *Vitis vinifera* (108 members) (Matus *et al.* 2008). Nevertheless, a systematic analysis of the *R2R3-MYB* gene family in *M. truncatula* is still lacking. It is therefore of interest to conduct a systematical analysis of *M. truncatula R2R3-MYB* genes, including gene classification, chromosomal locations, phylogenetic relationships, conserved motifs as well as expression patterns.

2. Materials and methods

2.1. Database search for R2R3-MYB protein coding genes in the *M. truncatula* genome

The sequences of 126 *A. thaliana* R2R3-MYB proteins were downloaded from The Arabidopsis Information Resource (TAIR, www.arabidopsis.org) (Berardini *et al.* 2015). These sequences were used as queries in a BLAST search against the *M. truncatula* proteome sequences provided by The J. Craig Venter Institute (JCVI) *M. truncatula* Sequencing Project v4.0 (www.jcvi.org/medicago/) with an E-value cutoff of 0.001. The obtained MYB sequences were confirmed based on the presence of two or more intact MYB domains using the Pfam Program (http://pfam.xfam.org/) (Finn *et al.* 2016) and SMART database with an E-value cutoff of 1E-10 (http://smart.embl-heidelberg.de/) (Finn *et al.* 2016). Information about coding sequences (CDS), full-length gene sequences and intron/exon distribution patterns were also downloaded from the *M. truncatula* genome website. We obtained information on the chromosome locations of the putative *R2R3-MYB* genes from the results of BLASTP searches in the JCVI *M. truncatula* Genome Browser (JBrowse 1.11.1).

2.2. Sequence conservation analysis

Multiple sequence alignments of the 155 identified R2R3-MYB proteins in *M. truncatula* were conducted using MUSCLE with default parameters. To analyze the sequence features of the MYB domains of these proteins, the conserved region of each alignment was trimmed with Gblocks v0.91b, and BioEdit v7.0.5 was used to highlight conserved amino acids. The sequence logos for R2 and R3 repeats were generated using the Weblogo v3 Online Program (http://weblogo.threeplusone.com/).

The motifs conserved among MtR2R3-MYB members