UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

THE REGULATORY MECHANISM OF SECONDARY CELL WALL BIOSYNTHESIS IN GRASSES

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

KANGMEI ZHAO
Norman, Oklahoma
2016

THE REGULATORY MECHANISM OF SECONDARY CELL WALL BIOSYNTHESIS IN GRASSES

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF MICROBIOLOGY AND PLANT BIOLOGY

BY

_____
Dr. Laura E. Bartley, Chair

_____
Dr. John P. Masly

_____
Dr. Ben F. Holt III

_____
Dr. Jizhong Zhou

_____
Dr. Marc Libault

Dedication

This dissertation is dedicated with gratitude to my parents,

Jianguo Zhao and Junrong Kang,

and my husband,

Junyi Du.

Acknowledgements

It would be impossible for me to write this dissertation without the mentoring of Dr. Laura E. Bartley. About five years ago, I came to the U.S by myself and Dr. Bartley was the only person I knew in this country. Since then, I have always felt lucky to have her as my advisor, as she is very encouraging, kind and intelligent. During my graduate training with Dr. Bartley, besides the knowledge and skills that I accumulated, she helped me cultivate the ability to be curious, solve problems and think critically, which will benefit my entire career. I also appreciate her being a kind friend and helping me understand American culture, teach me baking, and being open to discussing life and philosophy. In all, Dr. Bartley is a great mentor, who supported me through my challenging and exciting graduate studies.

I also would like to acknowledge the support of my committee members. I took Dr. Masly's course, R programming, during my first semester at OU. He was extremely encouraging and patient when I was struggling with programming, statistics, as well as English. I appreciate his help and understanding. Dr. Zhou gave critical comments on the network project and provided great opportunities for me to communicate with his lab members, Dr. Deng Ye and Shi Zhou, who have expertise on bioinformatics. I learned valuable concepts and ideas from the courses on molecular biology and genomics taught by Dr. Holt and Dr. Libault, which helped to strengthen my background. I am also thankful for the time and effort of my committee members to improve this dissertation.

I am grateful to members in the Bartley lab and my fellow graduate students, Dr. Matt Peck, Sandra Thibivilliers, Fan Lin, Chengcheng Zhang, David Ponder, Zhenzhen

Qiao, Jin Yuan, Daniel Jones and Zack Mayers. I am very glad to know them and have the chance to work with them. I appreciate their support, understanding and encouragement.

Last but not least, my deepest gratitude is to my family. I appreciate the sincere love and support from my parents, Jianguo Zhao and Junrong Kang. They taught me to be an independent person, to understand myself, and to explore what I want to be. They always have faith in me and support me through hard times. I am also very lucky to have Junyi Du as my life companion. I appreciate his love and understanding. We are best friends and will hold hands to explore our lives.

Finally, I would like to acknowledge the funding resources that have supported my research and travel, namely the National Science Foundation EPSCoR program under the grant No. EPS-0814361, the Department of Energy Plant Feedstock Genomics Program under grant No. DE-SC0006904, the University of Oklahoma College of Arts and Sciences, and the Department of Microbiology and Plant Biology.

Table of Contents

List of Tables

List of Figures

Abstract

Grass cell walls are environmentally and economically important, including being an abundant and sustainable carbon source to produce lignocellulosic biofuels. However, the crosslinked structure of cell walls limits polysaccharide extraction efficiency, which is a bottleneck for biofuel production. Based on knowledge in Arabidopsis, multiple transcription factors from various protein families can regulate cell wall biosynthesis by forming a series of feed-forward loops. Diverged from dicotyledonous plants approximately 150 million years ago, grasses have evolved different cell wall components and vascular bundle patterning in vegetative organs. In this dissertation, I aimed to characterize transcription factors and corresponding DNA binding sites that control cell wall biosynthesis in grasses. I hypothesized that unstudied grass cell wall transcription factors might fall into the following three categories: (1) orthologs of known dicot cell wall regulators that have conserved functions in regulating the cell wall network; (2) uncharacterized cell wall-associated transcription factors that also likely maintain similar functions with those in dicots; (3) uncharacterized grass cell wall-associated transcription factors that do not exist or have different functions in dicots.

In Chapter 2, to analyze conservation and divergence between known dicot cell wall-associated transcription factors and their orthologs in grasses, we examined the phylogeny of R2R3 MYB protein family across selected dicots and grasses. Though we observed dicot-specific, grass-specific, and two panicoid grass-expanded clades, in general, most R2R3 MYBs that regulate SCW in Arabidopsis show evidence of conservation in the grasses.

In Chapter 3, we developed a Rice Combined mutual Ranked (RCR) network to identify regulators of grass-specific genes and other uncharacterized cell wall-associated transcription factors in grasses. The RCR network covers approximately 90% of the rice genome and shows high quality in GO-term-based evaluations. Network prediction and further molecular genetic validation suggest that OsMYB61a can directly or indirectly regulate grass cell wall-specific genes, among others. The RCR network includes a cell wall sub-network with 96 novel transcription factors. Eight out of eleven of them altered expression of cell wall-related genes in a transient gene expression assays in rice protoplast.

In Chapter 4, I further examined the conservation of cell wall-associated *cis*-elements in grasses using comparative *de novo* motif discovery and explored various scenarios for incorporation of grass-specific genes into cell wall biosynthesis pathways. Firstly, we observed that known dicots cell wall-associated *cis*-elements, such as MYB and NAC DNA binding sites, are significantly enriched within the promoters of *CESA*, lignin biosynthesis genes, as well as grass cell wall-specific genes. This provides support for the generally held hypothesis that known dicot cell wall-associated *cis*-elements are conserved in grasses. In addition, *cis*-elements that are potentially associated with AP2/ERF, C2H2, C2C2, and homeodomain proteins are also significantly enriched within promoters of grass cell wall biosynthesis genes. These results support the prediction and characterization of novel cell wall-associated transcription factors and binding sites. In all, this dissertation provides guidance toward functional characterization of cell wall-associated regulatory elements in grasses, knowledge of which will promote terrestrial biofuel production.

Chapter 1 : Introduction

Biomass and Biofuel Production

Production of lignocellulosic biofuels from sustainably produced biomass is a promising alternative to ameliorate dependence on fossil fuels, such as petroleum (Tilman et al., 2009; Binod et al., 2010; Feltus and Vandenbrink, 2012). Besides arguably reducing greenhouse gas emission, biofuels generally release fewer combustion air pollutants and provide local fuel production rather than relying on imports (Hill et al., 2009; Tilman et al., 2009). Despite these promises, petroleum products are still the major current suppliers of U.S. transportation energy (DOE 2007). Biofuels only play a minor role in the energy supply, accounting for less than 10% of the total energy (DOE, 2007). One of the major limitations for biofuel production is the relatively low fuel conversion efficiency resulting in uncompetitive prices compared to fossil fuels (Somerville, 2007; Tilman et al., 2009; Youngs and Somerville, 2012). One of the strategies to improve the biochemical conversion efficiency is to enhance the access of enzymes to polysaccharide (Tilman, et al., 2009). Thus, it is critical to better understand the genetic mechanisms of bioenergy feedstock traits.

In addition to algae, biomass from terrestrial plants is one of the main feedstocks being developed to produce biofuels. Biomass includes a wide range of plant materials, including wood, agricultural residues, and herbaceous energy crops (Bartley and Ronald, 2009). The total annual sustainable biomass production in the U.S. is estimated to be about 1.3 billion tons (DOE 2007). In particular, ~55% of biomass is estimated to be obtainable from cultivated grasses, including from cereal crops, such as rice, maize and sorghum (DOE 2006). In recent years, bioenergy grasses, which are members of the

Poaceae family, have attracted academic and industrial interests, including the following five C4 photosynthesis species: *Zea mays* (maize); *Saccharum* spp. (sugarcane); *Sorghum bicolor* (sorghum); *Miscanthus* spp. (Miscanthus); and *Panicum virgatum* (switchgrass). These species exhibit great potential to produce biomass, in some cases on degraded or low-quality lands (Feltus and Vandenbrink, 2012). To facilitate biofuel production, one of the keys is to understand the components and structure of grass cell walls.

Structure of Plant Cell Walls

Polysaccharide-rich cell walls are the bulk of plant dry mass and play an important role in land plant adaptation and diversification (Popper et al., 2011). The earliest land plants may have evolved 450 million years ago during the Late Ordovician based on the fossil evidence (Stewart and Rothwell, 1993). Diversified metabolites from different biological pathways mediate plants responses to the environment and play a key role in plant adaptation (Chae et al., 2014). Cell walls may have allowed early land plants to survive under various stresses, including exposure to UV B, lack of water for support, and co-evolution with herbivores and pathogens (Raven, 1984; Weng et al., 2010; Popper et al., 2011). Due to their location, plant cell walls are also critical for intercellular communication and defense against biotic and abiotic stresses (Keegstra, 2010).

Almost all plant cells are surrounded by flexible primary walls during growth. Based on composition and associated plant taxa, angiosperm primary walls can be further divided into type I and type II (Popper et al., 2011). Most dicots and non-

commelinid monocots possess type I primary walls. Type II wall are associated commelinid monocots, which appear to have evolved around 120 million years ago (Bremer et al., 2006; Vogel, 2008; Burton and Fincher, 2012; Lockhart, 2015). In all angiosperms, primary walls are composed of cellulose, hemicellulose and pectin (Vogel, 2008; Burton and Fincher, 2012; Carpita, 2012). Cellulose is the most abundant polysaccharide on earth and is composed of linear chains of β-(1-4)-linked glucose. Complexes of cellulose synthase A proteins, from the glycosyltransferase (GT) 2 family, synthesize cellulose microfibrils at the plasma membrane (Somerville, 2006).

Hemicellulose encompasses a group of heterogeneous polysaccharides that contribute roughly one third of cell wall biomass (Scheller and Ulvskov, 2010; Pauly et al., 2013). For type I cell walls of most dicots and non-commelinoid monocots the major hemicelluloses are xyloglucan, xylans and mannans. In contrast, type II walls of commelinoid monocots utilize arabinoxylan as the major hemicellulose. In addition, mixed-linkage glucan (MLG) is an abundant hemicellulose within species in the order Poales, but very rare in dicotyledonous and other monocotyledonous species (Burton et al., 2006; Scheller and Ulvskov, 2010; Schwerdt et al., 2015). Proteins from the GT43, GT47 and GT75 families form xylan in the Golgi body for subsequent vesicle-mediated release to the cell wall (Oikawa et al., 2010; Oikawa et al., 2013). A number of GTs synthesize an tetrameric oligosaccharide found at the reducing end of dicot and gymnosperm xylan (Rennie et al., 2014); however, the tetramer has not been detected in grasses nor have the functions of the putative homologs of the tetramer-synthesis enzymes been examined in grasses (Vogel et al., 2008).

Pectin is a galacturonic acid-rich plant cell wall polysaccharide, including homogalacturonan, rhamnogalacturonan I, and the substituted galacturonans rhamnogalacturonan II (RG-II) and xylogalacturonan (XGA) (Mohnen, 2008). As the most structurally complex cell wall polysaccharide, 67 transferases are expected to function in pectin biosynthesis including glycosyl-, methyl-, and acetyltransferases (Seifert 2004). Pectins make up to ~35% and less than 10% of primary cell wall in dicots and grasses, respectively (Mohnen, 2008).

Secondary cell walls (SCWs), composed of cellulose, hemicellulose and lignin, are mostly present in the following plant cell types: tracheary elements (e.g. tracheids and vessels), fibers, and sclereids (Mauseth, 1988; Zhong and Ye, 2001). Cellulose and hemicellulose are the major polysaccharide within SCWs. Lignin is an aromatic polymer from the phenylpropanoid pathway and only present in the SCW (Vanholme et al., 2008; Voxeur et al., 2015). Covalently cross-linked lignin represents a major barrier to utilizing cell wall polysaccharides. The structure of lignin in terms of degree of branching depends on the relative proportion of guaiacyl (G), syringyl (S), and in grasses, hydroxyphenyl (H) subunits, which have different numbers of methoxyl groups on the aromatic ring (Dixon et al., 2001; Harrington et al., 2012). A series of genes are responsible for the synthesis of monolignol precursors from phenylalanine and the down regulation of their gene expression can decrease lignin content, such as phenylalanine ammonium ligase, hydroxycinnamoyl-CoA:shikimate transferase, and cinnamyl alcohol dehydrogenase. Lignin composition (e.g. the ratio of H, S, and G monolignol) can also be engineered by manipulating the expression of genes coding for

ferulate 5-hydroxylase and caffeic acid methyltransferase and cinnamoyl-CoA reductase (Bonawitz and Chapple, 2010; Vanholme et al., 2013).

Grasses have evolved major changes on cell wall composition and crosslinking besides the difference in hemicellulose composition (Chaw et al., 2004; Vogel, 2008a). Cellulose synthase-like (CSL) genes from the GT2 family that are only present in Poaceae and a few alge taxa, are responsible for the synthesis of MLG, such as *CSLF6*, *CSLF8* and *CSLH1* (Burton et al., 2006; Vega-Sa´nchez et al., 2012; Scheller et al., 2010). We refer them as grass cell wall-diverged genes comparing to dicots. Another cell wall feature that is specific to grasses and other recently evolved commelinid monocotyledonous plants is that cell wall polymers are esterified by hydrocinnamic acids (HCAs), which are phenylpropanoids on the monolignol biosynthesis pathway, including ferulic acid (FA) and *p*-coumaric acid *(p*CA). BAHD-acyltransferases form a large protein family with versatile catalytic abilities (D'Auria, 2006; Bontpart et al., 2015). In plants, functionally characterized BAHD-ATs from different species across dicots and monocots and found that they can be divided into five clades based on phylogeny (D'Aurian 2006). Mitchell et al. proposed a subclade belonging to the BAHD-AT Clade V, that may incorporate HCAs into grass cell wall components and we refer it as "Mitchell Clade". For example, *Brachypodium* p-Coumaroyl-CoA:monolignol transferase (BdPMT) can acylate monolignols with *p*CA to form the precursor for ligninfication (Petrik et al., 2014). OsAT10 behaves as the *p*-coumaroyl coenzyme A transferase that may be involved in glucuronoarabinoxylan modification (Bartley et al., 2010). OsAT5 appears to act as a Ferulate Monolignol Transferase

(FMT), increasing feruloylation of monolignols in rice overexpression mutants (Karlen et al., Submitted).

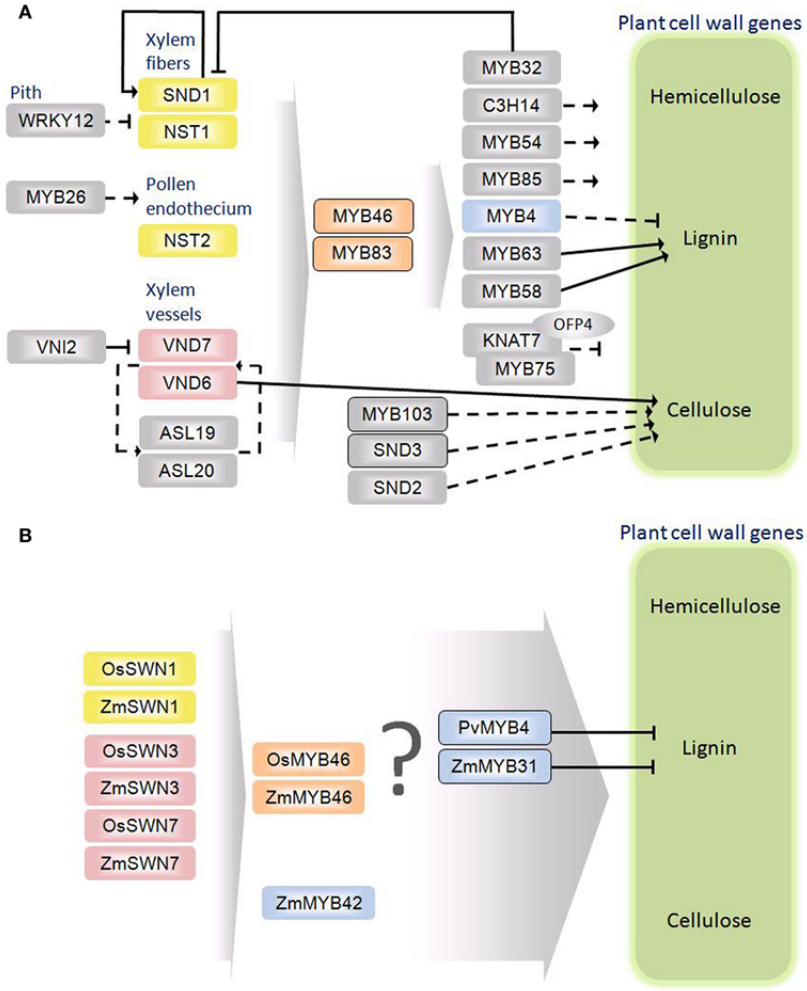


Figure 1-1 Schematic diagrams of the secondary cell wall regulatory networks in *Arabidopsis thaliana* (A) and grasses (B). Rectangles represent transcription factors. The oval indicates an interacting protein. Solid arrows and bordered rectangles signify evidence for direct interactions. Dashed arrows indicate no evidence for direct interaction. Orthology between A. thaliana and grasses is denoted by color (Handakumbura and Hazen, 2012). This figure has been published in Frontiers in Plant Science with open access, which allows referring in this dissertation.

Regulation of SCW Biosynthesis in Dicots and Grasses

In dicots, a series of transcription factors can regulate SCW biosynthesis and deposition in different tissues by forming a series of feed-forward loops, which top-level regulators can control both downstream transcription factors, as well as cell wall biosynthesis genes (Zhong et al., 2008; Taylor-Teeples et al., 2014). In Arabidopsis, 33 transcription factors from six protein families have been found to regulate SCW biosynthesis based on genetic or merely gene expression evidence (Figure 1-1). Among them, 11 and 16 members are from the NAC and R2R3 MYB protein families, respectively (Zhong et al., 2008; Zhao and Bartley, 2014; Zhong and Ye, 2014; Chai et al., 2015). A recent large-scale, yeast-one-hybrid screen identified 242 potential Arabidopsis transcription factors regulating root xylem SCW biosynthesis and expanded the number of protein families to 35; including over-representation by the AP2-EREBP, bHLH, C2H2, C2C2-GATA and GRAS transcription factor families (Taylor-Teeples et al., 2014). Some core Arabidopsis SCW regulators can control overall SCW biosynthesis genes as well as downstream regulators and result in dramatic phenotype in knockout or overexpression mutants. SECONDARY WALL-ASSOCIATED NAC PROTEIN/NAC SECONDARY WALL TRANSCRIPTION FACTOR (AtSND1/NST1) functions as a top-level SCW activator and controls overall SCW biosynthesis *via* the direct activation of downstream transcription factors (e.g. AtMYB46, AtMYB103, AtMYB58 etc.) as well as direct activation of cell wall biosynthesis genes (Mitsuda et al., 2005; Zhong et al., 2006; Zhong et al., 2007). Unique regulators control xylem vessel development, which is the major tissue for water and soluble mineral transportation. During xylem differentiation, VASCULAR-RELATED NAC-DOMAIN 6 and 7 (VND6/7) are key activators (Zhong

et al., 2008; Yamaguchi et al., 2010; Yamaguchi et al., 2011; Kondo et al., 2014; Zhou et al., 2014). Recently, Taylor-Teeples et al. (2014) identified that E2Fc, a member from the E2F DP family, acts upstream of AtVND6/7 as both activator and repressor based on the dosage in Arabidopsis root xylem. In addition, we still expect to identify additional transcription factors that mediate the response of cell wall biosynthesis pathways under hormonal and environmental stimuli.

Compared to SCW activation, less information is known about the repression of this pathway, which can alter cell wall composition or cell wall recalcitrance determined by the accessibility of sugars embedded in the wall (Carroll et al., 2009). In addition to AtWRKY12, other currently known SCW associated repressors are from the R2R3 MYB family, namely AtMYB4, AtMYB32 and AtMYB75. Orthologs of AtWRKY12 in *Miscanthus* and poplar behave as a pith cell wall formation repressor. AtMYB75 functions as a repressor of SCW biosynthesis and is also known as *PRODUCTION OF ANTHOCYANIN PIGMENT*1 (*PAP*1), with a role in positively regulating anthocyanin metabolism (Bhargava et al., 2010; Zhao and Dixon, 2011; Shin et al., 2013). AtMYB4 is a repressor of lignin biosynthesis and responses to ultraviolet B light (Jin et al., 2000). AtMYB4 has two paralogs, AtMYB32 and AtMYB7, which repress Arabidopsis pollen cell wall development and are down-regulated under drought stress, respectively (Jin et al., 2000; Preston et al., 2004; Ma and Bohnert, 2007). So far, it is not entirely clear how repressors switch-off SCW biosynthesis. The possible model may be that repressors can compete for downstream promoters of cell wall biosynthesis genes since different studies have shown that SCW activators and repressors from R2R3 MYB family may recognize similar DNA binding sites (Zhong et al., 2011). However,

we cannot rule out the possibility that repressors form dimers with activators to block their DNA binding ability.

So far, transcriptional control of grass cell wall biosynthesis has not been thoroughly examined (Handakumbura and Hazen, 2012; Hussey et al., 2013). Initial available data suggest that transcription factors responsible for the synthesis of cell wall major components (CESA, hemicellulose, and lignin) tend to be conserved in grasses in terms of binding specificity and *cis*-elements harbored within the promoters of grass cell wall biosynthesis genes (Zhong et al., 2011; Shen et al., 2012). For example, overexpression OsMYB46 or ZmMYB46 in Arabidopsis, orthologs of Arabidopsis core SCW activator MYB46, can promotes expression of CESA, lignin biosynthesis genes, and SCW-associated transcription factors in Arabidopsis (Zhong et al., 2011). In addition, PvMYB4, the ortholog of the Arabidopsis SCW repressor, AtMYB4M in switchgrass, can recognize known Arabidopsis MYB transcription factor binding sites, known as AC-elements, based on a yeast one-hybrid assay (Shen et al., 2012). In all, this suggests that dicots and grasses share a similar regulatory cascade to synthesize major cell wall components.

However, we still lack critical information about which transcription factors play predominant roles in grass cell wall biosynthesis. More importantly, no transcription factors are known to regulate grass cell wall-specific genes. Thus, we expect to characterize novel cell wall-associated transcription factors that can fall into the following two categories: (1) unstudied orthologs of Arabidopsis known SCW associated transcription factors in rice; (2) rice cell wall associated transcription factors that are not functionally examined on the regulation of cell wall biosynthesis pathway in

Arabidopsis. In this case, comparative phylogenomics and transcriptomics may identify conserved and grass-diverged SCW transcription factors by taking account of phylogeny across dicots and grasses, gene expression amounts, and gene network connections with grass cell wall-specific genes.

In this dissertation, I focus on identification of grass SCW regulators using phylogenetic analysis, genome-scale network and *de novo* comparative cell wall-associated *cis*-element discovery. Chapter 2 constitutes a comparative phylogenetic study of R2R3 MYB transcription factor families across five species, including dicots and grasses, to analyze the conservation and divergence of R2R3 MYBs during evolution. In Chapter 3, we constructed a novel, genome-scale network that is both more comprehensive and of similar or higher quality than existing networks. Taking advantage of the network, we addressed the following three questions: (1) what transcription factors are involved in grass cell wall biosynthesis; (2) do dicots and grass utilize the same group of predominant SCW regulators; (3) what regulators control the synthesis or incorporation of grass cell wall-specific components, such as HCAs and MLG. In Chapter 4, we applied comparative *de novo* motif analysis to predict potential DNA binding sites present within the promoters CESA and lignin biosynthesis genes and grass cell wall-specific genes. In Chapter 5, I summarized limitations of current studies and suggest strategies to further characterize novel cell wall-associated transcription factors in grasses, especially members controlling grass cell wall-specific features. In all, this work expands our understanding on transcriptional regulation of cell wall biosynthesis in grasses using rice as a model, which may further promote biofuel production.

References

Bartley, L.E., and Ronald, P.C. (2009). Plant and microbial research seeks biofuel production from lignocellulose. California Agriculture 63, 178-184.

Bartley, L.E., Peck, M.L., Kim, S.-R., Ebert, B., Manisseri, C., Chiniquy, D.M., Sykes, R., Gao, L., Rautengarten, C., Vega-Sánchez, M.E., Benke, P.I., Canlas, P.E., Cao, P., Brewer, S., Lin, F., Smith, W.L., Zhang, X., Keasling, J.D., Jentoff, R.E., Foster, S.B., Zhou, J., Ziebell, A., An, G., Scheller, H.V., and Ronald, P.C. (2013). Overexpression of a BAHD Acyltransferase, OsAt10, Alters Rice Cell Wall Hydroxycinnamic Acid Content and Saccharification. Plant Physiology 161, 1615-1633.

Bhargava, A., Mansfield, S.D., Hall, H.C., Douglas, C.J., and Ellis, B.E. (2010). MYB75 functions in regulation of secondary cell wall formation in the Arabidopsis inflorescence stem. Plant physiology 154, 1428-1438.

Binod, P., Sindhu, R., Singhania, R.R., Vikram, S., Devi, L., Nagalakshmi, S., Kurien, N., Sukumaran, R.K., and Pandey, A. (2010). Bioethanol production from rice straw: An overview. Bioresource Technology 101, 4767-4774.

Bonawitz, N.D., and Chapple, C. (2010). The genetics of lignin biosynthesis: connecting genotype to phenotype. Annu Rev Genet 44, 337-363.

Burton, R., and Fincher, G. (2012). Current challenges in cell wall biology in the cereals and grasses. Frontiers in Plant Science 3.

Burton, R.A., Wilson, S.M., Hrmova, M., Harvey, A.J., Shirley, N.J., Medhurst, A., Stone, B.A., Newbigin, E.J., Bacic, A., and Fincher, G.B. (2006). Cellulose Synthase-Like CslF Genes Mediate the Synthesis of Cell Wall (1,3;1,4)-ß-d-Glucans. Science 311, 1940-1942.

Carpita, N.C. (2012). Progress in the biological synthesis of the plant cell wall: new ideas for improving biomass for bioenergy. Current Opinion in Biotechnology 23, 330-337.

Chae, L., Kim, T., Nilo-Poyanco, R., and Rhee, S.Y. (2014). Genomic Signatures of Specialized Metabolism in Plants. Science 344, 510-513.

Chai, G., Kong, Y., Zhu, M., Yu, L., Qi, G., Tang, X., Wang, Z., Cao, Y., Yu, C., and Zhou, G. (2015). Arabidopsis C3H14 and C3H15 have overlapping roles in the regulation of secondary wall thickening and anther development. Journal of Experimental Botany.

Chaw, S.-M., Chang, C.-C., Chen, H.-L., and Li, W.-H. (2004). Dating the Monocot–Dicot Divergence and the Origin of Core Eudicots Using Whole Chloroplast Genomes. J Mol Evol 58, 424-441.

D'Auria, J.C. (2006). Acyltransferases in plants: a good time to be BAHD. Current Opinion in Plant Biology 9, 331-340.

Feltus, F., and Vandenbrink, J. (2012). Bioenergy grass feedstock: current options and prospects for trait improvement using emerging genetic, genomic, and systems biology toolkits. Biotechnology for Biofuels 5, 80.

Handakumbura, P.P., and Hazen, S.P. (2012). Transcriptional Regulation of Grass Secondary Cell Wall Biosynthesis: Playing Catch-Up with Arabidopsis thaliana. Front Plant Sci 3, 74.

Handakumbura, P.P., Matos, D.A., Osmont, K.S., Harrington, M.J., Heo, K., Kafle, K., Kim, S.H., Baskin, T.I., and Hazen, S.P. (2013). Perturbation of Brachypodium distachyon CELLULOSE SYNTHASE A4 or 7 results in abnormal cell walls. BMC Plant Biology 13, 1-16.

Harrington, M.J., Mutwil, M., Barrière, Y., Sibout, R., Jouanin, L., and Lapierre, C. (2012). Molecular biology of lignification in grasses. Advances in Botanical Research 61, 77-112.

Hussey, S.G., Mizrachi, E., Creux, N.M., and Myburg, A.A. (2013). Navigating the transcriptional roadmap regulating plant secondary cell wall deposition. Frontiers in Plant Science 4.

Jin, H., Cominelli, E., Bailey, P., Parr, A., Mehrtens, F., Jones, J., Tonelli, C., Weisshaar, B., and Martin, C. (2000). Transcriptional repression by AtMYB4 controls production of UV-protecting sunscreens in Arabidopsis. The EMBO journal 19, 6150-6161.

Karlen, S.D., Peck, M.L., Zhang, C., Smith, R.A., Padmakshan, D., Helmich, K.E., Free, H.C.A., Lee, S., Smith, B.G., Lu, F., Sedbrook, J.C., Sibout, R., Grabber, J.H., Runge, T.M., Mysore, K.S., Harris, P.J., Bartley, L.E., and Ralph, J. (Submitted). Monolignol Ferulate Conjugates are Naturally Incorporated into Plant Lignins. Science Advances.

Keegstra, K. (2010). Plant Cell Walls. Plant Physiology 154, 483-486.

Kondo, Y., Tamaki, T., and Fukuda, H. (2014). Regulation of xylem cell fate. Frontiers in Plant Science 5.

Lockhart, J. (2015). Uncovering the Unexpected Site of Biosynthesis of a Major Cell Wall Component in Grasses. The Plant Cell 27, 483.

Ma, S., and Bohnert, H. (2007). Integration of Arabidopsis thaliana stress-related transcript profiles, promoter structures, and cell-specific expression. Genome Biology 8, R49.

Mauseth, J.D. (1988). Plant anatomy. (Benjamin/Cummings Publ. Co.: Menlo Park, Calif).

Mutwil, M., Debolt, S., and Persson, S. (2008). Cellulose synthesis: a complex complex. Current Opinion in Plant Biology 11, 252-257.

Oikawa, A., Lund, C.H., Sakuragi, Y., and Scheller, H.V. (2013). Golgi-localized enzyme complexes for plant cell wall biosynthesis. Trends Plant Sci 18, 49-58.

Oikawa, A., Joshi, H.J., Rennie, E.A., Ebert, B., Manisseri, C., Heazlewood, J.L., and Scheller, H.V. (2010). An Integrative Approach to the Identification of Arabidopsis and Rice Genes Involved in Xylan and Secondary Wall Development. PLoS ONE 5, e15481.

Pauly, M., Gille, S., Liu, L., Mansoori, N., Souza, A., Schultink, A., and Xiong, G. (2013). Hemicellulose biosynthesis. Planta 238, 627-642.

Petrik, D.L., Karlen, S.D., Cass, C.L., Padmakshan, D., Lu, F., Liu, S., Le Bris, P., Antelme, S., Santoro, N., Wilkerson, C.G., Sibout, R., Lapierre, C., Ralph, J., and Sedbrook, J.C. (2014). p-Coumaroyl-CoA:monolignol transferase (PMT) acts specifically in the lignin biosynthetic pathway in Brachypodium distachyon. The Plant Journal 77, 713-726.

Popper, Z.A., Michel, G., Hervé, C., Domozych, D.S., Willats, W.G.T., Tuohy, M.G., Kloareg, B., and Stengel, D.B. (2011). Evolution and Diversity of Plant Cell Walls: From Algae to Flowering Plants. Annual Review of Plant Biology 62, 567-590.

Preston, J., Wheeler, J., Heazlewood, J., Li, S.F., and Parish, R.W. (2004). AtMYB32 is required for normal pollen development in Arabidopsis thaliana. The Plant Journal 40, 979-995.

Raven, J.A. (1984). Energetics and transport in aquatic plants. (AR Liss).
Scheller, H.V., and Ulvskov, P. (2010). Hemicelluloses. Annual Review of Plant Biology 61, 263-289.

Schwerdt, J.G., MacKenzie, K., Wright, F., Oehme, D., Wagner, J.M., Harvey, A.J., Shirley, N.J., Burton, R.A., Schreiber, M., Halpin, C., Zimmer, J., Marshall, D.F., Waugh, R., and Fincher, G.B. (2015). Evolutionary Dynamics of the Cellulose Synthase Gene Superfamily in Grasses. Plant Physiology 168, 968-983.

Seifert, G.J. (2004). Nucleotide sugar interconversions and cell wall biosynthesis: how to bring the inside to the outside. Curr Opin Plant Biol, 7, pp. 277–284.

Shen, H., He, X., Poovaiah, C.R., Wuddineh, W.A., Ma, J., Mann, D.G., Wang, H., Jackson, L., Tang, Y., and Neal Stewart Jr, C. (2012). Functional characterization of the switchgrass (Panicum virgatum) R2R3-MYB transcription factor PvMYB4 for improvement of lignocellulosic feedstocks. New Phytologist.

Shin, D.H., Choi, M., Kim, K., Bang, G., Cho, M., Choi, S.-B., Choi, G., and Park, Y.-I. (2013). HY5 regulates anthocyanin biosynthesis by inducing the transcriptional activation of the MYB75/PAP1 transcription factor in Arabidopsis. FEBS Letters 587, 1543-1547.

Somerville, C. (2006). Cellulose synthesis in higher plants. Annual review of cell and developmental biology 22, 53-78.

Somerville, C. (2007). Biofuels. Current Biology 17, R115-R119.

Somerville, C., Youngs, H., Taylor, C., Davis, S.C., and Long, S.P. (2010). Feedstocks for lignocellulosic biofuels. Science 329, 790-792.

Stewart, W.N., and Rothwell, G.W. (1993). Paleobotany and the evolution of plants. (Cambridge University Press).

Tilman, D., Socolow, R., Foley, J.A., Hill, J., Larson, E., Lynd, L., Pacala, S., Reilly, J., Searchinger, T., Somerville, C., and Williams, R. (2009). Beneficial Biofuels—The Food, Energy, and Environment Trilemma. Science 325, 270-271.

Vanholme, R., Morreel, K., Ralph, J., and Boerjan, W. (2008). Lignin engineering. Current Opinion in Plant Biology 11, 278-285.

Vanholme, R., Cesarino, I., Rataj, K., Xiao, Y., Sundin, L., Goeminne, G., Kim, H., Cross, J., Morreel, K., Araujo, P., Welsh, L., Haustraete, J., McClellan, C., Vanholme, B., Ralph, J., Simpson, G.G., Halpin, C., and Boerjan, W. (2013). Caffeoyl Shikimate Esterase (CSE) Is an Enzyme in the Lignin Biosynthetic Pathway in Arabidopsis. Science 341, 1103-1106.

Vogel, J. (2008a). Unique aspects of the grass cell wall. Curr. Opin. Plant Biol. 11, 301-307.

Vogel, J. (2008b). Unique aspects of the grass cell wall. Current Opinion in Plant Biology 11, 301-307.

Voxeur, A., Wang, Y., and Sibout, R. (2015). Lignification: different mechanisms for a versatile polymer. Current Opinion in Plant Biology 23, 83-90.

Weng, J.-K., Akiyama, T., Bonawitz, N.D., Li, X., Ralph, J., and Chapple, C. (2010). Convergent Evolution of Syringyl Lignin Biosynthesis via Distinct Pathways in the Lycophyte Selaginella and Flowering Plants. The Plant Cell 22, 1033-1045.

Withers, S., Lu, F., Kim, H., Zhu, Y., Ralph, J., and Wilkerson, C.G. (2012). Identification of Grass-specific Enzyme That Acylates Monolignols with p-Coumarate. Journal of Biological Chemistry 287, 8347-8355.

Yamaguchi, M., Mitsuda, N., Ohtani, M., Ohme-Takagi, M., Kato, K., and Demura, T. (2011). VASCULAR-RELATED NAC-DOMAIN 7 directly regulates the expression of a broad range of genes for xylem vessel formation. The Plant Journal 66, 579-590.

Yamaguchi, M., Goué, N., Igarashi, H., Ohtani, M., Nakano, Y., Mortimer, J.C., Nishikubo, N., Kubo, M., Katayama, Y., and Kakegawa, K. (2010). VASCULAR-RELATED NAC-DOMAIN6 and VASCULAR-RELATED NAC-DOMAIN7 effectively induce transdifferentiation into xylem vessel elements under control of an induction system. Plant physiology 153, 906-914.

Youngs, H., and Somerville, C. (2012). Development of feedstocks for cellulosic biofuels. F1000 biology reports 4, 10.

Zhao, K., and Bartley, L. (2014). Comparative genomic analysis of the R2R3 MYB secondary cell wall regulators of Arabidopsis, poplar, rice, maize, and switchgrass. BMC Plant Biology 14, 135.

Zhao, Q., and Dixon, R.A. (2011). Transcriptional networks for lignin biosynthesis: more complex than we thought? Trends Plant Sci 16, 227-233.

Zhong, R., and Ye, Z.-H. (2001). Secondary Cell Walls. In eLS (John Wiley & Sons, Ltd.

Zhong, R., and Ye, Z.-H. (2014). Complexity of the transcriptional network controlling secondary wall biosynthesis. Plant Science 229, 193-207.

Zhong, R., Lee, C., Zhou, J., McCarthy, R.L., and Ye, Z.H. (2008). A battery of transcription factors involved in the regulation of secondary cell wall biosynthesis in Arabidopsis. Plant Cell 20, 2763-2782.

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

Zhou, J., Zhong, R., and Ye, Z.-H. (2014). Arabidopsis NAC Domain Proteins, VND1 to VND5, Are Transcriptional Regulators of Secondary Wall Biosynthesis in Vessels. PLoS ONE 9, e105726.

Chapter 2 : Comparative genomic analysis of the R2R3 MYB secondary cell wall regulators of Arabidopsis, Poplar, Rice, Maize, and Switchgrass

Authors: Kangmei Zhao and Laura Bartley

Publication Status: This chapter has been published in BMC Plant Biology (doi:10.1186/1471-2229-14-135) with open access, which allows incorporation into this dissertation.

Authors Contribution: KZ and LEB conceived of and designed the study and wrote the manuscript. KZ carried out the analyses and created the figures. Both authors read and approved the final manuscript.

Abstract

R2R3 MYB proteins constitute one of the largest plant transcription factor clades and regulate diverse plant-specific processes. Several R2R3 MYB proteins act as regulators of secondary cell wall (SCW) biosynthesis in *Arabidopsis thaliana* (At), a dicotyledenous plant. Relatively few studies have examined SCW R2R3 MYB function in grasses, which may have diverged from dicots in terms of SCW regulatory mechanisms, as they have in cell wall composition and patterning. Understanding cell wall regulation is especially important for improving lignocellulosic bioenergy crops, such as switchgrass. Here, we describe the results of applying phylogenic, OrthoMCL, and sequence identity analyses to classify the R2R3 MYB family proteins from the annotated proteomes of Arabidposis, poplar, rice, maize and the initial genome (v0.0) and translated transcriptome of switchgrass (*Panicum virgatum*, Pv). We find that the R2R3 MYB proteins of the five species fall into 48 subgroups, including three dicot-specific, six grass-specific, and two panicoid grass-expanded subgroups. We observe four classes of phylogenetic relationships within the subgroups of known SCW-regulating MYB proteins between Arabidopsis and rice, ranging from likely one-to-one orthology (for AtMYB26, AtMYB103, AtMYB69) to no homologs identifiable (for AtMYB75). Microarray data for putative switchgrass SCW MYBs indicate that many maintain similar expression patterns with the Arabidopsis SCW regulators, though some of the switchgrass-expanded candidate SCW MYBs exhibit differences in gene expression patterns among paralogs consistent with subfunctionalization. Furthermore, some switchgrass representatives of grass-expanded clades have gene expression patterns consistent with regulating SCW development. Our analysis suggests that no

single comparative genomics tool is able to provide a complete picture of the R2R3 MYB protein family without leaving ambiguities, and establishing likely false-negative and -positive relationships, but that used together a relatively clear view emerges. Generally, we find that most R2R3 MYBs that regulate SCW in Arabidopsis are likely conserved in the grasses. This comparative analysis of the R2R3 MYB family will facilitate transfer of understanding of regulatory mechanisms among species and enable control of SCW biosynthesis in switchgrass toward improving its biomass quality.

Background

MYB proteins form one of the largest transcription factor families in plants. They regulate diverse processes including development, secondary metabolism, and stress responses (Du et al., 2009; Dubos et al., 2010). MYB proteins are typified by a conserved DNA binding domain consisting of up to four imperfect repeats (R) of 50 to 54 amino acids. Characterized by regularly spaced tryptophan residues, each repeat contains two α–helices that form a helix-turn-helix structure, and a third helix that binds the DNA major groove (Ogata et al., 1996; Dubos et al., 2010; Feller et al., 2011). MYB proteins are classified based on the sequence and number of adjacent repeats, with R1, R2R3, 3R and 4R proteins having one, two, three, and four repeats, respectively (Baranowskij et al., 1994; Jin and Martin, 1999; Kranz et al., 2001; Dubos et al., 2010). MYB proteins with one or more divergent or partial R repeat are classified as MYB-like or MYB-related (Riechmann et al., 2000). Two repeat domains, either covalently or non-covalently associated, appear to be necessary and sufficient for high-affinity DNA binding (Ogata et al., 1995).

In plants, the MYB R2R3 proteins are by far the most abundant of the MYB classes. R2R3 MYBs likely evolved from progenitor 3R MYB proteins by losing the R1 repeat (Rabinowicz et al., 1999). The family subsequently underwent a dramatic expansion after the origin of land plants but before the divergence of dicots and grasses (Rabinowicz et al., 1999; Dias et al., 2003; Chaw et al., 2004). The whole-genome complement of R2R3 MYB proteins has been investigated in several plant species, including Arabidopsis, rice (*Oryza sativa*), poplar (*Populus trichocarpa*), grapevine (*Vitis vinifera*), and maize (*Zea mays*), often with the goals of identifying orthologous

groups and species-diverged clades (Stracke et al., 2001; Yanhui et al., 2006; Wilkins et al., 2009; Du et al., 2012; Katiyar et al., 2012). The Arabidopsis genome encodes 126 R2R3 MYB proteins, most of which have been divided into 25 subgroups based on conserved motifs in the C-terminal protein regions (Stracke et al., 2001; Dubos et al., 2010). More recently, thirteen additional subgroups, for a total of 37 groups (G), were proposed based on comparative analysis of the R2R3 MYBs of *Arabidopsis* and maize (Du et al., 2012).

The function of R2R3 MYBs in regulating secondary cell wall (SCW) biosynthesis has garnered particular recent attention due to the importance of plant cell walls as a source of biomass for sustainable biofuel production (Bartley and Ronald, 2009; Youngs and Somerville, 2012). Secondary walls form around many cell types after cessation of plant cell growth. Genetic studies have clearly demonstrated that thickened and chemically cross-linked SCWs function in structural support, water transport, and stress resistance (Bonawitz and Chapple, 2010). SCWs are composed almost entirely of cellulose microfibrils encased by a network of glucurano-arabinoxylan and phenylpropanoid-derived lignin. Studies mostly undertaken in Arabidopsis, a eudicot, have shown that numerous R2R3 MYBs are part of the complex regulatory network controlling formation of SCWs (Zhong and Ye, 2007; Zhao and Dixon, 2011; Gray et al., 2012; Handakumbura and Hazen, 2012; Wang and Dixon, 2012). Figure 2-1 diagrams current understanding of the relationships among the 17 Arabidopsis R2R3 MYBs that have been identified so far to possibly function in SCW regulation. The network has multiple levels, though many higher-level regulators also directly regulate expression of genes encoding cell wall biosynthesis enzymes (Zhong

and Ye, 2007) (Figure 2-1). Table 2-1 summarizes the roles of individual Arabidopsis MYBs in SCW regulation and the initial forays into validating this regulatory network in grasses and poplar.

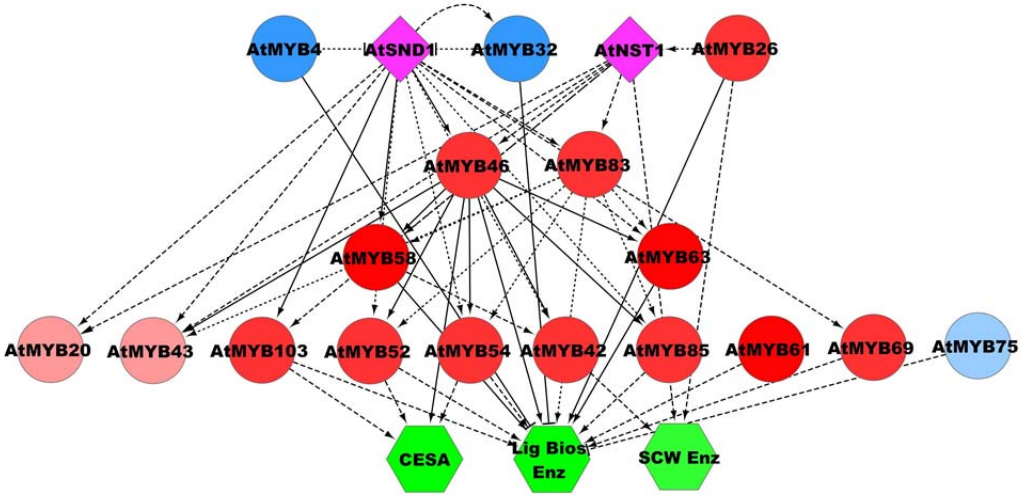


Figure 2-1 Transcription regulation network of Arabidopsis known secondary cell wall R2R3 MYB proteins. Pink and red symbols are positive regulators and blue are negative regulators. Nodes with darker shades show evidence of conservation in grasses that is absent for lighter shaded nodes (see text). MYBs are depicted by circles. Two crucial NAC-family transcriptional regulators, SND1, SECONDARY WALL-ASSOCIATED NAC DOMAIN PROTEIN1 and NST1, NAC SECONDARY WALL THINCKENING FACTOR 1, are depicted by diamonds. Other known regulators are excluded for simplicity (Zhong et al., 2010; Handakumbura and Hazen, 2012). Green hexagons represent genes that encode biosynthetic enzymes. Lig Bios Enz represents lignin biosynthesis enzymes, CESA is the cellulose synthases, and SCW Enz represents unspecified secondary cell wall synthesis enzymes. Solid edges represent direct interactions (i.e., evidence of physical promoter binding) and dashed edges represent indirect interactions (i.e., a change of gene expression with altered regulator expression). Indirect interactions may be direct, but not yet characterized. The figure was prepared with Cytoscape.

Table 2-1 Secondary cell wall (SCW)-associated R2R3 MYBs in dicots and grasses, organized based on phylogenetic tree topology.

Subgroup	Name	Function	Regulation and Phenotype	Reference
G29	AtMYB26	Activator	Overexpression results in ectopic induction of SCW thickening and lignification.	(Yang et al., 2007)
G30	AtMYB103	Activator	Loss of function mutant reduces syringyl lignin; Overexpression increases SCW thickening in fibers; Regulates pollen development.	(Higginson et al., 2003; Zhang et al., 2007; Öhman et al., 2012)
G21	AtMYB69	Activator	Dominant repression reduces SCW thickening in both interfascicular fibers and xylary fibers in stems.	(Zhong et al., 2008)
G31	AtMYB46	Activator	Dominant repression reduces SCW thickening of fibers and vessels; Overexpression mutant leads to ectopic deposition of secondary walls.	(Zhong et al., 2007; Zhong et al., 2008; Ko et al., 2009; Chiniquy et al., 2012; Zhong and Ye, 2012)
G31	AtMYB83	Activator	Functionally redundant with AtMYB46; Overexpression induces ectopic SCW deposition.	(McCarthy et al., 2009; Zhong and Ye, 2012) (Zhong and Ye, 2012)
G31	ZmMYB46	Activator	Overexpression in Arabidopsis induces ectopic deposition of lignin and xylan and an increases accumulation of cellulose in the walls of epidermis.	(Zhong et al., 2011)
G31	OsMYB46	Activator	Overexpression in Arabidopsis induces ectopic deposition of lignin and xylan and an increases accumulation of cellulose in the walls of epidermis.	(Zhong et al., 2011)
G31	PtrMYB20	Activator	Overexpression activates the biosynthetic pathway genes of cellulose, xylan and lignin.	(McCarthy et al., 2010)
G31	PtrMYB3	Activator	Overexpression activates the biosynthetic pathways genes of cellulose, xylan and lignin.	(McCarthy et al., 2010)
G8	AtMYB20	Activator	Activated by SND1 and NST1.	(Zhong et al., 2008)
G8	AtMYB43	Activator	Activated by SND1 and NST1.	(Zhong et al., 2008)
G8	AtMYB42	Activator	Activated by SND1 and NST1.	(Zhong et al., 2008)
G8	AtMYB85	Activator	Overexpression results in ectopic deposition of lignin in epidermal and cortical cells in stems; Dominant repression reduces SCW thickening in both stem interfascicular fibers and xylary fibers.	(Zhong et al., 2008)

Table 2-1 cont.,

G21	AtMYB52	Activator	Dominant repression reduces SCW thickening in both stem interfascicular fibers and xylary fibers.	(Zhong et al., 2008)
G21	AtMYB54	Activator	Dominant repression reduced SCW thickening in both stem interfascicular fibers and xylary fibers.	(Zhong et al., 2008)
G3.a	AtMYB58	Activator	Dominant repression reduces SCW thickening and lignin content; Overexpression causes ectopic lignification.	(Zhou et al., 2009)
G3.a	AtMYB63	Activator	Dominant repression reduces SCW thickening and lignin content; Overexpression causes ectopic lignification.	(Zhou et al., 2009)
G13.b	AtMYB61	Activator	Loss of function mutant reduces xylem vessels and lignification; Affects water and carbon allocation.	(Liang et al., 2005; Romano et al., 2012)
G4	AtMYB4	Repressor	Response to UV-B; Overexpression lines show white lesion in old leaves.	(Jin et al., 2000; Preston et al., 2004)
G4	AtMYB32	Repressor	Regulates pollen formation.	(Preston et al., 2004)
G4	ZmMYB31	Repressor	Overexpression reduces lignin content without changing composition.	(Fornalé et al., 2010)
G4	ZmMYB42	Repressor	Overexpression decreases S to G ratio of lignin.	(Sonbol et al., 2009; Fornalé et al., 2010)
G4	PvMYB4	Repressor	Overexpression represses lignin content.	(Shen et al., 2012)
G6	AtMYB75	Repressor	Represses lignin biosynthesis and cell wall thickening in xylary and interfascicular fibers.	(Bhargava et al., 2010)

Biomass from cereals and other grasses is of special interest as they constitute ~55% of the lignocellulosic material that can be sustainably produced in the U.S. (Perlack et al., 2005). Grass and eudicot SCWs have partially divergent compositions (Vogel, 2008; Handakumbura and Hazen, 2012; Bartley et al., 2014). In addition grasses and dicots have different patterns of vasculature, with its associated secondary wall, within leaves and stems. Grasses, as monocotoledenous plants, produce leaves

with parallel venation; whereas, dicot leaf venation is palmate or pinnate. In grasses with C4 photosynthesis, including maize and switchgrass, there is further cell wall thickening of the bundle sheath cells to support the separate phases of photosynthesis. Within stems, vascular bundles of dicots form in rings from the cambium; whereas, grass stems, which lack a cambium layer, exhibit a scattered (e.g., atactostele) pattern (Esau, 1977; Shen et al., 2009; Handakumbura and Hazen, 2012). Outside of the vasculature, the occurrence and patterning of extraxylary sclerenchyma cells, which are typified by thick cell walls, also varies between monocots and dicots (Esau, 1977). Grasses have, for example, a sclerenchyma layer circumscribing their root cortex that is absent in Arabidopsis and other dicots (Esau, 1977; Peret et al., 2009).

We postulate that the differences in composition and patterning of grass SCWs may have resulted in gains or losses of regulatory modules in grasses relative to dicots. The phylogenetic analysis of two dicots and three grasses presented here aims to refine this hypothesis. By comparing the R2R3 MYBs across diverse species, our goal is to identify conserved or expanded protein groups that may regulate grass SCW synthesis. Furthermore, examining the entire R2R3 MYB family will facilitate examination of MYB subgroups that regulate other important processes.

Our analysis is anchored on the relatively well-studied R2R3 MYBs of Arabidopsis (Dubos et al., 2010), which is in the eurosid I clade of eudicots (family Brassicaceae). We have also analyzed the angiosperm tree species poplar, which is an important species from an ecological context, is now used by the pulp and paper industry, and is also a major potential source of biomass for lignocellulosic biofuels. Poplar is in the family Salicaceae, which lies within the eurosid II clade, which shared a

common ancestor with Arabidopsis approximately 100 million years ago (Wang et al., 2009). The poplar genome has been sequenced for several years (Tuskan et al., 2006) and an early version was analyzed for R2R3 MYB content (Wilkins et al., 2009). To represent grasses, we have analyzed rice, maize, and switchgrass (*Panicum virgatum* L.). Rice is in the subfamily Erhardtoideae, whereas, maize and switchgrass are both in the Panicoideae (Kellogg, 2001). Rice was the first grass to have its genome sequenced (Matsumoto et al., 2005) and, among grasses, rice genomics and reverse genetic resources are arguably the best-developed (Jung et al., 2008). As a staple for about half of the human population, rice is an extremely important crop; consequently, its straw represents ~23% of global agriculture waste for which one potential use is lignocellulosic biofuels (Lal, 2005). Previous cataloging of rice R2R3 MYBs (Yanhui et al., 2006; Katiyar et al., 2012) had complementary foci to that presented here. Maize is also a very important food, feed, and first generation bioethanol crop with abundant genetic and genomic resources. Based on its recently sequenced genome (Schnable et al., 2009), Du et al. conducted a phylogenetic analysis of its R2R3 MYBs similar to that here and serving, in part, as validation. Lastly, we have examined the R2R3 MYB complement of the large-stature, C4 perennial grass, switchgrass, which is currently used for forage and in erosion control, and is being actively and widely developed as a bioenergy crop (McLaughlin and Adams Kszos, 2005; Bouton, 2007; Casler et al., 2011; Bartley et al., 2014). The tetraploid (1n=2x) genome size of lowlands and some upland switchgrass ecotypes is approximately 1.4 Mbp, which includes whole genome duplication approximately 1 million years ago (Lu et al., 2013). Switchgrass is an outcrossing species. In part due to the heterozygosity of the genome, a psuedomolecule

chromosomal assembly of the switchgrass genome was not available until recently (http:://www.phytozome.net/panicumvirgatum) (Casler et al., 2011).

Comparisons between model species, with their relatively small genomes, and non-models are often made more challenging due to whole genome and localized duplication events. To facilitate such translational science, multiple approaches have been developed for comparing the gene complement and genomic arrangement of whole genomes or particular biologically and economically relevant protein families (2013). Commonly employed methods include phylogentic analysis based on sequence alignments [e.g., (Wilkins et al., 2009; Du et al., 2012)], pair-wise quantitation of sequence identity [e.g., (Burton et al., 2006)], and more complex tools, like OrthoMCL [e.g., (Davidson et al., 2012; De Smet et al., 2013)]. Such approaches vary in their sophistication, underlying assumptions, and the level of time, attention, and bioinformatics acumen required. Another aim of this work is to analyze the apparent performance of commonly used tools at identifying individual genes for further study and manipulation.

Here, we present an investigation of the R2R3 MYB transcription factor family focusing on the non-model species switchgrass, using various comparative genomic approaches. We identified a total of 48 to 52 R2R3 MYB subgroups, most of which are common among all five species and similar to those previously described. Phylogenetic analysis reveals four patterns of conservation among proteins related to the known SCW R2R3 MYB regulators of Arabidopsis, ranging from one-to-one conservation between Arabidopsis and rice to unconserved between grasses and Arabidopsis, though most Arabidopsis SCW-regulating MYBs do appear to have orthologs in grasses. To clarify

which proteins from paralogous groups are more likely to act as functional orthologs, we also applied sequence identity and OrthoMCL analysis to the R2R3 MYB protein sequences. Moreover, switchgrass gene expression data provide evidence that particular paralogs are more likely to function in SCW regulation and that some novel grass-diverged MYB genes share similar expression patterns, illuminating avenues for improvement of economically important traits.

Results and discussion

Identification of R2R3 MYB proteins

R2R3 MYB proteins regulate diverse plant-specific processes, including secondary cell wall synthesis, stress responses, and development. To identify the R2R3 MYBs in the annotated genomes of poplar, rice, and maize, we used a Hidden Markov Model built from the R2R3 MYB proteins of Arabidopsis. We discarded identical sequences and loci that lack the complete R2R3 repeats following manual inspection and PROSITE characterization. Table 2-2 summarizes the number of unique putative R2R3 MYBs that we found in the genomes of each species, which are listed in Supporting Table 2-1. The species with smaller genomes, Arabidopsis and rice, possess similar numbers of R2R3 MYBs, whereas, organisms with larger genomes have greater numbers. Figure 2-2A and 2-2B show that our method may provide a more complete catalog of R2R3 proteins in rice and maize compared with recently published analyses (Du et al., 2012; Katiyar et al., 2012). The six sequences that Katiyar *et al*. identified from rice that are excluded from our list lack the R2R3 repeats compared with the PROSITE profile. The previous analysis in maize relied on BLASTP, which may be slightly less sensitive to distantly

related sequences (Finn et al., 2011). For poplar, Wilkins et al. (2009) identified 192 unique R2R3 MYBs, similar to the 202 that we were able to distinguish, and in keeping with the observation that poplar has undergone an enormous expansion in the number of R2R3 MYBs since its last common ancestor with Arabidopsis. The sequences used in the previous poplar analysis are not available, preventing a specific comparison with that work.

Table 2-2 R2R3 MYB proteins in analyzed species. Arabidopsis R2R3 MYB protein sequences were identified previously (Stracke et al., 2001).

Clade	Organism	Sequence Source	R2R3 MYBs
Eudicot	Arabidopsis	TAIR v.10	126
	Poplar	Phytozome v.3	202
Grass	Rice	Rice Genome Annotation v.7	125
	Maize	Phytozome v.2	162
	Switchgrass	Phytozome v.0.0 Switchgrass Functional Genomics Sever	230

For switchgrass, we combined the R2R3 MYBs that we identified from the annotated proteins in the DOE-JGI v0.0 genome with those from our translation of the unitranscript sequences available from the Switchgrass Functional Genomics Server. Figure 2-2C shows the distribution of the putative R2R3 MYBs from the two sources. Approximately twice as many proteins were identified from the translated unitranscripts than the v0.0 genome annotation. This is in part due to the fact that multiple genotypes were used to assemble the EST resource and about 10% of MYBs from the unitranscripts are attributed to the Kanlow cultivar. In addition, the presence of sequences within the genome that did not pass the protein annotation quality control

(see Methods) may decrease the protein complement of the v0.0 genome. That we identified more putative R2R3 MYBs from switchgrass than the other species likely reflects the recent whole genome duplication of switchgrass (Lu et al., 2013), though the total may be inflated by the heterozygous nature of the outcrossed genotypes sequenced and include alleles or unaligned splice-variants.

Comparative phylogenetic analysis of R2R3 MYB proteins in dicots and grasses

To examine broad conservation and divergence of R2R3 MYB proteins among the species examined, we inferred the phylogenetic relationships among the complete set of R2R3 MYB family proteins from Arabidopsis, poplar, rice, maize and switchgrass. We also accounted for the 25 published subgroups of Arabidopsis R2R3 MYB proteins and the more recently recognized 37 subgroups from a comparative analysis of R2R3 MYB family of Arabidopsis and maize (Stracke et al., 2001; Du et al., 2012). Proteins clustered in each subgroup of the phylogenetic tree frequently possess similar functions. On the other hand, general functions, such as regulation of specialized metabolism, are not isolated to specific or closely related subgroups. For example, characterized Arabidopsis R2R3 MYBs that regulate plant cell wall biosynthesis are spread among the subgroups G (or S) 3, G4, G6, G8, G13, G21 G29, G30, and G31 (Table 2-1).

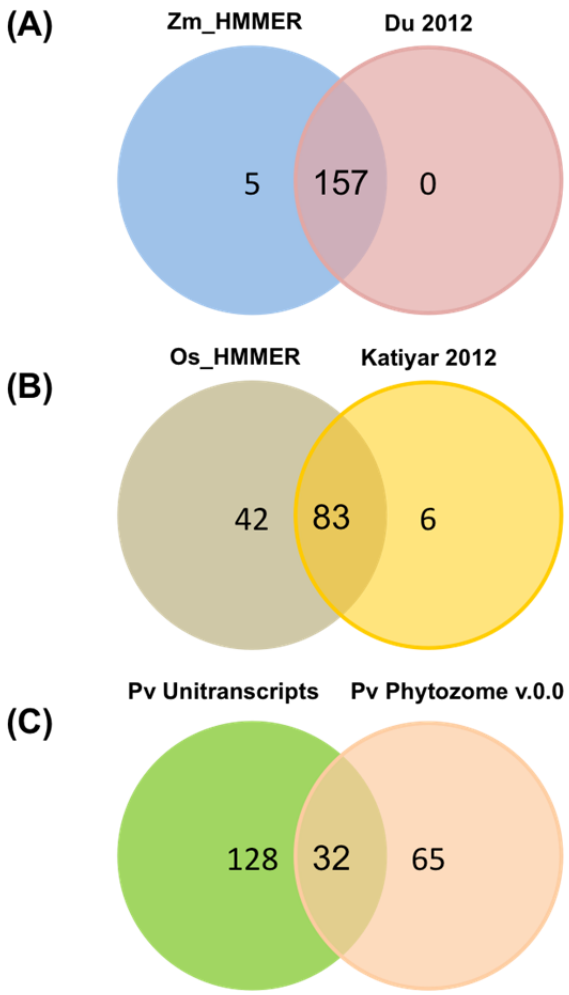


Figure 2-2 Summary of this study compared to previous ones on R2R3 MYBs and the source of switchgrass sequences. (A) Comparison of maize R2R3 MYB sequences identified here using HMMER and PROSITE prediction with previously published data from Du et al. 2012 (Du et al., 2012). (B) Comparison of rice R2R3 MYB sequences identified here using HMMER and PROSITE with published data from Katiyar et al. 2012 (Katiyar et al., 2012). (C) Sources of switchgrass R2R3 MYB family sequences used in this analysis.

We find that the R2R3 MYB proteins from the five species fall into approximately 48 subgroups (Table 2-3, Supporting Figure 2-1), with G38 to G48 emerging as novel groups in the five-species phylogeny. In addition, four of the previously described subgroups, G3, G13, G14 and G17, are poorly supported in our analysis and we have further subdivided them into a and b subclades. We identified three dicot-specific groups (G6, 10, 15) and six grass-specific groups (G27, G32, G35, G43, G45, G46) plus G3.b. These non-conserved groups likely evolved after the divergence of eudicots and grasses 140 to 150 million years ago (Rabinowicz et al., 1999; Dias et al., 2003; Chaw et al., 2004). In addition, poplar possesses four unique subgroups (G38, G39, G40, G48). Previous analysis showed that whole genome duplication and R2R3 MYB-specific expansions contributed to the evolution of MYBs in poplar (Wilkins et al., 2009). Though difficult to compare directly, Wilkins et al. did identify 6 subgroups in poplar that were not shared with Arabidopsis (Wilkins et al., 2009). We also find continued support for an Arabidopsis-specific subgroup, G12, which regulates glucosinolate biosynthesis and metabolism (Gigolashvili et al., 2007; Gigolashvili et al., 2008).

Table 2-3 Subgroups of R2R3MYB proteins from Arabidopsis (At), poplar (Ptr), rice (Os), maize (Zm) and switchgrass (Pv) defined by neighbor-joining phylogenetic reconstruction. Assignment to a subgroup is based on the 5-species neighbor-joining tree with 500 bootstraps (Supporting Figure 2-1). The brown and grey shading indicates grass- and dicot-diverged clades, respectively. The blue shading indicates an Arabidopsis-specific clade, orange poplar-specific clades, and purple a rice-specific clade. Open boxes indicate expansion in the Panicoid grasses, maize and switchgrass, relative to the other species. C-terminal conserved motifs were analyzed using MEME for each subgroup and compared to known motifs present in the 25 subgroups of Arabidopsis R2R3 MYB family. I: Previously identified; P: Partially previously identified; N: Not previously identified. The last column lists the Arabidopsis (At) secondary cell wall (SCW) regulators by their numeric names.

Subgroup	Bootstrap Score	At	Ptr	Os	Zm	Pv	Previous C-Terminal Motif Identification	Names of At SCW Regulators (AtMYB#)
G1	66	5	4	7	12	14	I	0
G2	37	3	4	3	5	8	P	0
G3.a	3	4	2	1	1	2	P	58, 63
G3.b	46	0	0	2	5	6	N	0
G4	14	6	7	8	10	22	I	5
G5	13	1	9	2	2	1	N	0
G6	7	4	8	0	0	0	I	75
G7	17	2	1	2	5	4	N	0
G8	89	4	6	5	8	17	P	20, 43, 42, 85
G9	51	2	4	3	4	7	P	0
G10	100	2	3	0	0	0	P	0
G11	92	4	6	1	2	0	I	0
G12	26	6	0	0	0	0	I	0
G13.a	21	1	2	1	2	5	P	0
G13.b	5	4	7	5	7	10	P	61
G14.a	33	2	5	2	2	1	N	0
G14.b	43	6	8	8	11	22	N	0
G15	39	4	5	0	0	0	I	0
G16	30	3	2	3	3	8	P	0
G17.a	93	2	2	3	5	4	N	0
G17.b	86	3	5	3	4	3	P	0
G18	7	7	5	2	2	3	N	0
G19	59	3	2	1	0	0	N	0
G20	88	6	8	5	13	10	I	0
G21	20	8	13	5	8	14	I	52, 54, 69

Table 2-3 cont.,

G22	62	4	6	3	5	12	P	0
G23	98	3	1	1	1	2	N	0
G24	88	3	4	3	3	5	I	0
G25	29	7	6	5	4	8	I	0
G26	80	1	4	2	3	0	N	0
G27	63	0	0	2	3	1	N	0
G28	25	1	7	1	1	0	N	0
G29	40	2	5	2	2	3	N	26
G30	100	1	2	1	1	2	N	103
G31	99	2	4	1	1	2	N	46, 83
G32	100	0	0	1	5	1	N	0
G33	100	1	3	1	3	4	N	0
G34	100	1	3	0	1	0	N	0
G35	42	0	0	2	4	6	N	0
G36	25	0	2	2	2	3	N	0
G37	100	2	2	1	1	1	N	0
G38	13	0	7	0	0	0	N	0
G39	86	0	3	0	0	0	N	0
G40	100	0	4	0	0	0	N	0
G41	100	1	5	7	3	0	N	0
G42	100	0	0	1	0	3	N	0
G43	75	0	0	2	1	5	N	0
G44	99	0	0	7	0	0	N	0
G45	100	0	0	1	1	3	N	0
G46	97	0	0	2	3	7	N	0
G47	21	0	5	0	1	0	N	0
G48	37	0	4	0	0	0	N	0

With MEME, we found that many of the subgroups designated in our analysis possess conserved C-terminal motifs, often supporting and extending those initially identified in the Arabidopsis R2R3 MYB subgroups (Table 2-3, Supporting Table 2-2). Located downstream of the N-terminal MYB DNA-binding domains, C-terminal motifs have been hypothesized to contribute to the biological functions of R2R3 MYB proteins (Stracke et al., 2001; Dubos et al., 2010). For example, the C-terminal motif, LNL[ED]L, of AtMYB4, found to be conserved in the analysis presented here, is

required for repression of the transcription at target promoters (Supporting Table 2-2) (Jin et al., 2000). The large number of sequences in our analysis apparently improved our sensitivity allowing identification of many motifs that were not apparent previously, including those of subgroup G23 (Stracke et al., 2001), and candidate motifs within the new subgroups (Supporting Table 2-2). Of the 25 original R2R3 MYB family subgroups of Arabidopsis, we found that all but 7 (G3.b, G5, G14.a and G14.b, G17.a, G18, G19 and G22) contain the same or similar motifs as identified previously in the corresponding Arabidopsis subgroups (Table 2-3, Supporting Table 2-2). Differences in identified motifs may stem from uncertainties in the subgroup designations. For the subgroups with different conserved motifs, two of them, G19 and G22, have bootstrap values higher than 50 in the five species phylogenetic tree; whereas, the phylogenies of subgroups G5 and G18, are poorly supported. The subdivided subgroups had variable effects on the identified motifs. Subgroups G3.a (but not G3.b) and G17.b (but not G17.a) possess the previously identified motifs. Both subgroups G13.a and .b contain the previously identified motif. In contrast, the original motif is not identifiable in either G14.a or .b.

Identification of putative orthologs of Arabidopsis SCW MYB across different species

To identify the putative SCW-associated R2R3 MYB proteins from each species, we performed a more focused analysis of the subgroups containing the known Arabidopsis SCW MYBs. For this, we identified related proteins from the multi-species neighbor-joining tree (as corroborated by dual Arabidopsis-other species trees), grouped closely related subgroups together, realigned these sequences, and inferred maximum

likelihood phylogenies. The results are summarized in Figures 2-3 to 2-7 and Table 2-4. We have sorted the R2R3 SCW MYB clades into four classes by comparing the relationships between the proteins of Arabidopsis and rice—the species with the smallest genomes. The classes are as follows: one-to-one relationships (class I), duplication in Arabidopsis and both of them are SCW regulators (class II), expansion in Arabidopsis with non-SCW R2R3 MYBs (class III), and no orthologs identifiable in the grasses examined (class IV). In addition to the in-depth phylogenetic analysis, we used OrthoMCL and sequence identity as alternatives for identifying orthologous groups of R2R3MYB proteins from the five species. OrthoMCL groups putative orthologs and paralogs based on BLAST scores across and within species and then resolves the many-to-many orthologous relationships using a Markov Cluster algorithm (Li et al., 2003). We analyzed sequence identity using alignments built with MUSCLE, which combines progressive alignment and iterative refinement (Edgar, 2004). Table 2-4 summarizes the results of all of these analyses.

To gain further support for our tentative identification of switchgrass SCW R2R3 MYBs, we examined their patterns of expression, as available, using the switchgrass gene expression atlas (Zhang et al., 2013). Of particular relevance, that study included gene expression of internode 4 of tillers at elongation stage 4, which is informative for the investigation of secondary development and recalcitrance in stem tissues (Shen et al., 2009; Saha et al., Submitted).

Table 2-4 Groups of homologous proteins from poplar, rice, maize and switchgrass relative to the Arabidopsis R2R3 MYB secondary cell wall (SCW) regulators. Classes and divisions among proteins are based on maximum likelihood phylogenic reconstruction. Shading indicates putative orthologous and paralogous relationships based on OrthoMCL. The boxes indicate Arabidopsis MYBs implicated in functions besides SCW regulation with higher sequence identity to proteins from the other species compared with the SCW MYBs in the same clade.

Class	Arabidopsis	Poplar POPTR_00	Sequence Identity (%)	Rice LOC_Os	Sequence Identity (%)	Maize GRMZM	Sequence Identity (%)	Switchgrass	Sequence Identity (%)
I	AtMYB26	01s20370	47	01g51260	45	2G0887834	45	AP13ISTG69224	44
I	AtMYB103	03s13190	60	08g05520	50	2G325907	48	AP13CTG15561	51
		01s09810	62					AP13ISTG58495	50
I	AtMYB69	07s04140	53	11g10130	47	5G803355	48	Pavirv00031864m	50
				05s06410	53			Pavirv00029353m	50
								Pavirv00020802m	49
II	AtMYB46[a]	PtrMYB3	58 [a]	OsMYB46	47 [a]	ZmMYB46	49 [a]	AP13ISTG55479	50 [a]
	AtMYB83[a]	PtrMYB20	57 [a]					AP13ISTG55477	51 [a]
		09s05860	53 [a]						
		01s26590	54 [a]						
II	AtMYB20[a]	04s08480	58 [a]	09g23620	54 [a]	2G169356	55 [a]	Pavirv00023586m	69 [a]
	AtMYB43[a]	17s02850	58 [a]	08g33150	56 [a]	2G126566	52 [a]	KanlCTG16207	53 [a]
								AP13ISTG67468	51 [a]
								Pavirv00053167m	60 [a]
								AP13ISTG57686	56 [a]
								Pavirv00069978m	56 [a]
								Pavirv00023587m	53 [a]
								Pavirv00051815m	57 [a]
								Pavirv00011866m	57 [a]

Table 2-4 cont.,

II	AtMYB42[a]	03s11360	61 [c]	09g36250	51 [a]	2G104551	52 [c]	AP13ISTG65795	52 [c]
	AtMYB85[c]	01s07830	61 [a]			2G138427	53 [c]	AP13CTG22878	52 [c]
		15s14600	55 [c]			2G037650	52 [c]	AP13CTG08064	53 [c]
		12s14540	57 [c]						
II	AtMYB52[a]	17s04890	55 [a]	03g51110	52 [a]	2G455869	53 [a]	AP13ISTG34280[c]	59 [a]
	AtMYB54[c]	15s05130	57 [c]			2G077147	52 [a]	AP13ISTG43780[c]	54 [c]
		12s03650	58 [c]					Pavirv0004892m[c]	54 [c]
		07s01430	53 [a]					Pavirv0004891m[c]	55 [c]
								Pavirv0000561 0m	52 [c]
III	AtMYB58[a]	07s08190	48 [a]	02g46780	49 [a]	5G833253	46 [a]	Pavirv0005504 5m	47 [c]
	AtMYB63[c]	05s09930	48 [a]	04g50770	48 [a]	2G097636	47 [a]	AP13ISTG56055	38 [a]
						2G097638	50 [a]	Pavirv0001995 0m	49 [a]
						2G038722	47 [a]	Pavirv0004704 0m	51 [c]
								AP13ISTG56056	49 [a]
								Pavirv0005341 5m	50 [a]
III	AtMYB61[a]	05s00340	53 [a]	05g04820	57 [c]	2G127490	56 [c]	AP13CTG04029	56 [c]
	AtMYB50[c]	13s00290	60 [a]	01g18240	57 [c]	2G171781	56 [c]	Pavirv0004249 5m	56 [c]
	AtMYB55[c]	02s18700	56 [c]			2G017520	56 [c]	Pavirv0002146 7m	56 [c]
		14s10680	57 [c]					Pavirv0003567 9m	58 [c]
								Pavirv0004131 2m	58 [c]
III	AtMYB4[a]	05s11410	67 [a]	09g36730	68 [a]	2G000818	75 [a]	AP13ISTG73550	68 [a]
	AtMYB32[c]	09s13640	66 [a]	08g43550	56 [c]	ZmMYB31	65 [a]	AP13ISTG73836	70 [a]
		04s18020	70 [a]			ZmMYB42	65 [a]	PvMYB4.a[c]	64 [a]
						2G084583	66 [a]	PvMYB4.b[c]	64 [a]
								PvMYB4.c[c]	64 [a]

Table 2-4 cont.,

				PvMYB4.d [a]	64 [a]
				PvMYB4.e [a]	64 [a]
IV	AtMYB75	05s14450	67 [a]		
	AtMYB90	05s14460	67 [a]		
	AtMYB113	05s14470	67 [a]		
	AtMYB114 [a]	05s14480	70 [a]		
		05s14490	72 [a]		

[a, b, c] Indicate proteins with highest sequence identity to the indicated Arabidopsis MYB.
[d]: MYBs that have ≥99% protein sequence similarity that are likely allelic to each other.

Class I: One-to-one relationships

Proteins in Class I show one-to-one conservation between Arabidopsis, rice, and maize and relatively modest expansion in poplar and switchgrass compared with other classes. The group consists of AtMYB26, AtMYB103 and AtMYB69 (Figures 2-3 and 2-5). For these and other classes, it remains a formal possibility that duplication and gene loss have occurred in other species relative to Arabidopsis resulting in pseudo-orthologs (Koonin, 2005). However, for the proteins in Class I, the expression patterns of the putative switchgrass orthologs support the hypothesis of conservation of function.

The only SCW MYB protein group with evidence of one-to-one conservation without duplication among all five species is those related to AtMYB26, which is also called *MALE STERILE35* (*MS35*). AtMYB26 was unclassified in the original subgroup analysis (Stracke et al., 2001) and is a member of the small subgroup, G29 (Du et al., 2012). AtMYB26 is a high-level activator of SCW thickening in anthers, functioning in the critical process of pollen dehiscence (Yang et al., 2007). Ectopic expression of A*tMYB26* upregulates *NST1* and *NST2* and causes SCW thickening, especially in epidermal tissues (Yang et al., 2007). We found one putative ortholog of AtMYB26 in each species, suggesting that the critical function of MYB26 in reproduction may be conserved across evolution (Figure 2-3). Consistent with this, *AP13ISTG69224*, the putative switchgrass ortholog of *AtMYB26*, is lowly expressed in the stems (i.e., node and internode samples) and leaves at the E4 (elongation 4) stage, but more highly expressed in the inflorescence (Figure 2-9). The absence of duplication in switchgrass is unexpected given its recent genome duplication and likely reflects the incomplete

genome sequence. On the other hand, sequence identity between AtMYB26 and its putative orthologs in grasses is relatively low, ~45%. Possibly due to that fact, OrthoMCL analysis did not identify AtMYB26 orthologs (Table 2-4). This amount of variation is consistent with divergence within this clade since the last common ancestor and sheds some doubt on the supposition of conservation of function in the absence of experimentation.

The other two clades included in Class I are those of AtMYB103 and AtMYB69, from subgroups G30 and G21, respectively. In Arabidopsis, these proteins are lower-level SCW activators, regulated by At*SND1* (Zhong et al., 2008). *AtMYB103* is mainly expressed in the stem, where cells are undergoing secondary wall thickening (Zhong et al., 2008). *AP13ISTG58495* also has high expression levels in the vascular bundle and internodes (Figure 2-9). Thus, both phylogentic analysis and gene expression are consistent with maintenance of the function of these proteins across grasses and eudicots. Sequence identity between AtMYB103 and the putative grass orthologs is intermediate, ranging from 48% to 51%, and OrthoMCL mostly supports the phylogenetic analysis, further evidence that *AP13ISTG58495* may be a SCW regulator in switchgrass (Table 2-4). In rice, a preliminary study reported that RNAi lines of *OsMYB103* show a severe dwarf phenotype and did not grow to maturity (Hirano et al., 2013); whereas, only altered tapetum, pollen and trichome morphology were observed in Arabidopsis *AtMYB103* silencing mutants (Higginson et al., 2003; Öhman et al., 2012).

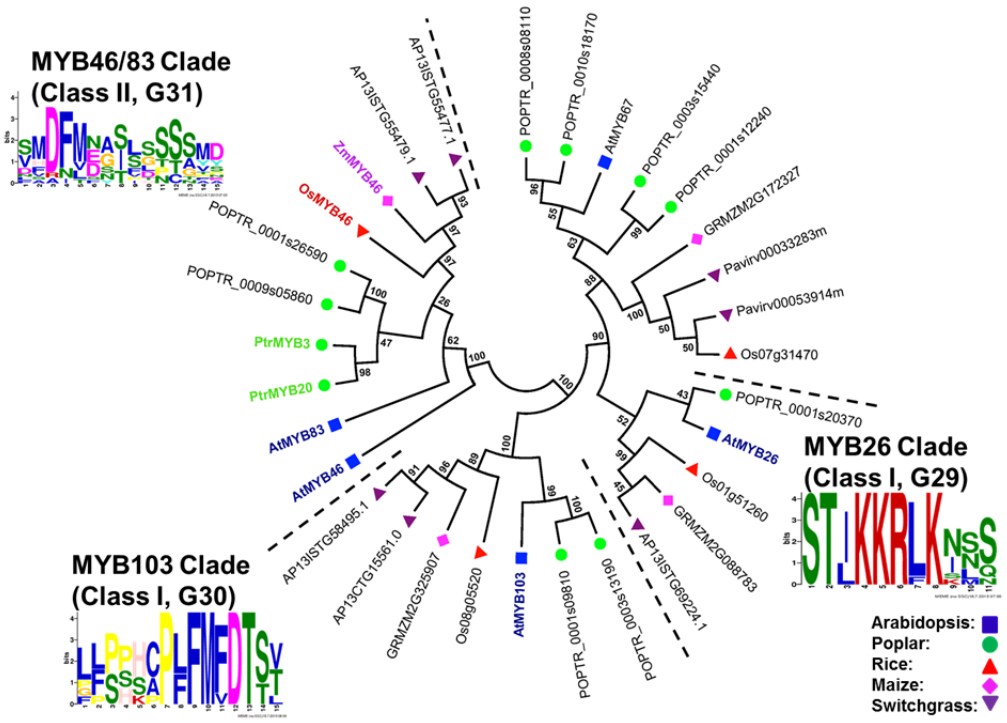


Figure 2-3 Maximum likelihood phylogenetic analysis of subgroups G29, G30, and G31 suggests that the function of the secondary cell wall (SCW) regulators, MYB46, MYB83, MYB103, and MYB26, are conserved between grasses and Arabidopsis. Poplar and switchgrass show gene duplication in the MYB46/83 and MYB103 clades. MYB proteins represented with bold and colored text are characterized SCW regulators in each species. Support values are from 1000 bootstrap analyses. Each logo is the C-terminal conserved motifs with the lowest E-value identified for the subgroup.

This difference in phenotypes caused by expression disruption of apparently orthologous genes between rice and Arabidopsis suggests differences in the SCW regulatory network between grasses and dicots not obvious from the phylogenetic relationships of the Class I proteins. For *AtMYB69*, of the three putative switchgrass co-orthologs, OrthoMCL identifies only *Pavirv00031864m* as an ortholog. These two proteins have 50% pairwise sequence identity and are similarly related to two other proteins in switchgrass (Table 2-4). No gene expression data for the three switchgrass co-orthologs are available to help resolve the question of whether there may be subfunctionalization in this family in switchgrass.

Class II: SCW related co-orthologs in Arabidopsis

R2R3 MYB proteins in Class II underwent duplication in the Arabidopsis lineage, though the duplicates have apparently retained roles in regulating SCW biosynthesis. This class consists of AtMYB46 and AtMYB83, AtMYB42 and AtMYB85, AtMYB52 and AtMYB54, and AtMYB20 and AtMYB43.

AtMYB46 and AtMYB83, from subgroup G31, function redundantly to activate SCW biosynthesis (McCarthy et al., 2009). AtMYB46 directly activates several genes related to cell wall synthesis and regulation, including *CESA*s, *AtMYB58*, *AtMYB63* and *AtMYB43* (Chiniquy et al., 2012; Zhong and Ye, 2012). Dominant repression of AtMYB46 reduces SCW accumulation, and simultaneous RNA interference of *AtMYB*46 and *AtMYB*83 deforms vessel and fibers (Zhong et al., 2007; McCarthy et al., 2009). Figure 2-3 shows the maximum likelihood phylogeny for this and closely related proteins and provides evidence that this group is part of a well-supported clade of likely

co-orthologs. Consistent with this, functional data on the named poplar proteins and the rice and maize co-orthologs show that these proteins phenocopy *AtMYB46* and *AtMYB83* when heterologously expressed in Arabidopsis (McCarthy et al., 2010; Zhong et al., 2011). We found two putative co-orthologs of *AtMYB46* and *AtMYB83* in switchgrass, *AP13ISTG55479* and *AP13ISTG55477*, which are likely regulators of SCW biosynthesis (Figure 2-3). *AtMYB46* and *AtMYB83* are predominantly expressed at the sites of SCW synthesis—interfascicular fibers, xylary fibers, and vessels (Zhong et al., 2007; Ko et al., 2009; McCarthy et al., 2009; Chiniquy et al., 2012). *AP13ISTG55479* and *AP13ISTG55477* also show relatively high expression in stems (Figure 2-9), with *AP13ISTG55477* being the more highly expressed of the two. OrthoMCL supports the orthologous relationship of grass MYB46-like proteins; however, the dicot sequences of the MYB46 clade do not cluster with those of the grasses, possibly due to the somewhat low sequence identity (47% to 50%; Table 2-4). The other three Class II R2R3 MYB protein pairs are AtMYB42 and AtMYB85, and AtMYB20 and AtMYB43, from subgroup G8 (Figure 2-4); and AtMYB52 and AtMYB54 from subgroup G21 (Figure 2-5). These genes are expressed mainly in stems and specifically, in tested cases, in fiber and xylem cells and downregulated in a line silenced for *AtSND1* and *AtNST1* (Zhong et al., 2008). Overexpression of *AtMYB85, AtMYB52*, or *AtMYB54* (but not of *AtMYB42, AtMYB20*, or *AtMYB43*) leads to ectopic deposition of lignin in epidermal and cortical cells in stems (Zhong et al., 2008). Moreover, RNAi of *OsMYB42/85* (*LOC_Os09g36250*) causes a severe dwarf phenotype (Hirano et al., 2013). The maximum likelihood phylogenic trees of each of these Arabidopsis protein pairs contains one or two rice proteins, one to three maize proteins

and two or more poplar proteins (Figure 2-4, Figure 2-5, Table 2-4). The OrthoMCL result for AtMYB42, AtMYB85, AtMYB52 and AtMYB54 largely supports the phylogenetic topology, though excludes paralogs from poplar and maize (Table 2-4). OrthoMCL analysis separates AtMYB20 and AtMYB43 into different groups and identifies proteins in switchgrass as (co-)orthologs for each of these (Table 2-4). Among the switchgrass genes in Class II, *AP13CTG22878* and *AP13ISTG65795*, co-orthologs of *AtMYB42* and *AtMYB85*, are also highly expressed in stems, consistent with conservation of function in SCW regulation and providing no evidence of subfunctionalization (Figure 2-9). In contrast, co-orthologs of *AtMYB20* and *AtMYB43*, namely *AP13ISTG67468*, *KanlCTG16207* and *AP13ISTG57686*, are all expressed at low levels. No expression data are available for the switchgrass genes encoding AtMYB52 and AtMYB54, four out of five of which may be putative alleles of each other due to high sequence identity (>99%; Table 2-4). In sum, though much of the phylogenic data are consistent with conserved function of other Class II proteins, for the three co-orthologs of AtMYB20 and AtMYB43, as well as the initial Arabidopsis genetic data, call into question the function of these proteins in SCW regulation.

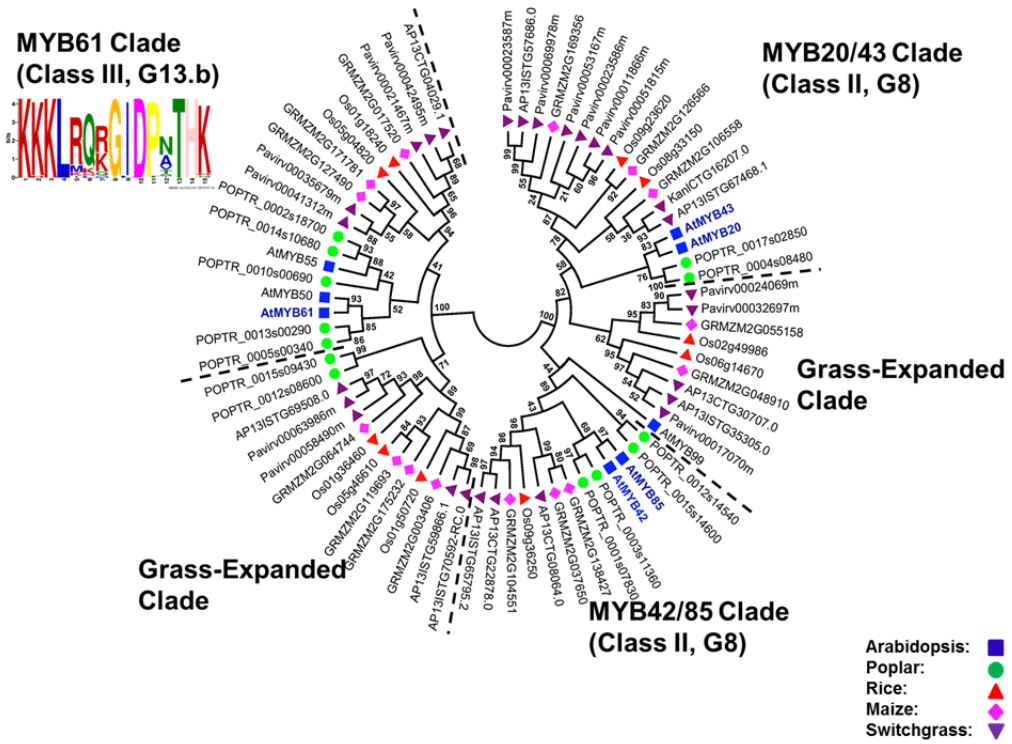


Figure 2-4 Maximum likelihood phylogenetic analysis of subgroups G8 and G13.b suggests gene duplication in dicots and grasses after divergence. MYB42/85 and MYB20/43 clades show expansion in maize and switchgrass. Two grass-expanded clades are indicated. MYB proteins shown with bold and colored text are characterized SCW regulators. Support values are from 1000 bootstrap analyses. Each logo is the C-terminal conserved motifs with the lowest E-value identified for the subgroup.

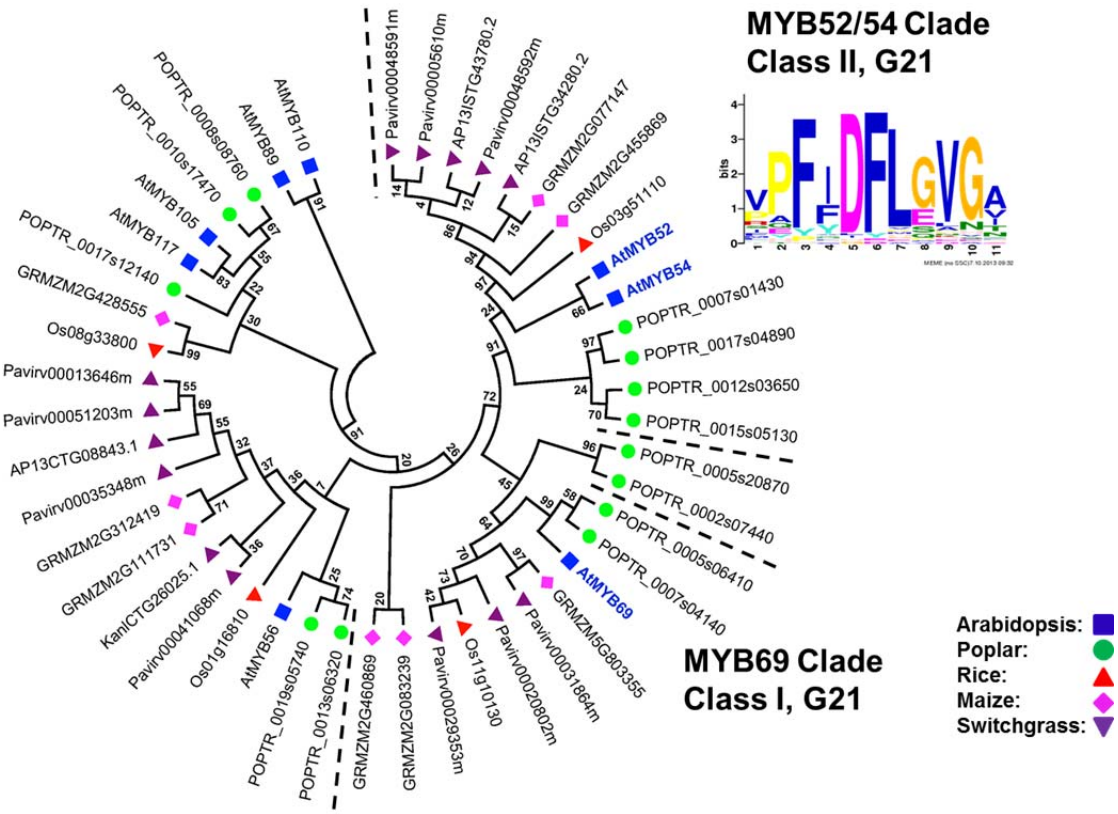


Figure 2-5 Maximum likelihood phylogenetic analysis of subgroup G21 suggests orthologous and paralogous relationships in dicots and grasses. The MYB69 clade is conserved across evolution with putative orthologs in the five species. The MYB52/54 clade shows expansion in poplar and switchgrass. MYB proteins shown with bold and colored text are characterized SCW regulators. Support values are from 1000 bootstrap analyses. Each logo is the C-terminal conserved motifs with the lowest E-value identified for the subgroup.

Class III: Non-SCW related paralogs in Arabidopsis

In Class III, the known Arabidopsis SCW regulators are closely related with other Arabidopsis R2R3 MYB proteins functioning in different biological processes. Thus, from phylogenetic analysis alone, it is difficult to hypothesize about the likely function of orthologs from other species. In this case, the amino acid identity within each clade and relationships identified by OrthoMCL aid in identification of likely functional orthologs (Östlund et al., 2010). Class III consists of AtMYB58 and AtMYB63, AtMYB61, and AtMYB4 and AtMYB32 (Figure 2-4, 2-6, and 2-7).

Functioning as lignin specific activators, *AtMYB58* and *AtMYB63* are regulated by *AtSND1* and its homologs, *AtNST1*, *AtNST2*, *AtVND6*, and *AtVND7*, and their target, *AtMYB46* (Zhou et al., 2009). As shown in Figure 2-6, AtMYB58 and AtMYB63 are in subgroup G3 and are paralogous with AtMYB10 and AtMYB72, which are involved in cesium toxicity tolerance and beneficial bacteria responses, respectively (Hampton et al., 2004; Segarra et al., 2009). This appears to be a case of neofunctionalization after gene duplication in the dicot lineage. Based on sequence similarity (Table 2-4), among the Arabidopsis proteins, AtMYB58 shares the highest similarity with those from other species; consistent with it being closest to the ancestral sequence and at least one homolog in other species having retained its function. *AtMYB58* and *AtMYB63* are predominantly expressed in vessels and fibers in Arabidopsis (Zhou et al., 2009). In contrast, their paralogs, *AtMYB10* and *AtMYB72* are mainly expressed in the

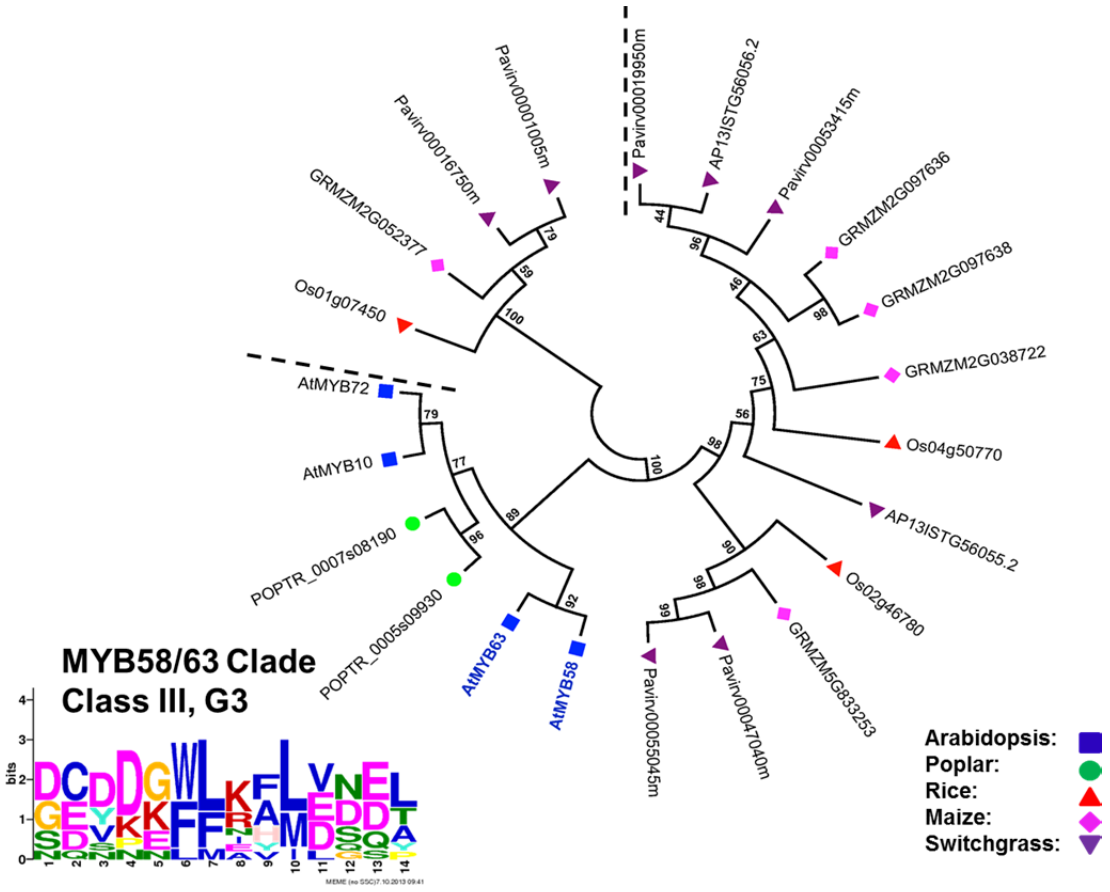


Figure 2-6 Maximum likelihood phylogenetic analysis of subgroup G3.a and G3.b suggests that MYB58/63 clade underwent expansion after the divergence of dicots and grasses. AtMYB10 and AtMYB72 are involved in cesium toxicity and pathogen resistance, which indicates neofunctionalization after duplication. MYB proteins shown with bold and colored text are characterized SCW regulators. Support values are from 1000 bootstrap analyses. Each logo is the C-terminal conserved motifs with the lowest E-value identified for the subgroup.

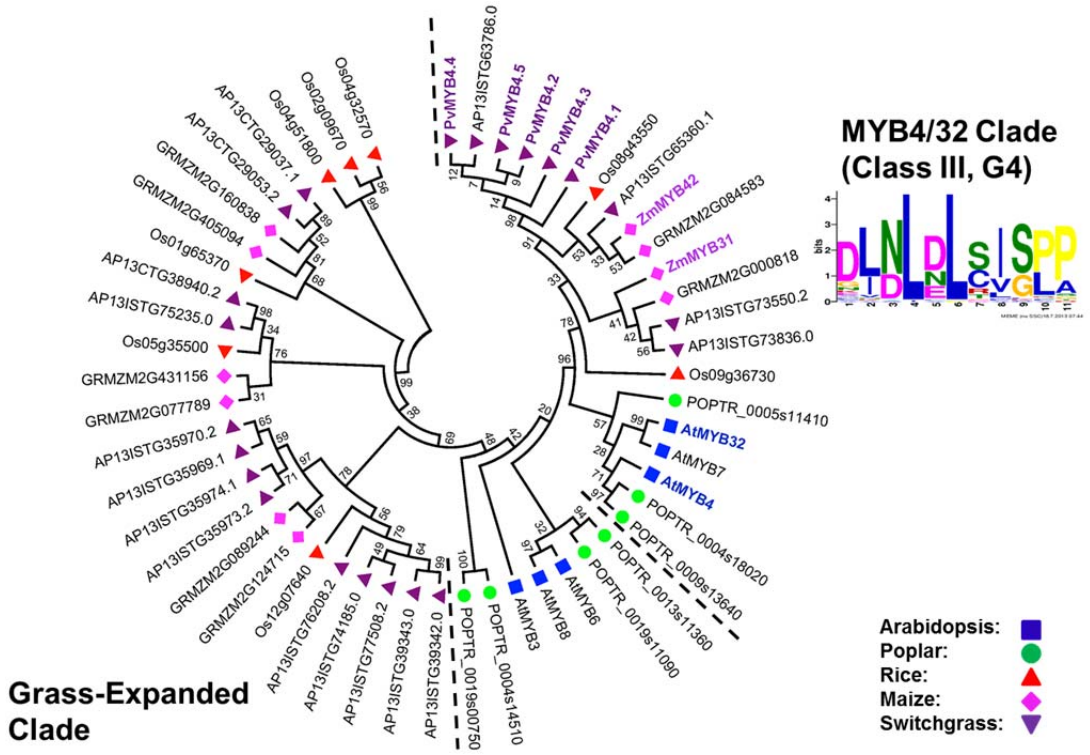


Figure 2-7 Maximum likelihood phylogenetic analysis of subgroup G4 suggests expansion of this group in both grasses and dicots since the last common ancestor. The MYB4/32 clade has many paralogs in dicots; however ZmMYB32, ZmMYB41 and PvMYB4.a have been shown to have function similarly to AtMYB4 and AtMYB32. PvMYB4.a to e are likely alleles among each other based on protein sequence similarity. MYB proteins shown with bold and colored text are characterized SCW regulators. Support values are from 1000 bootstrap analyses. Each logo is the C-terminal conserved motifs with the lowest E-value identified for the subgroup.

inflorescence (Yanhui et al., 2006). The switchgrass ortholog in this clade with gene expression data available, *AP13ISTG56055*, shows high expression in E4 vascular bundles and internodes, consistent with the possibility that they regulate SCW biosynthesis (Figure 2-9). Overexpression of the two *OsMYB58/63* was recently found to promote lignin deposition in rice stems, supporting their orthologous relationship with the AtMYB58 and 63 (Hirano et al., 2013). In the OrthoMCL analysis, AtMYB58 and AtMYB63 are paralogs and putative co-orthologs are found in the grasses. However, many related grass and poplar sequences are excluded from the orthologous relationship by OrthoMCL, possibly due to the somewhat low sequence identity (38% to 51%).

AtMYB61 is a SCW biosynthesis activator in subgroup G13.b that also belongs to Class III. AtMYB61 regulates water and sugar allocation and is mainly expressed in sink tissues. Loss-of-function mutants reduce xylem vessel formation and lignification (Romano et al., 2012). AtMYB61 is closely related to AtMYB50 and AtMYB55 (Figure 2-4). The function of AtMYB50, with 66% identity to AtMYB61, has not been studied in detail to our knowledge. Its transcript is upregulated during geminivirus infection (Ascencio-Ibáñez et al., 2008). Another paralog, *AtMYB55*, is involved in leaf development (Schliep et al., 2010). We found that this clade is expanded in poplar and switchgrass; whereas, rice and maize possess two paralogs (Figure 2-4). RNAi of the two *OsMYB61*s downregulates the expression of *OsCAD2*, which encodes a lignin biosynthesis enzyme (Hirano et al., 2013). *AtMYB61* is expressed in xylem, leaf and root. In contrast, *AtMYB50* and *AtMYB55* are broadly expressed in Arabidopsis (Riechmann et al., 2000; Romano et al., 2012). The ortholog in switchgrass for which

expression data are available, *AP13CTG04029*, also shows high expression in the stem (Figure 2-9). Based on this expression pattern, we conclude that AP13CTG04029 may regulate SCW formation. Despite these functional and expression results, from sequence identity analysis alone, AtMYB50 appears to be most similar to the ancestral sequence, with the co-orthologs from Arabidopsis and the other species ranging in identity with it from 53% to 58%. On the other hand, OthoMCL analysis groups all of the grass co-orthologs and two from poplar with AtMYB61 (Table 2-4).

The last pair of proteins in class III is AtMYB4 and AtMYB32, which negatively regulate SCW biosynthesis (Figure 2-1 and Figure 2-7). AtMYB4 is a repressor of lignin biosynthesis and ultraviolet B light responses (Jin et al., 2000). AtMYB4 has two paralogs, AtMYB32 and AtMYB7, which repress Arabidopsis pollen cell wall development and are downregulated under drought stress, respectively (Jin et al., 2000; Preston et al., 2004; Ma and Bohnert, 2007). In grasses, ZmMYB31, ZmMYB42 and PvMYB4a are all characterized orthologs of AtMYB4, that function as SCW biosynthesis repressors with somewhat paradoxically high expression in vascular tissues (Sonbol et al., 2009; Fornalé et al., 2010; Shen et al., 2012). The characterized PvMYB4a is closely related to four other predicted proteins with amino acid identity >99%, which are putative alleles or splice variants of each other (Shen et al., 2012). Among switchgrass ESTs, we found two additional orthologs of AtMYB4 that show high expression in vascular bundles, nodes, and internodes; whereas, the previously identified *PvMYB4d* is relatively lowly expressed (Figure 2-9). This difference in expression is consistent with subfunctionalization of PvMYB4d after gene duplication in switchgrass. Data for the other PvMYB4 alleles are lacking. Consistent with the gene

expression data, AtMYB4 is the most similar to the ancestral sequence, with orthologs from other species ranging in identity from 64% to 70% (Table 2-4). The MYB4/32 clade is disjointed in the OrthoMCL analysis. Most grass orthologs group with AtMYB4; however, ZmMYB42 and PvMYB4 cluster into two independent groups (Table 2-4).

Class IV: No clear homologs in grasses

AtMYB75 is the only SCW R2R3 MYB protein in Class IV, for which we found no evidence of orthologs in grasses. AtMYB75 functions as a repressor of SCW biosynthesis and is also known as *PRODUCTION OF ANTHOCYANIN PIGMENT*1 (*PAP*1), with a role in positively regulating anthocyanin metabolism (Bhargava et al., 2010; Zhao and Dixon, 2011; Shin et al., 2013). AtMYB75 belongs to the dicot-specific subgroup, G6, which includes AtMYB90, AtMYB113 and AtMYB114 (Table 2-2, Figure 2-8). Even when the relatively closely related G47 clade is included, our analysis separates AtMYB75 and the other members of G6 from all grass sequences. Among the G6 members, AtMYB114, which functions in nitrogen response, appears to be the most similar to the ancestral sequence, with the co-orthologs from Arabidopsis and poplar with identity ranging from 67% to 72% (Table 2-4) (Downie et al., 2004; Scheible et al., 2004). Thus, AtMYB75 may have resulted from gene duplication in the Arabidopsis lineage and is likely a dicot-specific SCW repressor. OrthoMCL analysis supports the phylogenetic topology and only identifies putative AtMYB75 co-orthologs from poplar (Table 2-4).

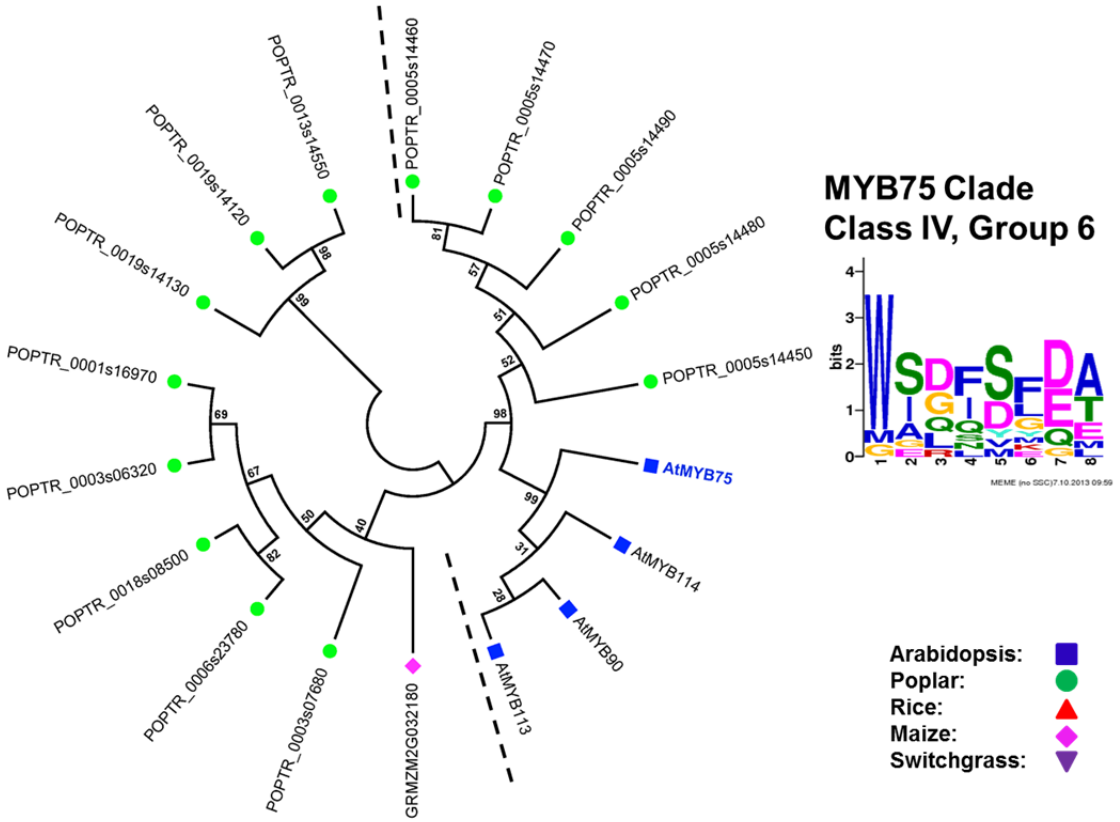


Figure 2-8 Maximum likelihood phylogenetic analysis of subgroups G6 and G47 suggests that AtMYB75 is a dicot-specific SCW repressor without homologs in grasses. MYB proteins shown with bold and colored text are characterized SCW regulators. Support values are from 1000 bootstrap analyses. Each logo is the C-terminal conserved motifs with the lowest E-value identified for the subgroup.

Expression of grass-expanded clades

In addition to putative (co-)orthologs of known SCW R2R3 MYBs, we noted the presence of grass-expanded clades in several of the subgroups that we examined in greater detail. As with the Class II proteins, these may have retained functions in SCW regulation or, as with Class III Arabidopsis proteins, developed new functions. Gene expression appears to be a useful indicator of their likely roles in secondary growth in vegetative tissues. Hence, we searched the database for expression of the switchgrass representatives of the grass-expanded clades. Figure 2-9 shows that three out of the nine genes for which data were available show strong expression in stems in general and vascular bundles in particular. Thus, these genes represent potential novel contributors to grass vegetative SCW regulation now under investigation.

Conclusions

A key element of translating basic research on model (or reference) species, such as Arabidopsis, to crops for food and fuel, is understanding the relative gene complement of the species in question, many of which, like switchgrass, possess a complex genome (Hirsch and Robin Buell, 2013). We have sought to address this need for the R2R3 MYB proteins. The three tools, phylogenetic, sequence identity and OrthoMCL analysis, for indicating orthologous relationships that we employed have various requirements for time and expertise. Multi-species phylogenetic analysis appears to be relatively inclusive in its groupings and is informative regarding the rough evolutionary history, such as the occurrence of gene or genome duplication and speciation. However,

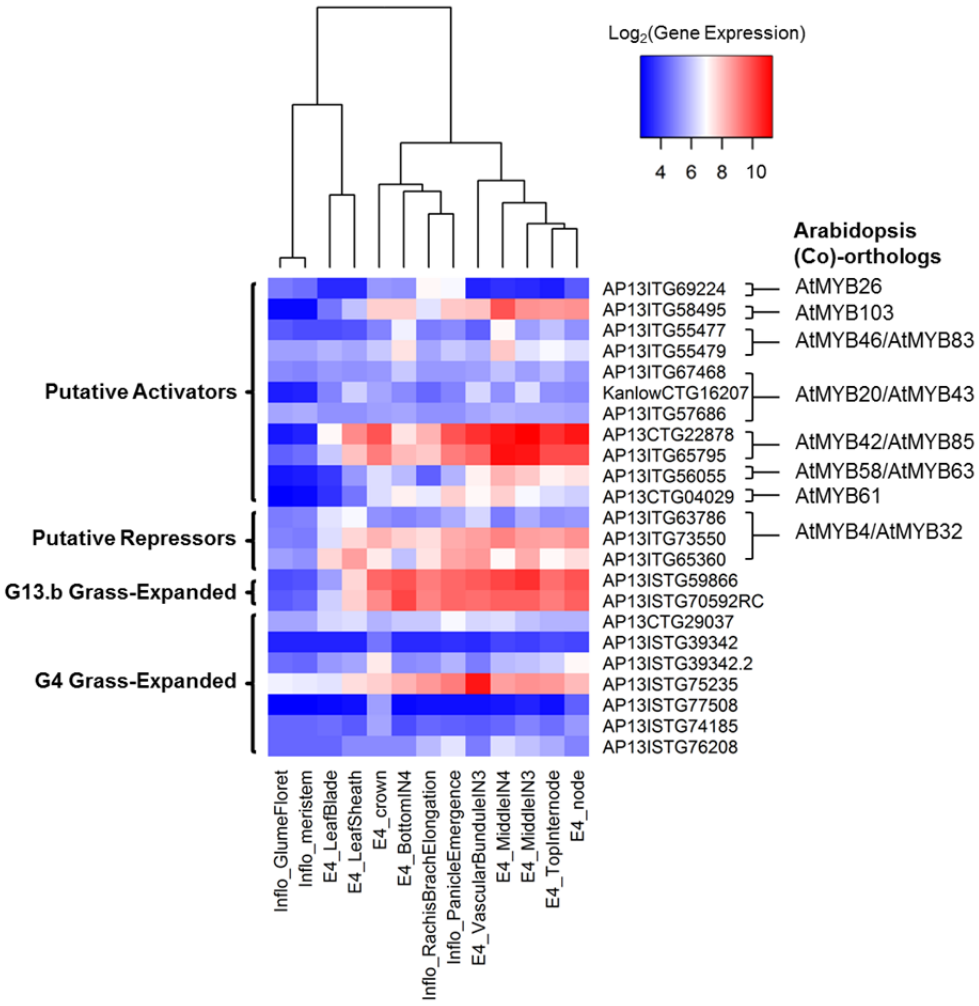


Figure 2-9 Gene expression analysis of switchgrass MYBs that are putative SCW-related activators or repressors and members of grass-expanded clades. The heatmap represents the log_2 of the expression data, which are normalized mean values of three biological replicates in the same experiments from the Switchgrass Functional Genomics Server (http://switchgrassgenomics.noble.org/). The blue indicates lower expression and red, higher expression. The relationships among columns are based on hierarchical clustering. The orthologs/co-orthologs from Arabidopsis are listed. Among the repressors with gene expression available, PvMYB4.d_AP13ITG63786 is one of the published homologs of AtMYB4/32 in switchgrass and it has 100% sequence similarity with PbMYB4.d with low expression in most of the tissues. The labels of tissues and developmental stages are abbreviated using the following scheme: from the inflorescence (Inflo) the meristem, glume floret, rachis branch during elongation, and panicle during emergence; from the tiller at elongation stage 4 (E4) the crown, leaf blade, leaf sheath, and stem the stem segments as follows: nodes, top internode, middle of internode (IN) 3, vascular bundle of IN 3, middle of IN 4, and the bottom of IN 4 (Zhang et al., 2013).

the topology of a phylogentic tree (1) can be model-dependent, especially for divergent sequences and (2) does not indicate which members of expanded groups are the most similar to those in other species, for example proteins in Class III that have expanded and functionally diverged in Arabidopsis. In addition, phylogenetic analysis is time consuming and, thus, infrequently used for genome-scale analysis.

In contrast to phylogenetic analysis, OrthoMCL, once implemented, can rapidly analyze multiple genomes. A previous comparative analysis of OrthoMCL and other similar large-scale ortholog identification methods found that OrthoMCL and the similar algorithm, InParanoid, have relatively high specificity and sensitivity on a "gold standard" data set (Chen et al., 2007). However, in the analysis presented here, OrthoMCL fails to identity known orthologs across dicots and grasses, as for the MYB46/83 and the MYB4/32 clades, though simple sequence identity supports the evidence of functional conservation across dicots and monocots in those clades. This indicates a problem with false negatives, if we select orthologs only based on OrthoMCL. Conversely, sequence similarity groups the grass co-orthologs in the MYB61/50 clade with the Cd^{2+}-tolerance regulator, AtMYB50, despite recent evidence consistent with their role in SCW regulation (Ascencio-Ibáñez et al., 2008). In that case, the OrthoMCL cluster is more consistent with the functional data. For both tools, the quantitation of similarity may not be generally applicable across the genome and lead to false grouping or grouping failure. Ideally, a genome-scale syntenic analysis across species could be an additional piece of information to assist in identifying orthologs when a more accurate switchgrass chromosomal assembly becomes available.

The switchgrass gene expression dataset, when available, appears to provide a much more nuanced guide of function among putative orthologs. For example, expression data suggest that among the switchgrass co-orthologs from the MYB46/83 and MYB42/85 clades, *AP13ISTG55479* and *AP13CTG22878*, are predominantly expressed and potentially better targets for reverse genetics compared with their paralogs. The gaps in the expression dataset provide support for applying and consolidating other transcriptomics approaches, such as RNA Seq (Li et al., 2013).

Comparative analysis of the R2R3 MYB family reinforces the assertion that though largely conserved, grass and dicot MYB families have undergone expansions and contractions (Table 2-3). With respect to SCW regulation, our analysis and emerging functional data (Shen et al., 2012; Hirano et al., 2013) are largely consistent with general but not complete, conservation of the Arabidopsis regulatory network (Figure 2-1). Phylogenetic and in some cases, gene expression data, for almost all of the AtMYBs grouped in classes I, II, and III, support conservation. This is despite the ambiguity of the class III proteins, which appear to have undergone expansion and neofunctionalization in the Arabidopsis lineage. This result is consistent with other global analyses of SCW regulation, such as based on maize gene expression data (Lai et al., 2010). Among established MYB SCW regulators, the repressor AtMYB75 is clearly not conserved and hence falls in class IV in our analysis. In addition, the MYB20/43 clade gene expression data in switchgrass and the reverse genetic data in Arabidopsis question the inclusion of these proteins among SCW regulators.

Differences between dicot and grass SCW regulation are likely to exist. In support of this, the gene expression data from switchgrass suggest that the expansion of

SCW R2R3 MYB proteins, either through whole genome duplication or more specific processes, has led to subfunctionalization in that species. For example, co-orthologs of *AtMYB4* and *AtMYB32*, namely, *AP13ISTG73550*, *AP13ISTG65360*, and *PvMYB4.d*, exhibit not just different expression amounts, but different expression patterns relative to each other (Figure 2-9). In addition, we identified several grass-expanded R2R3 MYB subgroups and clades (Table 2-3, Figure 2-4, and Figure 2-7) that may possess novel roles in grass-specific biology, including cell wall development. Some of these proteins are expressed highly in stems (Figure 2-9). Hence, this comparative analysis of the R2R3 MYB family will support the analysis of grass genomic data, providing particular insight into the emerging switchgrass genome. This information can be used to promote biofuel production from switchgrass and other grasses.

Methods

Identification of R2R3 MYB proteins

We used HMMER 3.0 (Finn et al., 2011) to identify the putative R2R3 MYB sequences in different species with an in-house Hidden Markov Model profile based on the 126 R2R3 MYB proteins in Arabidopsis (Dubos et al., 2010.) We mined the following genome annotation versions, which were current at the time of the analysis: *Oryza sativa*, MSU v7; *Populus trichocarpa*, Phytozome v3.0; *Zea mays*, Phytozome v2.0; *Arabidopsis thaliana*, TAIR v10; *Panicum virgatum*, Phytozome v0.0 DOE-JGI, (http:://www.phytozome.net/panicumvirgatum), and the unitranscripts dataset from the Switchgrass Functional Genomics Server (http://switchgrassgenomics.noble.org/) (Zhang et al., 2013). The switchgrass gene identifiers from Phytozome are “Pavirv” and

those from the Switchgrass Functional Genomics Server are “AP13” and “Kanl”. Only a few genes in the dataset have multiple known gene models, we used only gene model one (.1) for all analyses.

In our initial analysis of the switchgrass R2R3 MYBs in the v0.0 annotation, we noticed that expected sequences, namely, the recently characterized PvMYB4 proteins (Shen et al., 2012), were missing. A transcript with high homology was present in the v0.0 set of annotated coding sequences, suggesting that the omission was likely during the quality control of the protein annotation. To help to address this, we incorporated the proteins encoded by the unitranscripts in the Switchgrass Functional Genomics Server, which includes Sanger and 454 transcripts from Alamo (AP13) and Kanlow (Kanl) cultivars (Zhang et al., 2013). To identify switchgrass MYB proteins, we translated the transcripts, which are all the forward strands, using Bioperl, and screened them with the Arabidopsis R2R3 MYB Hidden Markov Model profile. The resulting putative MYB proteins were trimmed to remove the amino acids encoded by the RNA untranslated regions. The numeral (0, 1, 2) appended to the unitranscript sequence identifiers indicates the translation frames of the putative MYB, with “.0” indicating the +1 frame, etc. We compared the unitranscript-derived MYBs and the Phytozome switchgrass v0.0 protein datasets, and deleted the 100% redundant sequences from the Phytozome protein sequences for subsequent analysis. We also included the five sequences of the recently characterized PvMYB4 (Shen et al., 2012). Of those, PvMYB4.d is the only sequence that we found in the unitranscript dataset with the sequence identifier *AP13ISTG63786.0*.

We did an initial alignment of the R2R3 MYBs of each species using ClustalW2.0 and then removed sequences that lacked the R2R3 repeats. We also removed sequences that lacked two PROSITE (http://prosite.expasy.org/scanprosite/, PS50090) R repeats (Yanhui et al., 2006; Zhang et al., 2013). The final set of protein sequences and corresponding locus IDs or transcript identifiers used in this analysis is available in Supporting Table 2-1.

Phylogenetic and OrthoMCL analyses

We used CLUSTALW2.0 for all alignments, which we examined for quality, but did not need to edit. We randomly selected AmMYB6 from *Apis mellifera* as an outgroup. We used MEGA5.0 to infer phylogenetic relationships among the putative R2R3 MYB proteins. For the five species tree, we used the Neighbor-Joining algorithm with the default settings, except that gaps were treated by pair-wise deletion (Tamura et al., 2011). For the R2R3 MYB multispecies tree we used 500 bootstraps. For each of the SCW regulators, we inferred the relationships with the Maximum-Likelihood algorithm using 1000 bootstraps. The tree topologies were the same between Neighbor-Joining and Maximum-Likelihood algorithms. Within the SCW-related phylogenetic trees, we have identified SCW-protein containing and grass-expanded clades based on bootstrap scores of ≥50 and delimit these with dashed lines. In these trees, we define grass-expanded clades as having more members in rice than in either of the dicots. Most of these clades do not appear to be represented in Arabidopsis or poplar. To further examine homologous relationships among the R2R3 MYB proteins from the five species, we applied OrthoMCL analysis with the default settings (Li et al., 2003).

By convention, "homolog" is a general term for proteins that share a common origin and includes both "orthologs" and "paralogs." Orthologs derive from a single protein in the last common ancestor and tend to maintain similar function. Paralogs, on the other hand, are distinguished by being more similar to other proteins within the same genome and hence generated from expansion subsequent to the last common ancestor. Thus, it is harder to predict the function of paralogs across species, since expansion of the clade may have provided the opportunity for neo- or sub-functionalization (Koonin, 2005).

Sequence identity calculation and allelic diversity

Sequence similarity scores were calculated based on Multiple Sequence Alignment (MUSCLE) with the full-length protein sequences using DNA Subway (http://www.iplantcollaborative.org/discover/dna-subway). Through this analysis, some proteins appeared to have very high protein sequence similarity, consistent with being alleles or splice-variants of the same gene. There is no consensus on the criteria to identify alleles based on nucleotide or protein sequences similarity. Here, we highlight proteins with ≥99% similarity of amino acid sequences as possible alleles or splice-variants.

Conserved motifs

We analyzed the presence of conserved motifs in the full-length R2R3 MYB proteins from the 48 subgroups (and 4 sub-subgroups) separately with MEME (http://meme.nbcr.net/meme/intro.html) using the following parameters: distribution of

motif occurrences: one per sequence and present in all; number of different motifs: 10; minimum motif width: 6; maximum motif width: 15. Identified motifs C-terminal to the MYB domain with E-values lower than 1E-03 are listed in Supporting Table 2-2. To put our results in the context of the literature, the regular expression of each motif was compared to those previously identified for the Arabidopsis R2R3 MYB family (Stracke et al., 2001).

Gene expression

We used the gene expression data available from the Switchgrass Functional Genomics Server: http://switchgrassgenomics.noble.org/index.php (Zhang et al., 2013). The Gene Expression Atlas available through that server was assembled from Affymetrix microarray technology with 122,868 probe sets corresponding to 110,208 *Panicum virgatum* unitranscript sequences to measure gene expression in all major organs at one or more stages of development from germination to flowering (Zhang et al., 2013). Using heatmap.2 in R, we plotted the $\log_2$ of the Affymetrix hybridization signals, which represents the normalized mean values of three independent biological replicates for a given organ/stage/tissue. Data are available for only a subset of switchgrass gene models, presumably due to not being represented, at all or uniquely, on the Affymetrix array.

Abbreviations

SCW: Secondary Cell Wall; R: Repeat; G: Subgroup; At: *Arabidopsis thaliana*; Os: *Oryza sativa*; Pv: *Panicum virgatum*; Ptr: *Poplulus trichocarpa*; Zm: *Zea mays*; SND,

secondary wall-associated NAC domain protein; NST, NAC secondary wall thickening factor; E4: Elongation 4 stage; RNAi, RNA interference.

References

Ascencio-Ibáñez, J.T., Sozzani, R., Lee, T.-J., Chu, T.-M., Wolfinger, R.D., Cella, R., and Hanley-Bowdoin, L. (2008). Global Analysis of Arabidopsis Gene Expression Uncovers a Complex Array of Changes Impacting Pathogen Response and Cell Cycle during Geminivirus Infection. Plant Physiol. 148, 436-454.

Baranowskij, N., Frohberg, C., Prat, S., and Willmitzer, L. (1994). A novel DNA binding protein with homology to Myb oncoproteins containing only one repeat can function as a transcriptional activator. The EMBO journal 13, 5383.

Bartley, L., and Ronald, P.C. (2009). Plant and microbial research seeks biofuel production from lignocellulose. California Agriculture 63, 178-184.

Bartley, L.E., Tao, X., Zhang, C., Nguyen, H., and Zhou, J. (2014). Switchgrass Biomass Content, Synthesis, and Biochemical Conversion to Biofuels. In Switchgrass, H. Luo and Y. Wu, eds (Boca Raton, FL: Scioence Publishers), pp. 109-169.

Bhargava, A., Mansfield, S.D., Hall, H.C., Douglas, C.J., and Ellis, B.E. (2010). MYB75 functions in regulation of secondary cell wall formation in the Arabidopsis inflorescence stem. Plant Physiol. 154, 1428-1438.

Bonawitz, N.D., and Chapple, C. (2010). The genetics of lignin biosynthesis: connecting genotype to phenotype. Annu Rev Genet 44, 337-363.

Bouton, J.H. (2007). Molecular breeding of switchgrass for use as a biofuel crop. Curr. Opin. Genet. Dev. 17, 553-558.

Burton, R.A., Wilson, S.M., Hrmova, M., Harvey, A.J., Shirley, N.J., Medhurst, A., Stone, B.A., Newbigin, E.J., Bacic, A., and Fincher, G.B. (2006). Cellulose synthase-like CslF genes mediate the synthesis of cell wall (1, 3; 1, 4)-ß-D-glucans. Science 311, 1940-1942.

Casler, M.D., Tobias, C.M., Kaeppler, S.M., Buell, C.R., Wang, Z.-Y., Cao, P., Schmutz, J., and Ronald, P. (2011). The Switchgrass Genome: Tools and Strategies. Plant Gen. 4, 273-282.

Chaw, S.-M., Chang, C.-C., Chen, H.-L., and Li, W.-H. (2004). Dating the Monocot–Dicot Divergence and the Origin of Core Eudicots Using Whole Chloroplast Genomes. J Mol Evol 58, 424-441.

Chen, F., Mackey, A.J., Vermunt, J.K., and Roos, D.S. (2007). Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS ONE 2, e383.

Chiniquy, D., Sharma, V., Schultink, A., Baidoo, E.E., Rautengarten, C., Cheng, K., Carroll, A., Ulvskov, P., Harholt, J., Keasling, J.D., Pauly, M., Scheller, H.V., and Ronald, P.C. (2012). XAX1 from glycosyltransferase family 61 mediates xylosyltransfer to rice xylan. Proceedings of the National Academy of Sciences of the United States of America 109, 17117-17122.

Davidson, R.M., Gowda, M., Moghe, G., Lin, H., Vaillancourt, B., Shiu, S.-H., Jiang, N., and Robin Buell, C. (2012). Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. Plant J 71, 492-502.

De Smet, R., Adams, K.L., Vandepoele, K., Van Montagu, M.C.E., Maere, S., and Van de Peer, Y. (2013). Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. Proc. Natl. Acad. Sci. 110, 2898-2903.

Dias, A.P., Braun, E.L., McMullen, M.D., and Grotewold, E. (2003). Recently duplicated maize R2R3 Myb genes provide evidence for distinct mechanisms of evolutionary divergence after duplication. Plant Physiol. 131, 610-620.

Downie, A., Miyazaki, S., Bohnert, H., John, P., Coleman, J., Parry, M., and Haslam, R. (2004). Expression profiling of the response of Arabidopsis thaliana to methanol stimulation. Phytochemistry 65, 2305-2316.

Du, H., Feng, B.-R., Yang, S.-S., Huang, Y.-B., and Tang, Y.-X. (2012). The R2R3-MYB Transcription Factor Gene Family in Maize. PloS one 7, e37463.

Du, H., Zhang, L., Liu, L., Tang, X.-F., Yang, W.-J., Wu, Y.-M., Huang, Y.-B., and Tang, Y.-X. (2009). Biochemical and molecular characterization of plant MYB transcription factor family. Biochemistry (Mosc) 74, 1-11.

Dubos, C., Stracke, R., Grotewold, E., Weisshaar, B., Martin, C., and Lepiniec, L. (2010). MYB transcription factors in Arabidopsis. Trends Plant Sci 15, 573-581.

Edgar, R. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5, 113.

Esau, K. (1977). Anantomy of Seed Plants. (New York: John Wiley and Sons).

Feller, A., Machemer, K., Braun, E.L., and Grotewold, E. (2011). Evolutionary and comparative analysis of MYB and bHLH plant transcription factors. Plant J 66, 94-116.

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

Fornalé, S., Shi, X., Chai, C., Encina, A., Irar, S., Capellades, M., Fuguet, E., Torres, J.L., Rovira, P., and Puigdomènech, P. (2010). ZmMYB31 directly represses maize lignin genes and redirects the phenylpropanoid metabolic flux. Plant J 64, 633-644.

Gigolashvili, T., Yatusevich, R., Berger, B., Müller, C., and Flügge, U.-I. (2007). The R2R3-MYB transcription factor HAG1/MYB28 is a regulator of methionine-derived glucosinolate biosynthesis in Arabidopsis thaliana. Plant J 51, 247-261.

Gigolashvili, T., Engqvist, M., Yatusevich, R., Müller, C., and Flügge, U.-I. (2008). HAG2/MYB76 and HAG3/MYB29 exert a specific and coordinated control on the regulation of aliphatic glucosinolate biosynthesis in Arabidopsis thaliana. New Phytologist 177, 627-642.

Gray, J., Caparrós-Ruiz, D., and Grotewold, E. (2012). Grass phenylpropanoids: Regulate before using! Plant Science 184, 112-120.

Hampton, C.R., Bowen, H.C., Broadley, M.R., Hammond, J.P., Mead, A., Payne, K.A., Pritchard, J., and White, P.J. (2004). Cesium toxicity in Arabidopsis. Plant Physiol. 136, 3824-3837.

Handakumbura, P.P., and Hazen, S.P. (2012). Transcriptional regulation of grass secondary cell wall biosynthesis: playing catch-up with Arabidopsis thaliana. Frontiers in Plant Science 3.

Higginson, T., Li, S.F., and Parish, R.W. (2003). AtMYB103 regulates tapetum and trichome development in Arabidopsis thaliana. Plant J 35, 177-192.

Hirano, K., Kondo, M., Aya, K., Miyao, A., Sato, Y., Antonio, B.A., Namiki, N., Nagamura, Y., and Matsuoka, M. (2013). Identification of Transcription Factors Involved in Rice Secondary Cell Wall Formation. Plant Cell Physiol.

Hirsch, C.N., and Robin Buell, C. (2013). Tapping the Promise of Genomics in Species with Complex, Nonmodel Genomes. Annu. Rev. Plant Biol. 64, 89-110.

Jin, H., and Martin, C. (1999). Multifunctionality and diversity within the plant MYB-gene family. Plant molecular biology 41, 577-585.

Jin, H., Cominelli, E., Bailey, P., Parr, A., Mehrtens, F., Jones, J., Tonelli, C., Weisshaar, B., and Martin, C. (2000). Transcriptional repression by AtMYB4 controls production of UV-protecting sunscreens in Arabidopsis. The EMBO journal 19, 6150-6161.

Jung, K.H., An, G., and Ronald, P.C. (2008). Towards a better bowl of rice: assigning function to tens of thousands of rice genes. Nature reviews 9, 91-101.

Katiyar, A., Smita, S., Lenka, S.K., Rajwanshi, R., Chinnusamy, V., and Bansal, K.C. (2012). Genome-wide classification and expression analysis of MYB transcription factor families in rice and Arabidopsis. BMC genomics 13, 544.

Kellogg, E.A. (2001). Evolutionary History of the Grasses. Plant Physiol. 125, 1198-1205.

Ko, J.H., Kim, W.C., and Han, K.H. (2009). Ectopic expression of MYB46 identifies transcriptional regulatory genes involved in secondary wall biosynthesis in Arabidopsis. Plant J 60, 649-665.

Koonin, E.V. (2005). Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39, 309-338.

Kranz, H., Scholz, K., and Weisshaar, B. (2001). c-MYB oncogene-like genes encoding three MYB repeats occur in all major plant lineages. Plant J 21, 231-235.

Lai, J., Li, R., Xu, X., Jin, W., Xu, M., Zhao, H., Xiang, Z., Song, W., Ying, K., and Zhang, M. (2010). Genome-wide patterns of genetic variation among elite maize inbred lines. Nature genetics 42, 1027-1030.

Lal, R. (2005). World crop residues production and implications of its use as a biofuel. Environment International 31, 575-584.

Li, L., Stoeckert, C.J., and Roos, D.S. (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res. 13, 2178-2189.

Li, Y.-F., Wang, Y., Tang, Y., Kakani, V., and Mahalingam, R. (2013). Transcriptome analysis of heat stress response in switchgrass (Panicum virgatum L.). BMC Plant Biol. 13, 153.

Liang, Y.-K., Dubos, C., Dodd, I.C., Holroyd, G.H., Hetherington, A.M., and Campbell, M.M. (2005). AtMYB61, an R2R3-MYB Transcription Factor Controlling Stomatal Aperture in Arabidopsis thaliana. Curr. Biol. 15, 1201-1206.

Lu, F., Lipka, A.E., Glaubitz, J., Elshire, R., Cherney, J.H., Casler, M.D., Buckler, E.S., and Costich, D.E. (2013). Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. PLoS Genet 9, e1003215.

Ma, S., and Bohnert, H. (2007). Integration of Arabidopsis thaliana stress-related transcript profiles, promoter structures, and cell-specific expression. Genome Biol. 8, R49.

Matsumoto, T., Wu, J., Kanamori, H., Katayose, Y., Fujisawa, M., Namiki, N., Mizuno, H., Yamamoto, K., Antonio, B.A., and Baba, T. (2005). The map-based sequence of the rice genome. Nature 436, 793-800.

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

McCarthy, R.L., Zhong, R., Fowler, S., Lyskowski, D., Piyasena, H., Carleton, K., Spicer, C., and Ye, Z.-H. (2010). The poplar MYB transcription factors, PtrMYB3 and PtrMYB20, are involved in the regulation of secondary wall biosynthesis. Plant Cell Physiol. 51, 1084-1090.

McLaughlin, S.B., and Adams Kszos, L. (2005). Development of switchgrass (Panicum virgatum) as a bioenergy feedstock in the United States. Biomass and Bioenergy 28, 515-535.

Ogata, K., Morikawa, S., Nakamura, H., Hojo, H., Yoshimura, S., Zhang, R., Aimoto, S., Ametani, Y., Hirata, Z., and Sarai, A. (1995). Comparison of the free and DNA-complexed forms of the DMA-binding domain from c-Myb. Nat. Struct. Mol. Biol. 2, 309-320.

Ogata, K., Kanei-Ishii, C., Sasaki, M., Hatanaka, H., Nagadoi, A., Enari, M., Nakamura, H., Nishimura, Y., Ishii, S., and Sarai, A. (1996). The cavity in the hydrophobic core of Myb DNA-binding domain is reserved for DNA recognition and trans-activation. Nat. Struct. Mol. Biol. 3, 178-187.

Öhman, D., Demedts, B., Kumar, M., Gerber, L., Gorzsás, A., Goeminne, G., Hedenström, M., Ellis, B., Boerjan, W., and Sundberg, B. (2012). MYB103 is required for FERULATE-5-HYDROXYLASE expression and syringyl lignin biosynthesis in Arabidopsis stems. Plant J.

Östlund, G., Schmitt, T., Forslund, K., Köstler, T., Messina, D.N., Roopra, S., Frings, O., and Sonnhammer, E.L.L. (2010). InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. 38, D196-D203.

Peret, B., Larrieu, A., and Bennett, M.J. (2009). Lateral root emergence: a difficult birth. J Exp Bot 60, 3637-3643.

Perlack, R.D., Wright, L.L., Turhollow, A., Graham, R.L., Stokes, B.J., and Erbach, D.C. (2005). Biomass as Feedstock for a Bioenergy and Bioproducts Industry: The Technical Feasibility of a Billion-Ton Annual Supply (Oak Ridge, Tennessee: Oak Ridge National Laboratory).

Preston, J., Wheeler, J., Heazlewood, J., Li, S.F., and Parish, R.W. (2004). AtMYB32 is required for normal pollen development in Arabidopsis thaliana. Plant J 40, 979-995.

Rabinowicz, P.D., Braun, E.L., Wolfe, A.D., Bowen, B., and Grotewold, E. (1999). Maize R2R3 Myb genes: Sequence analysis reveals amplification in the higher plants. Genetics 153, 427-444.

Riechmann, J., Heard, J., Martin, G., Reuber, L., Keddie, J., Adam, L., Pineda, O., Ratcliffe, O., Samaha, R., and Creelman, R. (2000). Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. Science 290, 2105-2110.

Romano, J.M., Dubos, C., Prouse, M.B., Wilkins, O., Hong, H., Poole, M., Kang, K.-Y., Li, E., Douglas, C.J., Western, T.L., Mansfield, S.D., and Campbell, M.M. (2012). AtMYB61, an R2R3-MYB transcription factor, functions as a pleiotropic regulator via a small gene network. New Phytologist 195, 774-786.

Saha, P., Lin, F., Thibivilliers, S., Santoro, N., and Bartley, L.E. (Submitted). Correlations between cell wall properties and expression of lignin biosynthesis genes in differing genotypes of switchgrass. BioEnergy Res.

Scheible, W.-R., Morcuende, R., Czechowski, T., Fritz, C., Osuna, D., Palacios-Rojas, N., Schindelasch, D., Thimm, O., Udvardi, M.K., and Stitt, M. (2004). Genome-Wide Reprogramming of Primary and Secondary Metabolism, Protein Synthesis, Cellular Growth Processes, and the Regulatory Infrastructure of Arabidopsis in Response to Nitrogen. Plant Physiol. 136, 2483-2499.

Schliep, M., Ebert, B., Simon-Rosin, U., Zoeller, D., and Fisahn, J. (2010). Quantitative expression analysis of selected transcription factors in pavement, basal and trichome cells of mature leaves from Arabidopsis thaliana. Protoplasma 241, 29-36.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., Minx, P., Reily, A.D., Courtney, L., Kruchowski, S.S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S.M., Belter, E., Du, F., Kim, K., Abbott, R.M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S.M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M.J., McMahan, L., Van Buren, P., Vaughn, M.W., Ying, K., Yeh, C.T., Emrich, S.J., Jia, Y., Kalyanaraman, A., Hsia, A.P., Barbazuk, W.B., Baucom, R.S., Brutnell, T.P., Carpita, N.C., Chaparro, C., Chia, J.M., Deragon, J.M., Estill, J.C., Fu, Y., Jeddeloh, J.A., Han, Y., Lee, H., Li, P., Lisch, D.R., Liu, S., Liu, Z., Nagel, D.H., McCann, M.C., SanMiguel, P., Myers, A.M., Nettleton, D., Nguyen, J., Penning, B.W., Ponnala, L.,

Schneider, K.L., Schwartz, D.C., Sharma, A., Soderlund, C., Springer, N.M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T.K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J.L., Dawe, R.K., Jiang, J., Jiang, N., Presting, G.G., Wessler, S.R., Aluru, S., Martienssen, R.A., Clifton, S.W., McCombie, W.R., Wing, R.A., and Wilson, R.K. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science 326, 1112-1115.

Segarra, G., Van der Ent, S., Trillas, I., and Pieterse, C.M.J. (2009). MYB72, a node of convergence in induced systemic resistance triggered by a fungal and a bacterial beneficial microbe. Plant Biology 11, 90-96.

Shen, H., Fu, C.X., Xiao, X.R., Ray, T., Tang, Y.H., Wang, Z.Y., and Chen, F. (2009). Developmental Control of Lignification in Stems of Lowland Switchgrass Variety Alamo and the Effects on Saccharification Efficiency. BioEnergy Res 2, 233-245.

Shen, H., He, X., Poovaiah, C.R., Wuddineh, W.A., Ma, J., Mann, D.G., Wang, H., Jackson, L., Tang, Y., and Neal Stewart Jr, C. (2012). Functional characterization of the switchgrass (Panicum virgatum) R2R3-MYB transcription factor PvMYB4 for improvement of lignocellulosic feedstocks. New Phytologist.

Shin, D.H., Choi, M., Kim, K., Bang, G., Cho, M., Choi, S.-B., Choi, G., and Park, Y.-I. (2013). HY5 regulates anthocyanin biosynthesis by inducing the transcriptional activation of the MYB75/PAP1 transcription factor in Arabidopsis. FEBS Letters 587, 1543-1547.

Sonbol, F.-M., Fornalé, S., Capellades, M., Encina, A., Tourino, S., Torres, J.-L., Rovira, P., Ruel, K., Puigdomenech, P., and Rigau, J. (2009). The maize ZmMYB42 represses the phenylpropanoid pathway and affects the cell wall structure, composition and degradability in Arabidopsis thaliana. Plant Mol. Biol. 70, 283-296.

Stracke, R., Werber, M., and Weisshaar, B. (2001). The R2R3-MYB gene family in Arabidopsis thaliana. Curr Opin Plant Biol 4, 447-456.

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

Tuskan, G.A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R.R., Bhalerao, R.P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G.L., Cooper, D., Coutinho, P.M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Dejardin, A., dePamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y.,

Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjarvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leple, J.C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D.R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouze, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C.J., Uberbacher, E., Unneberg, P., Vahala, J., Wall, K., Wessler, S., Yang, G., Yin, T., Douglas, C., Marra, M., Sandberg, G., Van de Peer, Y., and Rokhsar, D. (2006). The Genome of Black Cottonwood, Populus trichocarpa (Torr. & Gray). Science 313, 1596-1604.

Vogel, J. (2008). Unique aspects of the grass cell wall. Curr. Opin. Plant Biol. 11, 301-307.

Wang, H., Moore, M.J., Soltis, P.S., Bell, C.D., Brockington, S.F., Alexandre, R., Davis, C.C., Latvis, M., Manchester, S.R., and Soltis, D.E. (2009). Rosid radiation and the rapid rise of angiosperm-dominated forests. Proc. Natl. Acad. Sci. 106, 3853-3858.

Wang, H.Z., and Dixon, R.A. (2012). On-off switches for secondary cell wall biosynthesis. Molecular plant 5, 297-303.

Wilkins, O., Nahal, H., Foong, J., Provart, N.J., and Campbell, M.M. (2009). Expansion and diversification of the Populus R2R3-MYB family of transcription factors. Plant Physiol. 149, 981-993.

Yang, C.Y., Xu, Z.Y., Song, J., Conner, K., Barrena, G.V., and Wilson, Z.A. (2007). Arabidopsis MYB26/MALE STERILE35 regulates secondary thickening in the endothecium and is essential for anther dehiscence. Plant Cell 19, 534-548.

Yanhui, C., Xiaoyuan, Y., Kun, H., Meihua, L., Jigang, L., Zhaofeng, G., Zhiqiang, L., Yunfei, Z., Xiaoxiao, W., and Xiaoming, Q. (2006). The MYB transcription factor superfamily of Arabidopsis: expression analysis and phylogenetic comparison with the rice MYB family. Plant Mol. Biol. 60, 107-124.

Youngs, H., and Somerville, C. (2012). Development of feedstocks for cellulosic biofuels. F1000 biology reports 4, 10.

Zhang, J.Y., Lee, Y.C., Torres-Jerez, I., Wang, M., Yin, Y., Chou, W.C., He, J., Shen, H., Srivastava, A.C., Pennacchio, C., Lindquist, E., Grimwood, J., Schmutz, J., Xu, Y., Sharma, M., Sharma, R., Bartley, L.E., Ronald, P.C., Saha, M.C., Dixon, R.A., Tang, Y., and Udvardi, M.K. (2013). Development of an integrated transcript sequence database and a gene expression atlas for gene discovery and analysis in switchgrass (Panicum virgatum L.). Plant J 74, 160-173.

Zhang, Z.B., Zhu, J., Gao, J.F., Wang, C., Li, H., Li, H., Zhang, H.Q., Zhang, S., Wang, D.M., and Wang, Q.X. (2007). Transcription factor AtMYB103 is required for anther development by regulating tapetum development, callose dissolution and exine formation in Arabidopsis. Plant J 52, 528-538.

Zhao, Q., and Dixon, R.A. (2011). Transcriptional networks for lignin biosynthesis: more complex than we thought? Trends Plant Sci 16, 227-233.

Zhong, R., and Ye, Z.H. (2007). Regulation of cell wall biosynthesis. Curr Opin Plant Biol 10, 564-572.

Zhong, R., and Ye, Z.-H. (2012). MYB46 and MYB83 bind to the SMRE sites and directly activate a suite of transcription factors and secondary wall biosynthetic genes. Plant Cell Physiol. 53, 368-380.

Zhong, R., Richardson, E.A., and Ye, Z.H. (2007). The MYB46 transcription factor is a direct target of SND1 and regulates secondary wall biosynthesis in Arabidopsis. Plant Cell 19, 2776-2792.

Zhong, R., Lee, C., and Ye, Z.-H. (2010). Evolutionary conservation of the transcriptional network regulating secondary cell wall biosynthesis. Trends in Plant Science 15, 625-632.

Zhong, R., Lee, C., Zhou, J., McCarthy, R.L., and Ye, Z.H. (2008). A battery of transcription factors involved in the regulation of secondary cell wall biosynthesis in Arabidopsis. Plant Cell 20, 2763-2782.

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

Zhou, J., Lee, C., Zhong, R., and Ye, Z.-H. (2009). MYB58 and MYB63 are transcriptional activators of the lignin biosynthetic pathway during secondary cell wall formation in Arabidopsis. The Plant Cell 21, 248-266.

Chapter 3 : A novel genome-scale network elucidates secondary cell wall regulators in rice

Authors: Kangmei Zhao, Fan Lin, Hyung-Jung Goh, Prasentjit Saha, Gynheung An, Ki-Hong Jung, Laura E. Bartley

Publication Status: This chapter is in preparation as a manuscript for publication.

Authors Contributions: K.Z., L.E.B., G.A, and K.-H.J. designed this study. K.Z., L.F., H.-J. G., and P.S. performed the experiments. K.Z analyzed the data. K.Z and L.E.B wrote the current draft of the manuscript.

Abstract

Grass secondary cell walls (SCW) compose the bulk of dry biomass and have evolved distinct composition and patterning compared to dicotyledonous plants. However, the regulation of SCW biosynthesis is relatively poorly understood in grasses. In this study, we use rice as a model to examine the conservation and divergence of cell wall transcription factors between dicots and grasses using network inference and comparative genomics. We developed a Rice Combined mutual Ranked (RCR) network that includes ~90% of the genome and shows high quality via GO-term-based evaluation. The RCR network includes a cell wall sub-network with 96 novel transcription factors, some connected only to secondary cell wall synthesis genes, and others only to primary wall synthesis genes. The edge ranking of transcription factors in the RCR and an Arabidopsis network are dramatically different for some TFs, suggesting that the structure of the cell wall-regulating network has diverged between these species. In the RCR network, OsMYB61a, a co-ortholog of a known Arabidopsis SCW activator, shares edges with several grass-specific cell wall synthesis enzymes. Reverse genetics, yeast one-hybrid, and rice protoplast-based assays confirm that OsMYB61a directly regulates grass-specific acyltransferase and cellulose synthenthase-like genes, as well as secondary cell wall synthesis enzymes. Transient gene expression analysis of other network transcription factors demonstrated positive or negative roles in cell wall regulation for ten out of 15 transcription factors, eight of which had not previously been examined for cell wall function. The RCR network and this study facilitate understanding of regulatory mechanisms in rice to enable control of grass SCW biosynthesis for improved fuel, feed, and fiber production.

Introduction

Grasses cover 20% of terrestrial land and include more than 1000 species including the major agriculturally important cereal crops, rice, maize, wheat, and sorghum (Kellogg, 2001). Residues of cultivated grasses are the most abundant sustainable biomass that can be produced in the U.S (Perlack et al., 2005). Among the cereals, rice (*Oryza sativa* L.) is the staple food for more than half of the world population and contributes 23% of agricultural residual (Lal, 2005). With a relatively small genome and abundant genetic and genomic resources, rice is a reference to study cereal genomes and gene function (Jung et al., 2008).

A great deal has been learned about the dicotyledenous reference plant, *Arabidopsis thaliana*; however, the challenge remains to translate that information to other species, like the economical important grasses. Phylogenetic and genomic synteny analysis provide initial support for translating information across species (Nishiyama et al., 2003; Davidson et al., 2012; Zhao and Bartley, 2014). Gene networks add an additional layer of functional data for comparing species, which have been widely used in different organisms to decipher regulatory pathways and complex traits (Mao et al., 2009; Mutwil et al., 2010; Movahedi et al., 2011; Yeung et al., 2011; Sarkar et al., 2014; Taylor-Teeples et al., 2014; Obertello et al., 2015). Comparative networks, such as, PlatNet, STARNET, ATTED II, CoP and PLANEX, incorporate transcriptomics datasets to facilitate analysis of coexpression associations across species (Ogata et al., 2010; Mutwil et al., 2011; Yim et al., 2013; Obayashi et al., 2014). Currently available genome-scale networks for rice built with microarray data include, Rice Oligonucleotide Array Database (ROAD), Oryza Express, RiceArrayNet, Rice GeneNet

Engine and RiceFREND (Lee et al., 2009; Hamada et al., 2011; Cao et al., 2012; Sato et al., 2012; Ficklin and Feltus, 2013). Among the expression-based rice networks, many have user-friendly web-based interfaces that facilitate single gene searches, but the whole networks are not available for download. The ROAD and the PlaNet networks are the exception to this, facilitating large-scale network comparison and cross-validation.

In addition to networks based on gene expression alone, Lee et al. (2011) developed Bayesian network RiceNet (version 1) using data from rice and other species to score evidence of a functional association. RiceNet incorporates 24 datasets including gene expression microarrays, protein-protein interactions, phylogenetic profile similarity, and diverse gene-gene associations from *yeast*, *Caenorhabditis elegans*, *humans*, *Arabidopsis thaliana*, and *Oryza sativa*. (Lee et al., 2004; Lee et al., 2010; Lee et al., 2011; Lee et al., 2015b). RiceNet v2 is a recent update and expansion of this network that incorporates additional rice microarray and RNA Seq transcriptomics data (Lee et al., 2015b). RiceNet v2 is also publicly available.

The publicly accessible rice genome-scale networks employ different scoring systems, show different genome coverage, and likely have different quality. ROAD incorporates gene expression across 1,867 publicly available rice microarray hybridizations based on Pearson Correlation Coefficient (PCC) scores and covers about 69% of the 41203 rice genes, but has not been experimentally validated (Cao et al., 2012). The rice network affiliated with PlaNet uses highest reciprocal rank (HRR) based on PCC calculated from 156 rice microarray experiments and covers 74% of rice genes (Mutwil et al., 2010; Mutwil et al., 2011). Rice PlaNet has not been

experimentally validated, though the preliminary validation rate for the Arabidopsis PlaNet was ~20% (i.e., 5 of 20 predicted essential genes have lethal phenotypes) (Mutwil et al., 2010; Mutwil et al., 2011). The functional network, RiceNet v1, only includes about 44% of the rice genome (Lee et al., 2011). However, RiceNet v1 has a high functional validation rate with thirteen out of fourteen high scoring nodes interacting with target genes in yeast two-hybrid assays and three out five of them functioning in pathogen response in RNAi and/or overexpression mutants (Lee et al., 2011). The recent update, RiceNet v2, expands rice genome coverage to 63% by incorporating additional rice datasets (Lee et al., 2015b). This suggests that RiceNet v2 is a high-quality functional network for inferring rice biological pathways.

The biological target for this work is the regulation and synthesis of grass cell walls. Influencing the shape and properties of each cell and organ, cell walls are integral to plant physiology and environmental adaptation, alter nutritional properties of cereal grains, and impact resistance of grass biomass to enzymatic breakdown (Niklas, 2004; Vogel, 2008a; Zhao and Dixon, 2014; Tenhaken, 2015; Kumar et al., 2016). Primary cell walls surround growing cells; whereas, secondary cell walls (SCWs) are laid down around many cells (e.g. tracheids, vessels and fibers) after cessation of growth. and constitute the bulk of plant biomass (Mauseth, 1988; Zhong and Ye, 2001; Niklas, 2004; Sørensen et al., 2010; Popper et al., 2011).

The major components of cell walls are cellulose; matrix polysaccharides, including hemicellulose and pectins; and lignin (Burton and Fincher, 2012; Carpita, 2012). Cellulose, the most abundant polysaccharide on the earth, is β-1,4-linked glucose (Mutwil et al., 2008; Handakumbura et al., 2013; Schwerdt et al., 2015). *Cellulose*

Synthase A (CESA) proteins are synthesize cellulose (Slabaugh et al., 2014). In rice, OsCESA4, OsCESA7, and OsCESA9 synthesize SCW cellulose, whereas OsCESA1, OsCESA2 and OsCESA8 are required for primary wall formation (Tanaka et al., 2003) (Handakumbura et al., 2013) (Mutwil et al., 2008). Hemicellulose encompasses a group of heterogeneous polysaccharide and contributes roughly one third of cell wall biomass (Scheller and Ulvskov, 2010; Pauly et al., 2013). The major hemicelluloses are xyloglucan, xylans, mixed linkage glucan, and mannans. Proteins from the glycosyltransferase (GT) GT43, GT47 and GT75 families form xylan in the Golgi body for subsequent vesicle-mediated release to the cell wall (Oikawa et al., 2010; Oikawa et al., 2013). Lignin is an aromatic polymer from the phenylpropanoid pathway,only present in SCWs (Vanholme et al., 2008; Bonawitz and Chapple, 2010; Voxeur et al., 2015). Covalently cross-linked lignin represents a major barrier to utilizing cell wall polysaccharides. The structure of lignin in terms of degree of branching depends on the relative proportion of guaiacyl (G), syringyl (S), and in grasses, hydroxyphenyl (H) subunits (Harrington et al., 2012). A series of enzymes synthesize the monolignol precursors from phenylalanine and tyrosine, reducing the 3-carbon tail (i.e., phenylalanine ammonium ligase, PAL; 4-coumaroyl ligase, 4CL, hydroxycinnamoyl-CoA:shikimate transferase, HCT, cinnamoyl-CoA reductase, CCR; and cinnamyl alcohol dehydrogenase, CAD) and modifying the phenyl ring (i.e., cinnamate 4-hydroxylase, C4H; p-coumaroyl shikimate 3´-hydroxylase, C3´H; caffeoyl-CoA methyltransferase, CCoAMT; ferulate 5-hydroxylase, F5H; caffeic acid methyltransferase, COMT) and caffeoyl shikimate esterase (CSE) (Bonawitz and

Chapple, 2010; Vanholme et al., 2013). Most of these enzymes are encoded by gene families in rice and other grasses.

Having diverged from dicotyledonous plants about 150 million years ago, grasses have evolved major differences in cell wall composition and vascular bundle patterning in leaves and stems (Chaw et al., 2004; Handakumbura and Hazen, 2012; Wang and Dixon, 2012; Zhong and Ye, 2015; Kumar et al., 2016). Grasses and other recently evolved, commelinid monocots use arabinoxylan as the most abundant hemicellulose in primary walls instead of xyloglucan. *OsXAX1* and *OsXAXl* are two known grass-specific arabinoxylan modifying enzymes from the GT61 family (Chiniquy 2012). In addition, commelinid monocots incorporate mixed-linkage glucan (MLG) into primary and secondary cell walls (Burton et al., 2006; Vogel, 2008b; Scheller and Ulvskov, 2010). Cellulose synthase-like (CSL) genes from grass-specific clades, lacking in dicots are responsible synthesis of MLG, including *OsCSLF6*, *OsCSLF8* and *OsCSLH1* (Burton et al., 2006; Vega-Sánchez et al., 2012; Kim et al., 2015). In addition, commelinid monocot lignin and arabinoxylan are esterified with hydroxycinnamic acids (HCAs). Several members of the so-called "Mitchell clade" of "BAHD" acyl-CoA acyltransferases (ATs) participate in cell wall modification in grasses (Withers et al., 2012; Bartley et al., 2013). For example, rice *p*-coumarate monolignol transferase, OsPMT1 (i.e., OsAT4) acylates monolignols with para-coumaric acid (*p*CA) (Withers et al. 2012). OsAT5 is a ferulate (FA) monolignol transferase that increases feruloylation of monolignols in overexpression mutants. OsAT10 appears to function in modifying glucuronoarabinoxylan with *p*-CA (Bartley et al., 2013). Though some phylogenetically unrelated taxa possess these cell wall

polymers/modifications, for the purposes of this work, we refer to the corresponding genes as “grass-specific” relative to Arabidopsis (Scheller and Ulvskov, 2010).

Scientists have made significant progress in understanding regulation of SCW synthesis in dicots, but only a few functional studies have been conducted in grasses (Zhong and Ye, 2007; Handakumbura and Hazen, 2012; Wang and Dixon, 2012; Taylor-Teeples et al., 2015; Zhong and Ye, 2015). In Arabidopsis, multiple transcription factor families regulate SCW biosynthesis and appear to form a series of hierarchical feed-forward loops (Taylor-Teeples et al., 2015). Approximately 33 dicot regulators of SCW development have been identified through forward and reverse genetic analysis (Zhong and Ye, 2007; Wang and Dixon, 2012; Zhong and Ye, 2015; Yang and Wang, 2016). A few NAC (named for NAM, ATAF1, 2 and CUC2) proteins are top-level activators in *Arabidopsis* and *Medicago* (Mitsuda et al., 2007; Zhong et al., 2010; Zhou et al., 2014). For example, Arabidopsis SECONDARY WALL ASSOCIATED NAC PROTEIN/ NAC SECONDARY WALL TRANSCRIPTION FACTOR (AtSND1/NST1) activates overall SCW biosynthesis, enhancing expression of downstream transcription factors and cell wall biosynthesis genes (Appenzeller et al., 2004; Mitsuda et al., 2007; Wang et al., 2011). Furthermore, the Arabidopsis VASCULAR-RELATED NAC-DOMAIN 6 and 7 (AtVND6 and 7) are key top-level activators of xylem vessel differentiation (Ohashi-Ito et al., 2010). Recently, Taylor-Teeples et al. (2015) reported yeast one-hybrid based screening for SCW transcription factors on regulatory and biosynthesis promoters expressed during Arabidopsis root xylem development. In addition to reporting a large number of binding interactions at SCW promoters, they identified AtE2Fc, a E2F DP family member, as another layer of

SCW regulation above AtVND6 and VND7 (Taylor-Teeples et al., 2015). On the other hand, known mid-level SCW regulators are mostly from the R2R3 MYB family (Zhao and Bartley, 2014). For example, AtMYB46 is a direct target of AtSND1 and can activate other downstream cell wall-associated transcription factors and biosynthesis genes (Zhong et al., 2007; Ko et al., 2009; McCarthy et al., 2009; Zhong and Ye, 2012; Kim et al., 2014). AtMYB61a can activate lignification and relocate water and resources by activating cell wall genes (Liang et al., 2005; Romano et al., 2012). However, we still lack the picture of cell wall regulators in grasses.

In this analysis, we developed a comprehensive rice genome-scale network, the Rice Combined mutual Ranked (RCR) network. We examined this high quality genome-scale network to understanding regulation of grass cell wall biosynthesis using rice as a model. Examination of rice homologs of known cell wall regulators suggests that grass cell wall-specific genes have been incorporated into regulatory pathways that include similar factors to those of Arabidopsis. In addition, we provide evidence that at least ten out of 15 predicted rice cell wall-related transcription factors regulate SCW biosynthesis.

Results

Development of a high coverage and quality rice gene network

Our goal is to understand cell wall biosynthesis and regulation through rice genome-scale networks, especially focusing on grass-specific cell wall genes and regulators. We first examined the grass-specific cell wall gene coverage in three fully publicly available rice networks, namely ROAD, PlaNet and RiceNet. We examined the

networks for inclusion of characterized and putative grass cell wall-specific genes, including 20 BAHD acyltransferase; three MLG cellulose synthase like genes (CSLs) *OsSLF6*, *OsCSLF8* and *OsCSLH1*; and two grass-specific arabinoxylan modifying genes, *OsXAX1* and *OsXAXl*, for a total of 25 genes (Supplementary Table 1). The two unvalidated co-expression networks, ROAD and PlaNet, are missing 16% and 4% of the grass cell wall gene list, respectively; however, the functional networks, RiceNet v1 and v2, lack 65% and 24% of these genes, respectively (Supplementary Figure 1). Thus, though improved over v1, the high-quality functional network, RiceNet v2 appears to be incomplete with respect to grass-diverged genes and pathways. On the other hand, the large coexpression networks may have low quality, i.e., low predictive power.

To overcome the depth and quality limitations of the existing networks, we sought to develop a high-quality network suitable for mining grass-diverged traits. Our strategy was to recalibrate the edge scores within ROAD and PlaNet to the scoring system of RiceNet to create a heuristic, but more complete, network. To scale the different scores to a similar range, we first calculated the inverse mutual rank for each network based on their original scores. Mutual rank improves overall performance of Pearson Correlation Coefficient-based co-expression networks (Obayashi and Kinoshita, 2009). Inverting the ranking makes greater scores reflect greater confidence. For ROAD, we used only the positive correlations; whereas, positive scores for RiceNet and PlaNet include both positive and negative gene expression correlations. We utilized a generalized linear model (GLM) to combine the datasets, thereby creating the Rice Combined mutual Ranked (RCR) network (see Equation I in the methods). We utilized a generalized linear model (GLM) to combine the datasets, creating a Rice Combined

mutual Ranked (RCR) network for each of the two RiceNet versions with Equation I and II, respectively.

To determine if we achieved our goal, we further assessed the size and features of the original and new networks. The RCR networks cover the highest number of rice genes (Figure 3-1, Table 3-1), including most of our list of functionally characterized and putative grass cell wall-specific genes (Figure 3-1C). This suggests that the RCR networks provide the opportunity to effectively study specialized genes or traits of rice. Moreover, we analyzed the topology of the networks by calculating fitness to the power law. Most biological networks have been found to be scale free and follow the power low, $P(k) \sim k^{-\gamma}$, in which a few nodes have a very large number of edges (Barabasi and Oltvai, 2004). All the networks fit the power law, though PlaNet displays relatively low fitness (Table 3-1).

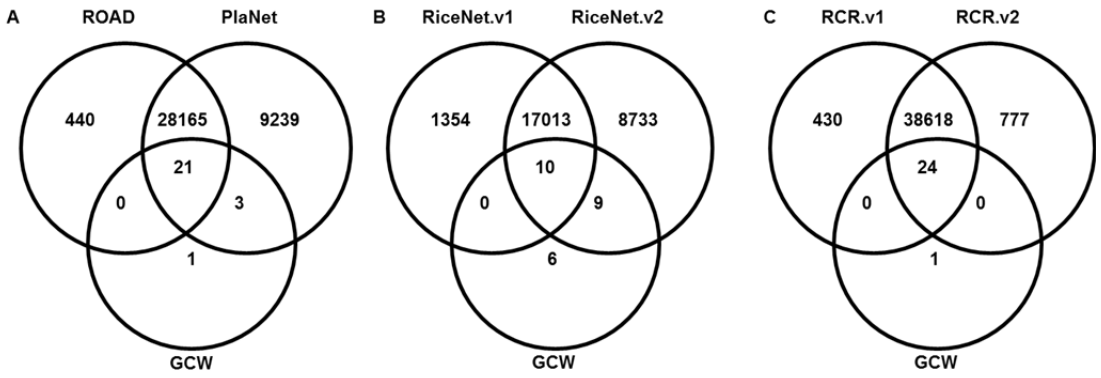


Figure 3-1 Representation of functionally characterized and putative "grass cell wall specific" genes shows that the high-quality RiceNet misses interactions for more than 50% of putative and known grass cell wall specific genes. The 25 so-called grass cell wall specific genes include the "Michal-clade" BAHD acyltransferases, MLG biosynthesis genes, *OsSLF6*, *OsCSLF8* and *OsCSLH1*; (3) grass-expanded arabinoxylan biosynthesis genes, *OsXAX1* and *OsXAXl*.

Table 3-1 Network size and fitness to the power law, $P(k) \sim k^{-\gamma}$.

Network	Total Nodes	Total Edges	R^2 of the power law
ROAD	28,626	8,520,163	0.76
PlaNet	30,428	3,310,397	0.54
RiceNet.v1	18,377	588,221	0.89
RiceNet v2	25,765	1,775,000	0.88
RCR. v1	36,072	12,243,093	0.75
RCR. v2	36,419	13,185,506	0.74

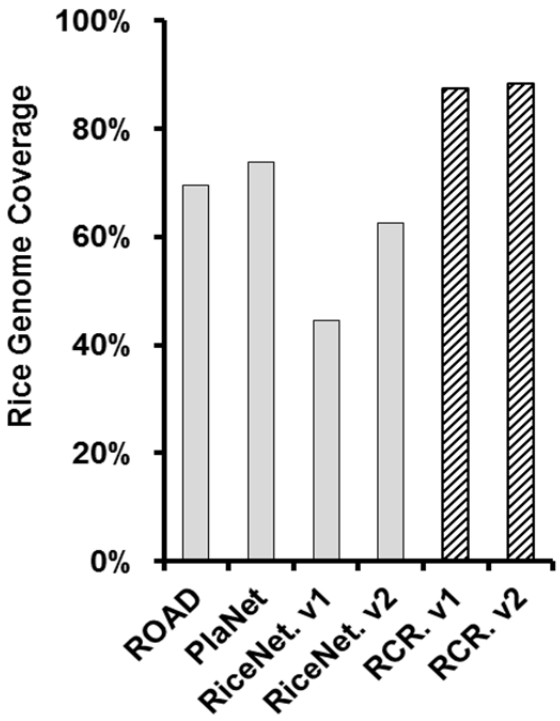


Figure 3-2 Genome coverage shows that RCR networks are more comprehensive than original ones.

We evaluated network quality based on Gene Ontology (GO) terms as genes involved in the same pathway tend to be co-expressed and co-regulated. To validate functional relatedness, we used the GO-Biological Process (BP) annotations from the Biofuel Feedstock Genomics Resources (BFGR). Forty percent of rice genes have been assigned GO-BP terms. As with RiceNet v2, we excluded ten common GO-BP terms to avoid bias with these generic terms. We measured the predictive power of each network with different edge scores using Receiver Operating Characteristic (ROC) curves and then calculated the area under the curves (AUC). The greater AUC of RCR v1 and v2, suggests that the novel, combined networks, have the highest network quality with AUCs of 0.68 and 0.69, respectively (Figure 3-3A and Supporting Figure 3-1). RiceNet v2 also shows increased network quality compared to v1. In addition, we examined

network performance using precision-recall analysis, which focuses on the evaluation of true positive predictions. In this analysis, precision is calculated as the proportion of true positive edges among all predictions at particular edge score cutoff. Recall represents the proportion of true positive edges relative to total true positives. The RiceNets and the RCR v1 and v2 networks exhibit a greater proportion of positive edges at most edge scores than the coexpression networks, with the RCRs identifying slightly more GO-BP-matched edges than than the RiceNets (Figure 3-3B). Our analyses indicated that the RCR v1 and v2 are relatively complete and high-quality networks. We used RCR network v2 to infer cell wall biosynthesis and regulatory genes of rice.

Systematic examination of known dicot SCW regulators in rice

Though useful for understanding many pathways, our main use of the RCR network has been to extend knowledge of cell wall regulation to grasses. Here, we sought to address the questions of which of the conserved secondary cell wall regulators grasses utilize and if grasses utilize novel, or previously unstudied, regulators. To test the recall of known cell wall-related interactions, we firstly extracted nodes from the RCR that share edges with 125 cell wall-associated "seed genes." The "seed genes" belong to the following three categories: (1) known rice cell wall biosynthesis genes including phenylpropanoid pathway genes, cellulose synthases (CESAs), "Mitchell-Clade" BAHD acyltransferases and xylan biosynthesis genes; (2) known rice cell wall-associated transcription factors; (3) putative orthologs of Arabidopsis known cell wall-associated transcription factors identified based on Inparanoid and phylogenetic

reconstruction (Supporting Table 3-1). The network among the 125 cell wall "seed" genes is highly interconnected, entailing 1177 interactions (i.e., edges) when considered without cut-offs (Figure 3-4, Table 3-2). This recalls 92% (97 out of 105) of characterized interactions based on the literature. However, when we include the recent yeast one-hybrid data with Arabidopsis xylem-expressed transcription factors from Taylor-Teeples et al. (2015), 26% (119 out of 460) of interactions are represented in the RCR network (Table 3-2).

Table 3-2 Recall of known interactions between transcription factors and cell wall biosynthesis genes in RCR v2.

Source	At TF-CW gene promoter interactions	Orthologous pairs in rice[b]	Present in RCR v2
Literature[#]	134	105	97
At root xylem Y1H[a]	623	355	22
Sum	757	460	119

[a] Yeast one hybrid screen of Arabidopsis root xylem genes (Taylor-Teeples et al. 2015)
[b] We transferred the known interactions from Arabidopsis to rice based on Inparanoid and only included the pairs when both orthologs exist in rice.

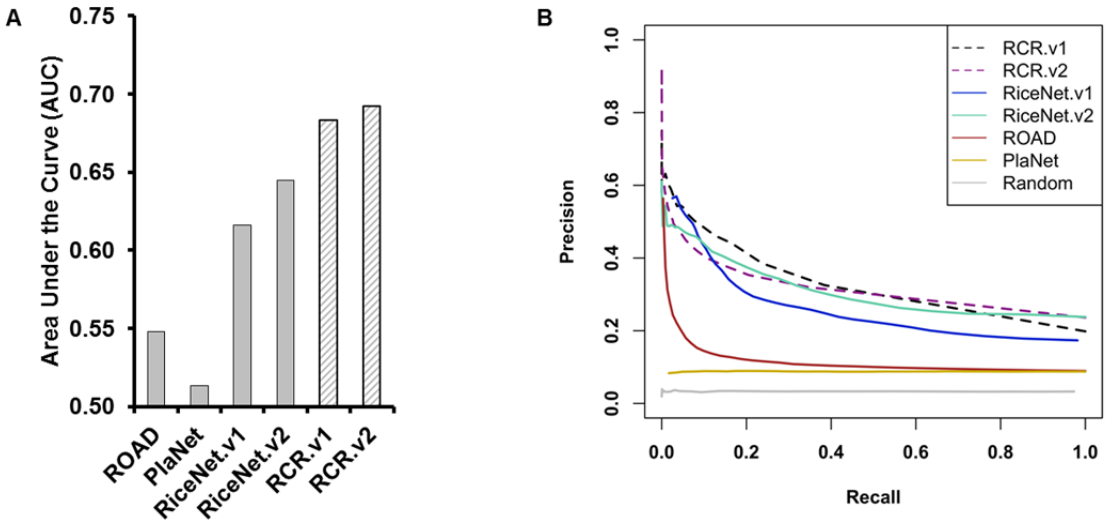


Figure 3-3 Network quality evaluation based on GO-Biological Process (BP) annotations. (A) The area under the Receiver Operating Characteristic curve indicates that RCR networks v1 and v2 have better prediction power compared to each original network. (B) Precision-recall analysis supports that RCR networks have better performance than original networks.

We further tested RCR network quality based on identification predominant cell wall biosynthesis gene family members versus minor members that do not have wide-spread function in cell wall synthesis (Figure 3-4). Though most of cell wall-related gene families have not been systematically examined, reverse genetics studies of a few lignin-related families have revealed their expansion in plants. In addition, only predominant members with relatively high expression level are responsible for the catalysis of biochemical reactions within each family. For example, CAD, CCR and COMT are gene families involved in the phenylpropanoid pathway and compose nine, five and five members in rice, respectively. Among them, OsCAD2, OsCCR1 and OsCOMT1 are the predominant members (Hirano et al., 2012). In RCR v2 network, OsCAD2, OsCCR1 and OsCOMT1 show higher degree (i.e., have more interactions) with other cell wall-related genes compared to their paralogs (Figure 3-4). In all, RCR v2 network shows high performance to infer cell wall interactions and can further predict predominant members functioning in rice.

To address the question of conservation of cell wall regulatory mechanisms between Arabidopsis and rice, we compared the relative network connectivity of Arabidopsis SCW regulators and rice orthologs. Based on molecular genetics studies in Arabidopsis, a few transcription factors function as the core regulators, and knockouts or dominant negative alleles dramatically decrease overall SCW biosynthesis. AtNST1/2 and AtMYB46/83 are four such genes. Within the ATTED II coexpression network, we observed that the core Arabidopsis SCW regulators possess relatively high numbers of edges (Figure 3-5), consistent with the observation that core regulators act as hub genes within biological networks. Several other transcription factors that have

confirmed, but more limited roles in cell wall regulation, such as AtMYB52, have relatively fewer edges. In addition, Arabidopsis gene expression atlas datasets reveal that the hub genes are relatively highly expressed in the stem compared to other tissues (Figure 3-6).

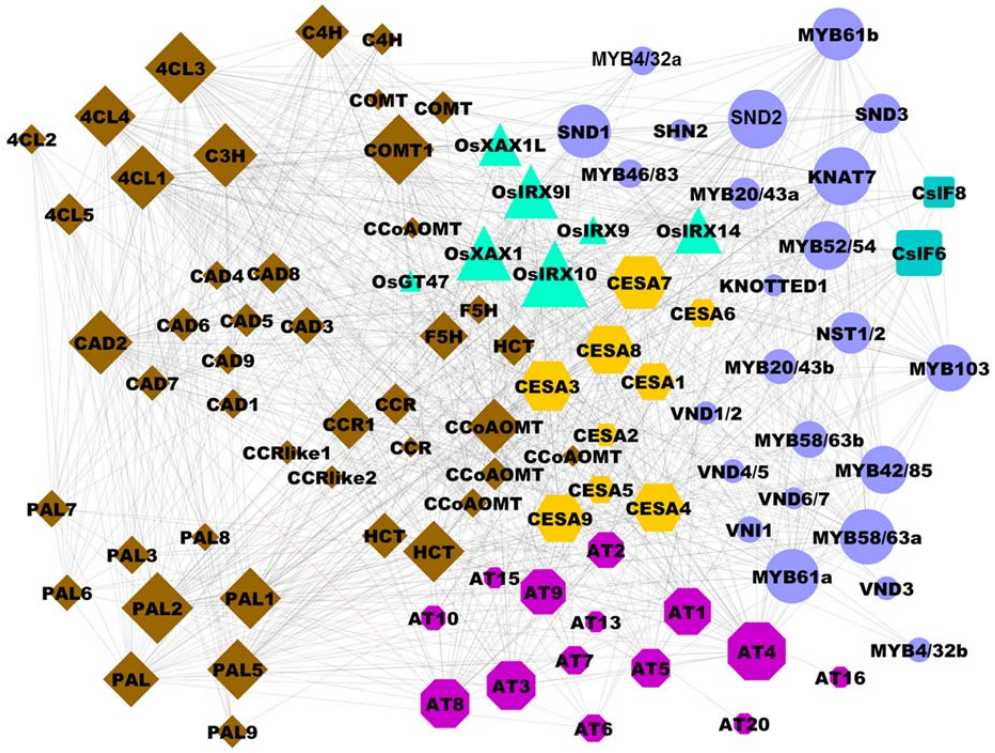


Figure 3-4 Interactions among rice cell wall-related genes within the RCR v2. Different colors and shapes represent various gene families involved in cell wall biosynthesis and regulation. The size of each node is proportional to the degree (number of connections).

A similar examination of rice orthologs in the RCR revealed that the rice transcription factors also cluster into two groups based on the number of edges with cell

wall biosynthesis genes (Figure 3-7 and Figure 3-8). Group i members are well connect with lignin, xylan biosynthesis genes, SCW associated CESAs (OsCESA4, OsCESA7 and OsCESA9) and primary cell wall CESAs. Group ii members show relatively fewer edges. We observed that OsNST1/2 still acts as a hub. Whereas, OsMYB46/83 show lower degree compared to their orthologs in Arabidopsis (Figure 3-7). Moreover, network connectivity and gene expression also suggest that OsMYB61a, OsMYB61b, OsSND2 and OsSND3 may also function as predominant regulators in rice.

Identification of novel rice cell wall-associated transcription factors

To identify novel transcription factors controlling rice cell wall biosynthesis, we assembled the Rice Cell Wall Network around the 125 cell wall "seed" genes. We observe 1790 non-bait nodes and 3139 edges with an RCR score ≥ 0.03 (Figure 3-9). Of these, 215 of 1790 of them are transcription factors. To better select cell wall-related transcription factors, we excluded 55% (118 out of 214) transcription factors that connect with fewer than five cell wall "seed" genes. The remaining 96 transcription factors are putative novel regulators of cell wall biosynthesis.

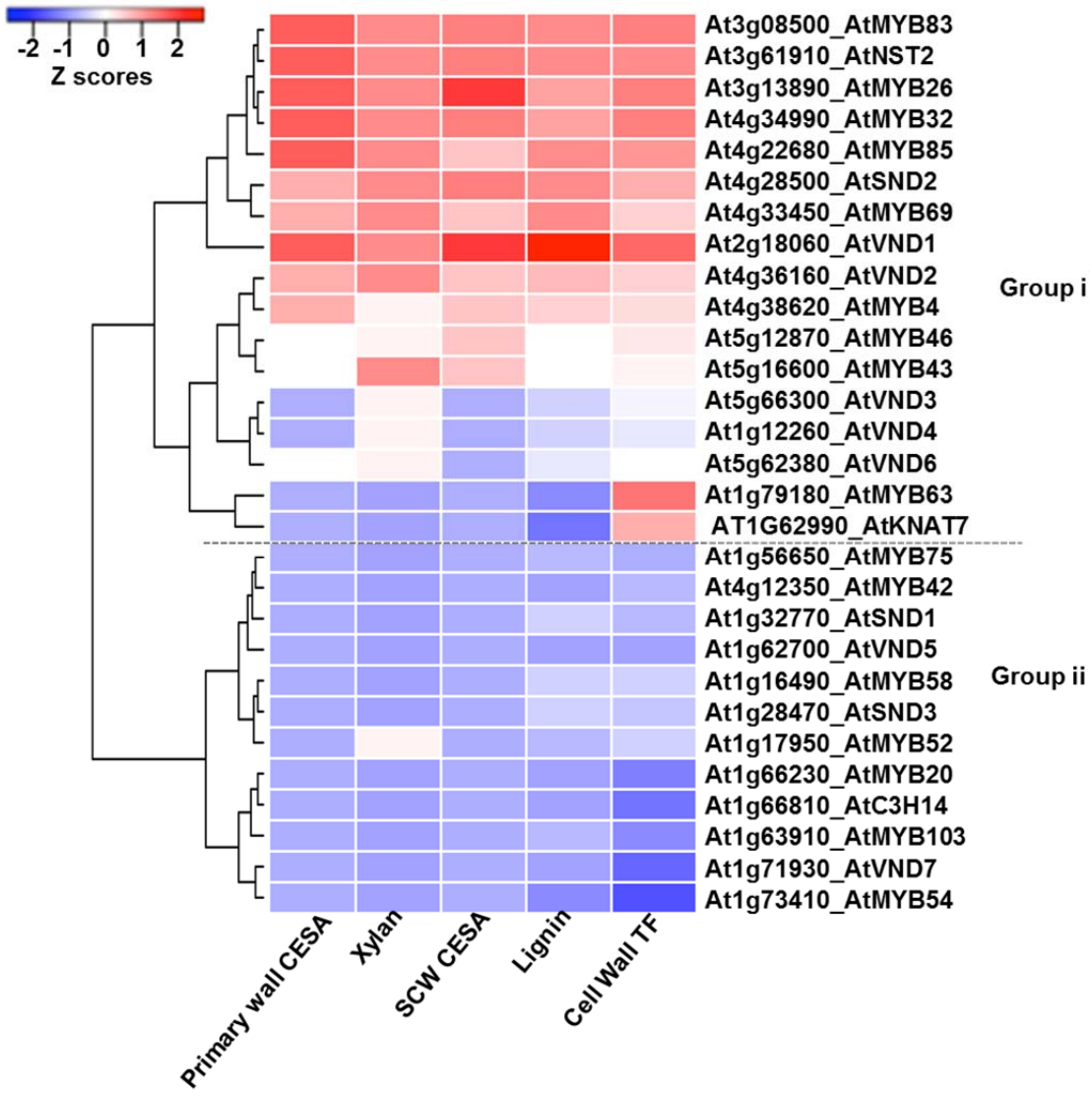


Figure 3-5 We extracted all interactions for Arabidopsis SCW transcription factors in ATTED II with default cutoff and only included interactions with Arabidopsis cell wall biosynthesis genes.

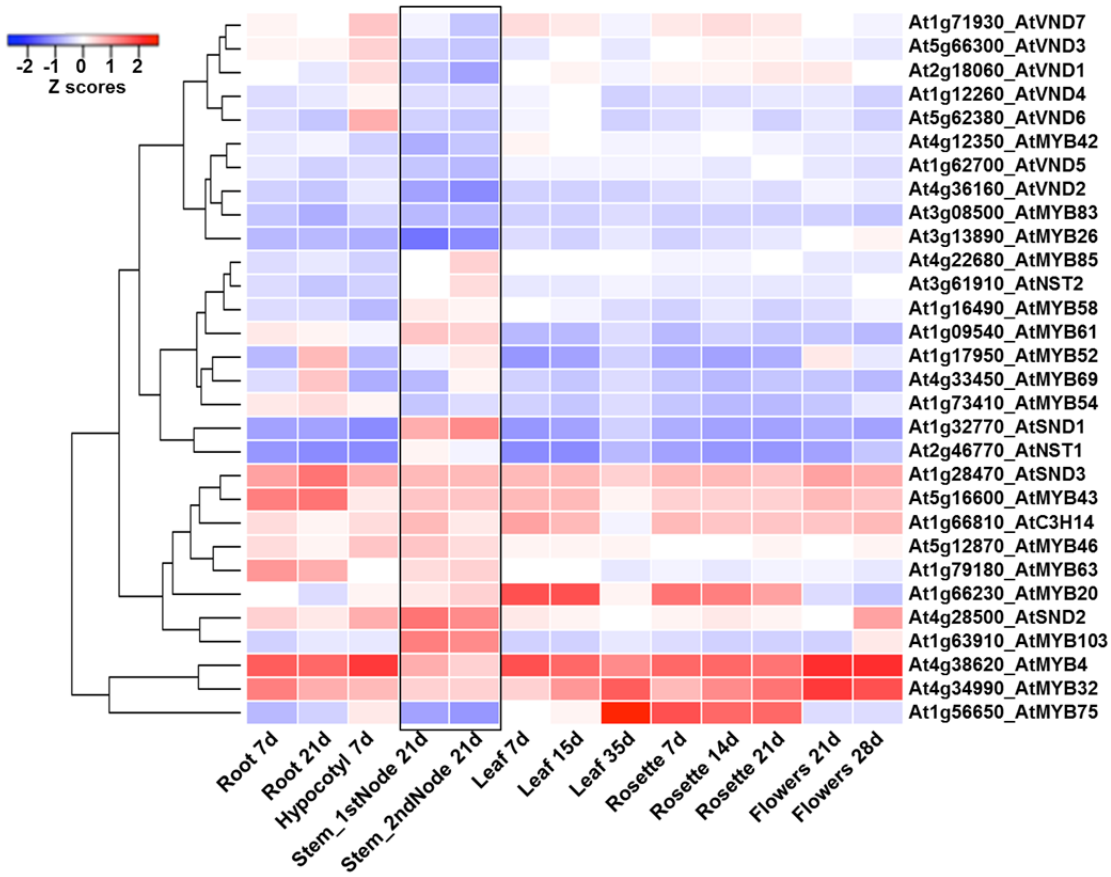


Figure 3-6 Expression pattern of Arabidopsis known SCW transcription factors during development. The data was extracted from the Arabidopsis expression atlas generated using Affymetrix microarrays (Schmid et al., 2005). The heatmap was built in r and rows were grouped by hierarchical clustering.

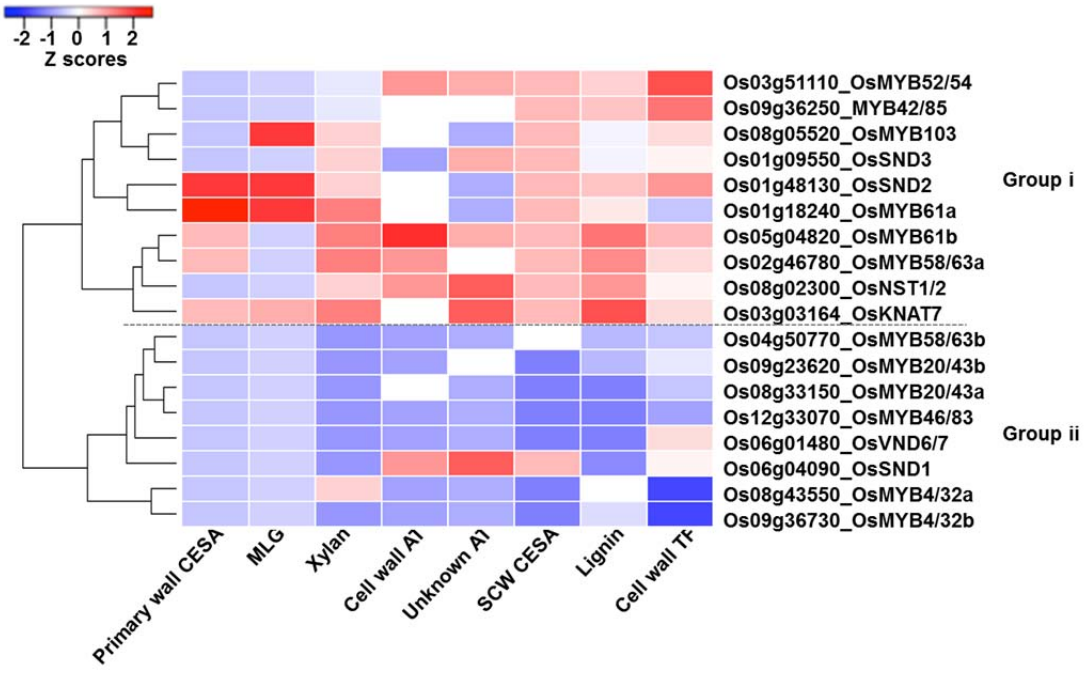


Figure 3-7 Summary of interactions in RCR v2 (no cut-off) between orthologs of known Arabidopsis SCW transcription factors and different classes of cell wall biosynthesis genes. Hierarchical clustering separates the transcription factors into two groups.

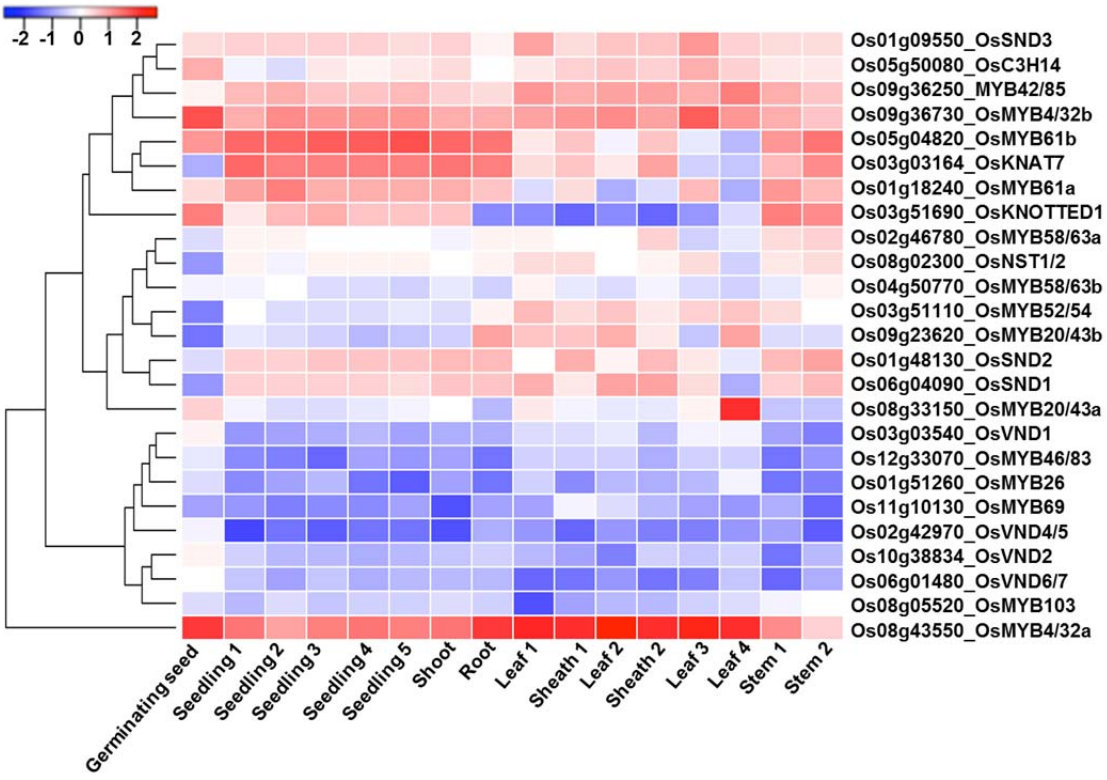


Figure 3-8 Expression pattern of rice SCW transcription factors included in as seed genes during development. The data was extracted from the rice expression atlas generated using Affymetrix rice microarrays (Wang et al., 2010). The heatmap was built in R and rows were grouped by hierarchical clustering.

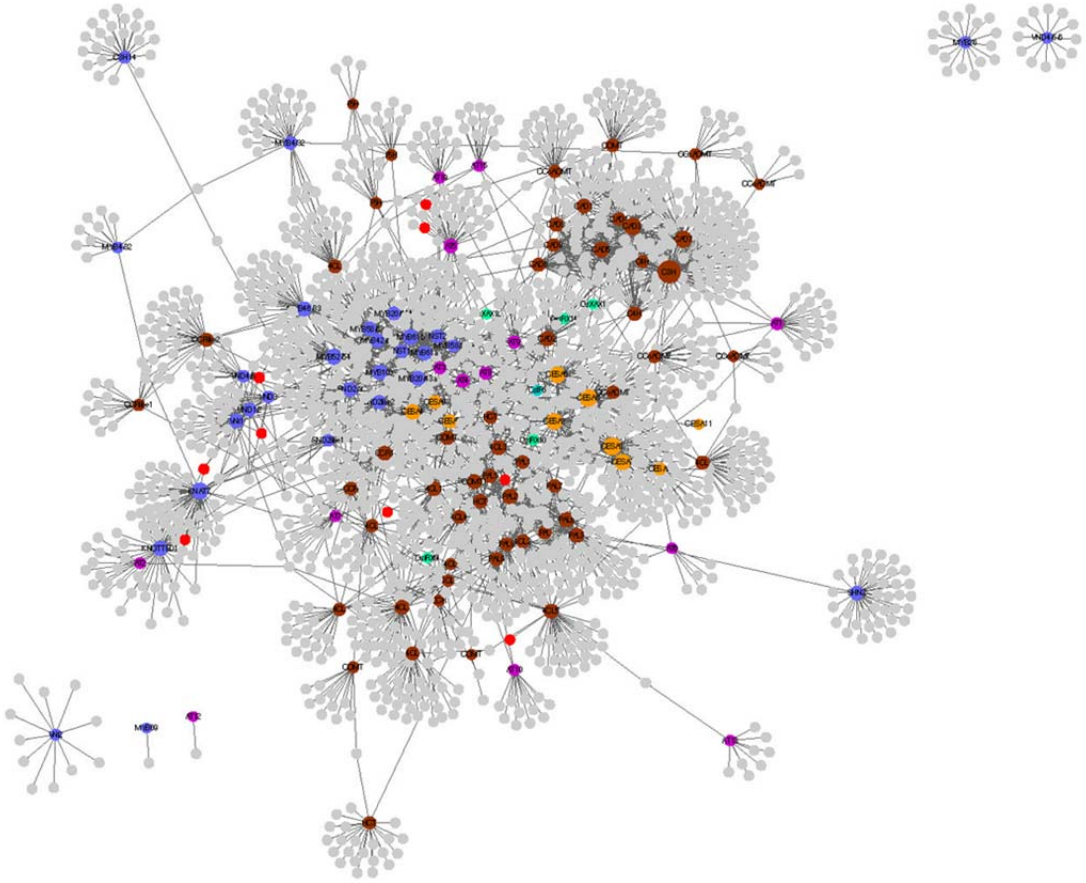


Figure 3-9 Rice cell wall network with edge scores ≥ 0.03. Blue, brown, purple and green coded nodes are cell wall related genes used as seed genes to construct this network. Red nodes represent putative novel transcription factors tested in transient gene expression analysis.

Besides expanding understanding of different protein families regulating rice cell wall biosynthesis, the putative cell wall related transcription factors are from 19 protein families, namely AP2/ERF, ARF, BES1, bHLH, bZIP, C2H2, C3H, DBB, Dof, GATA, GRAS, GRF, HD-ZIP, HSF, MIKC, MYB, NAC, TALE and WRKY (Table 3-3 and Supporting Tables 3-4). Based on their connection patterns with cell wall biosynthesis genes, putative novel cell wall associated transcription factors can be divided into three groups (Figure 3-10). Group i member are relatively well connected with most categories of cell wall genes except primary cell wall CESAs and MLG biosynthesis genes. Group ii members are relatively less connected; however, a few members within this group show specific connections with primary cell wall CESAs and MLG biosynthesis genes. The Group iii members connect primarily with other homologs of known Arabidopsis cell wall transcription factors.

Table 3-3 Summary of putative novel cell wall associated transcription factor families identified in the cell wall network based on RCR v2. In total, we identified 96 transcription factors from 19 protein families.

TF Family	Total TF
AP2/ERF	9
ARF	2
BES1	1
bHLH	6
bZIP	1
C2H2	4
C3H	1
DBB	5
Dof	1
GATA	3
GRAS	1
GRF	1
HD-ZIP	7
HSF	1
MIKC	2
MYB	19
NAC	16
TALE	10
WRKY	6

* indicates the function of transcription factors have been tested in rice.

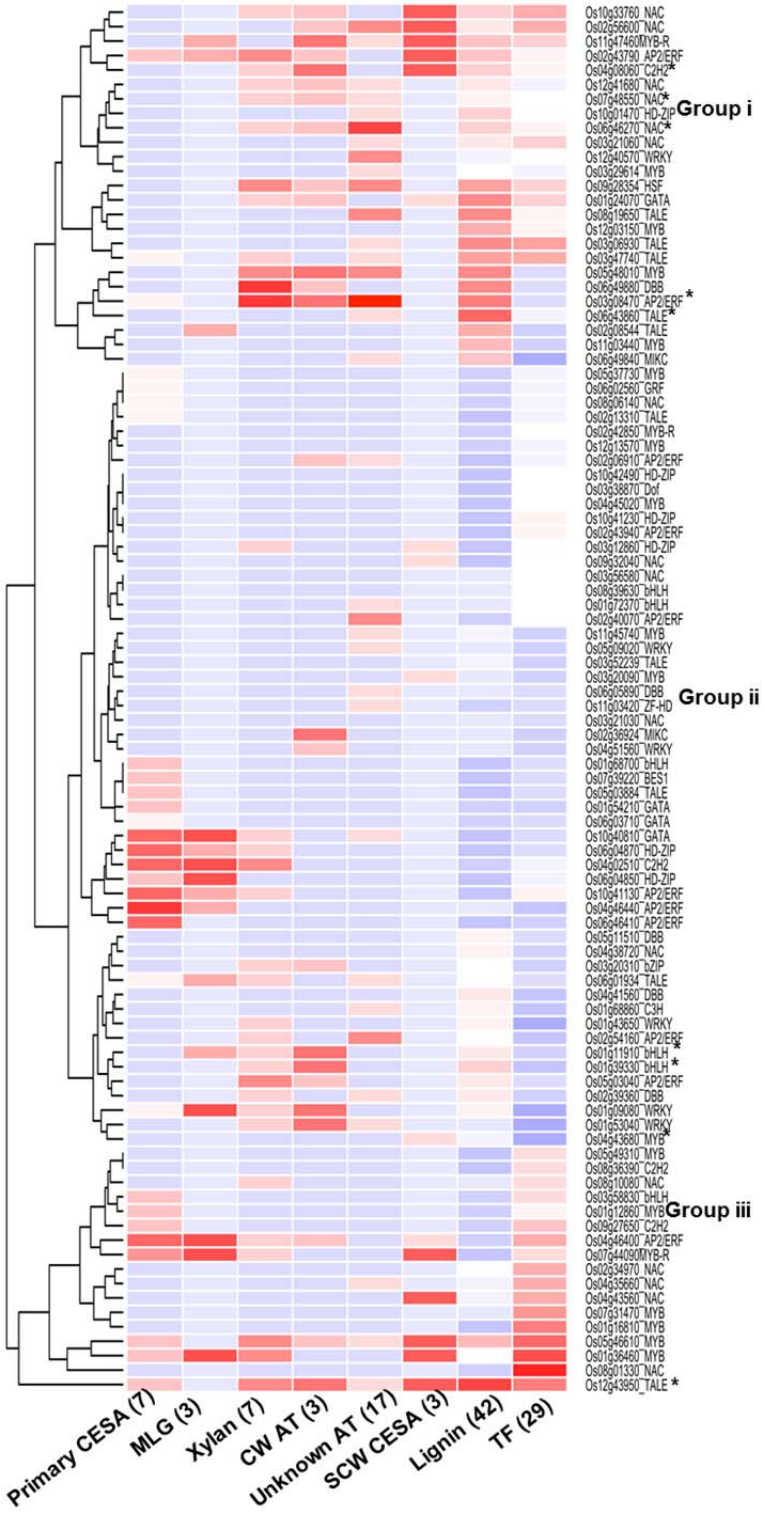


Figure 3-10 Interactions between putative novel SCW associated transcription factors and known cell wall related genes in RCR v2 without cutoff.

Reverse genetics supports diverged function of OsMYB61a

To begin to explore the function of transcription factors implicated by the network analysis, we focused on OsMYB61a, which is one of two co-orthologs of AtMYB61, a regulator of cell wall synthesis and possibly other carbon sink physiology (Romano 2012). The RCR v2 network suggests that in addition to regulating CESA and lignin biosynthesis genes as in Arabidopsis, OsMYB61a may have acquired the ability to activate grass cell wall-specific genes (Figure 3-11). We characterized a knockout mutant line, *myb61a,* which has a T-DNA insertion in the third exon (Figure 3-12A). Quantitative reverse transcriptase-PCR with primers designed near the 3' end of the transcript indicated that expression of *OsMYB61a* decreased five-fold in mature leaves of the mutant compared to those of the negative segregants (Figure 3-12B). *myb61a* plants also show a dwarf phenotype relative to the wild type (36% decrease, p-value < 10^{-5}), with each internode of *myb61a* being smaller than those from wild type plants (Figure 3-12C).

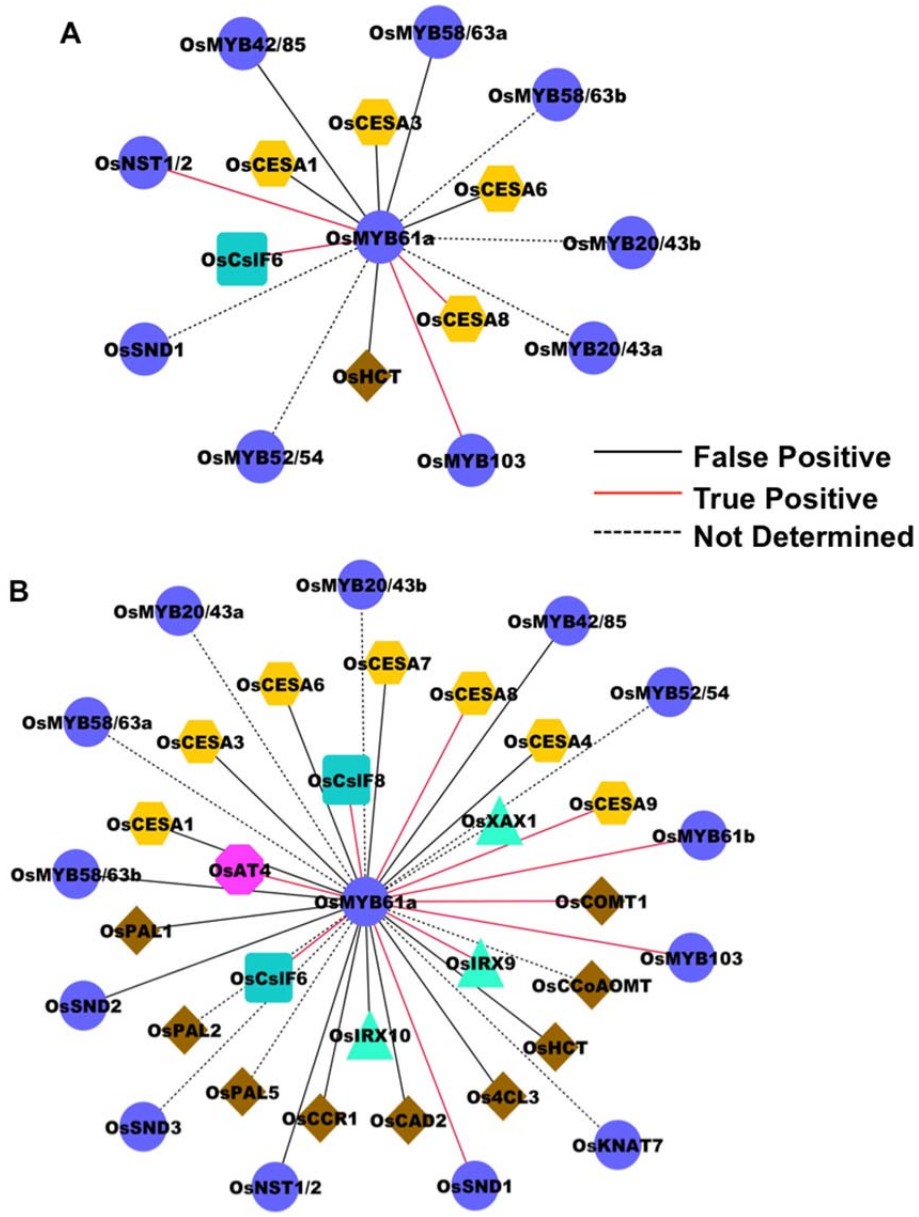


Figure 3-11 The RCR expands the potential interactions for OsMYB61a compared to RiceNet.v2. A. OsMYB61a subnetwork in RiceNet v2. B. OsMYB61a sub-network in the RCR network. False positive edges represent examined gene expression without significant change. True positive edges represent examined gene expression with significant change. Not determined edges represent untested gene expression in this analysis.

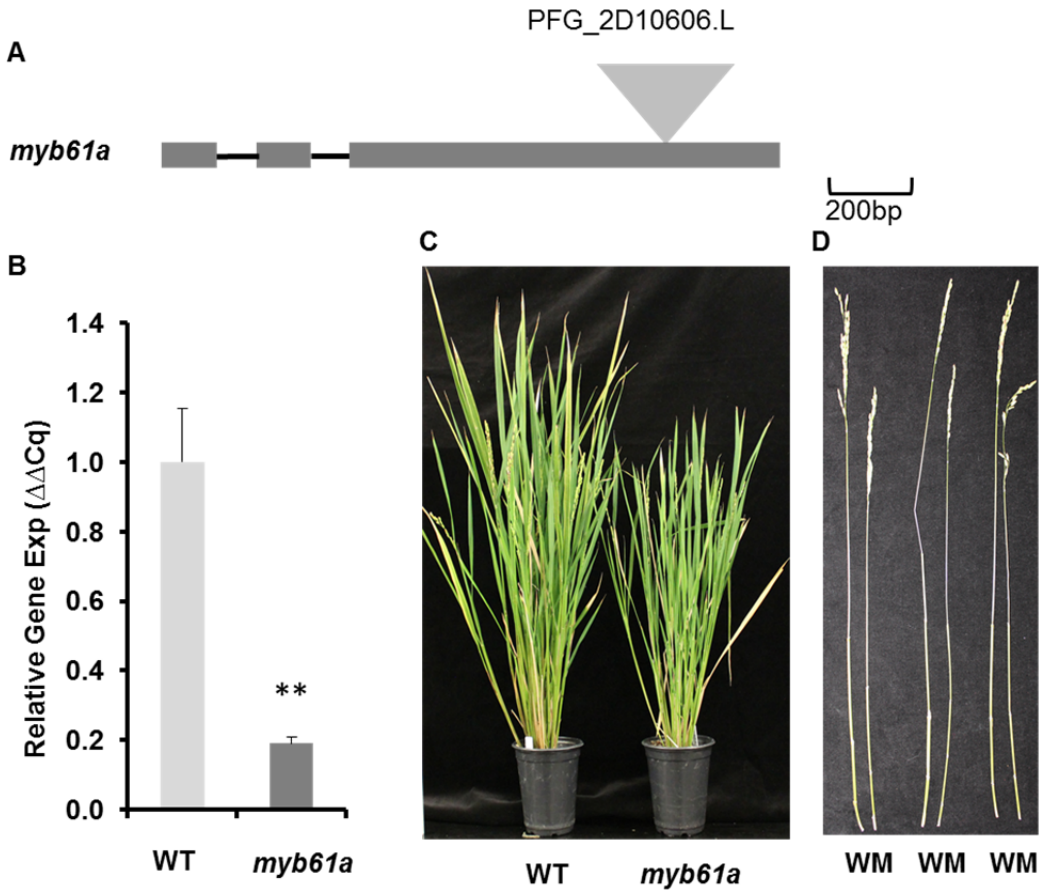


Figure 3-12 Summary of myb61a mutant genetic information and Phenotype. (A) Insertion map of *myb61a* with the line number: PFG_2D10906.L. (B) Fold change of *myb61a* in mutants comparing to the negative segragants. Two-month old leaf samples were used for gene expression analysis. Error bars represent standard deviation. ** represents two-tail student t test p value < 0.01. (C) *myb6a* plants show dwarf phenotype and each internode was shorter than negative segragants (D).

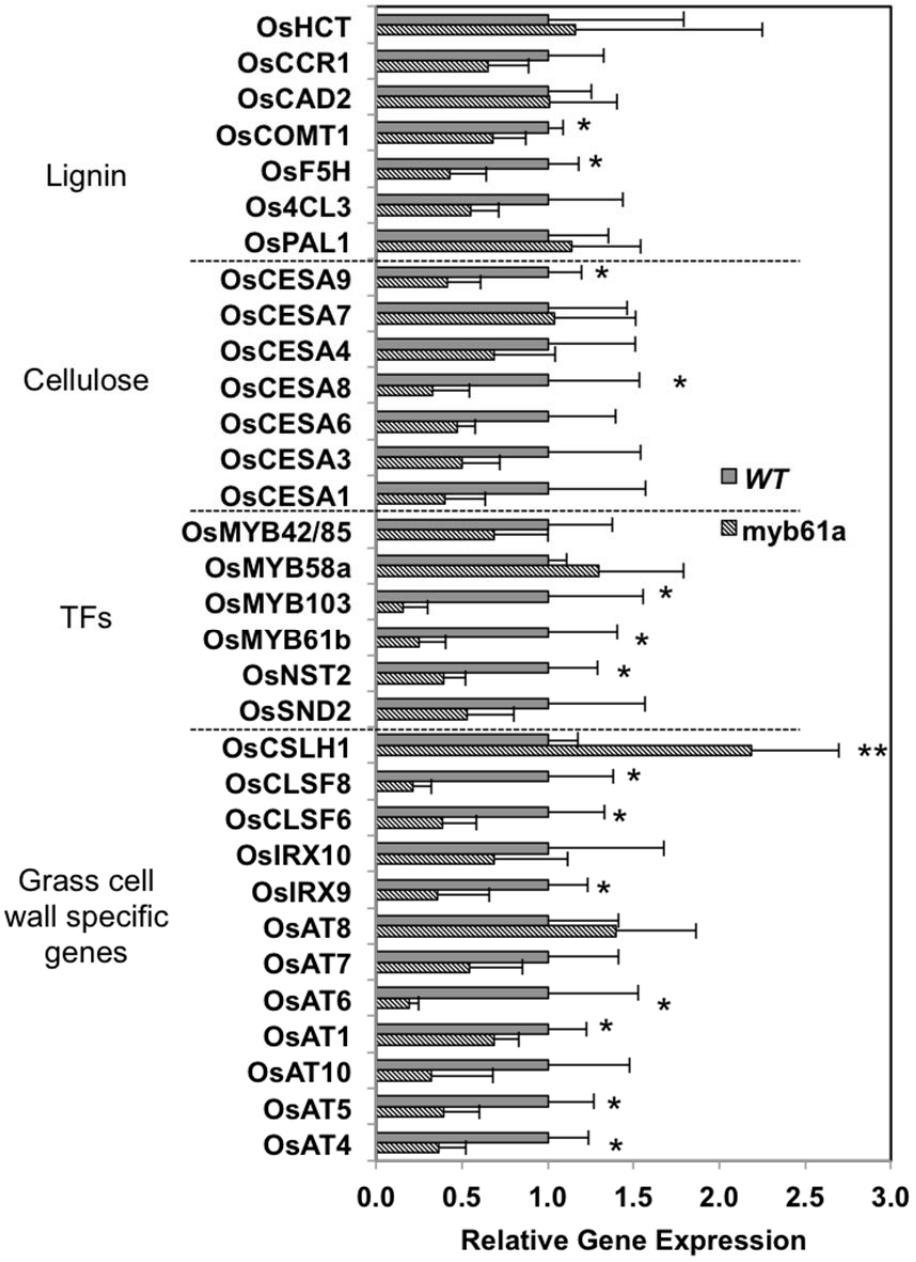


Figure 3-13 Expression analysis of cell wall-related genes in myb61a. In addition to reduced expression lignin biosynthesis and CESA genes, grass cell wall diverged genes also show reduced expression. Leaf samples were harvested from two-month old rice plants and five biological replicates were used in this experiment. Error bars represent standard deviation. * and ** represent two-tail t-test p-values lower than 0.05 and 0.01, respectively.

Further work revealed that OsMYB61a regulates cellulose and lignin biosynthesis gene expression and content (Figures 3-13 and 3-14). Following up on potential regulatory interactions inferred by edges in the RCR, we tested 30 cell wall genes for alterations in gene expression in the *myb61a* mutant (Figure 3-13). In *myb61a*, the expression of lignin biosynthesis genes, *OsCOMT1* and *OsF5H*, and SCW cellulose biosynthesis gene, *OsCESA9* are significantly reduced relative to wild-type plants. These data are consistent with a recent study demonstrating that OsMYB61a is able to activate SCW-associated CESAs when responding to gibberellin acid signaling (Huang et al. 2015). In addition, we measured expression of all cell wall-specific genes connected with OsMYB61a predicted by RCRv2, namely, *OsAT4*, *OsAT5*, *OsCSLF6*, *OsCSLF8* and *OsCSLH1*, and observed that all but *OsCSLH1* were significantly reduced (Figure 3-13). Surprisingly, *OsCSLH1* shows increased expression in *myb61a*, which may suggest that it has a different regulatory mechanism than *OsCSLF6* and *OsCSLF8*. Two uncharacterized BAHD-AT encoding genes, *OsAT1* and *OsAT6,* also show reduced expression in *myb61a*, this provides insights to further functionally characterize cell wall associated acyltransferase in grasses.

Using stringent criteria, Romano et al. (2012) identified and confirmed three direct targets of AtMYB61, *AtKNAT7*, a pectin methyl esterase, and a lignin biosynthesis gene. However, given the typical complexity of transcription factor binding, it seems likely that MYB61 has additional targets, whether direct or indirect. To begin to fill this gap in rice, we measured expression of six orthologs of Arabidopsis SCW-associated transcription factors selected based on the two following criteria: (1) They connect with OsMYB61a based on RCR v2 prediction; (2) They show relatively

high expression during rice development and are well connected with cell wall biosynthesis genes. *OsMYB61b*, *OsSND2* and *OsMYB103* show reduced expression in *myb61a* mutants; whereas *OsMYB52* and *OsMYB58a* did not (Figure 3-13). This suggests that OsMYB61a is upstream of OsMYB103 and OsMYB61b in the cell wall regulatory cascade of rice.

The gene expression measurements presented a further opportunity to compare the performance of RCR v2 and the high-quality functional network, RiceNet v2, to predict cell wall-related interactions. RiceNet v2 and RCR v2 predict 15 and 36 interactions between OsMYB61a and cell wall genes, respectively (Figure 3-11). Gene expression results show a similar positive rate between RiceNet v2 (40%; 4 out 9 interactions validated) and RCR v2 (39.1%, 9 out 23 interactions validated). We also examined false negative rate, which represents validated interactions in gene expression that are not predicted by the networks. The false negative percentage for RiceNet v2 is 53%, which is much higher than that of RCR v2 (8.3%). Especially, RiceNet v2 misses interactions with BAHD-ATs and rice xylan biosynthesis genes. In all, these results suggest that RCR v2 is a comprehensive, high quality network suitable for study biological process in rice, including those that are not conserved.

To explore the phenotypic consequences of disrupting OsMYB61a in rice, we measured the cell wall composition of *myb61a* and wild-type, negative segregant plants. We found that acytl bromide soluble lignin (ABSL) and cellulose contents are significantly reduced in *myb61a* by 20% and 21%, respectively (Figure 3-14). We also measured the content of grass cell wall-specific components, MLG and wall-associated HCAs. A lichenase assay showed that MLG is significantly reduced by 36% (two-tail

student t-test p-value 0.0012) in the stem of *myb61a* (Figure 3-14). To represent cell wall associated HCAs, we measured by alkali-labile hydroxycinnamoyl ester content in cell wall alcohol-insoluble residue (AIR) of leaf samples. FA and *p*CA are reduced by 24.9% and 12.2%, respectively (Figure 3-14). Taken together, these results provide evidence that grasses have evolved so that OsMYB61a is able to regulate, either directly or indirectly, the expression of grass-diverged cell wall enzymes that Arabidopsis and other dicots lack. Furthermore, we have observed regulation of grass-expanded acyltransferase and cellulose synthase like (CSL) genes coding MLG.

Functional analysis of rice orthologs of Arabidopsis cell wall transcription factors

To accelerate functional exploration of the network, we took advantage of a transient rice protoplast-based gene expression platform to rapidly identify transcription factors that regulate rice cell wall biosynthesis genes and regulators. This assay can detect both direct effects due to binding of a regulator to a promoter and indirect effects caused by a regulator altering the amount or affinity of another regulator that binds to a promoter. We report here the results of ectopic expression of 15 of the 96 high degree cell wall-associated transcription factors from the RCR v2. Expression was driven by the cauliflower mosaic virus 35S promoter, which is moderately strong in grass cells. In this section, we focus on orthologs or co-orthologs of Arabidopsis cell wall regulators that appear to have "predominant," i.e., hub, roles in rice cell wall regulation. In the next section, we discuss regulators that have not, to our knowledge, been shown to function in cell wall regulation. The validated relationships from the transient gene expression assays are summarized in Figure 3-15.

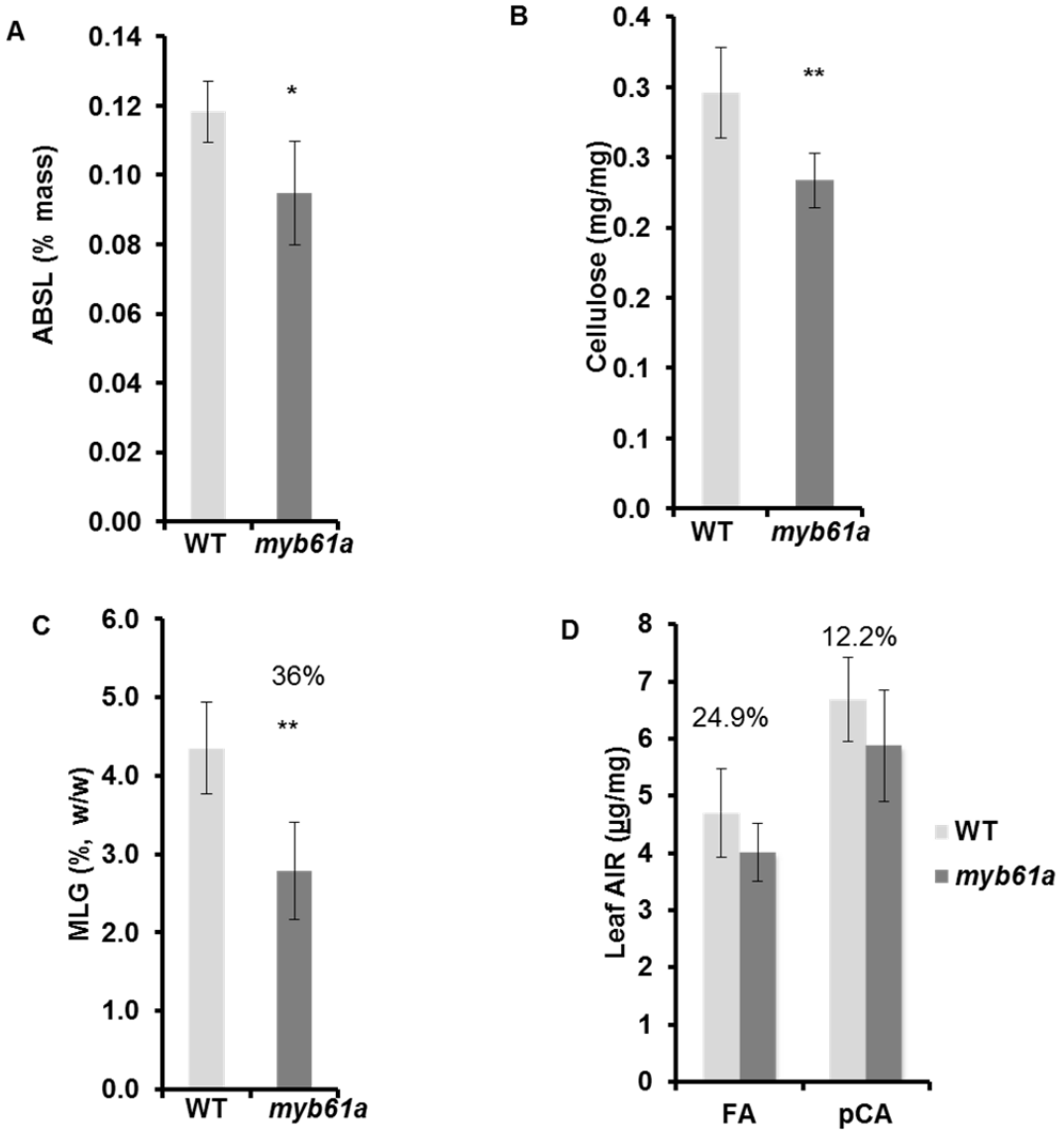


Figure 3-14 Examination of cell wall components alteration in myb61a plants. (A) Acetylbromide soluble lignin (ABSL) lignin content significantly reduced in mutants by 20%. (B) Cellulose content decreased 21% measured by antrone assays. (C) Mixed-linkage glucan significantly decreased 36% in mutants. Lichenase assay was used to measure MLG content. (D). The content of cell wall associated HCAs within leaf samples. Developmentally matched three-month old leaf or stem samples were used for *myb61a* cell wall composition analysis. We used five biological replicates for the assays. Error bars represent standard deviation. Two-tail student t test was used to compare wild type plants and mutants.

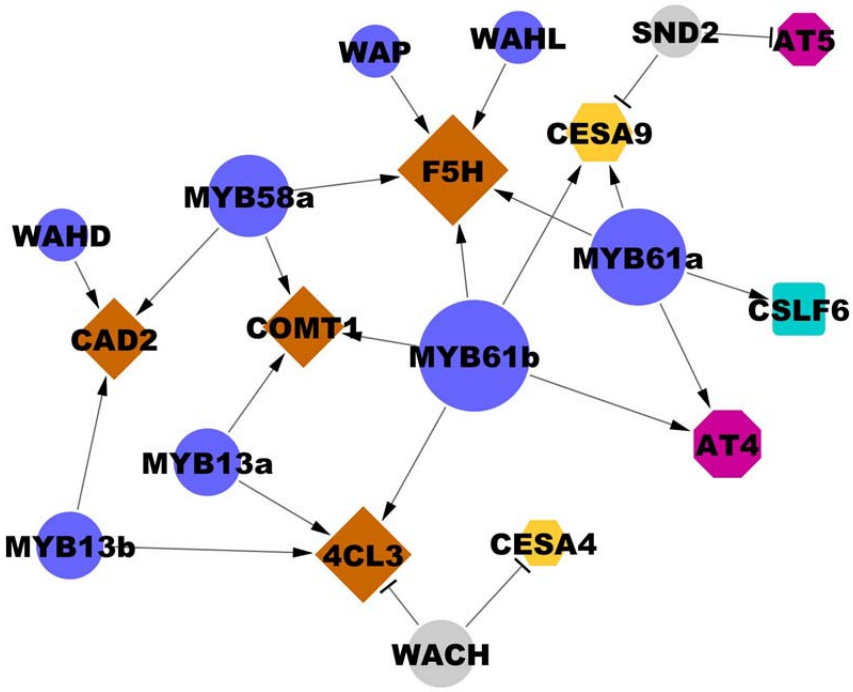


Figure 3-15 Transient gene expression analysis validated interactions between transcription factors and cell wall biosynthesis genes. Three biological replicates were used in the experiments. All the results were repeated independently. "→" represents activation and "–|" represents repression.

To test the sensitivity and specificity of the transient assay, we initially overexpressed OsMYB61a, and were able to recapitulate many cell wall gene expression changes expected from the whole plant studies (Table 3-4, Figure 3-15). We observed that four out of ten measurements in common between these two systems were consistently significantly altered in both whole plant down-regulation and protoplast up-regulation, including, *OsF5H*, *OsCESA9*, *OsCSLF6* and *OsAT4*. In contrast, OsCOMT1, OsCSLH1 and OsAT5, which exhibited significantly altered gene expression in *myb61a*, do not show a change in the transient overexpression assay. Thus, the transient over expression assay appears to be less sensitive than the stable whole plant system. Nonetheless, the results still support that conclusion that the grass

genome has evolved so that OsMYB61a activates grass cell wall-specific genes, either through a direct interaction or by regulating another transcription factor.

Table 3-4 Relative normalized gene expression results in rice protoplasts over expressing rice orthologs of Arabidopsis secondary cell wall (SCW) transcription factors (TFs). Data are the average and standard deviation of the fold-change of the listed transcript due to expression relative to controls of the regulator of interest under the control of the *35S* promoter. Three replicates were used in each assay and controls were transformed with an empty vector.

Category	Transcript	OsMYB58a	OsMYB61a	OsMYB61b	OsSND2
TFs	NA	589.3 ± 21.6**	191± 28.8*	324.4 ± 13.3**	268.1±12.4**
Lignin	Os4CL3	4.7 ± 0.3*	1.0 ± 0.2	1.4 ± 0.2*	ND
	OsCOMT1	5.1 ± 0.7*	1.3 ± 0.3	1.6 ± 0.1*	0.8 ± 0.1
	OsF5H	1.8 ± 0.1*	1.6 ± 0.2*	1.6 ± 0.2*	0.9 ± 0.1
	OsCCR1	1.2 ± 0.2	1.2 ± 0.20	1.3 ± 0.2	1.0 ± 0.1
	OsCAD2	2.5 ± 0.2*	ND	ND	1.3 ± 0.2
SCW CESA	OsCESA4	0.9 ± 0.1	ND	ND	0.4 ± 0.3
	OsCESA9	1.0 ± 0.3	1.9 ± 0.2*	2.1 ± 0.3*	0.3 ± 0.1*
MLG	OsCSLF6	ND	1.6 ± 0.1*	1.3 ± 0.1	ND
	OsCSLH1	ND	1.1 ± 0.1	1.1 ± 0.2	ND
HCA	OsAT4	1.8 ± 0.2*	1.8 ± 0.1**	1.7 ± 0.1*	0.3 ± 0.2*
	OsAT5	1.1 ± 0.1	1.3 ± 0.2	2.0 ± 0.3*	0.3 ± 0.1*

* : two-tail t-test p value lower than 0.05.
**: two-tail t-test p value lower than 0.01.
ND indicates the interactions were not determined in this assay since we examined interactions based on RCR network prediction.
Data are representative of a single experiment. All the experiments were repeated independently two to three times. Shaded boxes demarcate repeatable significant differences.

As a paralog of OsMYB61a, we found that OsMYB61b can also activate grass cell wall-specific genes in addition to lignin and cellulose biosynthesis genes (Table 3-4). In the transient gene expression assay, OsMYB61b activated *OsCOMT1*, *OsF5H*,

OsCESA9 and grass-specific cell wall genes, *OsAT4* and *OsAT5*. This indicates that OsMYB61a and OsMYB61b may function redundantly in SCW regulation; however, we cannot rule out the possibility that OsMYB61a and OsMYB61b can regulate different genes and untested within this screen.

A co-ortholog of Arabidopsis lignin biosynthesis gene transcriptional activator (Zhou et al., 2009), OsMYB58a connects with most of phenylpropanoid pathway genes in RCRv2 network (Figure 3-4). The statistically significant activation of four out of five tested lignin genes, is consistent with OsMYB58a sharing the conserved function with AtMYB58/63 (Table 3-4, Figure 3-15).

As a putative predominant rice SCW regulator from the NAC family, we screened OsSND2/NAC73 since its Arabidopsis orthologs show ambiguous function on cell wall biosynthesis (Zhong et al., 2008; Hussey et al., 2011). In the RCR network, OsSND2 connects with phenylpropanoid pathway genes and SCW related transcription factors (Figure 3-4). In addition, it also connects with five acyltransferase, including functionally characterized cell wall-related members (Figure 3-4). Transient gene overexpression assay shows that OsSND2/NAC73 may function as a repressor in rice by reducing the expression of *OsAT5* and *OsCESA9* (Table 3-4, Figure 3-15).

Functional analysis of novel cell wall regulators

To extend our understanding of SCW regulation, we selected eleven unstudied cell wall transcription factors for functional analysis using the protoplast transient gene expression platform. In addition to including novel members from well-studied cell wall transcription factors families (i.e. R2R3 MYB and NAC), we examined candidates

exhibiting high degree with cell wall biosynthesis genes from four other protein families (Figure 3-15, Table 3-5). For each transcription factor target, we tested for expression changes of the genes to which they connected in the RCR network-derived cell wall network. Based on their ability to significantly alter expression of cell wall genes when overexpressed in rice protoplast, we are able to validate 55% (6 out of 11) of the novel transcription factors examined (Figure 3-15, Table 3-5).

Among the novel regulators, five out six are activators from the AP2/ERF, homeodomain, basic helix-lop-helix and MYB families. The Wall-associated AP2/ERF family protein (WAP1), encoded by *LOC_Os03g08470*, significantly activated *OsF5H* (Figure 3-15, Table 3-5). Wall-Associated HomeoUomain (WAHD) encoded by *LOC_Os12g43950*, significantly activated *OsCAD2* among the nine cell wall genes examined (Figure 3-15, Table 3-5). An additional activator from the basic helix-lop-helix family (bHLH), Wall Associated bHLH (WAHL), encoded by *LOC_Os01g11910*, activated *OsAT4* and *OsF5H* i (Figure 3-15, Table 3-5). The two novel cell wall-associated transcription factors from the R2R3 MYB family are named based on their homolog in Arabidopsis, MYB13. OsMYB13a, encoded by *LOC_Os02g4151*, activated *Os4CL3* and *OsCOMT1* transcription. Whereas, OsMYB13b, encoded by *LOC_Os04g43680*, also activated *Os4CL3* and showed evidence of *OsCAD2* transcriptional increases examined (Figure 3-15, Table 3-5).

Based on the transient assay, we observed one repressor, Wall Associated C2H2 (WACH) encoded by *LOC_Os04g08060*, (Figure 9 and Table 4). In the RCR network

network, WACH connects with multiple CESAs, homologs of known cell wall transcription factors, and OsAT4 and OsAT5. Among the 11 tested rice cell wall-related genes, WACH repressed *Os4CL3* and SCW-associated *OsCESA4*. We were not able to detect signals with the other five transcription factors examined (Figure 3-15, Table 3-5).

Table 3-5 Transient gene expression changes due to over expression of results novel SCW transcription factors in rice protoplasts. Data are the average and standard deviation of the fold-change of the listed transcript due to expression relative to controls of the regulator of interest under the control of the *35S* promoter. Three replicates were used in each assay and controls were transformed with an empty vector.

Category	Transcript	02g41510 MYB13a	04g43680 MYB13b	04g08060 WACH	03g08470 WAP	12g43950 WAHD	10g39030 BLH	01g39330 bHLH	01g11910 WAHL	06g43860 KNOX	06g46270 NAC	07g48550 NAC
TF	NA	161.8 ± 12.4**	208.7 ± 30.5**	736.4 ± 49.2**	94.8 ± 16.4*	163.7 ± 12.2**	164.4 ± 19.5**	215.6 ± 18.6**	364.8 ± 33.7**	289.4 ± 36.5**	318.4 ± 18.7**	370.5 ± 25.9**
Lignin	Os4CL3	1.4 ± 0.1*	1.7 ± 0.1*	0.4 ± 0.1*	1.1 ± 0.1	1.0 ± 0.1	1.2 ± 0.1	1.3 ± 0.2	1.3 ± 0.1	1.0 ± 0.1	1.2 ± 0.1	1.0 ± 0.1
	OsCOMT1	1.6 ± 0.1*	1.1 ± 0.1	0.6 ± 0.1*	1.0 ± 0.2	0.9 ± 0.1	0.9 ± 0.1	0.9 ± 0.2	1.1 ± 0.1	1.0 ± 0.1	1.0 ± 0.1	1.1 ± 0.1
	OsF5H	2.4 ± 0.5*	1.3 ± 0.1	2.3 ± 0.3*	2.0 ± 0.2*	ND	0.8 ± 0.1	1.2 ± 0.1*	1.4 ± 0.1**	1.0 ± 0.1	1.2 ± 0.2	1.1 ± 0.1
	OsCCR1	1.6 ± 0.2	1.4 ± 0.2	1.0 ± 0.2	1.0 ± 0.4	1.0 ± 0.1	1.2 ± 0.1	1.1 ± 0.1	1.2 ± 0.1*	1.1 ± 0.1	1.2 ± 0.1	1.0 ± 0.1
	OsCAD2	0.8 ± 0.1	1.4 ± 0.1*	1.0 ± 0.2	0.9 ± 0.1	1.8 ± 0.2*	0.8 ± 0.1	ND	ND	ND	ND	ND
SCW	OsCESA4	ND	ND	0.2 ± 0.1*	1.2 ± 0.2	1.5 ± 0.3	0.9 ± 0.2	1.3 ± 0.1	1.3 ± 0.1	0.9 ± 0.1	1.0 ± 0.3	1.0 ± 0.1
CESA	OsCESA9	ND	ND	1.6 ± 0.4	1.2 ± 0.3	1.0 ± 0.1	1.2 ± 0.01	1.1 ± 0.1	ND	0.6 ± 0.2	1.6 ± 0.2	1.1 ± 0.1
MLG	OsCSLF6	ND	ND	0.9 ± 0.4	1.1 ± 0.2	1.0 ± 0.1	ND	ND	ND	ND	ND	ND
	OsCSLH1	ND	ND	1.1 ± 0.3	1.4 ± 0.2	1.1 ± 0.1	ND	ND	ND	ND	ND	ND
HCA	OsAT4	1.0 ± 0.1	1.0 ± 0.1	1.8 ± 0.1*	1.0 ± 0.1	1.2 ± 0.4	0.8 ± 0.2	1.0 ± 0.2	1.2 ± 0.1	ND	1.2 ± 0.3	0.7 ± 0.3
	OsAT5	1.0 ± 0.2	0.7 ± 0.2	1.2 ± 0.2	ND	ND	0.8 ± 0.1	0.9 ± 0.2	0.9 ± 0.2	1.1 ± 0.1	0.9 ± 0.2	0.9 ± 0.3

*: two-tail t-test p value lower than 0.05.
**: two-tail t-test p value lower than 0.01.
ND indicates the interactions that were not determined in this assay since we examined interactions based on RCR network prediction. Data are representative of a single experiment. All the experiments were repeated independently two to three times. Shaded boxes demarcate repeatable significant differences

Partitioning of grass-expanded gene families

We also used the RCR to expand understanding of unexplored members from grass cell wall-expanded families, such as BAHD acyltranferases. BAHD acyltransferase family has been expanded in grasses and other commelinids relative to dicots (Karlen, Submitted). Different members of ATs from the clade i may be likely to involve in the synthesis of different cell wall linkage in grasses based on their phylogenetic relationship, transcript abundance and preliminary screening of transgenic rice plants. However, the function and substrates of most acyltransferases have not yet been defined. To infer acyltransferases related biological pathways, we built a 1-step nework for each acyltransferase member. OsAT17 and OsAT18 are missing in this analysis due to poor connection and low gene expression. We then analyzed enrichment of biological process GO terms (hypergeometric p value < 0.05) within each acyltransferase sub-network and clustered the acyltransferases based on similarity of their enriched GO terms. Figure 3-16 summarizes the enriched GO-BP terms shared by at least four of AT sub-networks. The clade i members, OsAT1 to OsAT10 are relatively higher expressed compared to clade ii members, OsAT11 to OsAT20, thus they tend to have more enriched GO terms due to the larger number of edges. OsAT4 and OsAT5 are associated with lignin biosynthesis and both of their sub-networks are enriched in phenylpropanoid pathway genes (Figure 3-16). OsAT1 sub-network is also enriched with several cell wall related GO terms, including cellulose biosynthesis, glucose metabolic process and phenylpropanoid pathway (Figure 3-16). The GO term enrichment heatmap also provides insights on the unknown acyltranferase. For example,

OsAT9 sub-network is enriched with members like cellular glucan metabolic process and phenylpropanoid process, which could be an interesting candidate for reverse genetics studies. Moreover, most acyltransferases from clade ii representing OsAT11 to OsAT20, connect with limited genes. The enriched GO terms limited information on their putative function (Figure 3-16).

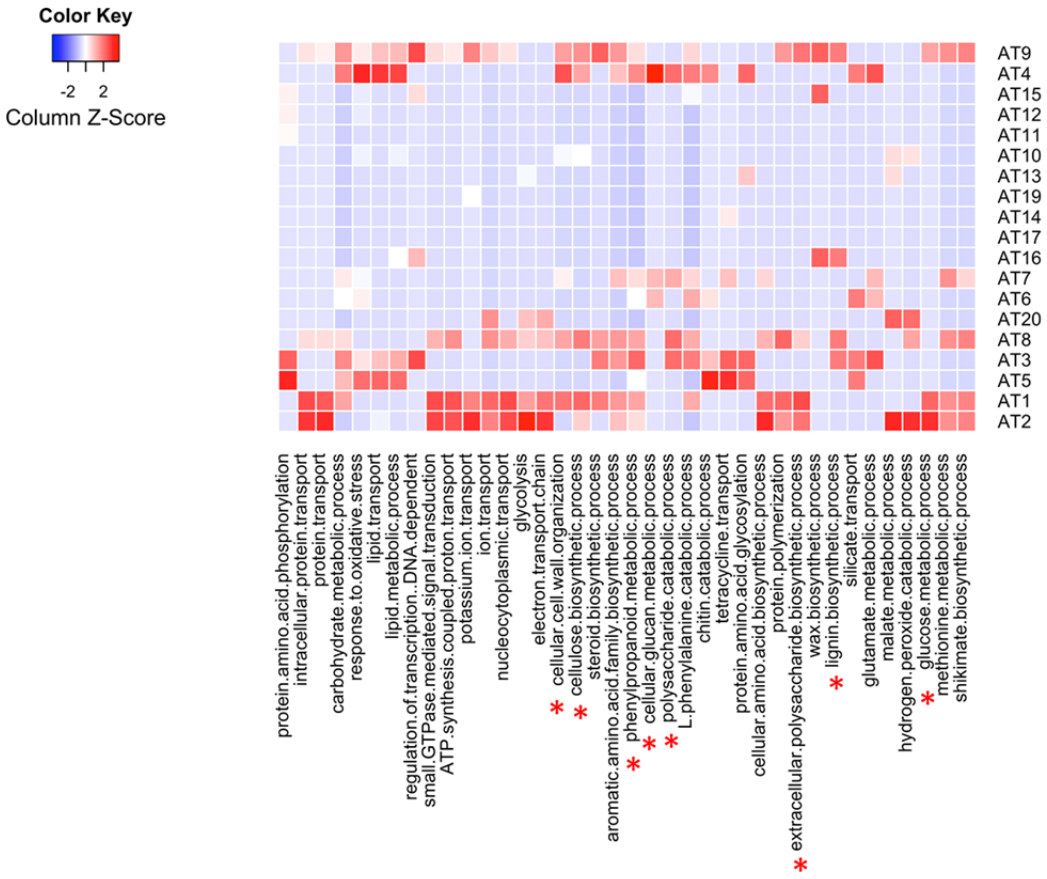


Figure 3-16 Enriched biological process GO terms of connected genes with each acyltransferase within the RCR v2 one-step network without cutoff. Red star represents cell wall related terms. To facilitate visualization, we only included GO terms enriched by ≥ 4 AT sub-networks.

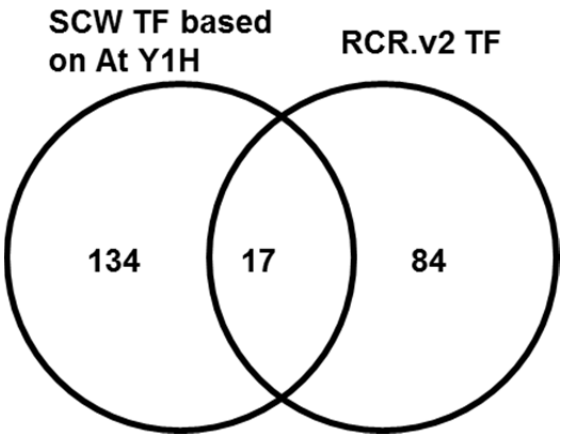


Figure 3-17 Comparison of putative novel SCW transcription factors predicted by RCR v2 to members revealed by Y1H screen of Arabidopsis root xylem SCW associated regulators. We identified orthologs of Arabidopsis genes in rice using Inparanoid.

Discussion

Superior quality of RCR networks promotes understanding of different biological pathways in rice

This work provides insights of quality comparison among networks built with different functional datasets and scoring systems using GO-BP terms based evaluations. Though relatively low depth, the Bayesian network, RiceNet v2, demonstrates better performance than conventional co-expression based networks, ROAD and the rice network incorporated into PlaNet. To construct RCR v2, we computed mutual rank for the three original networks and then took sum of inversed ranking score to weigh edges overlapped between networks. We observed superior quality of RCR v2 than original networks, which may be explained by the observation that mutual rank can improve reproducibility and overall performance of PCC-based co-expression networks (Obayashi and Kinoshita, 2009). The comparison of OsMYB61a sub-networks between RiceNet v2 and RCR v2 further supports the improved network performance of RCR v2 by eliminating false negatives while maintaining very similar true positive rates.

Prediction and functional screening of novel cell wall associated transcription factors may shed light on further improvement of gene annotation in grasses. Currently, Arabidopsis is the most well annotated plant species and approximately 40% of protein-coding genes have experimental validation based annotation (Rhee and Mutwil, 2014). The genome annotation of grass species is mostly based on sequence similarity comparing to genes in Arabidopsis and only 1% of genes in rice have experimental information validated annotations. In this analysis, we examined 11 novel transcription factors from six protein families and six of them show the ability altering the expression

of cell wall genes (Table 3-5). This information can be referred to assign biological process GO terms to undermined or non-conserved transcription factors. Besides cell wall biosynthesis pathway, the combined network, RCR v2, is suitable to expand our understanding on additional processes.

Regulation of cell wall biosynthesis in rice

In addition to maintain similar function to regulate CESA and lignin biosynthesis genes, functionally examined orthologs of Arabidopsis known cell wall transcription factors also show diverged function in rice. Reverse genetics and transient gene expression assays support that OsMYB61a and OsMYB61b can regulate CESA and lignin biosynthesis genes. In addition, we observed that OsMYB61a and OsMYB61b have evolved the ability to activate grass cell wall expanded genes. Interestingly, we observed that OsSND2 may function as a cell wall repressor by reducing the expression of *OsAT5* and *OsCESA9* in the transient overexpression assay. This suggests that OsSND2 may have undergone neo-functionalization in rice relative to Arabidopsis and function as a cell wall repressor. Previous studies show that AtSND2 can activate the promoter of AtCESA8 (Zhong et al., 2008; Hussey et al., 2011). However, the overexpression mutant of AtSND2 shows ambiguous effects, such as increased expression of Arabidopsis SCW activator, AtMYB103 and reduced SCW thickness (Hussey et al., 2011). In all, transcription factors controlling rice SCW biosynthesis may have evolved diverged function comparing to their orthologs in Arabidopsis.

Besides comparing the function of known SCW associated transcription factors between rice and Arabidopsis, we identified novel regulators from the protein that have

not been well examined on the regulation of cell wall biosynthesis pathway, including C2H2, AP2/ERF, homeodomain and bHLH families. C2H2-type zinc finger proteins have been known to repress defense and stress responses (Ciftci-Yilmaz and Mittler, 2008). We observed a C2H2 protein that may act as a cell wall repressor and its ortholog in Arabidopsis involves in stress response (Ciftci-Yilmaz and Mittler, 2008). AP2/ERF family is one of the largest protein families regulating diverse process throughout plant life cycle and development (Mizoi et al., 2012). So far, only one Arabidopsis AP2/ERF protein, SHINE/WAX INDUCER **(**SHN/WIN**)** is characterized as cell wall associated transcription factor and it can alter the leaf and stem cell wall composition when overexpression in rice (Ambavaram et al., 2011). However, the orthologs of AtSHN2 does not show higher number of connections with cell wall genes comparing to other transcription factors (Figure 3-4). Among the predicted novel rice SCW transcription factors, nine of them belong to the AP2/ERF family. The functionally screened member, *LOC_Os03g08470*, acts as a cell wall activator. Similar to AP2/ERF, plant hemoedomain proteins form a large family and largely undetermined in rice (Chan et al., 1998; Jain et al., 2008). Hirano et al screened the T_0 generation of OsBLH6, a bell- type homeodomain protein and the preliminary data suggests activation on lignin synthesis (Hirano et al., 2013). We predicted seven putative SCW associated homeodomain proteins and tested two bell-type members that are located within the neighboring clade of OsBLH6 in the whole family tree in rice. One of the two, *LOC_Os12g43950*, can significantly activate a lignin biosynthesis gene, *OsCAD2*. The last protein family we examined in this analysis is bHLH, which are present throughout the eukryotic lineages and especially expanded in land plants (Ogo, 2007;

Pires and Dolan, 2010; Xu et al., 2015). Li et al characterized 167 members in rice and 11 of them have been functionally studied using reverse genetics (Li et al., 2006). So far, none of them have been known to regulate cell wall synthesis. We tested two bHLH transcription factors and one of them, *LOC_Os01g11910,* can activate *OsAT4* and *OsF5H*. In all, we expected novel SCW associated transcription factors from the relatively undetermined protein family in rice.

We compared the putative novel rice SCW associated transcription factors to the Arabidopsis root xylem SCW regulators revealed by yeast one-hybrid screen (Y1H). The Y1H screening assays identify 197 transcription factors from 35 protein families with over-representation of AP2/ERE, bHLH, C2H2, C2C2-GATA and GRAS (Taylor-Teeples et al., 2015). We found 151 orthologs based on Inparanoid in rice. By comparing to the predicted 96 putative rice SCW transcription factors in this study, we only result in 17 common members (Table 3-3, Figure 3-17). Among them, two transcription factors were screened in transient assay, namely, *LOC_Os07g48550* and *LOC_Os04g08060*, which belongs to the NAC and C2H2 protein family. *LOC_Os07g48550* shows no effects on the expression of cell wall genes and *LOC_Os04g08060* acts as a novel cell wall repressor. Though we are aware of false positives and false negatives present in both studies, these results suggest that dicots and grasses are very likely to have their unique SCW regulators probably due to the genome-scale or tandem duplications after their divergence. For example, AtMYB75, a SCW repressor in Arabidopsis, does not have orthologs in grasses based on five-species phylogenetic study of R2R3 MYB family (Zhao and Bartley, 2014).

Incorporation of grass-expanded genes into cell wall biosynthesis pathway

Genome duplication has been a prevalent feature to drive pathway evolution and emergence of new genes in plants (Zhang, 2003; Paterson et al., 2006; Hollister, 2015). In grasses, both ancient whole genome duplication and lineage-specific duplication occurred after divergence with dicots, which also contributes to gene family expansion and emergence of new pathway (Paterson et al., 2004). For example, it has been shown that genes involved in C4 photosynthesis pathways in sorghum and maize are duplicates from a progenitor involved in C3 photosynthesis (Wang et al., 2009). For cell wall biosynthesis pathway, different studies have revealed the incorporation of grass-specific compounds. Correspondingly, genes from the grass-expanded clades/families are involved in the biosynthesis and incorporation process. For example, the known MLG biosynthesis genes, *OsCSLF6*, *OsCSLF8* and O*sCLSH1*, are from the Poaceae-expanded clade of the GT2 family (Burton et al., 2006b; Scheller and Ulvskov, 2010) Members of grass-expanded Mitchell clade of BAHD are able to incorporate HCAs to crosslink hemicellulose and lignin, which is lack in dicots (D'Auria, 2006; Bartley et al., 2013; Petrik et al., 2014; Bontpart et al., 2015).

To our knowledge, this study is the first time to identify SCW associated transcription factors that are able to control grass cell wall specific genes. We observed that orthologs of Arabidopsis known cell wall regulators may have acquired the ability to regulate cell wall related acyltransferase and MLG biosynthesis genes, such as OsMYB61a, OsMYB61b and OsSND2. On the other hand, novel cell wall transcription factors from the relatively unexamined protein families can also alter the expression of grass cell wall specific genes, such as *LOC_Os04g08060*, a C2H2 family protein. These

results suggest the following two possible models to explain the incorporation of grass-expanded genes into cell wall biosynthesis pathway: (1) The grass-expanded genes may maintain known cell wall associated DNA binding sites, thus orthologs of known SCW associated regulators can directly bind to their promoters; (2). Novel DNA binding sites may have evolved, thus different transcription factors can directly regulate grass cell wall expanded genes and they are also targets of identified transcription factors. This study not only identifies promising regulators controlling grass cell wall biosynthesis, but expands our understanding of the pathway evolution in plants.

Implication to understand grass genomes for improved biofuel production

Dicots and grasses have diverged approximately 150 million years ago and this relatively distant relationship limits the power of comparative genomics due to independent whole genome duplication events and following up chromosome rearrangements since their divergence (Nishiyama et al., 2003; Davidson et al., 2012). Lineage-specific gene family expansion and gene loss have been reported across Arabidopsis and rice. Thus, the comprehensive and high–quality rice network, RCR v2, can not only facilitate to the study grass genomics, but enable the exploration of grass or rice-specific pathways.

Currently, very few repressors have been identified comparing to the number of functionally characterized SCW associated activators. One of the well-known secondary cell wall repressors, AtMYB4, has been shown to maintain similar function by repressing overall lignin biosynthesis in maize and switchgrass (Sonbol et al., 2009; Fornalé et al., 2010; Du et al., 2012). Overexpression PvMYB4 demonstrated reduced

biomass recalcitrance (e.g. resistance to accessibility of sugars embedded in plant cell walls) and improve bioethanol production to 2.6 fold (Shen et al., 2012; Shen et al., 2013). In this analysis, the two co-orthologs of AtMYB4, OsMYB4a and OsMYB4b, are likely to maintain similar function by connecting with xylan and lignin biosynthesis genes based on RCR v2. We also observed two additional cell wall repressors, namely, OsSND2 and *LOC_Os04g08060*, a C2H2 family protein. Besides repressing a SCW associated CESA, OsCESA9, OsSND2 can also repress a grass cell wall specific gene, OsAT5, which is coding the enzyme responsible to synthesize feruloylation of monolignols. This information provides insight to characterize appropriate SCW repressors, which can particularly decrease biomass recalcitrance without dramatically affect plant growth to further promote biofuel production.

Methods

Generation of the rice combined ranked network

We constructed a comprehensive high-quality rice genome scale network based on three publically available networks, namely, ROAD, PlaNet and RiceNets. The goal of combining three networks is to expand the high quality network by covering more rice genes, which allows us to study grass-specific pathways. The three original rice networks, ROAD, PlaNet and RiceNets have three different score systems, Pearson Correlation Coefficiency (PCC), Highest Reciprocal Rank (HRR) and Log Likelihood Score (LLS). For ROAD, we only included positive correlations with the score from 0.5 to 1 from ROAD (Cao et al., 2012). PlaNet is a collection of different plant-species networks and we only included the rice dataset into our study. PlatNet was built based

on HRR and the score range is from 0 to 200 with the increase of 1 (Mutwil et al., 2010; Mutwil et al., 2011). RiceNet v1 and v2 used log likelihood score (LLS) to incorporate diverse proteomics, genomics and comparative genomics datasets likely related to rice biological process with score ranging from 1 to 5 (Lee et al., 2010; Lee et al., 2011; Lee et al., 2015). To combine the three rice networks, we scale different score systems using inversed mutual rank (IMR) following the equation: IMR=1/sqrt(rank (A, B) × rank(B, A)). To apply the RiceNet high quality score system to represent interactions between additional genes, we computed the coefficient score using generalized linear (GLM) model in R based on 1282 and 3389 common edges among ROAD, PlaNet and RiceNet v1 and v2, respectively. Then we followed equation I and II to generate the rice combined network v1 and v2.

$$\text{RCR v1} = \frac{1}{\text{RiceNet.v1\_MR}} + \frac{0.48}{\text{ROAD\_MR}} + \frac{0.04}{\text{PlaNet\_MR}} \quad \text{(Equation I)}$$

$$\text{RCR v2} = \frac{1}{\text{RiceNet v2\_MR}} + \frac{0.33}{\text{ROAD\_MR}} + \frac{0.025}{\text{PlaNet\_MR}} \quad \text{(Equation II)}$$

Receiver operating characteristics curve and area under the curve (ROC-AUC)

We evaluated the network quality based on Gene Ontology (GO) terms annotated by the Biofuel Feedstock Genomics Resources (BFGR). In all, 40% of rice genes have been assigned the GO-BP terms. As used in assessing RiceNet, we excluded 10 general GO-BP terms to avoid bias towards these common terms. We defined true positives (TP) as the number of edges with matched GO-BP terms with scores higher than a certain cutoff. True negatives (TN) are defined as the number of edges with unmatched GO-BP terms with scores lower than the cutoff. False positives (FP) are the number of edges unmatched GO-BP terms with scores higher than the cutoff. False negatives (FN) are

defined as the number of edges with matched GO-BP terms with scores lower than the cutoff. For each network included in the analysis, we applied 40 different cutoffs to generate the curves in each tested network.

In the precision-recall analysis, we sorted the RCR networks, ROAD, PlaNet and RiceNet v2 from high confident score to low, and only selected the same number of edges with RiceNet v1 to avoid the effects of network size on the evaluation. Then we define the total number of edges included with matched GO-BP terms within each whole network as the Total True. True positives are defined as the number of edges with matched GO-BP terms with scores higher than certain cutoff. The number (N) of predictions is defined as the number of edges within each network with particular cutoff. As a control for this analysis, we built a random network by randomly assigning edges between a pair of genes within the rice genome. Precision= TP/N predictions. Recall=TP/Total True.

Extracted interactions between cell wall related genes

We collected 125 rice cell wall related genes based on the following criteria: (1) 20 "Mitchell Clade" BAHD acyltransferase; (2) seven hemicellulose biosynthesis genes; (3) phenylpraponoid pathway genes; (4) homologs of cell wall related transcription factors in rice; (5) Functionally characterized rice cell wall transcription factors. These genes are summered in Supporting Table 3-1. We used Inparanoid to identify orthologs of Arabidopsis cell wall related genes in rice (http://inparanoid.sbc.su.se/cgi-bin/index.cgi). We identified 1177 interactions between 122 bait genes within RCR v2

without cutoff. Three baits genes were not directly connected with others. The network is displayed by Cytoscape version 3.2.1 (Shannon et al., 2003).

We extracted interactions between known Arabidopsis cell wall associated transcription factors and biosynthesis genes in ATTED II using the default cutoff (Obayashi et al., 2014).

Transcription factors expression pattern

The gene expression data for Arabidopsis cell wall associated transcription factors were extracted from Arabidopsis gene expression atlas, which systematically examined the transcripts abundance from different tissues during development (Schmid et al., 2005)

Rice transcription factors gene expression data were extracted from rice gene expression atlas, which analyzed transcript abundance of expressed genes using Affymetrix GeneChip with NCBI accession number GSE19024 (Wang et al., 2010)

The gene expression heatmaps were made by heatmap.2 function in R with the default hierarchical cluster for row dendrograms.

Construct rice cell wall network

To identify putative novel transcription factors controlling cell wall biosynthesis in rice, we constructed 1-step network 125 seed genes with the sum of inversed mutual rank score ≥ 0.03. This network includes 1790 nodes and 215of them are transcription factors. To better select candidates controlling rice cell wall biosynthesis, we excluded members maintaining less than five edges with cell wall seed genes. In all, we predicted

96 transcription factors from 19 protein families are putative novel regulators involved in cell wall biosynthesis and summarized in Supporting Table 3-5.

Characterization of OsMYB61a knockout mutants

We selected the available activation tagging mutant line 2D10906 for OsMYB61a with Dongjin background from Rice GE, the rice mutant flanking sequence database (An et al., 2005; Jeong et al., 2006). Natural day lengths less than 14 were supplemented with artificial lighting. Fertilizer was applied three times a week. We genotyped segragants from the initial imported line by harvesting ~20mg leaf samples from 30 progeny. The leaf samples were frozen with liquid nitrogen and ground with SPEX Genogrander using 1450rpm for 1 minute. DNA was collected using 200 µl extraction buffer containing 100 mM Tris-HCl (pH 9.5), 1 M KCl, and 10 mM EDTA (pH 8.0), incubated at 65°C for 30 min, diluted with 1 mL of water, and centrifuged for 10 min at the maximum speed. The supernatant was used as the template for PCR. Genotyping primers used are listed into Supporting Table 3-2. We measured gene expression using the 5th leaf (bottom to top) harvested from the two-month old plants. We attempted to choose morphologically and developmentally matched leaves for analysis, based on plant size, leaf length and expansion. We frozen half of the leaf for gene expression analysis and dry and ground the other half for cell wall assays.

For gene expression analysis, RNA is extracted with Zymo Quick RNA Extraction Kit. We used 1ug RNA to synthesize cDNA with Promega MMLV reverse transcriptase kit. We ran quantitative PCR with BioRad SYBR Green Master Mix and BioRad CFX96 thermocycler. The qPCR primers for all analyzed genes in this study are

summarized in Supporting Table 3-3. To analyze gene expression data, we firstly calculated the real-time primer efficiency with LinRegPCR (Ruijter et al., 2009). Gene expression data were normalized to two reference genes, *Cc55* and *Ubi5*, which show stable expression level during rice development. We repeated the measurements in the following generation and observed similar results.

Cell wall assays of myb61a

We harvested leaf and stem samples from three-month old wild type and *myb61a* plants, which are relatively developmentally matched. Five biological replicates were used for all the following cell wall assays. Alcohol insoluble residue (AIR) is used to examine the cell wall composition of negative segragant and *myb61a* plants. To prepare AIR, 2 mg ground tissue was treated with 95% ethanol (1:4, w/v) at 100°C for 30 min. After the treatment, the supernatant was removed by centrifugation (10,000g, 10 min), and the residue was subsequently washed three to five times with 70% ethanol and dried at approximately 35°C under vacuum. The dried powder obtained after 70% ethanol wash is designated as AIR.

Lignin content was measured by acetylbromide solubility followed quantification on the 96-wall plate using leaf AIR samples (Bartley et al., 2013). We used five biological replicates and three technical replications for this experiment.

We measured cellulose content using anthrone assay. Before the experiment, leaf AIR samples are destarched with amylase, amyloglucosidase and pullulanase as described by ØBro et al. (2004) and Bartley 2013. 2mg destarched AIR were used for anthrone assay.

Mixed-linkage glucan (MLG) is measured by an enzyme-based kit (Megazyme, K-BGLU) with 5 mg of stem AIR samples. We used five biological replicates for this experiment.

Cell wall associated hydroxinamic acids (e.g. FA and pCA) were examined in myb61a mutants and negative segragant plants. To release HCA from the cell walls, we treat the samples with 2 N NaOH for 24 h at 25°C and examined with HPLC as described in Bartley et al 2013. 2mg leaf AIR samples were used in this experiment.

Transient gene expression assay in rice protoplast

We extracted RNA from the leaf of one-month old rice wild type plant. We used 1ug RNA to synthesize cDNA with SuperScript III reverse transcriptase kit (Invitrogen). We cloned the coding sequence of examined transcription to pENTRY-D TOPO vector and all cloning primers are summarized in Supporting Table 3-2. The transcription factors were cloned to the destination vector, p2GW7, for overexpression. We extracted plasmid using Qiagen Plasmid Midi Prep kit.

Two-week old rice seedlings grown in the dark were used to make protoplast following the method of Bart 2011. Basically, we chopped and incubate the rice plant material in the fresh enzyme solutions for six hours. We used 5×10^5 cells and 8 μg plasmid for each transformation. The same amount of empty vectors was transformed in to control samples. After 20 hours incubate, RNA is extracted with Zymo Quick RNA Extraction Kit. We used 1ug RNA to synthesize cDNA with Promega MMLV reverse transcriptase kit. Gene expression data were normalized to two reference genes, *Cc55* and *Ubi5*. The qPCR primers for all analyzed genes in this study are summarized in

Supporting Table 3-3. All the transient gene expression analysis was repeated in another independent run.

Explore the putative function of unknown grass cell wall expanded genes

To understand the function of uncharacterized acyltranferase and predict the putative pathways they may be involved in, we build one-step sub-network for each AT without cutoff and 19 out of 20 are covered in RCR network v2. Then we analyzed the enriched biological process GO terms for each AT-subnetwork. In total, we identified 279 enriched GO terms (hypergeometric p value < 0.05); however, 240 of them (86%) only associated with four or less AT sub-networks. To better present the potential functional association of all acyltranferase, we only include the enriched GO terms shared by four or more AT sub-networks.

Reference

Ambavaram, M.M.R., Krishnan, A., Trijatmiko, K.R., and Pereira, A. (2011). Coordinated Activation of Cellulose and Repression of Lignin Biosynthesis Pathways in Rice. Plant Physiology 155, 916-931.

Appenzeller, L., Doblin, M., Barreiro, R., Wang, H., Niu, X., Kollipara, K., Carrigan, L., Tomes, D., Chapman, M., and Dhugga, K.S. (2004). Cellulose synthesis in maize: isolation and expression analysis of the cellulose synthase (CesA) gene family. Cellulose 11.

Barabasi, A.-L., and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. Nature reviews 5, 101-113.

Bartley, L.E., Peck, M.L., Kim, S.R., Ebert, B., Manisseri, C., Chiniquy, D.M., Sykes, R., Gao, L., Rautengarten, C., Vega-Sanchez, M.E., Benke, P.I., Canlas, P.E., Cao, P., Brewer, S., Lin, F., Smith, W.L., Zhang, X., Keasling, J.D., Jentoff, R.E., Foster, S.B., Zhou, J., Ziebell, A., An, G., Scheller, H.V., and Ronald, P.C. (2013). Overexpression of a BAHD acyltransferase, OsAt10, alters rice cell wall hydroxycinnamic acid content and saccharification. Plant Physiol 161, 1615-1633.

Bonawitz, N.D., and Chapple, C. (2010). The Genetics of Lignin Biosynthesis: Connecting Genotype to Phenotype. Annual Review of Genetics 44, 337-363.

Bontpart, T., Cheynier, V., Ageorges, A., and Terrier, N. (2015). BAHD or SCPL acyltransferase? What a dilemma for acylation in the world of plant phenolic compounds. New Phytologist, n/a-n/a.

Burton, R., and Fincher, G. (2012). Current challenges in cell wall biology in the cereals and grasses. Frontiers in Plant Science 3.

Burton, R.A., Wilson, S.M., Hrmova, M., Harvey, A.J., Shirley, N.J., Medhurst, A., Stone, B.A., Newbigin, E.J., Bacic, A., and Fincher, G.B. (2006a). Cellulose Synthase-Like CslF Genes Mediate the Synthesis of Cell Wall (1,3;1,4)-ß-d-Glucans. Science 311, 1940-1942.

Burton, R.A., Wilson, S.M., Hrmova, M., Harvey, A.J., Shirley, N.J., Medhurst, A., Stone, B.A., Newbigin, E.J., Bacic, A., and Fincher, G.B. (2006b). Cellulose synthase-like CslF genes mediate the synthesis of cell wall (1, 3; 1, 4)-ß-D-glucans. Science 311, 1940-1942.

Cao, P., Jung, K.-H., Choi, D., Hwang, D., Zhu, J., and Ronald, P. (2012). The Rice Oligonucleotide Array Database: an atlas of rice gene expression. Rice 5, 1-9.

Carpita, N.C. (2012). Progress in the biological synthesis of the plant cell wall: new ideas for improving biomass for bioenergy. Current Opinion in Biotechnology 23, 330-337.

Chan, R.L., Gago, G.M., Palena, C.M., and Gonzalez, D.H. (1998). Homeoboxes in plant development. Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression 1442, 1-19.

Chaw, S.-M., Chang, C.-C., Chen, H.-L., and Li, W.-H. (2004). Dating the Monocot–Dicot Divergence and the Origin of Core Eudicots Using Whole Chloroplast Genomes. J Mol Evol 58, 424-441.

Ciftci-Yilmaz, S., and Mittler, R. (2008). The zinc finger network of plants. Cellular and Molecular Life Sciences 65, 1150-1160.

D'Auria, J.C. (2006). Acyltransferases in plants: a good time to be BAHD. Current Opinion in Plant Biology 9, 331-340.

Davidson, R.M., Gowda, M., Moghe, G., Lin, H., Vaillancourt, B., Shiu, S.H., Jiang, N., and Robin Buell, C. (2012). Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. Plant J 71, 492-502.

Du, H., Feng, B.-R., Yang, S.-S., Huang, Y.-B., and Tang, Y.-X. (2012). The R2R3-MYB Transcription Factor Gene Family in Maize. PloS one 7, e37463.

Fornalé, S., Shi, X., Chai, C., Encina, A., Irar, S., Capellades, M., Fuguet, E., Torres, J.L., Rovira, P., and Puigdomènech, P. (2010). ZmMYB31 directly represses maize lignin genes and redirects the phenylpropanoid metabolic flux. Plant J 64, 633-644.

Gu, H., Zhu, P., Jiao, Y., Meng, Y., and Chen, M. (2011). PRIN: a predicted rice interactome network. BMC Bioinformatics 12, 161.

Handakumbura, P.P., and Hazen, S.P. (2012). Transcriptional Regulation of Grass Secondary Cell Wall Biosynthesis: Playing Catch-Up with Arabidopsis thaliana. Front Plant Sci 3, 74.

Handakumbura, P.P., Matos, D.A., Osmont, K.S., Harrington, M.J., Heo, K., Kafle, K., Kim, S.H., Baskin, T.I., and Hazen, S.P. (2013). Perturbation of Brachypodium distachyon CELLULOSE SYNTHASE A4 or 7 results in abnormal cell walls. BMC Plant Biology 13, 1-16.

Hirano, K., Aya, K., Kondo, M., Okuno, A., Morinaka, Y., and Matsuoka, M. (2012). OsCAD2 is the major CAD gene responsible for monolignol biosynthesis in rice culm. Plant Cell Rep 31, 91-101.

Hirano, K., Kondo, M., Aya, K., Miyao, A., Sato, Y., Antonio, B.A., Namiki, N., Nagamura, Y., and Matsuoka, M. (2013). Identification of Transcription Factors Involved in Rice Secondary Cell Wall Formation. Plant Cell Physiol.

Hollister, J.D. (2015). Polyploidy: adaptation to the genomic environment. New Phytol 205, 1034-1039.

Huang, D., Wang, S., Zhang, B., Shang-Guan, K., Shi, Y., Zhang, D., Liu, X., Wu, K., Xu, Z., Fu, X., and Zhou, Y. (2015). A Gibberellin-Mediated DELLA-NAC Signaling Cascade Regulates Cellulose Synthesis in Rice. The Plant Cell 27, 1681-1696.

Hussey, S.G., Mizrachi, E., Spokevicius, A.V., Bossinger, G., Berger, D.K., and Myburg, A.A. (2011). SND2, a NAC transcription factor gene, regulates genes involved in secondary cell wall development in Arabidopsis fibres and increases fibre cell area in Eucalyptus. BMC Plant Biology 11, 1-17.

Jain, M., Tyagi, A.K., and Khurana, J.P. (2008). Genome-wide identification, classification, evolutionary expansion and expression analyses of homeobox genes in rice. FEBS Journal 275, 2845-2861.

Jung, K.-H., An, G., and Ronald, P.C. (2008). Towards a better bowl of rice: assigning function to tens of thousands of rice genes. Nature reviews 9, 91-101.

Karlen, S.D., Peck, M.L., Zhang, C., Smith, R.A., Padmakshan, D., Helmich, K.E., Free, H.C.A., Lee, S., Smith, B.G., Lu, F., Sedbrook, J.C., Sibout, R., Grabber, J.H., Runge, T.M., Mysore, K.S., Harris, P.J., Bartley, L.E., and Ralph, J. . (Submitted). Monolignol Ferulate Conjugates are Naturally Incorporated into Plant Lignins. Science Advances.

Kellogg, E.A. (2001). Evolutionary History of the Grasses. Plant Physiology 125, 1198-1205.

Lee, I., Ambaru, B., Thakkar, P., Marcotte, E.M., and Rhee, S.Y. (2010). Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana. Nat Biotech 28, 149-156.

Lee, I., Seo, Y.-S., Coltrane, D., Hwang, S., Oh, T., Marcotte, E.M., and Ronald, P.C. (2011). Genetic dissection of the biotic stress response using a genome-scale gene network for rice. Proceedings of the National Academy of Sciences 108, 18548-18553.

Lee, T., Oh, T., Yang, S., Shin, J., Hwang, S., Kim, C.Y., Kim, H., Shim, H., Shim, J.E., Ronald, P.C., and Lee, I. (2015). RiceNet v2: an improved network prioritization server for rice genes. Nucleic Acids Research 43, W122-W127.

Li, X., Duan, X., Jiang, H., Sun, Y., Tang, Y., Yuan, Z., Guo, J., Liang, W., Chen, L., Yin, J., Ma, H., Wang, J., and Zhang, D. (2006). Genome-Wide Analysis of Basic/Helix-Loop-Helix Transcription Factor Family in Rice and Arabidopsis. Plant Physiology 141, 1167-1184.

Mao, L., Hemert, J.L., Dash, S., and Dickerson, J.A. (2009). Arabidopsis gene co-expression network and its functional modules. BMC Bioinformatics. 10.

Mauseth, J.D. (1988). Plant anatomy. (Benjamin/Cummings Publ. Co.: Menlo Park, Calif).

Mitsuda, N., Iwase, A., Yamamoto, H., Yoshida, M., Seki, M., Shinozaki, K., and Ohme-Takagi, M. (2007). NAC Transcription Factors, NST1 and NST3, Are Key Regulators of the Formation of Secondary Walls in Woody Tissues of Arabidopsis. The Plant Cell 19, 270-280.

Mizoi, J., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2012). AP2/ERF family transcription factors in plant abiotic stress responses. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms 1819, 86-96.

Movahedi, S., Van de Peer, Y., and Vandepoele, K. (2011). Comparative Network Analysis Reveals That Tissue Specificity and Gene Function Are Important Factors Influencing the Mode of Expression Evolution in Arabidopsis and Rice. Plant Physiology 156, 1316-1330.

Mutwil, M., Debolt, S., and Persson, S. (2008). Cellulose synthesis: a complex complex. Current Opinion in Plant Biology 11, 252-257.

Mutwil, M., Usadel, B., Schütte, M., Loraine, A., Ebenhöh, O., and Persson, S. (2010). Assembly of an interactive correlation network for the Arabidopsis genome using a novel heuristic clustering algorithm. Plant Physiol. 152.

Mutwil, M., Klie, S., Tohge, T., Giorgi, F.M., Wilkins, O., Campbell, M.M., Fernie, A.R., Usadel, B., Nikoloski, Z., and Persson, S. (2011). PlaNet: Combined Sequence and Expression Comparisons across Plant Networks Derived from Seven Species. The Plant Cell Online 23, 895-910.

Niklas, K.J. (2004). The Cell Walls that Bind the Tree of Life. BioScience 54, 831-841.

Nishiyama, T., Fujita, T., Shin-I, T., Seki, M., Nishide, H., Uchiyama, I., Kamiya, A., Carninci, P., Hayashizaki, Y., and Shinozaki, K. (2003). Comparative genomics of Physcomitrella patens gametophytic transcriptome and Arabidopsis thaliana: Implication for land plant evolution. Proc Natl Acad Sci USA 100.

Obayashi, T., and Kinoshita, K. (2009). Rank of Correlation Coefficient as a Comparable Measure for Biological Significance of Gene Coexpression. DNA Research 16, 249-260.

Obayashi, T., Okamura, Y., Ito, S., Tadaka, S., Aoki, Y., Shirota, M., and Kinoshita, K. (2014). ATTED-II in 2014: Evaluation of Gene Coexpression in Agriculturally Important Plants. Plant and Cell Physiology 55, e6.

Obertello, M., Shrivastava, S., Katari, M.S., and Coruzzi, G.M. (2015). Cross-Species Network Analysis Uncovers Conserved Nitrogen-Regulated Network Modules in Rice. Plant Physiology 168, 1830-1843.

Ogata, Y., Suzuki, H., Sakurai, N., and Shibata, D. (2010). CoP: a database for characterizing co-expressed gene modules with biological information in plants. Bioinformatics 26, 1267-1268.

Ogo, Y. (2007). The rice bHLH protein OsIRO2 is an essential regulator of the genes involved in Fe uptake under Fe-deficient conditions. Plant J 51.

Paterson, A.H., Bowers, J.E., and Chapman, B.A. (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proceedings of the National Academy of Sciences of the United States of America 101, 9903-9908.

Paterson, A.H., Chapman, B.A., Kissinger, J.C., Bowers, J.E., Feltus, F.A., and Estill, J.C. (2006). Many gene and domain families have convergent fates following

independent whole-genome duplication events in Arabidopsis, Oryza, Saccharomyces and Tetraodon. Trends in Genetics 22, 597-602.

Pauly, M., Gille, S., Liu, L., Mansoori, N., Souza, A., Schultink, A., and Xiong, G. (2013). Hemicellulose biosynthesis. Planta 238, 627-642.

Perlack, R.D., Wright, L.L., Turhollow, A.F., Graham, R.L., Stokes, B.J., and Erbach, D.C. (2005). Biomass as feedstock for a bioenergy and bioproducts industry: the technical feasibility of a billion-ton annual supply (DTIC Document).

Petrik, D.L., Karlen, S.D., Cass, C.L., Padmakshan, D., Lu, F., Liu, S., Le Bris, P., Antelme, S., Santoro, N., Wilkerson, C.G., Sibout, R., Lapierre, C., Ralph, J., and Sedbrook, J.C. (2014). p-Coumaroyl-CoA:monolignol transferase (PMT) acts specifically in the lignin biosynthetic pathway in Brachypodium distachyon. The Plant Journal 77, 713-726.

Pires, N., and Dolan, L. (2010). Origin and Diversification of Basic-Helix-Loop-Helix Proteins in Plants. Molecular Biology and Evolution 27, 862-874.

Popper, Z.A., Michel, G., Hervé, C., Domozych, D.S., Willats, W.G.T., Tuohy, M.G., Kloareg, B., and Stengel, D.B. (2011). Evolution and Diversity of Plant Cell Walls: From Algae to Flowering Plants. Annual Review of Plant Biology 62, 567-590.

Rhee, S.Y., and Mutwil, M. (2014). Towards revealing the functions of all genes in plants. Trends Plant Sci 19, 212-221.

Sarkar, N., Kim, Y.-K., and Grover, A. (2014). Coexpression network analysis associated with call of rice seedlings for encountering heat stress. Plant Molecular Biology 84, 125-143.

Sato, Y., Namiki, N., Takehisa, H., Kamatsuki, K., Minami, H., Ikawa, H., Ohyanagi, H., Sugimoto, K., Itoh, J.-I., Antonio, B.A., and Nagamura, Y. (2012). RiceFREND: a platform for retrieving coexpressed gene networks in rice. Nucleic Acids Research.

Scheller, H.V., and Ulvskov, P. (2010). Hemicelluloses. Annual Review of Plant Biology 61, 263-289.

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Scholkopf, B., Weigel, D., and Lohmann, J.U. (2005). A gene expression map of Arabidopsis thaliana development. Nat Genet 37, 501-506.

Schwerdt, J.G., MacKenzie, K., Wright, F., Oehme, D., Wagner, J.M., Harvey, A.J., Shirley, N.J., Burton, R.A., Schreiber, M., Halpin, C., Zimmer, J., Marshall, D.F., Waugh, R., and Fincher, G.B. (2015). Evolutionary Dynamics of the Cellulose Synthase Gene Superfamily in Grasses. Plant Physiology 168, 968-983.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research 13, 2498-2504.

Shen, H., He, X., Poovaiah, C.R., Wuddineh, W.A., Ma, J., Mann, D.G., Wang, H., Jackson, L., Tang, Y., and Neal Stewart Jr, C. (2012). Functional characterization of the switchgrass (Panicum virgatum) R2R3‐MYB transcription factor PvMYB4 for improvement of lignocellulosic feedstocks. New Phytologist.

Shen, H., Poovaiah, C., Ziebell, A., Tschaplinski, T., Pattathil, S., Gjersing, E., Engle, N., Katahira, R., Pu, Y., Sykes, R., Chen, F., Ragauskas, A., Mielenz, J., Hahn, M., Davis, M., Stewart, C.N., and Dixon, R. (2013). Enhanced characteristics of genetically modified switchgrass (Panicum virgatum L.) for high biofuel production. Biotechnology for Biofuels 6, 71.

Sonbol, F.-M., Fornalé, S., Capellades, M., Encina, A., Tourino, S., Torres, J.-L., Rovira, P., Ruel, K., Puigdomenech, P., and Rigau, J. (2009). The maize ZmMYB42 represses the phenylpropanoid pathway and affects the cell wall structure, composition and degradability in Arabidopsis thaliana. Plant Mol. Biol. 70, 283-296.

Sørensen, I., Domozych, D., and Willats, W.G.T. (2010). How Have Plant Cell Walls Evolved? Plant Physiology 153, 366-372.

Stracke, R., Werber, M., and Weisshaar, B. (2001). The R2R3-MYB gene family in Arabidopsis thaliana. Curr Opin Plant Biol 4, 447-456.

Taylor-Teeples, M., Lin, L., de Lucas, M., Turco, G., Toal, T.W., Gaudinier, A., Young, N.F., Trabucco, G.M., Veling, M.T., Lamothe, R., Handakumbura, P.P., Xiong, G., Wang, C., Corwin, J., Tsoukalas, A., Zhang, L., Ware, D., Pauly, M., Kliebenstein, D.J., Dehesh, K., Tagkopoulos, I., Breton, G., Pruneda-Paz, J.L., Ahnert, S.E., Kay, S.A., Hazen, S.P., and Brady, S.M. (2014). An Arabidopsis gene regulatory network for secondary cell wall synthesis. Nature advance online publication.

Taylor-Teeples, M., Lin, L., de Lucas, M., Turco, G., Toal, T.W., Gaudinier, A., Young, N.F., Trabucco, G.M., Veling, M.T., Lamothe, R., Handakumbura, P.P., Xiong, G., Wang, C., Corwin, J., Tsoukalas, A., Zhang, L., Ware, D., Pauly, M., Kliebenstein, D.J., Dehesh, K., Tagkopoulos, I., Breton, G., Pruneda-Paz, J.L., Ahnert, S.E., Kay, S.A., Hazen, S.P., and Brady, S.M. (2015). An Arabidopsis gene regulatory network for secondary cell wall synthesis. Nature 517, 571-575.

Vanholme, R., Morreel, K., Ralph, J., and Boerjan, W. (2008). Lignin engineering. Current Opinion in Plant Biology 11, 278-285.

Vogel, J. (2008a). Unique aspects of the grass cell wall. Current Opinion in Plant Biology 11, 301-307.

Vogel, J. (2008b). Unique aspects of the grass cell wall. Curr. Opin. Plant Biol. 11, 301-307.

Voxeur, A., Wang, Y., and Sibout, R. (2015). Lignification: different mechanisms for a versatile polymer. Current Opinion in Plant Biology 23, 83-90.

Wang, H., Zhao, Q., Chen, F., Wang, M., and Dixon, R.A. (2011). NAC domain function and transcriptional control of a secondary cell wall master switch. The Plant Journal 68, 1104-1114.

Wang, L., Xie, W., Chen, Y., Tang, W., Yang, J., Ye, R., Liu, L., Lin, Y., Xu, C., Xiao, J., and Zhang, Q. (2010). A dynamic gene expression atlas covering the entire life cycle of rice. Plant J 61, 752-766.

Wang, X., Gowik, U., Tang, H., Bowers, J.E., Westhoff, P., and Paterson, A.H. (2009). Comparative genomic analysis of C4 photosynthetic pathway evolution in grasses. Genome Biology 10, 1-18.
Xu, W., Dubos, C., and Lepiniec, L. (2015). Transcriptional control of flavonoid biosynthesis by MYB–bHLH–WDR complexes. Trends in plant science 20, 176-185.

Yeung, K.Y., Dombek, K.M., Lo, K., Mittler, J.E., Zhu, J., and Schadt, E.E. (2011). Construction of regulatory networks using expression time-series data of a genotyped population. Proceedings of the National Academy of Sciences of the United States of America 108.

Yim, W., Yu, Y., Song, K., Jang, C., and Lee, B.-M. (2013). PLANEX: the plant co-expression database. BMC Plant Biology 13, 83.

Zhang, J. (2003). Evolution by gene duplication: an update. Trends in Ecology & Evolution 18, 292-298.

Zhao, K., and Bartley, L. (2014). Comparative genomic analysis of the R2R3 MYB secondary cell wall regulators of Arabidopsis, poplar, rice, maize, and switchgrass. BMC Plant Biology 14, 135.

Zhong, R., and Ye, Z.-H. (2001). Secondary Cell Walls. In eLS (John Wiley & Sons, Ltd.

Zhong, R., and Ye, Z.-H. (2007). Regulation of cell wall biosynthesis. Current opinion in plant biology 10, 564-572.

Zhong, R., Lee, C., and Ye, Z.H. (2010). Functional characterization of poplar wood-associated NAC domain transcription factors. Plant Physiol 152, 1044-1055.

Zhong, R., Lee, C., Zhou, J., McCarthy, R.L., and Ye, Z.H. (2008). A battery of transcription factors involved in the regulation of secondary cell wall biosynthesis in Arabidopsis. Plant Cell 20, 2763-2782.

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

Zhou, J., Zhong, R., and Ye, Z.-H. (2014). Arabidopsis NAC Domain Proteins, VND1 to VND5, Are Transcriptional Regulators of Secondary Wall Biosynthesis in Vessels. PLoS ONE 9, e105726.

Zhou, J., Lee, C., Zhong, R., and Ye, Z.-H. (2009). MYB58 and MYB63 are transcriptional activators of the lignin biosynthetic pathway during secondary cell wall formation in Arabidopsis. The Plant Cell 21, 248-266.

Chapter 4 : Identification of putative grass cell wall-associated *cis*-elements using comparative *de novo* promoter analysis

Authors: Kangmei Zhao and Laura Bartley

This chapter is planned for publication upon experimental testing of some of the specific predictions of this work.

Authors Contribution: KZ and LEB conceived of and designed the study. KZ carried out the analyses, created the figures, and wrote the first draft of the text.

Abstract

Grass cell walls are environmentally and economically important; however, relatively limited information is available on grass cell wall associated regulators and their corresponding DNA binding sites. Many known Arabidopsis cell wall transcription factors from the R2R3 MYB and NAC families may maintain similar functions in grasses to regulate *CESA* and lignin biosynthesis genes. However, there has been no systematic examination of *cis*-elements present within promoters of grass CESA and lignin biosynthesis encoding genes in grasses. In addition, little information is available on the regulation of grass cell wall-specific genes, such as the Cellulose Synthase Like (*Csl) F* and *CslH*, and cell wall precursor modifying BAHD-acyltransferases, and their recruitment into cell wall regulatory pathways during evolution. In this study, we took advantage of two *de novo* motif algorithms, RAST and MEME, to identify potential DNA binding sites present within the promoters of three gene sets *CESA* and lignin biosynthesis genes, grass-diverged *Csl* gene and grass-expanded acyltransferase genes. We observed that known cell wall-associated *cis*-elements recognized by R2R3 MYBs and NAC proteins are significantly enriched within the promoters of all three gene-classes. This suggests that known dicot cell wall-associated *cis*-elements are conserved in grasses and that these elements have evolved (or been maintained) within the promoters of grass-specific cell wall genes. In addition, *cis*-elements potentially associated with AP2/ERF, C2H2, C2C2 and homeodomain proteins are also significantly enriched in grass cell wall biosynthesis genes. In all, this analysis provides guidance toward functional characterization of cell wall-associated regulatory elements in grasses, knowledge of which will promote food, fiber and biofuel production.

Introduction

Developmental and environmental responses are precisely controlled by transcription factors and their dynamic interactions with corresponding DNA binding sites, i.e., *cis*-elements (Wittkopp and Kalay, 2012; Voss and Hager, 2014; Rodriguez-Granados et al., 2016). Currently, genome-scale *cis*-element maps are available for large numbers of transcription factors from humans, *Drosophila melanogaster*, *Caenorhabditis elegans* and mouse based on chromatin immunoprecipitation-sequencing (ChIP-Seq) (Van Nostrand and Kim, 2013; Kheradpour and Kellis, 2014). However, information of *cis*-elements in plant genomes is more modest, with genome-scale ChIP data sets generally being limited to core *cis*-elements associated with flowing, hormone signaling and stress responses, such as the CAT-box, ABA-response elements (ABREs), G-box, and W-box, etc. (Priest et al., 2009; Zou et al., 2011; Walcher and Nemhauser, 2012; Boer et al., 2014; Liu et al., 2014). Recently, Franco-Zorrilla et al. (2014) used a protein-binding microarray to identify *cis*-elements for 65 transcription factors of Arabidopsis, representing 25 protein families,. However, for most expressed genes in plant genomes, the *cis*-elements present within promoters and the corresponding transcription factors remain to be revealed.

De novo prediction of *cis*-elements is a computational approach to identify potential binding sites that determine gene expression that has been applied in various organisms (Priest et al., 2009; Chen et al., 2012; Ding et al., 2012; Maruyama et al., 2012). For example, large-scale *de novo* motif discovery promoted identification of potential *cis*-elements associated with biotic and abiotic stress response and their copy number, location and combinations can be used to validate direct interactions between

transcription factors and analyzed promoters. This will not only facilitate the discovery of functional *cis*-elements associated with stress response pathways, but promote the prediction and validation of novel stress response genes (Zou et al., 2011)

Cis-element discovery algorithms can be divided into two classes, word-based and probabilistic sequence models. Word-based (or string-based) methods mostly rely on exhaustive enumeration, i.e., counting and comparing oligonucleotide frequencies. This strategy's strength is that it is fast and effective for constrained, shorter motifs. A popular tool incorporating is this algorithm is the Regulatory Sequence Analysis Tool (RSAT) (http://www.rsat.eu/) (Turatsinze et al., 2008). Probabilistic sequence models use maximum-likelihood estimates of position weight matrices. The strength of these algorithms is that they are relatively sensitive. One of the most popular tools that uses this algorithm is MEME (http://meme-suite.org/) (Bailey et al., 2006).

Here, we report the application of *de novo* motif analysis to discover potential *cis*-elements associated with cell wall biosynthesis genes in grasses. Grasses include the major cereal crops and contribute an estimated 55% of biomass that can be produced in the U.S. (Kellogg, 2001; Somerville, 2007; Bartley and Ronald, 2009; Binod et al., 2010). The bulk of plant dry mass, secondary cell walls include cellulose, hemicellulose and lignin (Keegstra, 2010; Carpita, 2012). Among the discovered Arabidopsis cell wall-associated regulators, core transcription factors from the NAC and R2R3 MYB protein families are relatively well examined. These proteins have been experimentally found to recognize the SNBE- (TACXTTXXXXATGA) and AC-elements (CC(A/T)A(A/C)(T/C)), respectively (Zhong and Ye, 2007; Zhong et al., 2010; Wang et al., 2011; Taylor-Teeples et al., 2015).

Currently, understanding of cell wall regulation largely relies on data for the dicotyledenouse plant Arabidopsis, with only a few studies in grasses (Handakumbura and Hazen, 2012; Taylor-Teeples et al., 2014; Nakano et al., 2015). Zhong et al. (2011) overexpressed rice and maize SCW-related NAC transcription factors in Arabidopsis and found that they are able to activate Arabidopsis *CESA* and lignin biosynthesis genes (Zhong et al., 2011). In switchgrass, PvMYB4, the ortholog of the Arabidopsis cell wall repressor AtMYB4, can bind to AC-elements based on yeast-one-hybrid. Recently, Hunag et al. showed that rice OsMYB61a, a co-ortholog of Arabidopsis SCW regulator, AtMYB61, can directly activate secondary cell wall-associated *CESA*s in rice and is able to respond to gibberellin (GA) signaling (Huang et al., 2015). In all, this suggests that dicot cell wall associated *cis*-elements may also be active in the promoters of grass *CESA* and lignin biosynthesis genes. In addition, grass orthologs of known dicot cell wall-associated transcription factors may maintain similar binding specificities to activate cellulose and lignin biosynthesis genes.

Compared to Arabidopsis, grasses incorporate or synthesize unique cell wall components besides conserved one, such as cellulose and lignin. The two grass cell wall features that we focus on here are the esterification of cell wall polymers with hydroxynamic acids (HCA), particularly ferulic acid and *para*-coumaric acid (*p*CA), and the synthesis of a novel hemicellulose, mixed-linkage glucan (MLG), which likely have different evolutionary origins (Vogel, 2008; Withers et al., 2012; Fincher and Burton, 2014). HCAs can crosslink hemicellulose and lignin and all evidence collected to date suggests that they are introduced into grass cell walls through the action of a subclade of BAHD-acyltransferases (Bartley et al., 2013). In plants, functionally

characterized BAHD-ATs from different species across dicots and monocots and found that they can be divided into five clades based on phylogeny (D'Aurian 2006). Mitchell et al. proposed a sub-clade belonging to the BAHD-AT Clade V, which are expanded in grasses and may be responsible for the incorporation of HCAs into grass cell wall components. Thus, we refer it as "Mitchell Clade". On the other hand, members of grass-diverged genes from the cellulose synthase-like (Csl) families, Clade F and H, synthesize MLG (Burton et al., 2006; Kim et al., 2015). The corresponding regulators of grass cell wall-specific genes remain to be revealed.

Though grass-expanded/-diverged BAHD-ATs and CslF/H clades both participate in cell wall biosynthesis pathways, their evolutionary origins seem to be different, which may have resulted in different mechanisms of recruitment into regulatory networks. BAHD-ATs form a large protein family with versatile catalytic abilities (D'Auria, 2006; Bontpart et al., 2015). Further phylogenetic analysis supports that this sub-clade has expanded in grasses and the most closely related acyltransferase does not show HCA phenotype in Arabidopsis (Rautengarten et al., 2012). In contrast, Csl and cellulose synthase (CESAs) belong to the GT2 family of gylcosyltransferase, which may be originating from cyanobacteria (Somerville, 2006; Popper et al., 2011; Kumar and Turner, 2015). Based on the phylogeny, Csl members have been assigned into different groups, namely, CESA, CslA, to Csl J(Schwerdt et al., 2015). Among them, CESAs have been involved in plant cell wall biosynthesis and also conserved during evolution (Somerville, 2006). Closely related to CESAs in phylogeny, most *Csl* genes are known to synthesize cell wall polysaccharide in plants, thus grass-diverged

Csl F and H are more likely to share similar regulatory machinery (Schwerdt et al., 2015).

We aim to elucidate the regulation of grass cell wall-specific genes by predicting putative DNA binding sites and their cognate transcription factors. Based on the information of *cis*-elements in dicots and grasses, as well as the origins of grass cell wall-specific genes, we focus on distinguishing among the following four non-mutually exclusive scenarios or models for grass-specific gene regulation (Figure 4-1): (1) Grass cell wall specific genes have maintained or evolved AC elements or NAC binding sites, and thus can be directly regulated by orthologs of Arabidopsis known cell wall transcription factors; (2) Grass cell wall specific genes have accumulated additional *cis*-elements within their promoters and orthologs of Arabidopsis known cell wall transcription factors have evolved different binding specificity to recognize them in grasses; (3) Grass cell wall specific genes have maintained or evolved known dicot cell wall *cis*-elements (e.g. AC elements or NAC binding sites); however, they can not be recognized by orthologs of Arabidopsis known transcription factors. Then, grass cell wall-specific genes are regulated separately comparing to CESAs and lignin biosynthesis genes by unknown transcription factors. (4) Grass cell wall-specific genes have evolved novel *cis*-elements due to the accumulation of mutations. Therefore, unknown transcription factors can directly regulate grass cell wall specific genes. We believe that model 2 and 3 are unlikely based on current information gathered for orthologs of Arabidopsis cell wall regulators. First, rice and maize SCW-related NAC transcription factors are able to activate Arabidopsis cell wall biosynthesis genes. In addition, PvMYB4, the ortholog of Arabidopsis cell wall repressor AtMYB4, can bind

to AC-elements based on Y1H. Second, based on molecular genetics and transient gene expression analysis in Chapter 3, OsMYB61a, a co-ortholog of AtMYB61, can activate grass cell wall-specific genes in a direct or indirect manner. Thus, we will focus on the following two questions in this analysis: (1) whether known dicot associated *cis*-elements are present within the promoters of grass cell wall-specific genes that can be recognized by orthologs of known dicot cell wall transcription factors (model 1); (2) whether novel *cis*-elements have been evolved within the promoters of grass cell wall-specific genes and what transcription factors can recognize them (Model 4).

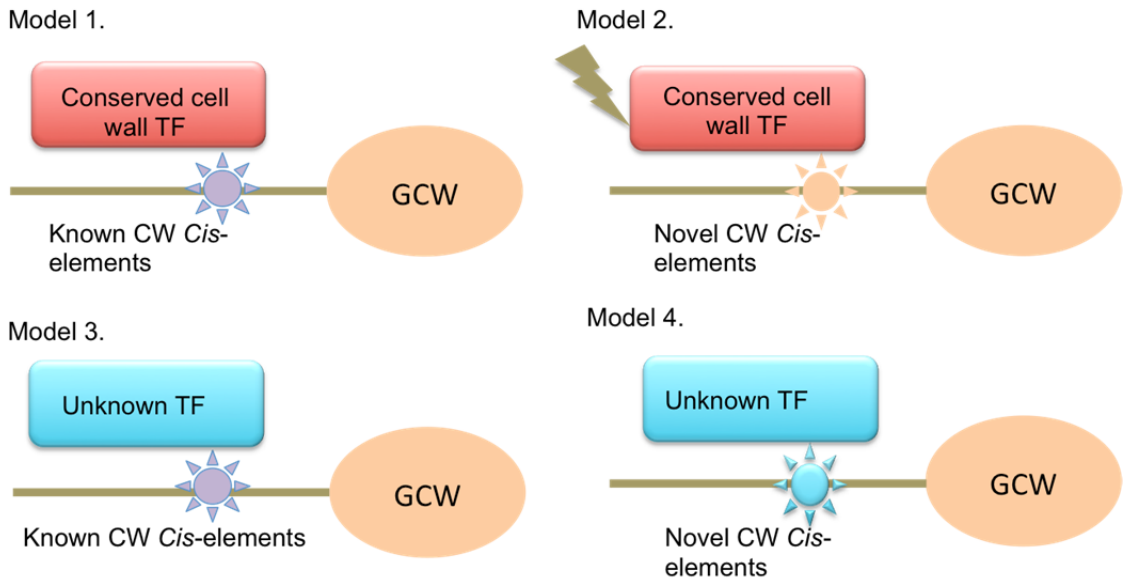


Figure 4-1 Summary of possible models representing the regulation of grass cell wall-specific genes in grasses. GCW represents grass cell wall-specific genes, including *ClsF/H* and "Mitchell Clade" *BAHD-ATs*. Conserved cell wall transcription factors represents known cell wall associated transcription factors in plants. CW is short for cell wall.

In this study, we identified potential DNA binding sites present within the promoters of cell wall biosynthesis genes including the following three groups: (1) both primary- and secondary-wall *CESA*s and lignin biosynthesis gens; (2) all *Csl* genes present in rice; (3) "Mitchell Clade" *BAHD-AT*s. We incorporated comparative *de novo* motif discovery to predict DNA motifs significantly enriched within the selected promoters. By comparing with known Arabidopsis *cis*-elements collections, we also predicted the putative transcription factor families associated with the enriched *cis*-elements. We observed that AC-elements and SNBE are significantly enriched within the promoters of *Csl* and *BAHD-ATs*. The results suggest that known cell wall-related MYB and NAC transcription factors may also be able to regulate Csl and cell wall-related BAHD-ATs, supporting Model 1. However, we also predict that novel regulators from the C2H2, C2C2 and ERF families, may also directly bind to promoters of *Csl* and *BAHD-ATs*, supporting Model 3. This analysis expands knowledge of cell wall-associated DNA binding sites in grasses, providing additional information for screening and functional characterization of novel cell wall-related transcription factors in grasses and promoting our understanding on the evolution of cell wall biosynthesis pathways in grasses.

Methods

De novo motif discovery

We employed comparative *de novo* motif discovery in rice and *Brachypodium distachyon* using MEME and RSAT (Bailey et al., 2006; Turatsinze et al., 2008). Locus IDs and gene names are summarized in Table 4-1. To improve the predictive power, we

included orthologs of rice cell wall-related genes from *Brachypodium* identified with Inparanoid (Östlund et al., 2010). We selected 1kb upstream sequences in this analysis, because previous plant promoter analysis suggests the majority of functional *cis*-elements are relatively local to the TSS (Velde et al., 2014). We downloaded rice 1 kb upstream sequences from Orisis with the MSU genome annotation v7 (Morris et al., 2008). We downloaded *Brachypodium* 1 kb upstream sequences from http://www.brachypodium.org/. The parameters for MEME were motif length: 6 to 10 bp with any number of replication within promoters. The parameters for RSAT were motif length: 6 to 8 bp. We took the union of discovered motifs from MEME and RSAT and motifs with high similarity (PWM Pearson correlation score > 0.8) were combined into one to represent the binding site. For each run of motif discovery, we also ran the same number of 1 kb upstream promoter sequences from randomly selected non-cell wall rice and *Brachypodium* as negative controls. We only report motifs that are discovered in the random promoters.

Table 4-1 Summary of sequences included in this analysis. The orthologs between rice and Brachypodium were identified based on Inparanoid.

Rice Gene Name	Locus ID (LOC_)	*Brachypodium* Gene Name	*Brachypodium* Locus ID
OsCESA1	Os05g08370	BdCESA1	Bradi2g34240
OsCESA2	Os03g59340	BdCESA2	Bradi1g04597
OsCESA8	Os07g10770	BdCESA8	Bradi1g54250
OsCESA4	Os01g54620	BdCESA4	Bradi2g49912
OsCESA7	Os10g32980	BdCESA7	Bradi3g28350
OsCESA9	Os09g25490	BdCESA9	Bradi4g30540
OsCCR1	Os08g34280	BdCCR1	Bradi3g36887
OsCOMT1	Os08g06100	NA	NA
OsCAD1	Os10g11810	BdCAD1	Bradi3g22980
OsCAD2	Os02g09490	NA	NA

OsCAD3	Os10g29470	NA	NA
OsPAL1	Os02g41630	NA	NA
OsPAL2	Os02g41650	BdPAL2	Bradi3g49260
OsF5H	Os10g36848	NA	NA
OsC3H	Os05g41440	BdC3H	Bradi2g21300
OsC4H	Os01g60450	BdC4H	Bradi2g31510
OsC4H	Os05g25640	Bd4CL3	Bradi3g05750

Table 4-1 cont.,

Os4CL1	Os08g14760	Bd4CL1	Bradi3g18960
Os4CL2	Os02g46970	NA	NA
Os4CL3	Os02g08100	NA	NA
Os4CL4	Os06g44620	BdF5H	Bradi3g30590
Os4CL5	Os08g34790	Bd4CL5	Bradi3g37300
OsHCT	Os02g39850	NA	NA
OsHCT	Os04g42250	BdHCT	Bradi3g48530
OsHCT	Os11g31090	NA	NA
Os_CSL_D1	Os10g42750	Bd_CSL_D1	Bradi3g34490
Os_CSL_D2	Os06g02180	Bd_CSL_D2	Bradi1g50170
Os_CSL_D3	Os08g25710	NA	NA
Os_CSL_D4	Os12g36890	Bd_CSL_D4	Bradi4g05027
Os_CSL_D5	Os06g22980	Bd_CSL_D5	Bradi2g03380
Os_CSL_E1	Os09g30120	Bd_CSL_E1	Bradi4g33080
Os_CSL_E2	Os02g49332	Bd_CSL_E2	Bradi3g56440
Os_CSL_E6	Os09g30130	Bd_CSL_E6	Bradi4g33090
Os_CSL_F4	Os07g36740	Bd_CSL_F4	Bradi1g25117
Os_CSL_F6	Os08g06380	Bd_CSL_F6	Bradi3g16307
Os_CSL_F9	Os07g36610	Bd_CSL_F9	Bradi3g45515
Os_CSL_H1	Os10g20090	Bd_CSL_H1	Bradi5g10130
Os_2026_CSL_F	Os10g20260	NA	NA
Os_2930_CSL_F	Os12g29300	NA	NA
Os_3502_CSL_H	Os04g35020	NA	NA
Os_3503_CSL_H	Os04g35030	NA	NA
Os_3663_CSL_F	Os07g36630	Bd_CSL_F	Bradi1g25107
Os_3669_CSL_F	Os07g36690	Bd_CSL_F	Bradi1g25117
Os_3670_CSL_F	Os07g36700	Bd_CSL_F	Bradi1g25117
Os_3675_CSL_F	Os07g36750	Bd_CSL_F	Bradi1g25130
OsAT1	Os01g42880	BdAT1	Bradi2g43520
OsAT10	Os06g39390	BdAT10	Bradi1g36990
OsAT11	Os04g11810	NA	NA
OsAT12	Os04g09590	BdAT12	Bradi3g22830

OsAT13	Os04g09260	NA	NA
OsAT14	Os10g01930	NA	NA
OsAT15	Os10g01920	NA	NA
OsAT16	Os10g02000	NA	NA
OsAT17	Os10g01800	NA	NA
OsAT18	Os10g03360	NA	NA
OsAT19	Os10g03390	NA	NA
OsAT2	Os01g42870	BdAT2	Bradi2g43510
OsAT20	Os06g48560	NA	NA

Table 4-1 cont.,

OsAT3	Os05g04584	BdAT3	Bradi2g36910
OsAT4	Os01g18744	BdAT4	Bradi5g01240
OsAT5	Os05g19910	BdAT5	Bradi4g06067
OsAT6	Os01g08380	NA	NA
OsAT7	Os05g08640	BdAT7	Bradi2g33980
OsAT8	Os06g39470	NA	NA
OsAT9	Os01g09010	BdAT9	Bradi2g05480

Motif similarity analysis

We matched discovered *cis*-elements to Arabidopsis protein binding microarray datasets to infer their potential associated transcription factors using TOMTOM, which is a motif comparison platform incorporated into MEME. The cutoff used to determined motif match is a p-value lower than E-04 (Gupta et al., 2007).

We compared the similarity among discovered motifs by the matrix comparison platform incorporated into RSAT. The heatmap was made using the heat.map2 function in R with hierarchical cluster.

Cell wall gene expression pattern

Rice cell wall gene expression data were extracted from the rice transcriptome atlas, NCBI accession number GSE19024 (Wang et al., 2010). The heatmap was made using the heatmap.2 in R and gene expression values were normalized using z-scores.

Transcription factors-cell wall genes network

We extracted interactions between cell wall genes and transcription factors in RCR v2, which is the currently most comprehensive genome scale network as reported in Chapter 3 of this dissertation. The networks were displayed by Cytoscape v.3.2.1 (Shannon et al., 2003).

Origin of grass cell wall expanded genes

Whole genome duplication (WGD) is based on identified duplication blocks in Plant Genome Duplication Database (http://chibba.agtec.uga.edu/uplication/). Tandem duplication between two or more genes is defined by the genes meeting the following criteria: 1. belonging to the same domain family; 2. located within 100 kb each other; 3. separated by less than 10 non-homologous spacer genes (Zou et al., 2009).

Results

Comparative de novo motif discovery

In this analysis, we incorporated comparative *de novo* motif discovery to distinguish among possible models for the evolution of grass cell wall biosynthesis promoters and to predict potential *cis*-elements present within the promoters of cell wall biosynthesis genes. Figure 4-2 outlines the procedure followed. To improve the prediction power and identify conserved functional DNA binding sites, we increased the number of input promoters in each gene class by including orthologs of rice cell wall-related genes from *Brachypodium* as identified by Inparanoid (Table 4-1). We utilized two motif discovery

tools, MEME and RSAT, which use probabilistic and word-based algorithm, respectively. We only report significant motifs enriched within cell wall promoters compared to ones associated within random promoters. We evaluated the discovered motifs based on their location (e.g. distance from the transcription start sites) and association with target gene expression level. Then, we took advantage of the following two functional datasets to associate *cis*-elements with potential transcription factors that recognize them: (a) match discovered *cis*-elements in this study to recently published Arabidopsis large-scale protein binding microarray dataset that represents DNA binding sites for 65 transcription factors using the motif comparison platform TOMTOM (Franco-Zorrilla et al., 2014); (b) extract interactions between cell wall related genes included in this study and transcription factors based on Rice Combined mutual Rank Network (RCR v2) to infer what specific transcription factors may recognize the *cis*-elements of cell wall genes in grasses. The RCR v2 is a comprehensive and high-quality rice network based on computational and experimental evaluations in Chapter 3.

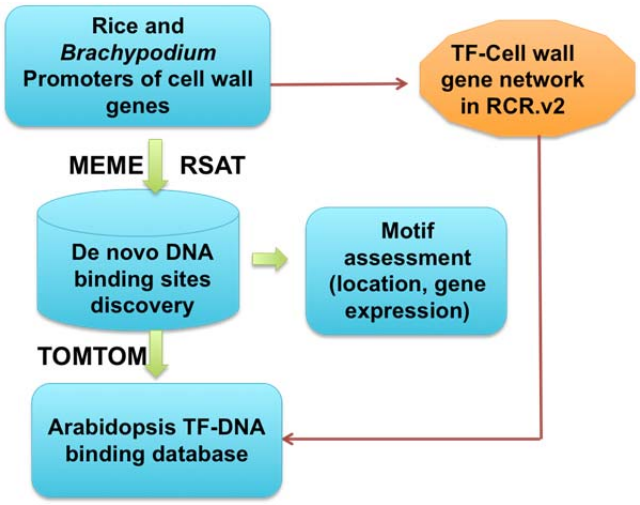


Figure 4-2 Workflow for comparative *de novo* motif discovery and rice gene network interrogation.

Promoter analysis of CESA and lignin biosynthesis genes

We first examined the putative *cis*-elements enriched within promoters of all *CESA* and lignin biosynthesis genes. Though expansions have occurred in these gene families in the grasses relative to Arabidopsis, genetic evidence demonstrates that the function of these genes are largely conserved between dicots and grasses (Boerjan et al., 2003; Somerville, 2006; Kumar and Turner, 2015; Zhong and Ye, 2015). In all, we observed six overrepresented motifs with hypergeometric p-values lower than 0.05 within the promoters of cellulose and lignin biosynthesis genes (Table 4-2; Hypergeometric p-values < 0.05 relative to 500 random rice promoters). We observed that three out of the six motifs tend to be associated with R2R3 MYB transcription factors, which may regulate cell wall and flavonoid biosynthesis or abiotic stress responses (Franco-Zorrilla et al., 2014). Particularly, the motif with the pattern, CACCAACC, represents a conserved Arabidopsis cell wall-associated *cis*-elements recognized by cell wall-associated R2R3 MYB proteins. This supports the model that regulation of the synthesis of the major SCW components (i.e., cellulose and lignin) tends to be conserved in grasses and that the R2R3 MYB transcription factor family is one of the predominant families that directly bind to promoters of CESA and lignin genes. We also discovered stress-related motifs recognized by the AP2/ERF and heat shock transcription factors (Table 4-2).

Table 4-2 Motifs discovered within the 1 kb upstream sequences of cell wall-related genes and putative associated transcription factors (TFs) and biological pathways.

Gene Clade	Motif	p-value*	Associated TF Family‡	Biological pathway
CESA. Lignin	CCCACCC	0.03	MYB55, MYB46; R2R3 MYB	Abiotic stress; cell wall
	CACCAACC	1.90E-11	MYB46; R2R3 MYB	Cell wall
	CCTACC	2.10E-07	MYB111, MYB46; R2R3 MYB	Cell wall; flavonoid
	AATTAATT	0.048	WOX13; Homeodomain proteins	Plant Development
	GCAGCAGC	0.0023	DREB2c; AP2/ERF	Abiotic stress; JA signaling
	GATCGATC	0.0036	HSFB; Heat shock transcription factor	Abiotic stress
Csls	CGGCG	0.0011	DEAR3; AP2/ERF	Abiotic stress response
	TAGCTA	5.45E-07	STY1; Zinc finger	Auxin synthesis
	ACCACC	4.16E-09	MYB46; R2R3 MYB	Cell wall
	ATGCAT	0.0042	ATERF; AP2/ERF	Biotic stress; ethylene signaling
	CCCCTC	0.0025	MYB111; R2R3 MYB	Flavonoid pathway
	CATGCAT	0.0002	NAC58; NAC	Cell wall
BAHD-ATs	CCTAGCTAGG	0.00012	STY1; Zinc finger	Auxin synthesis
	TCCCCTA	3.30E-11	MYB111, MYB55; R2R3 MYB	Flavonoid; Stress
	CAACTTG	0.0019	DAG2; C2C2	Unknown
	CAACAA	0.0029	MYB46; R2R3 MYB	Cell wall
	CATGCAT	2.54E-07	NAC58; NAC	Cell wall
	CTCCCTCCCC	3.58E-05	DOF5; C2C2	Unknown

* Hypergeometric p-values were calculated based on the occurrence of each motif within 500 randomly sampled promoters of rice.
‡ TF families associated with discovered motifs were matched to the Arabidopsis protein binding microarray datasets using TOMTOM (Gupta et al., 2007; Franco-Zorrilla et al., 2014).

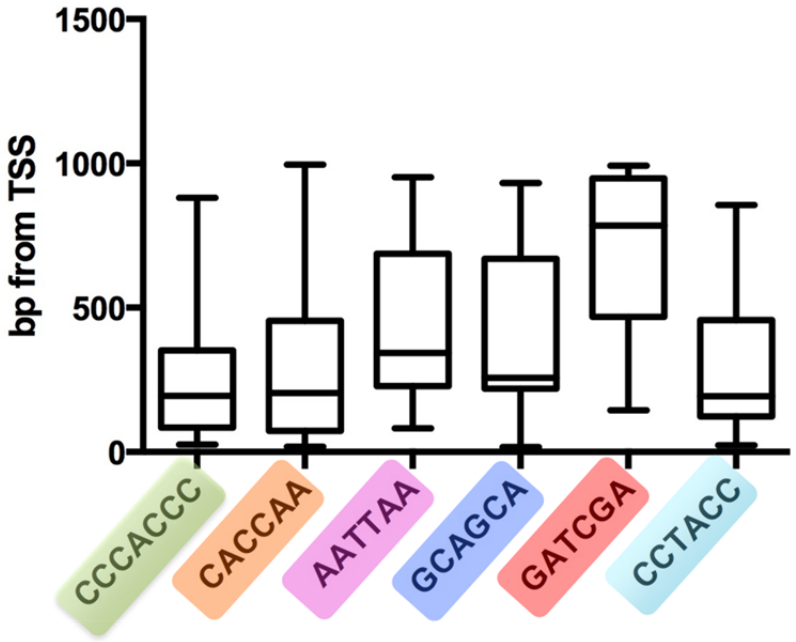


Figure 4-3 Location of discovered *cis*-elements within the 1kb promoters of *CESA* and lignin genes.

To distinguish likely active elements, we analyzed the motif location distribution relative to the transcription start sites (TSS), and the expression patterns of the downstream genes. Functional *cis*-elements tend to be located within 1kb from the TSS in plants (Velde et al., 2014). Five out of the six motifs are located within 500 bp of the TSS in rice promoters (Figure 4-3). The exception is the motif GATCGA, which is potentially bound by Heat shock-family transcription factors. We also observed that more copy numbers of particular *cis*-elements are associated with relatively high gene expression for genes from the same family. For example, we found the two motifs, CACCAA and CCCACCC, represented by the green and orange boxes in Figure 4-4, which are similar to known dicot cell wall associated *cis*-elements, are more abundant in the promoters of *OsCESA1* and *OsCESA8* (Figure 4-4).

Besides *de novo* motif prediction, we the comprehensive, high quality rice network, RCR, to provide hypotheses for which specific transcription factors might bind to the identified *cis*-elements.

We used the comprehensive, high-quality RCR network to independently establish associations between transcription factor families and cell wall target genes. We identified all transcription factors that interacted in the RCR v2 with *CESA* and lignin biosynthesis genes in with the cutoff higher than 0.003. Figure 4-5 shows the 32 transcription factor protein families that interact with the various SCW genes with the node sizes proportional to the number of connection (i.e., edges) between genes or gene family groups. Based on the number of edges, the top five families are MYB, NAC, ERF, bHLH and WRKY, which account for 58% of transcription factor-cell wall gene interactions. Among these top transcription factor families, we discovered *cis*-elements for MYB and ERF. This correspondence provides independent support that the de novo motif discovery approach is informative for identifying classes of direct regulators in this process.

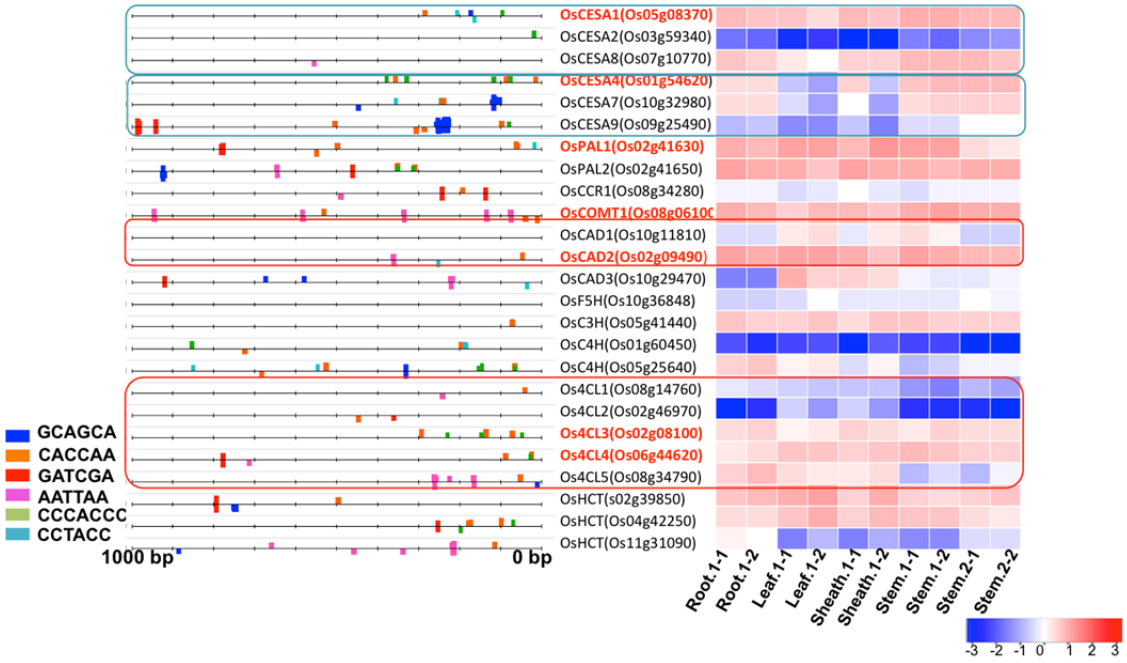


Figure 4-4 *Cis*-elements present within promoters of CESA and lignin genes and the expression pattern of corresponding genes during rice development. *Cis*-elements are represented by boxes and color-coded based on their consensus sequences with RSAT. Gene expression data are from the rice gene expression atlas (Wang et.al 2010).

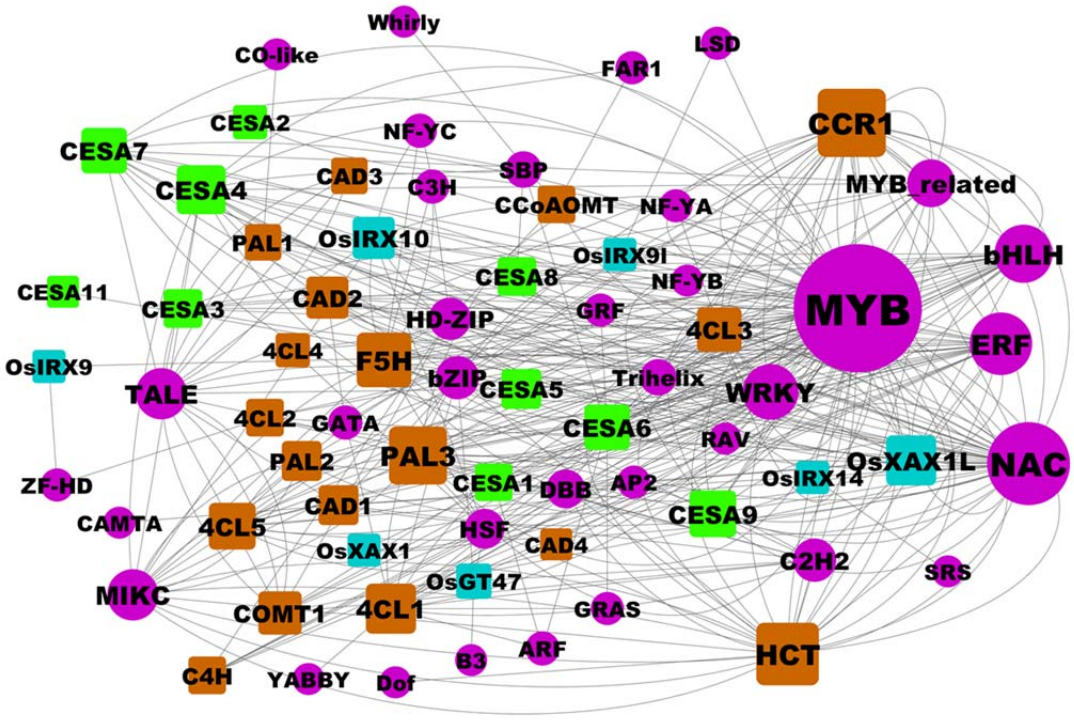


Figure 4-5 RCR network interactions between transcription factor families (magenta circles) CESAs (green squares), lignin biosynthesis genes (brown squares), and xylan biosynthesis genes (blue squares). Edges with scores higher than 0.003 were included in this sub-network. The size of nodes corresponds to the number of edges.

Promoter analysis of Csl genes

Csl and CESA belong to the glycosyltransferases (GT) 2 family and are closely related to each other based on phylogenetic analysis. In this analysis, we collected 20 rice *Csl* genes from the *CslD*, *CslE*, *CslF* and *CslH* groups (Schwerdt et al., 2015). We also added 15 orthologs in *Brachypodium*, which lacks a tandem expansion of the *CslF* family for a total of 35 promoters Csls promoters.

De novo motif discovery identified six significant motifs specifically enriched within the promoters of *Csl* genes (Table 4-2), with different distributions among the different *Csl* groups (Figure 4-6). Promoters of *CslD*s are enriched with CGaCGG and CCCCT, which are potential binding sites for DEAR transcription factors from the AP2/ERF family and MYB111 from the R2R3 MYB family, respectively. We also observed the motif, ACCACC, within every promoter of the *CslD* genes. CslE members are expanded via tandem duplication in rice and maintain fewer motifs compared to other Csl groups. The ACCACC and ATGCAT motifs in their promoters are putative binding sites for cell wall MYB and biotic stress response ERF proteins, respectively. CslF and CslH members are also duplicated *via* tandem duplication in rice. CslF6 is the ancient member within this group based on molecular clock analysis

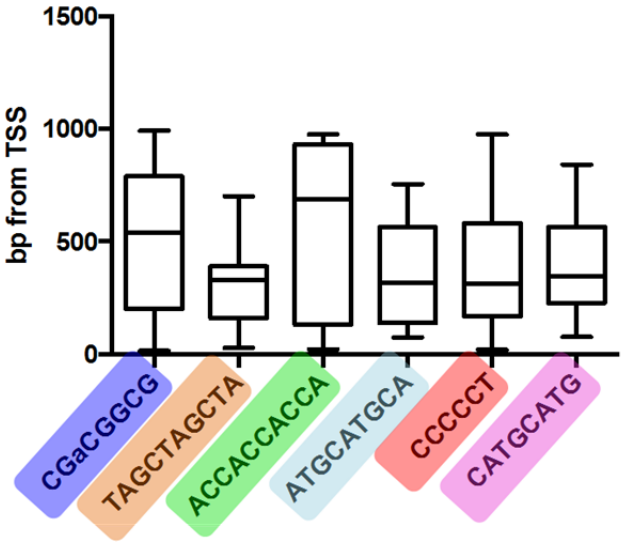


Figure 4-6 Location of motifs within the promoters of *Csl* genes.

(Schwerdt et al., 2015). Grass-diverged Csls tend to be enriched with the motifs, TAGCTAGC, CGaCGG and ACCACC within their promoters, which may be recognized by STY1, DEAR from the AP2/ERF family, and cell wall-associated R2R3 MYBs. The promoters of duplicated genes tend to maintain TAGCTAGC and ACCACC. However, we also observed that the duplicate members appear to have evolved the additional motifs, CGaCGG and ATGCABased on the Arabidopsis PBM datasets, the putative motifs may associate with transcription factors from R2R3 MYB, NAC, AP2/ERF and plant-specific zinc-finger protein families.

Among the putative DNA binding sites within *Csl* promoters, four out of six of them show relative proximity, average within 500 bp, to the transcription start sites. The potential binding sites for ERF and R2R3 MYB, CGGCG and ACCACC, respectively, have a broad distribution along the 1 kb upstream sequences (Figure 4-7).

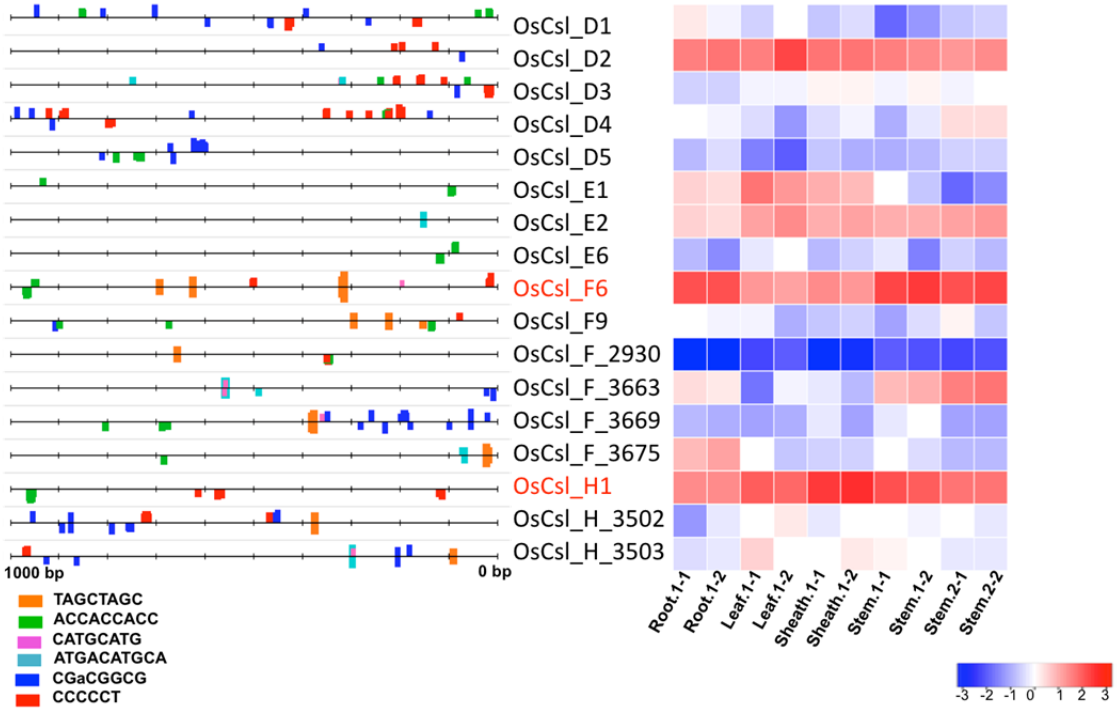


Figure 4-7 *Cis*-elements present within promoters of *Csl* genes and the expression pattern of corresponding genes during rice development. *Cis*-elements are represented by boxes and color-coded based on their consensus sequences with RSAT. Gene expression data were extracted from rice gene expression atlas (Wang et.al 2010).

We also extracted interactions between transcription factors and *Csl* genes from the RCR network to independently identify predominant transcription factor families that may bind to *Csl* promoters. Based on the number of edges, the protein families with highest degree are ERF, MYB, C2H2, bHLH and WRKY (Figure 4-8). Among them, ERF and MYBs are the two predominant families that correspond to the discovered DNA binding sites.

Promoter analysis of "Mitchell Clade" BAHD-Acyltransferases

Genetic and/or biochemical data have demonstrated roles for multiple members of the so-called "Mitchell Clade" BAHD acyltransferases (ATs) in incorporating HCAs into cell walls. The 20 "Mitchell Clade" ATs in rice divide into two sub-clades, with clade I being expressed more highly and possessing all members that have been characterized to date (Bartley et al., 2013). To identify potential regulators controlling grass cell wall specifics gene, we collected 20 promoters of "Mitchell Clade" ATs and nine orthologs from *Brachypodium* (Table 4-1).

Significant motifs enriched within AT promoters over random promoters are summarized in Table 4-2. We observed putative binding sites for R2R3 MYB transcription factors with the consensus patterns of TCCCCTA and CAACAA, which may be involved in cell wall biosynthesis, flavonoid, and stress response pathways. We also identified a putative motif that is similar to the known long and variable NAC cell wall transcription factors binding site, SNBE ([T/A]NN[C/T][T/C/G]TNNNNNNNA[A/C]GN[A/C/T][A/T]) (Wang et al., 2011).

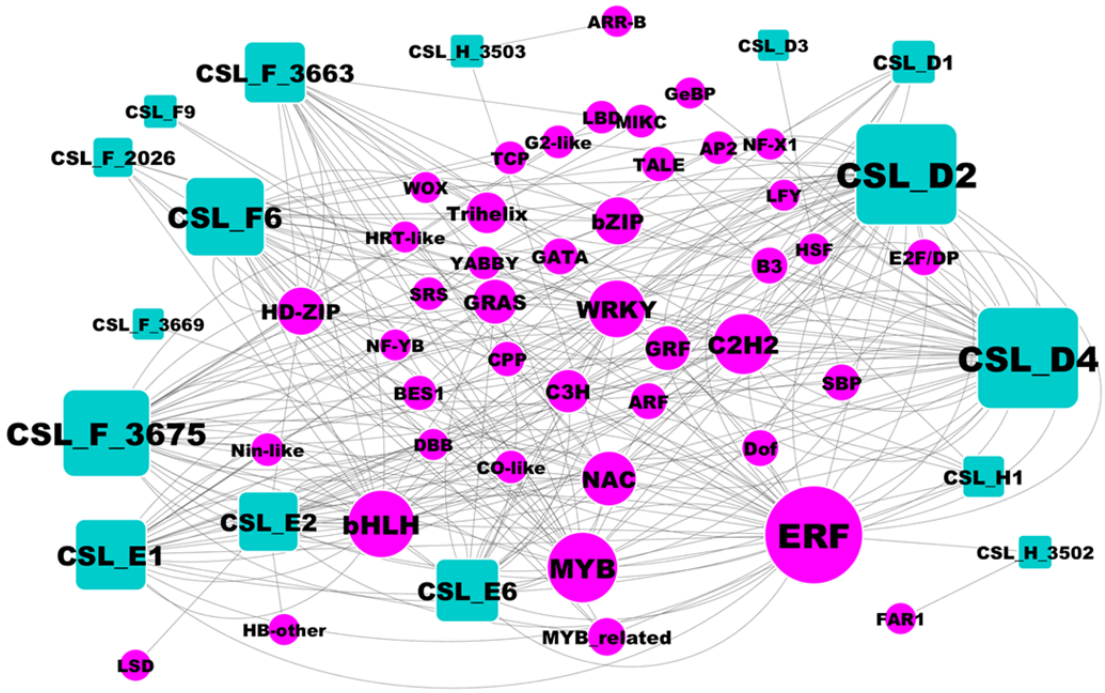


Figure 4-8 RCR network established interactions between transcription factors (magenta circles) and *Csl* genes (green rectangular). ERF, MYB, bHLH, WRKY and C2H2 are the transcription factors families that highly connected with *Csl* members. We only included edges with score higher than 0.003 in this sub-network. The size of nodes corresponds to the number of connections.

Another abundant motif is CCTAG, which may be recognized by a family of plant-specific zinc finger proteins that activate auxin biosynthesis. In addition, we discovered a CT-rich pattern, CTCCCTCCCC, which may be recognized by the C2C2 family proteins without further functional information. The discovered motifs within promoters of acyltransferases show broad distributions within 1kb upstream sequences (Figure 4-9).

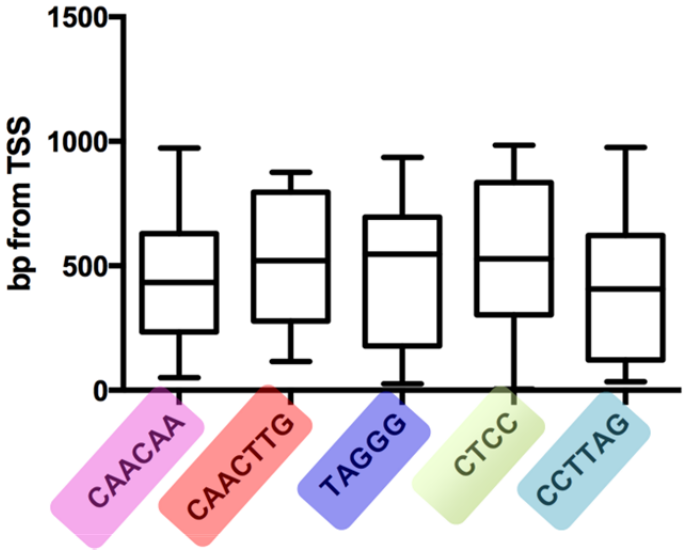


Figure 4-9 Location of motifs within the promoters of "Mitchell Clade" BAHD-ATs.

To infer whether the discovered motifs are associated with gene expression, we displayed the location, copy number of motifs and the expression pattern of corresponding genes (Figure 4-10). As previously noticed, members of subclade i, *OsAT1* to *OsAT10*, display relatively high expression in the major vegetative tissues. In contrast, subclade ii members, *OsAT11* to *OsAT20*, are lowly expressed.

Whole genome duplication (WGD) and tandem duplication contribute to 70% of the expansion of the "Mitchell Clade" ATs in rice (Figure 4-12). WGD and tandem duplication tend to maintain their original promoters. Particularly, four out five functionally characterized cell wall related ATs (OsAT1, OsAT4, OsAT7 and OsAT10, red font in Figure 4-9) result from WGD and tandem duplication. In contrast, tandem duplications tend to contribute to the expansion of Clade ii. Despite the dramatically different gene expression amounts; most discovered putative motifs are present within promoters from both subclades. The exception is CAACAA, represented by the light blue box (Figure 4-10), which is present only in clade i members of acyltransferases.

We also established interactions between transcription factors and acyltransferases inferred from the RCR (Figure 4-11). Based on the number of connections, we found that NAC family proteins have the most connections with ATs. The other four over-represented transcription factor families are MYB, AP2/ERF, WRKY and bZIP. Among them, the MYB and NAC families correspond to the DNA binding sites we found.

Comparison of discovered motifs

Next, we sought to compare the motifs identified for the three classes of grass cell wall genes. Our objective was to address the question of whether grass-specific genes have acquired (or maintained) cis-elements of conserved cell wall genes (Model 1), or if new (not previously cell wall-associated) TFs and corresponding regulatory elements have been recruited or retained by these genes (Model 4). We used TOMTOM to compare

the position weight matrices of each of the de novo motifs discussed above. Results are displayed as heat maps of Euclidian similarity normalized to the width of each motif.

The results show relatively high similarity between motifs discovered within CSL and "Mitchell Clade" BAHD-ATs with the ones from CESA and lignin genes promoters (Figure 4-13). We compared the position weight matrix (PWM) of CSL and "Mitchell Clade" BAHD-AT motifs with CESA and lignin motifs, respectively. For motifs discovered within CSL promoters, three out of six show high similarity with CESA and lignin motifs, which are potential binding sites for R2R3 MYBs and C2C2 proteins. For example, we observed that the most abundant motif discovered within CSL promoters, ACCACC, shows high similarity with known cell wall-related R2R3 MYB binding sites, ACCAACC and its variant, CCCACCC.

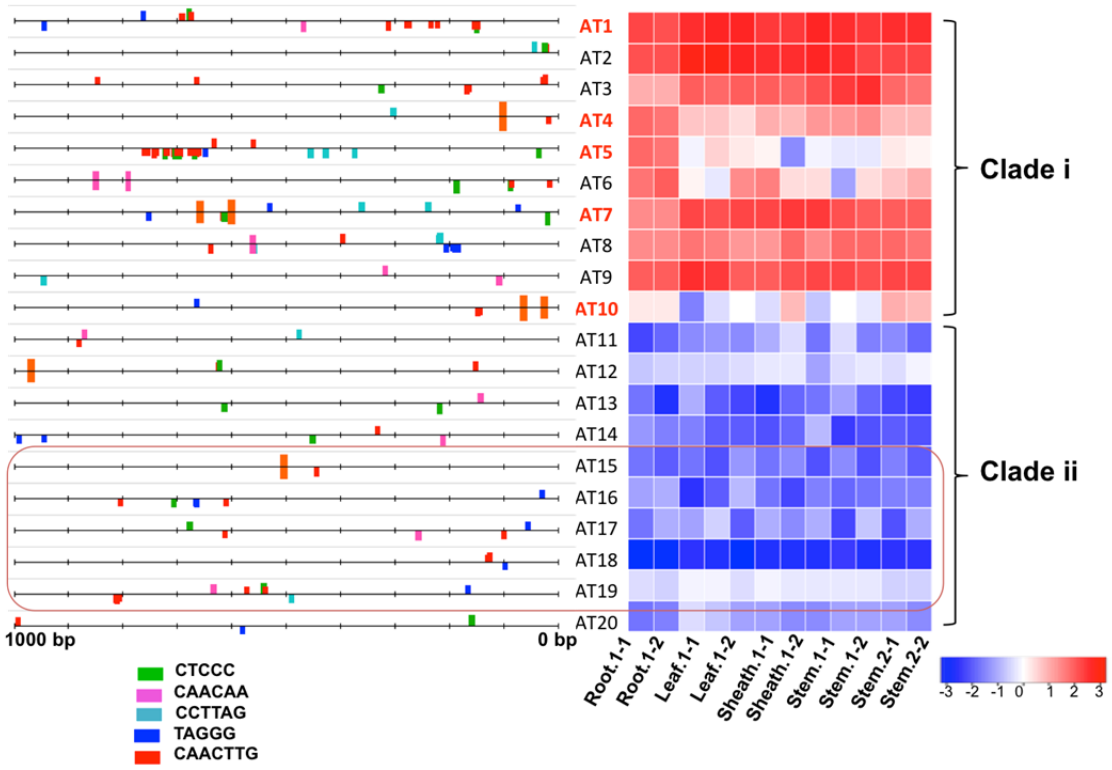


Figure 4-10 *Cis*-elements present within promoters of "Mitchell Clade" BAHD-ATs and the expression pattern of corresponding genes during rice development. *Cis*-elements are represented by boxes and color-coded based on their consensus sequences with RSAT. Gene expression data were extracted from rice gene expression atlas (Wang et.al., 2010). In rice, "Mitchell-clade" BAHD-ATs can be divided into two clades based on the phylogeny (Bartley et al., 2013)

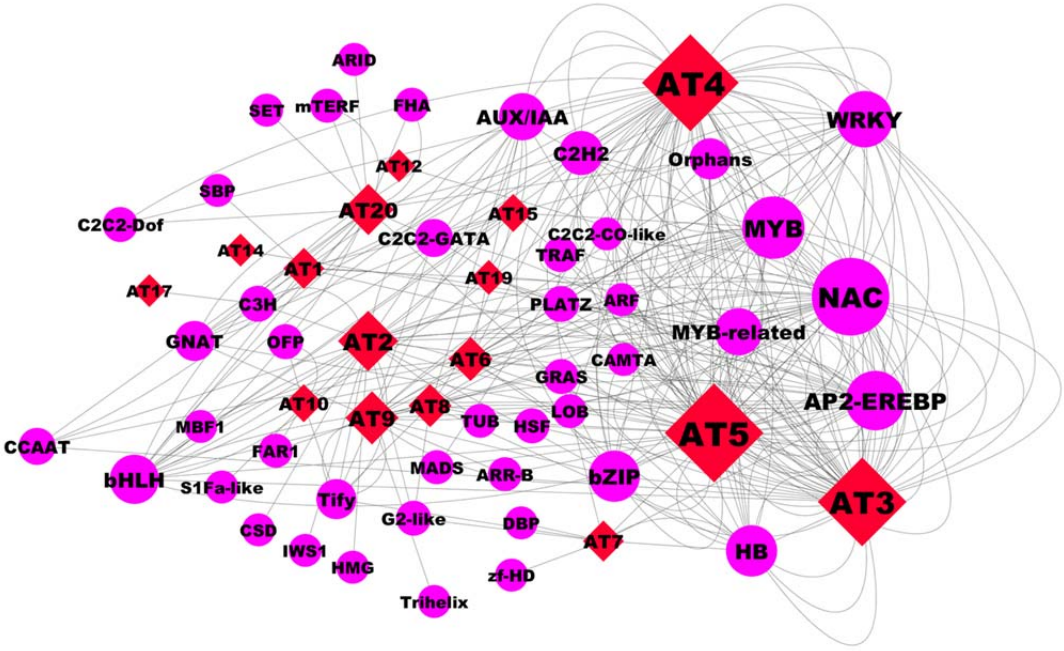


Figure 4-11 RCR v2 network established interactions between transcription factors (and BAHD-ATs. NAC, AP2/ERF, MYB, WRKY and bHLH represent highly connected protein families. We only included edges with score higher than 0.003 in this sub-network. The size of nodes corresponds to the number of connections.

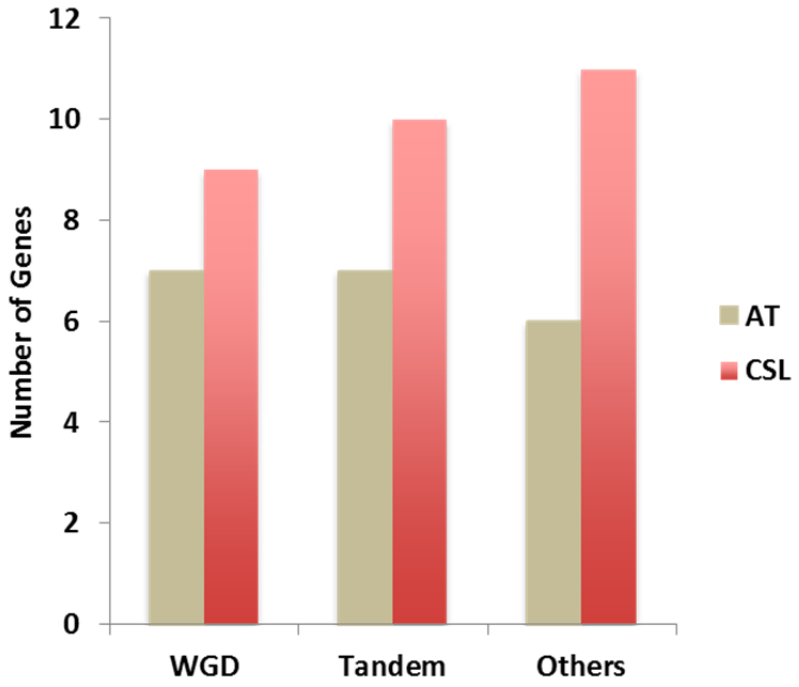


Figure 4-12 Origins of grass cell wall expanded genes in rice. Whole genome duplication (WGD) is based on identified duplication blocks in Plant Genome Duplication Database (http://chibba.agtec.uga.edu/uplication/).

Similarly, we compared discovered motifs within "Mitchell clade" *BAHD-ATs* with motifs present within promoters of lignin and *CESA* genes (Figure 4-14). The most abundant motifs present within the promoters of ATs are CTCCCT, which may be the potential binding site for C2C2 protein families. We also observed a motif, CAACAA, which is similar to the known cell wall associated R2R3 MYB binding site, CACCAA and its variant, CCCACCC. Putative binding sites for NAC and STY1 proteins were discovered in the promoters of CSL and ATs, but not CESA and lignin genes, supporting that the promoters of *CSL* and *ATs* may have evolved novel *cis*-elements, thus we expect to characterize additional transcription factors that can directly control the expression of grass cell wall-specific genes.

Discussion

Conservation of cell wall-associated cis-elements between dicots and grasses

This analysis systematically identified potential *cis*-elements present within promoters of rice cell wall biosynthesis genes. Via *de novo* motif analysis, we were able to recapitulate the presence of the known cell wall-associated DNA binding site, the AC-elements, within the promoters of *CESA* and lignin biosynthesis genes. The enrichment of this element within promoters, is consistent with previous observations that the switchgrass PvMYB4 protein is able to bind to AC-elements within the promoters of Arabidopsis known cell wall-associated DNA binding site, AC element, may be conserved in grasses (Zhong et al., 2011; Shen et al., 2013; Huang et al., 2015). In plants, the majority of plant MYB proteins that contain two MYB repeats can recognize the following sites: MBSI ((T/C)AAC(G/T)G(A/C/T) (A/C/T)), MBSII

(AGTTAGTTA), and MBSIIG ((C/T)ACC(A/T)A(A/C)C) (Prouse and Campbell, 2012). Most characterized cell wall-associated R2R3 MYB binding sites are similar to MBSIIG, which is also known as the AC-element. Though the binding specificity of transcription factors has not been well examined, the recently published Arabidopsis large-scale protein binding microarray analysis suggests R2R3 MYB may maintain unique binding affinity rather than recognizing multiple unrelated DNA binding sites (Franco-Zorrilla et al., 2014). In sum, CESA and lignin biosynthesis genes may maintain AC elements within their promoters across dicots and grasses. Thus, our analysis supports model 1, that both a regulatory element, the AC-element, and the general class of proteins that bind them, the MYBs have both been retained since the divergence of grasses and dicots.

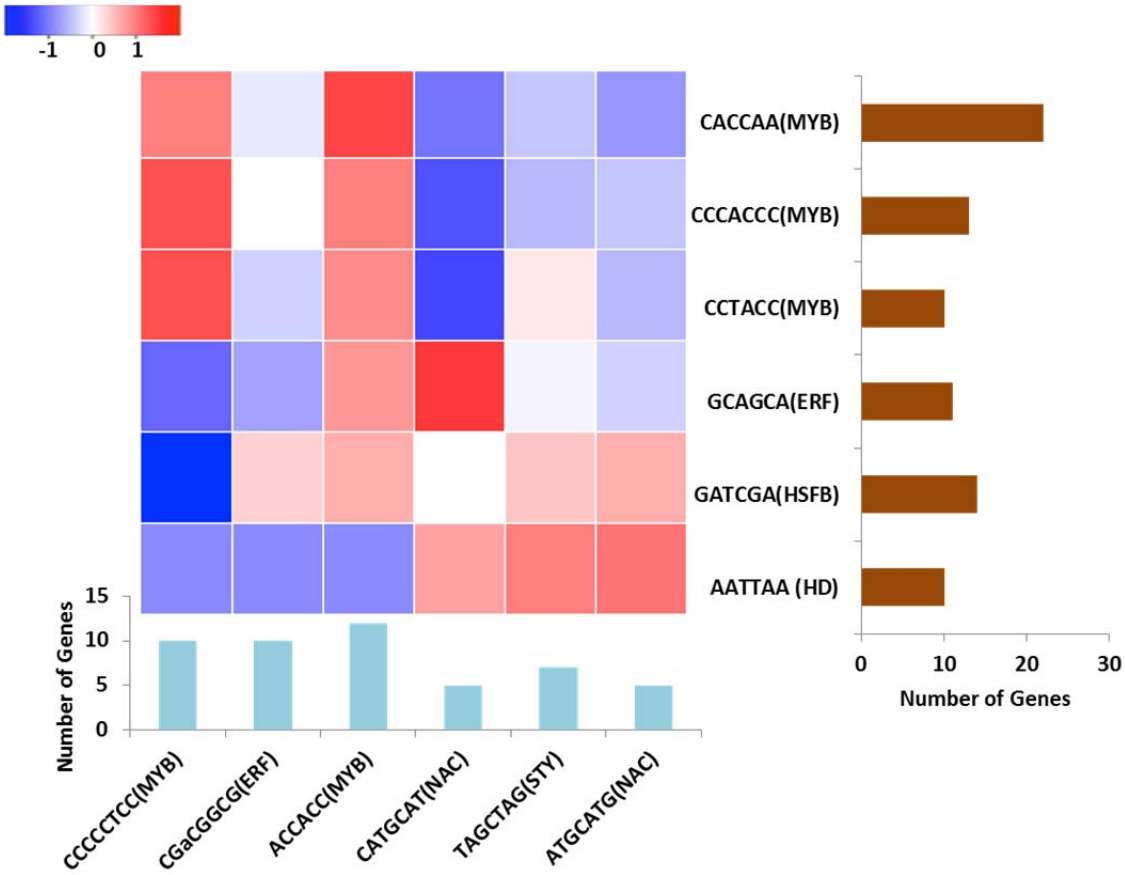


Figure 4-13 Comparison of motifs discovered within *CESA*, lignin promoters and *Csl* promoters. Euclidian similarity is calculated using RSAT and normalized to the width of a motif. Bar graphs show the presence of motifs within input promoter sequences.

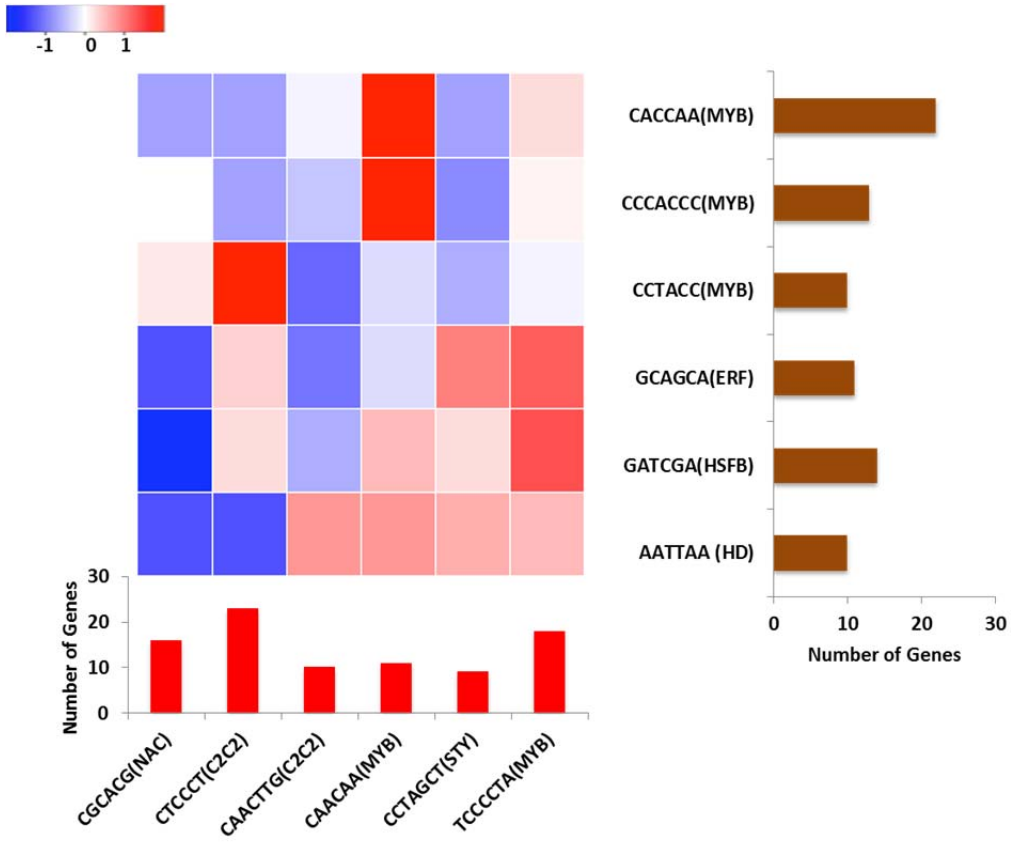


Figure 4-14 Comparison of motifs discovered within *CESA*, lignin and BAHD-ATs promoters. Euclidian similarity is calculated using RSAT and normalized to the width of a motif. Bar graphs show the presence of motifs within input promoter sequences.

Prediction of novel grass cell wall-associated regulators

The comparative *de novo* motif analysis also sheds light on additional transcription factor families that may directly bind the promoters of cell wall biosynthesis genes, such as AP2/ERF, C2C2, C2H2 and homeodomain proteins. AP2/ERF and homeodomain proteins are large proteins families that regulate multiple developmental processes, e.g. flower development, leaf epidermal cell identity, and embryo development (Chan et al., 1998; Ciftci-Yilmaz and Mittler, 2008; Pires and Dolan, 2010; Mizoi et al., 2012). However, only a few members have been characterized to control cell wall biosynthesis pathway (Zhong et al., 2008; Shen et al., 2012). In addition, C2C2 and C2H2 are two relatively understudied protein families, though a few members have been known to involve in stress response (Ciftci-Yilmaz and Mittler, 2008). In Arabidopsis, a recent yeast one-hybrid study using promoters of root xylem genes reveals that potential cell wall associated transcription factors from 35 families and overrepresented by AP2/ERF, bHLH, C2H2, C2C2-GATA and GRAS family regulators (Taylor-Teeples et al., 2014). Here, we observed potential binding sites for AP2/ERF, C2C2, C2H2 and bHLH that are significantly enriched within promoters of rice cell wall-associated genes. These results suggest that additional transcription factors associated with cell wall biosynthesis pathway remain to be revealed.

Incorporation of grass cell wall-specific genes into cell wall regulatory pathways

The similarity of *cis*-elements between *CESA* and lignin biosynthesis genes and *Csl* or *BAHD-ATs* suggests that the grass-specific genes have been incorporated into cell wall

regulatory pathways by maintaining or evolving known cell wall-associated DNA binding sites. In plants, duplication is a predominant feature of genomes and contributes dramatically to gene family expansion (De Smet and Van de Peer, 2012; De Smet et al., 2013). Duplicate genes may maintain their original function, or evolve partial or completely novel functions (i.e., sub- and neo-functionalization), thus promoting emergence of new pathways (Zhang, 2003; Ward and Durrett, 2004). Duplication events also have different effects on gene promoters. Usually, whole-genome scale duplication (WGD) and tandem duplication maintain original promoters and individual gene duplication induced by transposable elements may result in the loss of gene promoters (Paterson et al., 2004; Wang et al., 2012). Besides emergence of new genes, corresponding regulators, mainly transcription factors, may also need to "evolve the ability" to coordinate expression of different genes within the same pathway. *Csl* genes including members from F and H clades belong to GT2 family of glycosyltransferases together with *CESAs*. In rice, tandem duplications contribute to the emergence of grass-diverged *CslF* and *H* genes. Here, we found that promoters of *CslF* and *H* genes may also maintain AC-elements and be regulated by one or more R2R3 MYBs. A focus of the previous chapter, *CslF* gene expression correlated with up- and down-regulation of OsMYB61a in protoplast or whole plant in the *myb61a* mutant. Thus, OsMYB61a is a candidate for directly regulating other expression of grass diverged *CslF* genes.

On the other hand, "Mitchell Clade" ATs are grass-expanded and belong to Clade V of BAHD-ATs in plants (Mitchell et al., 2007). Mutants of the closest homolog of these genes members in Arabidopsis does not show an HCA phenotype (Rautengarten et al., 2012). In this study, we also observed known cell wall-associated

cis-elements, including R2R3 MYB and NAC DNA binding sites, within the promoters of grass-expanded ATs. This suggests that grass-diverged BAHD-ATs were recruited into cell wall related pathway by evolving corresponding cell wall-associated transcription factors binding sites. In Chapter 3 of this dissertation, molecular genetics revealed that OsMYB61a can alter the expression of grass-diverged *Csl F/H* genes and cell wall associated ATs in a direct or indirect manner. As with the Csl's, our observation of AC-elements within cell wall AT promoters, is consistent with direct binding of OsMYB61 and/or another R2R3 MYB protein to these promoters. In all, enriched *cis*-elements discovered within *Csl* and *BAHD-ATs* promoter supports model 1 as the incorporation mechanism to explain the recruitment of grass cell wall specific genes. Moreover, we also observed potential DNA binding sites associated with AP2/ERF, zinc finger and C2C2 proteins. Thus, it is possible that novel transcription factors may be able to directly regulated *CslF/H* and grass-expanded *BAHD-ATs*. Thus, model 4 may also be an additional mechanism to incorporated grass-expanded genes into cell wall biosynthesis pathway.

Challenges of De novo motif prediction

It is challenging to distinguish noisy signals with functional DNA binding sites. In this analysis, we took advantage of different informatics and biological strategies. Firstly, to avoid false negatives, we incorporated de novo motif discovery results from two widely used tools, namely, MEME and RSAT (Bailey et al., 2006; Turatsinze et al., 2008). MEME motifs are represented by position-dependent letter-probability matrics and the discovery process requires probabilistic sequence models based on position weight

matrix. MEME also considers local optimality rather than global by including an expectation maximization (EM) algorithm (Das and Dai, 2007). RSAT discovers local over-representation using a word-based (string-based) method that mostly relies on exhaustive enumeration. Further, we focused on putative DNA binding sites with the size between 6 to 10 bp and relatively local to the transcription start sites based on the features that have been observed in different studies of *cis*-elements in plant genomes (Wittkopp and Kalay, 2012; Hernandez-Garcia and Finer, 2014; Arsovski et al., 2015; Jiang, 2015). Based on these criteria, we are able to discover AC elements within promoters of cell wall biosynthesis genes, which suggests the prediction strategy is effective. Thus, the next critical step to functionally screen interactions between discovered potential *cis*-elements and grass cell wall associated transcription factors.

Conclusion

In this analysis, we took advantage of comparative *de novo* motif prediction to identify potential *cis*-elements present within promoters of rice *CESA* and lignin biosynthesis genes and grass-cell wall specific genes, including *Csl* and *BAHD-ATs*. The significantly enriched motifs suggest that known cell wall-associated DNA binding sites may be conserved in the promoters of rice *CESA* and lignin genes. We also observed R2R3 MYB and NAC binding sites within the promoters of grass cell wall-specific genes, including functionally characterized cell wall associated members (e.g. *Cslf6*, *Cslh1*, *AT4*, and *AT5*). This indicates that grass-expanded/diverged genes have been incorporated into cell wall biosynthesis pathway *via* maintaining or evolving known cell wall associated *cis*-elements. In addition, we do expect novel regulators from the

families of C2H2, C2C2, AP2/ERF and homeodomain, which have not been well examined in the regulation of cell wall biosynthesis in any species. This analysis will direct functional characterization of cell wall-associated regulators in grasses, which will facilitate the understanding of complex traits (e.g. cell wall) and their regulation, and further promote food, fiber, and biofuel production.

Reference

Ambavaram, M.M.R., Krishnan, A., Trijatmiko, K.R., and Pereira, A. (2011). Coordinated Activation of Cellulose and Repression of Lignin Biosynthesis Pathways in Rice. Plant Physiology 155, 916-931.

Arsovski, A.A., Pradinuk, J., Guo, X.Q., Wang, S., and Adams, K.L. (2015). Evolution of Cis-Regulatory Elements and Regulatory Networks in Duplicated Genes of Arabidopsis. Plant Physiology 169, 2982-2991.

Bailey, T.L., Williams, N., Misleh, C., and Li, W.W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Research 34, W369-W373.

Bartley, L., and Ronald, P. (2009). Plant and microbial research seeks biofuel production from lignocellulose. California agriculture 63, 178-184.

Bartley, L.E., Peck, M.L., Kim, S.R., Ebert, B., Manisseri, C., Chiniquy, D.M., Sykes, R., Gao, L., Rautengarten, C., Vega-Sanchez, M.E., Benke, P.I., Canlas, P.E., Cao, P., Brewer, S., Lin, F., Smith, W.L., Zhang, X., Keasling, J.D., Jentoff, R.E., Foster, S.B., Zhou, J., Ziebell, A., An, G., Scheller, H.V., and Ronald, P.C. (2013). Overexpression of a BAHD acyltransferase, OsAt10, alters rice cell wall hydroxycinnamic acid content and saccharification. Plant Physiol 161, 1615-1633.

Binod, P., Sindhu, R., Singhania, R.R., Vikram, S., Devi, L., Nagalakshmi, S., Kurien, N., Sukumaran, R.K., and Pandey, A. (2010). Bioethanol production from rice straw: An overview. Bioresource Technology 101, 4767-4774.

Boer, D.R., Freire-Rios, A., van den Berg, Willy A.M., Saaki, T., Manfield, Iain W., Kepinski, S., López-Vidrieo, I., Franco-Zorrilla, Jose M., de Vries, Sacco C., Solano, R., Weijers, D., and Coll, M. (2014). Structural Basis for DNA Binding Specificity by the Auxin-Dependent ARF Transcription Factors. Cell 156, 577-589.

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

Bontpart, T., Cheynier, V., Ageorges, A., and Terrier, N. (2015). BAHD or SCPL acyltransferase? What a dilemma for acylation in the world of plant phenolic compounds. New Phytologist, n/a-n/a.

Buanafina, M.M.d.O., Fescemyer, H.W., Sharma, M., and Shearer, E.A. (2016). Functional testing of a PF02458 homologue of putative rice arabinoxylan feruloyl transferase genes in Brachypodium distachyon. Planta 243, 659-674.

Burton, R.A., Wilson, S.M., Hrmova, M., Harvey, A.J., Shirley, N.J., Medhurst, A., Stone, B.A., Newbigin, E.J., Bacic, A., and Fincher, G.B. (2006). Cellulose synthase-like CslF genes mediate the synthesis of cell wall (1, 3; 1, 4)-ß-D-glucans. Science 311, 1940-1942.

Carpita, N.C. (2012). Progress in the biological synthesis of the plant cell wall: new ideas for improving biomass for bioenergy. Current Opinion in Biotechnology 23, 330-337.

Chan, R.L., Gago, G.M., Palena, C.M., and Gonzalez, D.H. (1998). Homeoboxes in plant development. Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression 1442, 1-19.

Chen, M.-J.M., Chou, L.-C., Hsieh, T.-T., Lee, D.-D., Liu, K.-W., Yu, C.-Y., Oyang, Y.-J., Tsai, H.-K., and Chen, C.-Y. (2012). De novo motif discovery facilitates identification of interactions between transcription factors in Saccharomyces cerevisiae. Bioinformatics 28, 701-708.

Ciftci-Yilmaz, S., and Mittler, R. (2008). The zinc finger network of plants. Cellular and Molecular Life Sciences 65, 1150-1160.

D'Auria, J.C. (2006). Acyltransferases in plants: a good time to be BAHD. Current Opinion in Plant Biology 9, 331-340.

Das, M., and Dai, H.-K. (2007). A survey of DNA motif finding algorithms. BMC Bioinformatics 8, S21.

De Smet, R., and Van de Peer, Y. (2012). Redundancy and rewiring of genetic networks following genome-wide duplication events. Current Opinion in Plant Biology 15, 168-176.

De Smet, R., Adams, K.L., Vandepoele, K., Van Montagu, M.C.E., Maere, S., and Van de Peer, Y. (2013). Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. Proc. Natl. Acad. Sci. 110, 2898-2903.

de Velde, J.V., Heyndrickx, K.S., and Vandepoele, K. (2014). Inference of Transcriptional Networks in Arabidopsis through Conserved Noncoding Sequence Analysis. The Plant Cell Online 26, 2729-2745.

Ding, J., Hu, H., and Li, X. (2012). Thousands of Cis-Regulatory Sequence Combinations Are Shared by Arabidopsis and Poplar. Plant Physiology 158, 145-155.

Fincher, G.B., and Burton, R.A. (2014). Evolution and Development of Cell Walls in Cereal Grains. Frontiers in Plant Science 5.

Franco-Zorrilla, J.M., López-Vidriero, I., Carrasco, J.L., Godoy, M., Vera, P., and Solano, R. (2014). DNA-binding specificities of plant transcription factors and their potential to define target genes. Proceedings of the National Academy of Sciences 111, 2367-2372.

Gui, J., Shen, J., and Li, L. (2011). Functional Characterization of Evolutionarily Divergent 4-Coumarate:Coenzyme A Ligases in Rice. Plant Physiology 157, 574-586.

Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., and Noble, W.S. (2007). Quantifying similarity between motifs. Genome Biology 8, 1-9.

Handakumbura, P.P., and Hazen, S.P. (2012). Transcriptional Regulation of Grass Secondary Cell Wall Biosynthesis: Playing Catch-Up with Arabidopsis thaliana. Front Plant Sci 3, 74.

Hernandez-Garcia, C.M., and Finer, J.J. (2014). Identification and validation of promoters and cis-acting regulatory elements. Plant Science 217–218, 109-119.

Huang, D., Wang, S., Zhang, B., Shang-Guan, K., Shi, Y., Zhang, D., Liu, X., Wu, K., Xu, Z., Fu, X., and Zhou, Y. (2015). A Gibberellin-Mediated DELLA-NAC Signaling Cascade Regulates Cellulose Synthesis in Rice. The Plant Cell 27, 1681-1696.

Inada, D.C., Bashir, A., Lee, C., Thomas, B.C., Ko, C., Goff, S.A., and Freeling, M. (2003). Conserved Noncoding Sequences in the Grasses4. Genome research 13, 2030-2041.

Jiang, J. (2015). The ‘dark matter’ in the plant genomes: non-coding and unannotated DNA sequences associated with open chromatin. Current Opinion in Plant Biology 24, 17-23.

Kakei, Y., Ogo, Y., Itai, R.N., Kobayashi, T., Yamakawa, T., Nakanishi, H., and Nishizawa, N.K. (2013). Development of a novel prediction method of cis-elements to hypothesize collaborative functions of cis-element pairs in iron-deficient rice. Rice 6, 1-14.

Karlen, S.D., Peck, M.L., Zhang, C., Smith, R.A., Padmakshan, D., Helmich, K.E., Free, H.C.A., Lee, S., Smith, B.G., Lu, F., Sedbrook, J.C., Sibout, R., Grabber, J.H., Runge, T.M., Mysore, K.S., Harris, P.J., Bartley, L.E., and Ralph, J. . (Submitted). Monolignol Ferulate Conjugates are Naturally Incorporated into Plant Lignins. Science Advances.

Keegstra, K. (2010). Plant Cell Walls. Plant Physiology 154, 483-486.

Kellogg, E.A. (2001). Evolutionary History of the Grasses. Plant Physiol. 125, 1198-1205.

Kheradpour, P., and Kellis, M. (2014). Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Research 42, 2976-2987.

Kiemle, S.N., Zhang, X., Esker, A.R., Toriz, G., Gatenholm, P., and Cosgrove, D.J. (2014). Role of (1,3)(1,4)-β-Glucan in Cell Walls: Interaction with Cellulose. Biomacromolecules 15, 1727-1736.

Kim, S.J., Zemelis, S., Keegstra, K., and Brandizzi, F. (2015). The cytoplasmic localization of the catalytic site of CSLF6 supports a channeling model for the biosynthesis of mixed-linkage glucan. Plant Journal 81, 537-547.

Kumar, M., and Turner, S. (2015). Plant cellulose synthesis: CESA proteins crossing kingdoms. Phytochemistry 112, 91-99.

Lal, R. (2005). World crop residues production and implications of its use as a biofuel. Environment International 31, 575-584.

Li, Y., Kim, J.I., Pysh, L., and Chapple, C. (2015). Four Isoforms of Arabidopsis 4-Coumarate:CoA Ligase Have Overlapping yet Distinct Roles in Phenylpropanoid Metabolism. Plant Physiology 169, 2409-2421.

Liu, J.-H., Peng, T., and Dai, W. (2014). Critical cis-Acting Elements and Interacting Transcription Factors: Key Players Associated with Abiotic Stress Responses in Plants. Plant Molecular Biology Reporter 32, 303-317.

Maruyama, K., Todaka, D., Mizoi, J., Yoshida, T., Kidokoro, S., Matsukura, S., Takasaki, H., Sakurai, T., Yamamoto, Y.Y., Yoshiwara, K., Kojima, M., Sakakibara, H., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2012). Identification of Cis-Acting Promoter Elements in Cold- and Dehydration-Induced Transcriptional Pathways in Arabidopsis, Rice, and Soybean. DNA Research 19, 37-49.

Mitchell, R.A.C., Dupree, P., and Shewry, P.R. (2007). A Novel Bioinformatics Approach Identifies Candidate Genes for the Synthesis and Feruloylation of Arabinoxylan. Plant Physiology 144, 43-53.

Mizoi, J., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2012). AP2/ERF family transcription factors in plant abiotic stress responses. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms 1819, 86-96.

Morris, R.T., O'Connor, T.R., and Wyrick, J.J. (2008). Osiris: an integrated promoter database for Oryza sativa L. Bioinformatics 24, 2915-2917.

Nakano, Y., Yamaguchi, M., Endo, H., Rejab, N.A., and Ohtani, M. (2015). NAC-MYB-based transcriptional regulation of secondary cell wall biosynthesis in land plants. Frontiers in plant science 6.

Östlund, G., Schmitt, T., Forslund, K., Köstler, T., Messina, D.N., Roopra, S., Frings, O., and Sonnhammer, E.L.L. (2010). InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. 38, D196-D203.

Paterson, A.H., Bowers, J.E., and Chapman, B.A. (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proceedings of the National Academy of Sciences of the United States of America 101, 9903-9908.

Petrik, D.L., Karlen, S.D., Cass, C.L., Padmakshan, D., Lu, F., Liu, S., Le Bris, P., Antelme, S., Santoro, N., Wilkerson, C.G., Sibout, R., Lapierre, C., Ralph, J., and Sedbrook, J.C. (2014). p-Coumaroyl-CoA:monolignol transferase (PMT) acts specifically in the lignin biosynthetic pathway in Brachypodium distachyon. The Plant Journal 77, 713-726.

Pires, N., and Dolan, L. (2010). Origin and Diversification of Basic-Helix-Loop-Helix Proteins in Plants. Molecular Biology and Evolution 27, 862-874.

Popper, Z.A., Michel, G., Hervé, C., Domozych, D.S., Willats, W.G.T., Tuohy, M.G., Kloareg, B., and Stengel, D.B. (2011). Evolution and Diversity of Plant Cell Walls: From Algae to Flowering Plants. Annual Review of Plant Biology 62, 567-590.

Priest, H.D., Filichkin, S.A., and Mockler, T.C. (2009). cis-Regulatory elements in plant cell signaling. Current Opinion in Plant Biology 12, 643-649.

Prouse, M.B., and Campbell, M.M. (2012). The interaction between MYB proteins and their target DNA binding sites. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms 1819, 67-77.

Rautengarten, C., Ebert, B., Ouellet, M., Nafisi, M., Baidoo, E.E.K., Benke, P., Stranne, M., Mukhopadhyay, A., Keasling, J.D., Sakuragi, Y., and Scheller, H.V. (2012). Arabidopsis Deficient in Cutin Ferulate Encodes a Transferase Required for Feruloylation of ω-Hydroxy Fatty Acids in Cutin Polyester. Plant Physiology 158, 654-665.

Rodriguez-Granados, N.Y., Ramirez-Prado, J.S., Veluchamy, A., Latrasse, D., Raynaud, C., Crespi, M., Ariel, F., and Benhamed, M. (2016). Put your 3D glasses on: plant chromatin is on show. J Exp Bot 67, 3205-3221.

Schwerdt, J.G., MacKenzie, K., Wright, F., Oehme, D., Wagner, J.M., Harvey, A.J., Shirley, N.J., Burton, R.A., Schreiber, M., Halpin, C., Zimmer, J., Marshall, D.F., Waugh, R., and Fincher, G.B. (2015). Evolutionary Dynamics of the Cellulose Synthase Gene Superfamily in Grasses. Plant Physiology 168, 968-983.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research 13, 2498-2504.

Shen, H., He, X., Poovaiah, C.R., Wuddineh, W.A., Ma, J., Mann, D.G., Wang, H., Jackson, L., Tang, Y., and Neal Stewart Jr, C. (2012). Functional characterization of the switchgrass (Panicum virgatum) R2R3‐MYB transcription factor PvMYB4 for improvement of lignocellulosic feedstocks. New Phytologist.

Shen, H., Poovaiah, C., Ziebell, A., Tschaplinski, T., Pattathil, S., Gjersing, E., Engle, N., Katahira, R., Pu, Y., Sykes, R., Chen, F., Ragauskas, A., Mielenz, J., Hahn, M., Davis, M., Stewart, C.N., and Dixon, R. (2013). Enhanced characteristics of genetically modified switchgrass (Panicum virgatum L.) for high biofuel production. Biotechnology for Biofuels 6, 71.

Sibout, R., Eudes, A., Mouille, G., Pollet, B., Lapierre, C., Jouanin, L., and Seguin, A. (2005). Cinnamyl alcohol dehydrogenase -C and -D are the primary genes involved in lignin biosynthesis in the floral stem of Arabidopsis. Plant Cell 17.

Somerville, C. (2006). Cellulose synthesis in higher plants. Annual review of cell and developmental biology 22, 53-78.

Somerville, C. (2007). Biofuels. Current Biology 17, R115-R119.

Taylor-Teeples, M., Lin, L., de Lucas, M., Turco, G., Toal, T.W., Gaudinier, A., Young, N.F., Trabucco, G.M., Veling, M.T., Lamothe, R., Handakumbura, P.P., Xiong, G., Wang, C., Corwin, J., Tsoukalas, A., Zhang, L., Ware, D., Pauly, M., Kliebenstein, D.J., Dehesh, K., Tagkopoulos, I., Breton, G., Pruneda-Paz, J.L., Ahnert, S.E., Kay, S.A., Hazen, S.P., and Brady, S.M. (2014). An Arabidopsis gene regulatory network for secondary cell wall synthesis. Nature advance online publication.

Taylor-Teeples, M., Lin, L., de Lucas, M., Turco, G., Toal, T.W., Gaudinier, A., Young, N.F., Trabucco, G.M., Veling, M.T., Lamothe, R., Handakumbura, P.P., Xiong, G., Wang, C., Corwin, J., Tsoukalas, A., Zhang, L., Ware, D., Pauly, M., Kliebenstein, D.J., Dehesh, K., Tagkopoulos, I., Breton, G., Pruneda-Paz, J.L., Ahnert, S.E., Kay,

S.A., Hazen, S.P., and Brady, S.M. (2015). An Arabidopsis gene regulatory network for secondary cell wall synthesis. Nature 517, 571-575.

Turatsinze, J.-V., Thomas-Chollier, M., Defrance, M., and van Helden, J. (2008). Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. Nat. Protocols 3, 1578-1588.

Van Nostrand, E.L., and Kim, S.K. (2013). Integrative analysis of C. elegans modENCODE ChIP-seq data sets to infer gene regulatory interactions. Genome research 23, 941-953.

Vega-Sánchez, M.E., Verhertbruggen, Y., Christensen, U., Chen, X., Sharma, V., Varanasi, P., Jobling, S.A., Talbot, M., White, R.G., Joo, M., Singh, S., Auer, M., Scheller, H.V., and Ronald, P.C. (2012). Loss of Cellulose Synthase-Like F6 Function Affects Mixed-Linkage Glucan Deposition, Cell Wall Mechanical Properties, and Defense Responses in Vegetative Tissues of Rice. Plant Physiology 159, 56-69.

Vogel, J. (2008). Unique aspects of the grass cell wall. Curr. Opin. Plant Biol. 11, 301-307.

Voss, T.C., and Hager, G.L. (2014). Dynamic regulation of transcriptional states by chromatin and transcription factors. Nature reviews 15, 69-81.

Walcher, C.L., and Nemhauser, J.L. (2012). Bipartite Promoter Element Required for Auxin Response. Plant Physiology 158, 273-282.

Wang, H., Zhao, Q., Chen, F., Wang, M., and Dixon, R.A. (2011). NAC domain function and transcriptional control of a secondary cell wall master switch. The Plant Journal 68, 1104-1114.

Wang, L., Xie, W., Chen, Y., Tang, W., Yang, J., Ye, R., Liu, L., Lin, Y., Xu, C., Xiao, J., and Zhang, Q. (2010). A dynamic gene expression atlas covering the entire life cycle of rice. Plant J 61, 752-766.

Wang, X., Haberer, G., and Mayer, K.F. (2009). Discovery of cis-elements between sorghum and rice using co-expression and evolutionary conservation. BMC genomics 10, 1-15.

Wang, Y., Wang, X., and Paterson, A.H. (2012). Genome and gene duplications and gene expression divergence: a view from plants. Annals of the New York Academy of Sciences 1256, 1-14.

Ward, R., and Durrett, R. (2004). Subfunctionalization: How often does it occur? How long does it take? Theoretical Population Biology 66, 93-100.

Withers, S., Lu, F., Kim, H., Zhu, Y., Ralph, J., and Wilkerson, C.G. (2012). Identification of Grass-specific Enzyme That Acylates Monolignols with p-Coumarate. Journal of Biological Chemistry 287, 8347-8355.

Wittkopp, P.J., and Kalay, G. (2012). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nature reviews 13, 59-69.

Youngs, H., and Somerville, C. (2012). Development of feedstocks for cellulosic biofuels. F1000 biology reports 4, 10.

Zhang, J. (2003). Evolution by gene duplication: an update. Trends in Ecology & Evolution 18, 292-298.

Zhong, R., and Ye, Z.-H. (2007). Regulation of cell wall biosynthesis. Current opinion in plant biology 10, 564-572.

Zhong, R., and Ye, Z.-H. (2015). Secondary Cell Walls: Biosynthesis, Patterned Deposition and Transcriptional Regulation. Plant and Cell Physiology 56, 195-214.

Zhong, R., Lee, C., and Ye, Z.H. (2010). Functional characterization of poplar wood-associated NAC domain transcription factors. Plant Physiol 152, 1044-1055.

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

Zou, C., Lehti-Shiu, M.D., Thomashow, M., and Shiu, S.-H. (2009). Evolution of Stress-Regulated Gene Expression in Duplicate Genes of *Arabidopsis thaliana*. PLoS genetics 5, e1000581.

Zou, C., Sun, K., Mackaluso, J.D., Seddon, A.E., Jin, R., Thomashow, M.F., and Shiu, S.-H. (2011). Cis-regulatory code of stress-responsive transcription in Arabidopsis thaliana. Proceedings of the National Academy of Sciences 108, 14992-14997.

Chapter 5 : Future directions for functional characterization of novel cell wall associated transcription factors and their corresponding binding sites in rice

In this dissertation, I aimed to identify transcription factors controlling grass cell wall biosynthesis by exploring the following three objectives: (1) Analyze the conservation of known dicot cell wall-associated transcription factors in grasses; (2) Predict and functionally examine novel grass cell wall-associated transcription factors, particularly focusing on members that may control grass cell wall-specific genes; (3) Identify potential *cis*-elements present within promoters of grass cell wall biosynthesis genes to further explore the incorporation of grass-expanded genes into cell wall biosynthesis regulatory pathways.

We addressed the first question in Chapter 2 and analyzed the phylogeny of one of the major cell wall associated protein families, R2R3 MYB, across dicots and grasses. The R2R3 MYB proteins have been designated into 48 subgroups and most R2R3 MYBs that regulate SCW in Arabidopsis are likely conserved in the grasses. Besides three dicot-specific subgroups, we identified six grass-specific and two panicoid grass-expanded subgroups. The grass-specific or –expanded R2R3 MYBs are promising candidates for control of cell wall biosynthesis that have yet to be examined. In addition, we observed uncertainty in infering orthologs across dicot and grass species, resulting from ambiguous phylogeny and relatively high false negative rate using Inparanoid and OrthoMCL. With the accumulation of high-quality annotated genomes in grass species, local synteny may further improve the power of comparative phylogenetic and genomic studies.

In Chapter 3, we reported a novel, high-depth and -quality gene network, RCR, and predicted 96 transcription factors. Reverse genetics of a co-ortholog of the Arabidopsis MYB61 transcription factor in rice revealed that OsMYB61a can directly regulate grass-specific cell wall genes. Whether this is due to a direct interaction between MYB61a and the promoters of the grass-specific genes was not resolved. In my previous studies, the large-scale validation of predicted novel cell wall transcription factors was limited for the following three reasons: (a) As a critical process for plant development, cell wall biosynthesis pathways will be affected by numerous signals, such as hormone, biotic and abiotic stress, which will complicate functional examination of individual cell wall regulators in plants; (b) Multiple transcription factors participate in the regulation of cell wall biosynthesis and a few have been known to be functionally redundant. Thus, knockout mutants may have limited role to examine gene function; (c) Rice reverse genetics is still a time- and labor-intensive process due to the difficulty of rice tissue culture and relatively long growth-cycle of many rice cultivars.

In addition to predicting the potential *cis*-elements significantly enriched within the promoters of rice cell wall genes, we also aim to explore how grass-expanded genes have been incorporated into cell wall biosynthesis pathways in Chapter 4. In addition to novel ones, we observed known cell wall associated *cis*-elements within the promoters of CESA, lignin biosynthesis genes and grass cell wall-specific genes. This suggests that known dicot cell wall-associated *cis*-elements may be conserved in grasses. Moreover, grass-specific genes have been incorporated into cell walls *via* maintaining or evolving known cell wall-associated DNA motifs. However, functional validation is

still required to further confirm the effect of discovered *cis*-elements on the interactions between rice cell wall transcription factors and biosynthesis genes in *vitro* and in *vivo*.

Based on this information, I suggest that integrating the phylogenomics, gene network and promoter analysis together would promote the systematical examination of the function of predicted cell wall-associated transcription factors in rice.

Gaining an overall picture of grass cell wall-associated regulators

Orthologs of known Arabidopsis cell wall core transcription factors may maintain similar function in rice. A few members of Arabidopsis NAC transcription factors are critical to activate secondary cell wall biosynthesis, such as AtNST1/2 and AtSND1 (Zhong et al., 2007; Zhong et al., 2010; Zhou et al., 2014). Zhong et al. (2011) overexpressed rice and maize SCW related NAC transcription factors in Arabidopsis, which are able to activate Arabidopsis CESA and lignin biosynthesis genes. Another example is AtMYB4, which is a cell wall repressor in Arabidopsis and conserved in grasses. Its orthologs in switchgrass, PvMYB4a-e, can bind to repeats of the AC-elements based on yeast-one-hybrid, which are known cell wall-MYB binding sites in dicots (Shen et al., 2012). In all, these studies suggest that orthologs of Arabidopsis known cell wall-associated transcription factors may maintain similar functions in grasses.

Despite possessing common regulators, rice and Arabidopsis may prefer different predominant regulators since known dicot cell wall transcription factors show different relative expression levels and different network connectivity compared with their rice orthologs. For example, AtMYB46 is one of the core activators of SCW

biosynthesis in Arabidopsis controlling both downstream transcription factors and cell wall biosynthesis genes (Zhong et al., 2007; Ko et al., 2009; Kim et al., 2012; Ko et al., 2012; Kim et al., 2014). Among all known Arabidopsis cell wall transcription factors, AtMYB46 shows relatively high expression level and performs as a network hub gene connecting with cell wall genes. In contrast, its orthologs in rice, OsMYB46 expresses lowly during rice development and has few connections with rice cell wall biosynthesis genes. This suggests that OsMYB46 may not play a major role in activation of rice cell wall biosynthesis. In addition to OsMYB46, our rice network analysis reveals a series of potential rice cell wall major regulators with different features compared to their Arabidopsis orthologs, such as OsSND2, OsSND3, OsMYB61a, OsMYB61b and OsNST1/2. These may be promising targets for further study to elucidate the overall picture of grass cell wall biosynthesis and regulation in rice.

We expect additional novel transcription factors from the known cell wall-associated protein families (e.g R2R3 MYB and NAC), especially grass-expanded members. After divergence with dicots, the ancestor of grass genomes underwent whole genome duplication (WGD), which may contribute to the expansion of different gene families (De Smet and Van de Peer, 2012; De Smet et al., 2013; Hollister, 2014; Geiser et al., 2016). Comparative phylogenetic study of the R2R3 MYB family reveals six grass-specific, and two panicoid grass-expanded clades. For example, we observed a grass-expanded clade related to the one containing AtMYB61. Molecular genetic analysis of one member, namely Secondary Wall Associated MYB 1 (SWAM1), from this clade shows that it can activate secondary cell walls biosynthesis in *Brachypodium*. Our further phylogenetic studies suggest that Brassicaceae may have lost SWAM1

homologs (Handakumbura et al., Under Revison). Based on the experience analyzing the phylogeny of R2R3 MYB proteins, we think transcription factors may function as grass cell wall regulators if they possess the following features: (1). Belong to a grass-expanded clade neighboring to known cell wall-related transcription factors; (2). Be highly expressed in rice stems, leaves or root, during periods of active synthesis of cell walls.

Members of transcription factors from AP2/ERF, C2C2, C2H2 and homeodomain families are also novel candidates for control of grass cell wall biosynthesis. The novel high-depth and -quality rice genome-scale network, RCR, expanded our understanding of potential families involved in regulation of cell wall biosynthesis and promoted the identification of novel cell wall-related transcription factors from relatively under-studied protein families. We predicted 96 putative novel rice cell wall-associated transcription factors from 19 protein families. This expands our perspectives for study of regulation of grass cell wall biosynthesis. On the other hand, promoter analysis of cell wall-related genes suggests transcription factors from relative undetermined protein families may directly bind to the promoters of cell wall genes, such as ERF, C2H2, C2C2 and bHLH. I suggest a focus for future studies on transcription factors from protein families covered in the rice network that also have evidence of *cis*-elements in grass cell wall-related genes. Candidates should be ranked based on their expression level during rice development and network connectivity with known cell wall-related genes.

Large-scale screening of novel cell wall-associated transcription factors

Molecular genetics based gene function characterization is a labor-intensive process in rice due to the relatively long plant growth cycle and tedious process to generate appropriate mutant lines. Thus, it is critical to screen candidates using rapid assays before careful characterization of transgenic plants. Based on my previous experience working on rice cell wall transcription factors, yeast 1-hybrid (Y1H) and DEX-inducible transient gene expression systems seem to be effective strategies to identify potential direct interactions between transcription factors and downstream targets.

Enhanced Yeast 1-Hybrid (eY1H) assays have been developed to screen individual transcription factors directly for binding to DNA baits using a robotic mating platform with a set of improved Y1H reagents and automated readout quantification (Gaudinier et al., 2011; Reece-Hoyes et al., 2011). This platform has been used to screen transcription factors related to root development and Arabidopsis xylem SCW biosynthesis (Taylor-Teeples et al., 2014). To obtain an overall picture of regulators controlling rice cell wall biosynthesis, I suggest screening transcription factor candidates against promoters of rice cell wall biosynthesis genes based on the following steps: (1) Test the effects on 1 kb native promoters of CESA, lignin biosynthesis genes, CSLF, CSLH and cell wall-related "Mitchell Clade" BAHD-ATs. In addition, gene families involve in phenylpropanoid pathways usually contain different members. Only promoters of predominant members with the highest expression should be selected. (2) Test the interactions between transcription factors and native promoters with mutated DNA binding sites. Promoter analysis reveals potential cell wall-associated DNA binding site that may interact with different transcription factor families. Examination

of mutated promoters can provide additional information to understand the regulatory mechanism of grass cell wall biosynthesis. (3) To further identify the interactions between transcription factor and DNA binding sites by examining the binding of transcription factor candidates on both normal and mutated *cis*-element repeating elements.

Dexamethasone (DEX) inducible transient gene expression system may assist the identification of novel cell wall-associated transcription factors in rice. Protoplast-based transient gene expression assays have been developed in different plant species. For example, Para et al. tested direct targets for nitrogen response transcription factors and their rapid and dynamic response under nitrogen induction (Shen et al., 2013b). I suggest transiently overexpressing target transcription factors fused with the CDS of glucocorticoid receptor (GR). The inducible system provides reasonable controls to explore the expression change of cell wall genes with and without DEX induction. With the additional treatment of cyclohexamide (CHX), a protein biosynthesis inhibitor in eukaryotic organisms, direct targets of transcription factors can be identified. We observed relatively small and variable signals in the transient gene expression screening and preliminary runs of DEX inducible assays in rice protoplast. Thus, if possible, I would suggest importing a GFP marker to the GR overexpression vector, which will allow us to sort protoplast and only collect transformed cells. This may control the gene expression background and provide relatively robust and significant signals.

Exploring the repression of cell wall biosynthesis

So far, few cell wall-associated repressors have been characterized in plants and it is relatively unknown how to "switch off" cell wall biosynthesis pathways. In Arabidopsis, AtMYB4 and AtMYB32 are paralogs that repress cell wall biosynthesis (Jin et al., 2000; Preston et al., 2004). Interestingly, they can recognize the same binding sites as R2R3 MYB cell wall activators, namely, AC-elements. Recent studies show that orthologs of AtMYB4 and AtMYB32 may also behave as cell wall repressors maintaining similar binding affinities in switchgrass and maize genomes (Sonbol et al., 2009; Shen et al., 2012). On the other hand, the ortholog of AtMYB4 in grape may be able to repress flavonoid biosynthesis pathway by forming dimers with known activators involved in flavonoid pathway (ref). This suggests the following two models to shut down cell wall biosynthesis:

Model I: Repressors may recognize similar DNA binding sites as activators and switch off cell wall biosynthesis by competing for DNA binding sites.

Model II: Repressors may form dimers with activators to block their DNA binding ability.

To test model I, I propose use of electrophoretic mobility shift assay (EMSA) to examine competition between known cell wall-associated R2R3 MYB activators and repressors. In rice, the two putative co-orthologs of AtMYB4 and AtMYb32 are highly expressed in most tissues during rice development, and we name then OsMYB4/32a and OsMYB4/32b. Firstly, I suggest examining the effect of OsMYB4/32a and

OsMYB4/32b on rice cell wall biosynthesis and selecting targets for further testing. Then, use EMSA to compare the binding affinity of repressors and selected activators with promoters of cell wall biosynthesis genes.

Yeast 2-hybrid (Y2H) can be used as the initial screen to test model II and examine the interactions between OsMYB4/32a and OsMYB4/32b with characterized rice cell wall activators. Since multiple transcription factors from different protein families are known as cell wall activators, select predominant activators that can regulate multiple rice cell wall biosynthesis genes.

Biomass engineering to improve biofuel production

Biomass is an abundant and sustainable resources produced in the U.S. with an estimated annual production of 1.3 billion tons (Perlack et al., 2005; Binod et al., 2010; Childs et al., 2012; Feltus and Vandenbrink, 2012). In recent years, the bioenergy grasses, members of the Poaceae family, have attracted academic and industrial interests, including the following five C4 photosynthesis species: *Zea mays* (maize); *Saccharum* spp. (sugarcane); *Sorghum bicolor* (sorghum); *Miscanthus* spp. (Miscanthus); and *Panicum virgatum* (switchgrass). These species exhibit great potential to produce biomass, especially in the rural areas. Manipulating cell wall-associated transcription factors, especially repressors, is a potentially effective approach to alter cell wall components. For example, overexpression of cell wall repressor, PvMYB4, can increase the cellulosic ethanol yield from switchgrass by 2.6-fold due to the decrease of lignin deposition (Shen et al., 2013a). Besides total lignin content, the ratio of S and G monolignols and hemicellulose content may also directly affect cell

wall recalcitrance. In all, the manipulation of transcription factors can shed lights on the understanding of cell wall recalcitrance and further engineering bioenergy crops to selectively accumulate cellulose while repressing lignin.

References

Binod, P., Sindhu, R., Singhania, R.R., Vikram, S., Devi, L., Nagalakshmi, S., Kurien, N., Sukumaran, R.K., and Pandey, A. (2010). Bioethanol production from rice straw: An overview. Bioresource Technology 101, 4767-4774.

Childs, K.L., Konganti, K., and Buell, C.R. (2012). The Biofuel Feedstock Genomics Resource: a web-based portal and database to enable functional genomics of plant biofuel feedstock species. Database 2012.

De Smet, R., and Van de Peer, Y. (2012). Redundancy and rewiring of genetic networks following genome-wide duplication events. Current Opinion in Plant Biology 15, 168-176.

De Smet, R., Adams, K.L., Vandepoele, K., Van Montagu, M.C.E., Maere, S., and Van de Peer, Y. (2013). Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. Proc. Natl. Acad. Sci. 110, 2898-2903.

Feltus, F., and Vandenbrink, J. (2012). Bioenergy grass feedstock: current options and prospects for trait improvement using emerging genetic, genomic, and systems biology toolkits. Biotechnology for Biofuels 5, 80.

Gaudinier, A., Zhang, L., Reece-Hoyes, J.S., Taylor-Teeples, M., Pu, L., Liu, Z., Breton, G., Pruneda-Paz, J.L., Kim, D., Kay, S.A., Walhout, A.J.M., Ware, D., and Brady, S.M. (2011). Enhanced Y1H assays for Arabidopsis. Nat Meth 8, 1053-1055.

Geiser, C., Mandáková, T., Arrigo, N., Lysak, M.A., and Parisod, C. (2016). Repeated Whole-Genome Duplication, Karyotype Reshuffling, and Biased Retention of Stress-Responding Genes in Buckler Mustard. The Plant Cell 28, 17-27.

Hollister, J.D. (2014). Polyploidy: adaptation to the genomic environment. New Phytologist, n/a-n/a.

Jin, H., Cominelli, E., Bailey, P., Parr, A., Mehrtens, F., Jones, J., Tonelli, C., Weisshaar, B., and Martin, C. (2000). Transcriptional repression by AtMYB4 controls production of UV-protecting sunscreens in Arabidopsis. The EMBO journal 19, 6150-6161.

Kim, W.-C., Ko, J.-H., and Han, K.-H. (2012). Identification of a cis-acting regulatory motif recognized by MYB46, a master transcriptional regulator of secondary wall biosynthesis. Plant molecular biology, 1-13.

Kim, W.-C., Kim, J.-Y., Ko, J.-H., Kang, H., and Han, K.-H. (2014). Identification of direct targets of transcription factor MYB46 provides insights into the transcriptional regulation of secondary wall biosynthesis. Plant Molecular Biology 85, 589-599.

Ko, J.-H., Kim, W.-C., Kim, J.-Y., Ahn, S.-J., and Han, K.-H. (2012). MYB46-Mediated Transcriptional Regulation of Secondary Wall Biosynthesis. Mol. Plant 5, 961-963.

Ko, J.H., Kim, W.C., and Han, K.H. (2009). Ectopic expression of MYB46 identifies transcriptional regulatory genes involved in secondary wall biosynthesis in Arabidopsis. Plant J 60, 649-665.

Perlack, R.D., Wright, L.L., Turhollow, A.F., Graham, R.L., Stokes, B.J., and Erbach, D.C. (2005). Biomass as feedstock for a bioenergy and bioproducts industry: the technical feasibility of a billion-ton annual supply (DTIC Document).

Preston, J., Wheeler, J., Heazlewood, J., Li, S.F., and Parish, R.W. (2004). AtMYB32 is required for normal pollen development in Arabidopsis thaliana. Plant J 40, 979-995.

Pubudu P. Handakumbura, K.B., Ian P. Whitney, Kangmei Zhao, Karen A. Sanguinet, Scott J. Lee, Michael J. Harrington, Michael T. Veling, Laura E. Bartley, Samuel P. Hazen. (Under Revison). Brachypodium distachyon SWAM1 is a positive regulator of secondary cell wall synthesis and biofuel feedstock attributes and is not found in the Brassicaceae. Plant Physiol.

Reece-Hoyes, J.S., Diallo, A., Lajoie, B., Kent, A., Shrestha, S., Kadreppa, S., Pesyna, C., Dekker, J., Myers, C.L., and Walhout, A.J.M. (2011). Enhanced yeast one-hybrid assays for high-throughput gene-centered regulatory network mapping. Nat Meth 8, 1059-1064.

Shen, H., He, X., Poovaiah, C.R., Wuddineh, W.A., Ma, J., Mann, D.G., Wang, H., Jackson, L., Tang, Y., and Neal Stewart Jr, C. (2012). Functional characterization of the switchgrass (Panicum virgatum) R2R3-MYB transcription factor PvMYB4 for improvement of lignocellulosic feedstocks. New Phytologist.

Shen, H., Poovaiah, C., Ziebell, A., Tschaplinski, T., Pattathil, S., Gjersing, E., Engle, N., Katahira, R., Pu, Y., Sykes, R., Chen, F., Ragauskas, A., Mielenz, J., Hahn, M., Davis, M., Stewart, C.N., and Dixon, R. (2013a). Enhanced characteristics of genetically modified switchgrass (Panicum virgatum L.) for high biofuel production. Biotechnology for Biofuels 6, 71.

Shen, H., Mazarei, M., Hisano, H., Escamilla-Trevino, L., Fu, C., Pu, Y., Rudis, M.R., Tang, Y., Xiao, X., Jackson, L., Li, G., Hernandez, T., Chen, F., Ragauskas, A.J., Stewart, C.N., Wang, Z.-Y., and Dixon, R.A. (2013b). A Genomics Approach to Deciphering Lignin Biosynthesis in Switchgrass. The Plant Cell Online 25, 4342-4361.

Sonbol, F.-M., Fornalé, S., Capellades, M., Encina, A., Tourino, S., Torres, J.-L., Rovira, P., Ruel, K., Puigdomenech, P., and Rigau, J. (2009). The maize ZmMYB42 represses the phenylpropanoid pathway and affects the cell wall structure, composition and degradability in Arabidopsis thaliana. Plant Mol. Biol. 70, 283-296.

Taylor-Teeples, M., Lin, L., de Lucas, M., Turco, G., Toal, T.W., Gaudinier, A., Young, N.F., Trabucco, G.M., Veling, M.T., Lamothe, R., Handakumbura, P.P., Xiong, G., Wang, C., Corwin, J., Tsoukalas, A., Zhang, L., Ware, D., Pauly, M., Kliebenstein, D.J., Dehesh, K., Tagkopoulos, I., Breton, G., Pruneda-Paz, J.L., Ahnert, S.E., Kay, S.A., Hazen, S.P., and Brady, S.M. (2014). An Arabidopsis gene regulatory network for secondary cell wall synthesis. Nature advance online publication.

Zhong, R., Richardson, E.A., and Ye, Z.H. (2007). The MYB46 transcription factor is a direct target of SND1 and regulates secondary wall biosynthesis in Arabidopsis. Plant Cell 19, 2776-2792.

Zhong, R., Lee, C., and Ye, Z.H. (2010). Functional characterization of poplar wood-associated NAC domain transcription factors. Plant Physiol 152, 1044-1055.

Zhou, J., Zhong, R., and Ye, Z.-H. (2014). Arabidopsis NAC Domain Proteins, VND1 to VND5, Are Transcriptional Regulators of Secondary Wall Biosynthesis in Vessels. PLoS ONE 9, e105726.

Supporting information

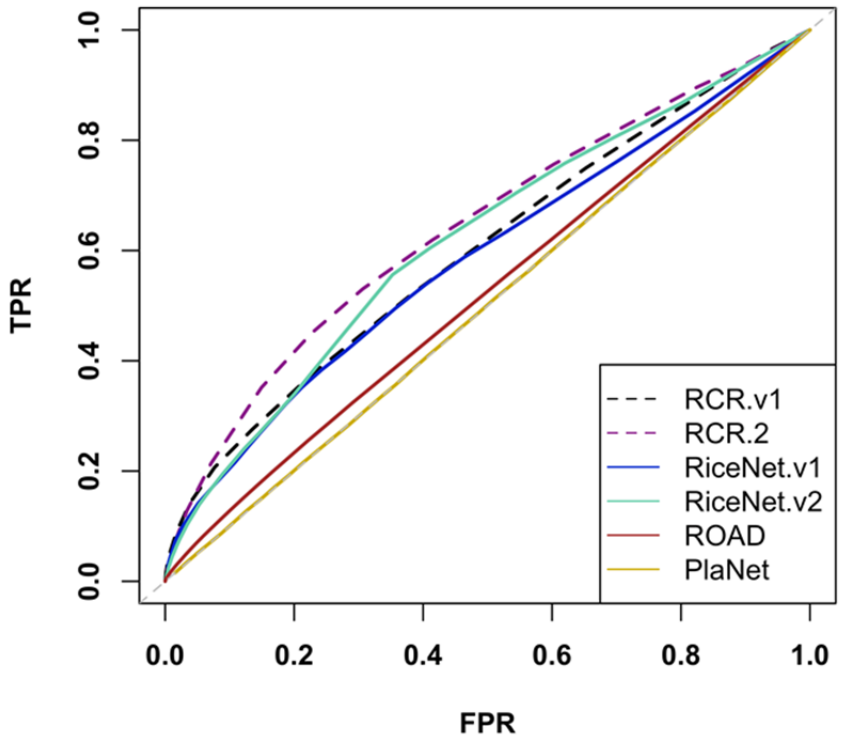
Supporting Figure 2-1 Neighbor-joining tree of R2R3 MYB family proteins from Arabidopsis, poplar, rice, maize and switchgrass with 500 bootstraps in .PNG format [84].

Supporting Table 2-1 R2R3 MYB protein sequences and names from Arabidopsis, poplar, rice, maize and switchgrass.

Supporting Table 2-2 C-terminal motif analysis of R2R3 MYB protein in designated subgroups.

Supporting Figure 2-1 and Supporting Tables 2-1 and 2-1 are not included in this dissertation duo to their large sizes, which can be accessed by the link:

http://bmcplantbiol.biomedcentral.com/articles/10.1186/1471-2229-14-135#Sec18.



Supporting Figure 3-1 Receiver operating characteristic curve (ROC) to plot True Positive Rate (TPR) vs. False Positive Rate (FPR) of Gene Ontology terms (biological process) based network quality evaluation.The grey dashed diagonal represents random prediction. False Positive Rate (FPR) = 1- specificity. The grey dashed diagonal represents random prediction.

Supporting Table 3-1 Rice cell wall network seed genes list.

Locus ID	Name
LOC_Os07g09050	OsGT47
LOC_Os01g70200	OsIRX10
LOC_Os06g27560	OsXAX1L
LOC_Os02g22380	OsXAX1
LOC_Os07g49370	OsIRX9
LOC_Os01g48440	OsIRX9l
LOC_Os06g47340	OsIRX14
LOC_Os08g06380	CslF6
LOC_Os07g36630	CslF8
LOC_Os10g20090	CslH1
LOC_Os05g08370	CESA1
LOC_Os03g59340	CESA2
LOC_Os07g24190	CESA3
LOC_Os01g54620	CESA4
LOC_Os03g62090	CESA5
LOC_Os07g14850	CESA6
LOC_Os10g32980	CESA7
LOC_Os07g10770	CESA8
LOC_Os09g25490	CESA9
LOC_Os06g39970	CESA11
LOC_Os12g29300	CESA10
LOC_Os10g42800	4CL
LOC_Os08g04770	4CL
LOC_Os07g17970	4CL
LOC_Os04g24530	4CL
LOC_Os03g05780	4CL
LOC_Os03g04000	4CL
LOC_Os01g67540	4CL
LOC_Os01g67530	4CL
LOC_Os08g14760	4CL1
LOC_Os02g46970	4CL2
LOC_Os02g08100	4CL3
LOC_Os06g44620	4CL4
LOC_Os08g34790	4CL5
LOC_Os01g42880	AT1
LOC_Os06g39390	AT10
LOC_Os04g11810	AT11
LOC_Os04g09590	AT12

LOC_Os04g09260	AT13
LOC_Os10g01930	AT14
LOC_Os10g01920	AT15
LOC_Os10g02000	AT16
LOC_Os10g01800	AT17
LOC_Os10g03360	AT18
LOC_Os10g03390	AT19
LOC_Os01g42870	AT2
LOC_Os06g48560	AT20
LOC_Os05g04584	AT3
LOC_Os01g18744	AT4
LOC_Os05g19910	AT5
LOC_Os01g08380	AT6
LOC_Os05g08640	AT7
LOC_Os06g39470	AT8
LOC_Os01g09010	AT9
LOC_Os05g41440	C3H
LOC_Os01g60450	C4H
LOC_Os05g25640	C4H
LOC_Os10g11810	CAD1
LOC_Os02g09490	CAD2
LOC_Os10g29470	CAD3
LOC_Os11g40690	CAD4
LOC_Os08g16910	CAD5
LOC_Os04g15920	CAD6
LOC_Os04g52280	CAD7
LOC_Os09g23530	CAD8
LOC_Os03g12270	CAD9
LOC_Os06g06980	CCoAOMT
LOC_Os08g05790	CCoAOMT
LOC_Os08g38900	CCoAOMT
LOC_Os08g38910	CCoAOMT
LOC_Os08g38920	CCoAOMT
LOC_Os09g30360	CCoAOMT
LOC_Os09g25150	CCR
LOC_Os03g60380	CCR
LOC_Os08g34280	CCR1
LOC_Os06g41810	CCRlike
LOC_Os06g41840	CCRlike
LOC_Os02g57760	COMT
LOC_Os04g01470	COMT

LOC_Os12g09770	COMT
LOC_Os12g13800	COMT
LOC_Os08g06100	COMT1
LOC_Os03g02180	F5H
LOC_Os10g36848	F5H
LOC_Os06g24180	F5H
LOC_Os02g39850	HCT
LOC_Os04g42250	HCT
LOC_Os11g31090	HCT
LOC_Os02g41630	PAL1
LOC_Os02g41650	PAL2
LOC_Os02g41670	PAL3
LOC_Os02g41680	PAL
LOC_Os04g43760	PAL5
LOC_Os04g43800	PAL6
LOC_Os05g35290	PAL7
LOC_Os11g48110	PAL8
LOC_Os12g33610	PAL9
LOC_Os08g05520	OsMYB103
LOC_Os08g33150	OsMYB20/43a
LOC_Os09g23620	OsMYB20/43b
LOC_Os01g51260	OsMYB26
LOC_Os09g36730	OsMYB4/32b
LOC_Os08g43550	OsMYB4/32
LOC_Os09g36250	OsMYB42/85
LOC_Os12g33070	OsMYB46/83
LOC_Os03g51110	OsMYB52/54
LOC_Os02g46780	OsMYB58/63a
LOC_Os04g50770	OsMYB58/63b
LOC_Os01g18240	OsMYB61a
LOC_Os05g04820	OsMYB61b
LOC_Os11g10130	OsMYB69
LOC_Os05g50080	OsC3H14
LOC_Os03g03164	OsKNAT7
LOC_Os03g51690	OsKNOTTED1
LOC_Os08g02300	OsNST1/2
LOC_Os06g04090	OsSND1
LOC_Os06g40150	OsSHN2
LOC_Os01g48130	OsSND2
LOC_Os01g09550	OsSND3
LOC_Os03g03540	OsVND1

LOC_Os10g38834	OsVND2
LOC_Os02g42970	OsVND4/5
LOC_Os06g01480	OsVND6/7
LOC_Os05g35170	VNI1
LOC_Os11g03300	VNI2

Supporting Table 3-2 Summary of primers used to genotype myb61a mutants and clone transcription factors in this analysis.

(A) Primers to genotype *myb61a* mutant line.

Line	Locus ID	5' primer	3' primer	primer pair for T-DNA::plant junction
2D-10906	Os01g18240	AAACTGGGGGCATCACAGTG	TGCTGGAGCACCGTTCCAA	5' primer / L0.5

(B) Primers to clone rice transcription factors.

MSU Locus ID	Name	5' Primer sequence	3' Primer sequence
LOC_Os01g11910	bHLH37	CACCATGGGCTCTGCTCCGTTCGG	CACCTTGATCTGCATGTCCTTTGG
LOC_Os01g39330	bHLH24	CACCATGGATGAGGTGTGGTGCAG	CACTCACTAATCAGCAATCA
LOC_Os03g08470	AP2	CACCATGTGTGGAGGCGCCATCCT	GCATCAATGGAGGAGTACA
LOC_Os04g08060	C2H2	CACCATGGGCGTCCAGGAGGAGG	TCTGTCGCTTGTGCTGCTAG
LOC_Os06g43860	KNOX	CACCATGGCGTTCCACTACCAGGAC	CTCCAACCACGTTCTACCAAG
LOC_Os06g46270	NAC1	CACCTCGATCCAGCATCTCAAGG	TCAACTGAGTGAGTTCCACA
LOC_Os07g48550	NAC87	CACCATGTCGGAGTCGGAGGTGTC	TCACCACAAACTCCATCATCA
LOC_Os10g39030	BLH10G	CACCCATCTCTCCTCTCCTCCACCT	CCCGTCAATCACATCAACCT
LOC_Os12g43950	BLH12G	CACCTGTCCTGCTGGAATCATCAA	TACAAATCCCATGCCCTTTC
LOC_Os01g48130	OsSND2	CTGCACGATCTCTCGTCTT	ATATACCTGCCCTGCCCTCT
LOC_Os04g50770	OsMYB58a	AGAGCAACACGAGCAAGGAG	TGGGGTTACTCGTGATGACA
LOC_Os02g41510	OsMYB13a	AGGAGGAAAGGTCGGCAATG	AAGAATAGTGGTGGTAAGAA
LOC_Os04g43680	OsMYB13b	GCAAGAGGAGCAGAGCAGTT	TGATTCGCTCATGGACACTC
LOC_Os05g04820	OsMYB61b	ATGGGGAGGCATTCTTGCTGCTAC	TAGCATTGCACCTAGATATGTTC
LOC_Os01g18240	OsMYB61a	TCTGCCATAAGCTTCCATC	TTCATGTGGTGCTCTGTTCC

Supporting Table 3-3 Summary of qPCR primers.

Locus	Name	5' qPCR primer sequence	3' qPCR primer sequence
LOC_Os01g18240	OsMYB61a	AGTAGCAGTACAGGGAGTAGTG	CCTCTAGTTGTGCTTGGCTAT
LOC_Os05g04820	OsMYB61b	ATGCAGAACCAGAGCCAATC	CTGGAACCAAGAGGCACATATC
LOC_Os09g36250	OsMYB42/85	GCTGCTCGACTACCAAGATTT	GCTAGAAGTTGGACCCATTTGA
LOC_Os08g02300	OsNST1/2	TCGACGATGATGGTGTTGATC	TCTCCTACCACGCCGTAC
LOC_Os02g46780	OsMYB58/63a	GATCACGTCCGTCCGATTT	GCAGTGGCCGATCTATCTT
LOC_Os04g50770	OsMYB58/63b	CTACGTGTAAGCTCGTCGTAAT	ATTCGGTCGAGTTGGGTAATC
LOC_Os12g33070	OsMYB46/83	AGCAGCAGCAACATGATCTA	CCACTCTCCCATGACATTCC
LOC_Os08g05520	OsMYB103	CAGTGGGATGAAGAAGAAGCA	GAGGGAATGAGCACCACTTT
LOC_Os03g51690	OsKNOTTED1	CAATGAGGTGGTTGTGGTTATG	CACTTAATTAGCAGCAGCAAGAG
LOC_Os08g43550	OsMYB4/32a	AAATGATCCTGCAAACCAATCC	GGACTAACACTAAGCAGTAACCA
LOC_Os02g41510	OsMYB13.a	ACAACACCACGGACAGTTT	GTGCTGTCCACCGTCATT
LOC_Os04g43680	OsMYB13.b	GCTTCCGAGGAGTTCCAGAT	GCTCCATGGACACGTCGTA
LOC_Os03g03164	KNAT7	CCACTGGCATATGCGTTGTA	GGTACACCCACGACAACAAA
LOC_Os12g43950	BLH12G	GCCTGAAGACAGCAGAGAAA	GGCATGTTATTCTCCCACTAGG
LOC_Os10g39030	BLH10G	CAAACAGGTGGTAGACCGAT	GCGCGAAATTGGACATTGA
LOC_Os07g48550	NAC87	TGGAAGGCCTGACCAGATTA	TCCTAGTTAACCCAAACCTTTGAC
LOC_Os06g46270	NAC1	CACCAAGTTTGAAGGCAATGT	TTGTGTGAAGCCACTCTCAG
LOC_Os06g43860	KNOXII	GAGATACTCCGAAAGCGAAGAG	AGTTGGGTATGGCCATTTAGAG
LOC_Os04g08060	C2H2	GCCTCCTCGAGCTGGATCT	TGACGGCGAAGGCGAAC
LOC_Os03g08470	AP2	CGACCCGTTCATGCTGTT	GGCGTTCATGTCGGTGTT
LOC_Os01g39330	bHLH24	GAATGCCTCAGGTGGATGAA	CACGATCTGGAGCTTGATGT
LOC_Os01g11910	bHLH37	GTCTCCACGTTCGAGTCATC	AATCACTGCAACCGATCAGA
LOC_Os02g41630	PAL1	CACCCGAGCAGAAACTCCTTCCAC	GTGGCTTGGTCAGATTGGTT
LOC_Os08g06100	COMT1	CACCAAGTAGGTGATTGGTGATCG	GCAAACGACATGATTACCCACGCT

LOC_Os02g08100	4CL3	GGAGACATCGGCTTCGTC	GGTGATTTCTGAGCCTTCTC
LOC_Os08g34280	CCR1	GCCGAGAAGACGGAGGAGGA	ACAGCACGGTAAACACTAACACGA
LOC_Os03g02180	F5H	CGGCAAGAATGGTTTGCAGGTGAT	AGGAAGCTAAGCTAGCCACACCCAA
LOC_Os02g09490	CAD2	TCCGCTACCGCTTCGTCGTCGA	AACGGATCATACGGATGCTGCCATGA
LOC_Os05g41440	C3H	ACGTCAACATGATGGAGTCCAACG	AGCAGGCAGCAAAGCAAACTGTTC
LOC_Os05g08370	CESA1	CATGGTGGCGGGTATATCAT	GATCACCCAGATGGAGAAGAAG
LOC_Os03g59340	CESA2	AAGATCGATGCAAGGAGGATG	TTGCCTTCTGGGTGATTGT
LOC_Os07g24190	CESA3	CTGATACCTCCAACCACTCTAC	CCCATGATTCATACCCGTTATTG
LOC_Os01g54620	CESA4	CATCCTGGTGCTCAACCTC	GAAGAACACCTTGCCGAAGA
LOC_Os03g62090	CESA5	GGGTGATTGTCCATCTGTATCC	CAGAGGAGCGAGAAGATTGAAG
LOC_Os07g14850	CESA6	CTGGGTGATTGTCCATCTATATCC	GTGAGAAGATTGAAGCCAGTAGA
LOC_Os10g32980	CESA7	TCATCCTCCACCTCTACCC	CCCAGACGAGGGAGAAGA
LOC_Os07g10770	CESA8	GTTGTTTGGGCAATCCTTCTG	TGCCACATGTTTGGGTATCT
LOC_Os09g25490	CESA9	ACGCCCACCATTGTTGT	CCTTGATGGTGAAGGGATCAA
LOC_Os08g06380	CSLF6	GTGGTGGTGCTCGTCTGGTG	AGCACGGCGGTGATGACG
LOC_Os07g36630	CSLF8	GGTGTTAGTTGGCTGCTA	TGGCGTTCATAGTGTATATCATA
LOC_Os10g20090	CSLH1	ACCGGTGCAAACTTGATCTTA	CCGCTCCAATGCTTCTACTT
LOC_Os07g49370	IRX9	GAGGAATACAACCAGTCGACG	GCCTAGAGCGTAGTTTGGATG
LOC_Os01g70200	IRX10	GATGTTATCCTGAGGAAGCAGAG	CGAGCGAGACCATTCAGTATC
LOC_Os01g18744	AT4	GACTTCAGGGAGGATCCGTACGAG	TAGTACGCGAAGGGTATGACGTGC
LOC_Os05g19910	AT5	CATCACTGCAGTAGCTGAATTGG	GCTTAGGTGGGTTCGGTATCAGC
LOC_Os05g08640	AT7	CGAGGAAGACAAGCTCATCCTGC	TGGCTGAACCTGAACCCCACCA
LOC_Os06g39470	AT8	AGTACCCGCTCATGGTGGAC	AACTGCGTGACCTGGACAA
LOC_Os06g39390	AT10	CTACGAGTCCGTGTACGTGTCGGA	TGAGGTCGCAGTTCACCAGC
LOC_Os01g22490	Ubi5	ACCACTTCGACCGCCACTACT	ACGCCTAAGCCTGCTGGTT
LOC_Os04g35910	Cc55	AAGGAGAAAGCCGAACAACG	TCCTCAAGTTTCTTCCTGTAGGC

Supporting Table 3-4 Summary of novel rice SCW transcription factors.

TF	Family	Cell Wall Network Degree
Os05g03040	AP2	11
Os02g40070	AP2	8
Os02g06910	ARF	6
Os06g46410	ARF	6
Os07g39220	BES1	5
Os01g11910	**bHLH**	10
Os01g39330	**bHLH**	10
Os03g58830	bHLH	10
Os01g72370	bHLH	8
Os08g39630	bHLH	7
Os01g68700	bHLH	5
Os03g20310	bZIP	8
Os04g08060	**C2H2**	18
Os09g27650	C2H2	12
Os04g02510	C2H2	9
Os08g36390	C2H2	7
Os01g68860	C3H	7
Os06g49880	DBB	16
Os02g39360	DBB	9
Os04g41560	DBB	8
Os05g11510	DBB	8
Os06g05890	DBB	6
Os03g38870	Dof	5
Os03g08470	**ERF**	24
Os02g43790	ERF	20
Os04g46400	ERF	16
Os10g41130	ERF	10
Os02g54160	ERF	8
Os04g46440	ERF	8
Os02g43940	ERF	6
Os01g24070	GATA	22
Os10g40810	GATA	8
Os01g54210	GATA	5
Os06g03710	GRAS	5
Os06g02560	GRF	6
Os10g01470	HD-ZIP	13
Os06g04870	HD-ZIP	7

Os03g12860	HD-ZIP	6
Os06g04850	HD-ZIP	6
Os10g41230	HD-ZIP	6
Os10g42490	HD-ZIP	5
Os09g28354	HSF	22
Os06g49840	MIKC	9
Os02g36924	MIKC	6
Os05g46610	MYB	31
Os01g36460	MYB	25
Os05g48010	MYB	19
Os01g16810	MYB	14
Os07g31470	MYB	14
Os11g03440	MYB	11
Os12g03150	MYB	11
Os01g12860	MYB	9
Os03g29614	MYB	9
Os05g49310	MYB	7
Os05g37730	MYB	6
Os11g45740	MYB	6
Os03g20090	MYB	5
Os04g43680	MYB	5
Os04g45020	MYB	5
Os12g13570	MYB	5
Os11g47460	MYB_related	23
Os07g44090	MYB_related	13
Os02g42850	MYB_related	6
Os07g48550	**NAC**	12
Os06g46270	**NAC**	17
Os02g56600	NAC	22
Os10g33760	NAC	21
Os08g01330	NAC	20
Os02g34970	NAC	15
Os03g21060	NAC	15
Os04g43560	NAC	15
Os04g35660	NAC	14
Os12g41680	NAC	12
Os08g10080	NAC	9
Os03g56580	NAC	7
Os04g38720	NAC	7
Os08g06140	NAC	6
Os09g32040	NAC	6

Os03g21030	NAC	5
Os12g43950	**TALE**	38
Os06g43860	**TALE**	20
Os03g06930	TALE	24
Os03g47740	TALE	23
Os08g19650	TALE	20
Os02g08544	TALE	12
Os06g01934	TALE	10
Os02g13310	TALE	5
Os03g52239	TALE	5
Os05g03884	TALE	5
Os12g40570	WRKY	10
Os01g09080	WRKY	8
Os01g43650	WRKY	5
Os01g53040	WRKY	5
Os04g51560	WRKY	5
Os05g09020	WRKY	5
Os11g03420	ZF-HD	5