Functional genomics of NAC transcription factor SND2 regulating secondary cell wall biosynthesis in *Arabidopsis* and *Eucalyptus*

by

Steven Grant Hussey

Submitted in partial fulfilment of the requirements for the degree

Philosophiae Doctor

In the Faculty of Natural and Agricultural Sciences

Department of Genetics

University of Pretoria

Pretoria

May 2014

Supervised by Prof. Alexander A. Myburg

Co-supervised by Prof. Dave K. Berger and Dr. Eshchar Mizrachi

© University of Pretoria

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

Declaration

I, the undersigned, declare that the thesis, which I hereby submit for the degree PhD at the University of Pretoria, is my own work and has not been previously submitted for a degree at this or any other tertiary institution.

____

Steven Grant Hussey

____

Date

© University of Pretoria

TABLE OF CONTENTS

© University of Pretoria

CHAPTER 2

© University of Pretoria

© University of Pretoria

© University of Pretoria

© University of Pretoria

© University of Pretoria

© University of Pretoria

© University of Pretoria

© University of Pretoria

CHAPTER 6

© University of Pretoria

Thesis summary

Functional genomics of NAC transcription factor SND2 regulating secondary cell wall biosynthesis in *Arabidopsis* and *Eucalyptus*

Steven G. Hussey

Supervised by ***Prof A. A. Myburg****,* ***Prof. D. K. Berger and Dr E. Mizrachi***

Submitted in partial fulfilment of the requirements for the degree **Philosophiae Doctor**

Department of Genetics

University of Pretoria

Wood formation is heavily exploited for the manufacturing of pulp, paper, sustainable biomaterials and, potentially, biofuels. *Eucalyptus* is a favourable fast-growing, short rotation plantation crop grown over millions of hectares globally for its superior fiber properties. Understanding the molecular biology of secondary cell wall (SCW) formation in trees, and in particular how it is transcriptionally and epigenetically regulated, lays the foundation for enhanced woody trait improvement strategies in tree biotechnology. Transcriptional networks regulating SCW biosynthesis have been discovered in the herbaceous model plant *Arabidopsis thaliana*, in which NAC domain transcription factors (TFs) play a prominent role. The functions of many NAC domain TFs remain to be resolved, and their regulatory roles and evolution in *Eucalyptus* is unknown. Functional genomics studies of *Eucalyptus* TFs are

© University of Pretoria

currently challenged by a lack of established high-throughput genomics techniques commonly applied to model organisms. In this study, we aimed to better understand NAC family evolution and the epigenetic regulation of xylogenesis in *E. grandis*, and characterize the role of NAC domain TF SND2 in transcriptional regulation of SCW biosynthesis in *A. thaliana* and *E. grandis*.

Comparative genomics and bioinformatics analyses of 189 curated gene models of the *E. grandis* NAC family, one of the largest described to date, revealed extensive tandem duplication in stress response-associated subfamilies, while SCW-associated subfamilies were generally conserved among angiosperms. Novel candidates for wood and tension wood formation as well as cold-stress tolerance were identified from transcriptional profiling in *E. globulus* and *E. grandis*. We identified the phenotypic effects and putative targets of the NAC domain TF SND2 in *A. thaliana* using microarray, microscopy and cell wall chemistry analyses. Moderate *SND2* overexpression upregulated genes involved in cellulose, xylan, mannan, signaling and lignin polymerization processes and affected mannose, rhamnose and lignin components of stem cell walls, while strong overexpression resulted in reduced interfascicular fiber SCW deposition. *SND2* overexpression in *Eucalyptus* somatic xylem sectors increased cross-sectional fiber cell area. We optimized a chromatin immunoprecipitation sequencing (ChIP-seq) approach and applied it to developing secondary xylem of mature *E. grandis* trees to identify the targets of the *E. grandis* ortholog of SND2, EgrNAC170. In validating the approach, we addressed the regulatory role of the epigenetic mark trimethylated lysine 4 of histone H3 (H3K4me3) in this tissue, showing a strong association with expressed loci, occupation of regions close to transcriptional start sites and

© University of Pretoria

tight correlation with transcript abundance, especially that of broadly expressed genes but also genes associated with SCW formation. A pilot study of EgrNAC170 targets was performed using the high-throughput ChIP-seq approach, identifying over 3,000 putative targets in *E. grandis* developing secondary xylem, but showing evidence that further ChIP-seq data are required for reliable target identification.

The results of this thesis contribute to science an understanding of the unique evolution of NAC proteins in *Eucalyptus*, knowledge of the function of SND2/EgrNAC170 as possible candidates for tree biotechnology, the first genomic profile of a histone modification in developing wood and a high-throughput ChIP-seq protocol for the study of native protein-DNA interactions in developing xylem.

© University of Pretoria

Preface

Mankind has relied on woody materials for tens of thousands of years. While still used primarily for timber and firewood in the twenty-first century, trees are now cultivated in plantations of remarkable scale to meet a growing demand for pulp, paper and biomaterials. *Eucalyptus*, a genus of over 700 species mostly native to Australia, has emerged as a favourable short-rotation commercial hardwood with over twenty million hectares grown worldwide in mainly Brazil, China and India. In addition to pulp and paper, the composite biopolymer known as lignocellulose, comprising the bulk of secondary cell walls in woody biomass, has become a promising biological resource for specialized manufactured materials such as nanocrystalline cellulose, cellophane, thickeners and stabilizers, adhesives, sanitary products and cellulosic biofuels. A diversifying market capitalizing on renewable materials derived from lignocellulose, encapsulated by what is generally known as the bio-economy, is creating new opportunities for the genetic improvement of wood and fiber properties.

The molecular basis of secondary cell wall biosynthesis and the genetic regulation of its complex development are not yet well understood. We know from transcriptional profiles of high-resolution sections through developing poplar xylem that gene transcripts are tightly regulated across spatial and developmental gradients. The difficulty in studying the genetics of trees – long generation times, demanding space requirements, often recalcitrant transformation tendencies and large genomes in the case of gymnosperms – has led to the adoption of the herbaceous plant *Arabidopsis thaliana* as the primary model for xylem development. Dozens of genes involved in the biosynthesis of cellulose, hemicellulose and lignin components of the secondary cell wall have been identified over the past few decades, and more recently

© University of Pretoria

transcriptional networks involved in their regulation have been discovered. It is now well established that certain transcription factors are “master regulators” of secondary cell wall deposition in various cell types, sufficient to initiate the cascading transcriptional networks that control secondary cell wall deposition and programmed cell death. The functions of the associated transcription factors have been studied in a range of model organisms including the tree model *Populus trichocarpa*, revealing an evolutionarily conserved mechanism behind the transcriptional regulation of secondary cell wall deposition.

It has been shown in multiple studies of model plants that some transcription factors regulate the deposition of very particular secondary cell wall biopolymers, are expressed in specific cell types or have the potential to alter woody traits when genetically manipulated. Such transcription factor candidates may be evaluated for their biotechnological potential in forest trees through overexpression, dominant repression or RNAi knock-out approaches. However, a more precise and effective strategy for the improvement of forest trees, facilitated by a thorough understanding of the structure and behaviour of transcriptional networks regulating fiber and vessel development, is to enhance desirable traits in relevant cell types through transcriptional network re-engineering. At the time of writing, at least two studies have shown how fiber traits can be specifically modified through transcription network rewiring in *Arabidopsis*. Deciphering the architecture of transcriptional networks in trees, however, is a daunting task that will rely on the discovery and painstaking functional analysis of candidate transcription factors, as well as comprehensive identification of their molecular interactions and gene targets.

© University of Pretoria

Towards these endeavors, this thesis enriches our understanding of the functions and evolution of the NAC domain family of transcription factors in *Eucalyptus grandis*, a prominent group of proteins regulating secondary cell wall deposition, with emphasis on NAC domain protein SND2. I extensively review the literature of the transcriptional regulation of secondary cell wall deposition in Chapter 1, compiling from *Arabidopsis* studies the most comprehensive transcriptional networks of this process yet published, assessing the evolutionary conservation of secondary cell wall transcriptional networks and reflecting on technological advances and shortcomings that influence the study of transcriptional regulation. Chapter 2 is a comparative genomics study of the NAC domain protein family of *E. grandis* that investigates their evolution relative to other angiosperms and identifies known and novel candidates for the transcriptional regulation of xylogenesis, tension wood formation and cold stress response through transcriptional profiling in *E. globulus*. The function and putative gene targets of SND2, a transcription factor previously linked to fiber secondary cell wall regulation in *Arabidopsis*, is further investigated in Chapter 3 through microarray, phenotypic and cell wall chemistry analysis of *SND2*-overexpressing *Arabidopsis* plants as well as hybrid *Eucalyptus* transgenic wood sectors. In Chapter 4 I optimized a chromatin immunoprecipitation sequencing (ChIP-seq) procedure to facilitate the identification of *in planta* genomic targets of DNA-binding proteins in developing secondary xylem of field-grown *E. grandis* trees. The approach was validated through the generation of genome-wide profiles of the activating histone H3K4me3 modification, which I show to be intimately associated with gene expression levels in developing secondary xylem and which plays a role in the epigenetic regulation of secondary cell wall-associated genes. Having firmly cemented a role for SND2 in secondary cell wall transcriptional regulation in Chapter 3, I used the ChIP-

© University of Pretoria

seq approach developed in Chapter 4 for the identification of the direct targets of a putative *E. grandis* ortholog of SND2, EgrNAC170, in secondary developing xylem (Chapter 5). In the concluding remarks of the thesis (Chapter 6), I consolidate the evidence for SND2 in the regulation of secondary cell wall biosynthesis in fibers, discuss the potential of this gene in tree biotechnology, reflect on new questions raised by the research herein and discuss aspects of the epigenetic regulation of secondary cell wall deposition.

Work included in this PhD thesis has resulted in the following research outputs:

Peer-reviewed publications:

Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D, Goodstein DM, Dubchak I, Poliakov A, Mizrachi E, Kullan ARK, van Jaarsveld I, Hussey SG, *et al.* The genome of *Eucalyptus grandis* - a global tree for fiber and energy. *Nature* (in press).

Hussey SG, Mizrachi E, Creux NM and Myburg AA. (2013). Navigating the transcriptional roadmap regulating plant secondary cell wall deposition. *Frontiers in Plant Science* **4**: 325.

Hussey SG, Mizrachi E, Spokevicius AV, Bossinger G, Berger DK and Myburg AA. (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**: 173.

© University of Pretoria

Peer-reviewed published conference proceedings:

Hussey SG, Mizrachi E, Berger DK and Myburg AA. (2011). The role of SND2 in the regulation of *Arabidopsis* fibre secondary cell wall formation. *BMC Proceedings* **5**(Suppl 7): P114.

Botha J, Pinard D, Creux N, Hussey SG, Maritz-Olivier C, Spokevicius A, Bossinger G, Mizrachi E and Myburg AA. (2011). Characterising the role of the *Eucalyptus grandis SND2* promoter in secondary cell wall biosynthesis. *BMC Proceedings* **5**(Suppl 7): P105.

Presentations at South African conferences:

Hussey SG, Groover A, Berger D and Myburg AA. (2012). Development of methods to map binding sites of two *Eucalyptus* transcription factors associated with wood formation using ChIP-seq and RNA-seq. South African Genetics and Bioinformatics Society Conference, Stellenbosch, 10-12 September 2012 (poster).

Hussey SG, Mizrachi E, Ranik M, Creux N and Myburg AA. (2010). Analysis of the SND2 transcription factor from *Arabidopsis* in secondary cell wall biogenesis and its evolutionary conservation in *Eucalyptus*. 21st Biennial Congress of the South African Genetics Society, Bloemfontein, 2010 (oral presentation).

© University of Pretoria

Presentations at international conferences:

Hussey SG, Singh P, Mizrachi E, Berger DK and Myburg AA. (2014). Genome-wide analysis of lysine 4-trimethylated histone H3 modification during xylogenesis in *Eucalyptus*. Plant and Animal Genome XII Conference, San Diego, CA, 11-15 Jan 2014 (poster).

Calvert M, Hussey SG, Mizrachi E and Myburg AA. (2014). Genetic dissection of gene expression variation of secondary cell wall related transcription factors in *Eucalyptus* hybrid populations. Plant and Animal Genome XII Conference, San Diego, CA, 11-15 Jan 2014 (poster).

Hussey SG, Saïdi MN, Hefer CA, Mizrachi E, Calvert M, Myburg AA and Grima-Pettenati J. (2013). The NAC domain family of transcription factors in *Eucalyptus*. IUFRO Tree Biotechnology Conference, Asheville, NC, 26 May – 1 Jun 2013 (poster).

Hussey SG, Mizrachi E, Hefer C, Groover A, Berger D and Myburg AA. 2012. Next-generation genomics tools for reconstructing transcriptional networks in *Eucalyptus*. Plant and Animal Genome XXII Conference, San Diego, CA, 14-18 Jan 2012 (oral presentation).

Hussey SG, Mizrachi E, Berger D, Myburg AA. (2011). The role of SND2 in the regulation of *Arabidopsis* fibre secondary cell wall formation. IUFRO Tree Biotechnology Conference: From Genomes to Integration and Delivery, Arraial d'Ajuda, Bahia, Brazil, 26 Jun – 2 Jul 2011 (poster).

© University of Pretoria

Acknowledgements

I convey my sincere thanks to the following people and organizations:

- My supervisor Prof. Zander Myburg, for his inspiring academic talent, exceptional mentorship, passion and generous financial support.
- My co-supervisor Prof. Dave Berger, for his transformational academic leadership, kind nature and keen attention to detail.
- My co-supervisor Dr. Eshchar Mizrachi, for his dedicated peer review and enlivening drive and passion. Your witty humour, keen intellect and great friendship is much appreciated.
- Prof. Andrew Groover, for his kind accommodation, collaboration and willingness to externally review my PhD upgrade application.
- Sappi, Mondi, the National Research Foundation, the Department of Science and Technology, and the Technologies and Human Resources for Industry Program for their generous research and student support and for recognizing the importance of research in *Eucalyptus* molecular genetics.
- The Department of Genetics and the University of Pretoria for their academic and financial support. Thanks in particular to former and current Heads of Department Prof. Henk Huismans and Prof. Paulette Bloomer for their excellent leadership.
- To all the members of FABI, under the leadership of Prof. Mike Wingfield, for providing a stimulating research environment, a true team atmosphere and for cultivating a diverse and welcoming ethos.
- Dr. Christine Maritz-Olivier, for her selfless troubleshooting advice.
- Dr. Nicky Creux, your friendship and mentorship since I joined FMG has been phenomenal. Thank you for your valuable advice, peer review and willingness to help when I got stuck.
- Marja O’Neil, for her friendship and for keeping FMG afloat through her dedicated administrative work.
- My friends and colleagues in Lab Z and Lab X, both past and present, for the many good memories we’ve shared.
- Karen van der Merwe and Pooja Singh, for their invaluable bioinformatics help. Thanks for putting up with my many requests and editing that script time after time.

© University of Pretoria

- Mmoledi Mphahlele, for his selfless assistance with *Eucalyptus* field trips.
- To our former cloning queen Minique de Castro, for her colourful personality and unfaltering willingness to share a "wyntjie".
- Charles Hefer, for his longsuffering bioinformatics assistance even while pursuing a fresh career abroad.
- The Mandela Rhodes Foundation, for their superb scholarship and personal development program, the opportunity to become part of a Community of driven African leaders, and for the privilege of meeting Madiba in 2010. Hamba Kahle Tata!
- To my beloved fiancé Marius Redelinghuys, for eight years of remarkable love and friendship. You were there when I needed you most, and I would be in a far worse-off space without you.
- My family – Jennifer, Richard, Medea, Michelle, Alan, Maxine and Troy – for their love, financial aid, encouragement and support that was crucial to achieve this final milestone.

© University of Pretoria

Dedicated to my wonderful parents Jennifer and Richard Hussey

and

in loving memory of my grandfather, Prof. Ernest Pereira (Head: Department of English, UNISA)

1930 - 1996

© University of Pretoria

CHAPTER 1

LITERATURE REVIEW

Navigating the transcriptional roadmap regulating plant secondary cell wall deposition

Steven G. Hussey, Eshchar Mizrachi, Nicky M. Creux, Alexander A. Myburg

Department of Genetics, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Private Bag X20, 0028, South Africa

This chapter has been written in manuscript format for submission to a peer-reviewed scientific journal. I conceived of the content and drafted the manuscript. Eshchar Mizrachi, Nicky Creux and Alexander Myburg assisted with draft reviewing and editing, and contributing to parts of Table 1.2. The chapter was accepted and published as a review paper in 2013 (Hussey *et al.*, *Frontiers in Plant Science* **4**:325).

© University of Pretoria

1.1. Summary

The current status of lignocellulosic biomass as an invaluable resource in industry, agriculture and health has spurred increased interest in understanding the transcriptional regulation of secondary cell wall (SCW) biosynthesis. The last decade of research has revealed an extensive network of NAC, MYB and other families of transcription factors regulating *Arabidopsis* SCW biosynthesis, and numerous studies have explored SCW-related transcription factors in other dicots and monocots. Whilst the general structure of the *Arabidopsis* network has been a topic of several reviews, they have not comprehensively represented the detailed protein-DNA and protein-protein interactions described in the literature, and an understanding of network dynamics and functionality has not yet been achieved for SCW formation. Furthermore the methodologies employed in studies of SCW transcriptional regulation have not received much attention, especially in the case of non-model organisms. In this review, we have reconstructed the most exhaustive literature-based network representations to date of SCW transcriptional regulation in *Arabidopsis*. We include a manipulable Cytoscape representation of the *Arabidopsis* SCW transcriptional network to aid in future studies, along with a list of supporting literature for each documented interaction. Amongst other topics, we discuss the various components of the network, its evolutionary conservation in plants, putative modules and dynamic mechanisms that may influence network function, and the approaches that have been employed in network inference. Future research should aim to better understand network function and its response to dynamic perturbations, whilst the development and application of genome-wide approaches such as ChIP-seq and systems genetics are in progress for the study of SCW transcriptional regulation in non-model organisms.

© University of Pretoria

1.2. Introduction

The bulk of plant biomass is comprised of secondary cell walls (SCWs), consisting of a cross-linked matrix of cellulose, hemicellulose and lignin biopolymers. The latter form the basic scaffold of fibers and vessels found in angiosperm xylem. In addition to providing mechanical support, SCWs facilitate critical biological processes, such as water and nutrient transport, anther dehiscence, silique shattering, plant organ movement and response to pathogens (Caño-Delgado *et al.*, 2003; Mitsuda *et al.*, 2005; Fratzl *et al.*, 2008; Mitsuda and Ohme-Takagi, 2008). Candidate genes involved in the biosynthesis of SCWs have been studied in both woody and herbaceous model species (e.g. Brown *et al.*, 2005; Mellerowicz and Sundberg, 2008). These structural genes are under strict transcriptional control during xylogenesis (Hertzberg *et al.*, 2001; Schrader *et al.*, 2004), highlighting the central role of transcription factors in this regard (Du and Groover, 2010). Understanding the regulation of SCW deposition is important because of (1) the widespread use of lignocellulosic biomass in pulp, paper and cellulose-derived products, (2) the potential of second-generation biofuel feedstocks such as short-rotation hardwoods (e.g. *Populus*, *Eucalyptus*) (Rockwood *et al.*, 2008; Carroll and Somerville, 2009; Hinchee *et al.*, 2010), and (3) the role of cell wall material in nutrition and health (Fincher, 2009; Doblin *et al.*, 2010; McCann and Rose, 2010). However, the challenges to studying transcriptional regulation in non-model organisms impede the improvement of lignocellulosic biomass for fiber, raw cellulose and biofuels.

Considerable progress has been made in understanding how TFs regulate SCW structural genes. To this end, various model organisms (*Arabidopsis*, *Oryza*, *Populus*) (e.g. Kubo *et al.*, 2005; Grant *et al.*, 2010; Zhong *et al.*, 2011a) as well as *Zinnia* and *Arabidopsis*

© University of Pretoria

(trans)differentiation systems (Fukuda and Komamine, 1980; Oda *et al.*, 2005) have been instrumental. In the last decade, studies in *Arabidopsis* in particular have revealed the existence of an extensive transcriptional network regulating SCW deposition in vessels, fibers, anther endothecium and structures (replum, endocarp, valve margin) within the silique (reviewed in Yamaguchi and Demura, 2010; Zhong *et al.*, 2010a). Whilst a considerable diversity of TF families participate in SCW transcriptional regulation, the most prominent families of TFs involved in this network appear to be the NAC (NAM/ATAF/CUC) and R2R3-type MYB (MYELOBLASTOSIS) family proteins, both characterized by conserved N-terminal DNA-binding domains and diverse C-termini that participate in transcriptional regulation (Ooka *et al.*, 2003; Dubos *et al.*, 2010).

General structures of SCW transcriptional networks have been illustrated in a number of reviews, based on knowledge of *Arabidopsis* (Umezawa, 2009; Zhong and Ye, 2009; Caño-Delgado *et al.*, 2010; Yamaguchi and Demura, 2010; Zhang *et al.*, 2010; Zhong *et al.*, 2010a; Wang and Dixon, 2011; Zhao and Dixon, 2011; Pimrote *et al.*, 2012; Schuetz *et al.*, 2013), and monocots (Handakumbura and Hazen, 2012). A few primary research articles also depict schematic representations of the *Arabidopsis* SCW network, incorporating data from *Populus* and limited knowledge of *Eucalyptus* and *Pinus* SCW transcriptional networks (Zhong *et al.*, 2008a; McCarthy *et al.*, 2009; Zhong *et al.*, 2011b). However, aside from Umezawa (2009) who focused on the cinnamate/monolignol pathway, these representations have not fully captured individual protein-DNA and protein-protein interactions reported in the literature. In addition, the regulatory dynamics of SCW transcriptional regulation are poorly understood compared to network structure (i.e. connectivity). Furthermore, the methodologies used to

© University of Pretoria

generate evidence lines for SCW network reconstruction have not been extensively reviewed. Here we comprehensively integrate and illustrate the complexity of known protein-DNA and protein-protein interactions in the *Arabidopsis* SCW transcriptional network. We discuss the roles of putative regulatory modules in the network, highlighting known and hypothetical balancing mechanisms that may influence network behaviour. Finally, we provide a critical review of the methodologies currently used to infer SCW transcriptional networks and recommend approaches for increasing reliability in inferring SCW transcriptional network structure.

1.3. Vascular patterning and differentiation

The deposition of SCWs and the initiation of programmed cell death (Bollhöner *et al.*, 2012) together represent the culmination of developmental signals that cue vascular tissue specification and cell fate determination (Fig. 1.1). This specification begins with establishing a population of meristematic cells known as the procambium via the combinatorial effect of hormones such as auxin, cytokinins and brassinosteroids (BRs). The procambium in turn gives rise to the primary vascular tissues (xylem, phloem) in the shoot vascular bundles and root vasculature (Turner *et al.*, 2007; Caño-Delgado *et al.*, 2010). In root and shoot tips, a pre-procambial state is established via PIN1-mediated polar auxin transport along files of parenchyma cells, effectively channeling auxin to what will become the procambium (Dettmer *et al.*, 2009). In leaf veins, a preprocambial state is associated with expression of *ATHB8*, which is directly activated by the auxin response factor MP/ARF5 (reviewed in Zhang *et al.*, 2010). In addition to procambium specification, auxin promotes cell division in the procambium in combination with cytokinins (reviewed in Caño-Delgado *et al.*, 2010). The

© University of Pretoria

vascular cambium, from which all secondary xylem and phloem tissues arise during secondary growth, develops from the procambium and interfascicular parenchyma (Plomion *et al.*, 2001; Baucher *et al.*, 2007). As per the convention of Dettmer *et al.* (2009), we generally refer to procambiums and (secondary) vascular cambiums as vascular meristems, which are thought to be regulated in a similar, but not identical, fashion to shoot and root apical meristems (Sanchez *et al.*, 2012; Milhinhos and Miguel, 2013) (Fig. 1.1).

The establishment of xylem and phloem cell fate is influenced by hormones, TFs, miRNAs, mobile peptides and proteoglycans acting on nascent mother cells produced in the vascular meristems (Fig. 1.1) (see Carlsbecker and Helariutta, 2005; Du and Groover, 2010; Zhang *et al.*, 2010; Schuetz *et al.*, 2013 for review). Auxin concentrations lower than those encountered at the vascular meristem promote xylem differentiation in the presence of cytokinin (Sorce *et al.*, 2013). In the root, xylem differentiation is in contrast thought to be promoted by high auxin concentrations, brought about by cytokinin-mediated activation of a phosphorylation cascade in the procambium that results in polar auxin transport towards the protoxylem (reviewed in Aichinger *et al.*, 2012). Five members of class III homeodomain leucine zipper (HD-ZIP III) TFs, including ATHB8, IFL1/REV, PHB and PHV, are induced by auxin and generally promote xylem differentiation (Zhong *et al.*, 1997; Baima *et al.*, 2001; Ohashi-Ito and Fukuda, 2003; Ilegems *et al.*, 2010; Schuetz *et al.*, 2013). However, some HD-ZIP III genes, such as *ATHB8* and *ATHB15*, appear to be antagonistic to *REV* in meristem formation, embryo patterning and interfascicular fiber development (Prigge *et al.*, 2005). For example, ATHB15 seems to negatively affect xylem development, while miR166-mediated cleavage of *ATHB15* transcript (see below) promotes xylem differentiation (Kim *et al.*, 2005).

© University of Pretoria

Xylogen, a secreted proteoglycan, has also been implicated in xylem specification (Motose *et al.*, 2004), while gibberellic acid (GA) promotes fiber elongation and general xylogenesis (Eriksson *et al.*, 2000; Israelsson *et al.*, 2003; Mauriat and Moritz, 2009). Brassinosteroids (BRs) have been associated with xylem differentiation in *Arabidopsis*, and in transdifferentiating *Zinnia* cell cultures BRs are required for the expression of a homolog of *ATHB8* (reviewed in Jung and Park, 2007). Ethylene is essential for *in vitro* tracheary element (TE) differentiation in cultured *Zinnia* cells (Pesquet and Tuominen, 2011). *In planta*, ethylene is thought to diffuse from its site of synthesis in maturing TEs through to the cambium (Pesquet and Tuominen, 2011), where it promotes cell division (Love *et al.*, 2009).

On the opposite side of the cambium, phloem differentiation occurs under the influence of APL, a MYB-related TF (Bonke *et al.*, 2003; Ilegems *et al.*, 2010), whilst KAN1/KAN2/KAN3/KAN4 TFs indirectly promote phloem differentiation by repressing (pro)cambium maintenance and restricting class III HD-ZIP TF expression through repression of polar auxin transport (Emery *et al.*, 2003; Izhaki and Bowman, 2007; Schuetz *et al.*, 2013). Phloem-expressed miR165/166, which are upregulated by SHR and SCR in roots, post-transcriptionally inhibit HD-ZIP III genes (Tang *et al.*, 2003; Mallory *et al.*, 2004; McHale and Koning, 2004; Zhong and Ye, 2004; Williams *et al.*, 2005; Zhong and Ye, 2007; Carlsbecker *et al.*, 2010). Ectopic xylem formation is inhibited by a dodecapeptide ligand TDIF/CLE41/CLE44, which is produced in the phloem and diffuses to the xylem side of the vascular meristem (Ito *et al.*, 2006). The peptide also co-ordinates the orientation of cell divisions in the cambium via the perception of a peptide concentration gradient by the LRR receptor-like kinase PXY in procambial cell membranes and induction of *WOX4* (Etchells and

© University of Pretoria

Turner, 2010; Hirakawa *et al.*, 2010). Xylem differentiation may be further suppressed in the phloem in part by XIP1, which is related to PXY (Bryan *et al.*, 2011) (Fig. 1.1).

Once xylem mother cell fate has been established and cell elongation has ceased in immature xylem, SCW deposition occurs. This is activated by the TFs VND6 and VND7 in the case of xylem vessels, and SND1 and NST1 in fibers. These "master regulators" initiate a SCW transcriptional network, successively activating at least two tiers of intermediate TFs which, in addition to the master regulators, activate the structural genes for SCW biosynthesis (Fig. 1.1, grey blocks). In the remainder of this review we focus on the SCW transcriptional network and the tools available to study its structure and function.

1.4. The SCW transcriptional network: Structure, evolution, and dynamics

A simplified representation of the SCW regulatory network is shown in Fig. 1.1, which depicts the putative positions of associated TFs and their direct or indirect targets. We have also reconstructed a SCW-regulating protein-DNA and protein-protein interaction network from the *Arabidopsis* literature using BioTapestry (Longabaugh *et al.*, 2005), showing cell type contexts where known (Fig. 1.2). Aside from the indicated exceptions, we represent only direct protein-DNA interactions, as elucidated using yeast one-hybrid, electrophoretic mobility shift assay, chromatin immunoprecipitation, or post-translationally induced protoplast transactivation (see section 1.5). Such interactions are referred to as direct regulation in this review. Finally, we provide as a supplementary file (Additional file 1.1) a more detailed network capturing the vast majority of demonstrated direct and indirect protein-

© University of Pretoria

DNA interactions and all known protein-protein interactions. This resource can be interactively visualized and manipulated with the freeware program Cytoscape (Shannon *et al.*, 2003), and is accompanied by a list of the literature supporting each of the 435 captured interactions (Additional file 1.2). To the best of our knowledge, this is the most exhaustive network representation compiled to date. The Cytoscape representation has several uses. First, it assists the generation of hypotheses related to biological function of poorly characterized proteins based on their interactions with known proteins. Second, additional attributes such as expression data may be integrated into the network to better understand network function and behaviour. This is further enhanced by the fact that the network layout can easily be converted into built-in or customized views, and new interactions added as they are reported in the literature. In future, researchers may be able to use the network to provide structural information for the building of probabilistic causal networks that integrate diverse types of data, as performed in yeast by Zhu *et al.* (2012). Third, the network serves as a reliable basis for template-based construction of SCW transcriptional networks in sequenced non-model organisms (Babu *et al.*, 2009).

At least three main tiers of TFs can be identified in the network that ultimately regulate a suite of structural genes involved in cellulose, hemicellulose and lignin biosynthesis, signal transduction, the cytoskeleton, programmed cell death and proteins with unknown functions (Fig. 1.1, 1.2). We designated TF tiers from the bottom upwards, relative to a reliable reference point, i.e. the structural genes. A similar convention has been adopted before (Jothi *et al.*, 2009). First-tier TFs are only known to directly regulate structural genes, second-tier TFs directly regulate first-tier TFs in additional to structural genes, and so forth. We stress that

© University of Pretoria

this assignment is not rigid and that TFs may be re-assigned, where possible, to a different tier as additional data arises. Furthermore, extensive feedback may occur between tiers.

SCW transcriptional networks in different cell types that synthesize SCWs are initiated by distinct, functionally redundant pairs of NAC proteins, which have been broadly referred to as secondary wall NACs (SWNs) (Zhong *et al.*, 2010c) (Fig. 1.2; third tier). Specifically, SCW deposition in xylary and interfascicular fibers (Mitsuda *et al.*, 2007; Zhong *et al.*, 2007b; Zhong *et al.*, 2008a) as well as silique valve endocarps and valve margins (Mitsuda and Ohme-Takagi, 2008) is redundantly regulated by NAC SECONDARY WALL THICKENING PROMOTING FACTOR1 (NST1) and SECONDARY WALL ASSOCIATED NAC DOMAIN PROTEIN1 (SND1). SND1 has also been referred to as NST3 and ANAC012 (Ko *et al.*, 2007; Mitsuda *et al.*, 2007; Mitsuda and Ohme-Takagi, 2008); to avoid confusion, we refer to this protein as SND1. In meta- and protoxylem vessels, SCW deposition is regulated by VASCULAR RELATED NAC DOMAIN6 (VND6) and VND7, respectively (Kubo *et al.*, 2005; Yamaguchi *et al.*, 2008; Zhong *et al.*, 2008a; Yamaguchi *et al.*, 2010a). NST1 and NST2 are SCW master regulators in the endothecium of anthers (Mitsuda *et al.*, 2005). To date, comparatively little data are available for the regulatory functions of NST2. MYB26 activates *NST1* and *NST2* in the endothecium through an as yet unknown mechanism (Yang *et al.*, 2007), suggesting the existence of a fourth tier (Fig. 1.2).

While the SCW master regulators in fibers, vessels, siliques and anther endothecia differ from one another, current data suggest that they regulate a common core transcriptional network (Fig. 1.2). VND6/VND7 and NST2 regulatory functions largely overlap with those of

© University of Pretoria

SND1/NST1, but a number of targets are unique to VND6, VND7 or SND1 (Fig. 1.2). Notably, vessel differentiation is distinguished from fiber development by strong VND6/VND7-mediated activation of genes involved in programmed cell death (PCD); in contrast, PCD gene activation by SND1/NST1 is weak (Ohashi-Ito *et al.*, 2010; Zhong *et al.*, 2010c) (Fig. 1.2). A second notable difference is the fact that VND6/VND7 participate in a positive feedback loop with ASYMMETRIC LEAVES2/LATERAL ORGAN BOUNDARIES DOMAIN TFs ASL19 and ASL20 (Soyano *et al.*, 2008), and VND7 additionally interacts with the transcriptional repressor protein VNI2 (Yamaguchi *et al.*, 2010b) (see section 1.4.4) which has not been identified in other cell types. VND7 also interacts with VND1, VND2 and VND3 (Additional file 1.1) which do not have clearly defined functions, whereas VND6 interacts primarily with itself and probably binds as a homodimer *in vivo* (Yamaguchi *et al.*, 2008).

Third-tier SWNs directly regulate common second-tier MYB domain TFs *MYB46*, *MYB83*, and *MYB103*, NAC domain TFs *XND1* and *SND3*, and ASYMMETRIC LEAVES2/LATERAL ORGAN BOUNDARIES DOMAIN TF *ASL11* (Zhong *et al.*, 2010c; Yamaguchi *et al.*, 2011) (Fig. 1.2). MYB46 and MYB83, which are functionally redundant, appear to form a common regulatory hub in the second tier that directly regulate first-tier TFs *MYB6*, *MYB43*, *MYB52*, *MYB54*, *MYB58* and *MYB63*, the functionally redundant trio *MYB4/MYB7/MYB32* (see section 1.4.4), a C3H-type zinc finger gene *C3H14* and homeobox TF *KNAT7* (Ko *et al.*, 2009; McCarthy *et al.*, 2009; Nakano *et al.*, 2010; Zhong and Ye, 2012) (Fig. 1.2). *KNAT7* is directly activated by all the SWNs (Zhong *et al.*, 2008a). In turn, KNAT7 represses cellulose, hemicelluloses and lignin biosynthetic genes directly or indirectly (Li *et*

© University of Pretoria

al., 2012a) (Fig. 1.1). KNAT7-mediated repression is dependent on protein-protein interactions with MYB75, a weak transcriptional activator which has no known targets or direct regulators (Bhargava *et al.*, 2010; Bhargava *et al.*, 2013) (Fig. 1.2). MYB20, associated with the regulation of lignin biosynthetic genes, is likely a first-tier candidate since it is an indirect SND1 target but downregulated in the *myb103* mutant (Zhong *et al.*, 2008a; Öhman *et al.*, 2012) (Fig. 1.1). A number of novel bZIP, homeodomain, BEL1-like and zinc finger TFs that have not been linked to SCW regulation were also listed as MYB46/MYB83 direct targets (Zhong and Ye, 2012) (Additional file 1.1, Additional file 1.2). Dominant repression of MYB52 and MYB54 result in reduced fiber SCW deposition (Zhong *et al.*, 2008a). Enhanced drought tolerance in MYB52 overexpression lines (Park *et al.*, 2011) suggests a pleiotropic role for this gene in both fiber development and abiotic stress response.

The first-tier TFs regulate various SCW biosynthetic genes although some members of the second tier (MYB46, MYB61, MYB83) and third tier (SND1, VND6 and VND7) also directly activate structural genes. BP, ATHB18, a C2H2-type zinc finger protein At3g46080, MYB20, MYB69, MYB79, MYB85 and the functionally redundant pair MYB58/MYB63 are known only to directly or indirectly regulate lignin biosynthetic genes (Mele *et al.*, 2003; Zhou *et al.*, 2009; Mitsuda *et al.*, 2010), whereas BES1 is the only TF currently shown to bind to *cellulose synthase* (*CesA*) genes in both primary and secondary cell walls (Xie *et al.*, 2011) (Fig. 1.1, Fig. 1.2). BP is a *KNOX* gene family member that maintains shoot apical meristems (Sanchez *et al.*, 2012) and strongly represses lignification in inflorescence stems (Mele *et al.*, 2003). MYB85 appears to specifically regulate the lignin pathway (Zhong *et al.*, 2008a) and appears to be regulated by MYB46/MYB83 (Fig. 1.1, Additional file 1.1). All other TFs

© University of Pretoria

regulate structural genes involved in the biosynthesis of more than one SCW biopolymer. SND2 has an unclear position in the network: it is known to be indirectly activated by SND1 (Zhong *et al.*, 2008a), it is downregulated in the *myb103* mutant (Öhman *et al.*, 2012), and appears to regulate genes related to signaling, hemicellulose and lignin polymerization in addition to the secondary wall *CesA* genes (Hussey *et al.*, 2011; Öhman *et al.*, 2012) (Fig. 1.1, Additional file 1.1). Therefore, we have tentatively placed it in tier 1.

1.4.1. Master regulators

SND1, NST1, NST2, VND6 and VND7 are considered master regulators of SCW formation because of their sufficiency for ectopic SCW deposition in some non-sclerified cell types when ectopically overexpressed (Mitsuda *et al.*, 2005; Zhong *et al.*, 2006; Mitsuda *et al.*, 2007; Yamaguchi *et al.*, 2010a). By this definition, MYB family proteins MYB46, MYB83, and their direct target *C3H14* (Kim *et al.* 2012a), are also master regulators, despite occurring directly underneath SND1/NST1 and VND6/VND7 in the network (Zhong *et al.*, 2007a; Ko *et al.*, 2009; McCarthy *et al.*, 2009). MYB83 is considered redundant with MYB46 since compromised functioning of both genes is required to visibly affect the phenotype (McCarthy *et al.*, 2009). Arguably, MYB85 is a master regulator of the lignin pathway because overexpression causes ectopic lignification (Zhong *et al.*, 2008a). Conversely, non-master regulators of SCW formation are recognized by subtle cell-specific phenotypes when overexpressed: for example, *SND2*, *SND3* and *MYB103* lie downstream of master regulator SND1 (Fig. 1.1), and their constitutive expression yields differences in SCW thickness only currently identified in fibers (Zhong *et al.*, 2008b; Hussey *et al.*, 2011). The factors rendering these TFs insufficient for ectopic SCW deposition are unclear, but a likely explanation is that

© University of Pretoria

auxiliary co-regulators are required for transcriptional activation or repression which are only expressed in the cells where a phenotype is observed. Discovery of these tissue-specific factors or protein complexes will advance the elucidation of the SCW transcriptional network.

Notably, the phenotypic importance of these master regulators does not correlate with their hierarchical position in the network: for example, *MYB46* and *MYB83* are subordinate to NST1 and SND1, but the double mutant of the subordinate pair yields a more extreme phenotype than the *snd1 nst1* double mutant (Zhong *et al.*, 2007b; McCarthy *et al.*, 2009). The genome-wide identification of direct gene targets of SND1 (Ko *et al.*, 2007; Zhong *et al.*, 2010c), VND6/VND7 (Ohashi-Ito *et al.*, 2010; Zhong *et al.*, 2010c; Yamaguchi *et al.*, 2011) and MYB46/MYB83 (Zhong and Ye, 2012) have revealed key regulatory features of these master regulators. First, they do not preferentially activate TFs located in the first subordinate tier, such that the signal is relayed to successive tiers and ultimately to the structural genes at the bottom of the network. Rather, they directly regulate structural genes in addition to subordinate TFs (Fig. 1.2). This pattern is consistent with the tendency of top and middle-tier TFs to act co-operatively in target gene regulation (Gerstein *et al.*, 2012). Second, functional redundancy between proteins as assessed through mutant and complementation studies need not imply that redundant homologs regulate the same gene targets: although this might be true of MYB46 and MYB83 (Zhong and Ye, 2012), SND1 and VND6 share only ~50% of their target genes (Ohashi-Ito *et al.*, 2010). SND1 and VND6/VND7 are quantitatively different in that PCD-related genes are upregulated strongly by vessel-associated VND6/VND7 but weakly, if at all, by fiber-associated SND1/NST1 (Ohashi-Ito *et al.*, 2010; Zhong *et al.*, 2010c) (Fig. 1.2).

© University of Pretoria

Induction of SND1 in undifferentiated transgenic *Arabidopsis* suspension culture cells is sufficient for smooth SCW deposition reminiscent of fibers, whereas induction of VND6 is sufficient for that resembling metaxylem vessels (Ohashi-Ito *et al.*, 2010). Similarly the complementation of SCW deposition of fibers in the *snd1 nst1* double mutant by VND7 driven by the *SND1* promoter resulted in vessel-like patterning of the fiber SCWs (Yamaguchi *et al.*, 2011). This suggests that SND1/NST1 and VND6/VND7 are sufficient for fiber- and vessel-specific differentiation. However, ectopic overexpression of SND1 only induced ectopic SCW deposition in particular cell types, with SCW patterning including smooth, banded, reticulated or helical deposition depending on the cell type (Mitsuda *et al.*, 2005; Zhong *et al.*, 2006). Poplar VND and NST homologs preferentially induce ectopic SCW deposition in hypocotyls, rather than leaves or roots, when constitutively expressed in *Arabidopsis* (Ohtani *et al.*, 2011). Additionally, whilst all SWNs can transactivate the promoter of the PCD-related gene *XCP1* in protoplasts, the gene is not expressed in fibers under the control of SND1/NST1 (Zhong *et al.*, 2010c). Together, these data suggest that whilst fiber- and vessel-associated SWNs preferentially confer SCW deposition patterns characteristic of these cell types, the action of other regulatory mechanisms between cell types may modify their gene targets.

1.4.2. Evolutionary conservation

The evolutionary history of SWN-mediated SCW regulation is not yet resolved. Although the moss *Physcomitrella* and primitive tracheophyte *Selaginella* possess multiple NAC proteins ancestral to the SWNs found in angiosperms, these proteins lack the extended C-terminal motifs found in derived SWNs (Shen *et al.*, 2009; Zhao *et al.*, 2010b; Zhu *et al.*, 2012). Whilst

© University of Pretoria

their functions are currently unknown, it is thought that these progenitor SWN proteins were adapted for the regulation of SCW deposition in advanced vascular plants (Zhong *et al.*, 2010a), mainly through the acquisition of C-terminal activation motifs, such as the WQ-box which is essential for SND1 transcriptional activation (Ko *et al.*, 2007). There is strong evidence that these C-terminal expansions preceded angiosperm radiation (Shen *et al.*, 2009).

The basis for the evolutionary conservation of functional redundancy between SND1-NST1, NST1-NST2 and VND6-VND7 pairs in different cell types in *Arabidopsis* and possibly other angiosperms is also poorly understood. Although postulated to be a backup mechanism to ensure SCW deposition ensues (Schuetz *et al.*, 2013), a wealth of theoretical models have been proposed to explain the persistence of functional redundancy in higher organisms (Nowak *et al.*, 1997; Krakauer and Nowak, 1999; Zhang, 2003). Redundancy appears to be a general characteristic of transcriptional regulators, as suggested by their underrepresentation amongst genes with single-copy status identified across twenty angiosperms (De Smet *et al.*, 2013). Interestingly, *Medicago* is the only angiosperm known to possess only one SWN, MtNST1. The *Mtnst1* mutant exhibits loss of fiber SCW deposition, reduced anther dehiscence and even defective guard cell functioning, but no apparent effect on vessels (Zhao *et al.*, 2010a). Thus, *Medicago* appears to have dispensed of the redundant homologs and may serve as a suitable candidate for the study of the evolutionary persistence of functional redundancy in other groups.

Numerous examples of functional conservation between *Arabidopsis* SCW-regulating TFs and their homologs in a variety of plants suggest that the SCW transcriptional network is

© University of Pretoria

largely conserved in angiosperms. Functional orthologs of *Arabidopsis* SWNs and MYB46 have been experimentally verified in the monocots *Brachypodium distachyon*, *Zea mays* and *Oryza sativa*, suggesting the establishment of the basic structure of the SCW transcriptional network at least prior to monocot-dicot divergence (Zhong *et al.*, 2011a; Valdivia *et al.*, 2013). Strong evidence also corroborates an *Arabidopsis*–like transcriptional cascade in woody angiosperm species. Whilst homologs of several TF candidates in Fig. 1.1 have been linked to xylem development in hybrid aspen (Bylesjö *et al.*, 2009; Courtois-Moreau *et al.*, 2009) and *Acacia* (Suzuki *et al.*, 2011), studies in *Populus trichocarpa* principally have demonstrated functional conservation of many SCW-regulating TF orthologs. A number of functionally redundant co-orthologs of SND1 from *P. trichocarpa*, referred to as wood-associated NAC domain TFs (PtrWNDs), are capable of ectopic SCW formation in *Arabidopsis* and can complement the *snd1 nst1* double mutant (Zhong *et al.*, 2010b). *Populus* orthologs of TFs regulated by SND1 in *Arabidopsis* (Zhong *et al.*, 2008a) are likewise regulated by the *Populus* PtrWNDs (Zhong *et al.*, 2011b), and a functional ortholog of KNAT7 has been described (Li *et al.*, 2012a). *Populus* PtrMYB3 and PtrMYB20 demonstrated similar master regulatory functions to their *Arabidopsis* homologs MYB46/MYB83 (McCarthy *et al.*, 2009; McCarthy *et al.*, 2010) and are sufficient for ectopic lignification in *Arabidopsis* (Zhong *et al.*, 2010b). *Eucalyptus gunni* also possesses an SND1 homolog, EgWND1, that displays functional conservation with *Populus* and *Arabidopsis* SWNs (Zhong *et al.*, 2010a; Zhong *et al.*, 2011b). EgMYB2, a close homolog of MYB46/MYB83 from *E. gunnii*, binds to promoters of lignin biosynthetic genes *EgCCR* and *EgCAD2* (Goicoechea *et al.*, 2005) and can complement the *myb46 myb83 Arabidopsis* mutant, suggesting functional orthology with MYB46/MYB83 (Zhong *et al.*, 2010a).

© University of Pretoria

The high degree of conservation in SCW-associated TF function between *Arabidopsis* and woody plants suggests that studies in the former are of direct relevance to SCW formation in other herbaceous and woody plants. In support of this, a genome-wide survey of *cis*-regulatory sequence combinations in promoters of *Arabidopsis* and *Populus* found that over 18,000 combinations are shared between these organisms and that most of these combinations are functional (Ding *et al.*, 2011). However, it is not yet clear whether network topology is equally conserved, and it is possible that *cis*-element evolution, which is both necessary and sufficient for network rewiring (Carroll, 2008), has occurred between species. In rice, for example, there appears to be functional divergence between an AP2/ERF TF known as SHINE (OsSHN), which has a SCW-regulatory function, and its closest *Arabidopsis* and barley homologs which regulate wax and lipid biosynthesis (Aharoni *et al.*, 2004; Broun *et al.*, 2004; Kannangara *et al.*, 2007; Taketa *et al.*, 2008). OsSHN is tightly co-expressed with homologs of SCW-associated TFs and biosynthetic genes. Interestingly, *Arabidopsis AtSHN1* was shown to directly repress rice homologs of *MYB58*/*MYB63*, *NST1*/*NST2*/*SND1* and *VND4*/*VND5*/*VND6* when overexpressed in rice (Ambavaram *et al.*, 2011). Rice, but not *Arabidopsis* plants, overexpressing *AtSHN1* showed increased sclerenchyma SCW thickness, decreased lignin and increased cellulose content (Kannangara *et al.*, 2007; Ambavaram *et al.*, 2011). The likely explanation for this phenotype is that, whilst the homologs of master regulators and lignin-associated TFs such as MYB85 and MYB58/MYB63 are repressed by AtSHN1, other TFs (including homologs of MYB20/MYB43) are upregulated which may specifically regulate cellulose deposition (Ambavaram *et al.*, 2011). Together, this data suggests that the differing SHN targets in rice (monocots) and *Arabidopsis* (dicots) have evolved through changes in *cis*-element composition in their promoters, rather than the SHN

© University of Pretoria

DNA-binding domain, since AtSHN1 can switch from wax to SCW pathway regulation depending on the genetic background.

1.4.3. DNA–protein interactions

TFs promote or inhibit transcription of target genes by binding to *cis*-elements in their promoters. General canonical binding sites for MYB and NAC domain TFs have been identified, and a number of *cis*-regulatory elements recognized by TFs involved in SCW regulation specifically have been described (Table 1.1). The *secondary wall NAC binding element* (SNBE) was discovered in the promoters of SND1 direct targets, existing as several related variants in target gene promoters (Zhong *et al.*, 2010c). It consists of 19 nucleotides and is semi-palindromic, as demonstrated by reverse complementation (Table 1.1). NST1, NST2, VND6 and VND7 all recognize the SNBE consensus sequence, but the differential ability of SWNs and their orthologs to activate naturally occurring variants of this element suggests that particular SWNs will preferentially activate SNBE elements of different promoters (Zhong *et al.*, 2010c; Zhong *et al.*, 2011a). The SNBE sequence is essential for SWN-mediated promoter activation, and *cis*-element copy number is correlated with the strength of promoter transactivation (Zhong *et al.*, 2010c). Recently, SWN homologs in the monocot *Brachypodium* were also shown to recognise the SNBE (Valdivia *et al.*, 2013). Wang *et al.* (2011) have identified a significantly more specific SNBE-like element bound by SND1, TACNTTNNNNATGA, which does not appear to be semi-palindromic (Table 1.1). Both the SNBE and SNBE-like elements appear superficially similar to the general NAC recognition sequence (NACRS), but neither contains the previously reported “canonical” CACG motif (Tran *et al.*, 2004) (Table 1.1). It has recently been revealed that NACs possess

© University of Pretoria

some degree of flexibility when binding as dimers, allowing for one monomer to bind to a strong canonical DNA element and the other monomer to a low-affinity element a variable number of bases away (Welner *et al.*, 2012). This may explain why SNBE is not a perfect palindrome.

TE-specific expression may be mediated by the eleven base pair *tracheary element-regulating cis-element* (TERE) (Pyo *et al.*, 2007). The element was identified in the promoters of sixty *Arabidopsis* genes upregulated during *in vitro* TE transdifferentiation (Kubo *et al.*, 2005; Pyo *et al.*, 2007). These included SCW-associated *CesA4* and *CesA7* promoters which were not identified as direct SND1 targets by Zhong *et al.* (2010c). It was suggested from protoplast transactivation experiments that VND6 and VND7 recognize the TERE elements of several SCW-associated genes (Ohashi-Ito *et al.*, 2010; Yamaguchi *et al.*, 2011). However, Zhong *et al.* (2010c) showed using electrophoretic mobility shift competition assays (EMSA) that VND6, VND7 and SND1 do not bind directly to the TERE element, and that most genes regulated by VND6 and VND7 do not contain recognizable TEREs (Ohashi-Ito *et al.*, 2010; Yamaguchi *et al.*, 2011). From *XCP1* promoter deletion experiments, Yamaguchi *et al.* (2011) postulated that the TERE is essential for basal transcription of VND7 targets, whilst their data supported the involvement of an additional element in enhancing VND7-mediated transactivation.

AC-rich elements are associated with various lignin biosynthetic genes (Raes *et al.*, 2003) and are thought to be generally bound by MYB proteins (Zhao and Dixon, 2011). SCW-regulating MYB proteins from various taxa have been shown to bind AC elements (Table

© University of Pretoria

1.1). The AC-like *cis*-element recognized by second-tier master regulators MYB46/MYB83 was independently identified by Zhong and Ye (2012; SMRE) and Kim *et al.* (2012a; MYB46RE), and the reported sequences are essentially identical following reverse complementation (Table 1.1, underlined). However three of eight functional variants of SMRE correspond to AC-I, AC-II and AC-III (Table 1.1) (Zhong and Ye, 2012), and MBSIIG, apparently a general MYB binding site recognized by *Arabidopsis* MYB proteins that are relatively distantly related from each other (Romero *et al.*, 1998), is identical to SMRE/MYB46RE (Table 1.1). In fact, diverse *Arabidopsis* R2R3-MYB proteins bind to similar, if not identical, sequences due to highly shared recognition specificities (Romero *et al.*, 1998; Prouse and Campbell, 2012). Despite this, MYB46RE is highly enriched amongst MYB46-regulated gene promoters compared to the genome frequency (Kim *et al.*, 2012a), suggesting that MYB46RE/SMRE/MBSIIG may be specifically associated with MYB TFs involved in lignin and/or cell wall regulation. Specificity may be conferred by the requirement of multiple instances of the motif at a promoter, as is the case for EgMYB2 binding to the *EgCAD* promoter, or additional elements such as the linked BSb element that confers cambium-specific expression (Table 1.1) (Rahantamalala *et al.*, 2010). Spatial expression specificity is discussed further in section 1.4.5. Notably F5H, which is required for the biosynthesis of S monolignols in Angiosperms, does not contain AC elements in the promoter region (Raes *et al.*, 2003). Zhao *et al.* (2010b) found that *Arabidopsis* SND1 could directly activate the *Medicago F5H* promoter. However, Öhman *et al.* (2012) were not able to demonstrate transactivation of the *Arabidopsis* F5H promoter by SND1.

© University of Pretoria

SCW-related canonical *cis*-elements have been identified *in vitro* through EMSA and *in vivo* through transactivation of naked DNA in GUS reporter constructs. However accurate characterization of *cis*-elements, which should preferably resemble a probability distribution, will require genome-wide knowledge of occupied sites *in planta*. Available binding sites in a given cell type are heavily influenced by chromatin structure and composition, and TF specificity may be dependent on post-translational modifications and protein-protein interactions. Using ChIP-seq and its high-resolution derivative, ChIP-exo (Rhee and Pugh, 2011) (see section 1.5), it will be possible to obtain statistical support for these motifs and assess single nucleotide dependencies, as has been done for MADS box TFs (Kaufmann *et al.*, 2009; Kaufmann *et al.*, 2010).

1.4.4. Network dynamics

Transcriptional networks are ultimately composed of small recurrent circuits known as network motifs, which are discrete patterns of interactions that occur more frequently than expected from randomized networks (Milo *et al.*, 2002; Walhout, 2006; MacNeil and Walhout, 2011). In contrast to sensory networks, transcriptional networks regulating developmental processes tend to act slowly and can irreversibly trigger a transient developmental instruction. Negative and positive feedback loops and long cascades of transcriptional regulation are a prominent feature of developmental networks (see Alon, 2007 for review). Here, we explore network motifs and possible functions of putative modules in the SCW transcription network. Since network modules have no consensus definition (Dong and Horvath, 2007), we define them in this section as a group of connected nodes that collectively determines a pattern of target gene regulation distinct from the regulatory effect of

© University of Pretoria

each individual node on the target gene(s). These modules should be understood as teams of transcriptional regulators that co-operate to achieve an appropriate transcriptional response of a target gene(s) following a perturbation in the expression of an individual regulator in the module from steady-state levels.

A negative feedback loop involving SND1, MYB46/MYB83 and a trio of repressors may prevent uncontrolled target gene activation during fiber SCW deposition. SND1 activates MYB32 directly, as well as indirectly through a coherent feed-forward loop involving SND1 targets MYB46/MYB83 which in turn activate MYB32 (Fig. 1.3a). *MYB4* and *MYB7* are also targets of MYB46/MYB83 and have a conserved repression protein motif in common with MYB32 (Preston *et al.*, 2004; Ko *et al.*, 2009). Overexpression of a maize homolog of *MYB4* in *Arabidopsis* results in downregulation of the lignin pathway and a patchy SCW deposition phenotype in interfascicular fibers (Sonbol *et al.*, 2009), supporting a repressive role for these proteins. SND1 and its poplar co-orthologs can self-activate their own promoters (Wang *et al.*, 2011; Zhong *et al.*, 2011b; Li *et al.*, 2012b), an arrangement that is generally associated with a slow transcriptional response (Mejia-Guerra *et al.*, 2012). MYB4, MYB7 and MYB32 in turn repress SND1 through an as yet unresolved mechanism (Fig. 1.3a) (Wang *et al.*, 2011). In addition, there is evidence from promoter transactivation experiments that MYB4, MYB7 and MYB32 repress their own promoters (Ko *et al.*, 2009). Such negative autoregulation tends to accelerate transcriptional responses (Rosenfeld *et al.*, 2002; Chalancon *et al.*, 2012) and reduce transcriptional noise (Kærn *et al.*, 2005; Alon, 2007). It could be postulated therefore that a combination of slow target gene activation by master regulator SND1, combined with a rapid MYB4/7/32-mediated negative feedback loop keeps SND1 activation in check, resulting

© University of Pretoria

in gradual target gene activation. This hypothesis is consistent with the prolonged lifespan and SCW deposition of fibers relative to vessels (Gorshkova *et al.*, 2012).

In addition to negative feedback loops, a number of repressors of SCW deposition may help to prevent runaway structural gene activation or "fine-tune" their regulation. *XYLEM NAC DOMAIN 1* (*XND1*) is an SND1-activated NAC domain TF that may negatively regulate tracheary element growth (Zhao *et al.*, 2008; Zhong *et al.*, 2010c). No XND1 direct targets are currently known (Fig. 1.2). Overexpression in *Arabidopsis* causes stunting, discontinuous or complete loss of xylem vessels, as well as a failure of xylem to undergo SCW deposition or PCD (Zhao *et al.*, 2008). *KNAT7*, a class II *KNOX* gene that is also a direct target of SND1 (Zhong *et al.*, 2008a), represses SCW deposition in xylary and interfascicular fibers through repression of cellulose, hemicellulose and lignin biosynthetic genes (Li *et al.*, 2011; Li *et al.*, 2012a) (Fig. 1.2). Surprisingly, *KNAT7* yields an *irx* phenotype in vessels of the null mutant (Brown *et al.*, 2005; Li *et al.*, 2012a), suggesting that KNAT7 may act as an activator in vessels (Schuetz *et al.*, 2013). *SND1* and *KNAT7* form a type 1 incoherent feed-forward loop, such that SND1 activates structural genes as well as *KNAT7*, after which KNAT7 represses the structural genes once its protein has been synthesized (Fig. 1.3b). This motif generates a pulse of target activation such that it reaches steady-state transcript levels faster than a simple regulation model, peaks and then declines to the stable target transcript abundance as the intermediate repressor becomes engaged (Alon, 2007). The response time of target gene activation is likely to be further accelerated by other TFs such as MYB46/MYB83 which also activate the structural genes. Thus, the putative module in Fig. 1.3b is hypothesized to cause a rapid burst of structural gene transcript levels followed by a return to a steady state. KNAT7

© University of Pretoria

additionally participates in protein-protein interactions with MYB75, a repressor of the lignin pathway that pleiotropically regulates anthocyanin biosynthesis (Bhargava *et al.*, 2010; Bhargava *et al.*, 2013), in addition to OFP1, OFP4 and MYB5 interaction (Wang *et al.*, 2007; Li *et al.*, 2011; Bhargava *et al.*, 2013). An interesting mechanism has been proposed whereby the differing fiber and vessel phenotypes observed in the *knat7* mutant depend on the composition and abundance of KNAT7-interacting proteins in different cell types (Li *et al.*, 2012a). Bhargava *et al.* (2013) propose that KNAT7 forms a complex with OFP proteins and MYB75 to repress lignin biosynthetic genes in stems, whereas it forms a complex with TT8, MYB5 and MYB75 which represses SCW biosynthetic genes in the seed coat.

In contrast to the negative regulatory loop regulating SND1 in fibers (Fig. 1.3a), VND6/VND7 master regulators of vessel SCW deposition are involved in a positive feedback loop with ASL/LBD family proteins (Iwakawa *et al.*, 2002; Shuai *et al.*, 2002). VND6/VND7 promote *ASL19/ASL20* upregulation through an unknown mechanism, and ASL19/ASL20 in turn promote VND6/VND7 upregulation such that they show similar expression patterns (Soyano *et al.*, 2008) (Fig. 1.3c). In addition, *ASL19* is downregulated by VNI2, a repressor that interacts with VND7 proteins to repress its function indirectly by competing with its heterodimerizing partners and possibly neutralizing VND7-mediated transcriptional activation (Yamaguchi *et al.*, 2010b). Since VNI2 is sensitive to the ubiquitin proteosome pathway (Yamaguchi *et al.*, 2010b), it has been postulated that the ASL19/ASL20/VND6/VND7 positive feedback loop promotes rapid and irreversible differentiation of vessel elements once VNI2 is proteolytically degraded (Ohashi-Ito and Fukuda, 2010).

© University of Pretoria

A similar yet distinct mechanism to the VNI2-VND6/VND7 interaction has been documented in *Populus*. Recently, Li *et al.* (2012b) discovered a naturally occurring splice variant of a poplar SND1 co-ortholog *PtrSND1-A2*. The intron-retaining transcript variant, *PtrSND1-A2*^{IR}, encodes a truncated protein lacking transactivation ability and a critical DNA-binding subdomain, but it retains its ability to form homo- and heterodimers. The dominant negative regulator represses *PtrSND1-A1*, *PtrSND1-B1* and *PtrSND1-B2* by interfering with their self-activation abilities through the formation of nonfunctional heterodimers (Li *et al.*, 2012b). The regulatory significance of this arrangement is not yet clear.

Network connectivity can only partially explain the behaviour of a transcriptional network. Whilst regulatory hubs and modules may be identified from physical interaction networks, protein-DNA interactions alone may not accurately predict the outcome of target gene transcriptional regulation, which is complex and highly combinatorial (Spitz and Furlong, 2012). Kinetic data are required to mathematically model the dynamic behaviour of a network (Bolouri and Davidson, 2002). New advances in network modeling allow for networks to be tested, quantified, and corrected (Sayyed-Ahmad *et al.*, 2007). Time-course expression data in particular can capture dynamic properties of transcriptional networks that steady-state transcript measurements cannot (Nelson *et al.*, 2004; Opper and Sanguinetti, 2010), and even time-course ChIP-seq data has been introduced into network models (Tang *et al.*, 2012). *Arabidopsis* and *Zinnia* transdifferentiation systems are potentially useful models for generating time-course transcript data relating to SCW regulation, but existing time-course data (e.g. Kubo *et al.*, 2005; Yamaguchi *et al.*, 2011) lacks the temporal resolution to test and model the dynamic behaviour of the SCW transcriptional network.

© University of Pretoria

1.4.5. Spatial specificity of fiber and vessel development

The preferential expression patterns of *SND1*/*NST1* and *VND6*/*VND7* in the *Arabidopsis* inflorescence and hypocotyl stems are remarkably fiber- and vessel-specific, respectively, and this expression pattern is consistent with the cell type showing a phenotype in loss-of-function mutants (Kubo *et al.*, 2005; Mitsuda *et al.*, 2007; Zhong *et al.*, 2007b; Yamaguchi *et al.*, 2008; Zhong *et al.*, 2008a). It is poorly understood how this cell-specific expression is achieved in xylem, but hypothetically cell type-specific signals direct *SND1*/*NST1* and *VND6*/*VND7* expression (Lucas *et al.*, 2013). One tonoplast-localized, membrane-spanning transporter protein was found to influence *SND1*/*NST1* expression through an unknown mechanism in *Arabidopsis*: the *WALLS ARE THIN1* (*WAT1*) T-DNA mutant demonstrated a marked reduction in SCW formation in interfascicular and xylary fibers as well as a reduction in inflorescence stem growth, without otherwise affecting fiber cell specification (Ranocha *et al.*, 2010). However, although *WAT1* transcripts are most prevalent in hypocotyls and inflorescence stems, the gene is almost ubiquitously expressed (Ranocha *et al.*, 2010). In addition to the generally minor effect on overall growth in the *wat1* mutant, these characteristics of *WAT1* are in conflict with the idea that signals regulating the master regulators are themselves cell-type specific. In fact, the observation that widely expressed transcription factors may participate in cell type-specific regulatory roles (Neph *et al.*, 2012) questions this expectation. The elucidation of a gene regulatory network of the *Arabidopsis* root stele showed that most TFs have a significantly broader expression pattern than their targets (Brady *et al.*, 2011), suggesting that the SWN regulators may also be more broadly expressed than expected.

© University of Pretoria

Examples of cell-to-cell signalling in the root may reveal clues to the specification of xylem cell types in vascular meristems. Protoxylem and metaxylem formation in the developing *Arabidopsis* root can be attributed to a gradient of class III HD-ZIP TFs such that high concentrations of these regulators promote metaxylem vessel formation and lower concentrations protoxylem vessel formation (reviewed in Caño-Delgado *et al.*, 2010; Hirakawa *et al.*, 2011). Specifically, the SHORT ROOT (SHR) TF is expressed in the developing stele, which moves into the endodermis to activate SCARECROW (SHR), both of which are involved in the endodermal expression of miRNA genes *MIR165A* and *MIR166B* (Carlsbecker *et al.*, 2010). Diffusion of the resulting miRNAs from the endodermis towards the centre of the stele results in a decreasing concentration gradient (reviewed in Aichinger *et al.*, 2012). Since miR165/166 post-transcriptionally inhibit HD-ZIP III TF *PHB*, an increasing gradient of *PHB* expression is created towards the centre of the stele, resulting in protoxylem formation at the stele periphery (i.e. low *PHB* concentration) and metaxylem vessel formation at the stele centre (i.e. high *PHB* expression) (Carlsbecker *et al.*, 2010; Miyashima *et al.*, 2011). Presumably, low *PHB* expression promotes *VND7* expression in protoxylem whilst high *PHB* expression drives *VND6* expression in metaxylem. However, this is yet to be investigated.

Whilst a miRNA concentration gradient model explains the formation of two distinct types of primary xylem cells in root, it cannot explain the pattern of fiber cells intercalated with vessel elements that is typically seen in secondary xylem. Such a system could be better explained, for example, by lateral polar auxin transport between adjacent cells, such that local foci of auxin maxima promote vessel differentiation whilst lower auxin concentrations

© University of Pretoria

promote fiber differentiation. Such a model is supported by the fact that, in root, lateral polar auxin transport determines the boundary between protoxylem and the procambium (reviewed in Milhinhos and Miguel, 2013), in stems the vessel density varies longitudinally as a function of the auxin concentration (reviewed in Sorce *et al.*, 2013), and that the radial expression of auxin carrier genes in stems is non-uniform (Schrader *et al.*, 2003). However, as discussed by Lucas *et al.* (2013), the localizations of auxin efflux proteins in stems and their distributions in fibers and vessels are currently unknown. It is most likely that a combination of hormones are involved: for example, the simultaneous presence of auxin, brassinosteroids and cytokinins was required for high expression of *VND6* and *VND7* (Kubo *et al.*, 2005).

Some spatial specificity in SCW deposition can be explained by the presence of transcriptional repressors in non-sclerenchymatous cells. For example, *WRKY12* is expressed in stem pith and cortex, where it inhibits SCW formation by directly repressing SCW master regulators such as *NST2* (Wang *et al.*, 2010) (Fig. 1.2). The *wrky12* mutant shows ectopic SCW formation in the pith of both *Arabidopsis* and *Medicago* inflorescence stems, suggesting that repression, rather than activation, of SCW master regulators in specific cell types contributes significantly to their specific spatial expression. Interestingly, in *Populus* many *PtrWND* genes have surprisingly widespread expression, even in shoot apices and nonvascular parts of leaves (Han *et al.*, 2011; Ohtani *et al.*, 2011). It can be postulated that a transcriptional repressor or nonfunctional splice variant is expressed in non-vascular tissues and cells that binds to PtrWND proteins to prevent them from initiating ectopic SCW deposition, in a similar way to *PtrSND1-A2*^{IR} (see section 1.4.4). Combined with the example above of the WRKY12 repressor that inhibits SCW initiation in some ground tissues in *Arabidopsis*, these

© University of Pretoria

data may point to an unexpected mechanism in which transcriptional activation of SCW deposition is a developmental program that is repressed in certain non-sclerified tissues, rather than simply induced in vascular tissues. Alternatively, co-factors required by these master regulators are not present in these nonvascular tissues, as evidenced by the failure of certain cells to ectopically deposit SCWs when the master regulators are overexpressed (see section 1.4.1).

The upstream regulators of SND1/NST1 and VND6/VND7 have not yet been reported, nor have the gene targets of xylem-regulating HD ZIPIII TFs (Fig. 1.1), which are good candidates for SWN regulation. Knowledge of the SWN regulators will greatly enhance our understanding of how cell type-specific SCW transcriptional networks are initiated. The techniques used to infer TF function, and the interpretation of specific assays, are an important aspect of gene regulation studies. Moreover, recent advances in our understanding of eukaryotic gene regulation through projects such as the Encyclopedia of DNA Elements (encodeproject.org), necessitates an increasingly single cell-level understanding of transcriptional networks. We turn now to an evaluation of the molecular tools that have been used to study and infer SCW transcriptional networks, and which approaches will best support such studies in the future.

1.5. Methodologies for the study of SCW transcriptional regulation

A number of techniques have been employed to study SCW-regulating TFs. We provide a summary of the advantages and challenges of common approaches used in the literature for TF functional annotation and SCW transcriptional network inference (Table 1.2). We have

© University of Pretoria

roughly arranged these techniques in increasing resolution of the regulatory information obtained in each; that is, increasing understanding of the *in vivo* direct gene targets of a given TF and its bound *cis*-element. Here, we discuss in greater detail the widespread use of reverse genetics and protoplast transfection approaches in model organisms in SCW regulation studies, and review approaches better suited to non-model organisms.

Classical reverse genetics approaches employing overexpression and knock-out mutagenesis have been central to the functional annotation of SCW TFs in *Arabidopsis* and *Populus* (e.g. Zhong *et al.*, 2007b; McCarthy *et al.*, 2009; Grant *et al.*, 2010). Direct or indirect targets of a TF subjected to knock-out or overexpression may be inferred under the premise that the transcriptional regulation of those targets is altered, leading to their differential expression relative to the wild type. However, we would like to highlight some problems associated with overexpression that have emerged in studies of SCW regulation, namely the level and site of overexpression.

SND1, now accepted as a master transcriptional activator of *Arabidopsis* SCW biosynthesis in fibers (Mitsuda *et al.*, 2007; Zhong *et al.*, 2007b; Zhong *et al.*, 2008a), was reported to suppress fiber SCW deposition when excessively overexpressed (Zhong *et al.*, 2006). *SND2*, an indirect target of SND1, exhibited increased fiber SCW thickness when overexpressed and a mirrored reduction in SCW deposition in dominant repression lines (Zhong *et al.*, 2008a). However, when our laboratory analyzed independent *Arabidopsis* lines overexpressing *SND2* (Hussey *et al.*, 2011), we observed a decrease in fiber SCW deposition which we attributed to *SND2* transcript levels far-exceeding those reported in the previous

© University of Pretoria

study. Such phenomena could be explained by transcriptional squelching, defined as the repressive effect of a transcriptional activator beyond a certain threshold of abundance, due to the sequestration of interacting co-regulators or general transcription factors (Cahill *et al.*, 1994; Orphanides *et al.*, 2006). Alternatively a "dosage balance" mechanism (Birchler *et al.*, 2005) holds that, for multi-subunit TF complexes, a relative increase in the abundance of one particular subunit does not lead to an increase in the yield of the assembled complex, but rather a stoichiometric reduction in the abundance of complete complexes and an increase in the abundance of non-functional sub-complexes (Birchler and Veitia, 2007, 2010). Together, these inconsistencies in functional studies of cell wall-related TFs suggest that overexpression differences may introduce indirect or even conflicting phenotypes.

Ectopic expression can also modify TF function. Regulating primary cell wall (PCW) deposition in the *Arabidopsis* root cap, three partially redundant TFs closely related to clade IIb NACs *NST1*, *SND1*, *VND6* and *VND7* have been described, namely *SOMBRERO* (*SMB*), *BEARSKIN1* (*BRN1*) and *BRN2* (Bennett *et al.*, 2010). When constitutively driven by the *35S CaMV* promoter, they are sufficient for ectopic deposition of lignified SCWs in several tissues, a phenotype resembling that of *NST1*, *VND6* and *VND7* overexpression (Bennett *et al.*, 2010). Since SCWs are not found in the root cap where the TFs normally function, ectopic expression resulted in a modification of the gene targets that *SMB* and *BRN1/2* naturally regulate, perhaps due to differences in co-regulators or other regulatory factors between tissues. This mechanism may also explain results reported by Bomal *et al.* (2008), where ectopic overexpression of the xylem-associated pine gene *PtMYB8* in spruce caused

© University of Pretoria

misregulation of flavonoid-associated transcripts, which have preferential expression in tissues that correspond to regions of low expression of native *PtMYB8* in pine.

The examples cited above of confounding effects due to the level and site of a candidate TF's expression provide compelling grounds to substantiate with additional evidence some of the conclusions arising from overexpression approaches. Such concerns have been echoed in a review of gain-of-function mutagenesis (Kondou *et al.*, 2010). To avoid these problems, loss-of-function mutagenesis and non-transgenic approaches such as ChIP-seq may be more reliable. Although conventional mutagenesis is frequently unsuitable for SCW-regulating TFs due to the high degree of functional redundancy between homologs, the use of chimeric repressor silencing technology (CRES-T; Hiratsu *et al.*, 2003) has circumvented this problem, at least for transcriptional activators. In CRES-T, dominant loss-of-function transgenic plants overexpress a candidate TF fused to a hexapeptide dominant repression domain (Hiratsu *et al.*, 2004; Mitsuda and Ohme-Takagi, 2008; Zhong *et al.*, 2008a). The hexapeptide repressor opposes the transcriptional activation function at loci bound by the TF, in addition to the complementation that functional homologs may exert at those loci. Ectopic noise arising from overexpression may also be reduced by the use of tissue-specific promoters or laser capture-microdissection to capture only those cell types where the TF candidate is naturally expressed. Similarly, inducible expression may limit knock-on effects of long-term overexpression.

Promoter transactivation by an induced candidate TF in *Arabidopsis* mesophyll protoplasts (Wehner *et al.*, 2011) has proved particularly useful in the identification of *Arabidopsis* gene targets, and may be used to complement approaches such as ChIP-seq which

© University of Pretoria

does not strictly indicate active target regulation (Table 1.2). The assay typically involves co-transfection with different plasmids, one harbouring a constitutively expressed candidate TF gene (the "effector"), a *promoter::GUS* reporter vector and a luciferase expression vector to allow for normalization of transfection efficiency. Inferring direct targets using this system is complicated by the potential ability of a candidate TF to induce transcription of an intermediate TF in the host cell that is responsible for activating a target gene. This has been addressed by translational fusion of the candidate TF to the regulatory region of the human estrogen receptor (Zuo *et al.*, 2000). The chimeric protein is post-translationally induced by β-estradiol, allowing for an inhibitor of protein synthesis to be added to the system prior to induction to block the translation of intermediate TFs. The activated chimeric TF is then able to regulate transcription of target genes using the existing transcriptional machinery of the cell, and direct targets can be inferred with *promoter::GUS* RT-qPCR analysis (Zhong *et al.*, 2008a; Zhou *et al.*, 2009) or microarray analysis of the host cell transcriptome (Zhong *et al.*, 2010c). For this reason, only those protoplast transactivation experiments that used post-translational induction were considered as evidence for protein-DNA interactions represented in Fig. 1.2, to avoid the possibility that putative targets may have been indirectly regulated.

Arabidopsis mesophyll protoplasts have also been used to assess *in vivo* promoter transactivation by *Populus* and *Eucalyptus* TFs using *GUS* reporters fused to candidate promoters from these species (Zhong *et al.*, 2011b). A genome-wide analysis of endogenous promoter transactivation of TFs from non-model species would require protoplasts derived from the same, or a related, species. Although the *promoter::GUS* approach suffers from much lower throughput, it mitigates the effects that the chromatin structure of the host

© University of Pretoria

protoplast may exert on the regulation of endogenous target genes. DNaseI hypersensitivity sites, an indicator of open chromatin marking most active TF binding sites, vary considerably across cell types (Thurman *et al.*, 2012). Therefore, protoplasts should ideally be sourced from the same tissue in which a candidate TF is expressed. The recent report of isolation and transfection of *Populus* secondary xylem protoplasts (Li *et al.*, 2012b) sets the stage for genome-wide analysis of poplar genes transactivated by poplar TFs involved in wood development.

Several approaches exist for *in vitro*, *in vivo* and *in planta* analysis of TFs from non-model organisms that are not yet easily transformed (Table 1.2). Systems genetics allows for gene regulatory networks to be reconstructed by co-expression analysis across large numbers of segregating progeny (Ayroles *et al.*, 2009). Microarray or RNA-seq expression data are obtained for tissues of interest from a structured segregating population. The addition of genetic markers allows for the identification of expression quantitative trait loci (eQTLs) (Jansen and Nap, 2001), which can be differentiated into those acting in *cis* or *trans*. Trans-eQTLs likely represent polymorphisms in transcriptional regulators, and due to their ability to affect expression of many genes, *trans*-eQTL "hotspots" may be mapped that contain significantly more eQTLs than the genome average. The combination of eQTLs and co-variation in transcript levels allows the prediction of causal relationships (Zhu *et al.*, 2007) and candidate regulators (Drost *et al.*, 2010) and is the basis on which regulatory networks can be constructed *a posteriori,* or hypothetical *a priori* networks tested (Kliebenstein, 2009). One considerable limitation of current systems genetics studies is the difficulty of studying the segregation of transcript abundance in specific cell types of organs. However, recent advances

© University of Pretoria

in obtaining high-throughput cell type-specific transcriptome data may make this challenge more feasible (Chitwood and Sinha, 2013).

Yeast one-hybrid (Y1H) has been widely used to identify TFs that interact directly with SCW-related promoters (Lin *et al.*, 2010; Kim *et al.*, 2012b). These assays can be performed either through direct cloning of candidate TF coding sequences and systematically testing interactions with different potential target promoters (Mitsuda *et al.*, 2010), or via screening of cDNA expression libraries which have the advantage of discovering novel interacting proteins (Lopato *et al.*, 2006). Recent advancements in Y1H screening, including smart pooling and robotics, have increased the generally low throughput of this technique (reviewed in Reece-Hoyes and Walhout, 2012). However, Y1H interactions occasionally fail independent validation assays. For example, although SPL8 was isolated from 20 of 72 yeast colonies showing a positive interaction with the *CCoAOMT1* promoter, it failed to activate the promoter in particle-bombarded *Arabidopsis* leaves (Mitsuda *et al.*, 2010). A significant disadvantage of Y1H is that protein-DNA interactions which require cofactors or bind as complexes will not be identified (Table 1.2).

A poorly researched area in SCW transcriptional regulation is the symplastic movement of regulatory proteins between cells. It is well known that some TFs may move from cell to cell through plasmodesmata (Burch-Smith *et al.*, 2011; Wu and Gallagher, 2012). Aside from the SHR example in section 1.4.5. (reviewed by Kurata *et al.*, 2005), in plants this has been found predominantly in meristems and involves mainly the KNOX (e.g. KNOTTED1) and MADS-box TF families (Zambryski and Crawford, 2000). However, at least one MYB-like

© University of Pretoria

protein is known to be non cell-autonomous (Wada *et al.*, 2002), and it is possible that some SCW-associated TFs act non-cell autonomously. This necessarily implies that the use of *in situ* hybridization, which has been widely used to study SCW-associated TF transcript abundance (e.g. Zhong *et al.*, 2010b), may not accurately reflect a candidate TF's biological function. For species that can be stably transformed, TF movement can be tracked by fluorescent protein fusion experiments. For example, Kim *et al.* (2003) expressed GFP~KN1 fusion proteins (where ~ denotes a linker sequence) in mesophyll or epidermal cells using tissue-specific promoters, and compared the movement of GFP~KN1 between the mesophyll and the epidermis with that of free GFP and GFP fused to a viral movement protein. Microinjection of fluorescently labeled recombinant TFs into the cytoplasm of cells of interest can also be performed, but this approach is technically cumbersome and limited to larger cells (Lucas *et al.*, 1995; Wang *et al.*, 2007). For non-model organisms, immunolocalization methods using an antibody against a TF of interest can be used to detect its presence *in planta*. Whilst low-abundance TFs may be difficult to detect using immunohistochemical methods, both alkaline phosphatase staining and immunogold labeling have been used to detect TF proteins at cellular and subcellular levels (Rodriguez-Uribe and O'Connell, 2006).

Chromatin immunoprecipitation combined with high-throughput sequencing (ChIP-seq) offers many advantages that are particularly suited to non-model organisms where genomic information is available (Table 1.2). It has been shown that even fragmented genome assemblies are acceptable for ChIP-seq read mapping (Buisine and Sachs, 2009), evading the need for genome assemblies on par with model plants. However ChIP-seq in plants currently suffers from a lack of protocols for isolation of sufficient amounts of chromatin from a wide

© University of Pretoria

range of tissues, and each tissue and species may require customized modifications to chromatin fixation, nuclei isolation and chromatin shearing (Haring *et al.*, 2007). To our knowledge no ChIP procedures have yet been applied to developing xylem from woody stems, but a report of successful mapping of the ARBORKNOX1 TF in poplar vascular cambium using ChIP-seq (Andrew Groover, personal communication) sets the stage for its implementation in xylem. A range of improvements have been made to the basic ChIP-seq principle (reviewed in Furey, 2012), amongst them the ability to amplify sufficient amounts of ChIP DNA for Illumina sequencing from limited cell numbers (Adli and Bernstein, 2011). The latter is a particularly exciting advancement as it may allow for ChIP to be applied for the first time to plant tissues where chromatin yield is poor or where a TF's expression is low.

While these and other technologies advance – especially those involving second-generation sequencing – systems approaches to study SCW transcriptional regulation are still lacking. Systems biology attempts to integrate various high-throughput datasets into a holistic biological model, or achieve meaningful dynamic modelling of extensive biological data. Despite considerable progress in plant systems biology (Yuan *et al.*, 2008) and the existence of several genome- and transcriptome-wide datasets relating to *Arabidopsis* SCW biosynthesis, few attempts have yet been made to integrate such data with other -omics platforms. The integration of metabolomic data into transcriptional networks, for example, can link modifications of TFs and their interactions to phenotypic outputs. Two model *Arabidopsis* studies, one using multiple knockout mutants of lignin biosynthetic pathway enzymes (Vanholme *et al.*, 2012) and another analyzing five TF overexpression lines involved in glucosinolate biosynthesis (Malitsky *et al.*, 2008), have integrated transcriptomic and

© University of Pretoria

metabolomic data to reveal novel aspects of metabolic pathway flux and regulation. In future, however, such analyses will have to be extended to cell-specific gene expression and interactions, especially in the field of transcriptional regulation. Overlaying cell type-specific expression profiles with Y1H and Y2H interaction data has been successfully achieved in the *Arabidopsis* root using enzymatic cell wall maceration and fluorescence-activated cell sorting of target cell protoplasts expressing a GFP marker (Brady *et al.*, 2011). Another approach developed in *Arabidopsis* involves the purification of tagged nuclei from specific cells for transcriptome and ChIP-seq analysis (Deal and Henikoff, 2010). These and other innovations will undoubtedly contribute to a systems-level understanding of SCW regulation in the near future.

1.6. Conclusion

In this review we aimed to provide a comprehensive summary of what is currently known about *Arabidopsis* SCW transcriptional regulation, highlighting current gaps in our understanding of the transcriptional network. We have also emphasized that an understanding of protein-protein interactions, spatial specificity and network dynamics (modules and hubs, regulatory motifs, and temporal regulation) is severely underdeveloped compared to what is known about the network's connectivity. The immediate goal of future research is to comprehensively identify the physical interactions (protein-DNA and protein-protein) involved in SCW transcriptional regulation. This includes the identification of not only interacting partners of known TFs, but also their cell-type context that might influence the functions of TFs in different ways. This goal will allow us to identify TFs and transcriptional modules that regulate genes involved in the biosynthesis of specific SCW biopolymers. This,

© University of Pretoria

together with systems approaches, will also reveal to what degree regulation of different genes and metabolic pathways is independent. Currently, only the lignin pathway seems to be specifically targeted by TFs such as MYB58, MYB63 and MYB85, and it may not be possible to uncouple the transcriptional regulation of cellulose and hemicellulose biopolymers. However, two recent studies have used components of the SCW transcriptional network to engineer plants with favourable biofuel properties by restoring vessel wall integrity in xylan (Petersen *et al.*, 2012) and lignin mutants (Yang *et al.*, 2013) or reinforcing polysaccharide deposition in fiber SCWs (Yang *et al.*, 2013).

The ability to predict the regulatory outcome of perturbations in transcriptional networks through network modeling is invaluable to the field of biotechnology. A detailed knowledge of the strength of interaction for each edge connecting two nodes and a mathematical understanding of how the network responds to perturbations in expression, as well as genetic and environmental modulation, has not yet been attained. Systems biology experiments in *Arabidopsis,* for which knock-out lines are readily available to quantify network dynamics in response to genetic perturbations, will contribute extensively in this regard. For non-model organisms, Y1H and ChIP-seq are expected to be two key techniques used to identify protein-DNA interactions in the near future. However systems genetics, which facilitates network reconstruction, modeling and quantification from perturbations caused by natural genetic variation, is gaining momentum in agronomically important species (Ingvarsson and Street, 2010; Mizrachi *et al.*, 2012). Identification of trait QTLs and eQTLs additionally allow for the assessment of phenotypic impact of expression variation in TFs, the strength of association of

© University of Pretoria

TFs with regulons of co-expressed genes, and the ability to apply molecular breeding strategies to populations.

An understanding of the integration of intercellular signals, miRNAs, chromatin changes and temporal dynamics in transcription during xylem development remains a future challenge, marred by a limited understanding of regulatory mechanisms. For example, the occurrence of alternative splicing as a form of SND1 regulation in *Populus* (Li *et al.*, 2012b) underscores an overlooked regulatory mechanism in SCW deposition, and there exists the possibility that certain RNA-binding proteins may participate in alternative splicing during xylogenesis. There is currently no data on cell type-specific chromatin modifications, DNA methylation or chromatin states during various aspects of fiber and vessel development that may influence availability of TF binding sites. We have no knowledge of the degree to which the SCW-associated TFs downstream of the HD-ZIP III TFs (Fig. 1.1) are post-transcriptionally regulated by miRNAs, or of the transcriptional changes associated with the transitions between S1, S2 and S3 layer deposition in SCWs. Finally, the findings that fibers in close proximity to vessels show a vessel-like lignin composition (Gorzsás *et al.*, 2011) and that lignification of tracheary elements may occur post-mortem due to monolignol transport from live cells (Pesquet *et al.*, 2013) highlights the need to better understand the role of cell non-autonomous regulation of xylogenesis. Clearly there are plenty of opportunities for further study in this exciting field.

© University of Pretoria

1.7. Acknowledgements

We thank the Mandela Rhodes Foundation and National Research Foundation of South Africa for student scholarship funding.

1.8. References

Adli M, Bernstein BE. 2011. Whole-genome chromatin profiling from limited numbers of cells using nano-ChIP-seq. *Nature Protocols* **6**(10): 1656-1668.

Aharoni A, Dixit S, Jetter R, Thoenes E, Arkel Gv, Pereira A. 2004. The SHINE clade of AP2 domain transcription factors activates wax biosynthesis, alters cuticle properties, and confers drought tolerance when overexpressed in *Arabidopsis*. *The Plant Cell* **16**: 2463-2480.

Aichinger E, Kornet N, Friedrich T, Laux T. 2012. Plant stem cell niches. *Annual Review of Plant Biology* **63**: 615-636.

Alon U. 2007. Network motifs: theory and experimental approaches. *Nature* **8**: 450-461.

Ambavaram MMR, Krishnan A, Trijatmiko KR, Pereira A. 2011. Coordinated activation of cellulose and repression of lignin biosynthesis pathways in rice. *Plant Physiology* **155**: 916-931.

Ayroles JF, Carbone MA, Stone EA, Jordan KW, Lyman RF, Magwire MM, Rollmann SM, Duncan LH, Lawrence F, Anholt RRH, Mackay TFC. 2009. Systems genetics of complex traits in *Drosophila melanogaster*. *Nature Genetics* **41**(3): 299-307.

Babu MM, Lang B, Aravind L. 2009. Methods to reconstruct and compare transcriptional regulatory networks. *Methods in Molecular Biology* **541**: 163-180.

Baima S, Possenti M, Matteucci A, Wisman E, Altamura MM, Ruberti I, Morelli G. 2001. The *Arabidopsis* ATHB-8 HD-Zip protein acts as a differentiation-promoting transcription factor of the vascular meristems. *Plant Physiology* **126**: 643-655.

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129**: 823-837.

Basso K, Margolin A, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. 2005. Reverse engineering of regulatory networks in human B cells. *Nature Genetics* **37**(4): 382-390.

Baucher M, Jaziri ME, Vandeputte O. 2007. From primary to secondary growth: origin and development of the vascular system. *Journal of Experimental Botany* **58**(13): 3485-3501.

© University of Pretoria

Bennett T, Toorn Avd, Sanchez-Perez GF, Campilho A, Willemsen V, Snel B, Scheres B. 2010. SOMBRERO, BEARSKIN1, and BEARSKIN2 regulate root cap maturation in *Arabidopsis*. *The Plant Cell* **22**: 640-654.

Bhargava A, Ahad A, Wang S, Mansfield SD, Haughn GW, Douglas CJ, Ellis BE. 2013. The interacting MYB75 and KNAT7 transcription factors modulate secondary cell wall deposition both in stems and seed coat in *Arabidopsis*. *Planta* **237**: 1199-1211.

Bhargava A, Mansfield SD, Hall HC, Douglas CJ, Ellis BE. 2010. MYB75 functions in regulation of secondary cell wall formation in the *Arabidopsis* inflorescence stem. *Plant Physiology* **154**(3): 1428-1438.

Birchler JA, Riddle NC, Auger DL, Veitia RA. 2005. Dosage balance in gene regulation: biological implications. *Trends in Genetics* **21**(4): 219-226.

Birchler JA, Veitia RA. 2007. The Gene Balance Hypothesis: from classical genetics to modern genomics. *The Plant Cell* **19**: 395-402.

Birchler JA, Veitia RA. 2010. The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. *New Phytologist* **186**: 54-62.

Bollhöner B, Prestele J, Tuominen H. 2012. Xylem cell death: emerging understanding of regulation and function. *Journal of Experimental Botany* **63**(3): 1081-1094.

Bolouri H, Davidson EH. 2002. Modeling transcriptional regulatory networks. *BioEssays* **24**: 1118–1129.

Bomal C, Bedon F, Caron S, Mansfield SD, Levasseur C, Cooke JEK, Blais S, Tremblay L, Morency M-J, Pavy N, Grima-Pettenati J, Séguin A, MacKay J. 2008. Involvement of *Pinus taeda MYB1* and *MYB8* in phenylpropanoid metabolism and secondary cell wall biogenesis: a comparative *in planta* analysis. *Journal of Experimental Botany* **59**(14): 3925-3939.

Bonke M, Thitamadee S, Mähönen AP, Hauser M-T, Helariutta Y. 2003. APL regulates vascular tissue identity in *Arabidopsis*. *Nature* **426**: 181-186.

Brady SM, Zhang L, Megraw M, Martinez NJ, Jiang E, Yi CS, Liu W, Zeng A, Taylor-Teeples M, Kim D, Ahnert S, Ohler U, Ware D, Walhout AJ, Benfey PN. 2011. A stele-enriched gene regulatory network in the *Arabidopsis* root. *Molecular Systems Biology* **7**: 459.

Broun P, Poindexter P, Osborne E, Jiang C-Z, Riechmann JL. 2004. WIN1, a transcriptional activator of epidermal wax accumulation in *Arabidopsis*. *Proceedings of the National Academy of Science* **101**(13): 4706–4711.

© University of Pretoria

Brown D, Zeef L, Ellis J, Goodacre R, Turner S. 2005. Identification of novel genes in *Arabidopsis* involved in secondary cell wall formation using expression profiling and reverse genetics. *Plant Cell* **17**: 2281-2295.

Bryan AC, Obaidi A, Wierzba M, Tax FE. 2011. XYLEM INTERMIXED WITH PHLOEM1, a leucine-rich repeat receptor-like kinase required for stem growth and vascular development in *Arabidopsis thaliana*. *Planta* **235**(1): 111-122.

Buisine N, Sachs L. 2009. Impact of genome assembly status on ChIP-Seq and ChIP-PET data mapping. *BMC Research Notes* **2**: 257.

Bulyk ML. 2007. Protein binding microarrays for the characterization of protein-DNA interactions. *Advances in Biochemical Engineering/Biotechnology* **104**: 65–85.

Burch-Smith TM, Stonebloom S, Xu M, Zambryski PC. 2011. Plasmodesmata during development: re-examination of the importance of primary, secondary, and branched plasmodesmata structure versus function. *Protoplasma* **248**: 61-74.

Bylesjö M, Nilsson R, Srivastava V, Gronlund A, Johansson AI, Jansson S, Karlsson J, Moritz T, Wingsle G, Trygg J. 2009. Integrated analysis of transcript, protein and metabolite data to study lignin biosynthesis in hybrid aspen. *Journal of Proteome Research* **8**(1): 199-210.

Cahill MA, Ernst WH, Janknecht R, Nordheim A. 1994. Regulatory squelching. *FEBS Letters* **344**: 105-108.

Caño-Delgado A, Lee J-Y, Demura T. 2010. Regulatory mechanisms for specification and patterning of plant vascular tissues. *Annual Review of Cell and Developmental Biology* **26**: 605–637.

Caño-Delgado A, Penfield S, Smith C, Catley M, Bevan M. 2003. Reduced cellulose synthesis invokes lignification and defense responses in *Arabidopsis thaliana*. *The Plant Journal* **34**(3): 351-362.

Carlsbecker A, Helariutta Y. 2005. Phloem and xylem specification: pieces of the puzzle emerge. *Current Opinion in Plant Biology* **8**(5): 512-517.

Carlsbecker A, Lee J-Y, Roberts CJ, Dettmer J, Lehesranta S, Zhou J, Lindgren O, Moreno-Risueno MA, Vatén A, Thitamadee S, Campilho A, Sebastian J, Bowman JL, Helariutta Y, Benfey PN. 2010. Cell signalling by microRNA165/6 directs gene dose-dependent root cell fate. *Nature* **465**: 316-321.

Carroll A, Somerville C. 2009. Cellulosic Biofuels. *Annual Review of Plant Biology* **60**: 165-182.

Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. *Cell* **134**: 25-36.

© University of Pretoria

Chalancon G, Ravarani CNJ, Balaji S, Martinez-Arias A, Aravind L, Jothi R, Babu MM. 2012. Interplay between gene expression noise and regulatory network architecture. *Trends in Genetics* **28**(5): 221-232.

Chitwood DH, Sinha NR. 2013. A census of cells in time: quantitative genetics meets developmental biology. *Current Opinion in Plant Biology* **16**: 92-99.

Courtois-Moreau CL, Pesquet E, Sjödin A, Muñiz L, Bollhöner B, Kaneda M, Samuels L, Jansson S, Tuominen H. 2009. A unique program for cell death in xylem fibers of *Populus* stem. *The Plant Journal* **58**(2): 260-274.

De Smet R, Adams KL, Vandepoele K, Van Montagu MCE, Maere S, van de Peer Y. 2013. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proceedings of the National Academy of Sciences of the USA* **110**(8): 2898-2903.

Deal RB, Henikoff S. 2010. The INTACT method for cell type-specific gene expression and chromatin profiling in *Arabidopsis thaliana*. *Nature Protocols* **6**(1): 56-68.

Deplancke B, Dupuy D, Vidal M, Walhout AJM. 2004. A Gateway-compatible yeast one-hybrid system. *Genome Research* **14**(10b): 2093–2101.

Dettmer J, Elo A, Helariutta Y. 2009. Hormone interactions during vascular development. *Plant Molecular Biology* **69**(4): 347-360.

Ding J, Hu H, Li X. 2011. Thousands of cis-regulatory sequence combinations are shared by *Arabidopsis* and poplar. *Plant Physiology* **158**: 145–155.

Doblin MS, Pettolino F, Bacic A. 2010. Plant cell walls: the skeleton of the plant world. *Functional Plant Biology* **37**: 357-381.

Dong J, Horvath S. 2007. Understanding network concepts in modules. *BMC Systems Biology* **1**: 24.

Drost DR, Benedict CI, Berg A, Novaes E, Novaes CRDB, Yu Q, Dervinis C, Maia JM, Yap J, Miles B, Kirst M. 2010. Diversification in the genetic architecture of gene expression and transcriptional networks in organ differentiation of *Populus*. *Proceedings of the National Academy of Sciences* **107**(18): 8492–8497.

Du J, Groover A. 2010. Transcriptional regulation of secondary growth and wood formation. *Journal of Integrative Plant Biology* **52**(1): 17–27.

Dubos C, Stracke R, Grotewold E, Weisshaar B, Martin C, Lepiniec L. 2010. MYB transcription factors in *Arabidopsis*. *Trends in Plant Science* **15**(10): 573-581.

Emery JF, Floyd SK, Alvarez J, Eshed Y, Hawker NP, Izhaki A, Baum SF, Bowman JL. 2003. Radial patterning of *Arabidopsis* shoots by class III HD-ZIP and KANADI genes. *Current Biology* **13**(20): 1768-1774.

© University of Pretoria

Eriksson ME, Israelsson M, Olsson O, Moritz T. 2000. Increased gibberellin biosynthesis in transgenic trees promotes growth, biomass production and xylem fiber length. *Nature Biotechnology* **18**(17): 784-788.

Etchells JP, Turner SR. 2010. The PXY-CLE41 receptor ligand pair defines a multifunctional pathway that controls the rate and orientation of vascular cell division. *Development* **137**: 767-774.

Fincher GB. 2009. Revolutionary times in our understanding of cell wall biosynthesis and remodeling in the grasses. *Plant Physiology* **149**(1): 27-37.

Fratzl P, Elbaum R, Burgert I. 2008. Cellulose fibrils direct plant organ movements. *Faraday Discuss* **139**: 275-282.

Fukuda H, Komamine A. 1980. Establishment of an experimental system for the study of tracheary element differentiation from single cells isolated from the mesophyll of *Zinnia elegans*. *Plant Physiology* **65**: 57-60.

Furey TS. 2012. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nature Reviews Genetics* **13**: 840-852.

Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan K-K, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, Min R, Alves P, Abyzov A, Addleman N, Bhardwaj N, Boyle AP, Cayting P, Charos A, Chen DZ, Cheng Y, Clarke D, Eastman C, Euskirchen G, Frietze S, Fu Y, Gertz J, Grubert F, Harmanci A, Jain P, Kasowski M, Lacroute P, Leng J, Lian J, Monahan H, O'Geen H, Ouyang Z, Partridge EC, Patacsil D, Pauli F, Raha D, Ramirez L, Reddy TE, Reed B, Shi M, Slifer T, Wang J, LinfengWu, Yang X, Yip KY, Zilberman-Schapira G, Batzoglou S, Sidow A, Farnham PJ, Myers RM, Weissman SM, Snyder M. 2012. Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**: 91-100.

Goicoechea M, Lacombe E, Legay S, Mihaljevic S, Rech P, Jauneau A, Lapierre C, Pollet B, Verhaegen D, Chaubet-Gigot N, Grima-Pettenati J. 2005. *Eg*MYB2, a new transcriptional activator from *Eucalyptus* xylem, regulates secondary cell wall formation and lignin biosynthesis. *The Plant Journal* **43**(4): 553-567.

Gorshkova T, Brutch N, Chabbert B, Deyholos M, Hayashi T, Lev-Yadun S, Mellerowicz EJ, Morvan C, Neutelings G, Pilate G. 2012. Plant fiber formation: state of the art, recent and expected progress, and open questions. *Critical Reviews in Plant Sciences* **31**(3): 201-228.

Gorzsás A, Stenlund H, Persson P, Trygg J, Sundberg B. 2011. Cell-specific chemotyping and multivariate imaging by combined FT-IR microspectroscopy and orthogonal projections to

© University of Pretoria

latent structures (OPLS) analysis reveals the chemical landscape of secondary xylem. *The Plant Journal* **66**: 903-914.

Grant EH, Fujino T, Beers EP, Brunner AM. 2010. Characterization of NAC domain transcription factors implicated in control of vascular cell differentiation in *Arabidopsis* and *Populus*. *Planta* **232**: 337-352.

Han X, He G, Zhao S, Guo C, Lu M. 2011. Expression analysis of two NAC transcription factors *PtNAC068* and *PtNAC154* from poplar. *Plant Molecular Biology Reporter* **30**: 370-378.

Handakumbura PP, Hazen SP. 2012. Transcriptional regulation of grass secondary cell wall biosynthesis: playing catch-up with *Arabidopsis thaliana*. *Frontiers in Plant Science* **3**: 74.

Haring M, Offermann S, Danker T, Horst I, Peterhansel C, Stam M. 2007. Chromatin immunoprecipitation: optimization, quantitative analysis and data normalization. *Plant Methods* **3**: 11.

Hatton D, Sablowski R, Vung M-H, Smith C, Schuch W, Bevan M. 1995. Two classes of *cis* sequences contribute to tissue-specific expression of a *PAL2* promoter in transgenic tobacco. *The Plant Journal* **7**(6): 859-876.

Hertzberg M, Aspeborg H, Schrader J, Andersson A, Erlandsson R, Blomqvist K, Bhalerao R, Uhlén M, Teeri TT, Lundeberg J, Sundberg B, Nilsson P, Sandberg G. 2001. A transcriptional roadmap to wood formation. *Proceedings of the National Academy of Sciences of the USA* **98**(25): 14732–14737.

Hinchee MAW, Mullinax LN, Rottmann WH. 2010. Woody biomass and purpose-grown trees as feedstocks for renewable energy. *Biotechnology in Agriculture and Forestry* **66**(2): 155-208.

Hirakawa Y, Kondo Y, Fukuda H. 2010. TDIF peptide signaling regulates vascular stem cell proliferation via the *WOX4* homeobox gene in *Arabidopsis*. *The Plant Cell* **22**: 2618-2629.

Hirakawa Y, Kondo Y, Fukuda H. 2011. Establishment and maintenance of vascular cell communities through local signaling. *Current Opinion in Plant Biology* **14**: 17-23.

Hiratsu K, Matsui K, Koyama T, Ohme-Takagi M. 2003. Dominant repression of target genes by chimeric repressors that include the EAR motif, a repression domain, in *Arabidopsis*. *The Plant Journal* **34**(5): 733-739.

Hiratsu K, Mitsuda N, Matsui K, Ohme-Takagi M. 2004. Identification of the minimal repression domain of SUPERMAN shows that the DLELRL hexapeptide is both necessary and sufficient for repression of transcription in *Arabidopsis*. *Biochemical and Biophysical Research Communications* **321**(1): 172-178.

© University of Pretoria

Hussey SG, Mizrachi E, Spokevicius AV, Bossinger G, Berger DK, Myburg AA. 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**: 173.

Ilegems M, Douet V, Meylan-Bettex M, Uyttewaal M, Brand L, Bowman JL, Stieger PA. 2010. Interplay of auxin, KANADI and Class III HD-ZIP transcription factors in vascular tissue formation. *Development* **137**: 975-984.

Ingvarsson PK, Street NR. 2010. Association genetics of complex traits in plants. *New Phytologist* **189**: 909-922.

Israelsson M, Eriksson ME, Hertzberg M, Aspeborg H, Nilsson P, Moritz T. 2003. Changes in gene expression in the wood-forming tissue of transgenic hybrid aspen with increased secondary growth. *Plant Molecular Biology* **52**(4): 893-903.

Ito Y, Nakanomyo I, Motose H, Iwamoto K, Sawa S, Dohmae N, Fukuda H. 2006. Dodeca-CLE peptides as suppressors of plant stem cell differentiation. *Science* **313**(5788): 842-845.

Iwakawa H, Ueno Y, Semiarti E, Onouchi H, Kojima S, Tsukaya H, Hasebe M, Soma T, Ikezaki M, Machida C, Machida Y. 2002. The *ASYMMETRIC LEAVES2* gene of *Arabidopsis thaliana*, required for formation of a symmetric flat leaf lamina, encodes a member of a novel family of proteins characterized by cysteine repeats and a leucine zipper. *Plant and Cell Physiology* **43**(5): 467-478.

Izhaki A, Bowman JL. 2007. KANADI and class III HD-Zip gene families regulate embryo patterning and modulate auxin flow during embryogenesis in *Arabidopsis*. *The Plant Cell* **19**: 495–508.

Jansen RC, Nap J-P. 2001. Genetical genomics: the added value from segregation. *Trends in Genetics* **17**(7): 388–391.

Jothi R, Balaji, Wuster A, Grochow JA, Gsponer J, Przytycka TM, Aravind L, Babu MM. 2009. Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture. *Molecular Systems Biology* **5**: 294.

Jung JH, Park CM. 2007. Vascular development in plants: specification of xylem and phloem tissues. *Journal of Plant Biology* **50**(3): 301-305.

Kærn M, Elston TC, Blake WJ, Collins JJ. 2005. Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics* **6**: 451-464.

Kannangara R, Branigan C, Liu Y, Penfield T, Rao V, Mouille G, Höfte H, Pauly M, Riechmann JL, Broun P. 2007. The transcription factor WIN1/SHN1 regulates cutin biosynthesis in *Arabidopsis thaliana*. *The Plant Cell* **19**(4): 1278-1294.

© University of Pretoria

Kaufmann K, Muiño JM, Jauregui R, Airoldi CA, Smaczniak C, Krajewski P, Angenent GC. 2009. Target genes of the MADS transcription factor SEPALLATA3: integration of developmental and hormonal pathways in the *Arabidopsis* flower. *PLoS Biology* **7**(4): e1000090.

Kaufmann K, Wellmer F, Muiño JM, Ferrier T, Wuest SE, Kumar V, Serrano-Mislata A, Madueño F, Krajewski P, Meyerowitz EM, Angenent GC, Riechmann JL. 2010. Orchestration of floral initiation by APETALA1. *Science* **328**: 85-89.

Kim J-Y, Yuan Z, Jackson D. 2003. Developmental regulation and significance of KNOX protein trafficking in *Arabidopsis*. *Development* **130**: 4351-4362.

Kim J, Jung J-H, Reyes JL, Kim Y-S, Kim S-Y, Chung K-S, Kim JA, Lee M, Lee Y, Kim VN, Chua N-H, Park C-M. 2005. microRNA-directed cleavage of *ATHB15* mRNA regulates vascular development in *Arabidopsis* inflorescence stems. *The Plant Journal* **42**: 84–94.

Kim W-C, Ko J-H, Han K-H. 2012a. Identification of a cis-acting regulatory motif recognized by MYB46, a master transcriptional regulator of secondary wall biosynthesis. *Plant Molecular Biology* **78**: 489–501.

Kim W-C, Ko J-H, Kim J-Y, Kim J-M, Bae H-J, Han K-H. 2012b. MYB46 directly regulates the gene expression of secondary wall-associated cellulose synthases in *Arabidopsis*. *The Plant Journal* **73**: 26-36.

Kliebenstein D. 2009. Quantitative genomics: analyzing intraspecific variation using global gene expression polymorphisms or eQTLs. *Annual Reviews of Plant Biology* **60**: 93–114.

Ko J-H, Kim W-C, Han K-H. 2009. Ectopic expression of MYB46 identifies transcriptional regulatory genes involved in secondary wall biosynthesis in *Arabidopsis*. *The Plant Journal* **60**(4): 649-665.

Ko J-H, Yang SH, Park AH, Lerouxel O, Han K-H. 2007. ANAC012, a member of the plant-specific NAC transcription factor family, negatively regulates xylary fiber development in *Arabidopsis thaliana*. *The Plant Journal* **50**(6): 1035-1048.

Kondou Y, Higuchi M, Matsui M. 2010. High-throughput characterization of plant gene functions by using gain-of-function technology. *Annual Review of Plant Biology* **61**: 373–393.

Krakauer DC, Nowak MA. 1999. Evolutionary preservation of redundant duplicated genes. *Seminars in Cell and Developmental Biology* **10**: 555-559.

Kubo M, Udagawa M, Nishikubo N, Horiguchi G, Yamaguchi M, Ito J, Mimura T, Fukuda H, Demura T. 2005. Transcription switches for protoxylem and metaxylem vessel formation. *Genes and Development* **19**: 1855-1860.

© University of Pretoria

Kurata T, Okada K, Wada T. 2005. Intercellular movement of transcription factors. *Current Opinion in Plant Biology* **8**: 600-605.

Li E, Bhargava A, Qiang W, Friedmann MC, Forneris N, Savidge RA, Johnson LA, Mansfield SD, Ellis BE, Douglas CJ. 2012a. The Class II *KNOX* gene *KNAT7* negatively regulates secondary wall formation in *Arabidopsis* and is functionally conserved in *Populus*. *New Phytologist* **194**(1): 102–115.

Li E, Wang S, Liu Y, Chen J-G, Douglas CJ. 2011. OVATE FAMILY PROTEIN4 (OFP4) interaction with KNAT7 regulates secondary cell wall formation in *Arabidopsis thaliana*. *The Plant Journal* **67**(2): 328-341.

Li JJ, Herskowitz I. 1993. Isolation of ORC6, a component of the yeast origin recognition complex by a one-hybrid system. *Science* **262**(5141): 1870-1874.

Li Q, Lin Y-C, Sun Y-H, Song J, Chen H, Zhang X-H, Sederoff RR, Chiang VL. 2012b. Splice variant of the SND1 transcription factor is a dominant negative of SND1 members and their regulation in *Populus trichocarpa*. *Proceedings of the National Academy of Sciences of the USA* **109**(36): 14699-14704.

Lin L, Young N, Handakumbura P, Breton G, Mockler TC, Kay SA, Hazen SP 2010. A protein-DNA interaction network for cell wall biosynthesis. *Second Annual TIMBR Conference on Cellulosic Biofuels*. Amherst: University of Massachusetts Amherst.

Lois R, Dietrich A, Hahlbrock K, Schulz W. 1989. A phenylalanine ammonia-lyase gene from parsley: structure, regulation and identification of elicitor and light responsive cis-acting elements. *The EMBO Journal* **8**(6): 1641-1648.

Longabaugh WJR, Davidson EH, Bolouri H. 2005. Computational representation of developmental genetic regulatory networks. *Developmental Biology* **283**: 1–16.

Lopato S, Bazanova N, Morran S, Milligan AS, Shirley N, Langridge P. 2006. Isolation of plant transcription factors using a modified yeast one-hybrid system. *Plant Methods* **2**: 3.

Love J, Björklund S, Vahala J, Hertzberg M, Kangasjärvi J, Sundberg B. 2009. Ethylene is an endogenous stimulator of cell division in the cambial meristem of *Populus*. *Proceedings of the National Academy of Sciences of the USA* **106**(14): 5984–5989.

Lucas WJ, Bouché-Pillon S, Jackson DP, Nguyen L, LucianBaker, Ding B, Hake S. 1995. Selective trafficking of KNOTTED1 homeodomain protein and its mRNA through plasmodesmata. *Science* **270**(5244): 1980-1983.

Lucas WJ, Groover A, Lichtenberger R, Furuta K, Yadav S-R, Helariutta Y, He X-Q, Fukuda H, Kang J, Brady SM, Patrick JW, Sperry J, Yoshida A, López-Millán A-F, Grusak MA,

© University of Pretoria

Kachroo P. 2013. The plant vascular system: evolution, development and functions. *Journal of Integrative Plant Biology* **55**(4): 294-388.

MacNeil LT, Walhout AJM. 2011. Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome Research* **21**: 645–657.

Malitsky S, Blum E, Less H, Venger I, Elbaz M, Morin S, Eshed Y, Aharoni A. 2008. The transcript and metabolite networks affected by the two clades of *Arabidopsis* glucosinolate biosynthesis regulators. *Plant Physiology* **148**: 2021-2049.

Mallory AC, Reinhart BJ, Jones-Rhoades MW, Tang G, Zamore PD, Barton MK, Bartel DP. 2004. MicroRNA control of *PHABULOSA* in leaf development: importance of pairing to the microRNA 5' region. *The EMBO Journal* **23**: 3356–3364.

Mauriat M, Moritz T. 2009. Analyses of *GA20ox*- and *GID1*-over-expressing aspen suggest that gibberellins play two distinct roles in wood formation. *The Plant Journal* **58**: 989–1003.

McCann M, Rose J. 2010. Blueprints for building plant cell walls. *Plant Physiology* **153**: 365.

McCarthy RL, Zhong R, Fowler S, Lyskowski D, Piyasena H, Carleton K, Spicer C, Ye Z-H. 2010. The poplar MYB transcription factors, PtrMYB3 and PtrMYB20, are involved in the regulation of secondary wall biosynthesis. *Plant and Cell Physiology* **51**(6): 1084–1090.

McCarthy RL, Zhong R, 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 and Cell Physiology* **50**(11): 1950-1964.

McHale NA, Koning RE. 2004. MicroRNA-directed cleavage of *Nicotiana sylvestris PHAVOLUTA* mRNA regulates the vascular cambium and structure of apical meristems. *The Plant Cell* **16**: 1730–1740.

Mejia-Guerra MK, Pomeranz M, Morohashi K, Grotewold E. 2012. From plant gene regulatory grids to network dynamics. *Biochimica et Biophysica Acta* **1819**: 454–465.

Mele G, Ori N, Sato Y, Hake S. 2003. The *knotted1*-like homeobox gene *BREVIPEDICELLUS* regulates cell differentiation by modulating metabolic pathways. *Genes and Development* **17**: 2088–2093.

Mellerowicz EJ, Sundberg B. 2008. Wood cell walls: biosynthesis, developmental dynamics and their implications for wood properties. *Current Opinion in Plant Biology* **11**: 293–300.

Milhinhos A, Miguel CM. 2013. Hormone interactions in xylem development: a matter of signals. *Plant Cell Reports* **32**(6): 867-883.

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. 2002. Network motifs: simple building blocks of complex networks. *Science* **298**: 824-827.

© University of Pretoria

Mitsuda N, Ikeda M, Takada S, Takiguchi Y, Kondou Y, Yoshizumi T, Fujita M, Shinozaki K, Matsui M, Ohme-Takagi M. 2010. Efficient yeast one-/two-hybrid screening using a library composed only of transcription factors in *Arabidopsis thaliana*. *Plant and Cell Physiology* **51**(12): 2145–2151.

Mitsuda N, Iwase A, Yamamoto H, Yoshida M, Seki M, Shinozaki K, 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.

Mitsuda N, Ohme-Takagi M. 2008. NAC transcription factors NST1 and NST3 regulate pod shattering in a partially redundant manner by promoting secondary wall formation after the establishment of tissue identity. *The Plant Journal* **56**(5): 768-778.

Mitsuda N, Seki M, Shinozaki K, Ohme-Takagi M. 2005. The NAC transcription factors NST1 and NST2 of *Arabidopsis* regulate secondary wall thickenings and are required for anther dehiscence. *Plant Cell* **17**: 2993–3006.

Miyashima S, Koi S, Hashimoto T, Nakajima K. 2011. Non-cell-autonomous microRNA165 acts in a dose-dependent manner to regulate multiple differentiation status in the *Arabidopsis* root. *Development* **138**: 2303-2313.

Mizrachi E, Mansfield SD, Myburg AA. 2012. Cellulose factories: advancing bioenergy production from forest trees. *New Phytologist* **194**: 54-62.

Motose H, Sugiyama M, Fukuda H. 2004. A proteoglycan mediates inductive interaction during plant vascular development. *Nature* **429**: 873-878.

Mukherjee S, Berger MF, Jona G, Wang XS, Muzzey D, Snyder M, Young RA, Bulyk ML. 2004. Rapid analysis of the DNA binding specificities of transcription factors with DNA microarrays. *Nature Genetics* **36**(12): 1331-1339.

Nakano Y, Nishikubo N, Goué N, Ohtani M, Yamaguchi M, Katayama Y, Demura T. 2010. MYB transcription factors orchestrating the developmental program of xylem vessels in *Arabidopsis* roots. *Plant Biotechnology* **27**: 267-272.

Nelson DE, Ihekwaba AEC, Elliott M, Johnson JR, Gibney CA, Foreman BE, Nelson G, See V, Horton CA, Spiller DG, Edwards SW, McDowell HP, Unitt JF, Sullivan E, Grimley R, Benson N, Broomhead D, Kell DB, White MRH. 2004. Oscillations in NF-κB signaling control the dynamics of gene expression. *Science* **306**: 704-708.

Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, Stamatoyannopoulos JA. 2012. Circuitry and dynamics of human transcription factor regulatory networks. *Cell* **150**: 1274–1286.

© University of Pretoria

Nowak MA, Boerlijst MC, Cooke J, Smith JM. 1997. Evolution of genetic redundancy. *Nature* **388**: 167-171.

Oda Y, Mimura T, Hasezawa S. 2005. Regulation of secondary cell wall development by cortical microtubules during tracheary element differentiation in *Arabidopsis* cell suspensions. *Plant Physiology* **137**: 1027-1036.

Ohashi-Ito K, Fukuda H. 2003. HD-Zip III homeobox genes that include a novel member, *ZeHB-13* (*Zinnia*)/ *ATHB-15* (*Arabidopsis*), are involved in procambium and xylem cell differentiation. *Plant and Cell Physiology* **44**(12): 1350–1358.

Ohashi-Ito K, Fukuda H. 2010. Transcriptional regulation of vascular cell fates. *Current Opinion in Plant Biology* **13**: 670–676.

Ohashi-Ito K, Oda Y, Fukuda H. 2010. *Arabidopsis* VASCULAR-RELATED NAC-DOMAIN6 directly regulates the genes that govern programmed cell death and secondary wall formation during xylem differentiation. *The Plant Cell* **22**: 3461–3473.

Öhman D, Demedts B, Kumar M, Gerber L, Gorzsás A, Goeminne G, Hedenström M, Ellis B, Boerjan W, Sundberg B. 2012. MYB103 is required for *FERULATE-5-HYDROXYLASE* expression and syringyl lignin biosynthesis in *Arabidopsis* stems. *The Plant Journal* **73**(1): 63–76.

Ohtani M, Nishikubo N, Xu B, Yamaguchi M, Mitsuda N, Goue N, Shi F, Ohme-Takagi M, Demura T. 2011. A NAC domain protein family contributing to the regulation of wood formation in poplar. *The Plant Journal* **67**(3): 499-512.

Ooka H, Satoh K, Doi K, Nagata T, Otomo Y, Murakami K, Matsubara K, Osato N, Kawai J, Carninci P, Hayashizaki Y, Suzuki K, Kojima K, Takahara Y, Yamamoto K, Kikuchi S. 2003. Comprehensive analysis of NAC family genes in *Oryza sativa* and *Arabidopsis thaliana*. *DNA Research* **10**(6): 239-247.

Opper M, Sanguinetti G. 2010. Learning combinatorial transcriptional dynamics from gene expression data. *Bioinfomatics* **26**(13): 1623–1629.

Orphanides G, Lagrange T, Reinberg D. 2006. The general transcription factors of RNA polymerase II. *Genes and Development* **10**: 2657-2683.

Park MY, Kang J-y, Kim SY. 2011. Overexpression of AtMYB52 confers ABA hypersensitivity and drought tolerance. *Molecules and Cells* **31**: 447-454.

Patzlaff A, McInnis S, Courtenay A, Surman C, Newman LJ, Smith C, Bevan MW, Mansfield S, Whetten RW, Sederoff RR, Campbell MM. 2003. Characterisation of a pine MYB that regulates lignification. *The Plant Journal* **36**(6): 743-754.

© University of Pretoria

Pesquet E, Tuominen H. 2011. Ethylene stimulates tracheary element differentiation in *Zinnia elegans* cell cultures. *New Phytologist* **190**: 138–149.

Pesquet E, Zhang B, Gorzsás A, Puhakainen T, Serk H, Escamez S, Barbier O, Gerber L, Courtois-Moreau C, Alatalo E, Paulin L, Kangasjärvi J, Sundberg B, Goffner D, Tuominena H. 2013. Non-cell-autonomous postmortem lignification of tracheary elements in *Zinnia elegans*. *The Plant Cell* **25**(4): 1314-1328.

Petersen PD, Lau J, Ebert B, Yang F, Verhertbruggen Y, Kim JS, Varanasi P, Suttangkakul A, Auer M, Loqué D, Scheller HV. 2012. Engineering of plants with improved properties as biofuels feedstocks by vessel-specific complementation of xylan biosynthesis mutants. *Biotechnology for Biofuels* **2012**(5): 84.

Pimrote K, Tian Y, Lu X. 2012. Transcriptional regulatory network controlling secondary cell wall biosynthesis and biomass production in vascular plants. *African Journal of Biotechnology* **11**(75): 13928-13937.

Plomion C, Leprovost G, Stokes A. 2001. Wood formation in trees. *Plant Physiology* **127**: 1513-1523.

Preston J, Wheeler J, Heazlewood J, Li SF, Parish RW. 2004. AtMYB32 is required for normal pollen development in *Arabidopsis thaliana*. *The Plant Journal* **40**: 979–995.

Prigge MJ, Otsuga D, Alonso JM, Ecker JR, Drews GN, Clark SE. 2005. Class III Homeodomain-Leucine Zipper gene family members have overlapping, antagonistic, and distinct roles in *Arabidopsis* development. *The Plant Cell* **17**: 61–76.

Prouse MB, Campbell MM. 2012. The interaction between MYB proteins and their target DNA binding sites. *Biochimica et Biophysica Acta* **1819**: 67–77.

Pyo H, Demura T, Fukuda H. 2007. TERE; a novel *cis*-element responsible for a coordinated expression of genes related to programmed cell death and secondary wall formation during differentiation of tracheary elements. *The Plant Journal* **51**: 955–965.

Raes J, Rohde A, Christensen JH, Peer YV, Boerjan W. 2003. Genome-wide characterization of the lignification toolbox in *Arabidopsis*. *Plant Physiology* **133**: 1051-1071.

Rahantamalala A, Rech P, Martinez Y, Chaubet-Gigot N, Grima-Pettenati J, Pacquit V. 2010. Coordinated transcriptional regulation of two key genes in the lignin branch pathway – *CAD* and *CCR* – is mediated through MYB- binding sites. *BMC Plant Biology* **10**: 130.

Ramírez V, Agorio A, Coego A, García-Andrade J, Hernández MJ, Balaguer B, Ouwerkerk PBF, Zarra I, Vera P. 2011. MYB46 modulates disease susceptibility to *Botrytis cinerea* in *Arabidopsis*. *Plant Physiology* **155**: 1920–1935.

© University of Pretoria

Ranocha P, Denancé N, Vanholme R, Freydier A, Martinez Y, Hoffmann L, Köhler L, Pouzet C, Renou J-P, Sundberg B, Boerjan W, Goffner D. 2010. *Walls are thin 1* (*WAT1*), an *Arabidopsis* homolog of *Medicago truncatula NODULIN21*, is a tonoplast-localized protein required for secondary wall formation in fibers. *The Plant Journal* **63**: 469–483.

Reece-Hoyes JS, Diallo A, Lajoie B, Kent A, Shrestha S, Kadreppa S, Pesyna C, Dekker J, Myers CL, Walhout AJM. 2011. Enhanced yeast one-hybrid (eY1H) assays for high-throughput gene-centered regulatory network mapping. *Nature Methods* **8**(12): 1059–1064.

Reece-Hoyes JS, Walhout AJM. 2012. Yeast one-hybrid assays: A historical and technical perspective. *Methods* **57**(4): 441-447.

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA. 2000. Genome-wide location and function of DNA binding proteins. *Science* **290**: 2306-2309.

Rhee HS, Pugh BF. 2011. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**: 1408–1419.

Rockwood DL, Rudie AW, Ralph SA, Zhu JY, Winandy JE. 2008. Energy product options for *Eucalyptus* species grown as short rotation woody crops. *International Journal of Molecular Sciences* **9**: 1361-1378.

Rodriguez-Uribe L, O'Connell MA. 2006. A root-specific bZIP transcription factor is responsive to water deficit stress in tepary bean (*Phaseolus acutifolius*) and common bean (*P. vulgaris*). *Journal of Experimental Botany* **57**(6): 1391-1398.

Romero I, Fuertes A, Benito MJ, Malpica JM, Leyva A, Paz-Ares J. 1998. More than 80 *R2R3-MYB* regulatory genes in the genome of *Arabidopsis thaliana*. *The Plant Journal* **14**(3): 273-284.

Rosenfeld N, Elowitz MB, Alon U. 2002. Negative autoregulation speeds the response times of transcription networks. *Journal of Molecular Biology* **323**: 785–793.

Sablowski RWM, Moyano E, A.Culianez-Macia F, Schuch W, Martin C, Bevan M. 1994. A flower-specific Myb protein activates transcription of phenylpropanoid biosynthetic genes. *The EMBO Journal* **13**(1): 128-137.

Sanchez P, Nehlin L, Greb T. 2012. From thin to thick: major transitions during stem development. *Trends in Plant Science* **17**(2): 113-121.

Sayyed-Ahmad A, Tuncay K, Ortoleva PJ. 2007. Transcriptional regulatory network refinement and quantification through kinetic modeling, gene expression microarray data and information theory. *BMC Bioinformatics* **8**: 20.

© University of Pretoria

Schrader J, Baba K, May ST, Palme K, Bennett M, Bhalerao RP, Sandberg G. 2003. Polar auxin transport in the wood-forming tissues of hybrid aspen is under simultaneous control of developmental and environmental signals. *Proceedings of the National Academy of Sciences of the USA* **100**(17): 10096-10101.

Schrader J, Nilsson J, Mellerowicz E, Berglund A, Nilsson P, Hertzberg M, Sandberg G. 2004. A high-resolution transcript profile across the wood-forming meristem of poplar identifies potential regulators of cambial stem cell identity. *The Plant Cell* **16**: 2278–2292.

Schuetz M, Smith R, Ellis B. 2013. Xylem tissue specification, patterning, and differentiation mechanisms. *Journal of Experimental Botany* **64**(1): 11–31.

Shannon P, Markiel A, Ozier O, Baliga N, Wang J, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* **13**(11): 2498-2504.

Shen H, He X, Poovaiah CR, Wuddineh WA, Ma J, Mann DGJ, Wang H, Jackson L, Tang Y, Jr CNS, Chen F, Dixon RA. 2011. Functional characterization of the switchgrass (*Panicum virgatum*) R2R3-MYB transcription factor *PvMYB4* for improvement of lignocellulosic feedstocks. *New Phytologist* **193**(1): 121–136.

Shen H, Yin Y, Chen F, Xu Y, Dixon RA. 2009. A bioinformatic analysis of *NAC* genes for plant cell wall development in relation to lignocellulosic bioenergy production. *BioEnergy Research* **2**: 217-232.

Shuai B, Reynaga-Peña CG, Springer PS. 2002. The *LATERAL ORGAN BOUNDARIES* gene defines a novel, plant-specific gene family. *Plant Physiology* **129**: 747-761.

Sonbol F-M, Fornalé S, Capellades M, Encina A, Touriño S, Torres J-L, Rovira P, Ruel K, Puigdomènech P, Rigau J, Caparrós-Ruiz D. 2009. The maize *ZmMYB42* represses the phenylpropanoid pathway and affects the cell wall structure, composition and degradability in *Arabidopsis thaliana*. *Plant Molecular Biology* **70**(3): 283-296.

Sorce C, Giovannelli A, Sebastiani L, Anfodillo T. 2013. Hormonal signals involved in the regulation of cambial activity, xylogenesis and vessel patterning in trees. *Plant Cell Reports* **32**(6): 885-898.

Soyano T, Thitamadee S, Machida Y, Chua N-H. 2008. *ASYMMETRIC LEAVES2-LIKE19/LATERAL ORGAN BOUNDARIES DOMAIN30* and *ASL20/LBD18* regulate tracheary element differentiation in *Arabidopsis*. *The Plant Cell* **20**: 3359-3373.

Spitz F, Furlong EEM. 2012. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics* **13**: 613-626.

© University of Pretoria

Sun H, Huang X, Xu X, Lan H, Huang J, Zhang H-S. 2011. ENAC1, a NAC transcription factor, is an early and transient response regulator induced by abiotic stress in rice (*Oryza sativa* L.). *Molecular Biotechnology* **52**(2): 101-110.

Suzuki S, Suda K, Sakurai N, Ogata Y, Hattori T, Suzuki H, Shibata D, Umezawa T. 2011. Analysis of expressed sequence tags in developing secondary xylem and shoot of *Acacia mangium*. *Journal of Wood Science* **57**: 40-46.

Taketa S, Amano S, Tsujino Y, Sato T, Saisho D, Kakeda K, Nomura M, Suzuki T, Matsumoto T, Sato K, Kanamori H, Kawasaki S, Takeda K. 2008. Barley grain with adhering hulls is controlled by an ERF family transcription factor gene regulating a lipid biosynthesis pathway. *Proceedings of the National Academy of Sciences of the USA* **105**(10): 4062–4067.

Tang B, Hsu H-K, Hsu P-Y, Bonneville R, Chen S-S, Huang TH-M, Jin VX. 2012. Hierarchical modularity in ERα transcriptional network is associated with distinct functions and implicates clinical outcomes. *Scientific Reports* **2**: 875.

Tang G, Reinhart BJ, Bartel DP, Zamore PD. 2003. A biochemical framework for RNA silencing in plants. *Genes and Development* **17**: 49-63.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield TK, Diegel M, Dunn D, Ebersol AK, Frum T, Giste E, Johnson AK, Johnson EM, Kutyavin T, Lajoie B, Lee B-K, Lee K, London D, Lotakis D, Neph S, Neri F, Nguyen ED, Reynolds HQAP, Roach V, Safi A, Sanchez ME, Sanyal A, Shafer A, Simon JM, Song L, Vong S, Weaver M, Yan Y, Zhang Z, Zhang Z, Lenhard B, Tewari M, Dorschner MO, Hansen RS, Navas PA, Stamatoyannopoulos G, Lieb VRIJD, Sunyaev SR, Akey JM, Sabo PJ, Kaul R, Furey TS, Dekker J, Crawford GE, Stamatoyannopoulos JA. 2012. The accessible chromatin landscape of the human genome. *Nature* **489**: 75-82.

Tran L-SP, Nakashima k, Sakuma Y, Simpson SD, Fujita Y, Maruyama K, Fujita M, Seki M, Shinozaki K, Yamaguchi-Shinozaki K. 2004. Isolation and functional analysis of *Arabidopsis* stress-inducible NAC transcription factors that bind to a drought-responsive *cis*-element in the *early responsive to dehydration stress 1* promoter. *The Plant Cell* **16**: 2481-2498.

Turner S, Gallois P, Brown D. 2007. Tracheary element differentiation. *Annual Reviews of Plant Biology* **58**: 407–433.

Umezawa T. 2009. The cinnamate/monolignol pathway. *Phytochemistry Reviews* **9**(1): 1-17.

© University of Pretoria

Valdivia ER, Herrera MT, Gianzo C, Fidalgo J, Revilla G, Zarra I, Sampedro J. 2013. Regulation of secondary wall synthesis and cell death by NAC transcription factors in the monocot *Brachypodium distachyon*. *Journal of Experimental Botany* **64**(5): 1333-1343.

Vanholme R, Storme V, Vanholme B, Sundin L, Christensen JH, Goeminne G, Halpin C, Rohde A, Morreel K, Boerjana W. 2012. A systems biology view of responses to lignin biosynthesis perturbations in *Arabidopsis*. *The Plant Cell* **24**: 3506–3529.

Wada T, Kurata T, Tominaga R, Koshino-Kimura Y, Tachibana T, Goto K, Marks MD, Shimura Y, Okada K. 2002. Role of a positive regulator of root hair development, *CAPRICE*, in *Arabidopsis* root epidermal cell differentiation. *Development* **129**: 5409-5419.

Walhout AJM. 2006. Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping. *Genome Research* **16**: 1445–1454.

Wang H-Z, Dixon RA. 2011. On–off switches for secondary cell wall biosynthesis. *Molecular Plant* **5**(2): 297-303.

Wang H, Avci U, Nakashima J, Hahn MG, Chen F, Dixona RA. 2010. Mutation of WRKY transcription factors initiates pith secondary wall formation and increases stem biomass in dicotyledonous plants. *Proceedings of the National Academy of Science* **107**(51): 22338–22343.

Wang H, Zhao Q, Chen F, Wang M, Dixon RA. 2011. NAC domain function and transcriptional control of a secondary cell wall master switch. *The Plant Journal* **68**(6): 1104–1114.

Wang M, Reed R. 1993. Molecular cloning of the olfactory neuronal transcription factor Olf-1 by genetic selection in yeast. *Nature* **364**: 121-126.

Wang S, Chang Y, Guo J, Chen J-G. 2007. *Arabidopsis* Ovate Family Protein 1 is a transcriptional repressor that suppresses cell elongation. *The Plant Journal* **50**(5): 858–872.

Wang Y-S, Yoo C-M, Blancaflor EB. 2007. Improved imaging of actin filaments in transgenic *Arabidopsis* plants expressing a green fluorescent protein fusion to the C- and N-termini of the fimbrin actin-binding domain 2. *New Phytologist* **177**(2): 525-536.

Wehner N, Hartmann L, Ehlert A, Böttner S, Oñate-Sánchez L, Dröge-Laser W. 2011. High-throughput protoplast transactivation (PTA) system for the analysis of *Arabidopsis* transcription factor function. *The Plant Journal* **68**(3): 560–569.

Welner DH, Lindemose S, Grossmann JG, Møllegaards NE, Olsen AN, Helgstrand C, Skriver K, Leggio LL. 2012. DNA binding by the plant-specific NAC transcription factors in crystal and solution: a firm link to WRKY and GCM transcription factors. *Biochemical Journal* **444**: 395–404.

© University of Pretoria

Williams L, Grigg SP, Xie M, Christensen S, Fletcher JC. 2005. Regulation of *Arabidopsis* shoot apical meristem and lateral organ formation by microRNA *miR166g* and its *AtHD-ZIP* target genes. *Development* **132**: 3657-3668.

Winzell A, Aspeborg H, Wang Y, Ezcurra I. 2010. Conserved CA-rich motifs in gene promoters of PtxtMYB021-responsive secondary cell wall carbohydrate-active enzymes in *Populus*. *Biochemical and Biophysical Research Communications* **394**: 848–853.

Wu S, Gallagher KL. 2012. Transcription factors on the move. *Current Opinion in Plant Biology* **15**: 645-651.

Xie L, Yang C, Wang X. 2011. Brassinosteroids can regulate cellulose biosynthesis by controlling the expression of *CESA* genes in *Arabidopsis*. *Journal of Experimental Botany* **62**(13): 4495-4506.

Yamaguchi M, Demura T. 2010. Transcriptional regulation of secondary wall formation controlled by NAC domain proteins. *Plant Biotechnology* **27**: 237-242.

Yamaguchi M, Kubo M, Fukuda H, Demura T. 2008. VASCULAR-RELATED NAC-DOMAIN7 is involved in the differentiation of all types of xylem vessels in *Arabidopsis* roots and shoots. *The Plant Journal* **55**: 652–664.

Yamaguchi M, Mitsuda N, Ohtani M, Ohme-Takagi M, Kato K, 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, Nadia G, Igarashi H, Ohtani M, Nakano Y, Mortimer JC, Nishikubo N, Kubo M, Katayama Y, Kakegawa K, Dupree P, Demura T. 2010a. VASCULAR-RELATED NAC-DOMAIN6 (VND6) and VND7 effectively induce transdifferentiation into xylem vessel elements under control of an induction system. *Plant Physiology* **153**: 906-914.

Yamaguchi M, Ohtani M, Mitsuda N, Kubo M, Ohme-Takagi M, Fukuda H, Demura T. 2010b. VND-INTERACTING2, a NAC domain transcription factor, negatively regulates xylem vessel formation in *Arabidopsis*. *Plant Cell* **22**(4): 1249-1263

Yang C, Xu Z, Song J, Conner K, Barrena GV, Wilson ZA. 2007. *Arabidopsis MYB26/MALE STERILE35* regulates secondary thickening in the endothecium and is essential for anther dehiscence. *The Plant Cell* **19**: 534-548.

Yang F, Mitra P, Zhang L, Prak L, Verhertbruggen Y, Kim J-S, Sun L, Zheng K, Tang K, Auer M, Scheller HV, Loqué D. 2013. Engineering secondary cell wall deposition in plants. *Plant Biotechnology Journal* **11**: 325-335.

Yuan JS, Galbraith DW, Dai SY, Griffin P, Jr. CNS. 2008. Plant systems biology comes of age. *Trends in Plant Science* **13**(4): 165-171.

© University of Pretoria

Zambryski P, Crawford K. 2000. Plasmodesmata: gatekeepers for cell-to-cell transport of developmental signals in plants. *Annual Review of Cell and Developmental Biology* **16**: 393-421.

Zhang J. 2003. Evolution by gene duplication: an update. *Trends in Ecology and Evolution* **18**(6): 292-298.

Zhang J, Elo A, Helariutta Y. 2010. *Arabidopsis* as a model for wood formation. *Current Opinion in Biotechnology* **22**: 1-7.

Zhao C, Avci U, Grant EH, Haigler CH, Beers EP. 2008. XND1, a member of the NAC domain family in *Arabidopsis thaliana*, negatively regulates lignocellulose synthesis and programmed cell death in xylem. *The Plant Journal* **53**(3): 425-436.

Zhao Q, Dixon RA. 2011. Transcriptional networks for lignin biosynthesis: more complex than we thought? *Trends in Plant Science* **16**(4): 227-233.

Zhao Q, Gallego-Giraldo L, Wang H, Zeng Y, Ding S-Y, Chen F, Dixon RA. 2010a. An NAC transcription factor orchestrates multiple features of cell wall development in *Medicago truncatula*. *The Plant Journal* **63**: 100–114.

Zhao Q, Wang H, Yin Y, Xu Y, Chen F, Dixon RA. 2010b. Syringyl lignin biosynthesis is directly regulated by a secondary cell wall master switch. *Proceedings of the National Academy of Science of the USA* **107**(32): 14496–14501.

Zhong R, Demura T, Ye ZH. 2006. SND1, a NAC domain transcription factor, is a key regulator of secondary wall synthesis in fibers of *Arabidopsis*. *Plant Cell* **18**(11): 3158-3170.

Zhong R, Lee C, McCarthy RL, Reeves CK, Jones EG, Ye Z-H. 2011a. Transcriptional activation of secondary wall biosynthesis by rice and maize NAC and MYB transcription factors. *Plant and Cell Physiology* **52**(10): 1856–1871.

Zhong R, Lee C, Ye Z-H. 2010a. Evolutionary conservation of the transcriptional network regulating secondary cell wall biosynthesis. *Trends in Plant Science* **15**(11): 625-632.

Zhong R, Lee C, Ye Z-H. 2010b. Functional characterization of poplar wood-associated NAC domain transcription factors. *Plant Physiology* **152**: 1044–1055.

Zhong R, Lee C, Ye Z-H. 2010c. Global analysis of direct targets of secondary wall NAC master switches in *Arabidopsis*. *Molecular Plant* **3**(6): 1087-1103.

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

© University of Pretoria

Zhong R, McCarthy RL, Lee C, Ye Z-H. 2011b. Dissection of the transcriptional program regulating secondary wall biosynthesis during wood formation in poplar. *Plant Physiology* **157**: 1452–1468.

Zhong R, Richardson EA, Lee C, Zhou J, McCarthy R, Ye Z-H. 2008. Transcriptional regulation of secondary wall biosynthesis in plants. *Microscopy and Microanalysis* **14**: 1504-1505.

Zhong R, Richardson EA, Y Z-H. 2007a. The MYB46 transcription factor is a direct target of SND1 and regulates secondary wall biosynthesis in *Arabidopsis*. *The Plant Cell* **19**: 2776-2792.

Zhong R, Richardson EA, Ye Z-H. 2007b. Two NAC domain transcription factors, SND1 and NST1, function redundantly in regulation of secondary wall synthesis in fibers of *Arabidopsis*. *Planta* **225**(6): 1603-1611.

Zhong R, Taylor JJ, Ye ZH. 1997. Disruption of interfascicular fiber differentiation in an *Arabidopsis* mutant. *The Plant Cell* **11**(12): 2159-2170.

Zhong R, Ye Z-H. 2004. *amphivasal vascular bundle 1*, a gain-of-function mutation of the *IFL1/REV* gene, is associated with alterations in the polarity of leaves, stems and carpels. *Plant and Cell Physiology* **45**(4): 369–385.

Zhong R, Ye Z-H. 2007. Regulation of *HD-ZIP III* genes by microRNA 165. *Plant Signalling and Bahaviour* **2**(5): 351-353.

Zhong R, Ye Z-H. 2009. Transcriptional regulation of lignin biosynthesis. *Plant Signalling and Bahaviour* **4**(11): 1028-1034.

Zhong R, 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 and Cell Physiology* **53**(2): 368–380.

Zhou J, Lee C, Zhong R, 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**(1): 248-266.

Zhu J, Sova P, Xu Q, Dombek KM, Xu EY, Vu H, Tu Z, Brem RB, Bumgarner RE, Schadt EE. 2012. Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation. *PLoS Biology* **10**(4): e1001301.

Zhu J, Wiener MC, Zhang C, Fridman A, Minch E, Lum PY, Sachs JR, Schadt EE. 2007. Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLoS Computational Biology* **3**(4): e69.

Zhu T, Nevo E, Sun D, Peng J. 2012. Phylogenetic analyses unravel the evolutionary history of NAC proteins in plants. *Evolution* **66**(6): 1833-1848.

© University of Pretoria

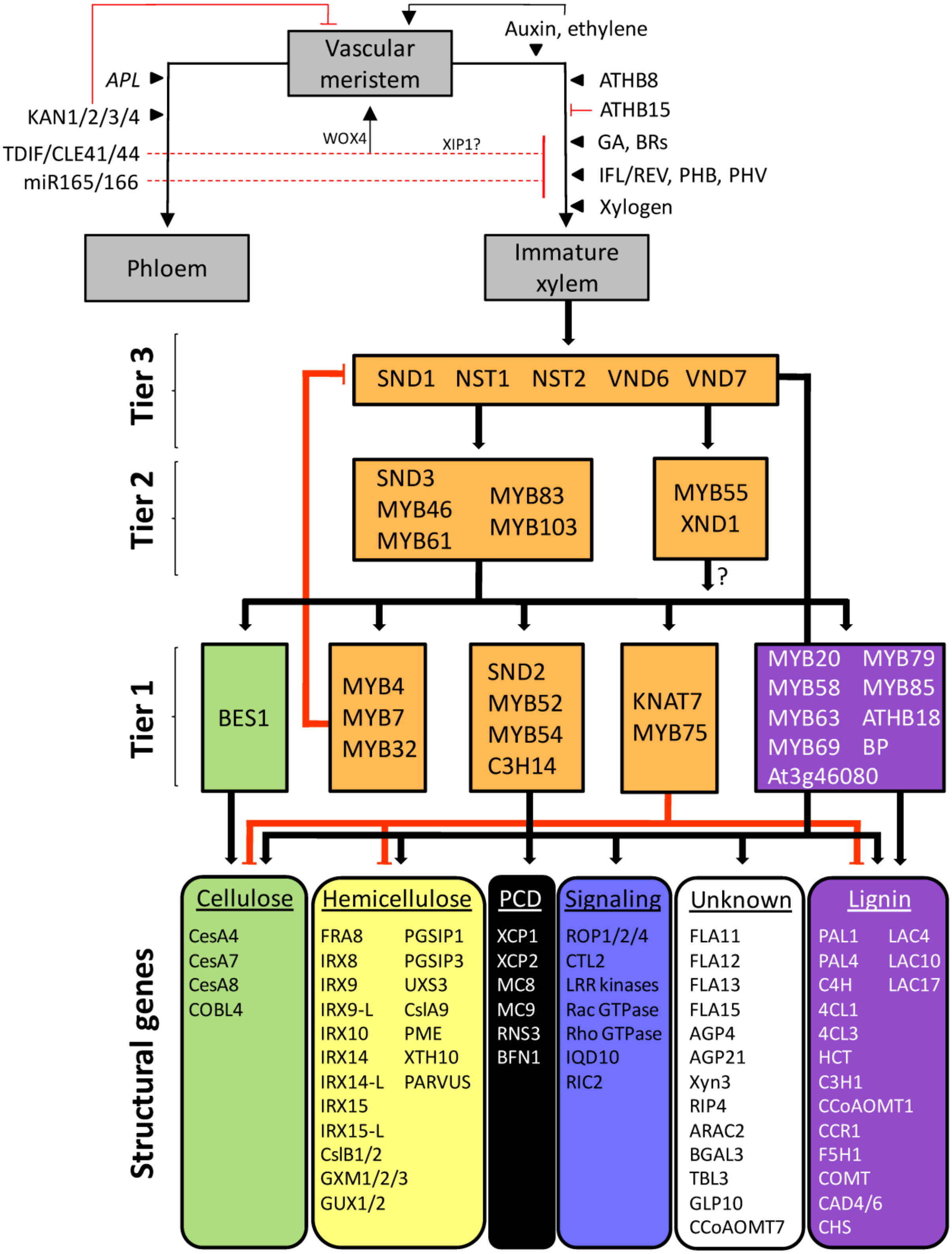
Zuo J, Niu Q-W, Chua N-H. 2000. An estrogen receptor-based transactivator XVE mediates highly inducible gene expression in transgenic plants. *The Plant Journal* **24**(2): 265–273.

© University of Pretoria

1.9. Figures

Fig. 1.1. The generalized *Arabidopsis* SCW transcriptional regulatory network in the light of vascular differentiation. Vascular meristems, representing procambiums or secondary cambiums, produce mother cells that differentiate into phloem and immature xylem tissue (grey boxes) under the influence of transcriptional, hormonal, peptide and miRNA regulators. Terminal differentiation of immature xylem cells into vessel elements and fibers is regulated by a tiered transcriptional network regulating genes associated with secondary cell wall cellulose, hemicellulose, programmed cell death (PCD), signaling, lignin and genes with unknown functions. Positive regulation is indicated by black arrows; negative regulation is represented by red edges. Block colours represent different biological function categories. TFs currently known to regulate only one functional category are colour-matched accordingly; orange blocks denote regulation of a combination of functional categories. The same colour scheme is used in Additional file 1.1.

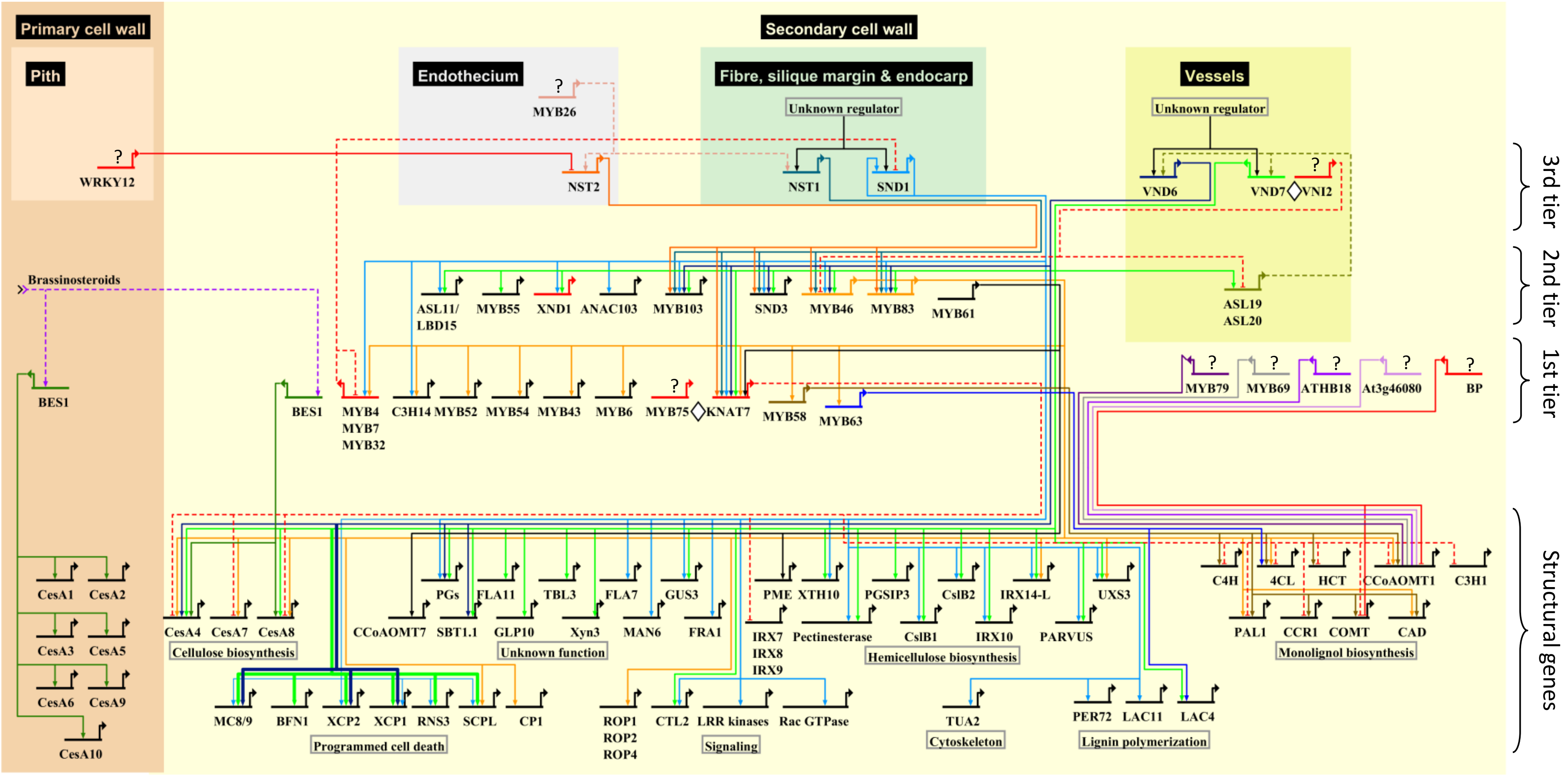
© University of Pretoria



© University of Pretoria

Fig. 1.2. Schematic representation of the protein-DNA interaction network underlying SCW biosynthesis in xylem fibers and vessels and anther endothecium in *Arabidopsis*. Interactions occurring specifically in primary cell wall tissues are also indicated. Direct protein-DNA interactions involving activation or repression are represented using solid edges, while known regulatory relationships in which the mechanism is unclear are represented with dashed edges. Repressors are denoted with red edges. Protein-protein interactions are represented as ◊; question marks represent unidentified upstream TFs; overlapping edges (MYB46, MYB83) represent redundancy. Target genes are arranged semi-hierarchically according to known functions. The complete list of supporting literature used to construct the network can be found in Additional file 1.2.

© University of Pretoria



© University of Pretoria

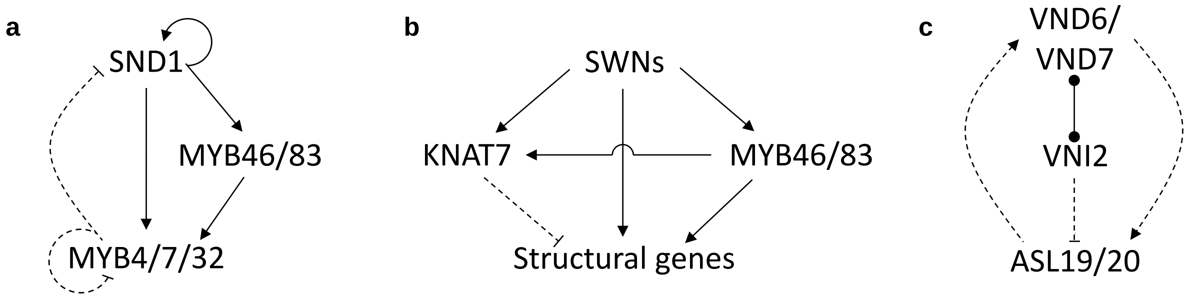


Fig. 1.3. Putative modules and motifs underlying SCW transcriptional regulation. (a) Negative feedback loop regulating SND1. **(b)** Negative regulation of structural genes by KNAT7. **(c)** Positive feedback loop regulating VND6/VND7. Dashed edges indicate unknown molecular mechanisms of protein-DNA interactions. Arrows indicate positive regulation, blunt ends indicate negative interactions. Dumbbells represent protein-protein interactions. Refer to section 1.4.4. for detailed discussion.

© University of Pretoria

1.10. Tables

Table 1.1. ***Cis*-regulatory elements that have been linked to SCW biosynthesis or which serve as general binding motifs for TF families involved in SCW transcriptional regulation.**

Element	Functional classification	Bound TF	Reference
Minimal NAC Recognition Sequence (NACRS; Tran *et al.*, 2004) TCNNNNNNNA**CACG**CATGT (core sequence in bold)	Abiotic stress response	ANAC19/55/72 ENAC1	Tran *et al.* (2004) Sun *et al.* (2011)
Secondary wall NAC Binding Element (SNBE) (T/A)NN(C/T)(T/C/G)TNNNNNNNA(A/ C)(G/)N(A/C/T)(A/T) =(T/A)NN(C)(T/ /G)TNNNNNNNA(A/G/C)(G/A)N(N)(A/T)[a]	Secondary cell wall biosynthesis	SND1, NST1, NST2, VND6, VND7 BdSWN5	Zhong *et al.* (2010c) Valdivia *et al.* (2013)
TACNTTNNNNATGA	Secondary cell wall biosynthesis	SND1	Wang *et al.* (2011)
Tracheary element-regulating cis-element (TERE) (Pyo *et al.*, 2007) CTTGAAAGCAA	Secondary cell wall biosynthesis	Possibly VND6/VND7	Ohashi-Ito *et al.* (2010);
AC elements (Lois *et al.*, 1989; Sablowski *et al.*, 1994; Hatton *et al.*, 1995) AC-I (SMRE8): ACCTACC AC-II (SMRE4): ACCAACC AC-III (SMRE7): ACCTAAC	Secondary cell wall biosynthesis/lignin biosynthesis	MYB58, MYB63, EgMYB2, PtMYB4, PttMYB021, PvMYB4	Patzlaff *et al.* (2003), Zhou *et al.* (2009), Rahantamalala *et al.* (2010), Winzell *et al.* (2010), Shen *et al.* (2011), Zhong and Ye (2012)
SMRE consensus ACC(A/T)A(A/C)(T/C)	Secondary cell wall biosynthesis/lignin biosynthesis	MYB46/MYB83	Zhong & Ye (2012)
M46RE (A/G)(G/T)T(T/A)GGT(A/G) =(T/C)ACC(A/T)A(A/C)(T/C)[a]	Secondary cell wall biosynthesis/lignin biosynthesis	MYB46	Kim *et al.* (2012a)
Element R GTTAGGT =ACCTAAC[a]	Disease resistance	MYB46	Ramírez *et al.* (2011)
MYB binding site IIG (MBSIIG) G(G/T)T(A/T)GGT(A/G) =(T/C)ACC(A/T)A(A/C)C[a]	General MYB binding?	MYB15, MYB84 EgMYB2	Romero *et al.* (1998) Goicoechea *et al.* (2005); Rahantamalala *et al.* (2010)
BSb CTGGTT	Cambium-specific expression	Unknown	Rahantamalala *et al.* (2010)

[a]Reverse complemented forms of the sequence. AC-related elements are underlined to highlight similarities between them.

© University of Pretoria

Table 1.2. Summary of techniques used to study transcriptional regulatory networks. Each technique is loosely arranged in order of increasing resolution of *in planta* protein-DNA associations.

		Advantages	**Challenges**
In vitro (trans)differentiation systems (Fukuda and Komamine, 1980; Kubo *et al.*, 2005; Oda *et al.*, 2005)		• Differentiation can be synchronized via hormonal induction • A high proportion of cultured cells differentiate into TEs • Time-course regulation of transcripts can be associated with developmental changes • *Arabidopsis* suspensions can be stably transformed (Ohashi-Ito *et al.*, 2010; Yamaguchi *et al.*, 2010a) • Provides temporal information to TE transcriptional regulation	• Currently only developed in *Zinnia* and *Arabidopsis* • Developmental *in planta* signals from neighbouring cells are lost
Reconstruction from co-expression data		• Co-regulated transcriptional modules can be identified • Direct interactions can be inferred from data transmission theory (Basso *et al.*, 2005) • Provides functional sets of genes for *in silico cis*-element identification where genome is available	• Transcriptomes from large numbers of diverse individuals, tissues and/or conditions required • Guilt by association suffers from type 1 errors
Reverse genetics		• Extensive catalog of mutant seedstocks available for *Arabidopsis* • Phenotypic relevance of candidate TFs can be assessed • Phenotypic effects of both gain- and loss-of-function mutants can be assessed	• Lethal knock-out and repression lines cannot be analyzed • Knock out lines not informative when TFs are functionally redundant • Over-expression can lead to unexpected knock-on and dosage balance effects • Generally suited to model organisms
Systems approaches	Systems biology	• Molecular interactions can be quantified and contextualized • Regulatory hubs can be identified and their regulatory effect assessed • Novel candidates can be identified using multiple omics data which may be missed using one-dimensional data	• Assumptions implicit to networks and modeling limit the biological accuracy of reconstructed networks • Requires large numbers of good quality high-throughput data • Generally more suited to model organisms
	Systems genetics	• The effect of allele substitution on regulatory networks can be quantified • Allows for the molecular basis of genetic associations to be understood • Co-expression clusters and eQTL analysis may identify potential master regulators • *Cis* and *trans* mechanisms of transcript regulation can be distinguished	• Constrained by the degree of expression polymorphism within the population under study • Large number of individuals required • Condition-specific co-expression may escape detection • Molecular basis of co-expression is unknown
Protein-binding microarrays (Mukherjee *et al.*, 2004; Bulyk, 2007)		• *Cis*-element sequences can be identified precisely • Oligonucleotide arrays are applicable across all taxa • *In vitro* results reportedly reflect *in vivo* binding	• Purified GST-tagged protein may need to be functionally validated (e.g. EMSA) prior to assay • Only dsDNA arrays can be used

© University of Pretoria

Table 1.2. *(continued)*

	Advantages	**Challenges**
Elecrophoretic mobility shift assay (EMSA)	• Direct method to detect protein binding • Can distinguish nucleotides essential for binding	• *In vitro* method • Low-throughput • Heterologously expressed protein may not be soluble
Yeast 1-hybrid (Y1H) (Li and Herskowitz, 1993; Wang and Reed, 1993)	• One of few gene-centred approaches available • High-throughput robotic screening possible (Reece-Hoyes *et al.*, 2011) • Gateway-compatible short DNA fragments or long gene promoters can be used as baits (Deplancke *et al.*, 2004) • Custom stringency control possible	• Prone to type 1 errors • Yeast-expressed proteins may lack essential post-translational modifications • Not suitable for TFs that require co-regulators to activate gene expression • Cell-type context of interaction cannot be inferred
Transient protoplast transactivation systems	• High-throughput (when combined with whole transcriptome analysis) • Circumvents the need for stable transformation • Little biological variation • *In vivo* method • Direct targets are inferred using post-translational induction in the presence of a protein synthesis inhibitor	• Currently restricted to *Arabidopsis* mesophyll and *Populus* secondary xylem protoplasts (Wehner *et al.*, 2011; Li *et al.*, 2012b) • Not suitable for TFs requiring tissue-specific co-factors (e.g. Bhargava *et al.*, 2010) • Possibility of false positives (misregulated genes) • Cells are exposed to high levels of stress, which may influence the assay
Chromatin immunoprecipitation • ChIP-on-chip (Ren *et al.*, 2000) • ChIP-seq (Barski *et al.*, 2007) • ChIP-exo (Rhee and Pugh, 2011) • Nano-ChIP-seq (Adli and Bernstein, 2011)	• High-throughput analysis of TF binding sites • *In planta* method • Can profile TFs that do not bind directly to DNA • Canonical binding sites can be identified (esp. using ChIP-exo)	• Critically dependent on antibody specificity and performance • Limited ability to assay TFs exhibiting low or cell-specific expression • Extensive optimization may be required for different tissues and organisms (Haring *et al.*, 2007) • Difficult to assign genes to TF binding sites, since not all binding sites are functional

© University of Pretoria

1.11. Additional files

The following files are available on the supplementary CD-ROM disk attached to this thesis:

1. **Additional file 1.1.cys** Cytoscape file (.cys) of direct and indirect protein-DNA interactions and protein-protein interactions involved in *Arabidopsis* SCW transcriptional regulation. Nodes indicate genes, edges indicate interactions. Red edges, known direct protein-DNA interactions; dark blue solid edges, known regulatory relationships of unknown nature; dark blue dashed edges; known indirect regulatory relationships; light blue edges, protein-protein interactions. Structural genes are indicated as square nodes, transcriptional activators as circular nodes, transcriptional repressors as diamond-shaped nodes, and transporter proteins (WAT1) as triangles. Nodes are colour-coded according to known biological processes: blue, signaling; green, cellulose biosynthesis; yellow, hemicelluloses biosynthesis; black, programmed cell death; purple, lignin biosynthesis; white, unknown function. Transcriptional regulators coloured in orange are involved in the regulation of more than one type of SCW biopolymer.

2. **Additional file 1.2.xlsx** Excel spreadsheet of supporting literature for each protein-DNA and protein-protein interaction depicted in Additional file 1.1.

© University of Pretoria

CHAPTER 2

Structural, evolutionary and functional analysis of the NAC domain protein family in *Eucalyptus*

Steven G. Hussey[1], Mohammed N. Saïdi[2,3], Charles A. Hefer[4], Alexander A. Myburg[1], Jacqueline Grima-Pettenati[3]

[1]Department of Genetics, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Private Bag X20, Pretoria, 0028, South Africa

[2]Laboratoire des Biotechnologies Végétales Appliquées à l'Amélioration des Cultures, Ecole Nationale d'Ingénieurs de Sfax, Route Soukra Km 4, B.P 1173, 3038 Sfax, Tunisia (present address)

[3]Laboratoire de Recherche en Sciences Végétales (LRSV), Université Toulouse, UPS, CNRS, BP 42617, F-31326, Castanet-Tolosan, France

[4]Department of Botany, University of British Columbia, 3529-6270 University Blvd, Vancouver, B.C., V6T 1Z4, Canada

This chapter has been written in manuscript format for submission to a peer-reviewed scientific journal. I drafted the majority of the manuscript and figures and conducted most of the bioinformatics analysis. Mohammed Saïdi identified the EgrNAC proteins, assisted with manual curation, conducted the chromosomal localization analysis and the Fluidigm RT-qPCR experiments in *E. globulus*. Charles Hefer remapped RNA-seq data to the curated gene models. Alexander A. Myburg and Jacqueline Grima-Pettenati participated in conceiving the project, supervised its progress and secured research funding. The manuscript has been submitted in its current state to the journal *New Phytologist* for peer review.

© University of Pretoria

2.1. Summary

NAC domain transcription factors regulate many developmental processes and stress responses in plants, but have been poorly characterized in non-model species. We analysed the characteristics and evolution of the NAC gene family of *Eucalyptus grandis*, a fast-growing forest tree in the rosid order Myrtales. NAC domain genes identified in the *E. grandis* genome were subjected to manual curation and amino acid sequence, phylogenetic and motif analyses. Transcript abundance in developing tissues and abiotic stress conditions in *E. grandis* and *E. globulus* was quantified using RNA-seq and RT-qPCR. 189 *E. grandis* NAC (EgrNAC) proteins, arranged into 22 subfamilies shared with other angiosperms, were extensively duplicated in subfamilies associated with stress response. Most *EgrNAC* genes formed tandem duplicate arrays that frequently carried signatures of purifying selection. Sixteen amino acid motifs were identified in EgrNAC proteins, eight of which were enriched in, or unique to, *Eucalyptus*. New candidates for the regulation of wood development and cold response were identified. This first description of a Myrtales NAC domain family advances our understanding of rosid NAC protein evolution. *EgrNAC* genes have a unique history of tandem duplication that has likely contributed to the adaptation of eucalypts to the challenging Australian environment.

© University of Pretoria

2.2. Introduction

Transcriptional regulators coordinate developmental processes and environmental responses in plants. A large family of NAC transcription factor proteins, defined by the conserved NAC (NAM/ATAF/CUC) DNA-binding domain in the N-terminal region, regulate diverse biological functions in terrestrial plants (Aida *et al.*, 1997; Olsen *et al.*, 2005). These proteins are chiefly involved in the response to biotic and abiotic stress (e.g. Tran *et al.*, 2004; Nakashima *et al.*, 2007; Jensen *et al.*, 2008; Meng *et al.*, 2009; Wu *et al.*, 2009), but also developmental processes (Guo & Gan, 2006; Yoo *et al.*, 2007; Willemsen *et al.*, 2008; Berger *et al.*, 2009; Morishita *et al.*, 2009), including secondary cell wall (SCW) biosynthesis during wood formation or xylogenesis (reviewed by Yamaguchi & Demura, 2010). It is thought that NAC proteins evolved over 400 million years ago and have to date only been identified in embryophytes (Zhu *et al.*, 2012). Most NAC subfamilies appeared before monocot-dicot divergence, with a few subfamilies restricted to tracheophytes, monocots, dicots or, rarely, specific plant families (Rushton *et al.*, 2008; Shen *et al.*, 2009; Zhu *et al.*, 2012).

The lignified SCWs that characterize vascular tissues of woody plants are rich in energy and biopolymers and therefore of significant agronomic importance (Plomion *et al.*, 2001; Mizrachi *et al.*, 2012). *Eucalyptus* is a woody plant genus encompassing some of the fastest growing plantation forest species. With over 20 million ha grown worldwide (Iglesias-Trabado & Wilstermann, 2008), it is a promising candidate for lignocellulosic biofuel production in addition to its extensive use in paper, pulp and raw cellulose products (Rockwood *et al.*, 2008; Carroll & Somerville, 2009). Lignified xylem fibers likely evolved through the sequential integration of independently evolved cellulosic cell wall thickenings,

© University of Pretoria

lignification (see Li & Chapple, 2010) and programmed cell death in a single cell type (Boyce *et al.*, 2003). NAC domain proteins feature prominently in the regulation of all these processes (Yamaguchi & Demura, 2010; Wang & Dixon, 2011). Knowledge of their transcriptional targets and biological functions provides a basis for developing approaches toward the improvement of wood and fiber properties. Considering the antiquity of xylogenesis, the apparent evolutionary conservation of most implicated NAC transcription factors (Zhong *et al.*, 2010a) is not surprising. Yet, certain NAC subfamilies display distinct patterns of evolution in particular plant lineages, associated with the unique evolutionary history of such lineages (Zhu *et al.*, 2012).

Almost all of the 700 known *Eucalyptus* species are endemic to Australia. Completing separation from Antarctica and drifting northwards around 50 Ma, the subcontinent's vegetation changed from tropical forest with high precipitation to a temperate, arid, grassland-dominated interior region during the Paleogene period (Kemp, 1981). *Eucalyptus* evolved by the Eocene (oldest fossils are ~52 Ma; Gandolfo *et al.*, 2011), diversifying in the cooler, drier conditions of the Paleogene and expanding throughout the continent (Beadle, 1981). While most eucalypts are adapted to xerophytic, fire-prone environments (Cary *et al.*, 2003), forest species such as *E. grandis* occupy wet, fertile regions of the eastern coast (Chippendale, 1988). The genetic basis for the successful adaption of eucalypts to the harsh Australian environment and their rich diversity of phytochemical products, such as essential oils, remains poorly understood.

© University of Pretoria

The evolutionary innovations allowing for the adaptation and unique properties of *Eucalyptus* could have involved diversification of transcription factors such as the NAC domain family. An understanding of how wood properties and environmental responses are transcriptionally controlled will help explain the adaptive potential of eucalypts, as well as their considerable capacity to produce woody biomass (Hinchee *et al.*, 2009). The structure, evolution, expression characteristics, and functions of the NAC domain family in *Eucalyptus* are currently unknown. We therefore analysed the gene and protein structure, phylogenetic relationships, and transcriptional dynamics of the *E. grandis* NAC domain family to elucidate the evolution and possible functions of NAC domain proteins in *Eucalyptus*. Within the core rosids (Eurosids), descriptions of NAC gene families have been reported for the Brassicales (Ooka *et al.*, 2003; Liu *et al.*, 2014; Ma *et al.*, 2014), Malvales (Shang *et al.*, 2013; Shah *et al.*, 2014), Sapindales (de Oliveira *et al.*, 2011), Fabales (Le *et al.*, 2011) and Malpighiales (Hu *et al.*, 2010). Comparative genomic analysis of NAC proteins in other angiosperms will facilitate a better understanding of NAC protein diversification and function. This study provides the first description of *NAC* gene family structure in the Myrtales, an order basal to the core rosids (Myburg *et al.*, in press), and contributes to our understanding of NAC domain evolution and function in the rosids.

2.3. Materials and Methods

2.3.1. Plant materials

E. globulus tissues were flash-frozen in liquid nitrogen. Juvenile and mature xylem samples (kindly provided by RAIZ, Portugal) were harvested from four- and ten-year-old trees (genotype VC9) respectively. Upright, tension and opposite xylem of two-year-old trees

© University of Pretoria

(genotypes GM52, BB3 and MB43, kindly provided by Altri Florestal, Portugal) were collected from main stems after three weeks of bending (45°). Fruit capsules and flower buds were harvested from genotype C33 (Altri Florestal, Portugal). For cold treatment experiments, one-year-old *E. globulus* genotype GM258 (Altri Florestal, Portugal) were subjected to cold (7°C) for 16 h in the dark. Control plants were maintained for 16 h in dark greenhouse conditions. Young and mature leaves, primary stems, secondary stems and roots were harvested. Each of three biological replicates consisted of bulked tissues from two trees.

2.3.2. RNA extraction and reverse transcription quantitative PCR (RT-qPCR)

Total RNA was extracted from frozen tissues as described elsewhere (Southerton *et al.*, 1998). cDNA synthesis, primer design and Fluidigm RT-qPCR analysis was conducted as described by Cassan-Wang *et al.* (2012). Primer set-specific PCR efficiencies and five control genes previously found to show stable expression across different tissues and various environments (Cassan-Wang *et al.*, 2012) were used for data normalization (Table S2.1). Expression data were subjected to QT clustering (Pearson correlation ≥ 0.5, minimum five genes per cluster) in TMEV (Saeed *et al.*, 2006).

2.3.3. Identification of NAC domain proteins

Genes in the *E. grandis* genome v.1.0 annotated with a Pfam NAM domain (PF02365; Punta *et al.*, 2012) were retrieved from Phytozome v7.0 (http://www.phytozome.net/Eucalyptus.php). All but the longest splice variants were removed. Some genes encoding proteins lacking initiation or termination codons were corrected with FGENESH (http://linux1.softberry.com/berry.phtml). Where possible,

© University of Pretoria

annotations were corroborated with RNA-seq data from Eucspresso (Mizrachi *et al.*, 2010) and EucGenIE (http://eucgenie.bi.up.ac.za/; Hefer *et al.* in preparation) databases. Gene models that could not be corrected were discarded. Twenty-two gene models were located on smaller "satellite" scaffolds that have not yet been mapped to the eleven *E. grandis* chromosome scaffolds: those with ≥ 95% nucleotide identity to other NAC genes were considered allelic. The presence of a NAC domain in the proteins was evaluated with a Hidden Markov Model (HMM) constructed from the NAC domain alignment of representative proteins from diverse species (Olsen *et al.*, 2005), using HMMER 3.0 (Finn *et al.*, 2011).

2.3.4. Phylogenetic analysis

For the EgrNAC protein tree, 189 curated EgrNAC protein sequences were aligned using MUSCLE (http://www.ebi.ac.uk/Tools/msa/muscle/; parameters pre-set). The alignment was trimmed with Gblocks (Castresana, 2000) using parameters: minimum sequences per conserved position, *n*/2 + 1; minimum sequences per flank position, *n*/2 + 1; maximum number of contiguous nonconserved positions, 10; minimum block length, 2; allowed gap positions, all. After visual inspection, poorly aligning sequences (EgrNAC91, EgrNAC187) were removed. For the five-species NAC gene tree, NAC protein sequences from *Arabidopsis thaliana*, *Oryza sativa*, *Populus trichocarpa*, *Vitis vinifera* and the 189 EgrNAC sequences (678 total) were similarly aligned with MUSCLE and trimmed with Gblocks. Two poorly aligning sequences (EgrNAC29, EgrNAC91) were removed. The alignments were submitted to PhyML 3.0 (Guindon *et al.*, 2010), initiated with a BIONJ tree using estimated Gamma distribution, proportion of invariable sites fixed at 0.0, four substitution rate categories, an LG

© University of Pretoria

substitution model with empirical equilibrium frequencies, and Shimodaira-Hasegawa-like aLRT branch support testing (Anisimova & Gascuel, 2006). Trees were visualized in MEGA 5.05 (Tamura *et al.*, 2011). The trees were rooted at the midpoint due to the lack of a known outgroup (Shen *et al.*, 2009; Zhu *et al.*, 2012).

2.3.5. Identification of conserved protein motifs

Sequences of the 189 aligned EgrNAC proteins were analysed using MEME v.4.7.0 (Bailey *et al.*, 2006) with parameters: distribution of motifs, zero or one per sequence; maximum number of motifs, 25; minimum number of sites, two; maximum number of sites, 189; minimum motif width, six; maximum motif width, 50. Overrepresented motifs were annotated using the PfamA and PfamB databases (http://pfam.sanger.ac.uk), and schematically represented using DomainDraw (Fink & Hamilton, 2007). HMMs were constructed from the MEME alignment of each motif using hmmbuild in HMMER 3.0 (Finn *et al.*, 2011). NAC protein sequences of *Populus trichocarpa*, *Glycine max*, *Carica papaya*, *Arabidopsis thaliana*, *Vitis vinifera*, *Oryza sativa* subsp. *japonica*, *Brachypodium distachyon* and *Zea mays* retrieved from the Plant Transcription Factor Database v.2.0 (Zhang *et al.*, 2011) were searched with the HMMs using HMMER 3.0. The TMHMM server v. 2.0 (http://www.cbs.dtu.dk/services/TMHMM/) was used for transmembrane helix (TMH) prediction, using a probability threshold of 1.0.

2.3.6. Gene structural analysis

Genomic sequences of the *E. grandis* v.1.0 annotation were downloaded from Phytozome v7.0 (http://www.phytozome.net/Eucalyptus.php), untranslated regions were removed and where applicable genomic sequences re-annotated corresponding to curated gene models.

© University of Pretoria

Coding sequences were aligned to genomic sequences and schematics generated using GSDS (http://gsds.cbi.pku.edu.cn) (Guo *et al.*, 2007).

2.3.7. Chromosomal localization and test for selection neutrality

MapChart 2.2 (Voorrips, 2002) was used for chromosomal linkage visualization. Coding sequences of genes in individual tandem duplicate blocks were aligned using MUSCLE in MEGA 5.05 (Tamura *et al.*, 2011). A codon-based Z-test was performed for each block in MEGA 5.05 using Pamilo-Bianchi-Li (Li, 1993; Pamilo & Bianchi, 1993) substitution model, bootstrap variance estimation method (1000 replicates), and pairwise deletion. Blocks were assessed for neutrality (H_A: dN ≠ dS), positive selection (H_A: dN/dS > 1.0) and purifying selection (H_A: dN/dS < 1.0). Only results that remained significant for most of the substitution models in MEGA were considered, and all blocks had more than ten synonymous or nonsynonymous substitutions as advised by Zhang *et al.* (1997).

2.3.8. *E. grandis* transcriptome analysis

EgrNAC coding sequences were aligned to the *E. grandis* (v.1.1) genome sequence using Exonerate (Slater & Birney, 2005), and the genome locations calculated. RNA-seq data of six tissues for three field-grown *E. grandis* individuals, and root samples prepared from young seedlings, were obtained from EucGenIE (http://eucgenie.bi.up.ac.za/, Hefer *et al.*, in preparation). The absolute transcript abundance values (FPKM) were obtained for the 189 NAC domain sequences with TopHat (Trapnell *et al.*, 2009) and Cufflinks (Trapnell *et al.*, 2010). The expression values were clustered using the QT clustering tool (Pearson correlation ≥ 0.75, minimum five genes per cluster) in the Multiple Array Viewer (Saeed *et al.*, 2006).

© University of Pretoria

2.4. Results

2.4.1. Identification of NAC domain genes in *E. grandis*

Gene models in the v.1.0 annotation of the *E. grandis* genome containing a NAC domain were identified using a superfamily search (see Methods), yielding 254 candidates (Table S2.2). Thirty-nine alternative splice variants were removed except for two splice variants that displayed only partial gene sequence overlap. Apart from these, only primary transcripts were considered. Sixteen genes on scaffolds other than those comprising the eleven main linkage groups (chromosomes) were considered putative alleles of gene models on the chromosome scaffolds since they showed ≥95% nucleotide similarity. One pair of adjacent gene models was collapsed into a single gene model (Table S2.2). Ten genes were removed following manual curation and 43 were corrected (see Additional note S2.1 for details). The remaining proteins were inspected for a significant match to the NAC domain (E-value < 0.001) using a Hidden Markov Model (HMM) of the NAC domain, rejecting one gene model (Eucgr.D00593.1). This process yielded 189 nonredundant candidates. They encoded proteins of 82 to 799 residues, a range similar to NAC proteins from other plants (Additional file 2.1). We sorted the gene symbols alphanumerically, i.e. in their order of appearance on chromosomes A through K and their sequential appearance in scaffolds not linked to the chromosomes, and renamed them *EgrNAC1* through *EgrNAC189* (Table S2.3).

2.4.2. Phylogeny of the EgrNAC proteins in relation to angiosperm NAC proteins

The evolution of the EgrNAC proteins was evaluated through maximum likelihood analysis incorporating well-described NAC domain sequences in the dicots *Arabidopsis*

© University of Pretoria

(http://www.arabidopsis.org/), *Vitis* (Shen *et al.*, 2009), and *Populus* (Hu *et al.*, 2010), and the monocot *Oryza* (Shen *et al.*, 2009). The non-*Eucalyptus* NAC proteins analysed are listed in Additional file 2.1. To improve the reliability of the phylogeny, only conserved positions in the alignment, represented by at least 50% of the sequences, were considered. Preliminary classification of the phylogeny according to the 21 subfamilies identified by an extensive analysis of the NAC protein family from nine lineages (Zhu *et al.*, 2012) revealed good agreement with the topology in our study. We therefore used the cited study to annotate subfamilies, but some differences should be noted. In our analysis (detailed dendrogram available in Additional file 2.2), subfamilies IIIa and IIIb could not be reliably dissociated and were combined into a single subfamily, IIIa/b. Four proteins in subfamily VIa (Zhu *et al.*, 2012) (ANAC084, PNAC134, PNAC135, VvNAC097) clustered with subfamily VIb in our phylogeny, and four rice genes previously assigned to VIb (ONAC001, ONAC005, ONAC139, ONAC041) clustered in VIa in our study. Three *Arabidopsis* proteins previously assigned to subfamily VIII (ANAC063, ANAC064, ANAC093; Zhu *et al.*, 2012) formed a well-supported clade with two *Arabidopsis* and three *Populus* NAC sequences that were unassigned to a subfamily by Zhu *et al.* (2012), allowing us to subdivide subfamily VIII into VIIIa and VIIIb. Eleven proteins unassigned by Zhu *et al.* (2012) (ANAC023, ANAC024, PNAC077, PNAC139, PNAC140, PNAC141, PNAC143, ONAC080, ONAC135, ONAC137, ONAC138) were incorporated into subfamily X. Finally, we defined an additional subfamily XI, from proteins unassigned by Zhu *et al.* (2012). Twenty-three proteins (~3%) remained unassigned. This culminated in 22 subfamilies with acceptable bootstrap support (>70) and distinguishable topologies (Fig. 2.1a). The biological functions of the respective members of the subfamilies, where known, were investigated in the literature. Generally, members of the

© University of Pretoria

same subfamily appeared to share similar biological processes (Table S2.4), as observed elsewhere (Shen *et al.*, 2009).

The phylogeny representation in Fig. 2.1b was linearized with respect to evolutionary distance. NAC proteins from different species were generally interspersed; however, EgrNAC proteins were overrepresented in subfamilies IVa, IVc, Vb and VII. Subfamily VII had previously been found to contain *Populus* and *Carica* sequences, but not those of *Vitis* or herbaceous plants (Zhu *et al.*, 2012). Consistent with this, in our phylogeny *Populus* but no *Arabidopsis*, *Oryza* or *Vitis* sequences constituted this apparently ancient subfamily, while we additionally identified several *Eucalyptus* NACs in subfamily VII (Fig. 2.1b). These sequences displayed greater intraspecific than interspecific homology, as previously observed in *Populus* and *Carica* (Zhu *et al.*, 2012). This indicates independent gain of subfamily VII NAC sequences in *Eucalyptus*, *Populus* and *Carica* as opposed to their loss in *Arabidopsis*, *Oryza* and *Vitis*, and suggests that parallel evolution in these genera may have facilitated the retention of duplicated genes in subfamily VII. Similarly, most of the *Eucalyptus* NACs overrepresented in subfamilies IVa, IVc and Vb appear to be lineage-specific paralogs (Fig. 2.1b).

Because of the importance of *Eucalyptus* as a wood fiber crop, we explored whether the *E. grandis* genome contains homologs of *Arabidopsis* NACs involved in SCW biosynthesis (reviewed by Yamaguchi & Demura, 2010; Zhong *et al.*, 2010b) using the phylogeny in Additional file 2.2. We found single putative *Eucalyptus* orthologs of *Arabidopsis* fiber-associated SND1, SND2 and NST1, vessel-associated VNI2, VND6 and VND7, and multiple

© University of Pretoria

co-orthologs of SND3 and XND1 (Table S2.5). Interestingly, we did not identify a *Eucalyptus* ortholog of NST2, which is associated with endothecium SCWs in *Arabidopsis* anthers (Mitsuda *et al.*, 2005). Overall, this suggests that NAC-mediated transcriptional regulation of SCW biosynthesis in *Eucalyptus* is relatively well conserved with, but not identical to, *Arabidopsis*.

2.4.3. Phylogenetic relationships and expression patterns of *EgrNAC* genes

A gene tree of 187 EgrNAC proteins was constructed using the maximum likelihood approach after removing two poorly aligning proteins (Fig. 2.2a). This phylogeny was used to assess the evolutionary conservation of gene expression patterns, amino acid motifs and gene structure of *EgrNAC* genes.

Tissue-specific transcript abundance is suggestive of a gene's biological function. To generate hypotheses of the functions of unknown *EgrNAC* genes, we examined their expression patterns in shoot tips, young leaves, mature leaves, flowers, roots, phloem and developing (secondary) xylem. RNA-seq data for three field-grown *E. grandis* individuals were obtained from the *Eucalyptus* Genome Integrative Explorer (EucGenIE; http://eucgenie.bi.up.ac.za/), and reads were re-mapped to the *EgrNAC* coding sequences to accommodate corrected gene models (data provided in Additional file 2.3). Transcripts of closely related genes showed broadly similar abundance profiles (Fig. 2.2b). No expression was detected for 19 (~10%) of the *EgrNAC* genes in the sampled tissues. Conversely, 93 genes (~50%) were expressed in all tissues.

© University of Pretoria

To identify transcripts with similar expression patterns, the expression data of the 189 *EgrNAC* genes were hierarchically clustered using a quality threshold algorithm, yielding 13 clusters (Fig. S2.1). Cluster 4 contained genes preferentially expressed in various stages of leaf development, enriched for paralogs belonging to stress response-associated subfamily Vb (Table S2.4). Transcripts preferentially expressed in tissues containing vascular cells (roots, phloem, developing xylem) were located in Clusters 6, 10 and 11, including (co-)orthologs of SND1, NST1, SND2 and XND1 (Fig. S2.1, Table S2.5). Based on their preferential expression in vascular tissues, *EgrNAC24*, *EgrNAC32*, *EgrNAC58*, *EgrNAC59*, *EgrNAC90*, *EgrNAC141* and *EgrNAC157* are novel candidates for the regulation of xylogenesis-related processes since they have no *Arabidopsis* orthologs associated with this process (Fig. S2.1, Additional file 2.2). Remarkably, of the 21 *EgrNAC* genes that were expressed in only one tissue, 14 were restricted to roots (Cluster 3). One gene was expressed only in mature leaves, three were restricted to flowers and another three to developing xylem (Fig. S2.1, unassigned cluster).

2.4.4. Conserved motifs in *Eucalyptus* NAC domain proteins and conservation of gene structure

Overrepresented amino acid motifs tend to represent functional regions that are evolutionarily conserved across or within specific lineages. We subjected the 189 EgrNAC sequences to motif overrepresention analysis using MEME (Bailey *et al.*, 2006). Sixteen significantly overrepresented motifs (E-value < 10^{-161}) of 11-50 residues were identified, present in 7-182 of the sequences (Table S2.6). The motifs are represented in parallel with the EgrNAC protein

© University of Pretoria

phylogeny in Fig. 2.2a, showing that motif composition and arrangement (Fig. 2.2c) were in good agreement with the gene tree.

Using HMMs describing subdomains A to E of the NAC domain (Ooka *et al.*, 2003), we assigned Motif 1 to subdomain A, Motif 2 to subdomain B, Motif 3 and Motif 4 to subdomain C, Motif 5 and Motif 6 to subdomain D, and Motif 7 to subdomain E. As expected, these motifs occurred in the N-terminal half of EgrNAC sequences (Fig. 2.2c). Because they were also found in the majority (> 65%) of EgrNAC sequences (Table S2.6), they were classified as "general motifs". The remaining "specific motifs" (Motif 8 through Motif 16) were restricted to groups of 7-20 EgrNAC proteins. With the exception of Motif 9, none of the specific motifs had any hits (E-value < 0.01) to Pfam-A or Pfam-B databases (http://pfam.sanger.ac.uk), and occurred in the diverse C-terminal region (Fig. 2.2c). Similarly, Motif 9 aside, none of the specific motifs matched those previously identified by Ooka *et al.* (2003), Fang *et al.* (2008), Jensen *et al.* (2010) or Hu *et al.* (2010).

To assess the distribution of the motifs in other plant genomes, HMMs designed from the *Eucalyptus* alignments of each motif in the MEME output were used to identify matching motifs (E-value < 0.01) in the NAC proteins from *Populus*, *Glycine*, *Arabidopsis*, *Carica*, *Vitis*, *Oryza*, *Brachypodium* and *Zea*. The EgrNAC protein set was also searched as a positive control for HMM performance. As expected, the general motifs corresponding to subdomains A through E of the NAC domain were detected at high frequencies in NAC proteins from all eight genomes (Table 2.1). Specific motifs, however, appeared enriched in *Eucalyptus* compared to other genomes, with the exception of Motif 9 (Table 2.1). The latter, which was

© University of Pretoria

found in a minority of NAC proteins in all nine genera, displayed significant homology (E-value = 8.4×10^{-5}) to the NAM domain (PF02365) in Pfam-A (http://pfam.sanger.ac.uk) and appears to replace Motifs 2, 3 and 4 (Fig. 2.2c). Motif 10 was present in all dicots, while 13 and 14 were only present in some dicots (Table 2.1). Motifs 12, 15 and 16 were exclusively found in EgrNAC proteins and may thus represent motifs unique to *Eucalyptus* (Table 2.1). It is unlikely that *Eucalyptus*-specific motifs were an artefact of HMMs built on *E. grandis* alignments, since no bias was observed in the cumulative frequency of general motifs identified in *E. grandis* compared to other genomes using HMMs built exclusively on alignments from this species (Fig. S2.2).

Some NAC proteins are tethered to the endoplasmic reticulum or plasma membrane via transmembrane helices (TMHs) (Chen *et al.*, 2008; Seo *et al.*, 2008). We identified seven putative membrane-tethered EgrNAC proteins, all with single C-terminal TMHs comprising at least twenty residues (Fig. 2.2c). All EgrNAC proteins with predicted TMHs had putative *Arabidopsis* orthologs known to contain TMHs (Kim *et al.*, 2010) (Table S2.7). Interestingly, multiple membrane-tethered *Arabidopsis* co-orthologs were found for each TMH-containing EgrNAC protein (Table S2.7). All except one TMH-containing EgrNAC protein occurred in subfamily IIIa/b, which prominently features NACs involved in stress response and development (Table S2.4).

We next analysed the conservation of intron and exon arrangements in the *EgrNAC* genes (Fig. 2.2d). An average of 3.3 exons was observed, similar to most *Arabidopsis* NAC genes (Duval *et al.*, 2002), ranging from one to eleven. The numbers of exons were similar

© University of Pretoria

between closely related genes. Intron phase was also well conserved (Fig. S2.3), as reported previously for *Populus* NAC domain genes (Hu *et al.*, 2010).

2.4.5. Physical distribution of *Eucalyptus* NAC genes

A significant proportion of genes in the *E. grandis* genome have expanded through tandem duplication (Myburg *et al.*, in press). We studied the distribution of the 185 *E. grandis* NAC domain genes located on the eleven main chromosome scaffolds in the v.1.0 annotation (Fig. 2.3). We defined tandem duplicates according to Hanada *et al.* (2008) as pairs of NAC gene models within 100 kb of each other, having ten or fewer non-homologous genes in-between. Based on this definition, 121 (~64%) of the NAC domain genes were distributed amongst 23 blocks of tandem duplicate arrays of 2-21 members (Fig. 2.3). In most cases, the members of each tandem array belonged to a single subfamily (Table S2.8).

The fate of tandem duplicates include nonfunctionalization, neofunctionalization, subfunctionalization and redundancy (Rastogi & Liberles, 2005). To assess if members of tandem duplicate arrays are under natural selection in *E. grandis*, we used a codon-based Z-test (Tamura *et al.*, 2011) based on the ratio of nonsynonymous to synonymous substitutions between all pairs of sequences in each tandem array. Out of the 23 tandem arrays, a test for overall purifying selection between pairs of genes in a given array was significant for 13 arrays ($P < 0.05$; Fig. 2.3). *P*-values for individual gene pairs in each of these arrays, which were not all significant alone, are shown in Additional file 2.4. No evidence of positive selection acting on any of the arrays overall or between any pairs of sequences in a given array was detected. These results suggest that most tandem arrays are under purifying selection.

© University of Pretoria

Next, we analysed the expression patterns of *EgrNAC* paralogs in the seven EucGenIE tissues (http://eucgenie.bi.up.ac.za/) and compared them to public expression data of close homologs in *Arabidopsis*, *Oryza*, and *Populus*. Here, we define paralogs as proteins that are more closely related to each other than to homologs from the other genomes (inferred from Additional file 2.2). *EgrNAC* paralogs were composed mostly of tandem duplicate arrays. Most groups of *EgrNAC* paralogs featured a "dominant" transcript at a marked level of expression, accompanied by paralogs with reduced overall expression of similar, slightly diverged or undetected transcript abundance (Fig. S2.4). Expression data for *Arabidopsis*, *Oryza* and *Populus* homologs from Genevestigator (Hruz *et al.*, 2008) and Poplar Expression Angler (Toufighi *et al.*, 2005; Wilkins *et al.*, 2009) are shown for similar tissues, where available, underneath each *EgrNAC* paralog group (Fig. S2.4).

Paralog groups (a), (b), (f), (i), (n), and (o) contained one or two "dominant" *EgrNAC* transcripts expressed similarly to at least one non-*Eucalyptus* homolog (Fig. S2.4). However, we observed conflicting expression patterns between *EgrNAC* genes and non-*Eucalyptus* homologs in groups (d), (h) and (m), suggesting functional divergence. Also, certain transcripts in group (b) (*EgrNAC24*), (e) (*EgrNAC43*, *141*, *154*, and *157*) and (i) (*EgrNAC50*, *59*, and several root-specific transcripts) showed expression patterns differing from *Eucalyptus* paralogs and non-*Eucalyptus* homologs (Fig. S2.4), three of which are wood development candidates (described above).

© University of Pretoria

2.4.6. Expression characteristics of *E. globulus* NAC domain genes

Besides having superior wood properties, *E. globulus* is more frost-tolerant than *E. grandis*, but still suffers from frost damage (Hasey & Connor, 1990; Skolmen & Ledig, 1990; Tibbits *et al.*, 2006). To better understand NAC gene functions in eucalypts, we profiled the expression patterns of *E. globulus* orthologs in various tissues and in response to cold and tension stress. We used Fluidigm RT-qPCR to analyse expression patterns of 33 *EgrNAC* orthologs in *E. globulus* ("*EglNAC*"). Transcript profiles across nine *E. globulus* tissues were hierarchically clustered, revealing three prominent expression clusters: xylem-enriched, xylem-deficient, and unassigned (Fig. 2.4). These are robust clusters, since tissues were sampled from trees of different ages, genotypes and sites (see Methods). *EglNAC* transcripts were also quantified in cold-treated trees relative to control in primary stems, secondary stems, young leaves and roots. Three genes in primary stems, two in secondary stems, five in young leaves and five in roots showed a significant transcriptional response to cold treatment (Fig. 2.5).

Candidate *EglNAC* genes involved in tension wood formation were identified by comparing selected *EglNAC* transcripts in upright stem xylem to those in xylem from tension and opposite wood in an *E. globulus* bending trial. Two genes were significantly upregulated (*EglNAC31*, *EglNAC152*) and two downregulated (*EglNAC44*, *EglNAC139*) in tension wood relative to upright control (Fig. 2.6). In opposite wood, *EglNAC139* and *EglNAC141* were downregulated and upregulated relative to upright control, respectively (Fig. 2.6). With the exception of *EglNAC44*, most of these genes were not preferentially expressed in xylem-enriched tissues (Fig. 2.4).

© University of Pretoria

We assessed the evolutionary conservation of NAC gene expression between *Eucalyptus* species by comparing the correlation of expression of *EglNAC* and *EgrNAC* orthologs. We compared the xylem/leaf expression ratio calculated from Fluidigm and RNA-seq data (see Methods) to account for developmental variation and environmental effects. These data correlated significantly (Fig. S2.5; $r = 0.69$, $P < 0.0001$), suggesting a high overall conservation in expression of *EgrNAC* and *EglNAC* orthologs. Amongst the exceptions, *EgrNAC142* expression was not detected using RNA-seq in *E. grandis*, while that of its ortholog *EglNAC142* was detected by RT-qPCR analysis in *E. globulus* secondary tissues (Fig. 2.4).

2.5. Discussion

As representative of the Myrtales, an order basal to the core rosids (Myburg *et al.*, in press), the *E. grandis* genome adds resolution to understanding gene family and function evolution in the rosids. We identified 189 nonredundant NAC domain proteins in the *E. grandis* genome, one of the largest NAC domain families known (Jin *et al.*, 2014). We modelled our subfamily annotation on that proposed by Zhu *et al.* (2012), which is based on nine diverse lineages and good bootstrap support for each subfamily. We included novel clustering such as the merging of subfamilies IIIa and IIIb, the subdivision of subfamily VIII and the annotation of a new subfamily, XI, resulting in 22 well-supported subfamilies in our study. The number of subfamilies proposed for the NAC domain family in different plants has been highly variable (Ooka *et al.*, 2003; Mitsuda *et al.*, 2005; Fang *et al.*, 2008; Rushton *et al.*, 2008; Pinheiro *et al.*, 2009; Hu *et al.*, 2010; Nuruzzaman *et al.*, 2010; Zhu *et al.*, 2012), and additional modifications will likely be proposed in the future as more plant genomes are analysed.

© University of Pretoria

However defined, most NAC subfamilies are represented in angiosperm genomes (Shen *et al.*, 2009; Zhu *et al.*, 2012), with only one example of a subfamily unique to a lineage, the Solanaceae (Rushton *et al.*, 2008; Singh *et al.*, 2013). We found no evidence of a *Eucalyptus*-specific subfamily, although expansion of EgrNAC proteins was apparent in subfamilies IVa, IVc, Vb and VII due to five, three, two and four arrays of tandem duplication, respectively (Fig. 2.1b, Table S2.8). Subfamily VII was hypothesized to represent a tree-specific expansion involved in the regulation of wood formation (Zhu *et al.*, 2012). Although no functional data is yet available, the apparent expansion of EgrNAC and PNAC sequences in this subfamily supports this hypothesis, with at least two *EgrNAC* members (*EgrNAC58*, *EgrNAC59*) specifically expressed in developing xylem (Fig. 2.2a) and one upregulated in *E. globulus* tension wood (*EglNAC31*, discussed below). Although most subfamily VII genes were expressed only in roots (Fig. 2.2b), these organs contain significant amounts of lignified vascular tissue. Some EgrNAC proteins in subfamily VII contain a novel *Eucalyptus*-specific amino acid motif (Motif 16) (Fig. 2.2c, Fig. S2.2b), attracting further attention to the significance of this group. Subfamilies firmly associated with transcriptional regulation of SCW formation also contained small-scale expansions, resulting in five SND3 co-orthologs in subfamily II, and four co-orthologs of XND1 in subfamily VIc (Table S2.5), some *E. globulus* orthologs of which we implicate in tension wood formation (*EglNAC44*, *EglNAC139*, *EglNAC152*; Fig. 2.6). In general, however, subfamilies with members known to regulate wood formation (Ic, II, VIc) did not exhibit notable expansion in *Eucalyptus*. Since the expanded subfamilies IVa and Vb contain members involved in stress response (Table S2.4), we hypothesize that the large blocks of tandem duplications contained within them reflect adaptations to environmental stress. A predominant stress response function for retained

© University of Pretoria

duplicates has been observed previously in model representatives across the three domains of life (Kondrashov *et al.*, 2002).

Over half of the tandem duplicate arrays showed significant overall purifying selection, suggesting that at least some of the retained duplicates are still functional and may provide adaptive advantages. Functional buffering, protein dosage benefits or subfunctionalization of paralogs could serve as the basis for this retention. Interestingly, paralogs with detected expression frequently exhibited dissimilar expression profiles (Fig. S2.4). This has previously been shown to be more common of small-scale duplicates than whole-genome duplications (Casneuf *et al.*, 2006), and is therefore an expected result due to the large number of tandem duplicates among these paralogs. Diverged expression may be explained by a transcriptional version of the duplication-degeneration-complementation (DDC) model of subfunctionalization, whereby duplicates accumulate deleterious mutations in regulatory regions that result in the complementary expression of functionally redundant copies in different tissues, and the subsequent retention of these complementary duplicates through purifying selection (Force *et al.*, 1999). For example, most *Arabidopsis* tandem duplicates have undergone rapid expression divergence in accordance with the DDC model (Haberer *et al.*, 2004). Protein sequence evolution of duplicated genes, rather than divergence in gene expression, is a more prominent mechanism towards morphological diversification (Hanada *et al.*, 2009). No evidence for positive selection (i.e. neofunctionalization of paralogs) was found amongst the tandem duplicates, but positive selection is rare, episodic and difficult to detect (Raes & van de Peer, 2003). Furthermore, subfunctionalization is considered a temporary state that ultimately facilitates the acquisition of novel functions (Rastogi & Liberles, 2005).

© University of Pretoria

Functional analysis of paralogous *EgrNAC* genes will help to unravel whether functional diversification has occurred.

The absence of detectable expression for 19 genes, all of which were found in tandem duplicate arrays, is suggestive of nonfunctionalization. However, most of these genes were predicted to encode full-length proteins containing the amino acid motifs present in related, expressed paralogs (Fig. 2.2c), suggesting that these genes may be functional. Five of these tandem duplicates tested statistically significant for purifying selection on their coding regions (Additional file 2.4), further suggesting a functional role. It is quite likely that these genes are expressed in tissues or conditions not sampled in this study. For example, removing just one tissue dataset, root, increases the proportion of non-expressed genes from ~10% to ~20%. Alternatively, the penetrance of paralog expression may differ between populations, genotypes or species. For example, *EgrNAC142* might be classified as a pseudogene from the *E. grandis* RNA-seq dataset, but a possible xylogenesis regulator from the *E. globulus* expression data (Fig. 2.2b, Fig. 2.4).

Novel conserved amino acid motifs were identified in the C-termini of EgrNAC proteins that were enriched or restricted to *Eucalyptus* (Table 2.1), suggesting that NAC domain proteins in the Myrtales, or the *Eucalyptus* lineage specifically, have acquired unique functions after divergence from the analysed taxa. The specific motifs were restricted to particular clades of EgrNAC proteins and are likely to participate in the transcriptional activation or repression activities of the associated members (Fig. 2.2c). Transmembrane helices (TMHs) are also relevant to transcriptional regulation because they facilitate rapid

© University of Pretoria

post-translational recruitment of transcription factors tethered to intracellular membranes (Seo *et al.*, 2008). In agreement with a predominant role of such proteins in stress response (Chen *et al.*, 2008), all EgrNAC proteins with predicted TMHs are homologous to *Arabidopsis* TMH-containing NACs (Table S2.7), and most belong to the stress and developmental process-associated subfamily IIIa/b (Table S2.4). Membrane tethering is therefore not a mechanism for regulatory novelty in *E. grandis*.

We showed that, between *E. grandis* and *E. globulus* tissues, orthologous NAC domain genes have correlated expression, suggesting that orthologous genes perform similar functions across *Eucalyptus* species. We implicated ten *EglNAC* genes in the transcriptional response to cold stress in leaves, stems and roots in *E. globulus*, none of which had close *Arabidopsis* homologs known to play a role in abiotic stress response. Five of these genes responded to cold treatment in more than one tissue, but two candidates (*EglNAC24* and *EglNAC168*) were differentially regulated in opposing directions in different tissues (Fig. 2.5). Interestingly, transcripts of *Arabidopsis* homologs of (secondary) cell wall-associated NAC genes *SND1* (*EgrNAC61*), *SND3* (*EglNAC64*) and *VND4*/*VND5* (*EglNAC50*) were also affected by cold treatment (Fig. 2.5). Other *E. globulus* homologs of *SND3* (*EglNAC44*) and *XND1* (*EglNAC139*, *EglNAC152*) showed transcriptional responses to tension and/or opposite wood formation (Fig. 2.6), as observed for *Populus* homologs of *SND3* and *XND1* (Grant *et al.*, 2010). Transcript profiles of *EglNAC31* (of subfamily VII discussed above) and *EglNAC141*, which have no close *Arabidopsis* homologs, suggest they are novel candidates for regulating tension and opposite wood, respectively (Fig. 2.6).

© University of Pretoria

2.6. Conclusion

The NAC gene family of *E. grandis* is one of the largest described to date. Pervasive tandem, and less frequent segmental, duplications have contributed significantly to *EgrNAC* expansion, a tenth of which appear to be transcriptionally silent in deeply sequenced *E. grandis* RNA-seq data. Although the functions of most *EgrNAC* genes remain to be investigated, limited duplication of homologs of known regulators of SCW biosynthesis suggests functional conservation, while subfamilies with paralog expansion appear to be associated with abiotic and biotic stress responses. It is thus postulated that duplication of *EgrNAC* genes has contributed to the adaptation of *Eucalyptus* to the diverse and often harsh Australian climate. Divergent expression and evidence of purifying selection acting on most groups of paralogs suggests a complex interplay of subfunctionalization, functional redundancy and nonfunctionalization in their evolution. However, we cannot rule out that neofunctionalization has already occurred in older duplications, or that apparent pseudogenes may be expressed in conditions or genotypes not sampled in this study. We report novel amino acid motifs overrepresented in EgrNAC proteins that are enriched in *Eucalyptus* or restricted to the genus entirely, which may influence transcriptional activation and repression. Similarly, several new candidates for vascular development, tension wood formation and cold response were found. Our study provides a first-level understanding of how one of the largest transcription factor families in plants may have contributed to the evolutionary success of the Myrtaceae and the accomplishment of *Eucalyptus* as a global fiber crop.

© University of Pretoria

2.7. Acknowledgements

The authors thank J.P Paiva and Altri-Florestal (Portugal) for kindly providing samples from an *E. globulus* tension wood experiment, and H. San Clemente and B. Savelli for bioinformatic assistance. Thanks to N. Ladouce and the Plateforme Génomique Génopole Toulouse/Midi-Pyrénées (Genotoul) for technical assistance with Biomark Fluidigm dynamic arrays. We thank Albe van der Merwe (UP) for his assistance with the phylogenetic analysis. JGP and MNS acknowledge the financial support of the ANR (Project "Tree For Joules" ANR-2010-KBBE-007-01; Labex Tulip ANR-10-LABX-41), the CNRS, and the University Toulouse III (UPS). SH and AAM acknowledge funding from the Bioinformatics and Functional Genomics Programme of the National Research Foundation (UID 71255), South Africa. *E. grandis* sequence data were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/ in collaboration with the user community.

2.8. References

Aida M, Ishida T, Fukaki H, Fujisawa H, Tasaka M. 1997. Genes involved in organ separation in *Arabidopsis*: an analysis of the *cup-shaped cotyledon* mutant. *The Plant Cell* **19**(6): 841-857.

Anisimova M, Gascuel O. 2006. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Systematic Biology* **55**(4): 539-552.

Bailey TL, Williams N, Misleh C, Li WW. 2006. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research* **34**: W369-W373.

Beadle NCW. 1981. *The Vegetation of Australia*. London: Cambridge University Press.

Berger Y, Harpaz-Saad S, Brand A, Melnik H, Sirding N, Alvarez JP, Zinder M, Samach A, Eshed Y, Ori N. 2009. The NAC-domain transcription factor GOBLET specifies leaflet boundaries in compound tomato leaves. *Development* **136**: 823-832.

Boyce CK, Cody GD, Fogel ML, Hazen RM, Alexander CMOD, Knoll AH. 2003. Chemical evidence for cell wall lignification and the evolution of tracheids in Early Devonian plants. *International Journal of Plant Sciences* **164**(5): 691-702.

© University of Pretoria

Carroll A, Somerville C. 2009. Cellulosic biofuels. *Annual Review of Plant Biology* **60**: 165-182.

Cary G, Lindenmayer D, Dovers S. 2003. *Australia burning: fire ecology, policy and management issues*. Melbourne: CSIRO Publishing.

Casneuf T, De Bodt S, Raes J, Maere S, van de Peer Y. 2006. Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant *Arabidopsis thaliana*. *Genome Biology* **7**: R13.

Cassan-Wang H, Soler M, Yu H, Camargo ELO, Carocha V, Ladouce N, Savelli B, Paiva JAP, Leplé J-C, Grima-Pettenati J. 2012. Reference genes for high-throughput quantitative reverse transcription–PCR analysis of gene expression in organs and tissues of *Eucalyptus* grown in various environmental conditions. *Plant and Cell Physiology* **53**(12): 2101-2116.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* **17**(4): 540-552.

Chen Y-N, Slabaugh E, Brandizzi F. 2008. Membrane-tethered transcription factors in *Arabidopsis thaliana*: novel regulators in stress response and development. *Current Opinion in Plant Biology* **11**: 695-701.

Chippendale GM 1988. *Eucalyptus*, *Angophora* (Myrtaceae). In: A. S. George ed. *Flora of Australia*. Melbourne: Brown Prior Anderson Pty Ltd, 470.

de Oliveira TM, Cidade LC, Gesteira AS, Filho MAC, Filho WSS, Costa MGC. 2011. Analysis of the NAC transcription factor gene family in citrus reveals a novel member involved in multiple abiotic stress responses. *Tree Genetics and Genomes* **7**: 1123-1134.

Duval M, Hsieh T-F, Kim SY, Thomas TL. 2002. Molecular characterization of *AtNAM*: a member of the *Arabidopsis* NAC domain superfamily. *Plant Molecular Biology* **50**: 237-248.

Fang Y, You J, Xie K, Xie W, Xiong L. 2008. Systematic sequence analysis and identification of tissue-specific or stress-responsive genes of NAC transcription factor family in rice. *Molecular Genetics and Genomics* **280**(6): 547-563.

Fink J, Hamilton N. 2007. DomainDraw: A macromolecular schematic drawing program. *In Silico Biology* **7**: 14.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research* **39**: W29-W37.

Force A, Lynch M, Pickett FB, Amores A, Yan Y-l, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**: 1531-1545.

Gandolfo MA, Hermsen EJ, Zamaloa MC, Nixon KC, González CC, Wilf P, Cúneo NR, Johnson KR. 2011. Oldest known *Eucalyptus* macrofossils are from South America. *PLoS ONE* **66**(6): e21084.

© University of Pretoria

Grant EH, Fujino T, Beers EP, Brunner AM. 2010. Characterization of NAC domain transcription factors implicated in control of vascular cell differentiation in *Arabidopsis* and *Populus*. *Planta* **232**: 337-352.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* **59**(3): 307-321.

Guo A, Zhu Q, Chen X, Luo J. 2007. GSDS: a gene structure display server. *Yi Chuan* **29**(8): 1023-1026.

Guo Y, Gan S. 2006. AtNAP, a NAC family transcription factor, has an important role in leaf senescence. *The Plant Journal* **46**(4): 601-612.

Haberer G, Hindemitt T, Meyers BC, Mayer KFX. 2004. Transcriptional similarities, dissimilarities, and conservation of *cis*-elements in duplicated genes of *Arabidopsis*. *Plant Physiology* **136**: 3009-3022.

Hanada K, Kuromori T, Myouga F, Toyoda T, Shinozaki K. 2009. Increased expression and protein divergence in duplicate genes is associated with morphological diversification. *PLoS Genetics* **5**(12): e1000781.

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu S-H. 2008. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiology* **148**: 993-1003.

Hasey JK, Connor JM. 1990. *Eucalyptus* shows unexpected cold tolerance. *California Agriculture* **44**(2): 25-27.

Hinchee M, Rottmann W, Mullinax L, Zhang C, Chang S, Cunningham M, Pearson L, Nehra N. 2009. Short-rotation woody crops for bioenergy and biofuels applications. *In Vitro Cellular & Developmental Biology - Plant* **45**(6): 619-629.

Hruz T, Laule O, Szabo G, Wessendorp F, Bleuler S, Oertle L, Widmayer P, Gruissem W, Zimmermann P. 2008. Genevestigator V3: A reference expression database for the meta-analysis of transcriptomes. *Advances in Bioinformatics* **2008**: 420747.

Hu R, Qi G, Kong Y, Kong D, Gao Q, Zhou G. 2010. Comprehensive analysis of NAC domain transcription factor gene family in *Populus trichocarpa*. *BMC Plant Biology* **10**: 145.

Iglesias-Trabado G, Wilstermann D 2008. *Eucalyptus universalis*. Global cultivated eucalypt forests map 2008 Version 1.0.1. In GIT Forestry Consulting's EUCALYPTOLOGICS: Information resources on *Eucalyptus* cultivation worldwide. Available online at http://www.git-forestry.com.

© University of Pretoria

Jensen M, Kjaersgaard T, Nielsen M, Galberg P, Petersen K, O'Shea C, Skriver K. 2010. The *Arabidopsis thaliana* NAC transcription factor family: structure-function relationships and determinants of ANAC019 stress signalling. *Biochemical Journal* **426**: 183-196.

Jensen MK, Hagedorn PH, de Torres-Zabala M, Grant MR, Rung JH, Collinge DB, Lyngkjaer MF. 2008. Transcriptional regulation by an NAC (NAM-ATAF1,2-CUC2) transcription factor attenuates ABA signalling for efficient basal defence towards *Blumeria graminis* f. sp. *hordei* in *Arabidopsis*. *Plant J* **56**(6): 867-880.

Jin J, He Zhang LK, Gao G, Luo J. 2014. PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Research* **42**(D1): D1182-D1187.

Kemp EM 1981. Tertiary palaeogeography and the evolution of the Australian climate. In: A. Keast ed. *Ecological Biogeography of Australia*. The Hague: Dr. W. Junk bv Publishers, 33-46.

Kim S-G, Lee S, Seo PJ, Kim S-K, Kim J-K, Park C-M. 2010. Genome-scale screening and molecular characterization of membrane-bound transcription factors in *Arabidopsis* and rice. *Genomics* **95**: 56-65.

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV. 2002. Selection in the evolution of gene duplications. *Genome Biology* **3**(2): research0008.0001–0008.0009.

Le DT, Nishiyama R, Watanabe Y, Mochida K, Yamaguchi-Shinozaki K, Shinozaki K, Tran L-SP. 2011. Genome-wide survey and expression analysis of the plant-specific NAC transcription factor family in soybean during development and dehydration stress. *DNA Research* **18**(4): 263-276.

Li W-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *Journal of Molecular Evolution* **36**: 96-99.

Li X, Chapple C. 2010. Understanding lignification: challenges beyond monolignol biosynthesis. *Plant Physiology* **154**: 449-452.

Liu T, Song X, Duan W, Huang Z, Liu G, Li Y, Hou X. 2014. Genome-wide analysis and expression patterns of NAC transcription factor family under different developmental stages and abiotic stresses in Chinese cabbage. *Plant Molecular Biology Reporter* **DOI**: 10.1007/s11105-11014-10712-11106.

Ma J, Wang F, Li M-Y, Jiang Q, Tan G-F, Xiong A-S. 2014. Genome wide analysis of the NAC transcription factor family in Chinese cabbage to elucidate responses to temperature stress. *Scientia Horticulturae* **165**: 82-90.

Meng C, Cai C, Zhang T, Guo W. 2009. Characterization of six novel NAC genes and their responses to abiotic stresses in *Gossypium hirsutum* L. *Plant Science* **176**(3): 352-359.

© University of Pretoria

Mitsuda N, Seki M, Shinozaki K, Ohme-Takagi M. 2005. The NAC transcription factors NST1 and NST2 of *Arabidopsis* regulate secondary wall thickenings and are required for anther dehiscence. *Plant Cell* **17**: 2993–3006.

Mizrachi E, Hefer CA, Ranik M, Joubert F, Myburg AA. 2010. *De novo* assembled expressed gene catalog of a fast-growing *Eucalyptus* tree produced by Illumina mRNA-Seq. *BMC Genomics* **11**: 681.

Mizrachi E, Mansfield SD, Myburg AA. 2012. Cellulose factories: advancing bioenergy production from forest trees. *New Phytologist* **194**(1): 54-62.

Morishita T, Kojima Y, Maruta T, Nishizawa-Yokoi A, Yabuta Y, Shigeoka S. 2009. *Arabidopsis* NAC transcription factor, ANAC078, regulates flavonoid biosynthesis under high-light. *Plant and Cell Physiology* **50**(12): 2210-2222.

Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D, Goodstein DM, Dubchak I, Poliakov A, Mizrachi E, Kullan ARK, van Jaarsveld I, Hussey SG, Pinard D, Merwe Kvd, Singh P, Silva-Junior OB, Togawa RC, Pappas MR, Faria DA, Sansaloni CP, Petroli CD, Yang X, Ranjan P, Tschaplinski TJ, Ye C-Y, Li T, Sterck L, Vanneste K, Murat F, Soler M, Clemente HS, Saidi N, Cassan-Wang H, Dunand C, Hefer CA, Bornberg-Bauer E, Kersting AR, Vining K, Amarasinghe V, Ranik M, Naithani S, Elser J, Boyd AE, Liston A, Spatafora JW, Dharmwardhana P, Raja R, Sullivan C, Romanel E, Alves-Ferreira M, Külheim C, Foley W, Carocha V, Paiva J, Kudrna D, Brommonschenkel SH, Pasquali G, Byrne M, Rigault P, Tibbits J, Spokevicius A, Jones RC, Steane DA, Vaillancourt RE, Potts BM, Joubert F, Barry K, Jr. GJP, Strauss SH, Jaiswal P, Grima-Pettenati J, Salse J, Peer YVd, Rokhsar DS, Schmutz J. in press. The genome of *Eucalyptus grandis* - a global tree for fiber and energy. *Nature*.

Nakashima K, Tran L-SP, Nguyen DV, Fujita M, Maruyama K, Todaka D, Ito Y, Hayashi N, Shinozaki K, Yamaguchi-Shinozaki K. 2007. Functional analysis of a NAC-type transcription factor OsNAC6 involved in abiotic and biotic stress-responsive gene expression in rice. *The Plant Journal* **51**: 617-630.

Nuruzzaman M, Manimekalai R, Sharoni AM, Satoh K, Kondoh H, Ooka H, Kikuchi S. 2010. Genome-wide analysis of NAC transcription factor family in rice. *Gene* **465**: 30-44.

Olsen AN, Ernst HA, Leggio LL, Skriver K. 2005. NAC transcription factors: structurally distinct, functionally diverse. *Trends in Plant Science* **10**(2): 79-87.

Ooka H, Satoh K, Doi K, Nagata T, Otomo Y, Murakami K, Matsubara K, Osato N, Kawai J, Carninci P, Hayashizaki Y, Suzuki K, Kojima K, Takahara Y, Yamamoto K, Kikuchi S.

© University of Pretoria

2003. Comprehensive analysis of NAC family genes in *Oryza sativa* and *Arabidopsis thaliana*. *DNA Research* **10**(6): 239-247.

Pamilo P, Bianchi NO. 1993. Evolution of the *Zfx* and *Zfy*, genes: Rates and interdependence between the genes. *Molecular Biology and Evolution* **10**: 271-281.

Pinheiro GL, Marques CS, Costa MDBL, Reis PAB, Alves MS, Carvalho CM, Fietto LG, Fontes EPB. 2009. Complete inventory of soybean NAC transcription factors: Sequence conservation and expression analysis uncover their distinct roles in stress response. *Gene* **444**: 10-23.

Plomion C, Leprovost G, Stokes A. 2001. Wood formation in trees. *Plant Physiology* **127**: 1513-1523.

Punta M, Coggill PC, Eberhardt RY, Tate JMJ, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer ELL, Eddy SR, Bateman A, Finn RD. 2012. The Pfam protein families database. *Nucleic Acids Research* **40**(D1): D290-D301.

Raes J, van de Peer Y. 2003. Gene duplication, the evolution of novel gene functions, and detecting functional divergence of duplicates in silica. *Applied Bioinformatics* **2**(2): 91-101.

Rastogi S, Liberles DA. 2005. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evolutionary Biology* **5**: 28.

Rockwood DL, Rudie AW, Ralph SA, Zhu JY, Winandy JE. 2008. Energy product options for *Eucalyptus* species grown as short rotation woody crops. *International Journal of Molecular Sciences* **9**: 1361-1378.

Rushton PJ, Bokowiec MT, Han S, Zhang H, Brannock JF, Chen X, Laudeman TW, Timko MP. 2008. Tobacco transcription factors: novel insights into transcriptional regulation in the Solanaceae. *Plant Physiology* **147**: 280-295.

Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, Howe EA, Li J, Thiagarajan M, White JA, Quackenbush J. 2006. TM4 Microarray Software Suite. *Methods in Enzymology* **411**: 134-193.

Seo PJ, Kim S-G, Park C-M. 2008. Membrane-bound transcription factors in plants. *Trends in Plant Science* **13**(10): 550-556.

Shah ST, Pang C, Hussain A, Fan S, Song M, Zamir R, Yu S. 2014. Molecular cloning and functional analysis of NAC family genes associated with leaf senescence and stresses in *Gossypium hirsutum* L. *Plant Cell, Tissue and Organ Culture* **117**(2): 167-186.

Shang H, Li W, Zou C, Yuan Y. 2013. Analyses of the NAC transcription factor gene family in *Gossypium raimondii* Ulbr.: Chromosomal location, structure, phylogeny, and expression patterns. *Journal of Integrative Plant Biology* **55**(7): 663-676.

© University of Pretoria

Shen H, Yin Y, Chen F, Xu Y, Dixon RA. 2009. A bioinformatic analysis of *NAC* genes for plant cell wall development in relation to lignocellulosic bioenergy production. *BioEnergy Research* **2**: 217-232.

Singh AK, Sharma V, Pal AK, Acharya V, Ahuja PS. 2013. Genome-wide organization and expression profiling of the NAC transcription factor family in potato (*Solanum tuberosum* L.). *DNA Research* **20**(4): 403-423.

Skolmen RG, Ledig TF 1990. *Eucalyptus globulus* Labill. Bluegum *Eucalyptus*. In: R. Burns B. Honkala eds. *Silvics of North America: 2. Hardwoods*. Washington, D.C.: US Department of Agriculture, Forest Service, 302.

Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**: 31.

Southerton SG, Marshall H, Mouradov A, Teasdale RD. 1998. Eucalypt MADS-box genes expressed in developing flowers. *Plant Physiology* **118**: 365-372.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: Molecular Evolutionary Genetics Analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* **28**: 2731-2739.

Tibbits WN, White TL, Hodge GR, Borralho NMG. 2006. Genetic variation in frost resistance of *Eucalyptus globulus* ssp. *globulus* assessed by artificial freezing in winter. *Australian Journal of Botany* **54**: 521-529.

Toufighi K, Brady SM, Austin R, Ly E, Provart NJ. 2005. The Botany Array Resource: e-Northerns, Expression Angling, and promoter analyses. *The Plant Journal* **43**: 153-163.

Tran L-SP, Nakashima k, Sakuma Y, Simpson SD, Fujita Y, Maruyama K, Fujita M, Seki M, Shinozaki K, Yamaguchi-Shinozaki K. 2004. Isolation and functional analysis of *Arabidopsis* stress-inducible NAC transcription factors that bind to a drought-responsive *cis*-element in the *early responsive to dehydration stress 1* promoter. *The Plant Cell* **16**: 2481-2498.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**(9): 1105-1111.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Baren MJv, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**: 511-515.

Voorrips RE. 2002. MapChart: Software for the graphical presentation of linkage maps and QTLs. *The Journal of Heredity* **93**(1): 77-78.

© University of Pretoria

Wang H-Z, Dixon RA. 2011. On–off switches for secondary cell wall biosynthesis. *Molecular Plant* **5**(2): 297-303.

Wilkins O, Nahal H, Foong J, Provart NJ, Campbell MM. 2009. Expansion and diversification of the *Populus* R2R3-MYB family of transcription factors. *Plant Physiology* **149**(2): 981-993.

Willemsen V, Bauch M, Bennett T, Campilho A, Wolkenfelt H, Xu J, Haseloff J, Scheres B. 2008. The NAC domain transcription factors FEZ and SOMBRERO control the orientation of cell division plane in *Arabidopsis* root stem cells. *Developmental Cell* **15**(6): 913-922.

Wu Y, Deng Z, Lai J, Zhang Y, Yang C, Yin B, Zhao Q, Zhang L, Li Y, Yang C, Xie Q. 2009. Dual function of *Arabidopsis ATAF1* in abiotic and biotic stress responses. *Cell Research* **19**: 1279-1290.

Yamaguchi M, Demura T. 2010. Transcriptional regulation of secondary wall formation controlled by NAC domain proteins. *Plant Biotechnology* **27**: 237-242.

Yoo SY, Kim Y, Kim SY, Lee JS, Ahn JH. 2007. Control of flowering time and cold response by a NAC-domain protein in *Arabidopsis*. *PLoS ONE* **2**(7): e642.

Zhang H, Jin JP, Tang L, Zhao Y, Gu XC, Gao G, Luo JC. 2011. PlantTFDB 2.0: update and improvement of the comprehensive plant transcription factor database. *Nucleic Acids Research* **39**: D1114-D1117.

Zhang J, S. K, Nei M. 1997. Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes. *Molecular Biology and Evolution* **14**(12): 1335-1338.

Zhong R, Lee C, Ye Z-H. 2010a. Evolutionary conservation of the transcriptional network regulating secondary cell wall biosynthesis. *Trends in Plant Science* **15**(11): 625-632.

Zhong R, Lee C, Ye Z-H. 2010b. Global analysis of direct targets of secondary wall NAC master switches in *Arabidopsis*. *Molecular Plant* **3**(6): 1087-1103.

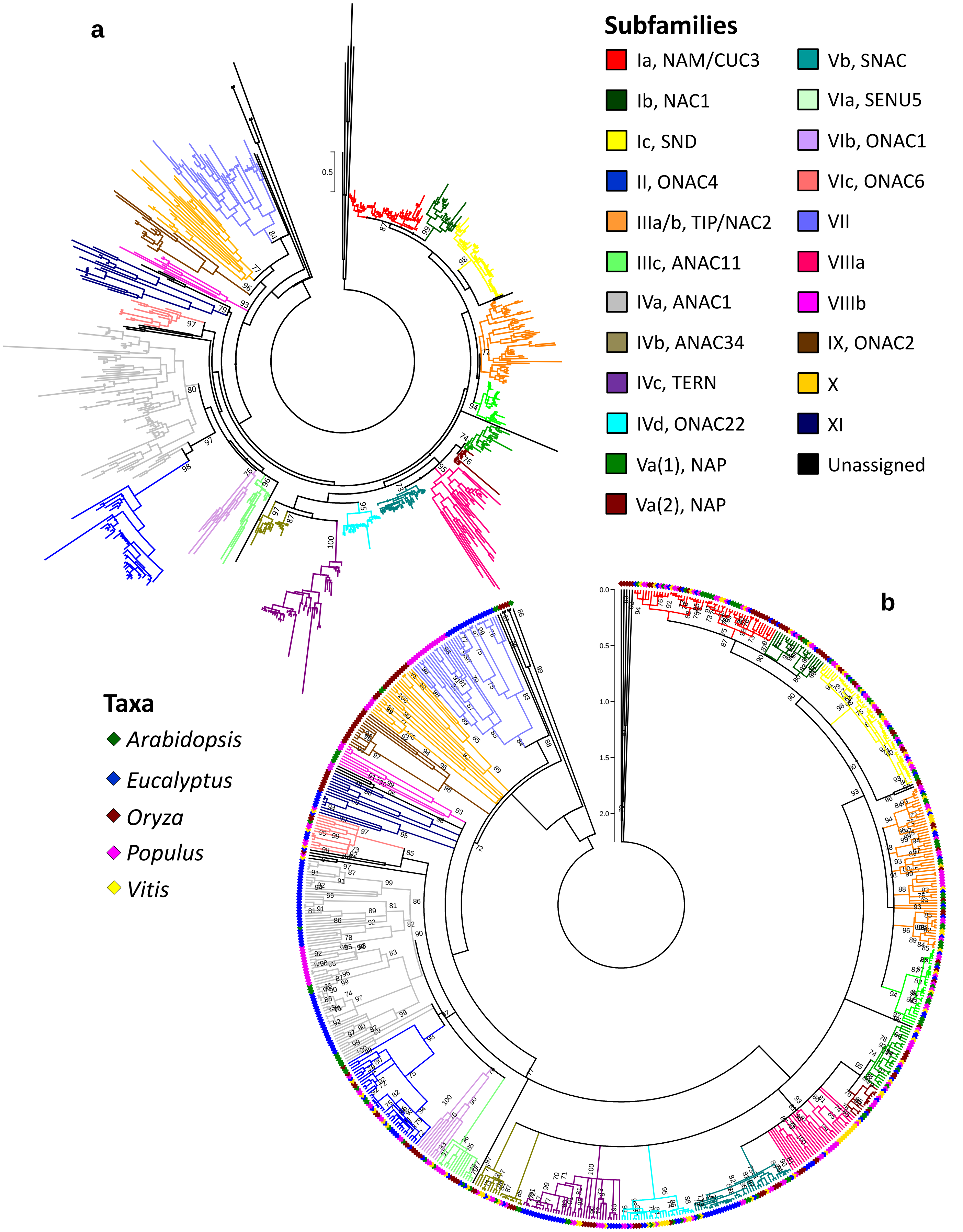
Zhu T, Nevo E, Sun D, Peng J. 2012. Phylogenetic analyses unravel the evolutionary history of NAC proteins in plants. *Evolution* **66**(6): 1833-1848.

© University of Pretoria

2.9. Figures

Fig. 2.1. Maximum likelihood phylogeny of the *Eucalyptus* NAC family in relation to angiosperm lineages *Arabidopsis*, *Populus*, *Oryza* and *Vitis*. Trees were rooted at the midpoint. **(a)** Circular phylogram showing subfamilies adapted from Zhu *et al.* (2012); aLRT branch support values for each subfamily are indicated. **(b)** Linearized circular representation, normalized with respect to evolutionary distance. aLRT values >70 are indicated and the organism of origin of each respective taxon is indicated with a diamond symbol. A detailed dendrogram is available in Additional file 2.2.

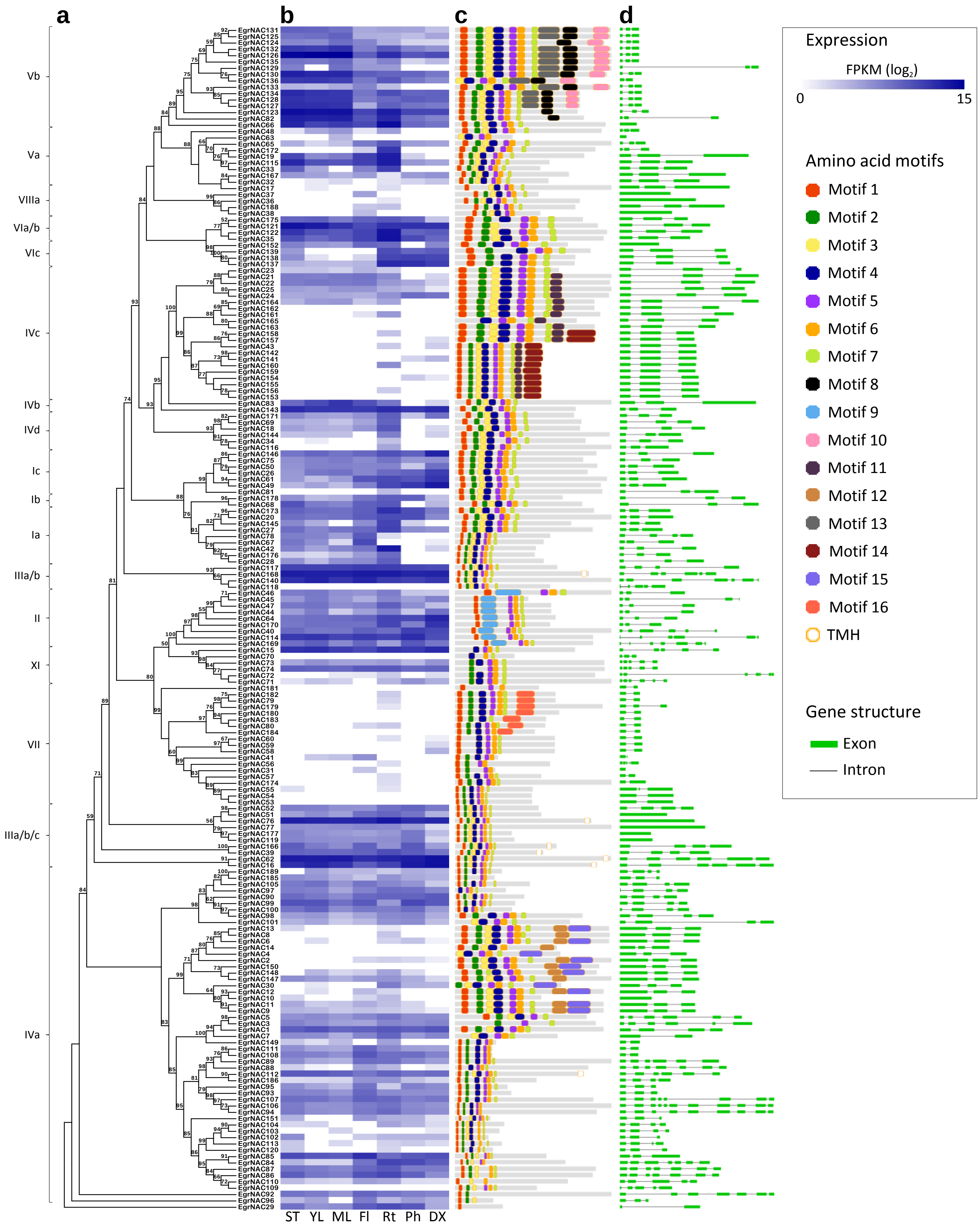
© University of Pretoria



© University of Pretoria

Fig. 2.2. Predicted EgrNAC protein phylogeny, transcript abundance profiles, conserved amino acid motifs and gene structure. From left to right: **(a)** unrooted maximum likelihood phylogeny of EgrNAC proteins, showing subfamily classification and aLRT values greater than 50. **(b)** RNA-seq transcript abundance of *EgrNAC* genes in shoot tip (ST), young leaf (YL), mature leaf (ML), flowers (Fl), root (Rt), phloem (Ph) and developing xylem (DX) of three field-grown *E. grandis* trees. Values are expressed as the $\log_2$ value of average fragments per kilobase of exon per million fragments mapped (FPKM) per tissue. **(c)** Composition and distribution of overrepresented amino acid motifs (see Table S2.6). Grey bars indicate relative protein lengths. **(d)** Position of exons and introns in the *EgrNAC* gene models.

© University of Pretoria



© University of Pretoria

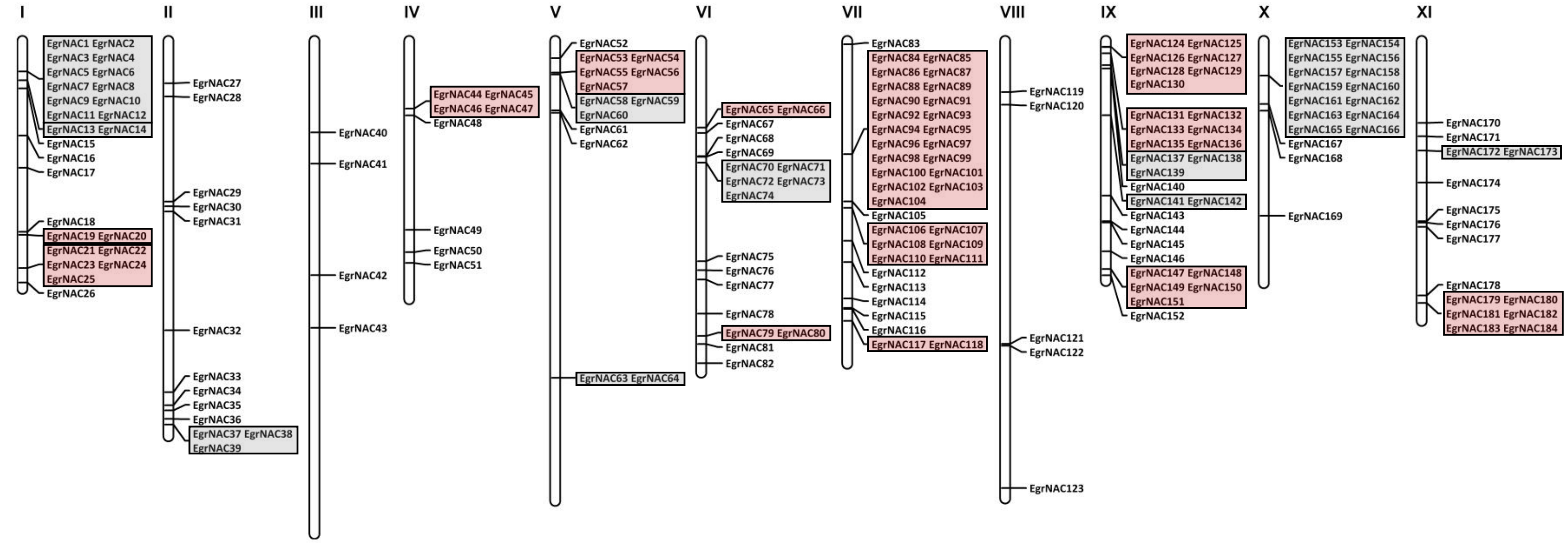


Fig. 2.3. Chromosomal locations of *EgrNAC* genes. Tandem duplicates are represented by shaded blocks; red shading indicates blocks with dN/dS < 1.0 (*P* < 0.05; codon-based Z-test). *P*-values for individual pairs of tandem duplicate pairs are available in Additional file 2.4.

© University of Pretoria

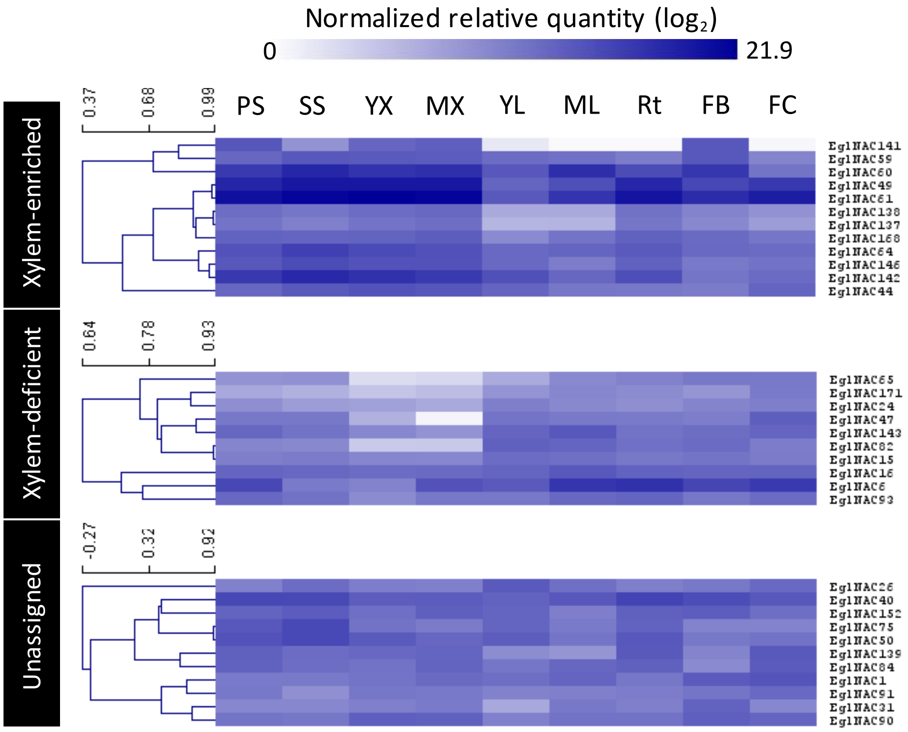


Fig. 2.4. Tissue and organ expression data for 33 *EgrNAC* orthologs from *E. globulus*, herein denoted *EglNAC*. Fluidigm RT-qPCR data were hierarchically clustered using a quality threshold (QT) algorithm. PS, primary stem; SS, secondary stem; YX, young xylem; MX, mature xylem; YL, young leaf; ML, mature leaf; Rt, root; FB, flower bud; FC, flower capsule. Branch distances indicate Pearson's correlation.

© University of Pretoria

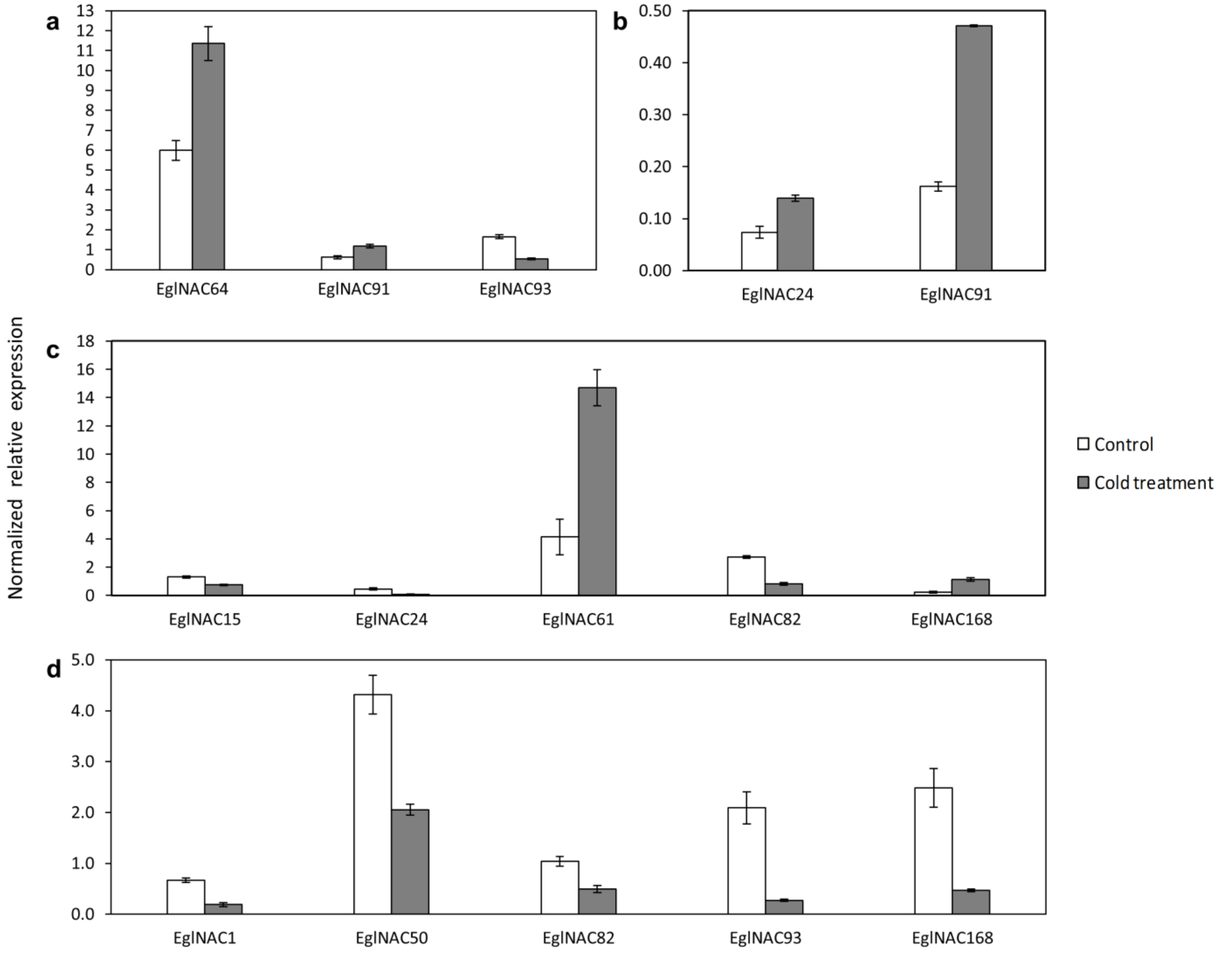


Fig. 2.5. ***EglNAC*** **genes showing a positive or negative transcriptional response towards cold treatment in primary stems (a), secondary stems (b), young leaves (c) and roots (d) in *E. globulus*.** Error bars indicate the standard deviation of three biological replicates. A Bonferroni-adjusted *P*-value of 0.05 was applied to a two-tailed Student's *t* test for each of 33 *EglNAC* genes. Only genes exhibiting significant responses are shown.

© University of Pretoria

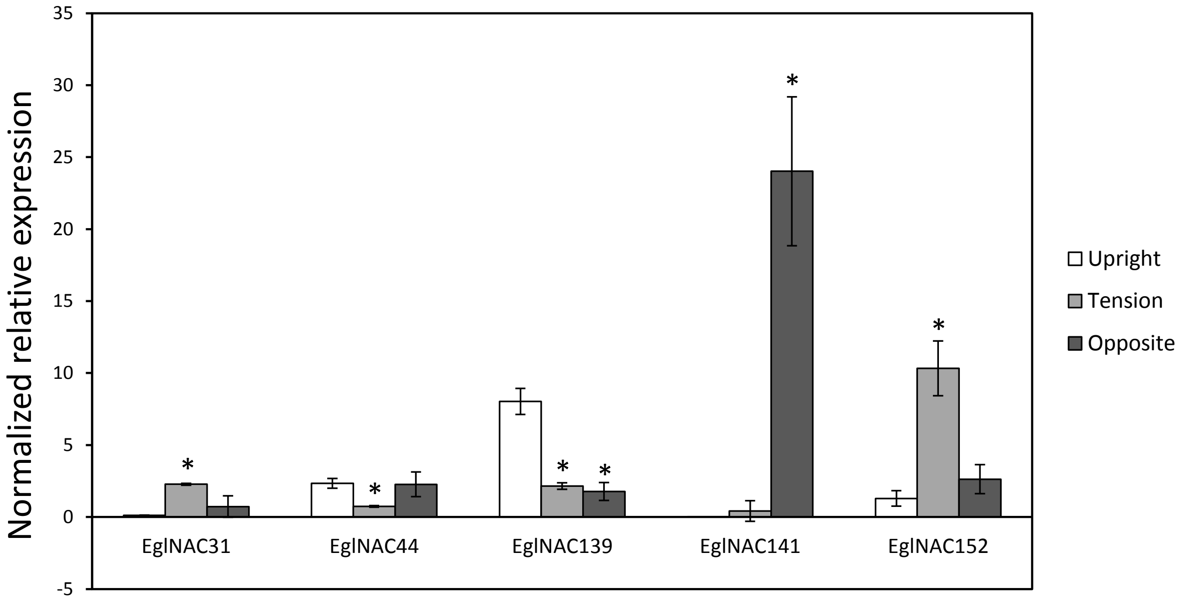


Fig. 2.6. Transcriptional profiles of *EglNAC* genes in xylem tissue from tension or opposite wood, relative to upright control in *E. globulus*. Error bars indicate standard deviation across three biological replicates. *Significant difference relative to upright control according to two-tailed Student's *t* test, using a Bonferroni-corrected *P*-value (*P** = 0.05/33).

© University of Pretoria

2.10. Tables

Table 2.1. Distribution of amino acid motifs in NAC domain proteins of dicot and monocot genomes. The total number of NAC proteins in each genome, according to the PlantTFDB (Zhang *et al.*, 2011), is indicated in parenthesis, while the number of NAC domain proteins in each genome with a match to a given motif is indicated in each row. The percentage of NAC proteins in each genome containing a particular motif is represented by a heat map.

Colour key (%): 100, 90, 80, 70, 60, 50, 40, 30, 20, 10, 0

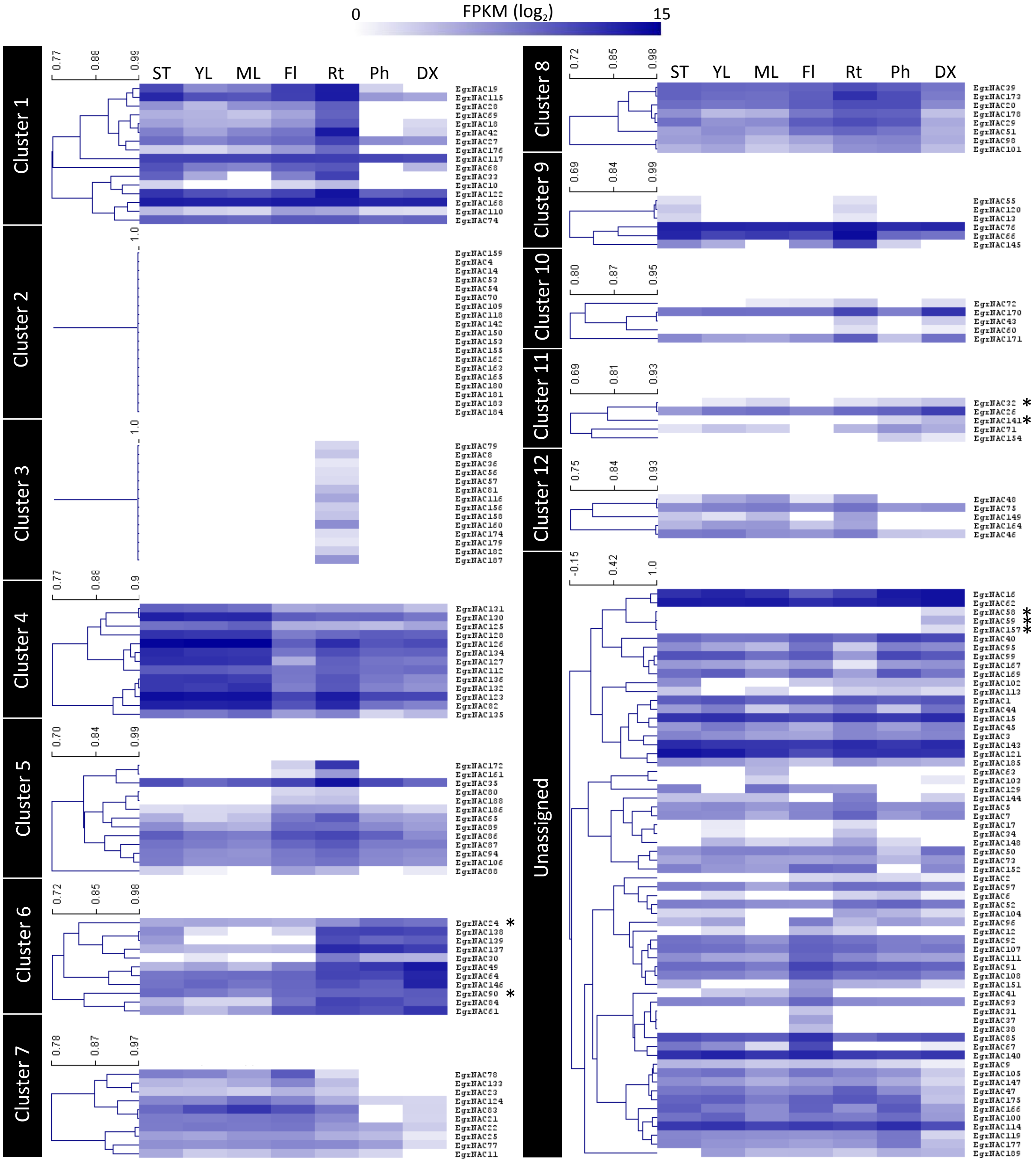
	Eucalyptus (190)	***Populus*** (192)	***Glycine*** (183)	***Carica*** (82)	***Arabidopsis*** (135)	***Vitis*** (142)	***Oryza*** (186)	***Brachypodium*** (100)	***Zea*** (190)
General motifs									
Motif 1	175	162	171	62	116	129	135	66	163
Motif 2	156	119	150	52	97	114	94	44	129
Motif 3	136	119	134	44	90	115	95	43	126
Motif 4	168	129	149	53	105	128	111	48	139
Motif 5	173	127	163	54	107	124	118	53	157
Motif 6	181	148	176	73	118	133	152	67	170
Motif 7	63	70	83	31	61	63	76	32	87
Specific motifs									
Motif 8	15	4	7	2	1	4	5	2	2
Motif 9	10	12	22	4	7	8	15	9	20
Motif 10	15	4	6	2	1	5	0	0	0
Motif 11	21	5	5	2	3	5	2	3	3
Motif 12	12	0	0	0	0	0	0	0	0
Motif 13	14	2	2	0	2	0	0	0	0
Motif 14	16	1	2	0	1	2	0	0	0
Motif 15	13	0	0	0	0	0	0	0	0
Motif 16	7	0	0	0	0	0	0	0	0

© University of Pretoria

2.11. Supplementary figures

Fig. S2.1. Quality threshold (QT) clustering of the 189 *EgrNAC* gene transcripts. Transcripts were clustered according to their expression profiles across shoot tips (ST), young leaves (YL), mature leaves (ML), flowers (Fl), roots (Rt) phloem (Ph) and developing (secondary) xylem (DX) in *Eucalyptus grandis.* Normalized RNA-seq transcript abundance data is expressed as the log2 value of fragments per kilobase of exon per million fragments mapped (FPKM). Branch lengths represent Pearson correlation coefficient, as indicated on the scale for each cluster. Asterisks indicate novel candidates potentially involved in the regulation of xylogenesis.

© University of Pretoria



© University of Pretoria

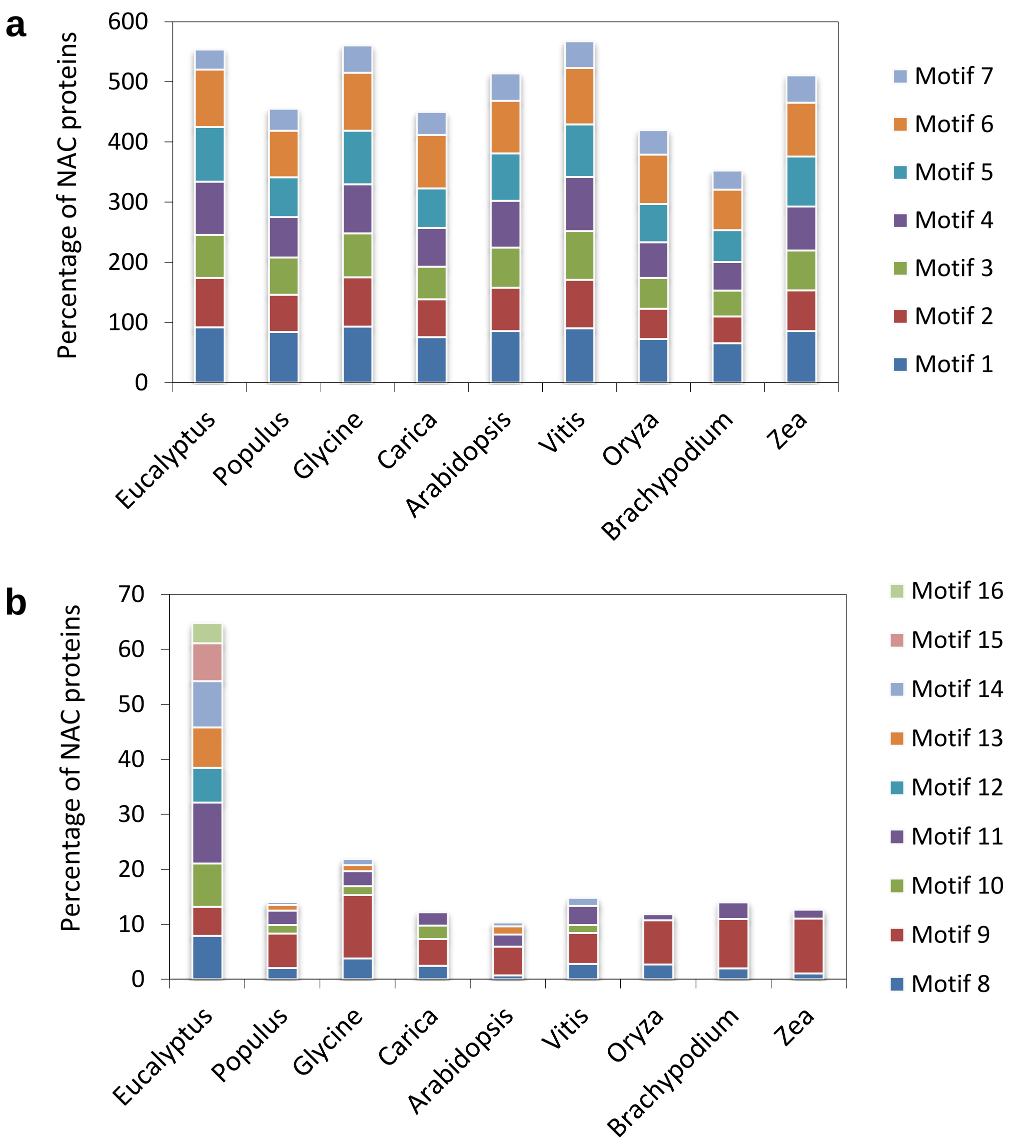


Fig. S2.2. Distribution of conserved protein motifs in eight plant genomes as identified using Hidden Markov Models built on *Eucalyptus* motif alignments. Motif frequencies are expressed as a percentage of NAC proteins containing a given motif out of the total NAC proteins in each genome. **(a)** General motifs, corresponding to subdomains A through E of the NAC domain, are distributed evenly across the genomes, showing that the Hidden Markov Models are not biased. **(b)** Specific motifs, showing enrichment in, or exclusive occurrence in, *Eucalyptus* NAC proteins with the exception of Motif 9.

© University of Pretoria

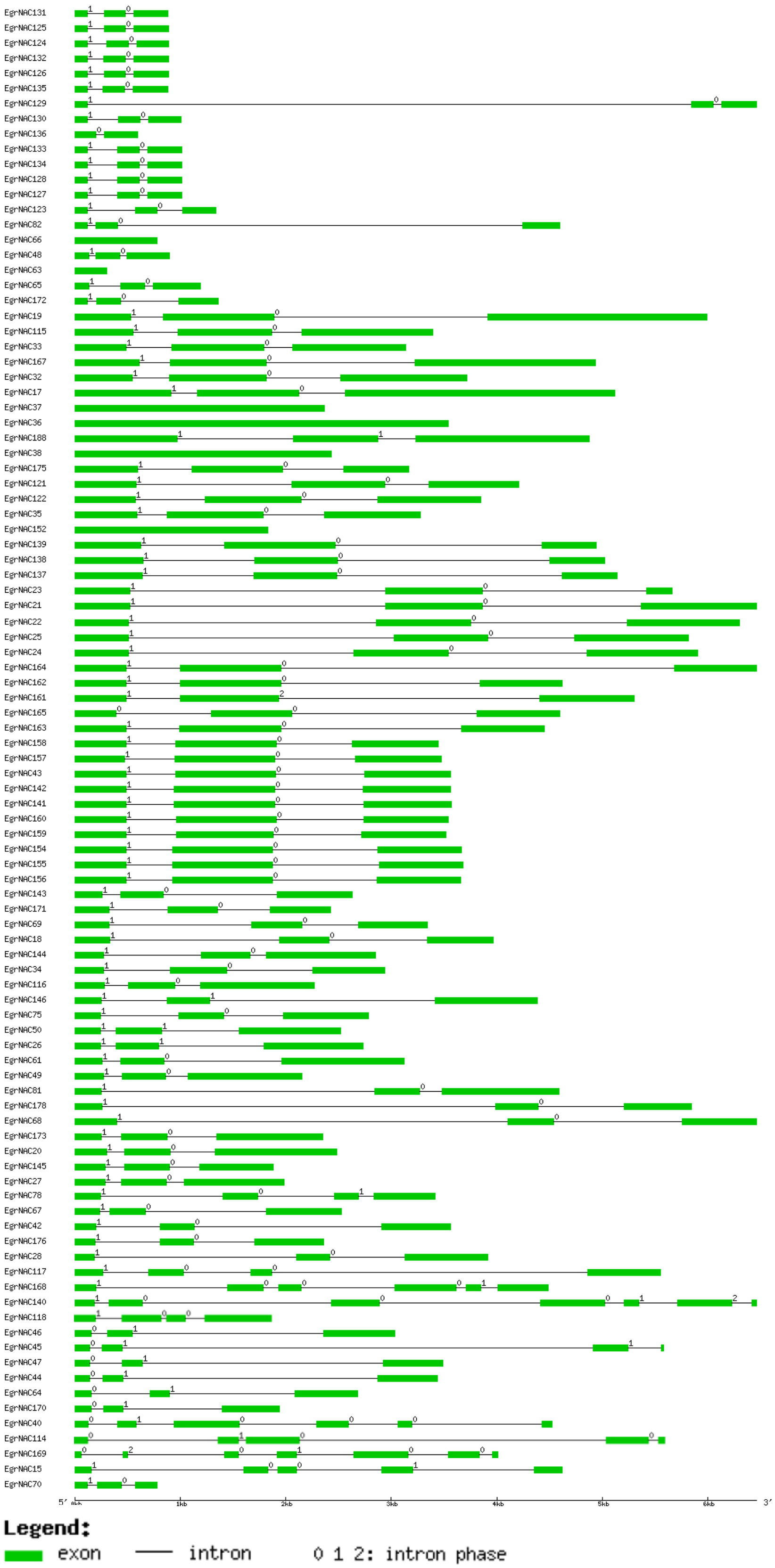


Fig. S2.3. Intron and exon structure of EgrNAC proteins, showing intron phase. The *EgrNAC* genes occur in the same order as those in Fig. 2.2 of the main manuscript.

© University of Pretoria

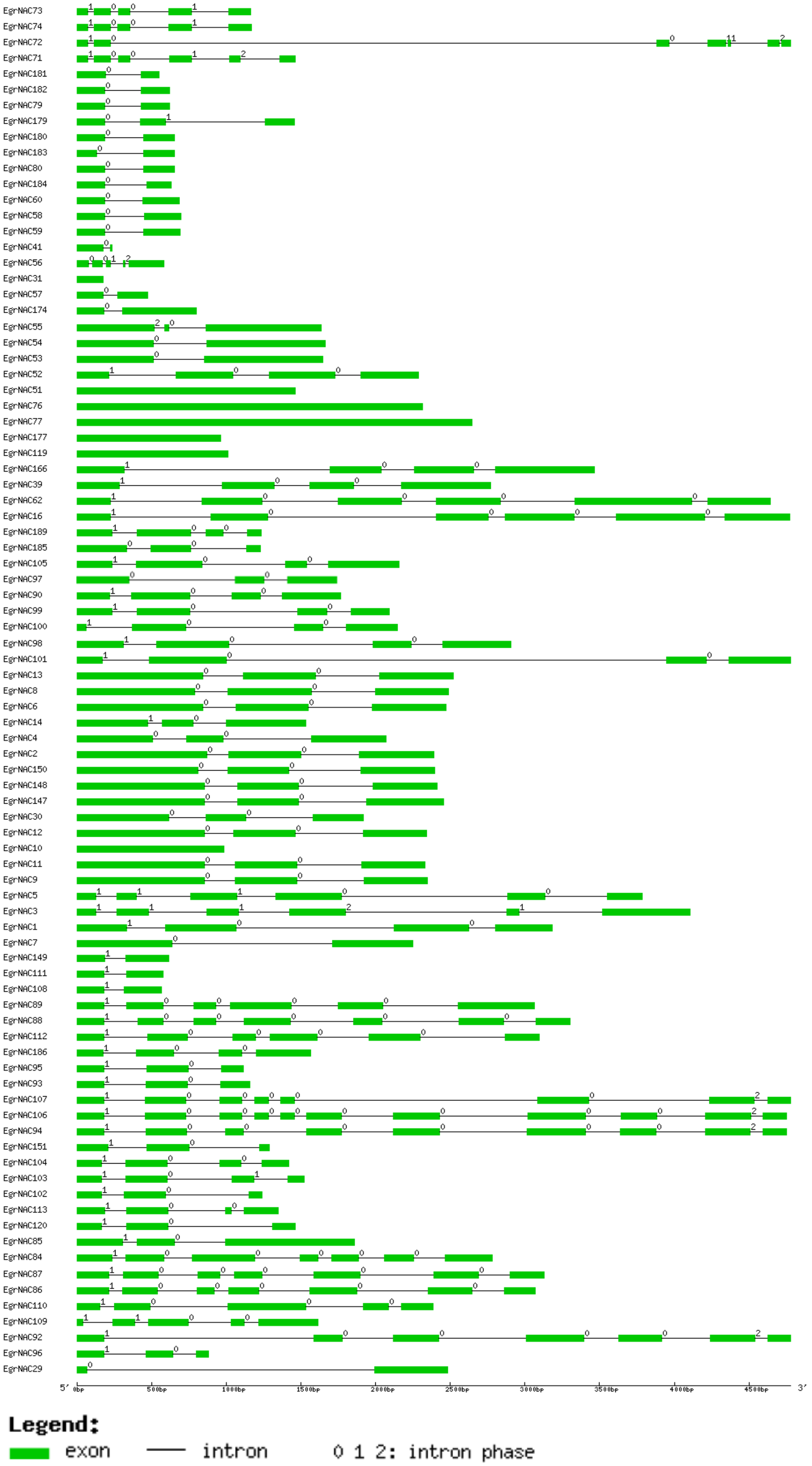
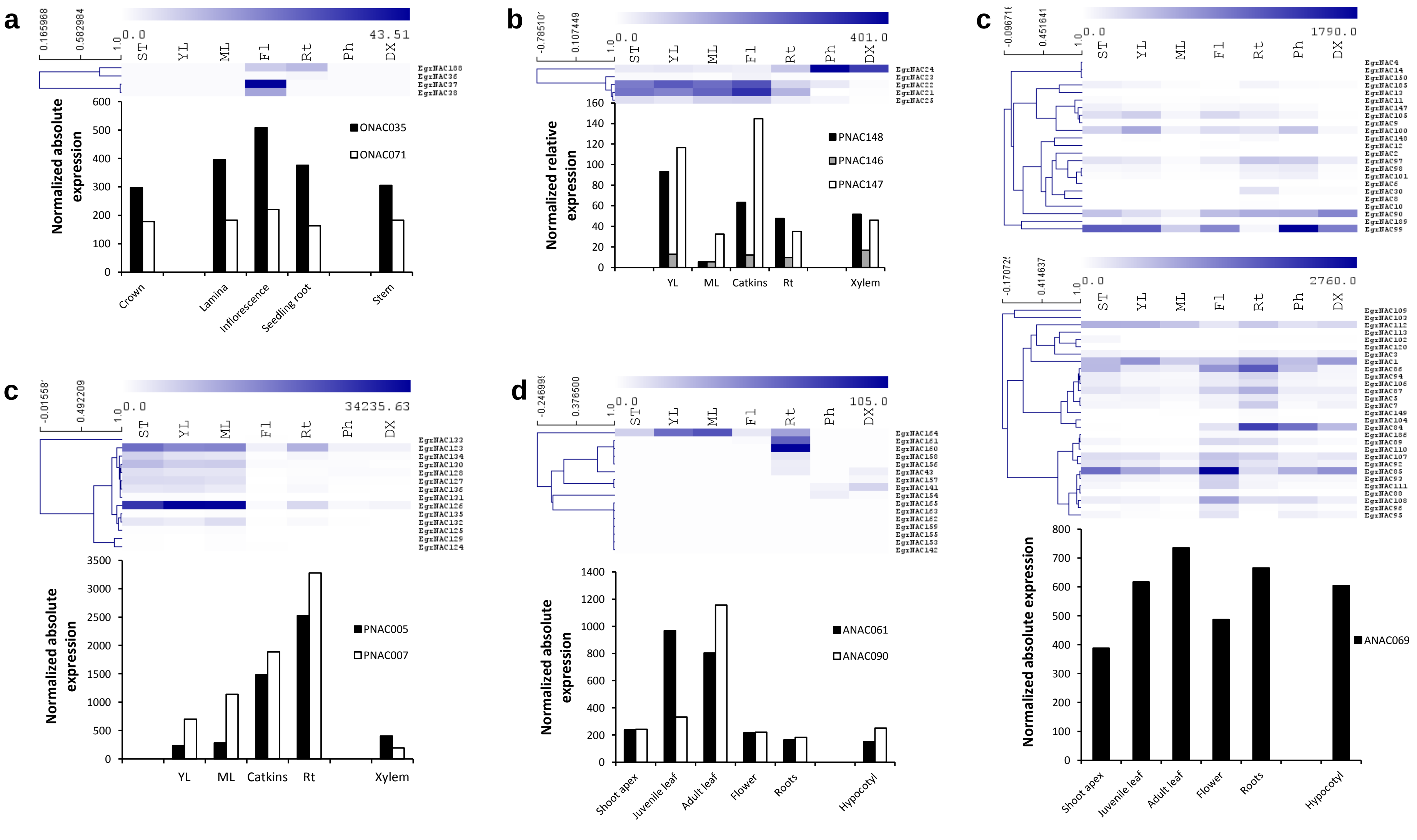


Fig. S2.3. (*continued*).

© University of Pretoria

Fig. S2.4. Comparison of tissue expression patterns of *EgrNAC* paralogs with those of their closest *Arabidopsis*, *Oryza* and/or *Populus* homologs. Absolute transcript levels from the EucGenIE database (http://eucgenie.bi.up.ac.za/) (FPKM; blue heatmap) are shown for groups of *EgrNAC* paralogous genes, arbitrarily labelled (a) through (o). Group (c) contains two blocks of paralogous genes which are phylogenetically distinct but share the same *Arabidopsis* homolog (see Additional file S2.2). Expression patterns were hierarchically clustered according to Pearson's correlation. Expression patterns of closest homologs inferred from Additional file S2.2 are shown in black and white for corresponding tissues, where available, of each *EgrNAC* paralog panel. These data were obtained from Genevestigator (Hruz *et al.*, 2008) (*Arabidopsis* and *Oryza*) or the Poplar Expression Angler developmental series (Toufighi *et al.*, 2005; Wilkins *et al.*, 2009). ST, shoot tip; YL, young leaf; ML, mature leaf; Fl, flower; Rt, root; DX, developing xylem.

© University of Pretoria



© University of Pretoria

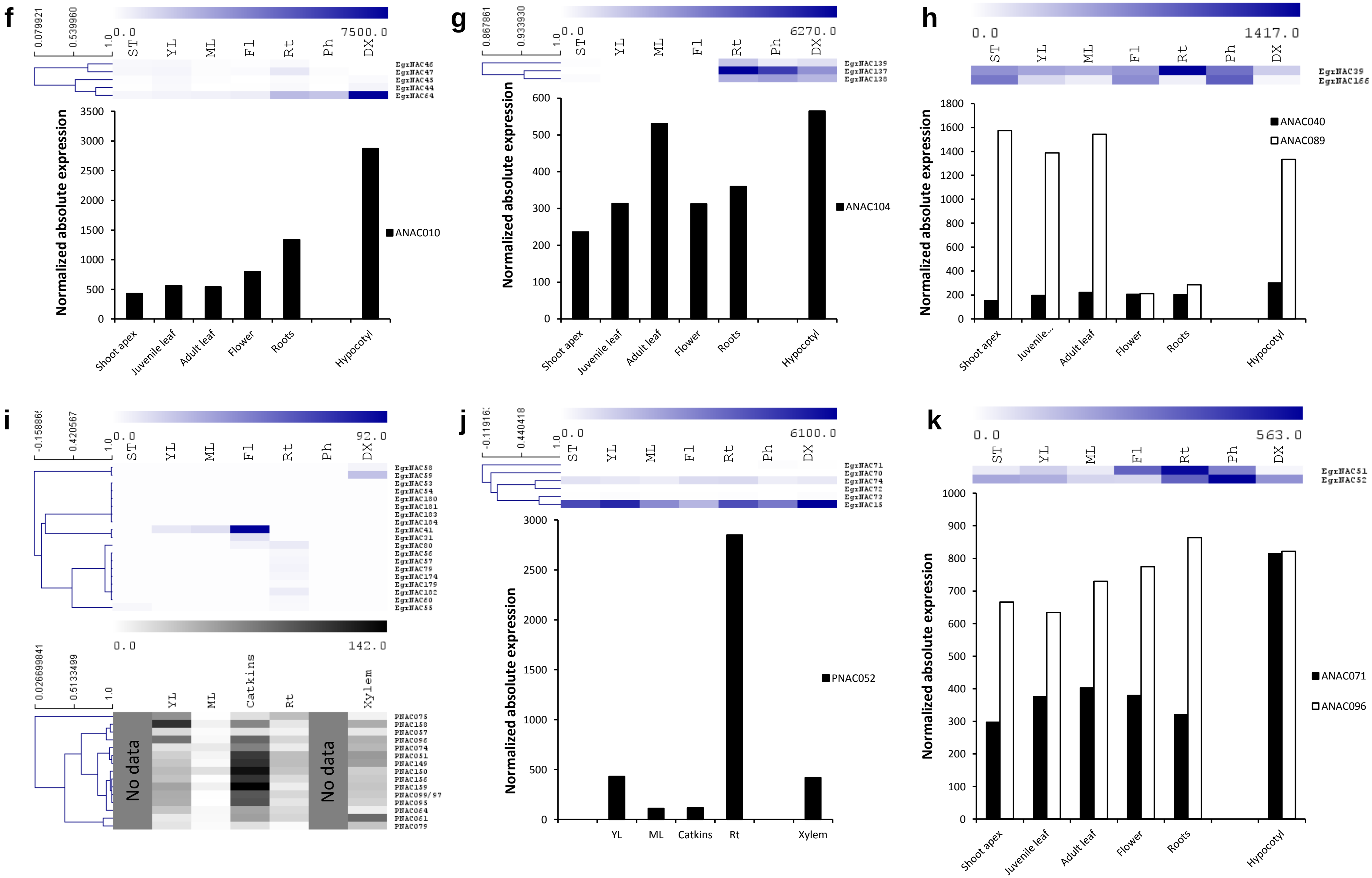


Fig. S2.4. *(continued).*

© University of Pretoria

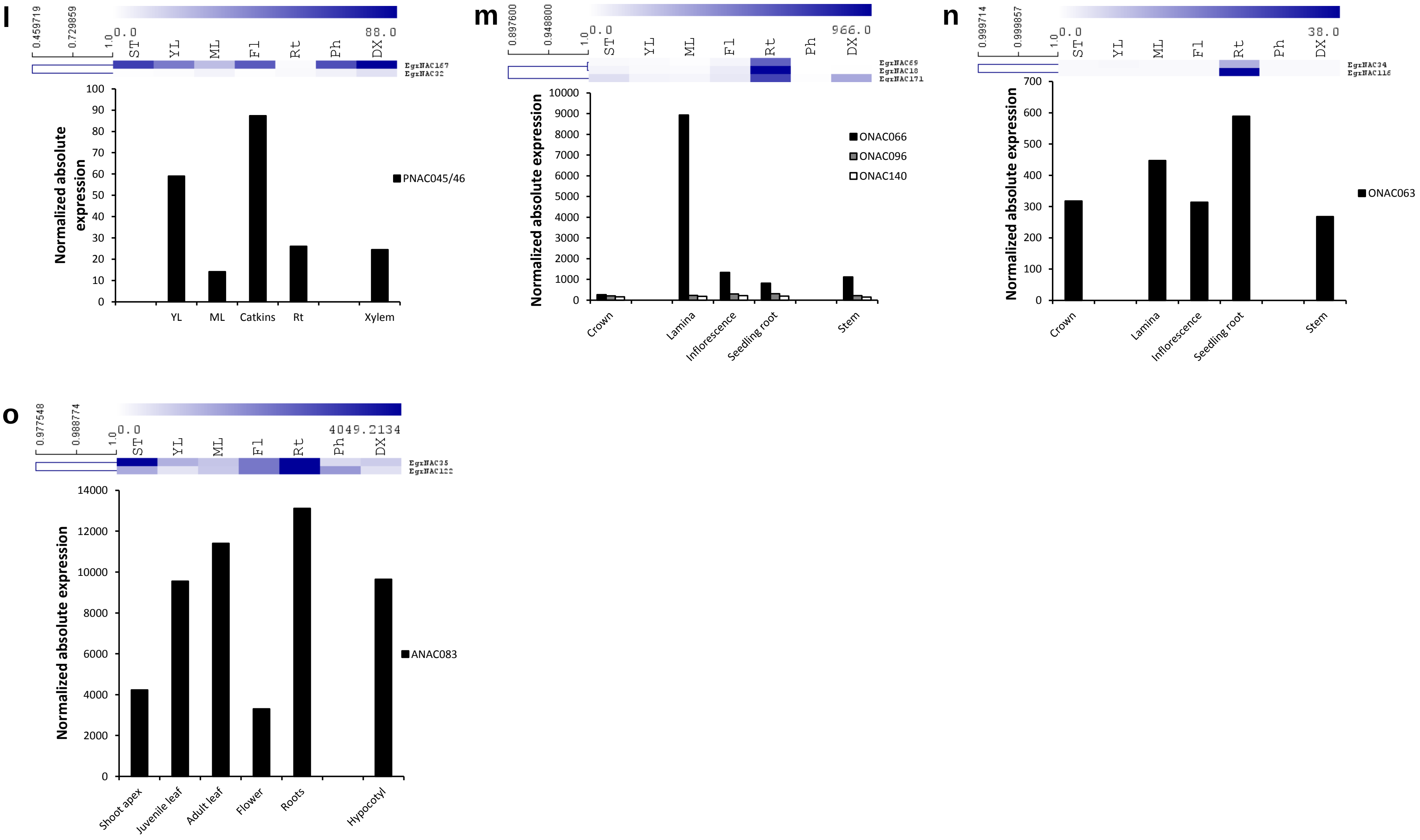


Fig. S2.4. *(continued).*

© University of Pretoria

r = 0.686
P < 0.0001

Fig. S2.5. Correlation of xylem/leaf expression ratio calculated with *E. grandis* RNA-seq data and *E. globulus* Fluidigm data for 33 NAC domain genes. A linear trendline is indicated in grey. The *P*-value represents the two-tailed probability for Pearson's correlation (*r*).

© University of Pretoria

2.12. Supplementary tables

Table S2.1. Primers pairs used for Fluidigm RT-qPCR analysis in *E. globulus*

Gene Name	Gene Symbol	Forward primer sequence [5'-3']	Reverse primer sequence [5'-3']	Amplicon length (bp)
Eucgr.A00357	EglNAC1	TTCCTCAAGTGCTGCAACTGTC	CCTGGCTAGGAAGTTGTTTGACTG	84
Eucgr.A00363	EglNAC6	GATCCACCAAACGGTCAAAC	GCCAGGTAAATCCCAAGATG	142
Eucgr.A00494	EglNAC15	GGGAAACAACAACCTGTCAACTGC	AACAGCCTGGTGGTTGCTTGTG	88
Eucgr.A00969	EglNAC16	TCAGGAAGCGACTGAAAGACAGAG	TGGAAGTTTCCTCGAGCCATCC	105
Eucgr.A02638	EglNAC24	AAGAGATCCCAGTGGTGCCAAG	TGATTTCTCCCATCGGTCTTCGG	102
Eucgr.A02887	EglNAC26	TGAGAACGGAACTGGTCAGGAAG	ATCCATGCTCGGTCATCTTGCG	92
Eucgr.B01624	EglNAC31	TCATGTTCGGTGACAAGTTCGG	GCCGCTTGAGTACTTGTGCTTC	74
Eucgr.C00958	EglNAC40	AGATTTGCGGATCCAAACAGTGC	TGCTGGTAGACCTAACCAATCCG	77
Eucgr.D00591	EglNAC44	CCAGAGAGACTTCCAGGAGTAAGC	TCTTGTGCCACCTTGTCTCAGTC	142
Eucgr.D00595	EglNAC47	TGATGTGGCCAAAGCAAGATGC	CCCTCTGATCGAACATTGGGAACC	144
Eucgr.D01671	EglNAC49	GAGCCATGGGATATCCAAGAGAGG	TTGTGGCTAAAGAAGTACCAGTCG	74
Eucgr.D02027	EglNAC50	TCAGGAGGAAGGATGGGTTGTG	AGGGCTGAGAAATCCTCCTTGG	149
Eucgr.E00574	EglNAC59	AGCATCCTCGCAACGAAACG	AGTGTCTGTGCCTTCTTTGCTC	148
Eucgr.E00575	EglNAC60	GGCGAACATGCAAGAAGATACCTG	TTCTTCAACCGATCCGCCCATTC	129
Eucgr.E01053	EglNAC61	TCGACTTGGACGTGATTCGTGAG	CCAATCCTGCACTTCTCTTGGATG	76
Eucgr.E03226	EglNAC64	AGAGGGAGAGAATGGGATCTGC	CCCATCTTTGCTCACTCCTGGTAG	64
Eucgr.F01091	EglNAC65	ATTCCCTGAGGCTGGATGACTG	TTCATGCCGTGAGGGTTCATCG	148
Eucgr.F02615	EglNAC75	AGAAACGACCACATCCCACTCC	AGGCGGTACACACGATTCCAAC	60
Eucgr.F04341	EglNAC82	AAGAACAGCTTGAGGCTTGACGAC	TTCTTCTCGATCGCGCCCTTCTTG	71
Eucgr.G01047	EglNAC84	GCATCCTGATGATACGGGCTTC	ACAGCTGCATAGTTATCTGCCTTC	78
Eucgr.G01061	EglNAC90	CGAGTACTTTGGCCAATTCAAGC	TTTCTTCCTCAGATGCCTGTGC	99
Eucgr.G01063	EglNAC91	GCCAGATGGCCTTTGTTCTCTGTC	TTCGTCGGATCGTCCATTTCCG	134
Eucgr.G01066	EglNAC93	TCAATGAACTGTTTCCACGTGCTC	CCTTGGTGCTTTAAGTGACAGACG	64
Eucgr.I00191	EglNAC137	GCTCCAACAGGTCAAGAGACAAAC	CCGGGTTCTTGTGTTGAATTAGCC	89
Eucgr.I00192	EglNAC138	CTACCACCTGGATTTCGGTTCTC	GGATGCAGAAAGTGAAGGACGAG	62
Eucgr.I00193	EglNAC139	TGTCTTCGCTCTATGCTCACTTGG	TCCATTCCACAGTGCCTTTCCG	89
Eucgr.I00583	EglNAC141	TGAACTCTCGCCGACCAATCTC	ATGCGAATTCACGCCTTAGCTC	74
Eucgr.I00587	EglNAC142	TGAGAACGGAACTGGTCAGGAAG	ATCCATGCTCGGTCATCTTGCG	92
Eucgr.I01494	EglNAC143	TGACTCGTCGCCCAAGGAAATG	TGGCGGCCTTATTCATGCCTTC	111
Eucgr.I02366	EglNAC146	ATTGCACCGAGTCTGCAAGC	TACACACGACTCCATCGTCTCC	66
Eucgr.I02695	EglNAC152	TATGATCCGTGGGAGCTTGAAGGG	ATAGGCTTCCAGTACCCGTTGC	110
Eucgr.J01038	EglNAC168	AAGGCTGGAATTCCGCAAGATG	GTTCTTTGGGCCAGAACCACTC	72
Eucgr.K01228	EglNAC171	TCCCTGGGATCTCCATGATGTTAG	CCGGATCCAGTCACTCTATTTGGC	114
Reference gene primers				
Eucgr.B02473	EF-1α	ATGCGTCAGACTGTGGCTGTTG	ATGCGTCAGACTGTGGCTGTTG	74
Eucgr.F02901	IDH	AATCGACCTGCTTCGACCCTTC	TCGACCTTGATCTTCTCGAAACCC	68
Eucgr.B03386	PP2A1	TCGAGCTTTGGACCGCATACAAG	ACCACAAGAGGTCACACATTGGC	62
Eucgr.B03031	PP2A3	CAGCGGCAAACAACTTGAAGCG	ATTATGTGCTGCATTGCCCAGTC	67
Eucgr.B02502	SAND	TTGATCCACTTGCGGACAAGGC	TCACCCATTGACATACACGATTGC	63
gDNA contamination assessment				
Intergenic	3’ of Eucgr.H02589	GCGGCTTTTAAGTCTCTTGCGAA	TTCGAAGCATAGCTTCGCCATATG	150

© University of Pretoria

Table S2.2. Classification of NAC domain genes identified in the v.1.0 *E. grandis* genome assembly (www.phytozome.net).

Manually curated *EgrNAC* gene models					
Eucgr.A00357.1	Eucgr.B03208.1	Eucgr.E03226.1	Eucgr.G01070.1	Eucgr.I00059.4	Eucgr.J00517.1
Eucgr.A00359.1	Eucgr.B03439.1	Eucgr.F01091.1	Eucgr.G01071.1	Eucgr.I00060.1	Eucgr.J00518.1
Eucgr.A00360.1	Eucgr.B03537.1	Eucgr.F01093.1	Eucgr.G01074.1	Eucgr.I00060.2	Eucgr.J00519.1
Eucgr.A00361.1	Eucgr.B03693.1	Eucgr.F01170.1	Eucgr.G01075.1	Eucgr.I00095.1	Eucgr.J00520.1
Eucgr.A00362.1	Eucgr.B03703.1	Eucgr.F01449.1	Eucgr.G01077.1	Eucgr.I00099.1	Eucgr.J00521.1
Eucgr.A00363.1	Eucgr.B03704.1	Eucgr.F01463.1	Eucgr.G01078.1	Eucgr.I00100.1	Eucgr.J00531.1
Eucgr.A00364.1	Eucgr.B03823.1	Eucgr.F01535.1	Eucgr.G01081.1	Eucgr.I00101.1	Eucgr.J00940.1
Eucgr.A00365.1	Eucgr.C00958.1	Eucgr.F01536.1	Eucgr.G01082.1	Eucgr.I00102.1	Eucgr.J01038.1
Eucgr.A00368.1	Eucgr.C01264.1	Eucgr.F01537.1	Eucgr.G01083.1	Eucgr.I00191.1	Eucgr.J02254.1
Eucgr.A00369.1	Eucgr.C02105.1	Eucgr.F01538.1	Eucgr.G01507.1	Eucgr.I00192.1	Eucgr.K01061.1
Eucgr.A00370.1	Eucgr.C02446.1	Eucgr.F01539.1	Eucgr.G01548.1	Eucgr.I00193.1	Eucgr.K01228.1
Eucgr.A00371.1	Eucgr.D00591.1	Eucgr.F02615.1	Eucgr.G01550.1	Eucgr.I00213.1	Eucgr.K01471.1
Eucgr.A00435.1	Eucgr.D00592.1	Eucgr.F02771.1	Eucgr.G01551.1	Eucgr.I00583.1	Eucgr.K01472.1
Eucgr.A00437.1	Eucgr.D00593.1	Eucgr.F02910.1	Eucgr.G01553.1	Eucgr.I00587.1	Eucgr.K01845.1
Eucgr.A00494.1	Eucgr.D00594.1	Eucgr.F03588.1	Eucgr.G01554.1	Eucgr.I01494.1	Eucgr.K02205.1
Eucgr.A00969.1	Eucgr.D00595.1	Eucgr.F03962.1	Eucgr.G01555.1	Eucgr.I01940.1	Eucgr.K02225.1
Eucgr.A01272.1	Eucgr.D00665.1	Eucgr.F03963.1	Eucgr.G01758.1	Eucgr.I01958.1	Eucgr.K02303.1
Eucgr.A02028.1	Eucgr.D01671.1	Eucgr.F04097.1	Eucgr.G01984.1	Eucgr.I02366.1	Eucgr.K03256.1
Eucgr.A02070.1	Eucgr.D02027.1	Eucgr.F04341.1	Eucgr.G02349.1	Eucgr.I02571.1	Eucgr.K03356.1
Eucgr.A02074.1	Eucgr.D02182.1	Eucgr.G00054.1	Eucgr.G02486.1	Eucgr.I02573.1	Eucgr.K03357.1
Eucgr.A02635.1	Eucgr.E00298.1	Eucgr.G01047.1	Eucgr.G02506.1	Eucgr.I02574.1	Eucgr.K03358.1
Eucgr.A02636.1	Eucgr.E00541.1	Eucgr.G01049.1	Eucgr.G02740.1	Eucgr.I02576.1	Eucgr.K03359.1
Eucgr.A02637.1	Eucgr.E00542.1	Eucgr.G01052.1	Eucgr.G02742.1	Eucgr.I02578.1	Eucgr.K03360.1
Eucgr.A02638.1	Eucgr.E00543.1	Eucgr.G01053.1	Eucgr.H00614.1	Eucgr.I02695.1	Eucgr.K03361.1
Eucgr.A02639.1	Eucgr.E00545.1	Eucgr.G01058.1	Eucgr.H00826.1	Eucgr.J00505.1	Eucgr.L00819.1
Eucgr.A02887.1	Eucgr.E00551.1	Eucgr.G01060.1	Eucgr.H03362.1	Eucgr.J00508.1	Eucgr.L01867.1
Eucgr.B00529.1	Eucgr.E00573.1	Eucgr.G01061.1	Eucgr.H03387.1	Eucgr.J00509.1	Eucgr.L02267.1
Eucgr.B00724.1	Eucgr.E00574.1	Eucgr.G01063.1	Eucgr.H05089.1	Eucgr.J00511.1	Eucgr.L02674.1
Eucgr.B01567.1	Eucgr.E00575.1	Eucgr.G01064.1	Eucgr.I00056.1	Eucgr.J00512.1	Eucgr.L03347.1
Eucgr.B01593.1	Eucgr.E01053.1	Eucgr.G01066.1	Eucgr.I00057.1	Eucgr.J00513.1	
Eucgr.B01624.1	Eucgr.E01095.1	Eucgr.G01067.1	Eucgr.I00058.1	Eucgr.J00514.1	
Eucgr.B02485.1	Eucgr.E03225.1	Eucgr.G01069.1	Eucgr.I00059.1	Eucgr.J00516.1	
Alternative splice variants					
Eucgr.A00357.2	Eucgr.D00593.2	Eucgr.F02771.5	Eucgr.G01548.2	Eucgr.I00213.2	Eucgr.L01924.2
Eucgr.A00494.2	Eucgr.E01095.2	Eucgr.F04341.2	Eucgr.G02486.2	Eucgr.I00213.3	Eucgr.L02268.2
Eucgr.A02028.2	Eucgr.E03226.2	Eucgr.G01047.2	Eucgr.G02740.2	Eucgr.I00213.4	
Eucgr.A02887.2	Eucgr.F01463.2	Eucgr.G01067.2	Eucgr.I00059.2	Eucgr.I00213.5	
Eucgr.B03537.2	Eucgr.F02771.2	Eucgr.G01067.3	Eucgr.I00059.3	Eucgr.I00213.6	
Eucgr.C00958.2	Eucgr.F02771.3	Eucgr.G01067.4	Eucgr.I00100.2	Eucgr.I02366.2	
Eucgr.C00958.3	Eucgr.F02771.4	Eucgr.G01067.5	Eucgr.I00191.2	Eucgr.K01228.2	
Putative alleles					
Eucgr.L01840.1	Eucgr.L02201.1	Eucgr.L02268.1	Eucgr.L02673.1	Eucgr.L02696.1	Eucgr.L03434.1
Eucgr.L01924.1	Eucgr.L02202.1	Eucgr.L02499.1	Eucgr.L02683.1	Eucgr.L02867.1	
Eucgr.L01925.1	Eucgr.L02266.1	Eucgr.L02501.1	Eucgr.L02695.1	Eucgr.L03094.1	
Failed manual curation					
Eucgr.A01274.1	Eucgr.G01265.1	Eucgr.H03391.1	Eucgr.L02177.1		
Eucgr.A01885.1	Eucgr.G01267.1	Eucgr.I02577.1			
Eucgr.G01073.1	Eucgr.G01448.1	Eucgr.J01735.1			
Collapsed into a single gene model					
Eucgr.I00097.1	Eucgr.I00098.1				

© University of Pretoria

Table S2.3. Nomenclature, lengths and coordinates of 189 NAC domain proteins identified in the draft *E. grandis* genome assembly (V1.0, www.phytozome.net).

Name	Gene	Protein length[a]	Locus[b]	Strand
EgrNAC1	Eucgr.A00357	308	scaffold_1:4972108-4973849	-
EgrNAC2	Eucgr.A00359	338	scaffold_1:4977347-4978655	-
EgrNAC3	Eucgr.A00360	292	scaffold_1:5011350-5013598	-
EgrNAC4	Eucgr.A00361	228	scaffold_1:5067016-5068149	-
EgrNAC5	Eucgr.A00362	273	scaffold_1:5073131-5075204	-
EgrNAC6	Eucgr.A00363	333	scaffold_1:5078319-5079673	-
EgrNAC7	Eucgr.A00364	213	scaffold_1:5102051-5103282	-
EgrNAC8	Eucgr.A00365	335	scaffold_1:5107090-5108452	-
EgrNAC9	Eucgr.A00368	308	scaffold_1:5180298-5181584	-
EgrNAC10	Eucgr.A00369	179	scaffold_1:5197166-5197706	-
EgrNAC11	Eucgr.A00370	308	scaffold_1:5211733-5213010	-
EgrNAC12	Eucgr.A00371	308	scaffold_1:5228116-5229398	-
EgrNAC13	Eucgr.A00435	333	scaffold_1:6414247-6415628	-
EgrNAC14	Eucgr.A00437	221	scaffold_1:6448797-6449635	-
EgrNAC15	Eucgr.A00494	374	scaffold_1:7719332-7723951	-
EgrNAC16	Eucgr.A00969	597	scaffold_1:15385239-15388740	-
EgrNAC17	Eucgr.A01272	466	scaffold_1:20607436-20609052	-
EgrNAC18	Eucgr.A02028	312	scaffold_1:31019047-31021637	-
EgrNAC19	Eucgr.A02070	385	scaffold_1:31501255-31503148	+
EgrNAC20	Eucgr.A02074	410	scaffold_1:31544167-31545788	+
EgrNAC21	Eucgr.A02635	266	scaffold_1:36895028-36897068	-
EgrNAC22	Eucgr.A02636	266	scaffold_1:36937941-36939979	-
EgrNAC23	Eucgr.A02637	177	scaffold_1:36965867-36967655	-
EgrNAC24	Eucgr.A02638	264	scaffold_1:36976003-36977913	-
EgrNAC25	Eucgr.A02639	268	scaffold_1:37001186-37003068	-
EgrNAC26	Eucgr.A02887	348	scaffold_1:39264076-39265859	+
EgrNAC27	Eucgr.B00529	361	scaffold_2:6904675-6905972	+
EgrNAC28	Eucgr.B00724	357	scaffold_2:9040994-9044251	+
EgrNAC29	Eucgr.B01567	184	scaffold_2:26103161-26105643	+
EgrNAC30	Eucgr.B01593	221	scaffold_2:26870281-26871330	-
EgrNAC31	Eucgr.B01624	127	scaffold_2:27726776-27727160	+
EgrNAC32	Eucgr.B02485	279	scaffold_2:47008154-47009327	-
EgrNAC33	Eucgr.B03208	255	scaffold_2:57054023-57055014	+
EgrNAC34	Eucgr.B03439	326	scaffold_2:59177735-59179654	+
EgrNAC35	Eucgr.B03537	253	scaffold_2:60000406-60001441	+
EgrNAC36	Eucgr.B03693	372	scaffold_2:61396803-61398876	-
EgrNAC37	Eucgr.B03703	248	scaffold_2:61459361-61460515	-
EgrNAC38	Eucgr.B03704	255	scaffold_2:61468791-61470297	-
EgrNAC39	Eucgr.B03823	371	scaffold_2:62322277-62324308	-
EgrNAC40	Eucgr.C00958	486	scaffold_3:14868676-14873196	+
EgrNAC41	Eucgr.C01264	135	scaffold_3:19938845-19939356	-
EgrNAC42	Eucgr.C02105	326	scaffold_3:38063370-38066334	+
EgrNAC43	Eucgr.C02446	242	scaffold_3:46604381-46605534	+
EgrNAC44	Eucgr.D00591	300	scaffold_4:10891985-10895422	-
EgrNAC45	Eucgr.D00592	229	scaffold_4:10923473-10929053	-
EgrNAC46	Eucgr.D00594	295	scaffold_4:10995874-10998399	-
EgrNAC47	Eucgr.D00595	300	scaffold_4:11010756-11014245	-
EgrNAC48	Eucgr.D00665	343	scaffold_4:12112914-12114131	+
EgrNAC49	Eucgr.D01671	383	scaffold_4:30694508-30695914	+
EgrNAC50	Eucgr.D02027	355	scaffold_4:34311770-34313414	+
EgrNAC51	Eucgr.D02182	357	scaffold_4:36060390-36062440	+
EgrNAC52	Eucgr.E00298	346	scaffold_5:2804334-2806009	-
EgrNAC53	Eucgr.E00541	318	scaffold_5:5139951-5141158	-
EgrNAC54	Eucgr.E00542	318	scaffold_5:5162361-5163579	-
EgrNAC55	Eucgr.E00543	322	scaffold_5:5171768-5172969	-
EgrNAC56	Eucgr.E00545	297	scaffold_5:5184896-5186151	-
EgrNAC57	Eucgr.E00551	269	scaffold_5:5278815-5279837	-
EgrNAC58	Eucgr.E00573	312	scaffold_5:5446847-5448354	+
EgrNAC59	Eucgr.E00574	312	scaffold_5:5454671-5456169	+
EgrNAC60	Eucgr.E00575	312	scaffold_5:5463114-5464594	+
EgrNAC61	Eucgr.E01053	399	scaffold_5:11289968-11292006	-
EgrNAC62	Eucgr.E01095	656	scaffold_5:11688351-11691755	-
EgrNAC63	Eucgr.E03225	136	scaffold_5:54730393-54730804	+
EgrNAC64	Eucgr.E03226	313	scaffold_5:54751853-54754536	-

© University of Pretoria

Table S2.3. *(continued)*

Name	Gene	Protein length[a]	Locus[b]	Strand
EgrNAC65	Eucgr.F01091	366	scaffold_6:14069990-14071611	-
EgrNAC66	Eucgr.F01093	353	scaffold_6:14089859-14091139	-
EgrNAC67	Eucgr.F01170	357	scaffold_6:14978729-14980833	-
EgrNAC68	Eucgr.F01449	334	scaffold_6:18724670-18728883	-
EgrNAC69	Eucgr.F01463	316	scaffold_6:18838114-18840295	-
EgrNAC70	Eucgr.F01535	185	scaffold_6:19715905-19716689	-
EgrNAC71	Eucgr.F01536	420	scaffold_6:19771807-19774964	-
EgrNAC72	Eucgr.F01537	377	scaffold_6:19785158-19795459	-
EgrNAC73	Eucgr.F01538	401	scaffold_6:19808330-19810843	-
EgrNAC74	Eucgr.F01539	402	scaffold_6:19823490-19826009	-
EgrNAC75	Eucgr.F02615	320	scaffold_6:35802610-35804425	-
EgrNAC76	Eucgr.F02771	565	scaffold_6:37242090-37245705	+
EgrNAC77	Eucgr.F02910	645	scaffold_6:38768675-38772557	+
EgrNAC78	Eucgr.F03588	383	scaffold_6:44293211-44296055	-
EgrNAC79	Eucgr.F03962	271	scaffold_6:47920826-47922168	-
EgrNAC80	Eucgr.F03963	282	scaffold_6:47930910-47932324	-
EgrNAC81	Eucgr.F04097	386	scaffold_6:49244683-49247676	+
EgrNAC82	Eucgr.F04341	302	scaffold_6:52357968-52364215	+
EgrNAC83	Eucgr.G00054	432	scaffold_7:561581-563631	+
EgrNAC84	Eucgr.G01047	566	scaffold_7:18148061-18150813	-
EgrNAC85	Eucgr.G01049	468	scaffold_7:18173499-18175338	-
EgrNAC86	Eucgr.G01052	528	scaffold_7:18187867-18190938	-
EgrNAC87	Eucgr.G01053	540	scaffold_7:18203111-18206238	-
EgrNAC88	Eucgr.G01058	505	scaffold_7:18290552-18293821	-
EgrNAC89	Eucgr.G01060	592	scaffold_7:18328319-18331355	-
EgrNAC90	Eucgr.G01061	291	scaffold_7:18350722-18352015	-
EgrNAC91	Eucgr.G01063	208	scaffold_7:18362215-18363154	-
EgrNAC92	Eucgr.G01064	599	scaffold_7:18367448-18372224	-
EgrNAC93	Eucgr.G01066	214	scaffold_7:18382203-18383350	-
EgrNAC94	Eucgr.G01067	726	scaffold_7:18399300-18404004	-
EgrNAC95	Eucgr.G01069	200	scaffold_7:18414360-18415464	-
EgrNAC96	Eucgr.G01070	148	scaffold_7:18422025-18422909	-
EgrNAC97	Eucgr.G01071	215	scaffold_7:18430028-18431574	-
EgrNAC98	Eucgr.G01074	276	scaffold_7:18459516-18461108	-
EgrNAC99	Eucgr.G01075	253	scaffold_7:18464202-18465735	-
EgrNAC100	Eucgr.G01077	241	scaffold_7:18472658-18474231	-
EgrNAC101	Eucgr.G01078	249	scaffold_7:18479499-18482113	-
EgrNAC102	Eucgr.G01081	173	scaffold_7:18489388-18490615	-
EgrNAC103	Eucgr.G01082	229	scaffold_7:18494774-18496279	-
EgrNAC104	Eucgr.G01083	252	scaffold_7:18501578-18502984	-
EgrNAC105	Eucgr.G01507	314	scaffold_7:26090034-26091615	+
EgrNAC106	Eucgr.G01548	799	scaffold_7:26977231-26981938	+
EgrNAC107	Eucgr.G01550	525	scaffold_7:26995211-26999940	+
EgrNAC108	Eucgr.G01551	142	scaffold_7:27102229-27102788	+
EgrNAC109	Eucgr.G01553	314	scaffold_7:27116820-27118435	+
EgrNAC110	Eucgr.G01554	433	scaffold_7:27156206-27158590	+
EgrNAC111	Eucgr.G01555	142	scaffold_7:27167842-27168414	+
EgrNAC112	Eucgr.G01758	490	scaffold_7:32483475-32486540	+
EgrNAC113	Eucgr.G01984	241	scaffold_7:35939865-35941201	-
EgrNAC114	Eucgr.G02349	429	scaffold_7:41845203-41851658	-
EgrNAC115	Eucgr.G02486	282	scaffold_7:43382045-43383118	+
EgrNAC116	Eucgr.G02506	390	scaffold_7:43532994-43534477	-
EgrNAC117	Eucgr.G02740	412	scaffold_7:45473081-45477702	-
EgrNAC118	Eucgr.G02742	383	scaffold_7:45483146-45484914	-
EgrNAC119	Eucgr.H00614	246	scaffold_8:8324740-8327663	-
EgrNAC120	Eucgr.H00826	196	scaffold_8:10366720-10368168	-
EgrNAC121	Eucgr.H03362	243	scaffold_8:49228861-49230191	+
EgrNAC122	Eucgr.H03387	259	scaffold_8:49527394-49528610	-
EgrNAC123	Eucgr.H05089	288	scaffold_8:72636728-72638548	-
EgrNAC124	Eucgr.I00056	281	scaffold_9:932401-933614	-
EgrNAC125	Eucgr.I00057	293	scaffold_9:945378-946589	-
EgrNAC126	Eucgr.I00058	294	scaffold_9:960712-961926	-
EgrNAC127	Eucgr.I00059	293	scaffold_9:982712-984095	-
EgrNAC128	Eucgr.I00059.4	293	scaffold_9:982711-984095	-
EgrNAC129	Eucgr.I00060	294	scaffold_9:1010872-1012226	-
EgrNAC130	Eucgr.I00060.2	286	scaffold_9:1010871-1012247	-

© University of Pretoria

Table S2.3. *(continued)*

Name	Gene	Protein length[a]	Locus[b]	Strand
EgrNAC131	Eucgr.I00095	293	scaffold_9:1990554-1991758	-
EgrNAC132	Eucgr.I00097(8)[c]	294	scaffold_9:2015283-2016494	-
EgrNAC133	Eucgr.I00099	293	scaffold_9:2029621-2031006	-
EgrNAC134	Eucgr.I00100	293	scaffold_9:2037886-2039269	-
EgrNAC135	Eucgr.I00101	294	scaffold_9:2054503-2055704	-
EgrNAC136	Eucgr.I00102	231	scaffold_9:2062222-2063030	-
EgrNAC137	Eucgr.I00191	204	scaffold_9:3901232-3902855	+
EgrNAC138	Eucgr.I00192	205	scaffold_9:3933513-3935099	+
EgrNAC139	Eucgr.I00193	231	scaffold_9:3942824-3944386	+
EgrNAC140	Eucgr.I00213	628	scaffold_9:4485737-4491109	-
EgrNAC141	Eucgr.I00583	244	scaffold_9:11983778-11984934	-
EgrNAC142	Eucgr.I00587	244	scaffold_9:12090444-12091597	-
EgrNAC143	Eucgr.I01494	301	scaffold_9:25146958-25148674	+
EgrNAC144	Eucgr.I01940	386	scaffold_9:29288937-29290797	+
EgrNAC145	Eucgr.I01958	305	scaffold_9:29495058-29496284	+
EgrNAC146	Eucgr.I02366	353	scaffold_9:34184572-34187433	-
EgrNAC147	Eucgr.I02571	324	scaffold_9:37055006-37056349	+
EgrNAC148	Eucgr.I02573	307	scaffold_9:37072424-37073743	+
EgrNAC149	Eucgr.I02574	156	scaffold_9:37077815-37078425	+
EgrNAC150	Eucgr.I02576	312	scaffold_9:37088684-37089995	+
EgrNAC151	Eucgr.I02578	184	scaffold_9:37097835-37099113	+
EgrNAC152	Eucgr.I02695	192	scaffold_9:38086280-38087232	+
EgrNAC153	Eucgr.J00505	240	scaffold_10:5382997-5384192	+
EgrNAC154	Eucgr.J00508	240	scaffold_10:5407603-5408789	+
EgrNAC155	Eucgr.J00509	240	scaffold_10:5438076-5439267	+
EgrNAC156	Eucgr.J00511	240	scaffold_10:5480784-5481969	+
EgrNAC157	Eucgr.J00512	241	scaffold_10:5500331-5501455	+
EgrNAC158	Eucgr.J00513	242	scaffold_10:5565803-5566917	+
EgrNAC159	Eucgr.J00514	237	scaffold_10:5644545-5645683	+
EgrNAC160	Eucgr.J00516	240	scaffold_10:5675996-5677141	+
EgrNAC161	Eucgr.J00517	249	scaffold_10:5699131-5700846	+
EgrNAC162	Eucgr.J00518	239	scaffold_10:5741072-5742566	+
EgrNAC163	Eucgr.J00519	240	scaffold_10:5770016-5771456	+
EgrNAC164	Eucgr.J00520	239	scaffold_10:5800427-5802517	+
EgrNAC165	Eucgr.J00521	209	scaffold_10:5822540-5824027	+
EgrNAC166	Eucgr.J00531	420	scaffold_10:5899713-5902252	+
EgrNAC167	Eucgr.J00940	340	scaffold_10:10271423-10272983	+
EgrNAC168	Eucgr.J01038	538	scaffold_10:11338856-11342587	-
EgrNAC169	Eucgr.J02254	430	scaffold_10:28401471-28405482	+
EgrNAC170	Eucgr.K01061	296	scaffold_11:13333715-13335656	-
EgrNAC171	Eucgr.K01228	297	scaffold_11:15492835-15494419	-
EgrNAC172	Eucgr.K01471	329	scaffold_11:17825556-17827411	+
EgrNAC173	Eucgr.K01472	365	scaffold_11:17839379-17840912	+
EgrNAC174	Eucgr.K01845	487	scaffold_11:23025448-23027173	+
EgrNAC175	Eucgr.K02205	218	scaffold_11:29289726-29290726	+
EgrNAC176	Eucgr.K02225	322	scaffold_11:29515773-29517736	+
EgrNAC177	Eucgr.K02303	235	scaffold_11:30215285-30218757	-
EgrNAC178	Eucgr.K03256	283	scaffold_11:41341485-41345296	+
EgrNAC179	Eucgr.K03356	397	scaffold_11:42561958-42565099	+
EgrNAC180	Eucgr.K03357	281	scaffold_11:42567237-42568645	+
EgrNAC181	Eucgr.K03358	226	scaffold_11:42572579-42573772	-
EgrNAC182	Eucgr.K03359	271	scaffold_11:42582884-42584228	+
EgrNAC183	Eucgr.K03360	245	scaffold_11:42604336-42605747	+
EgrNAC184	Eucgr.K03361	249	scaffold_11:42608150-42609513	+
EgrNAC185	Eucgr.L00819	168	scaffold_69:6684-7584	-
EgrNAC186	Eucgr.L01867	310	scaffold_423:7584-9133	-
EgrNAC187	Eucgr.L02267	82	scaffold_741:6537-7111	+
EgrNAC188	Eucgr.L02674	359	scaffold_1217:9229-10769	+
EgrNAC189	Eucgr.L03347	197	scaffold_2771:26-933	+

[a] In amino acids

[b] Excluding untranslated regions

[c] Collapsed gene models Eucgr.I00097 and Eucgr.I00098

© University of Pretoria

Table S2.4. Biological functions of functionally characterized NAC domain proteins occurring in the subfamilies annotated in Fig. 2.1.

Protein	Subfamily	General function	Specific function	References
ANAC054/CUC1	Ia	Development	Organ separation, gynoecium development	Aida *et al.*, 1997; Ishida *et al.*, 2000
ANAC098/CUC2			Organ separation, leaf development, axillary meristem formation	Aida *et al.*, 1997; Nikovics *et al.*, 2006; Peaucelle *et al.*, 2007; Raman *et al.*, 2008
ANAC031/CUC3			Organ separation, meristem initiation	Vroemen *et al.*, 2003; Hibara *et al.*, 2006
ANAC092			Leaf senescence	Oh *et al.*, 1997; Kim *et al.*, 2009; Balazadeh *et al.*, 2010
ANAC021	Ib	Development	Lateral root development, apical meristem specification	He *et al.*, 2005
ANAC030/VND7	Ic	Cell wall development	Secondary cell wall biosynthesis in xylem vessels	Kubo *et al.*, 2005; Yamaguchi *et al.*, 2010a; Zhong *et al.*, 2010; Yamaguchi *et al.*, 2011
ANAC101/VND6			Secondary cell wall biosynthesis in xylem vessels	Kubo *et al.*, 2005; Ohashi-Ito *et al.*, 2010; Yamaguchi *et al.*, 2010a
ANAC070/BRN2			Regulation of cell wall modification in root cap	Bennett *et al.*, 2010
ANAC015/BRN1			Regulation of cell wall modification in root cap	Bennett *et al.*, 2010
ANAC033/SMB			Regulation of cell wall modification in root cap	Bennett *et al.*, 2010
ANAC012/SND1			Secondary cell wall biosynthesis in fibres, endothecium, replum	Zhong *et al.*, 2006; Mitsuda *et al.*, 2007; Zhong *et al.*, 2007; Mitsuda & Ohme-Takagi, 2008; Zhong *et al.*, 2008
ANAC066/NST2			Secondary cell wall biosynthesis	Mitsuda *et al.*, 2005
ANAC043/NST1			Secondary cell wall biosynthesis	Mitsuda *et al.*, 2005; Mitsuda *et al.*, 2007; Zhong *et al.*, 2007; Mitsuda & Ohme-Takagi, 2008; Zhong *et al.*, 2008
ANAC008/SOG1	II	Cell wall development, response to DNA damage	Response to DNA damage	Preuss & Britt, 2003; Yoshiyama *et al.*, 2009
ANAC010/SND3			Regulation of secondary cell wall biosynthesis	Zhong *et al.*, 2008
ANAC073/SND2			Regulation of secondary cell wall biosynthesis	Zhong *et al.*, 2008; Hussey *et al.*, 2011

© University of Pretoria

Table S2.4. *(continued)*

Protein	Subfamily	General function	Specific function	References
ANAC013	IIIa/b	Response to stress, development	Response to red light and UV-B	Safrany *et al.*, 2008
ANAC016			Response to chitin	Libault *et al.*, 2007
ANAC040/NTL8			Salt regulation of seed germination	Kim *et al.*, 2008
ANAC053/NTL4			Drought-induced leaf senescence	Lee *et al.*, 2012
ANAC062/NTL6			Defence response, response to cold treatment	Libault *et al.*, 2007; Seo *et al.*, 2010
ANAC078			Regulation of flavonoid biosynthesis	Morishita *et al.*, 2009
ANAC089			Regulation of flower development	Li *et al.*, 2010
ONAC054/RIM1	IIIc		Response to biotic stress	Yoshii *et al.*, 2009
ANAC069/NTM2	IVa		Salt and auxin signalling pathways	Park *et al.*, 2011
ANAC035/LOV1	IVb	Development	Regulation of cold response and flowering time	Yoo *et al.*, 2007
ANAC036			Regulation of leaf cell growth	Kato *et al.*, 2010
ANAC068/NTM1			Cytokinin signalling during cell division	Kim *et al.*, 2006
ANAC009/FEZ	IVd		Regulation of periclinal cell division in root cap	Willemsen *et al.*, 2008
ONAC063	IVd		Response to salt stress	Yokotani *et al.*, 2009
ANAC029/NAP	Va(1)		Leaf senescence	Guo & Gan, 2006
ONAC010	Va(2)		Anther dehiscence	Distelfeld *et al.*, 2012
ANAC081/ATAF2	Vb	Stress response	Repression of *PR* genes	Delessert *et al.*, 2005
ANAC019 ANAC055			Abiotic stress response; regulation of jasmonic acid-induced gene expression	Tran *et al.*, 2004; Bu *et al.*, 2008; Jiang *et al.*, 2009
ANAC002/ATAF1			Drought response	Lu *et al.*, 2007
ANAC083/VNI2	VIa		Negative regulator of xylem vessel development	Yamaguchi *et al.*, 2010b
ANAC104/XND1	VIc		Regulation of secondary cell wall biosynthesis	Zhao *et al.*, 2008

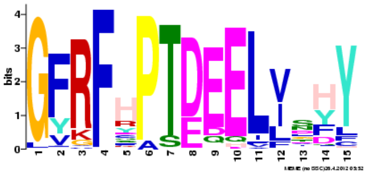
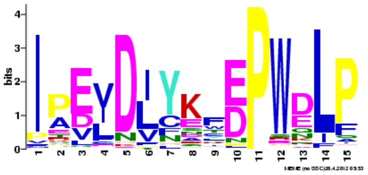
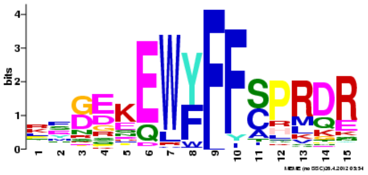
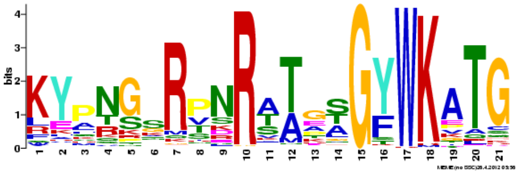
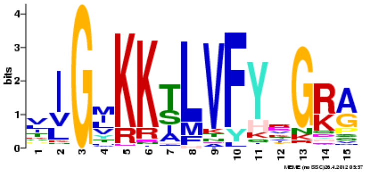
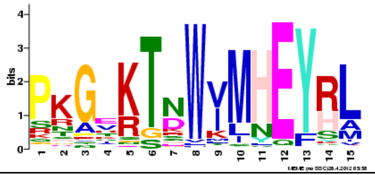
© University of Pretoria

Table S2.5. Putative *E. grandis* homologs of *Arabidopsis* NAC domain proteins known to be involved in regulating secondary cell wall biosynthesis.

ANAC protein	Synonym	Putative *Eucalyptus* (co-)ortholog
ANAC012	SND1	EgrNAC61
ANAC073	SND2	EgrNAC170
ANAC010	SND3	EgrNAC44 EgrNAC45 EgrNAC46 EgrNAC47 EgrNAC64
ANAC043	NST1	EgrNAC49
ANAC066	NST2	-
ANAC083	VNI2	EgrNAC122
ANAC104	XND1	EgrNAC137 EgrNAC138 EgrNAC139 EgrNAC152
ANAC037 ANAC076	VND1 VND2	EgrNAC146
ANAC105	VND3	-
ANAC007 ANAC026	VND4 VND5	EgrNAC50
ANAC101	VND6	EgrNAC26
ANAC030	VND7	EgrNAC75

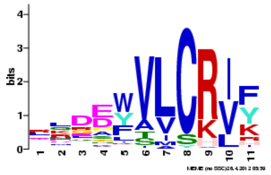
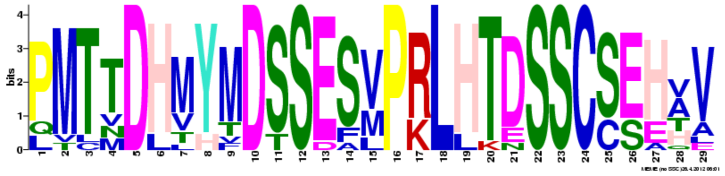
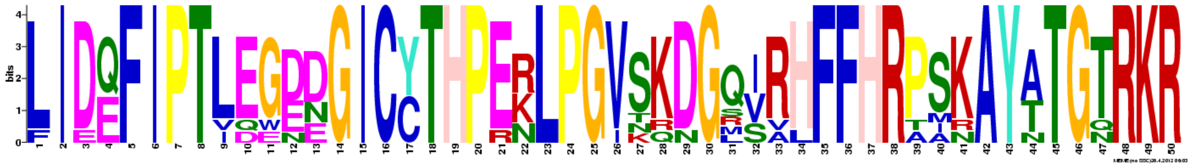
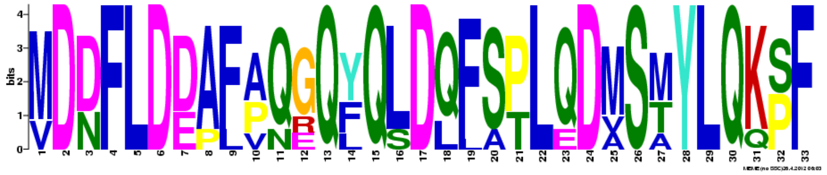
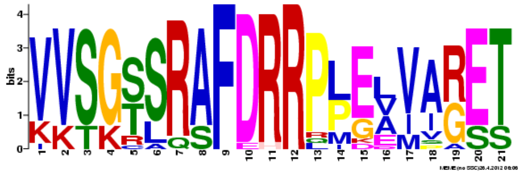
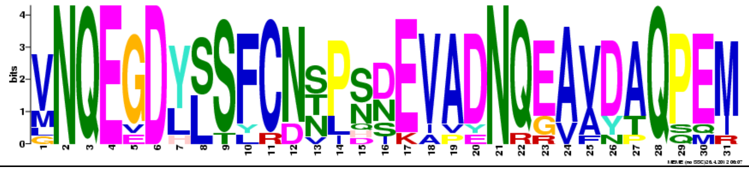
© University of Pretoria

Table S2.6. Amino acid sequence logos of sixteen overrepresented motifs identified in EgrNAC proteins using MEME. The E-value, number of proteins containing each motif and, where applicable, the annotation of each motif is indicated.

Motif name	Number of sites	Length	E-value	HMM annotation
Motif 1	173	15	2.7e-1645	NAC Subdomain A
				
Motif 2	159	15	1.3e-1161	NAC Subdomain B
				
Motif 3	122	15	5.6e-701	NAC Subdomain C
				
Motif 4	162	21	3.8e-1683	NAC Subdomain C
				
Motif 5	172	15	7.4e-1202	NAC Subdomain D
				
Motif 6	182	15	3.9e-1478	NAC Subdomain D
				

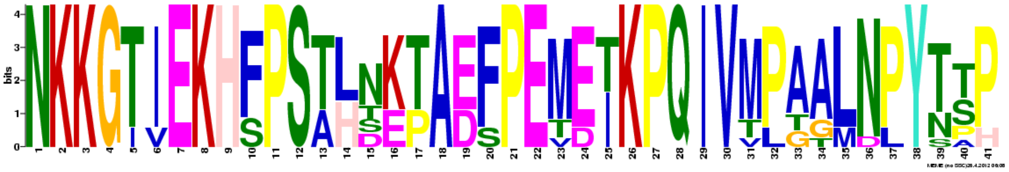
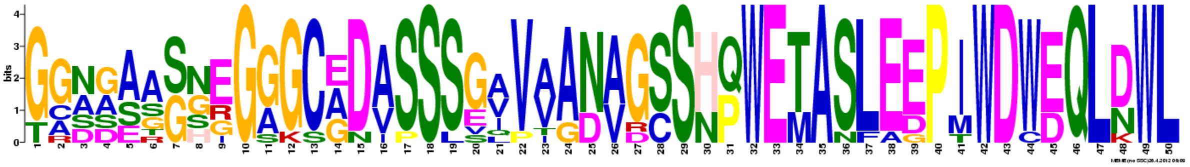
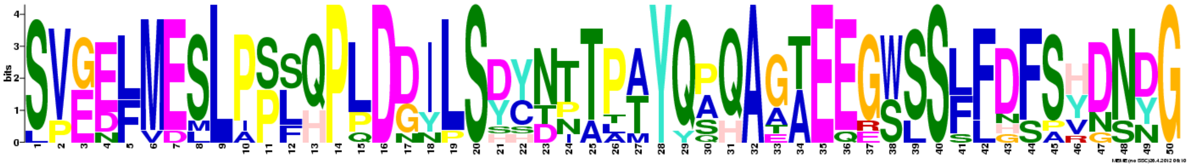
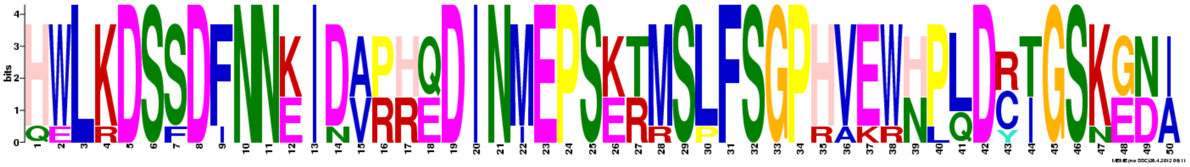
© University of Pretoria

Table S2.6. (*continued*)

Motif name	Number of sites	Length	E-value	HMM annotation
Motif 7	165	11	2.0e-553	NAC Subdomain E
				
Motif 8	15	29	5.20E-260	-
				
Motif 9	9	50	4.50E-271	-
				
Motif 10	12	33	3.80E-245	-
				
Motif 11	20	21	4.20E-193	-
				
Motif 12	11	31	1.80E-162	-
				

© University of Pretoria

Table S2.6. (*continued*)

<table>
<tr><th>Motif name</th><th>Number of sites</th><th>Length</th><th>E-value</th><th>HMM annotation</th></tr>
<tr><td>Motif 13</td><td>12</td><td>41</td><td>3.70E-306</td><td>-</td></tr>
<tr><td colspan="5"> </td></tr>
<tr><td>Motif 14</td><td>11</td><td>50</td><td>2.10E-288</td><td>-</td></tr>
<tr><td colspan="5"> </td></tr>
<tr><td>Motif 15</td><td>10</td><td>50</td><td>5.00E-216</td><td>-</td></tr>
<tr><td colspan="5"> </td></tr>
<tr><td>Motif 16</td><td>7</td><td>50</td><td>9.60E-183</td><td>-</td></tr>
<tr><td colspan="5"> </td></tr>
</table>

© University of Pretoria

Table S2.7. Putative EgrNAC membrane-tethered transcription factors (MTFs) and their corresponding *Arabidopsis* NAC MTF homologs as deduced from Additional file S2.2.

Putative EgrNAC MTF	Homologous *Arabidopsis* MTFs (Kim *et al.*, 2010)
EgrNAC16	ANAC062 ANAC091
EgrNAC39 EgrNAC166	ANAC040 ANAC060 ANAC089
EgrNAC62	ANAC014 ANAC062 ANAC091
EgrNAC76	ANAC016 ANAC017
EgrNAC168	ANAC053 ANAC078
EgrNAC112	ANAC068 ANAC069

© University of Pretoria

Table S2.8. ***EgrNAC* genes occurring in blocks of tandem duplications (Fig. 2.3).** The subfamily classification of each gene (Fig. 2.1) is also indicated.

Gene	Subfamily
Block 1	
EgrNAC1	IVa
EgrNAC2	
EgrNAC3	
EgrNAC4	
EgrNAC5	
EgrNAC6	
EgrNAC7	
EgrNAC8	
EgrNAC9	
EgrNAC10	
EgrNAC11	
EgrNAC12	
Block 2	
EgrNAC13	IVa
EgrNAC14	
Block 3	
EgrNAC19	Va
EgrNAC20	Ia
Block 4	
EgrNAC21	IVc
EgrNAC22	
EgrNAC23	
EgrNAC24	
EgrNAC25	
Block 5	
EgrNAC36	VIIIa
EgrNAC37	
EgrNAC38	
Block 6	
EgrNAC44	II
EgrNAC45	
EgrNAC46	
EgrNAC47	
Block 7	
EgrNAC53	VII
EgrNAC54	
EgrNAC55	
EgrNAC56	
EgrNAC57	
Block 8	
EgrNAC58	VII
EgrNAC59	
EgrNAC60	
Block 9	
EgrNAC63	Va(2)
EgrNAC64	II
Block 10	
EgrNAC65	Va
EgrNAC66	Vb
Block 11	
EgrNAC70	XI
EgrNAC71	
EgrNAC72	
EgrNAC73	
EgrNAC74	
Block 12	
EgrNAC79	VII
EgrNAC80	
Block 13	
EgrNAC84	IVa
EgrNAC85	
EgrNAC86	
EgrNAC87	
EgrNAC88	
EgrNAC89	
EgrNAC90	
EgrNAC91	
EgrNAC92	
EgrNAC93	
EgrNAC94	
EgrNAC95	
EgrNAC96	
EgrNAC97	
EgrNAC98	
EgrNAC99	
EgrNAC100	
EgrNAC101	
EgrNAC102	
EgrNAC103	
EgrNAC104	
Block 14	
EgrNAC106	IVa
EgrNAC107	
EgrNAC108	
EgrNAC109	
EgrNAC110	
EgrNAC111	
Block 15	
EgrNAC117	IIIa/b
EgrNAC118	
Block 16	
EgrNAC124	Vb
EgrNAC125	
EgrNAC126	
EgrNAC127	
EgrNAC129	
Block 17	
EgrNAC131	Vb
EgrNAC132	
EgrNAC133	
EgrNAC134	
EgrNAC135	
EgrNAC136	
Block 18	
EgrNAC137	IVc
EgrNAC138	
EgrNAC139	
Block 19	
EgrNAC141	IVc
EgrNAC142	
Block 20	
EgrNAC147	IVa
EgrNAC148	
EgrNAC149	
EgrNAC150	
EgrNAC151	Unassigned
Block 21	
EgrNAC153	IVc
EgrNAC154	
EgrNAC155	
EgrNAC156	
EgrNAC157	
EgrNAC158	
EgrNAC159	
EgrNAC160	
EgrNAC161	
EgrNAC162	
EgrNAC163	
EgrNAC164	
EgrNAC165	
EgrNAC166	IIIa/b
Block 22	
EgrNAC172	Va(1)
EgrNAC173	Ia
Block 23	
EgrNAC179	VII
EgrNAC180	
EgrNAC181	
EgrNAC182	
EgrNAC183	
EgrNAC184	

© University of Pretoria

2.13. Additional files

The following additional datasets are available in the supplementary CD-ROM disk attached to this thesis:

Additional file S2.1.xlsx: Lists of *Arabidopsis*, *Populus*, *Oryza* and *Vitis* NAC domain proteins used for phylogenetic analysis.

Additional file S2.2.pdf: Dendrogram and subfamily classification of NAC sequences from *Arabidopsis*, *Populus*, *Oryza*, *Vitis* and *Eucalyptus*. The dendrogram is based on the phylogeny shown in Fig. 1b.

Additional file S2.3.xlsx: *EgrNAC* RNA-seq data for developing xylem (DX), flowers (Fl), mature leaf (ML), phloem (Ph), roots (Rt), shoot tips (ST) and young leaves (YL) in three individual ramets. Values are expressed as average number of fragments per kilobase of transcript per million fragments mapped (FPKM).

Additional file S2.4.xlsx: Codon-based Z-test for purifying selection between pairwise comparisons of *EgrNAC* genes in each of twenty-three blocks of tandem duplicates. The P-value for each comparison is shown below the diagonal (*P*-values < 0.05 are indicated in bold); the Z-test statistic is shown above the diagonal.

2.14. Additional notes

The following additional note is available on the supplementary CD-ROM disk attached to this thesis:

Additional note S1.docx: Notes of manually curated and discarded EgrNAC gene candidates. Low confidence annotations refer to those included in the v.1.0 annotation (Phytozome v.7) but excluded from the v.1.1 annotation (Phytozome v.8) of the *E. grandis* genome at www.phytozome.net. Evidence for expression was obtained from Eucspresso (eucspresso.bi.up.ac.za/; Mizrachi *et al.* 2010) and EucGenIE (eucgenie.bi.up.ac.za; Hefer *et al.* in preparation).

© University of Pretoria

CHAPTER 3

SND2, a NAC transcription factor gene, regulates genes involved in secondary cell wall development in *Arabidopsis* fibers and increases fiber cell area in *Eucalyptus*

Steven G Hussey[1], Eshchar Mizrachi[1], Antanas V. Spokevicius[2], Gerd Bossinger[2], Dave K Berger[3], Alexander A Myburg[1]

[1]Department of Genetics, [3]Department of Plant Science, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, 0002, South Africa

[2]Department of Forest and Ecosystem Science, The University of Melbourne, Melbourne, 3363, Australia

This chapter has been written in manuscript format for submission to a peer-reviewed scientific journal. I performed most of the experimental work, performed all the data analysis and drafted the manuscript. Martin Ranik cloned the *SND2* gene and Gerd Bossinger and Antanas Spokevicius performed the ISSA experiment and drafted section 3.3.6. Eshchar Mizrachi, Dave Berger and Alexander Myburg reviewed and edited the manuscript, which was published in 2011 (Hussey *et al.*, *BMC Plant Biology* **11**:173). Some references have been updated since publication.

© University of Pretoria

3.1. Summary

NAC domain transcription factors initiate secondary cell wall biosynthesis in *Arabidopsis* fibers and vessels by activating numerous transcriptional regulators and biosynthetic genes. NAC family member *SND2* is an indirect target of a principal regulator of fiber secondary cell wall formation, SND1. A previous study showed that overexpression of *SND2* produced a fiber cell-specific increase in secondary cell wall thickness in *Arabidopsis* stems, and that the protein was able to transactivate the *cellulose synthase8* (*CesA8*) promoter. However, the full repertoire of genes regulated by *SND2* is unknown, and the effect of its overexpression on cell wall chemistry remains unexplored. We overexpressed *SND2* in *Arabidopsis* and analyzed homozygous lines with regards to stem chemistry, biomass and fiber secondary cell wall thickness. A line showing upregulation of *CesA8* was selected for transcriptome-wide gene expression profiling. We found evidence for upregulation of biosynthetic genes associated with cellulose, xylan, mannan and lignin polymerization in this line, in agreement with significant co-expression of these genes with native *SND2* transcripts according to public microarray repositories. Only minor alterations in cell wall chemistry were detected. Transcription factor *MYB103*, in addition to *SND1*, was upregulated in *SND2*-overexpressing plants, and we detected upregulation of genes encoding components of a signal transduction machinery recently proposed to initiate secondary cell wall formation. Several homozygous T4 and hemizygous T1 transgenic lines with pronounced *SND2* overexpression levels revealed a negative impact on fiber wall deposition, which may be indirectly attributable to excessive overexpression rather than co-suppression. Conversely, overexpression of *SND2* in *Eucalyptus* stems led to increased fiber cross-sectional cell area. This study supports a function for *SND2* in the regulation of cellulose and hemicellulose biosynthetic genes in addition of those involved in lignin polymerization and signalling. SND2 seems to occupy a subordinate but central

© University of Pretoria

tier in the secondary cell wall transcriptional network. Our results reveal phenotypic differences in the effect of *SND2* overexpression between woody and herbaceous stems and emphasize the importance of expression thresholds in transcription factor studies.

© University of Pretoria

3.2. Introduction

Plant fibers constitute a valuable renewable resource for pulp, paper and bioenergy production (Hinchee *et al.*, 2010). In angiosperms, the two principle sclerenchyma cell types that comprise secondary xylem are xylem vessels, which facilitate the transport of water, and xylary fibers, which provide mechanical strength and which make up the bulk of woody biomass (Plomion *et al.*, 2001). Wood density and chemical composition, fiber and vessel length, diameter and wall thickness, and even the proportion of axial and radial parenchyma heavily influence pulp yield, digestibility and quality, although the relative importance of each varies from species to species (Ona *et al.*, 2001; Ramirez *et al.*, 2009).

During xylogenesis in angiosperms, fibers differentiate from the vascular cambium, elongate, and deposit a lignified secondary cell wall (SCW). SCW formation is associated with a distinct form of programmed cell death (Turner *et al.*, 2007; Courtois-Moreau *et al.*, 2009). Much research has been devoted to the biosynthesis of SCW biopolymers, namely (in decreasing order of abundance) cellulose (Guerriero *et al.*, 2010; Endler & Persson, 2011), hemicellulose (Scheller & Ulvskov, 2010) and lignin (Humphreys & Chapple, 2002; Vanholme *et al.*, 2008). Complementing this, in recent years much of the transcriptional network underlying SCW biosynthesis has been deciphered, mainly exploiting *Arabidopsis thaliana* and the *Zinnia elegans* mesophyll-to-tracheary element *in vitro* transdifferentiation system (Fukuda & Komamine, 1980; Zhong *et al.*, 2010a). Genes involved in secondary xylem formation are regulated principally at the transcriptional level, accentuating the central significance of the SCW transcriptional network (Du & Groover, 2010). Manipulation of transcription factors (TFs) associated with the network presents the potential to enhance fiber properties through altering the regulation of a large number of biosynthetic genes.

© University of Pretoria

Kubo *et al.* (2005) first identified NAC domain TFs VASCULAR-RELATED NAC-DOMAIN7 (VND7) and VND6 as "master activators" of SCW formation in proto- and metaxylem vessels, respectively. It was later shown that VND6 and VND7 are functionally redundant, being sufficient for all vessel SCW formation (Yamaguchi *et al.*, 2008; Yamaguchi *et al.*, 2010). In xylem fibers, a similar transcriptional master switch was identified. NAC family proteins SECONDARY WALL-ASSOCIATED NAC DOMAIN1 (SND1) and NAC SECONDARY WALL THICKENING PROMOTING FACTOR1 (NST1) redundantly activate *Arabidopsis* fiber (and, to some extent, silique valve) SCW formation (Zhong *et al.*, 2006; Mitsuda *et al.*, 2007; Zhong *et al.*, 2007b; Mitsuda & Ohme-Takagi, 2008; Zhong *et al.*, 2008). In other cell types with secondary walls, such as the endothecium of anthers, NST1 was also found to activate SCW development, in this case redundantly with NST2 (Mitsuda *et al.*, 2005). Together, these studies support a role for NAC TFs as principal activators of SCW formation in fibers and vessels, acting in distinct combinations in each case.

Several studies suggest that SND1, NST1, VND6 and VND7 activate a conserved, cascading transcriptional network featuring, but by no means limited to, various NAC, MYB and homeodomain TFs (reviewed in Umezawa, 2009; Zhong *et al.*, 2010a). SND1, NST1, NST2, VND6 and VND7 regulate an overlapping set of targets (Zhong *et al.*, 2008; Ohashi-Ito *et al.*, 2010), supported by the ability of NST2, VND6 and VND7 to complement the *snd1 nst1* double mutant when ectopically expressed in fiber cells (Zhong *et al.*, 2010c; Yamaguchi *et al.*, 2011). For this reason, they have been collectively referred to as secondary wall NACs (SWNs) (Zhong *et al.*, 2010c). Amongst the downstream targets of SWNs, *SND3* and *MYB103* are directly activated by SND1/NST1 and VND6/VND7 (Zhong *et al.*, 2008; Ohashi-Ito *et al.*, 2010; Zhong *et al.*, 2010c;

© University of Pretoria

Yamaguchi *et al.*, 2011), although *SND3* has not consistently been detected as a VND6/VND7 direct target. *SND2* is indirectly regulated by SND1/NST1 (Ko *et al.*, 2007; Zhong *et al.*, 2008), but there exists no evidence for regulation by VND6/VND7. Loss- and gain-of-function mutagenesis of *SND2*, but interestingly also that of *SND3* and *MYB103*, produced a fiber-specific phenotype (Zhong *et al.*, 2008). While dominant repression (Hiratsu *et al.*, 2004) drastically reduced fiber-specific SCW thickness, individual overexpression of *MYB103*, *SND2* and *SND3* increased SCW thickness in interfascicular and xylary fibers, with no apparent impact on vessels. In stems a reduction in glucose, xylose and mannose cell wall sugars occurred during dominant repression of *MYB103*, *SND2* or *SND3*. Conversely, all three TFs could transactivate the SCW cellulose-associated *CesA8* gene promoter, but not representatives of hemicellulose (*IRX9*) or lignin (*4CL1*) biosynthesis (Zhong *et al.*, 2008).

The regulation and function of *SND2* may differ in herbaceous and woody plants, especially in woody tissues which possess greater proportions of fiber cells than stems of herbaceous plants. This may be facilitated by gene family expansion and specialization in woody plants (Tuskan *et al.*, 2006). As many as four putative *SND2* orthologs exist in poplar due to significant expansion of the NAC family (Grant *et al.*, 2010), some paralogs of which may have undergone subfunctionalization in *Populus* (Hu *et al.*, 2010). All four putative orthologs were found to be preferentially expressed in developing xylem and phloem fibers (Grant *et al.*, 2010). Overexpression of one of the putative orthologs, *PopNAC154*, resulted in a decrease in height and an increase in the proportion of bark to xylem in poplar trees, with no perceptible effect on SCW thickness (Grant *et al.*, 2010). This apparent conflict with the *SND2* overexpression phenotype in *Arabidopsis* (Zhong *et*

© University of Pretoria

al., 2008) illustrates that the regulatory function of SND2 homologs may differ between herbaceous and woody plants.

The observation that *SND2* overexpression led to enhanced SCW formation in *Arabidopsis* fibers and that it potentially regulates cellulosic genes are important findings, because evidence supports the existence of a similar transcriptional network regulating fiber SCW development in angiosperm trees (McCarthy *et al.*, 2010; Zhong *et al.*, 2010a; Zhong *et al.*, 2010b). However, several aspects of the biological function of *SND2* remain to be resolved before the biotechnological potential of the gene can be determined. The global targets of SND2 have not been identified and its position in the transcriptional network has not been established. The finding that SND2 regulates cellulose, but not xylan and lignin biosynthetic genes, was based on a single representative gene from each pathway (Zhong *et al.*, 2008). A greater knowledge of SND2 targets is required to confidently negate its regulation of hemicellulose and lignin biosynthesis. It is also unclear from the analysis by Zhong *et al.* (2008) whether *SND2* overexpression invariably leads to increased fiber SCW thickness, both in *Arabidopsis* and in woody taxa. Finally, the effect of *SND2* overexpression on cell wall chemistry has not yet been reported.

We aimed to further characterize the position and regulatory role of *SND2* in the fiber SCW transcriptional network, and confirm the phenotypic effects of *SND2* overexpression in *Arabidopsis* and *Eucalyptus* plants. Our objectives were to identify genes that are differentially expressed in *SND2*-overexpressing plants, and determine the overall effect on *Arabidopsis* development and biomass production, as well as fiber SCW formation in *Arabidopsis* and *Eucalyptus*. We describe novel regulatory roles for SND2 in

© University of Pretoria

fiber SCW development, and propose a model for the role of SND2 in the transcriptional network regulating SCW formation.

3.3. Materials and methods

3.3.1. Plant growth conditions

Arabidopsis thaliana Columbia (Col-0) plants were grown in peat moss bags (Jiffy Products International AS, Norway) under a 16h day artificial light regime, at ~22^{o}C and ~75% humidity with weekly fertilization. Where applicable, hygromycin selection was performed for ~14 days before transferral of seedlings to peat moss bags. The stated age of the plants is inclusive of the hygromycin selection period.

3.3.2. Generation of overexpression constructs and transformation

The coding sequence of *SND2* (AT4G28500) was amplified (forward primer, 5'-ATGACTTGGTGCAATGACCGTAG-3', reverse primer 5'-TTAAGGGATAAAAGGTTGAGAGTCAT-3') from *Arabidopsis thaliana* Col-0 inflorescence stem cDNA. The amplicon was gel-purified with the MinElute Gel Extraction Kit (Qiagen, Valencia, CA) and cloned into pCR8/GW/TOPO as per the manufacturer's instructions (Invitrogen, Carlsbad, CA). The sequenced insert was transferred to pMDC32 and pCAMBIA1305.1 (Curtis & Grossniklaus, 2003) using the Gateway LR Clonase^{TM} II Enzyme Mix (Invitrogen). The construct was introduced into *Agrobacterium tumefaciens* strains LBA4404 and AGL1 for pMDC32 and pCAMBIA1305.1 constructs, respectively, followed by *Agrobacterium*-mediated transformation of *Arabidopsis thaliana* Col-0. After surface sterilization, transgenic seed was selected on 0.8% agar containing 20 µg/ml Hygromycin B. The seeds were artificially stratified at 4^{o}C for 2-4 days prior to germination at 22^{o}C under artificial illumination.

© University of Pretoria

Homozygosity was assessed for the T3 generation based on a χ^2 test of hygromycin resistance in T4 seedlings from each plant grown on selective media (20 µg/ml Hygromycin B).

3.3.3. Microarray analysis

For the eight-week experiment, T4 seedlings were selected on hygromycin for two weeks and grown in peat moss bags for six weeks. For the four week experiment, no selection was employed; homozygous T4 seeds were germinated directly on peat moss. Each of three biological replicates consisted of ten or six plants in the four and eight week experiments, respectively. Stem tissues were collected on the same day between 08:30 and 11:00, flash-frozen in liquid nitrogen and stored at -80^{o}C. Total RNA extracted from the bottom 100 mm of the primary inflorescence stems was treated with the RNase-free DNase Set (Qiagen) and genomic DNA contamination assessed by PCR using intron-spanning primers. RNA integrity was quantified using the Experion^{TM} instrument (Bio-Rad Laboratories, Inc.). cDNA synthesis and cyanine dye coupling were performed as prescribed by the African Centre for Gene Technologies (ACGT) Microarray Facility (available at http://www.microarray.up.ac.za/MA008_indirect_labelling_version3.pdf).

Microarray hybridization was performed using the *Arabidopsis thaliana* 4x44k DNA microarray V3 (Agilent Technologies, Santa Clara, CA), as described by the manufacturer’s instructions, but substituting cRNA with cDNA. Dye-swaps were employed to correct for fluorophore bias. Slides were scanned using an Axon GenePix 4000B instrument (Axon Instruments, Foster City, CA, USA). Features were extracted using Axon GenePix Pro software (v6.0) and imported into limma (linear models for microarray data) (Wettenhall & Smyth, 2004). Data were normalized in R as described by

© University of Pretoria

Crampton *et al.* (2009), with linear models based on the comparison between SND2-OV(A) and the wild type, analyzing each time point independently. Significant DEGs were defined as those with *P**-value < 0.05, where *P** is the False Discovery Rate. Raw data files of all the microarray experiments are available from the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/projects/geo/), under accession number GSE29693.

Differentially expressed genes were subjected to an anatomical meta-analysis of expression in selected *Arabidopsis* tissues by hierarchical clustering (Pearson correlation) in the Genevestigator V3 public microarray database (Hruz *et al.*, 2008). Only high quality ATH1 22k arrays, and probe sets highly specific for a single gene, were selected for analysis.

3.3.3. Reverse Transcription Quantitative Polymerase Chain Reaction (RT-qPCR) analysis

The quality of total RNA extracted from lower inflorescence stems was assessed by Experion^{TM} analysis (Bio-Rad Laboratories, Hercules, CA). First-strand cDNA synthesis from genomic DNA-free RNA was performed using the Improm-II^{TM} Reverse Transcriptase cDNA synthesis kit (Promega, Madison, WI) and cDNA purified using the RNeasy Mini Kit (Qiagen). RT-qPCR reactions were quantified using the LightCycler 480 system [45 cycles of 95^{o}C denaturation (10s), 60^{o}C annealing (10s) and 72^{o}C extension (15s)] (Roche GmbH, Basel, Switzerland). Primer sequences that were used for each gene target are listed in Table S3.3. LightCycler 480 Software v. 1.5.0. (Roche) was used for second derivative maximum value calculation and melting curve analysis. Statistical analysis was performed with Biogazelle qBasePLUS (Hellemans *et al.*, 2007).

© University of Pretoria

3.3.4. Microscopy

For light microscopy, the lower ~5 mm of the primary inflorescence stem was fixed in formaldehyde/glutaraldehyde buffer (3.5% and 0.5% v/v, respectively) for up to five days and dehydrated in an ethanol series before embedding in LR White^{TM} resin. Stem sections of 0.5 µm thickness were visualized with Toluidine Blue. Micrograph measurements were performed using ImageJ software (National Institutes of Health, http://rsbweb.nih.gov/ij/), using the polygon tool for cell area measurements. For scanning electron microscopy (SEM), 90 nm thick epoxy-embedded samples were imaged following sodium methoxide etching for 1 min (Mayor *et al.*, 1961) using a LEO 1455 VP-SEM instrument (Carl Zeiss, Germany) at 5 kV.

3.3.5. Klason lignin and cell wall sugar analysis

Complete inflorescence stems from eight-week-old transgenic and wild type plants were stripped of siliques and cauline leaves and dried (100^{o}C, 24 h). Stems from up to 24 plants were pooled for each of three biological replicates. Cell wall sugar and Klason lignin analysis were performed essentially as described by Coleman *et al* (2008), using High Performance Liquid Chromatography (Dionex CarboPac PA1 4 x 250 mm) to determine carbohydrate concentrations. Triplicate technical repetitions were performed.

3.3.6. Induced Somatic Sector Analysis (ISSA)

ISSA was performed as described before (Spokevicius *et al.*, 2007) with some modifications. Eleven ramets of each of two hybrid clones, *E. grandis* x *E. camaldulensis* and *E. camaldulensis* x *E. globulus*, were selected in early summer on the basis of good form and growth for experimentation and ten 1 cm^{2} cambial windows were created on each plant. *Agrobacterium tumefaciens* AGL1 harbouring pCAMBIA1305.1 containing

© University of Pretoria

the *Arabidopsis SND2* CDS and the *β*-glucuronidase or ‘GUS’ reporter gene was injected into the cambial windows. Plants were fertilised after inoculation and maintained in the glasshouse in the same condition as described previously (Spokevicius *et al.*, 2007) until harvest. After 195-210 days cambial windows were excised from the main stem, the phloem portion was removed and the remaining xylem tissue was washed twice with 0.1 M $NaPO_4$ buffer (pH 7). Transgenic sectors were identified by GUS reporter staining. Eleven *SND2*-overexpressing and nine empty vector control sectors were analyzed. Transgenic sectors were excised in blocks of 1-3 mm^3 (from the cambium to wound parenchyma) using a single edge razor blade, so that the sector was located close to the middle of the block when viewed on the longitudinal tangential plane. Blocks were then sliced transversely through the middle of the sector to expose the transverse surface of the sector, and then mounted with conductive adhesive on SEM stubs. Transgenic sectors were delineated within the block by etching the borders of the GUS reporter stain with a razor blade. Blocks were desiccated overnight prior to SEM imaging. Cell morphology measurements were undertaken using the Quanta Environmental Scanning Electron Microscope (FEI, Hillsboro, Oregon) to investigate changes in cell wall thickness, cell wall area (total amount of cell wall), cell area and lumen area. Images were taken of both transgenic sector and directly adjacent non-transgenic tissue, twenty to fifty cells from the cambial surface, using the low vacuum mode. Images were then analysed using freeware Image-J (http://rsbweb.nih.gov/ij/) with ten fibers measured per micrograph. For the cell wall thickness, the mean of three measurements for each cell wall were used for cell wall thickness calculations, whilst for the remaining properties one value for each fiber was sufficient. Average values were calculated for each sector and their non-transgenic control tissues and converted into percentage change values. Percentage change values between

© University of Pretoria

SND2 overexpression sectors and empty vector control (EVC) were statistically assessed with the Student's t-test.

3.4. Results

3.4.1. Whole-transcriptome expression profiling of *SND2*-overexpressing *Arabidopsis* plants

SND2 was previously shown to transactivate the *CesA8* gene promoter in *Arabidopsis* protoplasts (Zhong *et al.*, 2008). In order to identify other genes regulated by SND2 *in planta*, we overexpressed *SND2* in *Arabidopsis* plants by cloning the *SND2* coding sequence into the overexpression vector pMDC32 (Curtis & Grossniklaus, 2003). We introduced the construct into *A. thaliana* Col-0 plants and randomly selected three homozygous T4 lines (A, B, and C), from a pool of T1 transgenic plants herein denoted "SND2-OV". The presence of the transgene in each line was assessed by PCR (Fig. S3.1), and the results of a χ^2 test for homozygosity based on hygromycin resistance is shown in Table S3.1. We confirmed that *SND2* was strongly upregulated in the T4 SND2-OV lines using RT-qPCR analysis (Fig. S3.2). We then tested the T4 SND2-OV lines for preliminary evidence of *CesA8* upregulation in lower inflorescence stems using RT-qPCR analysis. Interestingly, line A ("SND2-OV(A)") exclusively showed evidence for *CesA8* upregulation (not shown), and was therefore selected for transcriptome analysis.

In order to determine which genes were differentially expressed as a result of *SND2* overexpression in *Arabidopsis* stems, the transcriptome of SND2-OV(A) plants was compared to that of the wild type with respect to the bottom 100 mm of primary inflorescence stems. High quality total RNA (RQI > 9.3) was isolated from three biological replicates of eight-week-old wild type and SND2-OV(A) plant stems, and

© University of Pretoria

labelled cDNA hybridized to Agilent 4x44k *Arabidopsis* transcriptome arrays. Significantly differentially expressed genes (DEG) were identified as those with an experiment-wise false discovery rate below 0.05 and fold change > |±1.5|. This analysis identified a total of 155 upregulated and 68 downregulated genes in SND2-OV(A) relative to the wild type (Additional file 3.1).

In order to identify overrepresented gene ontology (GO) classes amongst the DEGs, the GOToolBox resource (Martin *et al.*, 2004) was interrogated with a hypergeometric test (Benjamini and Hochberg correction) using The *Arabidopsis* Information Resource (Rhee *et al.*, 2003) annotation set. Significantly enriched biological processes ($P < 0.01$) revealed a predominant role of the DEGs in (secondary) cell wall organization and biogenesis, carbohydrate metabolism, signalling and response to stimulus (Table S3.2).

3.4.2. Identification of putative SND2 targets

SND2 is preferentially expressed in xylem (Zhao *et al.*, 2005; Zhong *et al.*, 2008). We hypothesized that targets of SND2 would be co-expressed with endogenous *SND2* transcripts. The tissue-specific expression of DEGs identified in SND2-OV(A) (fold change > |±1.5|) was explored by observing the expression patterns across selected *Arabidopsis* tissues using the Genevestigator V3 (Hruz *et al.*, 2008) anatomy clustering tool. At the time of analysis, the Genevestigator database totalled 374 publicly available microarray studies for *Arabidopsis*, encompassing 6,290 samples. Of 223 genes in our SND2-OV(A) dataset, 190 were represented by unique probe sets on high quality ATH1 22k arrays. We examined the endogenous expression of these genes across 26 tissues based on results from 4,422 arrays, and subjected the genes to hierarchical clustering according to their absolute expression profiles. The majority of genes did not conform to a

© University of Pretoria

single expression pattern, with only ~9% of the genes displaying expression profiles clearly resembling that of native *SND2* transcript, i.e. with preferential expression in SCW-containing tissues (Fig. S3.3). Thus, the majority of genes differentially expressed as a result of *SND2* overexpression were not generally associated with SCW-containing tissues.

Novel targets arising from ectopic overexpression of cell wall-associated NAC TFs have been reported previously (Bennett *et al.*, 2010). It is possible that a similar phenomenon occurred in our study, since the bulk of the sampled transgenic stems comprised tissues where *SND2* is not naturally expressed. This may explain the small proportion of DEGs that were co-expressed with *SND2* in Fig. S3.3. To avoid this possibility, we stringently defined the putative authentic targets of *SND2* as those that were also a subset of SND1-regulated genes, the latter identified by microarray analysis of *SND1*-overexpressing *Arabidopsis* plants by Ko *et al.* (2007). The age of the plants in the cited study (~8.5 weeks) and the tissue sampled (lower 50 mm of the inflorescence stem) was similar to our experiment. *SND2*, a known indirect target of SND1 (Zhong *et al.*, 2008), was the most strongly upregulated TF in the SND1-overexpressing plant stems (Ko *et al.*, 2007), further justifying our approach.

We extracted genes common to the Ko *et al.* (2007) data and our significant SND2-OV(A) microarray data, without fold-change filtering. Seventy five genes were shared between the two datasets, herein denoted “SND2∩Ko”, ~79% of which were regulated in a consistent direction (Table 3.1). Amongst them, genes involved in transcription, (secondary) cell wall biosynthesis, cell wall expansion and modification, carbohydrate metabolism, stress response and proteins of unknown function were prominent (Table 3.1). There was notably no differential expression of monolignol biosynthetic genes.

© University of Pretoria

We independently assessed the possible function of *SND2* by identifying genes co-expressed with native *SND2* transcript from the AtGenExpress Plus Extended Tissue Set public microarray data using Expression Angler (Toufighi *et al.*, 2005), employing a stringent Pearson correlation coefficient threshold (R > 0.90). Genes associated with SCW biosynthesis (e.g. secondary wall *CesAs*, *IRX* genes) as well as TFs previously implicated in SCW regulation (*MYB103*, *SND1*), were amongst the 31 genes found to be co-expressed with *SND2* (Table 3.2), supporting a role of *SND2* in SCW regulation. 22 of the genes were differentially expressed in the SND2∩Ko data (Table 3.2).

The seventy five SND2∩Ko genes represented on the ATH1 22k array were subjected to hierarchical clustering across the Genevestigator V3 *Arabidopsis* anatomy database (Hruz *et al.*, 2008) as before to analyze their tissue specificity. Unique probe sets were found for all but one gene (AT5G24780). One cluster (a) contained 31 genes preferentially expressed in a similar fashion to *SND2*, namely in inflorescence nodes and stem, rosette stem and xylem, and silique (Fig. 3.1). Another cluster of 13 genes (b) appeared to exhibit preferential expression in inflorescence stems and nodes, rosette stems, and occasionally seedling hypocotyls, root steles and anther-containing stamens, all of which contain SCWs to some degree. Thus, compared to the original SND2-OV(A) dataset, a much higher percentage (59%) of genes in the SND2∩Ko dataset displayed preferential expression in tissues containing SCWs. Combined with the AtGenExpress co-expression analysis, these data support the role of *SND2* in SCW regulation and the validity of the SND2∩Ko dataset as the most likely direct or indirect targets of SND2.

The microarray results were validated by RT-qPCR analysis. We profiled fifteen genes based on the microarray RNA isolated from stems of eight-week-old SND2-OV(A)

© University of Pretoria

and wild type plants (Fig. 3.2). All RT-qPCR profiles agreed with the microarray data, and seven genes were significantly (P < 0.05) upregulated (including *CesA4*, *EXPA15*, *FLA12* and *MYB103*). We also confirmed that the endogenous *SND2* transcript showed no significant change in SND2-OV(A) stems, whereas total *SND2* transcript abundance (the sum of transgenic and endogenous transcripts) in SND2-OV(A) stems was ~180-fold that of the wild type (not shown). We obtained similar results for selected genes from plants grown in an independent trial (Fig. S3.4).

We were interested in the temporal effect of inflorescence stem development on the putative targets of SND2 when constitutively expressed. We therefore performed a second microarray analysis of SND2-OV(A) and wild type plants at four weeks of age, sampling inflorescence stems that were ~120 mm tall. Of the 21 upregulated and 24 downregulated DEGs, no SND2∩Ko candidates were present, nor were any SCW biosynthesis-associated genes (Additional file 3.2). This result suggests that an additional co-regulator(s), only expressed after four weeks, is required for SND2 to function in fiber SCW regulation.

3.4.3. Effect of *SND2* overexpression on *Arabidopsis* SCW thickness, biomass and SCW composition

Zhong *et al*. (2008) previously reported that *SND2* overexpression significantly increased SCW thickness in interfascicular fibers (IFs) of *Arabidopsis* inflorescence stems. However, among our homozygous SND2-OV lines, scanning electron microscopy (SEM) revealed no significant changes in fiber wall thickness for lines A and B, whilst line C had significantly thinner SCWs than the wild type (Fig. 3.3). These results were reproduced in an independent trial using light microscopy (Fig. S3.5).

© University of Pretoria

Fiber SCW thickness was additionally assessed in lower inflorescence stems of seven T1 SND2-OV and eight wild type plants using light microscopy. Representative micrographs are shown in Fig. S3.6. The T1 lines manifested a significant (21%, $P < 0.02$) decrease in mean IF SCW thickness (Fig. 3.4a) that resembled SND2-OV line C and the *SND2* dominant repression phenotype reported previously (Zhong *et al.*, 2008). Combined endogenous and transgenic *SND2* transcript abundance from T1 plants exceeded that of the wild type plants by ~435-fold, ruling out co-suppression as an explanation for the phenotype (Fig. 3.4b). Although no significant correlation could be found between *SND2* transcript abundance and SCW thickness, our data confirm that strong *SND2* overexpression reduces IF SCW thickness.

We hypothesized that *SND2* overexpression could influence overall inflorescence stem biomass, irrespective of IF SCW thickness. The entire inflorescence stems of eight-week-old T4 SND2-OV lines A, B and C were weighed to determine total biomass yield. Only in the most highly overexpressing line, SND2-OV line C, was biomass significantly different from the wild type, where fresh and dry biomass was decreased (Fig. S3.7). This was despite the fact that all SND2-OV lines appeared phenotypically normal and exhibited no stunting or dwarfing (results not shown). Biomass profiles in Fig. S3.7 were in agreement with the IF SCW thickness profile for each respective line (Fig. 3.3), suggesting a direct relationship between IF SCW thickness and biomass yield, and therefore a negative effect of excessive *SND2* overexpression on biomass yield.

The chemical composition of the inflorescence stems was investigated by Klason lignin analysis and quantification of monosaccharides following complete acid hydrolysis. SND2-OV(A) exhibited a nominal but statistically significant 2.5% relative decrease in

© University of Pretoria

total lignin (Table 3.3, *P* = 0.03). This was likely due to a reduction in insoluble lignin (Table 3.3). No changes were apparent in the relative abundance of glucose and xylose, and only mannose and rhamnose were significantly increased in line A (*P* < 0.05) by 7.4% and 5.4% respectively (Table 3.4). We also quantified the chemical composition of SND2-OV line C to investigate SCW composition when fiber wall thickness was reduced. However, no change in lignin or monosaccharide content was detected against the wild type (not shown).

3.4.4. Induced somatic overexpression of *SND2* in *Eucalyptus* stem sectors

Compared to herbaceous annuals such as *Arabidopsis*, woody perennials devote a larger proportion of carbon allocation to xylem formation. We therefore examined the effect of *Arabidopsis SND2* overexpression on xylem fiber characteristics in *Eucalyptus* trees by Induced Somatic Sector Analysis (Spokevicius *et al.*, 2007). Stems were transformed with a pCAMBIA1305.1 construct (containing the *β*-glucuronidase or 'GUS' reporter gene) overexpressing *SND2*. Tree stems were harvested after 195-210 days, transgenic sectors were identified in the cross-sections via GUS reporter staining, and etched to delineate the transgenic sectors prior to SEM analysis (Fig. S3.8).

Fiber dimensions were measured from SEM micrographs as a percentage change between eleven transgenic sectors and adjacent wild type sectors for the *SND2*-overexpressing gene construct, as well as nine empty vector control (EVC) sectors expressing only the GUS reporter. Fiber cell area (i.e. average fiber cross-section area) was significantly increased in SND2-OV sectors compared to EVC sectors (Table 3.5, *P* = 0.042), demonstrating that *SND2* influences fiber development in *Eucalyptus*. Fiber cell wall area and lumen area, which comprise fiber cell area, were marginally increased in

© University of Pretoria

SND2-overexpressing sectors relative to EVC sectors, but the differences were not statistically significant for these individual parameters. However, since the increase in cell wall area in *SND2*-overexpressing sectors was close to significant (*P* = 0.066), it is reasonable to suggest that the increase in fiber cell area was mainly due to a cell wall area increase rather than a lumen area contribution. Measurement of fiber cell area in the *Arabidopsis* T4 and T1 SND2-OV lines revealed no significant differences relative to the wild type (not shown).

3.5. Discussion

A role for *SND2* in regulating *Arabidopsis* fiber SCW formation was previously suggested by studies establishing it as an indirect target of SND1, a master regulator of fiber SCW development (Zhong *et al.*, 2006; Ko *et al.*, 2007; Mitsuda *et al.*, 2007; Zhong *et al.*, 2007b; Zhong *et al.*, 2008). In promoter transactivation experiments, *SND2* was implicated in the regulation of cellulose (*CesA8*), but a role in regulating hemicellulose or lignin biosynthesis seemed unlikely (Zhong *et al.*, 2008). A particularly interesting finding was a fiber cell-specific increase in SCW thickness when *SND2* was constitutively overexpressed, mirrored by decreased fiber SCW thickness in dominant repression lines (Zhong *et al.*, 2008). The proposed role of SND2 in *Arabidopsis* fiber SCW formation has not been independently validated and the full suite of genes regulated by SND2 has not been elucidated. To address this, we performed microarray analysis on a homozygous *SND2* overexpressing line, SND2-OV(A), which also exhibited significant upregulation of the *CesA8* gene.

TFs have been shown to activate novel targets when ectopically expressed. A striking example was described by Bennett *et al.* (2010) for NAC TFs regulating primary

© University of Pretoria

cell wall modification in the root cap. Overexpression in stems caused ectopic lignification and ectopic expression of SCW genes (Bennett *et al.*, 2010). Our microarray results therefore likely include direct and indirect targets of SND2, as well as genes misregulated due to the ectopic overexpression of *SND2*. To discriminate native targets of SND2, we defined a subset of genes (SND2∩Ko) regulated by SND1 (Ko *et al.*, 2007) that were also found to be differentially expressed in this study (Table 3.1). We reasoned that obtaining the SND1 subset of targets would be a robust approach for reducing ectopic noise, because *SND2* is indirectly, but strongly activated by SND1 (Ko *et al.*, 2007; Zhong *et al.*, 2008) and native SND2 targets should therefore be a subset of the SND1 targets. Further support for defining these seventy five genes as putative SND2 targets was provided by the finding that a large proportion (71%) of genes co-regulated (R > 0.9) with *SND2* in a large compendium of AtGenExpress microarray experiments were included in the SND2∩Ko set (Table 3.2). Recently, Zhong *et al.* (2011) demonstrated transactivation of poplar *CesA4*, *CesA8*, GT43 and GT47 family gene promoters by a poplar co-ortholog of *SND2*, providing a third line of evidence that *SND2* regulates SCW-associated genes.

The SND2∩Ko set (Table 3.1) prominently included genes involved in SCW biosynthesis, transcriptional regulation and signalling. Amongst the SCW-associated genes, *CesA4*, *CesA7* and *CesA8* are involved in SCW cellulose biosynthesis (Taylor *et al.*, 1999; Taylor *et al.*, 2000; Taylor *et al.*, 2003). *COBL4* and its orthologs also appear to be involved in SCW cellulose formation (Brown *et al.*, 2005; Sato *et al.*, 2010), and a homolog of *TRICHOME BIREFRINGENCE*, *TBL3* (AT5G01360), was shown to affect secondary cellulose deposition and possibly SCW structure through alterations in pectin methylesterification (Bischoff *et al.*, 2010). *PARVUS*, *IRX8* and *IRX10* are required for xylan biosynthesis in SCWs (Peña *et al.*, 2007; Persson *et al.*, 2007; Brown *et al.*, 2009;

© University of Pretoria

Wu *et al.*, 2009). *IRX15* and *IRX15L*, encoding functionally redundant DUF579 proteins, were shown to be essential for normal xylan biosynthesis (Brown *et al.*, 2011; Jensen *et al.*, 2011), but only the former was upregulated in SND2-OV(A) stems (Table 3.1). *PGSIP1* and *UXS3* are co-expressed with xylan synthases, with good evidence supporting a xylan α-glucuronosyltransferase function for *PGSIP1* (Oikawa *et al.*, 2010) and a UDP-xylose synthase function for *UXS3* (Harper & Bar-Peled, 2002; Oka & Jigami, 2006). As shown previously (Zhong *et al.*, 2008), *SND2* did not activate the xylan-associated *IRX9* gene in this study, nor did it activate lignin-associated *4CL1*. *LAC4* and *LAC17* encode laccases, an enzyme group that has been linked to SCW lignin polymerization (Mattinen *et al.*, 2008). *LAC4* and *LAC17* are regulated by lignin-specific TFs MYB58 and MYB63 (Zhou *et al.*, 2009) and were also recently shown to affect lignification in *Arabidopsis* xylem, with *LAC17* specifically implicated in G-lignin polymerization in IFs (Berthet *et al.*, 2011). Our results (Table 3.1) thus suggest an additional role for *SND2* in the regulation of lignification distinct from that of monolignol biosynthesis.

Several TFs were upregulated in the SND2∩Ko set (Table 3.1). ANAC019 regulates biotic and abiotic stress responses (Tran *et al.*, 2004; Bu *et al.*, 2008). AT4G17245 is a C3HC4 RING-type zinc finger gene of unknown function. However, at least one C3HC4 gene, AT1G72220, encodes an E3 ubiquitin ligase implicated in SCW formation (Brown *et al.*, 2005; Noda *et al.*, 2013). We observed upregulation of *RAP2.6L*, an ethylene response factor involved in shoot regeneration and abiotic stress response (Che *et al.*, 2006; Krishnaswamy *et al.*, 2011). The upregulation of *SND1* and *MYB103* in SND2(OV) plants was unexpected. SND1, a master activator of SCW biosynthesis in fibers (Zhong *et al.*, 2006; Ko *et al.*, 2007; Mitsuda *et al.*, 2007; Zhong *et al.*, 2007b; Zhong *et al.*, 2008), is expected to be upstream of SND2 in the transcriptional network. It also seems intuitive

© University of Pretoria

that *SND2* acts downstream of *MYB103*, since *SND2* is an indirect target of SND1, whilst *MYB103* is a direct target of *SND1* (Zhong *et al.*, 2008). A positive feedback loop may exist through which upregulation of *SND2*, or another TF (Table 3.1), promotes *SND1* expression.

A signal transduction pathway based on a mammalian signalling model was proposed for SCW biosynthesis in *Arabidopsis* and rice (Figure 4 in Oikawa *et al.*, 2010). Differentially expressed genes in SND2∩Ko included those encoding the principal proteins of this machinery (Table 3.1), namely *FLA11/FLA12*, *CTL2* (AT3G16920), the LRR kinase *PXC1* (AT2G36570), Rho GTPase activating protein (Rac; AT1G08340), IQ (*IQD10*, AT3G15050) and RIC (AT1G27380). *CHITINASE-LIKE 2* (*CTL2*), which lacks chitinase or chitin-binding activity (Hermans *et al.*, 2010), might interact with *FLA11/12* in a similar way to the interaction of mammalian chitinase-like protein SI-CLP with a fasciclin domain-containing transmembrane receptor, Stabilin-1 (Kzhyshkowska *et al.*, 2006; Oikawa *et al.*, 2010). The Rho GTPase activating protein (Rac) has been associated with a (glucurono)xylan biosynthesis functional module with SND2 (Heyndrickx & Vandepoele, 2012), and at least one Rho GTPase activating protein, ROP11, has been implicated in SCW pattern formation (Oda & Fukuda, 2012; Oda & Fukuda, 2013). Mutation of the LRR-RLK *PXC1* resulted in decreased SCW deposition in inflorescence stems during short-day conditions (Wang *et al.*, 2013), supporting a role in this pathway. Two additional kinases (AT1G09440 and AT1G56720, Table 3.1) could possibly be involved in this signalling cascade. Based on these findings, we propose a revised model for the role of SND2 in the transcriptional network underlying fiber SCW deposition (Fig. 3.5). Under this model, SND2 directly or indirectly upregulates the genes associated with

© University of Pretoria

this signalling machinery. The nature of this regulatory relationship remains to be resolved.

Despite the upregulation of the associated biosynthetic genes, we did not observe corresponding relative increases in glucose (i.e. cellulose) or xylose (i.e. xylan) content per unit mass (Table 3.4). There may not be a direct relationship between *CesA* expression and cellulose content, as evidenced when *SND1* is overexpressed (Ko *et al.*, 2007). However, we found that mannose and rhamnose content of stems were significantly increased in SND2-OV(A) by 7.4% and 5.4%, respectively (Table 3.4). The increase in mannose could be explained by the upregulation of *CslA9* (Table 3.1), since CSLA proteins encode β-mannan synthases (Dhugga *et al.*, 2004; Liepman *et al.*, 2007). Rhamnose and mannose were also reported to be increased due to *SND1* overexpression (Ko *et al.*, 2007).

Although we found no effect on fiber SCW thickness in homozygous SND2-OV lines A and B, the fiber SCW thickness of line C was significantly and reproducibly decreased relative to wild type (Fig. 3.3; Fig. S3.5). Because line C exhibited the highest *SND2* transcript abundance amongst the homozygous lines (Fig. S3.2), we confirmed using several T1 *SND2*-overexpressing lines, with *SND2* transcript far exceeding that of SND2-OV(A), that strong *SND2* overexpression reduces fiber SCW thickness (Fig. 3.4). This phenotype resembles the dominant repression phenotype of *SND2*, rather than the overexpression phenotype, reported previously (Zhong *et al.*, 2008). However, due to the stable expression of *SND2* transcript in all transgenic lines (Fig. S3.2; Fig. 3.4), this cannot be explained by co-suppression. Interestingly, a similar paradox has been observed for *SND1* overexpression (Zhong *et al.*, 2006; Ko *et al.*, 2007), where excess levels of this transcriptional activator were reported to have an indirect repressive effect. We suggest

© University of Pretoria

that this phenomenon could be attributed to gene dosage effects, where a stoichiometric increase in one TF protein leads to a decreased molar yield of a multi-protein complex, and greater yield of incomplete intermediates (reviewed by Birchler *et al.*, 2005). Such a phenomenon could also explain the observation that *CesA8* upregulation was restricted to the most moderate *SND2*-overexpressing line, SND2-OV(A). Notably, the transgenic lines in our study expressed *SND2* at least an order of magnitude greater than the ~16-fold expression levels reported for lines with increased fiber wall thickness by Zhong *et al.* (2008). This is likely due to a double, rather than a single, *CaMV 35S* promoter in the pMDC32 vector driving *SND2* overexpression in this study. Because we failed to identify SND2-OV lines with *SND2* abundance near the range of 16-fold, we cannot preclude that limited *SND2* overexpression may increase fiber SCW thickness.

Interestingly, when we overexpressed *Arabidopsis SND2* in *Eucalyptus* xylem (Table 3.5), we observed a phenotype in better agreement with that previously reported for *Arabidopsis* (Zhong *et al.*, 2008). *SND2* overexpression in *Eucalyptus* significantly increased fiber cell area, likely due to increased cell wall area (Table 3.5). Because our assessment of *SND2* overexpression in multiple independent events in both *Arabidopsis* and *Eucalyptus* contrast not only with each other but also with that of Zhong *et al.* (2008), our results suggest that the phenotypic effects of *SND2* gain-of-function mutagenesis are intrinsically variable. The positive effect of *SND2* overexpression on *Eucalyptus* fiber development could be the result of a greater tolerance in *Eucalyptus* to high-abundance SND2 and/or SND2 co-regulator levels in woody xylem, since more carbon is allocated to SCW biosynthesis in *Eucalyptus* than in *Arabidopsis*. Alternatively, *SND2* transcript levels remained moderate in *Eucalyptus*, a possibility that cannot be explored using the Induced Somatic Sector Analysis technique.

© University of Pretoria

In addition to the requirement of the appropriate level of SND2 abundance in *Arabidopsis*, spatial and temporal expression of a co-regulator(s) is a further requirement. The fiber-restricted SCW phenotype of *SND2* overexpression observed in *Arabidopsis* by Zhong *et al.* (2008) illustrates the requirement of a spatially regulated co-regulator(s) for SND2 to activate its targets, which is presumably also expressed in fibers. Our results support this observation. Due to the fact that none of the genes differentially expressed at eight weeks (Table 3.1) were differentially expressed in four week stems (Additional file 3.2), we further suggest that the co-regulator(s) is temporally regulated, and that the temporal regulation of the co-regulator(s) may be a limiting factor that constrains the ability of SND2 to activate its native target genes at four weeks.

3.6. Conclusion

Our results suggest that *SND2* regulates genes involved in cellulose, mannan, and xylan biosynthesis, cell wall modification, and lignin polymerization, but not monolignol biosynthesis. SND2 also promotes upregulation of a relatively small number of TFs, amongst them *MYB103* and *SND1*. We implicate *SND2* in the unexpected regulation of the machinery of a signal transduction pathway proposed for SCW development (Oikawa *et al.*, 2010) and propose a model in which SND2 occupies a subordinate but central position in the transcriptional regulatory network (Fig. 3.5), with possible indirect positive feedback to higher regulators and signalling pathways. Our data support the role of *SND2* in fiber SCW transcriptional regulation, but our study suggests that, at excessive levels of overexpression, *SND2* has a negative effect on IF SCW deposition. This phenomenon requires further investigation. We postulate that *SND2* overexpression could increase SCW deposition within a limited range of overexpression, relying in part on the abundance of additional regulator proteins. However, we show that *SND2* overexpression has the

© University of Pretoria

potential to enhance fiber development in *Eucalyptus* trees, an important commercial forestry crop.

3.7. Acknowledgements

We thank James Wesley-Smith of the University of Kwazulu-Natal Electron Microscopy Unit for conducting transmission and scanning electron microscopy, Nicky Olivier of the African Centre for Gene Technologies Microarray Facility at the University of Pretoria (UP) for assistance with data analysis, and L'zanne Jansen van Rensburg (UP) for assistance with cell wall chemistry analysis. Martin Ranik (UP) cloned the coding sequence of *SND2* and transformed *Arabidopsis thaliana*. This work was funded by Sappi and Mondi through the Forest Molecular Genetics (FMG) Programme at the University of Pretoria and further supported by funding from The Human Resources for Industry Programme (THRIP) (UID 71890) and the National Foundation for Research (NRF) of South Africa (UID 77178), as well as an Australian Research Council (ARC) Linkage Grant LP0668943 to the University of Melbourne.

3.8. References

Bennett T, Toorn Avd, Sanchez-Perez GF, Campilho A, Willemsen V, Snel B, Scheres B. 2010. SOMBRERO, BEARSKIN1, and BEARSKIN2 regulate root cap maturation in *Arabidopsis*. *The Plant Cell* **22**: 640-654.

Berthet S, Demont-Caulet N, Pollet B, Bidzinski P, Cézard L, Bris PL, Borrega N, Hervé J, Blondet E, Balzergue S, Lapierre C, Jouanin L. 2011. Disruption of LACCASE4 and 17 results in tissue-specific alterations to lignification of *Arabidopsis thaliana* stems. *The Plant Cell* **23**(3): 1124-1137.

Birchler JA, Riddle NC, Auger DL, Veitia RA. 2005. Dosage balance in gene regulation: biological implications. *Trends in Genetics* **21**(4): 219-226.

Bischoff V, Nita S, Neumetzler L, Schindelasch D, Urbain A, Eshed R, Persson S, Delmer D, Scheible W-R. 2010. *TRICHOME BIREFRINGENCE* and its homolog *AT5G01360*

© University of Pretoria

encode plant-specific DUF231 proteins required for cellulose biosynthesis in *Arabidopsis*. *Plant Physiology* **153**: 590–602.

Brown D, Wightman R, Zhang Z, Gomez LD, Atanassov I, Bukowski J-P, Tryfona T, McQueen-Mason SJ, Dupree P, Turner S. 2011. *Arabidopsis* genes *IRREGULAR XYLEM* (*IRX15*) and *IRX15L* encode DUF579-containing proteins that are essential for normal xylan deposition in the secondary cell wall. *The Plant Journal* **66**(3): 401–413.

Brown D, Zeef L, Ellis J, Goodacre R, Turner S. 2005. Identification of novel genes in *Arabidopsis* involved in secondary cell wall formation using expression profiling and reverse genetics. *Plant Cell* **17**: 2281-2295.

Brown DM, Zhang Z, Stephens E, Dupree P, Turner SR. 2009. Characterization of IRX10 and IRX10-like reveals an essential role in glucuronoxylan biosynthesis in *Arabidopsis*. *The Plant Journal* **57**(4): 732-746.

Bu Q, Jiang H, Li C-B, Zhai Q, Zhang J, Wu X, Sun J, Xie Q, Li C. 2008. Role of the *Arabidopsis thaliana* NAC transcription factors ANAC019 and ANAC055 in regulating jasmonic acid-signaled defense responses. *Cell Research* **18**: 756–767.

Che P, Lall S, Nettleton D, Howell SH. 2006. Gene expression programs during shoot, root, and callus development in *Arabidopsis* tissue culture. *Plant Physiology* **141**: 620–637.

Coleman HD, Park J-Y, Nair R, Chapple C, Mansfield SD. 2008. RNAi-mediated suppression of *p*-coumaroyl-CoA 3′-hydroxylase in hybrid poplar impacts lignin deposition and soluble secondary metabolism. *Proceedings of the National Academy of Sciences of the USA* **105**(11): 4501-4506.

Courtois-Moreau CL, Pesquet E, Sjödin A, Muñiz L, Bollhöner B, Kaneda M, Samuels L, Jansson S, Tuominen H. 2009. A unique program for cell death in xylem fibers of *Populus* stem. *The Plant Journal* **58**(2): 260-274.

Crampton BG, Hein I, Berger DK. 2009. Salicylic acid confers resistance to a biotrophic rust pathogen, *Puccinia substriata,* in pearl millet (*Pennisetum glaucum*). *Molecular Plant Pathology* **10**(2): 291–304.

Curtis MD, Grossniklaus U. 2003. A gateway cloning vector set for high-throughput functional analysis of genes *in planta*. *Plant Physiology* **133**: 462-469.

Dhugga KS, Barreiro R, Whitten B, Stecca K, Hazebroek J, Randhawa GS, Dolan M, Kinney AJ, Tomes D, Nichols S, Anderson P. 2004. Guar seed β-mannan synthase is a member of the cellulose synthase super gene family. *Science* **303**: 363-366.

Du J, Groover A. 2010. Transcriptional regulation of secondary growth and wood formation. *Journal of Integrative Plant Biology* **52**(1): 17–27.

Endler A, Persson S. 2011. Cellulose synthases and synthesis in *Arabidopsis*. *Molecular Plant* **4**(2): 199-211.

© University of Pretoria

Fukuda H, Komamine A. 1980. Establishment of an experimental system for the study of tracheary element differentiation from single cells isolated from the mesophyll of *Zinnia elegans*. *Plant Physiology* **65**: 57-60.

Grant EH, Fujino T, Beers EP, Brunner AM. 2010. Characterization of NAC domain transcription factors implicated in control of vascular cell differentiation in *Arabidopsis* and *Populus*. *Planta* **232**: 337-352.

Guerriero G, Fugelstad J, Bulone V. 2010. What do we really know about cellulose biosynthesis in higher plants? *Journal of Integrative Plant Biology* **52**(2): 161–175.

Harper AD, Bar-Peled M. 2002. Biosynthesis of UDP-xylose. Cloning and characterization of a novel *Arabidopsis* gene family, *UXS*, encoding soluble and putative membrane-bound UDP-glucuronic acid decarboxylase isoforms. *Plant Physiology* **130**: 2188–2198.

Hellemans J, Mortier G, Paepe AD, Speleman F, Vandesompele J. 2007. qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. *Genome Biology* **8**: R19.

Hermans C, Porco S, Verbruggen N, Bush DR. 2010. Chitinase-like protein CTL1 plays a role in altering root system architecture in response to multiple environmental conditions. *Plant Physiology* **152**: 904–917.

Heyndrickx KS, Vandepoele K. 2012. Systematic identification of functional plant modules through the integration of complementary data sources. *Plant Physiology* **159**: 884-901.

Hinchee MAW, Mullinax LN, Rottmann WH. 2010. Woody biomass and purpose-grown trees as feedstocks for renewable energy. *Biotechnology in Agriculture and Forestry* **66**(2): 155-208.

Hiratsu K, Mitsuda N, Matsui K, Ohme-Takagi M. 2004. Identification of the minimal repression domain of SUPERMAN shows that the DLELRL hexapeptide is both necessary and sufficient for repression of transcription in *Arabidopsis*. *Biochemical and Biophysical Research Communications* **321**(1): 172-178.

Hruz T, Laule O, Szabo G, Wessendorp F, Bleuler S, Oertle L, Widmayer P, Gruissem W, Zimmermann P. 2008. Genevestigator V3: A reference expression database for the meta-analysis of transcriptomes. *Advances in Bioinformatics* **2008**: 420747.

Hu R, Qi G, Kong Y, Kong D, Gao Q, Zhou G. 2010. Comprehensive analysis of NAC domain transcription factor gene family in *Populus trichocarpa*. *BMC Plant Biology* **10**: 145.

Humphreys JM, Chapple C. 2002. Rewriting the lignin roadmap. *Current Opinion in Plant Biology* **5**: 224–229.

Jensen JK, Kim H, Cocuron J-C, Orler R, Ralph J, Wilkerson CG. 2011. The DUF579 domain containing proteins IRX15 and IRX15-L affect xylan synthesis in *Arabidopsis*. *The Plant Journal* **66**(3): 387–400.

© University of Pretoria

Ko J-H, Kim W-C, Han K-H. 2009. Ectopic expression of MYB46 identifies transcriptional regulatory genes involved in secondary wall biosynthesis in *Arabidopsis*. *The Plant Journal* **60**(4): 649-665.

Ko J-H, Yang SH, Park AH, Lerouxel O, Han K-H. 2007. ANAC012, a member of the plant-specific NAC transcription factor family, negatively regulates xylary fiber development in *Arabidopsis thaliana*. *The Plant Journal* **50**(6): 1035-1048.

Krishnaswamy S, Verma S, Rahman MH, Kav NNV. 2011. Functional characterization of four APETALA2-family genes (*RAP2.6*, *RAP2.6L*, *DREB19* and *DREB26*) in *Arabidopsis*. *Plant Molecular Biology* **75**: 107–127.

Kubo M, Udagawa M, Nishikubo N, Horiguchi G, Yamaguchi M, Ito J, Mimura T, Fukuda H, Demura T. 2005. Transcription switches for protoxylem and metaxylem vessel formation. *Genes and Development* **19**: 1855-1860.

Kzhyshkowska J, Mamidi S, Gratchev A, Kremmer E, Schmuttermaier C, Krusell L, Haus G, Utikal J, Schledzewski K, Scholtze J, Goerdt S. 2006. Novel stabilin-1 interacting chitinase-like protein (SI-CLP) is up-regulated in alternatively activated macrophages and secreted via lysosomal pathway. *Blood* **107**(8): 3221-3228.

Liepman A, Nairn C, Willats W, Sørensen I, Roberts A, Keegstra K. 2007. Functional genomic analysis supports conservation of function among cellulose synthase-like A gene family members and suggests diverse roles of mannans in plants. *Plant Physiology* **143**: 1881-1893.

Martin D, Brun C, Remy E, Mouren P, Thieffr D, Jacq B. 2004. GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biology* **5**: R101.

Mattinen M-L, Suortti T, Gosselink R, Argyropoulos DS, Evtuguin D, Suurnäkki A, Jong Ed, Tamminen T. 2008. Polymerization of different lignins by laccase. *BioResources* **3**(2): 549-565.

Mayor HD, Hampton JC, Rosario B. 1961. A simple method for removing the resin from epoxy-embedded tissue. *Journal of Biophysical and Biochemical Cytology* **9**: 909-910.

McCarthy RL, Zhong R, Fowler S, Lyskowski D, Piyasena H, Carleton K, Spicer C, Ye Z-H. 2010. The poplar MYB transcription factors, PtrMYB3 and PtrMYB20, are involved in the regulation of secondary wall biosynthesis. *Plant and Cell Physiology* **51**(6): 1084–1090.

McCarthy RL, Zhong R, 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 and Cell Physiology* **50**(11): 1950-1964.

Mitsuda N, Iwase A, Yamamoto H, Yoshida M, Seki M, Shinozaki K, 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.

© University of Pretoria

Mitsuda N, Ohme-Takagi M. 2008. NAC transcription factors NST1 and NST3 regulate pod shattering in a partially redundant manner by promoting secondary wall formation after the establishment of tissue identity. *The Plant Journal* **56**(5): 768-778.

Mitsuda N, Seki M, Shinozaki K, Ohme-Takagi M. 2005. The NAC transcription factors NST1 and NST2 of *Arabidopsis* regulate secondary wall thickenings and are required for anther dehiscence. *Plant Cell* **17**: 2993–3006.

Noda S, Takahashi Y, Tsurumaki Y, Yamamura M, Nishikubo N, Yamaguchi M, Sakurai N, Hattori T, Suzuki H, Demura T, Shibata D, Suzuki S, Umezawa T. 2013. ATL54, a RING-H2 domain protein selected by a gene coexpression network analysis, is associated with secondary cell wall formation in *Arabidopsis*. *Plant Biotechnology* **30**: 169-177.

Oda Y, Fukuda H. 2012. Initiation of cell wall pattern by a Rho- and microtubule-driven symmetry breaking. *Science* **337**: 1333-1336.

Oda Y, Fukuda H. 2013. Rho of plant GTPase signaling regulates the behavior of *Arabidopsis* Kinesin-13A to establish secondary cell wall patterns. *The Plant Cell* **25**: 4439-4450.

Ohashi-Ito K, Oda Y, Fukuda H. 2010. *Arabidopsis* VASCULAR-RELATED NAC-DOMAIN6 directly regulates the genes that govern programmed cell death and secondary wall formation during xylem differentiation. *The Plant Cell* **22**: 3461–3473.

Oikawa A, Joshi HJ, Rennie EA, Ebert B, Manisseri C, Heazlewood JL, Scheller HV. 2010. An integrative approach to the identification of *Arabidopsis* and rice genes involved in xylan and secondary wall development. *PLoS ONE* **5**(11): e15481.

Oka T, Jigami Y. 2006. Reconstruction of *de novo* pathway for synthesis of UDP-glucuronic acid and UDP-xylose from intrinsic UDP-glucose in *Saccharomyces cerevisiae*. *Febs Journal* **273**: 2645–2657.

Ona T, Sonoda T, Ito K, Shibata M, Tamai Y, Kojima Y, Ohshima J, Yokota S, Yoshizawa N. 2001. Investigation of relationships between cell and pulp properties in *Eucalyptus* by examination of within-tree property variations. *Wood Science and Technology* **35**(3): 229-243.

Peña M, Zhong R, Zhou G-K, Richardson E, O'Neill M, Darvill A, York W, Ye Z-H. 2007. *Arabidopsis irregular xylem8* and *irregular xylem9*: implications for the complexity of glucuronoxylan biosynthesis. *Plant Cell* **19**: 549-563.

Persson S, Caffall K, Freshour G, Hilley M, Bauer S, Poindexter P, Hahn M, Mohnen D, Somerville C. 2007. The *Arabidopsis irregular xylem8* mutant is deficient in glucuronoxylan and homogalacturonan, which are essential for secondary cell wall integrity. *The Plant Cell* **19**: 237-255.

Plomion C, Leprovost G, Stokes A. 2001. Wood formation in trees. *Plant Physiology* **127**: 1513-1523.

© University of Pretoria

Ramirez M, Rodriguez J, Balocchi C, Peredo M, Elissetche JP, Mendonca R, Valenzuela S. 2009. Chemical composition and wood anatomy of *Eucalyptus globulus* clones: variations and relationships with pulpability and handsheet properties. *Journal of Wood Chemistry and Technology* **29**(1): 43-58.

Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, Miller N, Mueller LA, Mundodi S, Reiser L, Tacklind J, Weems DC, Wu Y, Xu I, Yoo D, Yoon J, Zhang P. 2003. The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Research* **31**(1): 224-228.

Sato K, Suzuki R, Nishikubo N, Takenouchi S, Ito S, Nakano Y, Nakaba S, Sano Y, Funada R, Kajita S, Kitano H, Katayama Y. 2010. Isolation of a novel cell wall architecture mutant of rice with defective *Arabidopsis COBL4* ortholog *BC1* required for regulated deposition of secondary cell wall components. *Planta* **232**: 257–270.

Scheller HV, Ulvskov P. 2010. Hemicelluloses. *Annual Review of Plant Biology* **61**: 10.11–10.27.

Spokevicius A, Southerton S, MacMillan C, Qiu D, Gan S, Tibbits J, Moran G, Bossinger G. 2007. β-Tubulin affects cellulose microfibril orientation in plant secondary fibre cell walls. *The Plant Journal* **51**: 717-726.

Taylor N, Laurie S, Turner S. 2000. Multiple cellulose synthase catalytic subunits are required for cellulose synthesis in *Arabidopsis*. *The Plant Cell* **12**: 2529-2539.

Taylor N, Scheible W-R, Cutler S, Somerville C, Turner S. 1999. The *irregular xylem3* locus of *Arabidopsis* encodes a cellulose synthase required for secondary cell wall synthesis. *The Plant Cell* **11**: 769-779.

Taylor NG, Howells RM, Huttly AK, Vickers K, Turner SR. 2003. Interactions among three distinct CesA proteins essential for cellulose synthesis. *Proceedings of the National Academy of Sciences of the USA* **100**(3): 1450–1455.

Toufighi K, Brady SM, Austin R, Ly E, Provart NJ. 2005. The Botany Array Resource: e-Northerns, Expression Angling, and promoter analyses. *The Plant Journal* **43**: 153-163.

Tran L-SP, Nakashima k, Sakuma Y, Simpson SD, Fujita Y, Maruyama K, Fujita M, Seki M, Shinozaki K, Yamaguchi-Shinozaki K. 2004. Isolation and functional analysis of *Arabidopsis* stress-inducible NAC transcription factors that bind to a drought-responsive *cis*-element in the *early responsive to dehydration stress 1* promoter. *The Plant Cell* **16**: 2481-2498.

Turner S, Gallois P, Brown D. 2007. Tracheary element differentiation. *Annual Reviews of Plant Biology* **58**: 407–433.

**Tuskan GA, 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 RR, Blaudez D,

© University of Pretoria

Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen G-L, Cooper D, Coutinho M, Couturier J, Covert S. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**(5793): 1596-1604.

Umezawa T. 2009. The cinnamate/monolignol pathway. *Phytochemistry Reviews* **9**(1): 1-17.

Vanholme R, Morreel K, Ralph J, Boerjan W. 2008. Lignin engineering. *Current Opinion in Plant Biology* **11**(3): 278-285.

Wang J, Kucukoglu M, Zhang L, Chen P, Decker D, Nilsson O, Jones B, Sandberg G, Zheng B. 2013. The *Arabidopsis* LRR-RLK, *PXC1*, is a regulator of secondary wall formation correlated with the TDIF-PXY/TDR-WOX4 signaling pathway. *BMC Plant Biology* **13**: 94.

Wettenhall JM, Smyth GK. 2004. limmaGUI: A graphical user interface for linear modeling of microarray data. *Bioinformatics* **20**(18): 3705–3706.

Wu A-M, Rihouey C, Seveno M, Hörnblad E, Singh SK, Matsunaga T, Ishii T, Lerouge P, Marchant A. 2009. The *Arabidopsis* IRX10 and IRX10-LIKE glycosyltransferases are critical for glucuronoxylan biosynthesis during secondary cell wall formation. *The Plant Journal* **57**(4): 718-731.

Yamaguchi M, Kubo M, Fukuda H, Demura T. 2008. VASCULAR-RELATED NAC-DOMAIN7 is involved in the differentiation of all types of xylem vessels in *Arabidopsis* roots and shoots. *The Plant Journal* **55**: 652–664.

Yamaguchi M, Mitsuda N, Ohtani M, Ohme-Takagi M, Kato K, 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, Nadia G, Igarashi H, Ohtani M, Nakano Y, Mortimer JC, Nishikubo N, Kubo M, Katayama Y, Kakegawa K, Dupree P, Demura T. 2010. VASCULAR-RELATED NAC-DOMAIN6 (VND6) and VND7 effectively induce transdifferentiation into xylem vessel elements under control of an induction system. *Plant Physiology* **153**: 906-914.

Zhao C, Craig JC, Petzold HE, Dickerman AW, Beers EP. 2005. The xylem and phloem transcriptomes from secondary tissues of the *Arabidopsis* root-hypocotyl. *Plant Physiology* **138**: 803–818.

Zhong R, Demura T, Ye ZH. 2006. SND1, a NAC domain transcription factor, is a key regulator of secondary wall synthesis in fibers of *Arabidopsis*. *Plant Cell* **18**(11): 3158-3170.

Zhong R, Lee C, Ye Z-H. 2010a. Evolutionary conservation of the transcriptional network regulating secondary cell wall biosynthesis. *Trends in Plant Science* **15**(11): 625-632.

Zhong R, Lee C, Ye Z-H. 2010b. Functional characterization of poplar wood-associated NAC domain transcription factors. *Plant Physiology* **152**: 1044–1055.

© University of Pretoria

Zhong R, Lee C, Ye Z-H. 2010c. Global analysis of direct targets of secondary wall NAC master switches in *Arabidopsis*. *Molecular Plant* **3**(6): 1087-1103.

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

Zhong R, McCarthy RL, Lee C, Ye Z-H. 2011. Dissection of the transcriptional program regulating secondary wall biosynthesis during wood formation in poplar. *Plant Physiology* **157**: 1452–1468.

Zhong R, Richardson EA, Y Z-H. 2007a. The MYB46 transcription factor is a direct target of SND1 and regulates secondary wall biosynthesis in *Arabidopsis*. *The Plant Cell* **19**: 2776-2792.

Zhong R, Richardson EA, Ye Z-H. 2007b. Two NAC domain transcription factors, SND1 and NST1, function redundantly in regulation of secondary wall synthesis in fibers of *Arabidopsis*. *Planta* **225**(6): 1603-1611.

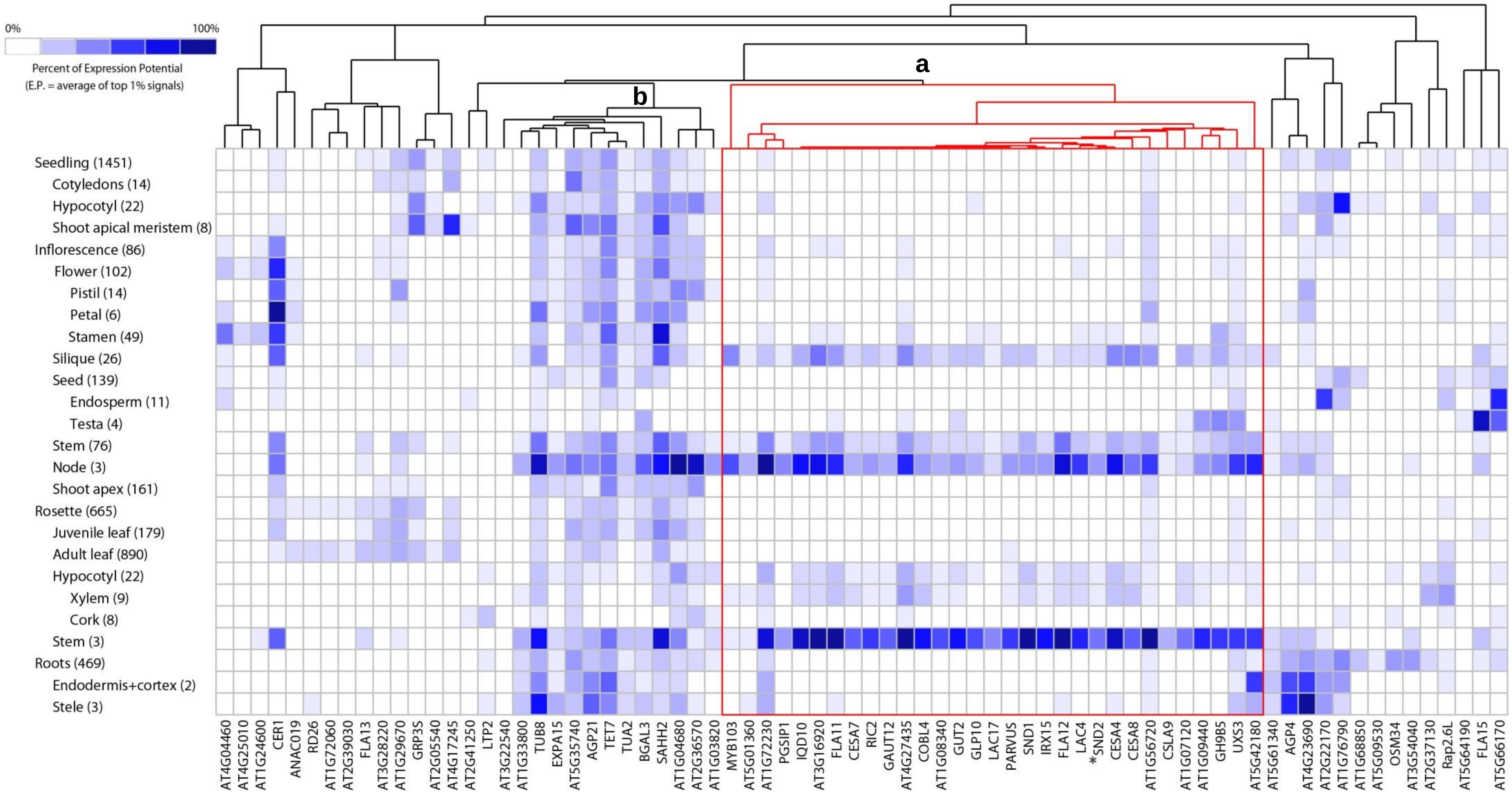
Zhou J, Lee C, Zhong R, Ye Z-H. 2009. MYB58 and MYB63 are transcriptional activators of the lignin biosynthetic pathway during secondary cell wall formation in *Arabidopsis*. *The Plant Cell* **21**(1): 248-266.

© University of Pretoria

3.9. Figures

Fig. 3.1. Absolute transcript abundance of SND2∩Ko genes represented on ATH1 22k arrays in *Arabidopsis* tissues and organs. Genevestigator V3 (Hruz *et al.*, 2008) was used for microarray data mining, and the anatomical cluster analysis tool was used to visualize and cluster the genes according to their tissue-specific expression patterns. Tissues/organs are staggered hierarchically, and the number of arrays on which the data are based is indicated in parentheses. Absolute transcript values are expressed as a percentage of their expression potential (E.P.), where E.P. is the mean of the top 1% of hybridization signals for a given probe set across all arrays. Cluster **(a)**, highlighted in red, is comprised of 31 genes, including SND2 (*), which displayed preferential expression in tissues and organs where SND2 is expressed. Cluster **(b)** encompasses of 13 genes which displayed preferential expression in inflorescence stems and nodes, rosette stems, and in some cases the stamen, seedling hypocotyl and/or vasculature (stele) of roots.

© University of Pretoria



© University of Pretoria

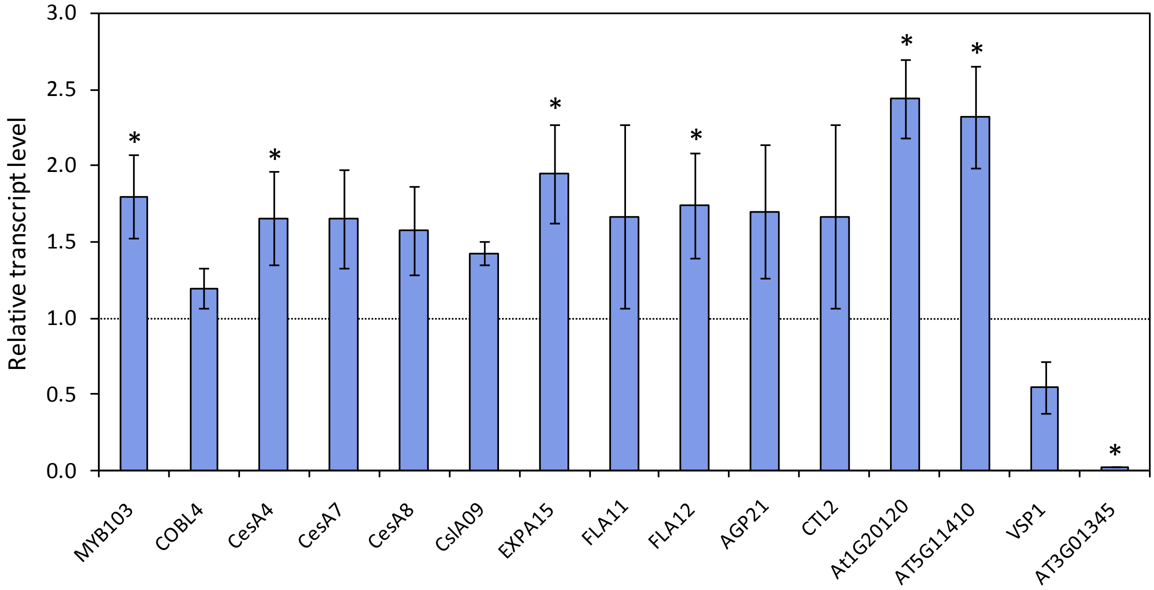


Fig. 3.2. RT-qPCR analysis of selected genes differentially expressed in inflorescence stems of eight-week-old SND2-OV(A) and wild type plants. SND2-OV(A) plants were grown alongside the -wild type in three biological replicate pairs, with primary stems from six plants pooled per sample. SND2-OV(A) transcript levels were normalized to the wild type in each replicate (assigned a value of 1, for each gene), hence error bars indicate the standard error of the deviation from wild type across biological replicates. Significance was evaluated by a one-tailed paired t-test, in accordance with the expected direction of response for each gene; **P* < 0.05.

© University of Pretoria

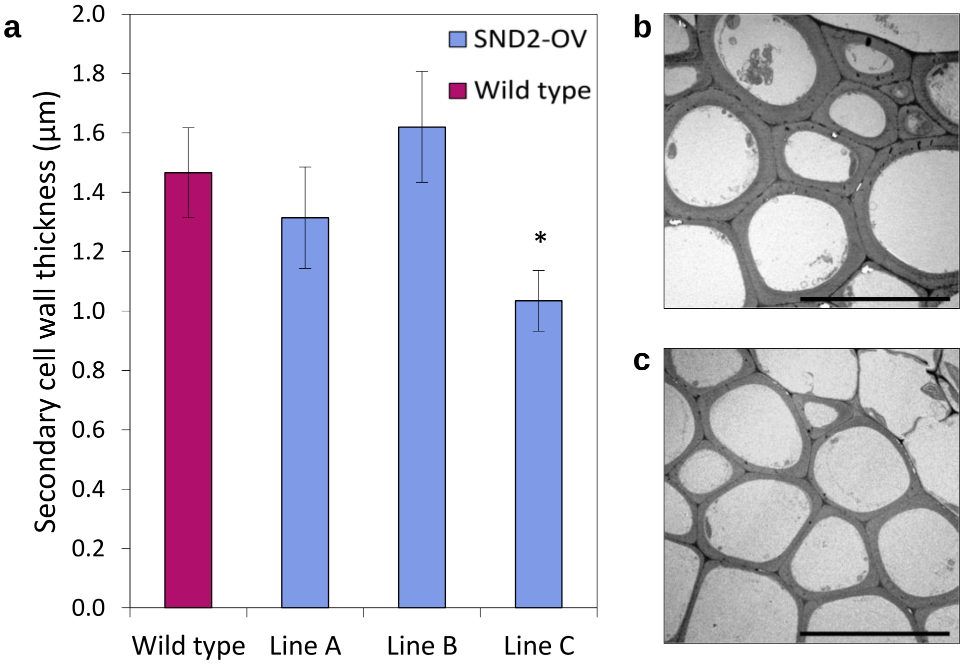


Fig. 3.3. SCW thickness in IFs of eight-week-old wild type and T4 homozygous SND2-OV lines A, B and C. (a) SCW thickness measurements based on scanning electron micrographs. Error bars indicate the standard error of the mean of three biological replicates (21-42 fibers were measured per line). *Significantly different from wild type according to homoscedastic two-tailed Student's t-test (*P* < 0.02). Transmission micrographs of representative IF regions of wild type and SND2-OV line C stems are shown in **(b)** and **(c)** respectively (scale bars = 20 µm).

© University of Pretoria

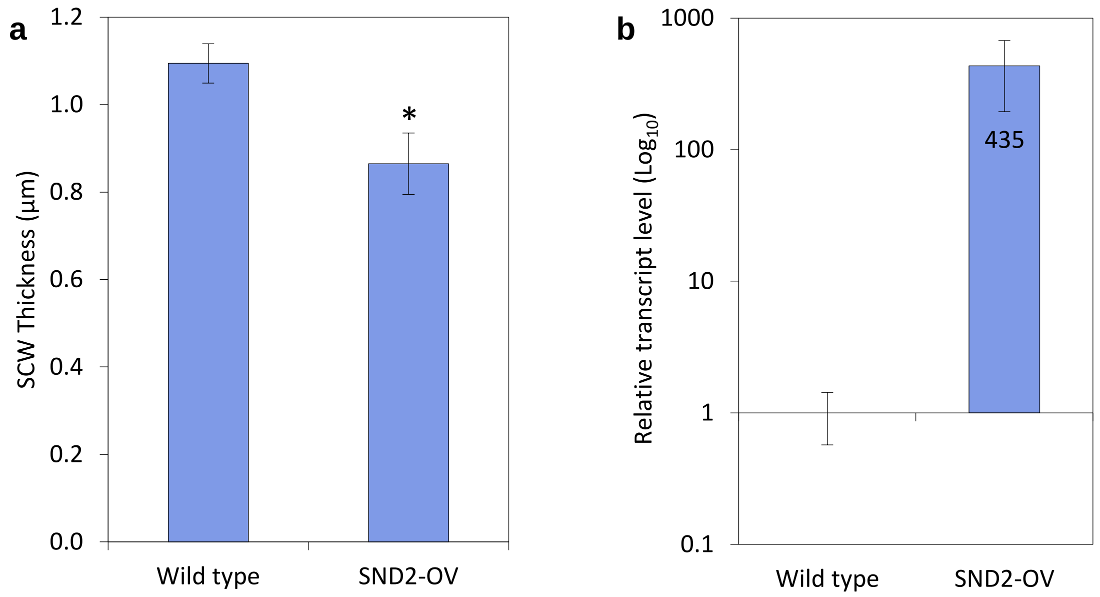


Fig. 3.4. Effect of SND2 overexpression on IF wall thickness in T1 generation stems. (a) Mean SCW thickness in IFs of eight-week-old wild type and T1 generation SND2-OV stems. Representative light microscopy images are shown in Fig. S3.6. Error bars indicate the standard error of the mean of eight wild type and seven T1 plants (26-48 fibers were measured per plant). *Significantly different from wild type based on homoscedastic two-tailed Student's t-test (P < 0.02). **(b)** Corresponding transcript abundance of total *SND2* transcript in lower stems of six wild type and six SND2-OV T1 plants used for SCW measurements, as measured by RT-qPCR. The primer pair quantifies endogenous and transgenic *SND2* transcript. Total *SND2* transcript is ~435-fold relative to the wild type, represented here on a $\log_{10}$ scale. Calibrated Normalized Relative Quantity (CNRQ) values were obtained by normalization against three control genes. Error bars indicate the standard error of the mean of six plants.

© University of Pretoria

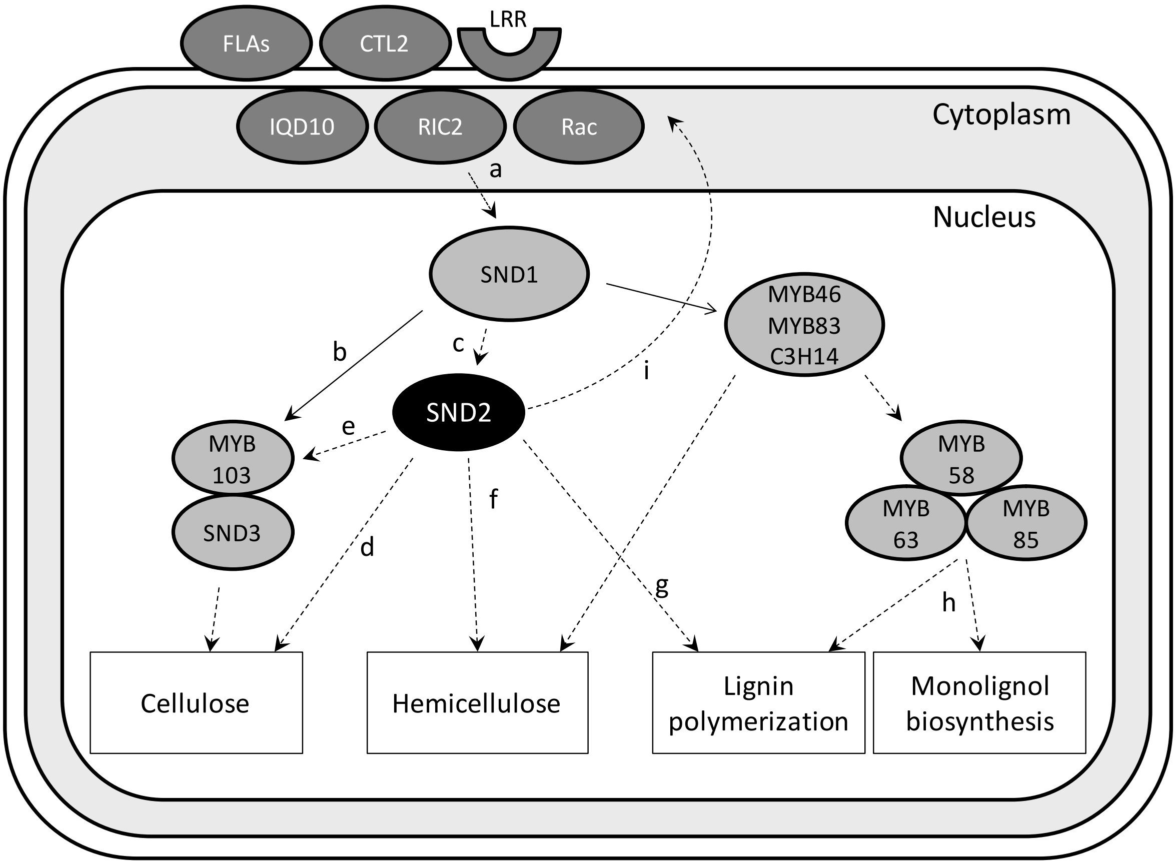


Fig. 3.5. Proposed model of SND2-mediated SCW regulation in IFs. Solid lines indicate known direct protein-DNA interactions. Dashed lines indicate direct or indirect protein-DNA interactions. Master regulator SND1 is activated by a signal transduction pathway proposed by Oikawa *et al.* (2010) (a). SND1 directly activates transcription of *MYB103* and *SND3* (b), and indirectly activates *SND2* through an unknown intermediate (c; Zhong *et al.*, 2008). SND2 activates cellulose-synthesizing *CesA*s, either directly (d) or through the activation of *MYB103* (e), which is known to activate SCW cellulose gene, *CesA8* (Zhong *et al.*, 2008). SND2 regulates hemicellulosic genes (f; Table 3.1), independently to a similar role played by direct SND1 targets *MYB46*, *MYB83* or *C3H14* (Zhong *et al.*, 2007a; Ko *et al.*, 2009; McCarthy *et al.*, 2009). SND2 plays a role in lignification through activation of lignin polymerization genes *LAC4* and *LAC17* (g; Table 3.1), but it does not regulate monolignol biosynthetic genes as is the case for MYB58, MYB63 and MYB85 (h) (Zhong *et al.*, 2008; Zhou *et al.*, 2009). SND2 activates transcription of GPI-anchored *FLA11/FLA12*, *CTL2* and other components of the signal transduction pathway (i), which leads to upregulation of *SND1* (a).

© University of Pretoria

3.10. Tables

Table 3.1. Subset of SND1-regulated genes (Ko *et al.*, 2007) also significantly differentially expressed in stems of eight-week-old SND2-OV(A) plants relative to wild type (SND2∩Ko).

Locus	Description	Fold change	*P*-value[a]
Transcription[b]			
AT4G28500	ANAC073/SND2 (*Arabidopsis* NAC domain containing protein 73); transcription factor	>100.00[c]	0.00E+00
AT1G63910	MYB103 (MYB DOMAIN PROTEIN 103); DNA binding / transcription factor	1.83	1.13E-11
AT1G52890	ANAC019 (*Arabidopsis* NAC domain containing protein 19); transcription factor	1.44	3.07E-04
AT1G32770	ANAC012/NST3/SND1 (*ARABIDOPSIS* NAC DOMAIN CONTAINING PROTEIN 12); transcription factor	1.42	6.93E-04
AT4G17245	Zinc finger (C3HC4-type RING finger) family protein	1.36	4.97E-03
AT5G13330	RAP2.6L (related to AP2 6L); DNA binding / transcription factor	1.36	4.87E-03
Secondary cell wall biosynthesis and cell wall modification			
AT2G03090	EXPA15 (EXPANSIN A15)	2.03	3.60E-16
AT5G44030	CESA4 (CELLULOSE SYNTHASE 4); transferase, transferring glycosyl groups	1.84	7.91E-12
AT5G60490	FLA12 (fasciclin-like arabinogalactan-protein 12)	1.83	1.37E-11
AT2G38080	IRX12/LAC4 (laccase 4); copper ion binding / oxidoreductase	1.77	2.51E-10
AT5G17420	CesA7/IRX3 (IRREGULAR XYLEM 3, MURUS 10); cellulose synthase	1.73	1.25E-09
AT5G03170	FLA11 (fasciclin-like arabinogalactan-protein 11)	1.72	2.80E-09
AT5G03760	CSLA09 (RESISTANT TO AGROBACTERIUM TRANSFORMATION 4); transferase, transferring glycosyl groups	1.66	4.33E-08
AT4G18780	CESA8 (CELLULOSE SYNTHASE 8); cellulose synthase/ transferase, transferring glycosyl groups	1.63	1.33E-07
AT5G15630	COBL4/IRX6 (COBRA-LIKE4)	1.62	2.12E-07
AT5G60020	LAC17 (laccase 17); copper ion binding / oxidoreductase	1.59	7.44E-07
AT3G18660	PGSIP1 (PLANT GLYCOGENIN-LIKE STARCH INITIATION PROTEIN 1); transferase, transferring glycosyl groups	1.58	1.21E-06
AT3G50220	IRX15; domain of unknown function 579 (DUF579)-containing protein	1.55	3.97E-06
AT5G54690	GAUT12/IRX8/LGT6 (GALACTURONOSYLTRANSFERASE 12); polygalacturonate 4-alpha-galacturonosyltransferase	1.39	1.81E-03
AT5G59290	UXS3 (UDP-GLUCURONIC ACID DECARBOXYLASE)	1.38	2.66E-03
AT1G19300	GATL1/GLZ1/PARVUS (GALACTURONOSYLTRANSFERASE-LIKE 1); polygalacturonate 4-alpha-galacturonosyltransferase	1.33	1.08E-02
AT5G01360	TBL3; domain of unknown function 231 (DUF231)-containing protein	1.31	2.08E-02
AT1G27440	GUT2/IRX10 (glucuronoxylan glucuronosyltransferase)	1.30	2.57E-02
Signal transduction			
AT3G16920	CTL2 (Chitinase -like protein 2)	1.71	4.23E-09
AT1G09440	Protein kinase family protein	1.47	1.21E-04
AT3G15050	IQD10 (IQ-domain 10); calmodulin binding	1.46	1.68E-04
AT1G27380	RIC2 (ROP-INTERACTIVE CRIB MOTIF-CONTAINING PROTEIN 2)	1.43	4.07E-04
AT1G56720	Protein kinase family protein	1.41	9.28E-04
AT1G08340	Rho GTPase activating protein, putative	1.38	2.29E-03
AT2G36570	PXC1 (Leucine-rich repeat transmembrane protein kinase)	1.32	1.47E-02

© University of Pretoria

Table S3.1. *(continued)*

Locus	Description	Fold change	*P*-value[a]
Carbohydrate metabolism			
AT5G35740	Glycosyl hydrolase family protein 17	1.57	2.54E-06
AT1G04680	Pectate lyase family protein	1.57	2.11E-06
AT4G36360	BGAL3 (beta-galactosidase 3); beta-galactosidase	1.45	2.66E-04
AT1G19940	GH9B5 (GLYCOSYL HYDROLASE 9B5); hydrolase, hydrolyzing O-glycosyl compounds	1.41	1.09E-03
Abiotic and biotic stress response			
AT5G42180	Peroxidase 64 (PER64) (P64) (PRXR4)	1.66	4.29E-08
AT1G72060	Serine-type endopeptidase inhibitor	1.52	1.59E-05
AT4G27410	RD26 (RESPONSIVE TO DESSICATION 26)	1.37	3.20E-03
AT2G37130	Peroxidase 21 (PER21) (P21) (PRXR5)	1.29	3.47E-02
AT4G23690	Disease resistance-responsive family protein / dirigent family protein	-1.34	7.80E-03
AT1G68850	Peroxidase, putative	-1.41	1.08E-03
AT4G11650	OSM34 (OSMOTIN 34)	-1.69	8.89E-09
AT5G24780	VSP1 (VEGETATIVE STORAGE PROTEIN 1); acid phosphatase	-2.38	3.08E-24
Cytoskeleton			
AT1G50010	TUA2 (tubulin alpha-2 chain)	1.47	1.26E-04
AT5G23860	TUB8 (tubulin beta-8)	1.36	4.58E-03
One-carbon metabolism			
AT3G23810	SAHH2 (S-ADENOSYL-L-HOMOCYSTEINE (SAH) HYDROLASE 2); adenosylhomocysteinase	1.41	9.68E-04
Lipid metabolism			
AT1G29670	GDSL-motif lipase/hydrolase family protein	1.28	4.39E-02
AT1G21360	GLTP2 (GLYCOLIPID TRANSFER PROTEIN 2)	-1.77	1.99E-10
Wax biosynthesis			
AT1G02205	CER1 (ECERIFERUM 1)	1.35	2.04E-03
Unknown function			
AT3G22540	Unknown protein	1.58	1.50E-06
AT1G33800	Unknown protein	1.55	4.24E-06
AT4G27435	Unknown protein	1.43	5.42E-04
AT5G64190	Unknown protein	1.42	6.86E-04
AT5G61340	Unknown protein	1.39	1.63E-03
AT1G07120	Unknown protein	1.32	1.47E-02
AT1G03820	Unknown protein	1.32	1.88E-02
AT1G24600	Unknown protein	-1.33	1.29E-02
AT5G66170	Unknown protein	-1.36	4.29E-03
Unassigned			
AT1G55330	AGP21 (ARABINOGALACTAN PROTEIN 21)	1.71	4.63E-09
AT4G28050	TET7 (TETRASPANIN7)	1.64	1.16E-07
AT3G54040	Photoassimilate-responsive protein-related	1.62	2.52E-07

© University of Pretoria

Table S3.1. *(continued)*

Locus	Description	Fold change	*P*-value[a]
AT2G41250	Haloacid dehalogenase-like hydrolase superfamily protein	1.57	2.23E-06
AT5G44130	FLA13 (FASCICLIN-LIKE ARABINOGALACTAN PROTEIN 13 PRECURSOR)	1.44	3.97E-04
AT2G05540	Glycine-rich protein	1.43	4.36E-04
AT3G62020	GLP10 (GERMIN-LIKE PROTEIN 10); manganese ion binding / metal ion binding / nutrient reservoir	1.40	1.46E-03
AT5G10430	AGP4 (ARABINOGALACTAN-PROTEIN 4)	1.38	2.29E-03
AT3G52370	FLA15 (FASCICLIN-LIKE ARABINOGALACTAN PROTEIN 15 PRECURSOR)	1.37	4.03E-03
AT2G05380	GRP3S (GLYCINE-RICH PROTEIN 3 SHORT ISOFORM)	1.37	3.58E-03
AT2G22170	Lipid-associated family protein	1.35	7.62E-03
AT1G72230	Plastocyanin-like domain-containing protein	1.32	1.70E-02
AT4G04460	Aspartyl protease family protein	-1.29	4.11E-02
AT1G76790	O-methyltransferase family 2 protein	-1.39	1.94E-03
AT3G28220	Meprin and TRAF homology domain-containing protein / MATH domain-containing protein	-1.60	5.82E-07
AT4G25010	Nodulin MtN3 family protein	-1.64	1.00E-07
AT2G39030	GCN5-related N-acetyltransferase (GNAT) family protein	-1.80	4.90E-11
AT5G09530	Hydroxyproline-rich glycoprotein family protein	-3.09	1.31E-41

[a] Adjusted *P*-value according to False Discovery Rate (FDR) method

[b] Genes are categorized by Gene Ontology classification according to The *Arabidopsis* Information Resource (www.arabidopsis.org), unless otherwise described in the main text

[c] Transgene. The fold change is likely an underestimate of the actual value because this target displayed a saturated hybridization signal

© University of Pretoria

Table 3.2. Genes co-expressed with endogenous *SND2* transcript. The R-value represents the Pearson correlation coefficient of co-expression, set to a threshold of R > 0.90. Co-expressed genes that were also differentially expressed in the SND2∩Ko subset of *SND2* overexpression data (Table 3.1) are indicated in the far right column.

Co-expressed gene	R-value	Description	SND2∩Ko
PGSIP1 (AT3G18660)	0.980	Plant glycogenin-like starch initiation protein 1	√
IQD10 (AT3G15050)	0.979	Calmodulin-binding protein	√
MYB103 (AT1G63910)	0.973	Secondary cell wall-associated transcription factor	√
IRX8 (AT5G54690)	0.972	Galacturonosyltransferase 12	√
COBL4 (AT5G15630)	0.972	COBRA-like protein	√
IRX15 (AT3G50220)	0.967	DUF579 protein required for normal xylan synthesis	√
IRX15-L (AT5G67210)	0.963	DUF579 protein required for normal xylan synthesis	
CesA7 (AT5G17420)	0.962	Secondary cell wall cellulose synthase protein	√
GLP10 (AT3G62020)	0.959	Germin-like protein 10	√
FLA11 (AT5G03170)	0.958	Fasciclin-like arabinogalactan protein	√
LAC4 (AT2G38080)	0.955	IRREGULAR XYLEM 12	√
LAC2 (AT2G29130)	0.953	Laccase	
AT1G08340	0.952	Rho GTPase activating protein	√
CTL2 (AT3G16920)	0.951	Chitinase-like protein 2	√
AT1G80170	0.950	Pectin lyase-like superfamily protein	
AT2G41610	0.950	Unknown protein	
SND1 (AT1G32770)	0.948	Secondary cell wall-associated transcription factor	√
RIC2 (AT1G27380)	0.948	ROP-interactive CRIB motif-containing protein	√
MAP65-8 (AT1G27920)	0.941	Microtubule-associated protein	
AT1G07120	0.938	Unknown protein	√
AT2G31930	0.936	Unknown protein	
CesA4 (AT5G44030)	0.934	Secondary cell wall cellulose synthase protein	√
AT4G27435	0.934	Protein of unknown function (DUF1218)	√
AT1G22480	0.933	Cupredoxin superfamily protein	
IRX10 (AT1G27440)	0.929	Glucuronoxylan glucuronosyltransferase	√
CesA8 (AT4G18780)	0.926	Secondary cell wall cellulose synthase protein	√
RWA3 (AT2G34410)	0.915	Polysaccharide O-acetyltransferase	
AT4G28380	0.915	Leucine-rich repeat (LRR) family protein	
LAC17 (AT5G60020)	0.914	Laccase	√
PARVUS (AT1G19300)	0.904	Polygalacturonate 4-α-galacturonosyltransferase	√
TBL3 (AT5G01360)	0.902	DUF231 protein involved in cellulose biosynthesis	√

© University of Pretoria

Table 3.3. Klason lignin content of SCW material of T4 SND2-OV(A) stems compared to the wild type control. Values are expressed as the mean of three biological replicates plus or minus the standard error of the mean. *P*-values are based on paired two-tailed Student's t-tests between SND2-OV(A) and the wild type.

Sample	**Total lignin (%)**	**Insoluble lignin (%)**	**Soluble lignin (%)**
SND2-OV(A)	21.06 ± 0.18	15.74 ± 0.20	5.32 ± 0.03
Wild type	21.61 ± 0.21	16.25 ± 0.32	5.43 ± 0.18
P-value	0.033	0.074	0.623

© University of Pretoria

Table 3.4. Monosaccharide composition of SCW material of T4 SND2-OV(A) stems compared to the wild type control. Values (mg/g dry weight) are expressed as the mean of three biological replicates plus or minus the standard error of the mean. *P*-values are based on paired two-tailed Student's t-tests between SND2-OV(A) and the wild type.

Sample	**Glucose**	**Xylose**	**Mannose**	**Galactose**	**Arabinose**	**Rhamnose**	**Fucose**
SND2-OV(A)	343.36 ± 1.42	109.85 ± 0.16	18.06 ± 0.04	18.59 ± 0.08	8.13 ± 0.03	12.83 ± 0.02	0.84 ± 0.01
Wild Type	344.94 ± 2.92	108.62 ± 2.71	16.82 ± 0.35	18.25 ± 0.41	7.88 ± 0.09	12.17 ± 0.18	0.88 ± 0.01
P-value	0.561	0.663	0.049	0.563	0.238	0.023	0.663

© University of Pretoria

Table 3.5. Change in fiber SCW thickness, cell wall area, fiber cell area and lumen area of hybrid *Eucalyptus* sectors overexpressing *Arabidopsis SND2*. *SND2*-overexpressing (SND2-OV) and empty vector control (EVC) sector values are expressed as a percentage change relative to non-transformed tissues. Measurements were obtained from 11 (SND2-OV) and 9 (EVC) transgenic-nontransgenic control sector pairs from two F1 *Eucalyptus* hybrids. *P* values are based on one-tailed Student's t-test.

Sample	Cell wall thickness (%)	Cell wall area (%)	Fiber cell area (%)	Lumen area (%)
SND2-OV	9.99 ± 2.34	14.60 ± 2.64	14.41 ± 2.44	9.68 ± 4.65
EVC	5.54 ± 4.33	6.16 ± 4.98	5.04 ± 4.84	3.78 ± 7.13
P-value	0.177	0.066	0.042	0.241

© University of Pretoria

3.11. Supplementary figures

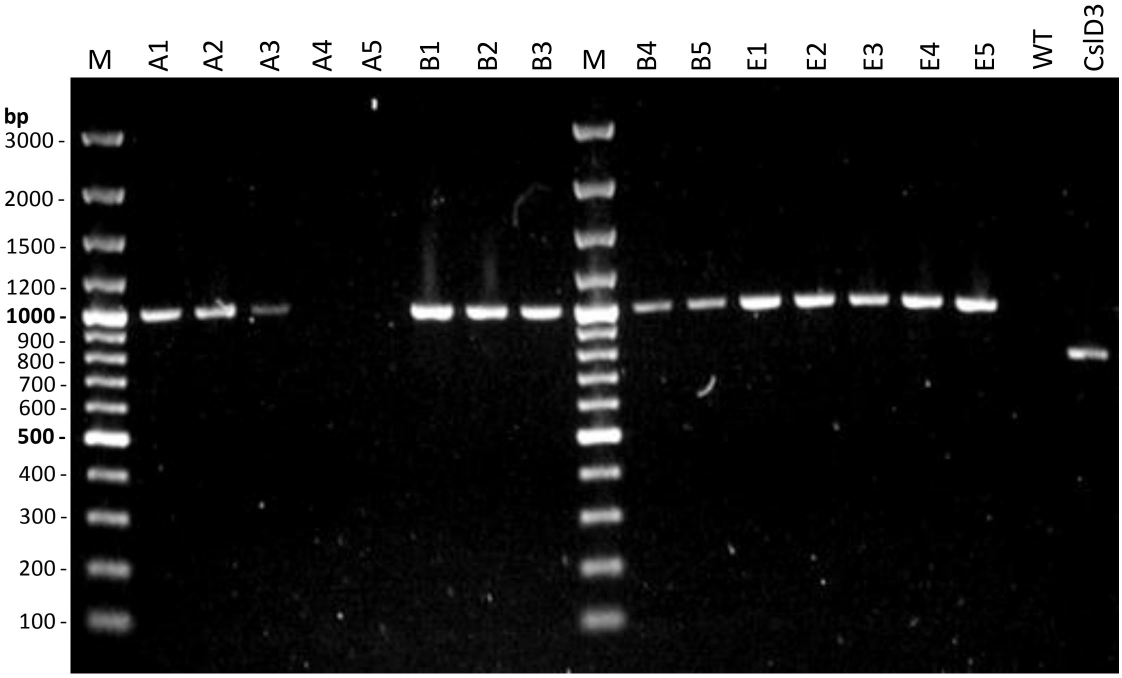


Fig. S3.1. PCR confirmation of the SND2 transgene in individual *A. thaliana* individuals (numbered) from lines A, B and E. The expected size of the transgene-specific amplicon is 972 bp. The endogenous *CslD3* promoter fragment (774 bp) was amplified as a positive control for gDNA template quality. WT, wild type Col-0; M, GeneRuler 100 bp DNA ladder plus.

© University of Pretoria

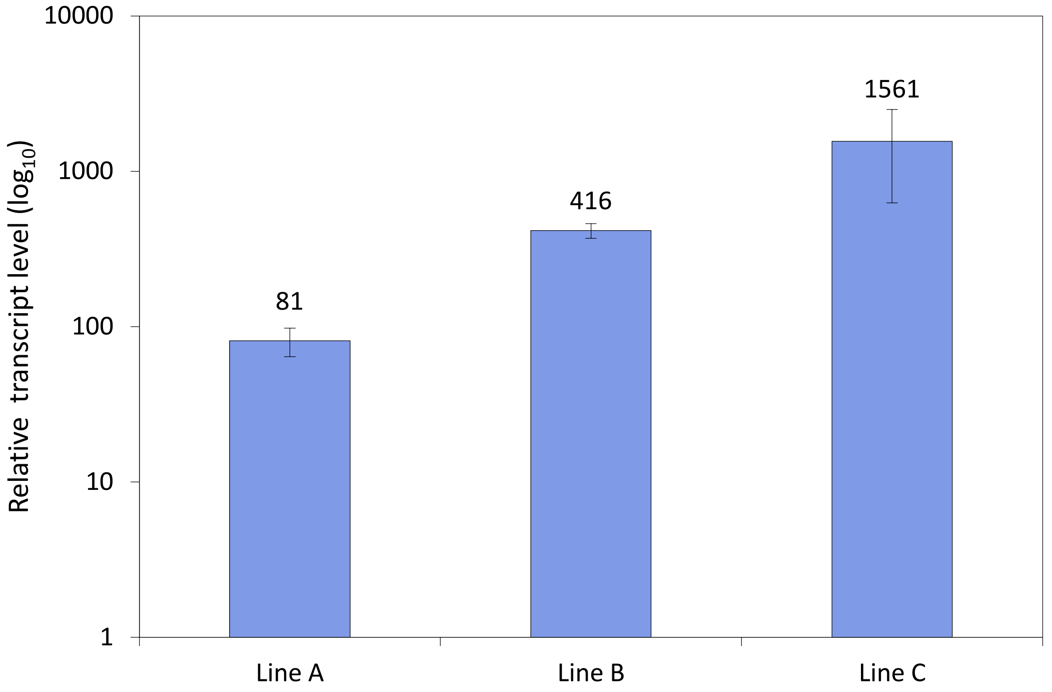
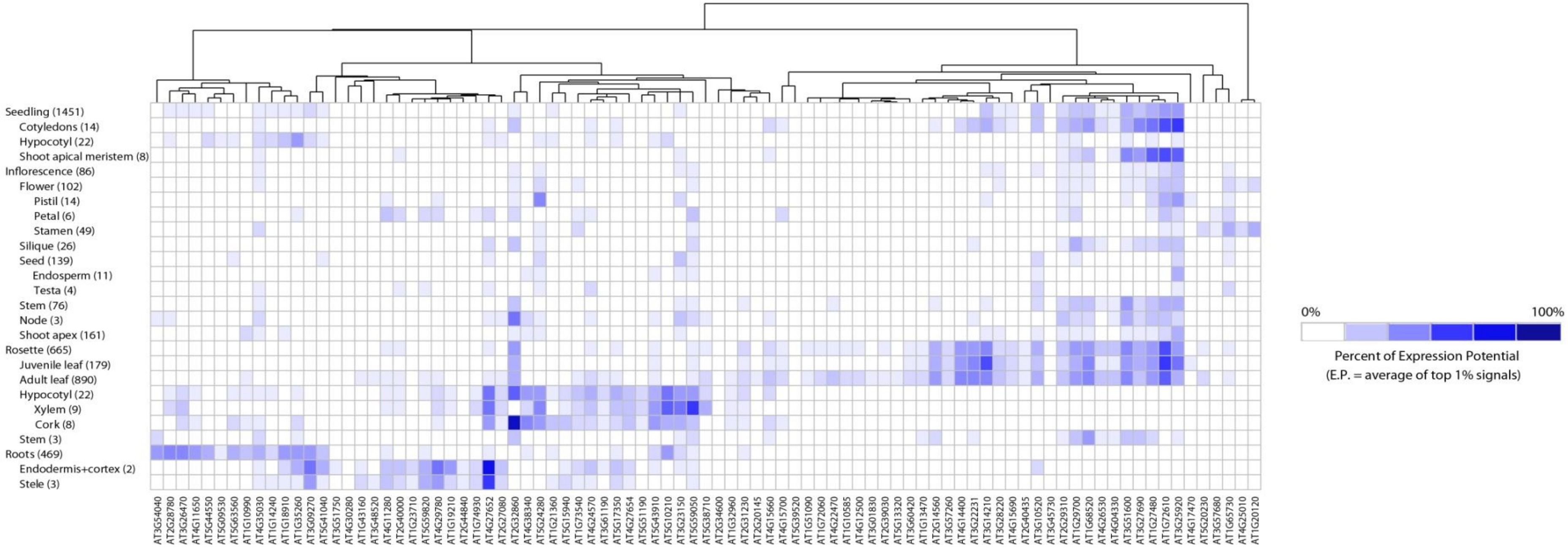


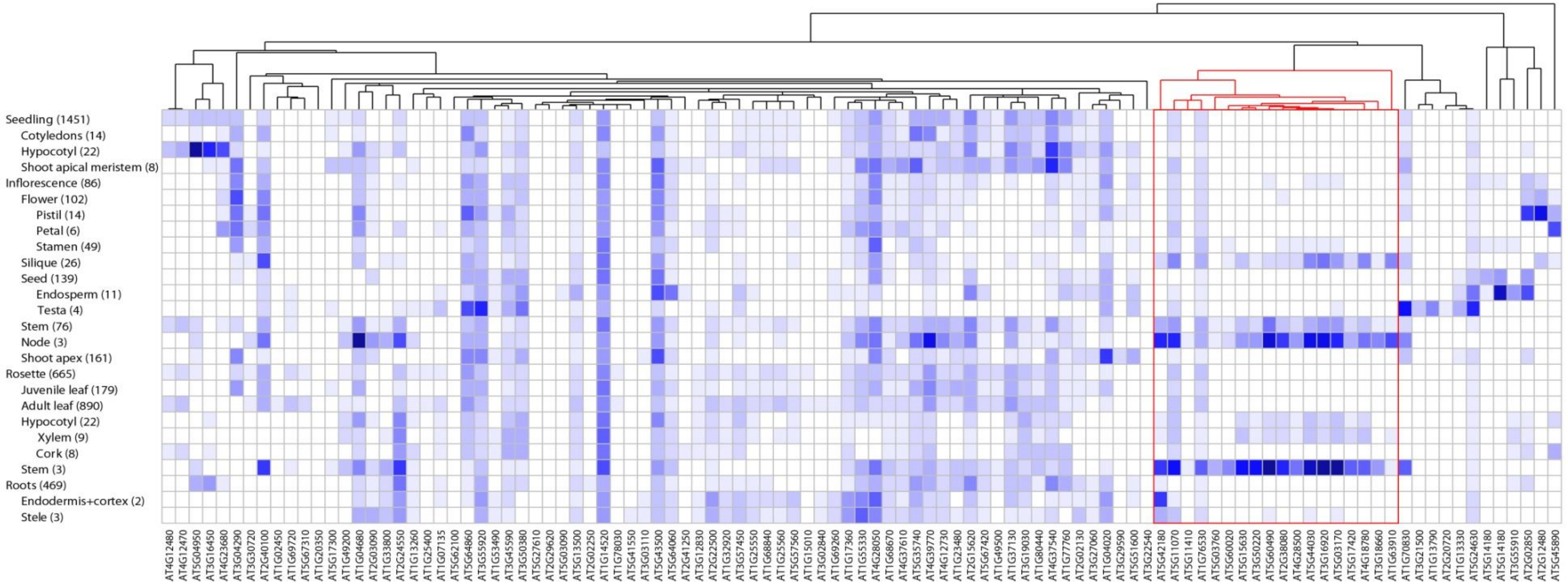
Fig. S3.2. RT-qPCR analysis of total *SND2* transcript in lower inflorescence stems of T4 SND2-OV lines A, B and C, relative to the wild type. The primer pair quantifies both endogenous and transgenic *SND2* transcript. Calibrated Normalized Relative Quantity (CNRQ) values were obtained by normalizing against three control genes (*ACTIN2*, *UBIQUITIN5*, *EF1α*). SND2-OV transcript levels were normalized to the wild type (assigned a value of 1) in each replicate, hence error bars indicate the standard error of the deviation from wild type across biological replicates. The means are indicated above each bar.

© University of Pretoria

Fig. S3.3. Heat map of absolute transcript abundance of genes represented on ATH1 22k arrays, that were differentially expressed by at least 1.5-fold in SND2-OV(A) stems relative to wild type, in various *Arabidopsis* tissues and organs. Genevestigator V3 was used for microarray data mining, and the anatomical cluster analysis tool was used to visualize and cluster the genes according to their tissue-specific expression patterns. Tissues/organs are staggered hierarchically, and the number of arrays on which the data are based is indicated in parenthesis. Absolute transcript values are expressed as a percentage of their expression potential (E.P.), where E.P. is the mean of the top 1% of hybridization signals for a given probe set across all arrays. The cluster highlighted in red is comprised of 18 genes, including *SND2* (AT4G28500), which displayed preferential expression in tissues and organs containing secondary cell walls.

© University of Pretoria





© University of Pretoria

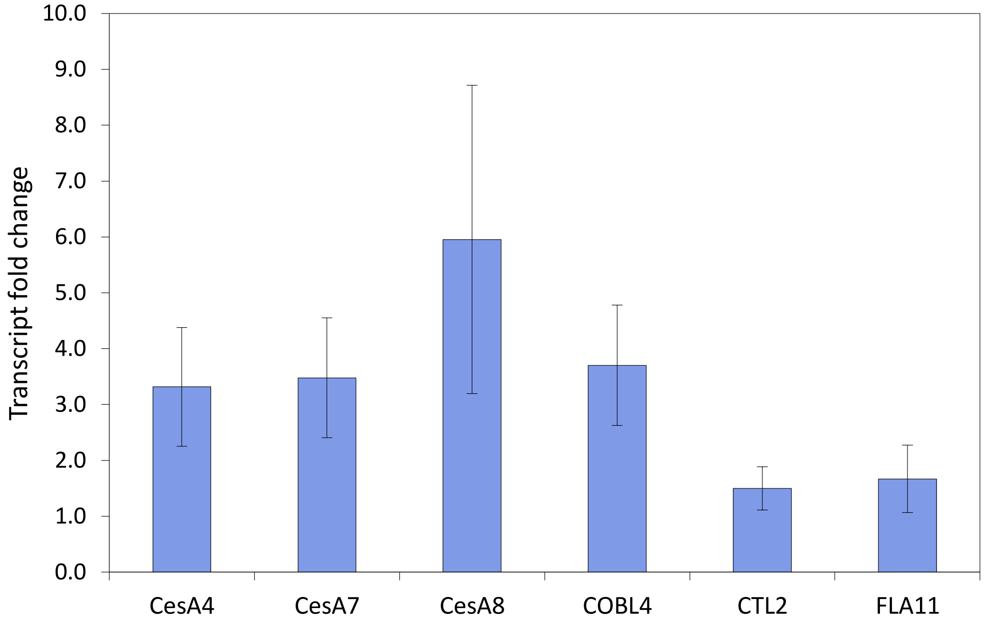


Fig. S3.4. Fold change of RT-qPCR analysis of selected genes differentially expressed in SND2-OV(A) inflorescence stems at eight weeks, from an independent trial to that of the microarray analysis. SND2-OV(A) plants were grown alongside the wild type in three paired biological replicates, where primary inflorescence stems from ten plants were pooled per replicate. SND2-OV(A) transcript levels were normalized to the wild type in each replicate (assigned a value of 1, for each gene), hence error bars indicate the standard error of the deviation from wild type across biological replicates.

© University of Pretoria

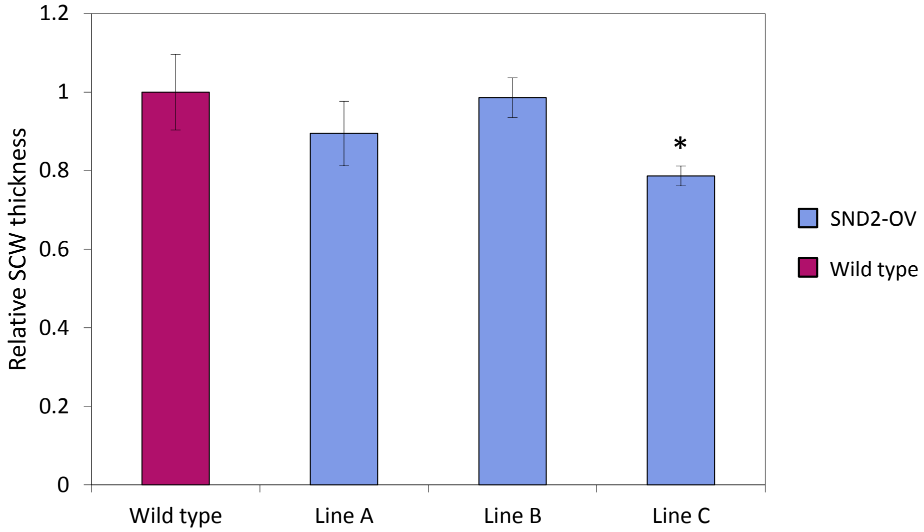


Fig. S3.5. Relative secondary cell (SCW) wall thickness in interfascicular fibers of eight-week-old wild type and T4 SND2-OV lines, determined by light microscopy. Measurements were obtained from Toluidine Blue-stained 400X images and are normalized to wild type. Error bars indicate the standard error of the mean of three plants, where for each plant between two and five interfascicular fiber regions were measured and an average value obtained across regions (10 to 66 fibers measured per plant). Samples were obtained from an independent trial to that of the SEM analysis. *Significantly different from the wild type based on a two-tailed Student's t-test ($P < 0.05$).

© University of Pretoria

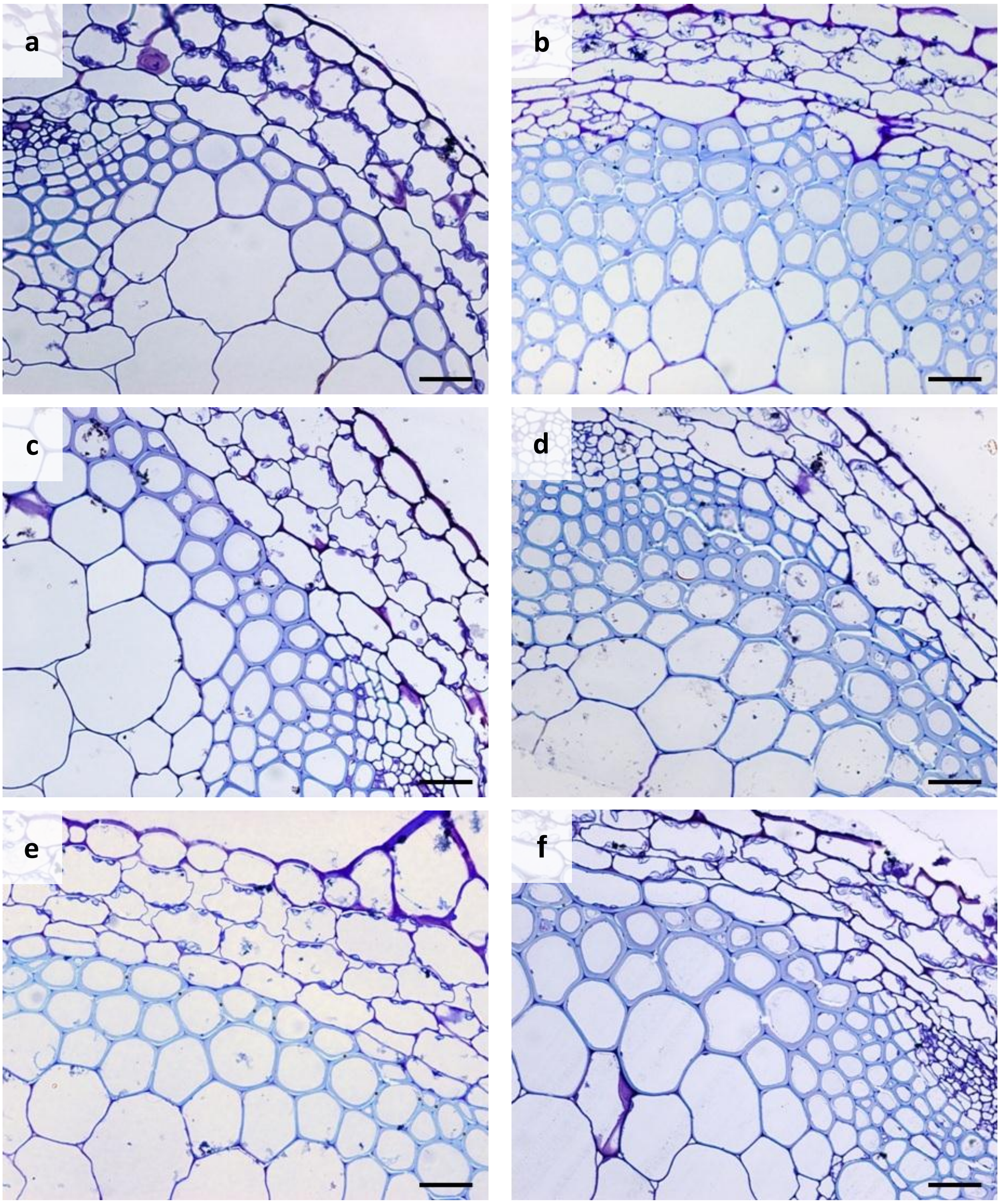


Fig. S3.6. Light microscopy images showing decreased interfascicular fiber SCW thicknesses in T1 SND2-OV plants. Sections were stained with toluidine blue. **a**, **c**, **e**, T1 SND2-OV inflorescence stem cross-sections; **b**, **d**, **f**, wild type. Scale bars = 20 µm.

© University of Pretoria

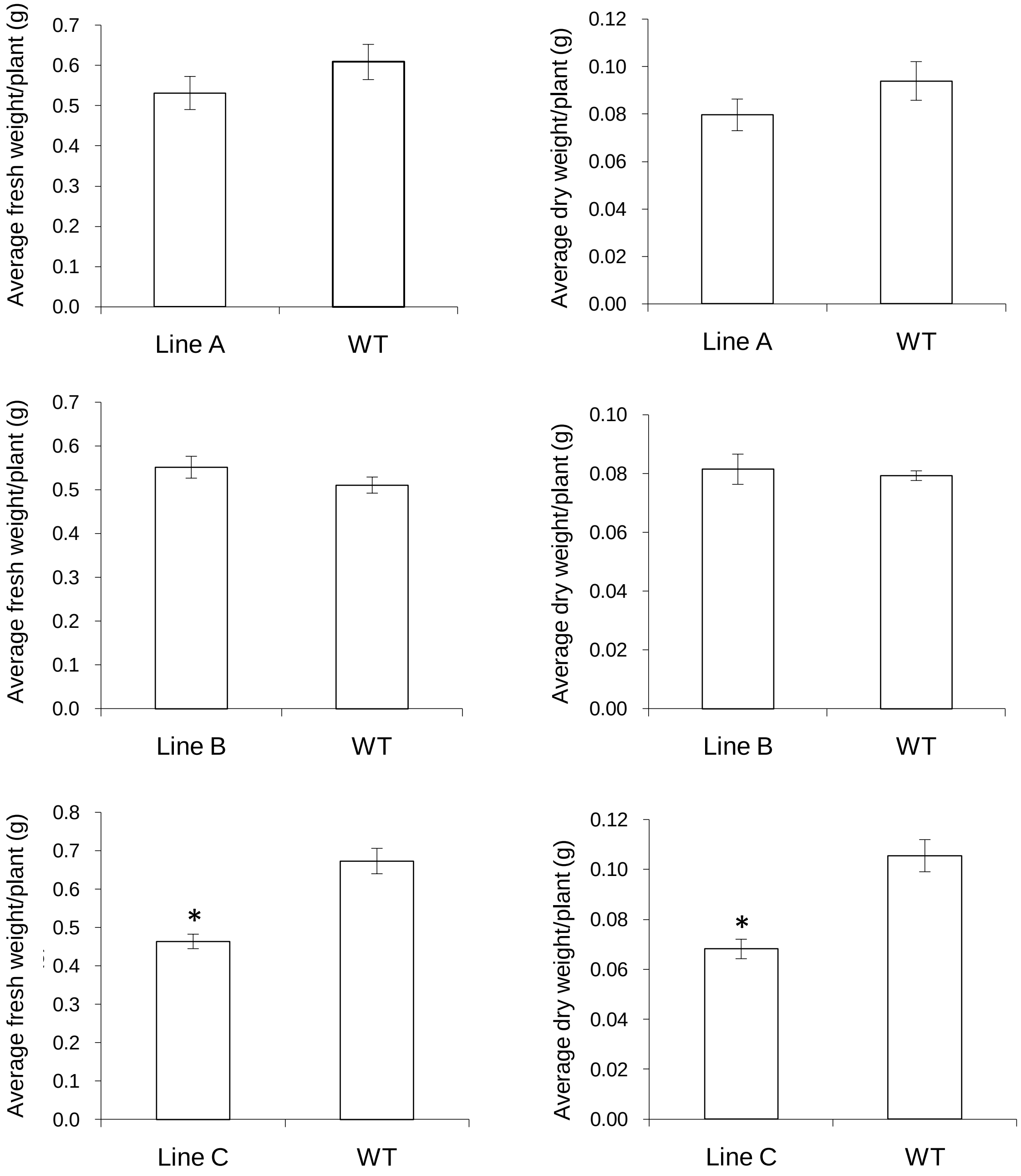


Fig. S3.7. Fresh and dry inflorescence stem biomass of eight-week-old wild type (WT) and T4 SND2-OV lines A, B, and C. Error bars indicate standard error of the mean of three biological replicates (stems from approximately nine plants were pooled per replicate). *Significantly different from the wild type according to homoscedastic two-tailed Student's t-test (*P* < 0.001). For lines A and C, similar results were obtained in an independent trial (data not shown).

© University of Pretoria

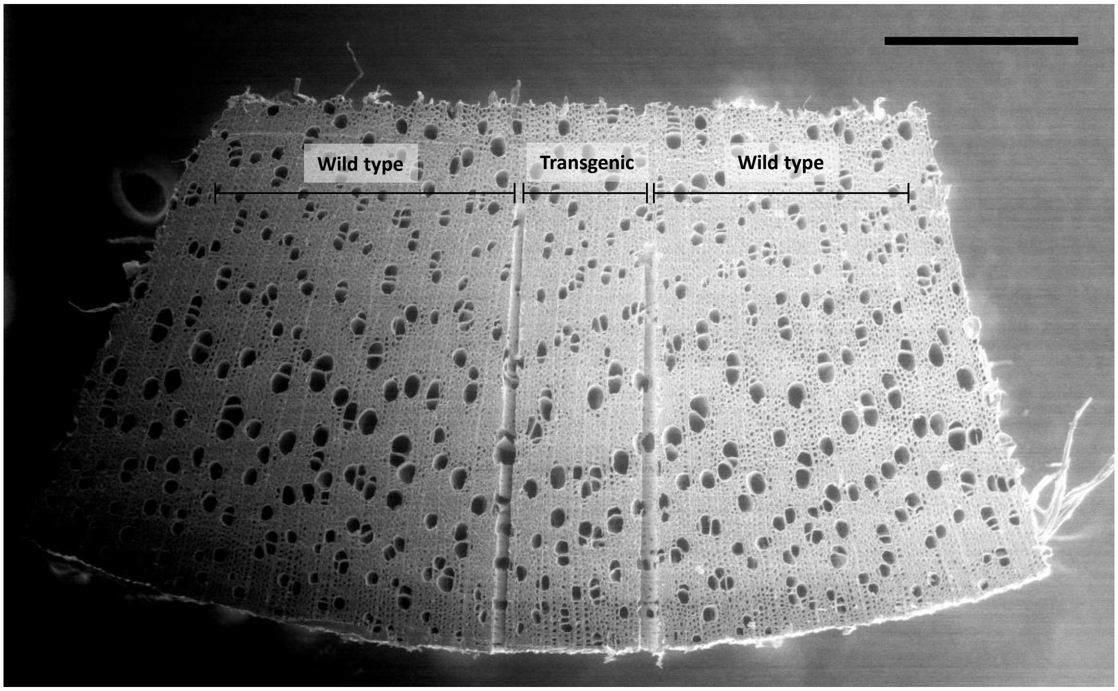


Fig. S3.8. Scanning electron micrograph of a hybrid *Eucalyptus* induced somatic sector. The transgenic sector has been marked by etching the sample either side of the GUS reporter stain (not visible). Scale bar = 500 µm.

© University of Pretoria

3.12. Supplementary tables

Table S3.1. χ2 analysis of SND2-OV T3 parent lines. Each χ^2 test is based on the number of T4 seedlings surviving germination on hygromycin selection medium. The null hypothesis (H_0) is that each parent line from which the T4 seeds were collected is hemizygous for the transgene. A significance threshold of 0.01 was adopted. Lines A1, B3 and E4 were used for all experiments.

SND2-OV T4 line	**Survival rate**	**χ^2 value**	**H_0 (hemizygous)**	**Deduced genotype**
A1	71/72	21.72	Reject	Homozygous
A3	70/77	10.40	Reject	Homozygous
B1	71/76	13.75	Reject	Homozygous
B3	89/93	21.25	Reject	Homozygous
E1	76/85	9.41	Reject	Homozygous
E3	56/67	2.63	Don't reject	Hemizygous
E4	86/88	24.24	Reject	Homozygous
Wild type[a]	109/110	34.05	Reject	Homozygous

[a] The wild type was grown on hygromycin-free medium. The test was used as a positive control for germination success.

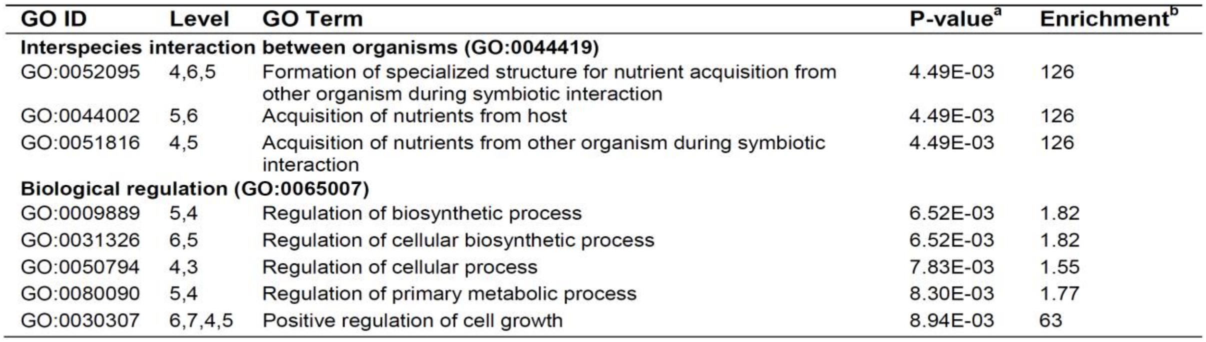
© University of Pretoria

Table S3.2. Enriched biological processes associated with genes differentially expressed in stems of eight-week-old SND2-OV(A) plants, relative to wild type. Terms exclusive to category levels higher than 6 or below level 4 were excluded for simplicity.

GO ID	Level	GO Term	P-value[a]	Enrichment[b]
Cell wall organization or biogenesis (GO:0071554)				
GO:0009834	5	Secondary cell wall biogenesis	3.25E-08	52.33
GO:0009832	4	Plant-type cell wall biogenesis	6.23E-06	19.63
GO:0009664	5	Plant-type cell wall organization	4.52E-03	9
GO:0010382	5	Cellular cell wall macromolecule metabolic process	8.94E-03	63
Carbohydrate metabolic process (GO:0005975)				
GO:0044042	6	Glucan metabolic process	1.51E-05	11.42
GO:0044264	5,6	Cellular polysaccharide metabolic process	5.91E-05	8.77
GO:0005976	5	Polysaccharide metabolic process	9.55E-05	8.02
GO:0044262	4,5	Cellular carbohydrate metabolic process	3.93E-04	3.96
GO:0000271	5,6	Polysaccharide biosynthetic process	2.17E-03	7.2
GO:0034637	5,6	Cellular carbohydrate biosynthetic process	2.60E-03	5.15
GO:0016051	4,5	Carbohydrate biosynthetic process	7.81E-03	3.88
Signalling process (GO:0023046)				
GO:0000160	6,5	Two-component signal transduction system (phosphorelay)	3.54E-05	13.65
GO:0007242	6,5	Intracellular signalling cascade	7.05E-03	2.7
Response to stimulus (GO:0050896)				
GO:0051707	3,4	Response to other organism	1.84E-10	6.78
GO:0009620	4,5	Response to fungus	7.08E-08	12.04
GO:0009611	4	Response to wounding	3.76E-07	11.98
GO:0006952	4	Defence response	1.99E-06	4.04
GO:0009617	4,5	Response to bacterium	3.46E-06	7.45
GO:0009725	4	Response to hormone stimulus	4.22E-06	4.03
GO:0009723	5	Response to ethylene stimulus	5.34E-06	10.23
GO:0009751	4	Response to salicylic acid stimulus	5.81E-06	10.23
GO:0050832	5,6	Defence response to fungus	8.69E-06	12.57
GO:0009816	6,5,7	Defence response to bacterium, incompatible interaction	1.59E-05	25.2
GO:0010200	6	Response to chitin	2.28E-05	10.47
GO:0042742	5,6	Defence response to bacterium	2.48E-05	8.15
GO:0009814	5,4,6	Defence response, incompatible interaction	3.20E-05	9.92
GO:0009743	5	Response to carbohydrate stimulus	2.71E-04	6.61
GO:0010033	4	Response to organic substance	3.01E-04	6.5
GO:0009409	5,4	Response to cold	7.00E-04	5.54
GO:0009753	4	Response to jasmonic acid stimulus	7.34E-04	6.98
GO:0032870	4,5	Cellular response to hormone stimulus	9.63E-04	5.16
GO:0009755	7,6,5	Hormone-mediated signalling	9.63E-04	5.16
GO:0009733	5	Response to auxin stimulus	2.09E-03	4.38
GO:0009612	4	Response to mechanical stimulus	2.26E-03	25.2
GO:0045087	4,5	Innate immune response	3.58E-03	3.93
GO:0016046	5,6	Detection of fungus	4.49E-03	126
GO:0009266	4	Response to temperature stimulus	5.10E-03	3.63
GO:0009416	5	Response to light stimulus	5.33E-03	3.14
GO:0009314	4	Response to radiation	6.16E-03	3.06
GO:0006979	4	Response to oxidative stress	6.37E-03	4.08
Nitrogen compound metabolic process (GO:0006807)				
GO:0046209	4	Nitric oxide metabolic process	1.19E-04	126
GO:0006809	6,5	Nitric oxide biosynthetic process	1.19E-04	126
GO:0042128	5	Nitrate assimilation	8.70E-04	42
GO:0042126	4	Nitrate metabolic process	1.06E-03	42
Transport (GO:0006810)				
GO:0006869	4,5	Lipid transport	2.01E-03	5.41

© University of Pretoria

Table S3.2. *(continued)*



GO ID	Level	GO Term	P-value[a]	Enrichment[b]
Interspecies interaction between organisms (GO:0044419)				
GO:0052095	4,6,5	Formation of specialized structure for nutrient acquisition from other organism during symbiotic interaction	4.49E-03	126
GO:0044002	5,6	Acquisition of nutrients from host	4.49E-03	126
GO:0051816	4,5	Acquisition of nutrients from other organism during symbiotic interaction	4.49E-03	126
Biological regulation (GO:0065007)				
GO:0009889	5,4	Regulation of biosynthetic process	6.52E-03	1.82
GO:0031326	6,5	Regulation of cellular biosynthetic process	6.52E-03	1.82
GO:0050794	4,3	Regulation of cellular process	7.83E-03	1.55
GO:0080090	5,4	Regulation of primary metabolic process	8.30E-03	1.77
GO:0030307	6,7,4,5	Positive regulation of cell growth	8.94E-03	63

[a] Benjamini and Hochberg correction

[b] Fold enrichment of each GO term is defined as the proportion of genes in the microarray dataset annotated by a GO term, relative to the genome-wide annotation of the GO term.

© University of Pretoria

Table S3.3. Primer sequences used for RT-qPCR analysis.

	Gene target	TAIR Locus	Primer sequence (5' → 3')
Target genes			
	SND2 (endogenous)	AT4G28500	Forward: TGATGAAGTTGTGAGCACTGAA Reverse: TGACAAGAGACCGGAAGTGA
	SND2 (total)	N/A	Forward: TCACTTCCGGTCTCTTGTCA Reverse: TGCGTCATCTCTTACCTTGC
	MYB103	AT1G63910	Forward: GTCGTCATCAACCGTCAGTA Reverse: TCGATGTTGTGGTGGTAGAG
	COBL4	AT5G15630	Forward: TAGAGTCCACTGGCACGTTA Reverse: CCTGAAGGTCCAGCTTCCAT
	CesA4	AT5G44030	Forward: TTGGTGTTGTTGCCGGAGTT Reverse: AACAGTCGACGCCACATTGC
	CesA7	AT5G17420	Forward: CGTTGTTGCAGGCATCTCAG Reverse: AGCAGTTGATGCCACACTTG
	CesA8	AT4G18780	Forward: CCGCAATCTTCATCATCGTC Reverse: CCGCCATTCTCCATAAGAGT
	CslA09	AT5G03760	Forward: ACACCAAGGTCATTGCATCT Reverse: TACACCGAGTTCCAACACAT
	EXPA15	AT2G03090	Forward: GTCCTCCTAACAACGCTCTT Reverse: CGCAACCGAATGAACATCTC
	FLA12	AT5G60490	Forward: ATGTCTACAGCGATGGACAG Reverse: CCATGCGAGCATTACACTCA
	AGP21	AT1G55330	Forward: ATGGAGGCAATGAAGATGAAG Reverse: AACATGGCAGCATCAGAAGTT
	AT1G20120	AT1G20120	Forward: ACCAGTTGTACCGGCATATT Reverse: TTGACATCGGTATCGCACTT
	AT5G11410	AT5G11410	Forward: CACGGGTCAAGATAGCCATA Reverse: GTGTTGTAGTGCTCGTCAAG
	VSP1	AT5G24780	Forward: AGTCCGGAGAATCAACTCCA Reverse: GTACACCACTTGCGTCAACT
	AT3G01345	AT3G01345	Forward: AATGGACGCCTTGCTATCAG Reverse: AGGCTTCGGTAACACCTACT
	FLA11	AT5G03170	Forward: GTGGCGATGATGGAGGAGAT Reverse: CAATGGCTGCAACGGTAGTG
	CTL2	AT3G16920	Forward: CTGCAACAGCGGATTCGATA Reverse: AGTCACCGAACCAGAGGTTA
Control genes			
	ACT2	AT3G18780	Forward: TGGAATCCACGAGACAACCT Reverse: TGGACCTGCCTCATCATACT
	EF1α	AT1G07920	Forward: ACAGGCGTTCTGGTAAGGAG Reverse: CCTTCTTGACGGCAGCCTTG

© University of Pretoria

3.13. Additional files

The following additional datasets are available on the supplementary CD-ROM disk attached to this thesis:

Additional file 3.1.xls Microarray data of SND2-OV(A) vs. wild type (8 weeks), fold change > |±1.5|. List of significantly differentially expressed genes of SND2-OV line A stems at 8 weeks, compared to the wild type, with fold change values larger than 1.5.

Additional file 3.2.xls Microarray data SND2-OV(A) vs. wild type (4 weeks). List of significantly differentially expressed genes of SND2-OV line A stems at 4 weeks, compared to the wild type.

© University of Pretoria

Chapter 4

Genome-wide mapping of histone H3K4 trimethylation in *Eucalyptus grandis* developing xylem using nano-ChIP-seq

Steven G. Hussey[1], Eshchar Mizrachi[1], Andrew Groover[2,3], Dave K. Berger[4], Alexander A. Myburg[1]

[1]Department of Genetics, [4]Department of Plant Science, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, South Africa

[2]US Forest Service, Pacific Southwest Research Station, Davis CA, USA

[3]Department of Plant Biology, University of California Davis, USA

This chapter has been written in manuscript format for submission to a peer-reviewed scientific journal. I conceived of the study, performed all the experimental work, and drafted the manuscript. Andrew Groover provided student training in ChIP-seq. Eshchar Mizrachi, Dave Berger and Alexander Myburg reviewed and edited the manuscript. We intend to further condense this chapter into a research paper for submission to *BMC Plant Biology* in May 2014.

© University of Pretoria

4.1. Summary

Histone modifications play an integral role in eukaryotic gene expression, but have been poorly studied in woody plants. While high-throughput chromatin immunoprecipitation techniques are ideal for studying genome-wide histone modifications *in vivo*, their application to new tissues such as developing secondary xylem requires extensive optimization. We aimed to understand the role of the modified histone H3K4me3 (trimethylated lysine 4 of histone H3) in wood formation in *Eucalyptus grandis* trees. Existing plant chromatin fixation and isolation protocols were optimized for direct fixation of nuclei from frozen, macerated developing xylem tissue collected from field-growing trees. A "nano-ChIP-seq" procedure was employed for ChIP DNA amplification. Over 9 million H3K4me3 ChIP-seq and 18 million control paired-end reads were mapped to the *E. grandis* reference genome for peak-calling using Model-based Analysis of ChIP-Seq. The 11,773 significant H3K4me3 peaks identified covered ~3% of the genome and overlapped some 9,760 genes and 30 noncoding RNAs. H3K4me3 library coverage, peaking ~600-700 bp downstream of the transcription start site, was highly correlated with gene expression levels. H3K4me3-enriched genes exhibited relatively low tissue specificity and were overrepresented for general cellular metabolism and development. However, many secondary cell wall-related genes with preferential expression in developing secondary xylem were positive for H3K4me3 as validated using ChIP-qPCR. In this first genome-wide analysis of a modified histone in a woody tissue, we have developed a ChIP-seq procedure suitable for frozen, field-collected developing xylem samples. *E. grandis* H3K4me3 profiles are consistent with known H3K4me3 functions in *Arabidopsis* and rice, while we show that this epigenetic mark is also associated with tightly regulated secondary cell wall biosynthetic genes. The H3K4me3 ChIP-seq data

© University of Pretoria

from this study complements RNA-seq evidence of gene expression for the future improvement of the *E. grandis* genome annotation.

4.2. Introduction

A wealth of histone modifications affect chromatin structure and/or gene activation and repression in eukaryotes (reviewed by Barrera & Ren, 2006; Kouzarides, 2007). Chromatin organization plays a crucial role in plant gene regulation, employing conserved as well as unique mechanisms to those of other eukaryotes (Pfluger & Wagner, 2007). In mammals, as well as plants (Deal & Henikoff, 2011), the presence of activating histones such as trimethylated lysine 4 of histone H3 (H3K4me3) and acetylated lysine 9 (H3K9Ac) at the transcription start site (TSS) are good predictors of gene expression (The ENCODE Project Consortium, 2012). For example, the degree of H3K4 trimethylation at the TSS is directly proportional to the transcript expression level (Barski *et al.*, 2007; Heintzman *et al.*, 2007). In mammals, monomethylated H3K4 (H3K4me1) is preferentially associated with enhancer elements, while dimethylated H3K4 (H3K4me2) is associated with both enhancers and promoters, as well as with "poised" genes that are expressed at defined developmental stages or in specific cell types (Heintzman *et al.*, 2007; Zhou *et al.*, 2011). H3K36 methylation, in contrast, is thought to mediate RNA polymerase II (Pol II) elongation and act as docking sites for transcript-processing enzymes (reviewed by Hampsey & Reinberg, 2003). In general, plants have a similar histone code to that of mammals, with some exceptions such as a higher abundance of H3K4me2 (reviewed by Liu *et al.*, 2010).

Lysine 4 of histone H3 is trimethylated by SET1 of the *Trithorax* protein complex COMPASS in yeast (Nagy *et al.*, 2002), with ATXR3 and to some extent ATX1 performing this function in *Arabidopsis* (Alvarez-Venegas *et al.*, 2003; Alvarez-Venegas

© University of Pretoria

& Avramova, 2005; Saleh *et al.*, 2008c; Berr *et al.*, 2011). In yeast, H3K4 trimethylation is predicated on Rad6-mediated ubiquitination of lysine 123 of H2B (uH2B-K123), a histone modification that is required for H3K4 methylation in gene regions (Dover *et al.*, 2002; Sun & Allis, 2002). The uH2B-K123 modification is critical for H3K4 methylation by SET1, possibly acting to open the chromatin structure for SET1 targeting (Sun & Allis, 2002). SET1 associates with the activated form of Pol II, in part through the PAF1 complex, ensuring that H2B ubiquitination and H3K4 methylation occur proximal to the pre-initiation complex (reviewed by Wood & Shilatifard, 2006). Thus, H3K4me3 appears to be established by active transcription itself, is reported to occur at over 90% of Pol II-enriched sites in human (Barski *et al.*, 2007) and is associated with transcription initiation but not necessarily transcription elongation in mammals (Guenther *et al.*, 2007). Since the H3K4me3 modification endures at previously active genes for up to several hours after silencing in yeast, it represents evidence of both active and recent transcription (Ng *et al.*, 2003). H3K4 methylation can, however, be dynamically reversed by histone demethylases (Shi & Whetstine, 2007; Liu *et al.*, 2010). The function of H3K4me3 is to recruit TFIID to active promoters and assisting in pre-initiation complex formation, which is enhanced in the presence of a TATA box (Lauberth *et al.*, 2013), via interaction with the TAF3 subunit (Vermeulen *et al.*, 2007; Ingen *et al.*, 2008). A number of other proteins are known to bind to H3K4me3 at specific loci, which are in turn tethered to, or recruit, enzymes that manipulate the local chromatin structure (Kouzarides, 2007).

At human TSSs, “open” chromatin regions that are hypersensitive to DNase I cleavage are followed by a prominent H3K4me3 signal immediately downstream; a relationship so strong that the pattern can be used to annotate TSSs and the direction of transcription (Thurman *et al.*, 2012). In plants, H3K4me3 histone modifications occur

© University of Pretoria

almost exclusively in genes and their promoters but preferentially occupy genic regions 250-600 bp (*Arabidopsis*) or 500-1000 bp (*Oryza*) downstream of the TSS (Li *et al.*, 2008; Zhang *et al.*, 2009; van Dijk *et al.*, 2010; Ha *et al.*, 2011). Genes occupied by H3K4me3, especially in the absence of H3K4me1 and H3K4me2, generally display low tissue specificity but high levels of constitutive expression in *Arabidopsis* (Zhang *et al.*, 2009; Ha *et al.*, 2011). However in two drought studies, H3K4me3 occupancy became broader in genes differentially expressed during drought stress in *Arabidopsis* (van Dijk *et al.*, 2010), and increased for a proportion of genes differentially expressed during drought stress in rice (Zong *et al.*, 2013), suggesting H3K4me3 can also be associated with tightly regulated pathways.

Despite their importance in growth and development, modified histones have been poorly studied in woody tissues. Developing secondary xylem (DSX) poses several challenges to obtaining crosslinked chromatin for chromatin immunoprecipitation (ChIP) studies. First, unlike herbaceous plant organs, it is impractical to fix intact DSX tissue of mature field-grown trees, a process that is normally performed by the submersion of entire organs in fixation buffer under vacuum. Second, fixation of freshly scraped DSX carries the risk of nuclei loss and proteolysis from proteases released during tissue scraping. Third, the fact that fibers are large, elongated cells, generally mononucleate in secondary xylem and possess elongated nuclei (Snegireva *et al.*, 2010; Gorshkova *et al.*, 2012) may pose a challenge to purifying nuclei in sufficient yields. Fourth, although slower when compared to vessel elements, fibers undergo programmed cell death within a narrow range of the cambium (650 - 1000 μm in poplar) (Courtois-Moreau *et al.*, 2009), limiting the sampling depth of live target cells. Finally, the presence of large quantities of secondary cell wall material in fiber and vessel cells may pose challenges to nuclei isolation.

© University of Pretoria

Here, we aimed to determine the role of the activating histone modification H3K4me3 in the epigenetic regulation of wood formation (xylogenesis) in field-growing *Eucalyptus grandis* trees. We hypothesized that H3K4me3 signals marking Pol II-transcribed genes, including those involved in wood formation, could accurately predict their corresponding transcript levels in developing xylem. We assessed and optimized existing protocols for the isolation of crosslinked chromatin from frozen field-collected tissue for use in ChIP-seq assays, and modified a nano-ChIP-seq protocol for the amplification of ChIP DNA. To the best of our knowledge, this is the first genome-wide study of the role of a modified histone in developing wood.

4.3. Materials and methods

4.3.1. Plant materials

ChIP-seq experiments were performed on *E. grandis* clone SA1 (Mondi Tree Improvement Research, Hilton, South Africa). DSX scrapings from seven-year-old ramets growing in a plantation in KwaMbonambi, KwaZulu-Natal Province, South Africa were sampled in September 2012. The bark was peeled off at breast height to expose the DSX tissue of two individuals, V5 and V11. 1-2 mm was lightly and uniformly scraped off using a razor, gently squeezed of excess sap and immediately flash-frozen in liquid nitrogen. Samples were stored at -80^{o}C until use.

4.3.2. Chromatin fixation, isolation and sonication

Nuclei were purified as described by Kaufmann *et al.* (2010), with modifications. Frozen DSX tissue was ground using a model A 11 B basic analytical mill (IKA, Germany) followed by fine grinding in liquid nitrogen using a mortar and pestle. Every five grams of frozen, ground DSX tissue was fixed in 25 ml M1 buffer supplemented with 1%

© University of Pretoria

formaldehyde, 1 mM EDTA and 1 mM phenylmethanesulfonyl fluoride (PMSF) on ice for 30 min. Fixation was quenched with 1/10 volume 1.25 M glycine for 5 min on ice, followed by addition of M1 buffer without formaldehyde to 50 ml. The suspension was filtered through 60 μm nylon mesh wetted with M1 buffer, changing the filter at least once per 50 ml suspension, and again through a double 60 μm nylon mesh. After centrifugation at 1,000 x *g* for 20 min (4^{o}C), the pellet was resuspended in 25 ml ice-cold M2 buffer supplemented with 1 mM PMSF and Complete Protease Inhibitor cocktail (CPIC; Roche), centrifuged at 1,000 x *g* for 10 min at 4^{o}C and resuspended in 25 ml ice-cold M3 buffer supplemented with 1 mM PMSF and CPIC. After centrifugation similarly for 10 min, the nuclear pellet was resuspended in ~1.5 ml sonic buffer supplemented with 1 mM PMSF and CPIC. Sonication was performed on 250 µl crude chromatin per 1.5 ml tube on ice using a Branson Sonifier 450 probe sonicator with 20 pulses of 10s duration on setting 1, and >30s rest on ice between pulses. Samples were mixed every ten cycles. After sonication, samples were centrifuged twice at 16,000 x *g* (10 min, 4^{o}C) and stored at -80^{o}C.

4.3.3. Micrococcal nuclease (S7) assay

Frozen DSX tissue (2 g) was ground fine in liquid nitrogen. Nuclei were isolated as described above, excluding formaldehyde crosslinking and the addition of sonic buffer. The crude nuclear pellet was resuspended in 350 µl nuclei digestion buffer (Zhao *et al.*, 2001) containing 400 µg RNase A. Samples were divided equally into four tubes and incubated with 0, 5, 10 or 20U of Nuclease S7 (Roche) at 37^{o}C for 15 min. Hydrolysis was terminated with 5 mM EDTA. Nuclei were lysed with 0.5% SDS and centrifuged (20,000 x *g*, 5 min) to clear. Soluble DNA was purified using the MN Nucleospin PCR purification kit.

© University of Pretoria

4.3.4. Fixation optimization

Ground, frozen DSX tissue (1 g) was fixed in 5 ml M1 buffer (Kaufmann *et al.*, 2010) containing 1% formaldehyde for 5, 15, 30, 45 or 60 min on ice. No formaldehyde was added to the control (0 min) sample. Samples were quenched with 500 µl 1.25M glycine for 5 min and M1 buffer was added to 12.5 ml. Chromatin was prepared as described above. Twenty rounds of sonication were performed in a 300 µl volume. To half of each sample, 500 µg/ml proteinase K was added and incubated at 37°C overnight followed by crosslink reversal at 65°C for ≥ 7 hrs. DNA was extracted with the DNeasy extraction kit (Qiagen, Limburg) and quantified with a Qubit HS dsDNA kit (Invitrogen, OR).

4.3.5. Protein extraction and Western blot analysis

Nuclei were purified according to the method of Kaufmann *et al.* (2010), with modifications. 11.5 g DSX was ground in liquid nitrogen and suspended in M1 buffer containing 1 mM PMSF and 1 mM EDTA at 5 ml per gram of tissue, for 30 min. The suspension was filtered twice through 60 µm nylon mesh and pelletized at 1000 x *g* (20 min, 4°C). The pellet was resuspended in 5 ml M2 buffer supplemented with 1 mM PSMF and CPIC, re-pelletized (1000 x *g*, 10 min, 4°C) and resuspended in 250 ul M3 buffer containing 1.7 M sucrose and CPIC. The suspension was overlaid on 1.5 ml 1.7 M sucrose in M3 buffer and centrifuged for 40 min at 16,000 x *g* (4°C). The pellet was resuspended in 1 ml M3 to wash, re-pelletized (12,000 x *g*, 5 min, 4°C) and the remaining pellet resuspended in 1 pellet volume of extraction buffer (10 mM sodium phosphate buffer pH 7.0, 150 mM NaCl, 0.1 mM EDTA, 5% glycerol, 10 mM β-mercaptoethanol, 0.1 mM PMSF, CPIC). The pellet was briefly sonicated with a Branson 450 sonicator (30s, 10% power output) and gently vortexed for 30 min at 4°C. Soluble protein in the supernatant from two rounds of centrifugation (16 000 x *g*, 10 min, 4°C) was quantified using the

© University of Pretoria

Qubit Protein Assay Kit (Invitrogen, OR), subjected to denaturing electrophoresis on a 12% SDS-PAGE gel and transferred to a nitrocellulose membrane using the semidry method. Blots were blocked with 5% nonfat milk, probed with 1:2000 dilution of anti-H3K4me3 antibody (Millipore #07-473) overnight (4°C) and incubated with horseradish peroxidase-conjugated goat anti-rabbit secondary antibody (Cappel Laboratories Inc., PA). Blots were treated with SuperSignal West Pico Chemiluminescent substrate (Thermo Scientific, Rockford, IL) and developed with CL-XPosure film (Thermo Scientific).

4.3.6. Chromatin immunoprecipitation, DNA amplification and sequencing

A minimum of 3 µg *E. grandis* DSX chromatin was incubated with 1 µg anti-H3K4me3 antibody (Millipore #07-473), or 1 µg naïve mouse IgG$_{2a}$ (sc-3878, Santa Cruz Biotechnology, CA) as negative control, overnight at 4°C. Chromatin immunoprecipitation was performed as described by Adli & Bernstein (2011) using 40 µl protein A-agarose beads, 25% slurry (sc-2001, Santa Cruz Biotechnology, CA). After crosslink reversal and DNA purification, the ChIP DNA was quantified with the Qubit HS dsDNA kit (Invitrogen, OR). A minimum of 1 ng ChIP or input DNA was amplified according to the protocol of Adli & Bernstein (2011), with modifications. We replaced the use of Sequenase v.2.0 DNA polymerase (Affymetrix, CA) with *Bsu* DNA polymerase, large fragment (NEB, MA), and substituted the corresponding Sequenase reaction buffer with NEB Buffer 2 (since this buffer already contains dithiothreitol, no additional dithiothreitol was added). We used 2 U of *Bsu* DNA polymerase per pre-amplification cycle, extended the pre-amplification extension time to 20 min and used 32 pmol P1 primer. Both the pre-amplification and PCR reactions were supplemented with 50 ng/µl tRNA. We applied a generic ExoSAP cocktail by adding 0.5 U rAPID alkaline phosphatase (Roche Applied Science, Ltd) and 5 U *E. coli* Exonuclease I (NEB), incubating at 37°C for 30 min and

© University of Pretoria

heat-inactivating the enzymes at 80°C for 20 min. For the Phusion PCR reactions, where five reactions were initiated from each pre-amplification sample, we used 4 ul 10 mM dNTPs and 0.5 ul Phusion DNA polymerase per 50 ul reaction. PCR extension time was reduced to 5 s. 20 ng template was used for Illumina library preparation and DNA sequencing (Beijing Genome Institute, Hong Kong), generating 50 nt paired-end sequences.

4.3.7. Bioinformatic analysis

Sequence data were trimmed of primer and adapter sequences and purged of low-quality reads (phred score <20 for ≥50% of the read, or reads with >10% "N" bases). In some cases further trimming of the 5' end was required to reduce overrepresented k-mers as identified using FastQC (http://www.bioinformatics.babraham.ac.uk/). The reads were mapped to the *E. grandis* v.1.1 reference genome (www.phytozome.net/Eucalyptus.php) using Bowtie2 (Langmead & Salzberg, 2012) in Galaxy (Giardine *et al.*, 2005; Goecks *et al.*, 2010), using parameters: "sensitive" preset option, 50-1,000 bp insert size for a valid PE pair, end-to-end alignment. BAM alignments of ChIP-seq and input libraries were converted to BED format and subjected to peak-calling analysis using MACS v.1 (Zhang *et al.*, 2008) in Nebula (Boeva *et al.*, 2012), with input libraries as controls and where paired-end reads were treated as single reads, duplicate reads were discarded, effective genome size was 640 Mb, *P*-value set at 10^{-5}, band width set to 300 bp and peak-calling model based on 10 - 30-fold enrichment. Significant H3K4me3 peaks were identified as those with a *P*-value < 10^{-5}, fold-enrichment ≥ 5, represented by at least 10 tags in both individuals, and having at least 100 bp overlap between peaks called in both individuals. To exclude false positive signals, 132 H3K4me3 peaks that overlapped the 732 peaks called in the IgG_{2a} ChIP-seq negative control by the same criteria were removed from all

© University of Pretoria

analyses. H3K4me3-enriched genes were defined as those overlapping a significant H3K4me3 peak by at least one base. Strand cross-correlation analysis was performed using SPP (Kharchenko *et al.*, 2008). Shannon entropy (Shannon, 1948) was calculated for each gene as described by Schug *et al.* (2005), using previously obtained RNA-seq data (http://eucgenie.bi.up.ac.za/; Hefer *et al.* in preparation). Genes were considered expressed if they had an FPKM value above 70. Tag density distributions across genomic coordinates were calculated using BEDTools (Quinlan & Hall, 2010) based on bulked BAM files of both individuals. For Gene Ontology analysis, the nearest *Arabidopsis* BLASTP hits of H3K4me3-enriched or DSX-expressed genes were analyzed for Biological Process enrichment analysis (Bonferroni-adjusted *P*-value < 0.01) using GOToolBox (Martin *et al.*, 2004). Significantly over- or underrepresented GO terms and their *P*-values were imputed into REVIGO (Supek *et al.*, 2011) for GO term summarization.

4.3.8. Quantitative polymerase chain reaction (qPCR)

The V11 ChIP-seq samples (prior to library preparation) were used for ChIP-qPCR analysis. Primers targeting selected genomic regions are listed in Table S4.1. Sample concentrations were quantified with the Qubit HS dsDNA kit (Invitrogen, OR) and equal quantities of input, H3K4me3 ChIP, and mock ChIP (i.e. IgG_{2a}) DNA added to triplicate reactions for qPCR quantification and melting curve analysis using the LightCycler 480 [50 cycles of 95°C denaturation (10s), 60°C annealing (10s) and 72°C extension (15s)] (Roche, Switzerland). Crossing points (Cp) were calculated using the second derivative maximum method and quantities relative to the input sample calculated using the formula $E^{\Delta Cp(input)\ -\ Cp(sample)}$, where E is the efficiency calculated from a standard curve of the relevant primer set.

© University of Pretoria

4.4. Results

4.4.1. Optimization of chromatin isolation from *E. grandis*

We regarded the isolation of a maximum quantity of chromatin from DSX tissue as a priority for ChIP-seq. We assessed DNA yield from the nuclear pellet using nuclei isolation buffers described by Kaufmann *et al*. (2010), McKeown *et al.* (2008), Saleh *et al.* (2008b) and a "woody plant buffer" (Loureiro *et al.*, 2007). We found that M1 buffer (Kaufmann *et al.*, 2010) facilitated the highest yield of nuclear DNA (Fig. S4.1a). The quality of DNA isolated using each buffer was similar (Fig. S4.1b). Based on this result, we used the protocol by Kaufmann *et al*. (2010) for nuclei isolations, with modifications as described in Methods. Increasing the buffer volume-to-tissue mass ratio further enhanced DNA yield from crude nuclei (Fig. S4.2). We also found that grinding xylem to the point of complete homogenization of fiber bundles (< 100 μm) did not increase nuclear DNA yield compared to particles > 100 μm (Fig. S4.3). Coarse preparations had the additional advantage that less cellular debris was co-purified with the nuclei, allowing for a reduced volume of sonication buffer to be added to the final nuclear pellet and hence more concentrated chromatin.

The ability to isolate intact chromatin is essential for the successful application of ChIP-seq. We investigated whether intact chromatin could be isolated from *E. grandis* DSX as assessed by the persistence of histone-DNA associations. Micrococcal nuclease (MNase) was used to detect nucleosomes by virtue of its ability to cleave the linker region between nucleosomes while rendering DNA packaged within nucleosomes intact. Whereas naked genomic DNA was complete degraded in the presence of 10U MNase, DSX chromatin exposed to up to 20U MNase was hydrolysed into distinct nucleosomal fragments (Fig. 4.1). This pattern is consistent with successive cleavage of linker DNA

© University of Pretoria

between nucleosomes, liberating nucleosomal fragments in decreasing dividends of ~195 bp to ~140 bp (Philipps & Gigot, 1977) (Fig. 4.1). These results indicate that intact chromatin was successfully isolated from DSX tissue.

4.4.2. Optimization of sonication and chromatin crosslinking

In cases where formaldehyde crosslinking is used to stabilize protein-DNA interactions, the chromatin is generally fragmented using sonication prior to ChIP (Das *et al.*, 2004). We optimized sonication parameters for DSX chromatin to produce an average fragment size ranging from 150 – 600 bp (Adli & Bernstein, 2011). Favouring a low sonication output and large number of pulses rather than high output with few pulses (Haring *et al.*, 2007), we found that twenty pulses of 10s with a probe sonicator at 10% power output produced the desired fragment range with an average of 300 bp (Fig. S4.4a). Analysis of DNA in the residual pellet showed that this sonication treatment released most of the chromatin (Fig. S4.4b).

The degree of formaldehyde crosslinking is critical since insufficient crosslinking may result in loss of bound proteins and crosslink reversal during sonication, while an excess of crosslinking will compromise protein-antibody recognition and reduce DNA yield after crosslink reversal (Haring *et al.*, 2007). Generally, a concentration of 1% formaldehyde is applied for 5 - 60 min; the optimum must be empirically determined for each system (Orlando *et al.*, 1997). To optimize crosslinking conditions, we fixed frozen and ground DSX tissue from field-grown *E. grandis* trees for varying durations in 1% cold formaldehyde buffer, and purified nuclei. We sonicated chilled samples to assess the degree of crosslink reversal as a by-product of sonication. We found that 30 min fixation

© University of Pretoria

gave the best trade-off between durable crosslinking and DNA yield following crosslink reversal (Fig. S4.5).

From our optimized fixation and sonication conditions and modifications to the Kaufmann *et al*. protocol (2010) we were able to obtain up to 7.5 µg chromatin per gram frozen DSX tissue for ChIP applications. We found that frequent filter changes during filtration of the tissue suspension and resuspension of the nuclear pellet in an adequate amount of sonic buffer (see Methods) was essential for high yields of chromatin.

4.4.3. Application of a ChIP DNA amplification protocol

The amount of DNA recovered from a ChIP enrichment is small, often only 1 - 10 ng from 50 µg of chromatin (Orlando *et al.*, 1997). We generally obtained 1 - 2 ng ChIP DNA from the modified protocol by Kaufmann *et al.* (2010). While single-molecule sequencing platforms have been successfully used to sequence ChIP libraries of as little as 50 pg (Goren *et al.*, 2010), Illumina sequencing generally requires ~10 ng of ChIP DNA. In order to perform Illumina sequencing with enough ChIP DNA to spare for qPCR validation, we adopted a ChIP DNA amplification protocol that was successfully developed for ChIP-seq of limited mammalian cell numbers (Adli *et al.*, 2010; Adli & Bernstein, 2011). The method generates several daughter fragments of various sizes for each parent fragment in a ChIP DNA sample, and circumvents material loss through column purification and enzymatic end-repair of ChIP fragments prior to Illumina adapter ligation. In short, the method involves a pre-amplification step using random primers containing a self-complementary adapter, such that a hairpin structure is formed to prevent primers from self-annealing. The adapter also contains a BciVI restriction site. After four rounds of pre-amplification using a strand-displacing DNA polymerase, the resulting

© University of Pretoria

adaptor-flanked fragments are PCR-amplified for up to 15 cycles using a second primer complementary to the adapters. Fragments are then digested with the BciVI restriction endonuclease, yielding 3' adenine overhangs that are suitable for Illumina adapter ligation and library preparation.

Trial amplification of a modified version of the Adli & Bernstein (2011) protocol (see Methods) using 500 – 2000 pg sonicated *E. grandis* genomic DNA was successful, but we observed an increase in average fragment size (Fig. S4.6b). Reducing the PCR extension time from 20s to 5s reduced the average fragment size of the amplicons, while pre-amplification extension time did not appear to influence fragment size (Fig. S4.6c). Together, our modifications to the Adli and Bernstein (2011) protocol produced surplus amounts of amplified DNA (up to several hundred nanograms), beginning with 1 ng template sonicated to ~300 bp. Although the amplified DNA fragment length was generally ≥ 500 bp, this size distribution is similar to that obtained by the protocol developers (Adli *et al.*, 2010).

4.4.4. ChIP-seq analysis of H3K4me3 in *E. grandis* developing secondary xylem

We conducted ChIP-seq analysis of the activating histone mark H3K4me3 to evaluate our modified ChIP-seq protocol and to better understand the role of this signature in developing xylem gene expression. We selected a commercial antibody for H3K4me3 which had been used previously in ChIP analyses in *Arabidopsis* (Alvarez-Venegas & Avramova, 2005; Pien *et al.*, 2008; Saleh *et al.*, 2008a; Luo *et al.*, 2013). Antibody recognition of the H3Kme3 signature in *Eucalyptus* immature xylem was confirmed by Western blot analysis of DSX nuclear extracts. Two independent blots showed that the

© University of Pretoria

antibody recognized a ~17 kDa band, corresponding to the predicted molecular weight of H3K4me3 (Fig. S4.7).

We isolated chromatin from frozen DSX collected from two field-grown *E. grandis* individuals (clonal ramets). In trial experiments, different amounts of anti-H3K4me3 antibody produced similar enrichments of candidate regions as assessed by ChIP-qPCR (Fig. S4.8). We performed ChIP enrichment using 1 ug antibody and generated over 30 million 50-base paired-end reads from both the H3K4me3-enriched and input sample sets (Table S4.2). The sequences were trimmed to remove contaminating primer sequences and mapped to the v.1.1. annotation of the *E. grandis* reference genome (www.phytozome.net). For one individual (V11), we additionally sequenced an IgG_{2a} negative control library to remove false positive peaks due to nonspecific antibody or protein A binding (see Methods). Despite a high degree of sequence duplication (owing to ChIP DNA amplification), 3.7 - 11.7 million read pairs mapped uniquely for each H3K4me3 and input replicate (Table S4.2). Input library sequence depths exceeded those of the ChIP libraries, which tends to increase peak-calling specificity (Chen *et al.*, 2012). Strand cross-correlation analyses showed that all H3K4me3 ChIP libraries were enriched to an efficiency well within ENCODE guidelines (Landt *et al.*, 2012) (Fig. S4.9). After peak calling with MACS (Zhang *et al.*, 2008), we identified 13,175 H3K4me3 peaks in individual V5 and 18,005 peaks in individual V11. 11,773 significant H3K4me3 peaks were common to both individuals sampled (Additional file 4.1), overlapping with some 9,760 genes (Additional file 4.2). Subsampling of various proportions of the mapped tags showed that the number of peaks called began to plateau (Fig. S4.10), suggesting that most of the H3K4me3 peaks had been detected at the reported sequencing depth. The peaks, which spanned a median interval of 1,534 bp (Fig. S4.11), covered 19.1 Mb of the

© University of Pretoria

assembled genome, ~85% of which overlapped annotated gene models and/or promoter regions within 1kb upstream of the TSS. Of the 1,905 peaks that did not overlap a gene model in the v.1.1. genome annotation, a further 234 overlapped some 235 low-confidence gene annotations that were removed from the first annotation (i.e. v.1.0), suggesting that some of these are *bona fide* gene models (Additional file 4.3).

On average, 42% of a given peak interval, defined here as the genomic span of a significant peak, overlapped intronic sequence within transcribed regions, and 25% overlapped exon sequence (Fig. 4.2). In intergenic regions, 9% of most peak intervals overlapped 1kb promoter regions of genes (Fig. 4.2). We also assessed the H3K4me3 enrichment of known and predicted noncoding RNA (ncRNA) elements in the *E. grandis* genome (Myburg *et al.*, in press). Disregarding ambiguous H3K4me3 peaks that overlapped with both ncRNAs and genes, ~13% of small nucleolar RNAs (snoRNAs) and ~4% of known or predicted microRNAs (miRNAs) were enriched for H3K4me3 whereas transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), small nuclear RNAs (snRNAs), antisense RNAs and small RNAs (sRNAs) showed little or no enrichment (Table 4.1). The enriched snoRNAs appeared to consist of at least 5 polycistronic clusters (not shown), a common arrangement in plants (Brown *et al.*, 2001). These data are consistent with the fact that miRNAs and many snoRNAs are transcribed by Pol II and might hence be expected to exhibit H3K4me3 modifications when expressed (Chen, 2005; Rodor *et al.*, 2010).

We assessed the binding profile of H3K4me3 relative to genic regions by calculating per-base coverage of H3K4me3 and input libraries across all annotated genes, as well as the upstream and downstream sequences, in a bin-wise manner. As expected, H3K4me3-

© University of Pretoria

enriched library coverage peaked shortly after the TSS (Fig. 4.3a). In contrast, input coverage was comparatively uniform across transcribed regions and their flanking non-coding sequences (Fig. 4.3a). Similarly, when absolute distance relative to the TSS or TTS (transcription termination site) was analyzed for H3K4me3 and input coverage across genes, the H3K4me3 profile yielded a prominent peak ~600-700 bp downstream of the TSS (Fig. 4.3b). The position of the peak was similar for genes of different lengths (Fig. S4.12)

4.4.5. Expression dynamics of H3K4me3-enriched genes

H3K4me3 enrichment of genes is tightly associated with their corresponding transcript abundances (Ruthenburg *et al.*, 2007). We compared H3K4me3-positive genes to their RNA-seq expression values in DSX tissue collected from a different trial (Mizrachi *et al.*, in preparation). On average, genes enriched for H3K4me3 were expressed almost two-fold higher than those with detected expression in DSX, and over five-fold more than those lacking the histone modification (Fig. S4.13). Less than one percent of H3K4me3-enriched genes had no expression evidence (not shown). Furthermore, the percentage of genes exhibiting H3K4 trimethylation increased with gene expression levels (Fig. 4.4a). Of the top 10% of genes expressed in DSX, ~73% were trimethylated at H3K4, compared to ~1.6% of genes with no detected expression (Fig. 4.4a). These results indicate that H3K4me3 enrichment of genes is indeed predictive of gene activation, where H3K4me3 is most often associated with genes expressed at high levels.

We next investigated whether local H3K4me3 tag density, which reflects the degree of enrichment of H3K4 trimethylation at a given locus, is related to transcript levels. Using the abovementioned RNA-seq data of DSX tissue from field-grown *E. grandis* trees, genes

© University of Pretoria

were ranked by transcript abundance and expressed genes were divided into ten ranked expression level categories of equal size. Average H3K4me3 ChIP-seq library coverage was calculated for each base around the 5' regions of genes in each category. As expected, we found that H3K4me3 enrichment was most pronounced around the 5' region of genes in the top expression level category, and that enrichment showed a concordant decrease with less abundant transcript levels (Fig. 4.4b). This relationship was maintained throughout the 2 kb region downstream of the TSS (Fig. 4.4b). These results confirm that the degree of H3K4 trimethylation at a locus is strongly associated with transcript abundance in *Eucalyptus* DSX.

In addition to an association with gene expression, it was reported in *Arabidopsis thaliana* that genes enriched for H3K4me3 tended to be less tissue-specific than those lacking the H3K4me3 modification, regardless of H3K4 mono- or dimethylation states (Zhang *et al.*, 2009). Shannon entropy (Shannon, 1948), a broad multidisciplinary concept, may be understood in the context of gene expressed as a measure of the evenness of relative transcript abundance across a set of tissues or conditions for a given gene (Schug *et al.*, 2005). To further explore the relationship between H3K4me3 modification and expression in *Eucalyptus*, entropy values of relative transcript abundance across seven tissues and organs (http://eucgenie.bi.up.ac.za/; Hefer *et al.* in preparation) was calculated for the 9,694 genes that were expressed in at least one tissue and significantly enriched for H3K4me3, and compared to entropy values for (1) all genes expressed in DSX, and (2) expressed genes that were not significantly enriched for H3K4me3. Genes enriched for H3K4me3 had significant higher entropy values (i.e., lower tissue specificity) compared to both the expressed, and expressed but lacking H3K4me3, gene sets (Kolmogorov-Smirnov test, $P < 2.2 \times 10^{-16}$) (Fig. 4.4c). Similarly, genes lacking the H3K4me3 mark were

© University of Pretoria

significantly more tissue-specific than all expressed genes in DSX ($P < 2.2 \times 10^{-16}$; Fig. 4.4c). Thus, H3K4me3-enriched genes tend not only to be highly expressed, but they are also expressed in many organs and tissues in *Eucalyptus*. It is noteworthy, however, that ~29% of H3K4me3-enriched genes had entropy values lower than the median of 2.1 for genes expressed in DSX (i.e. high tissue/organ-specificity). Therefore, while H3K4me3 enrichment is highly predictive of transcript abundance in *E. grandis* DSX, it is a poor indicator of tissue/organ-specificity.

4.4.6. The role of H3K4me3 modification in regulating wood-related biological processes

Since the 9,760 genes enriched for H3K4me3 in DSX comprise over 25% of those in the v.1.1. annotation and tend to be more broadly expressed than those lacking the modification, it was hypothesized that H3K4me3-enriched genes would be enriched for general biological processes rather than those specific to wood formation. Indeed, after simplifying 204 significantly overrepresented Gene Ontology (GO) terms using REVIGO (Supek *et al.*, 2011), H3K4-trimethylated genes were found to be enriched for biological processes largely comprising general cellular metabolism and developmental processes (Fig. S4.14). A considerable overlap was observed between significant biological function GO terms for H3K4me3-enriched genes and those of all genes expressed in DSX (Fig. S4.15), an expected result due to the correlation of H3K4me3 with gene expression.

While lower-level GO terms characteristic of xylogenesis, such as secondary cell wall biosynthetic processes, were not overrepresented among H3K4me3-enriched genes, H3K4 trimethylation at genes involved in xylogenesis provides insights into how they are epigenetically regulated. A number of putative functional homologs of secondary cell

© University of Pretoria

wall-associated biosynthetic genes, carbohydrate metabolism enzymes, and transcription factors (Myburg *et al.*, in press; Calvert *et al.*, unpublished) exhibited H3K4me3 modification, most of which were highly expressed and several showing preferential expression in DSX (Table 4.2).

Of the *Eucalyptus* phenylpropanoid pathway (Carocha *et al.*, in preparation), nearly all of the likely functional homologs of enzymes catalysing the conversion of phenylalanine to guaiacyl (G) lignin (Vanholme *et al.*, 2010) were marked by H3K4me3 in both individuals (Table 4.2), with the exception of C3H (Eucgr.A02190) which had a significant peak in only one of the individuals. Homologs of most of the genes required for xylan biosynthesis (reviewed by Mizrachi *et al.*, 2012), among them homologs of *IRX7*, *IRX8*, *IRX9*, *IRX9-L*, *IRX10*, *IRX10-L* and *IRX15-L*, as well as several enzymes involved in xylan chain substitution, were positive for H3K4me3 modification. Of the genes required for cellulose biosynthesis (Somerville, 2006), several primary cell wall-associated *CesA* homologs (*CesA1*, *CesA3*, *CesA6*, *CesA9*) overlapped H3K4me3 peaks, while of the secondary cell wall-associated CesAs, the putative ortholog of *Arabidopsis CesA4* (Eucgr.A01324) was positive for H3K4me3. Finally, a limited number of secondary cell wall-associated transcription factors (reviewed in Hussey *et al.*, 2013), among them homologs of fiber-associated SND1 and lignin-associated MYB85 showed evidence of H3K4 trimethylation (Table 4.2).

To validate the ChIP-seq data, we performed ChIP-qPCR analysis focusing on carbohydrate and secondary cell wall-associated loci with evidence of H3K4 trimethylation. This method evaluates enrichment directly against mock (nonspecific IgG) ChIP, whereas the ChIP-seq peak-calling algorithm uses input as negative control, thus

© University of Pretoria

providing an independent assessment of enrichment. Of the seven positive regions identified by ChIP-seq, all showed clear enrichment (8 - 165 fold) in the H3K4me3 ChIP-qPCR samples compared to mock ChIP (Fig. 4.5). We additionally investigated whether loci testing negative for H3K4me3 as assessed by MACS was enriched relative to mock ChIP. *CesA7* and *CesA8* orthologs, which were not detected as H3K4me3 targets using ChIP-seq, were found to be positive for H3K4me3 relative to mock ChIP (Fig. 4.5). For the S-lignin pathway, one *F5H* homolog and two *COMT* homologs were similarly enriched in at least one, if not both, of the trees sampled (Fig. S4.16). We included two controls for the qPCR analysis. In the first, we validated two IgG_{2a} (mock ChIP) peaks found by MACS overlapping homologs of SND2 and NST1, which were regarded as false positive peaks. These targets showed similar amplification between H3K4me3 and mock ChIP samples as expected (Fig. 4.5). Second, we profiled two intergenic negative control regions which showed negligible amplification in both H3K4me3 and mock ChIP samples, showing that there was no template loading bias in the H3K4me3 samples (Fig. 4.5; Fig. S4.16).

4.5. Discussion

In this study, we sought to understand the role of H3K4me3 in the epigenetic regulation of secondary xylem development in *E. grandis*, modifying and optimizing existing chromatin preparation protocols in order to perform ChIP-seq on this challenging tissue. We have shown that high-quality ChIP-seq profiles can be generated using our approach, revealing both known properties of trimethylated H3K4 as well as a novel role in the epigenetic regulation of various aspects of xylogenesis.

© University of Pretoria

Our crosslinked chromatin preparation procedure differs substantially from known methodologies, where fresh intact cells in culture or entire plant organs are submerged in formaldehyde solution to capture *in vivo* protein-DNA interactions. Since scraped DSX tissue is difficult to prepare in this way in the field, we opted for fixation after tissue freezing and maceration. Although crosslinking frozen tissue for ChIP applications has been reported (Morohashi, 2010), we are aware of no reports to fix nuclei directly. Over 90% of identified peaks overlapped between sampled individuals, showing that our approach successfully captured *in vivo* H3K4me3 binding sites despite fixing after tissue maceration. While the use of a ChIP DNA amplification step allowed for the preparation of Illumina sequencing libraries from only 1 - 2 ng, or less, of ChIP DNA in this study, the high duplication rate is undesirable. We have also frequently found that most of the DNA in amplified samples was >500 bp in length (Fig. S4.6), resulting in libraries with a small fraction of the DNA having the preferred insert size of 100 - 500 bp. These significant disadvantages may favour the preparation of Illumina libraries from suboptimal quantities of unamplified ChIP DNA, or the use of single-molecule sequencing approaches, in future.

Eucalyptus H3K4me3 data are consistent with our understanding of H3K4me3 in other plants, animals, and yeast. H3K4 trimethylation generally occurs ~600 - 700 bp downstream of the TSS (Fig. 4.3b), irrespective of the gene length. Of course, this range assumes that TSS predictions in the v.1.1 annotation of the *E. grandis* genome are accurate, and it is possible that revised TSS annotations will shift this range somewhat. Nonetheless, the range is similar to that reported in rice (Li *et al.*, 2008), but further from the TSS than that in *Arabidopsis* which mostly occurs within 500 bp of the TSS (Zhang *et al.*, 2009; Ha *et al.*, 2011). The vast majority of H3K4me3 peaks were gene-associated (Fig. 4.2), including those encoding noncoding RNAs that are predicted to be transcribed

© University of Pretoria

by Pol II (Table 4.1), and the H3K4me3 ChIP-seq tag coverage within the first kilobase after the TSS correlated well with transcript abundance (Fig. 4.4b). As transcript levels increased, a greater proportion of genes expressed at each level became enriched for H3K4me3 (Fig. 4.4a), supporting the known function of H3K4me3 in keeping expressed genes in a transcriptionally active state (Liu *et al.*, 2010; Lauberth *et al.*, 2013). The H3K4me3 signal could represent the number of cells of a particular cell type with H3K4me3 trimethylation at a given locus, or the proportion of cell types in the tissue that are H3K4-trimethylated at a locus, or both. ChIP-seq analysis of individual xylem cell types remains a future challenge.

H3K4me3 peaks predicted on-off states of target genes to a high degree of precision: over 99% of H3K4me3-enriched genes were expressed in DSX tissue, >85% of them above the median FPKM value, and less than two percent of genes without evidence of expression were positive for H3K4me3. Considering that our RNA-seq data originated from an independent trial, the exceptions to the rule are unsurprising. Conversely, gene transcript level was generally a poor predictor of H3K4me3 modification at a locus – even among the most highly expressed genes in DSX tissue, ~27% did not show evidence of H3K4me3 enrichment (Fig. 4.4a). Accurate prediction of mRNA abundance generally requires information for more than one histone modification mark (Kumar *et al.*, 2013) and depends largely on transcript quantification methods (e.g. CAGE, RNA-Seq) (Dong *et al.*, 2012). It is likely that partially functionally redundant histone modifications such as mono- or dimethylated H3K4, or lysine 9-acetylated histone H3, may be sufficient to promote an active chromatin configuration in the absence of H3K4me3. We also showed using ChIP-qPCR that several H3K4me3-negative sites according to MACS were indeed enriched relative to mock ChIP, suggesting a high rate of false negatives. It is unlikely that our

© University of Pretoria

criteria defining H3K4me3-enriched regions were too stringent, however, pooling replicate samples to increase peak-calling sensitivity identified a similar number of target genes at a false discovery rate threshold of 0.05 (not shown). Additional H3K4me3-enriched genes may be identified by increased sequencing depth, alternative peak-calling algorithms or using a mock ChIP-seq library as a negative control rather than input.

H3K4me3 alone does not preferentially regulate genes involved in specific biological pathways or functions (Mikkelsen *et al.*, 2007; Ha *et al.*, 2011). In agreement with this, overrepresented biological functions of H3K4me3-modified genes were similar to those of genes expressed in DSX tissue (Fig. S4.14, Fig. S4.15). It was reported in *Arabidopsis thaliana* that H3K4me3-modified genes tend to show little tissue-specificity compared to genes lacking the mark (Zhang *et al.*, 2009; Ha *et al.*, 2011), a trend we confirmed in *Eucalyptus* (Fig. 4.4c). Despite this tendency, we showed that H3K4me3 was present at several genes involved in secondary cell wall biosynthesis which were preferentially expressed in DSX (Table 4.2). These included a suite of cellulose, xylan and lignin biosynthetic genes, carbohydrate metabolism genes and transcription factors (Fig. 4.5, Fig. S4.16, Table 4.2). Thus, H3K4 trimethylation appears to play a prominent role in the epigenetic regulation of wood formation.

H3K4 trimethylation profiles, especially when combined with DNase-seq data (Thurman *et al.*, 2012), are a useful resource for annotating TSSs as well as direction of transcription (Hon *et al.*, 2009). Our H3K4me3 data suggest that 235 low-confidence gene models in the v.1.0 annotation that were removed from the v.1.1 annotation are true gene models. We have also found numerous examples of H3K4me3 peaks located at genomic regions that have not been previously annotated, but show clear RNA-seq expression

© University of Pretoria

coverage (see Fig. S4.17 for three examples). Thus, the H3K4me3 data from this study is an important line of evidence for future revisions of the *E. grandis* genome annotation.

4.6. Conclusion

ChIP-seq has proved to be a valuable technique for the high-throughput analysis of *in vivo* protein-DNA interactions in yeast, mammals, and to a lesser degree, plants. As this technology becomes more widespread, its application to novel and challenging tissues will require extensive optimization and testing. We successfully optimized and combined standard ChIP-seq with a nano-ChIP-seq protocol to developing secondary xylem tissue from *Eucalyptus*, showing that by directly crosslinking nuclei from frozen tissue, rather than intact fresh tissue, high-quality profiles of a modified histone could be produced for mature plantation trees. This approach thus allows for the study of protein-DNA interactions from tissues collected and frozen in the field. We identified 11,773 H3K4me3 peaks common to two individuals, 85% of which overlapped genes and their promoters. H3K4 trimethylation was common in the 5' vicinity of transcribed regions, the enrichment of which was strongly correlated with gene expression. While H3K4me3-enriched genes tend to be broadly expressed across tissues, we showed that this activating epigenetic mark plays a prominent role in the transcriptional regulation of several tissue-specific genes with crucial functions in wood formation. The H3K4me3-enriched miRNAs and splicing-associated snoRNAs identified in this study suggest that these noncoding RNAs are biologically active in developing secondary xylem, guiding future research into the poorly understood post-transcriptional regulation of wood formation. Finally, a number of H3K4me3 peaks located at unannotated genomic regions with transcriptional evidence, providing a valuable resource for improved annotation of the *E. grandis* genome sequence.

© University of Pretoria

4.7. Acknowledgements

Sequence data were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/ in collaboration with the user community. SH, EM and ZM acknowledge funding from the Department of Science and Technology (DST), South Africa, the National Research Foundation of South Africa (NRF) (UID 81111) and NRF Bioinformatics and Functional Genomics Program (UID 86936), Sappi and Mondi through the Forest Molecular Genetics (FMG) Program at the University of Pretoria, and The Human Resources for Industry Program (THRIP) (UID 80118). Opinions expressed, and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF. Mondi and Mmoledi Mphahlele (Mondi) are acknowledged for providing and collecting plant materials, respectively. AG acknowledges funding from USDA National Institute of Food and Agriculture and the Office of Science (BER), US Department of Energy.

4.8. References

Adli M, Bernstein BE. 2011. Whole-genome chromatin profiling from limited numbers of cells using nano-ChIP-seq. *Nature Protocols* **6**(10): 1656-1668.

Adli M, Zhu J, Bernstein BE. 2010. Genomewide chromatin maps derived from limited numbers of hematopoietic progenitors. *Nature Methods* **7**(8): 615–618.

Alvarez-Venegas R, Avramova Z. 2005. Methylation patterns of histone H3 Lys 4, Lys 9 and Lys 27 in transcriptionally active and inactive *Arabidopsis* genes and in *atx1* mutants. *Nucleic Acids Research* **33**(16): 5199-5207.

Alvarez-Venegas R, Pien S, Sadder M, Witmer X, Grossniklaus U, Avramova Z. 2003. ATX-1, an *Arabidopsis* homolog of trithorax, activates flower homeotic genes. *Current Biology* **13**: 627-637.

Barrera LO, Ren B. 2006. The transcriptional regulatory code of eukaryotic cells – insights from genome-wide analysis of chromatin organization and transcription factor binding. *Current Opinion in Cell Biology* **18**: 291-298.

© University of Pretoria

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129**: 823-837.

Berr A, Shafiq S, Shen W-H. 2011. Histone modifications in transcriptional activation during plant development. *Biochimica et Biophysica Acta* **1809**: 567-576.

Boeva V, Lermine A, Barette C, Guillouf C, Barillot E. 2012. Nebula – a web-server for advanced ChIP-seq data analysis. *Bioinfomatics* **28**(19): 2517-2519.

Brown JW, Clark GP, Leader DJ, Simpson CG, Lowe TODD. 2001. Multiple snoRNA gene clusters from *Arabidopsis*. *RNA* **7**: 1817-1832.

Chen X. 2005. microRNA biogenesis and function in plants. *FEBS Letters* **579**: 5923-5931.

Chen Y, Negre N, Li Q, Mieczkowska JO, Slattery M, Liu T, Zhang Y, Kim T-K, He HH, Zieba J, Ruan Y, Bickel PJ, Myers RM, Wold BJ, White KP, Lieb JD, Liu XS. 2012. Systematic evaluation of factors influencing ChIP-seq fidelity. *Nature Methods* **6**: 609-614.

Courtois-Moreau CL, Pesquet E, Sjödin A, Muñiz L, Bollhöner B, Kaneda M, Samuels L, Jansson S, Tuominen H. 2009. A unique program for cell death in xylem fibers of *Populus* stem. *The Plant Journal* **58**(2): 260-274.

Das PM, Ramachandran K, vanWert J, Singal R. 2004. Chromatin immunoprecipitation assay. *Biotechniques* **37**: 961-969.

Deal RB, Henikoff S. 2011. Histone variants and modifications in plant gene regulation. *Current Opinion in Plant Biology* **14**: 116-122.

Dong X, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, Gingeras TR, Gerstein M, Guigó R, Birney E, Weng Z. 2012. Modeling gene expression using chromatin features in various cellular contexts. *Genome Biology* **13**: R53.

Dover J, Schneider J, Tawiah-Boateng MA, Wood A, Dean K, Johnston M, Shilatifard A. 2002. Methylation of histone H3 by COMPASS requires ubiquitination of histone H2B by Rad6. *Journal of Biological Chemistry* **277**(32): 28368-28371.

Giardine B, Riemer C, Hardison R, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent W, Nekrutenko A. 2005. Galaxy: a platform for interactive large-scale genome analysis. *Genome Research* **15**(10): 1451-1455.

Goecks J, Nekrutenko A, Taylor J. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology* **11**(8): R86.

Goren A, Ozsolak F, Shoresh N, Ku M, Adli M, Hart C, Gymrek M, Zuk O, Regev A, Milos PM, Bernstein BE. 2010. Chromatin profiling by directly sequencing small quantities of immunoprecipitated DNA. *Nature Methods* **7**(1): 47–49.

© University of Pretoria

Gorshkova T, Brutch N, Chabbert B, Deyholos M, Hayashi T, Lev-Yadun S, Mellerowicz EJ, Morvan C, Neutelings G, Pilate G. 2012. Plant fiber formation: state of the art, recent and expected progress, and open questions. *Critical Reviews in Plant Sciences* **31**(3): 201-228.

Guenther MG, Levine SS, Boyer LA, Jaenisch R, Youn RA. 2007. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130**: 77-88.

Ha M, Ng DW-K, Li W-H, Chen ZJ. 2011. Coordinated histone modifications are associated with gene expression variation within and between species. *Genome Research* **21**: 590-598.

Hampsey M, Reinberg D. 2003. Tails of intrigue: phosphorylation of RNA polymerase II mediates histone methylation. *Cell* **113**(429-432).

Haring M, Offermann S, Danker T, Horst I, Peterhansel C, Stam M. 2007. Chromatin immunoprecipitation: optimization, quantitative analysis and data normalization. *Plant Methods* **3**: 11.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Calcar SV, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics* **39**(3): 311-318.

Hon GC, Hawkins RD, Ren B. 2009. Predictive chromatin signatures in the mammalian genome. *Human Molecular Genetics* **18**: R195-R201.

Hussey SG, Mizrachi E, Creux NM, Myburg AA. 2013. Navigating the transcriptional roadmap regulating plant secondary cell wall deposition. *Frontiers in Plant Science* **4**: 325.

Ingen Hv, Schaik FMAv, Wienk H, Ballering J, Rehmann H, Dechesne AC, Kruijzer JAW, Liskamp RMJ, Timmers HTM, Boelens R. 2008. Structural insight into the recognition of the H3K4me3 mark by the TFIID subunit TAF3. *Structure* **16**: 1245-1256.

Kaufmann K, Muiño J, Østerås M, Farinelli L, Krajewski P, Angenent GC. 2010. Chromatin immunoprecipitation (ChIP) of plant transcription factors followed by sequencing (ChIP-SEQ) or hybridization to whole genome arrays (ChIP-CHIP). *Nature Protocols* **5**: 457–472.

Kharchenko PV, Tolstorukov MY, Park PJ. 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature Biotechnology* **26**(12): 1351-1359.

Kouzarides T. 2007. Chromatin modifications and their function. *Cell* **128**: 693-705.

Kumar V, Muratani M, Rayan NA, Kraus P, Lufkin T, Ng HH, Prabhakar S. 2013. Uniform, optimal signal processing of mapped deep-sequencing data. *Nature Biotechnology* **31**(7): 615-622.

Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, Chen Y, DeSalvo G, Epstein C, Fisher-Aylor KI,

© University of Pretoria

Euskirchen G, Gerstein M, Gertz J, Hartemink AJ, Hoffman MM, Iyer VR, Jung YL, Karmakar S, Kellis M, Kharchenko PV, Li Q, Liu T, Liu XS, Ma L, Milosavljevic A, Myers RM, Park PJ, Pazin MJ, Perry MD, Raha D, Reddy TE, Rozowsky J, Shoresh N, Sidow A, Slattery M, Stamatoyannopoulos JA, Tolstorukov MY, White KP, Xi S, Farnham PJ, Lieb JD, Wold BJ, Snyder M. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research* **22**: 1813-1831.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357-359.

Lauberth SM, Nakayama T, Wu X, Ferris AL, Tang Z, Hughes SH, Roeder RG. 2013. H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation. *Cell* **152**: 1021-1036.

Li X, Wang X, He K, Ma Y, Su N, He H, Stolc V, Tongprasit W, Jin W, Jiang J, Terzaghi W, Li S, Deng XW. 2008. High-resolution mapping of epigenetic modifications of the rice genome uncovers interplay between DNA methylation, histone methylation, and gene expression. *The Plant Cell* **20**: 259-276.

Liu C, Lu F, Cui X, Cao X. 2010. Histone methylation in higher plants. *Annual Review of Plant Biology* **61**: 395-420.

Loureiro J, Rodriguez E, Doležel J, Santos C. 2007. Two new nuclear isolation buffers for plant DNA flow cytometry: A test with 37 species. *Annals of Botany* **100**: 875–888.

Luo C, Sidote DJ, Zhang Y, Kerstetter RA, Michael TP, Lam E. 2013. Integrative analysis of chromatin states in *Arabidopsis* identified potential regulatory mechanisms for natural antisense transcript production. *The Plant Journal* **73**: 77-90.

Martin D, Brun C, Remy E, Mouren P, Thieffr D, Jacq B. 2004. GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biology* **5**: R101.

McKeown P, Pendle AF, Shaw PJ 2008. Preparation of *Arabidopsis* Nuclei and Nucleoli. *The Nucleus: Nuclei and Subnuclear Components*: Humana Press, 65-74.

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim T-K, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**(7153): 553-560.

Mizrachi E, Mansfield SD, Myburg AA. 2012. Cellulose factories: advancing bioenergy production from forest trees. *New Phytologist* **194**: 54-62.

Morohashi K 2010. Chromatin immunoprecipitation for plant materials. Available online at http://arabidopsis.med.ohio-state.edu/NSF2010Project/Protocols/Chromatin-IP.pdf.

© University of Pretoria

Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D, Goodstein DM, Dubchak I, Poliakov A, Mizrachi E, Kullan ARK, van Jaarsveld I, Hussey SG, Pinard D, Merwe Kvd, Singh P, Silva-Junior OB, Togawa RC, Pappas MR, Faria DA, Sansaloni CP, Petroli CD, Yang X, Ranjan P, Tschaplinski TJ, Ye C-Y, Li T, Sterck L, Vanneste K, Murat F, Soler M, Clemente HS, Saidi N, Cassan-Wang H, Dunand C, Hefer CA, Bornberg-Bauer E, Kersting AR, Vining K, Amarasinghe V, Ranik M, Naithani S, Elser J, Boyd AE, Liston A, Spatafora JW, Dharmwardhana P, Raja R, Sullivan C, Romanel E, Alves-Ferreira M, Külheim C, Foley W, Carocha V, Paiva J, Kudrna D, Brommonschenkel SH, Pasquali G, Byrne M, Rigault P, Tibbits J, Spokevicius A, Jones RC, Steane DA, Vaillancourt RE, Potts BM, Joubert F, Barry K, Jr. GJP, Strauss SH, Jaiswal P, Grima-Pettenati J, Salse J, Peer YVd, Rokhsar DS, Schmutz J. in press. The genome of *Eucalyptus grandis* **-** a global tree for fiber and energy. *Nature*.

Nagy PL, Griesenbeck J, Kornberg RD, Cleary ML. 2002. A trithorax-group complex purified from *Saccharomyces cerevisiae* is required for methylation of histone H3. *Proceedings of the National Academy of Sciences of the USA* **99**(1): 90-94.

Ng HH, Robert F, Young RA, Struhl K. 2003. Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. *Molecular Cell* **11**: 709-719.

Orlando V, Strutt H, Paro R. 1997. Analysis of chromatin structure by *in vivo* formaldehyde cross-linking. *Methods* **11**: 205–214.

Pfluger J, Wagner D. 2007. Histone modifications and dynamic regulation of genome accessibility in plants. *Current Opinion in Plant Biology* **10**: 645-652.

Philipps G, Gigot C. 1977. DNA associated with nucleosomes in plants. *Nucleic Acids Research* **4**(10): 3617-3626.

Pien S, Fleury D, Mylne JS, Crevillen P, Inzé D, Avramova Z, Dean C, Grossniklaus U. 2008. *ARABIDOPSIS* TRITHORAX1 dynamically regulates *FLOWERING LOCUS C* activation via histone 3 lysine 4 trimethylation. *The Plant Cell* **20**: 580-588.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinfomatics* **26**(6): 841-842.

Rodor J, Letelier I, Holuigue L, Echeverria M. 2010. Nucleolar RNPs: from genes to functional snoRNAs in plants. *Biochemical Society Transactions* **38**(2): 672-676.

Ruthenburg AJ, Allis D, Wysocka J. 2007. Methylation of lysine 4 on histone H3: Intricacy of writing and reading a single epigenetic mark. *Molecular Cell* **25**(1): 15-30.

Saleh A, Alvarez-Venegas R, Avramova Z. 2008a. Dynamic and stable histone H3 methylation patterns at the *Arabidopsis FLC* and *AP1* loci. *Gene* **423**: 43-47.

© University of Pretoria

Saleh A, Alvarez-Venegas R, Avramova Z. 2008b. An efficient chromatin immunoprecipitation (ChIP) protocol for studying histone modifications in *Arabidopsis* plants. *Nature Protocols* **3**(6): 1018-1025.

Saleh A, Alvarez-Venegas R, Yilmaz M, Oahn-Le, Hou G, Sadder M, Al-Abdallat A, Xia Y, Lu G, Ladunga I, Avramova Z. 2008c. The highly similar *Arabidopsis* homologs of Trithorax ATX1 and ATX2 encode proteins with divergent biochemical functions. *The Plant Cell* **20**: 568-579.

Schug J, Schuller W-P, Kappen C, Salbaum JM, Bucan M, Jr CJS. 2005. Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biology* **6**: R33.

Shannon CE. 1948. A mathematical theory of communication. *The Bell System Technical Journal* **27**(3): 379-423.

Shi Y, Whetstine JR. 2007. Dynamic regulation of histone lysine methylation by demethylases. *Molecular Cell* **25**(1): 1-14.

Snegireva AV, Ageeva MV, Amenitskii SI, Chernova TE, Ebskamp M, Gorshkovaa TA. 2010. Intrusive growth of sclerenchyma fibers. *Russian Journal of Plant Physiology* **57**(3): 342–355.

Somerville C. 2006. Cellulose synthesis in higher plants. *Annual Review of Cell and Developmental Biology* **22**: 53–78.

Sun Z-W, Allis CD. 2002. Ubiquitination of histone H2B regulates H3 methylation and gene silencing in yeast. *Nature* **418**: 104-108.

Supek F, Bošnjak M, Škunca N, Šmuc T. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* **6**(7): e21800.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57-74.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield TK, Diegel M, Dunn D, Ebersol AK, Frum T, Giste E, Johnson AK, Johnson EM, Kutyavin T, Lajoie B, Lee B-K, Lee K, London D, Lotakis D, Neph S, Neri F, Nguyen ED, Reynolds HQAP, Roach V, Safi A, Sanchez ME, Sanyal A, Shafer A, Simon JM, Song L, Vong S, Weaver M, Yan Y, Zhang Z, Zhang Z, Lenhard B, Tewari M, Dorschner MO, Hansen RS, Navas PA, Stamatoyannopoulos G, Lieb VRIJD, Sunyaev SR, Akey JM, Sabo PJ, Kaul R, Furey TS, Dekker J, Crawford GE, Stamatoyannopoulos JA. 2012. The accessible chromatin landscape of the human genome. *Nature* **489**: 75-82.

van Dijk K, Ding Y, Malkaram S, Riethoven J-JM, Liu R, Yang J, Laczko P, Chen H, Xia Y, Ladunga I, Avramova Z, Fromm M. 2010. Dynamic changes in genome-wide histone

© University of Pretoria

H3 lysine 4 methylation patterns in response to dehydration stress in *Arabidopsis thaliana*. *BMC Plant Biology* **10**: 238.

Vanholme R, Demedts B, Morreel K, Ralph J, Boerjan W. 2010. Lignin biosynthesis and structure. *Plant Physiology* **153**: 895–905.

Vermeulen M, Mulder KW, Denissov S, Pijnappel WWMP, Schaik FMAv, Varier RA, Baltissen MPA, Stunnenberg HG, Mann M, Timmers HTM. 2007. Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell* **131**: 58-69.

Wood A, Shilatifard A. 2006. Bur1/Bur2 and the Ctk complex in yeast: the split personality of mammalian P-TEFb. *Cell Cycle* **5**(10): 1066-1068.

Zhang X, Bernatavichute YV, Cokus S, Pellegrini M, Jacobsen SE. 2009. Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in *Arabidopsis thaliana*. *Genome Biology* **10**: R62.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, Liu XS. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**: R137.

Zhao J, Morozova N, Williams L, Libs L, Avivi Y, Grafi G. 2001. Two phases of chromatin decondensation during dedifferentiation of plant cells. *The Journal of Biological Chemistry* **276**(25): 22772–22778.

Zhou VW, Goren A, Bernstein BE. 2011. Charting histone modifications and the functional organization of mammalian genomes. *Nature Reviews Genetics* **12**: 7-18.

Zong W, Zhong X, You J, Xiong L. 2013. Genome-wide profiling of histone H3K4-tri-methylation and gene expression in rice under drought stress. *Plant Molecular Biology* **81**: 175-188.

© University of Pretoria

4.9. Figures

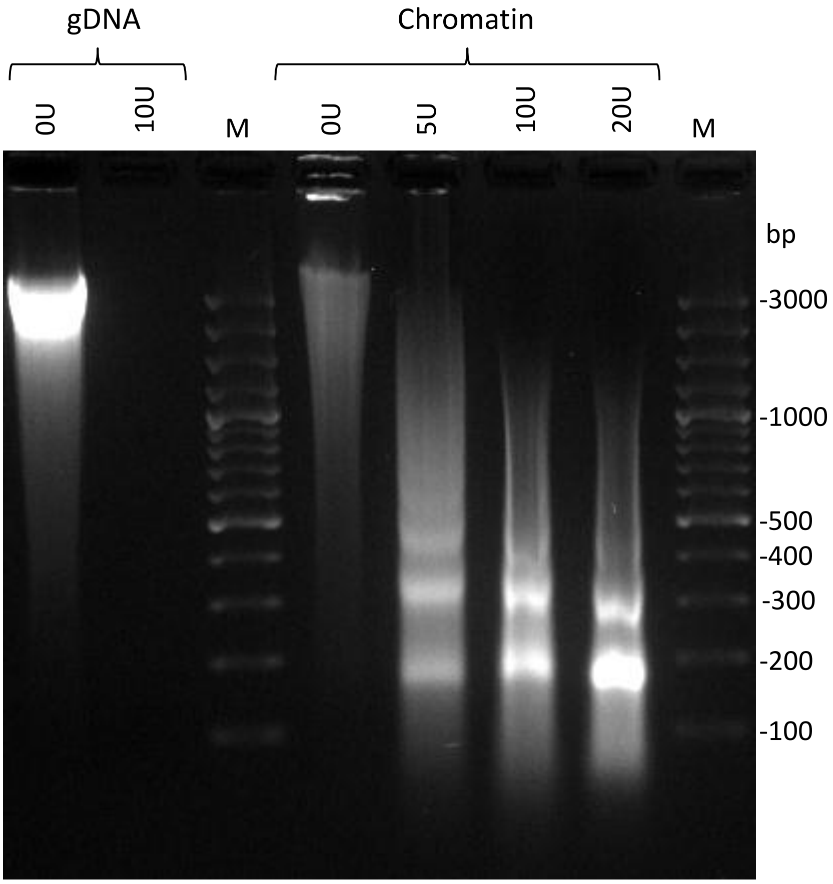


Fig. 4.1. Agarose gel electrophoresis analysis of quality of chromatin isolated from developing secondary xylem. Genomic DNA (gDNA) and developing secondary xylem chromatin were exposed to increasing units (U) of micrococcal nuclease. M, GeneRuler 100 bp plus DNA ladder (0.5 μg).

© University of Pretoria

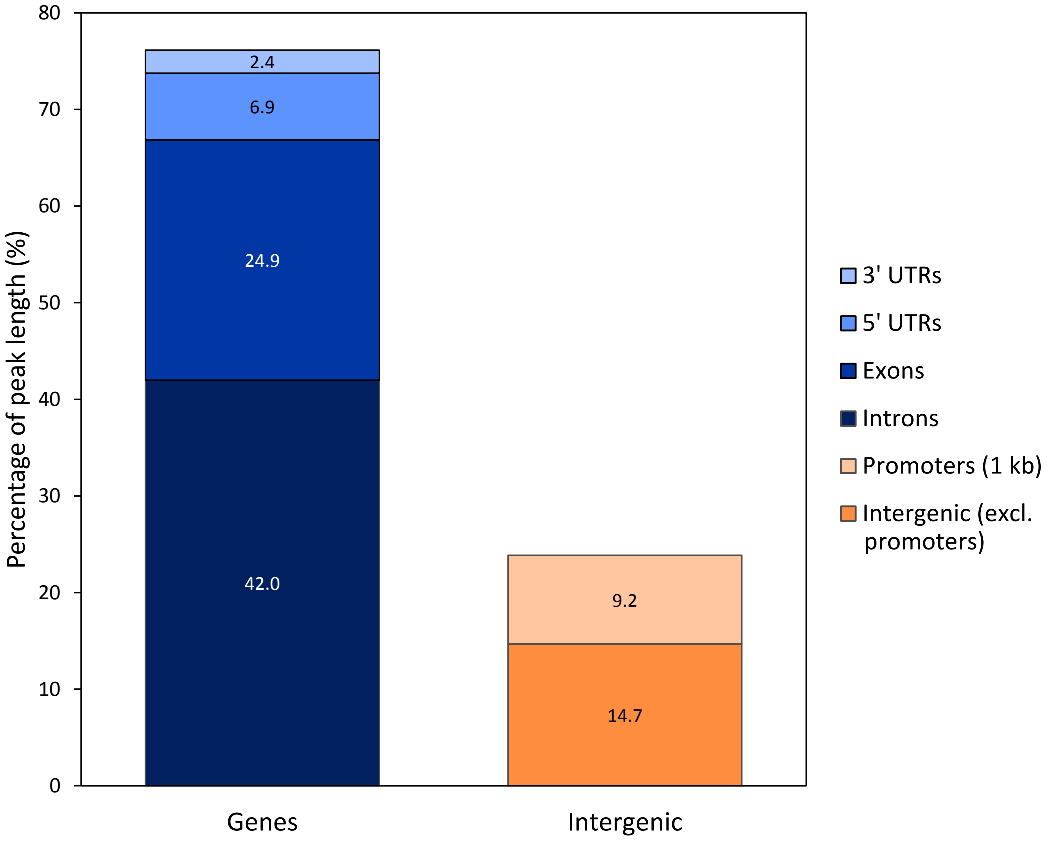


Fig. 4.2. Overlap of H3K4me3 ChIP-seq peaks with genic features. Values are expressed as the average percentage of all peak intervals (bp) that overlap with each feature class.

© University of Pretoria

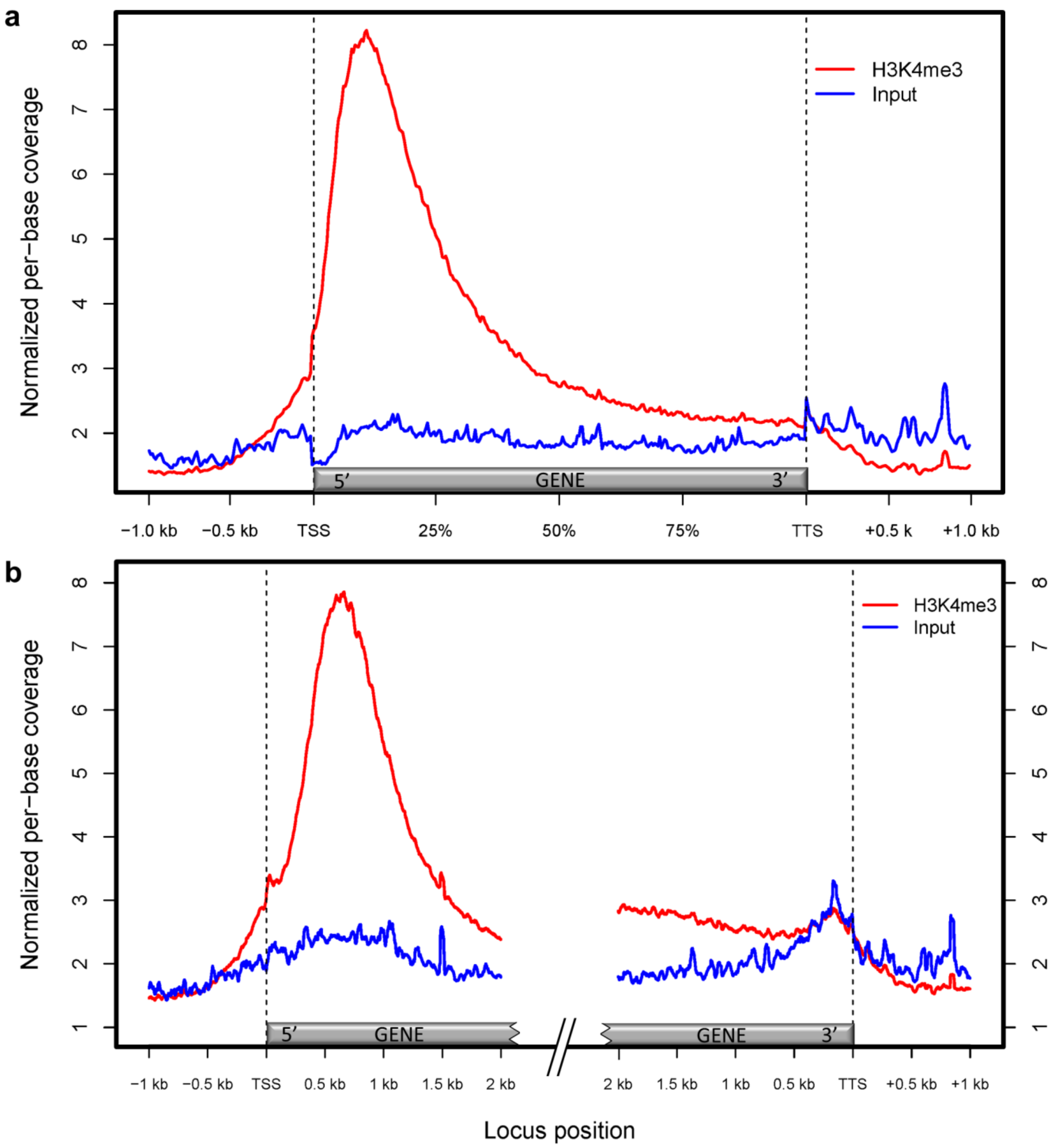
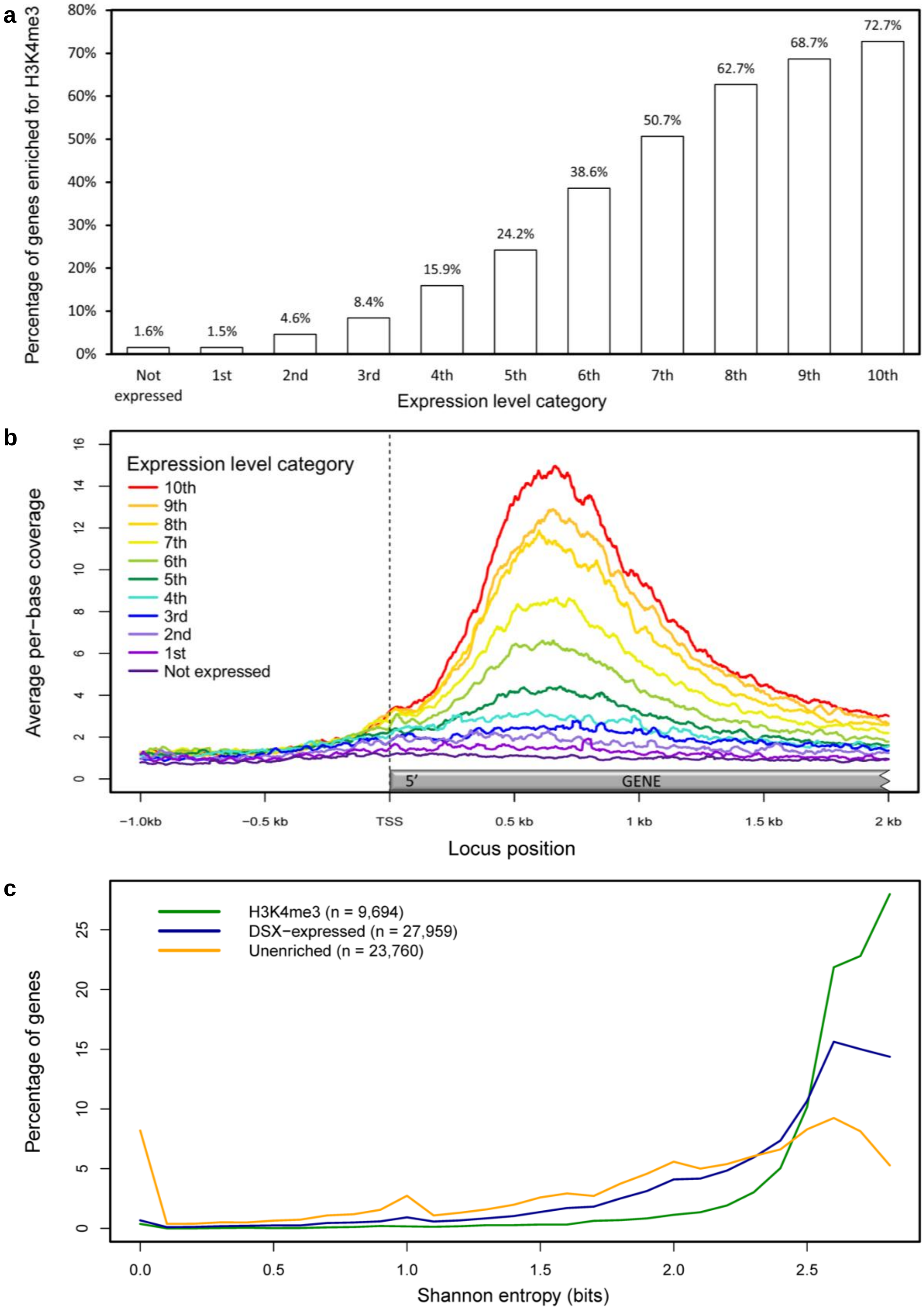


Fig. 4.3. Bulked H3K4me3 and Input ChIP-seq profiles across the 1 kb promoter, transcribed and 1 kb downstream regions of annotated loci. (a) Bin-wise, showing relative gene length. **(b)** Showing absolute distance anchored at the 5' and 3' ends of transcribed regions. Per-base coverage values were normalized between H3K4me3 and Input libraries. TSS, transcription start site; TTS, transcription termination site; gene, regions annotated as transcribed in *E. grandis* v.1.1.

© University of Pretoria

Fig. 4.4. Expression properties associated with H3K4me3 enrichment in developing secondary xylem tissue. (a) Percentage of genes enriched for H3K4me3 among non-expressed genes and genes with increasing expression levels, represented as ten ordinal categories of equal size. **(b)** H3K4me3 enrichment (measured as library coverage) at the 5' regions of transcribed genes, for each of the expression level categories in (a). Average per-base coverage values from 1 kb upstream to 2 kb downstream of the transcriptional start site (TSS) is shown for each expression level category. **(c)** Tissue specificity of genes enriched for H3K4me3 (green), genes expressed in developing secondary xylem regardless of histone modification status (blue), and genes expressed in developing secondary xylem but lacking H3K4me3 modification (orange), as measured by Shannon entropy. High entropy values indicate broad, even expression across tissues; low values indicate high tissue specificity. The maximum possible entropy value for this data is 2.81.

© University of Pretoria



© University of Pretoria

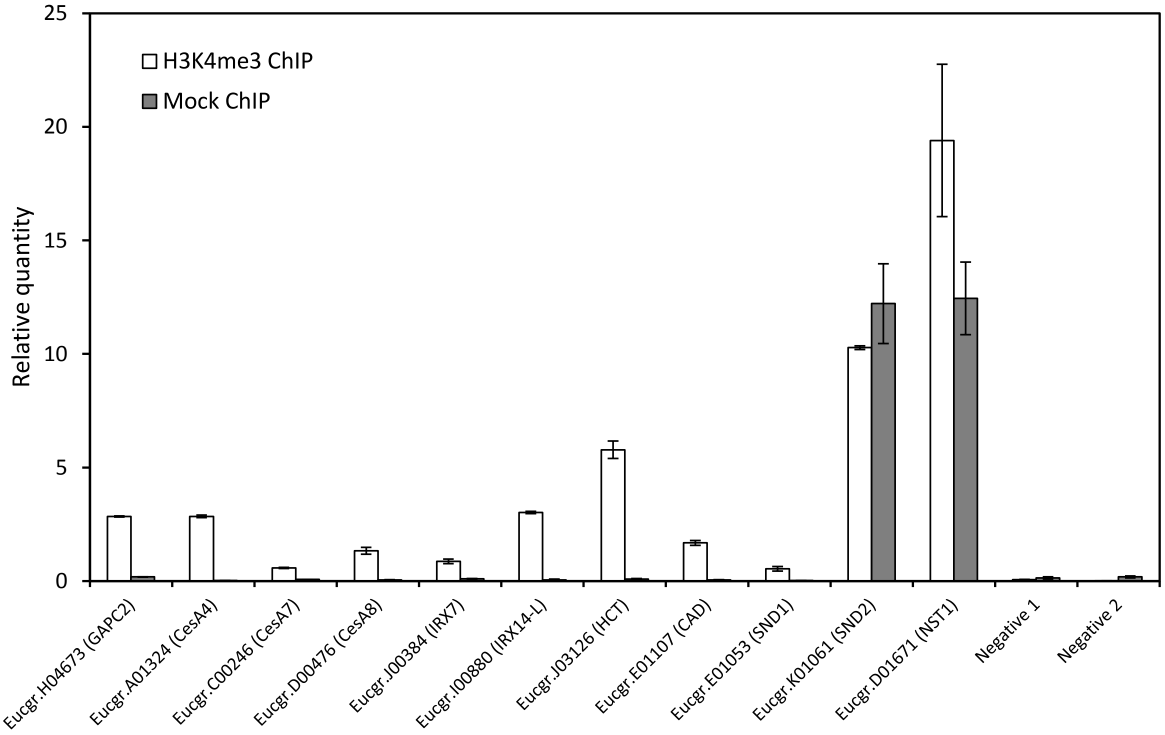


Fig. 4.5. ChIP-qPCR validation of H3K4me3-enriched and control loci. The putative *Arabidopsis* ortholog of each candidate is indicated in parenthesis. Eucgr.C00246 (*CesA7*) and Eucgr.D00476 (*CesA8*) were not identified as H3K4me3 targets using ChIP-seq. Eucgr.K01061 and Eucgr.D01671 serve as validations of false positives arising from nonspecific binding. Two intergenic negative control regions are included. Error bars indicate standard deviation of three technical replicates.

© University of Pretoria

4.10. Tables

Table 4.1. ncRNA elements enriched for H3K4me3. Putative targets of H3K4me3-enriched miRNAs are indicated in Table S4.3.

ncRNA class	H3K4me3-enriched[a]	Total annotations	% enriched[a]
Predicted snoRNAs	23 (61)	175	13.4 (35.5)
Predicted miRNAs	4 (5)	153	2.6 (3.3)
Known miRNAs	1 (2)	60	1.7 (3.3)
Predicted tRNAs	2 (19)	508	0.4 (3.74)
Predicted antisense RNAs	0 (1)	19	0.0 (5.3)
Predicted rRNAs	0 (0)	269	0.0 (0.0)
Predicted spliceosomal snRNA	0 (2)	125	0.0 (1.6)
Predicted sRNAs	0 (0)	80	0.0 (0.0)

[a] Excludes H3K4me3 peaks overlapping with annotated protein-coding genes. ncRNAs overlapping peaks that also overlap genes are indicated in parenthesis.

© University of Pretoria

Table 4.2. Putative functional homologs of secondary cell wall-associated genes exhibiting trimethylated H3K4 and their expression in developing secondary xylem.

	Protein family	Gene ID	Putative *A. thaliana* homolog	*A. thaliana* gene name	Relative expression (%)[a]	Absolute expression (FPKM)[b]
Cellulose biosynthesis						
	CesA	Eucgr.A01324	AT5G44030	CESA4, IRX5, NWS2	91	23,436,733
		Eucgr.I00286	AT2G21770	CESA9	12	3,558,470
		Eucgr.G03380	AT5G05170	CESA3, CEV1, IXR1	18	3,133,070
		Eucgr.C02801	AT4G32410	CESA1, RSW1	13	3,021,617
		Eucgr.F03635	AT2G21770	CESA9	10	1,356,514
		Eucgr.C01769	AT4G32410	CESA1, RSW1	21	1,205,310
		Eucgr.J01278	AT5G05170	CESA3, CEV1, IXR1	21	1,141,560
		Eucgr.F04216	AT5G64740	CesA6	84	789,956
		Eucgr.F04212	AT5G64740	CesA6	69	107,544
		Eucgr.H00939	AT4G32410	CESA1, RSW1	12	136,970
Hemicellulose biosynthesis						
	CslA	Eucgr.A01558	AT5G03760	CslA9	42	2,490,880
	GT8	Eucgr.F00995	AT5G54690	IRX8	88	5,863,770
		Eucgr.B02574	AT1G70090	GATL9, LGT8	13	248,814
		Eucgr.H01923	AT3G50760	GATL2	6	112,741
	GT43	Eucgr.A01172	AT2G37090	IRX9	90	9,605,213
		Eucgr.F02177	AT1G27600	I9H, IRX9-L	55	609,183
		Eucgr.F00463	AT1G27600	I9H, IRX9-L	35	494,170
		Eucgr.C00584	AT1G27600	I9H, IRX9-L	19	284,588
	GT47	Eucgr.G01977	AT1G27440	GUT1, GUT2, IRX10	88	11,264,120
		Eucgr.J00384	AT2G28110	FRA8, IRX7	56	2,854,840
		Eucgr.K02191	AT5G61840	GUT1, IRX10-L	30	1,304,447
	RWA	Eucgr.D00335	AT2G34410	RWA3	78	12,710,067
		Eucgr.B03976	AT3G06550	RWA2	28	2,686,777
	UXS	Eucgr.J00040	AT5G59290	UXS3	62	13,448,933
		Eucgr.H01112	AT3G62830	AUD1, UXS2	15	6,087,550
		Eucgr.A01221	AT3G53520	UXS1	49	2,712,707
	DUF579	Eucgr.I00888	AT5G67210	IRX15-L	67	7,135,410

© University of Pretoria

Carbohydrate metabolism					
SuSy	Eucgr.C03199	AT3G43190	SUS4	46	27,043,300
	Eucgr.C00769	AT3G43190	SUS4	53	6,929,210
	Eucgr.H01094	AT4G02280	SUS3	14	1,947,340
	Eucgr.F03879	AT1G80070	SUS2	10	682,244
	Eucgr.D02653	AT1G01040	SUS1	18	486,588
Hexokinase	Eucgr.B03711	AT1G50460	HKL1	27	1,335,271
	Eucgr.C03728	AT4G29130	HXK1	19	881,978
	Eucgr.F01647	AT1G47840	HXK3	20	759,403
	Eucgr.J00734	AT1G50460	HKL1	8	658,748
Phosphoglucomutase	Eucgr.G02157	AT1G70730	PGM2	36	3,172,777
	Eucgr.B02942	AT1G23190	PGM3	29	2,983,813
	Eucgr.J01084	AT5G17530		25	776,193
	Eucgr.K00185	AT5G51820	PGM1, STF1	14	608,346
PGSIP	Eucgr.F00232	AT4G33330	GUX2, PGSIP3	76	2,360,753
	Eucgr.H04942	AT3G18660	GUX1 ,PGSIP1	91	1,967,787
	Eucgr.H04216	AT5G18480	PGSIP6	20	736,868
Phenylpropanoids and lignin					
PAL	Eucgr.J01079	AT3G53260	PAL2	45	6,744,160
C4H	Eucgr.J01844	AT2G30490	C4H	54	18,372,167
4CL	Eucgr.C02284	AT1G51680	4CL1	57	12,572,167
HCT	Eucgr.J03126	AT5G48930	HCT	58	6,624,317
	Eucgr.F03978	AT5G48930	HCT	57	2,639,093
CCoAOMT	Eucgr.G01417	AT4G34050	CCoAOMT1	70	32,118,967
	Eucgr.I01134	AT4G34050	CCoAOMT1	66	16,420,333
CCR	Eucgr.J03114	AT1G15950	ATCCR1	41	5,899,387
CAD	Eucgr.E01107	AT1G72680	CAD1	21	351,410
	Eucgr.E01110	AT1G72680	CAD1	10	87,075
Transcription factors					
KNOTTED-LIKE	Eucgr.F00106	AT2G31390	BP	20	2,220,160
	Eucgr.D01935	AT1G62990	KNAT7	36	2,181,587
NAC	Eucgr.E01053	AT1G32770	SND1	59	1,036,140
MYB	Eucgr.G01774	AT4G38620	MYB4	47	1,361,414
	Eucgr.D02014	AT4G22680	MYB85	40	799,881
	Eucgr.C00721	AT2G16720	MYB7	22	252,578

[a] Relative to shoot tips, young leaves, mature leaves, flowers, roots and phloem

[b] The median fragments per kilobase per million fragments mapped (FPKM) value of genes expressed in developing secondary xylem tissue is 89,300

© University of Pretoria

4.11. Supplementary figures

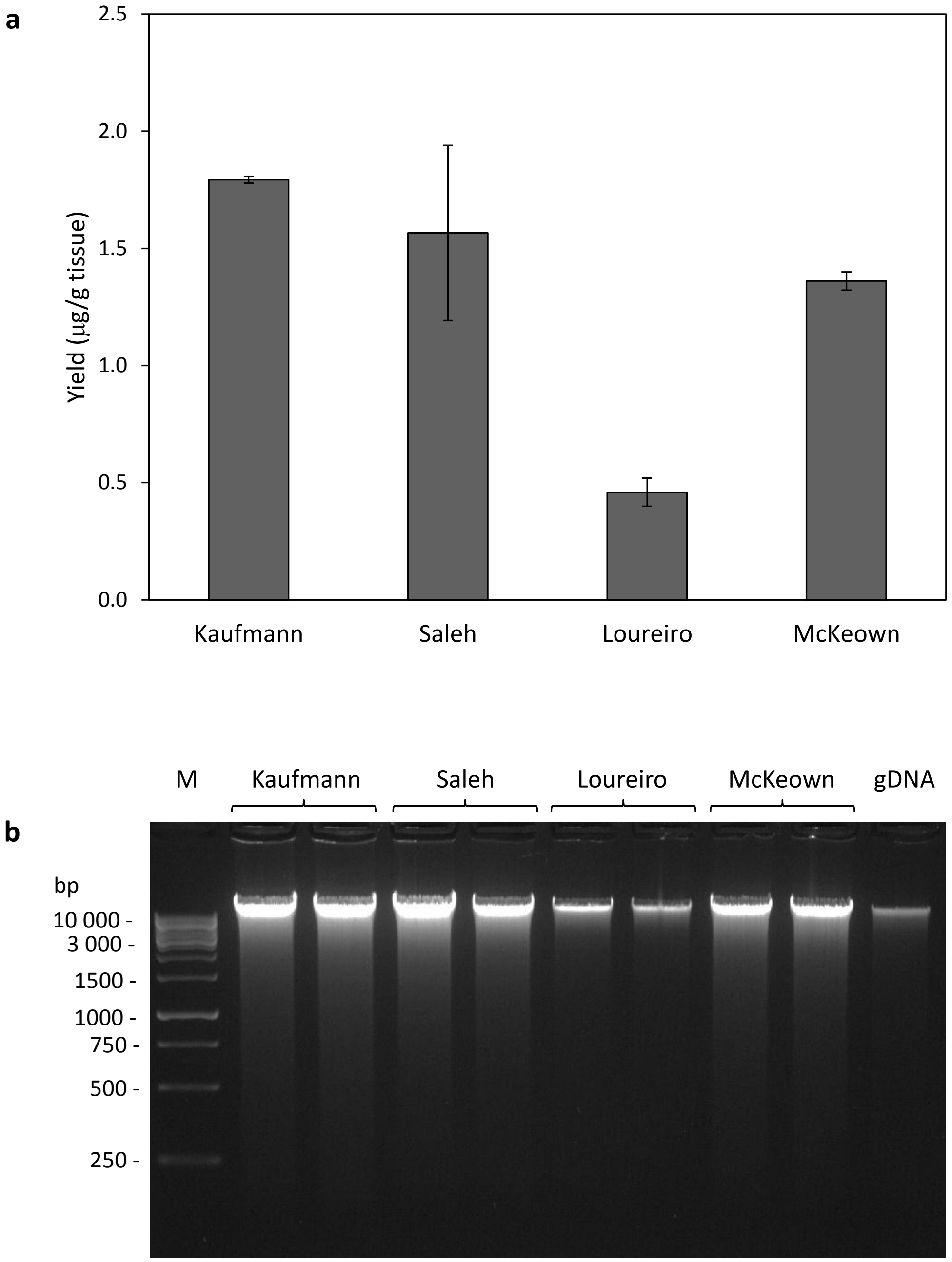


Fig. S4.1. Assessment of DNA yield from crude nuclei preparations of *E. grandis* developing secondary xylem tissue using different buffers. (a) Nuclear DNA yield from crude nuclear pellets. Error bars indicate the range of two extractions. **(b)** Agarose gel electrophoresis analysis of nuclear DNA quality. Duplicate extractions were performed for each buffer. M, GeneRuler 1kb DNA ladder (Fermentas). gDNA, genomic DNA extraction from whole developing secondary xylem tissue.

© University of Pretoria

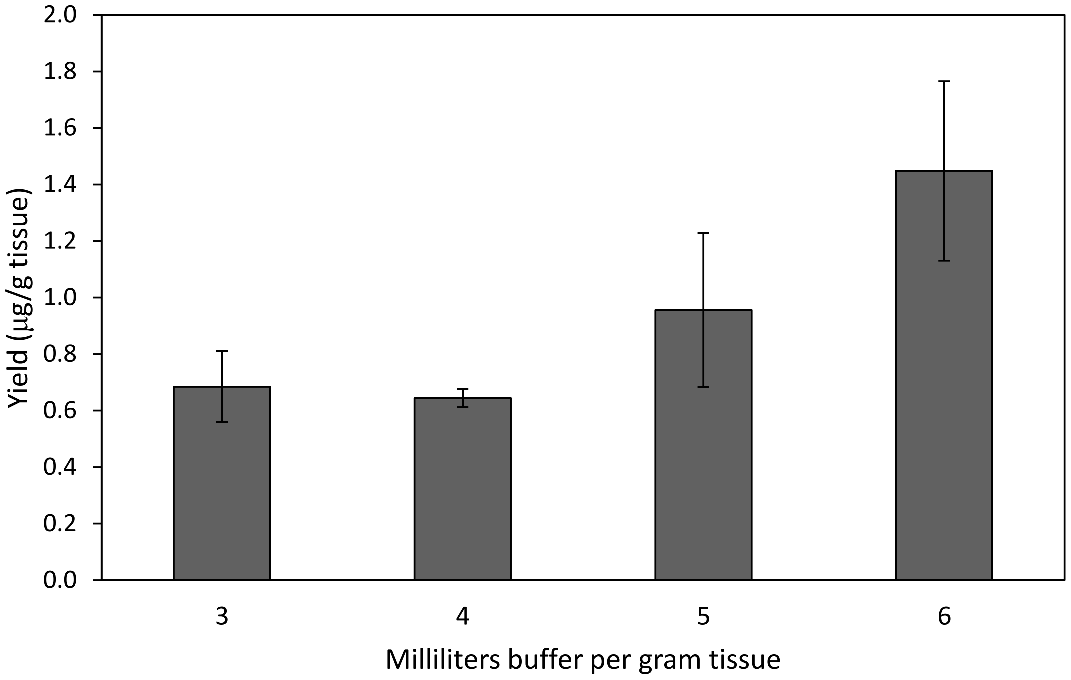


Fig. S4.2. Effect of M3 buffer to tissue mass ratio on yield of nuclear DNA extracted from the crude nuclear pellet. Error bars indicate the range of two independent extractions.

© University of Pretoria

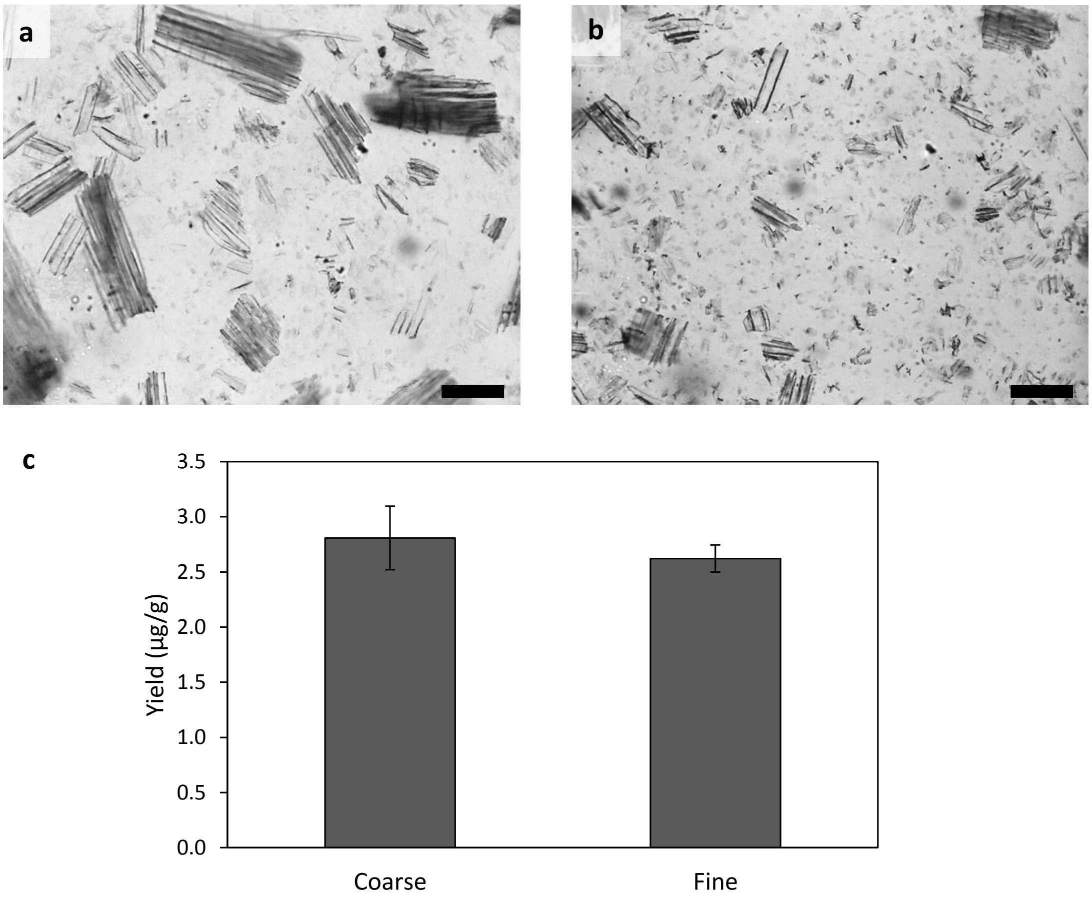


Fig. S4.3. Effect of grinding consistency on yield of nuclear DNA following nuclei isolation. Light micrographs show developing secondary xylem tissue ground to a coarse **(a)** or fine **(b)** consistency (bar = 100 µm). The yield of nuclear DNA is shown in **(c)**. Error bars represent the range of two technical replicates.

© University of Pretoria

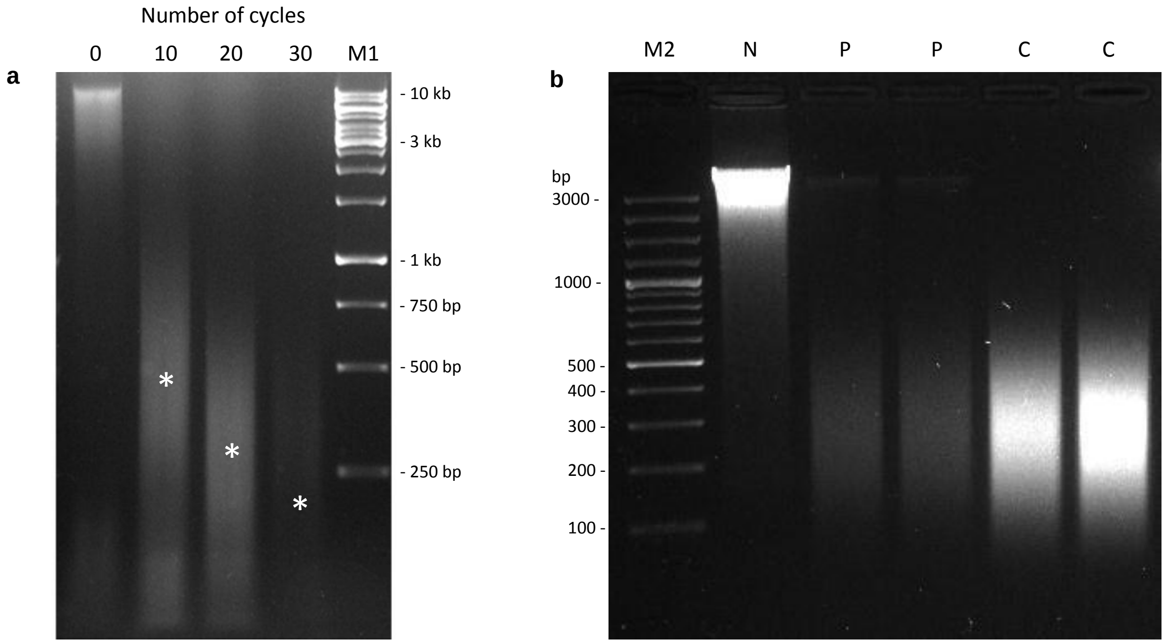


Fig. S4.4. Agarose gel electrophoresis analysis of sonication conditions. (a) Developing secondary xylem chromatin sonicated for various ten-second cycles. Asterisks indicate the average fragment size. M1, GeneRuler 1kb DNA ladder. **(b)** Analysis of residual DNA in the nuclear pellet following sonication. N, unsonicated nuclear pellet; P, residual nuclear pellet after sonication (two extractions); M2, GeneRuler 100 bp DNA ladder plus.

© University of Pretoria

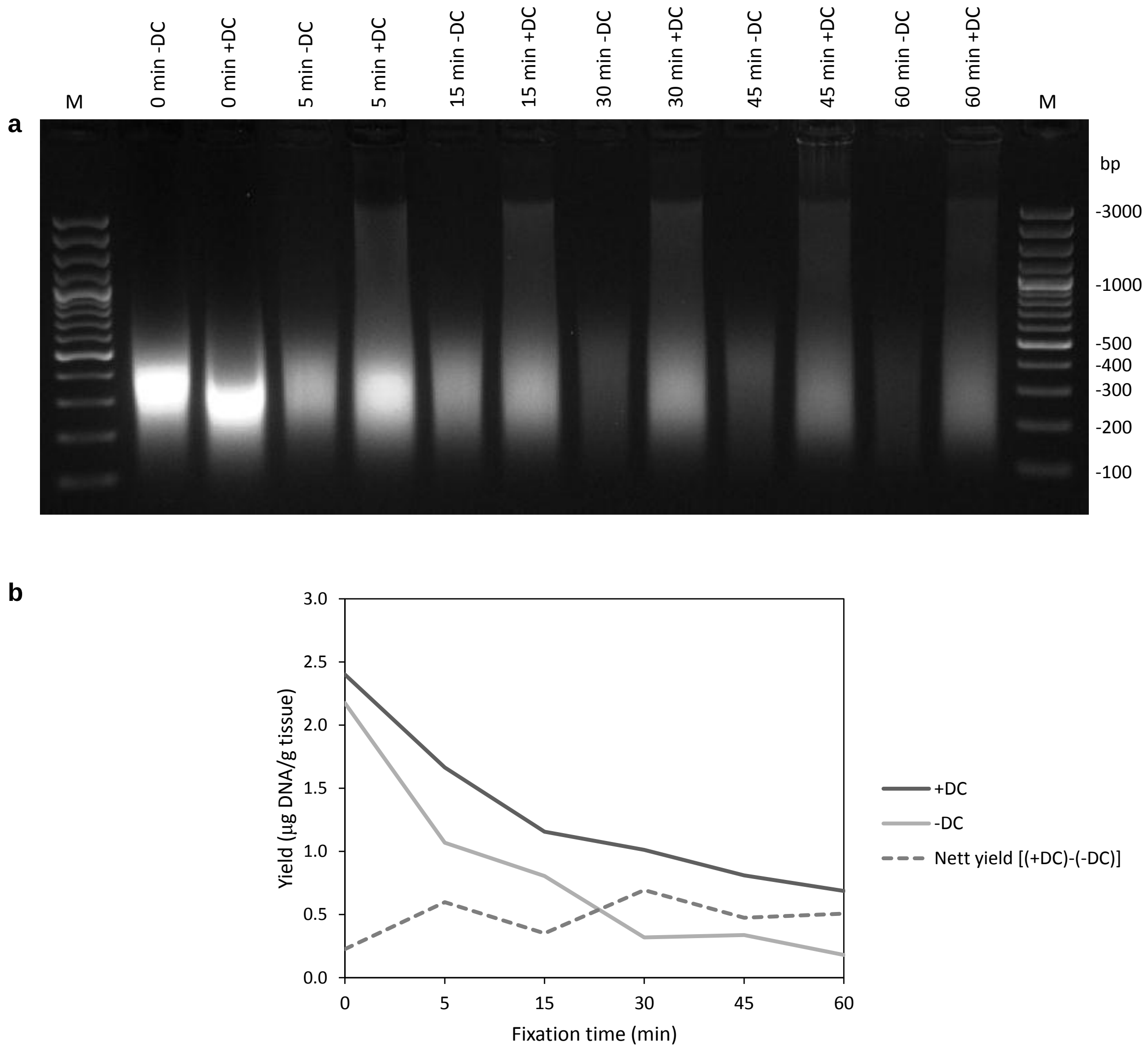


Fig. S4.5. Optimization of formaldehyde-mediated crosslinking of *E. grandis* developing secondary xylem samples. (a) Agarose gel electrophoresis analysis of DNA yield from samples crosslinked for 0, 5, 15, 30, 45 or 60 min and subjected to de-crosslinking (+DC) or no de-crosslinking (–DC). M, GeneRuler 100 bp DNA ladder plus. **(b)** Nett yield of DNA between de-crosslinked (+DC) samples and samples without de-crosslinking treatment (-DC). Samples fixed for 5 and 15 minutes show poor crosslink retention as measured by the ability to extract DNA without prior de-crosslinking, while more than 30 minutes of fixation led to compromised DNA yield after de-crosslinking.

© University of Pretoria

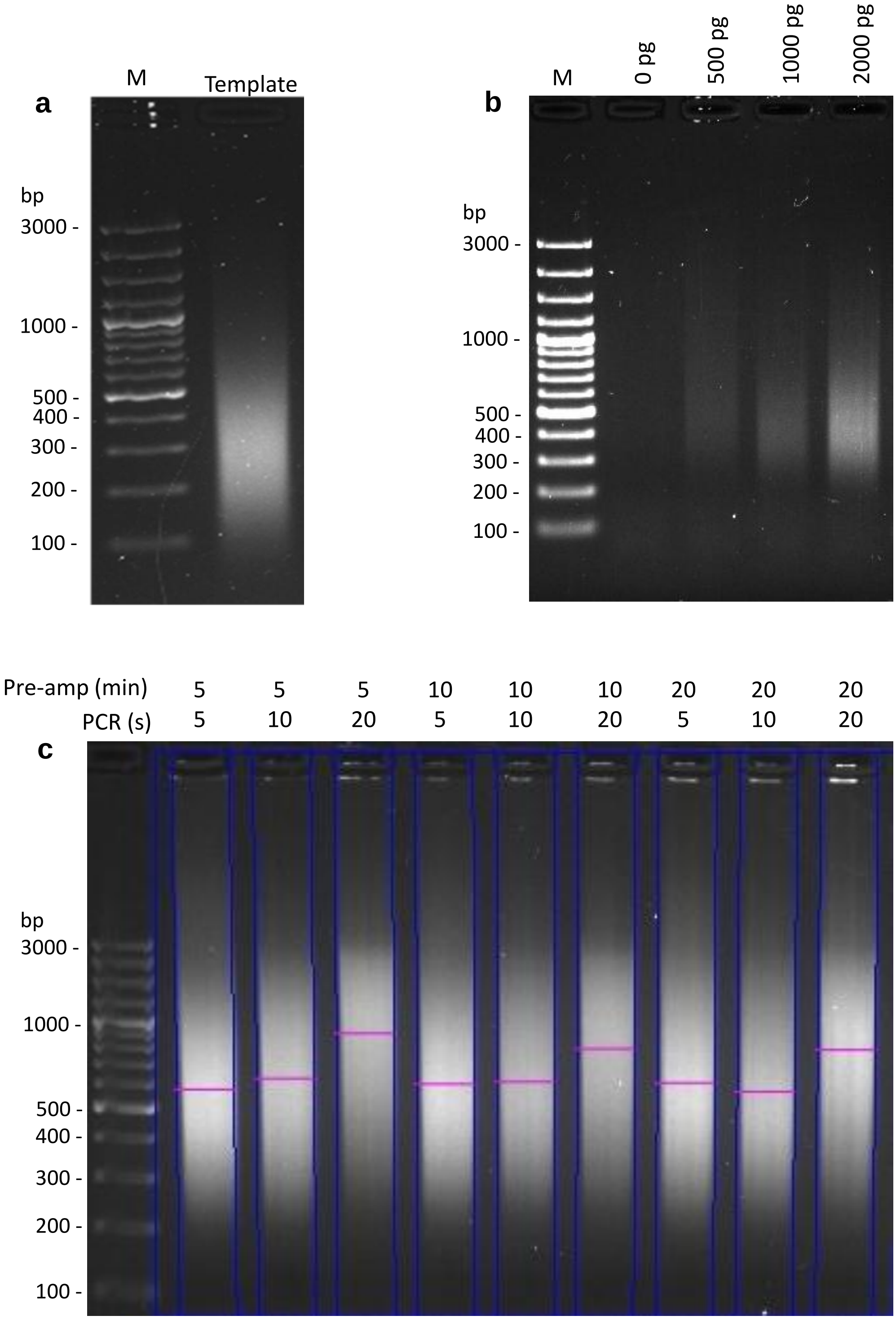


Fig. S4.6. Optimization of ChIP DNA amplification. (a) Fragment length distribution of concentrated template DNA. **(b)** Fragment length distribution of amplified DNA (15 cycles) starting from various template quantities. **(c)** Agarose gel electrophoresis analysis of amplified DNA fragment length using various pre-amplification and PCR extension times (seconds). M, GeneRuler 100 bp DNA ladder plus marker.

© University of Pretoria

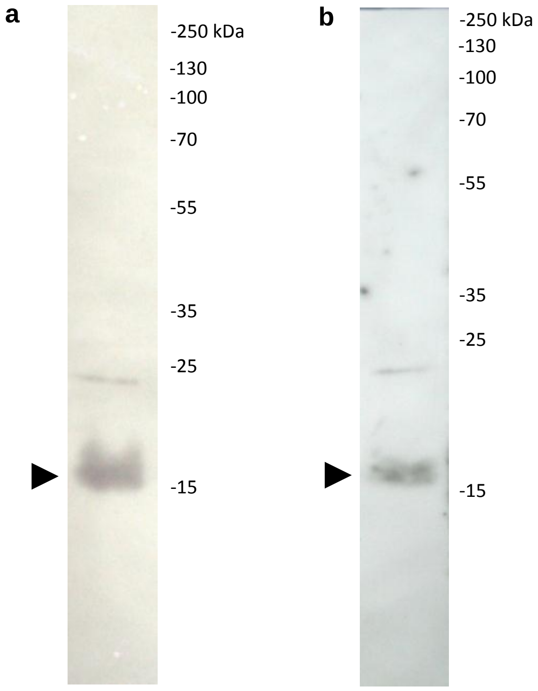


Fig. S4.7. Western blot analysis of *E. grandis* developing secondary xylem nuclear protein extracts using anti-H3K4me3 antibody. (a) ~60 µg protein extract, **(b)** ~30 µg protein extract. The 17 kDa target protein (arrows) comprises over 50% of the lane signal in each case and thus passes ENCODE requirements (Landt *et al.*, 2012).

© University of Pretoria

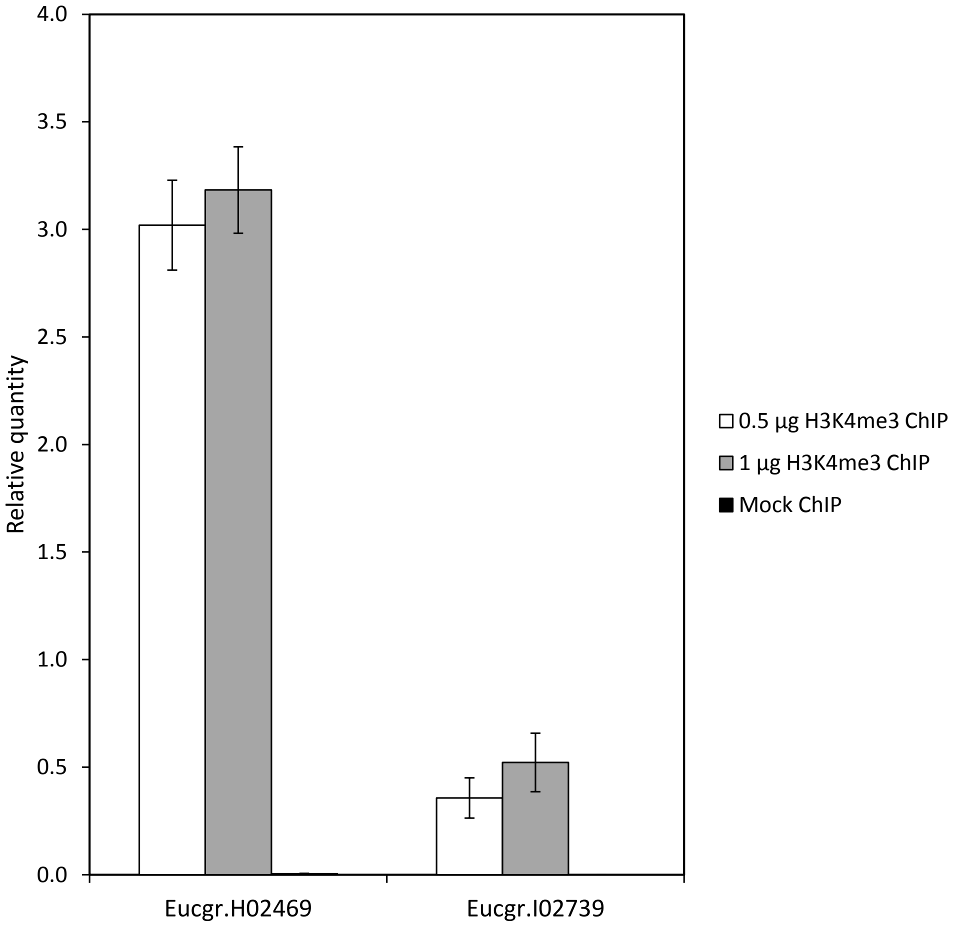
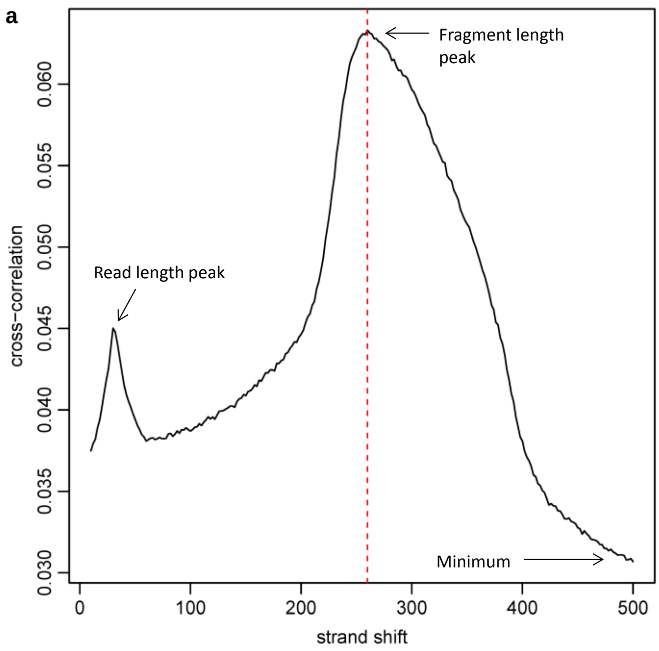


Fig. S4.8. ChIP-qPCR analysis of two candidate loci using different quantities of anti-H3K4me3 antibody. Values are expressed relative to input, where an equal quantity of template was used for qPCR analysis from each sample. Error bars indicate standard deviation of three technical replicates.

© University of Pretoria



b

$$\text{NSC} = \frac{\text{cc(fragment length)}}{\text{cc(min)}}$$

$$\text{RSC} = \frac{\text{cc(fragment length) - cc(min)}}{\text{cc(read length) - cc(min)}}$$

c

	NSC (1.05)	RSC (0.80)
V5 H3K4me3 (left reads)	4.33	2.28
V5 H3K4me3 (right reads)	5.33	2.29
V11 H3K4me3 (left reads)	6.33	1.72
V11 H3K4me3 (right reads)	7.33	1.60

Fig. S4.9. Strand cross-correlation (cc) analysis (Landt *et al.*, 2012) of mapped ChIP and input library reads. (a) H3K4me3 library (left reads) of individual V5 as representative of ChIP library samples, showing peaks corresponding to read length and fragment length. **(b)** Formulae for calculating normalized strand cross-correlation (NSC) and relative strand cross-correlation (RSC) values. **(c)** NSC and RSC values for all ChIP-seq samples, showing the minimum threshold preferred by ENCODE (Landt *et al.*, 2012) in parentheses. All H3K4me3 ChIP libraries yielded NSC and RSC values well above 1.05 and 0.8, respectively. Left and right reads were analyzed separately because paired-end reads yield high RSC values by default.

© University of Pretoria

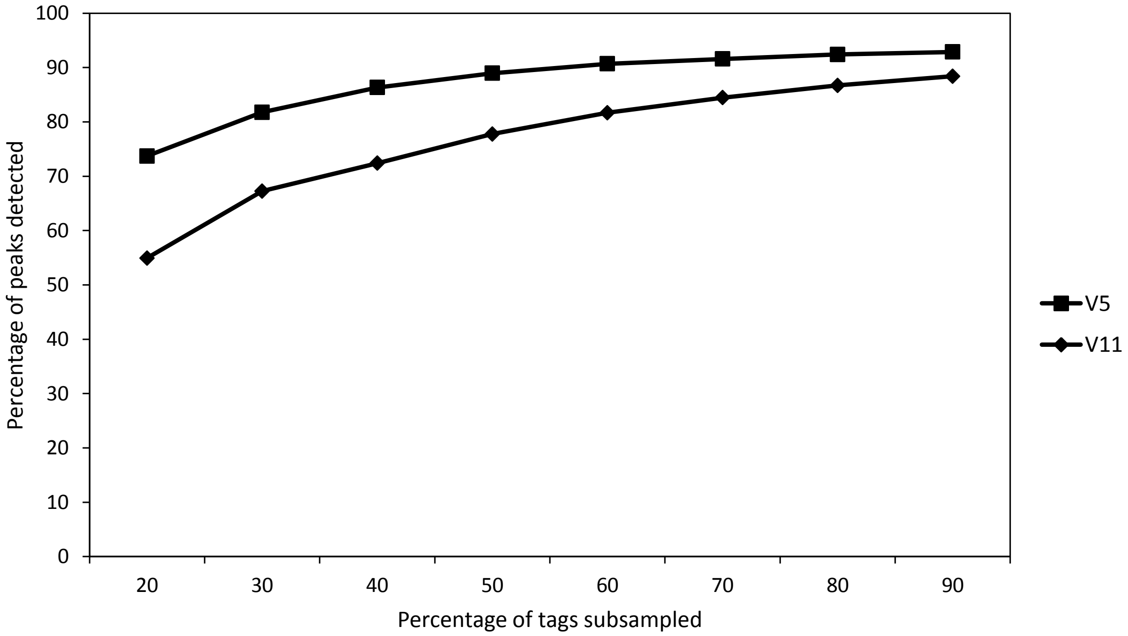


Fig. S4.10. MACS diagnostic analysis of H3K4me3 peaks (fold enrichment 5 - 15) detected for increasing proportions of subsampled tags. The percentage of peaks detected using all tags (*y*-axis) as a function of the proportion of the total tags used for peak detection (*x*-axis) is shown separately for individual V5 (blocks) and V11 (lines).

© University of Pretoria

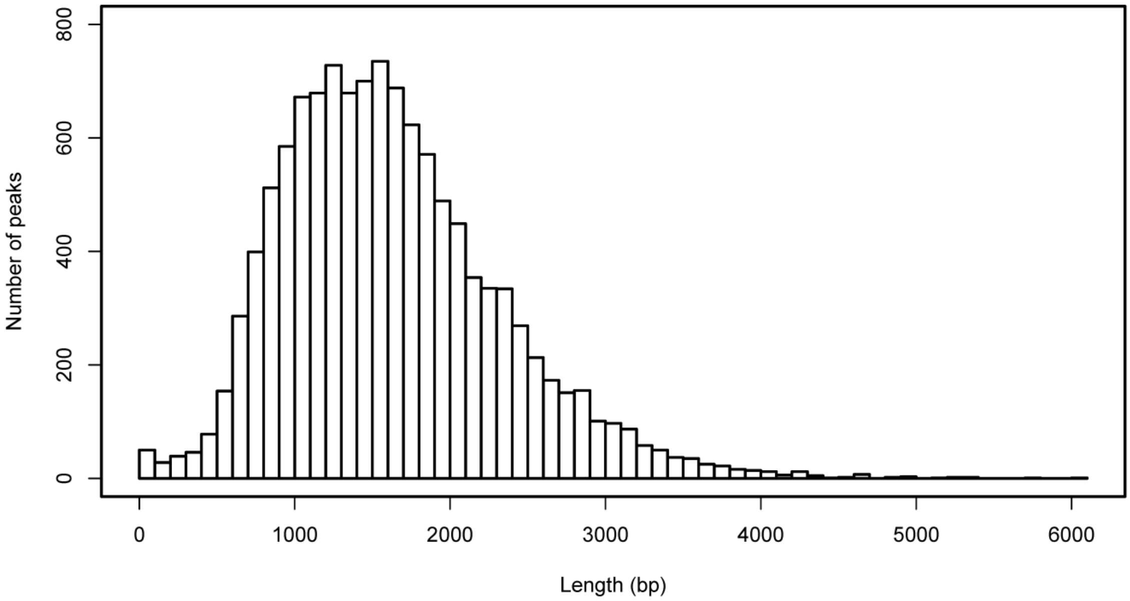


Fig. S4.11. Histogram of length distribution (bp) of significant H3K4me3 peaks.

© University of Pretoria

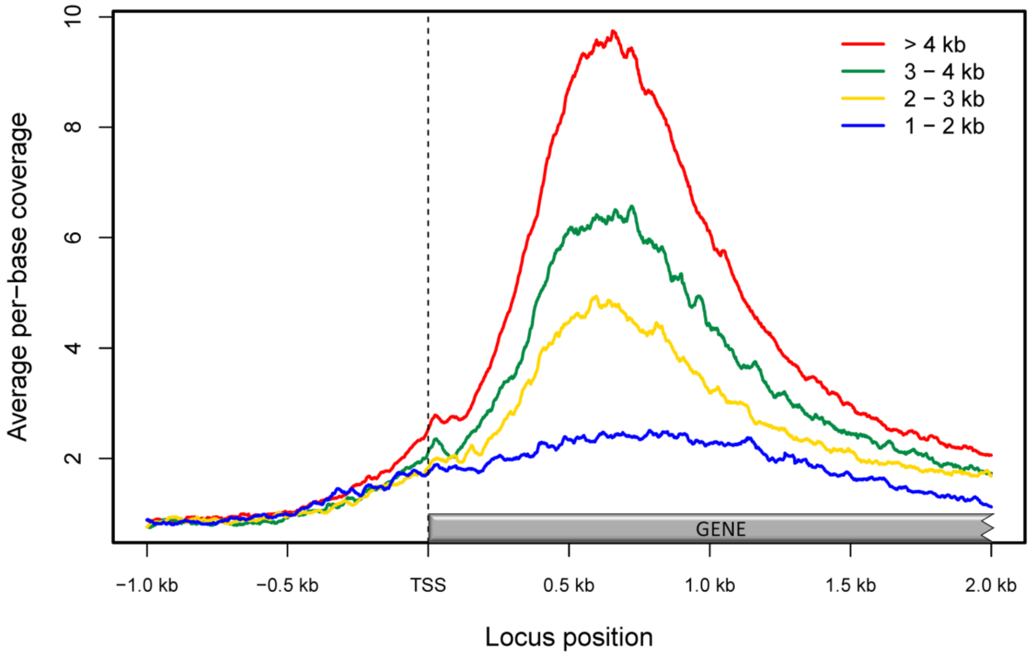


Fig. S4.12. Average H3K4me3 ChIP-seq library per-base read coverage across 5' gene regions of genes of various lengths. TSS, transcription start site.

© University of Pretoria

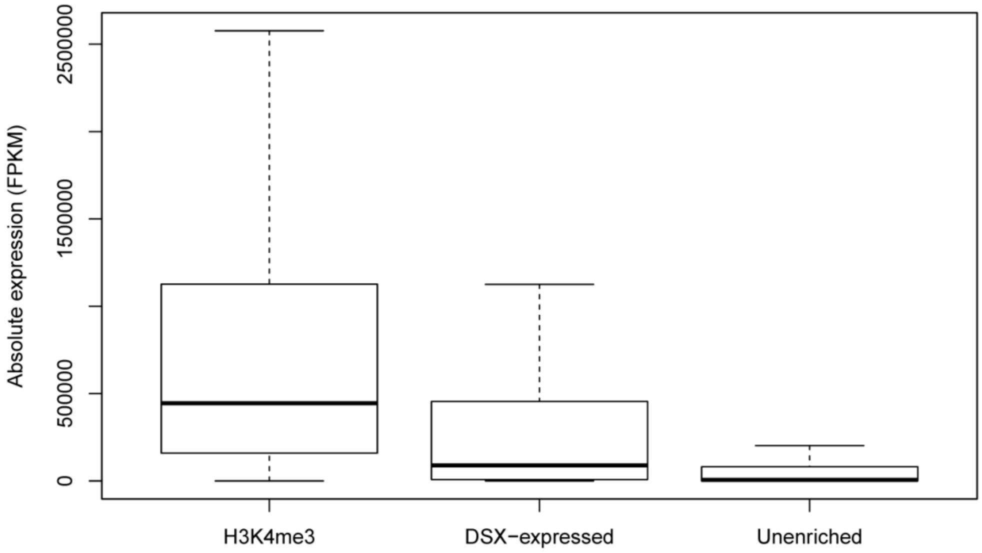


Fig. S4.13. Boxplot of absolute expression levels for H3K4me3-enriched and unenriched genes. Median FPKM (fragments per kilobase per million fragments mapped) values for genes trimethylated at H3K4 (H3K4me3; n = 9,760), genes expressed in developing secondary xylem (DSX-expressed; n = 27,595), or genes lacking H3K4me3 (Unenriched; n = 23,760) are indicated by the central bar. Outliers are not shown.

© University of Pretoria

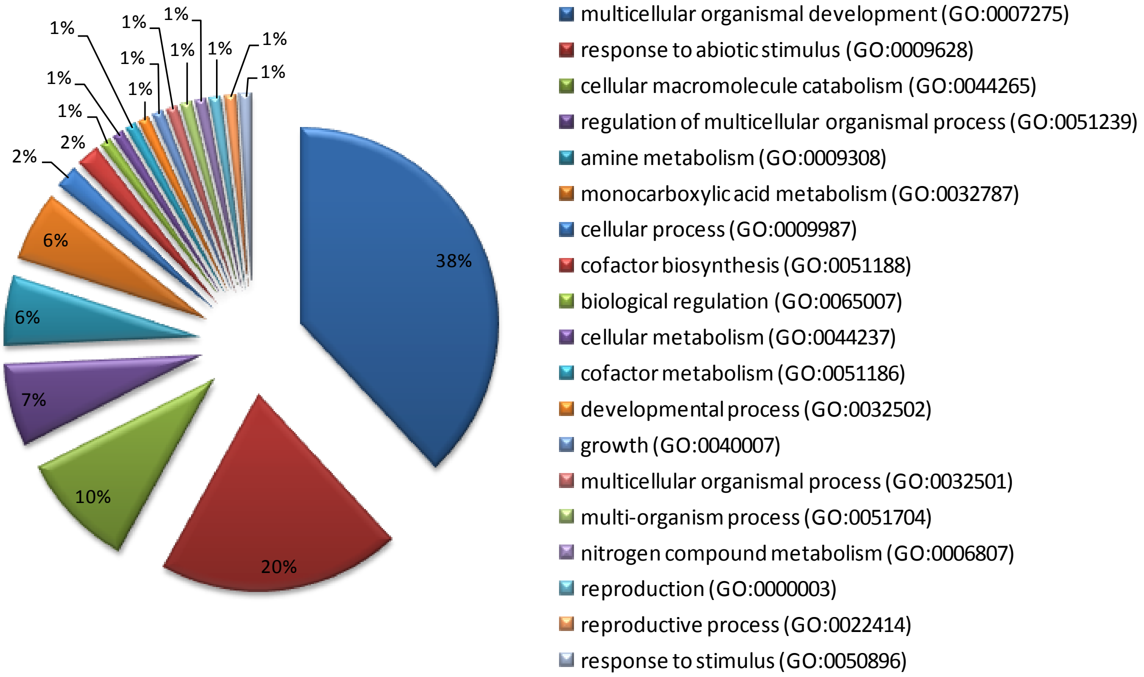


Fig. S4.14. REVIGO-summarized biological processes overrepresented among H3K4me3-enriched genes.

© University of Pretoria

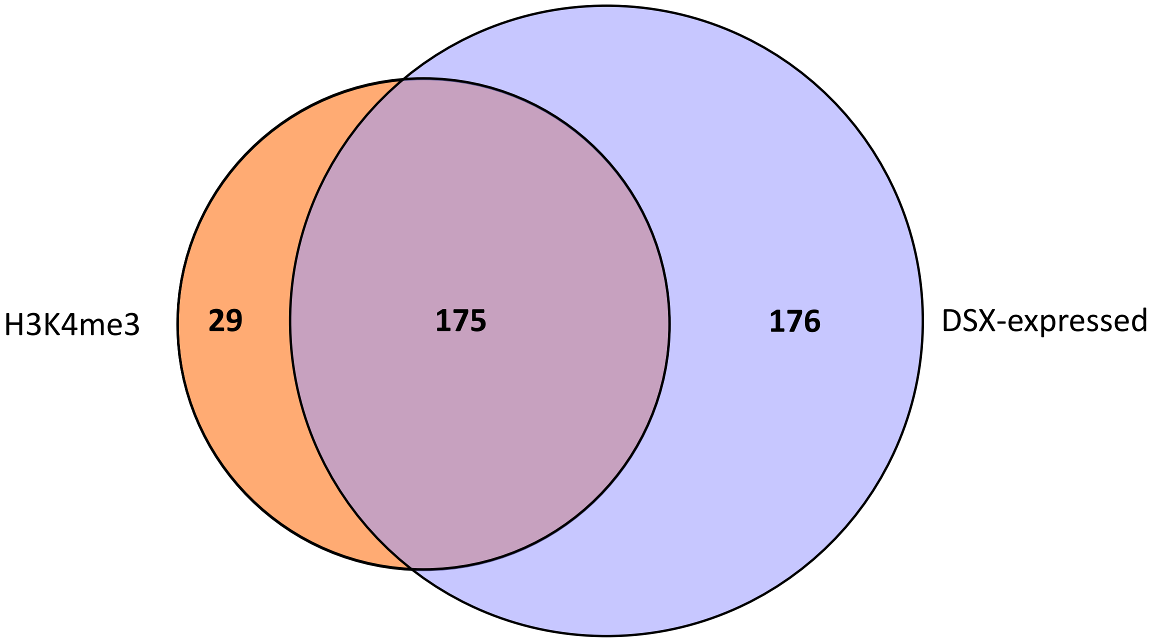


Fig. S4.15. Overlap of significantly overrepresented GO terms for biological function. Orange, terms overrepresented among H3K4-trimethylated genes; blue, terms overrepresented among developing secondary xylem (DSX)-expressed genes.

© University of Pretoria

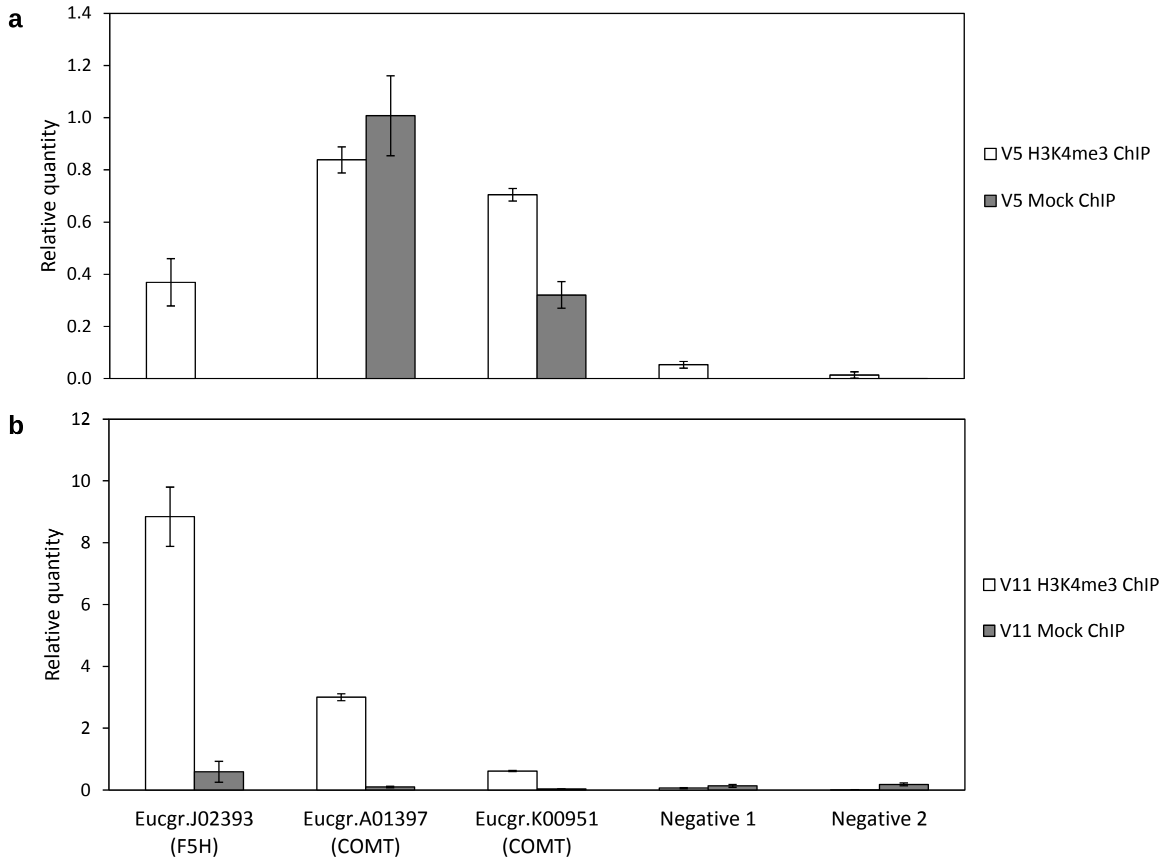


Fig. S4.16. ChIP-qPCR analysis of H3K4me3 at *E. grandis* homologs of *F5H* and *COMT*, **(a)** in individual V5, **(b)** in individual V11. The putative *Arabidopsis* ortholog of each candidate is indicated in parentheses. Two intergenic negative control regions are included. Error bars indicate standard deviation of three technical replicates.

© University of Pretoria

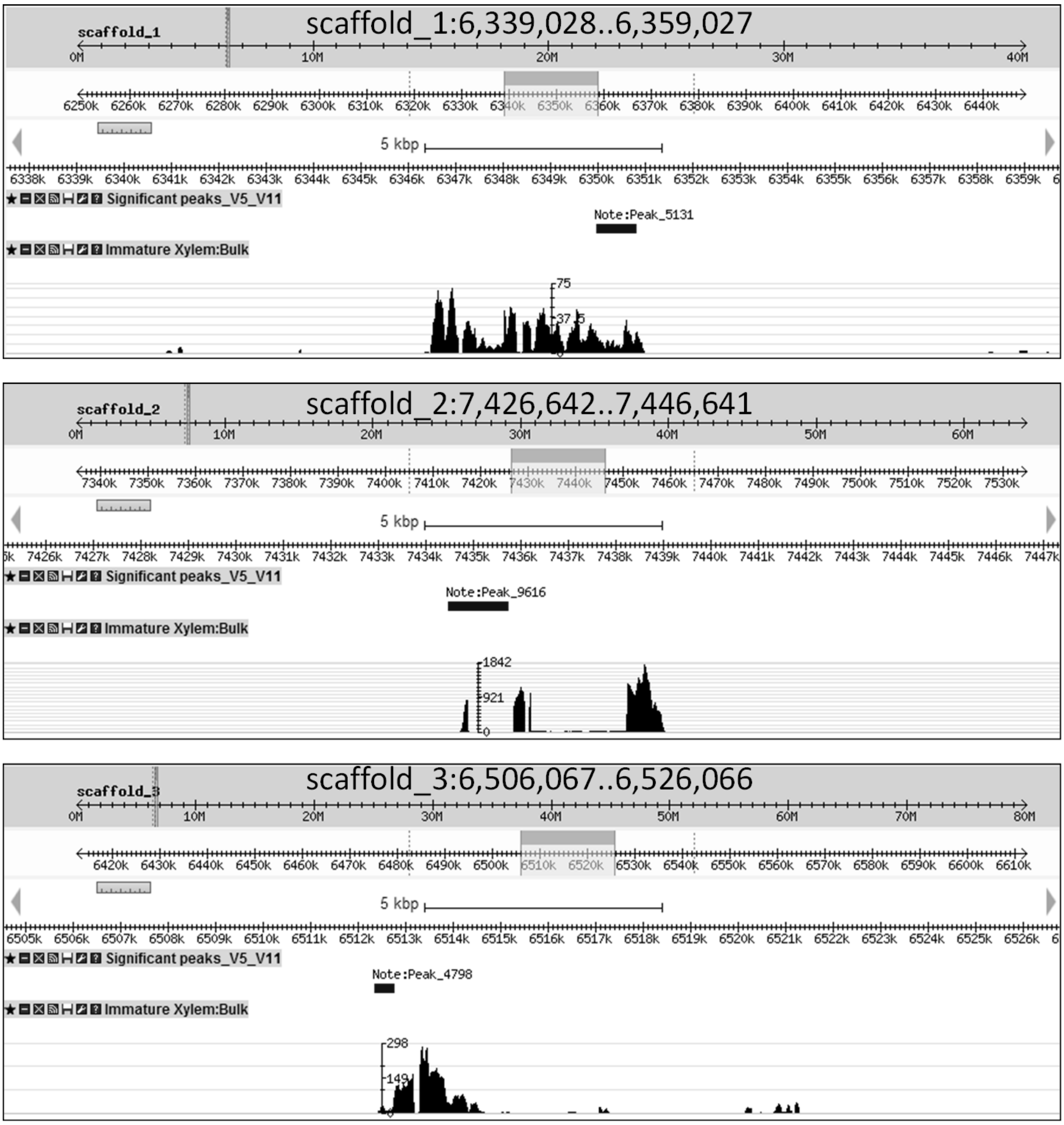


Fig. S4.17. Three examples of H3K4me3 peaks overlapping transcribed regions that have not yet been annotated. RNA-seq tags for developing secondary xylem (here referred to as "immature xylem") are indicated by the track "Immature Xylem: Bulk". H3K4me3 peaks are indicated by black bars in the "Significant peaks_V5_V11" track. Each window is 20 kb. Results were visualized in the EucGenIE browser (http://eucgenie.bi.up.ac.za/; Hefer *et al.* in preparation).

© University of Pretoria

4.12. Supplementary tables

Table S4.1. List of primers used for qPCR analysis

Locus	Annotation	Primers	Amplicon length (bp)
scaffold_8:33,795,824-33,795,938	Eucgr.H02469	5'-GATCGAGAGTTCGGCGCATA-3' 5'-CATACGCCACTGCAGGCAAT-3'	134
scaffold_9:38,403,997-38,403,868	Eucgr.I02739	5'-CATGCTCATTCTCCGCATGT-3' 5'-GTCACTTTCACCCTTCCTCTC-3'	149
scaffold_8:66,987,046-66,987,137	Eucgr.H04673	5'-CTGGCGTTGGATACAATGTT-3' 5'-TGGTGCTAAGAAGGTTGTCA-3'	111
scaffold_1:21,274,282-21,274,366	Eucgr.A01324	5'-GTGCAAATGACTCCCAAGAA-3' 5'-GATCAGATCACCGAGGACAA-3'	104
scaffold_3:5,079,522-5,079,592	Eucgr.C00246	5'-TTGAGCGTTATCGTCATCCT-3' 5'-GAGCAGAACAAGCACAAGTA-3'	90
scaffold_4:8,617,282-8,617,403	Eucgr.D00476	5'-GTTGCTCAGTCATGGCATTC-3' 5'-TAGGGCCTAAGACCAAACAC-3'	141
scaffold_10:3,764,427..3,764,590	Eucgr.J00384	5'-TAGCCGTGCAAGAGCCTCAT-3' 5'-TTCATCATCACCGCCATCGC-3'	183
scaffold_9:18,069,028..18,069,166	Eucgr.I00880	5'-GCACAATTCTGCTCCGATGA-3' 5'-AGTGCAAGGCTGTGAATCTC-3'	158
scaffold_10:38,339,144..38,339,239	Eucgr.J03126	5'-ATGGCCGCATTGAGATTGAC-3' 5'-AGCTTCCGAAGCTCCAATGT-3'	115
scaffold_5:11,792,297..11,792,357	Eucgr.E01107	5'-ATTCAGCGCATACACAACAA-3' 5'-GAAGTGTTCATGCGAGACAG-3'	80
scaffold_5:11,291,880..11,292,032	Eucgr.E01053	5'-TCACGTCCAAGTCGATCTTC-3’ 5'-GCTGAGCATACAGCTCGTTA-3'	172
scaffold_11:13,335,435..13,335,577	Eucgr.K01061	5'-ACCAAGACGACTATGCTAGAA-3' 5'-CACCGCCATCCAACAATAA-3'	161
scaffold_4:30,695,364..30,695,525	Eucgr.D01671	5'-GCAGCGTCCTGGATCAGATA-3' 5'-AATGTCCTCAAGCCGGTCTC-3'	181
scaffold_10:29,822,933..29,823,118	Eucgr.J02393	5'-GCGATCAAGATGTCGATACC-3' 5’-GGTGAACCGAGCAAGATTAG-3’	205
scaffold_1:22,485,326..22,485,482	Eucgr.A01397	5'-TGGTCCGCGTAATATGATGG-3' 5'-TGGTGGTGAGAATTGCAGAG-3'	176
scaffold_11:11,514,296..11,514,420	Eucgr.K00951	5'-ACTGCATCAGCATGTGGTAT-3' 5'-AGTGGCCGAGATCATTAAGT-3'	144
scaffold_8:58,195,203..58,195,428	Intergenic (Negative 1)	5'-CTCGACTGTGAAGAGCTATC-3' 5'-CAGAGTAGCCATTCTCAAGG-3'	245
scaffold_11:13,340,627..13,340,821	Intergenic (Negative 2)	5'-ATATGGTGTCACATTGCATCAG-3' 5'-ATCGGCTAATGTCTCAATCAAG-3'	216

© University of Pretoria

Table S4.2. ChIP-seq library sequence and mapping statistics. Suffixes V5 and V11 refer to the two individuals sampled.

Dataset	Read pairs after filtering	Read length after trimming	Read pairs uniquely mapped	% uniquely mapped
V5_Input	20,846,731	31 nt	11,464,953	55.0%
V5_H3K4me3	21,907,313	31 nt	5,707,963	26.1%
V11_Input	11,939,885	35 nt	6,709,264	56.2%
V11_H3K4me3	11,868,984	35 nt	3,741,400	31.5%
V11_IgG$_{2a}$	11,692,593	35 nt	1,144,387	9.8%

© University of Pretoria

Table S4.3. Putative targets of H3K4me3-regulated miRNAs in Table 4.1.

ncRNA class	ncRNA locus	Putative target locus	Target transcript	Strand	Description	Relative expression[a]
Known miRNA	scaffold_7:40120411-40120431					
Known miRNA	scaffold_7:46392607-46392626					
Predicted miRNA	scaffold_11:43275663-43275749	scaffold_10:7213656-7213730	-			
Predicted miRNA	scaffold_11:43275663-43275749	scaffold_8:33269803-33269878	-			
Predicted miRNA	scaffold_2:50779484-50779539	scaffold_7:44190105-44190125	Eucgr.G02574 (intron)	-	Nuclear-encoded CLP protease P7	0.10
Predicted miRNA	scaffold_3:4901554-4901642	scaffold_1:13630991-13631036	Eucgr.A00838 (exon)	-	Pentatricopeptide repeat (PPR) superfamily protein	0.13
Predicted miRNA	scaffold_7:40119641-40119697	scaffold_8:25401800-25401826	-			
Predicted miRNA	scaffold_9:28910581-28910654	scaffold_5:2945873-2945894	-			

[a] In *E. grandis* developing secondary xylem tissue

© University of Pretoria

4.13. Additional files

The following additional datasets are available on the supplementary CD-ROM disk attached to this thesis:

Additional file 4.1.bed (BED format): Genomic locations of significant H3K4me3 peaks.

Additional file 4.2.xlsx: Genomic locations of annotated genes overlapping with significant H3K4me3 peaks.

Additional file 4.3.xlsx: Genomic locations of low-confidence gene models overlapping with significant H3K4me3 peaks.

© University of Pretoria

Chapter 5

A pilot ChIP-seq analysis of the EgrNAC170 transcription factor, a homolog of SND2, in *Eucalyptus grandis* developing secondary xylem

Steven G. Hussey[1], Eshchar Mizrachi[1], Dave K. Berger[2], A.A. Myburg[1]

[1]Department of Genetics, [2]Department of Plant Science, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, South Africa

This chapter is part of an on-going study and has been formatted in accordance with the rest of the thesis. I performed most of the experimental work and data analysis and authored the manuscript. Afrah Khairalla, Marja O'Neil and Karen van der Merwe (University of Pretoria) performed the *E. grandis* clone SA1 genome resequencing, SNP calling and SNP substitution analyses. All co-authors have approved the manuscript.

© University of Pretoria

5.1. Summary

The transcriptional regulation of secondary cell wall biosynthesis is poorly understood in non-model woody species. The NAC domain transcription factor SND2 regulates a number of secondary cell wall biosynthetic genes in fibers of *A. thaliana*, but it is unknown whether it regulates similar genes in woody species. Here we aimed to identify genomic targets of EgrNAC170, a putative ortholog of SND2, in secondary developing xylem of *Eucalyptus grandis*. We employed the ChIP-seq approach optimized in Chapter 4 using two custom polyclonal anti-EgrNAC170 antibodies, mapped sequence reads to a custom *E. grandis* genome, identified ChIP-seq peaks and assigned peaks to putative gene targets. Both of the antibodies recognized the full-length recombinant protein, and despite low sequencing depth produced ChIP-seq binding peaks with a considerable degree of overlap. Of the 5,701 possible EgrNAC170 binding peaks, which were located preferentially in transcribed regions rather than the promoters of genes, over 3,200 target gene candidates were identified. However, these putative targets were enriched for few biological processes to a high degree of significance, and no clear biological function could be identified. Although EgrNAC170 putative targets were expressed significantly higher than all annotated genes in developing secondary xylem, they did not show evidence for preferential expression reflective of that of *EgrNAC170*, and only a small proportion had homologs that were differentially expressed following overexpression of *SND2* in *A. thaliana*. The preliminary results of this on-going study suggest a low signal to noise ratio and the requirement of further modification to the ChIP-seq procedure. Future work will aim to obtain biologically replicated, deeply sequenced ChIP-seq data and validate putative targets with using expression data and ChIP-qPCR.

© University of Pretoria

5.2. Introduction

Commercial forests are a major renewable source of timber, pulp, paper and chemical cellulose. *Eucalyptus* is the preferred commercial hardwood in tropical and temperate regions, occupying a global area exceeding 20 million hectares (Iglesias-Trabado & Wilstermann, 2008). While traditional and molecular breeding approaches have produced superior genetic clones of this recently domesticated genus (Grattapaglia *et al.*, 2012), genetic engineering of forest trees has also proved promising and potentially faster (Harfouche *et al.*, 2011). The latter approach requires a sound understanding of wood biology and its genetic regulation. Wood formation and transcriptional regulation studies in *Arabidopsis thaliana* (Zhang *et al.*, 2010; Hussey *et al.*, 2013) and *Populus trichocarpa* (Du & Groover, 2010; Zhong *et al.*, 2011), which are herbaceous and woody models respectively, have provided a good foundation for understanding xylem development within the context of contrasting evolutionary histories.

The development of secondary xylem involves a complex interplay between hormones, signalling proteins, transcriptional regulators and other regulating molecules (reviewed in Hussey *et al.*, 2013; Lucas *et al.*, 2013; Zhang *et al.*, 2014). It is now known that complex and highly interconnected transcriptional networks comprising NAC, MYB and other families of transcription factors (TFs) co-ordinate secondary cell wall (SCW) deposition and post-elongation maturation of fibers, vessels and other SCW-synthesizing cell types (reviewed in Zhong & Ye, 2007; Pimrote *et al.*, 2012). SCW transcriptional networks are largely conserved across angiosperms (Zhong *et al.*, 2010). However, the fact that some homologs of important SCW-associated NAC genes (e.g. *XND1*, *SND3*) have expanded in *Eucalyptus* (Chapter 2)

© University of Pretoria

and that many species in the genus, which is basal to the core rosids (Myburg *et al.*, in press), possess superior fiber properties, suggest that unique mechanisms regulating wood formation at the transcriptional level exists in this lineage.

It was previously shown that the *Arabidopsis* transcriptional activator SND2 regulates genes involved in, among others, cellulose and hemicellulose (xylan, mannan) biosynthesis, lignin polymerization and transcriptional regulation (Zhong *et al.*, 2008; Hussey *et al.*, 2011). Zhong *et al.* (2008) reported that SCW thickness of fibers, but not vessels, of *Arabidopsis* plants overexpressing or dominantly repressing *SND2* was increased and decreased respectively, while we observed a decrease in interfascicular fiber SCW thickness in overexpression lines (Hussey *et al.*, 2011; Chapter 3). In poplar trees, overexpression of a putative poplar *SND2* ortholog, *PopNAC154*, caused retarded growth, reduced xylem and increased cambium-phloem production (Grant *et al.*, 2010). The retarded growth phenotype resulting from *PopNAC154* overexpression was confirmed by Wang *et al.* (2013). When the protein was fused to a repression motif, Wang *et al.* (2013) also observed reduced growth, a reduction in xylem and phloem production as well as reduced SCW thickness, lignin and cellulose. Together, these results support two conclusions. First, that SND2 regulates SCW deposition in fibers in *Arabidopsis*, while a putative ortholog also plays a role in the regulation of SCW formation in *Populus.* Second, gain- and loss-of-function transgenic approaches, which are expected to yield contrasting phenotypes for transcriptional activators, have produced similar phenotypes in both *Arabidopsis* and *Populus* studies of *SND2* and *PopNAC154* respectively. This is thought to be related to a detrimental gene dosage effect brought about by overexpression (reviewed in Chapter 1; Hussey *et al.*, 2013), motivating the

© University of Pretoria

use of alternative approaches such as chromatin immunoprecipitation DNA sequencing (ChIP-seq) for functional genomics analysis of candidate TFs.

Based on a family-wide analysis of NAC domain proteins, it was previously shown that *E. grandis* likely possesses a single ortholog of SND2, EgrNAC170 (Chapter 2). Based on our knowledge of SND2 and the results of Chapter 3, we hypothesize that EgrNAC170 is evolutionarily conserved and regulates genes involved in SCW cellulose, xylan and mannan biosynthesis as well as lignin polymerization. In Chapter 4 we successfully applied the ChIP-seq method for studying the role of a modified histone, H3K4me3, in secondary developing xylem of *E. grandis* trees. Here, we aimed to use this ChIP-seq procedure to identify the genome-wide binding sites and gene targets of EgrNAC170 in *E. grandis* developing secondary xylem (DSX). The results of this on-going study suggest that further modification to the protocol and deeper library sequencing will be required to reliably infer binding events for EgrNAC170.

5.3. Materials and methods

5.3.1. Plant materials

This study used material from an *E. grandis* clone SA1 ramet (Mondi Tree Improvement Research, Hilton, South Africa), individual “V5”. A window was cut into the trunk of the seven-year-old individual growing in a plantation near KwaMbonambi, Kwazulu-Natal, South Africa at breast height, and the bark was removed. DSX tissue was scraped off to a depth of ~1 - 2 mm and flash-frozen in liquid nitrogen.

© University of Pretoria

5.3.2. Phylogenetic analysis

Predicted protein sequences of SND2 protein homologs in *E. grandis* and *A. thaliana* were obtained from Phytozome v.8 (www.phytozome.net) and aligned using MUSCLE in MEGA 5.05 (Tamura *et al.*, 2011). A maximum likelihood phylogeny was constructed in MEGA 5.05 using 5 discrete gamma rate categories, JTT amino acid substitution model, partial deletion of gaps and 1000 bootstrap iterations.

5.3.3. Generation of polyclonal peptide antibodies

Peptide sequences representing EgrNAC170 were designed by Genscript Inc. (Piscataway, NJ, USA). Criteria used for antigenic peptide generation included sequence conservation with other *E. grandis* homologs, surface accessibility, Kyte-Doolittle hydrophilicity (Kyte & Doolittle, 1982), Jameson-Wolf antigenicity index (Jameson & Wolf, 1988), secondary structure (Chou-Fasman method, Chou & Fasman, 1978; GOR method, Garnier *et al.*, 1996), and linear epitope conformation. Additionally, protein structure models of EgrNAC170 were generated using I-TASSER (Roy *et al.*, 2010) and the top ranking model used to visualize the three-dimensional location of predicted antigenic peptides using the PyMOL Molecular Graphics System, Version 1.7, Schrödinger, LLC. Keyhole Limpet Hemocyanin (KLH) carrier protein was conjugated via Cysteine-mediated amidation to the C- and N-terminus respectively of two synthetic peptides, NDNKSDEQRNESAT and SGHENANLKNN. Peptide synthesis and conjugation, rabbit inoculation, polyclonal antibody extraction, affinity purification and ELISA validation using the synthetic peptides were performed by Genscript, Inc.

© University of Pretoria

5.3.4. Heterologous protein expression and Western blot analysis

Primers (forward 5’-ATGACTTGGTGCAATAATGAC-3’, reverse 5’-TCAAGAGACAAAAGAAGACCCACCATGGA-3’) were designed to target the predicted coding sequence of EgrNAC170 (Eucgr.K01061.1; www.phytozome.net/Eucalyptus.php). cDNA from *E. grandis* DSX tissue was used as template for coding sequence amplification using high-fidelity Phusion *Taq* (NEB, MA), and the amplicon was cloned into the entry vector pCR8/GW/TOPO (Invitrogen, OR) as per manufacturer’s instructions. The insert was transferred to the pET160 expression vector (Invitrogen) using the Gateway LR Clonase^{TM} II Enzyme Mix (Invitrogen) and sequenced. Synonymous polymorphisms were permitted. *Eshcherichia coli* strain BL21Star (Invitrogen) was transformed with the plasmid as per the Champion pET Gateway Expression Kit instructions (Invitrogen), and expression was induced for 2 hours at 37°C with shaking using 0.1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG). The cell pellet was resuspended in protein extraction buffer (10 mM sodium phosphate buffer pH 7.0, 150 mM NaCl, 0.1 mM EDTA, 5% glycerol, 10 mM β-mercaptoethanol, 0.1 mM PMSF, Roche Complete Protease Inhibitor Cocktail), briefly sonicated to lyse the cells, and centrifuged at 16,000 × *g* for 10 min (4°C). The protein concentration of the supernatant was quantified using the Qubit Protein Assay Kit (Invitrogen). Western blot analysis was performed using ~30 µg protein as described in Chapter 4.

5.3.5. Whole genome resequencing

A custom *E. grandis* reference genome containing single nucleotide polymorphisms (SNPs) represented in clone SA1 but absent from the BRASUZ1 reference

© University of Pretoria

(http://www.phytozome.net/Eucalyptus.php), was constructed (Myburg *et al.*, unpublished). Briefly, 100-base paired-end reads (Beijing Genome Institute, Hong Kong) derived from separate leaf and xylem genomic DNA samples from clone SA1 were pooled and mapped to the BRASUZ1 reference (Phytozome V.8, http://www.phytozome.net) using the Burrows-Wheeler Alignment (BWA) tool (Li & Durbin, 2009), to a mean coverage of 15.9X. Local realignment (DePristo *et al.*, 2011) was performed to correct indel-induced misalignment, duplicate reads were removed (Mark Duplicate Reads v1.56.0; http://picard.sourceforge.net/index.shtml) and an alignment pileup (BCF format) was constructed with MPileup (Li *et al.*, 2009) followed by conversion to Variant Calling Format (VCF) using SAMtools (Li *et al.*, 2009). Called SNP variants were filtered using the VCF Tool varFilter using parameters: minimum Root Mean Square (RMS) mapping quality, 20; minimum read depth, 20. Where called SNP variants were homozygous in relation to the BRASUZ1 reference, or heterozygous but both alleles differed from the BRASUZ1 reference, the BRASUZ1 reference was substituted accordingly with the SA1 variant using a customized script.

5.3.6. ChIP-seq analysis

E. grandis DSX chromatin was crosslinked, purified and sonicated as described in Chapter 4 (section 4.3.2). Since the same samples from individual V5 were used as described in Chapter 4, the V5 input control library sequenced in Chapter 4 was used as the negative control. ChIP was performed as described in Chapter 4 (section 4.3.6), using two technical repetitions per ChIP which were then pooled after elution. ChIP DNA amplification using our modifications to the protocol by Adli and Bernstein (2011) and DNA sequencing (Beijing Genome Institute,

© University of Pretoria

Hong Kong) were performed as described in Chapter 4. 20 - 25 ng of amplified, 3' adenylated template was used for Illumina library preparation. Data filtering and read mapping were performed as described in Chapter 4 (section 4.3.7), using the resequenced *E. grandis* clone SA1 customized reference genome (Myburg *et al.*, unpublished; section 5.3.5) to improve mapping efficiency of sequenced DNA isolated from this clone. Only uniquely mapped reads were retained to avoid PCR-induced bias. ChIP-seq peaks were called using MACS (Zhang *et al.*, 2008), using a threshold of 3-fold peak enrichment relative to control, P-value < 10^{-5} and at least 10 mapped tags per peak. Only peaks located on scaffolds (chromosomes) one to eleven were considered.

5.3.7. Bioinformatics analysis

FastQC (http://www.bioinformatics.babraham.ac.uk/) was used to calculate overrepresented k-mers and sequence duplication rates of raw sequence data. SPP (Kharchenko *et al.*, 2008) was used for strand cross-correlation analysis. Genomic feature overlaps were identified using BEDTools (Quinlan & Hall, 2010). For gene ontology analysis, a reference control dataset was constructed from the top BLASTP hits of all primary *E. grandis* gene models (v.1.1. annotation; http://www.phytozome.net/Eucalyptus.php) to the *Arabidopsis thaliana* genome, in GOToolBox (Martin *et al.*, 2004). Significantly enriched biological processes (P < 0.05 after Hochberg correction for multiple testing) among EgrNAC170-associated genes were identified using a hypergeometric test against the reference control set in GOToolBox (Martin *et al.*, 2004).

© University of Pretoria

5.4. Results

5.4.1. EgrNAC170 is a putative *E. grandis* ortholog of SND2 from *Arabidopsis*

Based on a maximum likelihood phylogenetic analysis of the conserved NAC domains of *A. thaliana*, *E. grandis*, *P. trichocarpa*, *Vitis vinifera* and the monocot outgroup *Oryza*, we identified EgrNAC170 (Eucgr.K01061.1) as the putative ortholog of *Arabidopsis* SND2 (also known as ANAC073) in *E. grandis* (Chapter 2). A targeted phylogenetic analysis of the SND2/SND3 clade from Chapter 2 (Fig. S5.1a) confirmed that EgrNAC170 is the most likely possible ortholog of SND2 (Fig. S5.1b). RNA-seq coverage in DSX and bulked tissues, as visualized in EucGenIE (http://eucgenie.bi.up.ac.za/; Hefer *et al.* in preparation), supported the predicted gene model and showed no evidence for alternative splicing (Fig. S5.2). PCR amplification of the predicted EgrNAC170 coding sequence (891 bp) from *E. grandis* clone SA1 xylem and twig cDNA showed that only amplification from xylem cDNA was successful, yielding a single amplicon (Fig. S5.3). This amplicon was cloned, sequenced (Additional file 5.1) and found to contain three synonymous mismatches compared to the BRASUZ1 reference. Two of these were natural polymorphisms present in the *E. grandis* clone SA1 genome (discussed below); the other was a synonymous, likely PCR-induced mutation (not shown).

The EgrNAC170 coding sequence was predicted to encoded a 33.6 kDa, 296 amino acid residue protein with a theoretical isoelectric point of 9.29, showing 59.6% identity and 69.9% similarity to SND2 (Fig. 5.1). Subdomains A through E of the NAC domain, comprising EgrNAC170 residues 62 - 223, were largely conserved between the two putative orthologs, while the C-terminal region was variable and marked by indels in both EgrNAC170 and

© University of Pretoria

SND2 (Fig. 5.1). The first 58 residues in the N-terminal region preceding the NAC domain in EgrNAC170, which is unusually long in EgrNAC170 and its closest homologs compared to other EgrNAC proteins (see Chapter 2, Fig. 2.2) was poorly conserved with SND2 aside from the sequence TCPSCGH (positions 40 - 46; Fig. 5.1).

SND2 is preferentially expressed in tissue and cell types marked by SCW deposition in *Arabidopsis*, including inflorescence and hypocotyl stems, as well as protoplasts derived from the vasculature of the root (Fig. S5.4). The expression profile of *EgrNAC170* was similarly biased toward SCW-synthesizing DSX tissue of *E. grandis* trees growing in the field (Fig. 5.2). In this tissue, *EgrNAC170* transcript level was among the top 10% of genes (not shown). The expression profile of *EgrNAC170* suggests that this gene, like its putative ortholog *SND2*, plays a role in regulating SCW deposition.

5.4.2. Assessment of polyclonal anti-EgrNAC170 antibodies for ChIP-seq

We generated two polyclonal antibodies against EgrNAC170, anti-EgrNAC170-1 and anti-EgrNAC170-2, using two different synthetic peptides from regions of poor sequence conservation with other EgrNAC homologs (Fig. S5.5). Based on the predicted protein structure of EgrNAC170, the peptides were located in accessible loops of the protein, away from the conserved DNA-binding NAC domain (Fig. S5.6). Both antibodies recognized the full-length recombinant EgrNAC170 protein expressed in *Escherichia coli* as determined by Western blot analysis (Fig. S5.7). Additional assessment of antibody specificity via Western blotting is advisable for ChIP-seq. However, we could not detect a signal in Western blots of *E. grandis* DSX nuclear protein extracts, possibly because of low target protein abundance.

© University of Pretoria

Therefore, antibody fidelity was assessed by observation of a significant overlap between peaks called using the two different antibodies raised against EgrNAC170, as specified in the Encyclopedia of DNA Elements (ENCODE) Consortium guidelines (Gerstein *et al.*, 2012; Landt *et al.*, 2012).

5.4.3. Identification of EgrNAC170 binding sites in DSX tissue

A ChIP-seq analysis of *in vivo* EgrNAC170 genomic binding sites in DSX tissue was performed using the anti-EgrNAC170-1 and anti-EgrNAC170-2 antibodies. To avoid the high duplication rates and inefficient reference-based read mapping observed in H3K4me3 ChIP-seq libraries due to ChIP DNA amplification (Chapter 4), we first attempted library construction from ~2 ng ChIP DNA as template. Library construction was not successful using this strategy and the ChIP DNA amplification strategy explained in Chapter 4 was therefore adopted prior to library preparation (yields show in Table S5.1). Between 13.1 and 14.8 million total filtered reads were generated per ChIP library (Table S5.2). These were trimmed of primer sequences and further trimming was performed on a case-by-case basis when overrepresented k-mers were still observed at the 5’ end, after which reads were mapped to an *E. grandis* clone SA1 resequenced reference genome. Despite excellent read quality (Fig. S5.8), estimated sequence duplication rates were high and unique mapping efficiencies of the EgrNAC170 ChIP libraries were considerably lower than those obtained for H3K4me3 libraries in Chapter 4, ranging from 7.8% to 12.6% unique mapping (Table S5.2). Thus, a relatively low coverage of uniquely mapped reads was obtained for peak-calling analysis.

© University of Pretoria

A strand cross-correlation (SCC) analysis (Landt *et al.*, 2012) was performed to assess the ChIP efficiency for each dataset. This approach quantifies the degree of clustering of reads flanking true binding sites by shifting the reverse strand relative to the forward strand one base at a time and assessing the correlation in read coverage between the opposing strands. For a specified shift interval, the strand correlation of enriched reads flanking a true binding site results in an SCC peak corresponding to a shift equivalent to the average chromatin fragment length. In contrast, reads originating from non-enriched DNA will map more uniformly across the genome, yielding an SCC peak corresponding to a shift equivalent to the read length (Landt *et al.*, 2012). ChIP efficiency can be assessed by the relative strand cross-correlation (RSC), defined as the ratio of the background-subtracted fragment length SCC peak to that of the background-subtracted read length SCC peak (Landt *et al.*, 2012). In the EgrNAC170 ChIP-seq datasets, the ChIP efficiency was suboptimal but acceptable, with RSC values of ~0.52 for anti-EgrNAC170-1 and somewhat better for anti-EgrNAC170-2 (RSC ~0.63) ChIP datasets (Fig. S5.9).

Next, we investigated whether ChIP-seq peaks called separately for each anti-EgrNAC170 antibody overlapped significantly. Peaks were confined to those on the main assembled scaffolds one to eleven, and any peaks overlapping with peaks called in a nonspecific IgG ChIP control (Chapter 4) were excluded. For anti-EgrNAC170-1, 3,945 peaks were identified, all of which had a false discovery rate (FDR) < 2%, while for anti-EgrNAC170-2 some 5,126 peaks were called, all with FDR < 1%. The total proportion of the genome covered by peaks was ~0.58% and ~0.56%, respectively. 639 peaks common to anti-EgrNAC170-1 and anti-EgrNAC170-2 were identified, comprising ~16% of the anti-

© University of Pretoria

EgrNAC170-1 dataset. Furthermore, around 29% of the top 500 significance-ranked peaks identified using anti-EgrNAC170-1 were represented in the anti-EgrNAC170-2 dataset. These results demonstrated a considerable overlap in target sites identified by the two anti-EgrNAC170 antibodies, especially considering the low sequence coverage. Based on this observation, we pooled the mapped reads from anti-EgrNAC170-1 and EgrNAC170-2 libraries in order to increase the coverage for peak-calling and thus the reliability of the analysis. This treatment identified 5,701 peaks (FDR < 5%; median FDR = 1.94), capturing all of those identified using anti-EgrNAC170-1 and anti-EgrNAC170-2 datasets individually as well as 979 peaks not detected in the separate analyses. The median peak fold enrichment of ~8.6 is within the typical range of 5 – 13 for ENCODE data (Landt *et al.*, 2012). We performed a diagnostic analysis of the number of peaks detected as a function of increasing sequencing depth to assess the peak-calling sensitivity at the given sequencing depth. This relationship was largely linear and showed no sign of reaching a plateau at the maximum sequencing depth of this experiment (Fig. S5.10), suggesting that EgrNAC170 has significantly more binding sites in the genome that will require deeper sequencing to detect.

5.4.4. Identification of direct gene targets of EgrNAC170

The summits of the detected ChIP-seq peaks were mostly located in intergenic regions (~58%), but a surprisingly large percentage were located in introns and exons (~33%) (Fig. 5.3a). Only ~5% of peak summits were located in promoter regions, defined as 1 kb upstream from the transcription start site (TSS) (Fig. 5.3a). To guide criteria for assigning target genes to ChIP-seq peaks, the binding profile of EgrNAC170 was constructed to understand the binding preference of this TF relative to genes and the TSS. The percentage of ChIP-seq peaks

© University of Pretoria

overlapping at single-base positions spanning ten kilobases upstream and downstream of annotated TSSs in the *E. grandis* clone SA1 genome showed a prominent, sharp peak in the vicinity of the TSS, rising above the background level ~500 bp upstream of the TSS, and peaking shortly after the TSS within the gene model (Fig. 5.4). Based on this profile, it would not be appropriate to define target genes as those with peaks located in the promoter region alone. We therefore defined putative target gene models using a simple but reasonable criterion that any gene model, including two kilobases upstream and one kilobase downstream of the annotated transcribed region, overlapping a ChIP-seq peak was a potential direct target. Peaks overlapping more than one gene model were included, and genes overlapping with more than one peak were considered equally likely candidates as those overlapping only one peak. This procedure identified 3,234 candidate genes (Additional file 5.2).

To better understand the functions of these candidate targets, overrepresented biological processes among their closest *Arabidopsis* homologs were identified. Since several gene families have expanded in *E. grandis* relative to *Arabidopsis* (Myburg *et al.*, in press), the top BLASTP hits of the entire *E. grandis* annotation v.1.1 (Phytozome V.8, http://www.phytozome.net) against *A. thaliana* were used as the reference dataset to avoid potential bias. Among the 218 significantly enriched biological processes ($P* < 0.05$, where $P*$ is an adjusted P-value;), secondary metabolism and biotic and abiotic stress response gene ontology (GO) terms were among the top ten significant terms ($P* < 0.002$) among putative EgrNAC170 targets (Additional file 5.3). There was little evidence for specific significant enrichment of terms relating to the regulation of xylogenesis and SCW formation, although phenylpropanoid metabolism ($P* = 0.0079$) and biosynthesis ($P* = 0.0080$), ethylene

© University of Pretoria

metabolism (*P** = 0.013), lignin biosynthesis (*P** = 0.014), regulation of transcription (*P** = 0.017), response to brassinosteroid stimulus (*P** = 0.029), xylan catabolism (*P** = 0.03) and carbohydrate metabolism (*P** = 0.046) were enriched among all significantly enriched biological processes (Additional file 5.3). Terms related to (secondary) cell wall biosynthesis or modification were not significantly enriched. Results were similar when we analyzed the top 500 genes ranked according to the increasing *P*-value of their accompanying peaks, which should represent high-confidence targets (not shown). As a negative control, we analyzed the GO enrichment of 1,597 genes that we assigned to nonspecific IgG ChIP-seq peaks identified in Chapter 4 using the same criteria in this study. These IgG-associated genes were significantly enriched for secondary metabolism, several of the stress response terms, and phenylpropanoid metabolism/biosynthesis (not shown), suggesting that phenylpropanoid metabolism and biosynthesis is not specifically enriched among EgrNAC170 putative targets. Furthermore, most of the *E. grandis* homologs of phenylpropanoid-associated genes identified as possible EgrNAC170 targets had little or no expression in DSX tissue (Table S5.3), and are thus unlikely to be the functional homologs of enzymes of the monolignol pathway (Carocha *et al.*, in preparation). Among homologs of SCW-associated biosynthetic genes, several putative EgrNAC170 targets with high expression and/or preferential in DSX were detected, but these appeared to be isolated cases with no obvious representation of cellulosic or hemicellulosic biosynthetic pathways (Table S5.4). Together, these results shed some doubt on the reliability of the ChIP-seq data.

To further assess the plausibility of the putative EgrNAC170 gene targets, we analyzed their expression patterns using existing *E. grandis* RNA-seq data (http://eucgenie.bi.up.ac.za/;

© University of Pretoria

Hefer *et al.* in preparation). 469 (14.5%) of the putative EgrNAC170 target genes had no detected expression in DSX tissue, a considerably smaller proportion compared to the expression values of all *E. grandis* genes in this tissue (23.1%). The absolute transcript abundance of EgrNAC170 putative targets was significantly higher than those of all *E. grandis* genes in DSX tissue (Kolmogorov-Smirnov test; $P < 10^{-16}$), and appeared even higher among the top 500 EgrNAC170 putative targets ranked according to their assigned peak significance (Fig. 5.5a). We next analyzed the relative expression of EgrNAC170 putative targets in DSX tissue. Given that *EgrNAC170* is preferentially expressed in DSX (Fig. 5.2), we can expect that *bona fide* targets should also be preferentially expressed (if EgrNAC170 activates their expression) or significantly downregulated (if EgrNAC170 represses their expression) in DSX tissue relative to other tissues. For genes expressed in DSX tissue, we found no significant difference in relative expression of EgrNAC170 putative targets compared to all genes expressed in DSX tissue (Fig. 5.5b). The top 500 likely EgrNAC170 target candidates ranked according to the significance value of their assigned binding peaks appeared to have slightly higher preferentially expression in DSX compared to all expressed genes, but this too was not significant (Fig. 5.5b). Finally, we tested whether EgrNAC170 putative target genes showed greater tissue-specificity compared to all annotated genes, based on Shannon entropy of transcript abundance across seven tissues (Shannon, 1948; Schug *et al.*, 2005; see Chapter 4). The entropies of EgrNAC170 putative targets were not significantly different from those of known *E. grandis* genes (Fig. 5.5c).

As a final evaluation of the reliability of the putative gene targets identified using ChIP-seq, we calculated the proportion of genes with closest *A. thaliana* homologs that were

© University of Pretoria

differentially expressed as a result of *SND2* overexpression (Chapter 3). Only 96 *A. thaliana* homologs of putative EgrNAC170 targets were differentially expressed in the *SND2* overexpression data, comprising ~3% of the 3,234 putative targets. This value is negligibly larger than the proportion expected by chance (~2%). Applying more stringent filters to the ChIP-seq peaks, such as increasing the significance level, FDR threshold, number of tags or fold enrichment did not significantly improve the proportional overlap.

5.5. Discussion

The identification of *in planta* genomic targets of xylogenesis-associated TFs is a technically challenging prospect. In this study, we aimed to identify putative direct gene targets of EgrNAC170 in DSX tissue using ChIP-seq. We showed that EgrNAC170 is the most likely ortholog of *Arabidopsis* SND2, which regulates fiber SCW deposition (Chapter 3; Zhong *et al.*, 2008). The observation that *EgrNAC170* is preferentially expressed in DSX at a high level, similarly to the preferential expression of *SND2* in SCW-enriched tissues, suggests a role for EgrNAC170 in the regulation of xylogenesis and/or SCW biosynthesis. This is further supported by a previous study showing *E. grandis* homologs of several SCW-regulating TFs to bind the *EgrNAC170* promoter in yeast one-hybrid assays, among them homologs of VND6, MYB46, MYB83, SND3 and C3H14 (Botha *et al.*, in preparation). According to this hypothesis, we would expect an enrichment of biological functions relating to SCW biosynthesis among direct targets of EgrNAC170. In this pilot ChIP-seq study, we found little evidence for a successful experiment. Technical considerations and improvements for future ChIP-seq analyses are discussed further.

© University of Pretoria

Since library preparation from the small yield of eluted DNA obtained from the ChIP assay was not successful, we applied the nano-ChIP-seq DNA amplification procedure (Adli & Bernstein, 2011) modified in Chapter 4 to facilitate library preparation. The disadvantages of this approach, as echoed in Chapter 4, were also evident in this study, with high PCR-induced sequence duplication levels and low unique mapping rates (Table S5.2). Future experiments may have to rely on library construction using pooled template from several ChIP enrichments to ensure successful library preparation. While ChIP efficiency (as assessed using strand cross-correlation analysis) was also suboptimal, the cause is difficult to ascertain. It could lie with properties inherent to the antibodies (only around ~20% of antibodies perform well in ChIP-seq; Landt *et al.*, 2012), antigen exposure of the target protein or low sequencing depth. The anti-EgrNAC170-2 antibody appeared to produce greater enrichment and would be a better candidate for future ChIP-seq experiments. The number of EgrNAC170 binding peaks called and the number of potential target genes identified was surprisingly large considering these limitations. However, based on a critical assessment of these putative targets, it is possible that most of them represent random noise due to low quality data and poor representation of immuno-enriched DNA. Indicators suggesting this include (1) the negligible overlap with homologs differentially expressed during overexpression of the putative ortholog of *EgrNAC170, SND2*, in *Arabidopsis*, (2) the finding that EgrNAC170 putative targets were no more likely to be preferentially expressed in DSX tissue relative to expressed genes, despite considerable tissue-specific expression of EgrNAC170, and (3) the lack of related biological function GO terms with highly significant *P*-values, that were absent from the negative control dataset. The latter may not strictly suggest a failed experiment, since taking into account only gene-proximal peaks can result in underrepresentation of biologically

© University of Pretoria

relevant processes and bias resulting from non-uniform gene distribution (McLean *et al.*, 2010). Additionally, there were also indicators that at least some of the enriched regions identified represent actual binding sites, such as the significantly higher expression of EgrNAC170-enriched genes relative to those annotated in the genome (Fig. 5.5a) and the considerable overlap of peaks called independently using different anti-EgrNAC170 antibodies. Also, while the genomic locations of peak summits for EgrNAC170 (Fig. 5.3a) detracts from the conventional expectation for TFs to bind predominantly in promoter regions, the profile is similar to that of a recently published study analysing deeply sequenced ChIP-seq libraries of an unrelated soybean NAC protein (Shamimuzzaman & Vodkin, 2013) (Fig. 5.3b). Thus, it is likely that our experiment detected many actual EgrNAC170 binding sites, albeit at low sensitivity and accompanied by a large number of "noise" sites.

The expected number of target sites and target genes varies from one DNA-binding protein to another, and ChIP-seq data can be difficult to interpret given the fact that binding sites may not necessarily be functional (Graur *et al.*, 2013). Complementing information of putative binding sites with expression data from loss- or gain-of-function mutagenesis studies of the candidate gene can be used to identify functional binding sites. For example, in one of the first plant ChIP-seq studies, Kaufmann *et al.* (2010) identified 2,298 putative target genes for the MADS-box TF AP1, but less than half of these were differentially expressed following induced expression of AP1 in transgenic plants, and only ~250 were differentially expressed more than 1.8-fold. Thus, while library construction from non-amplified immuno-enriched DNA, increased sequencing depth and biological replication are likely to narrow down the

© University of Pretoria

number of regions occupied by EgrNAC170 *in planta*, the effective elucidation of its direct targets will require integration of expression data for EgrNAC170.

5.6. Conclusion

In this preliminary ChIP-seq study of EgrNAC170 direct gene targets, we identified thousands of binding sites and putative gene targets, but insufficient evidence to confidently describe the genome-wide binding sites of EgrNAC170. We attribute low sequencing depth and suboptimal ChIP enrichment to the generation of a possibly high degree of noise. Despite this, we found a considerable overlap between peaks identified using two independent antibodies against EgrNAC170, and EgrNAC170 putative targets showed significantly higher expression in DSX tissue compared to all known genes. We recommend removing the ChIP DNA amplification step prior to library preparation. Future work will aim to obtain deeply sequenced ChIP-seq data with biological replication, and integrate expression data for the identification of high-confidence target genes of EgrNAC170.

5.7. Acknowledgements

All authors acknowledge financial assistance from the Department of Science and Technology (DST), the National Research Foundation (NRF) Incentive fund (UID 81111) and NRF Bioinformatics and Functional Genomics Program (UID 86936), the Technology and Human Resources for Innovation Program (THRIP) (UID 80118), Sappi, Mondi, and the University of Pretoria. Mondi and Mmoledi Mphahlele (Mondi) are acknowledged for providing and collecting plant materials, respectively. Opinions expressed, and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the NRF.

© University of Pretoria

5.8. References

Adli M, Bernstein BE. 2011. Whole-genome chromatin profiling from limited numbers of cells using nano-ChIP-seq. *Nature Protocols* **6**(10): 1656-1668.

Chou PY, Fasman GD. 1978. Prediction of the secondary structure of proteins from their amino acid sequence. *Advances in Enzymology and Related Areas of Molecular Biology* **47**: 45–148.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**: 491-498.

Du J, Groover A. 2010. Transcriptional regulation of secondary growth and wood formation. *Journal of Integrative Plant Biology* **52**(1): 17–27.

Ernst HA, Olsen AN, Skriver K, Larsen S, Leggio LL. 2004. Structure of the conserved domain of ANAC, a member of the NAC family of transcription factors. *EMBO Reports* **5**(3): 297-303.

Garnier J, Gibrat J-F, Robson B. 1996. GOR method for predicting protein secondary structure from amino acid sequence. *Methods in Enzymology* **266**: 540-553.

Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan K-K, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, Min R, Alves P, Abyzov A, Addleman N, Bhardwaj N, Boyle AP, Cayting P, Charos A, Chen DZ, Cheng Y, Clarke D, Eastman C, Euskirchen G, Frietze S, Fu Y, Gertz J, Grubert F, Harmanci A, Jain P, Kasowski M, Lacroute P, Leng J, Lian J, Monahan H, O'Geen H, Ouyang Z, Partridge EC, Patacsil D, Pauli F, Raha D, Ramirez L, Reddy TE, Reed B, Shi M, Slifer T, Wang J, LinfengWu, Yang X, Yip KY, Zilberman-Schapira G, Batzoglou S, Sidow A, Farnham PJ, Myers RM, Weissman SM, Snyder M. 2012. Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**: 91-100.

Grant EH, Fujino T, Beers EP, Brunner AM. 2010. Characterization of NAC domain transcription factors implicated in control of vascular cell differentiation in *Arabidopsis* and *Populus*. *Planta* **232**: 337-352.

Grattapaglia D, Vaillancourt RE, Shepherd M, Thumma BR, Foley W, Külheim C, Potts BM, Myburg AA. 2012. Progress in Myrtaceae genetics and genomics: *Eucalyptus* as the pivotal genus. *Tree Genetics and Genomes* **8**: 463–508.

Graur D, Zheng Y, Price N, Azevedo RBR, Zufall RA, Elhaik E. 2013. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biology and Evolution* **5**(3): 578-590.

© University of Pretoria

Harfouche A, Meilan R, Altman A. 2011. Tree genetic engineering and applications to sustainable forestry and biomass production. *Trends in Biotechnology* **29**(1): 9-17.

Hruz T, Laule O, Szabo G, Wessendorp F, Bleuler S, Oertle L, Widmayer P, Gruissem W, Zimmermann P. 2008. Genevestigator V3: A reference expression database for the meta-analysis of transcriptomes. *Advances in Bioinformatics* **2008**: 420747.

Hussey SG, Mizrachi E, Creux NM, Myburg AA. 2013. Navigating the transcriptional roadmap regulating plant secondary cell wall deposition. *Frontiers in Plant Science* **4**: 325.

Hussey SG, Mizrachi E, Spokevicius AV, Bossinger G, Berger DK, Myburg AA. 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**: 173.

Iglesias-Trabado G, Wilstermann D 2008. *Eucalyptus universalis*. Global cultivated eucalypt forests map 2008 Version 1.0.1. In GIT Forestry Consulting's EUCALYPTOLOGICS: Information resources on *Eucalyptus* cultivation worldwide. Available online at http://www.git-forestry.com.

Jameson BA, Wolf H. 1988. The antigenic index: a novel algorithm for predicting antigenic determinants. *Bioinformatics* **4**(1): 181-186.

Kaufmann K, Wellmer F, Muiño JM, Ferrier T, Wuest SE, Kumar V, Serrano-Mislata A, Madueño F, Krajewski P, Meyerowitz EM, Angenent GC, Riechmann JL. 2010. Orchestration of floral initiation by APETALA1. *Science* **328**: 85-89.

Kharchenko PV, Tolstorukov MY, Park PJ. 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature Biotechnology* **26**(12): 1351-1359.

Kyte J, Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* **157**(1): 105-132.

Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, Chen Y, DeSalvo G, Epstein C, Fisher-Aylor KI, Euskirchen G, Gerstein M, Gertz J, Hartemink AJ, Hoffman MM, Iyer VR, Jung YL, Karmakar S, Kellis M, Kharchenko PV, Li Q, Liu T, Liu XS, Ma L, Milosavljevic A, Myers RM, Park PJ, Pazin MJ, Perry MD, Raha D, Reddy TE, Rozowsky J, Shoresh N, Sidow A, Slattery M, Stamatoyannopoulos JA, Tolstorukov MY, White KP, Xi S, Farnham PJ, Lieb JD, Wold BJ, Snyder M. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research* **22**: 1813-1831.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* **25**: 1754-1760.

© University of Pretoria

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinfomatics* **25**: 2078-2079.

Lucas WJ, Groover A, Lichtenberger R, Furuta K, Yadav S-R, Helariutta Y, He X-Q, Fukuda H, Kang J, Brady SM, Patrick JW, Sperry J, Yoshida A, López-Millán A-F, Grusak MA, Kachroo P. 2013. The plant vascular system: evolution, development and functions. *Journal of Integrative Plant Biology* **55**(4): 294-388.

Martin D, Brun C, Remy E, Mouren P, Thieffr D, Jacq B. 2004. GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biology* **5**: R101.

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of *cis*-regulatory regions. *Nature Biotechnology* **28**(5): 495-503.

Mizrachi E. 2013. *Functional genomics and systems genetics of cellulose biosynthesis in Eucalyptus.* PhD thesis, University of Pretoria, Pretoria.

Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D, Goodstein DM, Dubchak I, Poliakov A, Mizrachi E, Kullan ARK, van Jaarsveld I, Hussey SG, Pinard D, Merwe Kvd, Singh P, Silva-Junior OB, Togawa RC, Pappas MR, Faria DA, Sansaloni CP, Petroli CD, Yang X, Ranjan P, Tschaplinski TJ, Ye C-Y, Li T, Sterck L, Vanneste K, Murat F, Soler M, Clemente HS, Saidi N, Cassan-Wang H, Dunand C, Hefer CA, Bornberg-Bauer E, Kersting AR, Vining K, Amarasinghe V, Ranik M, Naithani S, Elser J, Boyd AE, Liston A, Spatafora JW, Dharmwardhana P, Raja R, Sullivan C, Romanel E, Alves-Ferreira M, Külheim C, Foley W, Carocha V, Paiva J, Kudrna D, Brommonschenkel SH, Pasquali G, Byrne M, Rigault P, Tibbits J, Spokevicius A, Jones RC, Steane DA, Vaillancourt RE, Potts BM, Joubert F, Barry K, Jr. GJP, Strauss SH, Jaiswal P, Grima-Pettenati J, Salse J, Peer YVd, Rokhsar DS, Schmutz J. in press. The genome of *Eucalyptus grandis* **-** a global tree for fiber and energy. *Nature*.

Pimrote K, Tian Y, Lu X. 2012. Transcriptional regulatory network controlling secondary cell wall biosynthesis and biomass production in vascular plants. *African Journal of Biotechnology* **11**(75): 13928-13937.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinfomatics* **26**(6): 841-842.

Ranik M, Myburg AA. 2006. Six new cellulose synthase genes from *Eucalyptus* are associated with primary and secondary cell wall biosynthesis. *Tree Physiology* **26**: 545–556.

© University of Pretoria

Roy A, Kucukural A, Zhang Y. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols* **5**: 725-738.

Schug J, Schuller W-P, Kappen C, Salbaum JM, Bucan M, Jr CJS. 2005. Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biology* **6**: R33.

Shamimuzzaman M, Vodkin L. 2013. Genome-wide identification of binding sites for NAC and YABBY transcription factors and co-regulated genes during soybean seedling development by ChIP-Seq and RNA-Seq. *BMC Genomics* **14**: 477.

Shannon CE. 1948. A mathematical theory of communication. *The Bell System Technical Journal* **27**(3): 379-423.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: Molecular Evolutionary Genetics Analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* **28**: 2731-2739.

Wang HH, Tang RJ, Liu H, Chen HY, Liu JY, Jiang XN, Zhang HX. 2013. Chimeric repressor of PtSND2 severely affects wood formation in transgenic *Populus*. *Tree Physiology* **33**(8): 878-886.

Zhang J, Elo A, Helariutta Y. 2010. *Arabidopsis* as a model for wood formation. *Current Opinion in Biotechnology* **22**: 1-7.

Zhang J, Nieminen K, Serra JAA, Helariutta Y. 2014. The formation of wood and its control. *Current Opinion in Plant Biology* **17**: 56-63.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, Liu XS. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**: R137.

Zhong R, Lee C, Ye Z-H. 2010. Evolutionary conservation of the transcriptional network regulating secondary cell wall biosynthesis. *Trends in Plant Science* **15**(11): 625-632.

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

Zhong R, McCarthy RL, Lee C, Ye Z-H. 2011. Dissection of the transcriptional program regulating secondary wall biosynthesis during wood formation in poplar. *Plant Physiology* **157**: 1452–1468.

Zhong R, Ye Z-H. 2007. Regulation of cell wall biosynthesis. *Current Opinion in Plant Biology* **10**: 564-572.

© University of Pretoria

5.9. Figures

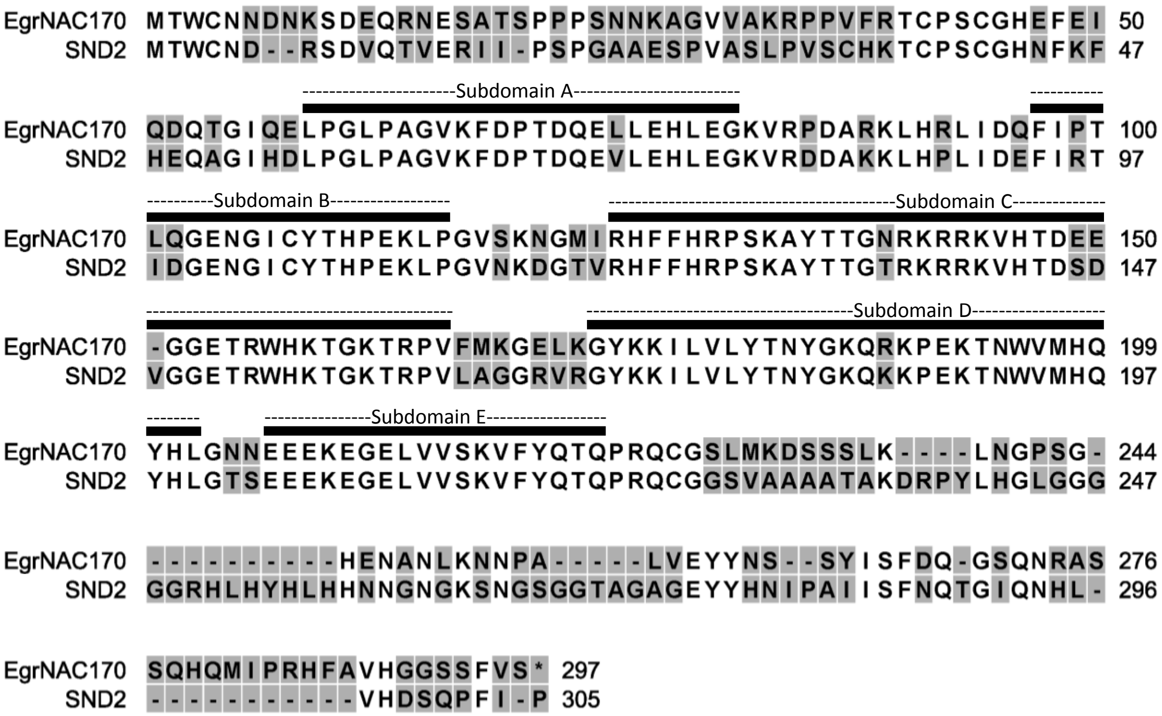

EgrNAC170 MTWCNNDNKSDEQRNESATSPPPSNNKAGVVAKRPPVFRTCPSCGHEFEI 50
SND2 MTWCND--RSDVQTVERII-PSPGAAESPVASLPVSCHKTCPSCGHNFKF 47
Subdomain A
EgrNAC170 QDQTGIQELPGLPAGVKFDPTDQELLEHLEGKVRPDARKLHRLIDQFIPT 100
SND2 HEQAGIHDLPGLPAGVKFDPTDQEVLEHLEGKVRDDAKKLHPLIDEFIRT 97
Subdomain B
Subdomain C
EgrNAC170 LQGENGICYTHPEKLPGVSKNGMIRHFFHRPSKAYTTGNRKRRKVHTDEE 150
SND2 IDGENGICYTHPEKLPGVNKDGTVRHFFHRPSKAYTTGTRKRRKVHTDSD 147
Subdomain D
EgrNAC170 -GGETRWHKTGKTRPVFMKGELKGYKKILVLYTNYGKQRKPEKTNWVMHQ 199
SND2 VGGETRWHKTGKTRPVLAGGRVRGYKKILVLYTNYGKQKKPEKTNWVMHQ 197
Subdomain E
EgrNAC170 YHLGNNEEEKEGELVVSKVFYQTQPRQCGSLMKDSSSLK----LNGPSG- 244
SND2 YHLGTSEEEKEGELVVSKVFYQTQPRQCGGSVAAAATAKDRPYLHGLGGG 247
EgrNAC170 ----------HENANLKNNPA-----LVEYYNS--SYISFDQ-GSQNRAS 276
SND2 GGRHLHYHLHHNNGNGKSNGSGGTAGAGEYYHNIPAIISFNQTGIQNHL- 296
EgrNAC170 SQHQMIPRHFAVHGGSSFVS* 297
SND2 -----------VHDSQPFI-P 305

Fig. 5.1. Amino acid sequence alignment of EgrNAC170 (*E. grandis*) and SND2 (*A. thaliana*) putative orthologs. Variable residues are indicated in grey. Subdomains A through E of the conserved NAC domain are indicated by black bars.

© University of Pretoria

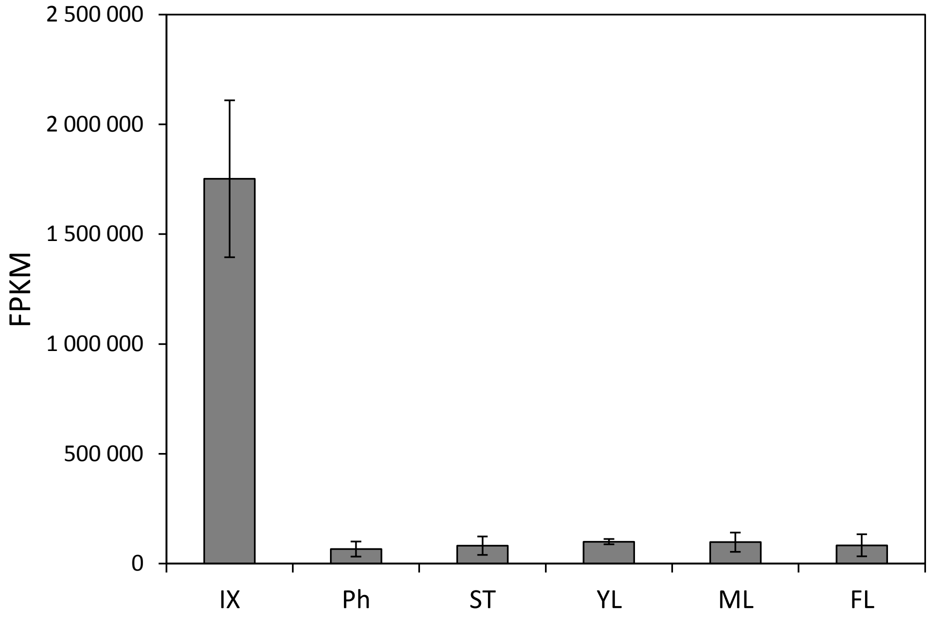


Fig. 5.2. Expression profile of *EgrNAC170* in various tissues and organs of mature *E. grandis* trees. Data were obtained from EucGenIE (http://eucgenie.bi.up.ac.za/; Hefer *et al.*, in preparation). FPKM; RNA-seq fragments per kilobase of transcript per million mapped RNA-seq fragments; DSX, developing secondary xylem; Ph, phloem; ST, shoot tips; YL, young leaves; ML, mature leaves; Fl, flowers. Error bars indicate standard deviation of three individuals.

© University of Pretoria

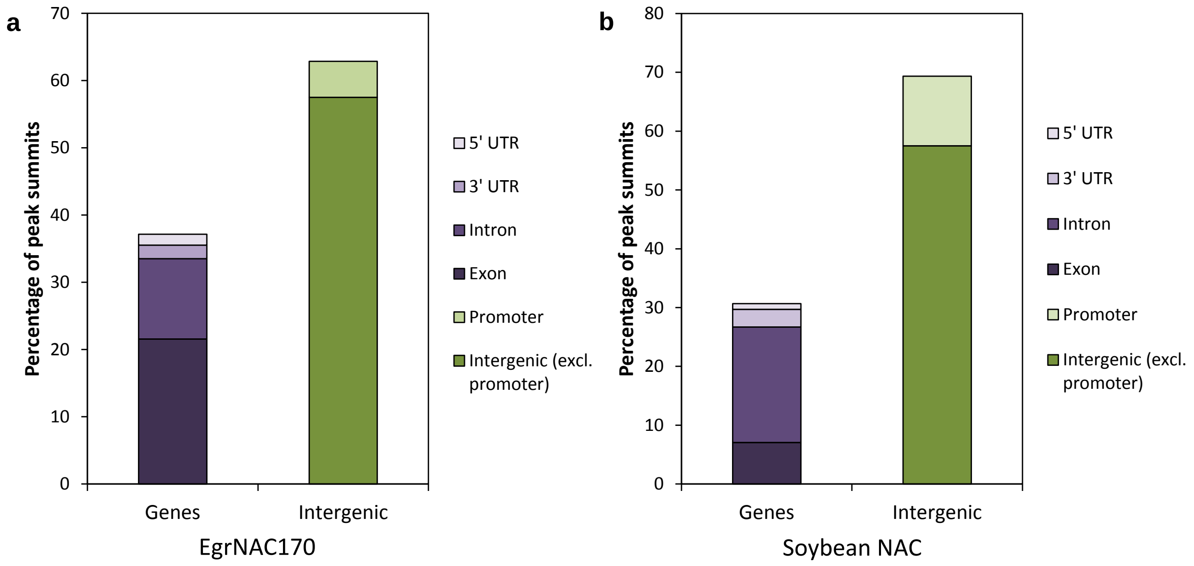


Fig. 5.3. Distribution of EgrNAC170 ChIP-seq peak summits with annotated genomic features (a), in comparison with a soybean NAC protein (b) (Shamimuzzaman & Vodkin, 2013). UTR, untranslated region. The promoter region is refined as 1000 bp upstream of the transcription start site. The statistics in (b) were calculated from data presented in the supplementary material of Shamimuzzaman & Vodkin (2013).

© University of Pretoria

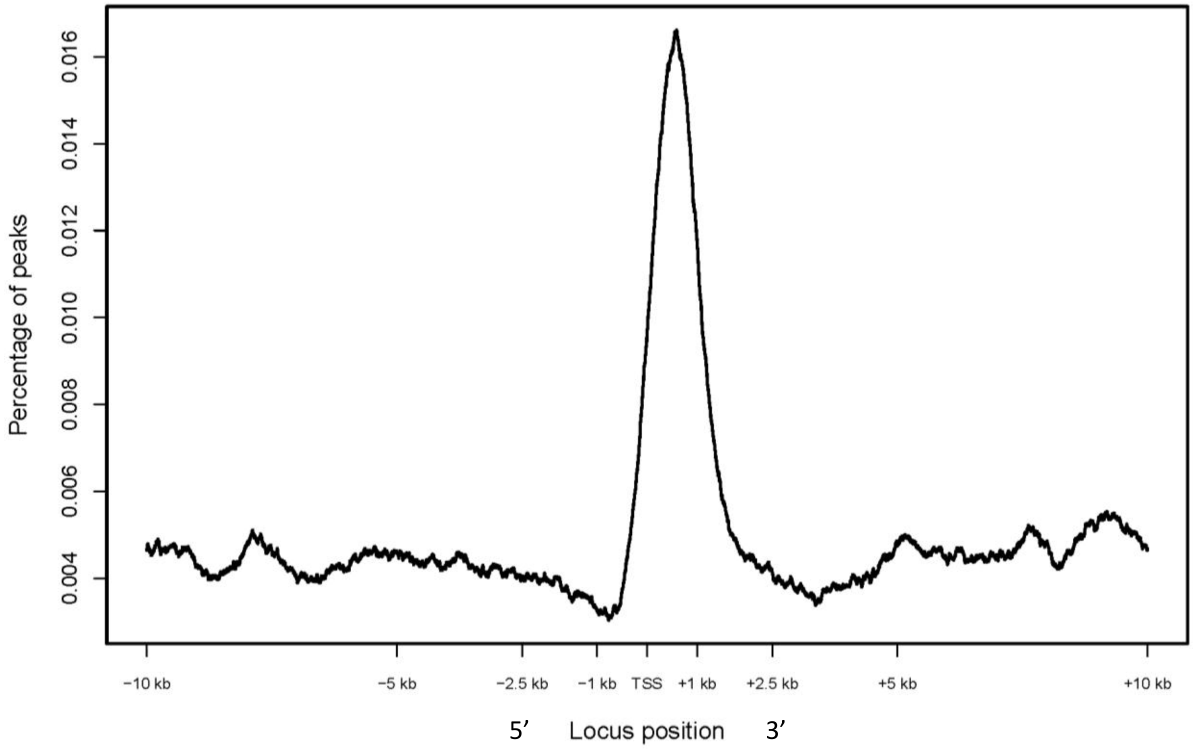
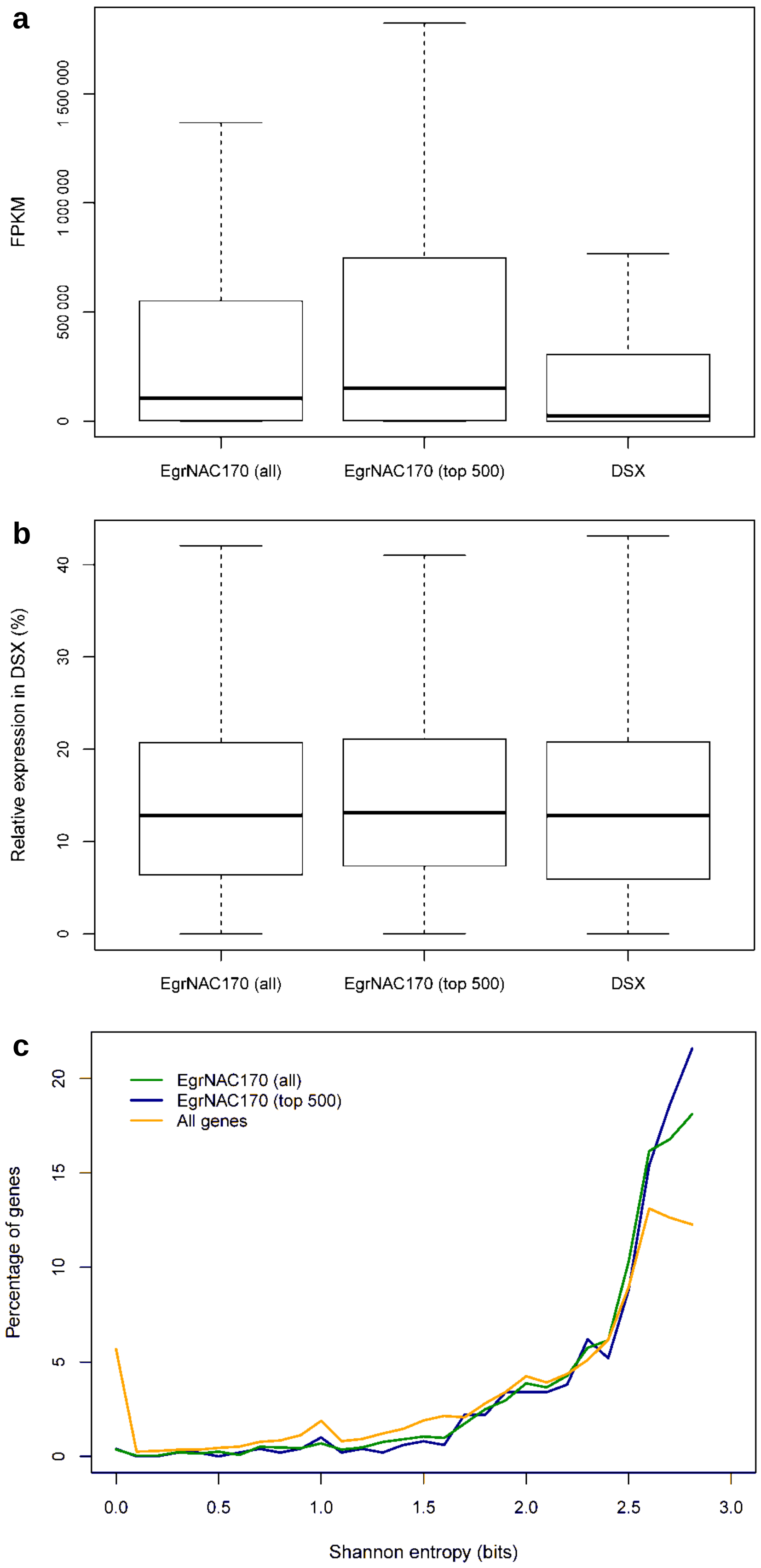


Fig. 5.4. Binding profile of EgrNAC170 with respect to the transcription start site (TSS) of all genes in the v.1.1. annotation of the *E. grandis* genome. The percentage of all peaks overlapping a given single-base position (binned across all genes) is shown on the *y*-axis.

© University of Pretoria

Fig. 5.5. Expression characteristics of EgrNAC170 putative targets in developing secondary xylem (DSX). (a) Absolute transcript levels of EgrNAC170 putative targets (n = 3,234) and those of the top 500 EgrNAC170 putative targets ranked according to the significance of their assigned binding peaks, compared to those of all genes ("DSX", n = 33,918). FPKM, fragments per kilobase of exon per million fragments mapped. **(b)** Relative expression in DSX tissue of the datasets represented in (a). **(c)** Shannon entropy values of tissue specificity for all expressed EgrNAC170 putative targets (n = 2,765) and those of the top 500 EgrNAC170 putative targets ranked according to the significance of their assigned binding peaks, compared to those of all genes (n = 31,403). RNA-seq data was obtained from EucGenIE (http://eucgenie.bi.up.ac.za/; Hefer *et al.*, in preparation).

© University of Pretoria



© University of Pretoria

5.10. Supplementary figures

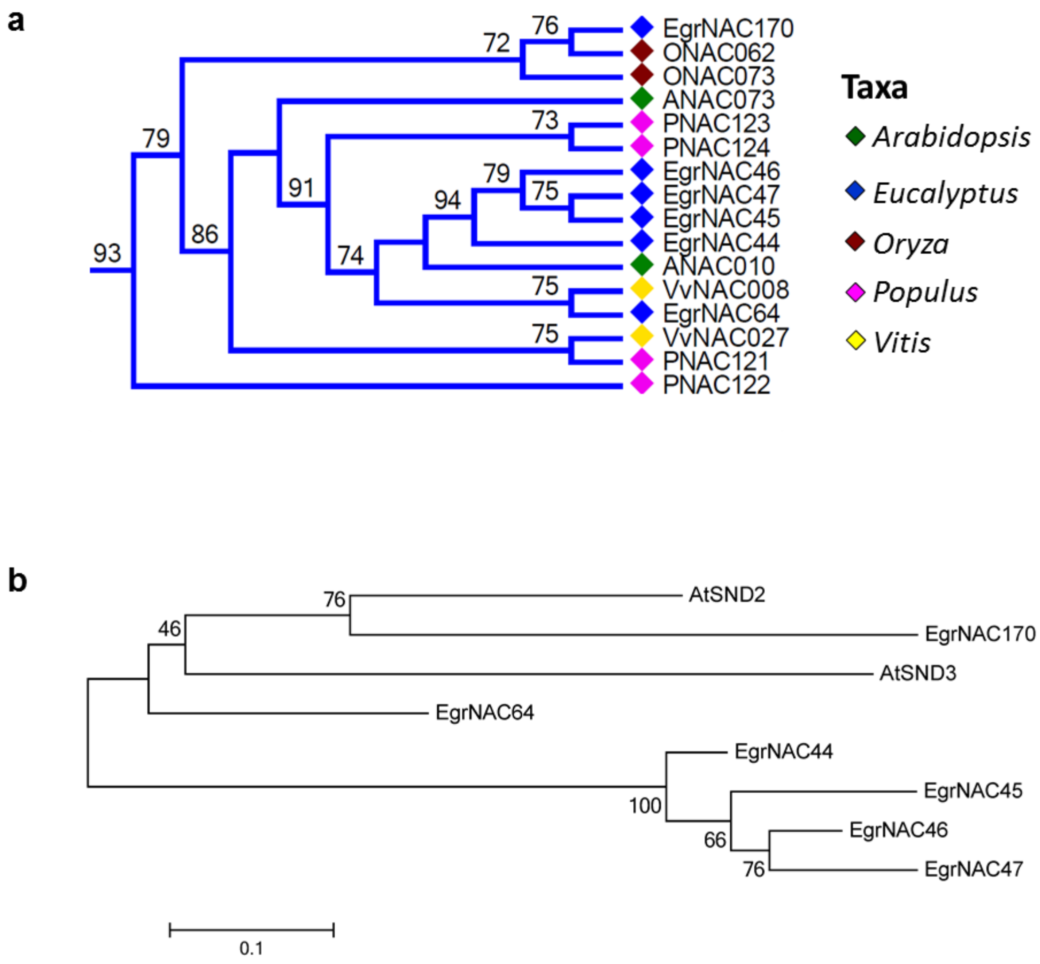


Fig. S5.1. Phylogenetic analysis of EgrNAC170 and related homologs in *E. grandis* and *A. thaliana*. (a) The SND2 (ANAC073) and SND3 (ANAC010) clade from the dendrogram constructed in Chapter 2 (Additional file 2.2), reproduced. **(b)** Bootstrapped, midpoint-rooted maximum likelihood phylogeny of SND2 and SND3 in relation to their closest *E. grandis* homologs. *Arabidopsis* proteins have been assigned the prefix “At” to distinguish them from *Eucalyptus* (prefix “Egr”) homologs.

© University of Pretoria

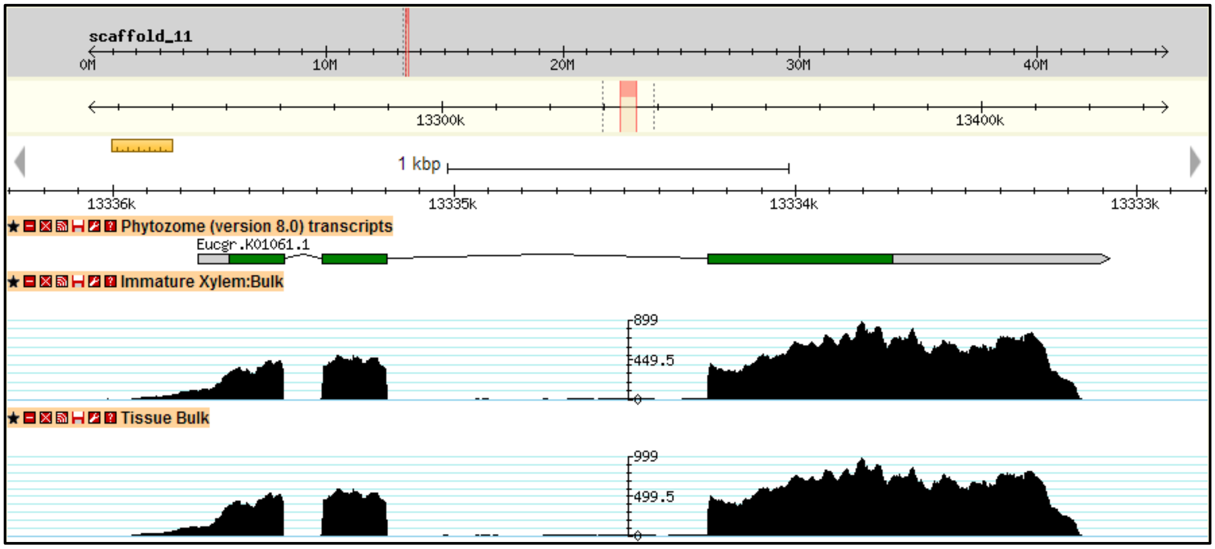


Fig. S5.2. Genome browser view of gene model Eucgr.K01061.1 in EucGenIE (http://eucgenie.bi.up.ac.za/; Hefer *et al.*, in preparation). RNA-seq coverage for Eucgr.K01061 (EgrNAC170) in immature xylem and bulked tissues is shown. The genome window corresponds to the locus "scaffold_11:13,332,910..13,336,069".

© University of Pretoria

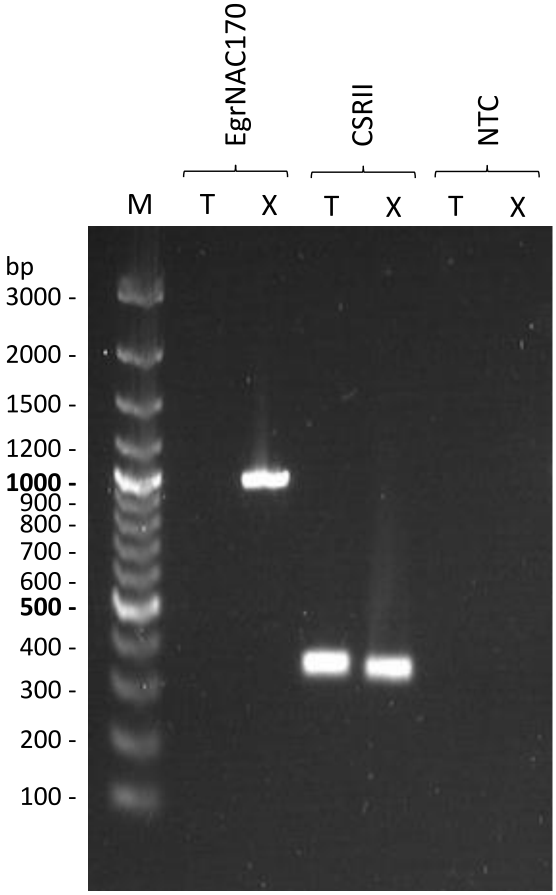


Fig. S5.3. Agarose gel electrophoresis analysis showing *EgrNAC170* coding sequence amplification using young twig (T) or mature xylem (X) cDNA as template. CSRII, amplification of the *EgCesA1* class-specific region II (Ranik & Myburg, 2006) as a positive control for cDNA quality; NTC, template-free control; M, GeneRuler 100 bp DNA ladder plus.

© University of Pretoria

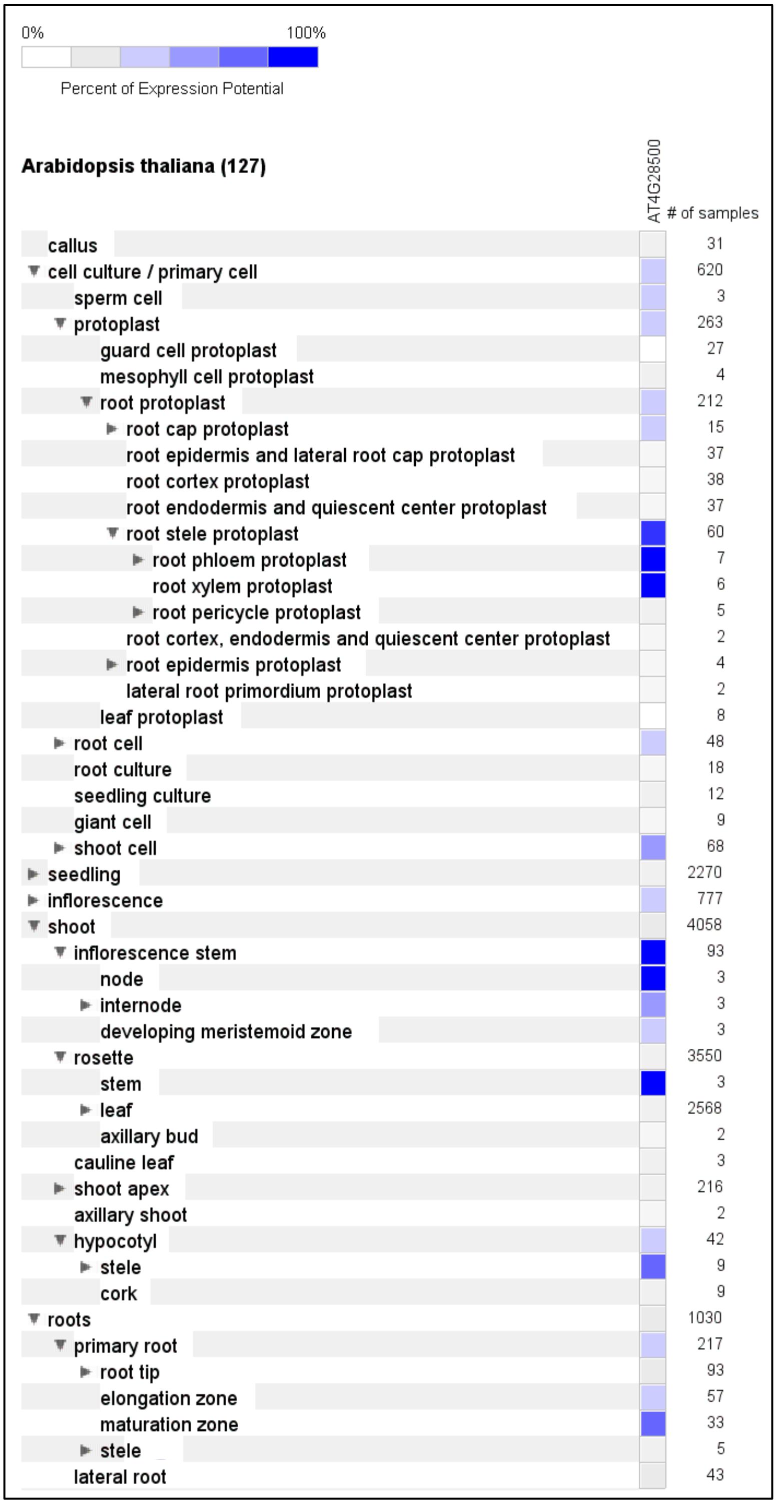
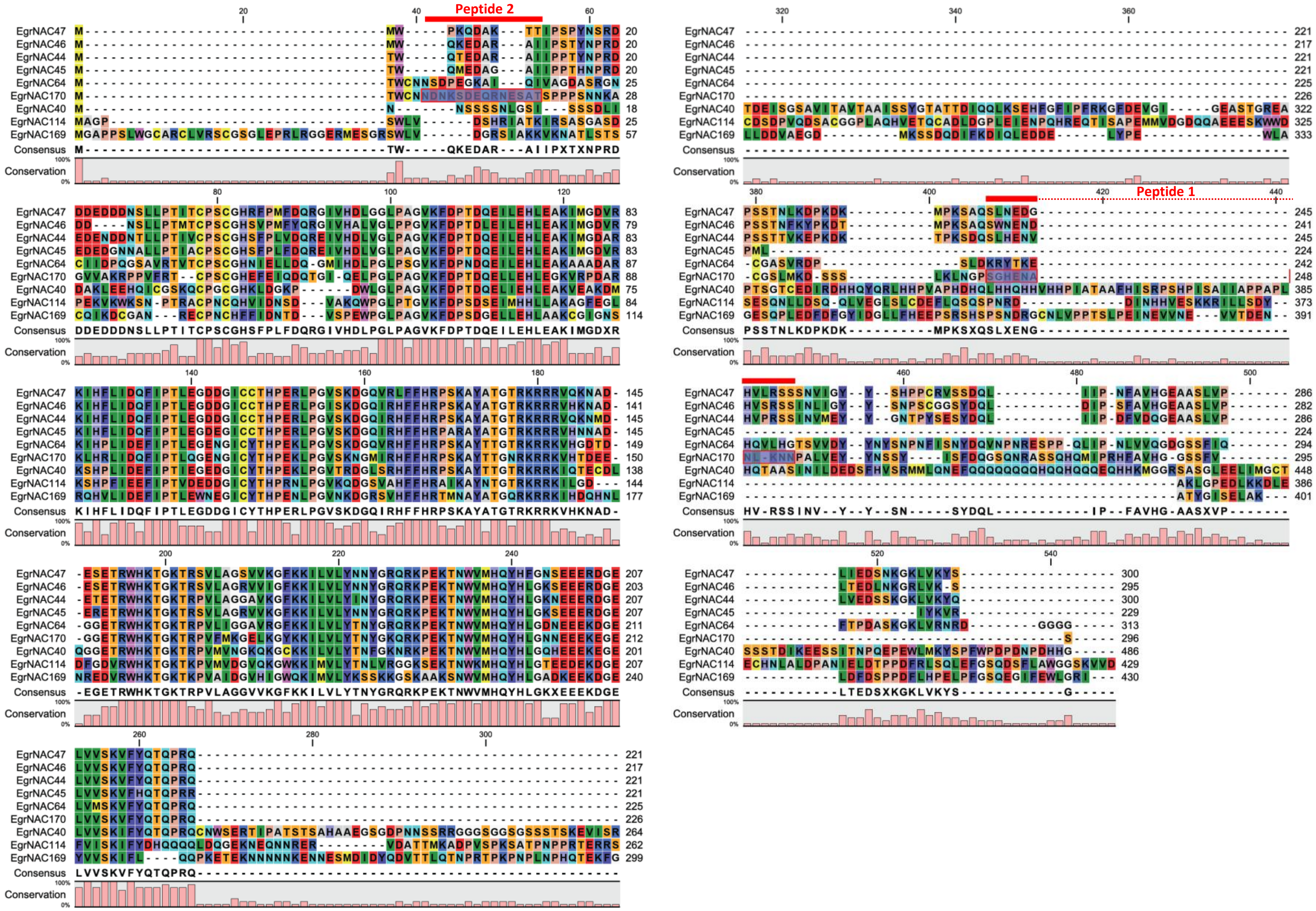


Fig. S5.4. Anatomical abundance of *SND2* (AT4G28500) transcripts visualized in Genevestigator (Hruz *et al.*, 2008).

© University of Pretoria

Fig. S5.5. Amino acid sequence alignment of EgrNAC170 and its closest *E. grandis* homologs. The two EgrNAC170 peptides used for custom antibody synthesis are indicated by red bars. Data were visualized in CLC Bio Main workbench 6 (http://www.clcbio.com).

© University of Pretoria



© University of Pretoria

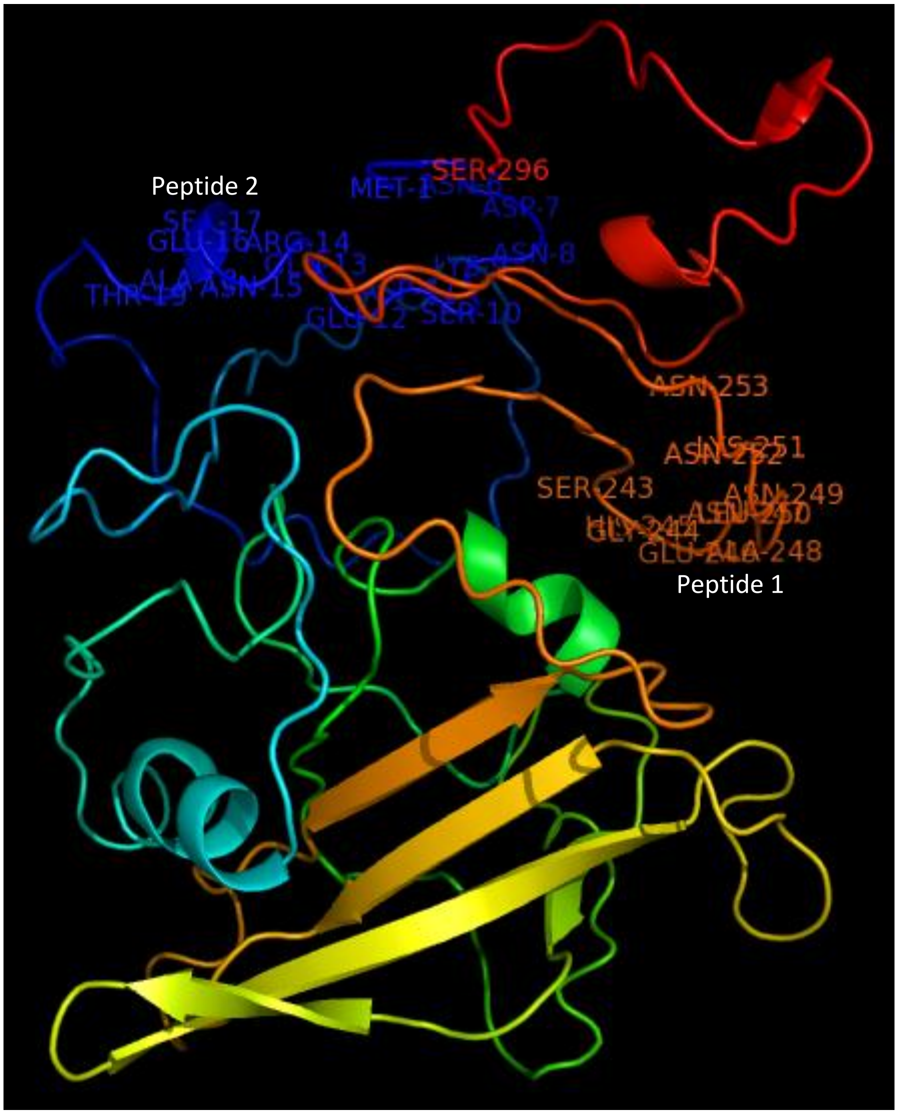


Fig. S5.6. Top-scoring I-TASSER model of EgrNAC170 showing the relative positions of two synthetic peptides designed for polyclonal antibody generation. The DNA-binding domain is visible in the lower third of the image as a twisted β-sheet flanked by two α-helices (Ernst *et al.*, 2004). N-terminal methionine and C-terminal serine residues are indicated. The C-score (C) for this model, a measure of model confidence $\{C \mid -5 \leq C \leq 2\}$, is -3.3.

© University of Pretoria

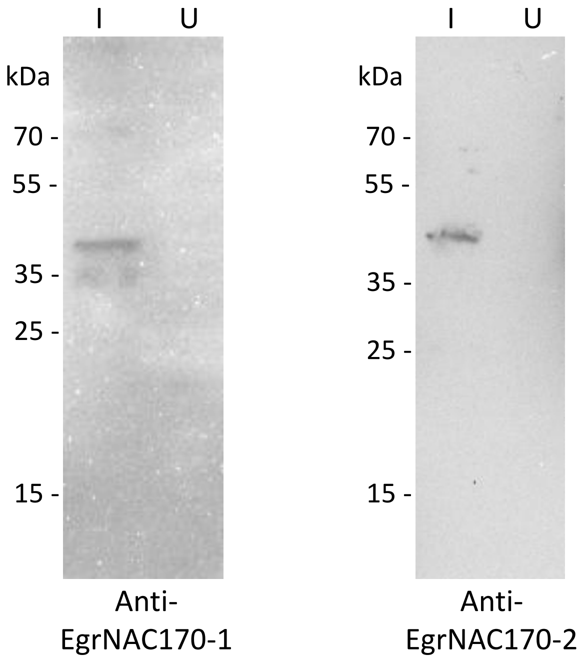


Fig. S5.7. Western blot analysis of recombinant EgrNAC170. Whole lysate of *E. coli* BL21Star transformed with pET160-EgrNAC170 were induced (I) or uninduced (U) with IPTG. Blots were independently probed anti-EgrNAC170-1 and anti-EgrNAC170-2 antibodies as indicated. The predicted molecular weight of the recombinant protein is ~38 kDa.

© University of Pretoria

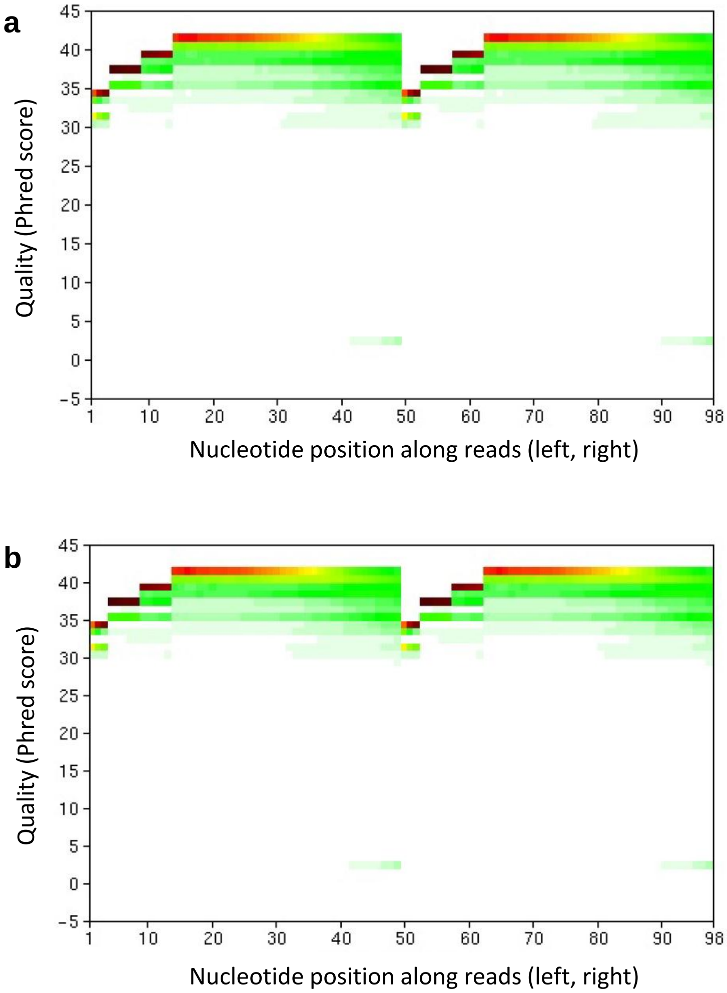


Fig. S5.8. Distribution of average per-base ChIP-seq library read quality (Phred score) along the sequence length. (a) Anti-EgrNAC170-1 ChIP-seq library. **(b)** Anti-EgrNAC170-2 ChIP-seq library. Graphics provided by Beijing Genome Institute.

© University of Pretoria

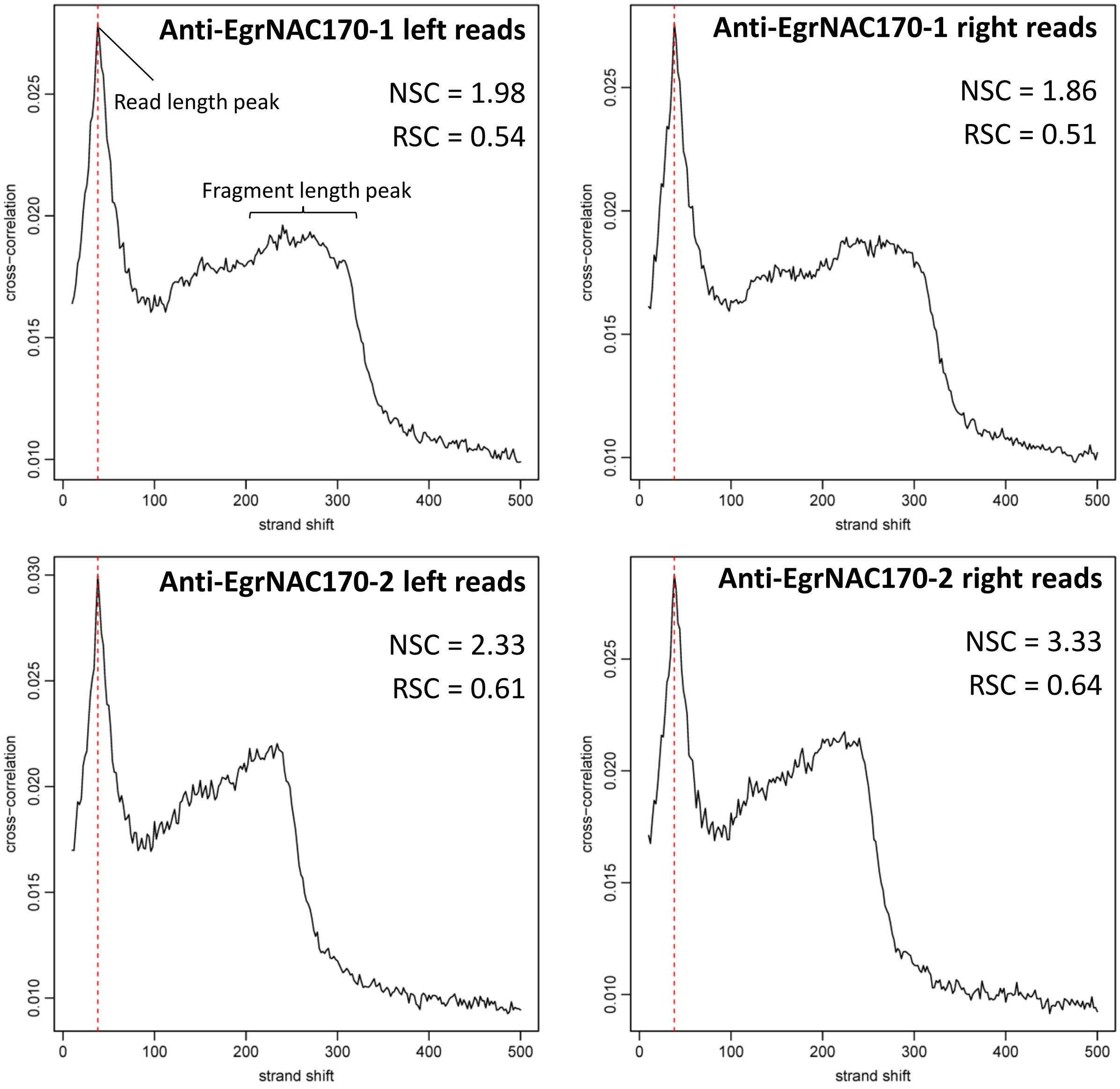


Fig. S5.9. Strand cross-correlation analysis of anti-EgrNAC170-1 and anti-EgrNAC170-2 ChIP-seq libraries. Normalized strand cross-correlation (NSC) and relative strand cross-correlation (RSC) values, defined in Chapter 4, are indicated. The read length peak and fragment length peaks are indicated in the top left panel.

© University of Pretoria

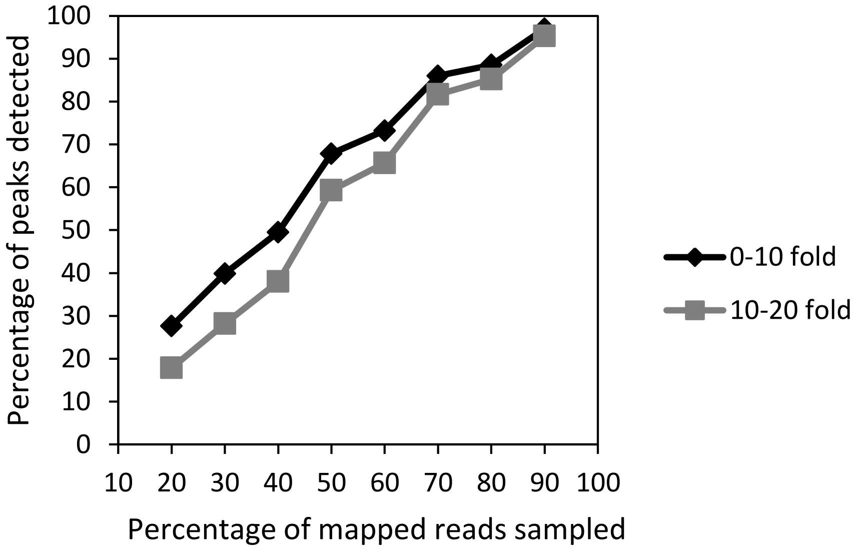


Fig. S5.10. Diagnostic report of peak detection saturation. Saturation curves are shown separately for low-enrichment peaks (0 - 10 fold) and high enrichment peaks (10 - 20 fold).

© University of Pretoria

5.11. Supplementary tables

Table S5.1. Yield of DNA obtained after ChIP DNA amplification.

	Yield after 15 PCR cycles
Anti-EgrNAC170-1 ChIP (replicate 1)	45.6 ng
Anti-EgrNAC170-1 ChIP (replicate 2)	4.5 ng
Anti-EgrNAC170-2 ChIP (replicate 1)	70.5 ng
Anti-EgrNAC170-2 ChIP (replicate 2)	75.3 ng
Template-free control	< 0.1 ng

© University of Pretoria

Table S5.2. Mapping efficiencies of ChIP-seq libraries

Dataset	Total clean reads produced	Read length after trimming	Estimated sequence duplication rate[a]	Uniquely mapped reads	% uniquely mapped
Anti-EgrNAC170-1	14 797 440	39 bp	81.6%	1 868 579	12.6
Anti-EgrNAC170-2	13 069 416	39 bp	85.4%	1 013 211	7.8
Input	20 846 731	31 bp	67.6%	11 464 953	55.0

[a] Estimate according to FastQC (http://www.bioinformatics.babraham.ac.uk/)

© University of Pretoria

Table S5.3. EgrNAC170 direct candidate target genes associated with lignin biosynthesis.

Annotated enzyme[a]	Gene model	FPKM[b]	Relative expression[c]
PAL	Eucgr.G02850	199 783	0.48
4CL	Eucgr.C02284	12 572 167	0.57
HCT	Eucgr.F03978	2 639 093	0.57
COMT	Eucgr.A01873	0	0.00
	Eucgr.F03794	160 802	0.04
	Eucgr.G00020	281	0.00
	Eucgr.H00351	0	0.00
	Eucgr.H03926	13 143	0.00
	Eucgr.K00957	38 356	0.10
CCR	Eucgr.C01240	598 960	0.17
	Eucgr.F03605	197 975	0.05
	Eucgr.G00052	303 772	0.15
	Eucgr.G02325	5 870	0.00
	Eucgr.I01552	4 890	0.01
CAD	Eucgr.D00471	0	0.00
	Eucgr.F01676	24 324	0.02
	Eucgr.F01677	436	0.03
	Eucgr.F01678	3 910	0.01
	Eucgr.F01679	4 997	0.01
	Eucgr.F01680	58 938	0.02
	Eucgr.H02433	0	0.00

[a] Carocha *et al.*, in preparation.

[b] Absolute expression (fragments per kilobase of exon per million fragments mapped) in developing secondary xylem tissue (http://eucgenie.bi.up.ac.za/; Hefer *et al.* in preparation). The median FPKM for this tissue is 90,000, the 90th percentile is 1,350,000 and the 99th percentile is 7,780,000 (Mizrachi, 2013).

[c] Relative expression in developing secondary xylem relative to six other tissues.

© University of Pretoria

Table S5.4. EgrNAC170 direct candidate target genes associated with secondary cell wall biosynthesis and transcriptional regulation. Genes associated with phenylpropanoid biosynthesis are shown separately in Table S5.3.

Gene model	FPKM[a]	Relative expression[b]	Tension wood[c]	At homolog	Gene name	Description
Eucgr.C01769	1 205 310	0.21		AT4G32410	CESA1	Cellulose synthase 1
Eucgr.D00476	25 758 933	0.92	√	AT4G18780	CESA8	Cellulose synthase A8
Eucgr.B02355	4 335 697	0.17	√	AT2G39700	EXPA4	Expansin A4
Eucgr.K02662	344 615	0.27		AT3G55820		Fasciclin-like arabinogalactan family protein
Eucgr.B03801	2 889 163	0.28		AT5G06390	FLA17	Fasciclin-like arabinogalactan protein 17 precursor
Eucgr.B00458	2 936 527	0.36		AT1G74690	IQD31	IQ-domain 31
Eucgr.F01629	2 685 367	0.25		AT1G19870	IQD32	IQ-domain 32
Eucgr.K03111	2 548 223	0.94		AT5G60020	LAC17	Laccase 17
Eucgr.K02996	20 269 233	0.89	√	AT2G38080	LAC4	Laccase 4
Eucgr.G03028	488 886	0.50	√	AT2G38080	LAC4	Laccase 4
Eucgr.I00012	110 326	0.08	√	AT4G38620	MYB4	Myb domain protein 4
Eucgr.I00213	1 212 588	0.12		AT5G09330	VNI1	NAC domain containing protein 82
Eucgr.F00232	2 360 753	0.75		AT4G33330	GUX2,PGSIP3	Plant glycogenin-like starch initiation protein 3
Eucgr.A02598	1 473 073	0.47	√	AT3G52480		Protein of unknown function
Eucgr.E01152	2 551 657	0.69	√	AT1G31720		Protein of unknown function (DUF1218)
Eucgr.H02217	19 256 333	0.87	√	AT5G67210		Protein of unknown function (DUF579)
Eucgr.E01053	1 036 140	0.59		AT1G32770	SND1	SECONDARY WALL-ASSOCIATED NAC DOMAIN1
Eucgr.K02955	173 202	0.46		AT2G38320	TBL34	TRICHOME BIREFRINGENCE-LIKE 34
Eucgr.B00451	192 504	0.51		AT5G59290	UXS3	UDP-glucuronic acid decarboxylase 3
Eucgr.G01174	148 987	0.43	√	AT2G14620	XTH10	Xyloglucan endotransglucosylase/hydrolase 10
Eucgr.A01968	30 120 967	0.45	√	AT3G23730	XTH16	Xyloglucan endotransglucosylase/hydrolase 16

[a] Absolute expression (fragments per kilobase of exon per million fragments mapped) in developing secondary xylem (http://eucgenie.bi.up.ac.za/; Hefer *et al.* in preparation). The median FPKM for this tissue is 90,000, the 90th percentile is 1,350,000 and the 99th percentile is 7,780,000 (Mizrachi, 2013).

[b] Relative expression in developing secondary xylem relative to six other tissues.

[c] Differential expression in tension wood (Mizrachi *et al.*, in preparation.

© University of Pretoria

5.12. Additional files

Additional file 5.1.txt Nucleotide sequence (Genbank format) of the EgrNAC170 coding region cloned in this study.

Additional file 5.2.xls *E. grandis* associated with EgrNAC170 ChIP-seq peaks, and their closest *Arabidopsis* homologs (Excel format).

Additional file 5.3.xls Biological function gene ontology (GO) terms significantly overrepresented among EgrNAC170-associated genes (Excel format).

© University of Pretoria

CHAPTER 6

CONCLUDING REMARKS

© University of Pretoria

A rapidly growing world population, accompanied by industrialization of developing countries on a vast scale, has seen an increase in demand for paper and products derived from forest biomass. Simultaneously, in light of anthropogenic climate change (PICC, 2013), trees have become an important carbon sink for the sequestration of atmospheric carbon. Today, tens of millions of hectares of *Eucalyptus* plantations are grown in tropical, subtropical and mild temperate regions of the globe to meet demand for pulp and paper (Iglesias-Trabado & Wilstermann, 2008). Eucalypts have become favourable pulping candidates due to their rapid growth and superior fiber qualities. However, the land area suitable for forestry plantations is declining, placing pressure on breeders to increase biomass productivity. At the same time, novel biomaterials being developed from lignocellulosic sources, such as cellulose nanocrystals and bioplastics (Granja *et al.*, 2006; Habibi *et al.*, 2010; Lagaron & Lopez-Rubio, 2011; Mishra & Mishra, 2011), and the potential use of lignocellulosic feedstocks for biofuel production (Weng *et al.*, 2008), has created opportunities for manipulating the physico-chemical properties of wood and secondary cell walls (SCWs).

Unlike biosynthetic genes involved in the biosynthesis of SCWs, the discovery of transcription factors and transcriptional networks underlying the regulation of the structural genes remained elusive until the last decade, with breakthrough studies by Kubo *et al.* (2005), Mitsuda *et al.* (2005) and Zhong *et al.* (2006) describing for the first time the NAC domain "master regulators" of SCW deposition. Our understanding of SCW transcriptional regulation has grown tremendously since then, with well over 400 regulatory relationships identified in the *Arabidopsis* literature (reviewed in Chapter 1) and functional studies performed in the tree model *Populus* (e.g. Zhong *et al.*, 2010; Li *et al.*, 2012). While SCW structural genes have

© University of Pretoria

been used, among others, to modify lignin content (reviewed by Chiang, 2006; Vanholme *et al.*, 2008) and increase growth rate and biomass (Coleman *et al.*, 2009; da Silveira, 2013) in trees, transcription factors are potentially powerful candidates for transgenic approaches to wood trait enhancement due to their ability to regulate the expression of many structural genes at a time. This was elegantly demonstrated in *A. thaliana* in a *c4h* mutant with decreased lignin content: in addition to restoring lignification in vessel walls to prevent their collapse, a positive feedback regulatory loop was engineered to enhance polysaccharide deposition in fibers by driving transcription factor *NST1* expression with the promoter of one of its targets, *IRX8* (Yang *et al.*, 2013).

The motivation for this PhD study began with an *Arabidopsis* report by Zhong *et al.* (2008) demonstrating the NAC domain protein SND2 to activate the *cellulose synthase8* (*CesA8*) gene promoter but not representatives of xylan or lignin biosynthesis, that overexpression and dominant repression of SND2 increased and decreased fiber (but not vessel) SCW deposition respectively, and that SND2 dominant repression led to reductions in glucose and xylose sugars derived from cell wall material. We interpreted these results as suggesting that SND2 may be a fiber-specific regulator of the cellulose biosynthetic machinery specifically, in which case it would be a valuable biotechnological candidate for increasing cellulose biosynthesis in plantation trees.

While further exploring whether a cellulose-specific regulatory role was performed by SND2 in *Arabidopsis*, in the design of this thesis I wished to link these findings to the organism of interest, *E. grandis*, for which a first draft DOE-JGI genome sequence and

© University of Pretoria

annotation had emerged by 2011 (http://www.phytozome.net/Eucalyptus.php). Considering that *Eucalyptus* is a fast-growing superior fiber crop, and that extensive tandem duplications as well as an independent whole genome duplication have occurred in this lineage (Myburg *et al.*, in press), I had some expectation that the number and functions of NAC domain proteins associated with SCW regulation in the *Eucalyptus* lineage may have diverged from that of other flowering plants. I first asked what was the evolutionary history, structural characteristics, and transcriptional dynamics of the NAC domain family of *E. grandis* in relation to other angiosperms (Chapter 2). In Chapter 3 I asked what genes are regulated by the NAC domain protein SND2 in *Arabidopsis*, and what are the phenotypic effects of its overexpression in *Arabidopsis* and *Eucalyptus* (Chapter 3). Having cemented a role for SND2 in regulating several aspects of SCW deposition in *Arabidopsis*, I asked what the direct gene targets are of EgrNAC170, the most likely *E. grandis* ortholog of SND2, in developing secondary xylem tissue (Chapter 5). To address this question, I aimed to optimize and apply the chromatin immunoprecipitation sequencing (ChIP-seq) approach to developing xylem tissue (Chapter 4). By profiling the well-studied histone modification lysine 4-trimethylated H3 (H3K4me3) as a validation of the approach, I took the opportunity to gain insight as to how xylogenesis is epigenetically regulated during secondary xylem development in *Eucalyptus*.

That NAC proteins have expanded significantly, mainly through tandem duplication, in subfamilies associated with biotic and abiotic stress response in *E. grandis*, while those associated with SCW regulation generally have one-to-one orthologous relationships with *Arabidopsis*, is an important finding of this thesis. These results suggest that the superior

© University of Pretoria

growth and wood properties of *Eucalyptus* are not likely a result of NAC protein diversification, thus justifying the extrapolation of NAC protein functions in model plants as directly relevant to *Eucalyptus*. That said, there are notable exceptions. First, the unique amino acid motifs identified within transcriptional activation and repression regions of EgrNAC proteins may reflect neofunctionalization mechanisms. Second, the novel NAC genes for the regulation of xylogenesis that were identified from transcript profiles across tissues and in response to trunk tension stress in *E. globulus* motivate future studies of their possible functions. Third, one major outstanding question from Chapter 2 pertains to the functions of NAC subfamily VII genes that have independently duplicated in the three tree models studied (*Eucalyptus*, *Populus* and *Carica*) but not *Vitis* or herbaceous angiosperms.

Apart from the wood formation focus of Chapter 2, the identification of ten NAC domain transcripts responding to cold stress in *E. globulus* contributes significantly to improved transgenic strategies for improving cold stress resistance in *Eucalyptus*. It is interesting that three of these transcripts are homologs of SCW-associated *SND1* (*EglNAC61*), *SND3* (*EglNAC64*) and *VND5/VND5* (*EglNAC50*), differentially expressed after cold treatment in young leaves, primary stem and root respectively. It is known that lignin content and composition, as well as the accumulation of phenylpropanoid pathway derivatives and increased phenylalanine ammonia-lyase (PAL) activity, is associated with cold stress (reviewed by Moura *et al.*, 2010). Possibly, these SCW-associated candidates prepare the tree for frost exposure through their direct or indirect regulation of the phenylpropanoid pathway. A more detailed analysis of these TFs will be an interesting line of future investigation.

© University of Pretoria

In Chapter 3, the hypothesis of a cellulose-specific regulatory function for SND2 in *Arabidopsis* SCW transcription regulation was rejected on the grounds that its overexpression was associated with the induction of cellulose, hemicellulose and lignin polymerization genes. *SND2* overexpression led to interfascicular fiber SCW thickness reduction in several independent lines, confirming a role in SCW regulation in these cells. While transcriptomic and phenotypic data support a role for SND2 in regulating fiber SCW deposition in *Arabidopsis*, including the regulation of a signaling complex with multiple lines of evidence supporting its role in SCW deposition and/or patterning (Chapter 3), the developmental consequences of *SND2* overexpression were minor, with small alterations in fiber SCW thickness and cell wall chemistry (mannose, rhamnose and total lignin). However, a significant increase in fiber cell cross-sectional area in hybrid *Eucalyptus* transgenic sectors overexpressing *SND2* supports its potential as a biotechnology candidate. Although considerable biological variation in this experiment obscured the basis for this phenotype, in Chapter 3 we attributed the fiber area increase as likely due, in part, to enhanced SCW deposition seeing that fiber cell wall area was near-significantly increased and SCW thickness was also qualitatively increased. There is also the possibility that SND2 influences cell wall expansion, perhaps through the cell wall loosening and cell growth function of its putative target *EXPANSIN A15* (Goh *et al.*, 2012).

One interesting hypothesis emanating from this thesis regarding *SND2*, introduced in Chapters 1 and 3 but developed further here, is that of an optimal range of expression that results in maximal fiber SCW deposition, presumably through optimal activation of SCW structural genes by SND2. In *Arabidopsis*, I could not reproduce the increased fiber SCW

© University of Pretoria

thickness phenotype in *SND2*-overexpressing plants reported by Zhong *et al.* (2008) and instead showed that overexpression of *SND2* had the opposite effect, at least in lines with strong *SND2* overexpression. While the gain-of-function phenotype reported by Zhong *et al.* (2008) remains to be independently verified, I attributed this contradiction to the indirect effect of high levels of overexpression observed in my study, a phenomenon also apparent for the fiber SCW master regulator SND1 as reviewed in Chapter 1 (section 1.5). This idea is further supported to some degree by two studies of the putative *Populus* ortholog of *SND2*, *PtrNAC154*. Grant *et al.* (2010) showed preferential expression of *PtrNAC154* in xylem cells actively depositing SCWs in *Populus*. *PtrNAC154* overexpression led to reduced height due to shortening of internodes, and reduced xylem production met with a relative increase in outer bark and phloem-cambium thickness (Grant *et al.*, 2010). Wang *et al.* (2013) showed that PtrNAC170 is a transcriptional activator, and that dominant repression of the protein resulted in reduced height and stem thickness, compromised secondary xylem and phloem production and thinner fiber SCW thickness. Cellulose and lignin were reduced, as were the transcript levels of cellulose-, xylan- and lignin-associated transcripts (Wang *et al.*, 2013). Thus, both overexpression of the native protein and overexpression of the chimeric dominant repressor reduce xylem production and internode elongation in *Populus*. Based on these data, I propose a model of dosage-dependent regulation of fiber SCW thickness and xylem production for *SND2* in *Arabidopsis* and *PtrNAC154* in *Populus* (Fig. 6.1). The mechanism of this dosage regulation is unknown, but may be explained by the gene balance hypothesis (Birchler & Veitia, 2007; Birchler & Veitia, 2010), general transcriptional squelching (Cahill *et al.*, 1994; Orphanides *et al.*, 2006) or negative feedback regulation. Considering that SND2 likely requires an essential co-factor to activate SCW genes, as discussed in Chapter 3, a greater

© University of Pretoria

knowledge of other proteins participating in the formation of protein complexes with SND2/EgrNAC170 will help to resolve the nature of its regulation.

The motivation to optimize a ChIP-seq procedure for developing secondary xylem (Chapter 4) was, in part, to study transcription factors in their native molecular context in mature field-grown trees, where gene dosage problems could be mitigated. ChIP-seq proved to be a challenging technique when applied to developing secondary xylem due to a relatively low yield of chromatin. However the global profiles of H3K4me3 I produced using a ChIP-verified antibody were of high quality as demonstrated by their efficiency (strand cross-correlation) statistics and good correlation with gene expression levels. Certainly the method described in this chapter remains a significant contribution to the field, which lacks established ChIP protocols for woody tissue. Lin *et al.* (2013) have recently reported ChIP-PCR results for poplar stem, but are yet to demonstrate the suitability of their approach for ChIP-seq. Where improvement in the method described in Chapter 4 is still essential, as echoed in Chapter 5, is the replacement of the ChIP DNA amplification step with an enhanced protocol, or library construction from the small amounts of DNA obtained after ChIP enrichment. Since the ChIP-seq results of Chapter 5 are inconclusive they will not be discussed in further detail here. Far more ChIP-seq experiments of candidate transcription factors in developing secondary xylem will need to be conducted to evaluate whether the existing protocol works as successfully for transcription factors as it does for identifying histone-DNA interactions.

© University of Pretoria

Apart from exploring the functions of SND2 and, to a limited extent, EgrNAC170, this thesis explored the epigenetic regulation of xylogenesis based on the activating histone mark H3K4me3 (Chapter 4). The data obtained in this chapter can also serve in improving transcription start site annotations of known and novel genes in future curation of the *E. grandis* genome, due to the strong association of H3K4me3 with transcription start sites (Hon *et al.*, 2009). Apart from that, I have described for the first time the participation of H3K4me3 in the regulation of genes in developing secondary xylem. Consistent with the well-known function of H3K4me3 as an activating histone mark, genes positive for H3K4me3 were almost always expressed, a relationship that was correlated with the degree of ChIP enrichment. While H3K4me3-enriched genes seem to be broadly expressed across tissues, as found in *Arabidopsis* (Zhang *et al.*, 2009), a somewhat surprising finding was that noncoding RNAs and many genes associated with xylogenesis with highly tissue-specific expression were also epigenetically regulated by H3K4me3. The generation of ChIP-seq profiles of diverse histone modifications will help to better understand and even predict how these genes are regulated by combinations of histone marks, or epigenetic states, as achieved in model organisms (e.g. Cheng *et al.*, 2011). Of course, a major challenge in epigenetics studies is to understand how histone-modifying enzymes are recruited to particular loci, and how this regulation may be manipulated to affect gene expression. In one of the most well-studied *Arabidopsis* models of this poorly understood process, several transcription factors, protein complexes and even long noncoding RNAs (Heo & Sung, 2011) appear to be involved in targeting histone-modifying enzymes to flowering loci such as *FLC* during vernalization (reviewed by Kim & Sung, 2014).

© University of Pretoria

The understanding of transcriptional and epigenetic regulation of SCW biosynthesis and xylogenesis in *Eucalyptus* will be greatly advanced through high-throughput genomics techniques such as ChIP-seq. One significant weakness of ChIP-seq is that it provides only limited information on actively regulated genes. Therefore, high-throughput overexpression systems serve to complement genomic occupancy data in identifying high-confidence direct gene targets. Recently, Lin *et al.* (2013) have developed a method for transfection of protoplasts isolated from *Populus* secondary xylem with candidate SCW-associated transcription factors, allowing identification of their gene targets using time-course RNA-seq analysis. The chromatin and co-factor environment of *Eucalyptus* xylem-derived protoplast cells could provide an ideal high-throughput platform for the rapid characterization of *Eucalyptus* candidate transcription factors linked to SCW biosynthesis, including those requiring co-factors to function (such as SND2, as postulated in Chapter 3). This, in combination with ChIP-seq, has the potential to aid in the reconstruction of extensive transcriptional networks of xylogenesis in *Eucalyptus* from a relatively small number of experiments.

References

Birchler JA, Veitia RA. 2007. The Gene Balance Hypothesis: from classical genetics to modern genomics. *The Plant Cell* **19**: 395-402.

Birchler JA, Veitia RA. 2010. The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. *New Phytologist* **186**: 54-62.

Cahill MA, Ernst WH, Janknecht R, Nordheim A. 1994. Regulatory squelching. *FEBS Letters* **344**: 105-108.

Cheng C, Yan K-K, Yip KY, Rozowsky J, Alexander R, Shou C, Gerstein M. 2011. A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biology* **12**: R15.

© University of Pretoria

Chiang VL. 2006. Monolignol biosynthesis and genetic engineering of lignin in trees, a review. *Environmental Chemistry Letters* **4**: 143–146.

Coleman HD, Yan J, Mansfield SD. 2009. Sucrose synthase affects carbon partitioning to increase cellulose production and altered cell wall ultrastructure. *Proceedings of the National Academy of Sciences of the United States of America* **106**(31): 13118-13123.

da Silveira E 2013. More cellulose per square centimeter: Transgenic *Eucalyptus* has 20% higher productivity than the conventional tree. Available online at http://revistapesquisa.fapesp.br/en/2013/04/01/more-cellulose-per-square-centimeter/.

Goh H-H, Sloan J, Dorca-Fornell C, Fleming A. 2012. Inducible repression of multiple expansin genes leads to growth suppression during leaf development. *Plant Physiology* **159**(4): 1759-1770.

Granja PL, Jéso BD, Bareille R, Rouais F, Baquey C, Barbosa MA. 2006. Cellulose phosphates as biomaterials. In vitro biocompatibility studies. *Reactive and Functional Polymers* **66**(7): 728-739.

Grant EH, Fujino T, Beers EP, Brunner AM. 2010. Characterization of NAC domain transcription factors implicated in control of vascular cell differentiation in *Arabidopsis* and *Populus*. *Planta* **232**: 337-352.

Habibi Y, Lucia LA, Rojas OJ. 2010. Cellulose nanocrystals: chemistry, self-assembly, and applications. *Chemical reviews* **110**: 3479-3500.

Heo JB, Sung S. 2011. Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. *Science* **331**: 76-79.

Hon GC, Hawkins RD, Ren B. 2009. Predictive chromatin signatures in the mammalian genome. *Human Molecular Genetics* **18**: R195-R201.

Iglesias-Trabado G, Wilstermann D 2008. *Eucalyptus universalis*. Global cultivated eucalypt forests map 2008 Version 1.0.1. In GIT Forestry Consulting's EUCALYPTOLOGICS: Information resources on *Eucalyptus* cultivation worldwide. Available online at http://www.git-forestry.com.

Kim D-H, Sung S. 2014. Genetic and epigenetic mechanisms underlying vernalization. *The Arabidopsis Book* **12**: e0171.

Kubo M, Udagawa M, Nishikubo N, Horiguchi G, Yamaguchi M, Ito J, Mimura T, Fukuda H, Demura T. 2005. Transcription switches for protoxylem and metaxylem vessel formation. *Genes and Development* **19**: 1855-1860.

Lagaron JM, Lopez-Rubio A. 2011. Nanotechnology for bioplastics: opportunities, challenges and strategies. *Trends in Food Science & Technology* **22**(11): 611-617.

© University of Pretoria

Li E, Bhargava A, Qiang W, Friedmann MC, Forneris N, Savidge RA, Johnson LA, Mansfield SD, Ellis BE, Douglas CJ. 2012. The Class II *KNOX* gene *KNAT7* negatively regulates secondary wall formation in *Arabidopsis* and is functionally conserved in *Populus*. *New Phytologist* **194**(1): 102–115.

Lin Y-C, Li W, Sun Y-H, Kumari S, Wei H, Li Q, Tunlaya-Anukit S, Sederoff RR, Chiang VL. 2013. SND1 transcription factor–directed quantitative functional hierarchical genetic regulatory network in wood formation in *Populus trichocarpa*. *The Plant Cell* **25**: 4324-4341.

Mishra AK, Mishra SB 2011. Cellulose based green bioplastics for biomedical engineering. In: S. Pilla ed. *Handbook of Bioplastics and Biocomposites Engineering Applications*. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Mitsuda N, Seki M, Shinozaki K, Ohme-Takagi M. 2005. The NAC transcription factors NST1 and NST2 of *Arabidopsis* regulate secondary wall thickenings and are required for anther dehiscence. *Plant Cell* **17**: 2993–3006.

Moura JCMS, Bonine CAV, Viana JdOF, Dornelas MC, Mazzafera P. 2010. Abiotic and biotic stresses and changes in the lignin content and composition in plants. *Journal of Integrative Plant Biology* **52**(4): 360-376.

Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D, Goodstein DM, Dubchak I, Poliakov A, Mizrachi E, Kullan ARK, van Jaarsveld I, Hussey SG, Pinard D, Merwe Kvd, Singh P, Silva-Junior OB, Togawa RC, Pappas MR, Faria DA, Sansaloni CP, Petroli CD, Yang X, Ranjan P, Tschaplinski TJ, Ye C-Y, Li T, Sterck L, Vanneste K, Murat F, Soler M, Clemente HS, Saidi N, Cassan-Wang H, Dunand C, Hefer CA, Bornberg-Bauer E, Kersting AR, Vining K, Amarasinghe V, Ranik M, Naithani S, Elser J, Boyd AE, Liston A, Spatafora JW, Dharmwardhana P, Raja R, Sullivan C, Romanel E, Alves-Ferreira M, Külheim C, Foley W, Carocha V, Paiva J, Kudrna D, Brommonschenkel SH, Pasquali G, Byrne M, Rigault P, Tibbits J, Spokevicius A, Jones RC, Steane DA, Vaillancourt RE, Potts BM, Joubert F, Barry K, Jr. GJP, Strauss SH, Jaiswal P, Grima-Pettenati J, Salse J, Peer YVd, Rokhsar DS, Schmutz J. in press. The genome of *Eucalyptus grandis* - a global tree for fiber and energy. *Nature*.

Orphanides G, Lagrange T, Reinberg D. 2006. The general transcription factors of RNA polymerase II. *Genes and Development* **10**: 2657-2683.

PICC 2013. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, by T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V.

© University of Pretoria

BexP. M. Midgley. (also available at http://www.climatechange2013.org/images/report/WG1AR5_Frontmatter_FINAL.pdf).

Vanholme R, Morreel K, Ralph J, Boerjan W. 2008. Lignin engineering. *Current Opinion in Plant Biology* **11**(3): 278-285.

Wang HH, Tang RJ, Liu H, Chen HY, Liu JY, Jiang XN, Zhang HX. 2013. Chimeric repressor of PtSND2 severely affects wood formation in transgenic *Populus*. *Tree Physiology* **33**(8): 878-886.

Weng J-K, Li X, Bonawitz ND, Chapple C. 2008. Emerging strategies of lignin engineering and degradation for cellulosic biofuel production. *Current Opinion in Biotechnology* **19**(2): 166-172.

Yang F, Mitra P, Zhang L, Prak L, Verhertbruggen Y, Kim J-S, Sun L, Zheng K, Tang K, Auer M, Scheller HV, Loqué D. 2013. Engineering secondary cell wall deposition in plants. *Plant Biotechnology Journal* **11**: 325-335.

Zhang X, Bernatavichute YV, Cokus S, Pellegrini M, Jacobsen SE. 2009. Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in *Arabidopsis thaliana*. *Genome Biology* **10**: R62.

Zhong R, Demura T, Ye ZH. 2006. SND1, a NAC domain transcription factor, is a key regulator of secondary wall synthesis in fibers of *Arabidopsis*. *Plant Cell* **18**(11): 3158-3170.

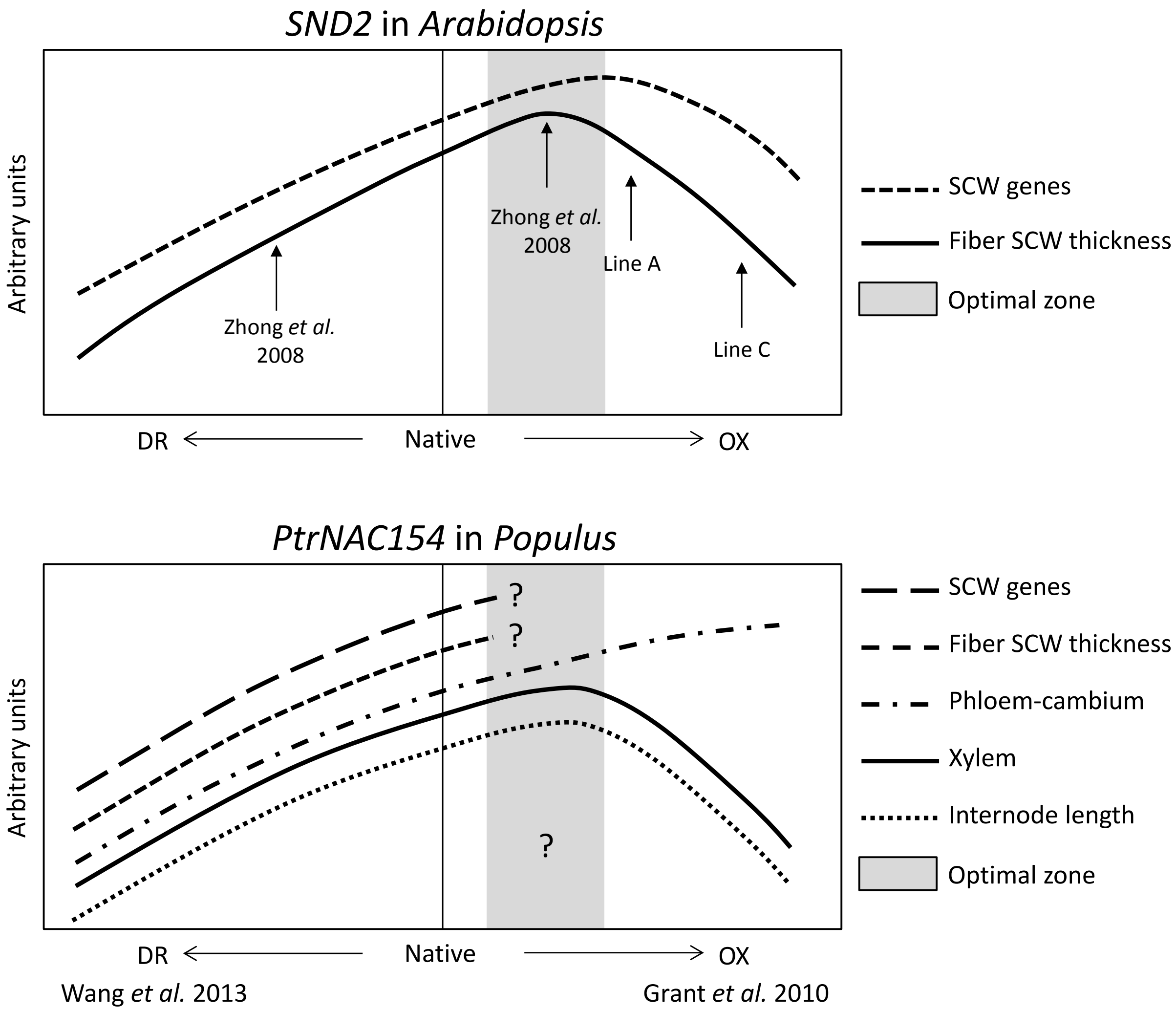
Zhong R, Lee C, Ye Z-H. 2010. Functional characterization of poplar wood-associated NAC domain transcription factors. *Plant Physiology* **152**: 1044–1055.

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

© University of Pretoria

Fig. 6.1. Integrated model of SND2-mediated regulation of SCW biosynthesis in *Arabidopsis* and *Populus*. The level of overexpression (OX) or dominant repression (DR) relative to the native transcript level is indicated on the *x*-axis for *SND2* in *Arabidopsis* (top panel) and its putative ortholog *PtrNAC154* in *Populus* (bottom panel). In *Arabidopsis*, limited *SND2* overexpression appears to increase fiber secondary cell wall (SCW) thickness, while dominant repression (Zhong *et al.*, 2008) and strong *SND2* overexpression (line C and T1 lines in Chapter 3) decreases fiber SCW thickness. In line A, where *SND2* was overexpressed more than that reported in Zhong *et al.* (2008), SCW thickness remained similar to the wild type, but upregulation of SCW biosynthetic genes was still apparent (Chapter 3). In *Populus*, a similar phenomenon appears to occur with regard to xylem production and internode length. Phloem-cambium production, which is not known to be regulated by *PtrNAC154*, appears to respond additively to PtrNAC154 activity. It is posulated that, in both *Arabidopsis* and *Populus*, an optimal range of *SND2*/*PtrNAC154* overexpression exists that leads to increased fiber SCW deposition and xylem production.

© University of Pretoria



SND2 in *Arabidopsis*

PtrNAC154 in *Populus*

© University of Pretoria