Functional genomics and systems genetics of cellulose biosynthesis in *Eucalyptus*

by

ESHCHAR MIZRACHI

Submitted in partial fulfillment of the requirements for the degree

Philosophiae Doctor

In the Faculty of Natural and Agricultural Sciences

Department of Genetics

University of Pretoria

Pretoria

September 2013

Supervisor: Prof. Alexander A. Myburg

Cosupervisors: Prof. Shawn D. Mansfield and Prof. David K. Berger

Declaration

I, the undersigned, hereby declare that the thesis submitted herewith for the degree Ph.D. to the University of Pretoria contains my own independent work and has not been submitted previously for any degree at any university.

_____

Eshchar Mizrachi

_____

Date

TABLE OF CONTENTS

THESIS SUMMARY

Functional genomics and systems genetics of cellulose biosynthesis in *Eucalyptus*

Eshchar Mizrachi

Supervised by ***Prof. A.A. Myburg***

Co-supervised by ***Prof. Shawn D. Mansfield*** *(University of British Columbia, Canada) and* ***Prof. David K. Berger*** *(University of Pretoria)*

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

Department of Genetics

University of Pretoria

The globally emerging bioeconomy demands rapid advancement in the sustainable production and utilization of bio-based raw materials for a multitude of downstream applications, particularly in the areas of food, health and bioenergy and biomaterials. These needs, particularly pertaining to plant productivity, quality and stress tolerance, will need to be addressed with advanced biotechnology strategies, which accelerate progress beyond what has been achieved with traditional breeding and cultivation methods. Woody biomass is a readily available source of renewable carbon, and trees from the genus *Eucalyptus*, displaying superior growth and wood properties and established agricultural practices worldwide, are attractive candidates as short-rotation (5-9 years) feedstocks for biofuels and biomaterials. Guiding advanced strategies in biotechnology in *Eucalyptus* and other biomass feedstocks requires a sophisticated understanding of the molecular underpinnings of carbon allocation and cell wall biology.

In the work presented here, we aimed to characterize the molecular biology of cellulose biosynthesis in *Eucalyptus* xylem (developing wood) and identify genes, processes and pathways that are linked to and possibly influence this process. We achieved this by detailed characterization of field-grown *Eucalyptus* hybrid trees, utilizing RNA-sequencing technology and metabolomics of xylem as well as measuring wood properties that are thought to impact the efficiency of industrial processing. Given the lack of information with regards to gene expression in *Eucalyptus* trees, a major aim was to characterize transcriptomes from various tissues and organs, including a cellulose-enriched form of xylem called tension wood. This involved challenging bioinformatics, which resulted in a high quality assembly and publication of a comprehensive gene catalogue for *Eucalyptus*, which was one of the first short-read RNA-sequencing based *de novo* assembly from a eukaryotic organism. We also characterized and modelled the properties of cellulose and xylan biosynthetic pathways as a biological system, the parts of which are segregating in *Eucalyptus* hybrid tree populations, which has generated novel insights into the allocation and partitioning of sequestered carbon between cellulose, xylan and lignin during active secondary cell wall deposition in woody stem tissues.

This research has made important contributions to the field of *Eucalyptus* biology, but also to the broader field of secondary cell wall biosynthesis in plants, specifically providing (i) resources for transcriptome analysis in a large woody perennial (ii) new biological insight into carbon allocation for polysaccharide biosynthesis in wood, and (iii) annotation and discovery of candidate genes and pathways that may influence wood chemical composition and structures. Importantly, we find that cellulose and xylan biosynthetic genes are transcriptionally hardwired in their co-regulation (along with other important processes for cellulose and xylan transport and deposition), likely due to the fact that they utilize a common source of sucrose-derived carbon for cell wall biosynthesis and the production of sufficient energy to do so. This co-regulation appears to be distinct from the regulation of other cell wall biopolymers. Furthermore, evidence from xylem gene expression and metabolite availability in xylem, as

well as from wood properties of field-grown trees, supports a model in which sucrose-derived cytosolic fructose is shunted to the production of lignin precursors during cellulose and xylan biosynthesis. This model parsimoniously explains a mechanism for trees to partition carbon between polysaccharide and lignin synthesis, and provides exciting new questions and potential strategies to influence carbon allocation in the secondary cell walls of woody plants.

PREFACE

The bulk of the biomass produced by woody plants is composed of cellulose (40-45%), hemicelluloses (20-25%) and lignin (25-35%), found in the secondary cell walls of tracheids (gymnosperms) and fiber cells (angiosperms). Most of the carbon in the secondary cell wall is therefore channeled towards synthesis of cellulose, a relatively simple polymer made up of repeating β-1,4 linked D-glucose molecules. It is ubiquitously found in cell walls of all vascular plants, and the mechanism of its synthesis is conserved in green plants, even in some lineages of green algae. Especially in secondary cell walls, cellulose can have a high degree of polymerization (up to 15,000 glucose molecules) and be highly crystalline. Its ordered deposition and orientation is a major determinant of plant cell form and ability to withstand internal and external pressures. Chemical cellulose, primarily derived from woody angiosperms such as *Eucalyptus* species, is already an important industrial product, and is the raw material for many high-value derivatives. In the future, more bio-based strategies for industrial applications will rely on carbon-based biopolymers such as cellulose, hemicelluloses and lignin for fuels, materials and chemical compounds.

The genus *Eucalyptus* is the most widely planted hardwood tree in the world (an estimated 20 million ha worldwide), and displays superior carbon sequestration and growth properties, as well as wood properties. As a hardwood crop for industrial applications, *Eucalyptus* is generally planted and grown outside its endemic (Australian) habitat, in a variety of tropical, subtropical and some temperate environments such as South America (especially Brazil), India, China, Central and Southern Africa and parts of the Mediterranean. Improvement of industrially and commercially relevant traits (growth and wood properties, resistance to biotic and abiotic stresses) has to date been approached from a breeding perspective, either through advancement of breeding populations within a species, or the clonal propagation of interspecific hybrid clones that display hybrid vigour (particularly as a strategy to overcome non-native pest and pathogen threats). Currently in South Africa, the world's largest chemical

cellulose producer (Sappi) relies almost completely on *Eucalyptus* as a feedstock. With limited arable land and water, maximizing the extractability of cellulose and other biopolymers from wood for a variety of industrial processes is a key target. Given today's scientific environment, the application of biotechnology solutions for feedstock improvement must be pursued to complement advances in engineering and feedstock processing.

Despite the importance of this biopolymer and decades of research, major questions in the biology of carbon allocation to and biosynthesis of cellulose remain unanswered. Fully formed wood, and the physical and chemical bonds within and between the carbon-based biopolymers, as well as proteins in the secondary cell wall, represents an evolutionarily optimized mechanical support structure for the plant. As such, it is incredibly difficult to study and observe the dynamics of cellulose biosynthesis in its native environment of complex physicochemical interactions, at least using current technologies. The application of both forward and reverse genetics approaches, mainly in the model plant *Arabidopsis thaliana*, has identified key genes and proteins and their roles in cellulose, xylan and lignin synthesis, though many genes remain unknown. What is less addressed, however, is understanding precisely how carbon is allocated from source (leaves) to sink (in trees – mainly wood), how it is partitioned for synthesis of the various biopolymers during secondary cell wall deposition, and how the deposition of these biopolymers is coordinated to produce the cell wall ultrastructure.

A major motivation for the work in this thesis was therefore the need to gain insight into the molecular biology of polysaccharide metabolism, especially cellulose and xylan, during wood formation in *Eucalyptus* trees, with the longer term aim of providing applied biotechnology solutions for wood and fiber improvement. The research reported on was structured to progressively build on current knowledge from model plants (mainly *Arabidopsis thaliana* and *Populus trichocarpa*) and characterize the roles of

genes, biological processes and pathways involved in or related to cellulose biosynthesis in *Eucalyptus*. This was done mainly by genome-wide transcriptome analysis of replicated *Eucalyptus* hybrid clones, as well as at the population level in segregating populations (progeny of interspecific backcrosses). At the individual tree level, these genes and pathways were studied in xylogenic (wood forming) and non-xylogenic tissues and organs (Chapter 2), as well as by studying transcriptome-level response and physicochemical changes in tension wood, a specialized cellulose-enriched wood of *Eucalyptus* (Chapter 3). In Chapter 4, I apply a systems genetics approach by looking at gene-gene and gene-metabolite correlations in xylem of tree populations to define a group of genes, processes and pathways that show coordinated activity during active cellulose deposition. Together this research has provided important insight into the utilization of carbon for polysaccharide and lignin biosynthesis during wood formation that was previously unknown. It has also identified genes and processes that play essential roles in wood formation. I conclude the thesis with motivation (supported by research herein) for how understanding wood as a complex system is essential in aiding future forest biotechnology strategies.

The following peer reviewed publications and conference presentations have emanated from this PhD work:

Publications in ISI rated Journals

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**(1): 681.

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

Myburg, AA, Grattapaglia, D, Tuskan, G, *et al*. Genome sequence of *Eucalyptus grandis*: A global tree crop for fiber and energy. (In Review – *Nature*).

Published conference Proceedings

Van Dyk MM, Kullan ARK, Mizrachi E , Hefer CA, Jansen van Rensburg L, Tschaplinski TJ, Cushman KC, Engle NE, Tuskan GA, Jones N, Kanzler A, Galloway G, Bayley A, Myburg AA. 2011. Genetic dissection of transcript, metabolite, growth and wood property traits in an F2 pseudo-backcross pedigree of *Eucalyptus grandis* × *E. urophylla*. *BMC Proceedings* **5**(7):7.

Presentations in national (South African) conferences

Mizrachi E, Hefer C, Ranik M, Joubert F, Myburg AA. 2009. De novo assembly and annotation of an expressed gene catalogue of a fast growing *Eucalyptus* tree hybrid and deep digital expression profiling using Illumina mRNA-Seq. Second Southern African Bioinformatics Workshop, Riverside Hotel and Conference, 11-12 October 2009 (poster).

Mizrachi E, Hefer C, Ranik M, Joubert F, Myburg AA. 2010. Sequencing and annotation of more than 18,000 expressed genes from a fast-growing *Eucalyptus* hybrid clone using Illumina mRNA-Seq and *de novo* assembly. South African Genetics Society 2010 Congress, Bloemfontein, 9-10 April 2010 (oral presentation).

Myburg AA, Mizrachi E, Van Dyk MM, Kullan ARK, Hefer CA and Joubert F. 2012. Genomics of wood formation in field-grown *Eucalyptus* hybrid trees. Southern African Plant Breeding Symposium, Protea Hotel Kruger Gate, Skukuza. 12-14 March 2012 (oral presentation).

Mizrachi E, Myburg AA, The Forest Molecular Genetics Group. 2012. The tree as a feedstock in the emerging bioeconomy. ASSAf-DST-NRF Third Annual South African Young Scientists' Conference: *Our Energy Future*. CSIR International Convention Centre, Pretoria, 16-18 October 2012. (oral presentation).

Presentations in International conferences

Ranik M, Mizrachi E, Hefer C, Uys P, Joubert F, Myburg AA. 2009. Whole-transcriptome sequencing of a *Eucalyptus* hybrid clone using Illumina RNA-Seq. Plant & Animal Genome XVII Conference W179, January 10-14, 2009. San Diego, CA. (poster).

Mizrachi E, Hefer C, Ranik M, Celton JM, Rees JG, Joubert F, Myburg AA. 2009. Whole transcriptome sequencing of an interspecific hybrid of *Eucalyptus* using Illumina mRNA-Seq: Preliminary assembly and challenges. CHI Next Generation Sequencing Conference, 17-19 March, 2009. San Diego, California (poster).

Mizrachi E, Hefer C, Ranik M, Celton JM, Rees JG, Joubert F, Myburg AA. 2009. Whole transcriptome sequencing of an interspecific hybrid of *Eucalyptus* using Illumina mRNA-Seq: Preliminary assembly and challenges. DOE-JGI User Meeting, 25-27 March, 2009, Walnut Creek, California (poster).

Mizrachi E, Hefer C, Ranik M, Joubert F, Myburg AA. 2009. *De novo* whole-transcriptome sequencing of an F1 interspecific hybrid of *Eucalyptus*. IUFRO Tree Biotechnology 2009 Meeting. June 28 – July 2, 2009. Whistler, British Columbia, Canada. (oral presentation).

Myburg AA, Ranik M, Kullan A, Creux NM, De Castro MH, Silberbauer J, Hussey S, Mizrachi E. 2009. *Eucalyptus* genomics for renewable energy and fibre production. International Tropical Crop Biotechnology Conference, July 22 – 25, 2009. Hazyview, South Africa. (plenary presentation).

Myburg AA, Ranik M, Creux NM, De Castro MH, Silberbauer J, Hussey S, Mizrachi E. 2009. Rosettes, transcriptomes and paper: Molecular biology and genetic regulation of cellulose biosynthesis in eucalypts. 9th IPMB (International Plant Molecular Biology) Congress, October 25-30, 2009. St. Louis, Missouri, USA (oral presentation).

Mizrachi E, Hefer CA, Ranik M, Joubert F, Myburg AA. 2010. Prelude to a Genome: *De novo* assembly, annotation and profiling of an expressed gene catalog of a fast-growing *Eucalyptus* hybrid clone using Illumina mRNA-Seq. Plant & Animal Genome XVIII Conference W233, January 9-13, 2010. San Diego, CA. (oral presentation).

Mizrachi E, Hefer C, Ranik M, Joubert F, Myburg AA. 2010. *De novo* assembly and annotation of more than 18,000 expressed genes derived from Illumina mRNA-Seq analysis of a fast-growing *Eucalyptus* plantation tree. Illumina user meeting, Sitges, Spain, 13-15 July 2010 (invited oral presentation).

Mizrachi E, Hefer C, Ranik M, Myburg AA, Joubert F. 2011. Towards a *Eucalyptus* gene expression atlas database. Plant & Animal Genome XIX Conference P779, January 15-19, 2011. San Diego, CA. (poster).

Hefer CA, Mizrachi E, Myburg AA, Joubert F. 2011. Eucspresso: A *Eucalyptus* gene expression database for next-generation transcriptome sequencing data. ISCB Africa ASBCB Conference on Bioinformatics, Cape Town, South Africa, 9-11 March, 2011. (poster).

Mizrachi E, Van Dyk MM, Kullan ARK, Hefer CA, Joubert F, Myburg AA. 2011. Systems genetics of wood formation in *Eucalyptus*. International Botanical Congress, Melbourne, Australia, 23-30 July, 2011. (invited oral presentation).

Mizrachi E, Myburg AA. 2011. *Eucalyptus* genome and transcriptome resources for pulp, paper and bioenergy. Agricultural Biotechnology International Conference (ABIC), Sandton, South Africa, Sept 6-9, 2011. (invited oral presentation).

Myburg AA, Van Dyk MM, Kullan ARK, Hefer CA and Mizrachi E. 2012. From Genome to Systems Genetics: Fast-tracking *Eucalyptus* Genomics and Biotechnology. International Congress on Plant Molecular Biology, 21-26 October 2012, JeJu, Korea (oral presentation).

Myburg AA, Mizrachi E, Grattapaglia D, Tuskan GA and The *Eucalyptus* Genome Network (EUCAGEN). 2013. Population-Wide Transcriptome Sequencing: Empowering Systems Genetics in *Eucalyptus*. Plant & Animal Genome XXI Conference W297, January 12-16, 2013. San Diego, CA. (oral presentation).

Myburg AA, Mizrachi E, Grattapaglia D, Tuskan GA and The *Eucalyptus* Genome Network (EUCAGEN) 2013. Systems Genetics and Genomics of Woody Biomass Production in *Eucalyptus*,a Global Fibre Crop. 8th Annual DOE JGI User Meeting: "*Genomics of Energy & Environment"*, March 26-28, 2013. Walnut Creek, CA. (oral presentation).

Myburg AA, Grattapaglia D, Tuskan GA, Mizrachi E, Coetzer N, The *Eucalyptus* Genome Network (EUCAGEN). 2013. The *Eucalyptus* genome sequence - from reference to population genomics. IUFRO Tree Biotechnology Meeting, Asheville, NC, 26 May-2 June, 2013. (oral presentation).

Mizrachi E, Verbeke L, Van der Merwe K, Fierro AC, Van Parys T, Mansfield SD, Marchal K, Van de Peer Y, Myburg AA. 2013. Systems genetics analysis of cell wall deposition in *Eucalyptus* xylem. IUFRO Tree Biotechnology Meeting, Asheville, NC, 26 May-2 June, 2013. (oral presentation).

ACKNOWLEDGEMENTS

I would like to sincerely thank the following people and institutions:

- Prof. Zander Myburg, my supervisor and mentor, for his passion, insight, expertise and patience, for always pushing boundaries, and supporting the development and independence of myself and other young scientists.

- My co-supervisor Prof. Shawn Mansfield, for his tremendous insight into wood biology, for constantly challenging me and for teaching me that only the highest scientific standards are an acceptable minimum expectation.

- My co-supervisor Prof. Dave Berger, for his perceptive insight and critique of the research.

- My wife, Jo, for her love, optimism, patience and encouragement that got me through the hardest times and quickly made them feel easy.

- My family (my parents, Elad, Rita, Danit, Meni, Itai, Jo, Greg, Marie, Margot, Michael and Nicky), for their love, support and encouragement.

- All named co-authors on the research presented in these chapters, especially Martin Ranik for introducing me to CesAs and cellulose synthesis, and Charles Hefer for his expertise, friendship, and all the patience required for a bioinformatician when working with a molecular biologist.

- Minique de Castro, Nicky Creux, Marja O'Neill and Dr. Vinet Coetzee, for their scientific insight and letting me test a million theories with them.

- Jabs, Jon, Omar, Riaz, Zain, Ritesh and Shani, and all my other friends for their support, friendship and comic relief in my life.

- All my friends, lab mates, colleagues and students under my supervision at FMG, for their spirit, passion and always finding a cause for celebration.

- The Department of Genetics and FABI at the University of Pretoria for providing a stimulating and supportive environment for research.

- Profs. Henk Huismans, Brenda Wingfield, Jaco Greef, Fourie Joubert, Bernard Slippers and Yves van de Peer for their wisdom and advice on my scientific development.

- Sappi, The University of Pretoria, the Department of Science and Technology, The Department of Trade and Industry and the National Research Foundation for funding the research, and for having the vision and insight to drive biotechnological innovations in the advancement of woody biomass.

- The entire team at Sappi – especially Dr. Arlene Bayley, Dr. Andrew Morris, Dr. Nicky Jones, Geoff Galloway, Dr. Arnulf Kanzler and Dr. Charlie Clarke– for committing so much of their time, effort and energy to this research and always being supportive of new ideas.

Dedicated to my parents – Hezi and Sari Mizrachi. Your love, support and belief in my abilities throughout my life have always made me feel I can achieve anything.

CHAPTER 1

LITERATURE REVIEW

Advancing forest tree biotechnology in the woody biomass industry

Eshchar Mizrachi[1], Shawn D Mansfield[2] and Alexander A Myburg[1]

[1]Department of Genetics, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Private bag X20, Pretoria, 0028, South Africa; [2]Department of Wood Science, University of British Columbia, Vancouver, BC, Canada, V6T 1Z4

This chapter has been prepared in the format of a manuscript for a peer-reviewed research journal. I conceived of and drafted the manuscript, and prepared all of the figures. Shawn Mansfield and Alexander Myburg helped draft the manuscript. This review was published as one of four reviews included in a special issue of New Phytologist titled “Bioenergy Trees”, under the title “Cellulose factories: Advancing bioenergy production from forest trees.” (Mizrachi *et al***.** *New Phytologist* **194**(1): 54-62).

1.1 Summary

Fast-growing, short-rotation forest trees, such as *Populus* and *Eucalyptus*, produce large amounts of cellulose-rich biomass that could be utilized for bioenergy and biopolymer production. Major obstacles need to be overcome before the deployment of these genera as energy crops, including the effective removal of lignin and the subsequent liberation of carbohydrate constituents from wood cell walls. However, significant opportunities exist to both select for and engineer the structure and interaction of cell wall biopolymers, which could afford a means to improve processing and product development. The molecular underpinnings and regulation of cell wall carbohydrate biosynthesis are rapidly being elucidated, and are providing tools to strategically develop and guide the targeted modification required to adapt forest trees for the emerging bioeconomy. Much insight has already been gained from the perturbation of individual genes and pathways, but it is not known to what extent the natural variation in the sequence and expression of these same genes underlies the inherent variation in wood properties of field-grown trees. The integration of data from next-generation genomic technologies applied in natural and experimental populations will enable a systems genetics approach to study cell wall carbohydrate production in trees, and should advance the development of future woody bioenergy and biopolymer crops.

1.2 Introduction

With the growing need for alternative sources of energy and raw materials, fast-growing plantation tree species, such as *Populus* and *Eucalyptus*, are important candidates for renewable sources of lignocellulosic biomass (for recent reviews on the feasibility of bioenergy production from wood biomass, refer to (Carroll & Somerville, 2009; Hinchee *et al.*, 2009; Mansfield, 2009; Richard, 2010; Somerville *et al.*, 2010; Séguin, 2011). These two genera, broadly representing the Northern and Southern Hemisphere, respectively, produce large amounts of woody biomass (> 50 m^3 ha^{-1} yr^{-1} for eucalypts in highly productive areas, such as Brazil) in relatively short rotation times and, in general, do not infringe on land dedicated to food crop production. In addition, contrary to agriculture-derived biomass, tree-derived lignocellulosics can be harvested all yearround to ensure a stable, predictable and constant supply of raw material for bioenergy or biofuel production. Establishment costs and carbon footprints of multiyear forest plantations are also lower than those of annually planted crops, especially for coppicing eucalypt species, which can be grown on marginal lands (Hinchee *et al.*, 2009). Well-established industrial breeding programmes already exploit the substantial inherent genetic variation available in these genera, which can be (and has been) expanded with interspecific hybridization, and ultimately captured in clonal plantations (Grattapaglia *et al.*, 2009). The processing of wood fibre and, especially, cellulose from woody biomass has been improved and optimized for decades, providing a technology base from which to develop processing plants for biofuels and biomaterials. One major consideration that is often overlooked when forecasting bioenergy feedstocks is that this bioenergy end-use will have to compete with the high-value products derived from chemical cellulose and its derivatives (Fig. 1), and the desired traits for many bioenergy applications are common to those desired for chemical cellulose production. Thus, the objective of improving feedstock characteristics in trees is complementary to current tree breeding programmes directed at traditional forest-reliant industries.

Cellulose-rich biomass derived from fast-growing tree species offers many advantages over agricultural feedstocks for bioenergy production, but the removal of lignin to facilitate the effective and efficient extraction of cell wall carbohydrates remains one of the primary hurdles (Studer *et al.*, 2011). To efficiently deconstruct lignocellulosic biomass, a detailed understanding of how wood cell walls are synthesized, deposited and modified in planta is required (Mansfield, 2009). Recent research has mainly focused on the modification of lignin, the most abundant natural biopolymer after cellulose (Vanholme *et al.*, 2008), but much remains to be learned about the possibilities for modifying and regulating the synthesis of cellulose, ultimately impacting on the overall chemistry and ultrastructure of wood cell walls. Although major advances have been made in understanding the biosynthesis of cellulose itself (Joshi & Mansfield, 2007), the underlying cellular and biochemical processes that influence cellulose properties in wood cell walls have not yet been fully dissected.

Most of our current knowledge of cellulose biosynthesis stems from studies in model herbaceous plants, such as *Arabidopsis thaliana*, and, to some extent, the extension of this knowledge to woody plant genera, such as *Populus* (Joshi *et al.*, 2011). The poplar genome sequence (Tuskan *et al.*, 2006) has been available for 5 yr and, as of 2011, the genome sequence of *Eucalyptus grandis* (Myburg *et al.*, in preparation) has also been publicly available (http://www.phytozome.net). These two landmark achievements have opened up new avenues for exploiting the genetic variation in forest trees, and strategically improving the physicochemical properties of woody biomass. The availability of a genome sequence is particularly important for *Eucalyptus*, the most widely grown hardwood crop in the world (c. 20 million ha). With advances in next-generation sequencing technologies, comparative genomics can now be applied to rapidly adopt the information learned from herbaceous models and other woody plants, such as poplars, to accelerate *Eucalyptus* improvement. However, with so many candidate genes known to influence xylogenesis, how does one prioritize targets when considering forest trees as bioenergy crops? How can

one expand the fundamental understanding of the biology and biosynthesis of cellulose and its interaction with other wood cell wall polymers?

Here, we provide a current summary of the general understanding of the molecular biology of cellulose production in plants, and discuss how the integration of emerging functional genomics technologies with the wealth of fundamental information on wood properties in tree breeding programmes could be used to accelerate the improvement of cellulose and bioenergy potential in trees.

1.3 An integrated view of the proteins involved in cellulose biosynthesis and deposition

Historically, the biosynthesis of cellulose has focused on the plasma membrane-located cellulose synthase (CESA) proteins that constitute the active synthesizing complex (CSC; cellulose synthase complex), which is ultimately responsible for producing the polymeric glucan chains that coalesce to form cellulose microfibrils in primary and secondary cell walls of plants (Delmer, 1999; Doblin *et al.*, 2002; Saxena & Brown, 2005; Somerville, 2006; Bessueille & Bulone, 2008; Taylor, 2008; Guerriero *et al.*, 2010). Building on these solid foundations, our current understanding requires an integrated view that incorporates a diverse set of proteins and regulatory mechanisms to fully understand this intricate biological process. Such a view should take into consideration the variety of cellular processes and metabolic fluxes that could, and do, influence the synthesis, deposition and physical properties of cellulose in the two distinctly different cell walls. This holistic view should also include the inherent and tightly regulated interactions of cellulose with other cell wall biopolymers, such as lignin and hemicellulose. For example, the biosynthesis and deposition of xylan, a major constituent of the dicot secondary cell wall (Scheller & Ulvskov, 2010), is closely coordinated with the deposition of cellulose (Hertzberg *et al.*, 2001; Schrader *et al.*, 2004). Thus, to advance our fundamental understanding and

further the biotechnological objectives of improving cellulose-rich resources, research areas to be explored should focus on the transcriptional regulation of xylem-forming genes, as well as post-translational modification, protein folding and protein complex assembly, substrate (metabolite) production, transport and availability, the transport of proteins and / or polysaccharides between organelles and to the plasma membrane, and signalling and feedback between the extracellular environment and the cytoplasm, organelles and nucleus.

Using *Arabidopsis* as the primary model, the current architecture of proteins and the cellular processes thought to be involved in, or influence, the biosynthesis and deposition of cellulose and xylan are illustrated in Fig. 1.2. At the level of transcriptional regulation, several transcription factors have been shown to directly regulate secondary cell wall CesA genes in *Arabidopsis* (Zhong *et al.*, 2008; Yamaguchi *et al.*, 2010; Xie *et al.*, 2011). Three of these – SND2, SND3 and MYB103 – appear to specifically regulate secondary cell wall CesA genes, but not xylan or lignin genes (Zhong *et al.*, 2008). These transcription factors are part of a complex transcriptional network regulating various aspects of xylogenesis, the extent of which is still being resolved in *Arabidopsis* (Kubo *et al.*, 2005; Zhong *et al.*, 2006; Demura & Fukuda, 2007; Zhong *et al.*, 2007; Zhong *et al.*, 2008), as well as, more recently, in *Populus* (McCarthy *et al.*, 2010; Zhong *et al.*, 2010; Zhong & Ye, 2010; Zhong *et al.*, 2011).

CESA proteins are synthesized and assembled into complexes in the endoplasmic reticulum (Rudolph, 1987) and, with the help of chaperones, packaged and delivered to the Golgi (Haigler & Brown Jr, 1986). The Golgi (Fig. 1.2) is also the site for xylan biosynthesis (Bolwell & Northcote, 1983), which can be divided, simplistically, into primer synthesis (PARVUS), chain elongation (IRX9, 10 and 14) and side chain modifications by IRX7, IRX8, PGSIP1, DUF579- and / or DUF231-containing proteins (Brown *et al.*, 2007; Lee *et al.*, 2007; York & O'Neill, 2008; Brown *et al.*, 2009; Wu *et al.*, 2009; Wu *et al.*, 2010;

Brown *et al.*, 2011; Jensen *et al.*, 2011). Once the CSCs are assembled, they are transported from the Golgi to the plasma membrane, via the trans-Golgi network, in specialized microtubule-associated compartments (MASCs; Crowell *et al.*, 2009) that interact with actin through MYOSIN (Wightman & Turner, 2008; Szymanski, 2009). At the plasma membrane, MASCs interact with cortical microtubules, possibly, but not conclusively, via KINESIN, and bud vesicles containing CSCs that fuse with and become embedded in the plasma membrane (Giddings *et al.*, 1980; Szymanski, 2009; Crowell *et al.*, 2010).

On the cytoplasmic face (Fig. 2), the CSCs associate with cortical microtubules, putatively through kinesin-like proteins, such as FRAGILE FIBER 1 (FRA1; Zhong *et al.*, 2002), CESA-interactive protein 1 (CSI1; Gu *et al.*, 2010) and other microtubule-associated proteins (MAPs). It is therefore apparent that cortical microtubule organization is extremely important in the regulation and deposition of cellulose, and the structure and orientation of said cortical microtubules are influenced by a variety of factors. From the assembly of α- and β-TUB at microtubule assembly sites containing γ-TUB and Gamma-complex proteins (Pastuglia & Bouchez, 2007; Cai, 2010), the growth and modification of the microtubules are influenced by strong association with actin via KCH (kinesin with calponin-homology domain) and MAP190 (Cai, 2010), association with other microtubules via MAP65-1, MAP 200, TBMP 200 (TOBACCO MICROTUBULE BUNDLING POLYPEPTIDE) and/or MICROTUBULE ORGANIZATION 1 (MOR1; Cai, 2010), and association with the plasma membrane via proteins such as END-BINDING 1 (EB1; Morrison, 2007), P-161 (Cai *et al.*, 2005), *A. thaliana* KINESIN 5 (ATK5) (Ambrose & Cyr, 2007; Pastuglia & Bouchez, 2007), SPIRAL 1 (SPR1; Nakajima *et al.*, 2004; Sedbrook *et al.*, 2004; Nakajima *et al.*, 2006), cytoplasmic linker proteins (CLIPs) and CLIP-associating proteins (CLASPs; Galjart, 2005; Ambrose & Wasteneys, 2008) and PHOSPHOLIPASE-D (Cai, 2010). Microtubule length and organization are also modified by KATANIN (McNally & Vale, 1993; Burk *et al.*, 2001; Stoppin-Mellet *et al.*, 2006; Sharma *et al.*, 2007), and therefore can have an impact on the

quality and quantity of cellulose. Transamination, tyrosylation or acetylation of microtubules can influence the binding of KINESIN proteins, whereas glutamination or glycylation of microtubules has been shown to influence KATANIN activity (Cai, 2010). These, and other as yet unidentified proteins, could all potentially have direct or indirect effects on cellulose deposition via their influence on cortical microtubule dynamics.

Movement of the CSC along the membrane is believed to be driven by the force of cellulose microfibril synthesis itself against the cell wall matrix (Diotallevi & Mulder, 2007), and is guided by the cortical microtubules (Paredez *et al.*, 2006), with membrane-associated sucrose synthase (SUSY) providing uridine diphosphate (UDP)-glucose as substrate for the CSC (Fig. 2). Towards the cell wall side, KORRIGAN (KOR; Lane *et al.*, 2001) and possibly other glycosyl hydrolases (GHs) edit elongating cellulose chains as they are synthesized, whereas COBRA/COBL and possibly other glycosylphosphatidylinositol (GPI)-anchored proteins, as well as the fasciclin-like arabinogalactan (FLA) proteins and/or other arabinogalactan proteins (AGPs), are thought to interact with cellulose as it is deposited, and concurrently relay signals back to the cytoplasm to regulate its synthesis (Zhang *et al.*, 2003; Seifert & Roberts, 2007; MacMillan *et al.*, 2010).

The mediation of cell wall feedback signalling is carried out by a number of pathways and, recently, the Rop/Rac guanosine triphosphatases (GTPases) (Fig. 1.2), which are regulated by RIC (ROP-INTERACTIVE CRIB MOTIF-CONTAINING PROTEIN) and ROPGEF (RHO GUANYL-NUCLEOTIDE EXCHANGE FACTOR), have been highlighted as playing an important role in cell wall signalling, together with IQD (IQ DOMAIN) and CTL (CHITINASE-LIKE) proteins, and wall-associated kinases (WAKs), such as leucine-rich repeat (LRR)-receptor kinases (Oikawa *et al.*, 2010). The LRR-receptor kinases include, amongst others, THESEUS (Hématy *et al.*, 2007) and

KOBITO/ELONGATION DEFECTIVE 1 (ELD1; Pagant *et al.*, 2002; Lertpiriyapong & Sung, 2003), both of which have been shown to have an impact on cell wall properties. In the secondary cell wall, laccases (LACs) and other peroxidases oxidize monolignols, leading to the random coupling of lignin monomers and resulting in the synthesis of the macromolecule lignin polymer (Boerjan *et al.*, 2003; Ralph *et al.*, 2004; Mattinen *et al.*, 2008; Berthet *et al.*, 2011), whereas other as yet unidentified GHs and carbohydrate binding module (CBM)-containing proteins appear to be involved in the mediation of cellulose-cellulose, cellulose-xylan, xylan-xylan or xylan-lignin interactions as the different biopolymers are synthesized, deposited and arranged.

In addition to the cellular processes and specific proteins involved in cellulose deposition itself, it is important to consider the metabolic flux and channelling to the various biochemical pathways that lead to the synthesis of cellulose and xylan. For example, a key metabolite is UDP-glucose, which is the immediate precursor for cellulose biosynthesis by CESA proteins. In addition, UDP-glucose can be readily converted to UDP-xylose for xylan biosynthesis (Fig. 1.3). UDP-glucose is produced directly via the hydrolysis of sucrose by sucrose SUSY or indirectly by invertase (Barratt *et al.*, 2009; Kleczkowski *et al.*, 2010), which cleaves sucrose to monomeric glucose and fructose. Monomeric glucose is then converted to UDP-glucose via phosphorylation of the 6' position (HEXOKINASE/GLUCOKINASE), followed by the substitution of the phosphate to the 1' position (PHOSPHOGLUCOMUTASE) and the subsequent substitution of the phosphate group with UDP by UTP-glucose-1-phosphate uridylyltransferase (UGP). UDP-glucose can be directly employed by CESA proteins for cellulose biosynthesis, or converted to UDP-xylose via conversion to UDP-D-glucuronate by UDP-glucose 6-dehydrogenase (UGD), followed by the removal of CO_2 by uridine-diphosphoglucuronate decarboxylase (UXS). UDP-xylose is then utilized as the backbone for xylan biosynthesis, with the addition of glucuronic acid (GlcA) and acetyl groups to the backbone or side chains to form heteroxylan.

Studies have shown that alterations in the metabolic flux of UDP-glucose can indeed affect the relative abundance and structure of cell wall polysaccharides. For example, up-regulation of SUSY in poplar trees resulted in an increase in cell wall thickness of fibres and the production of more cellulose that displayed enhanced crystallinity (Coleman *et al.*, 2009). The combination of SUSY and UGP overexpression in tobacco also resulted in a synergistic increase in plant height and biomass (Coleman *et al.*, 2006). It should be noted that the overall phenotypic effect of increased SUSY or UGP levels is dependent on the source and sink sugars and other metabolites (Haigler *et al.*, 2001; Coleman *et al.*, 2009; Meng *et al.*, 2009), which will vary in different plant species, and under an array of physiological conditions. These studies demonstrate that changes in metabolite levels, through intracellular and intercellular transport or enzymatic activity, could greatly influence the resulting abundance and/or structure of cell wall polysaccharides.

1.4 Towards systems genetics of cellulose production in trees

The scale of cellulose biosynthesis and biomass production in fast-growing plantation trees is vastly different from that in herbaceous models. There is an emphasis on large-scale cambial cell differentiation, cell elongation, secondary cell wall deposition and programmed cell death. The tremendous strength of the sink tissue means that the tree as a system must prioritize the channelling of carbon flow towards the synthesis of xylem biopolymers. Therefore, information cannot always be directly extended from herbaceous models to trees – good examples of this are the different outcomes that resulted from the overexpression of SUSY in tobacco plants (Coleman *et al.*, 2006) as opposed to poplar (Coleman *et al.*, 2009), and the fact that, for *Arabidopsis*, INVERTASE is necessary and sufficient for normal growth, whereas direct UDP-glucose production through SUSY is not (Barratt *et al.*, 2009). Recent findings have also suggested that the transcriptional network regulating cell wall biopolymer synthesis in woody plants

may be more complex and comprise novel transcription factors not previously linked to secondary cell wall formation in *Arabidopsis* (Zhong *et al.*, 2011). This implies the need to independently study the functions of secondary cell wall-related genes in trees. Some practical considerations are that very few commercial species and clonal genotypes have optimized transformation protocols, mature wood properties take several years to acquire and wood properties are complex traits affected by large numbers of genes. Rigorous glasshouse studies and field trials are required for each candidate, and these carry significant economical, ecological and regulatory burdens (for recent reviews on this issue, see Strauss *et al.*, 2009; Ahuja, 2011; Harfouche *et al.*, 2011). What is required is an approach that would prioritize genes or pathways that underlie variation in wood properties in mature, field-grown trees.

At our disposal is a rich history of tree breeding, resulting in large, structured populations, and large amounts of genetic diversity in these populations (Sederoff *et al.*, 2009; Neale & Kremer, 2011). These resources have been exploited through the application of molecular marker technologies and forward genetics approaches in multiple forest tree pedigrees, where high linkage disequilibrium (LD) has allowed the efficient identification of quantitative trait loci (QTLs; Grattapaglia & Kirst, 2008), as well as in large association populations where low LD has allowed the association of single genes with wood properties (Groover, 2007; Neale & Ingvarsson, 2008). Single gene associations detected in *Eucalyptus* and *Populus* (Thumma *et al.*, 2005; Thumma *et al.*, 2009; Wegrzyn *et al.*, 2010) have not always been intuitive – for example, the association between a lignin gene (cinnamoyl CoA reductase, CCR) and a physical cellulose property (microfibril angle) in *Eucalyptus* (Thumma *et al.*, 2005). This illustrates that our understanding of the causal relationship of genes and complex traits is still incomplete.

Phenotypic variation in tree breeding populations is influenced by a variety of intrinsic (and measurable) biological processes, mainly those of transcriptional and translational regulation of various biochemical

pathways (Du & Groover, 2010), as well as the flux of metabolic intermediates in these pathways (Mansfield, 2009). In addition, these biological processes are strongly influenced by environmental cues and seasonal variation over the lifetime of these long-lived organisms (Groover, 2007). A more holistic research approach encompassing genetic, biochemical and environmental variation must therefore be adopted to understand and improve wood property traits in trees.

Systems genetics (Fig. 1.4) connects the intermediate components of a complex phenotype (e.g. transcript, protein and metabolite levels) in related individuals to measurable phenotypic traits, such as wood properties or bioenergy potential, in the context of the underlying genetic variation in populations (MacKay *et al.*, 2009; Nadeau & Dudley, 2011). An extension of genetical genomics (Jansen & Nap, 2001), systems genetics is a network approach that explores the interconnectedness of the component levels of biological variation. It has been successfully applied in model organisms, such as *Drosophila* (Ayroles *et al.*, 2009; Morozova *et al.*, 2009; Jumbo-Lucioni *et al.*, 2010) and mouse (Farber *et al.*, 2011). It has also been applied in humans (Plaisier *et al.*, 2009; Romanoski *et al.*, 2010) and, importantly, in animal breeding (Kadarmideen *et al.*, 2006; Kadarmideen & Janss, 2007), which has many similarities to plant breeding. The power of systems genetics is that it reveals emergent properties of the system, providing insight into novel gene–gene, gene–trait and trait–trait relationships that would not be detected at the level of the individual. This often allows the reconstruction of complex directional gene regulatory networks and metabolic pathways (Kadarmideen *et al.*, 2006; Keurentjes *et al.*, 2007), adding insight into previously identified single gene associations and the molecular basis of QTLs. Systems genetics could also explain the biology underlying complex phenomena, such as G × E interactions, epigenetic control, biotic and abiotic interactions and hybrid vigour (heterosis), which are key themes to be addressed in tree improvement in the near future.

Tree breeding programmes already make use of structured pedigrees and populations replicated across environments, and therefore present an ideal starting place for systems genetics. Variation in transcriptomes has already been studied at the population level in *Eucalyptus* (Kirst *et al.*, 2005; Grattapaglia & Kirst, 2008) and *Populus* (Drost *et al.*, 2010). Transcriptome, proteome and metabolome profiling at the population level will allow integrated modelling of biomass production in trees. Systems genetics is complementary to fundamental biological investigations performed in model organisms and will also complement association genetics approaches and genomic selection strategies that are being implemented in forest tree breeding programmes (Grattapaglia & Resende, 2011). Moreover, systems genetics will allow the identification and prioritization of candidate genes for functional genetic testing in glasshouse and field trials of forest trees.

1.5 Conclusion

An understanding of how cellulose is deposited during xylogenesis in wood fibre cells has important implications for our ability to manipulate and select for industrially important traits in trees. We also need to understand the complex genetic relationships and biochemical interactions that underlie wood property variation in tree populations. The application of next-generation DNA and RNA sequencing (Mizrachi *et al.*, 2010), and the adoption of high throughput proteomics and metabolomics technologies in trees (Abril *et al.*, 2011; Plomion *et al.*, 2011; Robinson *et al.*, 2011), will allow integrated approaches to study complex relationships of genes, metabolites and wood (bio)chemistry traits at the population level. A systems genetics approach, which also includes the measurement of bioenergy potential, is a viable and increasingly cost-effective method to dissect complex phenotypes in trees, and will complement genomic selection efforts. It will also permit us to address the fundamental question of whether the same genes linked to cell wall biosynthesis by functional genetic studies in individual genotypes also influence cell wall properties in natural or experimental populations. In addition, the diversity of applications of next-

generation DNA sequencing will enable the investigation of other types of regulation, such as allele-specific expression, splice site variation, gene regulation by endogenous small RNAs or epigenetic modification, which may have an impact on the bioenergy potential of forest trees. Finally, the completion of additional tree genome sequences will permit comparative genomics approaches to dissect vital biosynthetic pathways important to industrial trait development, which should form the foundations of the emerging bio-based economy.

1.6 References

Abril N, Gion JM, Kerner R, Müller-Starck G, Cerrillo RMN, Plomion C, Renaut J, Valledor L, Jorrin-Novo JV. 2011. Proteomics research on forest trees, the most recalcitrant and orphan plant species. *Phytochemistry* **72**(10): 1219-1242.

Ahuja MR. 2011. Fate of transgenes in the forest tree genome. *Tree Genetics and Genomes* **7**(2): 221-230.

Ambrose CJ, Wasteneys GO. 2008. CLASP modulates microtubule-cortex interaction during self-organization of acentrosomal microtubules. *Molecular Biology of the Cell* **19**(11): 4730-4737.

Ambrose JC, Cyr R. 2007. The kinesin ATK5 functions in early spindle assembly in *Arabidopsis*. *Plant Cell* **19**(1): 226-236.

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.

Barratt DHP, Derbyshire P, Findlay K, Pike M, Wellner N, Lunn J, Feil R, Simpson C, Maule AJ, Smith AM. 2009. Normal growth of *Arabidopsis* requires cytosolic invertase but not sucrose synthase. *Proceedings of the National Academy of Sciences of the United States of America* **106**(31): 13124-13129.

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

Bessueille L, Bulone V. 2008. A survey of cellulose biosynthesis in higher plants. *Plant Biotechnology* **25**(3): 315-322.

Boerjan W, Ralph J, Baucher M 2003. Lignin Biosynthesis. *Annual Review of Plant Biology*. **51**(4): 519-546.

Bolwell GP, Northcote DH. 1983. Arabinan synthase and xylan synthase activities of *Phaseolus vulgaris*. Subcellular localization and possible mechanism of action. *Biochemical Journal* **210**(2): 497-507.

Brown D, Wightman R, Zhang Z, Gomez LD, Atanassov I, Bukowski JP, 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. *Plant Journal* **66**(3): 401-413.

Brown DM, Goubet F, Wong VW, Goodacre R, Stephens E, Dupree P, Turner SR. 2007. Comparison of five xylan synthesis mutants reveals new insight into the mechanisms of xylan synthesis. *Plant Journal* **52**(6): 1154-1168.

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*. *Plant Journal* **57**(4): 732-746.

Burk DH, Liu B, Zhong R, Morrison WH, Ye ZH. 2001. A katanin-like protein regulates normal cell wall biosynthesis and cell elongation. *Plant Cell* **13**(4): 807-827.

Cai G. 2010. Assembly and disassembly of plant microtubules: Tubulin modifications and binding to MAPs. *Journal of Experimental Botany* **61**(3): 623-626.

Cai G, Ovidi E, Romagnoli S, Vantard M, Cresti M, Tiezzi A. 2005. Identification and characterization of plasma membrane proteins that bind to microtubules in pollen tubes and generative cells of tobacco. *Plant and Cell Physiology* **46**(4): 563-578.

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

Coleman HD, Ellis DD, Gilbert M, Mansfield SD. 2006. Up-regulation of sucrose synthase and UDP-glucose pyrophosphorylase impacts plant growth and metabolism. *Plant Biotechnology Journal* **4**(1): 87-101.

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.

Crowell EF, Bischoff V, Desprez T, Rolland A, Stierhof YD, Schumacher K, Gonneau M, Höfte H, Vernhettes S. 2009. Pausing of golgi bodies on microtubules regulates secretion of cellulose synthase complexes in *Arabidopsis*. *Plant Cell* **21**(4): 1141-1154.

Crowell EF, Gonneau M, Stierhof YD, Höfte H, Vernhettes S. 2010. Regulated trafficking of cellulose synthases. *Current Opinion in Plant Biology* **13**(6): 700-705.

Delmer DP. 1999. CELLULOSE BIOSYNTHESIS: Exciting Times for A Difficult Field of Study. *Annual Review of Plant Physiology and Plant Molecular Biology* **50**: 245-276.

Demura T, Fukuda H. 2007. Transcriptional regulation in wood formation. *Trends in Plant Science* **12**(2): 64-70.

Diotallevi F, Mulder B. 2007. The cellulose synthase complex: A polymerization driven supramolecular motor. *Biophysical Journal* **92**(8): 2666-2673.

Doblin MS, Kurek I, Jacob-Wilk D, Delmer DP. 2002. Cellulose biosynthesis in plants: from genes to rosettes. *Plant and Cell Physiology* **43**(12): 1407-1420.

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 of the United States of America* **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.

Farber CR, Bennett BJ, Orozco L, Zou W, Lira A, Kostem E, Kang HM, Furlotte N, Berberyan A, Ghazalpour A, Suwanwela J, Drake TA, Eskin E, Wang QT, Teitelbaum SL, Lusis AJ. 2011. Mouse genome-wide association and systems genetics identify Asxl2 as a regulator of bone mineral density and osteoclastogenesis. *PLoS Genetics* 7(4): e1002038

Galjart N. 2005. CLIPs and CLASPs and cellular dynamics. *Nature Reviews Molecular Cell Biology* **6**(6): 487-498.

Giddings TH, Brower DL, Staehelin LA. 1980. Visualization of particle complexes in the plasma membrane of *Micrasterias denticulata* associated with the formation of cellulose fibrils in primary and secondary cell walls. *Journal of Cell Biology* **84**(2): 327-339.

Grattapaglia D, Kirst M. 2008. *Eucalyptus* applied genomics: From gene sequences to breeding tools. *New Phytologist* **179**(4): 911-929.

Grattapaglia D, Plomion C, Kirst M, Sederoff RR. 2009. Genomics of growth traits in forest trees. *Current Opinion in Plant Biology* **12**(2): 148-156.

Grattapaglia D, Resende MDV. 2011. Genomic selection in forest tree breeding. *Tree Genetics and Genomes* **7**(2): 241-255.

Groover AT. 2007. Will genomics guide a greener forest biotech? *Trends in Plant Science* **12**(6): 234-238.

Gu Y, Kaplinsky N, Bringmann M, Cobb A, Carroll A, Sampathkumar A, Baskin TI, Persson S, Somerville CR. 2010. Identification of a cellulose synthase-associated protein required for cellulose biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America* **107**(29): 12866-12871.

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.

Haigler CH, Brown Jr RM. 1986. Transport of rosettes from the golgi apparatus to the plasma membrane in isolated mesophyll cells of *Zinnia elegans* during differentiation to tracheary elements in suspension culture. *Protoplasma* **134**(2-3): 111-120.

Haigler CH, Ivanova-Datcheva M, Hogan PS, Salnikov VV, Hwang S, Martin K, Delmer DP. 2001. Carbon partitioning to cellulose synthesis. *Plant Molecular Biology* **47**(1-2): 29-51.

Harfouche A, Meilan R, Altmane A. 2011. Tree genetic engineering and applications to sustainable forestry and biomass production. *Trends in Biotechnology* **29**(1): 9-17.

Hématy K, Sado PE, Van Tuinen A, Rochange S, Desnos T, Balzergue S, Pelletier S, Renou JP, Höfte H. 2007. A receptor-like kinase mediates the response of Arabidopsis cells to the inhibition of cellulose synthesis. *Current Biology* **17**(11): 922-931.

Hertzberg M, Aspeborg H, Schrader J, Andersson A, Erlandsson R, Blomqvist K, Bhalerao R, Uhlen 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 United States of America* **98**(25): 14732-14737.

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 and Developmental Biology - Plant* **45**(6): 619-629.

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

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

Joshi CP, Mansfield SD. 2007. The cellulose paradox - simple molecule, complex biosynthesis. *Current Opinion in Plant Biology* **10**(3): 220-226.

Joshi CP, Thammannagowda S, Fujino T, Gou JQ, Avci U, Haigler CH, McDonnell LM, Mansfield SD, Mengesha B, Carpita NC, Harris D, Debolt S, Peter GF. 2011. Perturbation of wood cellulose synthesis causes pleiotropic effects in transgenic aspen. *Molecular Plant* **4**(2): 331-345.

Jumbo-Lucioni P, Ayroles JF, Chambers MM, Jordan KW, Leips J, Mackay TFC, De Luca M. 2010. Systems genetics analysis of body weight and energy metabolism traits in *Drosophila melanogaster*. *BMC Genomics* **11**(1).

Kadarmideen HN, Janss LL. 2007. Population and systems genetics analyses of cortisol in pigs divergently selected for stress. *Physiological Genomics* **29**(1): 57-65.

Kadarmideen HN, von Rohr P, Janss LL. 2006. From genetical genomics to systems genetics: potential applications in quantitative genomics and animal breeding. *Mammalian Genome* **17**(6): 548-564.

Kadarmideen HN, Von Rohr P, Janss LLG. 2006. From genetical genomics to systems genetics: Potential applications in quantitative genomics and animal breeding. *Mammalian Genome* **17**(6): 548-564.

Keurentjes JJB, Fu J, Terpstra IR, Garcia JM, Van Den Ackerveken G, Snoek LB, Peeters AJM, Vreugdenhil D, Koornneef M, Jansen RC. 2007. Regulatory network construction in *Arabidopsis* by using genome-wide gene expression quantitative trait loci. *Proceedings of the National Academy of Sciences of the United States of America* **104**(5): 1708-1713.

Kirst M, Basten CJ, Myburg AA, Zeng ZB, Sederoff RR. 2005. Genetic architecture of transcript-level variation in differentiating xylem of a *Eucalyptus* hybrid. *Genetics* **169**(4): 2295-2303.

Kleczkowski LA, Kunz S, Wilczynska M. 2010. Mechanisms of UDP-glucose synthesis in plants. *Critical Reviews in Plant Sciences* **29**(4): 191-203.

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**(16): 1855-1860.

Lane DR, Wiedemeier A, Peng L, Höfte H, Vernhettes S, Desprez T, Hocart CH, Birch RJ, Baskin TI, Burn JE, Arioli T, Betzner AS, Williamson RE. 2001. Temperature-sensitive alleles of rsw2 link the KORRIGAN endo-1,4-β-glucanase to cellulose synthesis and cytokinesis in *Arabidopsis*. *Plant Physiology* **126**(1): 278-288.

Lee C, Zhong R, Richardson EA, Himmelsbach DS, McPhail BT, Ye ZH. 2007. The PARVUS gene is expressed in cells undergoing secondary wall thickening and is essential for glucuronoxylan biosynthesis. *Plant and Cell Physiology* **48**(12): 1659-1672.

Lertpiriyapong K, Sung ZR. 2003. The elongation defective1 mutant of Arabidopsis is impaired in the gene encoding a serine-rich secreted protein. *Plant Molecular Biology* **53**(4): 581-595.

MacKay TFC, Stone EA, Ayroles JF. 2009. The genetics of quantitative traits: Challenges and prospects. *Nature Reviews Genetics* **10**(8): 565-577.

MacMillan CP, Mansfield SD, Stachurski ZH, Evans R, Southerton SG. 2010. Fasciclin-like arabinogalactan proteins: Specialization for stem biomechanics and cell wall architecture in *Arabidopsis* and *Eucalyptus*. *Plant Journal* **62**(4): 689-703.

Mansfield SD. 2009. Solutions for dissolution-engineering cell walls for deconstruction. *Current Opinion in Biotechnology* **20**(3): 286-294.

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

McCarthy RL, Zhong R, Fowler S, Lyskowski D, Piyasena H, Carleton K, Spicer C, Ye ZH. 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.

McNally FJ, Vale RD. 1993. Identification of katanin, an ATPase that severs and disassembles stable microtubules. *Cell* **75**(3): 419-429.

Meng M, Geisler M, Johansson H, Harholt J, Scheller HV, Mellerowicz EJ, Kleczkowski LA. 2009. UDP-glucose pyrophosphorylase is not rate limiting, but is essential in *Arabidopsis*. *Plant and Cell Physiology* **50**(5): 998-1011.

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**(1): 681.

Morozova TV, Ayroles JF, Jordan KW, Duncan LH, Carbone MA, Lyman RF, Stone EA, Govindaraju DR, Ellison RC, Mackay TFC, Anholt RRH. 2009. Alcohol sensitivity in *Drosophila*: Translational potential of systems genetics. *Genetics* **183**(2): 733-745.

Morrison EE. 2007. Action and interactions at microtubule ends. *Cellular and Molecular Life Sciences* **64**(3): 307-317.

Nadeau JH, Dudley AM. 2011. Systems genetics. *Science* **331**(6020): 1015-1016.

Nakajima K, Furutani I, Tachimoto H, Matsubara H, Hashimoto T. 2004. Spiral1 encodes a plant-specific microtubule-localized protein required for directional control of rapidly expanding *Arabidopsis* cells. *Plant Cell* **16**(5): 1178-1190.

Nakajima K, Kawamura T, Hashimoto T. 2006. Role of the SPIRAL1 gene family in anisotropic growth of *Arabidopsis thaliana*. *Plant and Cell Physiology* **47**(4): 513-522.

Neale DB, Ingvarsson PK. 2008. Population, quantitative and comparative genomics of adaptation in forest trees. *Current Opinion in Plant Biology* **11**(2): 149-155.

Neale DB, Kremer A. 2011. Forest tree genomics: Growing resources and applications. *Nature Reviews Genetics* **12**(2): 111-122.

Oikawa A, Joshi H, Rennie E. 2010. An integrative approach to the identification of *Arabidopsis* and rice genes involved in xylan and secondary wall development. *PLoS ONE* **5**: 263 - 679.

Pagant S, Bichet A, Sugimoto K, Lerouxel O, Desprez T, McCann M, Lerouge P, Vernhettes S, Höfte H. 2002. Kobito1 encodes a novel plasma membrane protein necessary for normal synthesis of cellulose during cell expansion in *Arabidopsis*. *Plant Cell* **14**(9): 2001-2013.

Paredez AR, Somerville CR, Ehrhardt DW. 2006. Visualization of cellulose synthase demonstrates functional association with microtubules. *Science* **312**(5779): 1491-1495.

Pastuglia M, Bouchez D. 2007. Molecular encounters at microtubule ends in the plant cell cortex. *Current Opinion in Plant Biology* **10**(6): 557-563.

Plaisier CL, Horvath S, Huertas-Vazquez A, Cruz-Bautista I, Herrera MF, Tusie-Luna T, Aguilar-Salinas C, Pajukanta P. 2009. A systems genetics approach implicates USF1, FADS3, and other causal candidate genes for familial combined hyperlipidemia. *PLoS Genetics* **5**(9): e1000642.

Plomion C, Vincent D, Bedon F, Joets J, Bonhomme L, Morabito D, Duplessis S, Nilsson R, Wingsle G, Larsson C, Jolivet Y, Renaut J, Pechanova O, Yuceer C 2011. Poplar Proteomics: Update and Future Challenges. *Genetics, Genomics, and Breeding of Poplar, C.P. Joshi, S. DiFazio, and C. Kole (eds.)*. New Hampshire USA: Science Publishers, 128-165.

Ralph J, Lundquist K, Brunow G, Lu F, Kim H, Schatz PF, Marita JM, Hatfield RD, Ralph SA, Christensen JH, Boerjan W. 2004. Lignins: Natural polymers from oxidative coupling of 4-hydroxyphenyl- propanoids. *Phytochemistry Reviews* **3**(1-2): 29-60.

Richard TL. 2010. Challenges in scaling up biofuels infrastructure. *Science* **329**(5993): 793-796.

Robinson A, Mansfield S, Joshi C. 2011. In: Joshi CP, Difazio S, Chittaranjan K, eds. *Genetics, genomics and breeding of poplar*. Enfield, NH, USA: CRC Press, Science Publishers, Inc. (Imprint of Edenbridge Ltd.), 166–192.

Romanoski CE, Lee S, Kim MJ, Ingram-Drake L, Plaisier CL, Yordanova R, Tilford C, Guan B, He A, Gargalovic PS, Kirchgessner TG, Berliner JA, Lusis AJ. 2010. Systems genetics analysis of gene-by-environment interactions in human cells. *American Journal of Human Genetics* **86**(3): 399-410.

Rudolph U. 1987. Occurrence of rosettes in the ER membrane of young *Funaria hygrometrica* protonemata. *Naturwissenschaften* **74**(9): 439.

Saxena IM, Brown RM. 2005. Cellulose biosynthesis: Current views and evolving concepts. *Annals of Botany* **96**(1): 9-21.

Scheller HV, Ulvskov P 2010. Hemicelluloses. *Annual Review of Plant Biology*. 263-289.

Schrader J, Moyle R, Bhalerao R, Hertzberg M, Lundeberg J, Nilsson P, Bhalerao RP. 2004. Cambial meristem dormancy in trees involves extensive remodelling of the transcriptome. *Plant Journal* **40**(2): 173-187.

Sedbrook JC, Ehrhardt DW, Fisher SE, Scheible WR, Somerville CR. 2004. The *Arabidopsis* SKU6/SPIRAL1 gene encodes a plus end-localized microtubule-interacting protein involved in directional cell expansion. *Plant Cell* **16**(6): 1506-1520.

Sederoff R, Myburg A, Kirst M 2009. Genomics, domestication, and evolution of forest trees. 303-317.

Séguin A. 2011. How could forest trees play an important role as feedstock for bioenergy production? *Current Opinion in Environmental Sustainability* **3**(1-2): 90-94.

Seifert GJ, Roberts K 2007. The biology of arabinogalactan proteins. *Annual Review of Plant Biology*. 137-161.

Sharma N, Bryant J, Wloga D, Donaldson R, Davis RC, Jerka-Dziadosz M, Gaertig J. 2007. Katanin regulates dynamics of microtubules and biogenesis of motile cilia. *Journal of Cell Biology* **178**(6): 1065-1079.

Somerville C. 2006. Cellulose synthesis in higher plants. *Annual Review of Cell and Developmental Biology* **22**: 53-78.

Somerville C, Youngs H, Taylor C, Davis SC, Long SP. 2010. Feedstocks for lignocellulosic biofuels. *Science* **329**(5993): 790-792.

Stoppin-Mellet V, Gaillard J, Vantard M. 2006. Katanin's severing activity favors bundling of cortical microtubules in plants. *Plant Journal* **46**(6): 1009-1017.

Strauss SH, Tan H, Boerjan W, Sedjo R. 2009. Strangled at birth? Forest biotech and the Convention on Biological Diversity. *Nature Biotechnology* **27**(6): 519-527.

Studer MH, DeMartini JD, Davis MF, Sykes RW, Davison B, Keller M, Tuskan GA, Wyman CE. 2011. Lignin content in natural *Populus* variants affects sugar release. *Proceedings of the National Academy of Sciences* **108**(15): 6300-6305.

Szymanski DB. 2009. Plant cells taking shape: new insights into cytoplasmic control. *Current Opinion in Plant Biology* **12**(6): 735-744.

Taylor NG. 2008. Cellulose biosynthesis and deposition in higher plants. *New Phytologist* **178**(2): 239-252.

Thumma BR, Matheson BA, Zhang D, Meeske C, Meder R, Downes GM, Southerton SG. 2009. Identification of a cis-acting regulatory polymorphism in a eucalypt COBRA-like gene affecting cellulose content. *Genetics* **183**(3): 1153-1164.

Thumma BR, Nolan MF, Evans R, Moran GF. 2005. Polymorphisms in cinnamoyl CoA reductase (CCR) are associated with variation in microfibril angle in Eucalyptus spp. *Genetics* **171**(3): 1257-1265.

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam M, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Dŕjardin A, DePamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehlting J, Ellis B, Gendler K, Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrissat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjärvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leplé JC, Locascio P, Lou Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson P, Rinaldi C, Ritland K, Rouzé P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai CJ, Uberbacher E, Unneberg P, Vahala J, Wall K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Van De Peer Y, Rokhsar D. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**(5793): 1596-1604.

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

Wegrzyn JL, Eckert AJ, Choi M, Lee JM, Stanton BJ, Sykes R, Davis MF, Tsai CJ, Neale DB. 2010. Association genetics of traits controlling lignin and cellulose biosynthesis in black cottonwood (*Populus trichocarpa*, Salicaceae) secondary xylem. *New Phytologist* **188**(2): 515-532.

Wightman R, Turner SR. 2008. The roles of the cytoskeleton during cellulose deposition at the secondary cell wall. *Plant Journal* **54**(5): 794-805.

Wu AM, Hörnblad E, Voxeur A, Gerber L, Rihouey C, Lerouge P, Marchant A. 2010. Analysis of the *Arabidopsis* IRX9/IRX9-L and IRX14/IRX14-L pairs of glycosyltransferase genes reveals

critical contributions to biosynthesis of the hemicellulose glucuronoxylan. *Plant Physiology* **153**(2): 542-554.

Wu AM, 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. *Plant Journal* **57**(4): 718-731.

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, Goué N, 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 and VASCULAR-RELATED NAC-DOMAIN7 effectively induce transdifferentiation into xylem vessel elements under control of an induction system. *Plant Physiology* **153**(3): 906-914.

York WS, O'Neill MA. 2008. Biochemical control of xylan biosynthesis - which end is up? *Current Opinion in Plant Biology* **11**(3): 258-265.

Zhang Y, Brown G, Whetten R, Loopstra CA, Neale D, Kieliszewski MJ, Sederoff RR. 2003. An arabinogalactan protein associated with secondary cell wall formation in differentiating xylem of loblolly pine. *Plant Molecular Biology* **52**(1): 91-102.

Zhong R, Burk DH, Morrison Iii WH, Ye ZH. 2002. A kinesin-like protein is essential for oriented deposition of cellulose microfibrils and cell wall strength. *Plant Cell* **14**(12): 3101-3117.

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 ZH. 2010. Functional characterization of poplar wood-associated NAC domain transcription factors. *Plant Physiology* **152**(2): 1044-1055.

Zhong R, Lee C, Zhou J, McCarthy RL, Ye ZH. 2008. A battery of transcription factors involved in the regulation of secondary cell wall biosynthesis in *Arabidopsis*. *Plant Cell* **20**(10): 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**(3): 1452-1468.

Zhong R, Richardson EA, Ye ZH. 2007. 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, Ye ZH. 2010. The poplar PtrWNDs are transcriptional activators of secondary cell wall biosynthesis. *Plant Signaling and Behavior* **5**(4): 469-472.

1.7 Figures

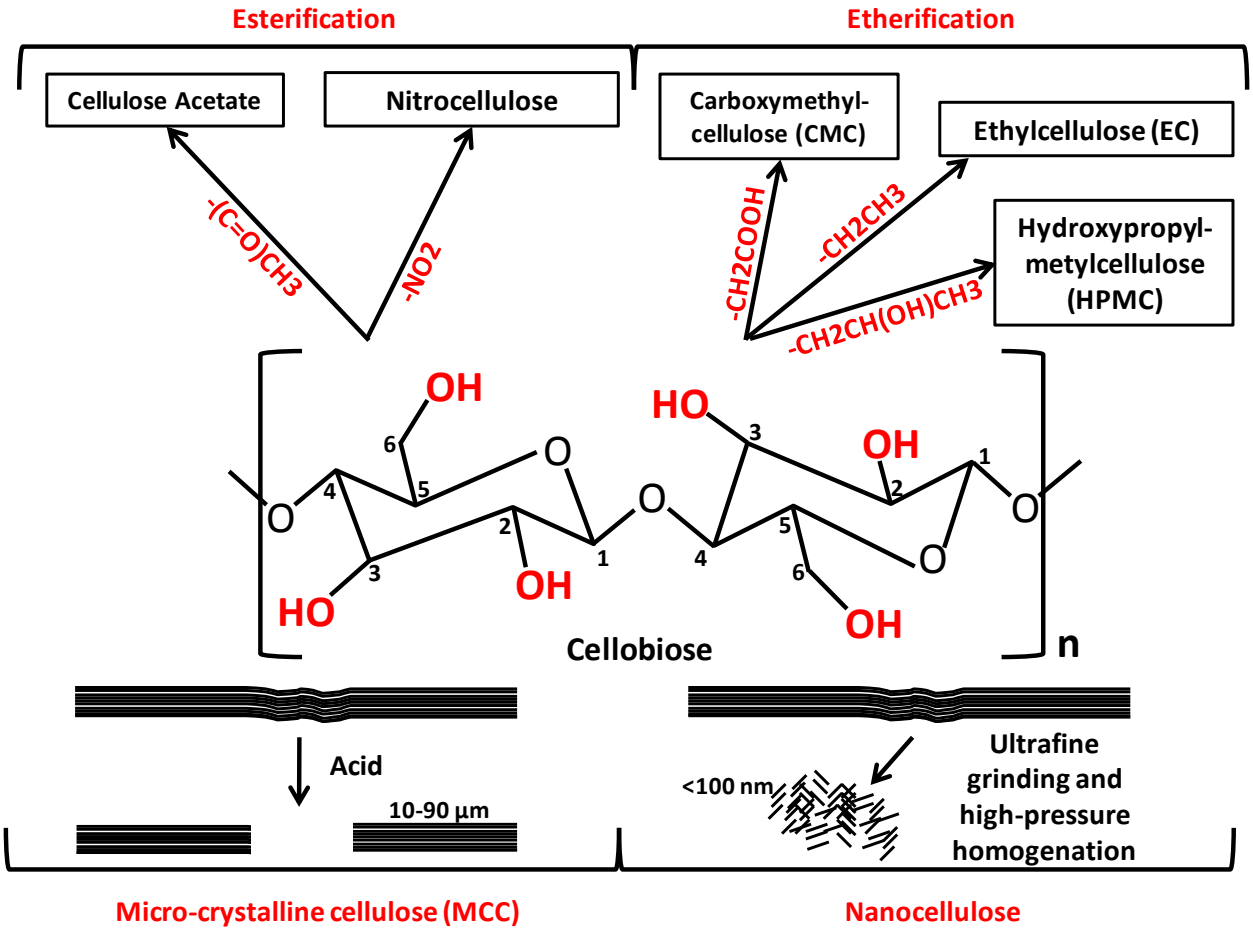


Fig. 1.1 Examples of the diversity of currently produced, high-value derivatives of wood-derived cellulose.

The structure of the repeating unit of cellulose – cellobiose – is shown in the middle, with a 'head-to-tail' arrangement of two glucose molecules bound via a β 1–4 linkage. The side-chain substitution of the hydroxyl groups from C2, C3 and/or C6 (highlighted in red) results in the production of a variety of unique physicochemical derivatives, all of which comprise diverse industrial and commercial products (top). Pure crystalline cellulose can also be broken up into micro-crystalline cellulose (bottom) by chemical disruption of the noncrystalline regions or, alternatively, the entire polymer can be separated into nanocellulose crystals.

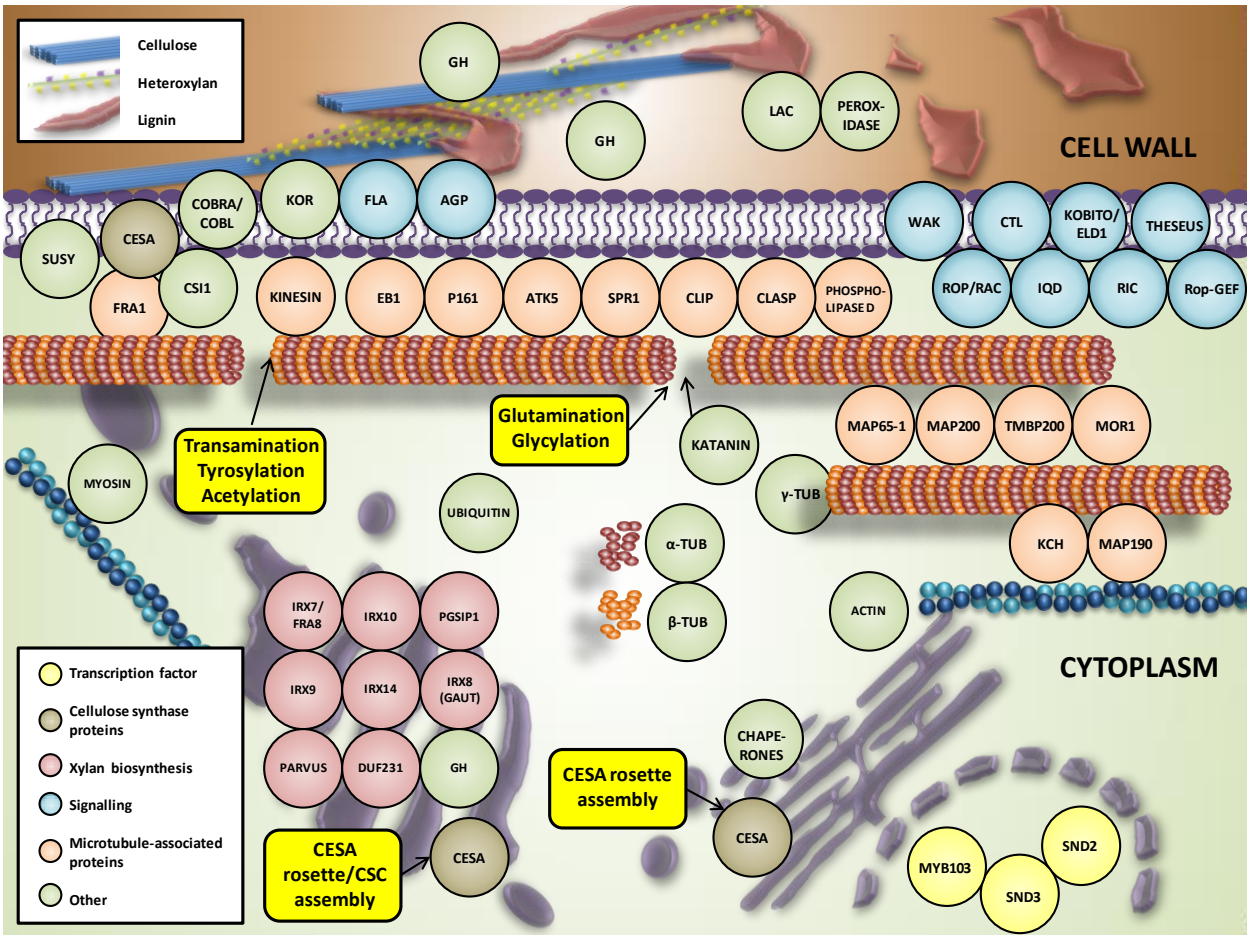


Fig. 1.2 An integrated view of currently known proteins and some cellular processes involved in cellulose and xylan biosynthesis.

Proteins are indicated as coloured circles in the cell areas with which they are associated, and classes of proteins are coloured as indicated by the legend at the bottom left of the figure. It should be noted that the proximity of proteins in the figure does not imply interaction. Actin (blue beads) and microtubules (red and orange tubes) are also shown. References for the inclusion of specific proteins and full protein names can be found in the text.

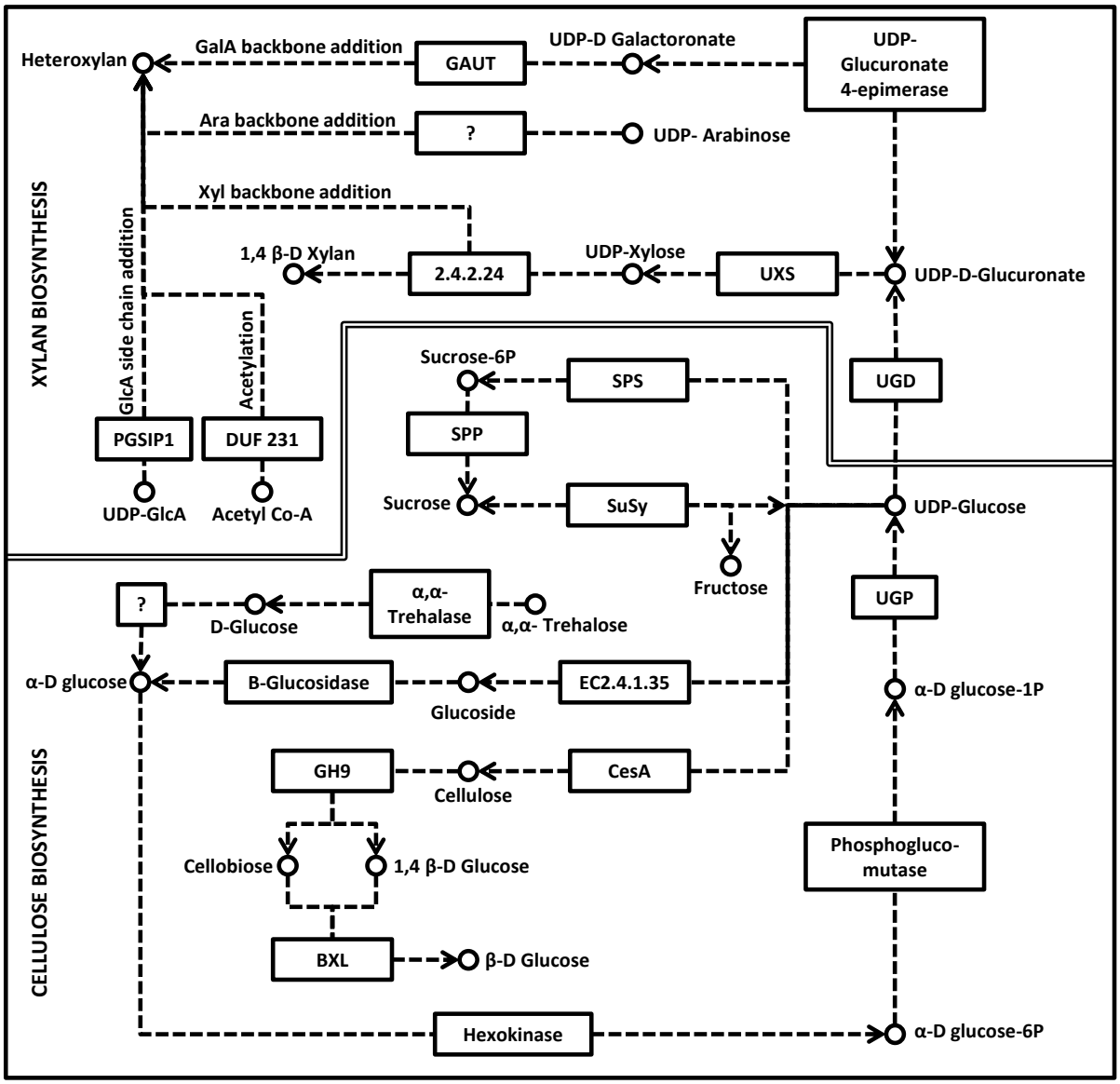


Fig. 1.3 Metabolic pathways and processes leading to cellulose and xylan biosynthesis.

Pathways are based on the Kyoto Encyclopedia of Genes and Genomes (KEGG; http://www.genome.jp/kegg/), as well as recent literature revealing putative biosynthetic enzymes involved in xylan biosynthesis (Brown et al., 2007, 2009; York & O'Neill, 2008; Oikawa et al., 2010). Metabolites are represented as circles, and enzymatic processes or known enzymes of interest as boxes. BGL, β-glucosidase; CESA, cellulose synthase; SPS, sucrose phosphate synthase; SPP, sucrose phosphate phosphatase; SUSY, sucrose synthase; UDP, uridine diphosphate; UGD, UDP-glucose 6-dehydrogenase; UGP, UTP-glucose-1-phosphate uridylyltransferase; UXS, uridine-diphosphoglucuronate decarboxylase.

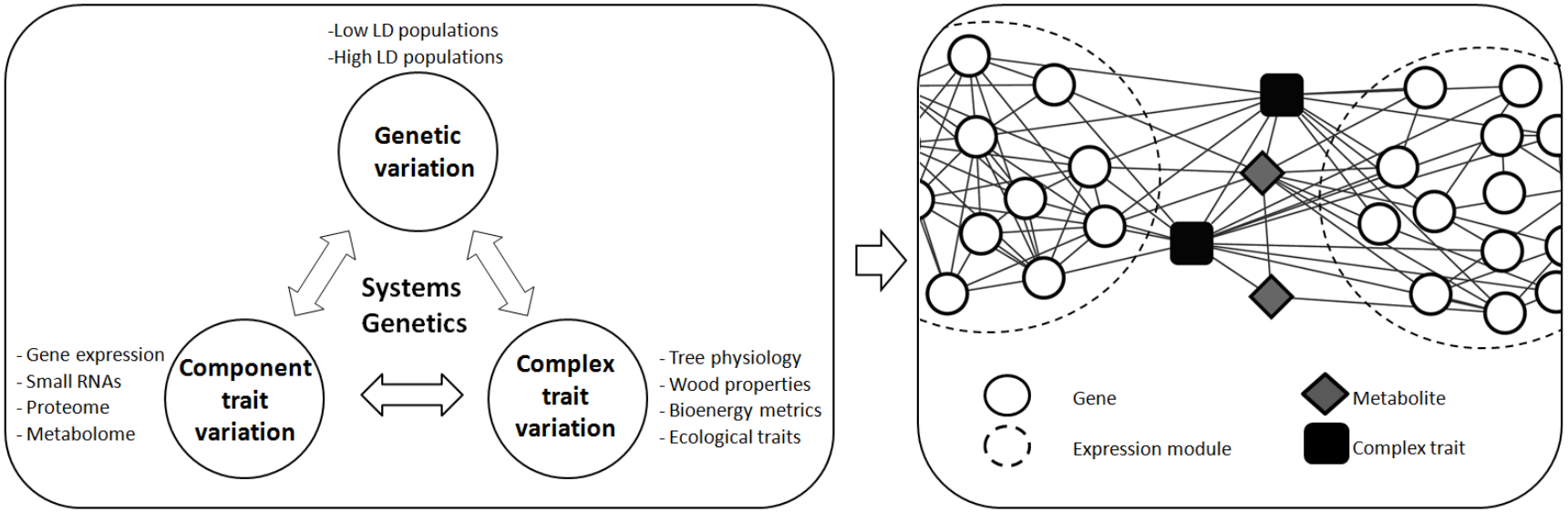


Fig. 1.4 A systems genetics approach to understanding the molecular basis of complex phenotypic traits in forest trees.

Left: systems genetics allows the molecular dissection of polygenic traits by relating phenotypic and genetic variation in experimental populations to measurable component traits (in developing cells, tissues and organs of trees) segregating in the same populations. Right: conceptual network resulting from the integration of the covariation of complex and component traits, revealing novel correlations among genes, expression modules, metabolites and complex wood phenotypes that would not be observed at the level of the individual.

CHAPTER 2

De novo assembled expressed gene catalogue of a fast-growing *Eucalyptus* tree produced by Illumina mRNA-Seq

Eshchar Mizrachi[1*], Charles A Hefer[2*], Martin Ranik[1], Fourie Joubert[2] and Alexander A Myburg[1]

[1] Department of Genetics, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Private bag X20, Pretoria, 0028, South Africa; [2] Bioinformatics and Computational Biology Unit, Department of Biochemistry, University of Pretoria, Private bag X20, Pretoria, 0028, South Africa. *These authors contributed equally to this work.

This chapter has been prepared in the format of a manuscript for a peer-reviewed research journal. A large part of this chapter was included in a manuscript which was published in BMC Genomics (Mizrachi *et al.*, *BMC Genomics* **11**(1): 618). This chapter contains some additional expression and pathway analyses I performed subsequent to publication of the paper. I drafted the manuscript, helped sample the material, helped prepare the libraries, participated in the *de novo* assembly and data analysis, and helped design Eucspresso. Charles Hefer performed the *de novo* assembly and automated annotation, participated in data analysis, designed the database Eucspresso, and helped draft the manuscript. Martin Ranik prepared the libraries, helped sample the material and participated in data analysis. Fourie Joubert participated in data analysis and the design of Eucspresso. Alexander Myburg conceived of and supervised the study, participated in its design and coordination, helped to draft the manuscript and participated in data analysis and design of Eucspresso.

2.1 Summary

- *De novo* assembly of transcript sequences produced by short-read DNA sequencing technologies offers a rapid approach to obtain expressed gene catalogues for non-model organisms. A draft genome sequence was produced in 2010 for a *Eucalyptus* tree species (*E. grandis*) representing the most important hardwood fibre crop in the world. Genome annotation of this valuable woody plant and genetic dissection of its superior growth and productivity will be greatly facilitated by the availability of a comprehensive collection of expressed gene sequences from multiple tissues and organs.

- We present an extensive expressed gene catalogue for a commercially grown *E. grandis x E. urophylla* hybrid clone constructed using only Illumina mRNA-Seq technology and *de novo* assembly. A total of 18,894 transcript-derived contigs, a large proportion of which represent full-length protein coding genes were assembled and annotated. Analysis of assembly quality, length and diversity show that this dataset represents the most comprehensive expressed gene catalogue for any *Eucalyptus* tree. mRNA-Seq analysis furthermore allowed digital expression profiling of all of the assembled transcripts across diverse xylogenic and non-xylogenic tissues, which is invaluable for ascribing putative gene functions and understanding the biology of wood formation.

- *De novo* assembly of Illumina mRNA-Seq reads is an efficient approach for transcriptome sequencing and profiling in *Eucalyptus* and other non-model organisms. The transcriptome resource (Eucspresso, http://eucspresso.bi.up.ac.za/) generated by this study will be of value for genomic analysis of woody biomass production in *Eucalyptus* and for comparative genomic analysis of growth and development in woody and herbaceous plants.

2.2 Introduction

Ultra-high-throughput second-generation DNA sequencing technologies from companies such as Roche (454 pyrosequencing), Illumina (sequencing by synthesis, Genome Analyzer) and Life Technologies (sequencing by ligation, SOLiD), are increasingly being used for novel exploratory genomics in small to medium-sized laboratories. "Short-read" (36 – 72 nt) technologies such as those of Illumina and Life Technologies have proven to be exceptionally successful in a wide variety of whole-transcriptome investigations (Cloonan *et al.*, 2008; Mortazavi *et al.*, 2008; Nagalakshmi *et al.*, 2008; Tang *et al.*, 2009; Wilhelm & Landry, 2009), but most of these studies have relied on prior sequence knowledge such as an annotated genome for qualitative and quantitative transcriptome analyses.

Genome assembly of short sequences without any auxiliary knowledge has primarily utilized 454 sequencing data, due to the longer individual read lengths of 150-400 base pairs (bp). However, short-read sequencing (Illumina GA and SOLiD) has been successfully used for *de novo* assembly of small bacterial genomes (2-5 Mbp), where 36 bp reads have been assembled (Hernandez *et al.*, 2008; Farrer *et al.*, 2009; Kozarewa *et al.*, 2009) and hybrid approaches, where genomes are *de novo* assembled using a combination of reads from multiple sequencing platforms to overcome the inherent limitations of each technology, have been used to successfully assemble genomes of up to 40 Mbp (DiGuistini *et al.*, 2009; Nowrousian, 2010). More recently, the sequencing of the giant panda genome was demonstrated (Li *et al.*, 2010) using *de novo* assembly of sequence derived from a single platform (Illumina), but utilizing a combination of different insert sizes allowing assembly of an estimated 94% of the genome (2.25 Gbp). *De novo* assembly of large, highly repetitive and highly heterozygous eukaryotic genomes from short-read data remains a challenge.

In transcriptome studies, 454 pyrosequencing has proven very useful for generating ESTs representing the majority of expressed genes. This has enabled gene discovery in a variety of previously uncharacterized eukaryotic organisms with little or no *a priori* DNA sequence information (Novaes *et al.*, 2008; Vera *et al.*, 2008; Dassanayake *et al.*, 2009; Hahn *et al.*, 2009; Meyer *et al.*, 2009). However, relatively few published studies have attempted *de novo* assembly of whole-transcriptome sequences from short-read data such as that generated by Illumina GA or SOLiD technologies. Assembly of short (36 – 72 bp) read data into accurate, contiguous transcript sequences has only recently been reported (Birol *et al.*, 2009; Gibbons *et al.*, 2009; Wu, T *et al.*, 2010) demonstrating that assembly of long, potentially full-length, transcript assemblies is indeed possible.

Eucalyptus tree species and hybrids presently constitute the most widely planted (≈20 Mha) and commercially important hardwood fibre crop in the world. They are mainly utilized for timber, pulp and paper production (Eldridge *et al.*, 1993). Their fast growth rates and wide adaptability may in future allow sustainable and cost efficient production of woody biomass for bioenergy generation (FAO, 2008; Hinchee *et al.*, 2009). *Eucalyptus* is only the second forest plantation genus (after *Populus*) for which a reference genome sequence was completed by end 2010 (Myburg, 2008). To support the genome annotation effort, there is much value in having a dataset of genes with strong transcriptional evidence across a range of tissues and developmental stages. Until recently, limited amounts of *Eucalyptus* EST/unigene data were available in public databases, mainly due to the fact that commercial interests have necessitated private EST collections (Hibino, 2009). As of March 2010, aside from a mixed-species collection of ≈56,000 nucleotide sequences on NCBI (≈37,000 of which are Sanger EST sequences) that contain extensive redundancy, the largest effort to date to generate a comprehensive catalogue of expressed genes in a single *Eucalyptus* species was based on 454 sequencing of cDNA fragments from *E. grandis* trees (Novaes *et al.*, 2008). While this study provided an excellent representation of expressed genes and gene ontology classes in *E. grandis*, the relatively short lengths of the assembled contigs (mean

length of 389 bp for all contigs longer than 200 bp) meant that very few complete gene models were represented. There remains therefore a fundamental need for a high-quality expressed gene catalog for *Eucalyptus*, to support genome annotation efforts and discern authentically expressed genes from predicted gene models, as well as for future genomics research, which will include transcriptome, proteome and metabolome profiling.

In this study we addressed three main questions: First, is it feasible to *de novo* assemble Illumina mRNA-Seq data into contiguous, near full-length gene model sequences for *Eucalyptus*? Second, what genes make up the expressed gene catalog for a fast-growing *Eucalyptus* plantation tree? Finally, can we re-use the mRNA-Seq data to create a tissue and organ-specific digital expression profile for each assembled contig? We addressed these questions by generating a comprehensive set of expressed gene sequences from a commercially grown *Eucalyptus* hybrid (*E. grandis* x *E. urophylla*) clone using Illumina mRNA-Seq technology and *de novo* short-read assembly. We report herein the complete annotation of the expressed gene catalog based on comparative analysis with the published *Arabidopsis thaliana* (Kaul *et al.*, 2000), *Populus trichocarpa* (Tuskan *et al.*, 2006) and *Vitis vinifera* (Jaillon *et al.*, 2007) protein-coding datasets. Additionally, we explore the dynamics of source-sink relationships in trees (leaves vs xylem) and identify pathways that show preferential investment in these tissues and organs. Genes likely to be involved in cellulose biosynthesis in *Eucalyptus* are also described. The assembly (transcripts, coding sequences and predicted protein sequences) has been made available on http://eucspresso.bi.up.ac.za/.

2.3 Materials and Methods

2.3.1 Plant tissue collection

Tissues from a six-year-old ramet of a commercially grown *E. grandis* x *E. urophylla* hybrid clone (GUSAP1, Sappi Forestry, Kwambonambi, South Africa) were collected in a clonal field trial and immediately frozen in liquid nitrogen, as previously described by Ranik and Myburg (2006). The following tissues were sampled from approximately breast height (1.35 m) on the main stem following bark removal: immature xylem (outer glutinous 1-2 mm layer comprising early developing xylem tissue) and xylem (after removal of the immature xylem layer, 2-mm-deep planing including xylem cells in advanced stages of maturity). Early developing phloem tissue including small amounts of cambial cells was collected by scraping the first 1-2 mm layer from the inner surface of the bark. Additionally, we sampled shoot tips (soft green termini of young crown tip branches containing shoot primordia and apical meristems), young leaves (rapidly-growing leaves in the process of unfolding) and mature leaves (older, fully expanded leaves of the current growth season).

2.3.2 Paired-end mRNA-Seq library preparation and sequence generation

Total RNA was extracted from the six tissues using the protocol described previously (Chang *et al.*, 1993). Total RNA quality and concentration were determined using the Agilent RNA 6000 Pico kit (Agilent, Santa Clara, CA) on a 2100 Bioanalyzer (Agilent). Enrichment of polyA+ RNA was performed using the Oligotex midi kit (Qiagen, Valencia, CA). Two hundred nanograms of polyA+ RNA were fragmented in 1X RNA fragmentation solution (Ambion, Austin, TX) at 70°C for 5 minutes. The fragmented RNA was precipitated with three volumes of ethanol and re-dissolved in water. Double-stranded cDNA was synthesized using the cDNA Synthesis System (Roche, Indianapolis, IN) according to the manufacturer's instructions using random hexamers (Invitrogen, Carlsbad, CA) to prime the first strand cDNA synthesis. Paired-end libraries with approximate average insert lengths of 200 base pairs were synthesized using the Genomic Sample Prep kit (Illumina, San Diego, CA) according to the manufacturer's instructions. Prior to cluster generation, library concentration and size were assayed using

the Agilent DNA1000 kit (Agilent) on a 2100 Bioanalyzer (Agilent). Libraries were sequenced on a Genome Analyzer equipped with a paired-end module (versions I, II and IIx, Illumina).

2.3.3 *De novo* assembly of mRNA-Seq data

After removing sequences containing low quality bases ('N's) or single base repeats and ribosomal RNA sequences, the 3.93 Gbp dataset was used for assembly and subsequent coverage per base (CPB) estimation for each assembled contig. We assembled the filtered Illumina paired-end (PE) reads using Velvet version 0.7.30 (Zerbino & Birney, 2008). Previous studies (Cloonan *et al.*, 2008; Lister *et al.*, 2008; Mortazavi *et al.*, 2008; Nagalakshmi *et al.*, 2008) have demonstrated that mRNA-Seq technology produces uneven coverage over a transcript, which prompted us to follow a coverage-assisted reference assembly strategy. Using Mosaik (http://bioinformatics.bc.edu/marthlab/Mosaik) to align the filtered Illumina PE sequences to the assembled contigs, the average coverage per contig was calculated. A custom script was then developed to extract the pairs of sequences that mapped to each contig, and using that contig as a template, each contig was re-assembled using Velvet with the associated expected coverage parameter set to the Mosaik average coverage value for that contig.

2.3.4 Contig validation

The degree to which the assembled contigs represented long, contiguous RNA transcript sequences, was evaluated by aligning 35 Velvet contigs and their respective predicted CDSs to full-length, cloned, Sanger-derived *Eucalyptus* reference sequences present in NCBI. CPB was calculated for the sequences using BWA (Li & Durbin, 2009) and a global pairwise alignment of the sequences was performed using the Needle package from EMBOSS (Rice *et al.*, 2000). Plots were constructed from the alignments with the CPB on the y-axis of the plot. Zero coverage values were assigned to gaps in the alignments. This

revealed where gaps and/or potentially misassembled regions were present in the assembled contigs, and to what depth these contigs were sequenced.

2.3.5 Coding sequence prediction

Coding sequence predictions were performed using GENSCAN (Burge & Karlin, 1997) and AUGUSTUS (Stanke & Waack, 2003), predicting 15,713 and 15,904 proteins respectively. The difference in coding sequences predicted could be attributed to the different training data sets used and inherent difficulty of predicting coding sequences from incomplete genomic sequences. The GENSCAN results (15,713 predicted proteins) were used in downstream analyses.

2.3.6 Annotation of assembled contigs

Homology searches were performed against public sequence databases. The newest versions as of February 2010 of the protein sequences of *Arabidopsis* (TAIR 9), *Vitis* (Sept 2009 build) and *Populus* (version 2.0, Phytozome) were used to construct the individual BLAST datasets. The *Eucalyptus* public dataset (EucAll) consisted of 45,442 entries in Genbank (downloaded March 2010), 13,930 entries from the *Eucalyptus* Wood unigenes and ESTs (Rengel *et al.*, 2009), *E. grandis* leaf tissue ESTs (120,661 entries from DOE-JGI-produced 454 sequences, http://eucalyptusdb.bi.up.ac.za/) and 190,106 Unigenes and singlets from *E. grandis* 454 data (Novaes *et al.*, 2008). The BLAST e-value threshold was set at $1e^{-10}$, with a minimum alignment length of 100 nucleotides (33 amino acids). Functional annotation (GO and KEGG) was performed using BLAST2GO (Conesa *et al.*, 2005), using the default annotation parameters (BLAST e-value threshold of $1e^{-06}$, Gene Ontology annotation threshold of 55). InterPro annotations were performed using InterProScan (http://www.ebi.ac.uk/Tools/InterProScan/).

2.3.7 Coverage and FPKM determination

Sequence depth and base coverage were calculated using BWA (Li & Durbin, 2009) and the FPKM values estimated by aligning the Illumina reads to the assembled transcriptome using Bowtie (Langmead *et al.*, 2009) and estimating the expression level of each predicted transcript (FPKM value) using Cufflinks (http://cufflinks.cbcb.umd.edu) (Trapnell *et al.*, 2010).

2.4. Results

2.4.1 *De novo* assembly, validation and annotation of contigs

In total, 62 million paired-end reads of raw mRNA-Seq data (6.90 Gbp) representing poly(A)-selected RNA from six *Eucalyptus* tissues and varying in lengths from 36 bp to 60 bp, were generated in 14 lanes on Illumina GA and GAII instruments. Following a sequence filtering process to exclude low quality and ribosomal RNA-derived reads, we assembled 36 million paired-end reads (3.93 Gbp, Table S2.1, Fig. S2.1, NCBI Sequence Read Archive accession SRA012408) of non-normalized mRNA sequence, using the Velvet short-read assembler (version 0.7.30, (Zerbino & Birney, 2008)). In total, 18,894 RNA-derived contigs were assembled (comprising 22.1 Mbp of transcriptome sequence) that were greater than 200 bp in length (mean = 1170 bp, Fig. 2.1 and Additional file 2.1), with a median coverage per base (CPB) per contig of 37X, ranging from 8X (minimum coverage cut-off for assembly) to 5,262X (Fig. S2.2).

We performed *ab inito* CDS prediction using GENSCAN (Burge & Karlin, 1997) and found that 15,713 contigs (83.2%) contained a predicted CDS (Table S2.2). Analysis of the predicted coding sequences using Anaconda (Pinheiro *et al.*, 2006) identified 6,208 contigs that contained putatively full-length CDSs (i.e. containing start and stop codons), 4,610 predicted to contain a start but no stop codon, 4,874 predicted to contain a stop but no start codon, and only 21 with neither. To ascertain the quality of Velvet

assembly of short reads into long contiguous coding sequences, we compared a subset of 35 of our transcript-derived contigs to corresponding Sanger-sequenced, full-length, cloned *Eucalyptus grandis* mRNA sequences in NCBI (Fig. 2.2 and Additional file 2.2). Paired reads were independently mapped to each Sanger reference sequence, the *de novo* assembled Velvet contig and its corresponding predicted CDS. A Needleman-Wunsch alignment of these three sequences was used for contiguity validation of the assembled contigs. Independently, each sequence had 100% coverage validation across the contig, except in cases of low quality assembly ('N's inserted by Velvet), which occurred in regions of coverage lower than 8X per base. Of the 35 transcript-derived contigs evaluated, 25 (71%) assembled completely with a 5' UTR, 3' UTR, as well as a contiguous coding sequence matching that of the reference mRNA sequence. We found several cases where, despite high coverage, our transcript-derived contigs differed from the Sanger reference sequence due to indels, but these were generally in the UTR regions and likely represent allelic differences between the F1 hybrid individual and the reference sequences (Additional file 2.2).

Of the 18,894 assembled contigs, 18,606 (98.48%) exhibited significant similarity (BLASTN, $<1e^{-10}$, (Altschul *et al.*, 1990)) to the preliminary draft 8X DOE-JGI *E. grandis* genome assembly (http://eucalyptusdb.bi.up.ac.za/) consistent with the origin of the mRNA contigs (an F1 hybrid of *E. grandis* and *E. urophylla*). We further characterized the assembled contigs by high stringency BLASTX analysis ($<1e^{-10}$ confidence, minimum 100 bp high scoring pair (HSP) match length) to protein datasets from three reference sequenced angiosperm genera (*Arabidopsis*, *Populus* and *Vitis*). Cumulatively, 15,055 contigs (79.68%) exhibited high similarity to *Arabidopsis* (14,235 contigs), *Populus* (14,769 contigs) or *Vitis* proteins (14,833 contigs, Fig. S2.3). Of the 15,055 contigs with high similarity to *Arabidopsis*, *Populus* or *Vitis* proteins, 13,806 (91.70%) also contained predicted coding sequences (Fig. 2.3A), while 1,249 (8.30%) did not (Fig. 2.3B), possibly due to low expression of these transcripts which would have resulted in lower coverage and shorter contigs that represented only a fraction of the open

reading frame (or mostly UTR sequence). Predicted codon usage and amino acid frequencies in the proteome represented by the *Eucalyptus* expressed gene catalog were very similar to those of expressed gene catalogs from *Arabidopsis* and *Populus* (Fig. S2.4, Fig. S2.5).

To compare the completeness of our expressed gene catalogue to that of all publicly available gene sequence data for *Eucalyptus*, we generated a separate dataset, termed EucALL, containing all publicly available *Eucalyptus* gene sequence data to date (March 2010). This included all NCBI unigenes and ESTs, assembled 454 EST data from *E. grandis* leaf tissue (DOE-JGI, http://eucalyptusdb.bi.up.ac.za/), assembled 454 EST data produced by Novaes and colleagues (Novaes *et al.*, 2008), and the EucaWood contig dataset (Rengel *et al.*, 2009). We compared the representation of *Arabidopsis* genes in the EucALL dataset and in our assembled *E. grandis x E. urophylla* (EGU) transcript dataset by BLASTX at significance levels of $<1e^{-05}$, $<1e^{-10}$ and $<1e^{-20}$ (Table S2.3). While the overall numbers of hits were higher in the EucALL dataset, these were mostly in the lower size ranges. For our *de novo* assembled contigs, a much higher number of significant hits in contigs larger than 2000 bp in size (6,602 compared to 1,940 at significance $<1e^{-10}$) indicating that a greater proportion of our contigs represent full-length gene models than the publicly available *Eucalyptus* gene sequence set (EucALL).

2.4.2 Functional annotation of the expressed gene catalog

The transcript-derived contig sequences were annotated according to several functional annotation conventions, including Gene Ontology (GO – http://www.geneontology.org/), KEGG (http://www.genome.jp/kegg/) and InterProScan (http://www.ebi.ac.uk/Tools/InterProScan/). The numbers and assortment of allocated GO categories provides a good indication of the large diversity of expressed genes sampled from the *Eucalyptus* transcriptome (Fig. 2.4). This was also reflected in the diversity of InterProScan categories identified (Fig. S2.6, Fig. S2.7), as well as the comprehensive

coverage of biochemical processes by KEGG annotation, which was similar to that of the entire *Arabidopsis* gene catalog (Fig. S2.8). Together these results provided evidence that the gene catalog was of high quality, sufficient for a transciptome-wide analysis and suitable for follow-up analysis of quantitative gene-expression using RNA-seq data.

2.4.3 Digital expression profiling

An accepted method of identifying large scale differences in gene expression is to use EST abundance as an indicator of transcript abundance. This method has been implemented and validated in numerous studies using Sanger-derived ESTs (Geisler-Lee *et al.*, 2006; Pavy *et al.*, 2008), as well as 454-pyrosequencing methods (Weber *et al.*, 2007; Hahn *et al.*, 2009; Hale *et al.*, 2009; Kristiansson *et al.*, 2009; Schwarz *et al.*, 2009). Quantitative transcriptome analysis using ultra-high-throughput sequencing technologies such as Illumina and SOLiD has been shown to be accurate and highly correlated with other quantitative methods such as RT-qPCR and microarray analysis (Cloonan *et al.*, 2008; Wilhelm & Landry, 2009). To quantify tissue-specific transcript abundance reflected in our short-read dataset, we combined data (multiple lanes in most cases) generated from the same tissues and mapped six tissue-specific datasets (Table S2.1) to the assembled gene catalog using Bowtie (Langmead *et al.*, 2009). Following this, we used the Cufflinks (Trapnell *et al.*, 2010) program (http://cufflinks.cbcb.umd.edu), which provides relative abundance values by calculating Fragments Per Kilobase of exon per Million fragments mapped (FPKM) as validated previously (Mortazavi *et al.*, 2008). This enabled the allocation of a tentative digital expression profile for each transcript-derived contig (Additional file 2.3).

To compare between two contrasting tissue types that are of interest for woody biomass production, we evaluated groups of genes whose FPKM values were greater than two-fold higher in woody (xylogenic, sink) tissues (average FPKM of immature xylem and xylem: 1,897 annotated contigs) or leaf (source)

tissues (average FPKM of shoot tips, young leaves and mature leaves: 1,531 annotated contigs). GO categories over-represented in the xylem-upregulated set compared to the leaf set (Fig. 2.5A) were representative of developing woody tissues, with significant enrichment ($p < 0.05$) in signaling ("kinase activity"), carbohydrate metabolism, and genes associated with the Golgi, cytoskeleton and the plasma membrane – consistent with an emphasis on delivery of biopolymers to the cell wall. In contrast, gene categories significantly enriched ($p < 0.05$) in leaf tissue compared to woody tissue (Fig. 2.5B) were associated with photosynthesis ("plastid", "thylakoid", "photosynthesis"), growth and energy production (precursor metabolites, "lipid biosynthesis", "amino acid metabolism").

We also interrogated our transcriptome data using the "core xylem gene set" identified in *Arabidopsis* by Ko and colleagues (2006). Of the 52 genes identified by the authors as markers of secondary xylem formation in *Arabidopsis*, 33 had putative homologues in the *Eucalyptus* transcriptome (BLASTX, $<1e^{-10}$) and in total 43 contigs were identified. Of these, 40 (93%) showed greater than two-fold "Xylem" to "Leaf" digital expression profile ratios and six were only detected in xylem tissues (Table S2.4). Most of the expression profiles were also highly correlated with that of secondary cell wall-specific *Eucalyptus* cellulose synthase genes, similar to the patterns previously observed in *Arabidopsis.* These results are comparable to the 80% (51 out of 63 genes) reported recently for the same set of *Arabidopsis* homologs in *Populus* (Dharmawardhana *et al.*, 2010), which provided further support for the biological validity of the short-read-based digital expression profiles associated with the *Eucalyptus* expressed gene catalog.

2.4.4 Processes and pathways differentially transcriptionally regulated in *Eucalyptus* xylem

We performed a metabolic pathway analysis to gain insight into the relative expression investment in source and sink tissues/organs during wood formation. Analysis of the *Arabidopsis* homologs of the

xylem or leaf differentially expressed contigs using MapMan (Thimm *et al.*, 2004) and KEGG (Kanehisa & Goto, 2000) showed that a wide range of core metabolic pathways was represented (Fig. 2.6, Fig. S2.9), with several categories (MapMan bins 4, 9, 10, 30, 31) showing a larger proportion of preferentially expressed genes in xylem (Fig. S2.10, Additional file 2.5). As expected, genes preferentially expressed in source tissues and organs (leaf) but not sink tissues (xylem) were related to photosynthesis (light reactions, Calvin cycle and photorespiration) as well as secondary metabolism of waxes, terpenes and flavonoids (Fig. 2.6). In contrast, many genes preferentially expressed in xylem were involved in cell wall metabolism, including cell wall precursor synthesis and cell wall proteins, as well as polysaccharide and phenylpropaoid/phenolics metabolism (Fig. 2.6, Additional file 2.5).

Perhaps more unexpected was the xylem-specific transcriptional investment in other aspects of plant metabolism such as sphingolipid and steroid metabolism (in contrast with fatty acid metabolism which was mainly leaf-specific), while metabolism of phospholipids was represented by both xylem and leaf-specific members (Fig. 2.6). Sphingolipids and sterols are associated with trafficking to the apoplast and cell surface (Borner *et al.*, 2005), biological processes that were also represented strongly in xylem under the "Cell" category (Mapman bin 31 – Fig. S2.10, Additional file 2.5). Other pathways predominantly represented in xylem were glycolysis, the pentose phosphate pathway and mitochondrial electron transport (ATP synthesis, Fig. 2.6, Fig. S2.11-S2.13, Table S2.5). Representative genes from all steps involved in mitochondrial electron transport and ATP synthesis were preferentially expressed in xylem compared to leaf (Fig. S2.13), with most genes being expressed within the top 25% and top 10% of expression in xylem (Fig. S2.13, refer to Table S2.6 for summary statistics of FPKM values). These findings indicate that there is additional investment in energy derivation from sugars that is proportionally higher in sink compared to source organs and tissues.

2.4.5 Identification of *Eucalyptus* homologs of genes related to cellulose biosynthesis

An objective of this investigation was to identify genes in *Eucalyptus* that are related to cellulose biosynthesis. Although the main cellulose synthase (*CesA*) genes have been previously described in *Eucalyptus* (Ranik & Myburg, 2006), there are many known genes involved in this biological process that are essential for cellulose biosynthesis. In *Arabidopsis* studies, several expression meta-data analyses have revealed genes that are commonly co-expressed with the secondary-wall specific *CesA* genes. We assembled several contigs having significant BLAST hits to *Arabidopsis CesA* proteins, with three (*EgCesA1*, *2* and *3* – orthologs of *AtCesA4*, *7* and *8*) showing marked xylem-specific expression (Fig. S2.14). The availability of transcriptome-wide data from *Eucalyptus* allowed the identification of additional genes that are likely influencing cellulose biosynthesis in *Eucalyptus*. To identify these, a combinatorial analysis was performed that integrated data from previous *Arabidopsis* studies as well as the xylem/leaf ratio of expression in *Eucalyptus*.

Arabidopsis gene IDs were extracted from six published co-expression meta-analyses that specifically used a the expression of secondary cell wall-specific *CesA* gene as an initial seed to identify highly co-expressed genes (Brown *et al.*, 2005; Persson *et al.*, 2005; Ko *et al.*, 2006; Mentzen & Wurtele, 2008; Mutwil *et al.*, 2009; Mutwil *et al.*, 2010). These were summarized to produce a non-redundant gene set of 208 genes (Additional file 2.6, "Arabidopsis cellulose genes"). *Eucalyptus* homologs of these were identified from the assembled transcriptome, which are also expressed preferentially in xylem compared to leaf. This resulted in a list of 86 *Arabidopsis* genes, which match 114 *Eucalyptus* homologs (Additional file 2.6).

Several genes were consistently found in all datasets, such as homologs of *COBL4*/*IRX6* (Brown *et al.*, 2005), FLA11/IRX13 (Persson *et al.*, 2005), AT1G09610 (GLUCURONOXYLAN

METHYLTRANSFERASE, GXM – Lee *et al.*, 2012a; Urbanowicz *et al.*, 2012), and AT4G27435 (DUF1218-containing protein of unknown function). Others found in at least five of the six datasets included homologs of the xylan biosynthetic genes IRX8 (Persson *et al.*, 2005; Brown *et al.*, 2007; Peña *et al.*, 2007), IRX9 (Brown *et al.*, 2007; Wu, AM *et al.*, 2010; Lee *et al.*, 2012b) and IRX10 (Brown *et al.*, 2009; Wu *et al.*, 2009) and LAC4/IRX12 (enzymes involved in lignification, Berthet *et al.*, 2011), as well as several proteins of unknown function (IQD10, GLP10, TBL3 [DUF231] and DUF579). The majority of these genes displayed extremely high (top 5%) expression in xylem and had very high xylem/leaf ratios of expression (16 out 144 genes had no detectible expression in leaf). Several transcription factors previously characterized as regulating secondary cell wall biosynthesis were also present in the dataset, including SND2, MYB20, MYB85, MYB103, and NST1 (reviewed in Hussey *et al.*, 2013), as well as two zinc-finger proteins that potentially play a role in transcriptional regulation.

2.4.6 Public data resource

We constructed a public data resource, Eucspresso (http://eucspresso.bi.up.ac.za), which provides a searchable interface to the assembled contigs. The database can be queried based on the closest homologous entry in the *Arabidopsis thaliana* (TAIR9), *Populus trichocarpa* (Version 2.0) and *Vitis vinifera* (Sept 2009 build) sequence data sets. Simple and compound keyword searches can be performed based on all of the functional annotation terms and the predicted coding and protein sequences can be obtained for all contigs. Finally, the tissue-specific (FPKM) digital expression profile and the location of each contig in the draft 8X *E. grandis* genome assembly (http://www.phytozome.net) can be viewed from within Eucspresso.

2.5 Discussion

We have assembled nearly 19,000 expressed gene sequences from xylogenic and non-xylogenic tissues of an actively growing *Eucalyptus* plantation tree using only Illumina mRNA-Seq technology and *de novo* short-read assembly. Quality control comparisons to full-length, cloned, Sanger-derived transcript sequences from *Eucalyptus*, as well as multiple lines of evidence such as CDS prediction and Pfam prediction showed that the transcript assemblies are robust and that thousands of full-length coding sequences and their respective 5' and/or 3' UTR regions were successfully assembled. Comparison of assembled gene models to gene catalogs of other angiosperm species by BLAST analysis and functional annotation (GO, InterProScan and KEGG category numbers and proportions – Fig. 2.4, Fig. S2.6, Fig. S2.7 and Fig. S2.8) indicate that we have sampled an expansive and diverse expressed gene catalog representing a large proportion of the genes expressed in mature *Eucalyptus* trees across a variety of woody and non-woody tissues. Comparison to all publicly available *Eucalyptus* DNA sequence suggests that we have sampled a more comprehensive set of genes, which is also more complete in length (Table S2.3) from a single eucalypt tree genotype than has been available to date for the entire genus. Additionally, using a validated approach to quantify mRNA-Seq data we have produced an informative database of transcript abundance across six *Eucalyptus* tree tissues, which, due to the depth of sequencing, provides a wider dynamic range than Sanger or 454-derived EST counts usually associated with this type of analysis.

A concern associated with *de novo* assembly of transcript sequences, be it Sanger derived (Rengel *et al.*, 2009) or 454 sequence derived (Novaes *et al.*, 2008) assemblies, is the contiguity of assembled sequences. This concern intuitively increases as the read length decreases, and may be one of the main reasons why most transcriptome *de novo* assembly approaches have utilized technologies with longer read lengths to date. We provide several lines of evidence which jointly support the contiguity of transcript sequences assembled in our study using short-read data. First, a high proportion of the contigs exhibited

high-confidence BLASTX similarity to protein sequences from annotated gene catalogs of three angiosperm species *Arabidopsis*, *Populus* and *Vitis* (Fig. 2.3). Second, a large proportion of the contigs contained long, near full-length, predicted CDSs (Fig. 2.3). Third, InterproScan analysis predicted 45,687 protein domains, which is indicative of contiguous, in-frame predicted protein. Finally, a random subset of the contigs, which represented a variety of length and read coverage, were validated by direct alignment to previously published, Sanger sequenced, full-length *Eucalyptus* genes that were directly cloned from cDNA (Additional file 2.2).

Assigning biological significance to *de novo* assembled contigs should be approached with caution. In our study, 13,806 assembled gene models (73.07% of the total assembled contigs, Fig. 2.3A) were considered high confidence annotations due to the presence of a significant high stringency BLAST hit in other angiosperm species, as well as a predicted CDS. These contigs had relatively high coverage per base (CPB) values (median 47X) as compared to contigs lacking a predicted CDS (median CPB of 20X or lower, Fig. 2.3B and 2.3D, Table S2.2). Thus, a lack of CDS prediction was generally associated with low gene expression level and low CPB, which resulted in 'N's inserted by Velvet in the contig sequences (Fig. 2.3B and D, Table S2.2). The annotation of these sequences will be improved in the future by even deeper sequencing, addition of more tissue types and mapping to an annotated genome sequence. Another possible source of error is the spurious prediction of CDSs in long, non-coding RNAs, which has been previously shown to occur (Clamp *et al.*, 2007; Dinger *et al.*, 2008). It is notable that of the 1,813 *Eucalyptus*-derived contigs with no significant BLAST hit to other angiosperms, but containing a predicted CDS (Fig. 2.3C), only 81 contigs had predicted InterProScan domains. Additionally, the median CDS to contig length ratio was 0.33, as compared to 0.62 in the 13,806 high confidence contigs in Fig. 2.3A, which suggests that many of these CDS predictions may be false positives. The ability to distinguish and classify the lower confidence annotations is, however, beyond the scope of this study, and can only be further resolved once a genome-based predicted gene catalog is available.

Validation of the digital expression (FPKM) profiles using the "core xylem gene set" identified in *Arabidopsis* (Ko *et al.*, 2006) has precedence in similar investigations in conifers (Pavy *et al.*, 2008), cotton (Betancur *et al.*, 2010) and poplar (Dharmawardhana *et al.*, 2010). This analysis, combined with the ontology and pathway analyses of differentially expressed genes, lend support to the biological significance of digital expression profiles derived from short-read sequencing technology, which will assist in the discovery and annotation of novel *Eucalyptus* genes (and using the genome sequence, promoters) playing key roles in growth and development, and particularly in woody biomass production.

An important finding in this study is the transcriptional investment in xylem in pathways related to energy metabolism during xylogenesis – an aspect not previously explored in the literature. Two genes in particular, transketolase (EC 2.2.1.1) and ribose 5-phosphate isomerase A (EC 5.3.1.6) have both xylem- and leaf-specific members. These genes are key components regulating carbon flux pertaining to both the Calvin cycle and the non-oxidative stage of the pentose phosphate pathway (Fig. S2.12), and the presence of both xylem and leaf-specific members highlights these genes' roles in the flexibility required during carbon metabolism in source and sink tissues for glycolysis, glucogenesis, energy metabolism and aromatic amino acid biosynthesis. The importance of plastidic transketolase in particular has been previously highlighted for its roles in both photosynthesis and phenylpropanoid biosynthesis (Henkes *et al.*, 2001).

Additionally, this study revealed many cellular transport and signaling related genes that were highly and preferentially expressed in xylem. Fiber cells involved in cell wall deposition would require rapid biosynthesis of polysacchcarides at the golgi and plasma membrane, as well as coordinated transport of proteins and polysaccharides by vesicles, guided by actin and cortical microtubules (see Chapter 1 for

review). Out of 397 assembled contigs whose *Arabidopsis* homologs fall into the "Cell" MapMan bin, 117 were either overrepresented in xylem or leaf and of these, 87 (74%) were overrepresented in xylem specifically. The classes of proteins represented by these 87 genes are mainly tubulins and other proteins related to SNARE-related transport, as well as kinesins, coatomer proteins, myosins, actins and actin binding proteins (Additional file 2.5). A number of genes related to signaling were also specifically upregulated in xylem, mainly from the cytoplasmic receptor-like kinases (RLKs) and leucine-rich repeat RLKs, as well as genes coding for G-proteins, and genes related to calcium signaling (Additional file 2.5). Understanding these signaling and transport mechanisms for polysaccharide and lignin deposition is still in its infancy (Wightman & Turner, 2008; Oikawa *et al.*, 2010).

2.5.1 Conclusion

Taking into consideration the number, length, coverage and quality of assembled gene models, as well as their digital expression profiles, this dataset surpasses several previous *de novo* transcriptome assemblies using Illumina (Birol *et al.*, 2009; Gibbons *et al.*, 2009) or 454 technology (Novaes *et al.*, 2008; Vera *et al.*, 2008; Hahn *et al.*, 2009; Meyer *et al.*, 2009). This can primarily be attributed to the amount of data generated (3.93 Gbp of non-rRNA derived reads), the diversity of tissues sampled and strategy of paired-end sequencing, as well as read-length (mostly 50-60 bp, compared to only 36 bp in earlier studies). Our dataset was generated using several generations of Illumina GA technology, but considering the current throughput of Illumina sequencing (up to 100 Gbp per flowcell), a gene catalog of this scale can now be produced using a single lane of Illumina mRNA-Seq. The use of non-normalized cDNA libraries enabled gene expression profiling and provided insight into source-sink specific investment in gene expression in *Eucalyptus*.

This study demonstrates the utility of transcriptome assembly and expression profiling using RNA-seq to rapidly identify homologs in *Eucalyptus* involved in crucial biological processes. *Eucalyptus* xylem-specific expression for the subset of 86 *Arabidopsis* genes out of the larger non-redundant set of 208 genes provides good supporting evidence for the likely roles of these genes in cellulose biosynthesis in wood, and provides a reduced list of candidate genes for future reverse genetics studies. The Eucspresso (now EucGenIE) online resource produced from this study, as well as future comparative analysis with other woody species such as *Vitis* and *Populus*, will be valuable for studying the unique biology of woody perennials. Thus, in addition to highlighting important pathways and biological processes specifically regulated in xylem such as energy metabolism, this study provides a good reference towards the construction of a detailed "expression atlas" for *Eucalyptus*, and increases our understanding of genes involved in cellulose biosynthesis during wood formation.

2.6 Acknowledgements

The authors would like to acknowledge J. Rees and J.-M. Celton of the University of the Western Cape (Cape Town, South Africa) for assistance with Illumina GA sequencing. Plant materials were kindly provided by Sappi Forestry (Kwambonambi, South Africa). This work was supported through a strategic research grant from the South African Department of Science and Technology (DST) and by research funding from Sappi and Mondi, through the Wood and Fibre Molecular Genetics (WFMG) Programme, the Technology and Human Resources for Industry Programme (THRIP) and the National Research Foundation (NRF) of South Africa.

2.7 References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**(3): 403-410.

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

Betancur L, Singh B, Rapp RA, Wendel JF, Marks MD, Roberts AW, Haigler CH. 2010. Phylogenetically distinct cellulose synthase genes support secondary wall thickening in *Arabidopsis* shoot trichomes and cotton fiber. *Journal of Integrative Plant Biology* **52**(2): 205-220.

Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, Stazyk G, Morin RD, Zhao Y, Hirst M, Schein JE, Horsman DE, Connors JM, Gascoyne RD, Marra MA, Jones SJM. 2009. *De novo* transcriptome assembly with ABySS. *Bioinformatics* **25**(21): 2872-2877.

Borner GH, Sherrier DJ, Weimar T, Michaelson LV, Hawkins ND, MacAskill A, Napier JA, Beale MH, Lilley KS, Dupree P. 2005. Analysis of detergent-resistant membranes in *Arabidopsis*. Evidence for plasma membrane lipid rafts. *Plant Physiology* **137**(1): 104-116.

Brown DM, Goubet F, Wong VW, Goodacre R, Stephens E, Dupree P, Turner SR. 2007. Comparison of five xylan synthesis mutants reveals new insight into the mechanisms of xylan synthesis. *Plant Journal* **52**(6): 1154-1168.

Brown DM, Zeef LAH, Ellis J, Goodacre R, Turner SR. 2005. Identification of novel genes in *Arabidopsis* involved in secondary cell wall formation using expression profiling and reverse genetics. *Plant Cell* **17**(8): 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*. *Plant Journal* **57**(4): 732-746.

Burge C, Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* **268**(1): 78-94.

Chang S, Puryear J, Cairney J. 1993. A simple and efficient method for isolating RNA from pine trees. *Plant Molecular Biology Reporter* **11**(2): 113-116.

Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES. 2007. Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences of the United States of America* **104**(49): 19428-19433.

Cloonan N, Forrest ARR, Kolle G, Gardiner BBA, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SM. 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods* **5**(7): 613-619.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**(18): 3674-3676.

Dassanayake M, Haas JS, Bohnert HJ, Cheeseman JM. 2009. Shedding light on an extremophile lifestyle through transcriptomics. *New Phytologist* **183**(3): 764-775.

Dharmawardhana P, Brunner AM, Strauss SH. 2010. Genome-wide transcriptome analysis of the transition from primary to secondary stem development in *Populus trichocarpa*. *BMC Genomics* **11**(1). 150.

DiGuistini S, Liao N, Platt D, Robertson G, Seidel M, Chan S, Docking TR, Birol I, Holt R, Hirst M. 2009. *De novo* genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome Biology* **10**(9): R94.

Dinger ME, Pang KC, Mercer TR, Mattick JS. 2008. Differentiating protein-coding and noncoding RNA: Challenges and ambiguities. *PLoS Computational Biology* **4**(11).

Eldridge K, Davidson J, Harwood C, Van Wyk G. 1993. *Eucalypt domestication and breeding*: Clarendon Press, Oxford.

FAO. 2008. Forests and Energy. *FAO Forestry Paper No.* **154**(Rome): (ISBN 978-992-975-105985-105982).

Farrer RA, Kemen E, Jones JDG, Studholme DJ. 2009. *De novo* assembly of the *Pseudomonas syringae* pv. syringae B728a genome using Illumina/Solexa short sequence reads: RESEARCH LETTER. *FEMS Microbiology Letters* **291**(1): 103-111.

Geisler-Lee J, Geisler M, Coutinho PM, Segerman B, Nishikubo N, Takahashi J, Aspeborg H, Djerbi S, Master E, Andersson-Gunneras S, Sundberg B, Karpinski S, Teeri TT, Kleczkowski LA, Henrissat B, Mellerowicz EJ. 2006. Poplar carbohydrate-active enzymes. Gene identification and expression analyses. *Plant Physiol* **140**(3): 946-962.

Gibbons JG, Janson EM, Hittinger CT, Johnston M, Abbot P, Rokas A. 2009. Benchmarking next-generation transcriptome sequencing for functional and evolutionary genomics. *Molecular Biology and Evolution* **26**(12): 2731-2744.

Hahn DA, Ragland GJ, Shoemaker DD, Denlinger DL. 2009. Gene discovery using massively parallel pyrosequencing to develop ESTs for the flesh fly *Sarcophaga crassipalpis*. *BMC Genomics* **10**(1). 234

Hale MC, McCormick CR, Jackson JR, DeWoody JA. 2009. Next-generation pyrosequencing of gonad transcriptomes in the polyploid lake sturgeon (*Acipenser fulvescens*): The relative merits of normalization and rarefaction in gene discovery. *BMC Genomics* **10**(1): 203.

Henkes S, Sonnewald U, Badur R, Flachmann R, Stitt M. 2001. A small decrease of plastid transketolase activity in antisense tobacco transformants has dramatic effects on photosynthesis and phenylpropanoid metabolism. *The Plant Cell Online* **13**(3): 535-551.

Hernandez D, François P, Farinelli L, Østerås M, Schrenzel J. 2008. De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. *Genome Research* **18**(5): 802-809.

Hibino T. 2009. "Post-genomics" research in *Eucalyptus* in the near future. *Plant Biotechnology* **26**(1): 109-113.

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 and Developmental Biology - Plant* **45**(6): 619-629.

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.

Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**(7161): 463-467.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **28**(1): 27-30.

Kaul S, Koo HL, Jenkins J, Rizzo M, Rooney T, Tallon LJ, Feldblyum T, Nierman W, Benito MI, Lin X, Town CD, Venter JC, Fraser CM, Tabata S, Nakamura Y, Kaneko T, Sato S, Asamizu E, Kato T, Kotani H, Sasamoto S, Ecker JR, Theologis A, Federspiel NA, Palm CJ, Osborne BI, Shinn P, Conway AB, Vysotskaia VS, Dewar K, Conn L, Lenz CA, Kim CJ, Hansen NF, Liu SX, Buehler E, Altafi H, Sakano H, Dunn P, Lam B, Pham PK, Chao Q, Nguyen M, Yu G, Chen H, Southwick A, Jeong Mi L, Miranda M, Toriumi MJ, Davis RW, Wambutt R, Murphy G, Düsterhoft A, Stiekema W, Pohl T, Entian KD, Terryn N, Volckaert G, Salanoubat M, Choisne N, Rieger M, Ansorge W, Unseld M, Fartmann B, Valle G, Artiguenave F, Weissenbach J, Quetier F, Wilson RK, De la Bastide M, Sekhon M, Huang E, Spiegel L, Gnoj L, Pepin K, Murray J, Johnson D, Habermann K, Dedhia N, Parnell L, Preston R, Hillier L, Chen E, Marra M, Martienssen R, McCombie WR, Mayer

K, White O, Bevan M, Lemcke K, Creasy TH, Bielke C, Haas B, Haase D, Maiti R, Rudd S, Peterson J, Schoof H, Frishman D. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**(6814): 796-815.

Ko JH, Beers EP, Han KH. 2006. Global comparative transcriptome analysis identifies gene network regulating secondary xylem development in *Arabidopsis thaliana*. *Molecular Genetics and Genomics* **276**(6): 517-531.

Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature Methods* **6**(4): 291-295.

Kristiansson E, Asker N, Förlin L, Joakim DGJ. 2009. Characterization of the *Zoarces viviparus* liver transcriptome using massively parallel pyrosequencing. *BMC Genomics* **10**(1): 305.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**(3).

Lee C, Teng Q, Zhong R, Yuan Y, Haghighat M, Ye ZH. 2012a. Three *Arabidopsis* DUF579 domain-containing GXM proteins are methyltransferases catalyzing 4-o-methylation of glucuronic acid on xylan. *Plant and Cell Physiology* **53**(11): 1934-1949.

Lee C, Zhong R, Ye ZH. 2012b. *Arabidopsis* family GT43 members are xylan xylosyltransferases required for the elongation of the xylan backbone. *Plant and Cell Physiology* **53**(1): 135-143.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754.

Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, Zhang Z, Zhang Y, Wang W, Li J, Wei F, Li H, Jian M, Nielsen R, Li D, Gu W, Yang Z, Xuan Z, Ryder OA, Leung FCC, Zhou Y, Cao J, Sun X, Fu Y, Fang X, Guo X, Wang B, Hou R, Shen F, Mu B, Ni P, Lin R, Qian W, Wang G, Yu C, Nie W, Wang J, Wu Z, Liang H, Min J, Wu Q, Cheng S, Ruan J, Wang M, Shi Z, Wen M, Liu B, Ren X, Zheng H, Dong D, Cook K, Shan G, Zhang H, Kosiol C, Xie X, Lu Z, Li Y, Steiner CC, Lam TTY, Lin S, Zhang Q, Li G, Tian J, Gong

T, Liu H, Zhang D, Fang L, Ye C, Zhang J, Hu W, Xu A, Ren Y, Zhang G, Bruford MW, Li Q, Ma L, Guo Y, An N, Hu Y, Zheng Y, Shi Y, Li Z, Liu Q, Chen Y, Zhao J, Qu N, Zhao S, Tian F, Wang X, Wang H, Xu L, Liu X, Vinar T, Wang Y, Lam TW, Yiu SM, Liu S, Huang Y, Yang G, Jiang Z, Qin N, Li L, Bolund L, Kristiansen K, Wong GKS, Olson M, Zhang X, Li S, Yang H. 2010. The sequence and de novo assembly of the giant panda genome. *Nature* **463**(7279): 311-317.

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. 2008. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**(3): 523-536.

Mentzen WI, Wurtele ES. 2008. Regulon organization of *Arabidopsis*. *BMC Plant Biology* **8**(1): 99.

Meyer E, Aglyamova GV, Wang S, Buchanan-Carter J, Abrego D, Colbourne JK, Willis BL, Matz MV. 2009. Sequencing and *de novo* analysis of a coral larval transcriptome using 454 GSFlx. *BMC Genomics* **10**(1): 219.

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**(1).

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**(7): 621-628.

Mutwil M, Ruprecht C, Giorgi FM, Bringmann M, Usadel B, Persson S. 2009. Transcriptional wiring of cell wall-related genes in *Arabidopsis*. *Molecular Plant* **2**(5): 1015-1024.

Mutwil M, Usadel B, Schütte M, Loraine A, Ebenhöh O, Persson S. 2010. Assembly of an interactive correlation network for the *Arabidopsis* genome using a novel Heuristic Clustering Algorithm. *Plant Physiology* **152**(1): 29-43.

Myburg AA, D. Grattapaglia, G.A. Tuskan, J. Schmutz, K. Barry, J. Bristow, and The Eucalyptus Genome Network. 2008. Sequencing the *Eucalyptus* genome: genomic resources for renewable

energy and fiber production. *Plant & Animal Genome XVI Conference W195, January 12-16, 2008. San Diego, CA.*

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**(5881): 1344-1349.

Novaes E, Drost DR, Farmerie WG, Pappas Jr GJ, Grattapaglia D, Sederoff RR, Kirst M. 2008. High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* **9**(1): 312.

Nowrousian M. 2010. Next-generation sequencing techniques for eukaryotic microorganisms: Sequencing-based solutions to biological problems. *Eukaryotic Cell* **9**(9): 1300-1310.

Obayashi T, Hayashi S, Saeki M, Ohta H, Kinoshita K. 2009. ATTED-II provides coexpressed gene networks for *Arabidopsis*. *Nucleic Acids Research* **37**(SUPPL. 1).

Obayashi T, Kinoshita K, Nakai K, Shibaoka M, Hayashi S, Saeki M, Shibata D, Saito K, Ohta H. 2007. ATTED-II: A database of co-expressed genes and cis elements for identifying co-regulated gene groups in *Arabidopsis*. *Nucleic Acids Research* **35**(SUPPL. 1).

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.

Pavy N, Boyle B, Nelson C, Paule C, Giguère I, Caron S, Parsons LS, Dallaire N, Bedon F, Bérubé H, Cooke J, Mackay J. 2008. Identification of conserved core xylem gene sets: Conifer cDNA microarray development, transcript profiling and computational analyses. *New Phytologist* **180**(4): 766-786.

Peña MJ, Zhong R, Zhou GK, Richardson EA, O'Neill MA, Darvill AG, York WS, Yeb ZH. 2007. *Arabidopsis* irregular xylem8 and irregular xylem9: Implications for the complexity of glucuronoxylan biosynthesis. *Plant Cell* **19**(2): 549-563.

Persson S, Wei H, Milne J, Page GP, Somerville CR. 2005. Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proceedings of the National Academy of Sciences of the United States of America* **102**(24): 8633-8638.

Pinheiro M, Afreixo V, Moura G, Freitas A, Santos MAS, Oliveira JL. 2006. Statistical, computational and visualization methodologies to unveil gene primary structure features. *Methods of Information in Medicine* **45**(2): 163-168.

Ranik M, Myburg AA. 2006. Six new cellulose synthase genes from *Eucalyptus* are associated with primary and secondary cell wall biosynthesis. *Tree Physiology* **26**(5): 545-556.

Rengel D, Clemente HS, Servant F, Ladouce N, Paux E, Wincker P, Couloux A, Sivadon P, Grima-Pettenati J. 2009. A new genomic resource dedicated to wood formation in *Eucalyptus*. *BMC Plant Biology* **9**(1): 36.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open software suite. *Trends in Genetics* **16**(6): 276-277.

Schwarz D, Robertson HM, Feder JL, Varala K, Hudson ME, Ragland GJ, Hahn DA, Berlocher SH. 2009. Sympatric ecological speciation meets pyrosequencing: Sampling the transcriptome of the apple maggot *Rhagoletis pomonella*. *BMC Genomics* **10**(1): 633.

Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**(SUPPL. 2):215-225.

Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* **6**(5): 377-382.

Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, Selbig J, Müller LA, Rhee SY, Stitt M. 2004. mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *The Plant Journal* **37**(6): 914-939.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, 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**(5): 511-515.

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam M, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Dŕjardin A, DePamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehlting J, Ellis B, Gendler K, Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrissat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjärvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leplé JC, Locascio P, Lou Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson P, Rinaldi C, Ritland K, Rouzé P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai CJ, Uberbacher E, Unneberg P, Vahala J, Wall K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Van De Peer Y, Rokhsar D. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**(5793): 1596-1604.

Urbanowicz BR, Peña MJ, Ratnaparkhe S, Avci U, Backe J, Steet HF, Foston M, Li H, O'Neill MA, Ragauskas AJ, Darvill AG, Wyman C, Gilbert HJ, York WS. 2012. 4-O-methylation of glucuronic acid in *Arabidopsis* glucuronoxylan is catalyzed by a domain of unknown function family 579 protein. *Proceedings of the National Academy of Sciences of the United States of America* **109**(35): 14253-14258.

Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH. 2008. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology* **17**(7): 1636-1647.

Weber APM, Weber KL, Carr K, Wilkerson C, Ohlrogge JB. 2007. Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing. *Plant Physiology* **144**(1): 32-42.

Wightman R, Turner SR. 2008. The roles of the cytoskeleton during cellulose deposition at the secondary cell wall. *Plant Journal* **54**(5): 794-805.

Wilhelm BT, Landry JR. 2009. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* **48**(3): 249-257.

Wu AM, Hörnblad E, Voxeur A, Gerber L, Rihouey C, Lerouge P, Marchant A. 2010. Analysis of the *Arabidopsis* IRX9/IRX9-L and IRX14/IRX14-L pairs of glycosyltransferase genes reveals critical contributions to biosynthesis of the hemicellulose glucuronoxylan. *Plant Physiology* **153**(2): 542-554.

Wu AM, 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. *Plant Journal* **57**(4): 718-731.

Wu T, Qin Z, Zhou X, Feng Z, Du Y. 2010. Transcriptome profile analysis of floral sex determination in cucumber. *Journal of Plant Physiology* **167**(11): 905-913.

Zerbino DR, Birney E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* **18**(5): 821-829.

2.8 Figures and Tables

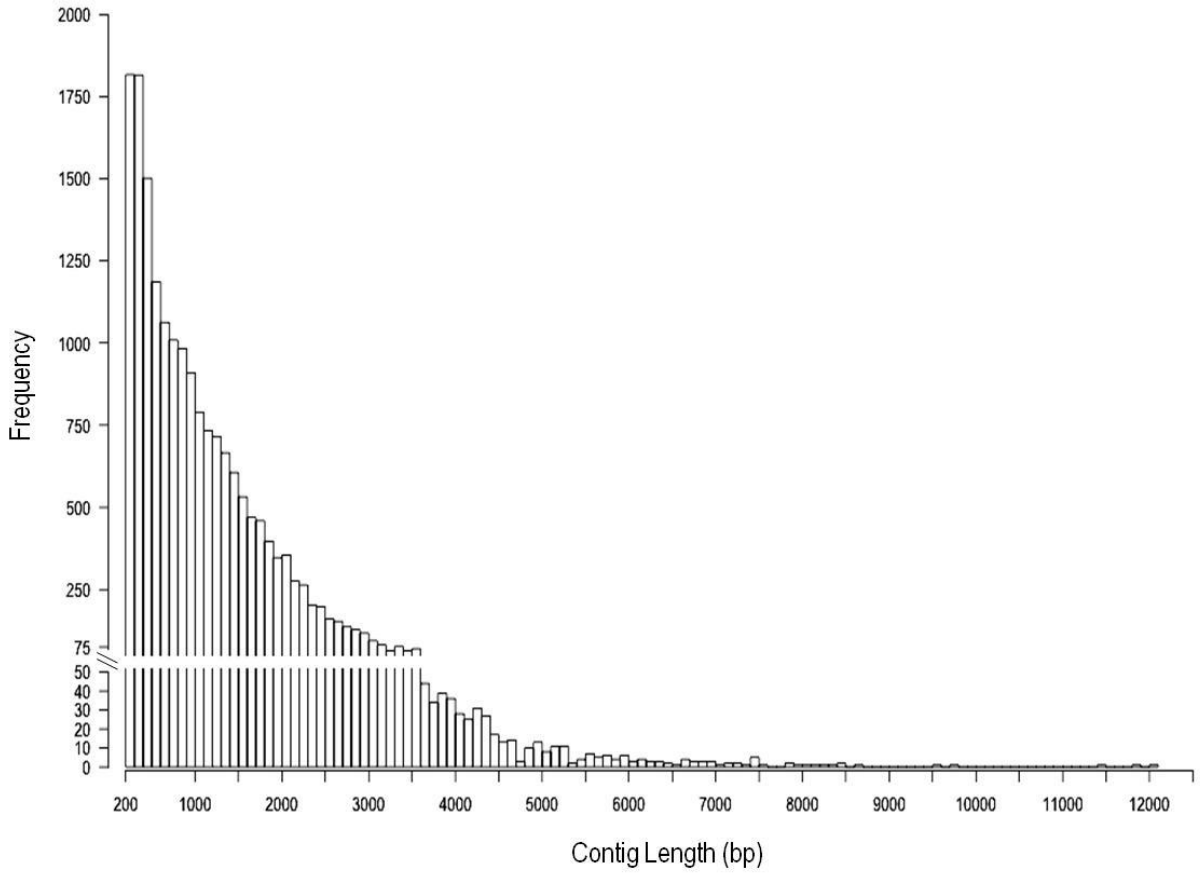


Fig. 2.1 Summary distribution of the lengths of the 18,894 assembled contigs (>200 bp, mean length = 1170 bp, N50 = 1,640 bp, Q3 = 1,573 bp, Max = 12,053 bp).

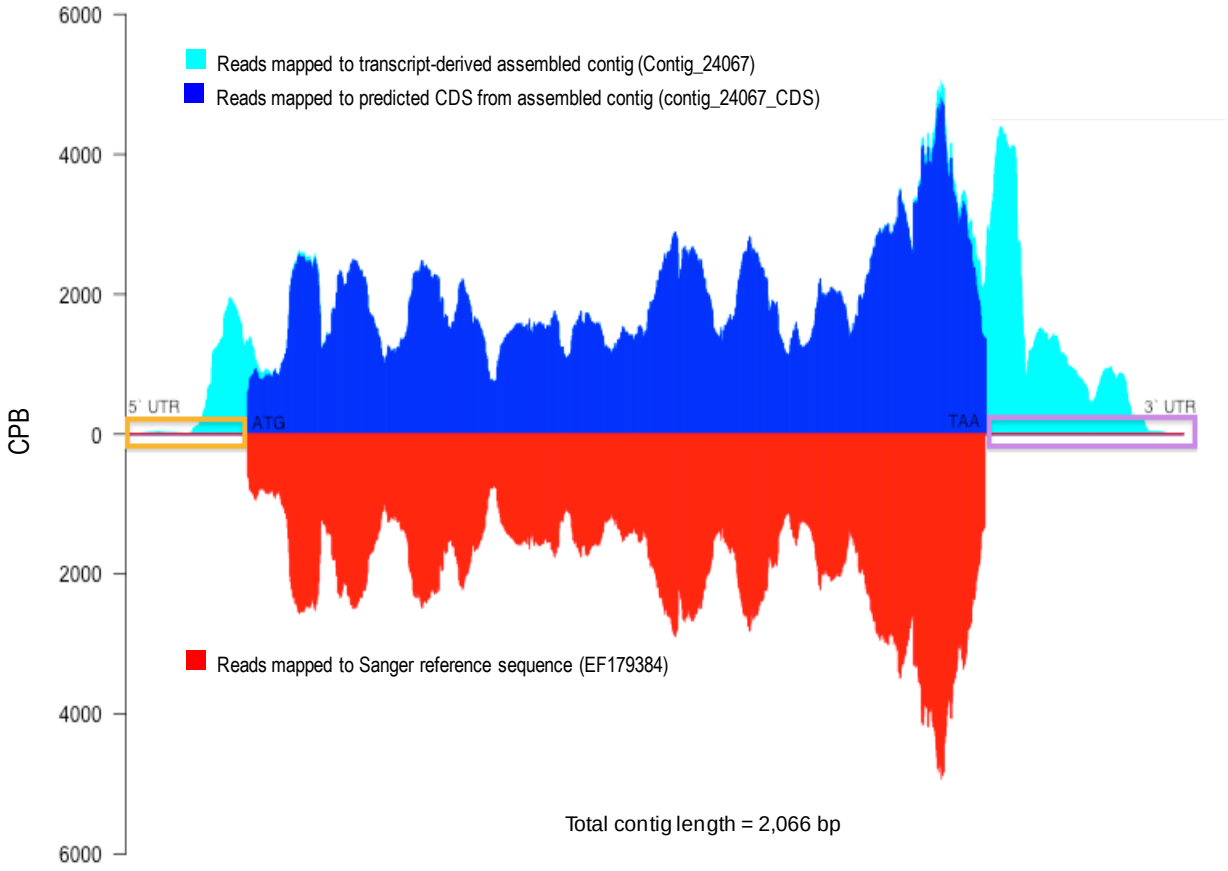


Fig. 2.2 Comparison of the *de novo* assembled contig of the *Eucalyptus grandis* UDP-glucose dehydrogenase (*UGDH*) transcript to a reference Sanger-based sequence (Genbank EF179384) for the same gene.

Peak height indicates coverage per base (CPB) of mapped short-reads across each sequence. CPB of the fully assembled contig is shown in cyan. CPB of the predicted CDS region is shown in dark blue. CPB of the Sanger reference sequence is shown in red. 5' UTR (orange box) and 3' UTR (purple box) regions are indicated.

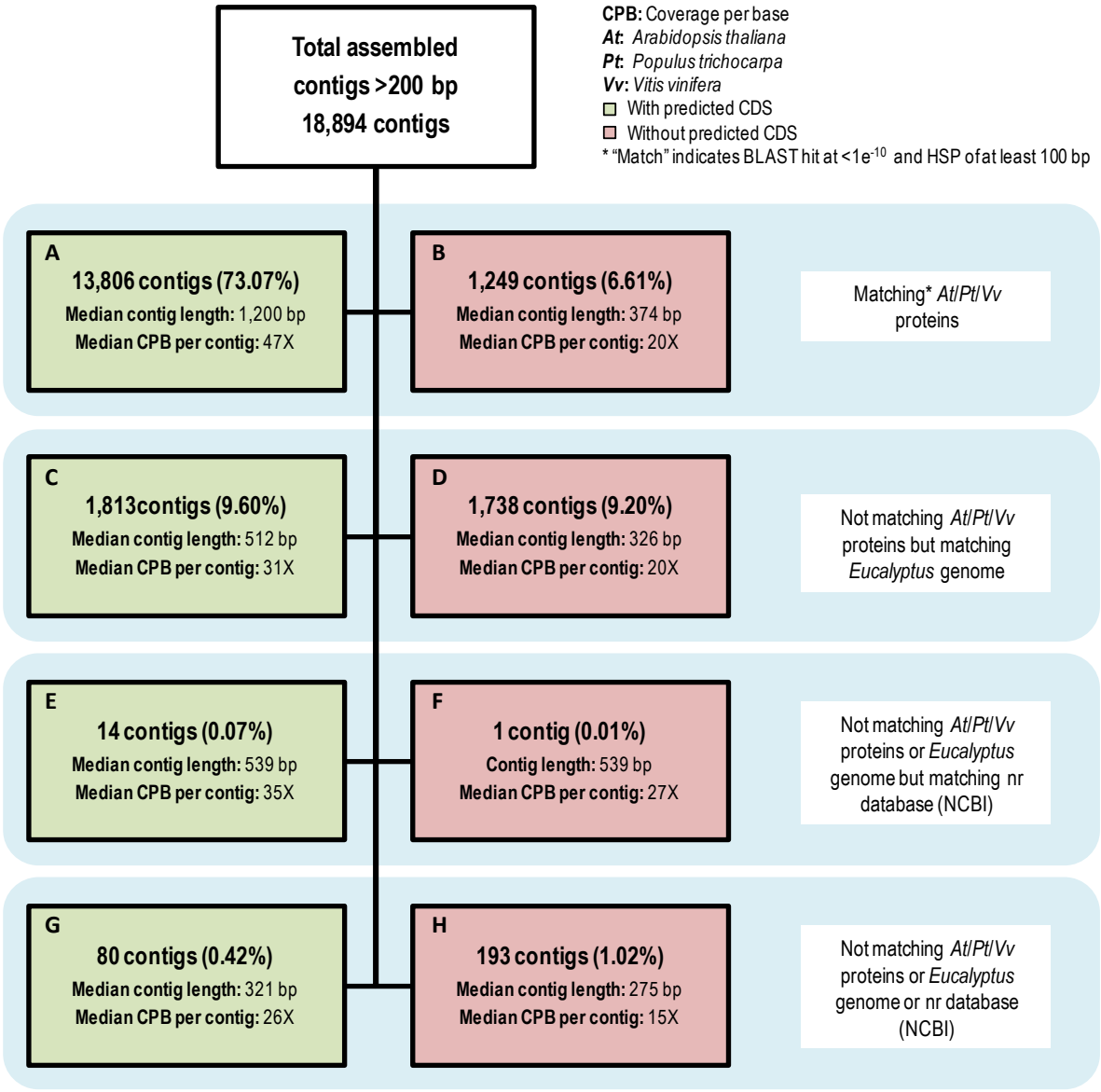


Fig. 2.3 Breakdown of annotation categories for all 18,894 transcript-derived contigs.

A large proportion (98.5%) of assembled contigs (A-D) had significant BLAST hits (<1e^{-10} confidence, minimum 100 bp HSP match length) to the draft *Eucalyptus* genome assembly (http://eucalyptusdb.bi.up.ac.za/), 80% of which (A, B) also exhibited significant similarity (BLASTX, <1e^{-10}, > 100 bp HSP) to coding sequences of *Arabidopsis*, *Populus* or *Vitis*.

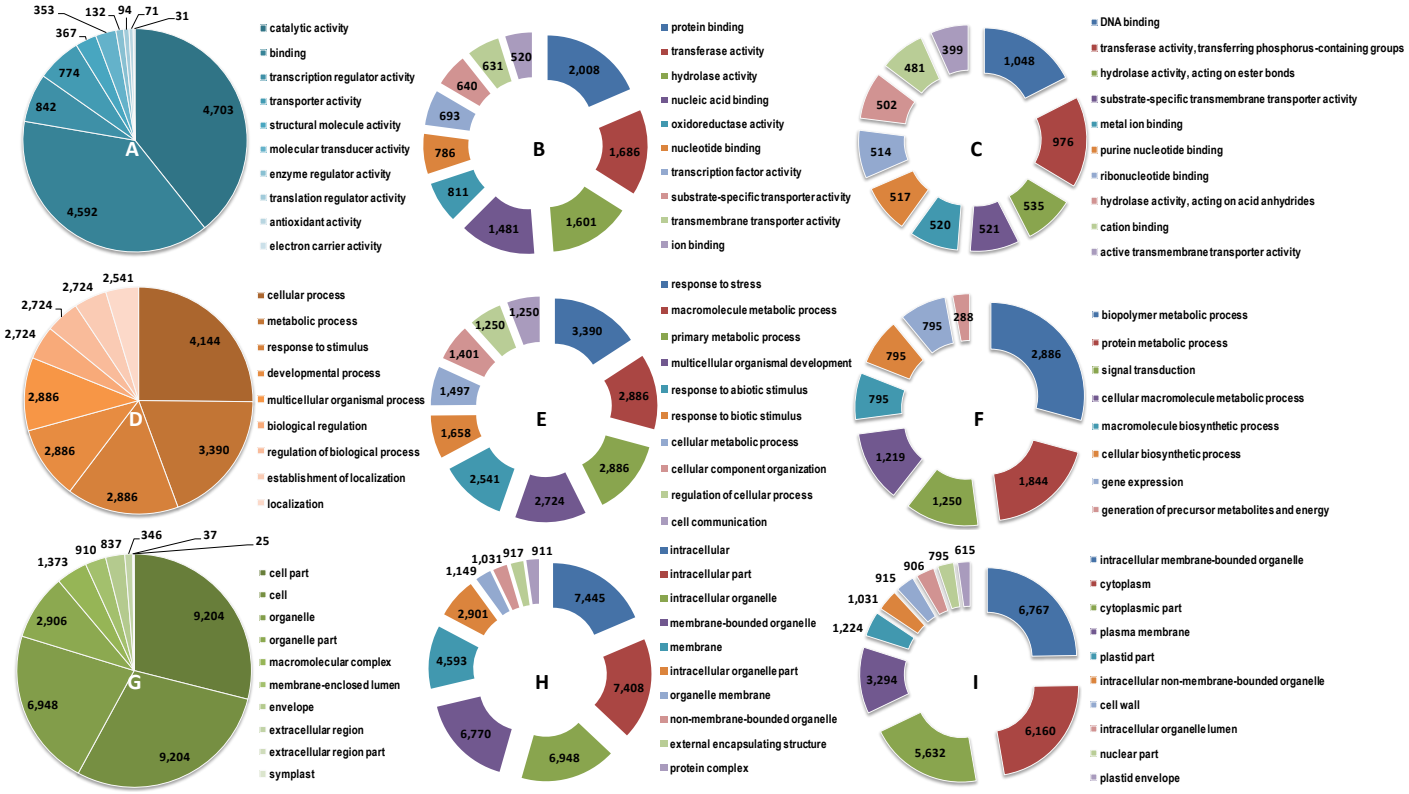
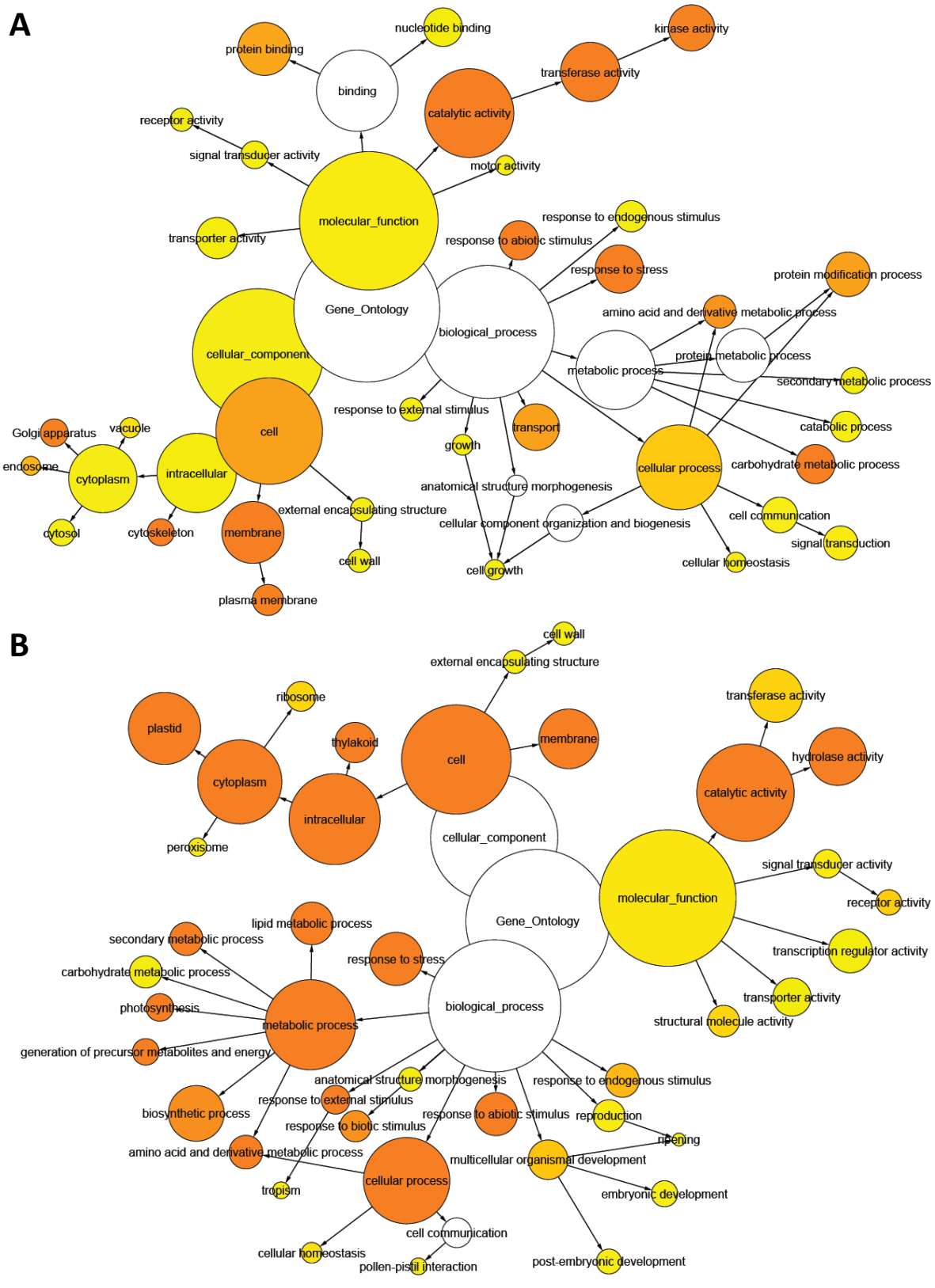


Fig. 2.4 Gene Ontologies represented in the gene catalog.

Top ten most represented GO categories under the “Molecular Function” (A-C), “Biological Process” (D-F) and “Cellular Compartment” (G-I) categories in level 2 (A, D and G), 3 (B, E and H) and 4 (C, F and I) are shown. The numbers and proportions in all categories reflect the diversity and complexity of genes expressed in multiple tissues sampled to make up the *Eucalyptus* gene catalog.

Fig. 2.5 Over-represented GO categories in xylem (A – 1,897 annotated contigs) and leaf (B – 1,531 annotated contigs) tissues.

All genes with a FPKM value more than two-fold higher in one tissue type versus the other were considered for the analysis. Samples were analysed using BiNGO (Maere et al. 2005). Node size is proportional to the number of genes in each category and colors shaded according to significance level (white – no significant difference, yellow – FDR = 0.05, Orange – FDR < 0.05).



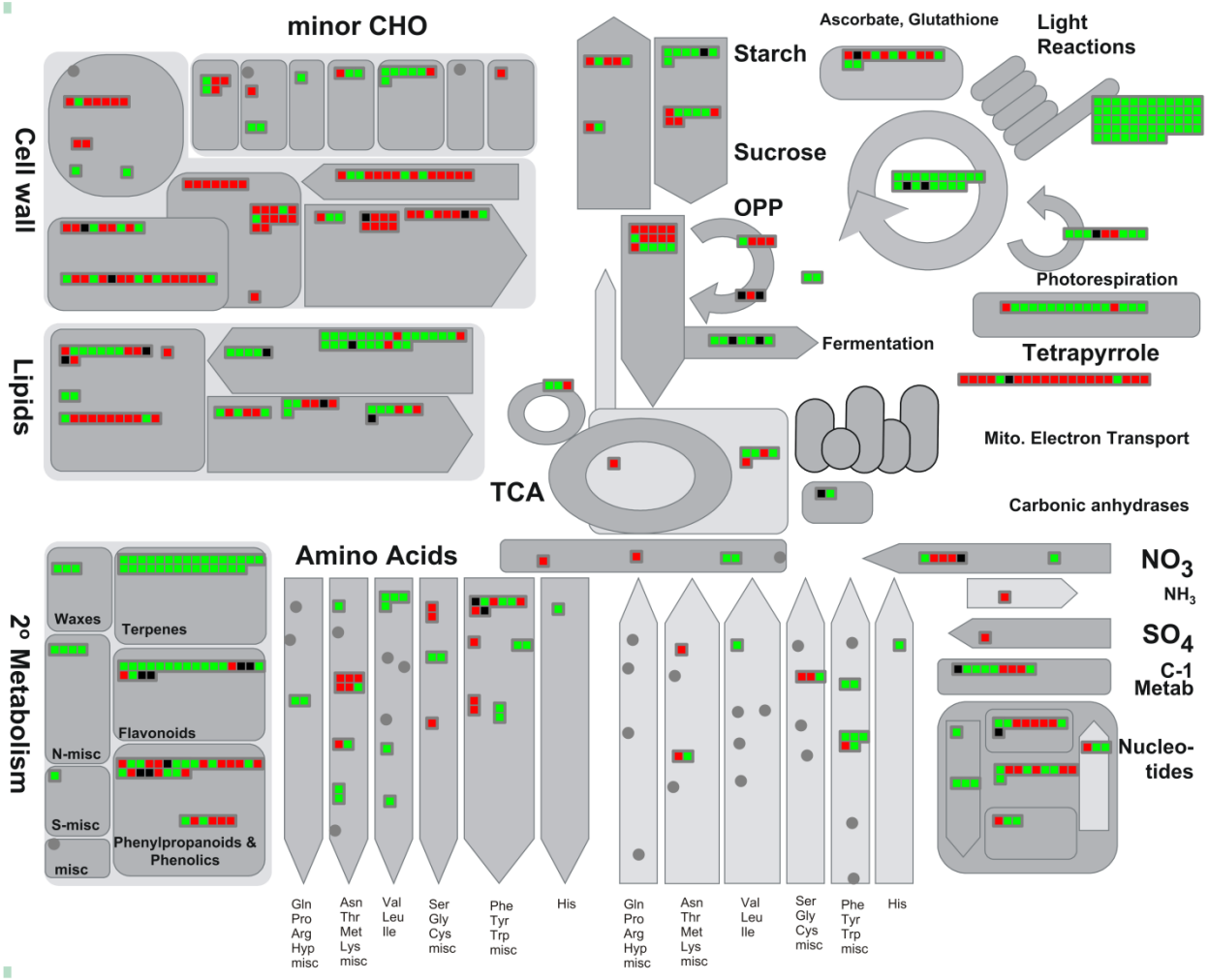


Fig. 2.6. Metabolism overview (MapMan) of annotated genes that are differentially expressed in xylem (red), leaf (green) or genes that have members differentially expressed in both xylem and leaf (black).

2.9 Additional files

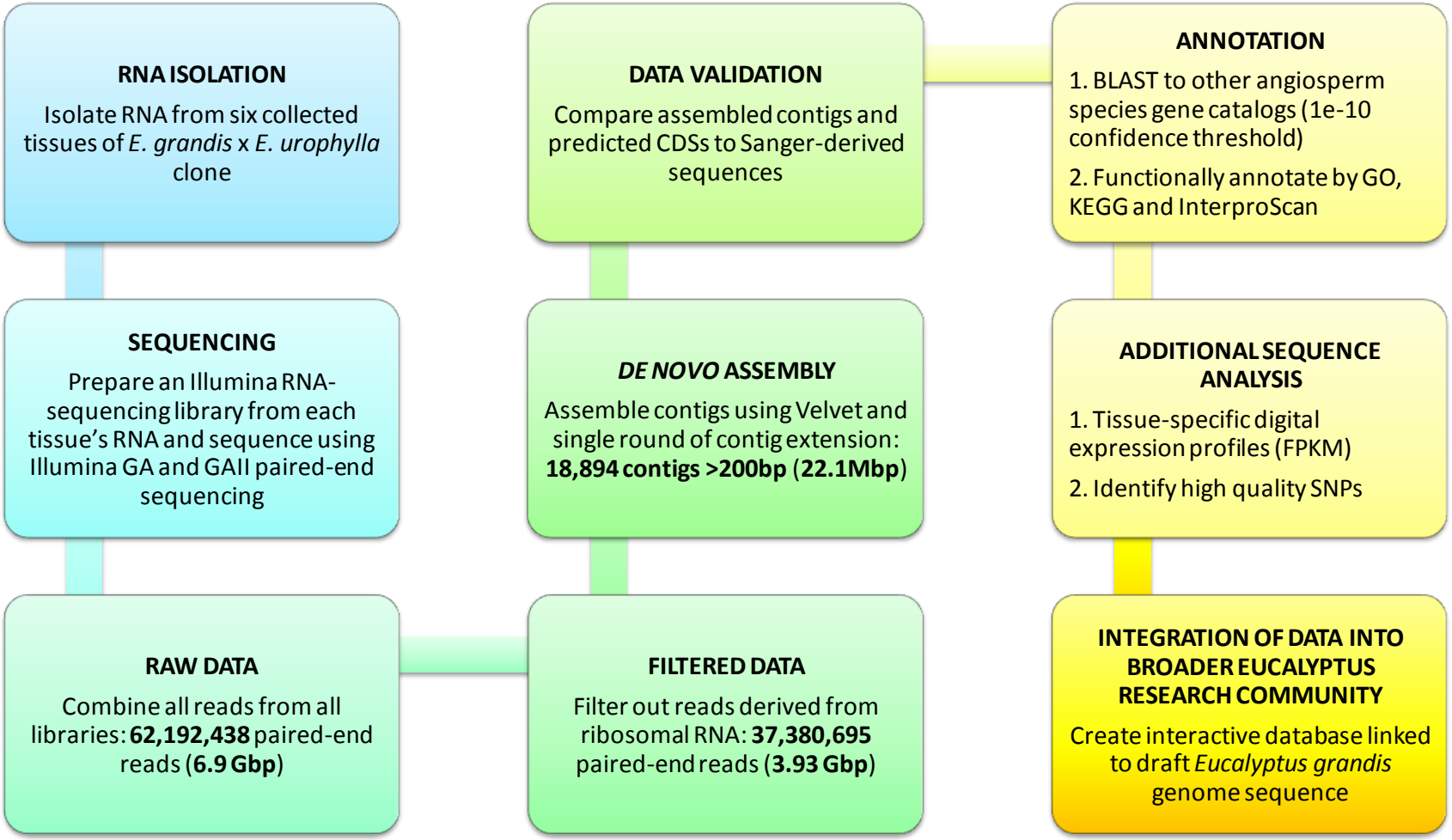
1. Additional file 2.1.fasta – FASTA formatted sequences of all 18,894 assembled contigs.
2. Additional file 2.2.doc – Contig validation, Needleman-Wunsch alignment Fig.s.
3. Additional file 2.3.xls – Table containing all 18,894 contig names and calculated FPKM values for six tissues (immature xylem, xylem, phloem, shoot-tips, young leaves and mature leaves).
4. Additional file 2.4.zip – Input data for MapMan and KEGG analysis. MapMan_input.txt: Codes for tissues apecificity are as follows: "3" – xylem-specific, "-3" – leaf-specific, "0" – specific members in both xylem and leaf. Input data for KEGG analysis that can be explored using the "KEGG_input.txt" file and the KeggMapper tool (http://www.genome.jp/kegg/tool/map_pathway2.html). Colour allocation indicate tissue specificity ("red" – xylem-specific, "green" – leaf-specific, blue – specific members in both xylem and leaf).
5. Additional file 2.5.xls – MapMan bins and *Arabidopsis* homolog IDs for major categories showing xylem-preferentially expressed members. Column F indicates tissue-specificity of expression ("3" – xylem-specific, "-3" – leaf-specific, "0" – specific members in both xylem and leaf).
6. Additional file 2.6.xlsx – Lists of *CesA* related genes from *Arabidopsis* literature ("*Arabidopsis* cellulose genes"), as well as *Eucalyptus* homologs of these genes that are preferentially expressed in xylem ("*Eucalyptus* cellulose genes"). Subcellular localization prediction is based on ATTED II annotations (Obayashi *et al.*, 2007; Obayashi *et al.*, 2009). Xylem (XY) and immature xylem (IX) FPKM are provided, as well as the xylem/leaf relative expression ratio (X/L) and each gene's correlation (column H) with a target gene *EgCesA3* (Contig 31).

Eucspresso (http://eucspresso.bi.up.ac.za/) - Online database with mRNA contig sequences and their Blast, GO, KEGG, Pfam annotations. Since the publication of a manuscript from this chapter (Mizrachi *et al.*, 2010), the data (assembled contigs as well as the reads re-mapped to the genome) contributed to the final annotation of the *E. grandis* genome (Myburg *et al.*, in preparation). Data from the produced Eucspresso dataset has also since been integrated into a more comprehensive *Eucalyptus* gene expression database we developed – EucGenIE (http://eucgenie.org/: Hefer, van der Merwe, Mizrachi, Joubert and Myburg, *in preparation*).

2.10 Supplemental data

Fig. S2.1 Summary of whole-transcriptome analysis strategy.

RNA was isolated from six tissues (Supplemental Table S1) of a *Eucalyptus grandis x E. urophylla* F1 hybrid clone. Tissue-specific Illumina RNA-Seq libraries were paired-end (PE) sequenced to generate a total of 6.9 Gbp of raw sequence. After filtering out ribosomal RNA derived and low quality reads, ≈36 million paired-end reads (3.93 Gbp) were *de novo* assembled using Velvet (version 0.7.30, Zerbino and Birney 2008) and a round of contig extension using custom scripts as explained in the METHODS section. Assembly quality was investigated by mapping reads to- and aligning a subsection of assembled contigs with their corresponding full-length reference Sanger gene sequences from NCBI, to evaluate assembly contiguity. Annotation was carried out by high stringency BLAST query of gene catalogs from three sequenced angiosperm species – *Arabidopsis thaliana*, *Populus trichocarpa* and *Vitis vinifera*. Functional annotation was performed by assigning Gene Ontology (GO - http://www.geneontology.org/), KEGG (http://www.genome.jp/kegg/) and InterProScan (http://www.ebi.ac.uk/interpro/) terms to each contig. A tissue-specific FPKM value for each contig was calculated using Cufflinks (Trapnell *et al.*, 2010), and SNPs detected across 13,806 high quality contigs in coding and non-coding regions using SAMtools (Li et al. 2009). All data was integrated into an interactive database, Eucspresso (http://eucspresso.bi.up.ac.za).



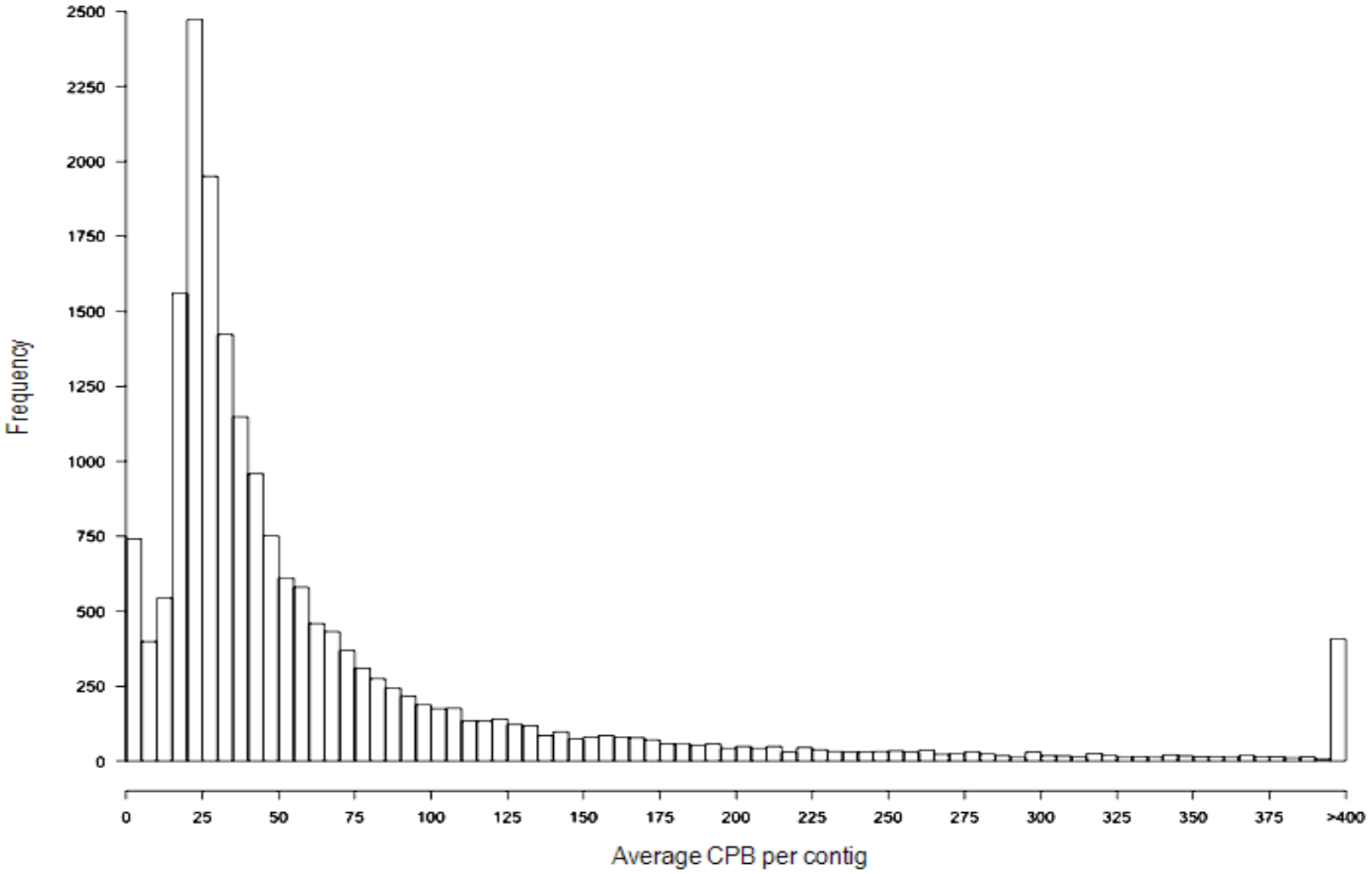


Fig. S2.2 Average coverage of transcript-derived short-read contigs.

Coverage per base (CPB) was calculated across each contig and a frequency histogram constructed from the coverage values of all contigs (Median CPB = 37X, Q3 = 72X, Max CPB = 5,262X).

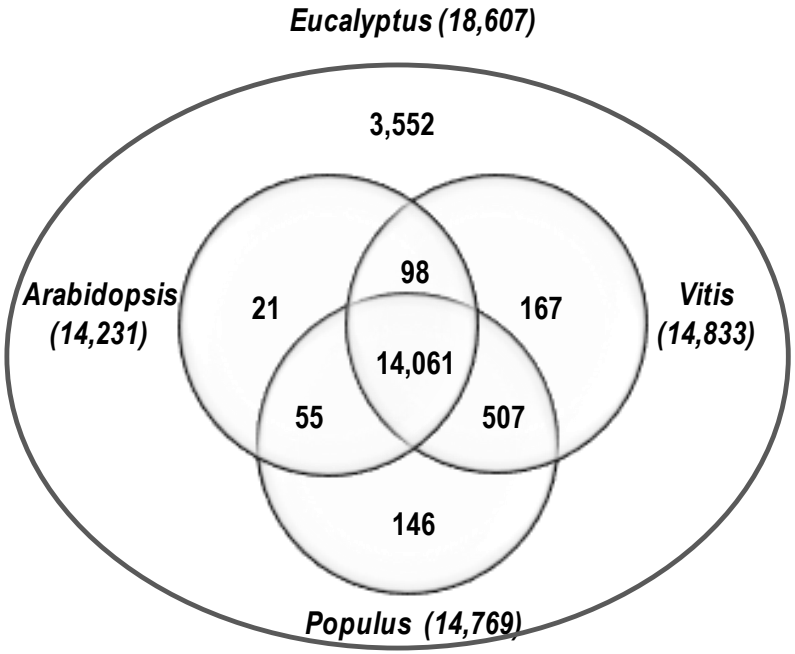


Fig. S2.3 High stringency BLAST analysis (<1e^{-10} confidence blastx, minimum 100 bp HSP match length) of the *Eucalyptus* transcript-derived contigs against protein datasets from three reference sequenced angiosperm genera (*Arabidopsis*, *Populus* and *Vitis*).

In total, 15,505 contigs (82.06% of the total contig dataset) exhibited similarity to *Arabidopsis* (14,231 contigs), *Populus* (14,769 contigs) or *Vitis* (14,833 contigs), while 3,552 did not show similarity to any of the three protein datasets at the chosen confidence threshold. A core set of 14,061 (74.4%) exhibited high similarity to all three protein sets.

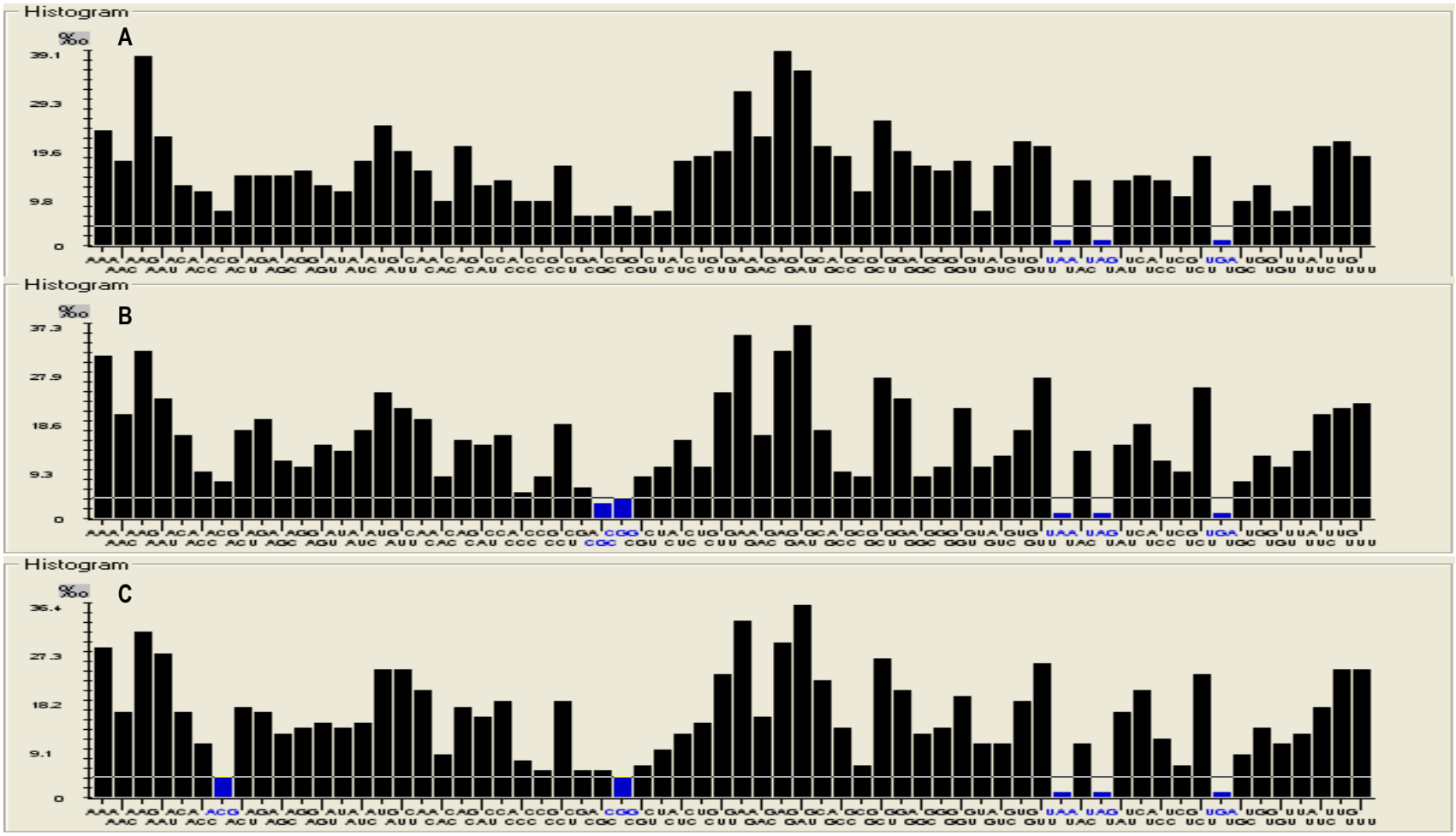


Fig. S2.4 Codon usage histogram for predicted coding sequences in the *Eucalyptus grandis* x *E. urophylla* hybrid (A), *Arabidopsis thaliana* (B) and *Populus tricocarpa* (C) gene catalogs.

Rare codons (<5%) are highlighted in blue. Analysis was performed using Anaconda 1.5 (Pinheiro et al. 2006). The y-axis shows frequency of codon usage.

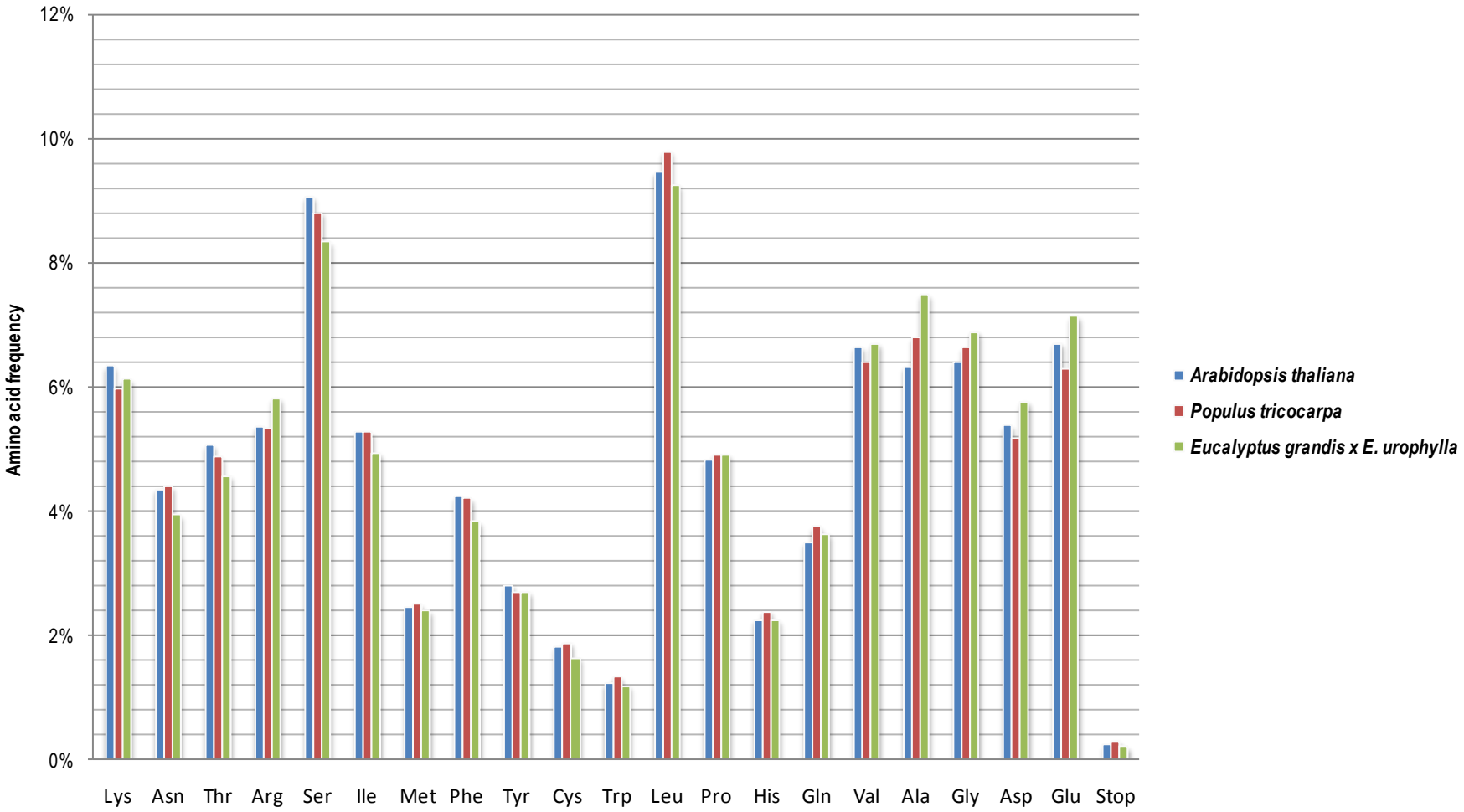


Fig. S2.5 Amino acid frequencies in the predicted proteomes of *Arabidopsis thaliana* and *Populus trichocarpa*, as compared to the predicted proteins from the expressed gene catalog of the *Eucalyptus grandis* x *E. urophylla* F1 hybrid.

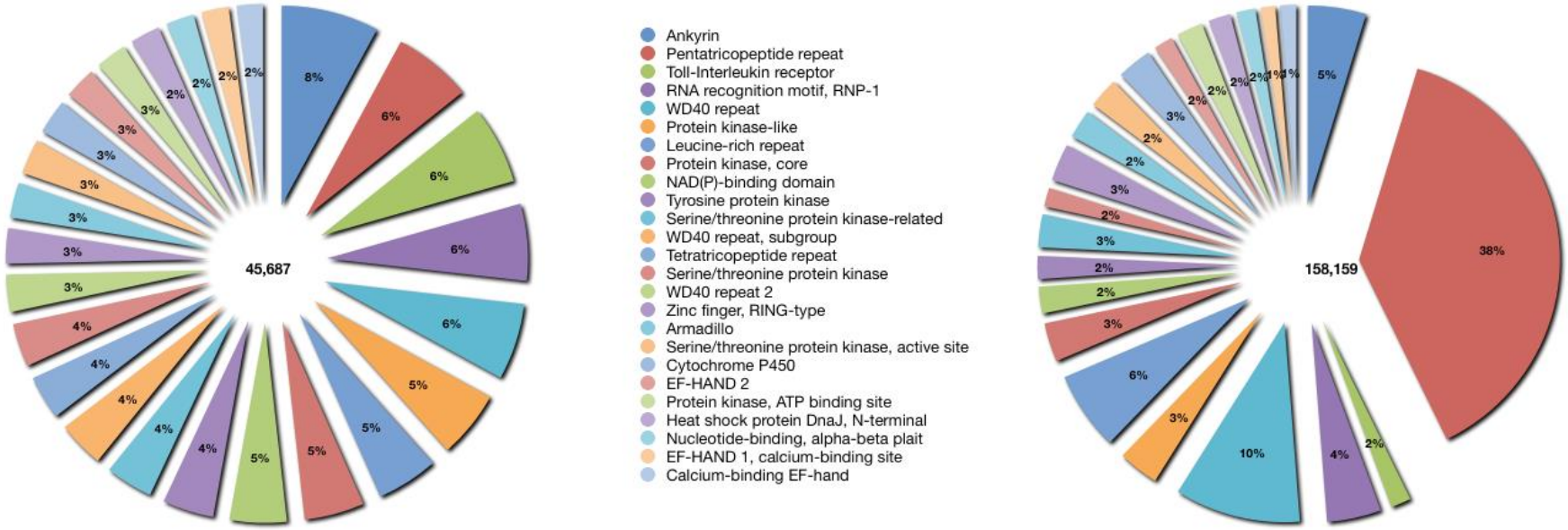
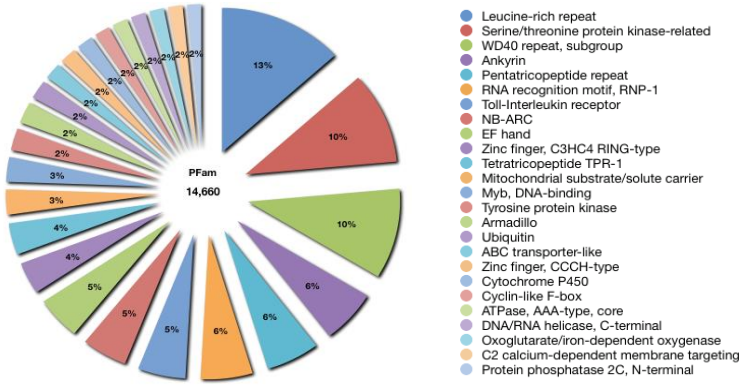


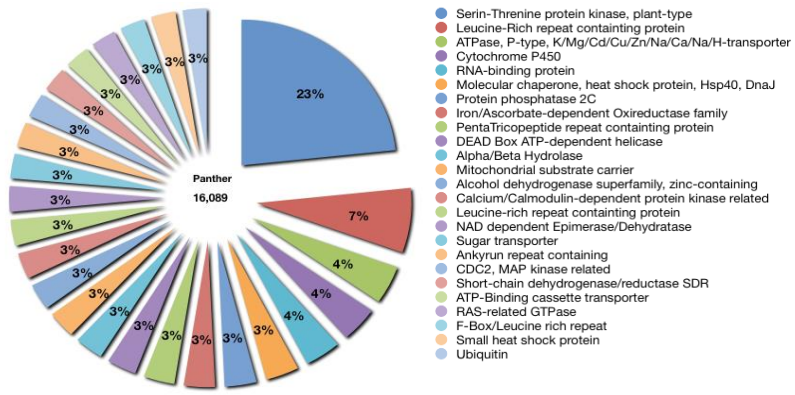
Fig. S2.6 Comparison of the 25 most abundant InterProScan categories present in the *Eucalyptus* gene catalogue (left) and their relative abundance in the complete *Arabidopsis* predicted protein coding gene catalog (right).

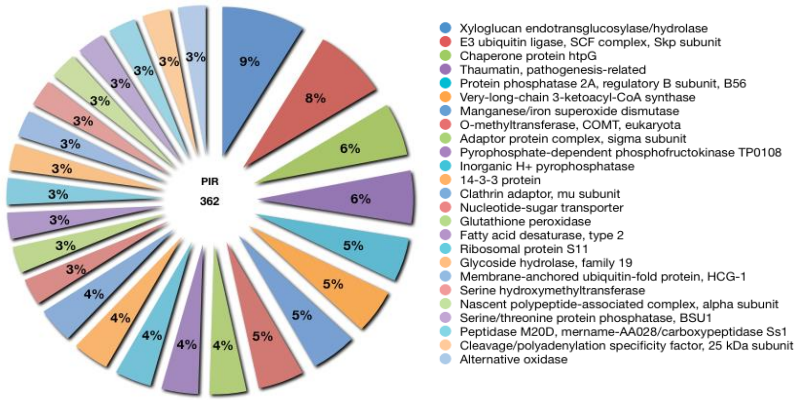
We annotated 45,687 domains by InterProScan in the *de novo* assembled *Eucalyptus* transcribed dataset (18,894 assembled contigs, 22.1 Mbp), as compared to 158,159 domains annotated in the complete TAIR 9 predicted coding gene dataset (39,640 genes, 87 Mbp).

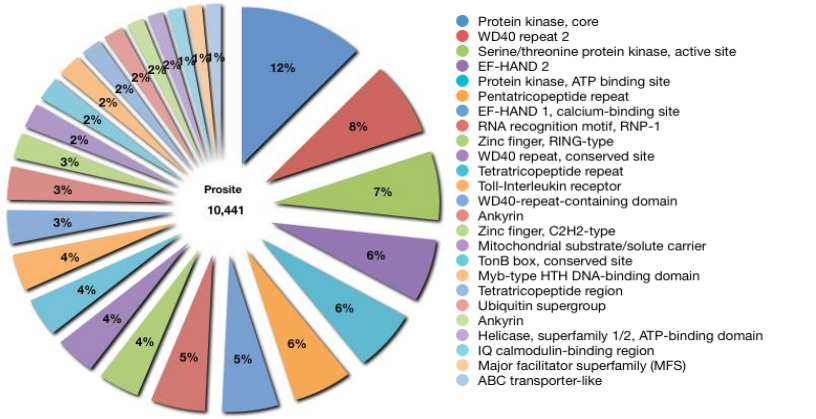
Fig. S2.7 Summary InterProScan statistics of the top 25 most populated categories in all domain based annotation of 18,894 *de novo* assembled contigs from *Eucalyptus*.

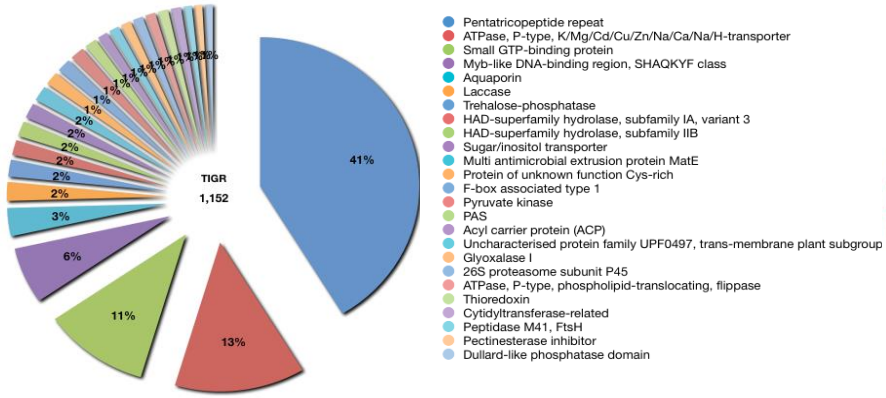
Numbers in the centre indicate the total number of domains identified for each annotation type. Details and links to InterProScan scanning methods and member databases can be obtained at ftp://ftp.ebi.ac.uk/pub/software/unix/iprscan/README.html#7).











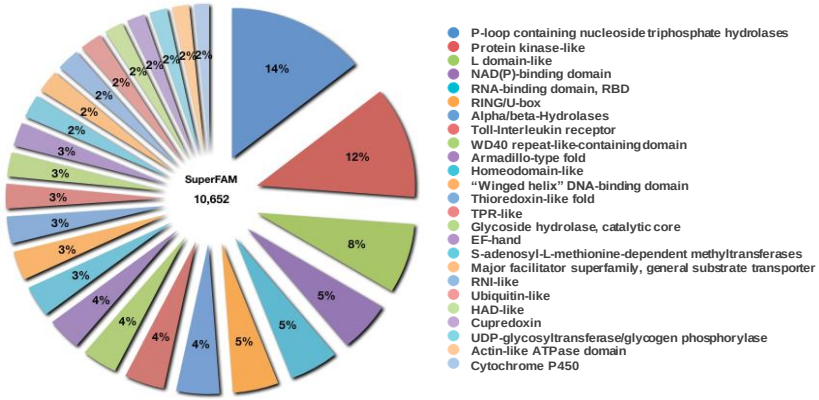
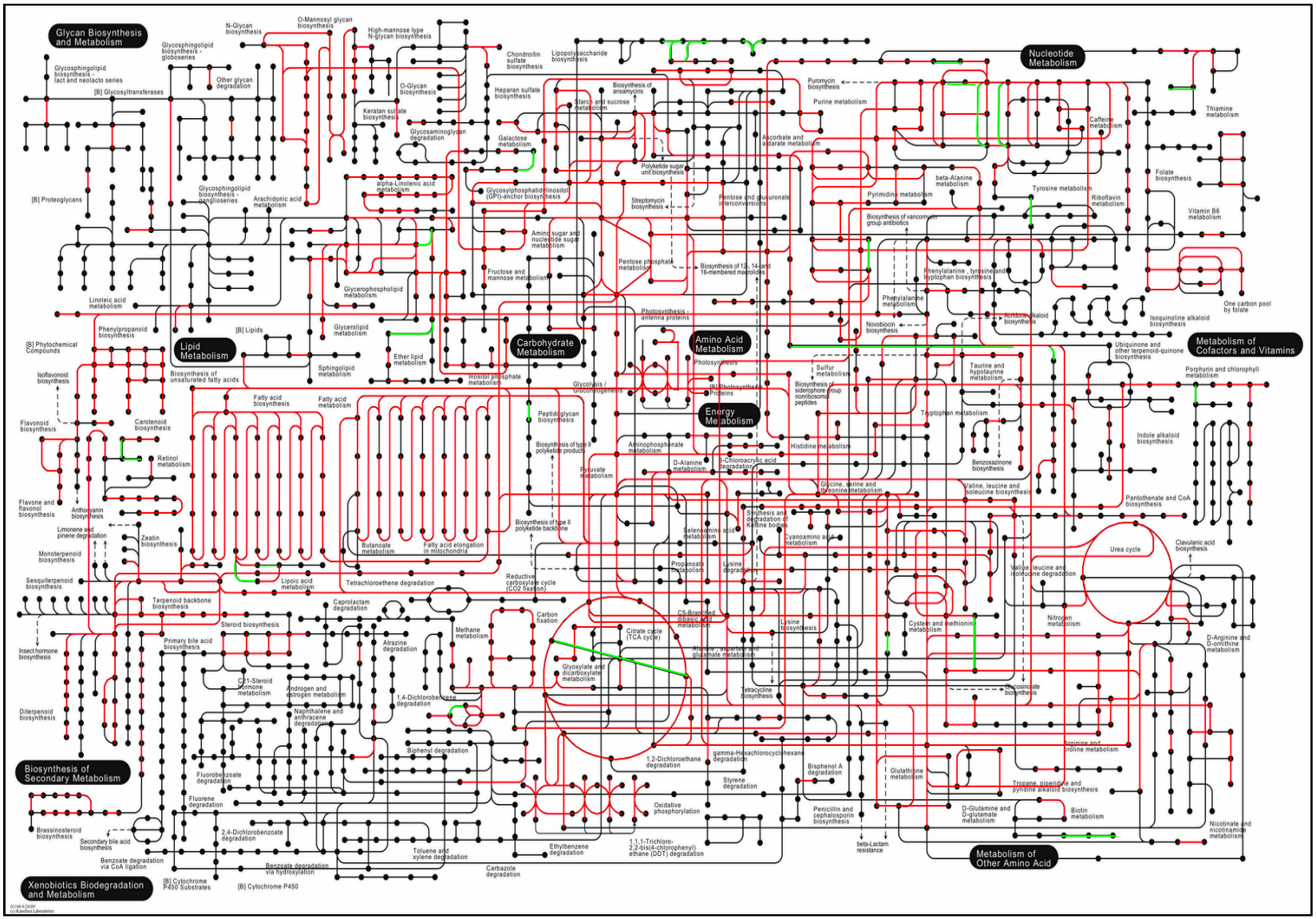


Fig. S2.8 Biochemical pathways represented in the *de novo* assembled gene catalog.

Arabidopsis accessions were obtained for all *Eucalyptus* genes with a significant BLAST hit (<1e^{-10}, minimum HSP of 100 bp), and plotted onto a biochemical pathways map using the KEGG resource (http://www.genome.jp/kegg/tool/color_pathway.html). Red edges indicate coverage of one or more genes in pathways represented in the *Eucalyptus* gene catalog that are shared with *Arabidopsis*, while green edges highlight the remaining *Arabidopsis* pathways not represented in the *Eucalyptus* gene catalog.



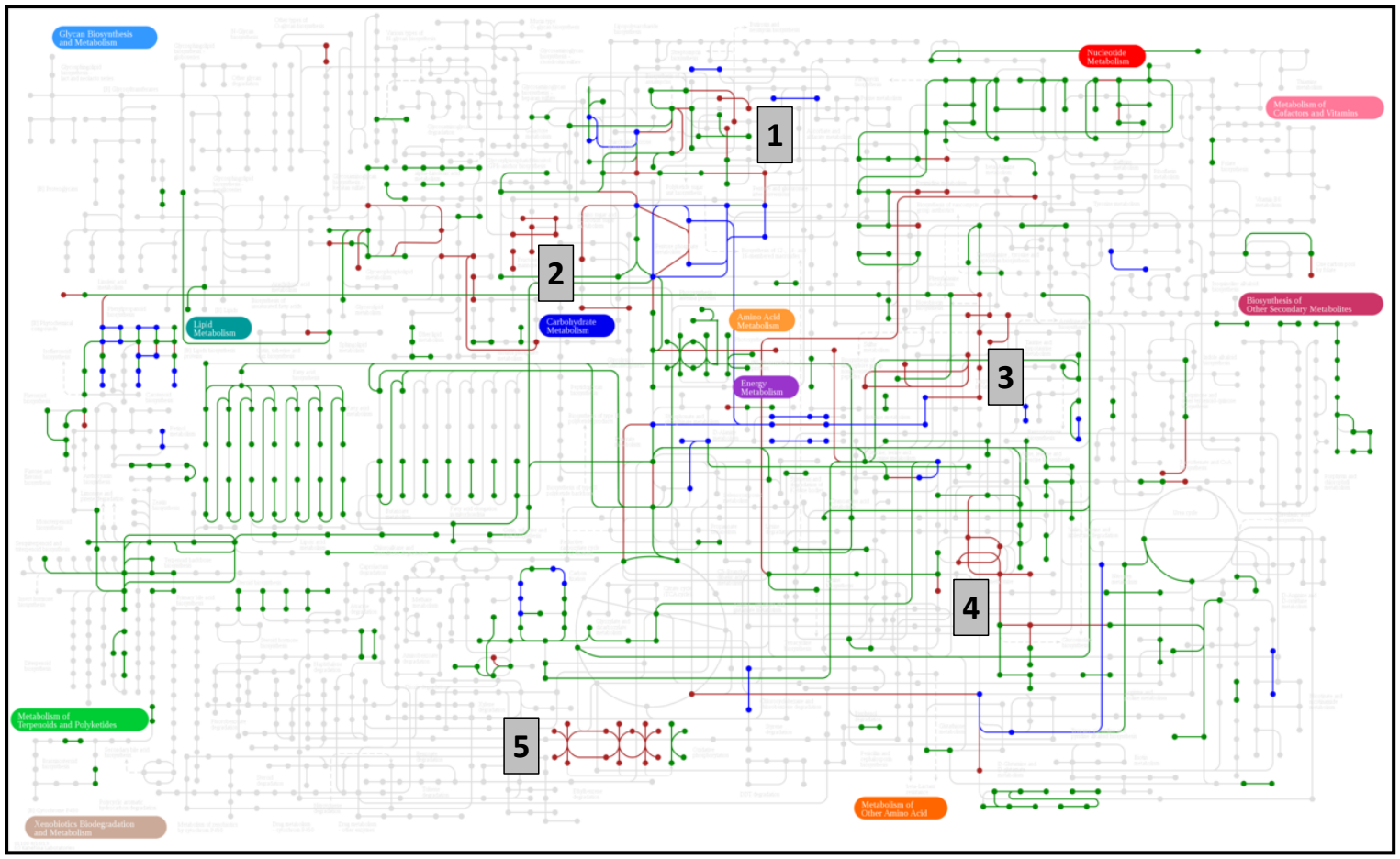


Fig. S2.9 Biochemical pathways represented by genes differentially expressed in xylem (red), leaf (green) or having members expressed differentially in xylem and leaf (blue).

Numbers indicate pathways where xylem-specific expression was predominant. 1. Starch and sucrose metabolism. 2. Fructose and mannose metabolism. 3. Shikimate pathway. 4. One carbon pool by folate. 5. Oxidative phosphorylation.

Fig. S2.10 MapMan overview of annotated genes that are differentially expressed in xylem (red), leaf (green) or genes that have members differentially expressed in both xylem and leaf (black).

Several bins including Glycolysis (4), Mitochondria (9), Cell wall (10), Signalling (30) and Cell (31) contained proportionally more genes that were preferentially expressed in xylem (Refere to Additional file 5 for detailed tables of genes in these categories.

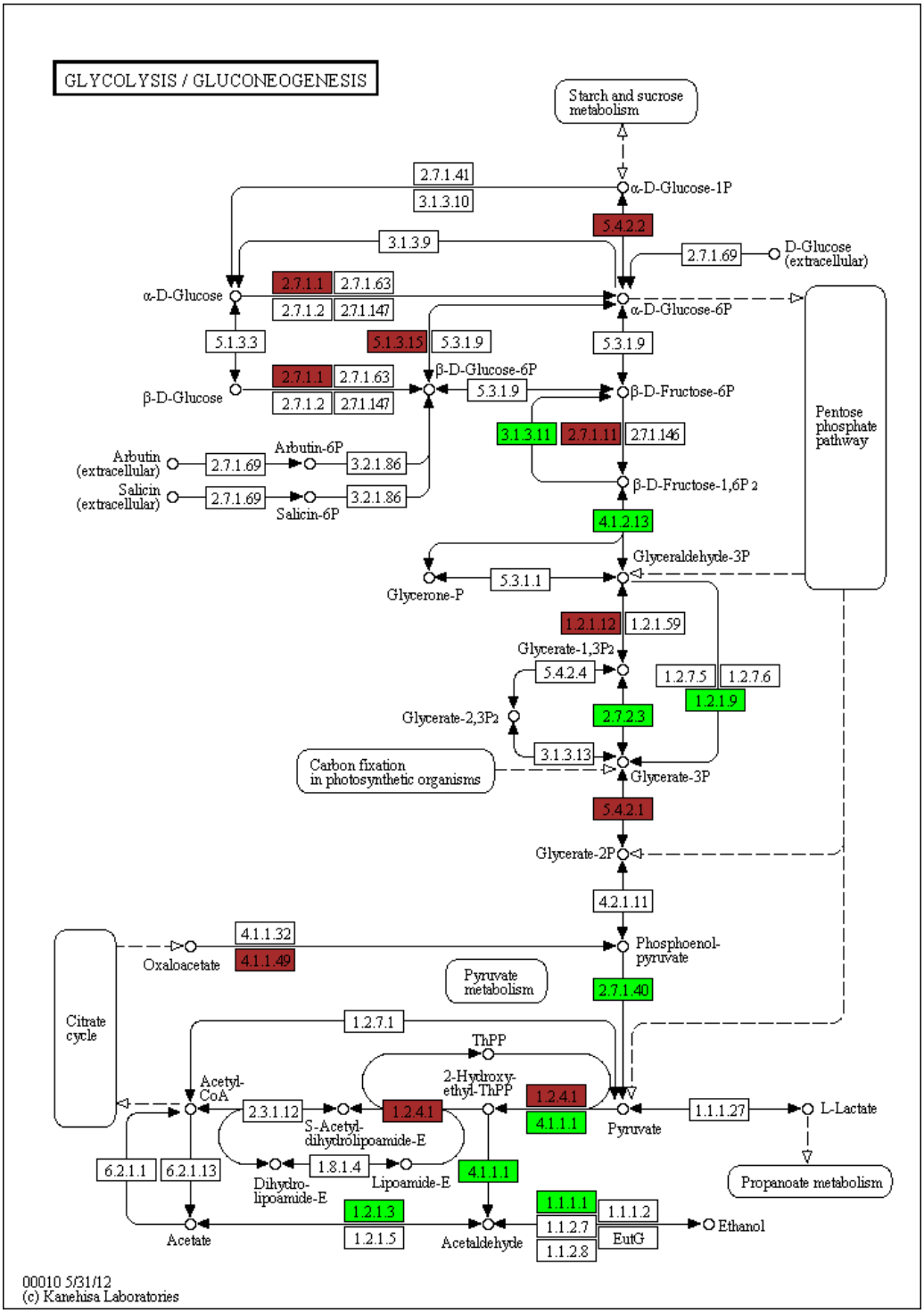


Fig. S2.11 Genes differentially expressed in xylem (red), leaf (green) involved in glycolysis/glucogenesis.

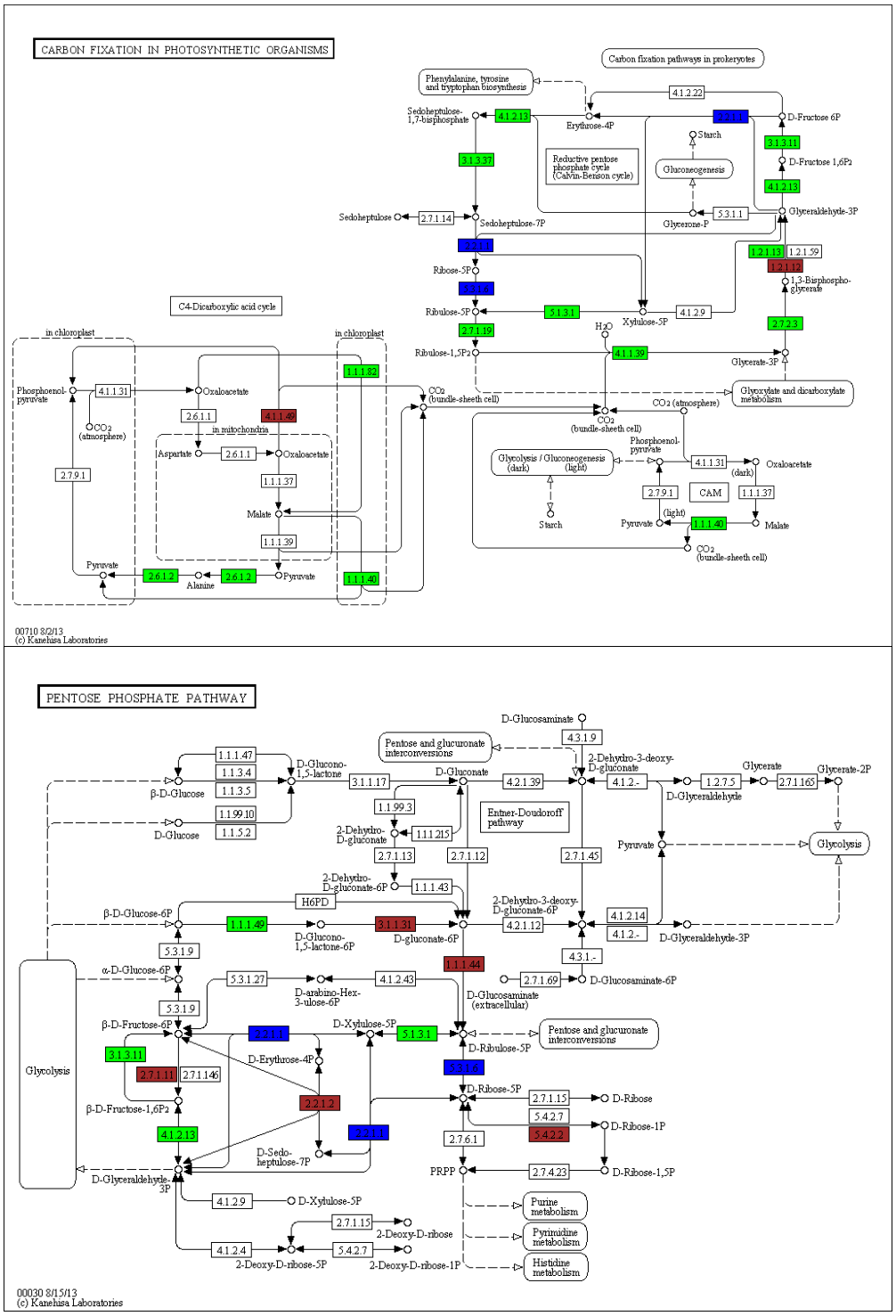


Fig. S2.12 Genes differentially expressed in xylem (red), leaf (green) or having members expressed differentially in xylem and leaf (blue) belonging to the pentose phosphate pathway. Two reactions, including EC 2.2.1.1. (Transketolase) and EC 5.3.1.6. (ribose 5-phosphate isomerase A) contained members that were both xylem specific and leaf specific.

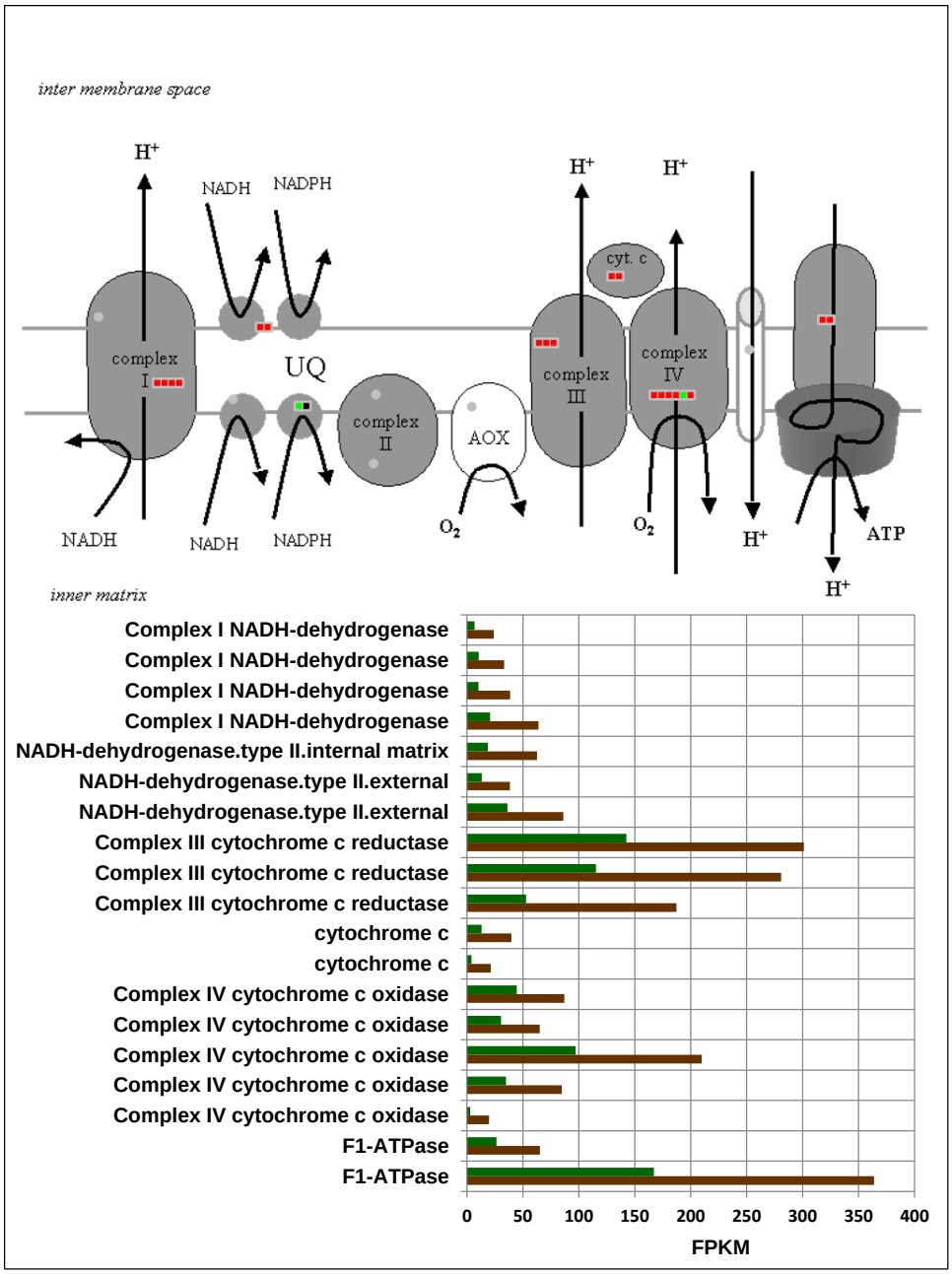


Fig. S2.13 Genes differentially expressed in xylem (red), leaf (green) or having members expressed differentially in xylem and leaf (black) involved in mitochondrial electron transport. Bar chart shows average of xylem and leaf expression for all xylem-specific genes.

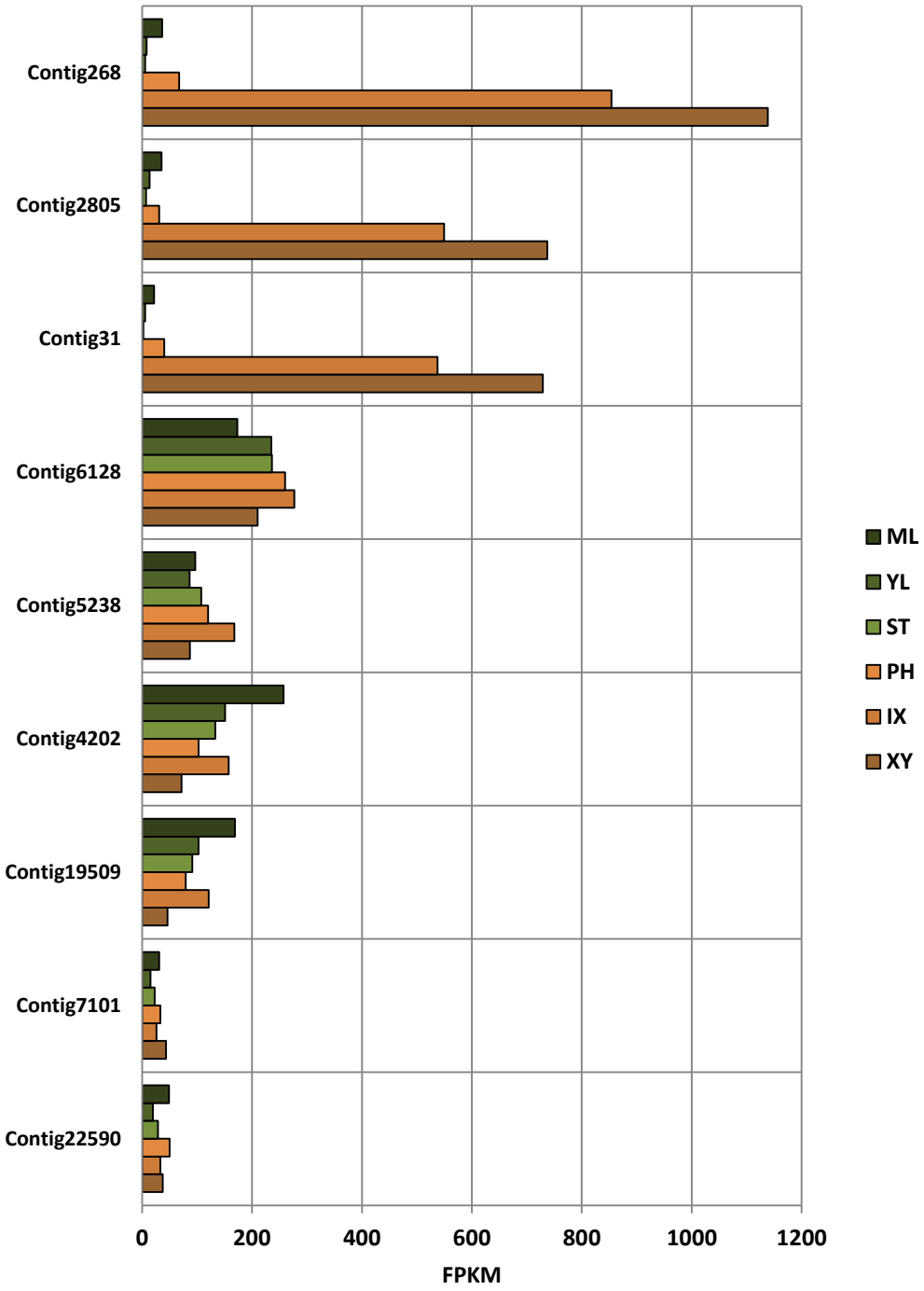


Fig. S2.14 Expression of *CesA* genes in *Eucalyptus*.

XY – xylem, IX – immature xylem, PH – phloem, ST – shoot tips, YL – young leaf, ML – mature leaf. Refer to Table S8 for contig IDs.

Table S2.1 Summary of filtered RNA-Seq data generated for *de novo* transcriptome assembly.

Fourteen RNA-Seq libraries were prepared and sequenced from RNA derived from six tissues of a *Eucalyptus grandis* x *E. urophylla* F1 hybrid clone and filtered to exclude low quality and ribosomal RNA-derived reads.

Tissue Type*	Dataset†	Reads	Read Length‡ (bp)	Total bp (Raw Data)
Xylem	ZMSR1	2,568,500	36-38	95,034,500
Xylem	ZMSR2	6,288,462	50-55	330,144,255
Immature Xylem	ZMSR3	2,228,286	36-38	82,446,582
Immature Xylem	ZMSR4	2,961,422	36-38	109,572,614
Immature Xylem	ZMSR5	3,243,376	50-55	170,277,240
Immature Xylem	ZMSR6	6,567,176	60-60	394,030,560
Immature Xylem	ZMSR7	6,799,600	60-60	407,976,000
Phloem	ZMSR8	6,875,592	50-55	360,968,580
Shoot Tips	ZMSR9	3,291.364	50-55	172,796,610
Shoot Tips	ZMSR10	8,263,698	60-60	495,821,88
Shoot Tips	ZMSR11	8,223,074	60-60	493,384,440
Young Leaf	ZMSR12	7,324,568	60-60	439,475,160
Young Leaf	ZMSR13	3,650,916	50-55	191,673,090
Mature Leaf	ZMSR14	3,466,122	50-55	181,971,405
TOTAL		**71,752,174**		**3,925,572,916**

*See METHODS section in main paper for sampling details. For FPKM calculations (Supplemental file SF3), six tissue-specific datasets were created by combining reads that were derived from the same tissue type.

†All raw data is available on NCBI SRA under accession SRA012408.

‡All sequencing was paired-end, with pairs ranging from 300-320 bp apart.

Table S2.2 Summary of *de novo* assembly statistics for different classes of annotated contigs.

	Subset of Reads	Number of Contigs	Median Contig Length (BP)	Median Coverage Per Base (CPB)	% Contigs Containing 'N's[†]	Median Length of Contigs Containing 'N's (bp)	Median % Ns in Contigs Containing 'N's	Median CPB in contigs containing 'N's
TOTAL DATASET OF ASSEMBLED CONTIGS	**With CDS***	15,713	1,090	44X	42.26%	1,369	1.89%	44X
	Without CDS	3,181	333	20X	41.62%	364	15.12%	19X
Matching At/Pt/Vv proteins	**With CDS**	13,806	1,200	47X	43.20%	1,453	1.79%	19X
	Without CDS	1,249	374	20X	40.72%	429	12.17%	20X
Not matching Angiosperm proteins but Matching *Eucalyptus* genome[‡]	**With CDS**	1,813	512	31X	35.21%	684	3.88%	33X
	Without CDS	1,738	326	20X	39.01%	357	15.03%	19X
Not Matching *Eucalyptus* genome but Matching NR	**With CDS**	14	539	35X	42.86%	680	3.49%	37X
	Without CDS	1	535	27X	0.00%	NA	NA	NA
Not matching *Eucalyptus* genome or NR	**With CDS**	80	321	26X	40.00%	336	16.94%	25X
	Without CDS	193	275	15X	41.45%	278	25.47%	17X

*CDS predicted by GenScan analysis (Burge and Karlin 1997)
[†]Any contig containing at least 1 'N' in its sequence was counted.
[‡]Draft 8X assembly, (http://eucalyptusdb.bi.up.ac.za/)
At – Arabidopsis thaliana, *Pt – Populus trichocarpa*, *Vv – Vitis vinifera.* A positive "match" indicates a positive BLAST hit at $<1e^{-10}$ and HSP of at least 100 bp in length.

Table S2.3 Quality assessment of assembled contigs by homology to *Arabidopsis thaliana*.

		EucAll						Velvet-Assembled Contigs					
		>200bp	**>300bp**	**>500bp**	**>1,000bp**	**>2,000bp**	**>3,000bp**	**>200bp**	**>300bp**	**>500bp**	**>1,000bp**	**>2,000bp**	**>3,000bp**
Arabidopsis	**$1e^{-05}$**	27,939	27,396	25,593	17,245	2,002	199	26,854	26,020	24,512	18,516	6,862	2,177
	$1e^{-10}$	26,587	26,202	24,662	16,903	1,940	199	25,538	24,757	23,390	17,744	6,602	2,114
	$1e^{-20}$	24,302	24,129	23,093	16,279	1,865	191	23,242	22,545	21,485	16,569	6,185	1,978

The complete gene catalog from *Arabidopsis thaliana* (TAIR 9) was used to query the EucALL dataset and the Velvet assembled dataset containing 18,894 contigs from this study. Significant BLAST hits were counted in incrementing contig size classes from the target datasets. In the Velvet-assembled dataset, the number of significant hits was 3.4 times higher in size categories greater than 2000 bp in length, indicating that the Velvet assembled contigs contained more full-length gene models than current publicly available coding sequence for *Eucalyptus*.

Table S2.4 List of 43 transcript-derived contigs homologous to 33 of the 52 "Core xylem genes" identified by Ko et al. (2006) in *Arabidopsis* and their relative xylem to leaf FPKM ratio.

In most cases, the expression profiles of these genes were highly positively correlated with that of *EgCesA1* (Contig268), a secondary cell wall-specific cellulose synthase gene (Ranik and Myburg, 2006).

At accession	Description	Contig*	*EgCesA1* Expression correlation†	Xylem/Leaf FPKM ratio‡
AT1G09610	Unknown protein	contig368	1.000	362.027
AT3G15050	IQD10	contig10671	0.999	No Leaf Expression Detected
AT5G03170	FLA11	contig2707	0.981	122.993
AT1G27440	IRX10	contig3811	0.999	126.595
AT4G18780	AtCesA8	contig268	1.000	57.736
AT5G17420	AtCesA7	contig31	1.000	62.810
AT1G22480	Plastocyanin-like domain-containing protein	contig6482	0.891	68.784
AT5G67210	unknown protein	contig3195	0.967	51.224
AT3G62020	GLP10	contig25018	0.937	36.847
AT2G03200	Aspartyl protease family protein	contig453	0.988	42.946
AT2G37090	IRX9	contig5622	0.994	47.563
AT5G44030	AtCesA4	contig2805	1.000	31.625
AT5G54690	IRX8	contig1569	0.98	34.588
AT4G28380	Leucine-rich repeat family protein	contig29940	0.794	221.689
AT5G01360	Unknown protein	contig8107	0.931	No Leaf Expression Detected
AT5G15630	IRX6	contig1665	0.99	32.632
AT5G67210	Unknown protein	contig5930	0.993	23.401
AT4G17220	MAP70-5	contig7003	0.666	No Leaf Expression Detected
AT4G22680	MYB85	contig3124	0.749	44.273
AT5G40020	Pathogenesis-related thaumatin family protein	contig22035	0.851	18.973
AT1G27920	MAP65-8	contig3070	0.891	16.325
AT1G63910	MYB103	contig16135	0.886	No Leaf Expression Detected
AT4G27435	Unknown protein	contig949	0.981	10.151
AT2G46770	NST1	contig44541	0.918	78.270
AT5G60720	Unknown protein	contig7972	0.935	17.281

AT3G18660	PGSIP1	contig19436	0.959	20.325
AT5G03170	FLA11	contig3257	0.79	No Leaf Expression Detected
AT5G01360	Unknown protein	contig5954	0.928	14.072
AT1G79620	Leucine-rich repeat transmembrane protein kinase, putative	contig53943	0.932	8.553
AT1G31720	Unknown protein	contig24491	0.922	11.887
AT4G28500	SND2	contig2382	0.935	11.174
AT4G33330	PGSIP3	contig14715	0.94	9.997
AT1G24030	Protein kinase family protein	contig20741	0.704	29.393
AT1G19300	PARVUS	contig26681	0.964	6.764
AT1G09440	Protein kinase family protein	contig6469	0.969	10.219
AT5G61340	Unknown protein	contig41480	0.623	No Leaf Expression Detected
AT1G66230	MYB20	contig10425	0.822	5.910
AT4G28500	SND2	contig21083	0.577	13.925
AT2G46770	NST1	contig21412	0.276	33.989
AT1G33800	Unknown protein	contig4765	0.734	2.778
AT1G80170	polygalacturonase, putative / pectinase, putative	contig2977	-0.469	1.002
AT5G67210	Unknown protein	contig53083	-0.85	0.130
AT5G05390	LAC12	contig92451	-0.624	No Xylem Expression Detected

* Contig (node) numbers originally assigned by Velvet during the short-read assembly. The complete list of transcript-derived contigs is available in Additional file 3.

† Correlation of digital expression profile to that of *EgCesA1* (Contig268 - orthologous to *AtCesA8*), a secondary cell wall associated cellulose synthase gene.

‡ Ratio of the average FPKM value for xylem and immature xylem to the average for shoot tips, young leaves and mature leaves. Cases where no xylem or no leaf expression were detected (average FPKM = 0) are indicated.

Table S2.5 **MapMan bin allocations and genes involved in glycolysis, pentose phosphate pathway and mitochondrial electron transport.**

BinCode	BinName	id	description	Specificity*
Glycolysis				
4.1.2	glycolysis.cytosolic branch.UGPase	at5g17310	AtUGP2	xylem
4.1.3	glycolysis.cytosolic branch.phosphoglucomutase (PGM)	at1g23190	PGM3	xylem
4.1.5	glycolysis.cytosolic branch.phosphofructokinase (PFK)	at4g26270	PFK3	xylem
4.1.6	glycolysis.cytosolic branch.pyrophosphate-fructose-6-P phosphotransferase	at1g76550		xylem
4.1.9	glycolysis.cytosolic branch.glyceraldehyde 3-phosphate dehydrogenase (GAP-DH)	at1g13440	GAPC-2	xylem
4.1.10	glycolysis.cytosolic branch.non-phosphorylating glyceraldehyde 3-phosphate dehydrogenase (NPGAP-DH)	at2g24270	ALDH11A3	leaf
4.1.13	glycolysis.cytosolic branch.phosphoglycerate mutase	at3g08590	iPGAM2	xylem
4.1.15	glycolysis.cytosolic branch.pyruvate kinase (PK)	at3g52990		xylem
4.1.17	glycolysis.cytosolic branch.phospho-enol-pyruvate carboxylase kinase (PPCK)	at1g08650	ATPPCK1	xylem
4.2.3	glycolysis.plastid branch.phosphoglucomutase (PGM)	at1g70820		xylem
4.2.6	glycolysis.plastid branch.pyrophosphate-fructose-6-P phosphotransferase	at1g12000		xylem
4.2.15	glycolysis.plastid branch.pyruvate kinase (PK)	at3g22960	PKP-ALPHA	leaf
4.2.15	glycolysis.plastid branch.pyruvate kinase (PK)	at1g32440	PKp3	leaf
4.3.13	glycolysis.unclear/dually targeted.phosphoglycerate mutase	at3g50520		leaf
4.3.13	glycolysis.unclear/dually targeted.phosphoglycerate mutase	at5g22620		leaf
Oxidative Pentose Phosphate Pathway				
7.1.1	OPP.oxidative PP.G6PD	at5g35790	G6PD1	leaf
7.1.1	OPP.oxidative PP.G6PD	at5g40760	G6PD6	xylem
7.1.2	OPP.oxidative PP.6-phosphogluconolactonase	at5g24400	EMB2024	xylem
7.1.3	OPP.oxidative PP.6-phosphogluconate dehydrogenase	at3g02360		xylem
Non-reductive Pentose phosphate pathway				
7.2.1	OPP.non-reductive PP.transketolase	at2g45290		xylem and leaf
7.2.2	OPP.non-reductive PP.transaldolase	at5g13420	TRA2	xylem
7.2.4	OPP.non-reductive PP.ribose 5-phosphate isomerase	at3g04790	EMB3119	xylem and leaf
Mitochondrial electron transport				
9.1.2	mitochondrial electron transport / ATP synthesis.NADH-DH.localisation not clear	atmg00665	NAD5	xylem
9.1.2	mitochondrial electron transport / ATP synthesis.NADH-DH.localisation not clear	atmg01320	NAD2	xylem
9.1.2	mitochondrial electron transport / ATP synthesis.NADH-DH.localisation not clear	at3g03100		xylem
9.1.2	mitochondrial electron transport / ATP synthesis.NADH-DH.localisation not clear	atmg00580	NAD4	xylem
9.2.1	mitochondrial electron transport / ATP synthesis.NADH-DH.type II.internal matrix	at1g07180	ATNDI1	leaf

9.2.1	mitochondrial electron transport / ATP synthesis.NADH-DH.type II.internal matrix	at2g29990	NDA2	xylem and leaf
9.2.2	mitochondrial electron transport / ATP synthesis.NADH-DH.type II.external	at4g05020	NDB2	xylem
9.2.2	mitochondrial electron transport / ATP synthesis.NADH-DH.type II.external	at4g28220	NDB1	xylem
9.5	mitochondrial electron transport / ATP synthesis.cytochrome c reductase	at3g10860		xylem
9.5	mitochondrial electron transport / ATP synthesis.cytochrome c reductase	at1g15120		xylem
9.5	mitochondrial electron transport / ATP synthesis.cytochrome c reductase	at3g52730		xylem
9.6	mitochondrial electron transport / ATP synthesis.cytochrome c	atmg00110	ABCI2	xylem
9.6	mitochondrial electron transport / ATP synthesis.cytochrome c	at3g51790	AtCCME	xylem
9.7	mitochondrial electron transport / ATP synthesis.cytochrome c oxidase	at5g40382		xylem
9.7	mitochondrial electron transport / ATP synthesis.cytochrome c oxidase	atmg00160	COX2	xylem
9.7	mitochondrial electron transport / ATP synthesis.cytochrome c oxidase	atmg01360	COX1	xylem
9.7	mitochondrial electron transport / ATP synthesis.cytochrome c oxidase	atmg00730	COX3	xylem
9.7	mitochondrial electron transport / ATP synthesis.cytochrome c oxidase	at1g28140		leaf
9.7	mitochondrial electron transport / ATP synthesis.cytochrome c oxidase	at4g37830		xylem
9.9	mitochondrial electron transport / ATP synthesis.F1-ATPase	atmg01190	ATP1	xylem
9.9	mitochondrial electron transport / ATP synthesis.F1-ATPase	atmg01080	ATP9	xylem

*preferential expression in xylem, leaf, or having both xylem- and leaf-specific members

Table S2.6 Summary statistics of expression levels (FPKM) for all contigs with FPKM>1 for each tissue sampled.

XY – xylem, IX – immature xylem, PH – phloem, ST – shoot tips, YL – young leaf, ML – mature leaf.

FPKM	XY	IX	PH	ST	YL	ML
Min	4	1	3	1	2	5
Q1	15	7	13	10	10	16
median	24	15	22	18	19	25
Q3	48	36	47	38	37	45
90th percentile	111	96	118	82	82	93
95th percentile	206	175	219	139	142	154
99th percentile	665	643	732	401	431	462
Max	53,656	40,622	45,913	23,068	41,372	75,728

Table S2.7 Contigs matching *Eucalyptus* cellulose synthase (*CesA*) genes.

Eucalyptus ("EgCesA") gene names and *Arabidopsis* ("AtCesA") homolog nomenclature according to (Ranik & Myburg, 2006) and TAIR (www.Arabidopsis.org).

Contig ID	CesA gene
Contig 40455	EgCesA5 (AtCesA1)
Contig 22590	EgCesA4 (AtCesA3)
Contig 7101	EgCesA5 (AtCesA1)
Contig 19509	EgCesA6 (AtCesA2/5/6/9)
Contig 4202	EgCesA5 (AtCesA1)
Contig 5238	EgCesA4 (AtCesA3)
Contig 6128	EgCesA6 (AtCesA2/5/6/9)
Contig 31	EgCesA3 (AtCesA7)
Contig 2805	EgCesA2 (AtCesA4)
Contig 268	EgCesA1 (AtCesA8)

References

Burge, C. and S. Karlin. 1997. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* **268:** 78-94.

Ko JH, Beers EP, Han KH. 2006. Global comparative transcriptome analysis identifies gene network regulating secondary xylem development in *Arabidopsis thaliana*. *Molecular Genetics and Genomics* **276:** 517-531.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078-2079.

Pinheiro, M., V. Afreixo, G. Moura, A. Freitas, M.A.S. Santos, and J.L. Oliveira. 2006. Statistical, computational and visualization methodologies to unveil gene primary structure features. *Methods of Information in Medicine* **45:** 163-168.

Ranik M and Myburg AA. 2006. Six new cellulose synthase genes from *Eucalyptus* are associated with primary and secondary cell wall biosynthesis. *Tree Physiology* **26:** 545-556.

Trapnell, C., B.A. Williams, G. Pertea, A.M. Mortazavi, G. Kwan, M.J. van Baren, S.L. Salzberg, B. Wold, and L. Pachter. Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. *Nature Biotechnology* **28**(5): 511.

Zerbino, D.R. and E. Birney. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* **18:** 821-829.

CHAPTER 3

The physiology of tension wood formation in *Eucalyptus*

Eshchar Mizrachi[1], Victoria J Maloney[3], Janine Silberbauer[1], Charles A Hefer[1], David K Berger[2], Shawn D Mansfield[3] and Alexander A Myburg[1*]

[1] Department of Genetics, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Private bag X20, Pretoria, 0028, South Africa. [2] Department of Plant Science, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, 0002, South Africa. [3] 4030-2424 Main Mall, Department of Wood Science, University of British Columbia, Vancouver, BC, Canada V6T 1Z4.

This chapter has been prepared in the format of a manuscript for a peer-reviewed research journal and is in preparation for submission. I drafted the manuscript, helped sample the material for transcriptome and wood phenotyping analysis, analysed the transcriptome data and the phenotypic data. Victoria Maloney performed all of the wood property analyses and drafted parts of the manuscript (sections 2.3.2-2.3.5). Janine Silberbauer helped sample the material and performed the RNA extraction and quality control. Charles Hefer contributed to RNA-seq data analysis. Alexander Myburg conceived of the study, and supervised the work. David Berger, Shawn Mansfield and Alexander Myburg helped draft and edit the manuscript.

3.1 Summary

- Tension wood has distinct physical and chemical properties, including changes in fibre properties, altered cellulose chemistry and ultrastructure, a reduction in lignin and a change in hemicellulose composition. For functional studies of secondary cell wall biosynthesis and wood formation, it serves as a good system for investigating the underlying genetic regulation of these processes. The reference genome sequence for *Eucalyptus* allows investigation of the global transcriptional reprogramming that accompanies tension wood formation in this tree.
- Here we report the first comprehensive analysis of physicochemical wood property changes in tension wood of *Eucalyptus* measured in a hybrid (*E. grandis* × *E. urophylla*) clone, as well as genome-wide gene expression changes in xylem tissues three weeks post-induction using RNA-sequencing.
- We find that *Eucalyptus* tension wood is characterized by an increase in glucose and alpha cellulose, a reduction in lignin, xylose and mannose, and a marked increase in galactose. Gene expression profiling in tension wood forming tissue showed downregulation of monolignol biosynthetic genes, as well as differential expression of several carbohydrate active enzymes.
- Our results allow us to hypothesize the roles of several carbohydrate and lignin biosynthetic genes and transcription factors involved in *Eucalyptus* tension wood formation. We conclude that based on at the level of transcript abundance, the alterations of cell wall traits induced by tension wood formation in *Eucalyptus* are a consequence of a combination of downregulation of lignin biosynthesis and hemicellulose remodelling, rather than the traditionally proposed upregulation of the cellulose biosynthetic pathways.

3.2 Introduction

Tension wood formation is a dicot-specific physiological reaction to mechanical or gravimetric stress on the tree. Industrially, the formation of tension wood is important as it alters wood composition, structure and homogeneity for downstream processing. The biology of tension wood is still not fully understood, but evidence suggests transcriptional and metabolic reprogramming contributes significantly to the altered structure and composition. Since tension wood formation is associated with a trend towards decreased lignin and increased cell wall polysaccharides (Al-Haddad *et al.*, 2013), it presents a relevant model to investigate the molecular underpinnings of carbon allocation and carbohydrate deposition in wood. While most genome-wide gene expression and wood physicochemical analyses have thus far been performed in *Populus* species and hybrids (Andersson-Gunnerås *et al.*, 2006), the availability of a genome from a second forest tree species, *Eucalyptus grandis* (Myburg et al., in preparation), facilitates comparative studies to better understand the biology of tension wood.

Tension wood possesses distinct physical and chemical properties. Reported physical changes in tension wood include longer vessels (Jourez *et al.*, 2001), longer and thinner fibres (Yoshizawa *et al.*, 2000; Jourez *et al.*, 2001) - possessing a thicker cell wall and relatively smaller lumen, and a higher fibre/vessel ratio compared to normal or opposite wood (Jourez *et al.*, 2001; Ruelle *et al.*, 2006). The vessels and fibres are also typically more compact in tension wood, such that the middle lamellae possess a smaller surface area (Bowling & Vaughn, 2008). Within the secondary cell walls (SCW) of tension wood fibres, there are marked differences in the cellulose properties. Typically, α-cellulose is increased in tension wood (Côté Jr *et al.*, 1969; Okuyama *et al.*, 1994; Yoshizawa *et al.*, 2000), the cellulose is more crystalline (Okuyama *et al.*, 1994; Müller *et al.*, 2006), and has a marked decrease in microfibril angle (MFA) (Okuyama *et al.*, 1994; Washusen *et al.*, 2005; Ruelle *et al.*, 2006; Ruelle *et al.*, 2010; Clair *et al.*, 2011).

In some cases, the traditional 3-layered (S_1, S_2, S_3) secondary cell wall layers are often reported to be replaced by a gelatinous layer (G-Layer), although this is not obligatory, and varies between and within angiosperm species (Washusen *et al.*, 2003; Clair *et al.*, 2006b; Qiu *et al.*, 2008; Ruelle *et al.*, 2010). It is also debatable whether the G-layer is the causal agent of generating the tension (Okuyama *et al.*, 1994; Yamamoto, 2004; Fang *et al.*, 2007; Fang *et al.*, 2008; Goswami *et al.*, 2008) or a physiological by-product of xylem formation in some species (Washusen *et al.*, 2003; Clair *et al.*, 2006a; Qiu *et al.*, 2008). However, in all cases tension wood is characterized by an increase in cell wall glucose in the form of cellulose, a decrease in lignin (Bentum *et al.*, 1969; Okuyama *et al.*, 1994; Aoyama *et al.*, 2001; Yoshida *et al.*, 2002), and an increase in the syringyl:guaiacyl (S:G) ratio of the lignin (Aoyama *et al.*, 2001; Yoshida *et al.*, 2002; Joseleau *et al.*, 2004). Similarly, the hemicellulose composition changes, and may be variable, especially with the presence/absence of a G-layer. Previous reports have highlighted the putative roles of xyloglucan (Nishikubo *et al.*, 2007; Mellerowicz *et al.*, 2008; Baba *et al.*, 2009), pectinacious compounds (Andersson-Gunnerås *et al.*, 2006; Goulao *et al.*, 2011) such as Rhamnogalacturonan I (RG I) (Bowling & Vaughn, 2008) and other galactose-containing polysaccharides such as those found in arabinogalactan proteins (AGPs) (Lafarguette *et al.*, 2004; Andersson-Gunnerås *et al.*, 2006; Bowling & Vaughn, 2008) in generating or facilitating tension wood formation. The orientation of cellulose microfibrils has been suggested to be influenced by galactan, which (either as a component of AGPs or high molecular weight galactan) has been implicated in cellulose orientation during S_2 deposition in other G-layer rich physiology, such as flax phloem fibres (Gorshkova & Morvan, 2006; Roach *et al.*, 2011).

The most detailed study of the physiological and molecular responses to tension wood to date has been performed by Andersson-Gunnerås *et al*. (2006), who examined transcriptional and metabolomic

responses of G-layer forming tension wood in *Populus tremula* three weeks post-tension wood induction. Primarily focusing on carbohydrate metabolism, these authors showed that while the expression of cellulose synthase genes was not necessarily affected at the transcriptional level, the differential regulation of genes belonging to several key pathways were indicative of the change in carbon allocation favouring cellulose biosynthesis over other cell wall moieties. For example, significant decreases in GDP sugar channelling to mannan biosynthesis and the pentose phosphate pathway were observed (Andersson-Gunnerås *et al.*, 2006).

Several studies have also consistently found evidence of key hormone signalling pathways affected in tension wood forming tissue, such as the activation of ethylene mediated pathways (Andersson-Gunnerås *et al.*, 2003; Andersson-Gunnerås *et al.*, 2006; Vahala *et al.*, 2013). Ethylene is known to be an inducer of cambial growth (Love *et al.*, 2009), and has been recently shown in *Zinnia elegans* to be mass produced in late maturing tracheary elements (TEs) and diffuse in a paracrine fashion (i.e. influencing any immediately surrounding cells) to modulate additional TE differentiation from the cambium, coordinating both axial and radial vascular development (Pesquet & Tuominen, 2011). Auxin maintains cambial initials in an undifferentiated form, while polar auxin transport and localized auxin suppression is associated with cambial differentiation (Moyle *et al.*, 2002; Ko *et al.*, 2004; Baba *et al.*, 2011). Gibberellins have also been shown to act synergistically with auxin to promote cambial differentiation and fibre elongation (Little & Savidge, 1987), and the production of gibberellic acid (GA) and auxins induce similar responses at the transcriptional level (Björklund *et al.*, 2007). The mechanism of synergy is likely through GA's influence on the polar transport of auxin (reviewed in Elo *et al.*, 2009). Several hormones, and particularly GA, have been linked to cortical microtubule arrangement during cellulose deposition, indirectly influencing cellulose properties (Lloyd, 2011). For example, recently jasmonic acid (JA) signalling has been linked with cells under tension; showing the upregulation of the mechano-inducible *JAZ10* gene in the interfascicular fibres of *Arabidopsis thaliana* (Sehr *et al.*, 2010). JA

signalling was shown to stimulate secondary growth of cambial initials (Sehr *et al.*, 2010), although this aspect has not yet been adequately studied in a woody species.

Despite the extensive research on tension wood formation, much has not been resolved about the physicochemical changes in tension wood in *Eucalyptus*, and it is not known whether the changes in tension wood would result in similar phenotype as that observed in other angiosperms, such as *Populus*. Previously, Paux *et al.* (2005) and Qiu *et al.* (2008) reported on the variation in gene expression profiles of 231 genes and a 4900 probe microarray, respectively, during tension wood formation in *Eucalyptus*. More recently, a focused study of 38 hemicellulose and pectin modifying candidate genes was also performed in *Eucalyptus globulus* (Goulao *et al.*, 2011). Considering the extent of gene duplication and the roles of potential paralogs, it is crucial to obtain a transcriptome-wide view of transcriptional response, as differential expression of individual paralogs may be misleading. Since xylogenesis is known to be regulated to a large extent at a transcriptional level, it would be interesting to see if tension wood formation results in similar transcriptional reprogramming that explain the changes in phenotype, and may highlight important genes or regulatory elements that have not previously been identified. This could be beneficial to industrial applications such as pulp, paper, timber and biofuel production, as they would offer targets for selection in breeding programs or candidates for genetic manipulation.

In this study we aimed to investigate the physical effects observed in mature tension wood forming tissue of a widely grown hybrid *Eucalyptus* genotype. We further aimed to provide a detailed, whole-transcriptome characterization of *Eucalyptus* tension wood forming tissue at three weeks post induction by mRNA-sequencing to investigate the transcriptional reprogramming that occurs during tension wood formation. We hypothesized that the transcriptional response in *Eucalyptus* should reflect the rapid differentiation of longer, thinner fibre cells, which should be accompanied by evidence of a flux of auxin

mediated pathways and the upregulation of ethylene and GA mediated pathways, as well as an increase in pectic degredation, rapid cell elongation and altered programmed cell death, to reflect this fibre phenotype.

3.3 Materials and methods

3.3.1 Sampling for wood property analysis

To characterize the physicochemical properties of tension wood in *Eucalyptus* trees, we collected basal sections of naturally leaning branches from five different ramets of the same 3-4 year-old F_1 *E. grandis* × *E. urophylla* hybrid clone (GUSAP1, Sappi Forest Research, KwaMbonambi, South Africa, Fig. S3.1). A section of the leaning stems of each of the five trees was analysed where the angle was at 45° in relation to the main trunk of the tree. For each tree, sections were compared between the side closer to the main trunk (top of the branch, tension wood forming) and the side opposite to that (opposite wood forming). This provided a sample of woody material produced by stable tension wood formation (*i.e.* not recently induced).

3.3.2 Klason lignin determination

Wood was ground in a Wiley mill to pass a 0.4 mm screen (40 mesh) and Soxhlet extracted overnight in hot acetone to remove extractives. Lignin and carbohydrate contents were determined with a modified Klason (Coleman *et al.*, 2009) method in which extracted ground stem tissue (100 mg) was treated with 3 mL of 72% H_2SO_4 and stirred every 10 min for 2 h. Samples were then diluted with 112 mL deionized (DI) water and autoclaved for 1 h at 121°C. The acid-insoluble lignin fraction was determined gravimetrically by filtration through a pre-weighed medium coarseness sintered-glass crucible, while the acid-soluble lignin component was determined spectrophotometrically by absorbance at 205 nm. Carbohydrate contents were determined by using anion exchange high-performance liquid

chromatography (Dx-600; Dionex, Sunnyvale, CA, USA) equipped with an ion exchange PA1 (Dionex) column, a pulsed amperometric detector with a gold electrode, and a SpectraAS3500 auto injector (Spectra-Physics).

3.3.3 α-cellulose content determination

Holocellulose was determined using a modified version of Browning (1967). Briefly, 200 mg of extracted ground wood was de-lignified by adding 3.5 mL buffer solution (60 mL of glacial acetic acid + 1.3 g NaOH/L) and 1.5 mL of 20% sodium chlorite solution (20 g $NaClO_2$ in 80 mL distilled water) then gently shaken at 50°C overnight (14-16 hrs). The following day the reaction was quenched by placing it into an ice bath and incubating it at 4ºC for several hours before the reaction solution was removed and a second reaction was performed overnight. Finally, the reacted wood meal was transferred and washed twice with 50 mL of 1% glacial acetic acid (under suction), followed by a wash with 10 mL acetone on a pre-weighed coarse sintered-glass crucible. The resulting holocellulose was permitted to dry in a 50°C oven overnight and percent of total (extracted ground wood) was determined gravimetrically. α-cellulose content was then determined by extracting 80 mg of the oven dried holocellulose with 4 mL of 17.5% sodium hydroxide for 30 min at room temperature then adding 4 mL water stirring for 1 min and leaving it to react for another 29 min. The reaction solution was then filtered through a pre-weighed coarse sintered-glass crucible, washed with DI water (3 x 50 mL), soaked in 1.0 M acetic acid for 5 minutes and washed again with DI water (3 x 50 mL). Finally, the samples were dried at 50ºC overnight and percent of total holocellulose was determined gravimetrically (Yokoyama *et al.*, 2002).

3.3.4 Microfibril angle determination

Microfibril angle estimates were generated by X-ray diffraction (Megraw *et al.*, 1998). The 002 diffraction spectra were screened for T value distribution and symmetry on a Bruker D8 discover X-ray

diffraction unit equipped with an area array detector (GADDS). Wide-angle diffraction was used in the transmission mode, and the measurements were performed with CuKα1 radiation (λ=1.54 Å). The X-ray source was fit with a 0.5-mm collimator, and the scattered photons were collected by a GADDS detector. Both the X-ray source and detector were set to theta=0°.

3.3.5 Calcofluor staining for cellulose

Samples were radially cut into 20 µm cross sections using a Leica SM2000r hand sliding microtome (Leica Microsystems, Wetzlar, Germany) and stored in dH_2O until needed. Sections were treated with 0.01% Calcofluor white for 3 min, then washed three times to remove excess stain (Falconer and Seagull, 1985). All sections were mounted onto glass slides and examined with a Leica DRM microscope (Leica Microsystems, Wetzlar, Germany) fitted with epifluorescence optics. Photos were taken with a QICAM camera (QImaging, Surrey, Canada) and OpenLab software (PerkinElmer Inc., Waltham, USA). Images were visualized and analysed using ImageJ software(Abràmoff *et al.*, 2004).

3.3.6 Tension wood induction and sampling of differentiating xylem for transcriptome analysis

A tree bending trial was conducted in a clonal trial near KwaMbonambi in Northern Kwazulu-Natal, South Africa (Sappi Forest Research) to induce tension wood formation in ramets of the same F1 clone (GUSAP1) used for wood property analyses. The main stems of three 18-month-old ramets of the clone were bent at an angle of approximately 45° for three weeks in the field. To avoid temporal variation in gene expression, sampling of all replicates was completed within three hours around noon on the same day under the same environmental conditions. Differentiating xylem tissue was collected by cutting out the section of the stem bent at 45° (approximately 50 cm), removing the bark and immediately scraping the exposed outer differentiating layers of xylem cells 4-5 mm deep. For each bent stem, the upper

(tension wood) side was scraped. Differentiating xylem was collected from the corresponding location (height from the base) on three unbent controls. All samples were immediately frozen in liquid nitrogen and stored at -80°C.

3.3.7 RNA Isolation, sequencing and analysis

Total RNA was isolated from the xylem samples using a cetyl trimethylammonium bromide (CTAB) based method (Chang *et al.*, 1993). Frozen wood samples were ground to a fine powder in liquid nitrogen using a high speed grinder (IKA-Werke, Staufen, Germany). Fifteen ml of extraction buffer was mixed with three grams of ground tissue. RNA quantity and purity were assessed by using a Nanodrop spectrophotometer (Nanodrop Technologies ND 1000, Wilmington DE), Agilent Bioanalyser 2100 RNA 6000 pico total RNA kits (Agilent Technologies, Santa Clara CA) and 1.5% RNase-free agarose gels. To qualify for DGE and mRNA-Seq library preparation, all RNA samples had to have RNA integrity (RIN) numbers (Schroeder *et al.*, 2006) of 8.0 or higher. In addition, the samples were tested for DNA contamination using an intron spanning PCR. Three biological replicates each of tension wood and upright control samples were sequenced and mapped to the *E. grandis* genome using Tophat (Trapnell *et al.*, 2009) version 1.3.1, with the JGI v.1.0. gene models as a reference (www.phytozome.net). Expression levels (Fragments Per Kilobase of coding sequence per Million mapped fragments – FPKM) and differentially expressed genes were calculated using the Cufflinks and Cuffdiff packages version 1.0.3. (Trapnell *et al.*, 2010).

3.4 Results

3.4.1 Physicochemical changes of tension wood in *Eucalyptus*

We assessed changes in wood properties in naturally occurring tension wood derived from plantation-grown trees. To do this we collected tension and opposite wood from leaning basal side branches of five different ramets of an *E. grandis* × *E. urophylla* clone (GUSAP1, Sappi Forest Research, Fig. S3.1). Physical, chemical and ultrastructural characteristics of the tension wood and opposite wood were measured and compared (Table 3.1). In accordance with the individually quantified fibre characteristics, a marked difference in cell wall thickness between tension wood and opposite wood was observable by microscopy (Fig. 3.1). In the tension wood, fibres were on average 20% longer and showed a 40% increase in fibre coarseness. Fibre width was not significantly changed, but the secondary cell walls were thicker, consistent with the coarseness estimates. The fibre/vessel ratio was also higher in the tension wood (vessel density was 33% lower in TW), while vessel length and width were not significantly different.

In addition to the physical wood measurements, chemical analysis of the wood was performed to quantify changes in cellulose properties (Table 3.2), as well as lignin and total cell wall carbohydrate differences compared to opposite wood (Table 3.3). Although wood density and holocellulose (total polysaccharide) content were not different between tension and opposite wood, there was a significant increase in glucose (approximately 6 mg/100mg or a 16% relative increase) in the tension wood. This was mainly due to an increase in cellulose, as reflected in the significant increase in the α-cellulose content of the tension wood (Table 3.2). Consistent with previous tension wood studies, we also found a relatively lower MFA in the tension wood (20%). The hemicellulose composition was also different between tension and opposite wood. In short, rhamnose and arabinose were not significantly different and showed the highest variation among trees, but xylose and mannose concentrations were significantly lower in tension wood. However,

the largest relative difference in tension wood hemicellulose was galactose content, which was 250% higher (from 0.58mg/100mg to 1.82mg/100mg dry weight) in tension wood. The insoluble lignin content was also significantly reduced in tension wood compared to opposite wood (4.34%, to 26.2g/100g dry weight). A summary of the relative differences in all wood properties can be seen in Fig. 3.2, Table S3.1 and Table S3.2.

3.4.2 Transcriptional response to induced tension wood formation

To profile differential gene expression in tension wood forming tissues, we collected xylem from three 18-month-old GUSAP1 trees that were bent for three weeks. This experimental design permitted comparison with previous tension wood profiling experiments in *Populus* (Andersson-Gunneråsa *et al.*, 2006; Jin *et al.*, 2011). However, the actual bending and sampling was performed in field-grown trees, in this case, as opposed to potted greenhouse-grown trees. Additional trees of the same age that had been bent for six months in the field demonstrated observable tension wood at a macroscopic level (Fig. S3.3). RNA was extracted and RNA-Seq (Illumina) data produced from the xylem of three biological replicates of tension wood forming trees (three weeks post-induction) and upright controls. Overall, we found 366 genes that were significantly ($q<0.05$) differentially expressed in tension wood compared to the upright control sample (176 upregulated, 190 downregulated, Fig. S3.4, Additional file 3.1). *Arabidopsis* gene IDs homologous to the *Eucalyptus* gene IDs according to the Phytozome annotation were used for analysis using the BiNGO (Maere *et al.*, 2005) and GOToolBox (Martin *et al.*, 2004) tools to identify overrepresentation of ontology terms in the differentially expressed genes (Additional file 3.2).

In general, the most enriched biological processes in tension wood were genes related to stress response (stress, chemical, abiotic and mechanical stimulus). In addition, and consistent with previous analyses of differentially regulated genes in tension wood, several genes coding for FASCICLIN-LIKE

ARABINOGALACTAN proteins (FLAs) were highly upregulated - *FLA11* (*Eucgr.B02486*), *FLA12* (*Eucgr.J00938*) and *FLA17* (*Eucgr.A02551*) homologs. Other cell wall signalling-related genes were upregulated, including two homologs of Leucine-rich repeat protein kinases (*Eucgr.F02727*, *Eucgr.L02854*), annexin (*Eucgr.F02423*), IQ-Domain10 (*IQD10*, *Eucgr.F01203*) and a RAB GTPase homolog (*Eucgr.B02741*). Homologs of several transcription factors that have previously been associated with SCW biosynthesis were also upregulated, including *KNAT7* (*Eucgr.D01935*), *MYB52* (*Eucgr.F02756*), a C3HC4-type RING finger zinc finger family protein (*Eucgr.I01697*) and two C2H2-like zinc finger proteins (*Eucgr.B02487*, *Eucgr.H00574*). The joint upregulation of *KNAT7* and *MYB52* is interesting, as these genes have been shown to be co-regulated in *Arabidopsis* and are both repressed by MYB7 (Ko *et al.*, 2009), an ortholog of which (*Eucgr.C00721*) was downregulated in tension wood (Additional file 3.1). A homolog of *MYB61* (*Eucgr.B02197*) was also upregulated. In *Arabidopsis* this gene has been shown to be expressed in sink tissues, and is essential for xylem formation (Romano *et al.*, 2012). Other than general stress response related ontologies, the only other categories enriched in the significantly upregulated genes were "positive regulation of cell death" (GO:0010942 and its child terms); "disaccharide metabolism" (GO:0005984 and its child terms), relating to sucrose and trehalose metabolism; and methionine biosynthesis (GO:0006555, Additional file 3.2).

In contrast, enriched ontologies for downregulated genes included phenylpropanoid and flavonoid biosynthesis, as well as the biosynthesis of phenylalanine, tyrosine and tryptophan (Additional file 3.2). Homologs of genes representing most of the enzymatic steps involved in the monolignol biosynthetic pathway were significantly downregulated (Fig. 3.3), including PAL (*Eucgr.J00907*), C4H (*Eucgr.J01844*), 4CL (*Eucgr.K00087*), C3H (*Eucgr.G03199*), F5H (*Eucgr.J02393*) and COMT (*Eucgr.A01397*). Although evidence of large expansions has been noted for some of these gene families in *Eucalyptus grandis*, 24 core "*bona fide* lignifying" genes have been identified (Carocha et al., in preparation). With the exception of the CAD genes, most family members of all *bona fide* monolignol

biosynthesis genes were downregulated, although statistical significance was lacking for some (Additional file 3.3).

Given that significant increases in glucose, putatively stemming from increased α-cellulose synthesis were observed, as well as changes in the relative proportion of hemicellulosic components, we specifically examined the differential expression of genes involved in carbohydrate metabolism (Table 3.4). Among the upregulated genes, a *sucrose synthase 4* homolog (*SUS4*, *Eucgr.C03199*), trehalose-phosphatase/synthase 9 (*Eucgr.B02686*) and a trehalose-phosphatase family protein (*Eucgr.B02686*) were significantly upregulated, which could provide the increased source of UDP-glucose needed to supply the cellulose synthase machinery. Additionally, a β-glucosidase GH1 coding gene (*Eucgr.B00859*) was upregulated, which could be involved in cellulose modification in the cell wall. Although we saw no significant differences in expression of any of the cellulose synthase (*CesA*) genes, we observed the expression of a tandem duplicate of *EgCesA3* (*Eucgr.C00246*, ortholog of *AtCesA7* (Ranik & Myburg, 2006) that has not been previously observed in *Eucalyptus*. The tandem duplicate gene codes for an in-frame copy of *EgCesA3*, and is expressed at a similar level to *EgCesA3* (Fig. S3.5).

Several other genes encoding for glycosyl transferases were also upregulated, including GT32 (α-1,4-glycosyltransferase family protein, *Eucgr.A00510*) and GT35 (glycogen or starch phosphorylase, *Eucgr.J01374*). Another CAZyme gene transcriptionally upregulated was *GME* (*GDP-D-mannose 3',5'-epimerase*), known to be involved in ascorbate biosynthesis, pectic polysaccharide biosynthesis, and general stress response (Wolucka & Van Montagu, 2003; Caffall & Mohnen, 2009; Smirnoff, 2011). Among the three identified possible enzymatic functions, the most commonly ascribed role (EC 5.1.3.18) is the catalytic conversion of GDP-D-mannose to GDP-L-galactose (Major et al., 2005). This is interesting, as tension wood displayed reduced mannose and increased galactose compared to opposite

wood (Fig. 3.2). In general, with the exception of homologs of *SUS4* and *GUX2* (*GLUCURONIC ACID SUBSTITUTION OF XYLAN 2, Eucgr.F00232*), most of the CAZyme genes significantly upregulated in tension wood showed very low or no expression in the upright control (Table 3.4), and there is no indication of transcriptional rewiring of CAZymes' roles normally associated with SCW polysaccharide metabolism, as previously demonstrated in *Populus* (Andersson-Gunneras *et al.*, 2006). The downregulation of *PARVUS* (*Eucgr.A00485*) is notable, as it could be responsible for the reduction in xylose content observed in tension wood.

Several genes indicative of changes in hormone metabolism were also differentially expressed during tension wood formation (Table 3.5). In short, genes associated with elevated activity in the synthesis of ethylene, gibberellic acid and jasmonic acid in tension wood relative to the upright control was apparent. One of the most highly upregulated genes in tension wood (50-fold upregulation) was a homolog of *ACC OXIDASE 4*/*ETHYLENE FORMING ENZYME* (*ACO4*/*EFE – Eucgr.D01368*), which functions as the final step in ethylene formation from methionine by converting 1-aminocyclopropane-1-carboxylate to ethylene. Interestingly, this was contrasted by an almost complete suppression of an *ACO1* homolog expressed in the upright control (*Eucgr.C03886*, Table 3.5), suggesting a different family member is recruited for ethylene synthesis in *Eucalyptus* tension wood compared to normal wood. The enzymatic function of ACC OXIDASE has previously been highlighted in tension wood forming tissues of *Populus*, in producing ethylene to stimulate asymmetrical cambial growth (Andersson-Gunneras *et al.*, 2003; Love *et al.*, 2009). Of note was the upregulation of an ethylene response transcription factor, *Eucgr.F03499*, Homolog of *Arabidopsis ERF72*. In *Populus*, the overexpression of ERF72 orthologs (named PtiERF34 and PtiERF35) in hybrid poplar caused significant increases in the diameter (either gene) and height (PtiERF35 only) of trangenic trees (Vahala *et al.*, 2013). Similarly, in GA signalling, a homolog of an *Arabidopsis* gibberellin-response protein, *Eucgr.B03366*, was highly upregulated. The largest group of genes related to hormonal response were auxin-response genes, which showed both significant up- and

down-regulation. Among these was the downregulation of an oxireductase (*CAROTENOID CLEAVAGE DIOXYGENASE 8/CCD8, Eucgr.C02930*) associated with polar auxin transport and the suppression of branching (Auldridge *et al.*, 2006). Given that auxin, GA and ethylene are expected to be globally increased in xylogenic tissue of tension wood, these results are consistent with previous reports of hormonal changes in tension wood formation in *Populus*, suggesting a conserved mechanism between species.

3.5 Discussion

Tension wood represents an important developmental state due to the altered transcriptional and hormonal regulation, and the coordination of cellular processes recruited to alter cell wall chemical constituents. Although several studies have looked at aspects of tension wood in Eucalyptus, most of the information, especially regarding gene expression and biological changes were based on studies in Populus. It was thus not known if the physiological processes in *Eucalyptus* would be similar to those observed in *Populus*, as no study has as yet looked at the detailed physiology of tension wood formation, or indeed profiled gene expression at whole transcriptome level during tension wood formation in this genus. In this study we report the first data investigating transcriptome-wide changes manifested during tension wood formation in field-grown *Eucalyptus* trees, and comprehensively describe the physicochemical changes that accompany this developmental response to mechanical stress in woody stems.

Despite the high variation in transcript abundance, which is expected in a field grown experimentation, trends in gene expression supported many of the observed physicochemical changes. The number of significantly differentially expressed genes (366) is similar to the 444 previously reported in a controlled, greenhouse-grown *Populus* study (Andersson-Gunnerås *et al.*, 2006). It is noteworthy that among the most significantly differentially expressed genes, several common observations could be made between

Eucalyptus and *Populus* (Table S3.3) in terms of upregulation of ethylene biosynthesis (*EFE/ACO4*) and response (*ERF72*), UDP-glucose production (*SUSY*), transcriptional regulators (*KNAT7*, *MYB52* and *At3g27330* homologs) and cell wall signalling genes (*FLA12*), the reduction in the expression of *PARVUS* which is associated with xylan synthesis, and the downregulation of most lignin-related genes. This suggests common mechanisms employed by these two woody dicots in regulating and forming tension wood.

In general, the data for morphological and chemical changes concur with those previously observed in tension wood of various angiosperm species, namely longer fibres, a higher fibre:vessel ratio, increased cellulose with a decrease MFA, and a decrease in lignin (Tables 3.1-3.3). In terms of hemicellulose biosynthesis, we show that in *Eucalyptus* tension wood, xylan is significantly reduced, and this may be a consequence of the 4-fold downregulation of the *PARVUS* gene (Table 3.4). The simultaneous strong upregulation of a homolog of *GUX2* (*Eucgr.F00232*) in tension wood should also be noted, since *GUX* genes are responsible for glucuronic acid (GlcA) side chain addition onto the xylan backbone (Mortimer *et al.*, 2010; Lee *et al.*, 2012). It would be interesting to further characterize tension wood xylan to see whether the side chain structure is modified and what effect that may have on the structural properties of the SCW (e.g. increasing wood flexibility and tolerance to mechanical perturbations).

Evidence for the increase in fibre cell formation in tension wood can be seen in the increase in proportion of fibres (Table 3.1), which at a molecular level is evident in the upregulation in methionine metabolism for ethylene production, GA and auxin signalling, as well as increased ontologies associated with programmed cell death (Additional file 3.2 Table 3.5). In terms of compositional changes in fibre SCW, changes the CAZymes observed to be upregulated in tension wood forming tissue (with the exception of GT8 and GT4) have been observed to be generally ubiquitously expressed across multiple tissues/organs

of *E. grandis* or specific to primary cell wall tissues (Table 3.4, Table S3.4). This observation could be due to the fact that tension wood is producing increased amounts of carbohydrates generally not found in high abundance in SCWs (such as galactose – Table 3.3, Fig. 3.2).

It is unclear in what form galacatose is present in the cell wall, as it could make up components of FLAs (Seifert & Roberts, 2007), pectic compounds such as the side chains of Rhamnogalacturonan I (Goubet *et al.*, 1995; Scheller *et al.*, 2007), or indeed as pure galactan. Nevertheless, the role of galactose and galactan in SCW and G-layer deposition, especially pertaining to orientation of cellulose microfibrils during cellulose deposition, has been previously highlighted (Gorshkova & Morvan, 2006; Roach *et al.*, 2011). Several studies have also reported the presence of β-(1→4) and increase in β-(1→6) galactans in tension wood, some of unique composition not usually found in upright wood (Meier, 1962). We identified a homolog of the gene *AT3G27330*, coding for a protein of unknown function that has recently been annotated as possessing a GT92 domain (Yin *et al.*, 2012), also present in all three recently characterized GALACTAN SYNTHASE 1, 2 and 3 proteins that were sufficient for increasing cell wall galactan content (Liwanag *et al.*, 2012). This gene is not normally expressed in *Eucalyptus* xylem (Table S3.4), but is upregulated in tension wood (Table 3.4), and would be a candidate for tension wood-specific β-(1→4) galactan synthesis.

The relative changes in cellulose quantity and properties are more complex, but are likely related to an increased carbon flux to UDP-glucose via SUSY or trehalose, and/or possible post-transcriptional/post-translational mechanism(s) that were not apparent in this study. The reduction in lignin can be attributed to a significant reduction in expression of the suite of monolignol biosynthetic genes (Fig. 3.3, Additional file 3.3), as well as those involved in shikimate biosynthesis (*Eucgr.H01214*) and phenylalanine metabolism (*Eucgr.J00428* – Table S3.2). Together with enriched ontologies represented by upregulated

genes, it is our conclusion that at a transcriptional level, the underlying molecular mechanism controlling *Eucalyptus* tension wood physiology is likely a reduction in lignin monomer production, xylan biosynthesis and synthesis of polysaccharides not usually occurring in wood, rather than the relative upregulation of pathways involved in secondary cell wall cellulose synthesis, as previously described.

Eucalyptus is a commercially important hardwood genus, which will in the future rely on more sophisticated biotechnology strategies to enhance woody biomass traits relevant to industry. These strategies depend on understanding the roles of genes and biological processes during xylogenesis, including those involved in hormonal changes, cellular patterning, carbohydrate composition and cell wall ultrastructure. In this study we have shown that gene expression and wood property changes during tension wood formation in field-grown *Eucalyptus* trees is consistent with previous results from model systems, highlighting key pathways and genes putatively involved in tension wood formation. Detailed microstructural studies will be needed to resolve any novel tension wood-specific polysaccharides, but this study suggests that the deposition of galactans not normally associated with secondary cell walls may play a role in tension wood formation or function in *Eucalyptus*. In the future, strategies to modify wood cell wall composition or ultrastructure will likely involve attenuation or overexpression of key genes, but could also involve the integration of novel biopolymers not normally found in wood.

3.6 Acknowledgements

The authors would like to acknowledge M. Ranik and M. O'Neill (University of Pretoria) for the mRNA-seq library preparations and the sequencing facility at Oregon State University for assistance with RNA-sequencing. Plant materials were kindly provided by Sappi Forest Research (KwaMbonambi, South Africa). This work was supported through a strategic research grant from the South African Department

of Science and Technology (DST) and by research funding from Sappi and Mondi, through the Forest Molecular Genetics Programme, the Technology and Human Resources for Industry Programme (THRIP, UID 80118) and the Bioinformatics and Functional Genomics Programme of the National Research Foundation (NRF, UID 18312) of South Africa.

3.7 References

Abràmoff MD, Magalhães PJ, Ram SJ. 2004. Image processing with ImageJ. *Biophotonics International* **11**(7): 36-42.

Al-Haddad JM, Kang K-Y, Mansfield SD, Telewski FW. 2013. Chemical responses to modified lignin composition in tension wood of hybrid poplar (*Populus tremula* × *Populus alba*). *Tree Physiology* **33**(4): 365-373.

Andersson-Gunnerås S, Hellgren JM, Björklund S, Regan S, Moritz T, Sundberg B. 2003. Asymmetric expression of a poplar ACC oxidase controls ethylene production during gravitational induction of tension wood. *Plant Journal* **34**(3): 339-349.

Andersson-Gunnerås S, Mellerowicz EJ, Love J, Segerman B, Ohmiya Y, Coutinho PM, Nilsson P, Henrissat B, Moritz T, Sundberg B. 2006. Biosynthesis of cellulose-enriched tension wood in *Populus*: Global analysis of transcripts and metabolites identifies biochemical and developmental regulators in secondary wall biosynthesis. *Plant Journal* **45**(2): 144-165.

Aoyama W, Matsumura A, Tsutsumi Y, Nishida T. 2001. Lignification and peroxidase in tension wood of *Eucalyptus viminalis* seedlings. *Journal of Wood Science* **47**(6): 419-424.

Auldridge ME, Block A, Vogel JT, Dabney-Smith C, Mila I, Bouzayen M, Magallanes-Lundback M, DellaPenna D, McCarty DR, Klee HJ. 2006. Characterization of three members of the *Arabidopsis* carotenoid cleavage dioxygenase family demonstrates the divergent roles of this multifunctional enzyme family. *The Plant Journal* **45**(6): 982-993.

Baba K, Karlberg A, Schmidt J, Schrader J, Hvidsten TR, Bako L, Bhalerao RP. 2011. Activity-dormancy transition in the cambial meristem involves stage-specific modulation of auxin response in hybrid aspen. *Proceedings of the National Academy of Sciences of the United States of America* **108**(8): 3418-3423.

Baba K, Park YW, Kaku T, Kaida R, Takeuchi M, Yoshida M, Hosoo Y, Ojio Y, Okuyama T, Taniguchi T, Ohmiya Y, Kondo T, Shani Z, Shoseyov O, Awano T, Serada S, Norioka N,

Norioka S, Hayashi T. 2009. Xyloglucan for generating tensile stress to bend tree stem. *Molecular Plant* **2**(5): 893-903.

Bentum ALK, Côté Jr WA, Day AC, Timell TE. 1969. Distribution of lignin in normal and tension wood. *Wood Science and Technology* **3**(3): 218-231.

Björklund S, Antti H, Uddestrand I, Moritz T, Sundberg B. 2007. Cross-talk between gibberellin and auxin in development of *Populus* wood: Gibberellin stimulates polar auxin transport and has a common transcriptome with auxin. *Plant Journal* **52**(3): 499-511.

Bowling AJ, Vaughn KC. 2008. Immunocytochemical characterization of tension wood: Gelatinous fibers contain more than just cellulose. *American Journal of Botany* **95**(6): 655-663.

Browning B. 1967. *Methods of wood chemistry*. New York, NY, USA: Wiley Interscience Publishers.

Caffall KH, Mohnen D. 2009. The structure, function, and biosynthesis of plant cell wall pectic polysaccharides. *Carbohydrate Research* **344**(14): 1879-1900.

Chang S, Puryear J, Cairney J. 1993. A simple and efficient method for isolating RNA from pine trees. *Plant Molecular Biology Reporter* **11**(2): 113-116.

Clair B, Alméras T, Pilate G, Jullien D, Sugiyama J, Riekel C. 2011. Maturation stress generation in poplar tension wood studied by synchrotron radiation microdiffraction. *Plant Physiology* **155**(1): 562-570.

Clair B, Alméras T, Yamamoto H, Okuyama T, Sugiyama J. 2006a. Mechanical behavior of cellulose microfibrils in tension wood, in relation with maturation stress generation. *Biophysical Journal* **91**(3): 1128-1135.

Clair B, Ruelle J, Beauchêne J, Prévost MF, Fournier M. 2006b. Tension wood and opposite wood in 21 tropical rain forest species. 1. Occurrence and efficiency of the G-layer. *IAWA Journal* **27**(3): 329-338.

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.

Côté Jr WA, Day AC, Timell TE. 1969. A contribution to the ultrastructure of tension wood fibers. *Wood Science and Technology* **3**(4): 257-271.

Elo A, Immanen J, Nieminen K, Helariutta Y. 2009. Stem cell function during plant vascular development. *Seminars in Cell and Developmental Biology* **20**(9): 1097-1106.

Fang CH, Clair B, Gril J, Alméras T. 2007. Transverse shrinkage in G-fibers as a function of cell wall layering and growth strain. *Wood Science and Technology* **41**(8): 659-671.

Fang CH, Clair B, Gril J, Liu SQ. 2008. Growth stresses are highly controlled by the amount of G-layer in poplar tension wood. *Iawa Journal* **29**(3): 237-246.

Gorshkova T, Morvan C. 2006. Secondary cell-wall assembly in flax phloem fibres: Role of galactans. *Planta* **223**(2): 149-158.

Goswami L, Dunlop JWC, Jungnikl K, Eder M, Gierlinger N, Coutand C, Jeronimidis G, Fratzl P, Burgert I. 2008. Stress generation in the tension wood of poplar is based on the lateral swelling power of the G-layer. *Plant Journal* **56**(4): 531-538.

Goubet F, Bourlard T, Girault R, Alexandre C, Vandevelde MC, Morvan C. 1995. Structural features of galactans from flax fibres. *Carbohydrate Polymers* **27**(3): 221-227.

Goulao LF, Vieira-Silva S, Jackson PA. 2011. Association of hemicellulose- and pectin-modifying gene expression with *Eucalyptus globulus* secondary growth. *Plant Physiology and Biochemistry* **49**(8): 873-881.

Humphreys JM, Chapple C. 2002. Rewriting the lignin roadmap. *Current Opinion in Plant Biology* **5**(3): 224-229.

Jin H, Do J, Moon D, Noh EW, Kim W, Kwon M. 2011. EST analysis of functional genes associated with cell wall biosynthesis and modification in the secondary xylem of the yellow poplar (*Liriodendron tulipifera*) stem during early stage of tension wood formation. *Planta* **234**(5): 959-977.

Joseleau JP, Imai T, Kuroda K, Ruel K. 2004. Detection in situ and characterization of lignin in the G-layer of tension wood fibres of *Populus deltoides*. *Planta* **219**(2): 338-345.

Jourez B, Riboux A, Leclercq A. 2001. Anatomical characteristics of tension wood and opposite wood in young inclined stems of poplar (*Populus euramericana* cv 'Ghoy'). *IAWA Journal* **22**(2): 133-157.

Ko JH, Han KH, Park S, Yang J. 2004. Plant body weight-induced secondary growth in *Arabidopsis* and its transcription phenotype revealed by whole-transcriptome profiling. *Plant Physiology* **135**(2): 1069-1083.

Ko JH, Kim WC, Han KH. 2009. Ectopic expression of MYB46 identifies transcriptional regulatory genes involved in secondary wall biosynthesis in *Arabidopsis*. *Plant Journal* **60**(4): 649-665.

Lafarguette F, Leplé JC, Déjardin A, Laurans F, Costa G, Lesage-Descauses MC, Pilate G. 2004. Poplar genes encoding fasciclin-like arabinogalactan proteins are highly expressed in tension wood. *New Phytologist* **164**(1): 107-121.

Lee C, Teng Q, Zhong R, Ye ZH. 2012. *Arabidopsis* GUX proteins are glucuronyltransferases responsible for the addition of glucuronic acid side chains onto xylan. *Plant and Cell Physiology* **53**(7): 1204-1216.

Little CHA, Savidge RA. 1987. 7. The role of plant growth regulators in forest tree cambial growth. *Plant Growth Regulation* **6**(1-2): 137-169.

Liwanag AJM, Ebert B, Verhertbruggen Y, Rennie EA, Rautengarten C, Oikawa A, Andersen MC, Clausen MH, Scheller HV. 2012. Pectin biosynthesis: GALS1 in *Arabidopsis thaliana* is a β-1, 4-galactan β-1, 4-galactosyltransferase. *The Plant Cell Online* **24**(12): 5024-5036.

Lloyd C 2011. Dynamic Microtubules and the Texture of Plant Cell Walls. *International Review of Cell and Molecular Biology* **287**: 287-329.

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 United States of America* **106**(14): 5984-5989.

Maere S, Heymans K, Kuiper M. 2005. BiNGO: A Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks. *Bioinformatics* **21**(16): 3448-3449.

Major LL, Wolucka BA, Naismith JH. 2005. Structure and function of GDP-mannose-3′,5′-epimerase: An enzyme which performs three chemical reactions at the same active site. *Journal of the American Chemical Society* **127**(51): 18309-18320.

Martin D, Brun C, Remy E, Mouren P, Thieffry D, Jacq B. 2004. GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biology* **5**(12).

Megraw R, Leaf G, Bremer D 1998. Longitudinal shrinkage and microfibril angle in loblolly pine. *Microfibril angle in wood. Proc. IAWA/IUFRO Intn. Workshop on the significance of microfibril angle to wood quality.. University of Canterbury: Christchurch, New Zealand*. 27-61.

Meier H. 1962. Studies on a galactan from tension wood of beech (*Fagus silvatica* L.). *Acta chem. scand* **16**(9): 14.

Mellerowicz EJ, Immerzeel P, Hayashi T. 2008. Xyloglucan: The molecular muscle of trees. *Annals of Botany* **102**(5): 659-665.

Mortimer JC, Miles GP, Brown DM, Zhang Z, Segura MP, Weimar T, Yu X, Seffen KA, Stephens E, Turner SR, Dupree P. 2010. Absence of branches from xylan in *Arabidopsis* gux mutants reveals potential for simplification of lignocellulosic biomass. *Proceedings of the National Academy of Sciences of the United States of America* **107**(40): 17409-17414.

Moyle R, Schrader J, Stenberg A, Olsson O, Saxena S, Sandberg G, Bhalerao RP. 2002. Environmental and auxin regulation of wood formation involves members of the Aux/IAA gene family in hybrid aspen. *Plant Journal* **31**(6): 675-685.

Müller M, Burghammer M, Sugiyama J. 2006. Direct investigation of the structural properties of tension wood cellulose microfibrils using microbeam X-ray fibre diffraction. *Holzforschung* **60**(5): 474-479.

Nishikubo N, Awano T, Banasiak A, Bourquin V, Ibatullin F, Funada R, Brumer H, Teeri TT, Hayashi T, Sundberg B, Mellerowicz EJ. 2007. Xyloglucan endo-transglycosylase (XET) functions in gelatinous layers of tension wood fibers in poplar - A glimpse into the mechanism of the balancing act of trees. *Plant and Cell Physiology* **48**(6): 843-855.

Okuyama T, Yamamoto H, Yoshida M, Hattori Y, Archer RR. 1994. Growth stresses in tension wood: Role of microfibrils and lignification. *Annales des Sciences Forestieres* **51**(3): 291-300.

Paux E, Carocha V, Marques C, De Sousa AM, Borralho N, Sivadon P, Grima-Pettenati J. 2005. Transcript profiling of *Eucalyptus* xylem genes during tension wood formation. *New Phytologist* **167**(1): 89-100.

Pesquet E, Tuominen H. 2011. Ethylene stimulates tracheary element differentiation in *Zinnia elegans* cell cultures. *New Phytologist* **190**(1): 138-149.

Qiu D, Wilson IW, Gan S, Washusen R, Moran GF, Southerton SG. 2008. Gene expression in *Eucalyptus* branch wood with marked variation in cellulose microfibril orientation and lacking G-layers. *New Phytologist* **179**(1): 94-103.

Ranik M, Myburg AA. 2006. Six new cellulose synthase genes from *Eucalyptus* are associated with primary and secondary cell wall biosynthesis. *Tree Physiology* **26**(5): 545-556.

Roach MJ, Mokshina NY, Badhan A, Snegireva AV, Hobson N, Deyholos MK, Gorshkova TA. 2011. Development of cellulosic secondary walls in flax fibers requires β-galactosidase. *Plant Physiology* **156**(3): 1351-1363.

Romano JM, Dubos C, Prouse MB, Wilkins O, Hong H, Poole M, Kang KY, Li E, Douglas CJ, Western TL. 2012. AtMYB61, an R2R3-MYB transcription factor, functions as a pleiotropic regulator via a small gene network. *New Phytologist* **195**(4): 774-786.

Ruelle J, Beauchêne J, Yamamoto H, Thibaut B. 2010. Variations in physical and mechanical properties between tension and opposite wood from three tropical rainforest species. *Wood Science and Technology*: 1-19.

Ruelle J, Clair B, Beauchêne J, Prévost MF, Fournier M. 2006. Tension wood and opposite wood in 21 tropical rain forest species 2. Comparison of some anatomical and ultrastructural criteria. *IAWA Journal* **27**(4): 341-376.

Scheller HV, Jensen JK, Sørensen SO, Harholt J, Geshi N. 2007. Biosynthesis of pectin. *Physiologia Plantarum* **129**(2): 283-295.

Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, Lightfoot S, Menzel W, Granzow M, Ragg T. 2006. The RIN: An RNA integrity number for assigning integrity values to RNA measurements. *BMC Molecular Biology* **7**(1): 3.

Sehr EM, Agusti J, Lehner R, Farmer EE, Schwarz M, Greb T. 2010. Analysis of secondary growth in the *Arabidopsis* shoot reveals a positive role of jasmonate signalling in cambium formation. *Plant Journal* **63**(5): 811-822.

Seifert GJ, Roberts K 2007. The biology of arabinogalactan proteins. *Annual Review of Plant Biology*. **58:**137-161.

Smirnoff N. 2011. Vitamin C: The metabolism and functions of ascorbic acid in plants. *Advances in Botanical Research* **59:**107-177.

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, van Baren MJ, 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**(5): 511-U174.

Vahala J, Felten J, Love J, Gorzsas A, Gerber L, Lamminmaki A, Kangasjarvi J, Sundberg B. 2013. A genome-wide screen for ethylene-induced Ethylene Response Factors (ERFs) in hybrid aspen stem identifies ERF genes that modify stem growth and wood properties. *New Phytologist* **200**(2): 511-522.

Washusen R, Evans R, Southerton S. 2005. A study of *Eucalyptus grandis* and *Eucalyptus globulus* branch wood microstructure. *Iawa Journal* **26**(2): 203-210.

Washusen R, Ilic J, Waugh G. 2003. The relationship between longitudinal growth strain, tree form and tension wood at the stem periphery of ten- to eleven-year-old *Eucalyptus globulus* labill. *Holzforschung* **57**(3): 308-316.

Wolucka BA, Van Montagu M. 2003. GDP-Mannose 3′,5′-Epimerase forms GDP-L-gulose, a putative intermediate for the *de novo* biosynthesis of vitamin C in plants. *Journal of Biological Chemistry* **278**(48): 47483-47490.

Yamamoto H. 2004. Role of the gelatinous layer on the origin of the physical properties of the tension wood. *Journal of Wood Science* **50**(3): 197-208.

Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. 2012. DbCAN: A web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research* **40**(W1): W445-W451.

Yokoyama T, Kadla J, Chang H. 2002. Microanalytical method for the characterization of fiber components and morphology of woody plants. *Journal of Agricultural and Food Chemistry* **50**(5): 1040-1044.

Yoshida M, Ohta H, Yamamoto H, Okuyama T. 2002. Tensile growth stress and lignin distribution in the cell walls of yellow poplar, *Liriodendron tulipifera* Linn. *Trees - Structure and Function* **16**(7): 457-464.

Yoshizawa N, Inami A, Miyake S, Ishiguri F, Yokota S. 2000. Anatomy and lignin distribution of reaction wood in two *Magnolia* species. *Wood Science and Technology* **34**(3): 183-196.

3.8 Figures and Tables

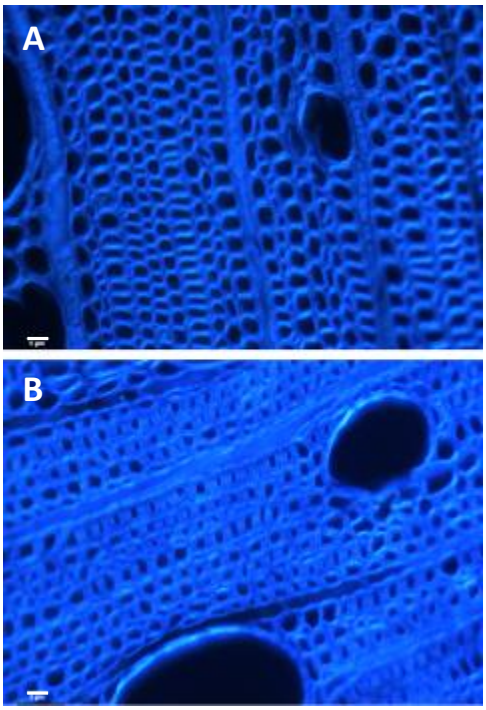


Fig. 3.1 Cell wall morphology of opposite wood (A) and tension wood (B) from an *E. grandis* × *E. urophylla* hybrid tree (ramet).

Comparisons from all sampled ramets (Fig. S3.1) can be seen Fig. S3.2 (Scale bar – 5 µm).

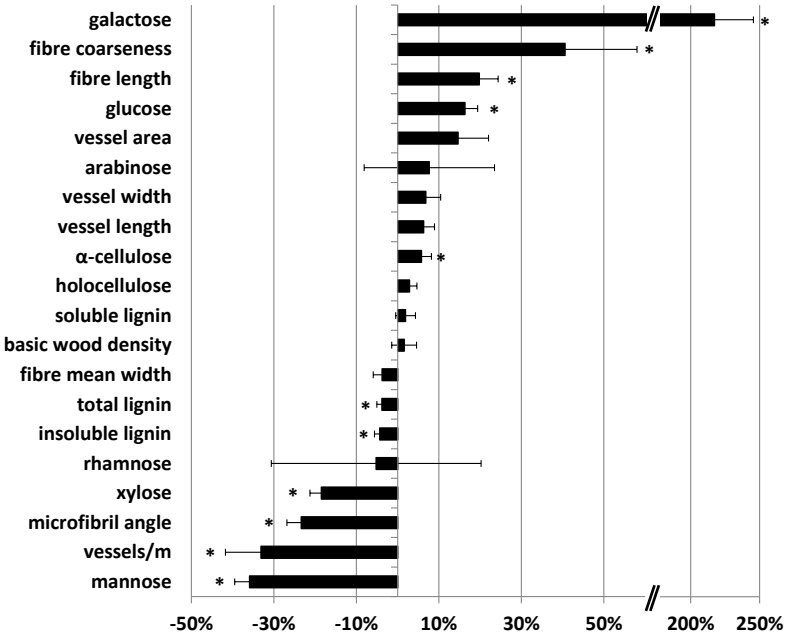


Fig. 3.2 Relative changes in wood properties between tension wood and opposite wood.

A positive change indicates a higher value in tension wood as compared to opposite wood, and a negative change indicates a lower value in tension wood compared to opposite wood. Error bars represent standard error (n=5), *denotes significance at $p \leq 0.05$ (paired, two-tailed t-test).

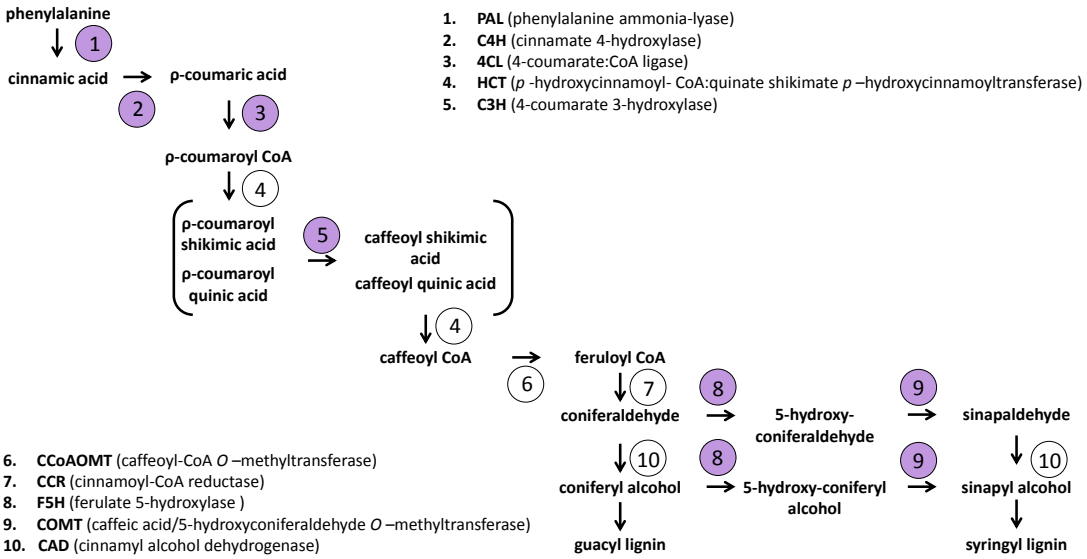


Fig. 3.3 Main steps of the monolignol biosynthetic pathway in *Eucalyptus* downregulated in tension wood.

Pathway construction is based on Humphreys & Chapple (2002). Enzymatic steps where one or more representative genes showed significant downregulation in tension wood xylem compared to the upright control in the *E. grandis* x *E. urophylla* hybrid clone are highlighted in purple. A full table of differential expression for all genes involved in the pathway (annotation according to Carocha *et al.*, in preparation) is available in Additional file 3.3.

Table 3.1 Fibre and vessel properties in tension and opposite wood of five ramets of *E. grandis* × *E. urophylla* F1 hybrid clone GUSAP1.

Vessels/m indicated the density of vessels, calculated as the number (n) of observed vessels *significant at $p \leq 0.05$, **significant at $p \leq 0.01$ (paired, two-tailed T-test).

Sample	Fibre length (mm)	Fibre width (µm)	Fibre coarseness (mg/m)	Vessel Area (mm^2)	Vessel length (mm)	Vessel width (µm)	Vessels/m (n)
Opposite wood 1	0.78	18.50	0.03	0.07	0.57	118.30	8.58
Opposite wood 2	0.69	19.20	0.04	0.06	0.54	107.20	10.12
Opposite wood 3	0.62	19.10	0.05	0.06	0.51	112.70	9.88
Opposite wood 4	0.77	19.20	0.06	0.06	0.54	114.20	7.99
Opposite wood 5	0.70	18.20	0.06	0.05	0.49	105.40	9.69
Tension wood 1	0.88	17.90	0.06	0.07	0.56	118.30	7.55
Tension wood 2	0.75	19.20	0.05	0.07	0.57	121.30	6.61
Tension wood 3	0.85	17.40	0.06	0.06	0.57	109.10	4.66
Tension wood 4	0.88	18.40	0.06	0.07	0.56	124.50	6.77
Tension wood 5	0.87	18.60	0.06	0.07	0.55	121.40	4.75
Opposite wood (MEAN ± SD)	**0.70 ± 0.06**	**18.84 ± 0.46**	**0.05 ± 0.01**	**0.06 ± 0.01**	**0.53 ± 0.03**	**111.56 ± 5.26**	**9.25 ±0.92**
Tension wood (MEAN +-SD)	**0.85 ± 0.05****	**18.30 ± 0.69**	**0.06 ± 0.00***	**0.07 ± 0.00**	**0.56 ± 0.01**	**118.92 ± 0.07**	**6.07 ± 1.29****

Table 3.2 Basic wood density, holocellulose, α-cellulose and microfibril angle in tension and opposite wood of five ramets of *E. grandis* x *E. urophylla* F1 hybrid clone GUSAP1.

*significant at p ≤ 0.05, **significant at p ≤ 0.01 (paired, two-tailed t-test).

Sample	Wood density (kg/m^3)	Holocellulose (mg/100mg)	α-Cellulose (mg/100mg)	Microfibril angle (°)
Opposite wood 1	460.28	64.35	38.29	21.23
Opposite wood 2	411.74	63.92	38.96	19.31
Opposite wood 3	504.38	65.28	37.29	18.42
Opposite wood 4	465.49	67.10	39.62	22.12
Opposite wood 5	531.19	65.24	40.47	18.55
Tension wood 1	483.04	64.34	41.86	15.23
Tension wood 2	413.49	62.81	39.66	16.25
Tension wood 3	516.17	66.68	41.93	14.33
Tension wood 4	507.82	71.61	42.04	14.59
Tension wood 5	483.78	69.97	40.14	15.48
Opposite wood (MEAN ± SD)	**474.62 ± 45.63**	**65.18 ± 1.22**	**38.93 ± 1.22**	**19.92 ± 0.74**
Tension wood (MEAN +-SD)	**480.86 ± 40.39**	**67.08 ± 3.70**	**41.13 ± 1.13***	**15.18 ± 0.34****

Table 3.3 Cell wall composition in tension and opposite wood of five ramets of *E. grandis* x *E. urophylla* F1 hybrid clone GUSAP1.

* significant at p ≤ 0.05, ** significant at p ≤ 0.01 (paired, two-tailed t-test).

	Mg/100mg								
	Acid-insoluble lignin	Acid-soluble lignin	Total lignin	Arabinose	Rhamnose	Galactose	Glucose	Xylose	Mannose
Opposite wood 1	26.66	2.74	29.40	0.44	0.30	0.51	42.45	13.39	2.25
Opposite wood 2	28.52	2.48	31.01	0.45	0.35	0.69	42.41	13.45	2.47
Opposite wood 3	26.97	2.76	29.74	0.37	0.33	0.59	41.86	13.45	2.90
Opposite wood 4	27.09	2.51	29.60	0.31	0.20	0.43	44.15	14.63	2.63
Opposite wood 5	27.94	2.55	30.49	0.20	0.31	0.67	39.29	11.23	2.39
Tension wood 1	26.67	2.86	29.53	0.41	0.08	1.38	48.27	11.19	1.47
Tension wood 2	26.61	2.68	29.29	0.32	0.25	1.75	49.26	11.16	1.82
Tension wood 3	25.67	2.76	28.42	0.35	0.27	1.77	48.07	10.47	1.54
Tension wood 4	26.18	2.59	28.77	0.36	0.36	1.72	48.20	10.73	1.56
Tension wood 5	26.03	2.40	28.44	0.34	0.35	2.47	50.17	10.07	1.65
Opposite wood (MEAN ± SD)	**27.44 ± 0.77**	**2.61 ± 0.13**	**30.05 ± 0.68**	**0.36 ± 0.10**	**0.30 ± 0.06**	**0.58 ± 0.11**	**42.03 ± 1.76**	**13.23 ± 1.23**	**2.53 ± 0.25**
Tension wood (MEAN +-SD)	**26.23 ± 0.42***	**2.66 ± 0.18**	**28.89 ± 0.50***	**0.36 ± 0.03**	**0.26 ± 0.11**	**1.82 ± 0.40****	**48.79 ± 0.90****	**10.73 ± 0.47****	**1.61 ± 0.13****

Table 3.4 Carbohydrate Active enZyme (CAZyme) genes significantly (q<0.05) differentially expressed in three-week tension wood forming xylem of *E. grandis* x *E. urophylla* trees.

***E. grandis* ID**	***Arabidopsis* homolog**	***Arabidopsis* protein**	**CAZyme annotation***	**Ln (fold change)**	**Average FPKM (upright)†**	**Average FPKM (TW)†**
Eucgr.A00510	AT2G38150	α-1,4-glycosyltransferase	GT32	4.24	0	43
Eucgr.H00343	AT1G68470	Exostosin family protein	GT47	3.87	2	143
Eucgr.I01697	AT3G27330	β-1,4-galactosyltransferase	GT92	3.71	3	153
Eucgr.B00354	AT1G23870	ATTPS9, TPS9	GT20	3	1	13
Eucgr.I01147	AT1G49710	FUCTB, FUT12	GT10	2.25	14	111
Eucgr.K00865	AT4G15240	Unknown (DUF604)	GT31	1.78	9	71
Eucgr.B00859	AT3G18080	BGLU44	GH1	1.51	10	26
Eucgr.B02686	AT1G68020.2	ATTPS6, TPS6	GT20	1.49	9	36
Eucgr.F03658	AT1G55740	AtSIP1, SIP1	GH36	1.48	10	41
Eucgr.E01169	AT4G19420	Pectinacetylesterase family protein	CE13	1.36	40	121
Eucgr.F01855	AT1G45130	BGAL5	GH35	1.19	5	14
Eucgr.J01374	AT3G29320	alpha-glucan phosphorylase	GT35	1.12	5	14
Eucgr.B02118	AT5G28840	GME		1.12	13	38
Eucgr.F00232	AT4G33330	GUX2, PGSIP3	GT8	0.94	101	386
Eucgr.H00536	AT3G17880	HIP, TDX	GT41	0.9	20	43
Eucgr.C03199	AT3G43190	SUS4	GT4	0.49	1 144	1 614
Eucgr.J02867	AT3G13750	BGAL1	GH35	-0.84	28	10
Eucgr.G02748	AT5G04310	Pectin lyase-like	PL1	-1	29	12
Eucgr.G02887	AT5G04310	Pectin lyase-like	PL1	-1.12	11	1
Eucgr.F02205	AT1G58370	ATXYN1, RXF12	CBM22	-1.26	16	5
Eucgr.K03600	AT4G13710	Pectin lyase-like	PL1	-1.26	22	8
Eucgr.A00780	AT5G01930	MAN6	GH5	-1.37	72	20
Eucgr.A00485	AT1G19300	GATL1, PARVUS	GT8	-1.51	196	46
Eucgr.C03207	AT3G43190	ATSUS4,SUS4	GT4	-2.74	10	1

***Annotation of CAZymes according to Yin *et al.* (2012)

†Average FPKM of three biological replicates

Table 3.5 Hormone-related genes significantly differentially (q<0.05) expressed in three-week tension wood forming xylem of *E. grandis* x *E. urophylla* trees.

***E. grandis* ID**	***Arabidopsis* homolog**	***Arabidopsis* protein**	**Ln (fold change)**	**Average FPKM* (upright)**	**Average FPKM* (TW)**	**Hormonal pathway**
Eucgr.B03366	AT5G14920		4.05	2	124	Gibberellic Acid
Eucgr.D01368	AT1G05010	EFE, ACO4,	3.68	15	807	Ethylene
Eucgr.E03916	AT5G47530		2.83	5	110	Auxin
Eucgr.H04545	AT1G56220		1.62	30	220	Auxin
Eucgr.H03965	AT3G16770	RAP2.3, ATEBP, ERF72, EBP	1.33	12	69	Ethylene
Eucgr.C03183	AT4G33150	LKR,\SDH	1.21	7	17	Jasmonic Acid
Eucgr.H03171	AT1G04240	SHY2, IAA3	-1.11	20	7	Auxin
Eucgr.H02914	AT2G33310	IAA13	-1.34	25	10	Auxin
Eucgr.G01769	AT2G21050	LAX2	-1.50	121	39	Auxin
Eucgr.C03886	AT2G19590	ACO1	-3.09	87	4	Ethylene
Eucgr.C02930	AT4G32810	CCD8,MAX4	-6.05	23	0	Auxin

*Average FPKM in three biological replicates

3.9 Additional Files

1. Additional File 3.1.xlsx – Genes significantly differentially expressed in induced tension wood xylem compared to upright control xylem.
2. Additional File 3.2.zip – Ontology summaries of upregulated and downregulated biological processes according to BiNGO ("cuffdiff_sig_up_BP.pdf" and "Cuffdiff_sig_down_BP.pdf", respectively) and GoToolBox ("GoToolBox Summary tables.docx").
3. Additional File 3.3.pdf – Differential expression values (FPKM) and relative fold-change in tension wood compared to upright control for all genes involved in the monolignol biosynthesis pathway in *Eucalyptus* (annotation according to Carocha et al., in preparation). Significant differences are highlighted and average expression colouring is scaled within gene families.

3.10 Supplemental data

Fig. S3.1 Trees used for analysis of physicochemical changes in tension wood properties**.**

Five individuals (ramets) of a commercial hybrid *E. grandis* × *E. urophylla* F1 hybrid clone (GUSAP1) were selected with naturally slanting branches emerging from the base of the tree. A transverse disc (bottom left) was cut from an area of the branch which was at approximately 45º from the longitudinal axis of the tree trunk.

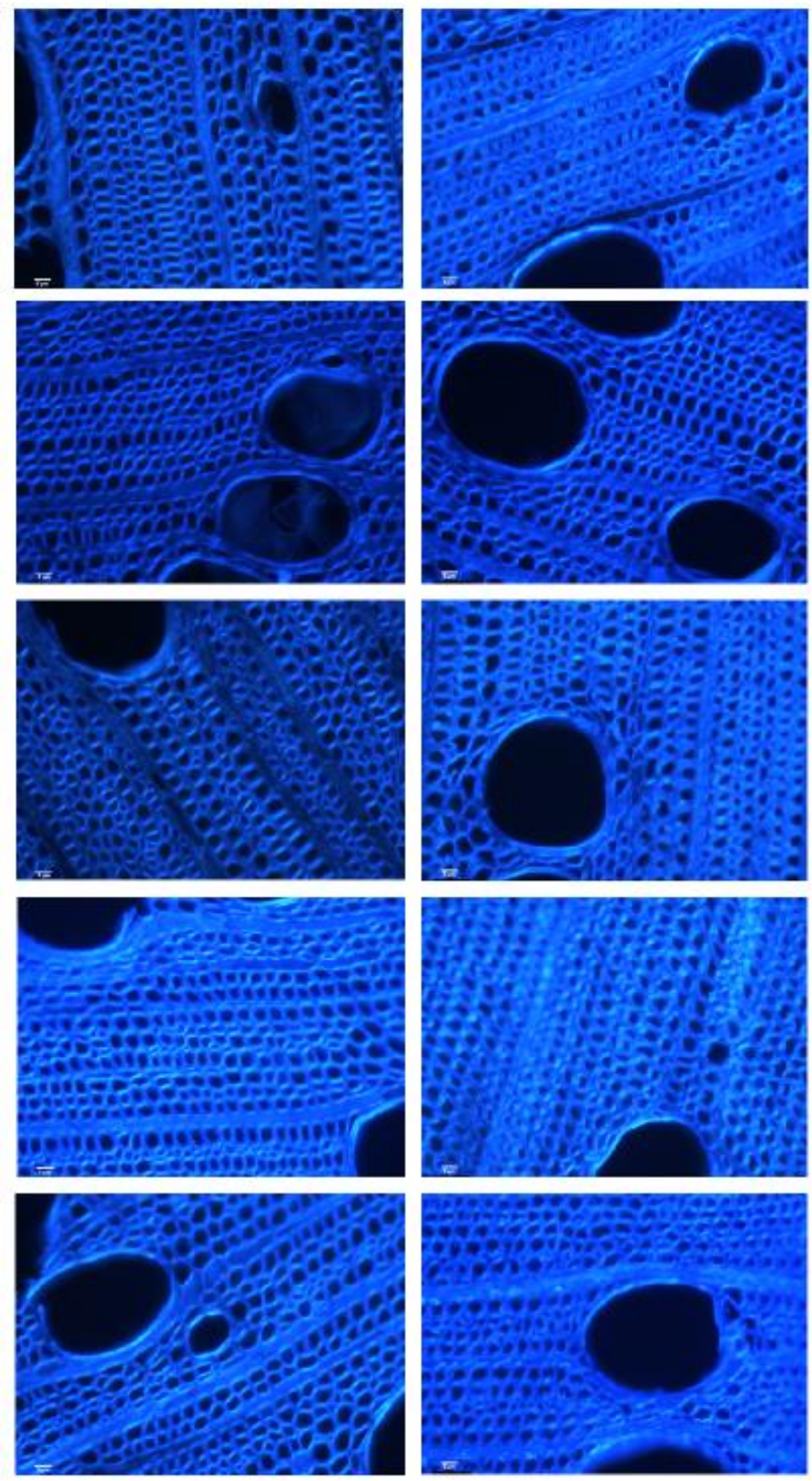


Fig. S3.2 Comparison of opposite wood (left) and tension wood (right) tissue from five ramets (top to bottom, 1-5 from Fig. S1) of an F_1 hybrid clone of *E. grandis* and *E. urophylla* (GUSAP1).

Fig. S3.3 Cross section of the main stem of an 18-month-old GUSAP1 (*E. grandis* x *E. urophylla*) tree after six months of bending. Tension wood can be seen at a macroscopic level on the top half of the bent trunk.

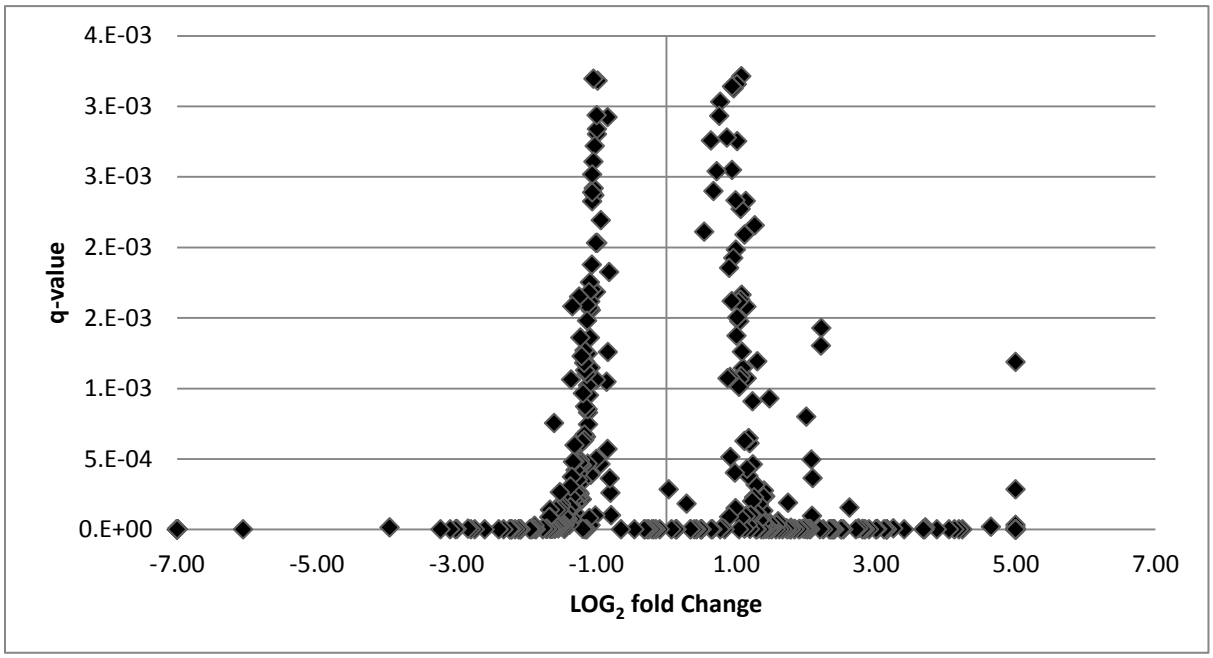


Fig. S3.4 Volcano plot of significantly differentially expressed genes in three-week tension wood vs. upright control.

Fig S3.5 Evidence for tension wood-specific expression of a tandem gene copy of a secondary cell wall cellulose synthase gene.

The genomic region of the paralogous cellulose synthase gene expressed in tension wood (top) and not in the upright control (bottom) is highlighted in the red box. The sequence of this paralog was interrogated and predicts a full length in-frame duplicate of *EgCesA3*, as annotated by Ranik and Myburg (2006).

Table S3.1 Summary of relative changes in MFA, lignin, and cell wall sugars in tension wood compared to opposite wood in five trees.

Tree	Δ MFA in TW	Δ% insol. lignin	Δ% lignin	ΔGa	ΔG	ΔX	ΔM
1	8.91%	0.04%	0.46%	170.37%	13.70%	-16.43%	-34.47%
2	-7.25%	-6.71%	-5.53%	151.83%	16.16%	-17.02%	-26.44%
3	-13.96%	-4.85%	-4.43%	198.28%	14.84%	-22.13%	-46.86%
4	-10.19%	-3.36%	-2.80%	298.24%	9.16%	-26.64%	-40.70%
5	-2.63%	-6.83%	-6.75%	267.49%	27.70%	-10.26%	-31.12%
MEAN	**-5.02%**	**-4.34%**	**-3.81%**	**217.24%**	**16.31%**	**-18.50%**	**-35.92%**
STDEV	8.82%	2.84%	2.79%	63.09%	6.89%	6.20%	8.02%
SE	3.95%	1.27%	1.25%	28.21%	3.08%	2.77%	3.59%

Table S3.2 Relative changes in holocellulose, α-cellulose and total glucose in tension wood compared to opposite wood in five trees.

Tree	Δ% holo	Δ% α-cellulose	Δ%Glucose
1	-0.02%	9.32%	13.70%
2	-1.73%	1.79%	16.16%
3	2.16%	12.44%	14.84%
4	6.71%	6.09%	9.16%
5	7.26%	-0.82%	27.70%
MEAN	**2.87%**	**5.77%**	**16.31%**
STDEV	4.00%	5.39%	6.89%
SE	1.79%	2.41%	3.08%

Table S3.3 *Arabidopsis thaliana* homologs significantly differentially expressed between tension wood and the upright control, in this study as well as in *Populus* (Andersson-Gunneråsa *et al.*, 2006).

For each ID, the description is provided as well as whether the direction of expression in tension wood relative to the upright control was shared between *Eucalyptus* and *Populus*.

Arabidopsis ID	Description	Shared poplar/*Eucalyptus*?	Direction
AT1G05010	EFE (ETHYLENE-FORMING ENZYME); 1-aminocyclopropane-1-carboxylate oxidase	YES	UP
AT1G17950	MYB52 (MYB DOMAIN PROTEIN 52); DNA binding / transcription factor	YES	UP
AT1G54100	ALDH7B4 (Aldehyde Dehydrogenase 7B4); 3-chloroallyl aldehyde dehydrogenase/ oxidoreductase	YES	UP
AT1G62990	KNAT7 (KNOTTED-LIKE HOMEOBOX OF ARABIDOPSIS THALIANA 7); DNA binding / transcription activator/ transcription factor	YES	UP
AT2G01940	nucleic acid binding / transcription factor/ zinc ion binding	YES	UP
AT3G27330	zinc finger (C3HC4-type RING finger) family protein	YES	UP
AT3G43190	SUS4; UDP-glycosyltransferase/ sucrose synthase/ transferase, transferring glycosyl groups	YES	UP
AT3G47690	zinc finger (GATA type) family protein	YES	UP
AT5G27030	FLA12	YES	UP
AT1G19300	PARVUS (PARVUS); polygalacturonate 4-alpha-galacturonosyltransferase/ transferase, transferring glycosyl groups / transferase, transferring hexosyl groups	YES	DOWN
AT2G38060	PHT4;2 (PHOSPHATE TRANSPORTER 4;2); carbohydrate transmembrane transporter/ inorganic phosphate transmembrane transporter/ organic anion transmembrane transporter/ sugar:hydrogen symporter	YES	DOWN
AT3G04730	IAA16; transcription factor	YES	DOWN

AT3G06350	MEE32 (MATERNAL EFFECT EMBRYO ARREST 32); 3-dehydroquinate dehydratase/ NADP or NADPH binding / binding / catalytic/ shikimate 5-dehydrogenase	YES	DOWN
AT3G21570	unknown protein	YES	DOWN
AT4G10270	MEE58 (MATERNAL EFFECT EMBRYO ARREST 58); adenosylhomocysteinase/ copper ion binding	YES	DOWN
AT4G13940	F5H (FERULIC ACID 5-HYDROXYLASE 1); ferulate 5-hydroxylase/ monooxygenase	YES	DOWN
AT4G36220	protein binding	YES	DOWN
AT5G07220	remorin family protein	YES	DOWN
AT5G23750	pathogenesis-related thaumatin family protein	YES	DOWN
AT5G37600	ATEB1A; microtubule binding	YES	DOWN
AT5G40020	PIP3 (PLASMA MEMBRANE INTRINSIC PROTEIN 3); water channel	YES	DOWN
AT5G60490	TPR3 (TOPLESS-RELATED 3)	YES	DOWN
AT1G06620	2-oxoglutarate-dependent dioxygenase, putative	NO	
AT2G38080	IRX12 (IRREGULAR XYLEM 12); laccase	NO	
AT3G54810	PLA IIIA (PATATIN-LIKE PROTEIN 6)	NO	
AT3G54950	wound-responsive family protein	NO	
AT4G35100	ATBAG3 (ARABIDOPSIS THALIANA BCL-2-ASSOCIATED ATHANOGENE 3); protein binding	NO	
AT5G14230	ATGSR1; copper ion binding / glutamate-ammonia ligase	NO	

Table S3.4 CAZymes upregulated in tension wood and their relative expression in seven tissues of *E. grandis*.

Data for relative expression is available on EucGeNIE (www.eucgenie.org). Colour intensity indicates relative tissue expression for each gene (0-white, 100%-red) in each of the seven listed tissues and organs.

Gene ID	Arabidopsis hits	Protein name	CAZyme	Young leaf	Shoot tips	Mature leaf	flowers	roots	phloem	Immature xylem
Eucgr.H00343	AT1G684701		GT47	0%	0%	0%	0%	0%	7%	93%
Eucgr.F00232	AT4G333301	GUX2 , PGSIP3	GT8	1%	3%	2%	3%	5%	11%	75%
Eucgr.C03199	AT3G431901	SUS4	GT4	12%	9%	8%	14%	2%	10%	46%
Eucgr.A00510	AT2G381501		GT32	0%	0%	0%	0%	66%	0%	34%
Eucgr.J01374	AT3G293201		GT35	2%	7%	2%	15%	17%	27%	30%
Eucgr.I01147	AT1G497101	FUCTB, FUT12	GT10	19%	16%	15%	11%	0%	9%	29%
Eucgr.H00536	AT3G178801	HIP, TDX	GT41	12%	11%	9%	14%	4%	25%	25%
Eucgr.K00865	AT4G152401		GT31	11%	24%	14%	13%	0%	15%	24%
Eucgr.F03658	AT1G557401	AtSIP1, SIP1	GH36	23%	13%	10%	9%	2%	23%	20%
Eucgr.E01169	AT4G194201		CE13	16%	5%	30%	10%	1%	19%	20%
Eucgr.B00354	AT1G238701	ATTPS9, TPS9	GT20	18%	20%	12%	9%	13%	9%	18%
Eucgr.B02686	AT1G680201	ATTPS6, TPS6	GT20	16%	24%	13%	15%	14%	8%	10%
Eucgr.F01855	AT1G451301	BGAL5	GH35	19%	13%	27%	19%	1%	11%	9%
Eucgr.I01697	AT3G273301		GT92	23%	19%	22%	21%	1%	7%	6%
Eucgr.B02118	AT5G288401	GME		12%	17%	12%	25%	0%	29%	6%
Eucgr.B00859	AT3G180801	BGLU44	GH1	22%	18%	21%	22%	0%	12%	5%

References

Andersson-Gunneråas S, Mellerowicz EJ, Love J, Segerman B, Ohmiya Y, Coutinho PM, Nilsson P, Henrissat B, Moritz T, Sundberg B. 2006. Biosynthesis of cellulose-enriched tension wood in *Populus*: Global analysis of transcripts and metabolites identifies biochemical and developmental regulators in secondary wall biosynthesis. *Plant Journal* **45**(2): 144-165.

Ranik M, Myburg AA. 2006. Six new cellulose synthase genes from *Eucalyptus* are associated with primary and secondary cell wall biosynthesis. *Tree Physiology* **26**(5): 545-556.

CHAPTER 4

Carbon partitioning for cellulose, hemicellulose and lignin biosynthesis during wood formation is transcriptionally hardwired

Eshchar Mizrachi[1], Karen van der Merwe[1], Charles A. Hefer[2], Gaby Mbanjo[1], Timothy J. Tschaplinski[3,4], Gerald A. Tuskan[3,4], Kathleen Marchal[5,7], Yves van de Peer[6,7], Shawn D. Mansfield[8], Alexander A. Myburg[1]

[1]Department of Genetics, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Private bag X20, Pretoria, 0028, South Africa. [2]Department of Botany, University of British Columbia, 3529-6270 University Blvd, Vancouver, V6T 1Z4, Canada. [3]US Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598 USA. [4]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, 37831, USA. [5]Department of Microbial and Molecular Systems, Katholieke Universiteit, Leuven, Belgium. [6]Department of Plant Systems Biology, VIB, Technologiepark 927, B-9052 Ghent, Belgium. [7]Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium. [8]Department of Wood Science, University of British Columbia, Vancouver, BC, Canada V6T 1Z4.

This chapter has been prepared in the format of a manuscript for a peer-reviewed research journal. The work forms part of a larger study directed by Alexander Myburg, conceived and designed by Alexander Myburg and myself. I conceived of this study, drafted the manuscript, helped sample the material used for transcriptome sequencing and metabolome profiling, and analysed the data. Karen van der Merwe mapped the RNA-seq data for the *E. grandis* BC population and helped with data analysis including R and JAVA scripting. Charles Hefer mapped the RNA-seq data for the *E. urophylla* BC population and helped with data analysis. Gaby Mbanjo and Alexander Myburg performed all trait, gene and metabolite QTL mapping. Timothy Tschaplinsky and Gerald Tuskan performed the metabolome GC-MS analysis, as well as wood property analysis including total lignin (Pyrolysis Molecular Beam Mass Spectrometry), syringyl/guaiacyl ratio, and total C5 and C6 sugars in wood samples. Shawn Mansfield helped sample the material and performed wood property analysis (density, alpha-cellulose). Kathleen Marchal, Yves van de Peer, Shawn Mansfield and Alexander Myburg helped draft and edit the manuscript.

4.1 Summary

- Understanding the biology and genetic architecture of carbon allocation and partitioning for biopolymer synthesis in plant secondary cell walls is a key priority for improvement of biomass feedstocks in the emerging bioeconomy. Here, we characterize the expression dynamics of cellulose and xylan biosynthetic genes in developing xylem across a population of field-grown *Eucalyptus* hybrid trees.
- Through combining genetic correlations of gene expression, metabolite and trait variation, we apply a gene-targeted systems genetics approach to define an important gene expression module in *Eucalyptus*, termed the “secondary cell wall (SCW) *CesA* regulon”, which encompasses a group of biologically related genes and biochemical pathways that tightly covary with SCW cellulose synthase expression at the population level.
- The analysis reveals transcriptional hardwiring of cellulose and heteroxylan related pathways, as well as pathways involved in primary metabolite production and intracellular transport. Our analysis also provides evidence of coordination of SCW polysaccharide biosynthesis and shikimate pathway activity for the production of lignin precursors, with fructose playing a key role in providing substrate. These results are supported by multiple levels of investigation of component and complex traits, including gene expression and metabolite variation in developing xylem, genome (QTL) mapping and corresponding wood and biomass-related tree phenotypes.
- This study is the first to demonstrate the extent of co-regulation and metabolic feedback within the SCW polysaccharide biosynthetic program in trees, and provides critical insight into carbon partitioning between polysaccharides and lignin in developing xylem.

4.2 Introduction

In woody plants the secondary cell walls (SCWs) of xylem fibre cells constitute the bulk of plant biomass and consists of three major biopolymers – cellulose, hemicellulose and lignin. The relative abundance, chemical composition and arrangement of these biopolymers is a major determinant of not only the mechanical properties required for the stature and longevity of trees (Lucas *et al.*, 2013; Schuetz *et al.*, 2013), but also the efficiency of their mechanical or chemical breakdown in industrial applications – including pulp, paper and chemical cellulose production in particular, and in the future second generation biofuels and biomaterials (Hinchee *et al.*, 2009; Mansfield, 2009; Sannigrahi *et al.*, 2010; Mizrachi *et al.*, 2012). Fibre SCW composition varies among tree genera, species and populations (Pauly & Keegstra, 2008; Carroll & Somerville, 2009), and is dependent on the interaction and regulation of multiple biochemical pathways and biological processes (Groover, 2005; Groover & Robischon, 2006; Spicer & Groover, 2010). This complexity is illustrated at the level of the genome, where many loci contribute to trait variation (Resende *et al.*, 2012). This also predicts that selection for- or manipulation of these traits at individual loci could impact many epistatic interactions, and could have unpredictable pleiotropic effects. Indeed, studies reporting manipulation of expression of individual biosynthetic genes related to cellulose or lignin biosynthesis in plants not only affect the trait of interest, but also often are associated with unanticipated detrimental phenotypes and responses at the level of transcriptome, metabolome, plant development and SCW properties (Taylor *et al.*, 2000; Vanholme *et al.*, 2008; Joshi *et al.*, 2011). However, influencing processes upstream of individual pathways (e.g. lignin or cellulose biosynthesis), for example increasing carbon allocation to polysaccharide synthesis, have demonstrated potential in improvement of biomass and/or SCW deposition in trees (Coleman *et al.*, 2006; Coleman *et al.*, 2009; Park *et al.*, 2009).

In plants, the acquisition, storage and utilization of carbon must be carefully regulated to optimize conditions for survival and growth (Smith & Stitt, 2007; Stitt & Zeeman, 2012). Especially in trees where

the bulk of sequestered carbon is channelled to secondary xylem formation, the allocation and partitioning of carbon for polysaccharide and lignin biosynthesis is an expensive process, and predicts that the regulation of this process would be under selection to ensure coordination of the necessary biological processes. This would include the full complement of enzymes responsible for synthesizing cellulose, hemicellulose and lignin. Conceivably, it would also include the pathways responsible for the production of the required precursor metabolites for these biopolymers, such as UDP-glucose, UDP-xylose and the monolignols. For example, it has been hypothesized that primary sugar metabolism for the production of sugar-nucleotide precursors (e.g. enzymes involved in producing and providing active metabolites such as UDP-glucose or UDP-xylose) would be co-regulated with SCW polysaccharide biosynthesis (Somerville, 2006). It has furthermore been suggested that in addition to being spatiotemporally co-regulated, these enzymes would be physically associated with the enzymatic complexes synthesizing cellulose and hemicellulose (Mansfield, 2009; Oikawa *et al.*, 2013). Some evidence for this has emerged in studies demonstrating membrane-bound isoforms of sucrose synthase (Amor *et al.*, 1995) – which are physically close to and probably interact with the cellulose synthase complex (CSC; Salnikov *et al.*, 2001) – as well as membrane bound, Golgi-localized forms of UDP-xylose-synthase (UXS), likely providing UDP-xylose directly to the xylan synthesizing protein complex (Pattathil *et al.*, 2005).

Much of the coordination of SCW biopolymer synthesis occurs at the transcriptional level. This has been one of the most important factors in accelerating gene discovery of SCW related genes, where transcript coexpression metadata (mainly from multiple conditions and/or organs/tissues of *Arabidopsis thaliana*) have been employed to identify genes and biological processes essential for cellulose (Brown *et al.*, 2005; Persson *et al.*, 2005; Mentzen & Wurtele, 2008; Mutwil *et al.*, 2009; Mutwil *et al.*, 2010; Ruprecht *et al.*, 2011), xylan (Brown *et al.*, 2005; Brown *et al.*, 2007; Oikawa *et al.*, 2010) and lignin (Vanholme *et al.*, 2012) biosynthesis. Homologous genes in woody species such as *Populus spp*. have also been identified (Hertzberg *et al.*, 2001; Schrader *et al.*, 2004; Geisler-Lee *et al.*, 2006), demonstrating the conserved

nature of these programs in angiosperms. Concomitantly, the regulators of these structural genes are being progressively identified, and demonstrate the cross-talk within the regulatory hierarchy for cellulose, xylan and lignin (Zhong *et al.*, 2008; Zhong & Ye, 2009; Yamaguchi & Demura, 2010; Zhao & Dixon, 2011; Hussey *et al.*, 2013), and the conservation of this regulatory network among angiosperms (Mutwil *et al.*, 2011; Zhong *et al.*, 2011; Hussey *et al.*, 2013).

Despite this, although the expression of individual genes involved in cellulose and xylan biosynthesis has been shown to be correlated using meta-analyses (Ma *et al.*, 2007; Mutwil *et al.*, 2009; Obayashi *et al.*, 2009; Mutwil *et al.*, 2011), the extent of co-regulation of these pathways is yet to be fully understood, especially at the population level where co-regulation would contribute to genetic correlation of biopolymer synthesis (see below). For example, it is unclear to what extent (both in terms of transcriptional regulation and carbon availability) cellulose and xylan pathways are co-regulated, or to what extent carbon partitioning to lignin is coordinated with these processes. Additionally, information is lacking with regards to the regulation of essential biological processes that contribute to SCW formation, such as primary metabolism, cortical microtubule arrangement, cell signalling and protein/carbohydrate transport (reviewed in Mizrachi *et al.*, 2012).

Complementary to meta-analysis of expression data from many diverse experiments in uniform genetic backgrounds is the interrogation of similar data from related, but genetically diverse individuals. Increasingly, the power of integrative analysis of quantitative component traits (e.g. gene expression, metabolite availability) from genetically segregating populations is providing insight into complex (*i.e.* multi-gene) biological traits such as development, behaviour and disease (Schadt *et al.*, 2008; Ayroles *et al.*, 2009; Edwards *et al.*, 2009; Drost *et al.*, 2010; Zhang *et al.*, 2013). A valuable application of this approach is the identification of known and unknown genes and/or metabolites using “guilt-by-

association", as well as understanding the interaction/interdependence between different biological pathways (Zhu *et al.*, 2008; Baldazzi *et al.*, 2012; Mizrachi *et al.*, 2012). Importantly, this "systems genetics" approach provides information on factors such as genotypic and phenotypic plasticity (Zhou *et al.*, 2012), pleiotropy (Zhu *et al.*, 2008; Ayroles *et al.*, 2009; Drost *et al.*, 2010) and epistasis (Huang *et al.*, 2012) with regards to transcriptional regulation, and where applicable, the impact of variation in component (e.g. molecular) traits on complex phenotypic traits.

In this study we applied a gene-centred systems genetics approach and identified the extent of transcriptional hardwiring of SCW polysaccharide biosynthesis (represented almost exclusively by cellulose and xylan biosynthetic genes) in the developing xylem of hybrid *Eucalyptus* trees. This was done by identifying genes significantly coexpressed in developing xylem tissue with a target secondary cell wall-specific cellulose synthase (*CesA*) gene, measured in 282 individuals from two previously reported (van Dyk *et al.*, 2011; Kullan *et al.*, 2012a) F2 hybrid (*E. grandis* x *E. urophylla*) backcross populations. We show that transcript abundance for all genes known to be necessary for cellulose, xylan and glucomannan biosynthesis, but not phenylpropanoid or primary cell wall polysaccharide biosynthesis, tightly covary in developing xylem in a regulon containing around 1.1% of the genes in the *Eucalyptus* genome. In addition, we highlight other important processes co-regulated with cellulose, xylan and glucomannan synthesis, which include cytoskeletal organization, intracellular transport and primary metabolite production for these biopolymers. Finally, evidence from gene expression and quantitative metabolite availability in xylem indicates that cytosolic fructose, which would be a major breakdown product of sucrose synthase (SUSY), could be a newly discovered link between the carbon being utilized for polysaccharide biosynthesis and the production of lignin precursors during active xylem development. Understanding the intrinsic programming of carbon allocation during SCW deposition is best demonstrated in woody plants, but will have widespread implications for all biomass-related research.

4.3 Materials and methods

4.3.1 Plant material and transcriptome profiling

For detailed description of sample preparation and RNA sequencing see Kullan *et al.* (2012b). Briefly, developing xylem tissue from two families containing F2 progeny of a previously described interspecific pseudo-backcross of an F_1 interspecific hybrid individual ("GUSAP1") of *E. grandis* and *E. urophylla* (van Dyk *et al.*, 2011; Kullan *et al.*, 2012a) was collected as previously described (Ranik *et al.*, 2006). Samples were collected from three-year-old trees over a 7½ hour period between 09:00 and 16:30. In total, 154 transcriptomes were sequenced from the developing xylem tissues of F_2 progeny from one family (*E. urophylla* x GUSAP1, henceforth referred to as "*E. urophylla* BC") and 128 progeny from a second family (*E. grandis* x GUSAP1, henceforth referred to as "*E. grandis* BC"). Expression values (Fragments Per Kilobase of coding sequence per Million mapped fragments – FPKM) were calculated genome-wide for each sequenced individual with Cufflinks V1.0.3 (Trapnell *et al.*, 2010) using the JGI v.1.1. *E. grandis* gene models as a reference (www.phytozome.net). Following FPKM calculations for all samples, we considered only genes with evidence of expression (FPKM>0) in at least 25% of the individuals in the two respective BC populations. This amounted to 27,337 genes and 27,894 genes (each termed "xylem transcriptome" henceforth) in the *E. urophylla* BC and *E. grandis BC* populations, respectively.

4.3.2 Metabolome profiling

Approximately 50 mg (fresh weight) of developing xylem tissue were twice extracted with 2.5 mL 80% ethanol overnight and then combined prior to drying a 1.0-ml aliquot in a nitrogen stream. Sorbitol was added (to achieve 15 ng/µL injected) before extraction as an internal standard to correct for differences in

extraction efficiency, subsequent differences in derivatisation efficiency and changes in sample volume during heating. Dried extracts were dissolved in 500 µL of silylation-grade acetonitrile followed by the addition of 500 µL N-methyl-N-trimethylsilyltrifluoroacetamide (MSTFA) with 1% trimethylchlorosilane (TMCS) (Thermo Scientific, Bellefonte, PA), and samples then heated for 1 h at 70 °C to generate trimethylsilyl (TMS) derivatives (Jung *et al.*, 2009; Li *et al.*, 2012). After 2 days, 1-µL aliquots were injected into an Agilent Technologies Inc. (Santa Clara, CA) 5975C inert XL gas chromatograph-mass spectrometer, fitted with an Rtx-5MS with Integra-guard (5% diphenyl/95% dimethyl polysiloxane) 30 m x 250 µm x 0.25 µm film thickness capillary column.

Metabolite peaks were extracted using a key selected ion, characteristic m/z fragment, rather than the total ion chromatogram, to minimize integrating co-eluting metabolites. Peaks were quantified by area integration and the concentrations were normalized to the quantity of the internal standard (sorbitol) recovered, amount of sample extracted, derivitized, and injected. A large user-created database (>1900 spectra) of mass spectral electron ionization (EI) fragmentation patterns of TMS-derivatized compounds, as well as the Wiley Registry 8th Edition combined with NIST 05 mass spectral database, were used to identify the metabolites of interest to be quantified. Unidentified metabolites were denoted by their retention time as well as key mass-to-charge (m/z) ratios.

4.3.3 QTL mapping and visualization

The *E. urophylla* BC parent and F1 hybrid (GUSAP1) genetic linkage maps previously developed by Kullan *et al.* (2012a) were used for trait dissection of selected chemical and physical wood properties. Wood properties included in this study were made available through Sappi Forest Research, Oak Ridge National Laboratory, University of British Columbia and the Forest Molecular Genetics Programme (University of Pretoria). These included wood density, cellulose, syringyl/guaiacyl ratio, total lignin, total

C5 sugar in wall, total C6 sugar in wall, as well as total cell wall carbohydrate (C5+C6) to lignin ratio. The chemical wood properties were assessed using different analytical methods including pyrolysis molecular beam mass spectrometry (pyMBMS) and near-infrared analysis (NIRA). In addition to the above-mentioned wood properties, selected metabolites thought to play a key role in polysaccharide biosynthesis were included for marker-trait association.

QTL detection was carried out using the composite interval mapping (CIM) module (Zeng, 1994) implemented in the software Windows QTL Cartographer version 2.5 (Wang *et al.*, 2007). CIM was conducted using model 6. Forward and backward stepwise regression (p = 0.1) was used to select the most significant cofactors to control for background segregation. QTL detection was performed in 1 cM map intervals and a window size of 10 cM was chosen for the test interval. Permutation tests (1000) were performed for each trait at $\alpha = 0.05$ and $\alpha = 0.01$ to empirically estimate the genome-wide significant logarithm of odds (LOD) threshold for declaring a QTL. QTL confidence intervals corresponding to a LOD score drop of 1.0 (LOD = 1.0) on either side of likelihood peak was automatically determined by the software and the values used to draw significant QTL in the maps. QTL graphs were drawn using MapChart 2.2 (Voorrips, 2002).

4.3.4 Statistical analysis, annotation and enrichment analysis

Transcript abundance correlations (Pearson correlation coefficients) were calculated using FPKM values of gene expression in the xylem transcriptome of each BC population by customized R scripts. Distribution of correlations was calculated using customized R and JAVA scripts, and visualized using SPSS V20. Statistical analyses were performed using SPSS V20. Gene Ontology over-representation analyses were carried out using the BiNGO plugin (Maere *et al.*, 2005) for Cytoscape (Smoot *et al.*, 2011). Pathway analysis was carried out using the KEGG (http://www.genome.jp/kegg/) online database

(Kanehisa & Goto, 2000). *Cis*-element analysis was carried out using the RSAT online database (Thomas-Chollier *et al.*, 2008).

4.4 Results

4.4.1 Identification, definition and expression dynamics of a "SCW *CesA* regulon" in *Eucalyptus* xylem

We sequenced the developing xylem transcriptomes of two families containing F2 progeny of an *E. grandis* x *E. urophylla* hybrid backcross (BC) pedigree (see Materials and Methods). From both the *E. urophylla* BC and the *E. grandis* BC data (full xylem transcriptome data from 156 and 126 individuals, respectively), the top 1% of genes for which the expression profile correlated to that of a target gene, *EgCesA3* – Eucgr.C00246, ortholog of *AtCesA7* in *Arabidopsis thaliana* (Ranik & Myburg, 2006) and *PtiCesA7A/B* in *Populus trichocarpa* (Kumar *et al.*, 2009) – were considered. To define the "SCW *CesA* regulon" the union of the top 1% of genes whose transcript abundance profiles correlated with that of *EgCesA3* from either datasets was taken. This regulon contained a non-redundant set of 422 genes (Table S4.1), of which 125 (30%) were shared in both datasets and 297 occurred in either of the datasets. Pearson correlation coefficients of these genes with *EgCesA3* ranged from 0.74-0.96 (*E. urophylla* BC) and 0.84-0.99 (*E. grandis* BC). These correlation values were above the 99.9th and 99.7th quantile of global (*i.e.* all vs all, Fig. S4.1) correlations within each respective dataset, and were therefore considered highly significant.

Given the experimental design, high correlation of transcript abundance (that is not random) could either be a function of common genetic regulators segregating in the F2 progeny, or coordination via environmental or cellular feedback mechanisms that act on the same transcriptional network, and hence

enhance the correlation of gene expression, or both. We hypothesized that if high correlation of transcript abundance is indeed due to co-regulation and not randomness, these genes would: i. share common *cis*- or *trans*-regulators, ii. share common *cis*-regulatory elements, and iii. share common biological functions or closely related steps in biochemical pathways. Indeed, quantitative analysis of expression of genes in the regulon revealed that although genes in the regulon were distributed throughout the genome, a large number of significant expression quantitative trait loci (eQTLs) mapped to shared genomic regions within each population, indicating they share segregating trans-regulators (Fig. S4.2). In total, genes in the SCW CesA regulon had 386 eQTLs (220 genes) in the *E. urphylla* BC, and 209 eQTLs (148 genes) in the *E. grandis BC*. The eQTLs were generally for different genes and mapped to different locations in the two backcross populations (mainly chromosomes 6, 8, 9 and 10 in the *E. urophylla* BC and chromosome 3 in the *E. grandis* BC – Fig. S4.2). The differences in eQTL locations between the populations may be indicative of different segregating components of a common transcriptional network.

We also searched the promoter regions (1000 bp upstream) of all genes in the SCW CesA regulon for shared *cis*-regulatory elements previously reported to be conserved in *CesA* promoters (Creux *et al.*, 2008; Creux *et al.*, 2013). We identified several important cis-regulatory elements that were overrepresented in promoters of the regulon genes compared to a random dataset (Fig. S4.3). In particular, the *cis*-acting elements PYRIMIDINEBOXOSRAMY1A (sugar repression, response to GA), CTRMCAMV35S (CT-rich enhancer element), CRPE31 (unknown function), CRPE8 (Anthocyanin regulatory element), and CRPE10 (vascular-specific expression) were overrepresented compared to a random dataset. The CRPE31 element occurred in 320 (80%) of the 422 genes, while the CTRMCAMV35S element occurred in 182 (43%) genes in the regulon. These elements, and particularly their joint spatial conservation when co-occurring, have previously been described for *CesA* promoters specifically in *Eucalyptus* species (Creux *et al.*, 2013).

Ontology overrepresentation analysis of the 422 genes revealed several categories that were significantly overrepresented (FDR-corrected $P < 0.05$, Fig. S4.4). The most significant of these included carbohydrate metabolism (specifically cellulose and glucuronoxylan biosynthesis), cell growth and cellular component organization (most significantly "plasma membrane" and "Golgi" related ontologies). The only ontologies significantly underrepresented included those relating to transcription and translation. Of the 422 genes in the regulon, 58 (14%) were Carbohydrate Active enZymes (CAZymes – Table 4.1), which is approximately double the proportion of total CAZymes in the genome (i.e. 7% – Pinard *et al*., in preparation). Analysis of Kegg Ontology (KO) terms highlighted specific pathways in the broad categories of starch/sucrose metabolism, lipid, carbohydrate, amino acid and energy metabolism (Fig. S4.5). Generally, these KO terms represented sequential steps in biochemical pathways. Together these results suggest that genes in the SCW *CesA* regulon are co-regulated structural genes that are biologically- and pathway-related, involved mainly in SCW polysaccharide biosynthesis, and whose co-variation can be at least partially be explained by common regulators segregating in the populations.

To examine the expression dynamics of this regulon we investigated the variation in expression of these genes, both in- and across the two backcross populations, as well as in a tree developmental context. We analysed the expression patterns of these genes in a previously established *E. grandis* expression dataset, in which we sequenced the transcriptomes of a diverse set of samples including tissue from shoot tips, young leaves, mature leaves, developing xylem, phloem, roots and flowers of a rotation-age *E. grandis* clone (www.eucgenie.org, Hefer *et al*., in preparation). Analysis of xylem specificity (*i.e.* relative expression in developing xylem compared to six other tissues and organs) of expression of these genes showed that a large proportion of the genes in the regulon were not necessarily xylem specific, although proportionally genes in the regulon did tend to have a higher specificity compared to the rest of the genes

in the genome (Fig. 4.1A). Additionally, the distribution of the coefficient of variation (CV – defined as the standard deviation to mean ratio $CV = \frac{\sigma}{\mu}$) values of gene expression in the regulon showed that the variation in expression of these genes in the two backcross families was lower relative to all genes in the xylem transcriptome (Fig. 4.1B and 4.1C). We also considered temporal dynamics of gene expression, since aspects of wood formation are known to be influenced by circadian regulation (Solomon *et al.*, 2010). The expression of three important clock genes, *LHY/CCA1*, *TOC1* and *GI*, demonstrated temporal variation throughout the period of sample collection (Fig. 2A) that was consistent with known relationships between these genes (e.g. LHY being a repressor of *TOC1* - Alabadí *et al.*, 2001), and previous investigation of these genes in two *Eucalyptus* hybrid clones (Solomon *et al.*, 2010); Fig. S4.6). In contrast, the three dominantly expressed SCW *CesA* genes did not show evidence of day time-related variation during the sampling period (Fig. 4.2B). The high correlation and low variation of expression of these genes in the two backcross families were consistent with genes essential for a conserved developmental process such as SCW polysaccharide deposition.

4.4.2 The SCW *CesA* regulon contains all known genes necessary for SCW polysaccharide (cellulose, xylan and glucomannan) biosynthesis, as well as those coding for the necessary sugar-nucleotide interconversion enzymes.

An interrogation of the closest *Arabidopsis thaliana* BLAST hits to *Eucalyptus* genes in the regulon revealed putative homologs for all known genes involved in secondary cell wall polysaccharide metabolism (Table S4.1, refer to Chapter 1 and Supplemental Note S4.1 for detailed information of genes involved in cellulose and xylan biosynthesis). This included homologs of all genes known to be necessary for cellulose biosynthesis (homologs of the three SCW-specific cellulose synthase genes *IRX1*, *IRX3* and *IRX5*, *IRX2/KOR1*, *IRX6/COBL*, *IRX13/FLA11*, *CSI1*, *CTL2*) and xylan primer and backbone synthesis (*PARVUS*, *IRX7*, *IRX8*, *IRX9*, *IRX10*, *IRX14*), as well as xylan acetyl (*RWA2*, *RWA3*, *ESK1*) and

glucuronic/methylglucuronic acid (*GXM*, *GUX1*, *GUX2*) side-chain modification. Homologs for the *CSLA9* gene, essential for glucomannan biosynthesis (Dhugga *et al.*, 2004; Liepman *et al.*, 2005), as well as the recently described *MSR2* gene (Wang *et al.*, 2013), were also present in the regulon. In the context of SCW polysaccharide biosynthesis, the genes comprise the full suite of genes known to be required for cellulose, heteroxylan and glucomannan biosynthesis. Although CSLD proteins have recently been shown to be essential for mannan synthesis (Verhertbruggen *et al.*, 2011), the absence of *CSLD* genes in the regulon supports the conclusions of Verhertbruggen and colleagues (2011) that CSLA-derived glucomannan is the dominant mannan type in SCW, while pure mannan (CSLD-derived) plays a minor role in the SCW. Importantly, no known genes involved in pectin biosynthesis (Atmodjo *et al.*, 2013), xyloglucan biosynthesis (Chou *et al.*, 2012) or xyloglucan remodelling (xyloglucan transglucosylase/hydrolase family members, see Eklöf & Brumer, 2010 for review) were present in the regulon, suggesting these genes are regulated in expression regulons different from that of SCW polysaccharide biosynthesis.

In addition to these polysaccharide biosynthetic genes, genes coding for key enzymes involved in carbon metabolism related to cellular sucrose regulation, sucrose catabolism and sugar nucleotide interconversion were also present in the SCW *CesA* regulon (notably *SUSY4*, *UGD* and *UXS*, the enzymatic products of which would enable the production of UDP-glucose, UDP-glucuronic acid and UDP-xylose, respectively). Two additional sucrose-related regulators whose function has not yet been resolved were also present in the regulon. First – SWEETIE – which has been shown in *Arabidopsis* to be essential in carbon utilization for growth and development, with *sweetie* mutants displaying hypersensitivity to sucrose and glucose (Veyres *et al.*, 2008). Second – SUCROSE TRANSPORTER 2 (SUT2) – which has been proposed to play a role either in sucrose loading from phloem into sink cells (Payyavula *et al.*, 2011; Milne *et al.*, 2013), or symplastic efflux of stored sucrose from the vacuole (Etxeberria *et al.*, 2012). In the case of the latter mechanism, the action of SUT2 would be particularly relevant if sucrose is in high

abundance in surrounding cells and it would be assimilated into cells *en masse* via mechanisms such as fluid phase endocytosis (Etxeberria *et al.*, 2009; Bandmann & Homann, 2012). This would be expected in strong sucrose sink tissues such as the xylem of woody plants, and indeed *SUT2* expression has mainly been observed in sink tissues (Barker *et al.*, 2000). The transcriptional coordination of these genes with polysaccharide biosynthetic genes provides evidence for a coordinated metabolite pool production for and during both cellulose and xylan biosynthesis.

4.4.3 Carbon allocation for polysaccharide biosynthesis is transcriptionally co-regulated with cytoskeleton organization and intracellular transport.

CSCs are regulated at various levels, and their trafficking to distinct regions of the plasma membrane is an intensively researched field, with evidence for essential roles of both actin and cortical microtubules (Wightman & Turner, 2008; Wightman *et al.*, 2009; Crowell *et al.*, 2010). The delivery and internalization of CSCs to and from the plasma membrane occurs through specific Golgi-derived bodies, termed MASCs/SMAccs (Crowell *et al.*, 2009; Gutierrez *et al.*, 2009), although the secretory pathway and proteins involved in this are still unknown. A likely model is thought to involve actin transport of Golgi bodies via the myosin XI-K (Peremyslov *et al.*, 2012), with KINESIN13A playing a critical role in recognition of the microtubule-associated protein RIP3/MIDD1 (Mucha *et al.*, 2010; Cai, 2011) to facilitate cortical microtubule depolymerisation. Depolymerisation of microtubules at both ends is a conserved feature of KINESIN13 family members (Asenjo *et al.*, 2013), and in cortical microtubules the proposed effect of this (microtubule scaffolding) is a key component of cortical microtubule-dependent delivery of CSCs to the membrane (Gutierrez *et al.*, 2009). Several actin arrangement-related genes with previously established cell wall associated phenotypes are apparent in the regulon, including *SAC1/FRA7* (Zhong *et al.*, 2005), *NET1A* (Deeks *et al.*, 2012) and *ITB1/SCAR2* (Basu *et al.*, 2005). The membership of myosin *XI-K*, *KINESIN13A* and *RIP3/MIDD1* in the SCW *CesA* regulon, as well as other important genes coding for the cortical microtubule-arrangement proteins AUG5, TUB6, SPR2/TOR1, MOR1,

RIC1, CLASP, KATANIN SUBUNIT B PROTEIN, ZWI/KCBP, TRM30 and AIR9 (Ambrose & Wasteneys, 2008; Ambrose *et al.*, 2011; Fishel & Dixit, 2013; Gardiner, 2013; Lin *et al.*, 2013) and known cortical microtubule-CSC interface proteins (FRA1, CSI1 - Zhong *et al.*, 2002; Bringmann *et al.*, 2012), indicates that actin organization, cortical microtubule arrangement, as well as transport and membrane-delivery of the CSCs are tightly transcriptionally co-regulated with the *CesA* genes. In addition, the necessary components of the TRAPP II complex (*TRS120* and *TRS130* - Qi *et al.*, 2011) and *VHA-a1* (Crowell *et al.*, 2009) were present, both of which would play a role in post-Golgi trafficking.

The role of clathrin mediated endocytosis (CME) in CSC-internalization has been debated, mainly because the sizes of clathrin coated vesicles would be smaller than the cytoplasmic dimensions of a CSC (Bowling & Brown Jr, 2008; Crowell *et al.*, 2009). However, recently Bashline and colleagues (2013) have shown that in primary cell wall biosynthesis, CME may play a partial role in the internalization of CSCs via interaction of the CSRII or P-CR region of the *At*CESA3 and *At*CESA6 proteins with the µ2 subunit of the AP2 adaptor complex (Bashline *et al.*, 2013). Several genes present in the regulon are involved in CME, including *EPSIN2*, *AP180* and genes coding for ENTH/ENTH-VHS domain proteins, clathrin heavy-chain linker proteins, *VAN7/GNOM* (Naramoto *et al.*, 2010) and AP2 complex α-subunit, suggesting that CME may be important for SCW cellulose biosynthesis during xylem formation as well. CME could be a potential recycling mechanism for faulty CSCs that need to be removed during the formation of CSC arrays. Other genes involved in endocytosis or secretion previously shown to adversely affect cell wall deposition were also apparent in the regulon, such as the dynamin-like phragmoplastins *DRP1A* and *DRP2C* (Konopka & Bednarek, 2008; Hirano *et al.*, 2010; Taylor, 2011), as well as components of the exocyst complex *SEC5*, *SEC8* and *EXO70A* and *DUF810* (Goonesekere *et al.*, 2010; Li *et al.*, 2013).

4.4.4 A proposed role for cytosolic fructose in lignin precursor synthesis during polysaccharide biosynthesis

Genes in the regulon involved in carbon allocation included SUSY, which is involved in sucrose catabolism for the availability of UDP-glucose. The other product of this reaction is cytosolic fructose, produced equimolar to UDP-glucose, which is thought to be converted to fructose-6-phosphate and recycled into UDP-glucose (Haigler *et al.*, 2001). Carbon availability and partitioning to lignin is not well established, though the roles of transaldolase (Vanholme *et al.*, 2012) and transketolase (Henkes *et al.*, 2001) have been previously highlighted to channel carbon to the shikimate pathway for phenylalanine (lignin precursor) synthesis. The 428 genes in the SCW *CesA* regulon contained only one gene from the phenylpropanoid pathway (*F5H*, Eucgr.J02393), as well as three genes that would be putatively involved in G- and S-lignin polymerization (Berthet *et al.*, 2011) (*LAC4/IRX12* and two *LAC17* homologs) that together account for around 80% of xylem laccase expression in the *E. grandis* genome (Table S4.2). However, several genes representing closely related steps early in the phenylalanine biosynthetic pathway were present (Fig. S4.7). If the transcription-level coordination of this pathway with SCW polysaccharide biosynthesis is reflective of coordination of these processes, the xylem sink tissue could be simultaneously utilizing a common carbon source for polysaccharide metabolism and phenylalanine (lignin precursor) synthesis. This hypothesis is especially attractive since the main carbon source for the shikimate pathway (erythrose-4 phosphate – E-4P) could be derived from fructose via the non-oxidative pentose phosphate pathway. Fructose would be a readily available product of catabolized sucrose during active polysaccharide biosynthesis. In support of this, genes coding for the two enzymatic steps required to produce E-4P (fruktokinase, Eucgr.A00095 and transketolase, Eucgr.D02466) were indeed present in the regulon.

To test the hypothesis that cytosolic fructose in xylem is utilized for the production of phenylalanine, we measured metabolite variation using Gas Chromatography-Mass Spectrometry (GC-MS) in the same

xylem samples used for transcriptomics from 154 individuals in the *E. urophylla* BC population. Among the metabolites measured were cytosolic sucrose, glucose, fructose and shikimic acid (an intermediate of phenylalanine biosynthesis). We hypothesized that co-variation (*i.e.* positive correlation) of fructose and shikimic acid in the population would provide an independent line of evidence that these metabolites are related, that they are simultaneously metabolised and that fructose is utilised (in part) for the shikimate pathway. A Principle Component Analysis (PCA) of the metabolite levels in the 154 individuals identified two major components with eigenvalues above 1 that cumulatively explain 79% of the variance (Table S4.3). Fructose and shikimic acid both loaded on the first principle components with values greater than 0.705 (Fig. 4.3, this threshold represents >50% overlapping variance and is considered "excellent" - Comrey & Lee, 1992). The second component contained mainly sucrose, with glucose loading similarly on both components. To see whether any relationship exists between the variation of expression of the SCW *CesA* regulon and these metabolites, we tested for correlation between *EgCesA3* (a proxy gene for the regulon) and the two principle components. We found a significant negative correlation ($N = 154$, $r = -0.424$, $P < 1e^{-4}$) between *EgCesA3* and PC1, but no correlation with PC2.

Transcriptomic and metabolomic evidence therefore supports a model of coregulation and coordination of SCW polysaccharide biosynthesis in xylem (Fig. 4.4). In this model SCW polysaccharide biosynthesis is tightly coordinated with essential substrate metabolism and intracellular transport, with at least some of the sucrose-derived fructose moieties being shunted to the shikimate pathway. While it is generally accepted that much of the cytosolic fructose is recycled for sucrose, glucose or UDP-glucose production (Haigler *et al.*, 2001), the utilization of the available fructose for energy and a shunt towards the shikimate pathway would be a parsimonious solution to carbon partitioning and highlights the role of erythrose-4-phosphate production, via transaldolase (Vanholme *et al.*, 2012) and transketolase (this study) in this process. Although a direct correlation between regulon expression and wood phenotype could not be identified in this study, we did identify shared quantitative trait loci for metabolite availability in

developing xylem and some resulting traits in wood, notably fructose (developing xylem) with wood density, and shikimic acid (developing xylem) with cellulose content in wood (Fig. S4.9). Together, these separate levels of evidence support the fact that the downstream fructose availability and metabolism during active polysaccharide biosynthesis is a key junction in carbon partitioning during xylem formation, a fact that will be useful for future strategies to modify these pathways.

4.5 Discussion

Central to the biology of cell wall biosynthesis in woody species is the fact that the main source of carbon for large-scale investment in polysaccharide and lignin biosynthesis, sucrose and its direct derivatives glucose, UDP-glucose and fructose, are also core metabolites for many downstream processes, including physiological and cellular homeostasis. At an organismal level, homeostasis must be maintained between storage and investment in growth and development (e.g. the synthesis and breakdown of sequestered carbon in plants - Smith & Stitt, 2007; Stitt & Zeeman, 2012), while at a cellular level these sugar precursors are selectively channelled towards the production of building blocks of increasing complexity such as more diverse sugar substrates, amino acids, nucleotides, fatty acids and co-factors (Csete & Doyle, 2004). Metabolism involving direct utilization of core metabolites such as these in both prokaryotic and eukaryotic organisms is usually central to a bow-tie architecture, at the centre of which are core precursors that are utilised to produce more specialised heterogeneous components as required by the organism (reviewed in Csete & Doyle, 2004). Genes and proteins interacting with these core metabolites generally display higher connectivity (e.g. coregulation and protein-protein interactions) and are generally as a group under selection for robustness, with an obligate trade-off that inherently results in disproportionate fragility (Carlson & Doyle, 2000; Kitano, 2004; Whitacre, 2012).

In this study we demonstrate that the characteristic properties of genes involved in carbon partitioning for SCW deposition are reflective of their dependence on sucrose and its derived metabolites in this architecture, namely: i. high transcriptional co-regulation of genes involved in cellulose and SCW hemicellulose biosynthesis, the products of which both rely on UDP-glucose as an initial input; ii. low variation in transcript abundance compared to other genes in the genome, indicating less fluctuation and tighter regulation; and iii. feedback between transcript abundance and core metabolites that their gene products would process, suggesting a homeostatic relationship between metabolites and transcript abundance. This relationship is consistent with a "closed-loop" behaviour (Csete & Doyle, 2002), and would be imperative to maintain the balance between cell wall synthesis and competing/complementary pathways central to cellular homeostasis, such as glycolysis, the pentose phosphate pathway and amino acid biosynthesis. The "fragility" of this system is demonstrated by the fact that a disproportionately high number of genes known to be essential for normal xylem formation were present in the same regulon. That is, out of 15 genes known to be essential for normal xylem formation (i.e. *IRX* genes), 13 were represented in this regulon. Additionally, the regulon contained many other genes whose homologs are known to have catastrophic effects on SCW or carbohydrate metabolism when mis-regulated or knocked out (e.g. *SWEETIE*, *CSI*, *FRA1*, etc.). The vast scale of carbon utilization for SCW biosynthesis must therefore require that these genes be tightly controlled and balanced against other cellular functions.

These findings have several implications. First, while gene-expression meta-analyses in *Arabidopsis* have captured important elements of xylem development, genes in these defined regulons often reflect a confounding effect of highly correlated developmental stages (cell patterning, maturation, development, SCW deposition, and PCD) and a low sampling resolution for stages of SCW development. By contrast, the genetic approach applied here, sampling across a large number of segregating genotypes, has provided the resolution and power required to decouple the SCW polysaccharide biosynthesis programme from other SCW related processes such as pectin, xyloglucan and lignin biosynthesis, and has demonstrated a

potential link between metabolic flux for polysaccharide and lignin deposition in secondary cell wall xylem. Second, a conventional systems biology approach to study and model SCW biosynthesis using single-gene perturbations may be at risk of misrepresenting/over-representing the roles and effects of individual genes, proteins, metabolites and other components during normal development, since complete perturbation of individual components frequently either leads to no effect or catastrophic failure. Pleiotropic effects on the plant in these conditions could reflect an imbalance in other tightly connected pathways, making it difficult to dissect direct effects. In this sense, a systems genetics approach in a segregating population of phenotypically "wild-type" individuals offers an attractive tool to better understand normal gene function and metabolic flux in an operational system, and should be seen as complementary to traditional systems biology approaches (e.g. Vanholme *et al.*, 2012). Finally, the findings in this study imply a potential canalisation of pathways central to several biomass feedstock traits that are industrially and commercially important. Any potential biotechnological manipulation of these pathways, including selective breeding, will need to carefully consider the inherent dynamics (resistance to perturbation, highly interconnected closed-loop system), as well as the fragility of this system.

4.6 Acknowledgements

The authors would like to acknowledge Vinet Coetzee (University of Pretoria) for assistance and consultation on statistical methods, and Marja O'Neill (University of Pretoria) for assistance with xylem sample preparation for RNA-seq analysis. Plant materials and NIRA measurements for glucose and lignin content in wood were kindly provided by Sappi Forestry Research (KwaMbonambi, South Africa). This work was supported through a strategic research grant from the South African Department of Science and Technology (DST) and by research funding from Sappi and Mondi, through the Forest Molecular Genetics Programme, the Technology and Human Resources for Industry Programme (THRIP, UID

80118) and the Bioinformatics and Functional Genomics Programme of the National Research Foundation (NRF, UID 71255 of South Africa).

4.7 References

Alabadí D, Oyama T, Yanovsky MJ, Harmon FG, Mas P, Kay SA. 2001. Reciprocal regulation between TOC1 and LHY/CCA1 within the *Arabidopsis* circadian clock. *Science* **293**(5531): 880-883.

Ambrose C, Allard JF, Cytrynbaum EN, Wasteneys GO. 2011. A CLASP-modulated cell edge barrier mechanism drives cell-wide cortical microtubule organization in *Arabidopsis*. *Nature Communications* **2**: 430.

Ambrose CJ, Wasteneys GO. 2008. CLASP modulates microtubule-cortex interaction during self-organization of acentrosomal microtubules. *Molecular Biology of the Cell* **19**(11): 4730-4737.

Amor Y, Haigler CH, Johnson S, Wainscott M, Delmer DP. 1995. A membrane-associated form of sucrose synthase and its potential role in synthesis of cellulose and callose in plants. *Proceedings of the National Academy of Sciences of the United States of America* **92**(20): 9353-9357.

Asenjo AB, Chatterjee C, Tan D, DePaoli V, Rice WJ, Diaz-Avalos R, Silvestry M, Sosa H. 2013. Structural model for tubulin recognition and deformation by Kinesin-13 microtubule depolymerases. *Cell reports* **3**: 759-768.

Atmodjo MA, Hao Z, Mohnen D 2013. Evolving views of pectin biosynthesis. 747-779.

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.

Baldazzi V, Bertin N, De Jong H, Génard M. 2012. Towards multiscale plant models: integrating cellular networks. *Trends in Plant Science*. **17**(12): 1360-1385.

Bandmann V, Homann U. 2012. Clathrin-independent endocytosis contributes to uptake of glucose into BY-2 protoplasts. *Plant Journal* **70**(4): 578-584.

Barker L, Kühn C, Weise A, Schulz A, Gebhardt C, Hirner B, Hellmann H, Schulze W, Ward JM, Frommer WB. 2000. SUT2, a putative sucrose sensor in sieve elements. *Plant Cell* **12**(7): 1153-1164.

Bashline L, Li S, Anderson CT, Lei L, Gu Y. 2013. The endocytosis of cellulose synthase in *Arabidopsis* is dependent on μ2, a clathrin mediated endocytosis adaptin. *Plant Physiology*. **163**(1): 150.

Basu D, Le J, El-Essal SE-D, Huang S, Zhang C, Mallery EL, Koliantz G, Staiger CJ, Szymanski DB. 2005. DISTORTED3/SCAR2 is a putative *Arabidopsis* WAVE complex subunit that activates the Arp2/3 complex and is required for epidermal morphogenesis. *The Plant Cell Online* **17**(2): 502-524.

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

Bowling A, Brown Jr R. 2008. The cytoplasmic domain of the cellulose-synthesizing complex in vascular plants. *Protoplasma* **233**(1-2): 115-127.

Bringmann M, Li E, Sampathkumar A, Kocabek T, Hauser M-T, Persson S. 2012. POM-POM2/cellulose synthase interacting1 is essential for the functional association of cellulose synthase and microtubules in Arabidopsis. *The Plant Cell Online* **24**(1): 163-177.

Brown DM, Goubet F, Wong VW, Goodacre R, Stephens E, Dupree P, Turner SR. 2007. Comparison of five xylan synthesis mutants reveals new insight into the mechanisms of xylan synthesis. *Plant Journal* **52**(6): 1154-1168.

Brown DM, Zeef LAH, Ellis J, Goodacre R, Turner SR. 2005. Identification of novel genes in Arabidopsis involved in secondary cell wall formation using expression profiling and reverse genetics. *Plant Cell* **17**(8): 2281-2295.

Cai G. 2011. How do microtubules affect deposition of cell wall polysaccharides in the pollen tube? *Plant Signaling & Behavior* **6**(5): 732-735.

Carlson JM, Doyle J. 2000. Highly optimized tolerance: Robustness and design in complex systems. *Physical Review Letters* **84**(11): 2529.

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

Chou YH, Pogorelko G, Zabotina OA. 2012. Xyloglucan xylosyltransferases XXT1, XXT2, and XXT5 and the glucan synthase CSLC4 form Golgi-localized multiprotein complexes. *Plant Physiology* **159**(4): 1355-1366.

Coleman HD, Ellis DD, Gilbert M, Mansfield SD. 2006. Up-regulation of sucrose synthase and UDP-glucose pyrophosphorylase impacts plant growth and metabolism. *Plant Biotechnology Journal* **4**(1): 87-101.

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.

Comrey AL, Lee HB. 1992. *A first course in factor analysis*: Routledge.

Creux NM, De Castro MH, Ranik M, Maleka MF, Myburg AA. 2013. Diversity and cis-element architecture of the promoter regions of cellulose synthase genes in *Eucalyptus*. *Tree Genetics & Genomes*: 1-16.

Creux NM, Ranik M, Berger DK, Myburg AA. 2008. Comparative analysis of orthologous cellulose synthase promoters from *Arabidopsis*, *Populus* and *Eucalyptus*: evidence of conserved regulatory elements in angiosperms. *New Phytologist* **179**(3): 722-737.

Crowell EF, Bischoff V, Desprez T, Rolland A, Stierhof YD, Schumacher K, Gonneau M, Höfte H, Vernhettes S. 2009. Pausing of golgi bodies on microtubules regulates secretion of cellulose synthase complexes in Arabidopsis. *Plant Cell* **21**(4): 1141-1154.

Crowell EF, Gonneau M, Stierhof YD, Höfte H, Vernhettes S. 2010. Regulated trafficking of cellulose synthases. *Current Opinion in Plant Biology* **13**(6): 700-705.

Csete M, Doyle J. 2004. Bow ties, metabolism and disease. *Trends in Biotechnology* **22**(9): 446-450.

Csete ME, Doyle JC. 2002. Reverse engineering of biological complexity. *Science* **295**(5560): 1664-1669.

Deeks MJ, Calcutt JR, Ingle EKS, Hawkins TJ, Chapman S, Richardson AC, Mentlak DA, Dixon MR, Cartwright F, Smertenko AP, Oparka K, Hussey PJ. 2012. A superfamily of actin-binding proteins at the actin-membrane nexus of higher plants. *Current Biology* **22**(17): 1595-1600.

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 beta-mannan synthase is a member of the cellulose synthase super gene family. *Science* **303**(5656): 363-366.

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 of the United States of America* **107**(18): 8492-8497.

Edwards AC, Ayroles JF, Stone EA, Carbone MA, Lyman RF, Mackay TFC. 2009. A transcriptional network associated with natural variation in *Drosophila* aggressive behavior. *Genome Biology* **10**(7). 76.

Eklöf JM, Brumer H. 2010. The XTH gene family: An update on enzyme structure, function, and phylogeny in xyloglucan remodeling. *Plant Physiology* **153**(2): 456-466.

Etxeberria E, Gonzalez P, Pozueta J. 2009. Evidence for two endocytic transport pathways in plant cells. *Plant Science* **177**(4): 341-348.

Etxeberria E, Pozueta-Romero J, Gonzalez P. 2012. In and out of the plant storage vacuole. *Plant Science* **190**: 52-61.

Fishel EA, Dixit R. 2013. Role of nucleation in cortical microtubule array organization: variations on a theme. *The Plant Journal*. **9**(6): 571-578.

Gardiner J. 2013. The evolution and diversification of plant microtubule-associated proteins. *The Plant Journal*. **75**(2): 219-229.

Geisler-Lee J, Geisler M, Coutinho PM, Segerman B, Nishikubo N, Takahashi J, Aspeborg H, Djerbi S, Master E, Andersson-Gunneras S, Sundberg B, Karpinski S, Teeri TT, Kleczkowski LA, Henrissat B, Mellerowicz EJ. 2006. Poplar carbohydrate-active enzymes. Gene identification and expression analyses. *Plant Physiol* **140**(3): 946-962.

Goonesekere NC, Shipely K, O'Connor K. 2010. The challenge of annotating protein sequences: The tale of eight domains of unknown function in Pfam. *Computational Biology and Chemistry* **34**(3): 210-214.

Groover A, Robischon M. 2006. Developmental mechanisms regulating secondary growth in woody plants. *Current Opinion in Plant Biology* **9**(1): 55-58.

Groover AT. 2005. What genes make a tree a tree? *Trends in Plant Science* **10**(5): 210-214.

Gutierrez R, Lindeboom JJ, Paredez AR, Emons AMC, Ehrhardt DW. 2009. *Arabidopsis* cortical microtubules position cellulose synthase delivery to the plasma membrane and interact with cellulose synthase trafficking compartments. *Nature Cell Biology* **11**(7): 797-806.

Haigler CH, Ivanova-Datcheva M, Hogan PS, Salnikov VV, Hwang S, Martin K, Delmer DP. 2001. Carbon partitioning to cellulose synthesis. *Plant Molecular Biology* **47**(1-2): 29-51.

Henkes S, Sonnewald U, Badur R, Flachmann R, Stitt M. 2001. A small decrease of plastid transketolase activity in antisense tobacco transformants has dramatic effects on photosynthesis and phenylpropanoid metabolism. *The Plant Cell Online* **13**(3): 535-551.

Hertzberg M, Aspeborg H, Schrader J, Andersson A, Erlandsson R, Blomqvist K, Bhalerao R, Uhlen 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 United States of America* **98**(25): 14732-14737.

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 and Developmental Biology - Plant* **45**(6): 619-629.

Hirano K, Kotake T, Kamihara K, Tsuna K, Aohara T, Kaneko Y, Takatsuji H, Tsumuraya Y, Kawasaki S. 2010. Rice BRITTLE CULM 3 (BC3) encodes a classical dynamin OsDRP2B essential for proper secondary cell wall synthesis. *Planta* **232**(1): 95-108.

Huang W, Richards S, Carbone MA, Zhu D, Anholt RRH, Ayroles JF, Duncan L, Jordan KW, Lawrence F, Magwire MM, Warner CB, Blankenburg K, Han Y, Javaid M, Jayaseelan J, Jhangiani SN, Muzny D, Ongeri F, Perales L, Wu YQ, Zhang Y, Zou X, Stone EA, Gibbs RA, Mackay TFC. 2012. Epistasis dominates the genetic architecture of *Drosophila* quantitative traits. *Proceedings of the National Academy of Sciences of the United States of America* **109**(39): 15553-15559.

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.

Joshi CP, Thammannagowda S, Fujino T, Gou JQ, Avci U, Haigler CH, McDonnell LM, Mansfield SD, Mengesha B, Carpita NC, Harris D, Debolt S, Peter GF. 2011. Perturbation of wood cellulose synthesis causes pleiotropic effects in transgenic aspen. *Molecular Plant* **4**(2): 331-345.

Jung HW, Tschaplinski TJ, Wang L, Glazebrook J, Greenberg JT. 2009. Priming in systemic plant immunity. *Science* **324**(5923): 89-91.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **28**(1): 27-30.

Kitano H. 2004. Biological robustness. *Nature Reviews Genetics* **5**(11): 826-837.

Konopka CA, Bednarek SY. 2008. Comparison of the dynamics and functional redundancy of the *Arabidopsis* dynamin-related isoforms DRP1A and DRP1C during plant development. *Plant Physiology* **147**(4): 1590-1602.

Kullan ARK, van Dyk MM, Hefer CA, Jones N, Kanzler A, Myburg AA. 2012a. Genetic dissection of growth, wood basic density and gene expression in interspecific backcrosses of *Eucalyptus grandis* and *E. urophylla*. *BMC Genetics* **13**(1): 60.

Kullan ARK, van Dyk MM, Jones N, Kanzler A, Bayley A, Myburg AA. 2012b. High-density genetic linkage maps with over 2,400 sequence-anchored DArT markers for genetic dissection in an F2 pseudo-backcross of *Eucalyptus grandis* × *E. urophylla*. *Tree Genetics and Genomes* **8**(1): 163-175.

Kumar M, Thammannagowda S, Bulone V, Chiang V, Han KH, Joshi CP, Mansfield SD, Mellerowicz E, Sundberg B, Teeri T, Ellis BE. 2009. An update on the nomenclature for the cellulose synthase genes in *Populus*. *Trends in Plant Science* **14**(5): 248-254.

Li S, Chen M, Yu D, Ren S, Sun S, Liu L, Ketelaar T, Emons A-MC, Liu C-M. 2013. EXO70A1-Mediated Vesicle Trafficking Is Critical for Tracheary Element Development in *Arabidopsis*. *The Plant Cell* DOI 10.1105/tpc.113.112144.

Li Y, Tschaplinski TJ, Engle NL, Hamilton CY, Rodriguez M, Liao JC, Schadt CW, Guss AM, Yang Y, Graham DE. 2012. Combined inactivation of the Clostridium cellulolyticum lactate and malate dehydrogenase genes substantially increases ethanol yield from cellulose and switchgrass fermentations. *Biotechnology for Biofuels* **5**(2): 1-13.

Liepman AH, Wilkerson CG, Keegstra K. 2005. Expression of cellulose synthase-like (Csl) genes in insect cells reveals that CslA family members encode mannan synthases. *Proceedings of the National Academy of Sciences* **102**(6): 2221-2226.

Lin D, Cao L, Zhou Z, Zhu L, Ehrhardt D, Yang Z, Fu Y. 2013. Rho GTPase signaling activates microtubule severing to promote microtubule ordering in *Arabidopsis*. *Current Biology*.

Lucas WJ, Groover A, Lichtenberger R, Furuta K, Yadav SR, Helariutta Y, He XQ, Fukuda H, Kang J, Brady SM, Patrick JW, Sperry J, Yoshida A, López-Millán AF, Grusak MA, Kachroo P. 2013. The plant vascular system: evolution, development and functions. *Journal of Integrative Plant Biology* **55**(4): 294-388.

Ma S, Gong Q, Bohnert HJ. 2007. An *Arabidopsis* gene network based on the graphical Gaussian model. *Genome Research* **17**(11): 1614-1625.

Maere S, Heymans K, Kuiper M. 2005. BiNGO: A Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics* **21**(16): 3448-3449.

Mansfield SD. 2009. Solutions for dissolution-engineering cell walls for deconstruction. *Current Opinion in Biotechnology* **20**(3): 286-294.

Mentzen WI, Wurtele ES. 2008. Regulon organization of *Arabidopsis*. *BMC Plant Biology* **8**(1): 99.

Milne RJ, Byrt CS, Patrick JW, Grof CP. 2013. Are sucrose transporter expression profiles linked with patterns of biomass partitioning in *Sorghum* phenotypes? *Frontiers in Plant Science* **4**(223): 12.

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

Mucha E, Hoefle C, Hückelhoven R, Berken A. 2010. RIP3 and AtKinesin-13A - A novel interaction linking Rho proteins of plants to microtubules. *European Journal of Cell Biology* **89**(12): 906-916.

Mutwil M, Klie S, Tohge T, Giorgi FM, Wilkins O, Campbell MM, Fernie AR, Usadel B, Nikoloski Z, Persson S. 2011. PlaNet: Combined sequence and expression comparisons across plant networks derived from seven species. *Plant Cell* **23**(3): 895-910.

Mutwil M, Ruprecht C, Giorgi FM, Bringmann M, Usadel B, Persson S. 2009. Transcriptional wiring of cell wall-related genes in *Arabidopsis*. *Molecular Plant* **2**(5): 1015-1024.

Mutwil M, Usadel B, Schütte M, Loraine A, Ebenhöh O, Persson S. 2010. Assembly of an interactive correlation network for the *Arabidopsis* genome using a novel Heuristic Clustering Algorithm. *Plant Physiology* **152**(1): 29-43.

Naramoto S, Kleine-Vehn J, Robert S, Fujimoto M, Dainobu T, Paciorek T, Ueda T, Nakano A, Van Montagu MC, Fukuda H. 2010. ADP-ribosylation factor machinery mediates endocytosis in plant cells. *Proceedings of the National Academy of Sciences* **107**(50): 21890-21895.

Obayashi T, Hayashi S, Saeki M, Ohta H, Kinoshita K. 2009. ATTED-II provides coexpressed gene networks for *Arabidopsis*. *Nucleic Acids Research* **37**(SUPPL. 1).

Oikawa A, Joshi H, Rennie E. 2010. An integrative approach to the identification of *Arabidopsis* and rice genes involved in xylan and secondary wall development. *PLoS ONE* **5**: 263 - 679.

Oikawa A, Lund CH, Sakuragi Y, Scheller HV. 2013. Golgi-localized enzyme complexes for plant cell wall biosynthesis. *Trends in Plant Science* **18**(1): 49-58.

Park JY, Canam T, Kang KY, Unda F, Mansfield SD. 2009. Sucrose phosphate synthase expression influences poplar phenology. *Tree Physiology* **29**(7): 937-946.

Pattathil S, Harper AD, Bar-Peled M. 2005. Biosynthesis of UDP-xylose: Characterization of membrane-bound AtUXS2. *Planta* **221**(4): 538-548.

Pauly M, Keegstra K. 2008. Cell-wall carbohydrates and their modification as a resource for biofuels. *The Plant Journal* **54**(4): 559-568.

Payyavula RS, Tay KH, Tsai CJ, Harding SA. 2011. The sucrose transporter family in *Populus*: the importance of a tonoplast *Pta*SUT4 to biomass and carbon partitioning. *The Plant Journal* **65**(5): 757-770.

Peremyslov VV, Klocko AL, Fowler JE, Dolja VV. 2012. *Arabidopsis* myosin XI-K localizes to the motile endomembrane vesicles associated with F-actin. *Frontiers in plant science* **3**.

Persson S, Wei H, Milne J, Page GP, Somerville CR. 2005. Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proceedings of the National Academy of Sciences of the United States of America* **102**(24): 8633-8638.

Qi X, Kaneda M, Chen J, Geitmann A, Zheng H. 2011. A specific role for *Arabidopsis* TRAPPII in post-Golgi trafficking that is crucial for cytokinesis and cell polarity. *The Plant Journal* **68**(2): 234-248.

Ranik M, Creux NM, Myburg AA. 2006. Within-tree transcriptome profiling in wood-forming tissues of a fast-growing *Eucalyptus* tree. *Tree Physiology* **26**(3): 365-375.

Ranik M, Myburg AA. 2006. Six new cellulose synthase genes from *Eucalyptus* are associated with primary and secondary cell wall biosynthesis. *Tree Physiology* **26**(5): 545-556.

Resende MFR, Muñoz P, Acosta JJ, Peter GF, Davis JM, Grattapaglia D, Resende MDV, Kirst M. 2012. Accelerating the domestication of trees using genomic selection: Accuracy of prediction models across ages and environments. *New Phytologist* **193**(3): 617-624.

Ruprecht C, Mutwil M, Saxe F, Eder M, Nikoloski Z, Persson S. 2011. Large-scale co-expression approach to dissect secondary cell wall formation across plant species. *Frontiers in plant science* **2**(23): 1-13

Salnikov VV, Grimson MJ, Delmer DP, Haigler CH. 2001. Sucrose synthase localizes to cellulose synthesis sites in tracheary elements. *Phytochemistry* **57**(6): 823-833.

Sannigrahi P, Ragauskas AJ, Tuskan GA. 2010. Poplar as a feedstock for biofuels: A review of compositional characteristics. *Biofuels, Bioproducts and Biorefining* **4**(2): 209-226.

Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, Kasarskis A, Zhang B, Wang S, Suver C, Zhu J, Millstein J, Sieberts S, Lamb J, GuhaThakurta D, Derry J, Storey JD, Avila-Campillo I, Kruger MJ, Johnson JM, Rohl CA, van Nas A, Mehrabian M, Drake TA, Lusis AJ, Smith RC, Guengerich FP, Strom SC, Schuetz E, Rushmore TH, Ulrich R. 2008. Mapping the genetic architecture of gene expression in human liver. *PLoS Biology* **6**(5).

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. *Plant Cell* **16**(9): 2278-2292.

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

Smith AM, Stitt M. 2007. Coordination of carbon supply and plant growth. *Plant, Cell and Environment* **30**(9): 1126-1149.

Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T. 2011. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**(3): 431-432.

Solomon OL, Berger DK, Myburg AA. 2010. Diurnal and circadian patterns of gene expression in the developing xylem of *Eucalyptus* trees. *South African Journal of Botany* **76**(3): 425-439.

Somerville C. 2006. Cellulose synthesis in higher plants. *Annu Rev Cell Dev Biol* **22**: 53-78.

Spicer R, Groover A. 2010. Evolution of development of vascular cambia and secondary growth. *New Phytologist* **186**(3): 577-592.

Stitt M, Zeeman SC. 2012. Starch turnover: Pathways, regulation and role in growth. *Current Opinion in Plant Biology* **15**(3): 282-292.

Taylor NG. 2011. A role for *Arabidopsis* dynamin related proteins DRP2A/B in endocytosis; DRP2 function is essential for plant growth. *Plant Molecular Biology* **76**(1-2): 117-129.

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

Thomas-Chollier M, Sand O, Turatsinze J-V, Defrance M, Vervisch E, Brohée S, van Helden J. 2008. RSAT: regulatory sequence analysis tools. *Nucleic Acids Research* **36**(suppl 2): W119-W127.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, 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**(5): 511-U174.

van Dyk MM, Kullan A, Mizrachi E, Hefer C, Jansen van Rensburg LZ, Tschaplinski T, Cushman K, Engle N, Tuskan G, Jones N, Kanzler A, Myburg A. 2011. Genetic dissection of transcript, metabolite, growth and wood property traits in an F2 pseudo-backcross pedigree of *Eucalyptus grandis* × *E. urophylla*. *BMC Proceedings* **5**(Suppl 7): O7.

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

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*. *Plant Cell* **24**(9): 3506-3529.

Verhertbruggen Y, Yin L, Oikawa A, Scheller HV. 2011. Mannan synthase activity in the CSLD family. *Plant Signaling and Behavior* **6**(10): 1620-1623.

Veyres N, Danon A, Aono M, Galliot S, Karibasappa YB, Diet A, Grandmottet F, Tamaoki M, Lesur D, Pilard S, Boitel-Conti M, Sangwan-Norreel BS, Sangwan RS. 2008. The *Arabidopsis* sweetie mutant is affected in carbohydrate metabolism and defective in the control of growth, development and senescence. *Plant Journal* **55**(4): 665-686.

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

Wang S, Basten C, Zeng Z. 2007. Windows QTL cartographer 2.5. *Department of Statistics, North Carolina State University, Raleigh, NC.*

Wang Y, Mortimer JC, Davis J, Dupree P, Keegstra K. 2013. Identification of an additional protein involved in mannan biosynthesis. *Plant Journal* **73**(1): 105-117.

Whitacre JM. 2012. Biological robustness: paradigms, mechanisms, and systems principles. *Frontiers in Genetics* **3**(67): 1-25.

Wightman R, Marshall R, Turner SR. 2009. A cellulose synthase-containing compartment moves rapidly beneath sites of secondary wall synthesis. *Plant and Cell Physiology* **50**(3): 584-594.

Wightman R, Turner SR. 2008. The roles of the cytoskeleton during cellulose deposition at the secondary cell wall. *Plant Journal* **54**(5): 794-805.

Yamaguchi M, Demura T. 2010. Transcriptional regulation of secondary wall formation controlled by NAC domain proteins. *Plant Biotechnology* **27**(3): 237-242.

Zeng Z-B. 1994. Precision mapping of quantitative trait loci. *Genetics* **136**(4): 1457-1468.

Zhang F, Gao B, Xu L, Li C, Hao D, Zhang S, Zhou M, Su F, Chen X, Zhi H, Li X. 2013. Allele-Specific Behavior of Molecular Networks: Understanding Small-Molecule Drug Response in Yeast. *PLoS ONE* **8**(1): e53581

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

Zhong R, Burk DH, Morrison Iii WH, Ye ZH. 2002. A kinesin-like protein is essential for oriented deposition of cellulose microfibrils and cell wall strength. *Plant Cell* **14**(12): 3101-3117.

Zhong R, Burk DH, Nairn CJ, Wood-Jones A, Morrison WH, Ye Z-H. 2005. Mutation of SAC1, an *Arabidopsis* SAC domain phosphoinositide phosphatase, causes alterations in cell morphogenesis, cell wall synthesis, and actin organization. *The Plant Cell Online* **17**(5): 1449-1466.

Zhong R, Lee C, Zhou J, McCarthy RL, Ye ZH. 2008. A battery of transcription factors involved in the regulation of secondary cell wall biosynthesis in *Arabidopsis*. *Plant Cell* **20**(10): 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***(3), 1452-1468.

Zhong R, Ye ZH. 2009. Transcriptional regulation of lignin biosynthesis. *Plant Signaling & Behavior* **4**(11): 1028-1034.

Zhou S, Campbell TG, Stone EA, Mackay TFC, Anholt RRH. 2012. Phenotypic plasticity of the *Drosophila* transcriptome. *PLoS Genetics* **8**(3): e1002593.

Zhu J, Zhang B, Smith EN, Drees B, Brem RB, Kruglyak L, Bumgarner RE, Schadt EE. 2008. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genetics* **40**(7): 854-861.

4.8 Figures and Tables

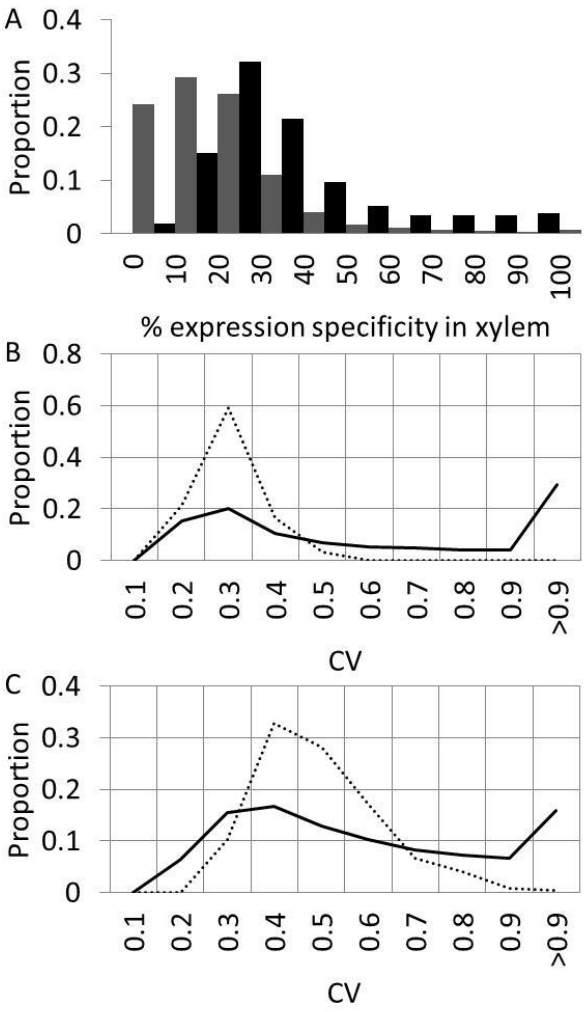


Fig. 4.1 Dynamics of expression of genes in the SCW *CesA* regulon. Dynamics of expression are shown in the context of tissue specificity (A), or variation of gene expression in the xylem samples of populations (B and C).

(A) Xylem specificity of expression of the SCW *CesA* regulon genes (black bars) and all genes in the genome (grey bars). (B) Coefficient of Variation (CV) distributions of gene expression for all genes (solid line) and genes in the SCW CesA regulon (dotted lines) in the *E. urophylla* BC. (C) CV values in the *E. grandis* BC.

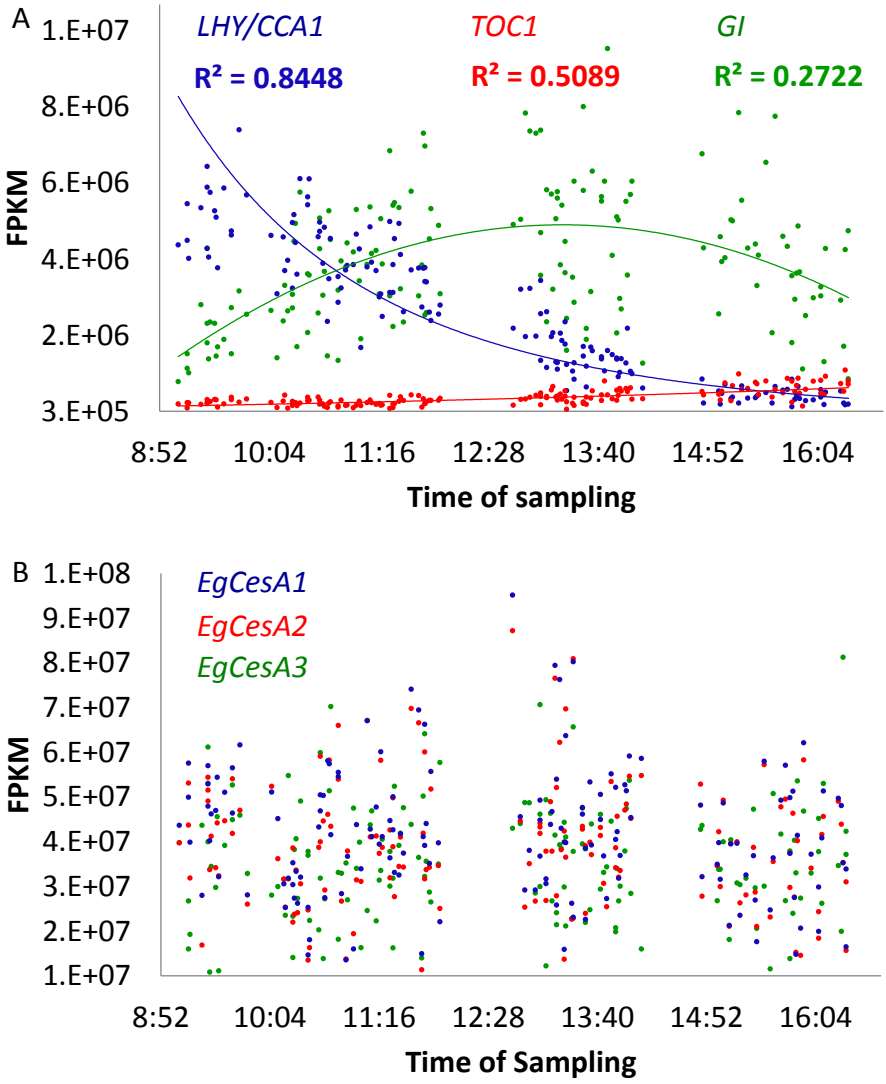


Fig. 4.2 Temporal dynamics of SCW *CesA* gene expression during sampling period.

Each dot shows the indicated gene's expression level in one of 154 sampled trees in the *E. urophylla* BC population, which were sampled over a 7½ hour period. (A) Gene expression of circadian rhythm-related genes during sample collection in the *E. urophylla* BC (see also Fig. S6). (B) Expression of the three *Eucalyptus* SCW-related *CesA* genes during sample collection. The expression of *EgCesA1*, *2* and *3* genes, as well as other genes in the regulon, do not display any variation related to time of collection.

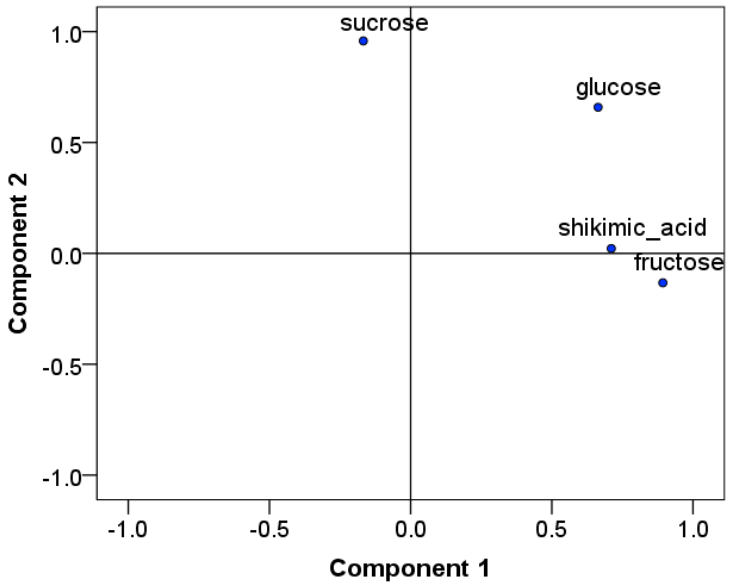


Fig 4.3 Principle component plot showing relationship between the four related metabolites in the xylem of *E. urophylla* BC population (N=154).

Refer to Table S4.3 for eigen values and component loadings.

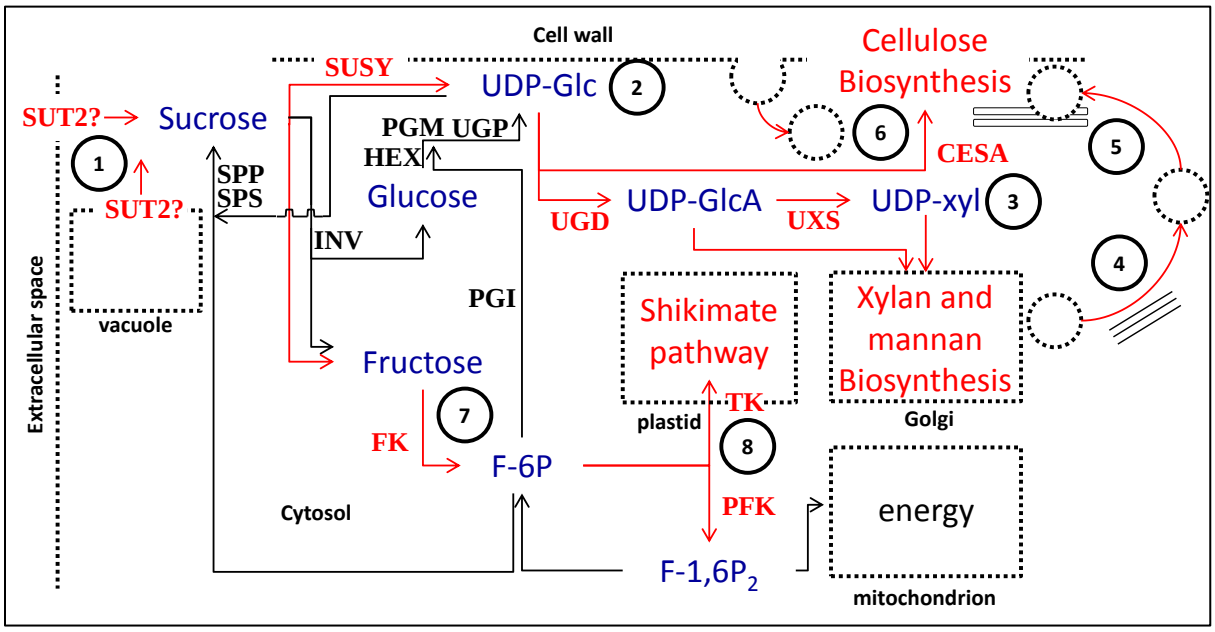


Fig. 4.4 Systems level reconstruction of transcriptionally co-regulated biological processes and pathways in the SCW *CesA* regulon.

Red lines and labels represent important reactions and biological processes that are transcriptionally co-ordinated, and blue labels indicate main metabolites derived and utilized from the source sugar (sucrose). Black arrows and enzymes indicate the other main reactions possible with these metabolites in developing xylem, but that are not represented in the regulon. Sucrose is imported into the extracellular space by SUT2 from the extracellular space or vacuole (①). Sucrose is the main source of carbon for the production of UDP-glucose via SUSY (②), which is utilized for cellulose biosynthesis or converted via UGD and UXS to UDP-glucuronic acid and UDP-xylose for xylan biosynthesis (③). Golgi-derived vesicles containing CSCs travel along actin skeleton (④) and CSCs are delivered to the membrane with the aid of cortical microtubules (⑤). Clathrin mediated endocytosis (⑥) is also co-regulated. Cytosolic fructose is converted to F-6P via FK (⑦), which is shunted to the shikimate pathway via TK (⑧) or converted to F-1,6P_2.

Table 4.1 Carbohydrate Active enZmes (CAZymes) in the SCW *CesA* Regulon.

E. grandis gene	*A. thaliana* homology	*A. thaliana* Protein name	CAZyme domain/s*
Eucgr.A01324	AT5G44030	CESA4,IRX5,NWS2	GT2
Eucgr.C00246	AT5G17420	ATCESA7,CESA7,IRX3,MUR10	GT2
Eucgr.D00476	AT4G18780	ATCESA8,CESA8,IRX1,LEW2	GT2
Eucgr.C02801	AT4G32410	AtCESA1,CESA1,RSW1	GT2
Eucgr.F04216	AT5G64740	CESA6,E112,IXR2,PRC1	GT2
Eucgr.F04212	AT5G64740	CESA6,E112,IXR2,PRC1	GT2
Eucgr.H00646	AT2G21770	CESA09,CESA9	GT2
Eucgr.A01558	AT5G03760	ATCSLA09,ATCSLA9,CSLA09,CSLA9,RAT4	GT2
Eucgr.F02219	AT4G07960	ATCSLC12,CSLC12	GT2
Eucgr.C03199	AT3G43190	ATSUS4,SUS4	GT4
Eucgr.G02730	AT5G04480	UDP-Glycosyltransferase superfamily protein	GT4
Eucgr.A00485	AT1G19300	ATGATL1,GATL1,GLZ1,PARVUS	GT8
Eucgr.F01531	AT3G50760	GATL2	GT8
Eucgr.B01494	AT1G06780	GAUT6	GT8
Eucgr.F00995	AT5G54690	GAUT12,IRX8,LGT6	GT8
Eucgr.I02091	AT3G02350	GAUT9	GT8
Eucgr.H04942	AT3G18660	GUX1,PGSIP1	GT8
Eucgr.F00232	AT4G33330	GUX2,PGSIP3	GT8
Eucgr.C00129	AT4G30060	Core-2/I-branching beta-1,6-N-acetylglucosaminyltransferase family protein	GT14
Eucgr.F03095	AT5G11730	Core-2/I-branching beta-1,6-N-acetylglucosaminyltransferase family protein	GT14
Eucgr.J02142	AT5G15050	Core-2/I-branching beta-1,6-N-acetylglucosaminyltransferase family protein	GT14
Eucgr.B03217	AT1G26810	GALT1	GT31
Eucgr.E02455	AT4G21060	GALT2	GT31
Eucgr.A01123	AT1G05170	Galactosyltransferase family protein	GT31
Eucgr.F03473	AT1G05170	Galactosyltransferase family protein	GT31
Eucgr.I01797	AT3G27960	KLCR2	GT41, GT41, GT41
Eucgr.J02210	AT3G27960	KLCR2	GT41, GT41, GT41
Eucgr.B00370	AT1G27500	KLCR3	GT41, GT41, GT41
Eucgr.A01172	AT2G37090	IRX9	GT43
Eucgr.F00463	AT1G27600	I9H,IRX9-L	GT43
Eucgr.I00880	AT5G67230	I14H,IRX14-L	GT43
Eucgr.H02219	AT5G67230	I14H,IRX14-L	GT43
Eucgr.J00384	AT2G28110	FRA8,IRX7	GT47
Eucgr.G01977	AT1G27440	ATGUT1,GUT2,IRX10	GT47
Eucgr.B00504	AT1G74680	Exostosin family protein	GT47
Eucgr.B00160	AT2G36850	ATGSL08,ATGSL8,CHOR,GSL08,GSL8	GT48
Eucgr.E02763	AT1G29200	O-fucosyltransferase family protein	GT65
Eucgr.A00648	AT1G04910	O-fucosyltransferase family protein	GT65
Eucgr.J03159	AT1G51630	MSR2	GT68
Eucgr.J02587	AT5G50420	O-fucosyltransferase family protein	GT68

Eucgr.I02168	AT5G50420	O-fucosyltransferase family protein	GT68
Eucgr.A02349	AT2G41770	Unknown function, contains DUF288	GT75
Eucgr.H04490	AT1G09010	hydrolyzing O-glycosyl compounds	GH2
Eucgr.H02409	AT5G10560	xylan 1,4-beta-xylosidase activity	GH3
Eucgr.G00035	AT5G49720	ATGH9A1,DEC,GH9A1,IRX2,KOR,KOR1,RSW2,TSD1	GH9
Eucgr.F01640	AT1G75680	AtGH9B7,GH9B7	GH9
Eucgr.H04034	AT3G16920	ATCTL2,CTL2	GH19
Eucgr.C00786	AT1G19170	Pectin lyase-like superfamily protein	GH28
Eucgr.H00038	AT2G32810	BGAL9	GH35
Eucgr.C03896	AT4G26140	BGAL12	GH35
Eucgr.J02166	AT5G14950	ATGMII,GMII	GH38
Eucgr.K02497	AT5G61250	AtGUS1,GUS1	GH79
Eucgr.C04395	AT5G34940	AtGUS3,GUS3	GH79
Eucgr.A01823	AT2G20680	Endo-beta-mannase	GH113, GH5
Eucgr.H02851	AT3G24180	Beta-glucosidase	GH116
Eucgr.I02786	AT5G49900	Beta-glucosidase	GH116
Eucgr.A02084	AT1G64760	1,3 B-glucan degradation	CBM43, GH17
Eucgr.B03672	AT5G12950	Unknown function	CBM42

*Pinard *et al*., in preparation.

4.9 Supplemental data

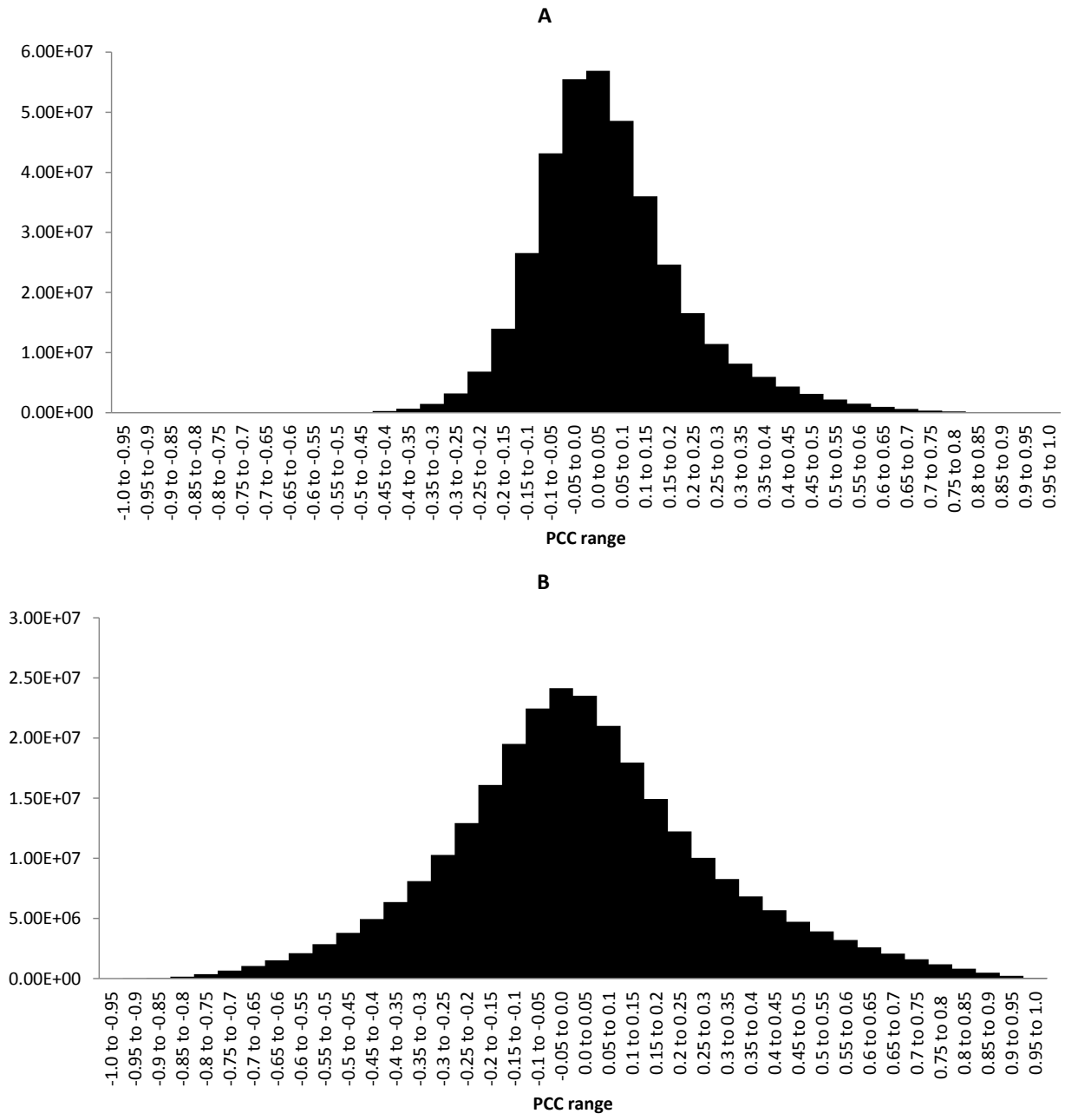


Fig. S4.1 Global correlation coefficient distributions for expressed genes in backcross populations.

A. *E. urophylla* BC population (27,337 expressed genes). B. *E. grandis* BC population (27,894 expressed genes). PCC – Pearson Correlation Coefficient.

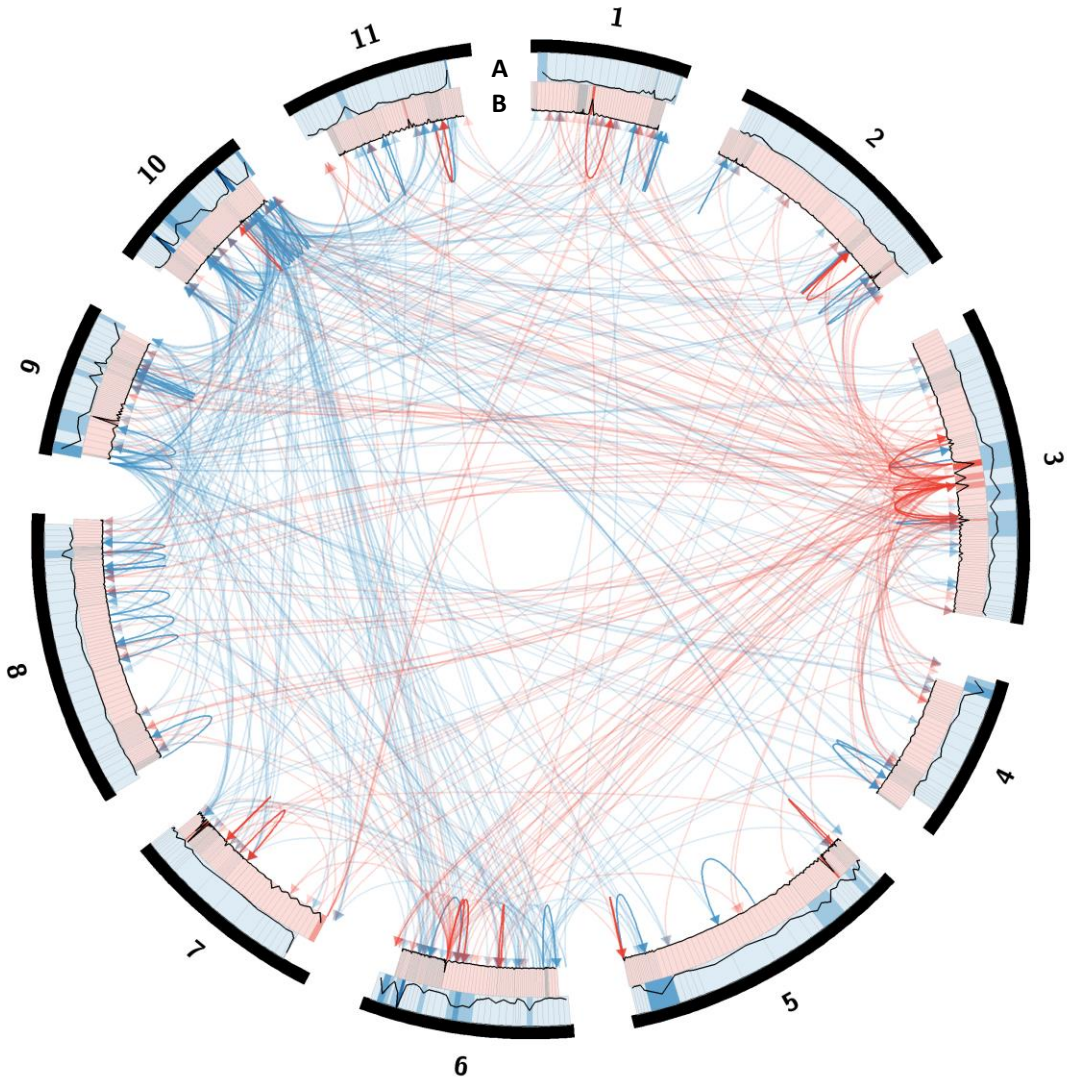


Fig. S4.2 Expression quantitative trait loci (eQTLs) for genes in the SCW *CesA* xylem regulon in the two backcross populations.

Chromosomes 1-11 are indicated. Arrows originate at gene loci location, and arrowheads indicate eQTL peak locations for eQTLs in the *E. urophylla* BC (blue lines) and *E. grandis* BC (red lines) populations. Internal tracks describe global *trans*-eQTL hotspots in the *E. urophylla* BC population (A) and *E. grandis* BC population (B), averaged across 1 cM bins. Arrowhead density in common locations is indicative of shared trans-eQTL locations for multiple genes, likely a result of segregating trans-regulators of these genes in these genomic regions.

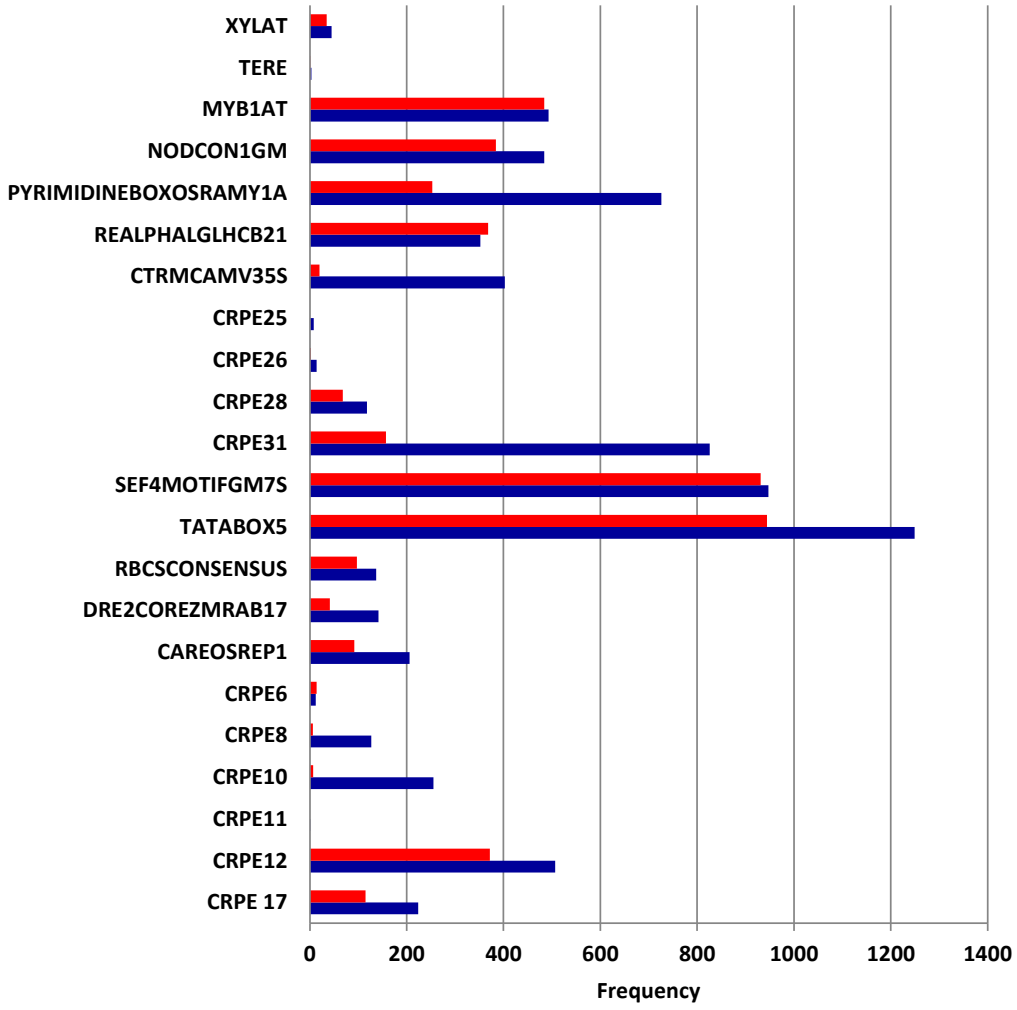


Fig. S4.3 Analysis of *CesA* associated cis-elements across genes in the SCW *CesA* regulon.

Frequency of *cis*-regulatory elements in promoter regions (1000 bp upstream of TSS) of genes in the SCW *CesA* xylem regulon (blue bars) compared to frequency of these elements in a Markov-Model generated random dataset (based on *Arabidopsis thaliana*) with identical nucleotide composition as the promoters from the CesA regulon (red bars). Twenty-two previously described motifs were interrogated, named and described in a recent promoter analysis of the *CesA* gene family in *Eucalyptus grandis* (Creux *et al.*, 2013).

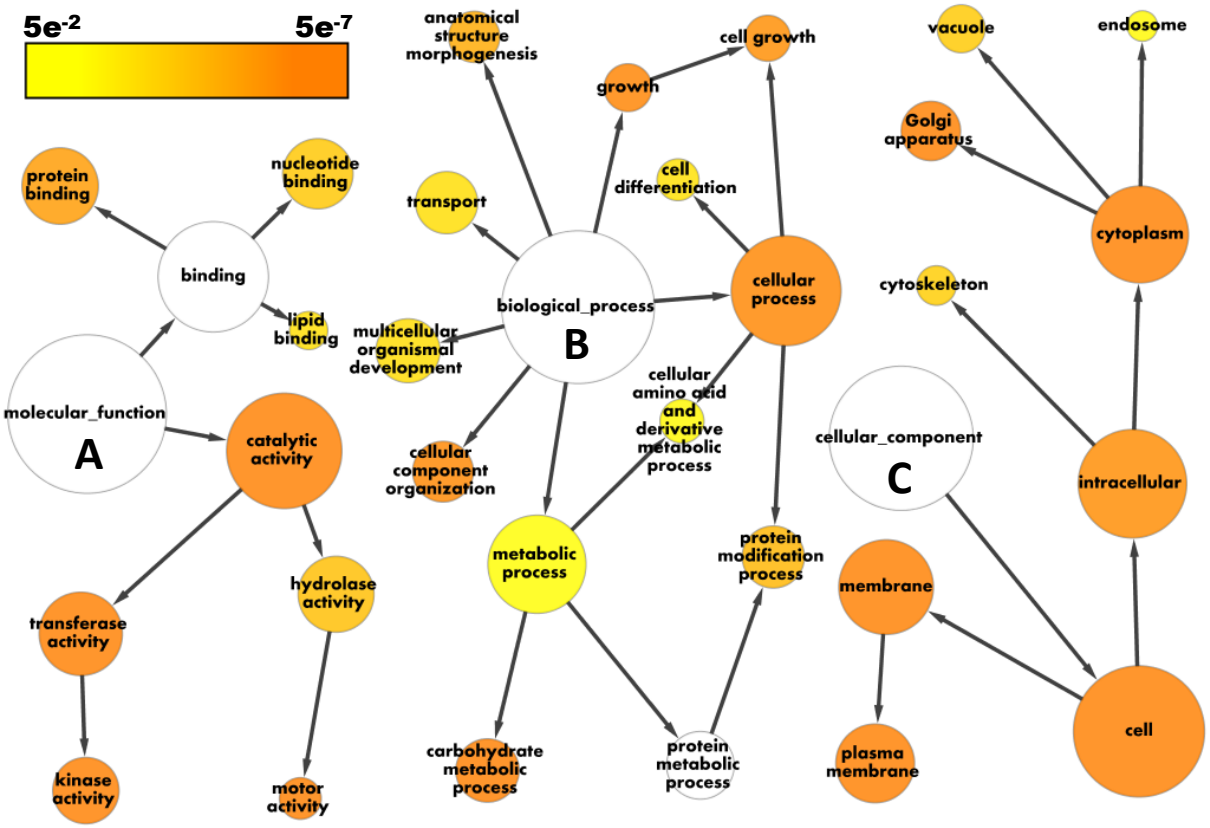


Fig. S4.4 GO terms significantly overrepresented in the SCW *CesA* regulon.

Gene Ontology (GOSLiM) terms significantly ($P< 5e^{-2}$) overrepresented in the SCW *CesA* xylem regulon in the Molecular function (A), Biological process (B) and Cellular component (C) categories. Colour scale indicates *P*-value (Hypergeometric test, FDR correction).

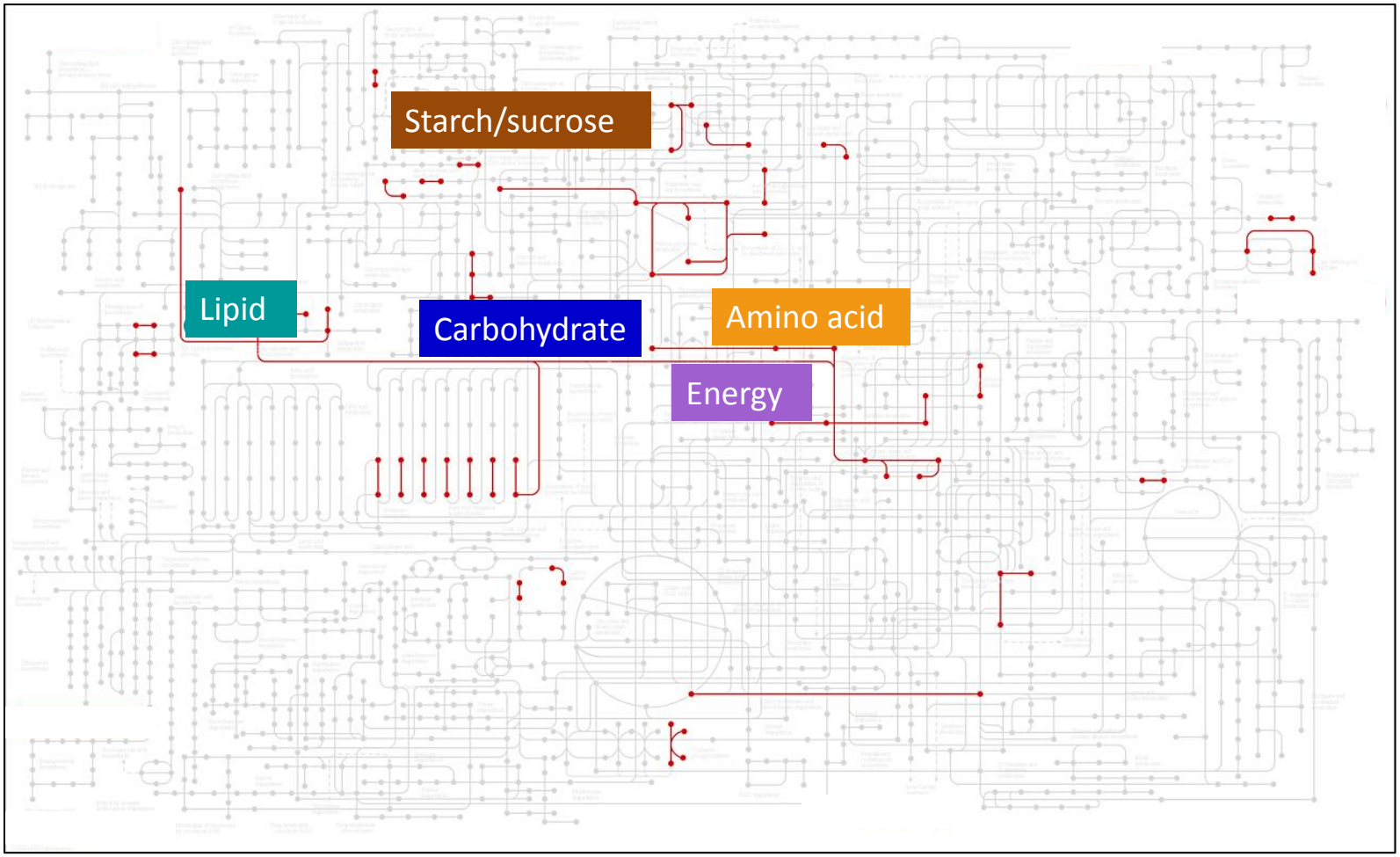


Fig. S4.5 Biochemical pathways and steps represented in the SCW *CesA* xylem regulon.

Data was mapped using KEGG Ontology (KO) terms of genes and the KEGG Mapper function (http://www.genome.jp/kegg/tool/map_pathway2.html).

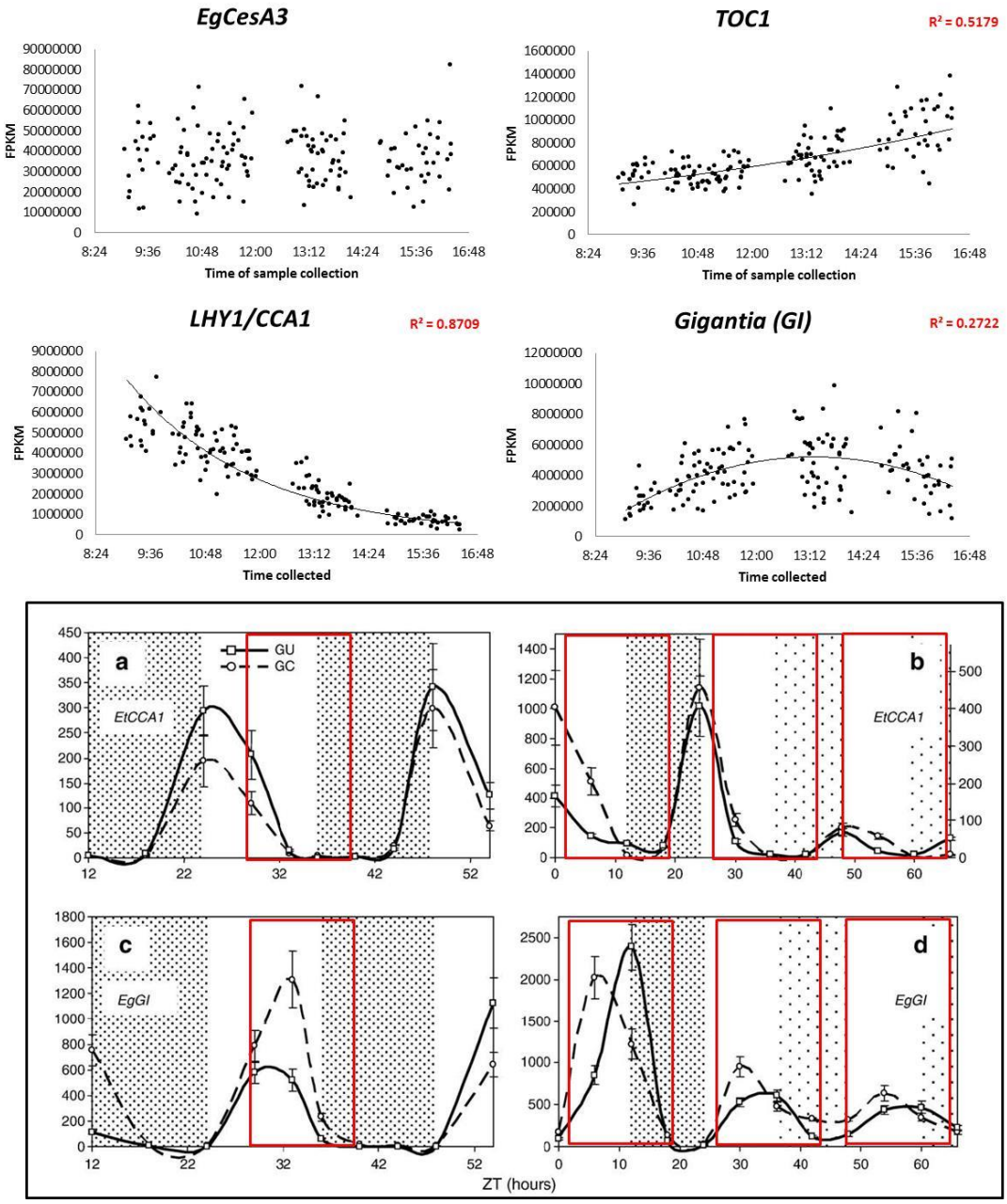


Fig. S4.6 Temporal dynamics of SCW *CesA* and circadian rhythm gene expression during sampling period.

Top: Variation of gene expression for *EgCesA3* and the clock genes *TOC1*, *LHY1/CCA1* and *Gigantia* (*GI*) over the period of sample collection in this study. **Bottom (box)**: Reproduction of Fig. 4 from (Solomon *et al.*, 2010), showing circadian variation of the *CCA1* and *GI* genes in field grown (a, c) and replicated potted ramets (b, d) of *Eucalyptus* hybrid clones. Intervals of sample collection times equivalent to this study are highlighted in red boxes.

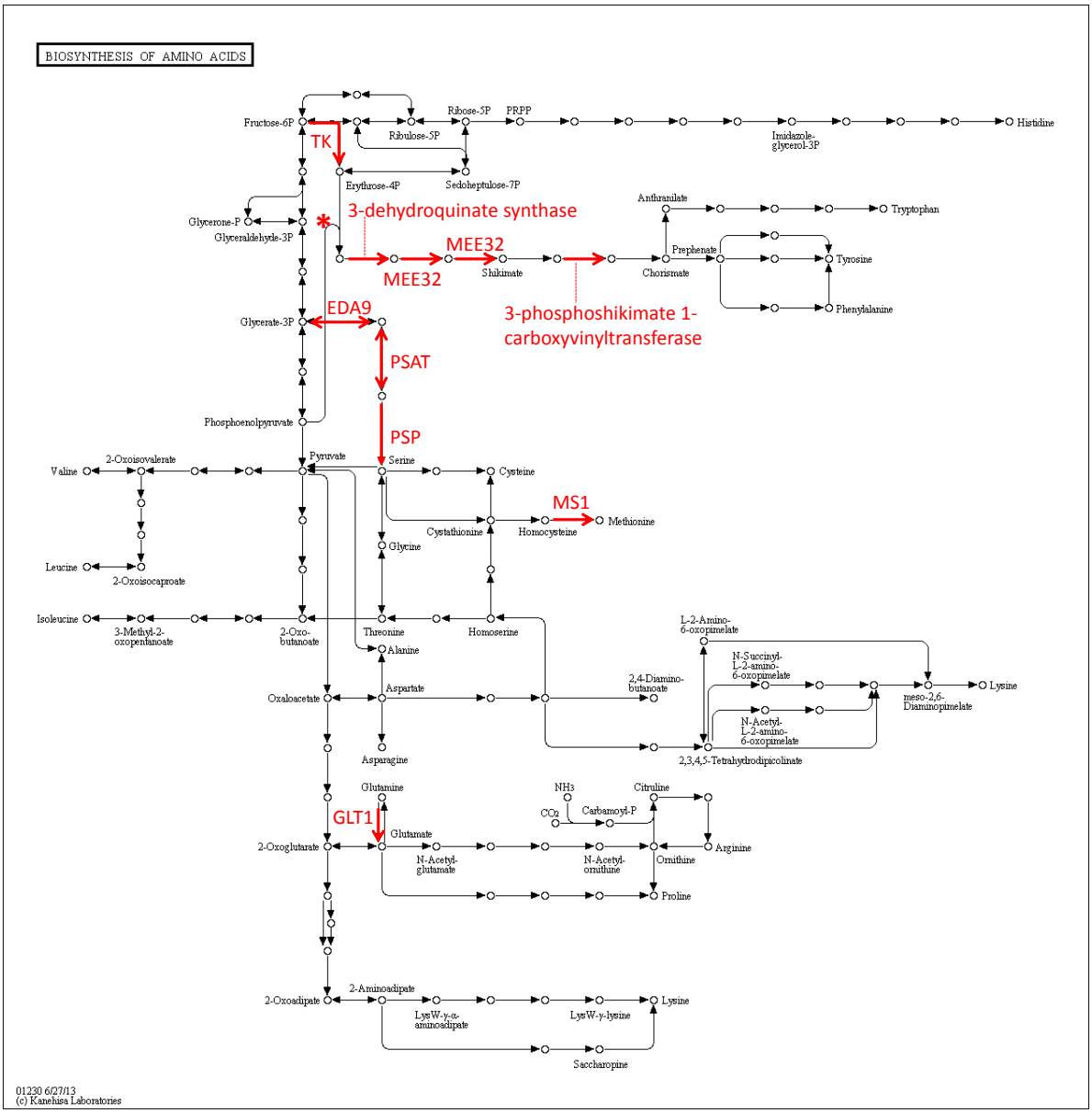


Fig. S4.7 Enzymatic reactions involved in amino acid biosynthesis represented in the SCW *CesA* regulon described in this study.

TK, Transketolase; MEE32, shikimate dehydrogenase; EDA9, D-3-phosphoglycerate dehydrogenase; PSAT, phosphoserine aminotransferase; PSP, phosphoserine phosphatase; MS1, methionine synthase 1; GLT1, glutamate synthase 1. *DAHP synthetase was not in the *CesA* regulon according to the definition of this study but was highly correlated with *EgCesA7* (PCC=0.71).

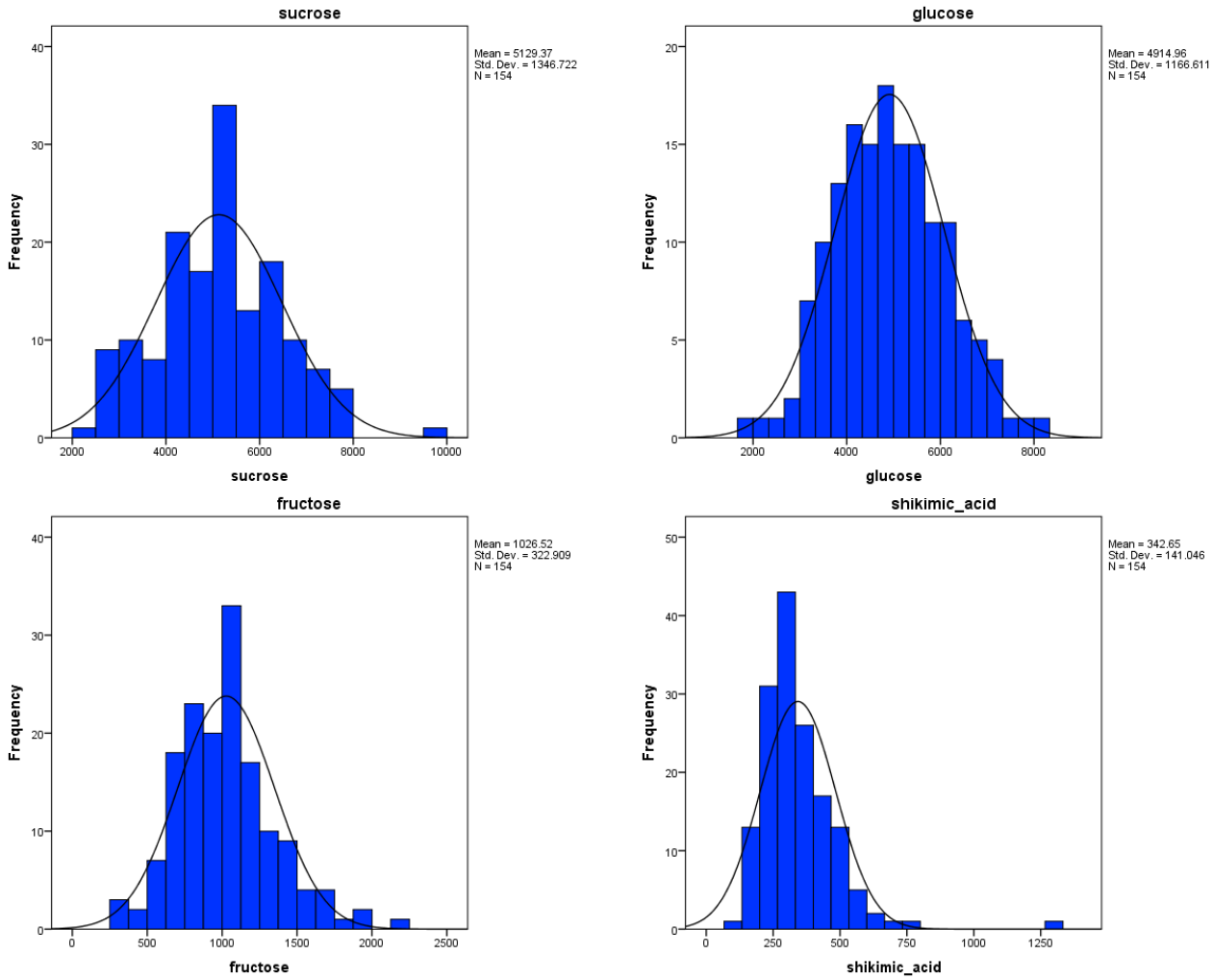
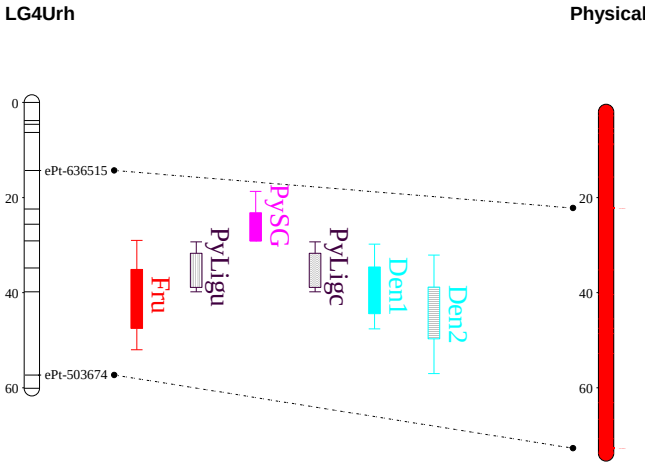


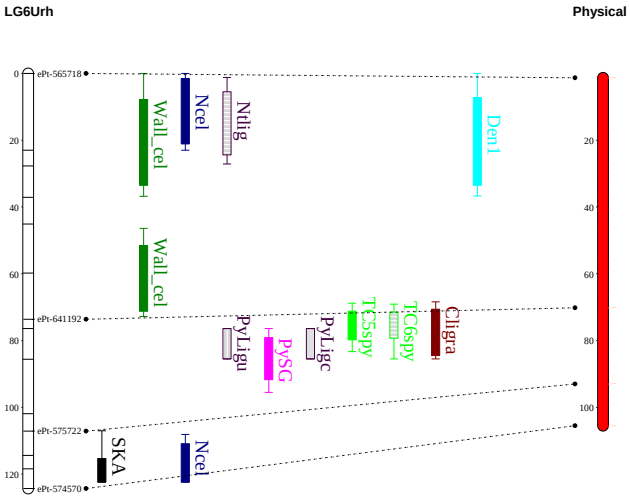
Fig. S4.8 Variation of cytosolic sucrose, glucose, fructose and shikimic acid in the developing xylem of 154 *E. urophylla* BC individuals.

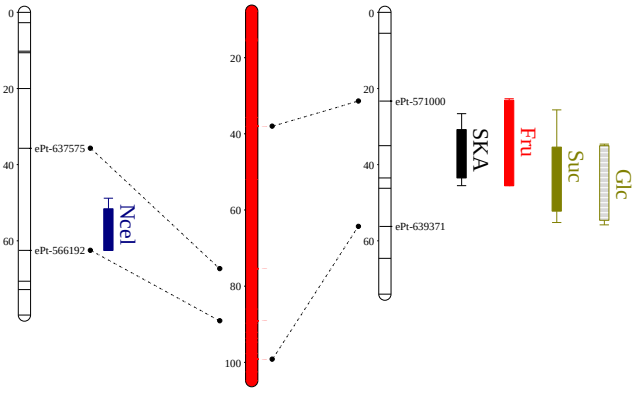
Quantities (µg/g) are shown on the x-axes.

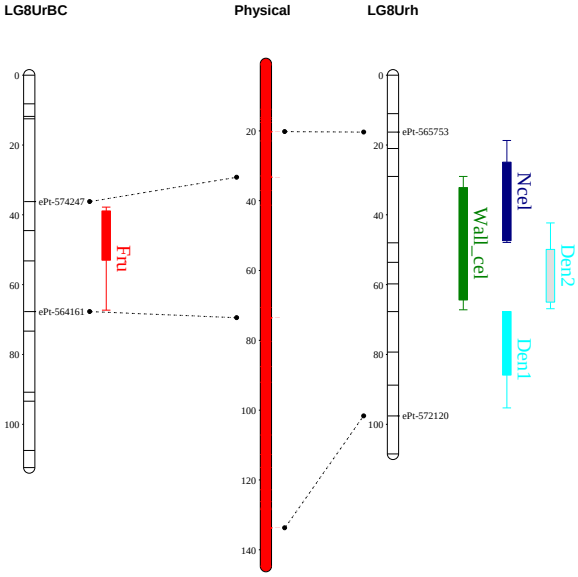
Fig. S4.9 Genetic location of putative QTLs detected in *E. urophylla* BC (UrBC) and *E. urophylla* F_1 hybrid (Urh) in selected linkage groups (LGs).

Trait QTLs are represented with the rectangular bars projected onto the parental genetic maps. Bars of a given category are filled with the same color. The vertical line for each QTL corresponds to confidence interval. QTL flanking markers are used for projection of detected QTLs and their confidence interval to the physical map (indicated in red) by dashed lines. Co-localization of trait QTL (e.g. fructose QTL with density QTLs; (LG4Urh)) as well as mQTLs (e.g. shikimic acid with cellulose (LG6Urh); shikimic acid with fructose (LG7Urh)) is observed. The trait names are abbreviated as follows: Shikimic acid (Shik), Fructose (Fru), UP-Cellulose (Wall_cel), NIRA-Cellulose (Ncel), UP_total_lignin (UPtlig), NIRA_total lignin (Ntlig), pyMBMS_lignin uncorrected (PyLigu), pyMBMS_syringyl and guaiacyl ratio (PySG), pyMBMS_lignin corrected to aspen (PyLigc), pyMBMS_total C_5 sugar in walls (TC5spy), pyMBMS_total C_6 sugar in walls (TC6spy), carbohydrate (C_5+C_6) and lignin ratio (Cligra), and density (Den).









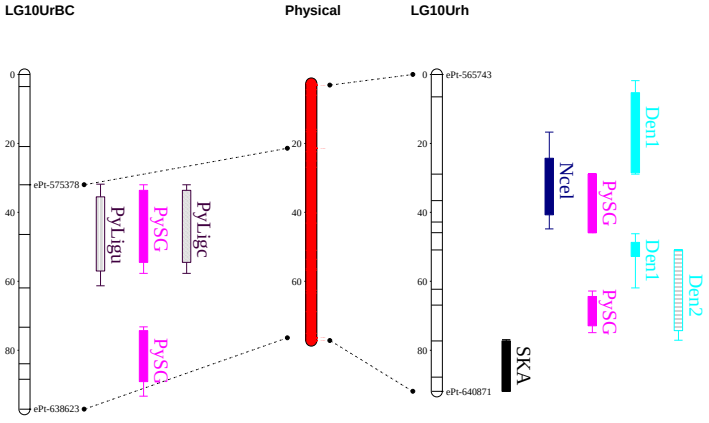


Table S4.1 Genes included in the SCW *CesA* regulon.

E. grandis gene	Correlation (*E. urophylla* BC)*	Correlation (*E. grandis* BC)*	*A. thaliana* homology	Protein
Eucgr.A00095	0.75	0.49	AT2G31390	pfkB-like
Eucgr.A00394	0.89	0.85	AT2G44160	MTHFR2
Eucgr.A00485	0.77	0.84	AT1G19300	PARVUS
Eucgr.A00519	0.74	0.86	AT3G51850	CPK13
Eucgr.A00530	0.79	0.42	AT5G01360	TBL3
Eucgr.A00648	0.82	0.93	AT1G04910	protein
Eucgr.A00750	0.83	0.82	AT1G61670	
Eucgr.A00785	0.72	0.85	AT5G58300	
Eucgr.A01005	0.45	0.87	AT1G02520	PGP11
Eucgr.A01123	0.78	0.31	AT1G05170	
Eucgr.A01172	0.87	0.77	AT2G37090	IRX9
Eucgr.A01221	0.78	0.90	AT3G53520	UXS1
Eucgr.A01282	0.80	0.87	AT5G60020	LAC17
Eucgr.A01324	0.96	0.99	AT5G44030	CESA4
Eucgr.A01471	0.82	0.90	AT5G53460	GLT1
Eucgr.A01547	0.63	0.85	AT5G04560	DME
Eucgr.A01558	0.82	0.87	AT5G03760	ATCSLA09
Eucgr.A01601	0.86	0.95	AT1G22060	
Eucgr.A01602	0.91	0.92	AT1G63300	Myosin heavy chain-related protein
Eucgr.A01823	0.78	0.63	AT2G20680	MAN2
Eucgr.A01947	0.83	0.81	AT1G06470	
Eucgr.A01974	0.79	0.78	AT2G42880	MPK20
Eucgr.A02084	0.62	0.84	AT1G64760	
Eucgr.A02119	0.76	0.94	AT3G03790	Ankyrin repeat protein
Eucgr.A02126	0.91	0.92	AT5G17920	ATMS1
Eucgr.A02137	0.71	0.85	AT1G58030	CAT2
Eucgr.A02349	0.87	0.89	AT2G41770	DUF288 protein
Eucgr.A02384	0.68	0.86	AT2G41900	
Eucgr.A02551	0.84	0.89	AT5G06390	FLA17
Eucgr.A02575	0.59	0.84	AT5G16280	
Eucgr.A02653	0.75	0.31	AT5G22400	
Eucgr.A02754	0.76	0.82	AT5G52510	SCL8
Eucgr.A02759	0.77	0.58	AT5G52430	
Eucgr.A02920	0.78	0.50	AT5G07620	
Eucgr.A02985	0.83	0.85	AT5G16730	DUF827 protein
Eucgr.B00047	0.75	0.67	AT1G72220	
Eucgr.B00160	0.51	0.85	AT2G36850	GSL8
Eucgr.B00215	0.76	0.67	AT3G48260	WNK3
Eucgr.B00248	0.66	0.84	AT4G24680	MOS1
Eucgr.B00297	0.75	0.74	AT3G48195	

Eucgr.B00370	0.78	0.64	AT1G27500	KLCR3
Eucgr.B00504	0.76	0.69	AT1G74680	
Eucgr.B00549	0.62	0.87	AT5G07740	
Eucgr.B00550	0.34	0.85	AT5G07740	
Eucgr.B00563	0.88	0.81	AT5G15430	
Eucgr.B00592	0.68	0.89	AT5G61040	
Eucgr.B00997	0.66	0.87	AT1G18260	
Eucgr.B01111	0.79	0.71	AT4G14220	RHF1A
Eucgr.B01358	0.74	0.88	AT3G05940	DUF300 protein
Eucgr.B01361	0.64	0.86	AT5G24710	
Eucgr.B01494	0.77	0.87	AT1G06780	GAUT6
Eucgr.B01725	0.47	0.88	AT5G27970	
Eucgr.B01931	0.69	0.86	AT3G05420	ACBP4
Eucgr.B02001	0.75	0.73	AT5G27410	
Eucgr.B02027	0.82	0.81	AT1G54710	ATG18H
Eucgr.B02225	0.54	0.85	AT5G58510	
Eucgr.B02229	0.88	0.92	AT4G18640	MRH1
Eucgr.B02316	0.68	0.85	AT5G60020	LAC17
Eucgr.B02458	0.79	0.88	AT2G28520	VHA-A1
Eucgr.B02486	0.76	0.50	AT5G03170	FLA11
Eucgr.B02511	0.79	0.79	AT2G28310	
Eucgr.B02576	0.72	0.87	AT1G27190	
Eucgr.B02594	0.60	0.86	AT4G32640	
Eucgr.B03179	0.67	0.86	AT1G69340	
Eucgr.B03196	0.67	0.87	AT1G13980	GN
Eucgr.B03217	0.86	0.96	AT1G26810	GALT1
Eucgr.B03381	0.66	0.85	AT1G24190	SNL3
Eucgr.B03390	0.64	0.87	AT1G59820	ALA3
Eucgr.B03410	0.68	0.87	AT3G17900	
Eucgr.B03418	0.82	0.87	AT2G02860	SUT2
Eucgr.B03582	0.79	0.77	AT1G55550	
Eucgr.B03608	0.86	0.93	AT4G08810	SUB1
Eucgr.B03672	0.70	0.86	AT5G12950	
Eucgr.B03677	0.82	0.86	AT5G12950	
Eucgr.B03736	0.87	0.85	AT5G19390	similar to REN1
Eucgr.B03801	0.80	0.84	AT5G06390	FLA17
Eucgr.B03827	0.76	0.85	AT2G27350	OTLD1
Eucgr.B03912	0.71	0.88	AT2G27900	
Eucgr.B03976	0.70	0.87	AT3G06550	RWA2
Eucgr.B04005	0.68	0.87	AT5G07120	SNX2b
Eucgr.C00094	0.85	0.64	AT5G57110	ACA8
Eucgr.C00095	0.89	0.70	AT4G29900	ACA10
Eucgr.C00121	0.79	0.83	AT4G29950	Ypt/Rab-GAP domain of gyp1p superfamily protein
Eucgr.C00129	0.78	0.71	AT4G30060	

Eucgr.C00246	1.00	1.00	AT5G17420	IRX3
Eucgr.C00247	0.70	0.84	AT5G57870	eIFiso4G1
Eucgr.C00344	0.68	0.86	AT5G25100	
Eucgr.C00422	0.76	0.90	AT5G35160	EMP70
Eucgr.C00639	0.76	0.85	AT1G17720	ATB BETA
Eucgr.C00646	0.93	0.95	AT4G33010	GLDP1
Eucgr.C00756	0.63	0.88	AT3G43300	ATMIN7
Eucgr.C00786	0.82	0.68	AT1G19170	
Eucgr.C00849	0.75	0.31	AT4G29810	MKK2
Eucgr.C01068	0.72	0.84	AT4G13350	NIG
Eucgr.C01246	0.71	0.87	AT5G57740	XBAT32
Eucgr.C01579	0.62	0.86	AT5G11490	Adaptin family protein
Eucgr.C01726	0.69	0.88	AT4G32160	
Eucgr.C01893	0.85	0.55	AT5G25050	
Eucgr.C02132	0.55	0.84	AT4G16130	ARA1
Eucgr.C02153	0.65	0.84	AT5G58100	
Eucgr.C02330	0.78	0.87	AT4G30210	ATR2
Eucgr.C02506	0.73	0.86	AT4G30710	QWRF8
Eucgr.C02574	0.54	0.85	AT2G25050	
Eucgr.C02604	0.79	0.87	AT5G10840	EMP1
Eucgr.C02726	0.78	0.93	AT5G11040	TRS120
Eucgr.C02768	0.78	0.87	AT2G25430	ENTH-domain protin
Eucgr.C02801	0.56	0.84	AT4G32410	CESA1
Eucgr.C02936	0.90	0.89	AT2G25800	DUF 810 protein
Eucgr.C03047	0.75	0.53	AT4G33010	GLDP1
Eucgr.C03199	0.76	0.72	AT3G43190	SUS4
Eucgr.C03303	0.86	0.80	AT5G20680	TBL16
Eucgr.C03404	0.61	0.86	AT5G20490	XIK
Eucgr.C03702	0.58	0.85	AT5G11700	
Eucgr.C03784	0.74	0.73	AT2G25800	
Eucgr.C03822	0.80	0.84	AT3G24550	PERK1
Eucgr.C03891	0.87	0.83	AT4G29680	
Eucgr.C03896	0.80	0.83	AT4G26140	BGAL12
Eucgr.C04018	0.68	0.85	AT2G24640	UBP19
Eucgr.C04138	0.67	0.85	AT2G26890	GRV2
Eucgr.C04382	0.61	0.85	AT1G68030	
Eucgr.C04383	0.85	0.90	AT1G15690	AVP1
Eucgr.C04395	0.80	0.82	AT5G34940	GUS3
Eucgr.C04398	0.70	0.88	AT5G20350	TIP1
Eucgr.D00027	0.86	0.89	AT5G44790	RAN1
Eucgr.D00034	0.72	0.86	AT1G79570	
Eucgr.D00052	0.62	0.85	AT2G35110	GRL
Eucgr.D00188	0.75	0.88	AT1G30470	SIT4
Eucgr.D00335	0.84	0.92	AT2G34410	RWA3
Eucgr.D00434	0.75	0.77	AT1G29340	PUB17

Eucgr.D00476	0.96	0.98	AT4G18780	IRX1
Eucgr.D00483	0.71	0.87	AT3G07100	ERMO2
Eucgr.D00489	0.82	0.87	AT1G29170	WAVE2
Eucgr.D00618	0.81	0.85	AT1G28240	DUF616 protein
Eucgr.D00655	0.71	0.92	AT2G35630	MOR1
Eucgr.D01025	0.83	0.64	AT1G30620	MUR4
Eucgr.D01125	0.62	0.86	AT3G54280	RGD3
Eucgr.D01612	0.78	0.59	AT5G12250	TUB6
Eucgr.D01685	0.75	0.86	AT4G10730	Protein kinase superfamily protein
Eucgr.D01794	0.82	0.92	AT4G23640	TRH1
Eucgr.D01842	0.64	0.86	AT1G63640	
Eucgr.D01865	0.84	0.90	AT1G63430	LRR-kinase
Eucgr.D01870	0.62	0.86	AT1G63300	
Eucgr.D01935	0.85	0.66	AT1G62990	KNAT7
Eucgr.D01962	0.76	0.79	AT1G12460	
Eucgr.D02081	0.73	0.90	AT5G47490	
Eucgr.D02202	0.75	0.35	AT4G17890	AGD8
Eucgr.D02276	0.80	0.83	AT4G11610	
Eucgr.D02329	0.83	0.91	AT3G61570	GDAP1
Eucgr.D02465	0.78	0.79	AT2G45300	
Eucgr.D02466	0.81	0.90	AT2G45290	TK
Eucgr.D02556	0.79	0.87	AT4G19180	GDA1/CD39 nucleoside phosphatase family protein
Eucgr.D02574	0.82	0.75	AT5G45290	
Eucgr.D02596	0.79	0.83	AT4G19110	
Eucgr.D02641	0.78	0.77	AT2G46620	
Eucgr.E00020	0.79	0.75	AT2G46710	
Eucgr.E00041	0.63	0.88	AT2G46560	
Eucgr.E00054	0.88	0.76	AT3G61750	
Eucgr.E00070	0.74	0.46	AT1G01430	TBL25
Eucgr.E00109	0.68	0.88	AT5G28350	RIC1
Eucgr.E00490	0.46	0.86	AT1G12430	ARK3
Eucgr.E00571	0.85	0.88	AT1G63300	Myosin heavy chain related protein
Eucgr.E00578	0.69	0.87	AT1G63440	HMA5
Eucgr.E00596	0.78	0.60	AT4G11450	
Eucgr.E00832	0.75	0.83	AT4G23740	
Eucgr.E01009	0.68	0.87	AT4G12780	
Eucgr.E01151	0.72	0.87	AT1G31730	Adaptin family protein
Eucgr.E01352	0.60	0.86	AT2G45540	
Eucgr.E01413	0.62	0.85	AT3G60860	
Eucgr.E01549	0.69	0.85	AT1G02120	VAD1
Eucgr.E02414	0.72	0.85	AT5G46210	CUL4
Eucgr.E02455	0.60	0.88	AT4G21060	GALT2

Eucgr.E02482	0.59	0.85	AT2G34680	AIR9
Eucgr.E02763	0.78	0.71	AT1G29200	
Eucgr.E03521	0.75	0.88	AT1G05500	NTMC2T2.1
Eucgr.E03535	0.78	0.58	AT5G35200	ENTH domain protein
Eucgr.E03840	0.76	0.69	AT5G47750	D6PKL2
Eucgr.E03898	0.80	0.63	AT3G07810	
Eucgr.E03976	0.60	0.88	AT4G12560	CPR30
Eucgr.E03983	0.82	0.78	AT1G12000	
Eucgr.E04303	0.60	0.85	AT5G44800	CHR4
Eucgr.E04317	0.63	0.85	AT2G07360	
Eucgr.E04321	0.92	0.84	AT1G30900	VSR6
Eucgr.F00232	0.76	0.75	AT4G33330	PGSIP3
Eucgr.F00435	0.55	0.84	AT1G58250	SAB
Eucgr.F00463	0.77	0.66	AT1G27600	IRX9-L
Eucgr.F00965	0.81	0.90	AT4G27060	TOR1
Eucgr.F00995	0.76	0.35	AT5G54690	GAUT12
Eucgr.F01094	0.87	0.85	AT4G27430	CIP7
Eucgr.F01096	0.61	0.86	AT5G54200	
Eucgr.F01158	0.78	0.69	AT3G15220	
Eucgr.F01182	0.79	0.71	AT3G15070	
Eucgr.F01374	0.77	0.75	AT3G14720	MPK19
Eucgr.F01531	0.69	0.87	AT3G50760	GATL2
Eucgr.F01564	0.87	0.92	AT5G42710	TRM30
Eucgr.F01568	0.78	0.88	AT1G19430	DUF248 protein
Eucgr.F01595	0.77	0.90	AT2G21520	Sec14p-like phosphatidylinositol transfer family protein
Eucgr.F01629	0.78	0.79	AT1G19870	iqd32
Eucgr.F01640	0.86	0.79	AT1G75680	GH9B7
Eucgr.F01786	0.74	0.89	AT1G20760	
Eucgr.F01823	0.84	0.91	AT1G76550	PFK
Eucgr.F01981	0.74	0.85	AT5G56270	WRKY2
Eucgr.F02084	0.86	0.92	AT1G19835	DUF869 protein
Eucgr.F02167	0.79	0.48	AT5G26780	SHM2
Eucgr.F02219	0.77	0.75	AT4G07960	CSLC12
Eucgr.F02237	0.61	0.87	AT1G10130	ECA3
Eucgr.F02263	0.76	0.87	AT1G49890	QWRF2
Eucgr.F02315	0.85	0.94	AT5G47820	FRA1
Eucgr.F02323	0.79	0.87	AT5G42940	RING/U-box superfamily protein
Eucgr.F02367	0.66	0.87	AT4G34310	
Eucgr.F02384	0.76	0.30	AT1G47310	
Eucgr.F02476	0.75	0.35	AT3G14170	
Eucgr.F02484	0.49	0.86	AT1G67140	SWEETIE
Eucgr.F02704	0.81	0.85	AT4G32010	HSL1
Eucgr.F02727	0.77	0.72	AT5G48940	
Eucgr.F02986	0.75	0.82	AT1G72180	

Eucgr.F03001	0.71	0.86	AT1G22620	ATSAC1
Eucgr.F03028	0.65	0.86	AT1G15240	
Eucgr.F03041	0.86	0.91	AT1G79830	GC5
Eucgr.F03072	0.88	0.97	AT1G59870	PEN3
Eucgr.F03095	0.75	0.58	AT5G11730	
Eucgr.F03170	0.75	0.77	AT3G14010	CID4
Eucgr.F03269	0.61	0.84	AT3G14920	
Eucgr.F03272	0.81	0.79	AT3G14920	
Eucgr.F03341	0.76	0.89	AT5G54440	CLUB
Eucgr.F03342	0.77	0.42	AT5G54400	
Eucgr.F03415	0.79	0.63	AT5G54590	CRLK1
Eucgr.F03419	0.88	0.92	AT5G54670	ATK3
Eucgr.F03473	0.80	0.60	AT1G05170	
Eucgr.F03616	0.65	0.87	AT1G20970	
Eucgr.F03747	0.75	0.85	AT1G21980	PIP5K1
Eucgr.F03990	0.75	0.79	AT1G16180	
Eucgr.F04026	0.87	0.86	AT1G78880	Ubiquitin-specific protease family C19-related protein
Eucgr.F04075	0.74	0.89	AT1G22870	
Eucgr.F04107	0.62	0.90	AT1G33360	
Eucgr.F04116	0.78	0.82	AT1G22610	
Eucgr.F04212	0.49	0.86	AT5G64740	CESA6
Eucgr.F04216	0.74	0.89	AT5G64740	CESA6
Eucgr.F04242	0.67	0.86	AT1G21170	SEC5B
Eucgr.G00035	0.91	0.98	AT5G49720	GH9A1
Eucgr.G00082	0.77	0.87	AT1G59610	DL3
Eucgr.G00444	0.76	0.80	AT4G34450	
Eucgr.G00451	0.61	0.86	AT1G49340	ATPI4K ALPHA
Eucgr.G00556	0.63	0.86	AT2G25170	PKL
Eucgr.G00871	0.75	0.87	AT5G16300	Vps51/Vps67 family protein
Eucgr.G01643	0.79	0.67	AT2G21520	
Eucgr.G01695	0.75	0.69	AT4G39140	
Eucgr.G01703	0.87	0.87	AT4G34610	BLH6
Eucgr.G01708	0.73	0.87	AT1G62020	
Eucgr.G01711	0.70	0.85	AT1G62020	
Eucgr.G01717	0.68	0.85	AT1G62020	
Eucgr.G01875	0.81	0.90	AT1G21630	Calcium-binding EF hand family protein
Eucgr.G01977	0.86	0.62	AT1G27440	GUT2
Eucgr.G01983	0.60	0.85	AT5G13390	NEF1
Eucgr.G02047	0.73	0.87	AT2G02040	PTR2
Eucgr.G02064	0.78	0.88	AT2G01970	EMP70
Eucgr.G02102	0.82	0.76	AT1G14830	DL1C
Eucgr.G02183	0.57	0.88	AT2G01460	
Eucgr.G02192	0.70	0.91	AT1G71010	FAB1C

Eucgr.G02416	0.65	0.85	AT2G03890	PI4K GAMMA 7
Eucgr.G02451	0.67	0.87	AT5G18520	
Eucgr.G02621	0.87	0.79	AT5G23430	KATANIN subunit B
Eucgr.G02649	0.74	0.90	AT5G04930	ALA1
Eucgr.G02730	0.68	0.86	AT5G04480	GT4
Eucgr.G02864	0.83	0.65	AT3G54850	PUB14
Eucgr.G03021	0.78	0.78	AT3G11320	
Eucgr.G03055	0.76	0.90	AT2G40070	
Eucgr.G03056	0.83	0.76	AT3G08510	PLC2
Eucgr.G03181	0.63	0.84	AT5G06120	
Eucgr.G03281	0.70	0.85	AT3G63460	
Eucgr.H00038	0.80	0.81	AT2G32810	BGAL9
Eucgr.H00308	0.75	0.28	AT3G59690	IQD13
Eucgr.H00341	0.87	0.88	AT1G08760	DUF936 protein
Eucgr.H00432	0.72	0.89	AT2G43160	EPSIN2
Eucgr.H00557	0.71	0.87	AT1G31930	XLG3
Eucgr.H00588	0.73	0.87	AT2G01970	
Eucgr.H00646	0.20	0.85	AT2G21770	CESA9
Eucgr.H00656	0.77	0.50	AT1G73390	
Eucgr.H00823	0.71	0.90	AT2G20190	CLASP
Eucgr.H00921	0.83	0.93	AT1G22060	
Eucgr.H00929	0.59	0.84	AT4G02030	
Eucgr.H00952	0.61	0.87	AT1G20110	
Eucgr.H01314	0.80	0.67	AT1G56720	
Eucgr.H01395	0.66	0.85	AT3G02750	
Eucgr.H02219	0.75	0.24	AT5G67230	IRX14-L
Eucgr.H02267	0.79	0.74	AT5G43100	
Eucgr.H02409	0.78	0.82	AT5G10560	
Eucgr.H02469	0.68	0.87	AT1G24560	
Eucgr.H02617	0.75	0.90	AT1G05820	SPPL5
Eucgr.H02678	0.76	0.52	AT4G13940	MEE58
Eucgr.H02851	0.74	0.83	AT3G24180	
Eucgr.H02900	0.78	0.85	AT2G33290	SUVH2
Eucgr.H03220	0.38	0.85	AT4G34200	EDA9
Eucgr.H03269	0.82	0.89	AT5G27030	TPR3
Eucgr.H03277	0.82	0.85	AT2G47500	Calponin homology domain containing protein
Eucgr.H03411	0.86	0.93	AT1G04200	
Eucgr.H03424	0.89	0.91	AT3G22790	NET1A
Eucgr.H03536	0.74	0.60	AT4G14950	
Eucgr.H03550	0.52	0.84	AT4G19600	CYCT1;4
Eucgr.H03604	0.83	0.92	AT3G23590	RFR1
Eucgr.H03711	0.79	0.85	AT3G06330	C3HC4 RING-type
Eucgr.H03918	0.87	0.92	AT3G16630	KINESIN-13A
Eucgr.H04034	0.79	0.64	AT3G16920	CTL2

Eucgr.H04118	0.86	0.85	AT1G56720	Protein kinase superfamily protein
Eucgr.H04219	0.75	0.87	AT3G04350	DUF946 protein
Eucgr.H04490	0.71	0.90	AT1G09010	
Eucgr.H04654	0.71	0.87	AT1G67510	
Eucgr.H04665	0.85	0.92	AT1G27850	
Eucgr.H04679	0.82	0.82	AT3G26000	
Eucgr.H04698	0.63	0.88	AT1G24460	
Eucgr.H04721	0.83	0.91	AT4G38050	Xanthine/uracil permease family protein
Eucgr.H04786	0.77	0.42	AT4G39870	
Eucgr.H04942	0.85	0.91	AT3G18660	PGSIP1
Eucgr.H05072	0.89	0.83	AT2G03200	
Eucgr.H05077	0.77	0.61	AT3G05270	
Eucgr.H05116	0.79	0.89	AT4G12770	Chaperone DnaJ-domain superfamily protein
Eucgr.I00278	0.82	0.76	AT2G32850	
Eucgr.I00330	0.72	0.86	AT5G10020	
Eucgr.I00378	0.84	0.95	AT2G22125	CSI1
Eucgr.I00540	0.77	0.86	AT5G65290	LMBR1-like membrane protein
Eucgr.I00602	0.65	0.85	AT2G22660	
Eucgr.I00650	0.81	0.94	AT4G37820	
Eucgr.I00687	0.77	0.67	AT3G49810	
Eucgr.I00880	0.80	0.88	AT5G67230	IRX14-L
Eucgr.I01020	0.73	0.90	AT2G23460	XLG1
Eucgr.I01266	0.74	0.43	AT4G34500	
Eucgr.I01293	0.83	0.88	AT4G34610	BLH6
Eucgr.I01311	0.86	0.92	AT4G39050	Kinesin motor family protein
Eucgr.I01329	0.75	0.73	AT2G21300	
Eucgr.I01543	0.62	0.88	AT4G38200	
Eucgr.I01571	0.77	0.76	AT3G26670	
Eucgr.I01665	0.65	0.86	AT3G01780	TPLATE
Eucgr.I01694	0.79	0.84	AT3G26020	
Eucgr.I01797	0.86	0.86	AT3G27960	KCLR2
Eucgr.I01911	0.78	0.43	AT5G62670	HA11
Eucgr.I01941	0.77	0.77	AT5G39785	
Eucgr.I02064	0.82	0.91	AT5G38880	41491
Eucgr.I02091	0.79	0.74	AT3G02350	GAUT9
Eucgr.I02092	0.82	0.71	AT3G02360	
Eucgr.I02168	0.76	0.64	AT5G50420	GT68 O-fucosyltransferase family protein
Eucgr.I02176	0.68	0.86	AT5G65930	ZWI
Eucgr.I02231	0.75	0.84	AT4G35630	PSAT
Eucgr.I02404	0.83	0.85	AT5G66420	
Eucgr.I02712	0.79	0.87	AT2G01070	

Eucgr.I02740	0.74	0.76	AT1G22930	
Eucgr.I02785	0.78	0.76	AT1G09610	GXM1/3
Eucgr.I02786	0.51	0.85	AT5G49900	
Eucgr.J00108	0.68	0.85	AT2G40730	
Eucgr.J00170	0.77	0.43	AT2G40320	TBL33
Eucgr.J00193	0.66	0.85	AT3G11130	Clathrin, heavy-chain linker
Eucgr.J00196	0.87	0.83	AT3G55990	ESK1
Eucgr.J00199	0.80	0.85	AT5G05570	Transducin family protein
Eucgr.J00204	0.83	0.92	AT3G55950	CCR3
Eucgr.J00263	0.78	0.78	AT3G06350	MEE32
Eucgr.J00290	0.75	0.28	AT5G04840	
Eucgr.J00322	0.72	0.87	AT5G23450	LCBK1
Eucgr.J00384	0.75	0.89	AT2G28110	FRA8
Eucgr.J00394	0.84	0.93	AT2G27950	Ring/U-Box superfamily protein
Eucgr.J00415	0.72	0.85	AT5G22780	AP2 complex alpha-subunit
Eucgr.J00613	0.63	0.87	AT5G17920	ATMS1
Eucgr.J00717	0.48	0.86	AT2G13370	CHR5
Eucgr.J00960	0.69	0.93	AT2G28520	VHA-A1
Eucgr.J01016	0.63	0.90	AT5G04560	DME
Eucgr.J01029	0.78	0.68	AT5G04510	PDK1
Eucgr.J01098	0.78	0.55	AT3G07950	
Eucgr.J01245	0.84	0.86	AT5G13820	TBP1
Eucgr.J01372	0.83	0.63	AT5G15490	UGD
Eucgr.J01393	0.91	0.86	AT5G15630	IRX6
Eucgr.J01604	0.76	0.89	AT5G16590	LRR1
Eucgr.J01821	0.68	0.86	AT4G08810	SUB1
Eucgr.J01953	0.74	0.85	AT1G07380	
Eucgr.J02120	0.75	0.83	AT3G45630	
Eucgr.J02142	0.77	0.90	AT5G15050	Core-2/I-branching beta-1,6-N-acetylglucosaminyltransferase family protein
Eucgr.J02166	0.69	0.85	AT5G14950	GMII
Eucgr.J02210	0.89	0.89	AT3G27960	KCLR2
Eucgr.J02316	0.82	0.91	AT5G43230	
Eucgr.J02393	0.89	0.89	AT4G36220	FAH1
Eucgr.J02401	0.87	0.90	AT3G51150	ATP binding microtubule motor family protein
Eucgr.J02403	0.60	0.89	AT3G51150	ATP binding microtubule motor family protein
Eucgr.J02467	0.80	0.76	AT5G66120	
Eucgr.J02485	0.79	0.79	AT2G17760	
Eucgr.J02528	0.70	0.85	AT5G66030	ATGRIP
Eucgr.J02587	0.76	0.72	AT5G50420	GT68 O-fucosyltransferase family protein
Eucgr.J02604	0.78	0.72	AT3G14205	

Eucgr.J02838	0.82	0.87	AT5G20490	XIK
Eucgr.J02949	0.82	0.85	AT3G16270	ENTH-VHS domain protein
Eucgr.J03026	0.93	0.94	AT1G52780	DUF2921 protein
Eucgr.J03159	0.80	0.58	AT1G51630	MSR2
Eucgr.J03186	0.69	0.90	AT1G16780	VHP2;2
Eucgr.K00067	0.80	0.77	AT1G64990	GTG1
Eucgr.K00157	0.77	0.87	AT4G25230	RIN2
Eucgr.K00227	0.56	0.84	AT5G62090	SLK2
Eucgr.K00712	0.81	0.82	AT5G35700	FIM2
Eucgr.K00914	0.67	0.84	AT1G12470	
Eucgr.K00963	0.88	0.86	AT2G20780	Major facilitator superfamily protein
Eucgr.K01043	0.78	0.82	AT2G20650	
Eucgr.K01080	0.75	0.81	AT1G06290	ACX3
Eucgr.K01262	0.76	0.90	AT2G38440	SCAR2
Eucgr.K01267	0.65	0.86	AT3G08850	RAPTOR1
Eucgr.K01290	0.60	0.86	AT5G04930	ALA1
Eucgr.K01508	0.77	0.75	AT5G17920	ATMS1
Eucgr.K01670	0.76	0.77	AT3G10760	
Eucgr.K01852	0.77	0.88	AT3G58050	
Eucgr.K02187	0.63	0.89	AT3G10380	SEC8
Eucgr.K02201	0.77	0.81	AT5G23670	LCB2
Eucgr.K02213	0.75	0.64	AT1G49050	
Eucgr.K02293	0.76	0.71	AT2G01970	
Eucgr.K02348	0.71	0.90	AT4G16340	SPK1
Eucgr.K02492	0.77	0.83	AT5G61340	
Eucgr.K02497	0.77	0.64	AT5G61250	GUS1
Eucgr.K02531	0.74	0.29	AT1G18640	PSP
Eucgr.K02541	0.87	0.85	AT5G15630	IRX6
Eucgr.K02733	0.75	0.89	AT5G50380	EXO70F1
Eucgr.K02840	0.76	0.89	AT4G32180	PANK2
Eucgr.K02974	0.78	0.62	AT5G01360	TBL3
Eucgr.K02996	0.88	0.91	AT2G38080	IRX12
Eucgr.K03212	0.83	0.45	AT2G37080	RIP3
Eucgr.K03323	0.51	0.86	AT1G03060	SPI
Eucgr.K03451	0.60	0.87	AT3G62900	
Eucgr.K03590	0.69	0.92	AT2G20190	CLASP
Eucgr.L01272	0.67	0.88	AT4G14920	
Eucgr.L02369	0.60	0.85	AT5G64070	PI-4KBETA1
Eucgr.L03337	0.62	0.85	AT2G13370	CHR5

*Correlations with *EgCesA3* (Eucgr.C00246) in specific BC population.

Table S4.2 Expression of laccase gene homologs in immature xylem of *Eucalyptus grandis*.

Expression data is the xylem expression average from three biological replicates of field grown mature *E. grandis* clone (data from Hefer *et al*., in preparation). The three laccase genes present in the SCW *CesA* regulon are indicated in bold.

E. grandis gene	A. thaliana homology	FPKM (xylem)	Relative *LAC* expression*
Eucgr.K02996	*IRX12/LAC4*	20,269,233	39.8%
Eucgr.A01282	*LAC17*	17,231,500	33.8%
Eucgr.B02316	*LAC17*	4,124,857	8.1%
Eucgr.G03098	*LAC5*	3,434,833	6.7%
Eucgr.K03111	*LAC17*	2,548,223	5.0%
Other Laccase genes (n=81)	*NA*	3,304,319	6.5%

*proportion of FPKM values of the total xylem FPKM of all 86 annotated *LAC* genes in *E. grandis* genome.

Table S4.3 Principle component extraction of variation of metabolite levels (cytosolic sucrose, gluclose, fructose and shikimic acid) in the developing xylem of individuals in the *E. urophylla* BC population (N=154).

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Loadings	Component*	
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %		1	2
1	1.84	45.91	45.91	1.84	45.91	45.91	Fructose	0.893	-0.133
2	1.31	32.63	78.55	1.31	32.63	78.55	Shikimic acid	0.711	
3	0.69	17.33	95.88				Glucose	0.664	0.659
4	0.17	4.12	100.00				Sucrose	-0.168	0.958

*Rotated component matrix (Varimax with Kaiser Normalization, 3 iterations).

Supplemental Note S4.1: Genome-wide and expression analysis of cellulose and xylan biosynthesis genes in the *Eucalyptus grandis* genome

The following analysis was performed to identify *Eucalyptus* homologs of the known essential genes and gene families involved in cellulose and xylan biosynthesis, as well as metabolism of sucrose for the formation of UDP-glucose (the precursor for both cellulose and xylan). Expression data (mRNA-sequencing) was utilized from the transcriptomes of a diverse set of samples including tissue from shoot tips, young leaves, mature leaves, developing xylem, phloem, roots and flowers of a rotation-age *E. grandis* clone (http://eucgenie.org/: Hefer, Van der Merwe, Mizrachi, Joubert and Myburg, *in preparation*) to provide a biological context for each gene's role, similar to the approach taken in Chapter 2 of this dissertation. This analysis was also included in the *Eucalyptus grandis* genome manuscript ("Genome sequence of *Eucalyptus grandis*: A global tree crop for fiber and energy". Myburg *et al.*, *in review*).

Polysaccharide metabolism in secondary cell walls of fibre cells in woody species is primarily geared towards channelling sucrose into UDP-glucose, which is either incorporated directly to form cellulose or converted via a two-step reaction to form UDP-xylose (the precursor for the xylan backbone). In wood, xylose is an important monomeric sugar for both xylan and xyloglucan, and it has previously been hypothesized that enzymes involved in producing and providing these monosaccharide metabolites may be physically associated and spatio-temporally regulated in coordination with the enzymatic complexes synthesizing cellulose and hemicellulose (Mansfield, 2009). An analysis performed in Chapter 2 highlighted important genes that may be involved in cellulose and xylan synthesis. Several important missing links, especially in xylan biosynthesis, have been published since the work. This, combined with the release of a fully annotated *Eucalyptus grandis* genome presented the opportunity to comprehensively catalogue and analyse known genes involved in cellulose and xylan biosynthesis. Using a combination of finding closest annotated genes in *Arabidopsis* as well as pfam domain analysis (Kersting, Mizrachi, *et al.*, *in preparation*), I identified putative homologs of previously characterized *Arabidopsis* genes that are functionally involved in the downstream metabolism of sucrose to form cellulose and xylan. Expression level and specificity have been used to prioritize putative functional homologs of different enzymatic steps.

We considered 18 enzymatic reactions involved in four major processes, (Fig. S4.10) including i. The breakdown of sucrose to produce UDP-glucose (either directly through SUSY [EC: 2.4.1.13] or indirectly through INV [EC: 3.2.1.26] → HEX [EC: 2.7.1.1] → PGM [EC: 5.4.2.2] → UGP [EC: 2.7.7.9]), ii. UDP-glucose utilization directly into cellulose (CESA), iii. UDP-glucose conversion into UDP-glucuronate (UGD [EC: 1.1.1.22]), followed by conversion to UDP-xylose (UXS [EC: 4.1.1.35]), and iv. The biosynthesis of xylan backbone and side chain-additions. All possible family members were identified. Gene expression was considered in terms of xylem specificity of expression relative to other tissues/organs in a biologically replicated *E. grandis* tree experiment (Hefer *et al.*, in preparation). I also

took into account each gene's expression relative to other family members/isoforms in xylem, as well as relative to the median (90,000 FPKM), 90th (1.35 million FPKM), 95th (2.57 million FPKM) and 99th (7.78 million FPKM) percentiles of xylem expression in the entire transcriptome. Considering each gene's relative and absolute expression levels, all members expressed in xylem over median expression (100,000 FPKM) were noted (refer to Supplemental Note S4.1 and Additional file 4.1 for complete tables with annotations and FPKM values). The functional importance of these genes in xylem is highlighted by the fact that all 18 enzymatic steps contained at least one gene member expressed in the 90th percentile of xylem gene expression, and most (15 steps) contained at least one member expressed in the 95th and 99th percentile (11 steps). All steps, with the exception of the alternative pathway to UDP-glucose production from sucrose via INV → HEX → PGM, contained at least one member showing highly xylem-specific expression (>50% of expression in xylem compared to other tissues).

There are, however, two *UGP* members specifically and highly expressed in xylem (*Eucgr.F02905*, homolog of *UGP2*, and *Eucgr.E04308*, which codes for a homolog of N-acetylglucosamine-1-phosphate uridylyltransferase (UDP-GlcNAc), an enzyme also able to catalyse the reaction glucose-6-1P → UDP-glucose). These enzymes have previously been considered in their role of providing a source of UDP-glucose from sucrose to cellulose (Amor *et al.*, 1995; Ciereszko *et al.*, 2001; Coleman *et al.*, 2006; Meng *et al.*, 2009), and the heterologous expression of UGP from *Acetobacter xylinum* in poplar resulted in an increase in growth and cellulose content (Coleman *et al.*, 2006). There are ten sucrose synthase genes expressed above median level in xylem, most are homologs of *AtSUS4* (*AT3G43190*) and are located in close proximity on chromosome 3. However, two *SUS4* homologs, both on chromosome 3 (*Eucgr.C00769* and *Eucgr.C03199*) are expressed at much higher levels than the others in xylem and together account for 18% and 70% of expression, respectively, of *SUSY* expression in xylem. This suggests that in xylem of *Eucalyptus*, the production of UDP-glucose for cellulose biosynthesis is maintained through both direct and indirect sucrose catabolism, with four clear candidates for the

enzymatic roles of both SUSY (Eucgr.C00769 and Eucgr.C03199) and UGP (Eucgr.E0430 and Eucgr.F02905).

In terms of cellulose biosynthesis the main cellulose synthase genes expressed in *Eucalyptus grandis* have been previously described (Ranik & Myburg, 2006), and the secondary cell wall *CesA* genes – orthologs of *AtCesA4* (*Eucgr.A01324*), *AtCesA7* (*Eucgr.C00246*) and *AtCesA8* (*Eucgr.D00476*) – are indeed the three dominant, and most highly and specifically expressed cellulose synthase genes in xylem. Four other *CesA* genes (*Eucgr.C02801*, *Eucgr.F03635*, *Eucgr.G03380* and *Eucgr.I00286*) are also highly, though not specifically, expressed in xylem, as they are mainly involved in primary cell wall biosynthesis.

In the conversion of UDP-glucose to UDP-xylose, there are two potential UDP-glucose dehydrogenase (UGD, [EC: 1.1.1.22]) coding genes involved in UDP-glucuronate production, but only one (*Eucgr.J01372*) is predominantly expressed in xylem and is the highest expressed gene in this pathway. This is interesting as UDP-glucuronate is a key metabolite for the production of several important primary and secondary cell wall monosaccharides including arabinose, xylose and galacturonic acid, and in *Arabidopsis* is a process with some redundancy in the form of two main enzymes, UGD2 and UGD3 (Reboul *et al.*, 2011). Evidence for only one dominantly (and extremely highly) expressed gene for this process suggests more specific downstream regulation for the production of UDP-D-xylose [EC 4.1.1.35] and UDP-D-Arabinose [EC: 5.1.3.5], or UDP-Galacturonate [EC: 5.1.3.6] from UDP-glucuronate. Several annotated epimerases in the *Eucalyptus grandis* genome show specific and/or high expression in xylem and are candidates in *Eucalyptus* for these downstream processes (www.eucgenie.org). In contrast to *UGP* gene homologs, there are four highly expressed *UXS* [EC 4.1.1.35] genes: one is a homolog of *UXS3*, predicted to code for a cytosolic form of UXS, and three are homologs of *UXS1*, *UXS2* and *UXS6*, which are predicted to code for golgi-localized, membrane bound forms of UXS (Pattathil *et al.*, 2005).

The expression of multiple isoforms has suggested that specific isoforms are responsible for providing UDP-xylose to specific xylosyltransferases for different biopolymers such as xylan and xyloglucan (Pattathil *et al.*, 2005). The high expression level and specificity in xylem of *Eucgr.G02921*, as well as the predicted membrane localization of the protein, makes it a very likely candidate for producing UDP-xylose for xylan in Eucalyptus wood (note that *Eucgr.A01221* would also be a good candidate under these criteria). *Eucgr.H01112*, which is more ubiquitously expressed across tissues and organs, could potentially be involved in xyloglucan synthesis.

Although the full complement of enzymes involved in xylan biosynthesis have not yet been described, much progress has been made in their identification in recent years and most of the genes and their functions have been described. I considered all homologs of *IRX7* (Brown *et al.*, 2007), *IRX8* (Brown *et al.*, 2007; Peña *et al.*, 2007), *IRX9/IRX9-like* and *IRX14/IRX14-like* genes (Brown *et al.*, 2007; Peña *et al.*, 2007; Brown *et al.*, 2009; Wu *et al.*, 2010; Lee *et al.*, 2012c), *IRX10/IRX10-Like* (Brown *et al.*, 2009; Wu *et al.*, 2009), as well as *PARVUS* (Lee *et al.*, 2007) and other galacturonosyltransferases, *GUX* (Mortimer *et al.*, 2010; Lee *et al.*, 2012a), *IRX15/15-Like* (Brown *et al.*, 2011; Jensen *et al.*, 2011) *GXM* (Lee *et al.*, 2012b; Urbanowicz *et al.*, 2012) and *RWA* (Lee *et al.*, 2011). In addition to potential homologs of *IRX15/15-L* and *GXM* genes, all other genes whose protein products contain the predicted DUF579 domain were considered. Additionally, although the specific function of DUF231 proteins has not been elucidated, they have been proposed in multiple studies to be related to xylan/xyloglucan acetylation (Oikawa *et al.*, 2010; Gille *et al.*, 2011), as well as cellulose biosynthesis (Bischoff *et al.*, 2010a), likely through their proposed roles in pectin methylesterification (Bischoff *et al.*, 2010b). A recent study has also proved that DUF231 protein ESK1 is involved in xylan acetylation (Yuan *et al.*, 2013). All genes coding for proteins containing DUF231 were therefore included in the analysis.

It is interesting to note that *IRX7* and *IRX8* appear to be single-copy in *Eucalyptus*, as they are in *Arabidopsis*. Single copy genes conserved across genomes are thought to be involved in processes where there is dosage balance sensitivity (De Smet *et al.*, 2013), although further analyses across other species would be required to investigate this. In contrast, there are 56 loci coding for DUF231-containing proteins, of which 33 were expressed above median expression levels in xylem and several (homologs of TBL 3, 25, 29/*ESK1*, 31 and 33) were highly and specifically expressed in xylem. There are two *GUX* homologs that are dominantly expressed, at approximately equivalent levels (*Eucgr.H04942* and *Eucgr.F00232*). The two are homologs of both *GUX1* and *GUX2*, which have been proposed to differently and distinctly substitute glucuronic acid and methyl-glucuronic acid on xylan side chains (Bromley *et al.*, 2013). Two *GXM* homologs were identified and expressed, but one (*Eucgr.I02785*) was noticeably dominant. In terms of xylan acetylation, only two RWA homologs (*RWA2* and *RWA3*, *Eucgr.B03976* and *Eucgr.D00335*, respectively) were identified, of which RWA3 was the dominantly expressed member.

Based on the relative and absolute expression levels measured by mRNA sequencing, many of the key enzymatic steps leading to sucrose catabolism, cellulose and xylan biosynthesis involve only one or two functional and active homologs in immature xylem. The genes identified in this analysis are prime candidates to be the functional homologs of the core biosynthetic machinery of cellulose and xylan, and provide a valuable reference for future studies. In the future comparative genomics and functional genetics studies could help add insight as to the roles of some of these genes. In particular, modifications such as acetylation influence cell wall biosynthesis and ultrastructure, either through polysaccharide modification or through post-translational modification of proteins that affect important biological processes such as microtubule-facilitated trafficking (Gardiner *et al.*, 2007; Cai, 2010), and it will be important to resolve especially the role of DUF231 containing proteins and their role in this biological process.

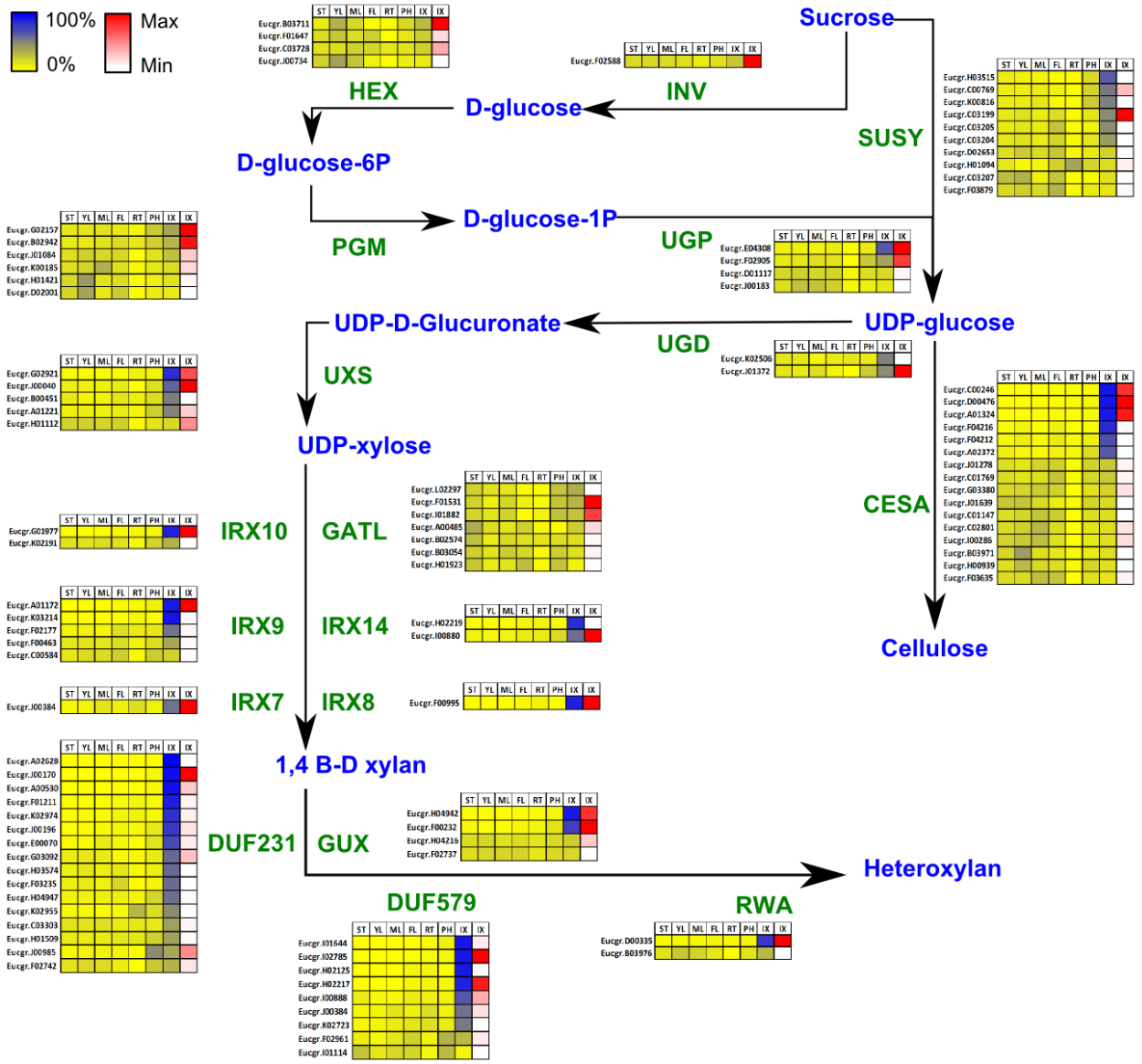


Fig. S4.10 Genes involved in cellulose and xylan biosynthesis in wood-forming tissues of *Eucalyptus*.

Relative (yellow-blue scale) and absolute (white-red scale) expression profiles of secondary cell wall related genes implicated in cellulose and xylan biosynthesis. ST, shoot tips; YL, young leaves; ML, mature leaves; FL, floral buds; RT, roots; PH, phloem, IX, immature xylem. Absolute expression level (FPKM) is only shown for immature xylem, the target secondary cell wall producing tissue. Refer to Additional file 4.1 for complete tables with annotations and FPKM values.

References

Amor Y, Haigler CH, Johnson S, Wainscott M, Delmer DP. 1995. A membrane-associated form of sucrose synthase and its potential role in synthesis of cellulose and callose in plants. *Proceedings of the National Academy of Sciences of the United States of America* **92**(20): 9353-9357.

Bischoff V, Nita S, Neumetzler L, Schindelasch D, Urbain A, Eshed R, Persson S, Delmer D, Scheible WR. 2010a. *TRICHOME BIREFRINGENCE* and its homolog AT5G01360 encode plant-specific DUF231 proteins required for cellulose biosynthesis in arabidopsis. *Plant Physiology* **153**(2): 590-602.

Bischoff V, Selbig J, Scheible WR. 2010b. Involvement of TBL/DUF231 proteins into cell wall biology. *Plant Signaling and Behavior* **5**(8): 1057-1059.

Bromley JR, Busse-Wicher M, Tryfona T, Mortimer JC, Zhang Z, Brown D, Dupree P. 2013. GUX1 and GUX2 glucuronyltransferases decorate distinct domains of glucuronoxylan with different substitution patterns. *The Plant Journal* **74**(3):423–434.

Brown D, Wightman R, Zhang Z, Gomez LD, Atanassov I, Bukowski JP, 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. *Plant Journal* **66**(3): 401-413.

Brown DM, Goubet F, Wong VW, Goodacre R, Stephens E, Dupree P, Turner SR. 2007. Comparison of five xylan synthesis mutants reveals new insight into the mechanisms of xylan synthesis. *Plant Journal* **52**(6): 1154-1168.

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*. *Plant Journal* **57**(4): 732-746.

Cai G. 2010. Assembly and disassembly of plant microtubules: Tubulin modifications and binding to MAPs. *Journal of Experimental Botany* **61**(3): 623-626.

Ciereszko I, Johansson H, Kleczkowski LA. 2001. Sucrose and light regulation of a cold-inducible UDP-glucose pyrophosphorylase gene via a hexokinase-independent and abscisic acid-insensitive pathway in *Arabidopsis*. *Biochemical Journal* **354**(1): 67-72.

Coleman HD, Ellis DD, Gilbert M, Mansfield SD. 2006. Up-regulation of sucrose synthase and UDP-glucose pyrophosphorylase impacts plant growth and metabolism. *Plant Biotechnology Journal* **4**(1): 87-101.

Creux NM, De Castro MH, Ranik M, Maleka MF, Myburg AA. 2013. Diversity and cis-element architecture of the promoter regions of cellulose synthase genes in *Eucalyptus*. *Tree Genetics & Genomes*: 1-16.

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 United States of America* **110**(8): 2898-2903.

Gardiner J, Barton D, Marc J, Overall R. 2007. Potential role of tubulin acetylation and microtubule-based protein trafficking in familial dysautonomia. *Traffic* **8**(9): 1145-1149.

Gille S, de Souza A, Xiong G, Benz M, Cheng K, Schultink A, Reca IB, Pauly M. 2011. O-acetylation of Arabidopsis hemicellulose xyloglucan requires AXY4 or AXY4L, proteins with a TBL an DUF231 domain. *Plant Cell* **23**(11): 4041-4053.

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

Lee C, Teng Q, Zhong R, Ye ZH. 2011. The four *Arabidopsis REDUCED WALL ACETYLATION* genes are expressed in secondary wall-containing cells and required for the acetylation of xylan. *Plant and Cell Physiology* **52**(8): 1289-1301.

Lee C, Teng Q, Zhong R, Ye ZH. 2012a. *Arabidopsis* GUX proteins are glucuronyltransferases responsible for the addition of glucuronic acid side chains onto xylan. *Plant and Cell Physiology* **53**(7): 1204-1216.

Lee C, Teng Q, Zhong R, Yuan Y, Haghighat M, Ye ZH. 2012b. Three *Arabidopsis* DUF579 domain-containing GXM proteins are methyltransferases catalyzing 4-o-methylation of glucuronic acid on xylan. *Plant and Cell Physiology* **53**(11): 1934-1949.

Lee C, Zhong R, Richardson EA, Himmelsbach DS, McPhail BT, Ye ZH. 2007. The *PARVUS* gene is expressed in cells undergoing secondary wall thickening and is essential for glucuronoxylan biosynthesis. *Plant and Cell Physiology* **48**(12): 1659-1672.

Lee C, Zhong R, Ye ZH. 2012c. Arabidopsis family GT43 members are xylan xylosyltransferases required for the elongation of the xylan backbone. *Plant and Cell Physiology* **53**(1): 135-143.

Mansfield SD. 2009. Solutions for dissolution-engineering cell walls for deconstruction. *Current Opinion in Biotechnology* **20**(3): 286-294.

Meng M, Geisler M, Johansson H, Harholt J, Scheller HV, Mellerowicz EJ, Kleczkowski LA. 2009. UDP-glucose pyrophosphorylase is not rate limiting, but is essential in *Arabidopsis*. *Plant and Cell Physiology* **50**(5): 998-1011.

Mortimer JC, Miles GP, Brown DM, Zhang Z, Segura MP, Weimar T, Yu X, Seffen KA, Stephens E, Turner SR, Dupree P. 2010. Absence of branches from xylan in *Arabidopsis* gux mutants reveals potential for simplification of lignocellulosic biomass. *Proceedings of the National Academy of Sciences of the United States of America* **107**(40): 17409-17414.

Oikawa A, Joshi H, Rennie E. 2010. An integrative approach to the identification of *Arabidopsis* and rice genes involved in xylan and secondary wall development. *PLoS ONE* **5**: 263 - 679.

Pattathil S, Harper AD, Bar-Peled M. 2005. Biosynthesis of UDP-xylose: Characterization of membrane-bound AtUXS2. *Planta* **221**(4): 538-548.

Peña MJ, Zhong R, Zhou GK, Richardson EA, O'Neill MA, Darvill AG, York WS, Yeb ZH. 2007. *Arabidopsis irregular xylem8* and *irregular xylem9*: Implications for the complexity of glucuronoxylan biosynthesis. *Plant Cell* **19**(2): 549-563.

Ranik M, Myburg AA. 2006. Six new cellulose synthase genes from *Eucalyptus* are associated with primary and secondary cell wall biosynthesis. *Tree Physiology* **26**(5): 545-556.

Reboul R, Geserick C, Pabst M, Frey B, Wittmann D, Lütz-Meindl U, Léonard R, Tenhaken R. 2011. Down-regulation of UDP-glucuronic acid biosynthesis leads to swollen plant cell walls and severe developmental defects associated with changes in pectic polysaccharides. *Journal of Biological Chemistry* **286**(46): 39982-39992.

Solomon OL, Berger DK, Myburg AA. 2010. Diurnal and circadian patterns of gene expression in the developing xylem of *Eucalyptus* trees. *South African Journal of Botany* **76**(3): 425-439.

Urbanowicz BR, Peña MJ, Ratnaparkhe S, Avci U, Backe J, Steet HF, Foston M, Li H, O'Neill MA, Ragauskas AJ, Darvill AG, Wyman C, Gilbert HJ, York WS. 2012. 4-O-methylation of glucuronic acid in *Arabidopsis* glucuronoxylan is catalyzed by a domain of unknown function family 579 protein. *Proceedings of the National Academy of Sciences of the United States of America* **109**(35): 14253-14258.

Wu AM, Hörnblad E, Voxeur A, Gerber L, Rihouey C, Lerouge P, Marchant A. 2010. Analysis of the *Arabidopsis* IRX9/IRX9-L and IRX14/IRX14-L pairs of glycosyltransferase genes reveals critical contributions to biosynthesis of the hemicellulose glucuronoxylan. *Plant Physiology* **153**(2): 542-554.

Wu AM, 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. *Plant Journal* **57**(4): 718-731.

Yuan Y, Teng Q, Zhong R, Ye Z-H. 2013. The *Arabidopsis* DUF231 domain-containing protein ESK1 mediates 2-O-and 3-O-acetylation of xylosyl residues in xylan. *Plant and Cell Physiology* **54**(7): 1186-1199

CHAPTER 5

CONCLUDING REMARKS

The current and future bioeconomy is going to increasingly rely on a sustainable supply of woody biomass from fast-growing tree crops suitable for processing. Today, *Eucalyptus* is the most planted hardwood genus (approximately 20M Ha worldwide). *Eucalyptus* inherently has relatively short rotation time, small genome size, tremendous genetic diversity and established breeding populations, pedigrees and clones, and as such is ideally suited for biotechnological improvement (Myburg, 2008; Hinchee *et al.*, 2009; Sederoff *et al.*, 2009; Grattapaglia *et al.*, 2012). This applies to trait improvement for current applications in fibre and chemical cellulose (mainly concerning growth, wood and biotic/abiotic stress-related traits), but in today's environment it is also important to be prepared for advanced genetic engineering and novel synthetic biology applications that can take advantage of a strong, relatively homogenous carbon sink.

A common challenge in all downstream applications, however, is the recalcitrance of woody biomass to mechanical, chemical and enzymatic breakdown, which remains a major hurdle despite decades of wood-related research (Hinchee *et al.*, 2009; Mansfield, 2009; Pu *et al.*, 2011). This is largely related to the inherent complexity of growth and wood property traits (*i.e.* controlled by hundreds of genes), which can be highly inter-related. In the case of selective breeding and improvement of populations and species, methods such as genomic selection (Grattapaglia, 2008; Grattapaglia & Resende, 2011; Resende *et al.*, 2012) currently show the most promise in being able to capture genetic variation explaining complex traits in successive generations. However, the efficacy of these methods is still subject to each trait's heritability, and the relative impact of genetic factors (pleiotropy, epistasis), environmental effects, and interactions between these (G × E). Improvement of these complex traits for any downstream application is therefore going to benefit from insight gleaned from reverse engineering of secondary cell wall formation and biopolymer deposition during xylogenesis. By understanding the molecular components and their interactions underlying these traits, the limitations and potential points of improvement can be identified in wood formation as a biological system, which should benefit both breeding and transgenic

biotechnology strategies. Given the conservation of many of the programs in secondary cell wall biosynthesis across a variety of plant lineages, it will also have broader implications for other biomass feedstocks.

Defining, modelling and understanding any biological system requires (i) detailed knowledge of its parts, (ii) understanding the dynamics of their interactions, and (iii) understanding the genetic and environmental variation and interactions, and their relative impact on these dynamics. Cataloging the "parts" involves definition and quantitation of all possible "sub-phenotypes" – a term more commonly used in disease research, but one that is appropriate for any measurable trait (molecular or emergent) that contributes to and could be predictive of the eventual phenotype. In the context of tree developmental biology and biotechnology the eventual phenotype can be defined as amenability to processing as required by the application (broadly fitting into physical, chemical and/or enzymatic processes). The "parts" include (in increasing complexity) transcript, protein and metabolite quantity, the physicochemical properties of cell wall components (simple molecules, biopolymers, and proteins) and the arrangement and homogeneity of plant cell types in wood. Between these levels are additional measurable levels of finer scale resolution, e.g. allele-specific transcript levels, alternative splice variation, a diversity of non-coding RNA types, translation efficiency, post translational modification of proteins, tissue patterning, etc. Perhaps more important are the dynamics of how these parts vary and interact. This applies across the entire system, for example transcriptional regulatory and feedback mechanisms, signalling cascades and metabolic activation/inhibitory effects. In biological systems the myriad of components and the dynamics of their interactions can be represented in the form of networks that describe parts (nodes) and dynamics of interactions (edges). Biological networks are large, scale free networks, where connectivity follows a power law distribution (Barabási & Albert, 1999; Jeong *et al.*, 2000; Strogatz, 2001). The structure reflects the fact that biological systems are built for robustness, with an obligate tradeoff of fragility (Csete & Doyle, 2002; Csete & Doyle, 2004; Whitacre, 2012).

In the broader field of understanding secondary cell wall biosynthesis, approaches to date have been largely reductionist, exploring single gene associations or phenotypic effects. This has been extremely valuable in understanding the function of individual genes (reviewed in Boerjan *et al.*, 2003; Mellerowicz & Sundberg, 2008; Vanholme *et al.*, 2008; Scheller & Ulvskov, 2010; Mizrachi *et al.*, 2012; Oikawa *et al.*, 2013; Pauly *et al.*, 2013), and tremendous advances have been made (especially in understanding xylan biosynthesis) over the past few years. Despite this, increased knowledge about individual genes has had very weak translational application in the production of stable transgenic trees with altered cell walls that offer a viable alternative to wild type trees at the plantation scale. This is mainly because key genes identified in cell wall biopolymer synthesis are generally essential for normal growth and development of the plant. Some progress in lignin engineering has been made, focusing more on modifying the homogeneity of lignin composition and structure rather than quantity/relative abundance, which is essential for development (Vanholme *et al.*, 2008). The limited success that has been demonstrated in the field of transgenic improvement of polysaccharide synthesis in wood has mainly been through the modification of sucrose flux to UDP-glucose production (Coleman *et al.*, 2006; Coleman *et al.*, 2009; Park *et al.*, 2010). Additionally, some recent strategies involving re-engineering cell-specific expression to accommodate dramatic changes in fibre cell walls alone have proven particularly promising (Petersen *et al.*, 2012; Yang *et al.*, 2012), although these studies still need to be validated in a woody plant. A major point of contention is still the fact that those transgenics that do show potential are often only observed in greenhouse studies, and relatively little is known about the persistence of these modified phenotypes in field-grown trials.

Part of the motivation for the research presented in this thesis is the hypothesis that a more holistic approach, involving a system-wide analysis, could provide novel insight that can guide future strategies in

tree biotechnology, especially with regards to cellulose and xylan. With relatively little known about the molecular biology, genes or gene expression in *Eucalyptus* at the start of this research, and virtually all knowledge relying on studies in the model plants *Arabidopsis thaliana* and *Populus trichocarpa*, the thesis was designed to address a number of fundamental questions about wood developmental biology in *Eucalyptus*. First, what is the diversity and nature of genes expressed during non-reproductive development (primary and secondary growth) in actively growing *Eucalyptus* trees, and which of these genes are most likely involved in wood formation and cellulose biosynthesis in *Eucalyptus*? (Chapter 2). Second, what are the physicochemical characteristics of cellulose-rich tension wood formed by this plantation tree, and does the transcriptome-wide reprogramming of gene expression reflect the changes in biochemical pathways leading to these characteristics? (Chapter 3). Third, building on some preliminary evidence from *Arabidopsis* gene expression meta-analyses, what is the extent of co-regulation of SCW cellulose biosynthetic genes with other biological functions? (Chapter 4). Finally, does transcript variation in the genes and pathways involved in cellulose and xylan biosynthesis in field grown trees influence (and thus predict) variation of cellulose and xylan in the wood of these trees? (Chapter 4). Previous studies highlighting the importance of transcription-level regulation of biosynthetic genes during wood formation in *Populus* (Hertzberg *et al.*, 2001; Schrader *et al.*, 2004; Geisler-Lee *et al.*, 2006) showed that transcript abundance (although a single component of complex biology) is a good indicator of gene processes and pathways important for wood formation. We therefore employed second generation RNA sequencing technologies to provide this insight. Where possible, other traits such as metabolite variation and wood property traits were also measured to add support for biological inferences made using transcript abundance.

Research from this study has had wide impact in the *Eucalyptus* research community and the forestry industry, as well as contributing to fundamental knowledge of secondary cell wall and wood biology by providing, primarily (i) resources for transcriptome analysis (ii) new biological insight into carbon

allocation for polysaccharide biosynthesis in wood, and (iii) candidate genes and pathways that may influence wood chemical composition and structure, some of which are currently being investigated in industry-supported research projects in the University of Pretoria and University of British Columbia. There have also been other important outputs, including one of the first published studies to produce a high quality *de novo* assembled gene catalog from short-read second generation sequencing technology, a viable strategy for rapid genomic characterization of other non-model species. Additionally, detailed phenotypic analysis of *Eucalyptus* wood (Chapter 3) has provided valuable comparative data for future studies. Finally, an important contribution was to the annotation of important polysaccharide metabolic pathways in the *Eucalyptus* genome (Myburg *et al.*, in preparation).

A novel approach applied here in characterizing xylogenesis has been the utilization of trees in segregating populations from F2 interspecific backcrosses, which maximize linkage disequilibrium of complex and component traits involved in wood formation. As discussed in Chapter 4, this systems genetics approach has several advantages, importantly that of observing wide phenotypic variation, and the underlying mechanisms thereof, within the constraints of a normal functioning system (field-grown trees). New insight provided by this study is the important role of energy metabolism, and the metabolism and flux of carbon between polysaccharide metabolism, glycolysis and the pentose phosphate pathway during cellulose and xylan biosynthesis in wood. Perhaps most important is the insight that when considering xylogenesis as a system, the allocation of carbon towards polysaccharide and lignin biosynthesis is transcriptionally hardwired, and is tightly coordinated in homeostasis with metabolite availability (Chapter 4). Although this will arguably require additional proof in future studies, the model of carbon flux from sucrose to UDP-glucose for simultaneous cellulose and xylan production, and the utilization of the sucrose-derived fructose moieties for additional UDP-glucose, energy production and the shikimate pathway (Chapter 4, Fig. 4) is especially attractive. This, combined with evidence for

metabolic homeostatic feedback in these pathways, provides critical insight into the predicted limitations and opportunities in modifying these genes.

Work presented in Chapter 4 has also served as proof-of-concept for a systems genetics approach in studying complex traits related to wood, and a system-wide analysis involving all measured genes, metabolites and wood traits is already underway at the time of writing. This will provide a more holistic view of processes and pathways, and crucially will reveal pleiotropic associations of genes and regulons with wood property traits measured in the population. A challenge in the future will be to find a suitable model system in which to test these models. While *Populus* is a relatively efficient woody model system, molecular interactions predicted from the model such as gene-gene or gene-metabolite relationships could be tested in heterologous plant expression systems such as *in vitro* trans-differentiation (Kubo *et al.*, 2005; Yamaguchi *et al.*, 2010). Some hypotheses could potentially be tested in *Arabidopsis thaliana*, though many effects may only be observed in systems involving a stronger xylem carbon sink.

Already, many new questions arise from this study that have not been asked or addressed. For example, what are the distinct or redundant roles of transaldolases and transketolase in directing carbon flux between carbohydrate, lignin and energy production in fibre cells during xylogenesis? If these represent key points in determining carbon allocation, what effect would targeted engineering to alter this flux have on cell wall properties? Given the centrality of pathways involved in energy production and aromatic amino acid synthesis, what is the relative contribution of mitochondria and plastids to xylogenesis, and would genetic variation in these organelles contribute to wood trait variation? At an organismal level how is sucrose produced, transported and how is its distribution regulated in xylogenic tissue, and what is the relative feedback to stored sugars (starch)? Are the major laccase genes truly involved in lignin polymerization, and are they transcriptionally wired to cellulose and xylan genes during xylogenesis?

Does this reflect the presence of protein and if so, is there a reason why laccases need to be produced during cellulose and xylan biosynthesis, and not with phenylpropanoid synthesis? (e.g. positioning of laccases in the vicinity of polysaccharides in the cell wall before the extrusion of monolignols to the forming secondary cell wall). Since the properties of tension wood could be partially explained by potential xylan modification and addition of galactose rich polymers, what effect would altering these properties in normal wood through genetic engineering approaches have?

Perhaps most importantly, how do the complex traits measured in wood influence the ultimate trait (processing)? Understanding this can drive more focused research – given our expectation that some traits (e.g. polysaccharide synthesis) are predicted to be under selection for robustness to resist perturbation, an expectation is that the genetic component should explain a relatively lower proportion of the total variation in the trait. Determining the genetic architecture of these traits can guide more effective biotechnology strategies, which may differ completely depending on the trait.

Over the past three years tremendous advances have taken place in *Eucalyptus* genomics, to the point where in 2013 a fully sequenced and annotated *Eucalyptus grandis* genome is available, along with a multitude of supporting tools for biotechnological applications, including an inter-species genome wide high density SNP chip and a gene-expression atlas consisting of multiple RNA-seq datasets from various species, tissues, organs and stress responses. Over the coming years, constructing and refining the model of xylogenesis using genetically variable populations will increase our understanding of this process, and should guide rational engineering approaches to improve wood-related traits. This will be supported by broader system-wide analyses, across different unrelated populations, with clonal replications that allow approximation of broad-sense heritability, as well as G × E interactions. Ideally, new models should include additional levels of component traits such as non-coding RNAs, proteins and more detailed

metabolomics (e.g. fully characterizing sugar nucleotide dynamics in xylem). To complement these studies, advances must be made in technologies to allow single-cell resolution, as well as the development of robust *Eucalyptus* xylem *in vitro* trans-differentiation protocols. In terms of genetic modification, results presented in this thesis suggest that a rational strategy, likely involving multiple genes to facilitate any reallocation of carbon, would need to consider and compensate for any resulting metabolic imbalances.

References

Barabási A-L, Albert R. 1999. Emergence of scaling in random networks. *Science* **286**(5439): 509-512.

Boerjan W, Ralph J, Baucher M 2003. Lignin biosynthesis. *Annual Review of Plant Biology*. 519-546.

Coleman HD, Ellis DD, Gilbert M, Mansfield SD. 2006. Up-regulation of sucrose synthase and UDP-glucose pyrophosphorylase impacts plant growth and metabolism. *Plant Biotechnology Journal* **4**(1): 87-101.

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.

Csete M, Doyle J. 2004. Bow ties, metabolism and disease. *Trends in Biotechnology* **22**(9): 446-450.

Csete ME, Doyle JC. 2002. Reverse engineering of biological complexity. *Science* **295**(5560): 1664-1669.

Geisler-Lee J, Geisler M, Coutinho PM, Segerman B, Nishikubo N, Takahashi J, Aspeborg H, Djerbi S, Master E, Andersson-Gunneras S, Sundberg B, Karpinski S, Teeri TT, Kleczkowski LA, Henrissat B, Mellerowicz EJ. 2006. Poplar carbohydrate-active enzymes. Gene identification and expression analyses. *Plant Physiol* **140**(3): 946-962.

Grattapaglia D. 2008. Perspectives on genome mapping and marker-assisted breeding of eucalypts. *Southern Forests* **70**(2): 69-75.

Grattapaglia D, Resende MDV. 2011. Genomic selection in forest tree breeding. *Tree Genetics and Genomes* **7**(2): 241-255.

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 & Genomes* **8**(3): 463-508.

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 United States of America* **98**(25): 14732-14737.

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 and Developmental Biology - Plant* **45**(6): 619-629.

Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási A-L. 2000. The large-scale organization of metabolic networks. *Nature* **407**(6804): 651-654.

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**(16): 1855-1860.

Mansfield SD. 2009. Solutions for dissolution-engineering cell walls for deconstruction. *Current Opinion in Biotechnology* **20**(3): 286-294.

Mellerowicz EJ, Sundberg B. 2008. Wood cell walls: biosynthesis, developmental dynamics and their implications for wood properties. *Current Opinion in Plant Biology* **11**(3): 293-300.

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

Myburg AA, D. Grattapaglia, G.A. Tuskan, J. Schmutz, K. Barry, J. Bristow, and The Eucalyptus Genome Network. 2008. Sequencing the *Eucalyptus* genome: genomic resources for renewable energy and fiber production. *Plant & Animal Genome XVI Conference W195, January 12-16, 2008. San Diego, CA.*

Oikawa A, Lund CH, Sakuragi Y, Scheller HV. 2013. Golgi-localized enzyme complexes for plant cell wall biosynthesis. *Trends in Plant Science* **18**(1): 49-58.

Park S, Baker JO, Himmel ME, Parilla PA, Johnson DK. 2010. Cellulose crystallinity index: Measurement techniques and their impact on interpreting cellulase performance. *Biotechnology for Biofuels* **3**: 1-10.

Pauly M, Gille S, Liu L, Mansoori N, de Souza A, Schultink A, Xiong G. 2013. Hemicellulose biosynthesis. *Planta*: 1-16.

Petersen PD, Lau J, Ebert B, Yang F, Verhertbruggen Y, Kim JS, Varanasi P, Suttangkakul A, Auer M, Loqué D. 2012. Engineering of plants with improved properties as biofuels feedstocks by vessel-specific complementation of xylan biosynthesis mutants. *Biotechnology for Biofuels* **5**(1): 84.

Pu Y, Kosa M, Kalluri UC, Tuskan GA, Ragauskas AJ. 2011. Challenges of the utilization of wood polymers: how can they be overcome? *Applied Microbiology and Biotechnology* **91**(6): 1525-1536.

Resende MFR, Muñoz P, Acosta JJ, Peter GF, Davis JM, Grattapaglia D, Resende MDV, Kirst M. 2012. Accelerating the domestication of trees using genomic selection: Accuracy of prediction models across ages and environments. *New Phytologist* **193**(3): 617-624.

Scheller HV, Ulvskov P 2010. Hemicelluloses. *Annual review of plant biology*. 263-289.

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. *Plant Cell* **16**(9): 2278-2292.

Sederoff R, Myburg A, Kirst M 2009. Genomics, domestication, and evolution of forest trees. 303-317.

Strogatz SH. 2001. Exploring complex networks. *Nature* **410**(6825): 268-276.

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

Whitacre JM. 2012. Biological robustness: paradigms, mechanisms, and systems principles. *Frontiers in genetics* **3**(67): 1-15.

Yamaguchi M, Goué N, 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 and VASCULAR-RELATED NAC-DOMAIN7 effectively induce

transdifferentiation into xylem vessel elements under control of an induction system. *Plant Physiology* **153**(3): 906-914.

Yang F, Mitra P, Zhang L, Prak L, Verhertbruggen Y, Kim JS, Sun L, Zheng K, Tang K, Auer M. 2012. Engineering secondary cell wall deposition in plants. *Plant Biotechnology Journal* **11**: 325-335.