Genetic architecture of transcription factor expression variation in developing secondary xylem of *Eucalyptus grandis* x *E. urophylla*

by

Megan Calvert

Submitted in partial fulfilment of the requirements for the degree

Magister Scientiae

In the Faculty of Natural and Agricultural Sciences

Department of Genetics

University of Pretoria

Pretoria

November 2016

Under the supervision of Prof Alexander A. Myburg

and co-supervision of Dr Eshchar Mizrachi and Dr Nanette Christie

Declaration

I, the undersigned, hereby declare that the dissertation submitted herewith for the degree of M.Sc. (Genetics) to the University of Pretoria, contains my own independent work and has not been submitted for any other degree at any other University.

Megan Calvert

February 2017

Thesis Summary

Genetic architecture of transcription factor expression variation in developing secondary xylem of *Eucalyptus grandis* x *E. urophylla*

Megan Calvert

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

Co-supervised by **Drs Eshchar Mizrachi** and **Nanette Christie**

Submitted in partial fulfilment of the requirements for the degree *Magister Scientiae*

Department of Genetics

University of Pretoria

The xylem secondary cell walls (SCW) are a strong carbon sink in woody plants and part of an irreversible developmental commitment which ends in programmed cell death. The biosynthetic processes involved in fibre SCW wall formation are guided by intrinsic developmental programmes and extrinsic signals, such as mechanical stress in woody stems, which are tightly co-ordinated. The model plants *Arabidopsis* and *Populus* have been used to study many of the transcription factors which are involved in the complex network responsible for the regulation of the formation of SCW. It is thought that variation in the expression of SCW transcription factor genes will affect cell wall deposition and wood properties in commercially important wood fibre crops such as *Eucalyptus* tree species and hybrids. To date the transcriptional control of SCW deposition in woody plants has not been fully resolved nor is it known which components of the transcriptional network regulating SCW exhibit natural variation in expression, or what the genetic architecture of such regulatory variation might be. In this study expression quantitative trait locus (eQTL) mapping and expression correlation analyses was performed for SCW-related transcription factor (transcription factor) homologs predicted in the *Eucalyptus grandis* genome. The expression of 353 candidate xylem-expressed transcription factor genes was quantified in the immature xylem of 156 and 127 F_2 backcross progeny of an *E. grandis* x *E. urophylla* F_1 hybrid respectively using Illumina RNA-Seq expression profiling. Many of the SCW related transcription factor genes shared *trans*-acting eQTLs or had shared *cis*- and *trans*-eQTLs, indicating the presence of polymorphisms affecting the expression of these SCW transcription factors and their downstream targets in

the transcriptional network. Using information from expression correlation across 283 individuals and from shared transcription factor eQTLs, regulatory modules of the *Eucalyptus* SCW transcriptional network were partially reconstructed allowing the identification of known and novel candidate transcription factor genes and their genetic interactions that represent novel targets for functional studies in woody plants.

Preface

The main component of woody biomass in trees is the secondary cell walls (SCWs) of the xylem tissue in stems and branches. One of the major plantation crops that is planted for its woody biomass comprise a group of fast-growing *Eucalyptus* tree species and their hybrids. The formation of this tissue is controlled by a complex regulatory and metabolic network which has been the topic of many recent studies. The purpose of these studies is to determine which genes and transcription factors are involved in the formation of wood, with the aim of increasing the amount and improving the properties of woody biomass. Unravelling the regulatory network of transcription factors directing the formation of the woody biomass may allow for the modification or enhancement of expression of several biosynthetic genes at the same time and rewiring of the network to modify the regulation of particular biopolymers in the SCWs of wood fibre cells.

The majority of traits that are of agronomic importance are quantitative traits. The dissection of these traits will allow for the improvement of crop and plantation species, as well as allowing an understanding of the genetic architecture of the traits. The expression profiles of genes (measured as transcript abundance) in a population can be treated as quantitative traits and as such the genomic regions contributing to variation in expression can be mapped in a similar manner to traditional quantitative trait loci (QTLs), an approach referred to as eQTL mapping. These eQTLs can then be used, with additional supporting information, to infer modules (groups of interacting transcription factors) or parts of modules of the transcriptional networks regulating the expression of the genes. This is done by classifying the eQTLs as either *cis*- or *trans*-eQTLs and identifying transcription factor genes that share *trans*-eQTLs, evidence for underlying regulatory polymorphisms affecting two or more functionally related or interacting transcription factor genes. A *cis*-eQTL implies that the polymorphism that is altering the expression of the gene is close to or at the same position as the gene itself. A *trans*-eQTL implies that the polymorphism is located elsewhere in the genome and affecting the expression of the gene in a trans-acting fashion.

In this study, I examined the expression profiles of *Eucalyptus* homologs of transcription factors that are thought to be part of the SCW regulatory network. This study was performed on an F_2 pseudo backcross population generated in 2005 and first analysed by Kullan et al. (2011). The study of these SCW transcription factor homologs allows for the characterisation of some of the underlying genetic architecture of woody biomass formation. Chapter 1 reviews the literature on QTLs, eQTLs and transcriptional network inference as well as the economic importance of *Eucalyptus* as a crop species. Chapter 2 is written in the format of a manuscript for a peer reviewed journal as is the convention of our research program (Forest Molecular Genetics). In Chapter 2, I report the selection of SCW transcription factor homologs, their gene expression profiles, eQTL mapping and the generation of partial regulatory networks using all of the available information. My study revealed that there are large differences in genetic architecture of transcript abundance between the two backcross families which can be attributed to the different alleles segregating from the F_1 hybrid used as parent in both backcrosses. The MSc study also identified several novel interactions that appear to play a role in SCW regulation.

The following conference proceedings have emanated from this MSc work:

National:

M. Calvert, S.G. Hussey, E. Mizrachi, A.A. Myburg. "Genetic architecture of secondary cell wall related transcription factor networks in wood-forming tissues of a pseudo-backcross pedigree of *E. grandis* x *E. urophylla*." South African Genetics & Bioinformatics Society Conference, Tshwane, South Africa, September 23-26, 2014.

(Poster Presentation)

International:

M. Calvert, S.G. Hussey, E. Mizrachi, A.A. Myburg. "Genetic dissection of gene expression variation of secondary cell wall related transcription factors in *Eucalyptus* hybrid populations." Plant and Animal Genome XXII, San Diego, CA, USA, January 11-15, 2014.

(Poster Presentation)

Myburg AA, Calvert M, Singh P, Mbanjo G, Van der Merwe K, O'Neill MM, Reynolds M, Christie N, Hussey SG, Mizrachi E. 2014. Population genomics unravels genetic diversity and regulation of growth and development in *Eucalyptus*. Australasian Genomics Technology Association 2014 Meeting, 12-15 October 2014, Crowne Promenade, Melbourne, Australia (Invited Plenary Presentation by AA Myburg).

Acknowledgements

I would like to acknowledge my supervisors, Prof Myburg, Dr Mizrachi, and Dr Nanette Christie for all of their support and assistance during this project. It would never have come together and been completed without them. The bioinformatics support that I received from Dr Nanette Christie and Ms. Karen van der Merwe was amazing. They are unbelievably patient individuals who dealt with a lot of questions. I would also like to thank Dr Steven Hussey for sharing his knowledge about transcription factors. I have to thank the Forest Molecular Genetics Programme, the Department of Genetics and Forestry and Agricultural Biotechnology Institute (FABI) for the use of their facilities, social support and for providing a rich learning environment. I would also like to thank the NRF, DST, Sappi, THRIP and other funding bodies without whom none of this work would have been possible. Finally I have to acknowledge my family and friends for all of their support and encouragement while working on this project. It was a lot easier to do knowing that there were people behind me every step of the way.

Table of Contents

List of Tables and Figures

Supplementary Materials

Chapter 1

Literature Review:

Quantitative genetics of transcriptional regulation in plants

1.1 Introduction

One of the most widely cultivated plantation tree genera is *Eucalyptus*. This is due to the favourable growth properties, such as height and wood density, of many species in the genus and the interspecific hybrids that they are able to form (Ladiges et al., 2003). They are also capable of growing in a variety of different environments and are used to manufacture a large variety of products, for example pulp, paper and specialised cellulose. Recently it has been identified that *Eucalyptus* is a potentially valuable source of biomass for the production of biofuels (Saxena et al., 2009, Shepherd et al., 2011).

The main source of woody biomass in *Eucalyptus* comes from the secondary cell walls (SCW) of the xylem tissue, which is formed by a process known as xylogenesis, i.e. wood formation (Mizrachi et al., 2012, Scheller & Ulvskov, 2010, Delmer & Amor, 1995). The SCW of the xylem tissue are composed of cellulose, hemi-celluloses (such as xylan) and lignin. These molecules are closely linked, and together provide strength and support to the xylem fibres and tracheary elements. Over the past few years much more information has become available on genes involved in the biosynthetic pathways producing these molecules (Mizrachi et al., 2012, Oikawa et al., 2010, Scheller & Ulvskov, 2010, Hussey et al., 2011, Zhong et al., 2008, Zhong et al., 2011). While there have been advances in understanding the transcriptional control of these pathways (Demura & Fukuda, 2007, Hussey et al., 2013, Zhong et al., 2011), they are still not fully understood in *Eucalyptus*, and it is not completely understood how variation in transcriptional regulators affects the cell wall properties.

The use of expression Quantitative Trait Loci (eQTL) analysis to study regulatory pathways and reconstruct transcriptional networks is becoming more common (Hansen et al., 2008, Kloosterman et al., 2012, Mackay et al., 2009). This approach uses the gene expression variation of a gene, measured by microarray or RNA-Seq analysis in a segregating population, as a quantitative trait. This allows for the identification of *cis*-acting eQTLs, polymorphisms that are at the same physical

location as the gene, and *trans*-acting eQTLs, that are at a different physical location in the genome to that of the gene. These *trans*-acting eQTLs can be used to identify genes which influence the expression of others and as such may be genetically upstream of the target gene(s), i.e. it may be a transcriptional regulator. In this way it is possible to identify regions of the genome, and as such candidate genes, that were not previously known to be associated with a gene or trait. As gene expression is under transcriptional regulation it is a way of elucidating the transcriptional network regulating a specific pathway or phenotype. The *trans*-eQTL for different genes involved in similar functions or the same pathways may form clusters known as eQTL hotspots. These hotspots may be formed by the presence of a transcriptional regulator which is segregating in the population under study and influencing multiple genes either directly or indirectly. It is important to note that the term "regulator" can refer to a multitude of molecular mechanisms, such as transcription factors, miRNAs, lncRNAs and siRNAs, and can also be the result of polymorphisms in a metabolic gene which results in the metabolic regulation of many related genes to compensate for the polymorphism.

1.2 Quantitative Trait Loci

A quantitative trait is a trait that can show a continuous distribution in a population, depending on which genes are polymorphic in that specific population, i.e. it is not just present or absent, is affected by multiple genes and is influenced by environmental interactions. Often these traits comprise those that are of economic importance in crops, traits such as growth, yield, disease resistance and nutrient composition (composition of proteins, carbohydrates and oils). This makes the study of quantitative traits of great importance in terms of improving livestock, agricultural and plantation crop species.

Quantitative trait loci (QTLs) are regions of the genome that are statistically associated with a quantitative trait. These traits are analysed with QTL mapping, which is the localisation of a quantitative trait with molecular markers based on its genetic linkage to a specific marker locus

(Mackay et al., 2009). Lander and Botstein (1989) first described QTL mapping using the Interval Mapping (IM) approach. This approach uses one marker on either side of the proposed QTL, and by means of LOD score analysis, determines whether the association exceeds that which is expected by chance (Lander & Botstein, 1989). Subsequently Composite Interval Mapping (CIM) and Multiple Interval Mapping (MIM) were developed to reduce the background noise and more precisely identify smaller regions associated with the quantitative trait (Kao et al., 1999, Zeng, 1993, Zeng, 1994). CIM was developed to minimize the bias created by several QTLs on the same chromosome and increase the power of QTL detection (Zeng, 1993, Zeng, 1994). CIM restricts the statistical analysis to a defined interval; minimizing the effect of other QTLs segregating outside of this interval (Zeng, 1994). MIM, however analyses multiple intervals at the same time, increasing the statistical power and the precision with which the QTL is mapped (Kao et al., 1999). One of the requirements for QTL mapping is a genetic linkage map of the molecular markers for that species or population. In the initial paper by Lander and Botstein (1989) a genetic linkage map consisting of Restriction Fragment Length Polymorphisms (RFLPs) was used to develop the statistical approach. There are now many different types of molecular markers, such as polymorphic insertions or deletions (indels), simple sequence repeats (microsatellites) or single nucleotide polymorphisms (SNPs) that are used to generate genetic linkage maps (Mackay et al., 2009).

Two types of QTL mapping studies can be performed, linkage mapping and association mapping. In linkage mapping the QTLs are mapped in a specific pedigree produced by crossing divergent strains or species. In association mapping the QTLs are mapped in unrelated individuals from the same population. Genome-Wide Association Studies, or GWAS, are an example of association mapping (Mackay et al., 2009). Associations discovered by GWAS are not considered as QTLs even though they can refer to the same type of polymorphism.

The success of QTL detection depends on different aspects of the experimental procedure. The power to detect QTLs is dependant not only on the size of the effect of the QTL but also on allele frequencies in the population and the size of the experimental population under study (Jansen &

Nap, 2001, Mackay et al., 2009). As the effect size of the QTL decreases more individuals are required for successful detection, since it is easier to detect QTLs with larger effects than those with smaller effects. This has resulted in effect size of QTLs often being overestimated (Beavis et al., 1994). As such, an increase in the number of individuals used in QTL mapping experiments has resulted in the detection of more QTLs of smaller effect. Another factor on the detection of QTLs is the presence of linkage in the population under study. If there are large linkage blocks in the population fewer markers are required, however the regions of the genome associated with the trait will be very large, as there is more distance to the nearest molecular marker, which may hamper the identification of candidate genes (Sutter & Ostrander, 2004). One method of further narrowing the selection of candidate genes in QTL mapping is through the use of expression QTLs (eQTLs).

1.2.1 Expression QTLs

Expression QTLs are similar to QTLs, except the quantitative phenotype being measured is the abundance of transcripts produced by the gene, or the expression of the gene. The use of eQTLs to better understand quantitative traits was first proposed by Jansen and Nap (Jansen & Nap, 2001). They proposed using the genetic variation observed in related segregating progeny to identify polymorphisms affecting quantitative traits, and how this is related to other quantitative phenotypic traits (Jansen & Nap, 2001). eQTL analysis is based on the same principles as conventional QTL analysis, yet can provide more information about the biological processes of interest. The combination of QTL and eQTL analysis has the potential to narrow the number of candidate genes identified, as well as possibly showing which other genes have an effect on a specific trait (Jansen & Nap, 2001, Mackay et al., 2009).

1.2.2 *Cis*- and *trans*-eQTLs

There are two different types of eQTLs, namely *cis*-eQTLs and *trans*-eQTLs (Jansen & Nap, 2001; Figure 1.1). The definition of *cis*-eQTLs and *trans*-eQTLs is dependent on the resolution of the genetic map and the number of individuals used in the study, as an eQTL may be classified as a *cis*-

eQTL when it is within a certain number of centiMorgans (cM) from the gene. The detection of molecular markers depends on the type of molecular markers to be used, the species and the population under study.

It has been noted that *cis*-eQTLs are generally of a larger effect than *trans*-eQTLs (Mackay et al., 2009, Drost et al., 2010). This is possibly due to the nature of the polymorphism responsible for generating the eQTL. A *cis* sequence polymorphism has a direct effect of the expression of a gene, as such making the effect larger. As transcriptional abundance is often regulated by multiple factors, a polymorphism in one regulator is likely to only have a small effect on the expression of the genes under its regulation. As such the polymorphism underlying a *trans*-eQTL is likely to affect multiple genes. As multiple genes are often affected, a *trans*-eQTL can be considered pleiotropic. Large-effect mutations in pleiotropic genes are often harmful, which may also be a constraint on the effect size of *trans*-eQTLs (Hansen et al., 2008)

The target genes of a transcription factor could have *trans*-eQTLs at the position of the transcription factor, which may itself have a *cis*-eQTL (Hansen et al., 2008). The variation in expression caused by the *cis*-eQTL can affect multiple downstream target genes, and as such cause phenotypic variation. There are multiple different causes of *cis*-eQTLs, as promoter polymorphisms, indels, splicing variants or differential RNA degradation can all generate differential transcript abundance (Hansen et al., 2008).

There are two methods by which to draw networks from eQTL data. The first method is *a priori* network analysis, where the network being drawn or tested is known or predicted. The other method of network analysis is *a posteriori* network analysis which is used for the identification of novel networks and novel regulation. This method generally uses the correlation of expression patterns of genes or the sharing of eQTLs to identify gene clusters or networks (Hansen et al., 2008). These networks can use different definitions for the nodes and edges depending on the main purpose of the study (Keurentjes et al., 2007, Kliebenstein, 2009)

.

There have been multiple studies done using eQTL analysis. One of the first eQTL studies that was performed on a plantation tree species was done by Kirst et al. (2005) on a *Eucalyptus grandis* x *E. globulus* hybrid. In this study they examined how transcript abundance was regulated among individuals of an interspecific segregating population. They also examined the conservation of regulation of orthologous gene transcripts (Kirst et al., 2005). They determined that the high environmental variation in tree plantations affects the ability to detect eQTLs by lowering the statistical power of the study. They also determined that the eQTLs clustered in specific genomic regions. The interaction between these loci plays an important role in the control of transcript abundance (Kirst et al., 2005). The eQTL hotspots that were identified may be in gene-rich regions or they may be overlying a transcriptional regulator that controls multiple downstream genes. These regions may be used to identify genetic loci that control the flux through metabolic or regulatory pathways.

A study by Drost et al. (2010) in *Populus* was one of the first to draw transcriptional networks using eQTL data. This study focused on how gene expression and the related transcriptional networks were involved in organ differentiation. This was done because it is known that differential gene expression is one of the critical ways in which phenotypic diversity is created in eukaryotes (Jansen & Nap, 2001). One of the main discoveries in this study was that *trans*-eQTLs play a larger role in cell or organ differentiation than when compared to a *cis*-eQTL for the coding region (Drost et al., 2010). This is because *trans*-eQTLs typically affect multiple downstream genes which are functionally related, and therefore more likely to be involved in a pathway. *Cis*-eQTLs however are by default stochastic, i.e. the result of random mutation in or near genes, and therefore we do not expect strong functional enrichment.

1.3 Genetic Architecture

Genetic architecture is the description of the genetic interactions and structure that underlie a specific phenotypic trait. This will include the number of loci involved with the production of that phenotype, the interaction between and regulation of those loci, the alleles that potentially cause variation in the loci and any other factors that can affect the genetic underpinnings of a phenotype (Mackay et al., 2009, Mackay, 2001).

It is often difficult to get a comprehensive view of the genetic architecture of any one phenotypic trait, especially those that are quantitative in nature (Mackay, 2001). There are many ways of determining genetic architecture ranging from experimental approaches, such as experimental crosses and linkage mapping, to computational approaches, such as QTL and eQTL mapping and genetic network analysis (Civelek & Lusis, 2014, Mackay, 2001, Porth et al., 2013). It is often difficult to prescribe a specific variation in phenotype to one specific gene or transcript. This is due to the nature of these types of study; the effects that are discovered are often specific to the population and the environmental conditions of the experiment (Mackay, 2001).

There have been severally studies done that look at the genetic architecture of specific traits. The majority of these studies do not elucidate the full nature of the trait, as there is not enough information available to provide a complete examination (Mackay, 2001). One of these studies was recently performed by Porth et al. (2013). In this study they looked at the genetic architecture for 6 different wood traits and their interaction networks. This was done by looking at the differentially expressed transcripts and SNPs that were present or absent in different phenotypic extremes in the population. They identified multiple transcripts that were differentially expressed and SNPs that were dependent on the specific phenotypes. They determined that due to the complexity of the genetic architecture of the specific wood phenotypes under examination, it will be challenging to produce any genetic improvements for economically important traits using molecular markers as opposed to phenotypic selection (Porth et al., 2013).

1.4 Transcription Factors

Transcription factors (transcription factors) are involved in the activation, repression and modulation of transcription of genes. As such they are instrumental in the regulation of the expression of genes in different tissues and conditions. Multiple transcription factors are involved in the regulation of one gene, these transcription factors are in turn regulated by other transcription factors; as such transcriptional networks are formed. Transcriptional networks contain transcription factors, represented by nodes, and interactions, whether activation or repression, represented as edges (Hussey et al., 2013). They help to show which transcription factors interact with each other and in turn with which genes, as well as the manner in which they interact. This all contributes to a better understanding of the phenotype that is produced by the gene or pathway and what influences it.

Multiple families of transcription factors exist which have different characteristics and modes of action depending on the family. These characteristics are usually defined by several specific domains in the protein, such as the activation and the binding domains. These domains are characterised by specific DNA and protein sequences. This allows for the prediction of transcription factors, from genomic sequence, as well as of target genes and modes of action (Perez-Rodriguez et al., 2010). A database of the transcription factors in plants Planttranscription factorDB 3.0, is available online. They contain information on the different transcription factor families, as well as the protein and nucleotides sequences in different plant species (Jin et al., 2014, Perez-Rodriguez et al., 2010, Riano-Pachon et al., 2007). They contain the same seven taxonomic groups, the Bangiophyceae, Prasinophyceae, Chlorophyceae, Bryophyte, Lycopodiophyta, Monocot and Eudicot, and 83 species and are divided into 84 and 58 gene families (Jin et al., 2014, Perez-Rodriguez et al., 2010). They are an exceptionally useful resource for the plant research community as they can assist with the identification of transcription factors and their sequence characteristics in different species.

1.4.1 Transcriptional Networks

Transcriptional networks are an intricate component of gene regulation. These types of networks often share similar features, such as structure and organisation (Bhardwaj et al., 2010). The position of a transcription factor in the network may provide prior expectations about its regulation and expression. Often transcription factors within a specific hierarchical layer will have similar properties to other transcription factors in that layer, yet those properties will differ depending on the specific layer (Jothi et al., 2009). This allows for variation in the network and may confer a selective advantage under certain conditions, such as when it is energetically expensive to maintain the expression of a specific transcription factor and as such that transcription factor will require a very low concentration to achieve the desired effect or the pathway it activates may be activated by another transcription factor as well (Jothi et al., 2009). The core or middle layer transcription factors of a transcriptional network is well known to have the most connections to other transcription factors in the network. Genes in this layer are often able to regulate within the same layer and exhibit co-regulation of downstream targets (Bhardwaj et al., 2010, Jothi et al., 2009). It is also know that the higher up in the hierarchy a transcription factor is, the more conserved and versatile (involved in multiple pathways) that transcription factor (Jothi et al., 2009). This is often seen because the top-layer, or master, transcription factors are involved in several different networks. It has also been determined that post-translational regulation or modification of transcription factors is important in ensuring the correct amount of transcription factor is available in the cell when needed (Jothi et al., 2009). This shows that the level of gene expression of a transcription factor may not be a complete indication of its level of activity.

When constructing transcriptional networks it is important to consider the information available and how to utilise it in the most appropriate way. There are multiple different models available that can be used to construct these networks, such as Bayesian network construction, Correlation network construction, Gaussian graphical model-based network construction and Maximum Likelihood

network construction (Schadt et al., 2005, Zhang et al., 2010, Zhu et al., 2012, Koo et al., 2014). All of these methods have their own specific advantages and disadvantages, and are often dependent on the computational resources available, as well as the type of data and available resources for the species. It is important to remember that over-fitting of the data to a specific model (having a model that is to stringent) can result in incorrect or skewed results, as such procedures like 'leave-one-out' analysis can be useful (Porth et al., 2013).

Transcriptional networks can be said to be democratic or autocratic in nature (Bhardwaj et al., 2010). Most networks show a combination of these two approaches to governing expression. An autocratic network is a network in which several main regulators regulate their own targets, or sets of targets, and have chains of influence. A democratic network is a network in which many regulators regulate many other targets cooperatively. This results in multiple genes working cooperatively to get a specific response. These two different types of regulation result in the network being organised into neat hierarchies, in the case of autocratic regulation, and a lack of specific hierarchies or structure, in the case of democratic regulation (Bhardwaj et al., 2010).

1.5 Tree Development

Trees are eukaryotic organisms with multiple different industrial uses, such as pulp, paper and specialised cellulose. The type of tree planted for a specific product is determined by the stature of the tree, which refers to the size, and the architecture, which refers to the form of the tree (Grattapaglia et al., 2009). Another set of characteristics that are important in plantation trees are the growth rate of the tree and the wood density (Grattapaglia et al., 2009). Tree growth comes from the division and expansion of cells in the apical (primary) and cambial (secondary) meristems, which results in growth in height and diameter respectively. Wood density is the relative proportion of cell wall thickness and cell size. These are the characteristics that are most often considered when selecting plantation trees for a specific industrial purpose.

The most important part of the tree with regards to industrial uses is the woody biomass of the tree, which is composed of the secondary cell walls (SCW) of the xylem tissue. The difference between the primary and SCWs is the lignification that takes place in the SCWs and composition of the cellulose, hemicelluloses and pectins in the cell wall (Scheller & Ulvskov, 2010, Lee et al., 2011). The xylem tissue, which is composed of xylary fibres, vessels and tracheary elements, forms the 'wood' of the tree as opposed to the bark, which is formed by the phloem tissue. Xylem is responsible for the transport of water and mineral salts in the tracheary elements, and is formed during primary growth by the procambium and during secondary growth by the vascular cambium.

1.5.1 Secondary Cell Walls

The SCWs of terrestrial plants are mainly composed of cellulose, hemi-celluloses (such as xylan, xyloglucans, mannans and glucomannans) and lignin (Lee et al., 2011, Mizrachi et al., 2012, Oikawa et al., 2010, Scheller & Ulvskov, 2010, York & O'Neill, 2008). The composition of hemicelluloses in pant cell walls varies between different plant species and cell types, an example of which is β-(1→3, 1→4)-glucans which are found only in Poales and several other groups (Scheller & Ulvskov, 2010).

Wood is one of the richest sources of carbon-rich biomass available, which has led to interest in its possible use in the production of biofuels (Saxena et al., 2009). Cellulose is the main component of SCWs which make up the wood. Traditionally the research regarding cellulose biosynthesis has focused on cellulose synthase proteins (CESA) located on the plasma membrane, which make up the cellulose synthase complex (CSC). Cellulose is mainly composed of cellobiose, which is a chain of glucose molecules bound via a β 1-4 linkage to form β 1-4-D-glucan (Delmer & Amor, 1995). The chains that are formed are linear and extended. This means that they can interact with each other, and other molecules, in a specific manner to form a rigid structure (Delmer & Amor, 1995). These interactions are generally in the form of hydrogen bonds and the resulting chains are referred to as microfibrils. In SCW these microfibrils often join together in what is referred to as a macrofibril

(Delmer & Amor, 1995). The side-chain substitutions of the hydroxyl groups of C_2, C_3 and C_6 alter the properties of the cellulose and as such its industrial applications (Mizrachi et al., 2012). There are multiple different enzymes involved in cellulose biosynthesis, some of these enzymes can be seen in Figure 1.2 which gives a schematic representation of the formation of cellulose and xylan.

Xylan is the major hemicellulose found in dicotyledonous plants (Scheller & Ulvskov, 2010). It is composed of 1,4-linked-β-D-xylopyranosyl residues that can contain multiple different side chain residues, such as glucuronic acid or 4-O-methyl glucuronic acid, arabinose or a combination of neutral and acid sugars which forms glucuronoxylan, arabinoxylan and glucuronoarabinoxylan respectively (Scheller & Ulvskov, 2010, York & O'Neill, 2008). In plants that contain SCW there is a unique sequence of glycosyl residues at the reducing end which is required for the formation of xylan in the Golgi bodies (York & O'Neill, 2008). There are three main stages to xylan biosynthesis, backbone biosynthesis, side-chain modification and reducing end biosynthesis (Lee et al., 2012a, Lee et al., 2012b, Oikawa et al., 2010, Scheller & Ulvskov, 2010, York & O'Neill, 2008). The biosynthetic pathway of xylan, including some of the enzymes involved, can be seen in Figure 1.2.

Lignin biosynthesis is a more complicated biosynthetic process that cellulose or xylan biosynthesis, as there are multiple different ways in which carbon can move through the phenylpropanoid pathway. After cellulose lignin is the most abundant biopolymer found in nature (Boerjan et al., 2003). It is derived from three hydroxycinnamyl alcohol monomers that differ in their degree of methoxylation, *p*-coumaryl, coniferyl and sinapyl alcohols. These monomers produce, *p*-hydroxyphenyl (H), guaiacyl (G) and syringyl phenylpropanoid units (S) respectively (Boerjan et al., 2003, Vanholme et al., 2008). As with hemicelluloses the composition of the different monolignol units differs among plant species, for example in hardwood trees the lignin is mainly composed of G and S units with small amounts of H units, whilst softwood trees have lignin that is mainly composed of G units with small amounts of H units (Vanholme et al., 2008, Boerjan et al., 2003). An example of the biosynthetic pathway that produces the different monolignol units is given in Figure 1.3.

In the nuclei of the developing xylem cells, specific families of transcription factors that have been identified as being involved in the regulation of SCW biosynthesis. These families include the NAM/ATAF/CUC (NAC) family, MYELOBLASTOSIS (MYB) family, AUXIN RESPONSE FACTOR (ARF) family, CLASS II HOMEODOMAIN-LEUCINE ZIPPER (HD-ZIPIII) family and the KANADI (KAN) family (Demura & Fukuda, 2007, Yamaguchi et al., 2010, Zhong et al., 2008). These transcription factors are controlled by different regulatory signals, often in the form of hormones, at different stages of the plants development.

The main regulatory transcription factors involved in the formation of SCWs are NAC transcription factors, which are often referred to as the secondary wall NACs (Zhong et al., 2011). These NAC proteins are functionally redundant and often function in pairs. They include NAC SECONDARY WALL THICKENING PROMOTING FACTOR1 (NST1) and SECONDARY WALL ASSOCIATED NAC DOMAIN PROTEIN1 (SND1) as a pair, and VASCULAR RELATED NAC DOMAIN6 (VND6) and VND7 as another pair (Kubo et al., 2005, Yamaguchi et al., 2010, Zhong et al., 2006, Zhong et al., 2008, Zhong et al., 2007). These pairs are found in the nuclei of the developing xylem within the plant, with NST1 and SND1 being found in the xylary and interfascicular fibres, the siliques valve endocarps and valve margins, and VND6 and VND7 being found in the vessels. These pairs of SCW NACs co-regulate a common set of targets and as such VND6 and VND7 are able to complement the *snd1 nst1* double mutant (Yamaguchi et al., 2010, Zhong et al., 2011). They do not co-regulate all the same targets however, as can be seen in the differing ability to activate certain programmed cell death genes. If these genes are not activated, the vessels are unable to reach maturity (Ohashi-Ito et al., 2010). A recent review of SCW regulatory mechanisms can be found in Hussey et al., (2013).

1.6 *Eucalyptus*

The hardwood genus *Eucalyptus* forms part of the Myrtaceae family and is the most used short-rotation hardwood plantation species (Myburg et al., 2014, Ladiges et al., 2003). The genus evolved

in Australia and its surrounding islands, and grows in tropical, sub-tropical and temperate regions depending on the specific species (Ladiges et al., 2003). As the different species grow in such a wide variety of environments, there is significant genomic variation within and between the species, yet there is high genome interspecific collinearity between the species (Myburg et al., 2014). The different species readily form hybrids which allows for the selection of specifically desired growth, wood quality and defence traits, and as such the tailoring of the plantation tree to a specific environment and purpose. Species that are commonly grown as plantation species are *Eucalyptus grandis, E. globulus, E. camaldulensis, E. urophylla, E. nitens* and hybrids of these species (Myburg et al., 2011, Myburg et al., 2014, Poke et al., 2005). These species grow well in plantation regions and have desired wood and growth properties. In tropical and subtropical regions, hybrids of *E. grandis* and *E. urophylla* are commonly grown as this hybrid shows excellent growth properties in these regions, as well as fungal disease resistance which is contributed buy *E. urophylla* (Ladiges et al., 2003, Myburg et al., 2011, Myburg et al., 2014, Poke et al., 2005). In temperate regions the most commonly grown species are *E. globulus* and *E. nitens.* Together these species are the targets of breeding programs to enhance tree quality and production (Poke et al., 2005).

A draft sequence of the *E. grandis* genome was produced at the DOE Joint Genomes Institute (JGI), funded by the US Department of Energy (DOE) in collaboration with the *Eucalyptus* Genome Network (EUCAGEN, *E. grandis* V1.1, JGI, www.phytozome.net (Myburg et al., 2011, Myburg et al., 2014). The tree that was sequenced was a partially inbred, i.e. one generation of self-fertilization, 17 year-old *E. grandis* tree from Brazil (BRASUZ1). The specimen had to be partially inbred due to the difficulty in sequencing and assembling a highly heterozygous genome. This new resource has made investigation of the genome of *Eucalyptus* much easier, as it provides the opportunity to perform in depth transcriptome analysis and is extremely valuable with regards to marker-assisted breeding. There is evidence for the duplication events that are known to have taken place in the rosid clade, as well as unique lineage-specific genome duplications in *Eucalyptus*. This can often be seen when candidate genes are selected for analysis. There have already been several studies

utilising QTL and eQTL data in *Eucalyptus* species, an example of which is the study by Kirst et al. (2005) which was discussed previously.

1.7 Conclusion

Phenotypes are often determined by the proteins that contribute to growth and development. These proteins are in turn controlled by the regulation of the genes which are responsible for their production. Transcriptional networks ultimately underlie desirable and non-desirable phenotypic characteristics. These networks can be highly complex and inter-connected. A small manipulation of one transcription factor has the potential to influence multiple nodes of a network and the biological systems associated with these nodes.

The study of transcriptional networks and their genetic architecture has become more comprehensive through the use of transcriptomics, eQTLs and other molecular techniques such as yeast-two-hybrid approaches. This allows for a better understanding of the effects a specific manipulation may have on the network and resulting phenotype. Once a more comprehensive understanding of a transcriptional network is achieved, a way to engineer a desired phenotype with minimal disruption of the network or other related traits can be determined. The resulting plant will be tailored to the manufacture of a specifically desired product. In terms of forestry plantations this will increase the productivity of the plantations, thereby decreasing the amount of land required to meet production goals.

This study will contribute to our understanding of the regulatory network underlying SCW biosynthesis in *Eucalyptus* hybrids. As the majority of the work on the regulatory networks of SCWs has been done in *Arabidopsis* and occasionally in *Populus*, the expansion of this work to *Eucalyptus* has the potential to determine novel interactions. This could then be applied to other tree species, as well as confirming the validity of the discoveries made in other species.

1.8 References

Beavis WD, Smith OS, Grant D, Fincher R, 1994. Identification of Quantitative Trait Loci using a small sample of topcrossed and F4 progeny from maize. *Crop Science* **34,** 882-96.

Bhardwaj N. YK, Gerstein M.B., 2010. Analysis of diverse regulatory networks in a hierarchical context shows consistent tendencies for collaboration in the middle levels. *Proceedings of the National Academy of Sciences* **107**, 6841-6.

Boerjan W, Ralph J, Baucher M, 2003. Lignin biosynthesis. *Annual Review Plant Biology* **54**, 519-46.

Brown DM, Goubet F, Wong VW, et al., 2007. Comparison of five xylan synthesis mutants reveals new insight into the mechanisms of xylan synthesis. *The Plant Journal: for cell and molecular biology* **52**, 1154-68.

Brown DM, Zeef LA, 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**, 2281-95.

Civelek M, Lusis AJ, 2014. Systems genetics approaches to understand complex traits. *Nature Reviews Genetics* **15**, 34-48.

Delmer DP, Amor Y, 1995. Cellulose Biosynthesis. *Plant Cell* **7**, 987-1000.

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

Drost DR, Benedict CI, Berg A, et al., 2010. Diversification in the genetic architecture of gene expression and transcriptional networks in organ differentiation of *Populus*. *Proceedings of the National Academy of Sciences* **107**, 8492-7.

Grattapaglia D, Plomion C, Kirst M, Sederoff RR, 2009. Genomics of growth traits in forest trees. *Current Opinion in Plant Biology* **12**, 148-56.

Grattapaglia D, Resende M, 2011. Genomic selection in forest tree breeding. *Tree Genetics & Genomes* **7**, 241-55.

Hansen BG, Halkier BA, Kliebenstein DJ, 2008. Identifying the molecular basis of QTLs: eQTLs add a new dimension. *Trends in Plant Science* **13**, 72-7.

Hefer C, Mizrachi E, Joubert F, Myburg AA, 2011. The *Eucalyptus* genome integrative explorer (EucGenIE): a resource for *Eucalyptus* genomics and transcriptomics. *BMC Proceedings* **5**, O49.

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

Hussey SG, Mizrachi E, Spokevicius AV, Bossinger G, Berger DK, Myburg AA, 2011. *SND2*, a *NAC* transcription factor gene, regulates genes involved in secondary cell wall development in *Arabidopsis* fibres and increases fibre cell area in *Eucalyptus*. *BMC Plant Biology* **11**, 173.

Jansen RC, Nap J-P, 2001. Genetical genomics: the added value from segregation. *Trends in Genetics* **17**, 388-91.

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

Jothi R. BS, Wuster A., Grochow J.A,Gsponer J., Przytycka T.M., Aravind L., Babu M.M., 2009. Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture. *Molecular Systems Biology* **5**, 1-15.

Kao C-H, Zeng Z-B, Teasdale RD, 1999. Multiple interval mapping for quantitative trait loci. *Genetics* **152**, 1203-16.

Keurentjes JJ, Fu J, Terpstra IR, et al., 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**, 1708-13.

Kirst M, Basten CJ, Myburg AA, Zeng Z-B, Sederoff RR, 2005. Genetic architecture of transcript-level variation in differentiating xylem of a *Eucalyptus* hybrid. *Genetics* **169**, 2295-303.

Kliebenstein D, West M, Van Leeuwen H, Loudet O, Doerge R, St Clair D, 2006. Identification of QTLs controlling gene expression networks defined *a priori*. *BMC Bioinformatics* **7**, 308.

Kliebenstein DJ, 2009. Quantitative Genomics: Analyzing intraspecific variation using global gene expression polymorphisms or eQTLs. *Annual Review of Plant Biology* **60**, 93-114.

Kloosterman B, Anithakumari A, Chibon P-Y, et al., 2012. Organ specificity and transcriptional control of metabolic routes revealed by expression QTL profiling of source-sink tissues in a segregating potato population. *BMC Plant Biology* **12**, 17.

Koo I, Yao S, Zhang X, Kim S, 2014. Comparative analysis of false discovery rate methods in constructing metabolic association networks. *Journal of Bioinformatics and Computational Biology* **12,** 1450018.

Kubo M, Udagawa M, Nishikubo N, et al., 2005. Transcription switches for protoxylem and metaxylem vessel formation. *Genes and Development* **19**, 1855-60.

Kullan AR, Van Dyk MM, Jones N, Kanzler A, Bayley A, Myburg AA, 2011. 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**, 163-75.

Kullan AR, Van Dyk M, Hefer C, Jones N, Kanzler A, Myburg A, 2012. Genetic dissection of growth, wood basic density and gene expression in interspecific backcrosses of *Eucalyptus grandis* and *E. urophylla*. *BMC Genetics* **13**, 60.

Ladiges PY, Udovicic F, Nelson G, 2003. Australian biogeographical connections and the phylogeny of large genera in the plant family Myrtaceae. *Journal of Biogeography* **30**, 989-98.

Lander ES, Botstein D, 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185-99.

Lee C, Teng Q, Zhong R, Ye Z-H, 2011. Molecular dissection of xylan biosynthesis during wood formation in poplar. *Molecular Plant* **4**, 730-47.

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

Lee C, Zhong R, Ye Z-H, 2012b. *Arabidopsis* family *GT43* members are xylan xylosyltransferases required for the elongation of the xylan backbone. *Plant and Cell Physiology* **53**, 135-43.

Mackay TF, Stone EA, Ayroles JF, 2009. The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics* **10**, 565-77.

Mackay TF, 2001. The genetic architecture of quantitative traits. *Annual Reviews Genetics* **35**, 303-39.

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

Myburg A, Grattapaglia D, Tuskan G, et al., 2011. The *Eucalyptus grandis* Genome Project: Genome and transcriptome resources for comparative analysis of woody plant biology. *BMC Proceedings* **5**, 1-2.

Myburg AA, Grattapaglia D, Tuskan GA, et al., 2014. The genome of *Eucalyptus grandis*. *Nature* **510**, 356-362.

Ohashi-Ito K, Oda Y, Fukuda H, 2010. *Arabidopsis VASCULAR-RELATED NAC-DOMAIN6* directly regulates the genes that govern programmed cell death and secondary wall formation during xylem differentiation. *Plant Cell* **22**, 3461-73.

Oikawa A, Joshi HJ, Rennie EA, et al., 2010. An integrative approach to the identification of *Arabidopsis* and rice genes involved in xylan and secondary wall development. *PLoS ONE* **5**, e15481.

Pérez-Rodríguez P, Riaño-Pachón DM, Corrêa LGG, Rensing SA, Kersten B, Mueller-Roeber B, 2010. Plant transcription factorDB: updated content and new features of the plant transcription factor database. *Nucleic Acids Research* **38**, D822-D7.

Poke F, Vaillancourt R, Potts B, Reid J, 2005. Genomic research in *Eucalyptus*. *Genetica* **125**, 79-101.

Porth I, Klapste J, Skyba O, et al., 2013. Network analysis reveals the relationship among wood properties, gene expression levels and genotypes of natural *Populus trichocarpa* accessions. *New Phytologist* **200**, 727-42.

Riano-Pachon D, Ruzicic S, Dreyer I, Mueller-Roeber B, 2007. Plant transcription factorDB: an integrative plant transcription factor database. *BMC Bioinformatics* **8**, 42.

Saxena RC, Adhikari DK, Goyal HB, 2009. Biomass-based energy fuel through biochemical routes: A review. *Renewable and Sustainable Energy Reviews* **13**, 167-78.

Schadt EE, Lamb J, Yang X, et al., 2005. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics* **37,** 710-7.

Scheller HV, Ulvskov P, 2010. Hemicelluloses. *Annual Review of Plant Biology* **61**, 263-89.

Shepherd M, Bartle J, Lee D, Brawner J, Bush D, 2011. Eucalypts as a biofuel feedstock. *Biofuels* **2**, 639-57.

Sutter NB, Ostrander EA, 2004. Dog star rising: the canine genetic system. *Nature Reviews Genetics* **5**, 900-10.

Vanholme R, Morreel K, Ralph J, Boerjan W, 2008. Lignin engineering. *Current Opinion in Plant Biology* **11**, 278-85.

Vanholme R, Storme V, Vanholme B, et al., 2012. A systems biology view of responses to lignin biosynthesis perturbations in *Arabidopsis*. *Plant Cell* **24**, 3506-29.

Yamaguchi M, Goué N, Igarashi H, et al., 2010. *VASCULAR-RELATED NAC-DOMAIN6* and *VASCULAR-RELATED NAC-DOMAIN7* effectively induce transdifferentiation into xylem vessel elements under control of an induction system. *Plant Physiology* **153**, 906-14.

York WS, O'Neill MA, 2008. Biochemical control of xylan biosynthesis — which end is up? *Current Opinion in Plant Biology* **11**, 258-65.

Zeng ZB, 1993. Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 10972-6.

Zeng ZB, 1994. Precision mapping of quantitative trait loci. *Genetics* **136**, 1457-68.

Zhang W, Zhu J, Schadt EE, Liu JS, 2010. A Bayesian partition method for detecting pleiotropic and epistatic eQTL modules. *PLoS Computational Biology* **6**, e1000642.

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**, 3158-70.

Zhong R, Lee C, Zhou J, McCarthy RL, Ye Z-H, 2008. A battery of transcription factors involved in the regulation of secondary cell wall biosynthesis in *Arabidopsis*. *Plant Cell* **20**, 2763-82.

Zhong R, McCarthy RL, Lee C, Ye Z-H, 2011. Dissection of the transcriptional program regulating secondary wall biosynthesis during wood formation in poplar. *Plant Physiology* **157**, 1452-68.

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**, 1603-11.

Zhu J, Sova P, Xu Q, et al., 2012. Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation. *PLoS Biology* **10**, e1001301.

Tables and Figures

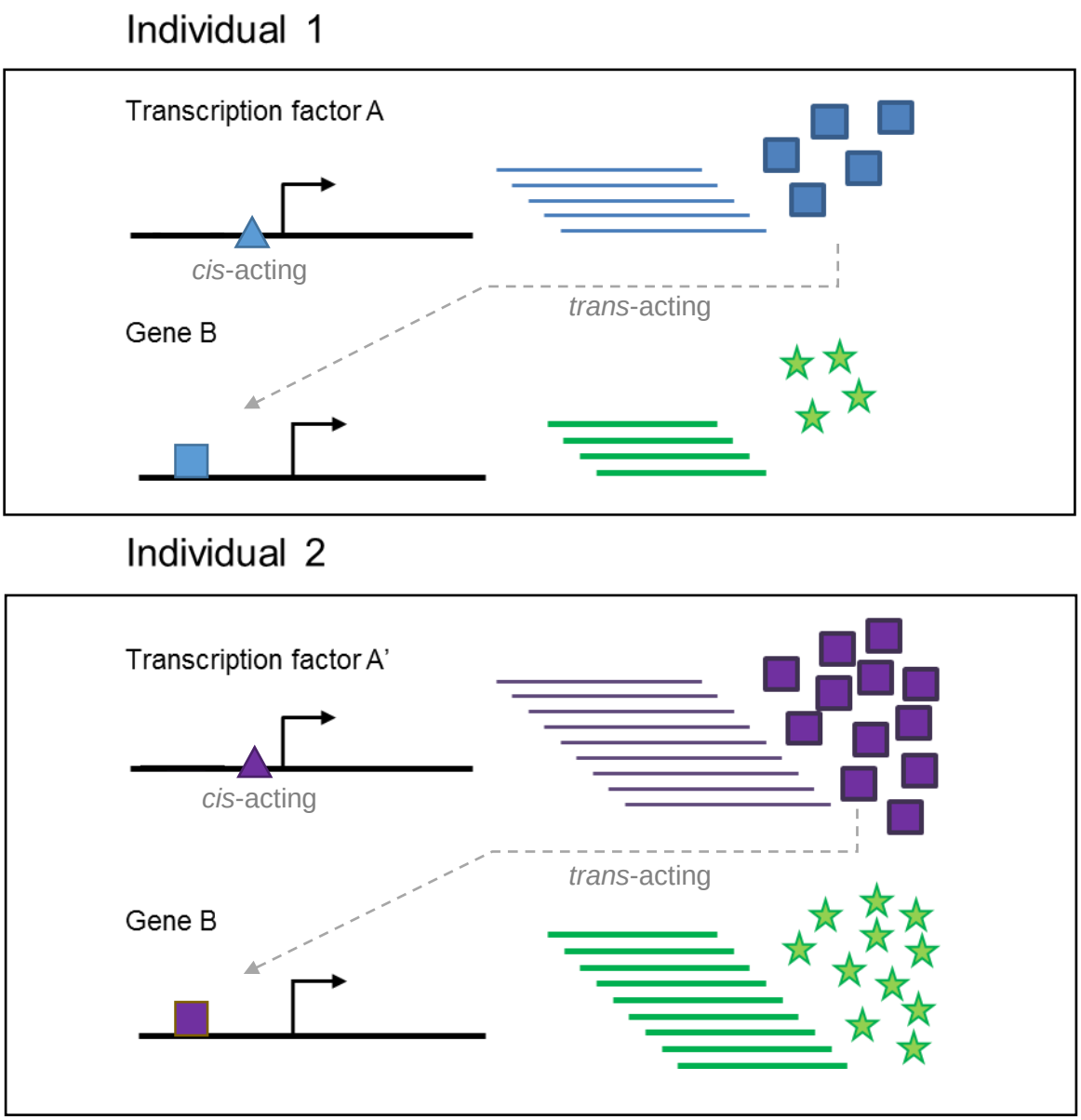

Figure 1.1. Schematic explanation of the difference between cis- and trans-eQTLs. A *cis*-acting polymorphism is represented by the blue and purple triangles in the promoter regions of alleles A and A' of Transcription factor A in individual 1 and 2, respectively. This polymorphism results in higher transcription and higher protein levels of allele A' (purple squares) compared to allele A (blue squares) in the two individuals (ignoring in this example that the other A allele carried by each of the individuals may also be polymorphic). If Transcription factor A regulates Gene B, the same polymorphism in the promoter of Transcription factor A may constitute a *trans*-acting polymorphism resulting in higher expression of Gene B in individual 2 vs individual 1. Note that the trans-acting polymorphism in Transcription factor A is not dependent on any polymorphism in Gene B, although it is likely that Gene B could also be polymorphic and have a cis-acting polymorphism in highly outbred organisms like *Eucalyptus*.

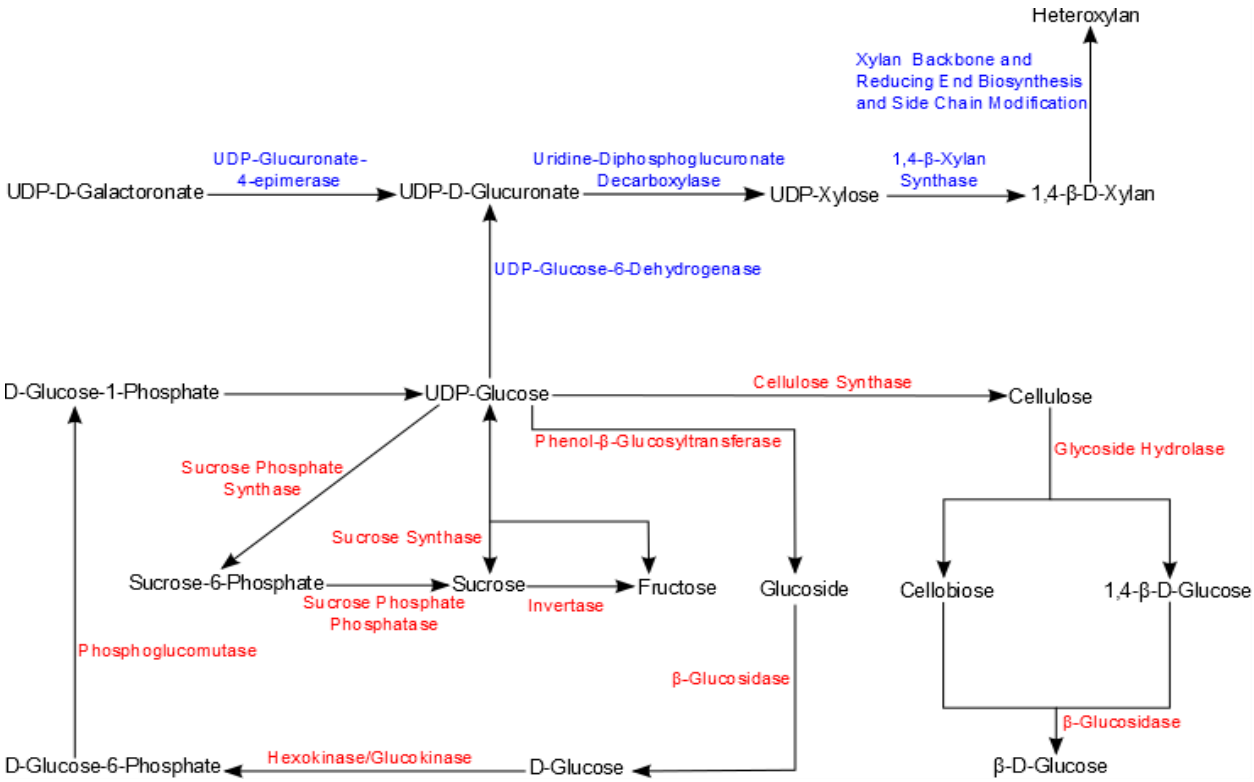

Figure 1.2. Schematic representation of cellulose and xylan biosynthesis. Xylan biosynthetic pathway enzymes are represented in blue and cellulose biosynthetic pathway enzymes are represented in red. Pathway redrawn based on previous reviews (Brown et al., 2007, Brown et al., 2005, Mizrachi et al., 2012, Scheller & Ulvskov, 2010, York & O'Neill, 2008).

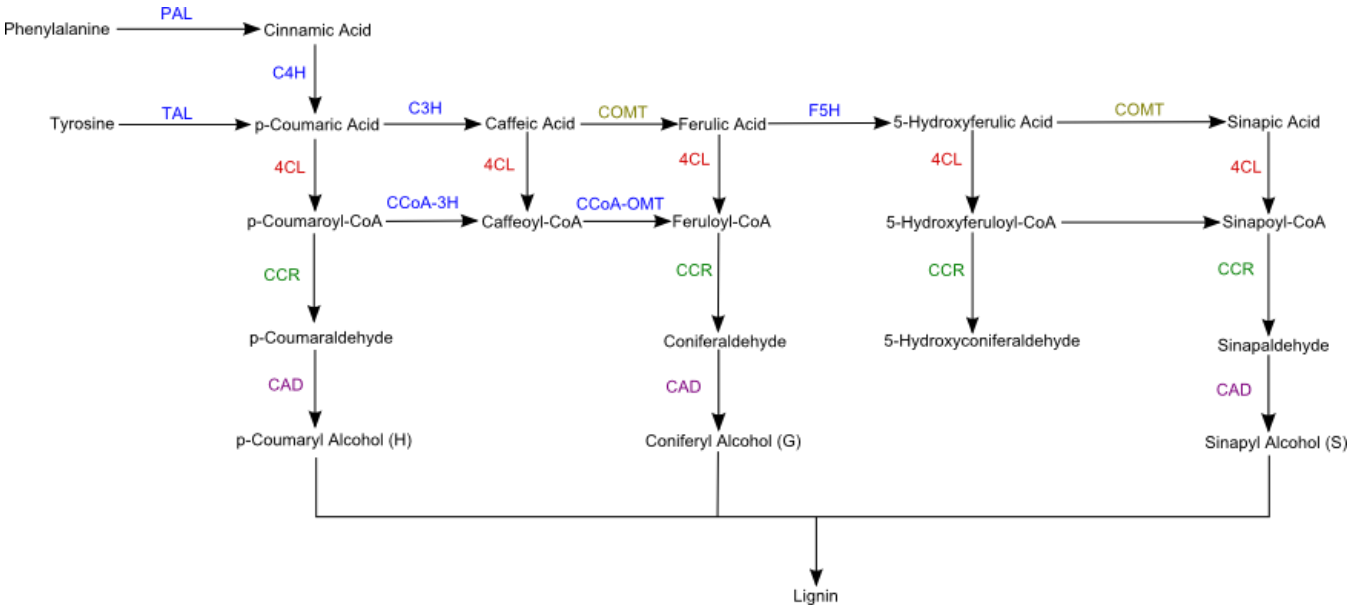

Figure 1.3. A schematic representation of the main lignin biosynthetic pathway. The enzymes that appear once in the pathway are coloured in blue, while the enzymes that appear multiple times each have their own specific colour. Pathway redrawn based on previous reviews (Vanholme et al., 2008, Vanholme et al., 2012).