Genetic architecture of gene expression during xylogenesis in *Eucalyptus* interspecific hybrids

by

LIZETTE LOUBSER

Submitted in partial fulfilment of the requirements for the degree

Magister Scientiae

In the Faculty of Natural and Agricultural Sciences

Department of Biochemistry, Genetics and Microbiology

University of Pretoria

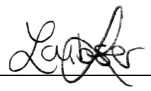
December 2019

Under the supervision of **Prof Alexander A. Myburg**

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

DECLARATION

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



Lizette Loubser

9 December 2019

TABLE OF CONTENTS

CHAPTER 2

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTARY FILES

DISSERTATION SUMMARY

Genetic architecture of gene expression during xylogenesis in *Eucalyptus* interspecific hybrids

Lizette Loubser

Supervised by **Prof Alexander A. Myburg**

Co-supervised by **Prof Eshchar Mizrachi** and **Dr Nanette Christie**

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

Department of Biochemistry, Genetics and Microbiology

University of Pretoria

Xylogenesis is a complex biological process involving thousands of genes that leads to the formation of woody biomass, which is one of the largest sources of renewable raw materials for many industrial applications, such as construction, bioenergy, and biomaterials. This process is a strong carbon sink that needs to be kept under strict regulation at the transcriptional level. Better understanding of the key regulators underlying environmentally or industrially desirable phenotypes will allow us to improve woody biomass traits for bioprocessing and biorefinery. Variation in these phenotypes is associated with many genes segregating at population level, particularly in highly outbred populations such as *Eucalyptus* interspecific hybrids. *Eucalyptus* trees, which have a large capacity to produce woody tissue with superior structural and chemical qualities and relatively short rotation, are important models for wood formation research.

In this study, we aimed to characterise the genetic architecture of gene expression during xylogenesis in an interspecific (*E. grandis* x *E. urophylla*) *Eucalyptus* F_2 backcross population and to determine

the conservation of the genetic architecture across age. This was done by analysing RNA-seq data from xylem tissues of 156 and 100 field-grown *Eucalyptus* trees at juvenile (three years) and rotation (eight years) age respectively. Transcriptome-derived SNPs were used to construct a gene-based genetic linkage map for eQTL detection using 236 high confidence markers in highly expressed genes. We identified co-expression modules for 25,267 genes, which were used to construct a co-expression network. This network allowed identification of the main biological functions of each module, and dissemination of transcriptional coordination of metabolic functions and development during xylogenesis. Global eQTL analyses of co-regulated genes led us to the identification of 22 *trans*-eQTL hotspots, which are major regulatory perturbations that can change the structure of the co-expression network. To characterise the genetic architecture of gene expression variation during xylem development, the co-expression and co-regulation results were integrated into systems genetics models, which revealed a major shift in the transcriptional regulation architecture from juvenile to mature age, evidenced by new *trans*-eQTL hotspots detected in mature trees.

This study provides a new method for rapid genetic dissection of gene expression variation from population-wide transcriptome data alone and provides insight into the regulation of xylem genes across age. The observed changes in the genetic architecture of wood formation genes, as well as those observed in genes related to abiotic stress response, suggests that there are multiple contributing factors associated with variation in transcript abundance, including developmental or age-related changes, stress-related changes and other yet unknown biological effects. By combining results from eQTL and co-expression analyses into systems genetics models, we identified a genetic basis for coordinated gene expression responses regulating biological processes in xylem. These results will enable us to analyse the genetic architecture underlying complex wood biorefinery traits and identify interacting genes and pathways. This can then be used to engineer or breed for complex wood property traits while avoiding negative effects on plant growth.

PREFACE

Woody biomass, also known as secondary xylem, is formed during a process called xylogenesis, where thousands of genes are involved in the formation of xylem cells that largely consist of thick secondary cell walls. Wood is not only one of the main sources of renewable raw materials to produce timber, pulp, paper, and biorefinery derivatives, it is also a major carbon sink inside plant cells that is important for climate regulation and requires strict transcriptional control at the tissue and organ level. The rate at which carbon is deposited into secondary cell walls requires strict coordination of xylem cell metabolism, cell wall formation, and coordination of carbon (sugar) transport via the phloem to xylem cells. *Eucalyptus* tree species and hybrids are some of the most widely planted hardwood crops in South Africa (and globally), due to their superior wood quality, wide adaptability, fast growth and quick rotation times that can be as short as five to seven years. Environmentally and industrially important wood property traits have complex genetic architectures, particularly in highly outbred populations such as *Eucalyptus* interspecific hybrids, where large numbers of genes are segregating in the population and contributing to trait variation.

Many studies have used systems biology approaches to better understand pathways related to development and stress responses and identify associated key regulators associated with them. However, such studies are often based on extreme phenotypes that are expressed as a result of multiple reactions to a major perturbation (such as gene knock-out) with an adverse effect. To fully understand complex quantitative traits, systems genetics approaches need to be used to link naturally occurring genetic polymorphisms to trait variation, which will assist in the discovery of novel strategies to manipulate phenotypic variation and provide a better understanding of the mechanisms underlying complex quantitative trait variation. Systems genetics approaches also enable population-wide modelling of interactions among thousands of genes involved in important biological pathways and their response to environmental factors. Mapping expression quantitative trait loci (eQTLs)

detects significant associations between genomic loci and the variation in molecular intermediate phenotypes, such as transcript abundance. This allows the identification of genes that are associated with quantitative traits. However, these traits are influenced by many genes and therefore integrative systems genetics models can be used to account for this. Network-based systems genetics approaches can be used to prioritise candidate genes associated with complex traits in highly outbred populations and to identify possible key regulators underlying variation in these traits.

This dissertation describes the results of a genome-wide study of the genetic architecture of gene expression during xylogenesis in a *E. grandis* x *E. urophylla* interspecific backcross population, with the future objective of prioritising candidate genes underlying complex biorefinery traits for genetic engineering. Chapter 1 is a literature review which aims to provide a comprehensive background on systems biology, systems genetics, genetical genomics, and the state-of-the-art in systems biology dissection of complex traits and high-throughput RNA-seq analysis in plant populations. Chapter 2 describes the development of integrative systems genetics models that allowed the characterisation of the genetic architecture of xylem expressed genes over time (three to eight years of growth). This approach allowed us to achieve the following objectives:

- Population-wide mapping and quantification of xylem transcriptome profiles
- Determination of the heritability of xylem transcriptome profiles in clonal replicates
- Construction of a robust genetic linkage map for the *E. grandis* x *E. urophylla* F_1 hybrid parent
- Systems genetics modelling and genetic architecture characterisation of xylem development at rotation age
- Comparison of transcript abundance and genetic architecture in juvenile (three-years-old) and mature (eight-years-old) trees

The results provide insight into the regulation of xylem genes and how this regulation changes across age and in response to abiotic stresses.

Chapter 2 has been prepared in the format of a draft research manuscript for submission to a peer-reviewed journal (e.g. *New Phytologist*). The work done for this dissertation was part of a larger project within the Forest Molecular Genetics (FMG) programme, in the Department of Biochemistry, Genetics and Microbiology (BGM) and the Forestry and Agricultural Biotechnology Institute (FABI), at the University of Pretoria from January 2018 until November 2019. The following conference outputs were gained from the M.Sc.:

Loubser L, van der Merwe K, Ployet R, Christie N, Mizrachi E, Myburg AA. Age-to-age correlation and heritability of transcriptome variation in *Eucalyptus*. South African Society for Bioinformatics (SASBi) / South African Genetics Society (SAGS) Conference, October 2018. Golden Gate Conference Centre, Clarens, South Africa (Poster).

Loubser L, Ployet R, Christie N, Mizrachi E, Myburg AA. Rapid genetic dissection of xylem gene expression variation in a *Eucalyptus* interspecific backcross population. International Union of Forest Research Organizations (IUFRO) Tree Biotechnology 2019 Meeting, June 2019. Raleigh, NC, USA (Poster).

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to the following people and institutions:

* My main supervisor, **Prof Zander Myburg**, thank you for giving me the opportunity to work on this project and for all your patience, guidance and advice that helped me get through it. Thank you for always making time in your busy schedule for meetings and feedback and for encouraging me through your positive attitude and sometimes quirky sense of humour. I am honoured to have been part of your team.

* My co-supervisor, **Dr Nanette Christie**, thank you for your support and guidance throughout this project and for all the extra time and effort you put in to help me when I was struggling or needed feedback. Thank you for our weekly Skype meetings and for always being so friendly and willing to help. This project would not have been possible without all the valuable resources you provided.

* My committee member, **Dr Raphael Ployet**, thank you for your willingness to help wherever you could throughout this project and for being a great mentor. Thank you for all the meetings that you attended, the support you provided, and for your reassurance when I was doubting myself.

* My co-supervisor, **Prof Eshchar Mizrachi**, thank you for the valuable input that you provided in our meetings and for your ideas and suggestions that showed us what was missing and helped to further improve this project.

* My committee member, **Mrs Marja O'Neill**, thank you for all your effort in arranging for the samples to be sequenced and for your assistance with analysing the SNP chip genotypes. Thank you also for being a ray of sunshine and keeping our spirits up.

* **Prof Fourie Joubert**, **Mrs Karen van der Merwe**, and **Mr Johann Swart**, thank you for your support and willingness to always help with any bioinformatics questions and issues. A special thank you to **Prof Fourie Joubert** for making the lab such a welcoming place and for all your mentorship and advice since we joined the lab as wide-eyed-and-terrified honours students.

* The **National Research Fund (NRF)** and the **Department of Science and Technology (DST)**, thank you for awarding me the DST-NRF Innovation Master's Scholarship that allowed me to pursue this MSc degree. Your investment allowed us to promote competitive, high quality research and innovation, as well as applying sophisticated, cutting-edge computational techniques throughout this project.

* The **Forest Molecular Genetics (FMG) Programme** and the **Forestry and Agricultural Biotechnology Institute (FABI)**, thank you for providing your students with wonderful opportunities to learn and share their findings and for the constant support you provide us. I feel very privileged to have been part of such an excellent group of people and will always carry the memories, skills and knowledge I have gained with me wherever I go.

* **Sappi Forest Research**, thank you for developing and maintaining the *E. grandis* x *E. urophylla* backcross population that we analysed in this project. Thank you also for the financial support that you provide to FMG that allows for the generation of valuable data to improve our understanding of wood formation.

* The various funding bodies, including the **Department of Science and Technology (DST)**, the **Technology and Human Resources for Industry Programme (THRIP)**, and the **Technology Innovation Agency (TIA)** for additional funding provided to FMG students. Your support throughout the year lifts a great weight from the shoulders of postgraduate students and it is very much appreciated.

* Thank you to my **family** for a lifetime of support and encouragement and also to my friends from the Bioinformatics lab: **Monique**, **Hannes**, **Alisa**, **Werner**, **Chris**, **Greg** and **Caryn**; from FMG: **Julia**, **Luke**, **Thandeka**, **Marja** and **Raph**; life-long pals: **Wihanli**, **Tarien**, **Melissa** and **Alessandri**; and favourite cousins: **Mel** and **Carli**, for always being supportive, helpful and entertaining. I appreciate every one of you and the light that you bring into my life.

* My fiancé, **Marcel van Staden**, thank you for your constant support, love, encouragement and reassurance. You remind me to believe in myself, to never take myself too seriously, and you have always kept me grounded. I am very lucky to have you in my life and I appreciate everything you do for me. Thank you for being you, I love you.

* My parents, **Linda and Manie Loubser**, thank you for always encouraging me to do my best and for teaching me from a very young age that I can achieve anything I set my mind to. You have always inspired me, and my goal will forever be to make you proud of everything I do. Thank you for your love and prayers that have carried me through life. I love you both very much.

* Finally, I would like to thank **God** for the many blessings and opportunities He has given me in life. Without His unconditional love and grace, I would not be where I am today. All honour and glory to Him.

Happy reading!

CHAPTER 1

LITERATURE REVIEW

Systems Genetics and the Genetic Architecture of Complex Quantitative Traits

1.1 Introduction

One of the longest-standing questions that geneticists have struggled with is understanding how phenotypic variation is affected by variation in the genotype (Boyle *et al.*, 2018). Quantitative genetic traits are observed as different phenotypes across individuals in a population and can be controlled by environmental influences or inherited through the genotype. However, these traits are typically determined by a combination of both factors, known as genotype-environment (GxE) interactions (Falconer & Mackay, 1996). These quantitative characteristics can either be monogenic or complex. A monogenic trait is controlled by variation in a single gene, whereas complex traits are controlled by the variation in multiple genes and their interaction with environmental elements. Complex traits are known to have non-Mendelian patterns of inheritance that are not readily predictable and usually display many different phenotypes (Colbert *et al.*, 2011). It is important for us to study the genetic architecture of complex traits as it enables the use of marker-assisted selection for breeding, but the mapping of polymorphisms that regulate the variation in these traits presents a key challenge in biology (Mackay, 2001; Thavamanikumar *et al.*, 2013).

Most quantitative traits display some degree of heritability and tend to have lower heritabilities than qualitative traits, as they can be greatly influenced by the environment. Heritability is a statistical estimation of the amount of phenotypic variation observed between individuals within a population due to variation in their genotypes. It can be divided into two categories, namely narrow-sense heritability (h^2) and broad-sense heritability (H^2), depending on the type of genetic variation in question (Falconer & Mackay, 1996). Narrow-sense heritability is a parameter generally used for predicting responses to selection and correlation between relatives, as it is concerned with genotypic variation that is caused by purely additive genetic factors (Visscher *et al.*, 2008). Broad-sense heritability, also known as the repeatability, refers to the amount of phenotypic variation explained by the total genetic variance, which includes additive genetic factors and genetic variance. This genetic variance can be due to dominance, where one allele masks the phenotypic effect of the other,

and epistasis, where one gene masks the effect of another (Kruijer *et al.*, 2014). Knowing the heritability of a trait is valuable, as traits with greater heritabilities can be modified more easily by selection and breeding (Xu *et al.*, 2017).

The genetic architecture of complex quantitative traits refers to the underlying genetic basis and variation observed within these traits (Hansen, 2006). This can become quite complicated, as the observed variation in a phenotype is essentially due to alleles segregating at several loci (Mackay, 2001). The variation in gene expression affecting these traits may be regulated by several factors, such as quantitative trait loci (QTLs), varying strengths of the effects of different loci, random distribution of genes, pleiotropic effects, and gene interactions. Genes can either interact with environmental elements or with other genes, resulting in additive, dominant, or epistatic effects (Wu *et al.*, 2007). Variance in gene expression due to epistatic effects results from the interactions between genes at different loci that all have an effect on a complex quantitative trait, but this variance is generally only significant at high levels of heterozygosity (Mäki-Tanila & Hill, 2014). Genetical genomics can help us to better understand variation in gene expression levels and how it affects complex traits.

Genetical genomics is a concept that was first proposed by Jansen and Nap in 2001, where transcriptome mapping is performed to analyse gene expression across an entire genome. The identification and mapping of gene expression QTLs (eQTLs), which control the level of variation in a transcriptome, allows us to connect the variation in genotypes to phenotypic variation in a segregating population. The regulatory relationships between an eQTL and its respective structural genes can be divided into two categories: (i) *cis*-eQTLs that map within the same chromosome as the gene itself and (ii) *trans*-eQTLs that map elsewhere in the genome (Jansen & Nap, 2001). One of the most important elements of genetical genomics is eQTL hotspots, where many genes that map to one locus are influenced by a single polymorphism that leads to large biological effects. Identifying these

hotspots allows for the construction of gene regulatory networks through systems biology approaches (Breitling *et al.*, 2008).

Systems biology is a holistic approach to understand the complex interactions within biological systems. This can be done by disturbing a system, studying its responses, and integrating the data to form models which can describe the system and its responses to variation caused by mutations. These models are also able to predict how a system will change over time and under varying conditions. Systems biology is an interdisciplinary field with a focus on complex, scale-free biological networks of molecular and physical interactions (Civelek & Lusis, 2014). An important component of systems biology is systems genetics, which aims to understand complex interactions and the flow of biological information underlying complex traits by quantifying intermediate phenotypes. These intermediate phenotypes can be transcript levels, interactions across multiple biological scales, or even large biological networks (Civelek & Lusis, 2014). Systems genetics ultimately links genes and gene networks to specific traits and integrates systems biology methods with genetics methods to connect genotypes and phenotypes in complex traits at a population level (Nadeau & Dudley, 2011).

The process of wood formation is a complex biological system that is important for us to model and understand. Apart from obvious applications in pulp and paper production, wood is also an economically important renewable energy source and provides energy-efficient building materials as environmentally cost-effective alternatives (Plomion *et al.*, 2001). Hardwood species, such as *Eucalyptus* trees, are important models for wood formation research as they have a large capacity to produce woody tissue. The formation of wood, or secondary xylem largely composed of fibres and vessels with thick secondary cell walls, is a continuous dynamic process that is regulated by internal and external factors and involves thousands of genes (Zhang *et al.*, 2014; Mizrachi and Myburg, 2016). There are five key steps in wood formation: (i) division of cells; (ii) expansion of cells; (iii) thickening of secondary cell walls; (iv) cell wall lignification; and (v) programmed cell death which

leads to the formation of an empty tube with secondary cell walls (Plomion *et al.*, 2001; Demura & Fukuda, 2007).

This review aims to provide a comprehensive background on systems biology, genetical genomics and systems genetics. It also aims to discuss the state-of-the-art in systems biology dissection of complex traits and high-throughput RNA-seq analysis in plant populations. We want to gain an understanding of the genetic variations underlying complex phenotypic variations for wood growth and development and determine how genetic variation is expressed via its effect on systems components. This will ultimately allow us to determine the phenotypic variation in segregating plant populations. For more in-depth reviews, please refer to Mackay (2001), Hansen, Halkier and Kliebenstein (2008), Visscher and Goddard (2010), Civelek and Lusis (2014), Feltus (2014), and Goddard *et al.* (2016).

1.2 Genetic Dissection of Quantitative Traits in Plants

1.2.1 Genetical genomics

Genetical genomics is a strategy first described by Jansen and Nap (2001) to analyse the expression profile data of an entire genome in combination with the genetic variation observed between individuals in a segregating population (**Figure 1.1**). It specifically aims to identify QTLs for gene expression data, referred to as expression QTLs (eQTLs), and map these at a global level (Breitling *et al.*, 2008; Zhu *et al.*, 2009). QTL analyses describe the natural variation in a population by searching for regions in the genome where genetic variation and phenotypic variation are correlated, and mapping QTLs ultimately allows the identification of genes responsible for the variation in a trait. eQTLs, which result from segregating polymorphisms that affect the level of transcript variation, such as single nucleotide polymorphisms (SNPs), can be used to detect correlations (genomic

overlaps) between the polymorphisms and QTLs. Such overlaps can be used to identify candidate genes that are responsible for the observed phenotypic variation in a trait (Hansen *et al.*, 2008).

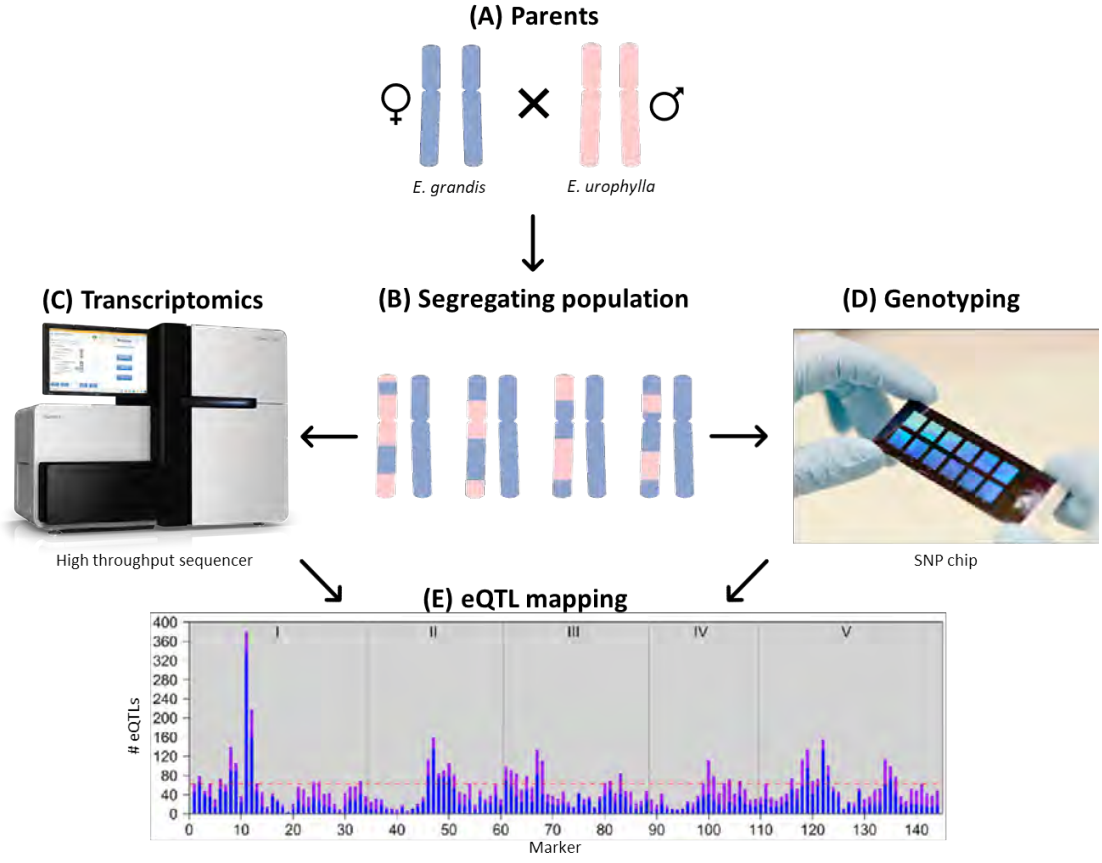


Figure 1.1: Genetical genomics. Combining the expression profiles of individuals in a segregating population with the analysis of their molecular markers allows us to identify candidate genes associated with complex trait phenotypes using QTL analyses. **(A)** Two parents from different ecospecies are crossed to produce an F_1 progeny. **(B)** The progeny can self-cross or undergo backcrossing to produce a segregating population. Individuals in the segregating population are analysed through **(C)** expression profiling, using transcriptomics techniques, and **(D)** molecular marker analysis, e.g. SNP genotyping. **(E)** eQTL mapping uses genotype and expression data to identify significant associations between SNPs and molecular traits (Jansen & Nap, 2001).

eQTLs arise due to the structural variation of DNA which can be caused by several events, including sequence variation such as SNPs and small indels, rearrangements in the genome due to insertions, deletions and translocations, copy number variation, differences in the stability of mRNAs and allelic variants of transcription factors. eQTLs can be divided into two classes, namely *cis*- and *trans*-eQTLs, based on the location of the polymorphism responsible for the variation relative to the gene being expressed (Wolen & Miles, 2012). An example of each class is illustrated in **Figure 1.2**. A *cis*-eQTL regulates a gene if the gene's expression level is associated with a nearby polymorphism located on

the same chromosome and this class generally makes up about 30-50% of the total eQTLs identified in a population (Joosen *et al.*, 2009). Candidate genes associated with the variation in trait phenotypes can be prioritised when they have a *cis*-eQTL at the same genomic location as a QTL for the phenotypic trait. This is because the observed variation in the phenotypic trait is linked to a specific locus, where a *cis*-acting polymorphism causes the associated gene to be produced in variable quantities (Wolen & Miles, 2012).

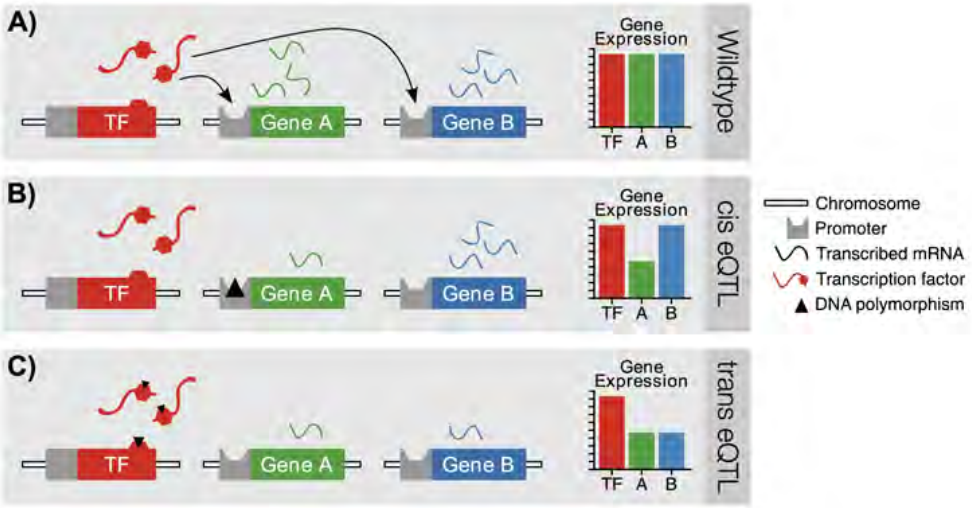


Figure 1.2: **Classification of eQTLs.** eQTLs are classified based on the position of a polymorphism (black triangle) relative to the gene being expressed. **A)** In the wildtype, a transcription factor (TF) (red) binds to the promoter regions (grey) of gene A (green) and gene B (blue) and activates their transcription. The bar graph indicates that all the genes are fully transcribed. **B)** A polymorphism is present in the promoter region of gene A, preventing the TF from binding. This results in a reduced transcription rate for gene A, whereas gene B remains fully transcribed. Gene A is therefore under regulation of a *cis*-eQTL. **C)** A polymorphism is present in the DNA binding region of the TF, preventing the TF from binding to any downstream gene promoter regions. Both gene A and gene B have reduced transcription rates and all downstream genes under regulation of the mutated TF will be linked to a *trans*-eQTL at the site of this polymorphism. Image from Wolen and Miles (2012).

In contrast to *cis*-eQTLs, a *trans*-eQTL regulates a gene if the expression level is associated with a polymorphism located elsewhere on the genome, far away from the gene itself (Wolen & Miles, 2012). A *cis*-eQTL is essentially only associated with the gene in which it occurs, whereas multiple *trans*-eQTLs can arise throughout the whole genome from only one causal gene. Key-regulators, such as transcription factors, can explain the identification of *trans*-eQTL hotspots throughout the genome as they have major effects on the expression of downstream genes. This can play an important role in

disease susceptibility, for example, as conventional breeding strategies can select for or against segregating *trans*-eQTLs (Kadarmideen *et al.*, 2006). However, QTL studies have indicated only a few examples where hotspots with significant effects on the biological system were detected, suggesting that most of the variation in the genotype is buffered against in the phenotype (Joosen *et al.*, 2009).

The application of genetical genomics approaches to identify eQTLs allows us to study gene expression and gain insight into the basis of complex traits. In a study done by Drost *et al*. (2010), genetical genomics was used to identify transcriptional networks in three different tissues (xylem, leaf, and root) of an interspecific hybrid population of *Populus*. The authors detected pleiotropic hotspots, where one gene affects multiple phenotypic traits, and used these to construct co-expression networks that showed significant enrichment for genes in gene ontology (GO) functional categories and regulatory elements for transcription. When the transcriptional networks of the three tissues were compared, the authors found that the topology was commonly conserved, but that the transcriptional networks were regulated by different loci in each tissue. The authors also found that where shared eQTLs (i.e. eQTLs with peaks at the same locus) were identified in two organs, less than a third of the genes were regulated by the same locus in both organs. From this study, it appears that the genetic architecture of gene expression is significantly different between the different tissues.

1.2.2 eQTL studies in plants and trees

eQTLs are genomic regions that contribute to the variation in gene expression levels between individuals in a population. Candidate genes underlying specific phenotypic traits can be identified through correlation analyses of eQTLs with their respective trait QTLs and the construction of gene regulatory networks allows us to elucidate the basis of variation in phenotypic traits. The variation in these traits are caused by genetic polymorphisms (e.g. SNPs) and result from qualitative and quantitative differences in gene expression. In a population of plants, one can measure transcript

levels and treat the variation in each gene as a heritable trait that can be analysed through genetical genomics approaches, however, these analyses rely on natural genetic variation in the population being studied. The use of eQTL studies in plants also allow the analysis of candidate genes underlying quantitative traits without having to go through the time-consuming process of positional cloning (Druka *et al.*, 2010).

One important aspect of eQTL analysis in plant populations is experimental design, which can be divided into three major categories: (i) the type of population under study; (ii) the size of the population; and (iii) how replication and randomization is organized. Typically, when plant populations are studied, F_1 hybrids or F_2 backcross hybrids are used with a population size of 200 or more lines (although many studies using 100 lines have been successful) (Schön *et al.*, 2004; West *et al.*, 2007; Kadarmideen, 2008). eQTL studies need to have a suitable measure of replicate error to prove that genetic variation exists. Replicates of individual genotypes are also necessary as they allow the prediction of true within line variation. It is important to consider limited pleiotropy in plants, as it may play a significant role in adaptive traits such as quantitative age-related resistance. It may also indicate the presence of transcription factors that influence gene expression in a stage- or tissue-specific manner (Druka *et al.*, 2010).

eQTL studies have been performed on many different plant species such as wheat (Jordan *et al.*, 2007), rice (Wang *et al.*, 2010, 2014), maize (Shi *et al.*, 2007; Swanson-Wagner *et al.*, 2009; Holloway *et al.*, 2011; Li *et al.*, 2013; Christie *et al.*, 2017), potato (Kloosterman *et al.*, 2012), and *Arabidopsis* (Keurentjes *et al.*, 2007; Wentzell *et al.*, 2007; West *et al.*, 2007). Several eQTL studies have also been done on hardwood species (Kirst *et al.*, 2004, 2005; Drost *et al.*, 2010; Mähler *et al.*, 2017; Mizrachi *et al.*, 2017; Zhang *et al.*, 2018), to identify genetic factors that are involved in complex traits (e.g. growth) and to determine how they interact in networks that predict the response of a system to changes in the genetic structure. In addition to studying changes in the genetic

architecture of gene expression in *Populus* during organ differentiation, Drost *et al.* (2015) also applied genetical genomics methods to identify a candidate gene and genetic elements that are responsible for regulation of variation in leaf morphology. In *Eucalyptus* species, genetical genomics studies have generally focused on how growth and lignin are regulated, identifying candidate genes involved in variation of wood traits and gene expression, and analysis of the genetic architecture of mRNA abundance in developing xylem tissue (Kirst *et al.*, 2004; Kullan *et al.*, 2012; Mizrachi *et al.*, 2017).

1.2.3 Systems genetics and the genetic architecture of complex quantitative traits

Systems genetics can be defined as the global analysis of molecular elements that are associated with the variation of physiological and phenotypic traits between all individuals within a population. Systems genetics analyses do not only study interactions between genes and their interaction with the environment, but also focus on intermediate phenotypic traits affected by variations in the DNA, such as levels of gene expression, proteins, and metabolites. Different systems genetics approaches are illustrated in **Figure 1.3**. One of the main benefits of using a systems genetics approach is the ability to study interactions on a molecular level by analysing multiple genetic perturbations, which represents natural populations under normal selective pressures and is most pertinent to the specific trait under study. It also allows one to analyse the genetic architecture of complex quantitative traits to identify the different properties of individual genes associated with variation in these traits (Civelek & Lusis, 2014).

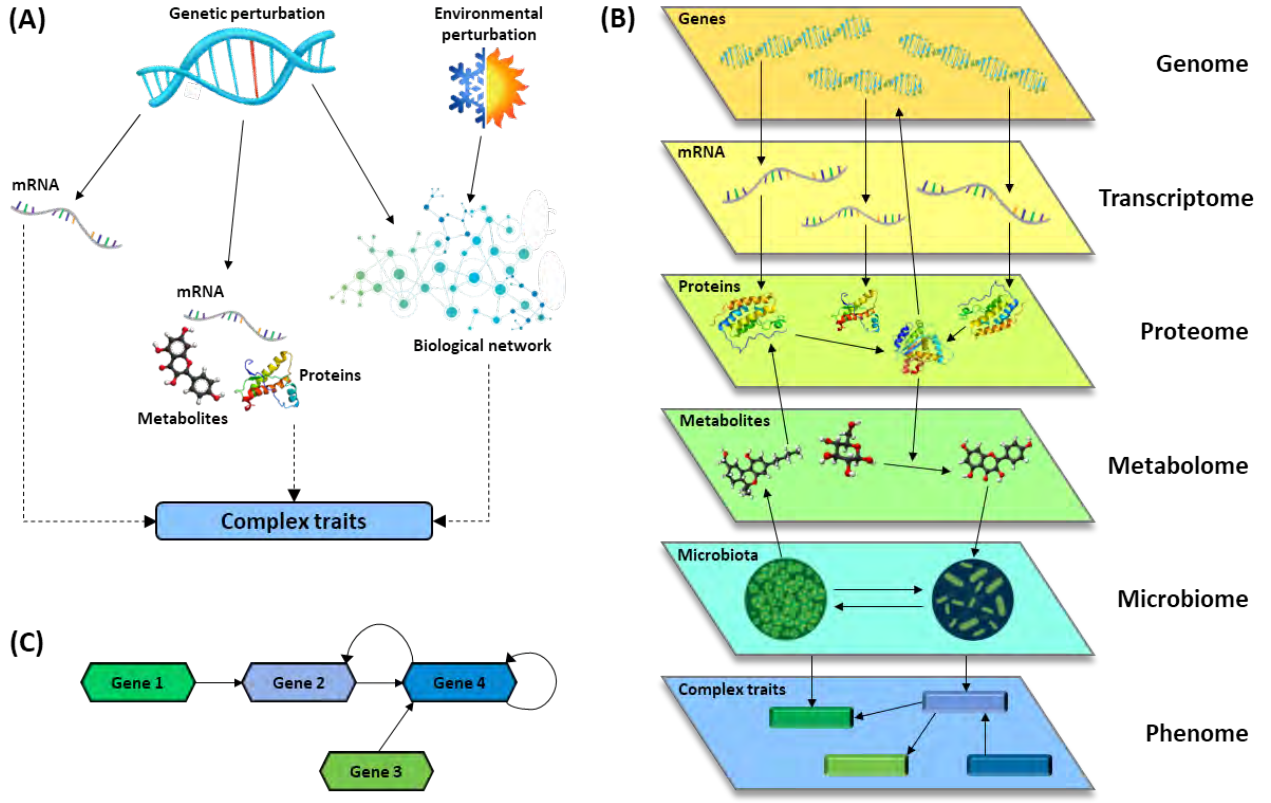


Figure 1.3: Systems genetics approaches. Systems genetics studies can have simple or more complex designs based on the type of intermediate phenotype under study. **(A)** An example of a simple design would involve quantifying a single type of intermediate phenotype within a population and integrating it with complex traits through correlation analyses or mapping to chromosomes. In more complex designs, many intermediate phenotypes can be studied, which enables the investigation of interactions across different biological spaces. These phenotypes can also be used to model biological networks that are affected by environmental factors. **(B)** Intermediate phenotypes across different biological scales can interact with each other (indicated by arrows). These interactions can be used to construct a map based on natural variation within a population. **(C)** Biological networks can be modelled from trait correlations between individuals within a population. In this example, a directional expression network is constructed between four genes on the basis of their natural variation. Image adapted from Civelek and Lusis (2014).

In order to have a complete understanding of the genetic architecture underlying complex quantitative traits, it is important to gain information on several factors: (i) the identities and total number of genes involved in all biological processes associated with the trait; (ii) the mutation rate at each locus containing these genes; (iii) the identities and total number of loci associated with the variation of the trait across populations, as well as within a single population, and across different species; (iv) how the trait is affected by any new mutations or segregating alleles; (v) any effects caused by epistatic interactions; (vi) any pleiotropic effects; (vii) all polymorphisms that define QTL alleles; (viii) QTL allele frequencies; and (ix) the mechanism underlying changes in the phenotype of the specific trait.

There is currently no quantitative trait that has been described at such a high resolution yet, therefore the combination of many different fields of genetics is required to work towards this (Mackay, 2001).

Determining the genetic architecture underlying complex quantitative traits is a key challenge of systems genetics. This is because the variation observed between phenotypes is caused by interactions between many alleles that are sensitive to changes in the environment. Many studies have implemented systems genetics approaches to address the prevailing question of the underlying genetic basis of gene interactions. Ayroles *et al.* (2009) quantified the phenotypes and genome-wide transcript abundance of six ecologically important complex traits in *Drosophila melanogaster*. The study allowed them to predict genetic networks and gene functions, as well as to identify multiple candidate genes associated with variation in stress responses, life span, and behaviour. Park *et al.* (2011) integrated fear phenotypes, genotype information and transcriptome data from hippocampus and striatum tissue in mice to gain insight into the underlying genetic basis of memory and learning. The study enabled them to prioritise key markers and genes associated with fear phenotypes and to gain an increased understanding of genetic networks underlying behaviour.

1.2.4 Co-expression network topology and the genetic architecture of gene expression

Biological networks are scale-free networks that are buffered against random mutations and can be analysed to determine the relationship between the genetic architecture of gene expression and network topology. Mähler *et al*. (2017) used gene expression data for the construction of a co-expression network to determine this relationship and relate it to signatures of selection. The authors performed eQTL mapping on a natural population of *Populus tremula* and identified thousands of significant eQTLs associated with unique genes and SNPs. The study showed that these unique genes were abundant in the network periphery, but underrepresented in module cores, and that the effect sizes of eQTLs had a negative correlation with network connectivity (i.e. the correlation of each gene

with all other genes in the network). The authors also found that connectivity was linked to signatures of selection, where it suggested that the genetic architecture of natural variation in gene expression is regulated by purifying selection and that connectivity within the network is associated with the strength of this selection.

There are several factors that can influence the functional connectivity and topology of co-expression networks. Ballouz, Verleyen and Gillis (2015) examined RNA-seq co-expression data generated from 1,970 samples with a Guilt-By-Association framework, where genes were evaluated to determine their tendency to reflect shared function through co-expression. The authors found two important factors that can have significant effects on functional connectivity, namely the number of samples and the read depth. Larger sample sizes and increased read depths allow networks to perform better and greatly increase the functional connectivity. The authors also saw that the functional connectivity of the network increased when multiple networks were combined and that there was a decrease in the amount of expression variation noise. The functional connectivity was not influenced by different machine learning methods, however, there was an effect on network topology. And finally, the authors determined that the network topology is affected by the type of data used (microarray vs. RNA-seq), due to changes that occur in the correlation of expression variation noise in the different technologies.

In a study done by Fagny *et al*. (2017), eQTL analyses were used to construct networks that illustrate the relationships between gene expression levels and genetic variants in 13 human tissues. The authors found that three elements of the topologies of these networks inform regulatory function at tissue-level; (i) communities, which are highly modular groups of genes and SNPs, are enriched for genes with related functions, as well as for regulatory pathways and SNPs located in tissue-specific active chromatin regions; (ii) community hubs, also known as core SNPs which have a high connection to genes that are in the same community, are enriched for active chromatin regions located

close to transcriptional start sites; and (iii) global hubs, which are linked to multiple genes in the network, are enriched for distal elements. This study used eQTL analyses to produce complex networks of relationships that represent the polygenic architecture underlying different tissues, which aids in the understanding of the effects of genetic variants at a tissue-specific level.

1.2.5 Systems genetics of wood formation in forest trees

Wood formation is a complex process involved in the growth and development of woody plants. To improve wood properties of trees with industrially important phenotypes, this process can be modelled as a biological system to identify and better understand the roles of its key transcriptional and metabolic regulators. The biosynthesis of xylem tissue, which is responsible for water and nutrient transport in plants, is known as xylogenesis and is an extremely strong carbon sink under strong transcriptional and metabolic control. The regulation of these processes ultimately governs the development and properties of wood, which is of economical, ecological, and evolutionary importance (Mizrachi and Myburg, 2016).

Modelling wood formation as a biological system involves the integration of several sets of information (**Figure 1.4**). Firstly, trees in a population can be genotyped to outline a genetic map and DNA markers can be used to tag the genetic variation between individuals in a population. This is done based on the level of linkage disequilibrium (LD), which defines the non-random statistical associations of alleles at different loci (Hansen, 2006). By determining metabolite levels and obtaining quantitative gene expression profiles through the profiling of transcriptomes, QTLs can be mapped and QTL- and co-expression networks can be constructed. *A priori* information, such as known regulatory relationships and targets for transcription factors, can be included along with wood properties and growth to infer directionality of interactions in networks. Integration of all this information allows for the construction of an effective systems model of wood formation (Mizrachi and Myburg, 2016).

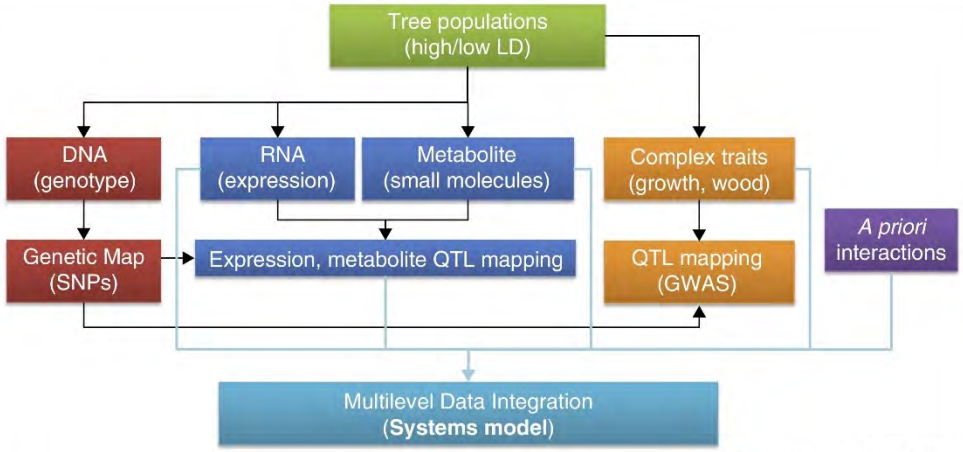


Figure 1.4: Systems genetics of wood formation. Incorporating several sets of information into a systems model allows the construction of a working model of wood formation. This information includes the genotype, levels of gene expression and metabolites, and determining the genetic signal for their variation through mapping of eQTLs and metabolite QTLs. Prior knowledge of interactions and relationships of component data, along with wood property traits can be integrated to construct a systems model. Image from Mizrachi and Myburg (2016).

Eucalyptus is an important renewable feedstock for manufacturing paper, pulp, timber, and more recently, biorefinery. It is one of the most widely planted hardwood species for forestry in South Africa and has superior wood qualities to most other hardwood species (Mizrachi *et al.*, 2017). In 2014, Myburg *et al.* published the complete genome assembly of *Eucalyptus grandis,* which has provided a powerful tool for modelling wood formation and gaining insights into the complex interactions between genotypic and phenotypic variation in *Eucalyptus*. The *Eucalyptus* genome integrative explorer (EucGenIE) was developed as a resource for genomics and transcriptomics studies in *Eucalyptus* (Hefer *et al.*, 2011). As part of this resource, qtlXplorer was developed as a tool for querying and visualising systems genetics data using an interactive version of Circos (Christie *et al.*, in preparation; Krzywinski *et al.*, 2009). This provides a valuable resource to show how the integration of systems genetics with functional genomics will allow us to extract meaningful biological information from available data.

Mizrachi *et al.* (2017) applied a network-based data integration (NBDI) method to developing xylem tissues from a segregating *Eucalyptus* hybrid population to gain a systems-level understanding of the

genes underlying important wood property traits, as well as the biological processes and pathways associated with these traits. By using an NBDI approach, the authors were able to prioritise genes and pathways associated with phenotypic traits by combining genotype data, gene expression data, and prior network information. The authors found that the phenotypic variation observed in these traits could be explained by the genetic variation segregating in the population, as this affected genes and pathways underlying wood property traits. Wang *et al.* (2018) performed a multi-omics integrative study on lignin formation in *Populus* to improve wood properties. The analysis determined the effect of gene expression changes on several phenotypes, such as protein abundance and wood property traits. Predictions could then be made on improvements in these traits through genetic engineering. This multi-omics approach also provides a strategy for analysing other biological pathways to better understand the processes of growth, metabolism, and adaptation in plants.

1.2.6 Machine learning in systems biology

Machine learning is a branch of artificial intelligence which aims to develop and apply computer algorithms that can learn and improve with subsequent runs, using training models. Machine learning algorithms are robust tools for classifying large data sets from complex systems and some of the most popular uses for them include speech, text and image recognition (Mohri, 2012; Kim & Kim, 2018). There are generally three stages in a machine learning process: (i) an algorithm is developed; (ii) a training data set is provided, which consists of known (labelled) and unknown (unlabelled) data, after which the labelled data are processed and stored in a model; and (iii) subsequent unlabelled data are provided for which the model then predicts labels. Machine learning algorithms can be supervised, unsupervised, or semi-supervised. Supervised learning methods make use of algorithms that are trained by labelled data to predict labels of unknown data, whereas unsupervised learning methods do not require labelled data and the aim is to find structure in the data.

One of the key objectives of systems biology is to gain a systems-level understanding through computational biology. Computational systems biology includes several elements of bioinformatics and is concerned with the identification of patterns in large collections of data, subsequent formation of hypotheses, and performing simulation-based analyses to test these hypotheses *in silico* and predict system dynamics (see review by Kitano, 2002). Four key elements provide insight into this systems-level understanding: (i) structure of the system, such as gene interaction networks; (ii) dynamics of the system, indicating how the system reacts under varying conditions; (iii) method of control, where mechanisms that control the cell state can be targeted; and (iv) method of design, where simulations can be used to plan strategies for modifying the system (Kitano, 2007). Machine learning approaches have been applied in genetics and genomics studies for a wide variety of uses, such as identifying locations of transcription start sites (Ohler *et al.*, 2002), predicting genomic susceptibility to cancer (Kim & Kim, 2018), predicting promoter regions (Oubounyt *et al.*, 2019), and predicting disease risks from SNPs for precision medicine (Sik *et al.*, 2019). Expression data from RNA-seq, can also be used as input by machine learning algorithms to differentiate between different phenotypes and to detect possible important biomarkers, such as those indicating diseases (Sun & Markey, 2011). These algorithms are also largely used for assigning functional annotations to genes, mostly in the form of GO terms (Ashburner *et al.*, 2000; Libbrecht & Noble, 2015).

As reviewed by Libbrecht and Noble (2015), machine learning algorithms are frequently used in systems genetics approaches to predict shared functional relationships between genes to construct co-functional networks, where nodes represent genes and edges represent shared functions. These algorithms can also be used to train network models which can model gene expression across the entire genome, allowing us to better understand the underlying biological mechanisms of gene expression. The prediction of genes and proteins that are essential to the survival of an organism is another important application of machine learning and integration of network topology properties, as attributes for algorithm training greatly improves the prediction of these genes (Zhang *et al.*, 2016).

Machine learning methods also allows us to identify candidate genes and predict the functions of co-expression networks for further downstream analyses (Lee *et al.*, 2009).

1.3 Bioinformatics Strategy Towards Systems Genetics Analysis

1.3.1 Transcriptome profiling using next-generation sequencing

RNA sequencing

RNA sequencing is an approach that makes use of deep-sequencing technologies to profile the transcriptome. Compared to older microarray methods, this profiling technique has two new features that are important for eQTL studies: firstly, it provides information on allele-specific expression (ASE), where hybrids express one parental allele over the other, and secondly, it produces exceptionally rich data with high coverage that enables the study of RNA-isoform expression. In addition, when comparing co-expression networks from microarray and RNA-seq data, major differences are observed in their topology where there is little overlap from each network between hub-like (highly connected) genes. These differences are caused by changes that occur in the correlation of expression noise in the different technologies (Ballouz *et al.*, 2015). To generate raw RNA-seq reads, mRNA is purified and extracted where it is either first fragmented and then reverse transcribed to cDNA, or first reverse transcribed and then the cDNA is fragmented. Finally, short double-stranded cDNA is produced and adapters are ligated for next-generation sequencing (Sun & Hu, 2013). Because current protocols can often not precisely quantify samples containing a large amount of degraded RNA, Miller *et al.* (2017) developed a protocol known as complete transcriptome RNA-seq that allows for the collection of qualitative data and aims to produce quantitative stranded data for the whole transcriptome.

Tag-seq, also known as Digital Gene Expression (DGE-seq), is a tag-based approach that makes use of deep-sequencing methods to profile gene expression. To generate tag-seq data, mRNA is attached

to beads via poly(A) tails, cDNA synthesis occurs on the beads, double-stranded cDNA is digested with a frequent cutting restriction enzyme, and the residual 3' fragments are ligated to the 5' end adapter that has a binding site for the tagging enzyme. The cDNA is cleaved by the tagging enzyme and produces a 21 bp tag that is ligated to an adapter at the 3' end before PCR amplification and sequencing of the cDNA. A popular tag-based method is 3' end sequencing, which detects transcripts based on either the differences in their 3'-terminal exon or the length of their 3' untranslated region. This approach simplifies data processing and allows for the detection of rare transcripts, but its reduced library complexity raises the issue of PCR duplicates that distort gene expression levels (De Klerk *et al.*, 2014).

Gene expression can also be measured at the single-cell level to account for the loss of co-expression patterns between genes and the presence of different cell types when cells are bulked. This technique can be challenging, as single cells need to be isolated and the methods for cDNA library preparation are very sensitive with low RNA inputs (Hrdlickova *et al.*, 2017). State-of-the-art technologies, such as MR-seq that measures a single cell repeatedly, are able to characterize transcriptomes inside single cells and can measure each cell's transcriptome repeatedly to statistically assess the technical variation, allowing the identification of differentially expressed genes between just two single cells (Yang *et al.*, 2017). The multiplex sequencing technique uniquely tags samples from each cell with short identifying sequences called barcodes and are then sequenced together in a single lane. Before analysis can begin, the sequences are sorted by their barcodes so that the transcriptome of each individual cell can be assembled separately. This allows for large numbers of samples to be sequenced together, making complete genome studies much more affordable (Wong *et al.*, 2013).

Mapping algorithms

Once RNA-seq reads have been obtained and processed they need to be mapped to the reference genome with an alignment software to determine where the genes are located. There are two possible

algorithms available for aligning RNA-seq reads to the reference, namely "spliced" or "unspliced" aligners. Unspliced aligners allow reads to map to a reference without large gaps and are ideal for quantification purposes, but do not allow for the identification of novel exons or splice junctions (Garber *et al.*, 2011). Two popular software currently available for unspliced alignment (alignment of continuous reads) are Bowtie2 (Langmead & Salzberg, 2012) and the Burrows-Wheeler Aligner (BWA) (Li & Durbin, 2009). Both make use of a Burrows-Wheeler transform method to create an index of the genome, while others are based on Needleman-Wunsch or Smith-Waterman algorithms that are more sensitive but much slower, as reviewed by Porter, Berkhahn and Zhang (2015).

Unbiased or spliced aligners typically make use of a Smith-Waterman algorithm to determine the exact spliced alignment for each read, allowing for the identification of novel exons and splice junctions. Examples of available spliced aligners include GSNAP (Wu & Nacu, 2010), STAR (Dobin *et al.*, 2013), HISAT (Kim *et al.*, 2015) and TopHat (Trapnell *et al.*, 2009). HISAT uses a global index, similar to suffix arrays, as well as multiple small indexes to enable the effective alignment of reads that span multiple exons, whereas TopHat first aligns unspliced reads with Bowtie and then identifies splice junctions from the initial unmapped reads. GSNAP significantly reduces mapping bias, where non-reference alleles do not match to the reference sequence, and is useful for ASE studies, digital gene expression, and genotyping of SNPs and indels. STAR uses uncompressed suffix arrays for the alignment of full-length, long and short RNA-seq reads and can detect canonical junctions, non-canonical splices, and chimeric transcripts.

RNA-seq reads that contain non-reference alleles have a tendency of not mapping to the reference genome sequence, which means that gene quantification results may be biased and there is a possibility of false positive eQTL correlations arising. To assess the effect of allelic mapping biases on the discovery of eQTLs, Panousis *et al.* (2014) simulated RNA-seq read alignment with BWA over common variants and analysed the mapping bias rate for reference vs. non-reference reads. The

authors found that removing RNA-seq reads that were possibly biased had little effect on the discovery of eQTLs and gene quantifications, with only a few potential false positive eQTLs being detected. The results suggest that the detection of eQTLs and quantification of RNA-seq data is typically not affected by mapping bias. However, when working with RNA-seq and eQTL data this mapping bias could possibly imitate a biological signal indicating an association between the genotype and gene expression. It is therefore purely good practise to use an unbiased aligner when studying eQTL data.

Transcript quantification

RNA-seq technologies produce millions of short reads that provide digital counts when mapped to the reference genome and therefore allows the quantification of transcripts. Inferring transcript abundance is a necessary step when performing quantitative analyses, such as eQTL detection, as the observed read counts only offer an average of all the sequences expressed at a given locus (Bohnert *et al.*, 2009). A few examples of popular software available for transcript quantification include Cufflinks (Trapnell *et al.*, 2010), Kallisto (Bray *et al.*, 2016), Salmon (Patro *et al.*, 2017) and StringTie (Pertea *et al.*, 2015). To avoid alignment of bases to the reference, Kallisto implements the concept of "pseudoalignments", which produces a list of transcripts compatible for each read. These pseudoalignments are essentially the relationship between each read and a set of compatible transcripts, generated through the use of *k*-mers and de Bruijn graphs. Salmon produces "quasi-alignments" produced by an ultra-fast internal aligner and makes use of *k*-mer based counting to directly quantify transcripts. It is the first software for whole-transcriptome quantification that is able to correct for GC-content bias in fragments, which increases the accuracy of transcript abundance estimates (Bray *et al.*, 2016; Patro *et al.*, 2017; Zhang *et al.*, 2017).

A comparative study done by Zhang *et al.* (2017) showed that recently-developed software implementing alignment-free methods, such as Kallisto and Salmon, generally had decreased

computation times with relatively equal or even higher accuracies compared to those using alignment-dependent methods, such as Cufflinks. For gene-level quantification, which is generally more accurate than transcript-level quantification, other software that was not included in the comparative study (e.g. StringTie) can be considered. StringTie provides the option of applying two approaches: (i) transcriptome assemblies are built from genome-guided alignments; and (ii) *de novo* transcriptome assemblies are used to reconstruct transcripts. Unlike other software, StringTie assembles transcripts and estimates their expression levels simultaneously. The software was compared to other leading transcript assembly software at that time, such as Cufflinks, and was shown to produce a more complete assembly at faster computation times (Pertea *et al.*, 2015; Salzberg, 2016).

1.3.2 Genetic markers and eQTL mapping

In past studies, microarrays were used to measure the expression levels of genes across segregating populations to conduct association analyses across the entire genome, referred to as eQTL mapping. It is a statistical method that identifies possible genomic regions responsible for regulating transcript expression by correlating polymorphisms segregating in a population with quantitative measurements of mRNA expression. Individuals in the population are genotyped across a panel of genetic markers within a genetic linkage map, which represent differences between individual species, but it may be difficult to determine which factors are responsible for variation in expression due to linkage disequilibrium and the distances between markers, as each marker may be located close to multiple genes (Li & Deng, 2010).

Genetic markers can either be dominant (such as Diversity Array Technology (DArT)-based PCR markers) or co-dominant (such as SNP markers), based on whether or not they are able to distinguish between homozygous and heterozygous genotypes. A dominant marker has only two alleles and is most commonly either random amplified polymorphic DNA (RAPD) or an amplified fragment length polymorphism (AFLP). Co-dominant markers, on the other hand, can have many different alleles and

are most commonly restriction fragment length polymorphisms (RFLPs) or simple sequence repeats (SSRs), also known as microsatellites. Some advantages and disadvantages of each type of DNA marker are listed in **Table 1.1**. Genetic linkage maps can be constructed by calculating the linkage between DNA markers, using logarithm of odds (LOD) scores, which represents the ratio of linkage to no linkage between markers (Collard *et al.*, 2005). These genetic linkage maps are useful for examining patterns of inheritance of complex quantitative traits across the genome and can be constructed by popular programs such as JoinMap (Stam, 1992).

Table 1.1: Advantages and disadvantages of common molecular DNA markers in QTL analyses

Molecular marker	Dominant or Co-dominant	Advantages	Disadvantages
RAPD	Dominant	Quick, simple, and inexpensive. Requires small amounts of DNA and multiple loci from a single primer are possible.	Has problems with reproducibility and are generally not transferrable.
RFLP	Co-dominant	Robust, reliable, and transferrable across populations.	Time-consuming and expensive. Requires large amounts of DNA and has limited polymorphisms, especially in related lines.
AFLP	Dominant	Multiple loci possible with high levels of polymorphisms generated.	Requires large amounts of DNA and has a complicated methodology.
SSRs	Co-dominant	Technically simple, robust, reliable, and transferrable across populations.	Requires a lot of time and labour to produce primers and requires polyacrylamide electrophoresis.

In more recent studies, microarrays have been replaced by RNA-seq, allowing the use of ASE in eQTL mapping to compare the expression of both alleles at a heterozygous SNP. ASE of a gene refers to the allele-specific transcript abundance, as each gene in a diploid individual has one paternal and one maternal allele. Generally, a *cis*-eQTL can modify the expression of the two alleles in different ways, therefore ASE can be used to distinguish between *cis*- and *trans*-eQTLs (Sun & Hu, 2013). Wang, Richard and Pan (2016) proposed a guided eQTL mapping method that makes use of data-driven prior knowledge to identify candidate genes. This method uses QTL analyses to identify SNP markers linked to specific traits, after which co-expressed gene modules that are significantly linked

with those traits are selected. The authors used the least absolute shrinkage and selection operator (Lasso) to perform eQTL mapping, where a single gene is associated with multiple SNPs. Other popular tools that are available for the mapping of QTLs are compared in **Table 1.2** (Basten, Weir and Zeng, 1994; Shabalin, 2012; Ongen *et al.,* 2015).

Table 1.2: Review and comparison of popular software available for mapping of QTLs

	FastQTL	QTL cartographer	Matrix eQTL
General information	Interactive package for mapping QTLs (multi-trait or multi-environment complexes) in real data and extensive simulations	A suite of programs in C (programming language) used for mapping quantitative trait loci	An intuitive MS-Windows user interface software used for mapping quantitative trait loci in experimental populations
Mode of Operation	Windows-based, flexible, extremely user-friendly graphic interface for research and all levels of teaching. A classroom version is underway	UNIX based, command-line interface (the option of interactive mode presents the user with a menu of numbered options)	Graphical user interface allows an intuitive ease of use
Known limitations	Composite interval mapping option is not available y*et al*though more important multiple interval mapping (MIM) is available	The Windows version of QTL Cartographer either does the conversions on a Windows machine, or ftp the files to the Windows machine as text	Relatively poor spectrum of main types of analysis (e.g. no multiple trait analysis combined with MIM)
Mapping population	Backcross, F2, Dihaploids, RIL selfing, RIL sib mating, F2/F3, F2/4 and multiple families for all these. Underway: double backcross, recombinant intercross, F3, outbred full-sib	Backcross and F2 designs. Analysis of either simulated or real data. Other experimental designs can be incorporated	Experimental population types: BC1, F2, RIL, DH and outbred full-sib family
Model fitted	Marker or interval analysis, single- or two-linked QTLs, single, two- or multiple-trait analysis, QTL-E; all these with MIM. Multiple families; selective genotyping	Single-marker analysis, interval mapping, composite interval mapping, Bayesian interval mapping, MIM, MIM, multiple-trait analysis QTL-E, pleiotropic effects	Interval mapping, MQM mapping, nonparametric mapping with the Kruskal-Wallis rank sum test per marker
Compatibility	Map orders can be imported from several formats or from the new powerful package MultiPoint (fast ordering of many hundreds of markers per chromosome with verification of the map by jack-knifing; allows building consensus maps)	Genetic linkage map data can be imported from MAPMAKER, results can be displayed graphically, or printed using Gnuplot, or imported into any graphics package on any computing platform	Input in plain text files with a flexible layout of the QTL data, marker genotypes and the (precalculated) linkage map; map and molecular marker data files are compatible with JoinMap
Significance testing	Permutation and bootstrap tools; FDR for total analysis	Permutation analysis	Permutation analysis
Operating systems	MS-windows; because of very fast algorithms can also run on Macintosh under PowerPC	Unix, Macintosh, and Microsoft Windows	MS-windows

1.3.3 Network construction

Gene co-expression networks are undirected graphs consisting of nodes and edges. Each node represents a gene and nodes are connected if there is a significant co-expression between the corresponding genes across a sample. Edges therefore represent functional relationships (i.e. interactions) between genes. Zhang and Horvath (2005) proposed a framework for 'soft' thresholding where numerous adjacency functions convert the measure of co-expression to a connection weight which is assigned to each pair of genes. Each resulting co-expression network matches up with an adjacency matrix that specifies the strength of connection between a pair of genes and therefore defines the node connectivity. The weighted networks present node connectivity measures, a measure of topological overlap, and a clustering coefficient that is not inversely related to node connectivity. A flow diagram listing the steps of co-expression network construction is shown in **Figure 1.5**.

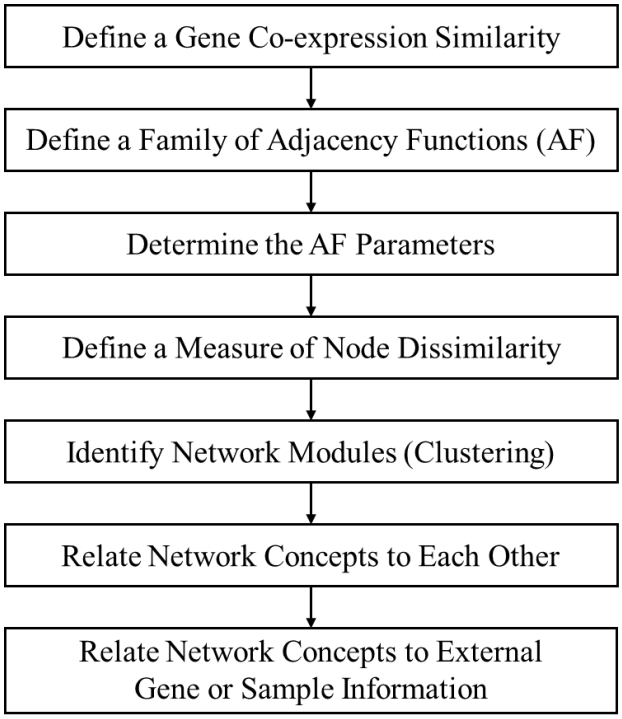


Figure 1.5: Weighted gene co-expression network analysis. Firstly, a gene co-expression similarity needs to be defined, which measures the consistency of gene expression across samples. The next step is to define a family of adjacency functions to determine whether the network will be weighted or not and to construct an adjacency matrix. The adjacency function parameters then need to be determined, which will regulate the node connection strength sensitivity and specificity. Gene modules, which are groups of nodes that have a high topological overlap between them (i.e. are highly co-expressed), are identified by defining a node dissimilarity measure. Clustering methods, along with the dissimilarity measure, are used to group genes into modules and construct the network. Network concepts can then be related to each other and to external information on genes and samples using standard statistical methods. Image from Zhang and Horvath (2005).

A posteriori approaches can use eQTL data to identify novel gene networks and the regulators underlying them, either by determining the correlation between expression patterns or utilizing the co-localization of eQTL locations to identify gene clusters or networks. This approach requires the network under study to be known or predicted before analyses can be done. After novel networks have been identified, eQTLs can be categorized as either *cis* or *trans* and assessed to determine whether they are involved in network regulation (Hansen *et al.*, 2008). Some emergent properties of networks include: (i) the association between connectivity and the strength of selection pressures; (ii) the ability to infer biological knowledge from network topology and gene modules; and (iii) the association between criticality, network topology, and the phenotypic landscape evolvability (Torres-Sosa *et al*., 2012; Mähler *et al*., 2017; Vella *et al.*, 2017).

1.4 Conclusion

In this review, we discussed the state-of-the-art in systems biology for complex trait dissection and high throughput RNA-seq analysis in plant populations. One of the most noticeable trends recently observed in systems biology is the modelling and reconstruction of gene networks, as they are responsible for carrying out and regulating important biological pathways and functions. Studying the variation in topology of these networks is necessary, as it can most likely lead to variations in complex traits and phenotypes. The advent of genome-wide association studies (GWAS) and QTL studies have allowed the identification of genetic variations that are linked to these traits and phenotypes through the analysis of thousands of SNPs.

Approaches to complex data in the future will need to look at each biological system as a whole, by combining information from multiple biological scales including genes, transcripts, proteins, metabolites, microbiomes, and physiological traits. To gain a systems-level understanding of these components, computational models of the different scales can be built and integrated to identify where there is a lack of knowledge or to predict new responses of the biological system. These models

should include the expression of DNA and genes, intracellular networks, transmembrane signals, signals between cells, and possibly even models at the organ level (Finkelstein *et al.*, 2004). The importance of machine learning and artificial intelligence in systems biology is constantly growing, as it is vital for analysing large amounts of complex data and used to train network models that can, for example, allow the identification of candidate genes and predict functions of co-expression networks.

It is important for us to determine the genetic architecture underlying complex wood property traits in *Eucalyptus* hybrid trees, as this will allow us to identify genes and pathways that interact to determine complex trait outcomes. We can use this knowledge to engineer or breed for these traits with the aim of gaining specific biotechnology outcomes, while avoiding negative effects on plant growth. In this study, we aim to gain a systems biology understanding of the genetic architecture of gene expression variation underlying complex phenotypic variation for wood growth and development in segregating populations of mature field-grown *Eucalyptus* trees. We can compare systems genetics data between a three-year-old and eight-year-old population to determine the conservation of eQTLs and the genetic architecture across age in the same trees and to establish whether data from different ages can be used to prioritise candidate genes for genetic engineering.

1.5 References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald Martin, Rubin GM, Sherlock G. 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**: 25–29.

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**: 299–307.

Ballouz S, Verleyen W, Gillis J. 2015. Guidance for RNA-seq co-expression network construction and analysis: Safety in numbers. *Bioinformatics* **31**: 2123–2130.

Basten CJ, Weir BS, Zeng Z-B. 1994. Zmap-a QTL cartographer. *In Proceedings of the 5th World Congress on Genetics Applied to Livestock Production: Computing Strategies and Software* **22**: 65–66.

Bohnert R, Behr J, Rätsch G. 2009. Transcript quantification with RNA-Seq data. *BMC Bioinformatics* **10**: 1–2.

Boyle EA, Li YI, Pritchard JK. 2018. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**: 1177–1186.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* **34**: 525–527.

Breitling R, Li Y, Tesson BM, Fu J, Wu C, Wiltshire T, Gerrits A, Bystrykh L V., De Haan G, Su AI, Jansen RC. 2008. Genetical genomics: Spotlight on QTL hotspots. *PLoS Genetics* **4**: 2–5.

Christie N, Myburg AA, Joubert F, Murray SL, Carstens M, Lin YC, Meyer J, Crampton BG, Christensen SA, Ntuli JF, Wighard SS, Van de Peer Y, Berge DK. 2017. Systems genetics reveals a transcriptional network associated with susceptibility in the maize–grey leaf spot pathosystem. *Plant Journal* **89**: 746–763.

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

Colbert RA, Thompson SD, Glass DN. 2011. Integrative genomics. In: Textbook of Pediatric Rheumatology. Philadelphia: W.B. Saunders, 60–70.

Collard BCY, Jahufer MZZ, Brouwer JB, Pang ECK. 2005. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica*: 169–196.

De Klerk E, Den Dunnen JT, 'T Hoen PAC. 2014. RNA sequencing: From tag-based profiling to resolving complete transcript structure. *Cellular and Molecular Life Sciences* **71**: 3537–3551.

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

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

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

Drost DR, Puranik S, Novaes E, Novaes CRDB, Dervinis C, Gailing O, Kirst M. 2015. Genetical genomics of *Populus* leaf shape variation. *BMC Plant Biology* **15**: 1–10.

Druka A, Potokina E, Luo Z, Jiang N, Chen X, Kearsey M, Waugh R. 2010. Expression quantitative trait loci analysis in plants. *Plant Biotechnology Journal* **8**: 10–27.

Fagny M, Paulson JN, Kuijjer ML, Sonawane AR, Chen C-Y, Lopes-Ramos CM, Glass K, Quackenbush J, Platig J. 2017. Exploring regulation in tissues with eQTL networks. *Proceedings of the National Academy of Sciences* **114**: E7841–E7850.

Falconer DS, Mackay TFC. 1996. *Introduction to quantitative genetics*. Harlow, Essex, UK: Longmans Green.

Feltus FA. 2014. Systems genetics: A paradigm to improve discovery of candidate genes and mechanisms underlying complex traits. *Plant Science* **223**: 45–48.

Finkelstein A, Hetherington J, Margoninski O, Saffrey P, Seymour R, Warner A. 2004. Computational challenges of systems biology. *IEEE Computer Society* **37**: 26–33.

Garber M, Grabherr MG, Guttman M, Trapnell C. 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods* **8**: 469–477.

Goddard ME, Kemper KE, MacLeod IM, Chamberlain AJ, Hayes BJ. 2016. Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture. *Proceedings of the Royal Society B: Biological Sciences* **283**: 20160569.

Hansen TF. 2006. The evolution of genetic architecture. *Annual Review of Ecology, Evolution, and Systematics* **37**: 123–157.

Hansen BG, Halkier BA, Kliebenstein DJ. 2008. Identifying the molecular basis of QTLs: eQTLs add a new dimension. *Trends in Plant Science* **13**: 72–77.

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

Holloway B, Luck S, Beatty M, Rafalski J-A, Li B. 2011. Genome-wide expression quantitative trait loci (eQTL) analysis in maize. *BMC Genomics* **12**: 336.

Hrdlickova R, Toloue M, Tian B. 2017. RNA-Seq methods for transcriptome analysis. *Wiley Interdisciplinary Reviews: RNA* **8**.

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

Joosen R, Ligterink W, Hilhorst H, Keurentjes J. 2009. Advances in genetical genomics of plants. *Current Genomics* **10**: 540–549.

Jordan MC, Somers DJ, Banks TW. 2007. Identifying regions of the wheat genome controlling seed development by mapping expression quantitative trait loci. *Plant Biotechnology Journal* **5**: 442–

453.

Kadarmideen HN. **2008**. Genetical systems biology in livestock : application to gonadotrophin releasing hormone and reproduction. *IET Systems Biology* **2**: 429–441.

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**: 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* **104**: 1708–1713.

Kim B-J, Kim S-H. **2018**. Prediction of inherited genomic susceptibility to 20 common cancer types by a supervised machine-learning method. *Proceedings of the National Academy of Sciences*: 201717960.

Kim D, Langmead B, Salzberg SL. **2015**. HISAT: A fast spliced aligner with low memory requirements. *Nature Methods* **12**: 357–360.

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**: 2295–2303.

Kirst M, Myburg AA, De Leon JPG, Kirst ME, Scott J, Sederoff R. **2004**. Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of *Eucalyptus*. *Plant Physiology* **135**: 2368–2378.

Kitano H. **2002**. Computational systems biology. *Nature* **420**: 206–210.

Kitano H. **2007**. Systems biology: A brief overview. *Science* **1662**: 1662–1665.

Kloosterman B, Anithakumari AM, Chibon P-Y, Oortwijn M, van der Linden GC, Visser RGF, Bachem CWB. **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.

Kruijer W, Boer MP, Malosetti M, Flood PJ, Engel B, Kooke R, Keurentjes JJB, Van Eeuwijk FA. **2014**. Marker-based estimation of heritability in immortal populations. *Genetics* **199**: 379–398.

Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. **2009**. Circos: An information aesthetic for comparative genomics. *Genome Research* **19**: 1639-1645.

Kullan ARK, van Dyk MM, Hefer CA, Jones N, Kanzler A, Myburg AA. **2012**. Genetic dissection of growth, wood basic density and gene expression in interspecific backcrosses of *Eucalyptus grandis* and *E. urophylla*. *BMC Genetics* **13**: 1–12.

Langmead B, Salzberg SL. **2012**. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357.

Lee S-I, Dudley AM, Drubin D, Silver PA, Krogan NJ, Pe'er D, Koller D. **2009**. Learning a prior on regulatory potential from eQTL data. *PLoS Genetics* **5**: 1–24.

Li H, Deng H. **2010**. Systems genetics, bioinformatics and eQTL mapping. *Genetica* **138**: 915–924.

Li H, Durbin R. **2009**. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Li L, Petsch K, Shimizu R, Liu S, Xu WW, Ying K, Yu J, Scanlon MJ, Schnable PS, Timmermans MCP, Springer NM, Muehlbauer GJ. **2013**. Mendelian and non-Mendelian regulation of gene expression in maize. *PLoS Genetics* **9**: e1003202.

Libbrecht MW, Noble WS. **2015**. Machine learning applications in genetics and genomics. *Nature Reviews Genetics* **16**: 321–332.

Mackay TFC. **2001**. The genetic architecture of quantitative traits. *Annual Review of Genetics* **35**: 303–339.

Mähler N, Wang J, Terebieniec BK, Ingvarsson PK, Street NR, Hvidsten TR. **2017**. Gene co-expression network connectivity is an important determinant of selective constraint. *PLoS Genetics* **4**: 1-33.

Mäki-Tanila A, Hill WG. **2014**. Influence of gene interaction on complex trait variation with multilocus models. *Genetics* **198**: 355–367.

Miller DFB, Yan P, Fang F, Buechlein A, Kroll K, Frankhouser D, Stump C, Stump P, Ford JB, Tang H, Michaels S, Matei D, Huang TH, Chien J, Liu Y, Rusch DB Nephew KP. 2017. Complete Transcriptome RNA-Seq. In: *Cancer Gene Networks*. New York: Humana Press, 141-162.

Mizrachi E, Myburg AA. **2016**. Systems genetics of wood formation. *Current Opinion in Plant Biology* **30**: 94–100.

Mizrachi E, Verbeke L, Christie N, Fierro AC, Mansfield SD, Davis MF, Gjersing E, Tuskan GA, Van Montagu M, Van de Peer Y, Marchal K, Myburg AA. **2017**. Network-based integration of systems genetics data reveals pathways associated with lignocellulosic biomass accumulation and processing. *Proceedings of the National Academy of Sciences* **114**: 1195-1200.

Mohri M, Rostamizadeh A, Talwalkar A. **2018**. *Foundations of machine learning*. MIT press.

Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D, Goodstein DM, Dubchak I, Poliakov A, Mizrachi E, Kullan ARK, Hussey SG, Pinard D, van der Merwe K, Singh P, Van Jaarsveld I, Silva-Junior OB, Togawa RC, Pappas MR, Faria DA, Sansaloni CP, Petroli CD, Yang X, Ranjan P, Tschaplinski TJ, Ye C, Li T, Sterck L, Vanneste K, Murat F, Soler MM, Clemente HS, Saidi N, Cassan-Wang H, Dunand C, Hefer CA, Bornberg-Bauer E, Kersting AR, Vining K, Amarasinghe V, Ranik M, Naithani S, Elser J, Boyd AE, Liston A, Spatafora JW, Dharmwardhana P, Raja R, Sullivan C, Romanel E, Alves-Ferreira M, Külheim CK, Foley W, Carocha V, Paiva J, Kudrna D, Brommonschenkel SH, Pasquali G, Byrne M, Rigault P, Tibbits J, Spokevicius A, Jones RC, Steane DA, Vaillancourt RE, Potts BM, Joubert F, Barry K, Pappas Jr GJ, Strauss SH, Jaiswal P, Grima-Pettenati J, Salse J, Van de Peer Y, Rokhsar DS, Schmutz J. **2014**. The genome of *Eucalyptus grandis*. *Nature* **510**: 356.

Nadeau JH, Dudley AM. **2011**. Systems genetics. *Science* **331**: 1015–1016.

Ohler U, Liao G, Niemann H, Rubin GM. **2002**. Computational analysis of core promoters in the *Drosophila* genome. *Genome Biology* **3**: research0087--1.

Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O. **2016**. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**: 1479–1485.

Oubounyt M, Louadi Z, Tayara H, Chong KT. **2019**. DeePromoter: Robust promoter predictor using deep learning. *Frontiers in Genetics* **10**: 1–9.

Panousis NI, Gutierrez-Arcelus M, Dermitzakis ET, Lappalainen T. **2014**. Allelic mapping bias in RNA-sequencing is not a major confounder in eQTL studies. *Genome Biology* **15**: 467.

Park CC, Gale GD, de Jong S, Ghazalpour A, Bennett BJ, Farber CR, Langfelder P, Lin A, Khan AH, Eskin E, Horvath S, Lusis AJ, Ophoff RA, Smith DJ. **2011**. Gene networks associated with conditional fear in mice identified using a systems genetics approach. *BMC Systems Biology* **5**: 43.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. **2017**. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14**: 417–419.

Pertea M, Pertea GM, Antonescu CM, Chang T, Mendell JT, Salzberg SL. **2015**. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* **33**: 209.

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

Porter J, Berkhahn J, Zhang L. **2015**. A comparative analysis of read mapping and indel calling pipelines for next-generation sequencing data. *Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology* 521-35.

Salzberg SL. **2016**. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* **33**: 290–295.

Schön CC, Utz HF, Groh S, Truberg B, Openshaw S, Melchinger AE. **2004**. Quantitative trait locus mapping based on resampling in a vast maize testcross experiment and its relevance to quantitative genetics for complex traits. *Genetics* **498**: 485–498.

Shabalin AA. **2012**. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**: 1353–1358.

Shi C, Uzarowska A, Ouzunova M, Landbeck M, Wenzel G, Lübberstedt T. **2007**. Identification of candidate genes associated with cell wall digestibility and eQTL (expression quantitative trait loci) analysis in a Flint x Flint maize recombinant inbred line population. *BMC Genomics* **8**: 22.

Sik D, Ho W, Schierding W, Wake M, Saffery R, Sullivan JO, Sullivan JO. **2019**. Machine learning SNP based prediction for precision medicine. *Frontiers in Genetics* **10**: 1–10.

Stam, P. **1993**. Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. *The Plant Journal* **3**: 739-744.

Sun W, Hu Y. **2013**. eQTL Mapping Using RNA-seq Data. *Statistics in Biosciences* **5**: 1–25.

Sun CS, Markey MK. **2011**. Recent advances in computational analysis of mass spectrometry for proteomic profiling. *Journal of Mass Spectrometry* **46**: 443–456.

Swanson-Wagner RA, DeCook R, Jia Y, Bancroft T, Ji T, Zhao X, Nettleton D, Schnable PS. **2009**. Paternal dominance of *trans*-eQTL influences gene expression patterns in maize hybrids. *Science* **326**: 1118–1120.

Thavamanikumar S, Southerton SG, Bossinger G, Thumma BR. **2013**. Dissection of complex traits in forest trees - opportunities for marker-assisted selection. *Tree Genetics and Genomes* **9**: 627–639.

Torres-Sosa C, Huang S, Aldana M. **2012**. Criticality is an emergent property of genetic networks that exhibit evolvability. *PLoS Computational Biology* **8**: e1002669.

Trapnell C, Pachter L, Salzberg SL. **2009**. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**: 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**: 511–515.

Vella D, Zoppis I, Mauri G, Mauri P, Di Silvestre D. **2017**. From protein-protein interactions to protein co-expression networks: a new perspective to evaluate large-scale proteomic data. *EURASIP Journal on Bioinformatics and Systems Biology* **2017**: 6.

Visscher PM, Goddard ME. **2010**. Systems genetics: The added value of gene expression. *HFSP Journal* **4**: 6–10.

Visscher PM, Hill WG, Wray NR. **2008**. Heritability in the genomics era - Concepts and misconceptions. *Nature Reviews Genetics* **9**: 255–266.

Wang JP, Matthews ML, Williams CM, Shi R, Yang C, Tunlaya-anukit S, Chen H, Li Q, Liu J, Lin C, Naik P, Sun Y, Loziuk PL, Yeh T, Kim H, Gjersing E, Shollenberger T, Shuford CM, Song J, Miller Z, Huang Y, Edmunds CW, Liu B, Sun Y, Lin YJ, Li W, Chen H, Peszlen I, Ducoste JJ, Ralph J, Chang H, Muddiman DC, Davis MF, Smith C, Isik F, Sederoff R, Chiang VL. **2018**. Improving wood properties for wood utilization through multi-omics integration in lignin biosynthesis. *Nature Communications* **9**: 1579.

Wang Y, Richard R, Pan Y. **2016**. Prior knowledge guided eQTL mapping for identifying candidate genes. *BMC Bioinformatics* **17**: 531.

Wang J, Yu H, Weng X, Xie W, Xu C, Li X, Xiao J, Zhang Q. **2014**. An expression quantitative trait loci-guided co-expression analysis for constructing regulatory network using a rice recombinant inbred line population. *Journal of Experimental Botany* **65**: 1069–1079.

Wang J, Yu H, Xie W, Xing Y, Yu S, Xu C, Li X, Xiao J, Zhang Q. **2010**. A global analysis of QTLs for expression variations in rice shoots at the early seedling stage. *The Plant Journal* **63**: 1063–1074.

Wentzell AM, Rowe HC, Hansen BG, Ticconi C, Halkier BA, Kliebenstein DJ. **2007**. Linking metabolic QTLs with network and *cis*-eQTLs controlling biosynthetic pathways. *PLoS Genetics* **3**: 1687–1701.

West MAL, Kim K, Kliebenstein DJ, Van Leeuwen H, Michelmore RW, Doerge RW, St. Clair DA. **2007**. Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in *Arabidopsis*. *Genetics* **175**: 1441–1450.

Wolen A, Miles MF. **2012**. Identifying gene networks underlying the neurobiology of ethanol and alcoholism. *Alcohol Research: Current Reviews.* **34**: 306–317.

Wong KH, Jin Y, Moqtaderi Z. **2013**. Multiplex Illumina sequencing using DNA barcoding.

Current Protocols in Molecular Biology **101**: 7–11.

Wu R, Ma C-X, Casella G. **2007**. *Statistical Genetics of Quantitative Traits*. Springer US.

Wu TD, Nacu S. **2010**. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**: 873–881.

Xu Y, Li P, Zou C, Lu Y, Xie C, Zhang X, Prasanna BM. **2017**. Enhancing genetic gain in the era of molecular breeding. *Journal of Experimental Botany* **68**: 2641–2666.

Yang L, Ma Z, Cao C, Zhang Y, Wu X, Lee R, Hu B, Wen L, Ge H, Huang Y, Lao K, Tang F. **2017**. MR-seq: measuring a single cell's transcriptome repeatedly by RNA-seq. *Science Bulletin* **62**: 391–398.

Zhang X, Acencio ML, Lemke N. **2016**. Predicting essential genes and proteins based on machine learning and network topological features: A comprehensive review. *Frontiers in Physiology* **7**: 1–11.

Zhang B, Horvath S. **2005**. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology* **4**.

Zhang J, Nieminen K, Serra JAA, Helariutta Y. **2014**. The formation of wood and its control. *Current Opinion in Plant Biology* **17**: 56–63.

Zhang J, Yang Y, Zheng K, Xie M, Feng K, Jawdy SS, Gunter LE, Ranjan P, Singan VR, Engle N, Lindquist E, Barry K, Schmutz J, Zhao N, Tschaplinski TJ, LeBoldus J, Tuskan GA, Chen JG, Muchero W. **2018**. Genome-wide association studies and expression-based quantitative trait loci analyses reveal roles of HCT2 in caffeoylquinic acid biosynthesis and its regulation by defense-responsive transcription factors in *Populus*. *New Phytologist* **220**: 502–516.

Zhang C, Zhang B, Lin LL, Zhao S. **2017**. Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics* **18**: 1–11.

Zhu M, Yu M, Zhao S. **2009**. Understanding quantitative genetics in the systems biology era. *International Journal of Biological Sciences* **5**: 161–170.

CHAPTER 2

Rapid Genetic Dissection of Xylem Gene Expression Variation in a *Eucalyptus* Interspecific Backcross Population

Lizette Loubser, Marja O'Neill, Raphael Ployet, Nanette Christie, Eshchar Mizrachi, Alexander A. Myburg

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

This chapter has been prepared in the format of a draft research manuscript for submission to a peer-reviewed journal (e.g. *New Phytologist*). I performed all the analyses in this manuscript, prepared the manuscript and developed a new pipeline for RNA-seq expression profiling, SNP calling, and genetic linkage map construction. Nanette Christie provided the analysis pipelines, base files, and supporting scripts for eQTL mapping, eQTL backend analysis, eQTL overlap analysis, co-expression analysis, Fisher's test, GO enrichment analysis, and generating Circos plots. She also gave many valuable suggestions, advice, and guidance throughout the study and provided valuable manuscript revisions. Raphael Ployet assisted with the evaluation and filtering of samples, network analysis using WGCNA, and network construction in Cytoscape. He also provided valuable suggestions, advice, and guidance throughout the study. Marja O'Neill arranged the shipping, library preparation, and sequencing of samples, provided technical assistance with the extraction of SNP chip genotypes and provided valuable manuscript revisions. Eshchar Mizrachi provided valuable suggestions and advice that enhanced and extended the scope of the study. Alexander Myburg devised the study, gave many valuable suggestions, advice and guidance throughout the study, helped with the drafting of the manuscript and provided many valuable manuscript revisions.

2.1 Summary

Systems genetics approaches aim to understand the flow of biological information underlying complex traits by quantifying intermediate phenotypes, such as gene expression levels. To characterise the genetic architecture of genes involved in xylogenesis and analyse the age-to-age correlation of xylem gene expression, we performed whole-transcriptome sequencing of developing xylem tissue from 156 individuals in an interspecific *Eucalyptus* hybrid population sampled at juvenile (3.5 years) and rotation age (8.5 years). For the first time, we were able to construct a robust genetic linkage map by extracting SNP genotypes from population-wide transcriptome data of highly expressed genes. The genetic framework map consisted of 236 markers distributed over 11 linkage groups, spanning 1,008 cM with an average interval size of 4.5 cM. This genetic map, together with existing bioinformatic pipelines for eQTL data analysis and co-expression analysis, allowed us to rapidly identify biologically enriched eQTL hotspots and gene expression modules to build systems models of gene expression variation in xylem tissue from transcriptome data alone. This approach not only allowed us to identify regulatory polymorphisms affecting gene expression at different developmental stages, it also enabled us to determine the genetic correlation for a variety of complex traits and establish whether the underlying genetic basis for variation remains conserved across age. Age-to-age correlation of systems genetics data in the same population allowed us to identify novel and conserved regulatory polymorphisms acting at different ages. This also enabled us to compare the genetic architecture across age, which will ultimately establish whether systems genetics data from different ages can be used to prioritise candidate genes for genetic engineering.

2.2 Introduction

The development of secondary xylem (xylogenesis) is a complex process involving thousands of genes that determine the chemical and physical properties of wood, as well as intricate molecular networks underlying phenotypic variation (Kirst *et al.*, 2004; Dharanishanthi & Ghosh, 2016). Apart from providing raw materials such as paper, pulp and timber, wood is also an economically important renewable energy source and provides energy-efficient building materials (Plomion *et al.*, 2001; Mizrachi & Myburg, 2016). The quality of these products is dependent on the combined woody biomass properties of the tree, including wood density, growth, and wood chemistry, which are difficult to improve in long-lived species such as *Eucalyptus*. *Eucalyptus* plantations have emerged as key renewable feedstocks due to their fast growth, exceptional wood properties and high adaptability and *Eucalyptus* hybrids with specific genotypes have been extensively propagated due to their unique hybrid properties (Mizrachi *et al.*, 2017). These hybrids are especially appropriate for characterising the genetic architecture underlying complex wood property traits and the natural variation in transcript abundance (Kirst *et al.*, 2005).

To better understand the mechanisms underlying variation in complex wood property traits, systems genetics approaches can be used to gain insight into the relationship between the genotype and the phenotype to understand how these traits are regulated at a genetic level and to dissect the molecular networks underlying phenotypic variation (Blein-Nicolas *et al.*, 2019). These approaches can combine population-wide genetical genomics and gene co-expression analyses to produce systems genetics models that characterises the genetic architecture of transcript abundance (Feltus, 2014). Genetical genomics involves the identification and mapping of expression quantitative trait loci (eQTLs), which describe the association of genotypic variation with variation in transcript abundance in a segregating population (Jansen & Nap, 2001). Gene co-expression analyses attempt to divide genes with highly correlated expression patterns into modules, which suggests that genes in the same module are under the same transcriptional control and that each module can be enriched for one or

more related biological processes (Zhang & Horvath, 2005). Phenotypic data, such as wood biorefinery traits, can also be integrated into systems genetics models to identify candidate genes associated with variation in the phenotype (Christie *et al.*, 2017). Understanding the genetic architecture underlying complex wood property traits will guide the application of marker-assisted selection for breeding, which is especially desirable in forest trees due to their long reproductive cycles and the time it takes for them to express mature traits (Thavamanikumar *et al.*, 2013).

As reviewed by Myburg *et al.* (2019), many studies have explored the genetic architecture of gene expression variation in trees and have reported different results in natural populations (Mähler *et al.*, 2017; Zhang *et al.*, 2018) and pedigrees (Kirst *et al.*, 2005; Drost *et al.*, 2010), which suggests that the genetic background of a population is an important factor to take into account when inferences are made from the evidence in a wider context. As far as we know, only one true systems genetics study on *Eucalyptus* has been reported (Mizrachi *et al.*, 2017), with others only focusing on genetical genomics approaches (Drost *et al.*, 2015). Although much research has been done on the relationship between variation in complex wood property traits and genes involved in xylogenesis, there is still much to uncover about the regulatory effect of the genetic variation underlying these traits. Gaining a systems biology understanding of the genetic architecture of gene expression variation underlying growth and wood property traits in *Eucalyptus* hybrids will allow us to identify genes and pathways that interact to determine complex trait phenotypes. This will allow us to engineer or breed for these traits with the aim of gaining specific biotechnology outcomes, while avoiding negative effects on plant growth.

Here, we characterise the genetic architecture of xylem gene expression at rotation age and perform a comparison of the genetic architecture of genes involved in xylogenesis between juvenile and mature trees using an integrative systems genetics approach. We identify possible developmental and stress-related changes between juvenile and mature trees, which provides new insight into the

regulation of genes related to abiotic stress response in developing xylem. The three-year-old population mentioned here has already been analysed and discussed extensively (Mizrachi et al., 2017; Christie *et al*., in preparation), whereas the eight-year-old population was sequenced recently. We repeated all analyses on the three-year-old population in parallel with the eight-year-old population using the same bioinformatics pipeline to eliminate as many technical differences or inconsistencies as possible that may affect the accuracy of the comparison. This also allows us, for the first time, to generate a genetic map derived from over 250 xylem transcriptome SNPs that is the same for both populations. Our paper mainly focuses on the genetic architecture of xylem gene expression for the eight-year-old population and concludes with an age-to-age comparison of the genetic architecture between juvenile and mature trees. Due to the fact that the eight-year-old trees were clonally propagated, we report for the first time on the heritability of the xylem transcriptome.

2.3 Materials and Methods

2.3.1 Plant materials

The population under study is an interspecific F_2 backcross population that was derived from a *Eucalyptus grandis x E. urophylla* F_1 hybrid used as pollen parent (GUSAP1, Sappi, South Africa) and an *E. urophylla* seed parent (USAP1), both previously described by Kullan *et al*. (2012). The population consists of 308 seedling progeny and 375 clonal progeny that were sampled at three and eight years of growth respectively. The trial site is located near Kwambonambi in KwaZulu-Natal (Sappi, South Africa) which is on a flat coastal land with deep sandy soils and little spatial variation (Kullan *et al.*, 2012).

2.3.2 mRNA library preparation, sequencing, and expression profiling

Immature xylem samples were collected as described by Mizrachi *et al.* (2017) from 156 and 144 trees at three and eight years of growth respectively. Three-year-old samples were sent to the

genomics service provider, Beijing Genomics Institute (BGI), on dry ice for total RNA extraction and RNA sequencing (30 million, Illumina PE50, BGI Hong Kong). Eight-year-old samples were sent to Novogene on dry ice for mRNA extraction, eukaryotic strand-specific RNA-seq library preparation, and RNA sequencing (30 million, Illumina HiSeq-PE150, Novogene Hong Kong). Four random samples from the eight-year-old population were sequenced in duplicate to gain insight into the technical repeatability of the RNA sequencing. Quality control of the raw RNA-seq reads was performed with FastQC v0.11.7 (Andrews S., 2010) and reports for all the samples were aggregated into a single report with MultiQC v1.6 (Ewels *et al.*, 2018). Reads were aligned to the reference genome assembly of *E. grandis* v2.0 (Myburg *et al.*, 2014) using STAR v2.6 (Dobin *et al.*, 2013) and transcripts were quantified in Transcript Per Kilobase Million (TPM) values using StringTie v1.3.4 (Pertea *et al.*, 2015). The genetic architecture was compared across age based on correlations within each dataset, therefore technical differences due to different generations of sequencers were ignored.

2.3.3 SNP calling on RNA-seq data and sample filtering

Transcript-derived SNPs were extracted using the Genome Analysis Toolkit (GATK) best practices pipeline for SNP calling in RNA-seq data (van der Auwera *et al.*, 2013). Read groups were first added to aligned reads to allow genotyping of each sample individually and PCR duplicates were marked using Picard Tools v2.17.11 (http://broadinstitute.github.io/picard). The GATK *SplitNCigarReads* tool was used to split reads into exons and clip sequences overhanging into introns. Variant calling was performed using the *HaplotypeCaller* tool and indels were removed using the *SelectVariants* tool. Hard filters were then applied to the resulting set of SNPs with the *VariantFiltration* tool's default parameters and converted to a user-friendly table format using the *VariantsToTable* tool. For the eight-year-old population, an identity-by-descent (IBD) analysis was performed with the SNP & Variation Suite v8.8.3 (Golden Helix, Inc., Bozeman, MT, www.goldenhelix.com) to confirm whether all the samples belonged to the same backcross family and eliminate clonal or technical replicates, as this dataset had not been analysed previously. This was done by comparing genotypic

data previously generated (unpublished) using the EuCHIP60K SNP chip (Silva-Junior *et al*., 2015) with the transcript-derived SNPs, as well as doing an all-by-all comparison with the transcript-derived SNPs. Finally, the Weighted Gene Co-expression Network Analysis (WGCNA v1.66) package in R (Langfelder & Horvath, 2008) was used to cluster samples based on their expression profiles to identify outliers in the dataset and confirm technical and clonal replicates.

2.3.4 Genetic linkage map construction from transcriptome data

Engelbrecht *et al.* (unpublished) identified around 2000 SNPs within highly expressed genes in the three-year-old population that could be used as testcross markers to construct a genetic linkage map for the F_1 hybrid parent. Testcross markers have a 1:1 segregation pattern (nnxnp), where the marker is homozygous in the *E. urophylla* (maternal) parent and heterozygous in the F_1 hybrid (paternal) parent. The genotypes for these SNPs were extracted for the eight-year-old individuals using the same method described previously for deriving SNPs from transcriptome data. A minimum coverage of 10 and a call rate of 0.75 per marker was used as cut-off for constructing a high-density genetic map, whereas a more stringent call rate of 0.9 was used for the framework map. Around 800 markers were analysed using JoinMap 4.1 (Van Ooijen, 2006) where a genotype frequency was calculated for each locus to determine the segregation distortion (i.e. significant deviation from expected Mendelian ratios) using a Chi-square test (χ^2). Markers were sorted into linkage groups based on an independence logarithm of odds (LOD) score calculated for the recombination frequency and a LOD score threshold of four was used. Regression mapping was then performed on each linkage group using Kosambi's mapping function with default mapping parameters. Markers with a nearest neighbour (NN) fit (cM) value ≥ 3 were manually removed and mapping was repeated until we were left with a high-density map where all the markers were in the correct order. A genetic framework map was then derived from the high-density map by pruning markers to achieve a spacing of 4-5 cM. Markers with a call rate below 0.9, poor goodness-of-fit, or incorrect ordering were manually replaced one-by-one until all the markers fit well within the linkage group. The genetic positions of the framework map were

aligned to the physical map (*E. grandis* v2.0 assembly) and visualised with MapChart (Voorrips, 2002).

2.3.5 Gene co-expression analysis

Co-expression networks were constructed with WGCNA by assigning genes with highly correlated expression profiles to co-expression modules, based on Pearson correlations. A soft-thresholding power of four was selected to build an adjacency matrix. The scale-free topology criterion suggests that a soft-thresholding power (adjacency function parameter) should be chosen that will result in a network with a scale-free topology model fit of at least $R^2 > 0.8$, a high mean connectivity, and a saturated curve for the relationship between the adjacency function parameter and R^2. Genes were then assigned to modules based on the workflow described by Zhang and Horvath (2005) and an average gene profile was calculated for each module, referred to as the module eigengene. The module eigengene network for our eight-year-old dataset was visualised with Cytoscape v3.7.1 (Shannon *et al.*, 2003) to show the biggest cluster of correlated gene modules (absolute correlation > 0.5; p-value < 0.00001).

2.3.6 eQTL analysis

eQTL mapping, classification, hotspot detection, and overlap analyses were performed on the three- and eight-year-old datasets with a robust pipeline for eQTL analysis developed by Christie *et al.* (2017). QTL Cartographer's composite interval mapping (CIM) approach was used to map eQTLs with a forward regression step and a backward elimination step, with p-value < 0.1 for both (Basten *et al.*, 2005). The average genome-wide likelihood ratio (LR) threshold of 11.5 was previously calculated through permutation analyses by Mizrachi *et al.* (2017), to correct for multiple traits and genome-wide markers (LR = LOD/0.217). User-specified parameters include a window size of 10 and walking speed of 1 cM. The classification of an eQTL as either *cis* or *trans* was based on the position of the eQTL relative to the gene with which it is associated. A *cis*-eQTL is located at a

position less than half the average size of an eQTL away from its linked gene, on the same chromosome, whereas a *trans*-eQTL is further away and is usually located on a different chromosome. Genome-wide eQTL frequency was calculated in bins of size 1 cM and normalised for local gene density to identify global *trans*-eQTL hotspots (i.e. where more *trans*-eQTLs occur than is expected by chance). Hotspot names indicate the chromosome where the hotspot is located, as well as a Mb bin where many eQTL peaks are located for that hotspot (e.g. HS_1.39 is located on chromosome 1 with many peaks at the 39 Mb position). To compare the conservation of eQTLs for the same genes between the three- and eight-year-old datasets, a pairwise eQTL overlap score was calculated on the shared genetic map for genes with eQTLs in both datasets, as described by Mizrachi *et al.* (2017). An overlap score of 1.0 indicates a complete overlap between the eQTLs. A score between 0.5 and 1.0 indicates a partial overlap between the eQTLs, where they either start and end in the same 1 cM bin with the peaks being in different bins, or they start or end in different bins with high overlap between peaks in the same bin. Scores below 0.5 were not considered significant.

2.3.7 Integrative systems genetics analysis

eQTL hotspots and co-expression data were integrated in systems genetics models to illustrate how groups of co-expressed genes can be linked to shared regulatory loci. A Fisher's exact test (Upton, 1992) was used to determine the significance of association (shared membership) between gene modules and *trans*-eQTL hotspots, with a false discovery rate (FDR) adjusted p-value to correct for multiple comparisons. eQTL hotspots were also split based on the direction of the estimated additive effects of the eQTLs (i.e. whether higher transcript abundance is associated with the *E. grandis* or *E. urophylla* parental allele) and tested for significant overlaps with gene modules. Circos plots (Krzywinski *et al.*, 2009) and Cytoscape networks were used to visualise and compare the genetic architecture (i.e. module membership and shared eQTLs) of genes involved in the general phenylpropanoid, cellulose, and xylan biosynthesis pathways between three- and eight-year-old samples (Christie *et al*., in preparation).

2.3.8 Gene ontology enrichment analysis

Gene ontology (GO) enrichment analyses (Ashburner *et al.*, 2000) were performed for all the genes in gene modules, *trans*-eQTL hotspots and split hotspots, as well as for all the significantly overlapping module-hotspot gene sets. This was done using the R package TopGO v2.34 (Alexa & Rahnenführer, 2018) and a gene-to-GO annotation file, previously created with annotations from Plaza (Proost *et al.*, 2015), Phytozome (Goodstein *et al.*, 2012) and a study by Kersting *et al.* (2015). User-specified parameters include the "elim" algorithm for calculating enrichment, the Fisher's exact test statistic, FDR adjusted p-values to correct for multiple testing, a minimum node size of five, and listing the top 150 significant GO terms identified by the algorithm.

2.4 Results

2.4.1 Background

We analysed transcriptome data from developing xylem tissues for 156 three-year-old and 100 eight-year-old individuals of the *E. grandis* x *E. urophylla* backcross to *E. urophylla* population. The RNA-seq data quality control and expression profiling is discussed in **Supplementary Note 2.1**. Genes were filtered for downstream analyses based on their expression in at least 25% of the population (i.e. TPM > 0 in 25% or more of the population), resulting in a final set of 24,861 expressed genes for the three-year-old population and 25,267 expressed genes for the eight-year-old population. Descriptive statistics for both populations used in downstream analyses are summarised in **Table 2.1**.

Table 2.1: Descriptive statistics of three- and eight-year-old population datasets used in downstream analyses

Population	Number of individuals	Genes	Gene overlap[a]	Unique genes[b]	Min TPM	Max TPM	Mean TPM	Median TPM	SD[c]
3-year-old	156	24,861	92.1%	1,976	0	22,024	40.21	6.19	188
8-year-old	100	25,267	90.5%	2,382	0	22,674	39.57	4.55	219.6

[a] Percentage of genes overlapping between the three- and eight-year-old population
[b] Genes not expressed in the other population
[c] Standard deviation of TPM values

2.4.2 Broad-sense heritability analysis in clones

To analyse the broad-sense heritability (H^2) of the transcriptome, also known as the individual heritability or repeatability when studying clonal populations, we identified 20 clonal pairs in the eight-year-old population and confirmed that the genotypes of each pair were identical, based on transcriptome-derived SNPs (**Supplementary Figure 2.1**). Trees were planted in a common garden, therefore limiting the amount of environmental variation that can influence variation in the transcriptome. If there was no effect from the environment, we would expect $H^2 = 1$, as there was no genetic variation between clonal partners. Spearman rank correlations (Spearman, 1904) were calculated as a proxy for heritability across 25,307 genes, where the distribution of H^2 values for transcript-level variation ranged from -0.60 to 0.99 and the H^2 across the entire transcriptome per clonal pair ranged from 0.92 to 0.98, with a mean of 0.97 (**Figure 2.1; Supplementary File 2.2**). This explains the proportion of the transcriptome that is controlled by genetic variation and allows us to discriminate between genetic and environmental effects. The heritability of each gene was also calculated across the 20 pairs and the calculation was repeated ten times with 20 random non-clonal pairs to determine an estimate of the average full-sib family heritability (**Supplementary Figure 2.2**). We did not expect the repeatability of non-clonal pairs to be zero, as there was still a certain amount of relatedness within the family. The transcriptome-wide correlations per gene for non-clonal pairs were centered around zero (mean = -0.02), while the correlations per gene for 20 confirmed clonal pairs were centered around 0.5 (mean = 0.45). The difference in repeatability between clones and siblings was also calculated to assess the impact of non-additive variation on the population (ΔH^2 = mean H^2_{clonal} – mean $H^2_{\text{non-clonal}}$ = 0.47).

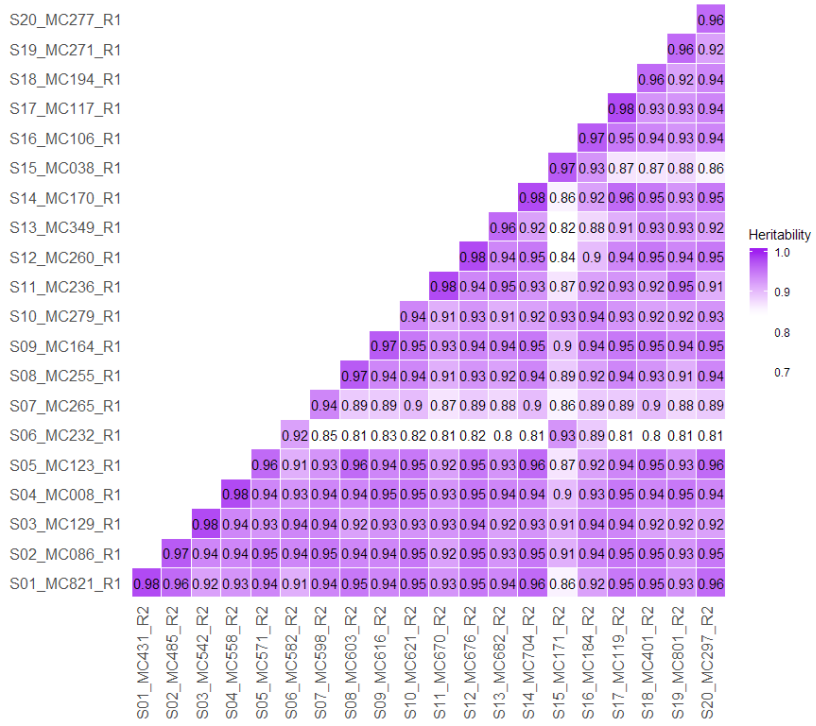


Figure 2.1: Overall heritability of transcriptome variation based on 20 clonal pairs. Spearman rank correlations were used to calculate the overall broad-sense heritability (H^2) of gene expression between two clonal replicates (first row of diagonal values), ranging from 0.92 to 0.98 with a mean of 0.97. Sample pairs are indicated by the "S" prefix and the biological replicate number is indicated by the "R" suffix.

We also investigated the types of genes that tend to have very high or very low repeatability, by performing a GO enrichment analysis on genes in the top (10%) and bottom (10%) of the list of H^2 values (**Supplementary File 2.2**). We expected to see that key developmental genes will tend to have very high H^2 values, whereas environmental response genes will have very low H^2 values. We found that genes with high values ($H^2 > 0.765$) were significantly enriched for defense response and signal transduction processes, whereas genes with low values ($H^2 < 0.119$) were significantly enriched for respiratory burst involved in defense response and RNA methylation processes. These processes may not be an accurate representation of genes that are less repeatable, as lower expressed genes that were not expressed in all 40 clonal individuals tend to have lower repeatabilities that is likely due to the inconsistent expression profiles between pairs. To account for this, we repeated the analysis with only genes that have a mean TPM of at least 10 and found that they were enriched for other significant GO terms as well, such as ribosome biogenesis and ethylene biosynthesis. Finally, to assess the repeatability of genes involved in xylogenesis, we calculated the average heritability of lignin (H^2 =

0.425), cellulose (H^2 = 0.575) and xylan (H^2 = 0.499) genes. The average heritability of all three sets of genes was calculated as 0.512, suggesting that variation in expression of these genes is associated with genetic and environmental effects.

2.4.3 Genetic linkage map construction from transcriptome data

To perform eQTL mapping on the three- and eight-year-old population using the same genetic map, we constructed a new genetic framework map by extracting SNPs in highly expressed genes from transcriptome data (**Table 2.2; Figure 2.2**). The map consisted of 236 markers distributed over 11 linkage groups, spanning 1,008 cM with an average interval size of 4.5 cM. The genome coverage (p) of the map was estimated as 99.1% with the equation $p = 1\text{-}e^{-2dn/L}$, where d is an average interval of 10 cM between markers, n is the number of markers, and L is the length of the linkage map in cM. The genetic position of each marker was compared to its physical position to ensure that all the markers were in the correct order within each linkage group (**Supplementary Figure 2.3**). A minimum marker site coverage of 10 was allowed and the average coverage across all markers was calculated as 346 reads per marker (Note: This is the down-sampled coverage reported by GATK that was used to calculate the genotype of each marker, not the actual read depth per marker which was much higher; **Supplementary File 2.3**). To test whether segregation deviated significantly from the expected 1:1 ratio, we performed a Chi-square test (χ^2) at a 0.05 level of significance (**Figure 2.3**). Of the 236 markers, 35 (14.8%) deviated significantly from the expected ratio in the three-year-old population and 21 (8.9%) deviated significantly from the expected ratio in the eight-year-old population. Noticeably, markers with significant segregation distortion tend to cluster together (i.e. linkage to the same segregation distortion factor), resulting in regions of distortion that we expect to see in interspecific hybrids. As described by Myburg *et al.* (2003), markers were rephased to allow the direction of distortion to indicate the presence or absence of the *E. grandis* maternal allele of the F_1 hybrid.

Table 2.2: Summary of the genetic linkage map

Linkage group	Markers	Length (cM)	Length (Mb)	Average gap (cM)
LG1	20	94.62	43.22	4.98
LG2	24	96.47	57.73	4.19
LG3	24	104.44	82.50	4.54
LG4	15	60.86	40.03	4.35
LG5	23	99.22	72.88	4.31
LG6	27	123.02	56.59	4.73
LG7	21	78.80	54.35	3.94
LG8	24	101.8	64.55	4.43
LG9	17	75.93	34.75	4.75
LG10	21	86.75	37.41	4.34
LG11	20	86.49	41.61	4.54
Average	**21**	**91.67**	**53.24**	**4.46**
Total	**236**	**1,008.4**	**585.62**	**-**

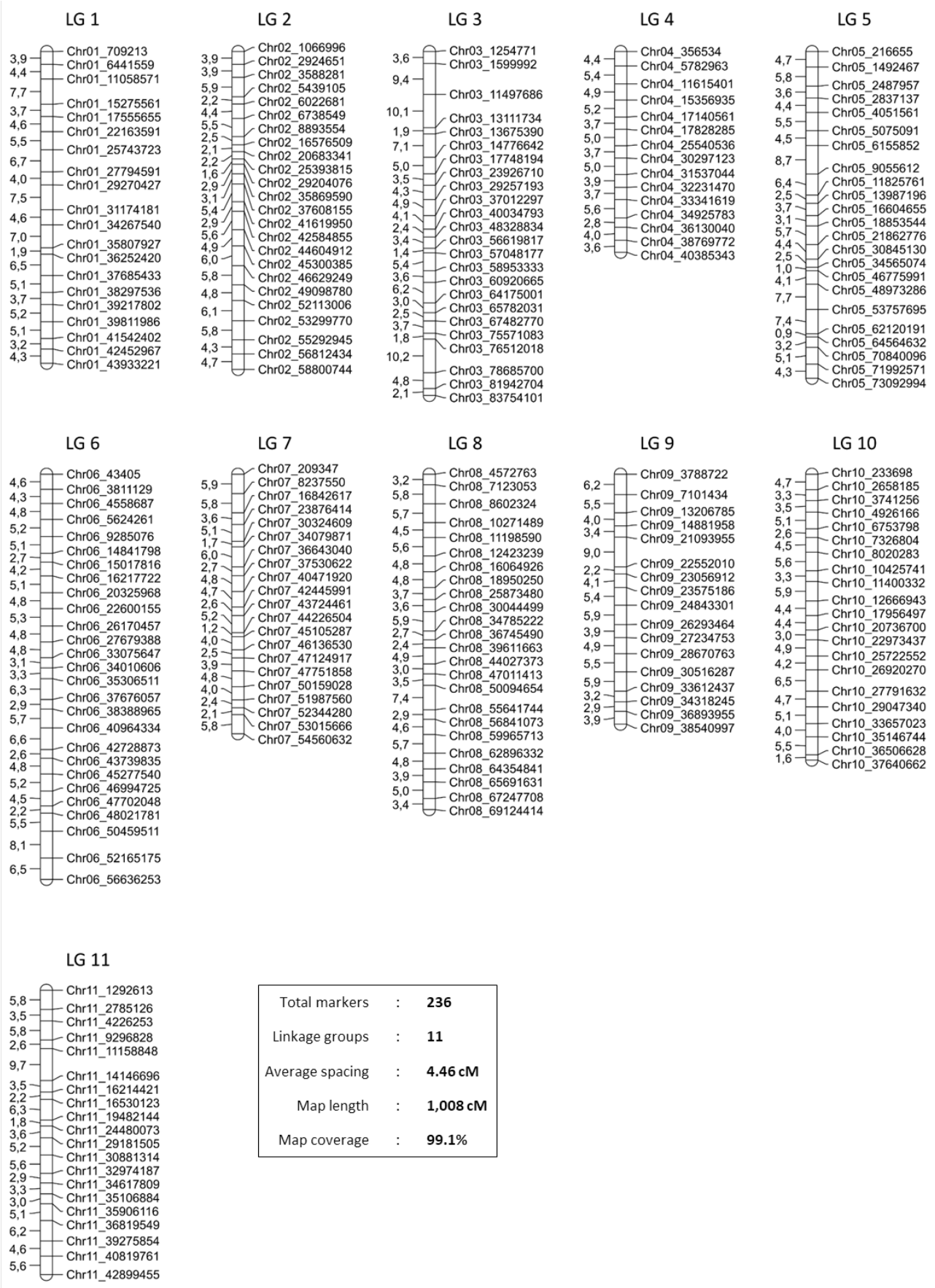


Figure 2.2: Genetic framework map for the *E. grandis* x *E. urophylla* F_1 hybrid parent. The map consists of 236 markers distributed over 11 linkage groups, spanning 1,008 cM with an average interval size of 4.5 cM. The map covers 99.1% of the genome at an average marker interval of 10 cM. Each bar represents a linkage group (chromosome), with the gap sizes between markers in cM (left) and marker positions in bp (right) (**Supplementary File 2.3**).

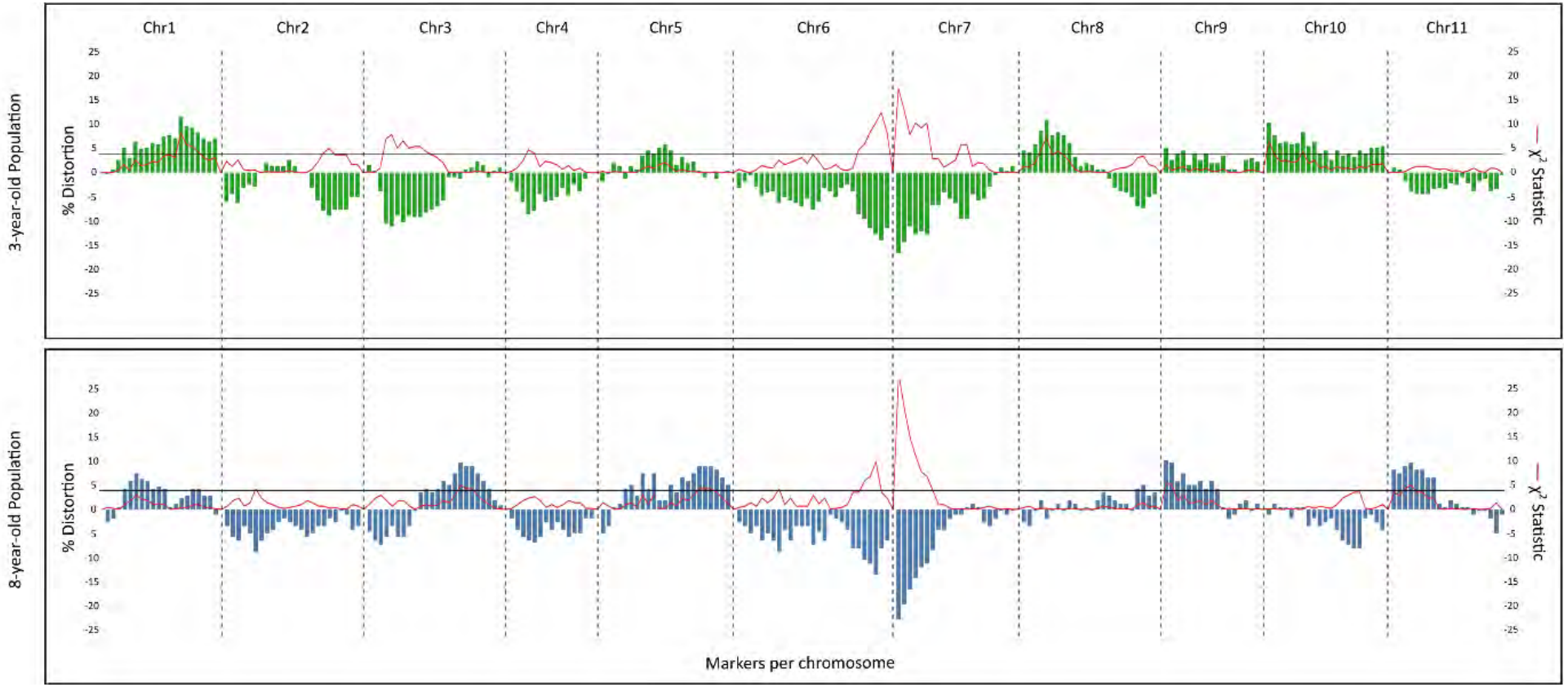


Figure 2.3: Segregation distortion observed in the genetic framework map for the three- and eight-year old population. Markers are indicated by bars and linkage groups (chromosomes) are indicated by dotted lines. The χ^2 test statistic is indicated by the red line and the critical value of the χ^2 test statistic at a 0.05 level of significance (3.841) is indicated by a horizontal black line. Significant distortion occurs where the χ^2 test statistic exceeds the critical value. Markers were rephased to represent the *E. grandis* allele based on the negative direction of distortion (**Supplementary File 2.3**).

2.4.4 Co-expression, co-regulation, and systems genetics analysis of gene expression during xylogenesis in mature trees

A weighted gene co-expression network analysis (WGCNA) was performed to assign highly correlated genes to gene co-expression modules. Genes expressed in the eight-year-old population were assigned to 54 co-expression modules and the average expression pattern of each co-expression module was represented by a module eigengene (Langfelder & Horvath, 2007). The average expression patterns were used to calculate the correlation between modules and visualise the largest cluster (absolute correlation > 0.5; p-value < 0.00001) of the co-expression network as a module eigengene network (**Figure 2.4**). GO enrichment analysis was performed for each module to identify the biological processes represented in the network, with 65% of modules having significant enrichment for GO terms (**Supplementary File 2.5**). Several of the highly correlated modules were significantly enriched for processes involved in secondary cell wall formation, indicating that it was important for developmental genes to be co-expressed at this age. Some highly correlated modules were also significantly enriched for stress and defense responses, which showed that these trees were not only investing energy into regulating growth and development, but also regulating xylem gene expression in response to environmental stresses that can influence growth and mortality.

A global eQTL analysis identified 29,860 eQTLs in the eight-year-old population, of which 6,743 (23%) were classified as *cis*-eQTLs and 21,662 (73%) were classified as *trans*-eQTLs. We identified 22 *trans*-eQTL hotspots (**Table 2.3**), of which 13 were significantly enriched for GO terms (**Supplementary File 2.5**). Within each hotspot region, we were able to identify possible candidate regulators by identifying transcription factors with *cis*-eQTLs (**Supplementary File 2.6**). If we consider highly correlated co-expression models to be a phenotype representing gene regulation that was achieved despite genetic and environmental variation, then we can ask whether genetic variation acts as a mechanism driving the co-expression and co-regulation of these models. By combining the

results from our co-expression and eQTL analyses into systems genetics models, we identified a possible genetic basis for coordinated expression responses regulating xylem biological processes. **Figure 2.5** shows an example of such a model, where the association between three *trans*-eQTL hotspots and gene co-expression modules is illustrated in an integrated network. Two modules (brown and blue) are highlighted, as there was a significant number of genes overlapping between these modules and the hotspots. The groups of overlapping genes were significantly enriched for defense responses in the brown module and regulation of G2/M transition of the mitotic cell cycle in the blue module.

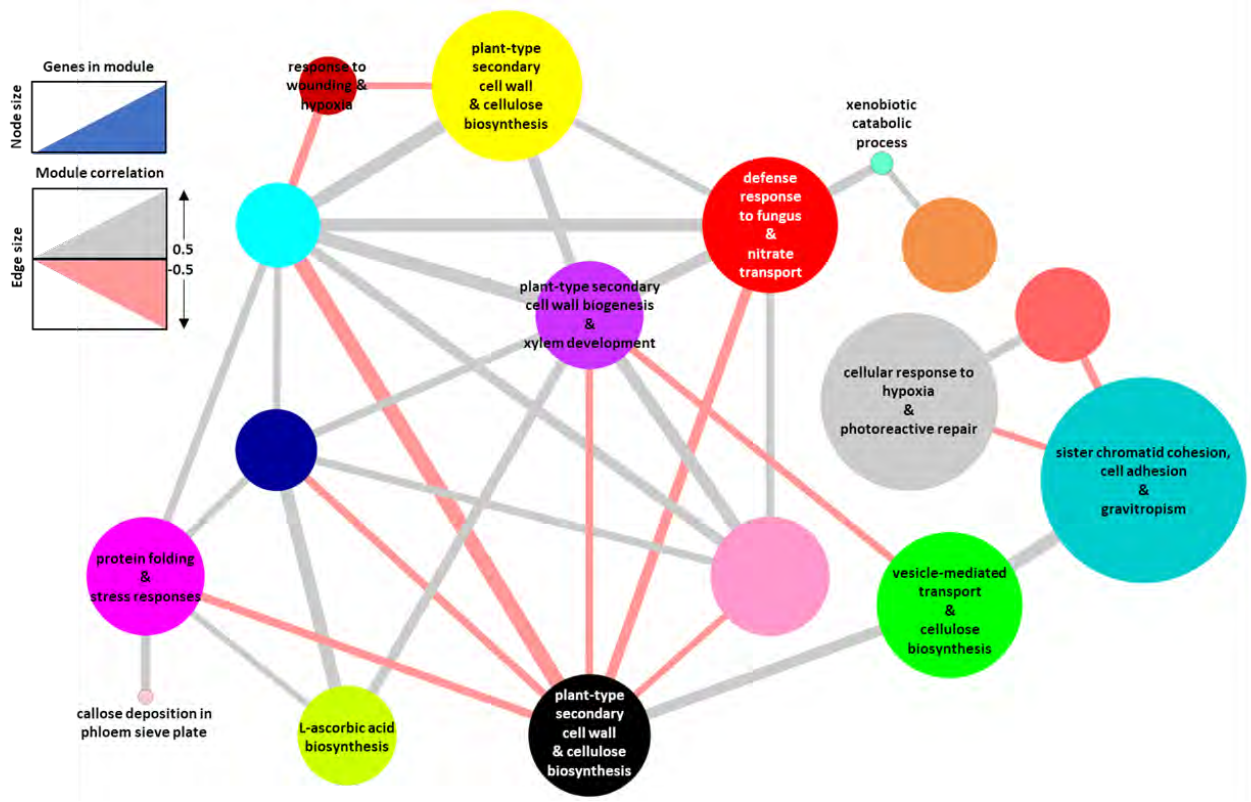


Figure 2.4: Module eigengene network for the eight-year-old population. Biological processes involved in xylogenesis are highlighted. Only the biggest cluster of correlated gene modules is shown (absolute correlation > 0.5), where nodes represent gene modules, node size represents the number of genes in a module, edges represent correlations between modules, edge width represents the absolute correlation value, and edge colour represents the correlation sign.

We further divided the eQTL hotspots into so called "split"-hotspots, based on the direction of their additive effects (i.e. whether higher transcript abundance was associated with the *E. grandis* or *E. urophylla* parental allele) (**Supplementary Figure 2.4**). We observed that some hotspots had strong

hotspot-module overlaps for genes in the same modules that were associated with eQTLs of opposite additive effects. To examine the associations between split-hotspots and modules involved in xylogenesis, we analysed the association between three secondary cell wall-related gene modules (black, purple and green) and hotspots with which at least two of the modules had a significant number of overlapping genes (**Figure 2.6; Supplementary File 2.5**). The green module had a group of highly correlated genes forming a tight cluster within the module network, for which the higher expression levels were associated with the *E. urophylla* parental allele in HS_10.37 and the *E. grandis* parental allele in HS_3.72. This cluster of genes was significantly enriched for vesicle-mediated transport and pollen germination and was highly correlated with a group of genes in the black module that were also associated with the *E. urophylla* parental allelic effect in HS_10.37. For HS_10.37 the two split-hotspots were significantly enriched for different biological processes.

Table 2.3: Summary of *trans*-eQTL hotspots in eight-year-old population

Hotspot name	Chr	Average peak LOD[a]	Average R^2	Significant GO-terms[b]	*trans*-eQTLs in hotspot	Genes in hotspot region	*cis*-eQTLs in hotspot region	TF genes[c] with *cis*-eQTLs
HS_1.39	1	3.19	0.10	4	344	57	13	1
HS_1.44	1	3.22	0.10	1	362	122	34	2
HS_2.44	2	4.11	0.16	0	196	19	1	0
HS_2.54	2	3.32	0.11	0	194	55	14	0
HS_2.59	2	3.31	0.11	16	1,494	367	108	7
HS_3.5	3	3.60	0.11	0	297	72	16	0
HS_3.11	3	3.62	0.12	3	296	79	35	3
HS_3.37	3	3.47	0.12	1	457	176	60	2
HS_3.58	3	3.69	0.12	6	385	225	81	6
HS_3.72	3	3.61	0.12	4	1,656	225	78	4
HS_6.51	6	3.52	0.12	0	191	89	19	1
HS_7.35	7	3.51	0.11	0	229	117	31	4
HS_7.45	7	3.80	0.13	2	277	51	8	2
HS_7.53	7	3.24	0.11	35	610	200	51	1
HS_9.25	9	3.27	0.11	0	500	189	49	1
HS_9.28	9	3.42	0.12	0	227	85	21	0
HS_9.37	9	3.41	0.12	8	915	346	91	8
HS_10.37	10	3.41	0.11	1	672	123	36	1
HS_11.36	11	3.51	0.11	5	933	247	72	2
HS_11.43	11	3.45	0.12	3	422	187	50	1
Total					**10,657**	**3,031**	**868**	**46**

[a] LOD = LR x 0.217

[b] Refer to Supplementary File 2.5

[c] Transcription factors in *trans*-eQTL hotspot region

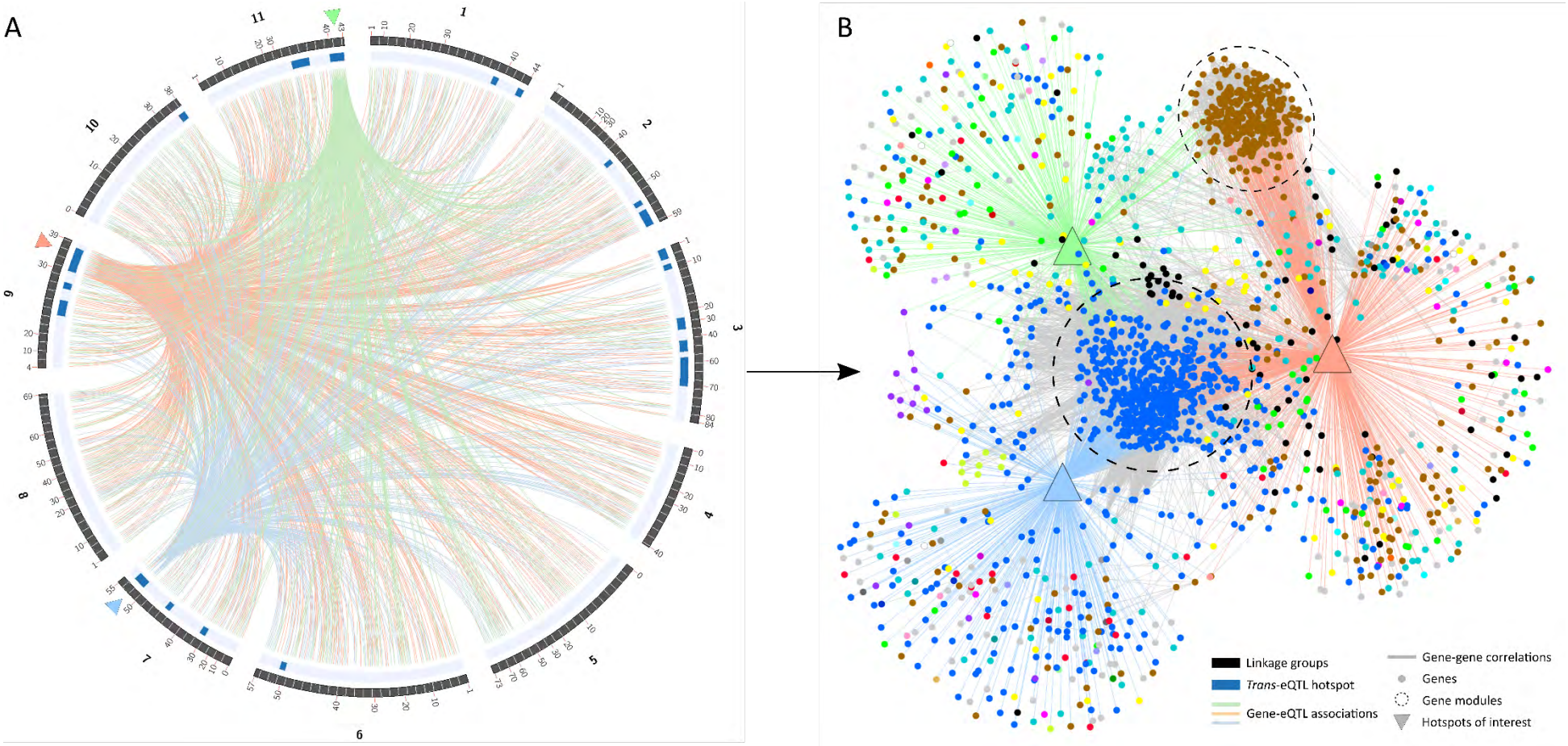


Figure 2.5: Integrated systems genetics model showing the association between eQTL hotspots and gene co-expression modules for xylem expressed genes in the eight-year-old population. Three hotspots of interest are visualised to demonstrate how combining results from eQTL and co-expression analyses into a systems genetics model allows the identification of a genetic basis for coordinated expression responses regulating xylem biological processes. **A**) Circos plot showing the gene-eQTL associations for each hotspot. **B**) Integrated network showing how genes in hotspots overlap with genes in the same module (module membership = colour). The brown and blue modules are highlighted, for which the hotspot overlapping genes were significantly enriched for defense responses and regulation of G2/M transition of the mitotic cell cycle respectively. Refer to **Supplementary File 2.5** for all hotspot-module overlaps and GO enrichment results.

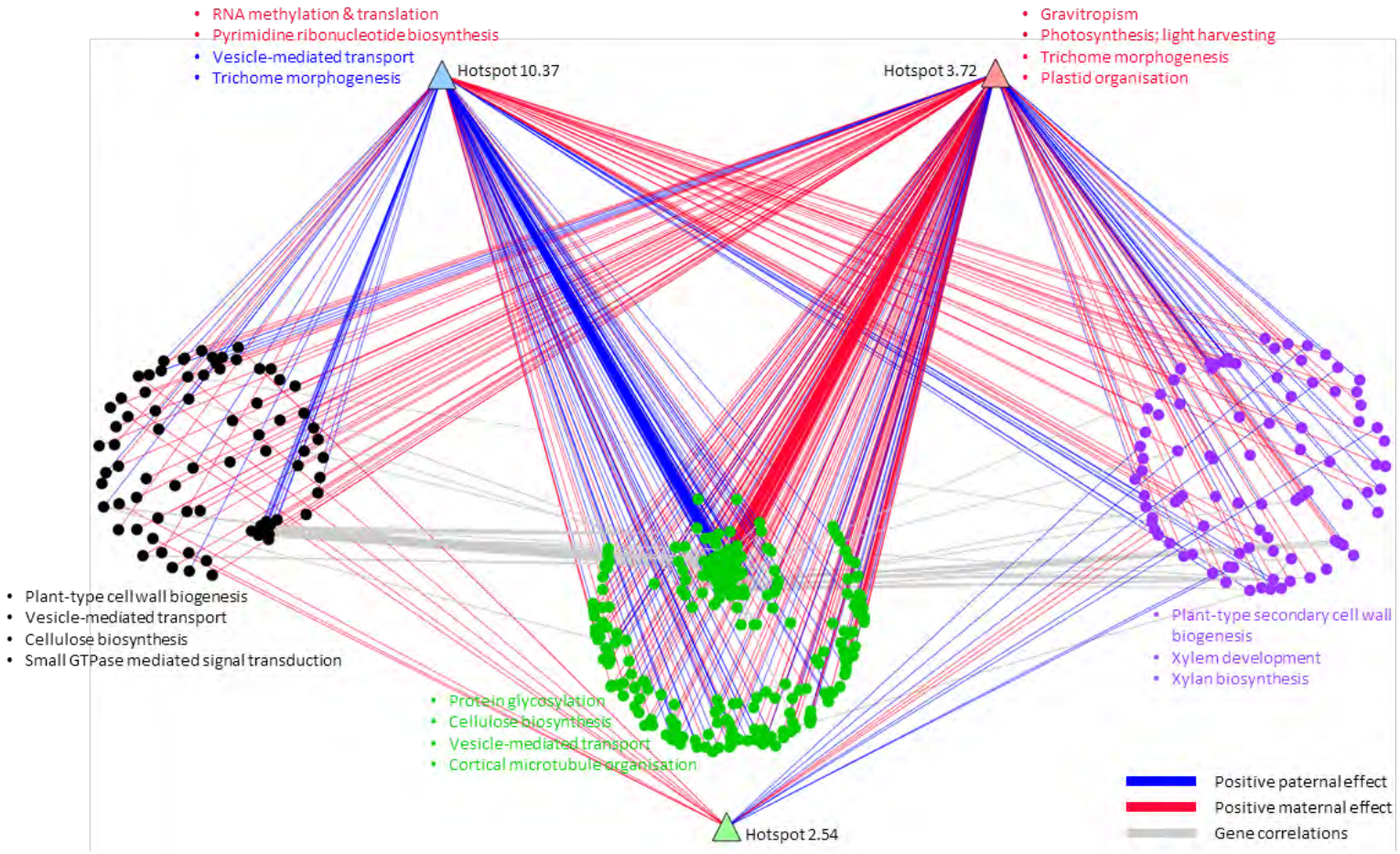


Figure 2.6: Association of secondary cell wall-related modules with split-hotspots. The network shows the association of three secondary cell wall-related gene modules (black, green, and purple circles) with three significantly overlapping split-hotspots (hotspots 2, 5, and 1 on chromosomes 2, 3, and 10 respectively). Grey lines indicate absolute correlations > 0.8 between genes and genes within each module were clustered based on their correlation values. Module-hotspot overlap significance was determined with a Fisher's exact test and significant GO enrichment terms for these overlapping genes are listed, with the text colour corresponding to the colour of the module or split-hotspot. Hotspot 2.54 did not have any significant enrichment, but may still contain important regulatory genes. For split-hotspots, higher transcript abundance associated with the *E. urophylla* allele is represented by blue lines and for the *E. grandis* allele by red lines. Refer to **Supplementary File 2.5** for all hotspot-module overlaps and GO enrichment.

2.4.5 Age-to-age comparison of the genetic architecture of xylem gene expression

To determine the age-to-age conservation of absolute gene expression levels, a Pearson correlation of genes expressed in both the three- and eight-year-old populations was calculated as 0.78. Since we were looking specifically at xylogenesis, we wanted to know whether wood formation genes were highly correlated in their expression between the three- and eight-year-old population. To determine the age-to-age correlation of absolute xylem gene expression values, six groups of genes (Myburg *et al.*, 2014; Ployet *et al.*, 2019; Christie *et al.*, in preparation) were chosen from both populations: (i) secondary cell wall genes (142); (ii) cellulose synthase (CesA) genes (21); (iii) bona fide lignin genes (18); (iv) cellulose/xylan pathway genes (113); (v) shikimate/lignin/flavonoid pathway genes (71); and (vi) the top 100 highest expressed genes in each population (130). For each gene, the mean transcript abundance was compared between the two ages and a Pearson correlation was calculated per group (**Figure 2.7**). The expression of genes involved in xylogenesis (group i-v) remained relatively conserved across age ($r \geq 0.9$), whereas the top expressed genes varied more ($r = 0.52$). To determine whether the relative expression of these genes was conserved across age (e.g. two genes that were most correlated at three years were still most correlated at eight years). Spearman rank correlations of transcript abundance were compared between the three- and eight-year-old population (**Supplementary Figure 2.5**). The correlations for gene groups ranged from 0.59 to 0.85, with a median of 0.64 and a mean of 0.68, suggesting that these relationships were not highly conserved and that gene co-expression patterns may vary across age.

We also wanted to know whether genes, for which the relative expressions remained conserved across age, fell within the same co-expression modules. To do this, we performed a WGCNA on both populations to analyse the age-to-age changes in gene co-expression patterns. For the three-year-old population, 24,861 genes were assigned to 101 co-expression modules (min: 10 genes; max: 7,091 genes), whereas for the eight-year-old population, 25,267 genes were assigned to 54 co-expression modules (min: 16 genes; max: 6,946 genes) using the same parameters (**Supplementary Figure 2.6**).

Both networks had good scale-free topology fit where a soft thresholding power of four was selected (three-year-old population: $R^2 > 0.85$; eight-year-old population: $R^2 > 0.90$). For each gene, we identified its nearest neighbour (i.e. the same two genes were nearest neighbours at both ages). The proportion of gene pairs that were conserved across age was 8% for genes expressed in both populations (**Supplementary File 2.6**). For these gene pairs that remained, the age-to-age changes in correlation ranged from 0.0002 to 0.519. We also determined the proportion of gene pairs occurring together within the same co-expression module. In the three-year-old population, 65% of gene pairs fell within the same co-expression module and in the eight-year-old population, 71% fell within the same co-expression module.

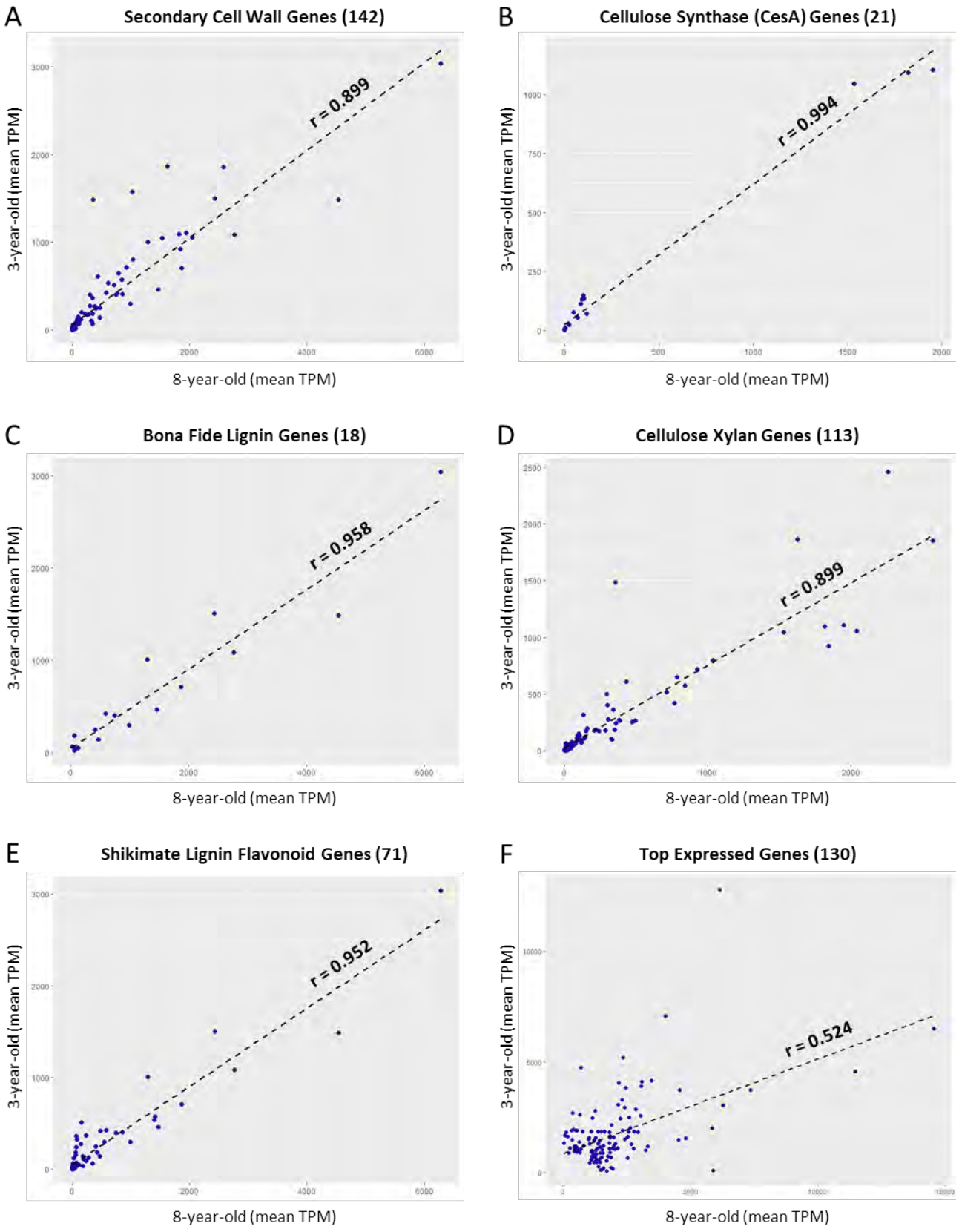


Figure 2.7: Age-to-age correlation of mean expression values for six different gene groups. A) Genes involved in secondary cell wall processes, **B**) CesA genes, **C**) bona fide lignin genes, **D**) genes involved in cellulose/xylan pathways, **E**) genes involved in shikimate/lignin/flavonoid pathways, and **F**) the top 100 expressed genes in each population.

Next, we considered the role of genetic variation (i.e. eQTLs) on the conservation of gene expression patterns and correlations (i.e. co-expression networks). A global eQTL analysis was performed on the three- and eight-year-old populations to determine which genes have variation in their expression patterns that can be associated with variation in the genotype (**Table 2.4**; **Supplementary Table 2.1)**. For genes with high enough levels of expression (i.e. TPM > 0 in 25% or more of the population), 70.6% and 72.5% had eQTLs in the three- and eight-year-old population, respectively. The three-year-old population have a higher proportion of *cis*-eQTLs and lower proportion of *trans*-eQTLs relative to the eight-year-old population. An average of 1.5 *trans*-eQTLs per gene was observed in both populations, with some genes having up to five or six *trans*-eQTLs. Genes in the eight-year-old population had more *trans*-eQTLs per gene than those in the three-year-old population (**Supplementary Figure 2.7**). Overall, *cis*-eQTLs explained about 18% of the variation observed in the transcriptome, whereas *trans*-eQTLs explained only about 9% for genes that had these eQTLs.

Table 2.4: Summary of eQTLs in the three-and eight-year-old population

Population	Three-year-old	Eight-year-old
Total number of eQTLs	27,284	29,860
Average peak LR	30.3	20.8
Average R^2	0.14	0.15
Number of *cis*-eQTLs	8,960 (33%)	6,743 (23%)
Average peak LR (*cis*-eQTLs)	53.2	35
Average R^2 (*cis*-eQTLs)	0.23	0.23
Number of *trans*-eQTLs	17,063 (63%)	21,662 (73%)
Average peak LR (*trans*-eQTLs)	17.4	16
Average R^2 (*trans*-eQTLs)	0.08	0.12

The genome-wide location of genes and eQTLs were compared between the two age groups to provide an overview of the genetic architecture of xylem transcript abundance and how it changed over time (**Figure 2.8**). Expressed gene and *cis*-eQTL densities remained fairly conserved across the two ages, but there were many instances where the *trans*-eQTL densities varied significantly, resulting in different *trans*-eQTL hotspots between the two ages (**Supplementary Figure 2.8; Supplementary Figure 2.9**). To determine whether the conserved regions represented the same

eQTLs, an eQTL overlap score was calculated for each gene that had an eQTL at both ages on the same chromosome, based on the distance between their peaks (**Supplementary File 2.4**). As expected, *cis*-eQTLs tended to be more conserved with higher overlap scores than *trans*-eQTLs (**Supplementary Figure 2.10; Supplementary Figure 2.11**).

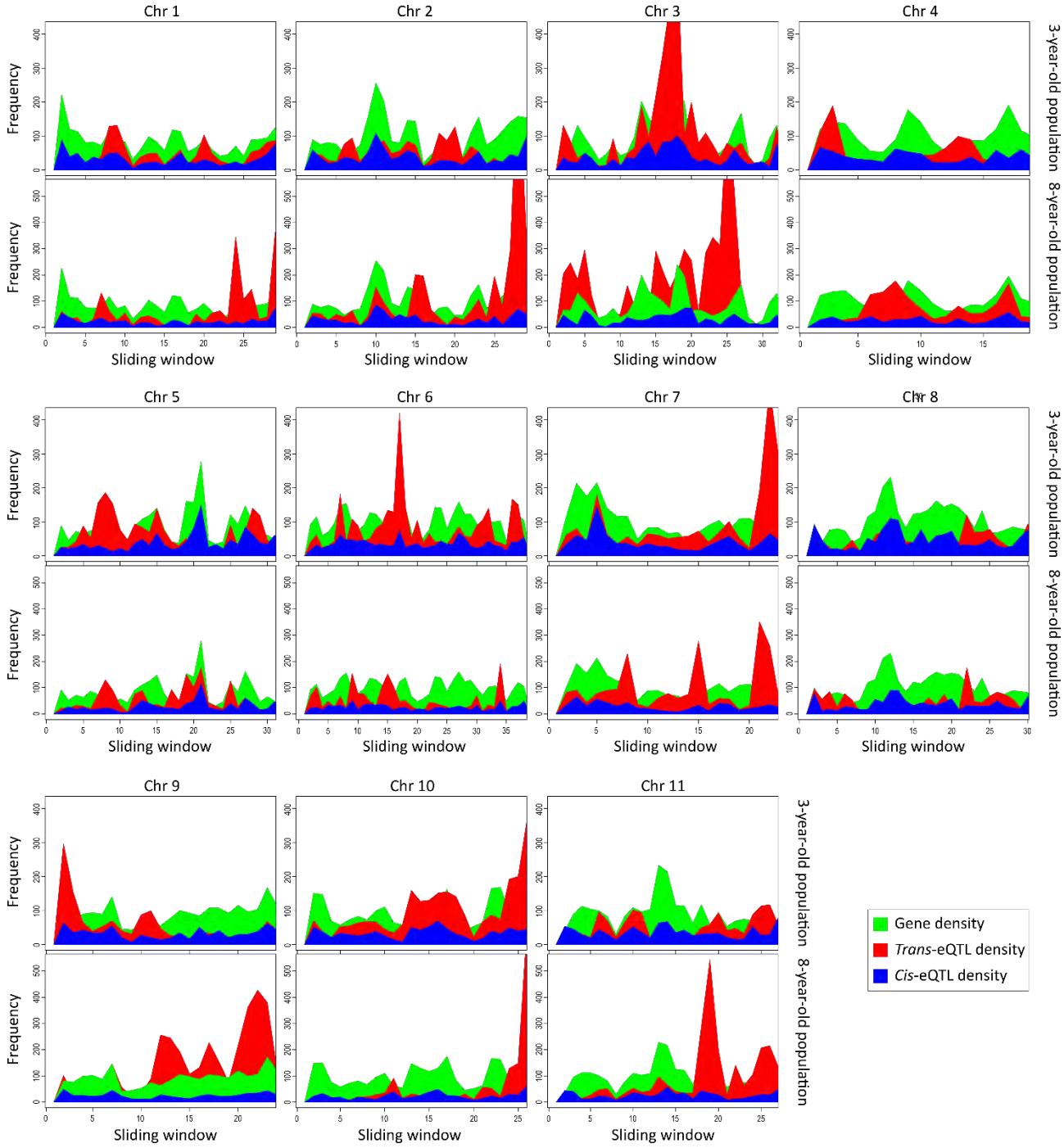


Figure 2.8: Frequency plots showing the density of xylem expressed genes and eQTLs at different ages. A genome-wide overview of gene density (green), *cis*-eQTL density (blue) and *trans*-eQTL density (red) per chromosome is visualised to give us an idea of the changes in the genetic architecture across age. Each chromosome was divided into sliding window bins based on the genetic map.

To study the genetic architecture of genes directly involved in xylogenesis, we analysed the gene module membership and shared eQTLs of genes (with average TPM > 10) involved in the general phenylpropanoid (18), cellulose (26), and xylan (58) biosynthesis pathways between the three- and eight-year-old populations. Module colours were assigned arbitrarily and did not correspond between the two populations (figures only show whether the same genes were correlated at both ages). **Figure 2.9** shows that most lignin genes were co-expressed in the three-year-old population (black module), whereas in the eight-year-old population these genes were divided into two co-expression modules (tan and purple) that were not highly correlated with each other based on their module eigengenes (*r* = -0.22). Some genes shared eQTLs with each other at both ages, however, from our overlap analysis we determined that it was not the same eQTL that had been conserved across age. The same dramatic shift in the genetic architecture was seen for cellulose genes (**Supplementary Figure 2.12**) and xylan genes (**Supplementary Figure 2.13**), however, some shared eQTLs between xylan genes did remain conserved across age.

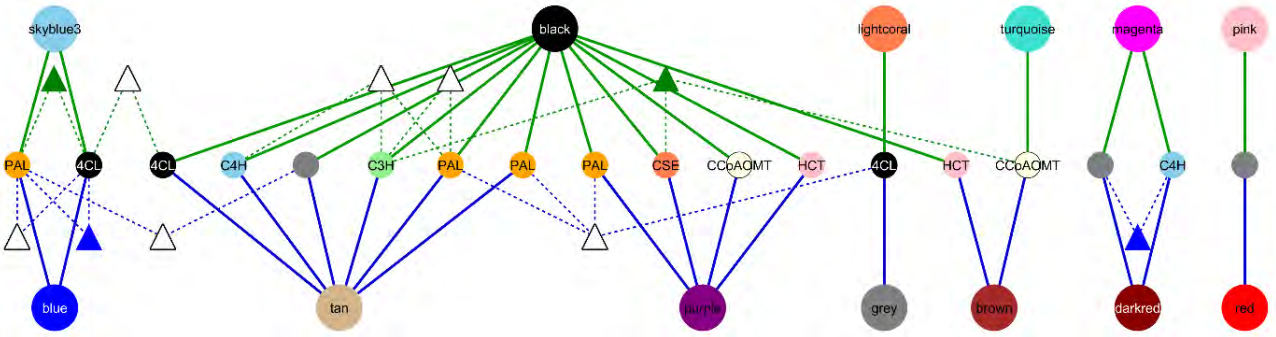


Figure 2.9: Genetic architecture of lignin genes across age. Gene module memberships (larger circles) and shared eQTLs (triangles) are shown for lignin genes (smaller circles) between three-year-old individuals (green edges) and eight-year-old individuals (blue edges). Genes are labelled and coloured according to their enzyme names in the general phenylpropanoid pathway. Coloured triangles indicate that the shared eQTLs fall within *trans*-eQTL hotspots for some of the genes.

It is important to consider that, apart from age, other factors may be involved in the differences seen in the genetic architectures of our two populations. The eight-year-old population was sampled after a drought period of around 24 months (years six and seven), which led us to believe that stress response may have had a significant effect on the regulation of developmental genes in this

population. Ployet *et al.*, (2019) identified a set of genes that are potentially involved in secondary cell wall remodelling in response to abiotic stress. We compared the genetic architecture of 15 of these known transcription factor genes between the three- and eight-year-old population to show how their regulation changed across age under stressful conditions (**Figure 2.10**). We also tested whether these genes were significantly differentially expressed between the two populations and found that 14 of the 15 genes were up-regulated in the eight-year-old population, with a median log2 fold change of 1.9.

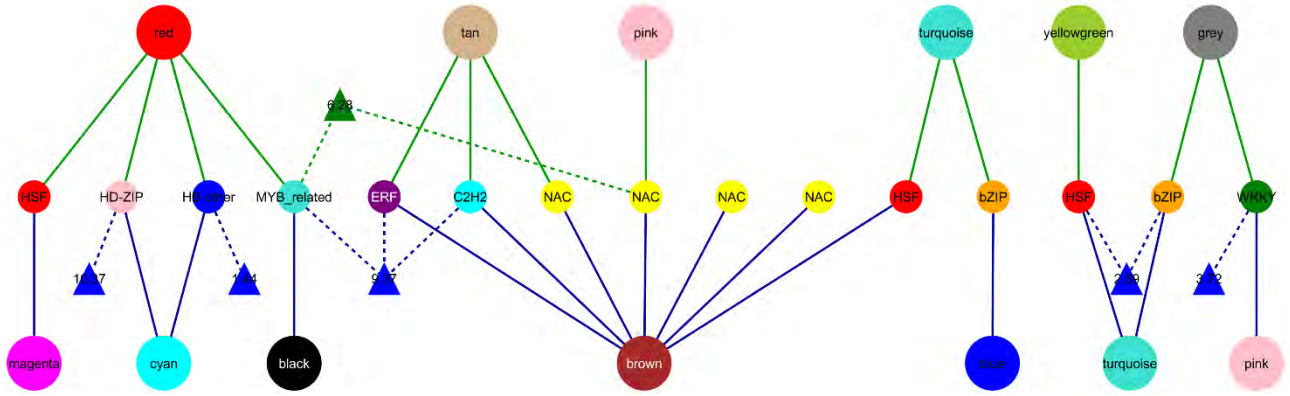


Figure 2.10: Genetic architecture of transcript**ion factor genes associated with response to abiotic stress.** Gene module membership (larger circles) and eQTL hotspot membership (triangles) is shown for transcription factor genes associated with response to abiotic stress (smaller circles) between three-year-old individuals (green edges) and eight-year-old individuals (blue edges).

To dissect how genetic variation affects co-expression of genes and therefore contributes to the observed shift in the genetic architecture, we first had to determine what biological functions were gained or lost between juvenile and mature trees in terms of only gene co-expression (i.e. independent of eQTLs). We performed a GO enrichment analysis on genes that were unique to each population and genes that were significantly differentially expressed with a log2 fold-change > 2 (**Supplementary File 2.5**): (i) genes unique to the three-year-old population were enriched for defense response, protein phosphorylation and signal transduction; (ii) genes unique to the eight-year-old population were enriched for response to ethylene, protein phosphorylation and salicylic acid biosynthesis; (iii) genes up-regulated in the eight-year-old population were enriched for response to chitin, ethylene biosynthesis and protein folding; and (iv) genes down-regulated in the eight-year-old

population were enriched for defense response, photosynthesis and cell proliferation. We performed a Fisher's exact test on gene co-expression modules and module GO terms between the three- and eight-year-old population to identify which modules and processes were either conserved across age or were unique to each developmental stage (**Supplementary File 2.5**). We found that GO terms for the eight-year-old magenta module (553 genes), which was enriched for stress responses, significantly overlaps with GO terms for the yellowgreen (76 genes) and darkgreen (113 genes) three-year-old modules. Similarly, GO terms for the eight-year-old brown module (3,087 genes), which was enriched for defense responses, significantly overlapped with GO terms for the tan (295 genes), darkorange (105 genes) and plum (37 genes) modules in the three-year-old population. Two of the new hotspots that were present in the eight-year-old population, HS_9.37 and HS_11.36, were significantly enriched for defense and stress responses respectively (**Supplementary File 2.5**). Interestingly, these hotspots had a significant number of genes that respectively overlap with the brown and magenta modules enriched for the same biological functions. Out of the 117 genes that respond to abiotic stress (Ployet *et al.*, 2019), 14 had *trans*-eQTLs that fall within HS_11.36 (**Supplementary File 2.6**).

2.5 Discussion

The aim of this study was to characterise the genetic architecture of xylem gene expression during wood formation at rotation age in a full-sib family of *E. grandis* x *E. urophylla* F_2 hybrid backcross trees and to perform a comparative analysis of the genetic architecture between juvenile and mature trees. For the first time in one of our studies, we had access to a clonal mapping population that allowed us to measure the heritability of transcriptome profiles. High broad-sense heritability estimates suggest that genetic variation contributes substantially to the phenotypic variation observed in transcript levels. This was also the first time where we made use of population-wide transcriptome-derived SNPs to construct a robust genetic linkage map that could be used for eQTL mapping in juvenile and mature trees. This allowed for a more accurate and direct comparison of the genetic

architecture of gene expression between the two ages, based on xylem genes expressed at both ages. We showed that by combining the results from co-expression and eQTL analyses into systems genetics models, it is possible to identify a potential genetic basis for coordinated expression responses regulating xylem biological processes at population level. Finally, we were able to perform an age-to-age comparison of the absolute and relative transcript abundances of xylem expressed genes and investigate the conservation of their genetic architecture across age. We discovered that there was a major shift in the genetic architecture for genes involved in xylogenesis, which may be a result of transcriptional rewiring due to developmental stage or age, stress responses, other unknown biological factors, or likely a combination of the above.

Before any major analyses could be performed, we identified and removed problematic samples (transcriptomes) that could influence the accuracy of downstream results (**Supplementary Note 2.1**). This resulted in a smaller sample set that was available for the eight-year-old population (100 samples) compared to the three-year-old population (156 samples), which decreased the statistical power of our analyses. Due to this, it is also possible that the WGCNA method could not separate some highly correlated modules well enough in the eight-year-old population, as a higher number of samples allows for better cluster separation. We have recently expanded the transcriptome sequencing to have equivalent amounts of data, but there was not enough time to include these results in this manuscript. It will, however, be included in the scientific paper that we plan to publish. For this study, we only mapped eQTLs for one of the two parents (the F_1 hybrid), as we were mostly interested in the interspecific differentiation between the *E. grandis* or *E. urophylla* haplotypes in the F_1 hybrid that are segregating in the progeny. This means that only half of the potential genetic variation was explored in this study and that these analyses must be expanded to include the *E. urophylla* parent as well, in order to reflect the full genetic architecture. The major changes we observe in the genetic architecture between the two populations may be due to at least four contributing factors that we cannot rule out: (i) development or age-related changes; (ii) stress related changes; (iii) unknown

biological reasons; and (iv) limited statistical power. For the first three, the loss or gain of an eQTL could be explained by different polymorphic transcription factors or repressors affecting gene expression, whereas for the fourth, the eQTL signal might be just above or below the detection threshold in three-year-old vs. eight-year-old trees. We can possibly start to address this by asking what was conserved across age in terms of gene co-expression and co-regulation (eQTLs). If the genetic variation remains the same (typically *cis*-eQTLs), the polymorphism still has the same knock-on effect on the expression of target genes, and we can assume that these genes are probably not developmental stage dependent, although their absolute expression levels can be higher or lower in juvenile vs. mature trees.

Heritability of xylem transcriptome profiles in *Eucalyptus*

One of the most persistent questions that geneticists have struggled with is understanding how phenotypic variation is affected by variation in the genotype (Boyle *et al.*, 2018). To estimate the proportion of xylem transcript level variation that was affected by genetic variation, we calculated the broad-sense heritability, also estimated by the transcriptome repeatability, across 25,307 xylem expressed genes for 20 pairs of clonal replicates planted in a common garden. The distribution of H^2 values for transcript-level variation ranged from -0.60 to 0.99 (**Supplementary Figure 2.2**), suggesting that some genes are highly affected by environmental factors, whereas others are highly heritable and therefore mainly affected by genotype. According to Steinsaltz *et al.* (2017), negative heritability should not be ignored, as it could suggest that individuals with similar genotypes are likely to have more trait variation than those with very different genotypes. Estimating the heritability allows us to discriminate between genetic and environmental effects on transcript abundance, which is in line with the results from previous studies that could detect genetic control of gene expression levels through genetical genomics approaches in *Eucalyptus* (Kirst *et al.*, 2004; Kirst *et al.*, 2005; Kirst and Yu, 2007) and *Populus* (Drost *et al.*, 2010; Mähler *et al.*, 2017). The average heritability estimates for lignin ($H^2 = 0.425$), cellulose ($H^2 = 0.575$), and xylan ($H^2 = 0.499$) genes are in line with

results from a recent study by Vaillant *et al.* (2018), where the authors used a variance component ratio (VCR) based on the concept of broad-sense heritability to evaluate the genetic effect in leaf and xylem transcriptome variation among *E. urophylla x E. grandis* hybrids. The authors found that the mean VCR for phenylpropanoid (lignin) genes was higher in leaf tissue (VCR = 0.37) than in xylem tissue (VCR = 0.28), whereas the mean VCR for cellulose and xylan genes were higher in xylem tissue (VCR = 0.40) than in leaf tissue (VCR = 0.29). The authors conclude that the results suggest that the genetic control of gene expression within these pathways is not associated with the genes that constitute the pathways, but that the control is related to regulatory genes. It is important to show significant heritability for transcript abundance, as the statistical power that is needed to detect genetic variants affecting gene expression (eQTLs) is dependent on heritability (Visscher *et al.*, 2008).

Simultaneous genetic mapping and transcriptome profiling

We wanted to use a single genetic map for mapping eQTLs in the two populations (three- and eight-year-old backcross progeny from the same cross) in order to avoid as much technical variation as possible, which would allow for a more accurate and direct comparison of the genetic architecture of xylem expressed genes. For the three-year-old samples, we previously only had access to DArT markers (Kullan *et al.*, 2012), and for the eight-year-old population, we had SNP chip genotypes generated from genomic DNA for one clonal ramet of each genotype using the EuCHIP60K SNP chip (Silva-Junior *et al.*, 2015). However, in some cases the SNP data was from a different ramet to what was used for RNA collection, which created the possibility of mismatches between genotypes and transcriptomes. As we had RNA-seq data available for both populations, we could simply extract robust SNP genotypes from transcriptome data of highly expressed genes to construct a new genetic linkage map (**Figure 2.2**). This allowed us to have a 1:1 correspondence between the genetic map and the transcriptome. Requirements such as a minimum coverage of 10 (per individual) and a minimum call rate of 0.9 per marker ensured that the map was based on high confidence SNP calls within highly expressed genes (Engelbrecht *et al.*, unpublished; **Supplementary File 2.3**). Recent

studies that have applied this method of constructing a genetic linkage map from transcriptome-derived SNPs were done by Shearman *et al.* (2015) in rubber trees, Galpaz *et al.* (2018) in melons, Santos *et al.* (2018) in peas, and Singh *et al.* (2018) in potato beans. This transcriptome-derived genetic map, together with existing bioinformatics pipelines for eQTL data analysis (Christie *et al.*, 2017) and co-regulation analysis (Christie *et al*., in preparation), allowed us to propose a method for rapid genetic dissection of gene expression variation from population-wide transcriptome data alone.

Systems genetics modelling of xylem development at rotation age

To characterise the genetic architecture of xylem gene expression at rotation age, we studied the relationship between genetic variation (e.g. *trans*-eQTL hotspots) and the co-expression of genes in modules. We know that, in the absence of genetic variation (e.g. in clonal or inbred populations), the transcriptional network will regulate expression modules in response to development and environmental factors. When a layer of genetic variation is added, genes are perturbed ("pulled out") from their normal expression patterns, affecting biological processes and networks that may be subject to natural selection. One such example is demonstrated in a study by Mähler *et al.* (2017) on *Populus* leaf buds, where the authors found that the key mechanism shaping the genetic architecture of gene expression variation is purifying selection and that the strength of this selection influences co-expression network connectivity. However, because our population consisted of a novel genetic construct (F_1 hybrid of non-overlapping species) that does not exist in nature, we are studying much more genetic variation than would normally be present in natural populations. Combinations of genes or expression patterns that are deleterious in this cross would have died early due to natural selection, as is expected for interspecific backcross families. Thus, the co-expression modules that we observed were those that could exist (i.e. permissible) despite the large amount of genetic variation in the population (**Figure 2.4**). Combining results from our eQTL and co-expression analyses into systems genetics models therefore allows us to identify a genetic basis for coordinated expression responses regulating xylem biological processes.

Age-to-age comparison of transcript abundance and genetic architecture

To determine if wood formation genes were highly correlated in their expression between the three- and eight-year-old population, we analysed the age-to-age correlation of absolute (**Figure 2.7)** and relative (**Supplementary Figure 2.5**) xylem transcript abundance. Absolute values remained relatively conserved across age ($r \geq 0.9$), whereas relative values were less conserved ($0.59 \leq r \leq 0.85$), suggesting that, even though these genes were still expressed at the same levels in juvenile and mature trees, there are differences in the way in which their co-expression was regulated across individuals of the population. To further analyse the degree of conservation of the genetic architecture of genes directly involved in xylogenesis, we compared the gene module membership, shared *trans*-eQTLs and *trans*-eQTL hotspot membership of genes involved in the lignin (**Figure 2.9)**, cellulose (**Supplementary Figure 2.12**) and xylan (**Supplementary Figure 2.13**) pathways between juvenile and mature trees. We found that the transcriptional regulation architecture, dominated by major regulatory perturbations, exhibited a large shift across age manifested by new *trans*-eQTL hotspots detected in the eight-year-old population (**Figure 2.8; Supplementary Figure 2.9**). So how was xylem development maintained with such a big shift in the genetic architecture in terms of *trans*-eQTL hotspots, which are considered major drivers of gene expression? Are there factors other than age/developmental stage that could explain the drastic shift in genetic architecture?

It is important to consider that, apart from age, other factors may be involved in the differences seen in the xylem genetic architectures of the two populations. The eight-year-old siblings were sampled after a severe drought period of around 24 months (years six and seven), which had led us to believe that stress responses may have had a significant effect on the regulation of developmental genes in this population. We expected to see some developmental changes, but due to the exceptional stress effect that affected the mature trees, we hypothesise that the shift was potentially due to stress-related transcription factors that were causing, or were regulated by, new *trans*-eQTL hotspots and thereby played a major role in xylogenesis in mature (stressed) trees (**Figure 2.10**). If such transcription

factors were segregating in juvenile trees, the polymorphisms may not have affected xylem gene expression much, because the trees were not stressed. However, in mature trees the polymorphisms (*E. grandis* or *E. urophylla* alleles), may have had a knock-on effect via the transcription factor on other stress-related genes. The dramatic shift in genetic architecture was possibly due to a combination of developmental and stress responses, as observed in a study by Taylor-Teeples *et al.* (2015), who showed that secondary cell wall gene regulation is tightly interwoven with responses to abiotic stress and that different stresses are capable of promoting functional adaptation by perturbing cell wall genes. Similarly, Ployet *et al.* (2019) implemented a network-based approach which showed that co-regulated genes and metabolite modules involved in wood formation play a fundamental role in the trade-off between production of biomass and response to stress.

To dissect how genetic variation affects co-expression of genes and therefore contributes to the observed shift in the genetic architecture, we first had to determine what biology was gained or lost between juvenile and mature trees in terms of gene co-expression (i.e. independent of eQTLs). We found that two relatively large modules in the eight-year-old trees (brown and magenta), which were enriched for stress and defense responses, significantly overlap in terms of biology with several smaller modules in the three-year-old population (**Supplementary File 2.5**). It is possible that these genes that were in different modules in juvenile trees, but were then co-expressed in mature trees, could have been subject to a new *trans*-eQTL that was coordinating their expression. If the gene that was subject to a new *trans*-eQTL was a regulator, the polymorphism will have had a knock-on effect on all the target genes, which could have created an eQTL hotspot. Interestingly, two of the new hotspots that arose in the eight-year-old population (HS_9.37 and HS_11.36) were biologically enriched for defense and stress responses, respectively, and a significant number of genes present in these hotspots overlapped with the two modules enriched for the same biological functions (**Supplementary File 2.5**). We were particularly interested in the dehydration responsive element binding protein gene (DREB2), which is an ethylene response factor (ERF) transcription factor that

plays an important role in response to drought-related stress. In the eight-year-old population, this gene was co-expressed in the brown module, which was enriched for response to chitin, and was associated with HS_3.97, which was enriched for ethylene biosynthesis. Interestingly, the subset of genes that overlap between this module and hotspot were significantly enriched for processes involved in stress and defense response and were specifically associated with eQTLs for which the higher gene expression was associated with the *E. grandis* allele. This shows that the difference between the *E. grandis* and *E. urophylla* allele suddenly became important in mature trees, as it had a bigger outsize effect on the outcome of gene expression.

2.6 Conclusion and Future Prospects

This study provides a genome-wide overview of regulatory loci and polymorphisms associated with variation in transcript abundance for xylem expressed genes in mature, field-grown trees and characterises the molecular genetic architecture of wood formation in *Eucalyptus* hybrids. We observe a dramatic shift in the genetic architecture of genes involved in xylogenesis, which could be a result of developmental and stress responses working in combination to keep the trees alive during the severe drought period. Combining the results from our eQTL analyses with important wood property trait QTLs will allow us to identify positional candidate genes that act as regulators of gene expression. This will enable us to analyse the genetic architecture underlying complex wood biorefinery traits and to identify genes and pathways that interact to determine complex trait outcomes, which can be used to engineer or breed for these traits while avoiding negative effects on plant growth. It is important to develop effective strategies to validate these genes and their roles in complex biological processes involving many genes. It is also important for more research to be done on how plants can be engineered to adapt to environmental changes caused by climate change and global warming, as abiotic stress has substantial effects on cell wall metabolism and wood formation (Le Gall *et al.*, 2015; Ployet *et al.*, 2017).

Blein-Nicolas *et al.* (2019) used a systems genetics approach that included genomics, proteomics and phenotypic data on maize hybrids subjected to water deprivation. The authors showed that drought response has a strong effect on many proteins by inducing proteome remodelling and reprogramming the genetic control of protein abundance that could influence the phenotype. This can be an important avenue to explore in trees, as it could help to further elucidate the molecular mechanisms underlying drought response and tolerance. Although there has been considerable progress in understanding the effects of abiotic stress on cell wall metabolism, the complex mechanisms underlying the response can be better understood by combining intermediate phenotypes across different biological scales through systems biology approaches. One of the major objectives of systems biology is to produce models that are able to predict how a system will react to genetic variation (or artificial selection) and how this will affect growth and development (Drost *et al.*, 2010). As more data becomes available and systems genetics analyses become more comprehensive, future studies will have to adopt multidisciplinary approaches, such as the incorporation of machine learning algorithms for computational modelling, to extract meaning from the data and to produce testable models that are predictive to some extent (Harfouche *et al.*, 2019).

2.7 Acknowledgements

The work performed in this study was funded in part by the National Research Foundation (NRF) of South Africa – Bioinformatics and Functional Genomics Programme (BFG Grant UID 86936 and 97911), the Technology and Human Resources for Industry Programme (THRIP Grant UID 80118 and 96413), the Department of Science and Technology (Strategic Grant for the *Eucalyptus* Genomics Platform) and by Sappi Forest Research through the Forest Molecular Genetics Programme at the University of Pretoria. The authors acknowledge postgraduate scholarship support from the NRF.

2.8 References

Alexa A, Rahnenführer J. 2018. topGO: enrichment analysis for Gene Ontology. cran.r-project.org.

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald Martin, Rubin GM, Sherlock G. 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**: 25–29.

Basten CJ, Weir BS, Zeng ZB. 2005. QTL cartographer, version 1.17. *Department of Statistics, North Carolina State University, Raleigh* **189**.

Blein-Nicolas M, Negro SS, Balliau T, Welcker C, Bosquet LC, Nicolas SD, Charcosset A, Zivy M. 2019. A proteomics-based systems genetics approach reveals environment-specific loci modulating protein co-expression and drought-related traits in maize. *bioRxiv* 636514.

Boyle EA, Li YI, Pritchard JK. 2018. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**: 1177–1186.

Christie N, Myburg AA, Joubert F, Murray SL, Carstens M, Lin YC, Meyer J, Crampton BG, Christensen SA, Ntuli JF, Wighard SS, Van de Peer Y, Berge DK. 2017. Systems genetics reveals a transcriptional network associated with susceptibility in the maize–grey leaf spot pathosystem. *Plant Journal* **89**: 746–763.

Dharanishanthi V, Ghosh M. 2016. Construction of co-expression network based on natural expression variation of xylogenesis-related transcripts in *Eucalyptus tereticornis*. *Molecular Biology Reports* **43**: 1129–1146.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

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

Drost DR, Puranik S, Novaes E, Novaes CRDB, Dervinis C, Gailing O, Kirst M. 2015. Genetical genomics of *Populus* leaf shape variation. *BMC Plant Biology* **15**: 1–10.

Ewels P, Lundin S, Max K. 2018. Data and text mining MultiQC : summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**: 3047–3048.

Feltus FA. 2014. Systems genetics: A paradigm to improve discovery of candidate genes and mechanisms underlying complex traits. *Plant Science* **223**: 45–48.

Le Gall H, Philippe F, Domon JM, Gillet F, Pelloux J, Rayon C. 2015. Cell wall metabolism in response to abiotic stress. *Plants* **4**: 112–166.

Galpaz N, Gonda I, Shem-Tov D, Barad O, Tzuri G, Lev S, Fei Z, Xu Y, Lombardi N, Mao L, Jiao C, Harel-Beja R, Doron-Faigenboim A, Tzfadia O, Bar E, Meir A, Sa'ar U, Fait A, Halperin E, Kenigswald M, Fallik E, Kol G, Ronen G, Burger Y, Gur A, Tadmor Y, Portnoy

V, Schaffer AA, Lewinsohn E, Giovannoni JJ, Katzir N. **2018**. Deciphering genetic factors that determine melon fruit-quality traits using RNA-Seq-based high-resolution QTL and eQTL mapping. *The Plant Journal* **94**: 169-191.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS. **2012**. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research* **40**: D1178–D1186.

Harfouche AL, Jacobson DA, Kainer D, Romero JC, Harfouche AH, Scarascia Mugnozza G, Moshelion M, Tuskan GA, Keurentjes JJB, Altman A. **2019**. Accelerating climate resilient plant breeding by applying next-generation artificial intelligence. *Trends in Biotechnology*.

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

Kersting AR, Mizrachi E, Bornberg-bauer E, Myburg AA. **2015**. Protein domain evolution is associated with reproductive diversification and adaptive radiation in the genus *Eucalyptus*. *New Phytologist* **206**: 1328-1336.

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**: 2295–2303.

Kirst M, Myburg AA, De Leon JPG, Kirst ME, Scott J, Sederoff R. **2004**. Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of *Eucalyptus*. *Plant Physiology* **135**: 2368–2378.

Kirst M, Yu Q. **2007**. Genetical genomics: Successes and prospects in plants. *Genomics-Assisted Crop Improvement* **1**: 245–265.

Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: An information aesthetic for comparative genomics. *Genome Research* **19**: 1639-1645.

Kullan ARK, van Dyk MM, Hefer CA, Jones N, Kanzler A, Myburg AA. **2012**. Genetic dissection of growth, wood basic density and gene expression in interspecific backcrosses of *Eucalyptus grandis* and *E. urophylla*. *BMC Genetics* **13**: 1–12.

Kullan KR, Van Dyk MM, Myburg AA. **2012**. High-density genetic linkage maps with over 2 , 400 sequence-anchored DArT markers for genetic dissection in an F_2 pseudo-backcross of *Eucalyptus grandis* × *E . urophylla*. *Tree Genetics & Genomes*, **8**: 163-175.

Langfelder P, Horvath S. **2007**. Eigengene networks for studying the relationships between co-expression modules. *BMC Systems Biology* **1**: 1–17.

Langfelder P, Horvath S. **2008**. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* **9**: 1–14.

Mähler N, Wang J, Terebieniec BK, Ingvarsson PK, Street NR, Hvidsten TR. 2017. Gene co-expression network connectivity is an important determinant of selective constraint. *PLoS Genetics* **4**: 1-33.

Mizrachi E, Myburg AA. **2016**. Systems genetics of wood formation. *Current Opinion in Plant Biology* **30**: 94–100.

Mizrachi E, Verbeke L, Christie N, Fierro AC, Mansfield SD, Davis MF, Gjersing E, Tuskan

GA, Van Montagu M, Van de Peer Y, Marchal K, Myburg AA. **2017**. Network-based integration of systems genetics data reveals pathways associated with lignocellulosic biomass accumulation and processing. *Proceedings of the National Academy of Sciences* **114**: 1195-1200.

Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D, Goodstein DM, Dubchak I, Poliakov A, Mizrachi E, Kullan ARK, Hussey SG, Pinard D, van der Merwe K, Singh P, Van Jaarsveld I, Silva-Junior OB, Togawa RC, Pappas MR, Faria DA, Sansaloni CP, Petroli CD, Yang X, Ranjan P, Tschaplinski TJ, Ye C, Li T, Sterck L, Vanneste K, Murat F, Soler MM, Clemente HS, Saidi N, Cassan-Wang H, Dunand C, Hefer CA, Bornberg-Bauer E, Kersting AR, Vining K, Amarasinghe V, Ranik M, Naithani S, Elser J, Boyd AE, Liston A, Spatafora JW, Dharmwardhana P, Raja R, Sullivan C, Romanel E, Alves-Ferreira M, Külheim CK, Foley W, Carocha V, Paiva J, Kudrna D, Brommonschenkel SH, Pasquali G, Byrne M, Rigault P, Tibbits J, Spokevicius A, Jones RC, Steane DA, Vaillancourt RE, Potts BM, Joubert F, Barry K, Pappas Jr GJ, Strauss SH, Jaiswal P, Grima-Pettenati J, Salse J, Van de Peer Y, Rokhsar DS, Schmutz J. **2014**. The genome of *Eucalyptus grandis*. *Nature* **510**: 356.

Myburg AA, Griffin AR, Sederoff RR, Whetten RW. **2003**. Comparative genetic linkage maps of *Eucalyptus grandis*, *Eucalyptus globulus* and their F_1 hybrid based on a double pseudo-backcross mapping approach. *Theoretical and Applied Genetics* **107**: 1028-1042.

Myburg AA, Hussey SG, Wang JP, Street NR. **2019**. Systems and synthetic biology of forest trees: A bioengineering paradigm for woody biomass feedstocks. *Frontiers in Plant Science* **10**: 775.

Van Ooijen JW. **2006**. JoinMap4, Software for the calculation of genetic linkage maps in experimental populations. *Kyazma BV, Wageningen* **33**.

Pertea M, Pertea GM, Antonescu CM, Chang T, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* **33**: 209.

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

Ployet R, Labate TV, Cataldi TR, Christina M, Morel M, Grima-Pettenati J, Mounet F, Labate CA, Chaix G. **2019**. A systems biology view of wood formation in *Eucalyptus grandis* trees submitted to different potassium and water regimes. *New Phytologist* **223**: 766-782.

Ployet R, Soler M, Carocha V, Ladouce N, Alves A, Rodrigues J, Harvengt L, Marque C, Teulières C, Grima-pettenati J, Mounet F. **2017**. Long cold exposure induces transcriptional and biochemical remodelling of xylem secondary cell wall in *Eucalyptus*. *Tree Physiology* **38**: 409-422.

Proost S, Van Bel M, Vaneechoutte D, Van de Peer Y, Inzé D, Mueller-Roeber B, Vandepoele K. **2015**. PLAZA 3.0: an access point for plant comparative genomics. *Nucleic Acids Research* **43**: D974–D981.

Santos C, Almeida NF, Alves ML, Horres R, Krezdorn N, Leitão ST, Aznar-Fernández T, Rotter B, Winter P, Rubiales D, Vaz Patto MC. **2018**. First genetic linkage map of *Lathyrus cicera* based on RNA sequencing-derived markers: Key tool for genetic mapping of disease resistance. *Horticulture Research* **5**: 45.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. **2003**. Cytoscape: a software environment for integrated models of biomolecular interaction

networks. *Genome Research* **13**: 2498–2504.

Shearman JR, Sangsrakru D, Jomchai N, Ruang- P. 2015. SNP identification from RNA sequencing and linkage map construction of rubber tree for anchoring the draft genome. *PLoS ONE* **10**: 1–12.

Singh J, Kalberer SR, Belamkar V, Assefa T, Nelson MN, Farmer AD, Blackmon WJ, Cannon SB. 2018. A transcriptome-SNP-derived linkage map of *Apios americana* (potato bean) provides insights about genome re-organization and synteny conservation in the phaseoloid legumes. *Theoretical and Applied Genetics* **131**: 333–351.

Spearman C. 1904. The proof and measurement of association between two things. *American Journal of Psychology* **15**: 72–101.

Steinsaltz D, Dahl A, Wachter KW. 2017. On negative heritability and negative estimates of heritability. *bioRxiv* 232843.

Taylor-Teeples M, Lin L, de Lucas M, Turco G, Toal TW, Gaudinier A, Young NF, Trabucco GM, Veling MT, Lamothe R, Handakumbura PP, Xiong G, Wang C, Corwin J, Tsoukalas A, Zhang L, Ware D, Pauly M, Kliebenstein DJ, Dehesh K, Tagkopoulos I, Breton G, Pruneda-Paz JL, Ahnert SE, Kay SA, Hazen SP, Brady SM. 2015. An *Arabidopsis* gene regulatory network for secondary cell wall synthesis. *Nature* **517**: 571–575.

Thavamanikumar S, Southerton SG, Bossinger G, Thumma BR. 2013. Dissection of complex traits in forest trees - opportunities for marker-assisted selection. *Tree Genetics and Genomes* **9**: 627–639.

Upton GJG. 1992. Fisher's Exact Test. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **155**: 395–402.

Vaillant A, Honvault A, Bocs S, Summo M, Makouanzi G, Vigneron P, Bouvet J-M. 2018. Genetic effect in leaf and xylem transcriptome variations among *Eucalyptus urophylla* x *grandis* hybrids in field conditions. *Silvae Genetica* **67**: 57–65.

Van Der Auwera GA , Carneiro MO, Hartl C, Poplin R, Angel G, Levy-moonshine A, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, Depristo MA. 2013. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics* **43**: 11-10.

Visscher PM, Hill WG, Wray NR. 2008. Heritability in the genomics era - Concepts and misconceptions. *Nature Reviews Genetics* **9**: 255–266.

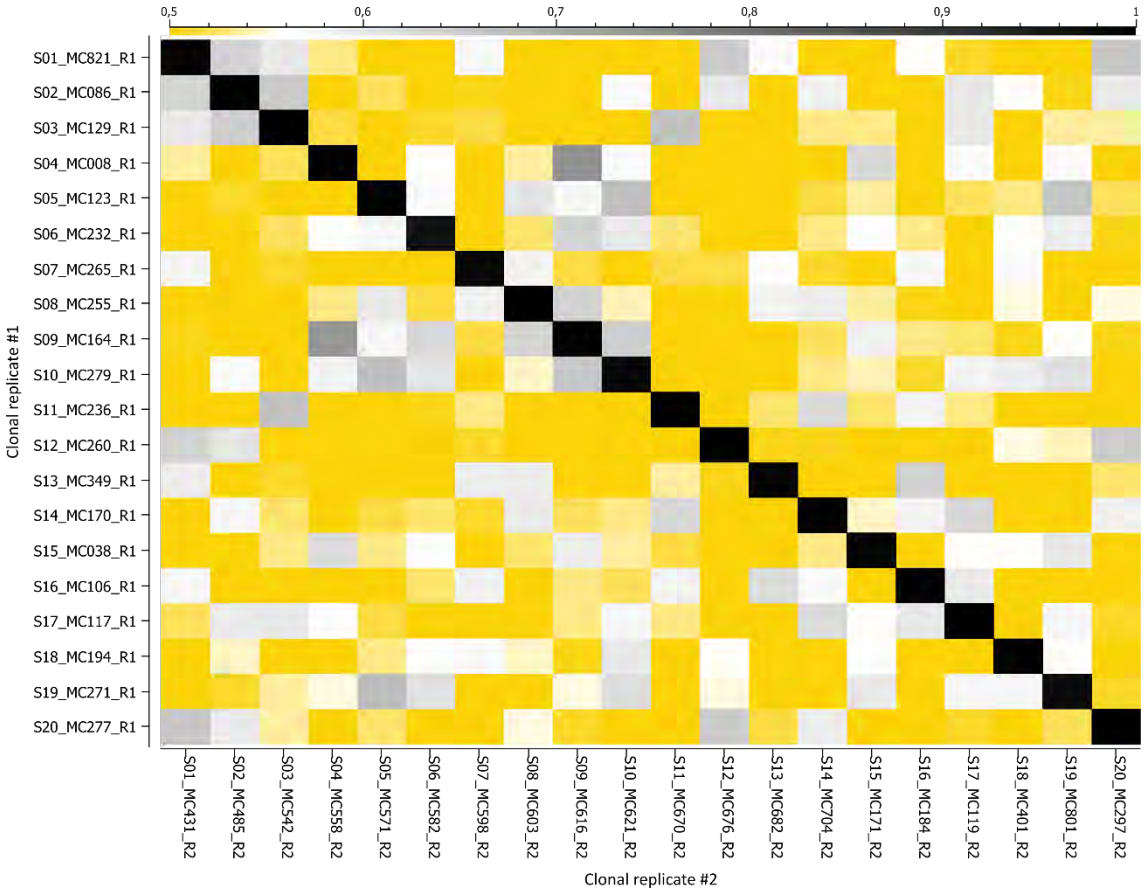
Voorrips RE. 2002. MapChart: software for the graphical presentation of linkage maps and QTLs. *Journal of Heredity* **93**: 77–78.

Zhang B, Horvath S. 2005. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology* **4.**

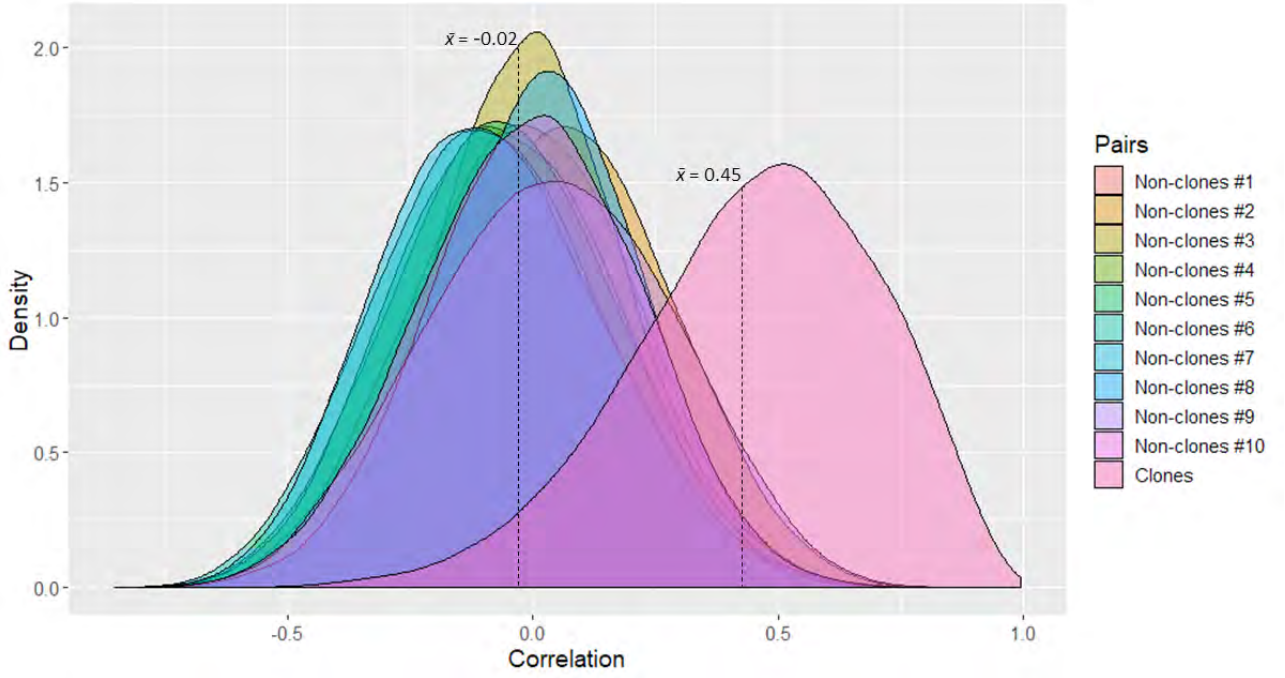
Zhang J, Yang Y, Zheng K, Xie M, Feng K, Jawdy SS, Gunter LE, Ranjan P, Singan VR, Engle N, Lindquist E, Barry K, Schmutz J, Zhao N, Tschaplinski TJ, LeBoldus J, Tuskan GA, Chen JG, Muchero W. 2018. Genome-wide association studies and expression-based quantitative trait loci analyses reveal roles of HCT2 in caffeoylquinic acid biosynthesis and its regulation by defense-responsive transcription factors in *Populus*. *New Phytologist* **220**: 502–516

2.9 Supplementary Information

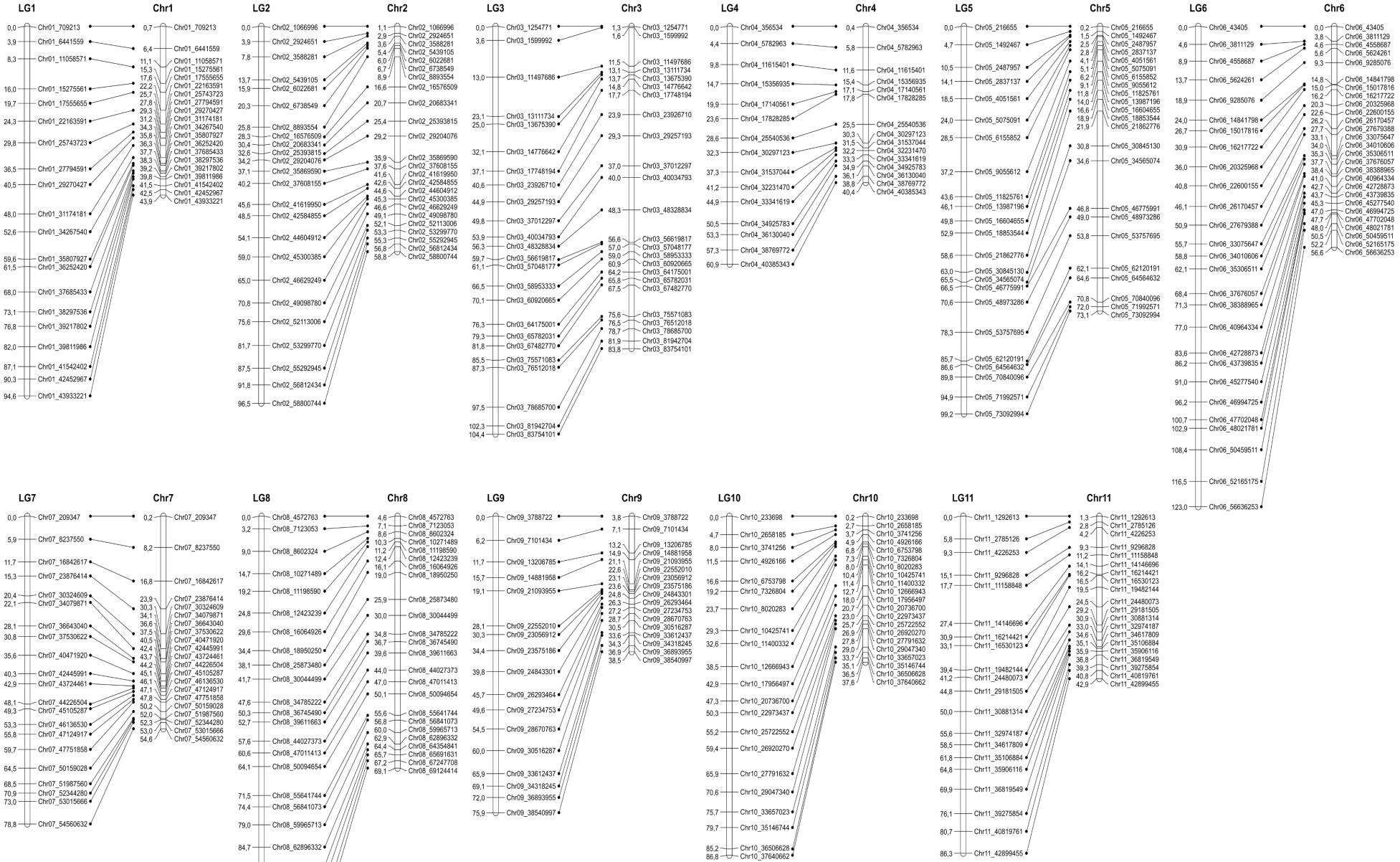
2.9.1 Supplementary Figures and Tables



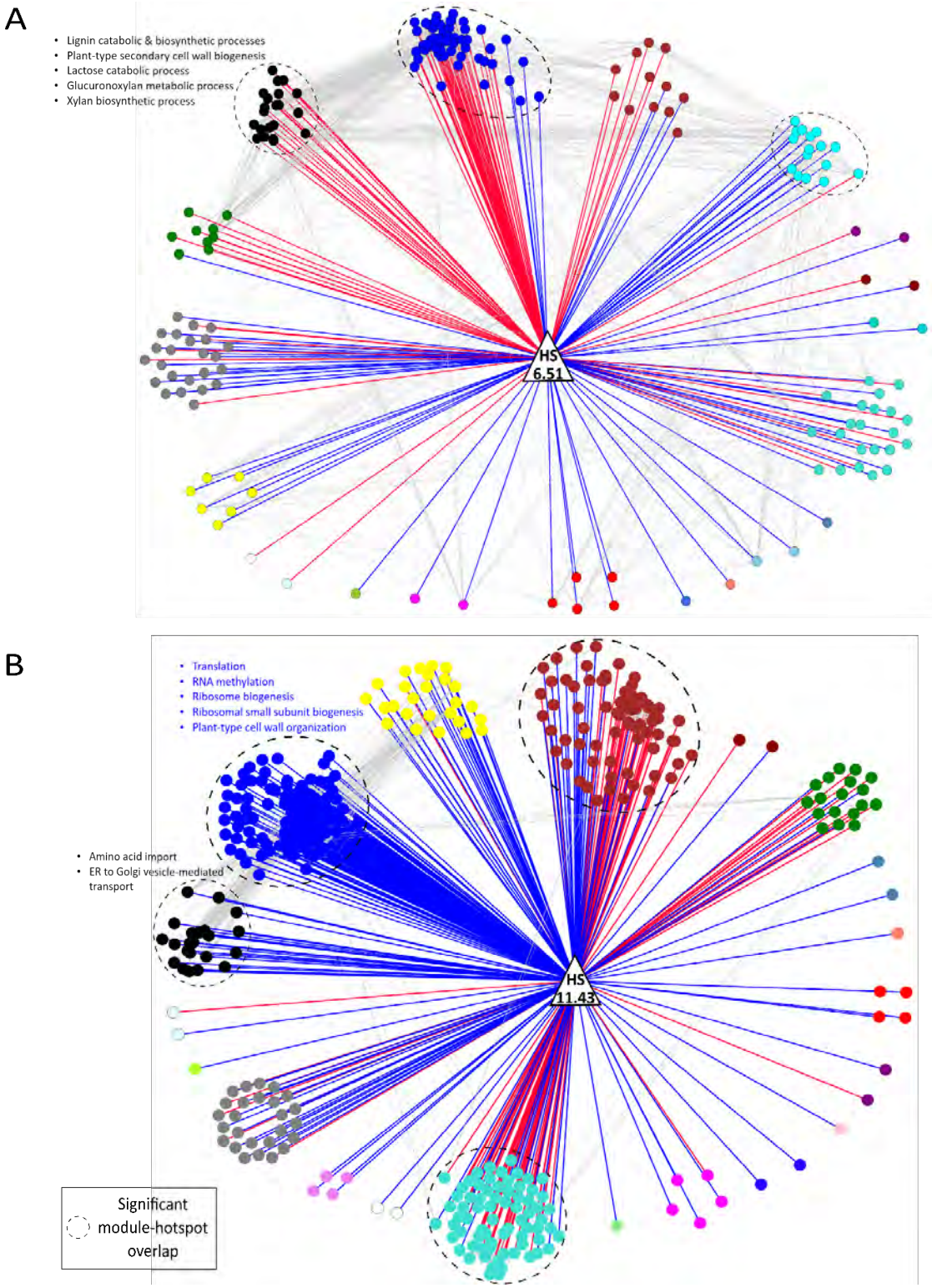
Supplementary Figure 2.1: Pairwise identity-by-decent (IBD) analysis of 234 transcriptome-derived SNP genotypes for 20 pairs of clonal replicates. Black squares indicate a full or approximately 1:1 match between the RNA-seq genotypes and yellow squares indicate an approximate 50% match, as expected from full-siblings. Sample pairs are indicated by the "S" prefix and the biological replicate number is indicated by the "R" suffix.



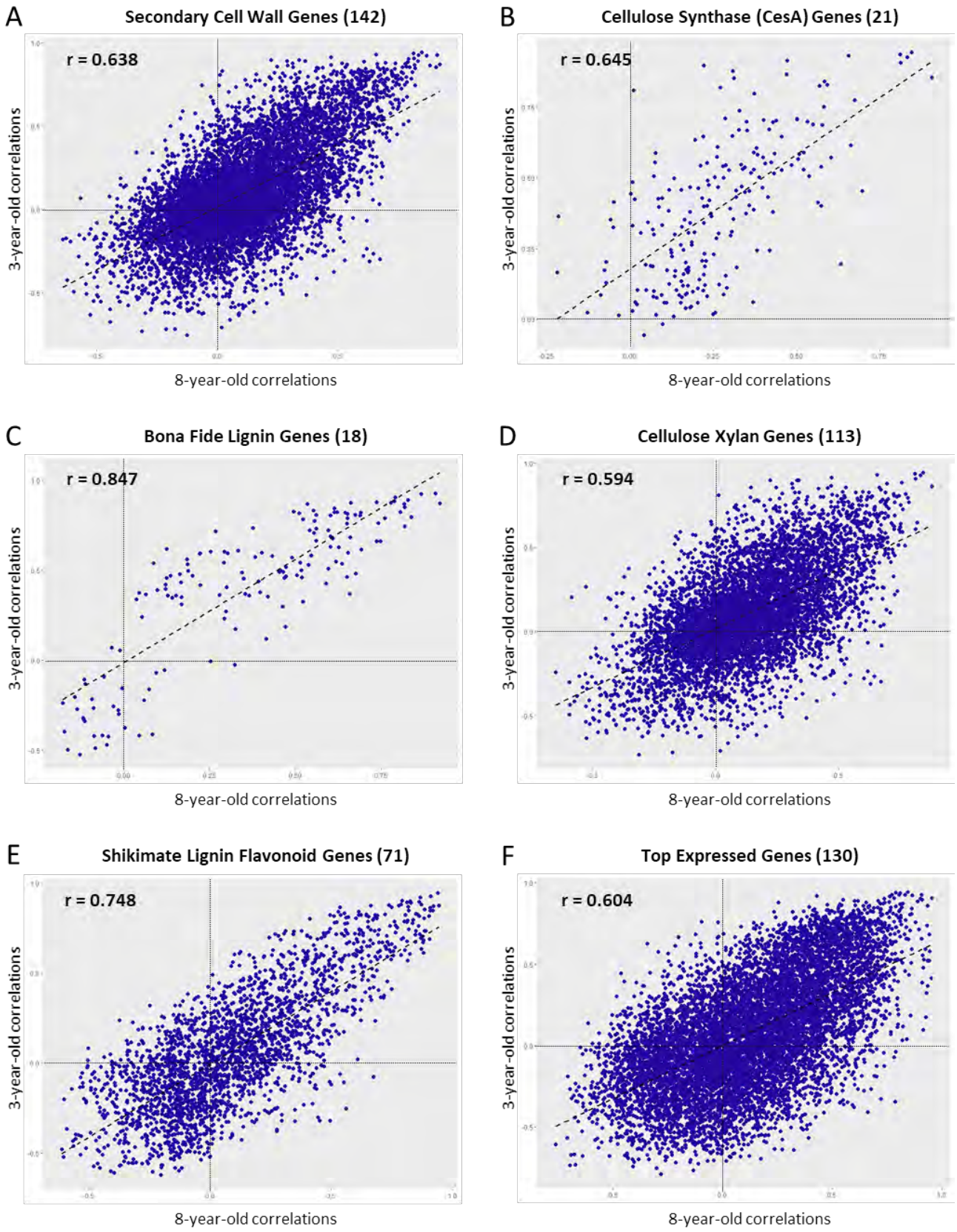
Supplementary Figure 2.2: Distribution of correlation values for 25,307 genes in eight-year-old developing xylem tissue. The density plots show the distribution of Spearman rank correlations of transcript abundance for 20 random non-clonal pairs (repeated ten times) and 20 clonal pairs. The transcriptome-wide correlations for non-clonal pairs were centered around zero ($\bar{x}$ = -0.02), while the correlations for 20 confirmed clonal pairs were centered around 0.5 ($\bar{x}$ = 0.45).


LG1
0,0 Chr01_709213
3,9 Chr01_6441559
8,3 Chr01_11058571
16,0 Chr01_15275561
19,7 Chr01_17555655
24,3 Chr01_22163591
29,8 Chr01_25743723
36,5 Chr01_27794591
40,5 Chr01_29270427
48,0 Chr01_31174181
52,6 Chr01_34267540
59,6 Chr01_35807927
61,5 Chr01_36252420
68,0 Chr01_37685433
73,1 Chr01_38297536
76,8 Chr01_39217802
82,0 Chr01_39811986
87,1 Chr01_41542402
90,3 Chr01_42452967
94,6 Chr01_43933221
Chr1
0,7 Chr01_709213
6,4 Chr01_6441559
11,1 Chr01_11058571
15,3 Chr01_15275561
17,6 Chr01_17555655
22,2 Chr01_22163591
25,7 Chr01_25743723
27,8 Chr01_27794591
29,3 Chr01_29270427
31,2 Chr01_31174181
34,3 Chr01_34267540
35,8 Chr01_35807927
36,3 Chr01_36252420
37,7 Chr01_37685433
38,3 Chr01_38297536
39,2 Chr01_39217802
39,8 Chr01_39811986
41,5 Chr01_41542402
42,5 Chr01_42452967
43,9 Chr01_43933221
LG2
0,0 Chr02_1066996
3,9 Chr02_2924651
7,8 Chr02_3588281
13,7 Chr02_5439105
15,9 Chr02_6022681
20,3 Chr02_6738549
25,8 Chr02_8893554
28,3 Chr02_16576509
30,4 Chr02_20683341
32,6 Chr02_25393815
34,2 Chr02_29204076
37,1 Chr02_35869590
40,2 Chr02_37608155
45,6 Chr02_41619950
48,5 Chr02_42584855
54,1 Chr02_44604912
59,0 Chr02_45300385
65,0 Chr02_46629249
70,8 Chr02_49098780
75,6 Chr02_52113006
81,7 Chr02_53299770
87,5 Chr02_55292945
91,8 Chr02_56812434
96,5 Chr02_58800744
Chr2
1,1 Chr02_1066996
2,9 Chr02_2924651
3,6 Chr02_3588281
5,4 Chr02_5439105
6,0 Chr02_6022681
6,7 Chr02_6738549
8,9 Chr02_8893554
16,6 Chr02_16576509
20,7 Chr02_20683341
25,4 Chr02_25393815
29,2 Chr02_29204076
35,9 Chr02_35869590
37,6 Chr02_37608155
41,6 Chr02_41619950
42,6 Chr02_42584855
44,6 Chr02_44604912
45,3 Chr02_45300385
46,6 Chr02_46629249
49,1 Chr02_49098780
52,1 Chr02_52113006
53,3 Chr02_53299770
55,3 Chr02_55292945
56,8 Chr02_56812434
58,8 Chr02_58800744
LG3
0,0 Chr03_1254771
3,6 Chr03_1599992
13,0 Chr03_11497686
23,1 Chr03_13111734
25,0 Chr03_13675390
32,1 Chr03_14776642
37,1 Chr03_17748194
40,6 Chr03_23926710
44,9 Chr03_29257193
49,8 Chr03_37012297
53,9 Chr03_40034793
56,3 Chr03_48328834
59,7 Chr03_56619817
61,1 Chr03_57048177
66,5 Chr03_58953333
70,1 Chr03_60920665
76,3 Chr03_64175001
79,3 Chr03_65782031
81,8 Chr03_67482770
85,5 Chr03_75571083
87,3 Chr03_76512018
97,5 Chr03_78685700
102,3 Chr03_81942704
104,4 Chr03_83754101
Chr3
1,3 Chr03_1254771
1,6 Chr03_1599992
11,5 Chr03_11497686
13,1 Chr03_13111734
13,7 Chr03_13675390
14,8 Chr03_14776642
17,7 Chr03_17748194
23,9 Chr03_23926710
29,3 Chr03_29257193
37,0 Chr03_37012297
40,0 Chr03_40034793
48,3 Chr03_48328834
56,6 Chr03_56619817
57,0 Chr03_57048177
59,0 Chr03_58953333
60,9 Chr03_60920665
64,2 Chr03_64175001
65,8 Chr03_65782031
67,5 Chr03_67482770
75,6 Chr03_75571083
76,5 Chr03_76512018
78,7 Chr03_78685700
81,9 Chr03_81942704
83,8 Chr03_83754101
LG4
0,0 Chr04_356534
4,4 Chr04_5782963
9,8 Chr04_11615401
14,7 Chr04_15356935
19,9 Chr04_17140561
23,6 Chr04_17828285
28,6 Chr04_25540536
32,3 Chr04_30297123
37,3 Chr04_31537044
41,2 Chr04_32231470
44,9 Chr04_33341619
50,5 Chr04_34925783
53,3 Chr04_36130040
57,3 Chr04_38769772
60,9 Chr04_40385343
Chr4
0,4 Chr04_356534
5,8 Chr04_5782963
11,6 Chr04_11615401
15,4 Chr04_15356935
17,1 Chr04_17140561
17,8 Chr04_17828285
25,5 Chr04_25540536
30,3 Chr04_30297123
31,5 Chr04_31537044
32,2 Chr04_32231470
33,3 Chr04_33341619
34,9 Chr04_34925783
36,1 Chr04_36130040
38,8 Chr04_38769772
40,4 Chr04_40385343
LG5
0,0 Chr05_216655
4,7 Chr05_1492467
10,5 Chr05_2487957
14,1 Chr05_2837137
18,5 Chr05_4051561
24,0 Chr05_5075091
28,5 Chr05_6155852
37,2 Chr05_9055612
43,6 Chr05_11825761
46,1 Chr05_13987196
49,8 Chr05_16604655
52,9 Chr05_18853544
58,6 Chr05_21862776
63,0 Chr05_30845130
65,5 Chr05_34565074
66,5 Chr05_46775991
70,6 Chr05_48973286
78,3 Chr05_53757695
85,7 Chr05_62120191
86,6 Chr05_64564632
89,8 Chr05_70840096
94,9 Chr05_71992571
99,2 Chr05_73092994
Chr5
0,2 Chr05_216655
1,5 Chr05_1492467
2,5 Chr05_2487957
2,8 Chr05_2837137
4,1 Chr05_4051561
5,1 Chr05_5075091
6,2 Chr05_6155852
9,1 Chr05_9055612
11,8 Chr05_11825761
14,0 Chr05_13987196
16,6 Chr05_16604655
18,9 Chr05_18853544
21,9 Chr05_21862776
30,8 Chr05_30845130
34,6 Chr05_34565074
46,8 Chr05_46775991
49,0 Chr05_48973286
53,8 Chr05_53757695
62,1 Chr05_62120191
64,6 Chr05_64564632
70,8 Chr05_70840096
72,0 Chr05_71992571
73,1 Chr05_73092994
LG6
0,0 Chr06_43405
4,6 Chr06_3811129
8,9 Chr06_4558687
13,7 Chr06_5624261
18,9 Chr06_9285076
24,0 Chr06_14841798
26,7 Chr06_15017816
30,9 Chr06_16217722
36,0 Chr06_20325968
40,8 Chr06_22600155
46,1 Chr06_26170457
50,9 Chr06_27679388
55,7 Chr06_33075647
58,8 Chr06_34010606
62,1 Chr06_35306511
68,4 Chr06_37676057
71,3 Chr06_38388965
77,0 Chr06_40964334
83,6 Chr06_42728873
86,2 Chr06_43739835
91,0 Chr06_45277540
96,2 Chr06_46994725
100,7 Chr06_47702048
102,9 Chr06_48021781
108,4 Chr06_50459511
116,5 Chr06_52165175
123,0 Chr06_56636253
Chr6
0,0 Chr06_43405
3,8 Chr06_3811129
4,6 Chr06_4558687
5,6 Chr06_5624261
9,3 Chr06_9285076
14,8 Chr06_14841798
15,0 Chr06_15017816
16,2 Chr06_16217722
20,3 Chr06_20325968
22,6 Chr06_22600155
26,2 Chr06_26170457
27,7 Chr06_27679388
33,1 Chr06_33075647
34,0 Chr06_34010606
35,3 Chr06_35306511
37,7 Chr06_37676057
38,4 Chr06_38388965
41,0 Chr06_40964334
42,7 Chr06_42728873
43,7 Chr06_43739835
45,3 Chr06_45277540
47,0 Chr06_46994725
47,7 Chr06_47702048
48,0 Chr06_48021781
50,5 Chr06_50459511
52,2 Chr06_52165175
56,6 Chr06_56636253
LG7
0,0 Chr07_209347
5,9 Chr07_8237550
11,7 Chr07_16842617
15,3 Chr07_23876414
20,4 Chr07_30324609
22,1 Chr07_34079871
28,1 Chr07_36643040
30,8 Chr07_37530622
35,6 Chr07_40471920
40,3 Chr07_42445991
42,9 Chr07_43724461
48,1 Chr07_44226504
49,3 Chr07_45105287
53,3 Chr07_46136530
55,8 Chr07_47124917
59,7 Chr07_47751858
64,5 Chr07_50159028
68,5 Chr07_51987560
70,9 Chr07_52344280
73,0 Chr07_53015666
78,8 Chr07_54560632
Chr7
0,2 Chr07_209347
8,2 Chr07_8237550
16,8 Chr07_16842617
23,9 Chr07_23876414
30,3 Chr07_30324609
34,1 Chr07_34079871
36,6 Chr07_36643040
37,5 Chr07_37530622
40,5 Chr07_40471920
42,4 Chr07_42445991
43,7 Chr07_43724461
44,2 Chr07_44226504
45,1 Chr07_45105287
46,1 Chr07_46136530
47,1 Chr07_47124917
47,8 Chr07_47751858
50,2 Chr07_50159028
52,0 Chr07_51987560
52,3 Chr07_52344280
53,0 Chr07_53015666
54,6 Chr07_54560632
LG8
0,0 Chr08_4572763
3,2 Chr08_7123053
9,0 Chr08_8602324
14,7 Chr08_10271489
19,2 Chr08_11198590
24,8 Chr08_12423239
29,6 Chr08_16064926
34,4 Chr08_18950250
38,1 Chr08_25873480
41,7 Chr08_30044499
47,6 Chr08_34785222
50,3 Chr08_36745490
52,7 Chr08_39611663
57,6 Chr08_44027373
60,6 Chr08_47011413
64,1 Chr08_50094654
71,5 Chr08_55641744
74,4 Chr08_56841073
79,0 Chr08_59965713
84,7 Chr08_62896332
89,5 Chr08_64354841
93,4 Chr08_65691631
98,4 Chr08_67247708
101,8 Chr08_69124414
Chr8
4,6 Chr08_4572763
7,1 Chr08_7123053
8,6 Chr08_8602324
10,3 Chr08_10271489
11,2 Chr08_11198590
12,4 Chr08_12423239
16,1 Chr08_16064926
19,0 Chr08_18950250
25,9 Chr08_25873480
30,0 Chr08_30044499
34,8 Chr08_34785222
36,7 Chr08_36745490
39,6 Chr08_39611663
44,0 Chr08_44027373
47,0 Chr08_47011413
50,1 Chr08_50094654
55,6 Chr08_55641744
56,8 Chr08_56841073
60,0 Chr08_59965713
62,9 Chr08_62896332
64,4 Chr08_64354841
65,7 Chr08_65691631
67,2 Chr08_67247708
69,1 Chr08_69124414
LG9
0,0 Chr09_3788722
6,2 Chr09_7101434
11,7 Chr09_13206785
15,7 Chr09_14881958
19,1 Chr09_21093955
28,1 Chr09_22552010
30,3 Chr09_23056912
34,4 Chr09_23575186
39,8 Chr09_24843301
45,7 Chr09_26293464
49,6 Chr09_27234753
54,5 Chr09_28670763
60,0 Chr09_30516287
65,9 Chr09_33612437
69,1 Chr09_34318245
72,0 Chr09_36893955
75,9 Chr09_38540997
Chr9
3,8 Chr09_3788722
7,1 Chr09_7101434
13,2 Chr09_13206785
14,9 Chr09_14881958
21,1 Chr09_21093955
22,6 Chr09_22552010
23,1 Chr09_23056912
23,6 Chr09_23575186
24,8 Chr09_24843301
26,3 Chr09_26293464
27,2 Chr09_27234753
28,7 Chr09_28670763
30,5 Chr09_30516287
33,6 Chr09_33612437
34,3 Chr09_34318245
36,9 Chr09_36893955
38,5 Chr09_38540997
LG10
0,0 Chr10_233698
4,7 Chr10_2658185
8,0 Chr10_3741256
11,5 Chr10_4926166
16,6 Chr10_6753798
19,2 Chr10_7326804
23,7 Chr10_8020283
29,3 Chr10_10425741
32,6 Chr10_11400332
38,5 Chr10_12666943
42,9 Chr10_17956497
47,3 Chr10_20736700
50,3 Chr10_22973437
55,2 Chr10_25722552
59,4 Chr10_26920270
65,9 Chr10_27791632
70,6 Chr10_29047340
75,7 Chr10_33657023
79,7 Chr10_35146744
85,2 Chr10_36506628
86,8 Chr10_37640662
Chr10
0,2 Chr10_233698
2,7 Chr10_2658185
3,7 Chr10_3741256
4,9 Chr10_4926166
6,8 Chr10_6753798
7,3 Chr10_7326804
8,0 Chr10_8020283
10,4 Chr10_10425741
11,4 Chr10_11400332
12,7 Chr10_12666943
18,0 Chr10_17956497
20,7 Chr10_20736700
23,0 Chr10_22973437
25,7 Chr10_25722552
26,9 Chr10_26920270
27,8 Chr10_27791632
29,0 Chr10_29047340
33,7 Chr10_33657023
35,1 Chr10_35146744
36,5 Chr10_36506628
37,6 Chr10_37640662
LG11
0,0 Chr11_1292613
5,8 Chr11_2785126
9,3 Chr11_4226253
15,1 Chr11_9296828
17,7 Chr11_11158848
27,4 Chr11_14146696
30,9 Chr11_16214421
33,1 Chr11_16530123
39,4 Chr11_19482144
41,2 Chr11_24480073
44,8 Chr11_29181505
50,0 Chr11_30881314
55,6 Chr11_32974187
58,5 Chr11_34617809
61,8 Chr11_35106884
64,8 Chr11_35906116
69,9 Chr11_36819549
76,1 Chr11_39275854
80,7 Chr11_40819761
86,3 Chr11_42899455
Chr11
1,3 Chr11_1292613
2,8 Chr11_2785126
4,2 Chr11_4226253
9,3 Chr11_9296828
11,2 Chr11_11158848
14,1 Chr11_14146696
16,2 Chr11_16214421
16,5 Chr11_16530123
19,5 Chr11_19482144
24,5 Chr11_24480073
29,2 Chr11_29181505
30,9 Chr11_30881314
33,0 Chr11_32974187
34,6 Chr11_34617809
35,1 Chr11_35106884
35,9 Chr11_35906116
36,8 Chr11_36819549
39,3 Chr11_39275854
40,8 Chr11_40819761
42,9 Chr11_42899455

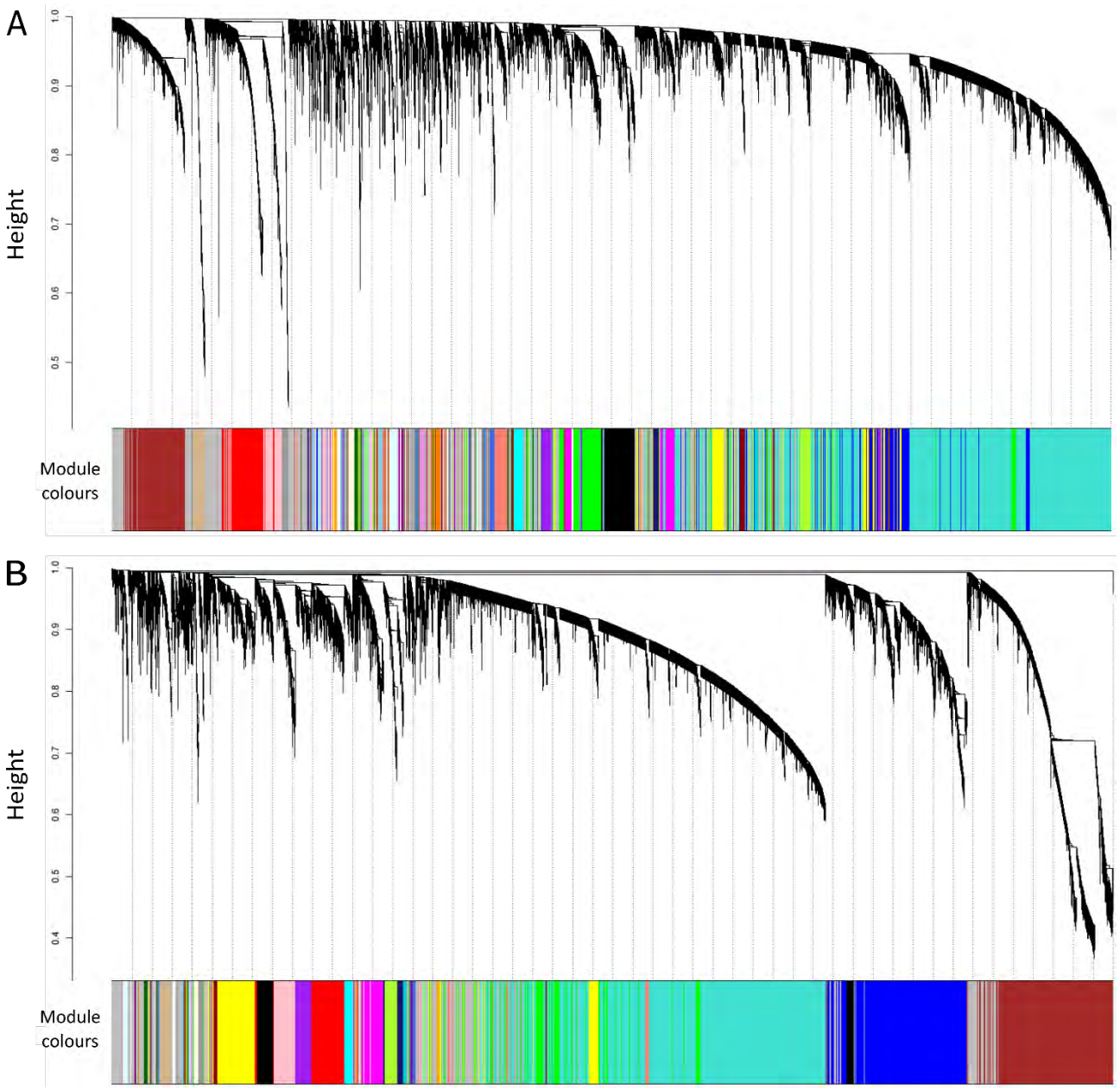
Supplementary Figure 2.3: Genetic framework map vs. physical map for the *E. grandis* x *E. urophylla* F_1 hybrid parent. For each pair, the bar on the left represents a linkage group (LG) on the genetic map, with the genetic position in cM (left) and marker positions in bp (right). For each pair, the bar on the right represents a chromosome (Chr) on the physical map (*E. grandis* v2.0 assembly), with physical position in Mbp (left) and marker positions in bp (right).



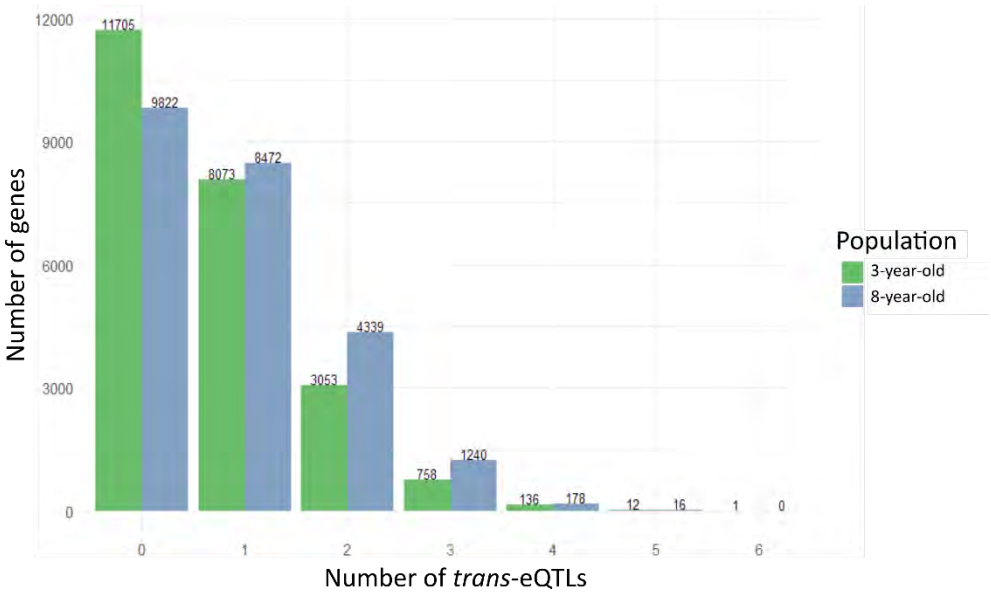
Supplementary Figure 2.4: Association of split-hotspots with gene modules. Networks show the association of gene modules (circles) with split-hotspots for two *trans*-eQTL hotspots (triangles), **(A)** HS_6.51 and **(B)** HS_11.43, to determine if opposite allelic effects are enriched for different biological processes. Module-hotspot overlap significance was determined with a Fisher's exact test and significant GO enrichment terms for these overlapping genes are listed. Grey lines represent correlation between genes. For split-hotspots, higher transcript abundance associated with the *E. urophylla* allele is represented by blue lines and for *E. grandis* with red lines. Genes in HS_6.51 strongly overlap with the black and blue modules and are associated with a maternal allelic effect. Genes in HS_11.43 strongly overlap with these same modules and are associated with a paternal allelic effect. These genes are enriched for different biological processes than those in HS_6.51.



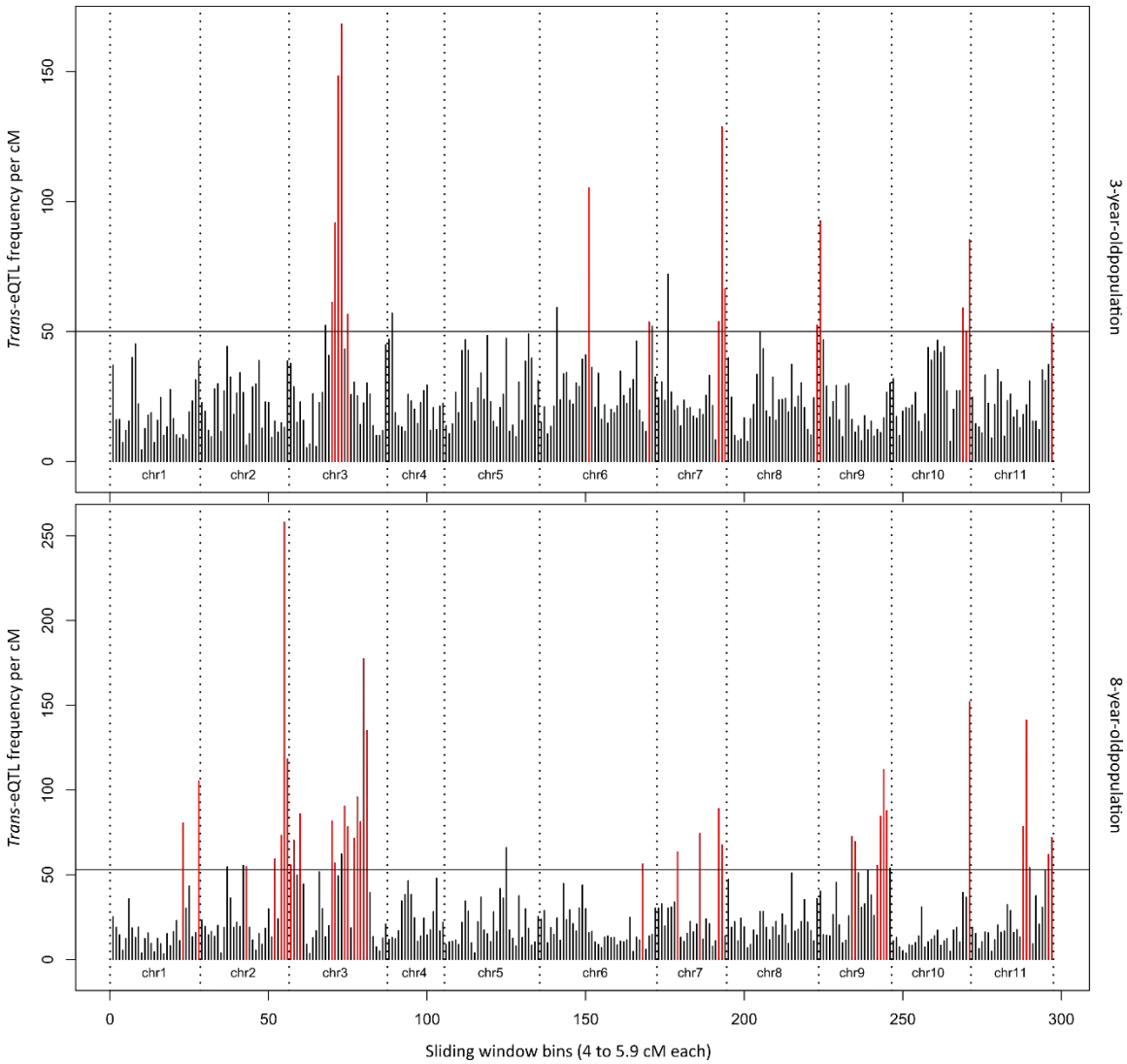
Supplementary Figure 2.5: Age-to-age correlation of gene pairs for six different gene groups. The correlation of each gene with all other genes in the same group was compared between different ages. **A**) Genes involved in secondary cell wall processes, **B**) CesA genes, **C**) bona fide lignin genes, **D**) genes involved in cellulose/xylan pathways, **E**) genes involved in shikimate/lignin/flavonoid pathways, and **F**) the top 100 expressed genes in each population.



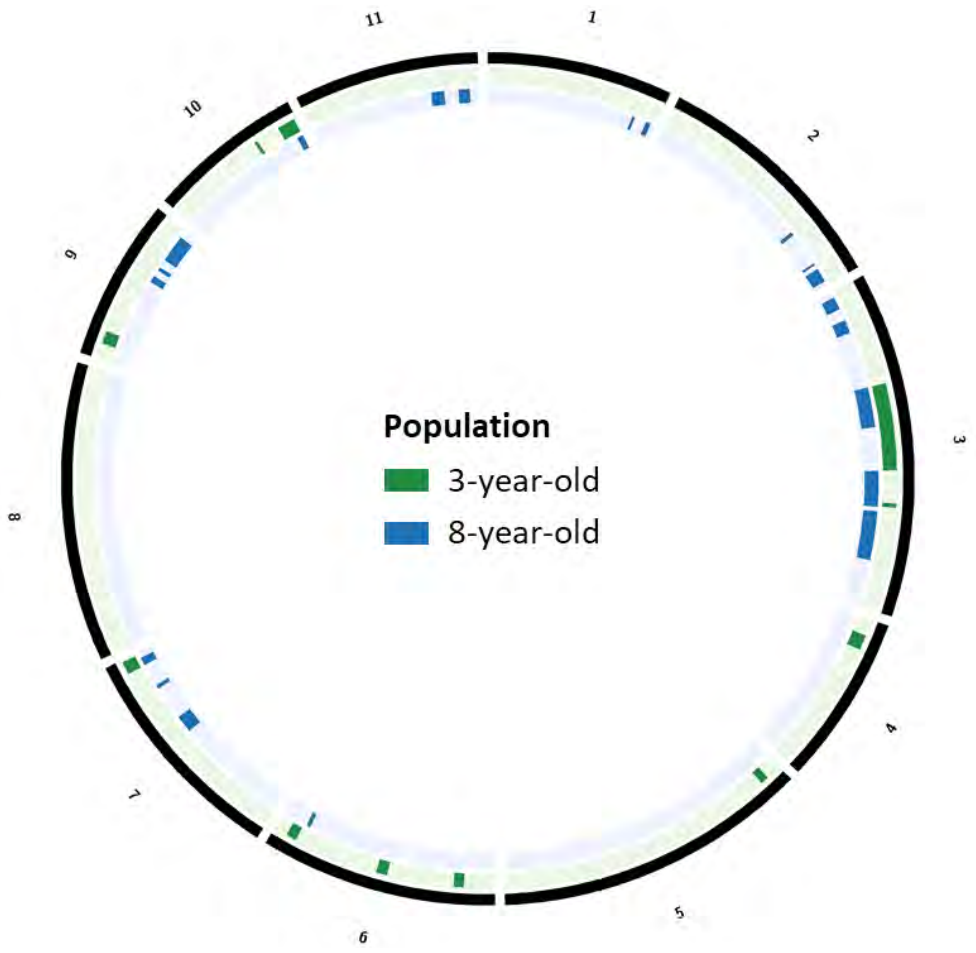
Supplementary Figure 2.6: Gene dendrograms and module colours. Genes are represented by vertical lines (leaves) and branches group together genes that are highly co-expressed and interconnected. Modules are identified by individual branches using the Dynamic Tree Cut approach. **(A)** Genes are divided into 101 modules for the three-year-old population. **(B)** Genes are divided into 54 modules for the eight-year-old population. Refer to **Supplementary File 2.6** for gene module memberships.



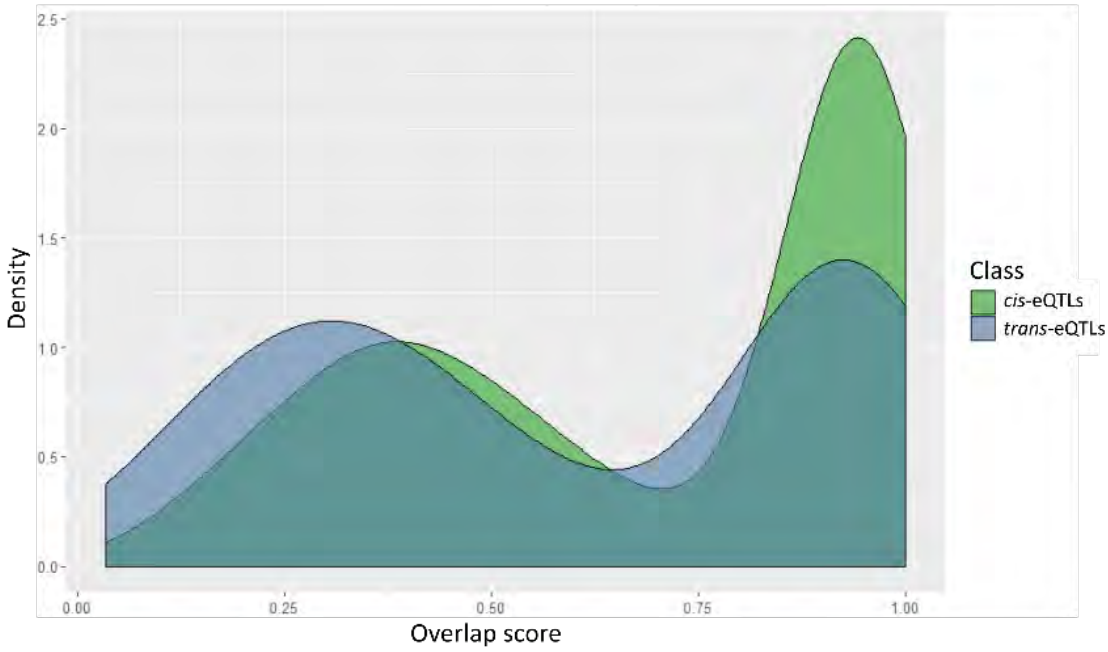
Supplementary Figure 2.7: Number of *trans*-eQTLs per gene in three- and eight-year-old xylem expressed genes. The number of *trans*-eQTLs per gene is compared between the three-year-old (green) and eight-year-old (blue) population.



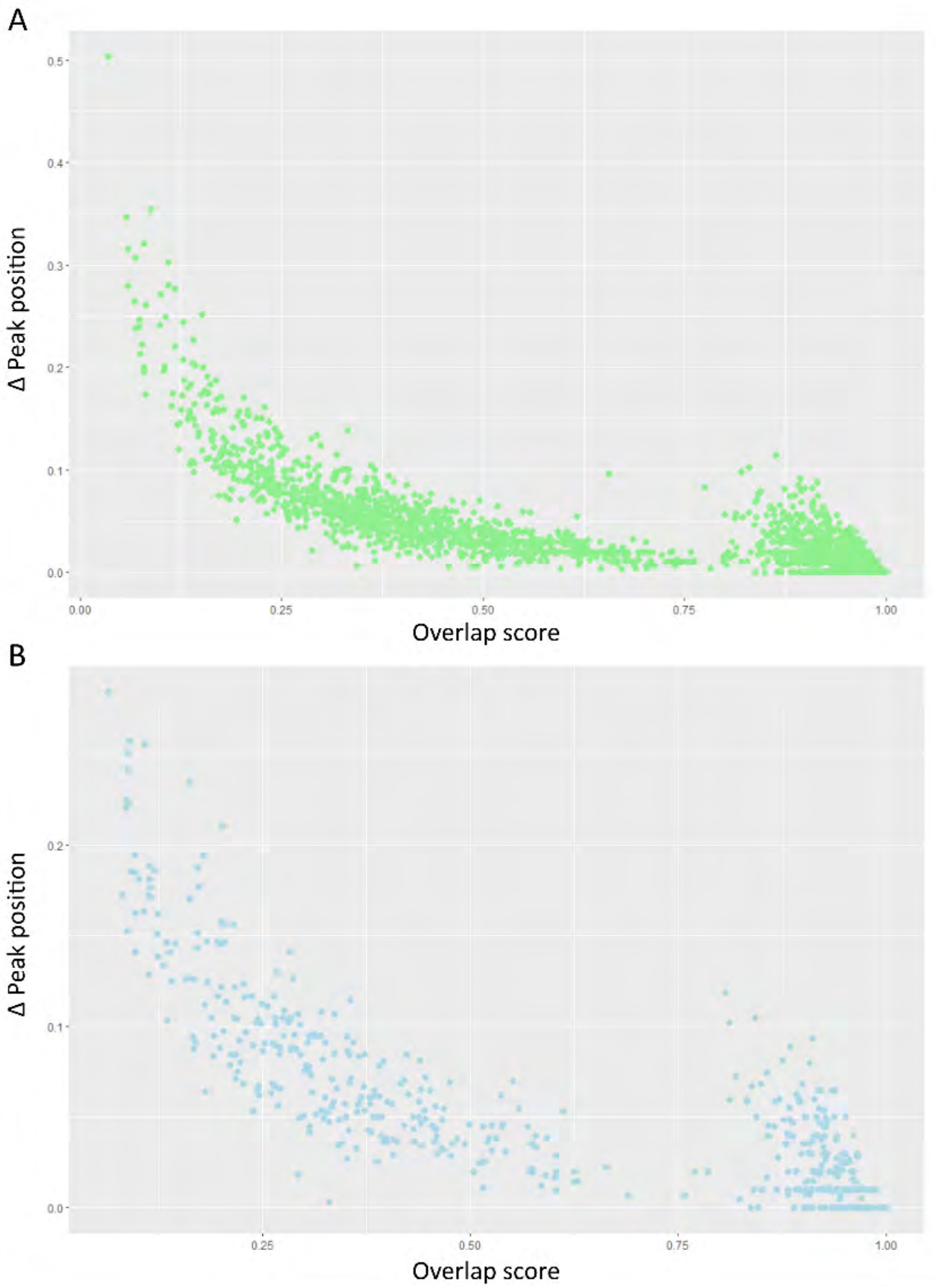
Supplementary Figure 2.8: Genome-wide *trans*-eQTL hotspots per population. The frequency of *trans*-eQTLs per cM is indicated by vertical bars. Red bars indicate possible *trans*-eQTL hotspots, where the frequency of eQTLs at that position is higher than expected to occur by chance (above a genome-wide permutation threshold, indicated by a horizontal black line).



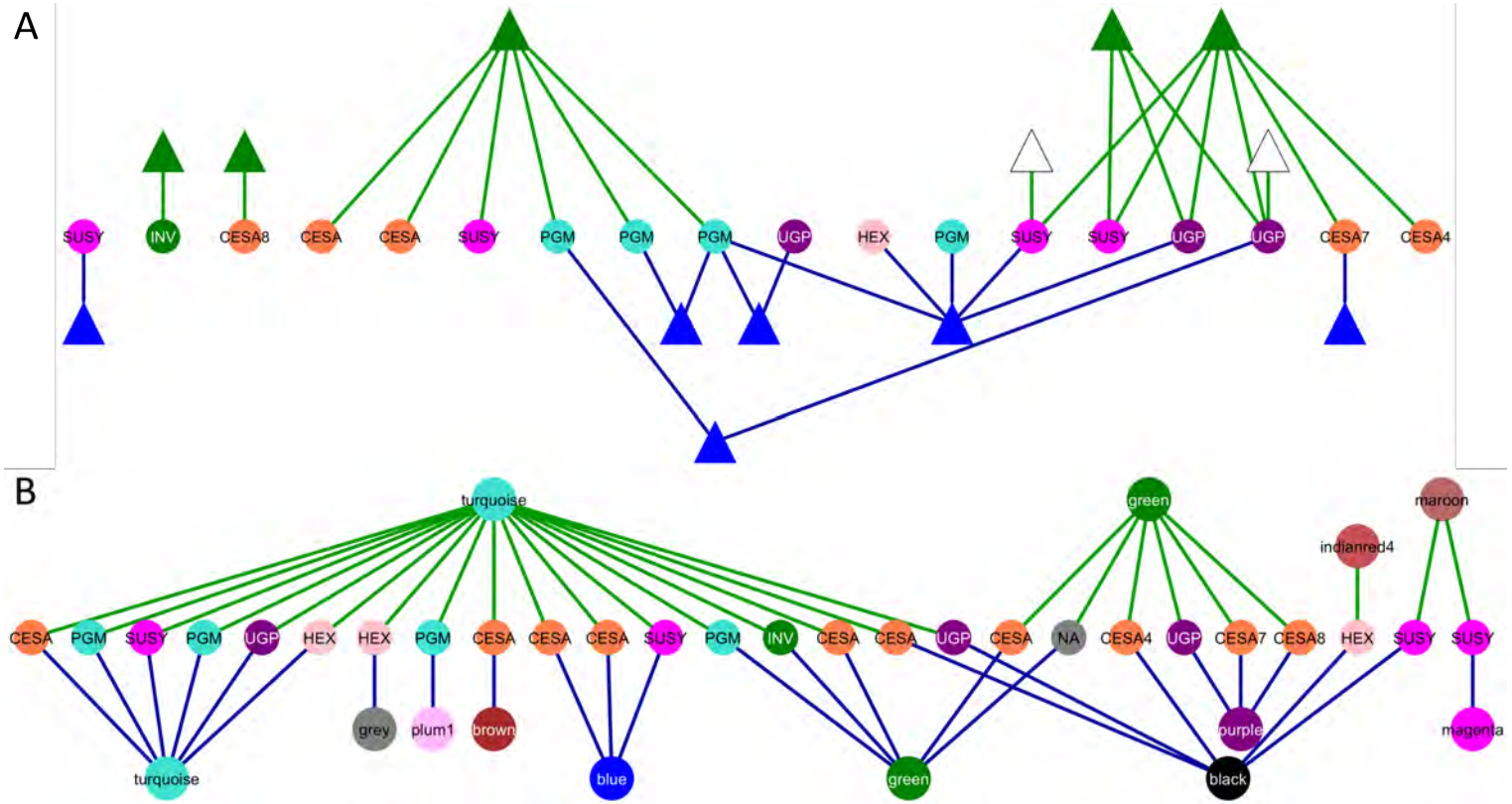
Supplementary Figure 2.9: *Trans*-eQTL hotspot density at two different ages. The Circos plot shows the density of *trans*-eQTL hotspots for 11 linkage groups (black) for the three-year-old (green) and eight-year-old (blue) population.



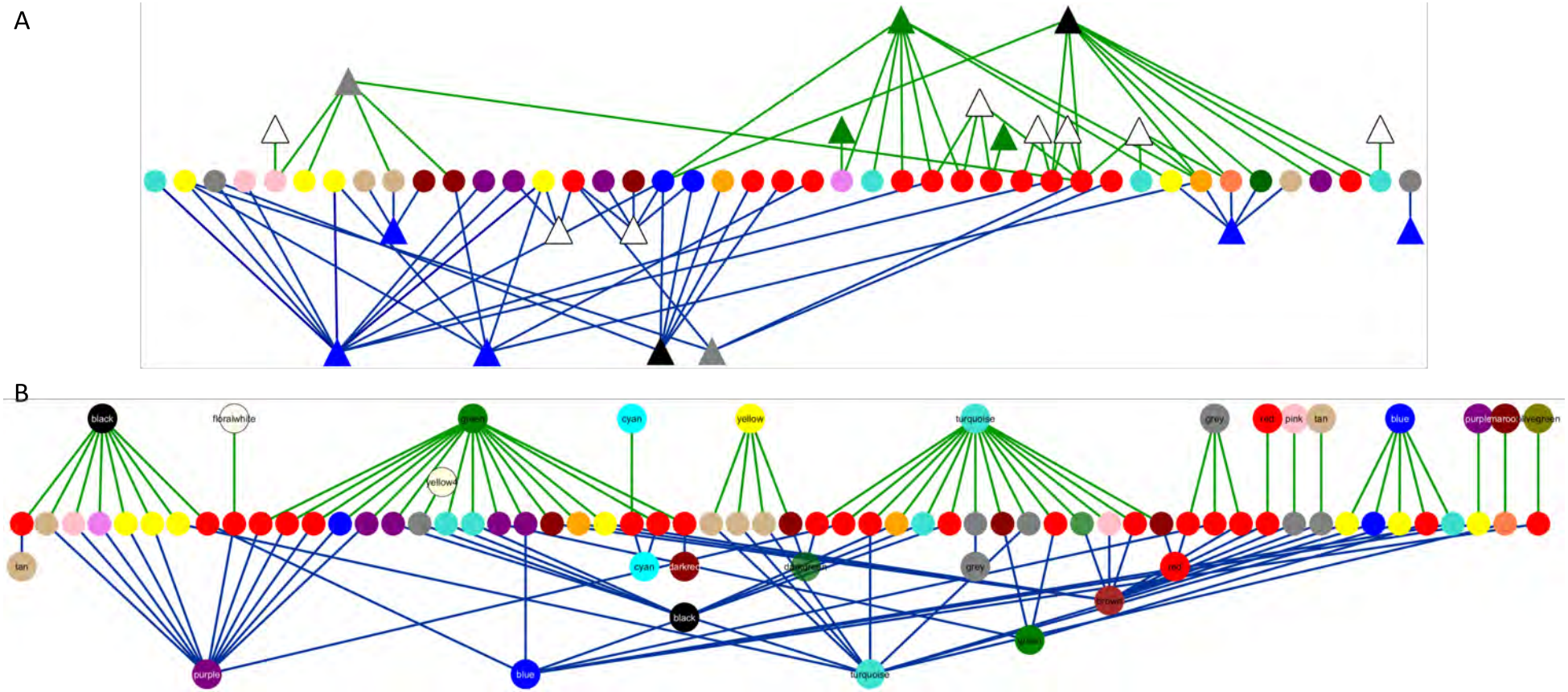
Supplementary Figure 2.10: Distribution of eQTL overlap scores between the three- and eight-year-old population per class. The density plots show the distribution of overlap scores that were calculated per gene between the three- and eight-year-old population for *cis*-eQTLs (green) and *trans*-eQTLs (blue). Overall, *cis*-eQTLs are more conserved across age than *trans*-eQTLs. The bimodal distribution can be explained by the relationship between the change in the eQTL peak position and the overlap score, as illustrated in the following figure.



Supplementary Figure 2.11: Relationship between overlap scores and the changes in peak positions between three- and eight-year-old individuals. For every conserved eQTL, an overlap score was calculated and compared to the change (Δ) in the peak position of the eQTL across age, for **(A)** *cis*-eQTLs (2,589) and **(B)** *trans*-eQTLs (506).



Supplementary Figure 2.12: Genetic architecture of cellulose genes across age. (A) Shared eQTLs (triangles) and **(B)** gene module membership (larger circles) are shown for cellulose genes (smaller circles) between three-year-old individuals (green edges) and eight-year-old individuals (blue edges). Genes are labelled according to their enzyme names in the cellulose pathway. Coloured triangles indicate that the shared eQTLs fall within *trans*-eQTL hotspots for some of the genes. CESA 4, 7 and 8 are secondary cell wall related proteins (Taylor-Teeples *et al.*, 2015).



Supplementary Figure 2.13: Genetic architecture of xylan genes across age. (A) Shared eQTLs (triangles) and **(B)** gene module membership (larger circles) are shown for xylan genes (smaller circles) between three-year-old individuals (green edges) and eight-year-old individuals (blue edges). Genes are labelled according to their enzyme names in the xylan pathway. Coloured triangles indicate that the shared eQTLs fall within *trans*-eQTL hotspots for some of the genes. The two black triangles represent the same eQTL on chromosome 10 in both populations and the two grey triangles represent the same eQTL on chromosome 6 in both populations.

Supplementary Table 2.1: Summary of genes with eQTLs in the three-and eight-year-old population

Population	**Three-year-old**		**Eight-year-old**	
Number of genes with eQTLs	17,559	(70.6%)	18,308	(72.5%)
Number of genes with *cis* or *trans* eQTLs	16,735	(67.3%)	17,411	(68.9%)
• Number of genes with only *cis*-eQTL	8,911	(53.2%)	6,695	(38.5%)
• Number of genes with only *trans*-eQTL	12,033	(71.9%)	14,245	(81.8%)
• Number of genes with *cis*- and *trans*-eQTLs	4,209	(25.2%)	3,529	(20.3%)
Average number of eQTLs per gene	1.6		1.6	
Average number of *cis*-eQTLs per gene with *cis*-eQTLs	1		1	
Average number of *trans*-eQTLs per gene with *trans*-eQTLs	1.4		1.5	

2.9.2 Supplementary Notes

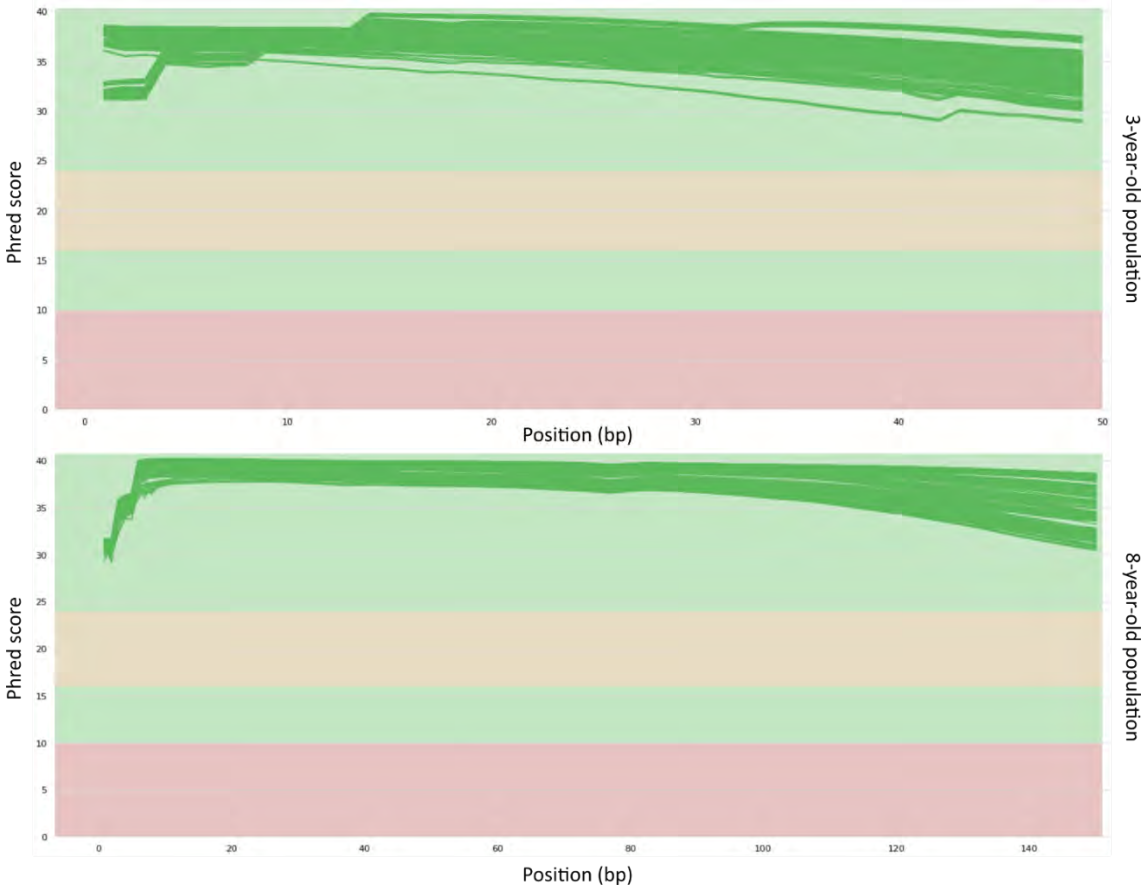
Supplementary Note 2.1: RNA-seq data quality control and expression profiling

We analysed transcriptome data from developing xylem tissues for 156 three-year-old and 144 eight-year-old individuals of the *E. urophylla* backcross population. All samples passed a quality test with very good Phred scores (**Supplementary Note Figure 2.1**), eliminating the need for read trimming. An average of 86.1% and 90.5% of reads mapped uniquely to the *E. grandis* reference genome for the three- and eight-year-old samples respectively (**Supplementary Note Table 2.1**). To gain insight into the technical repeatability of our sample set, four random samples from the eight-year-old population were sequenced in duplicate. A Pearson correlation value close to 1.0 was observed for the two replicates of all four individuals (**Supplementary Note Figure 2.2**), showing that the transcriptome data was quantified in a repeatable fashion, which allows us to analyse biological variation without the need to adjust for variation caused by technical factors.

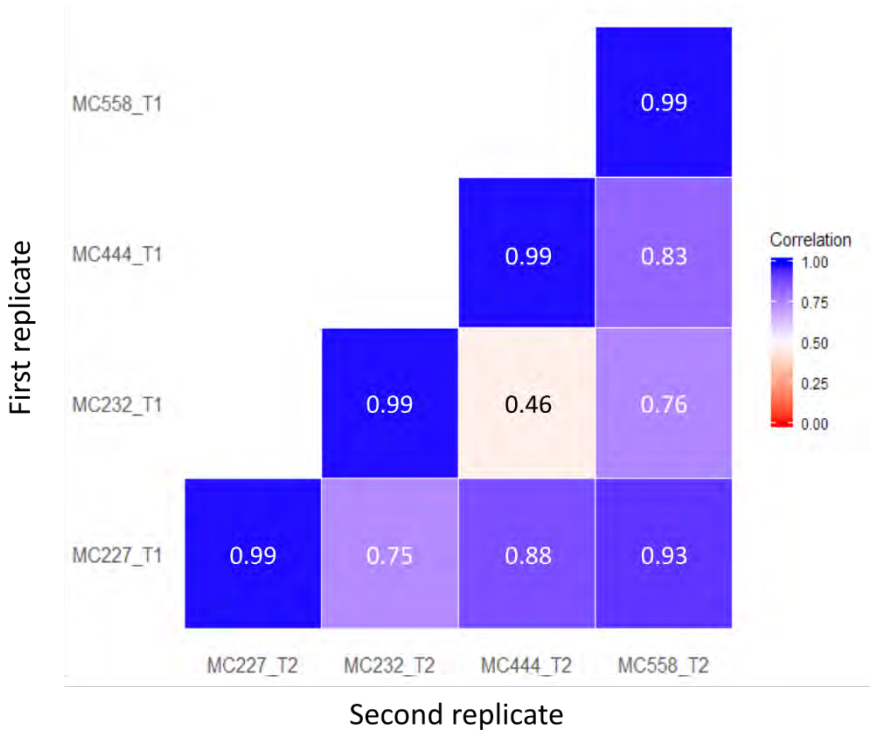
As the transcriptome data of the eight-year-old population had not been analysed before, an identity-by-descent (IBD) analysis was performed to confirm whether all the samples belonged to the same backcross family and to eliminate any possible clonal replicates. This was done by comparing previously generated SNP chip (EuCHIP60K) (Silva-Junior *et al*., 2015) genotype data with the transcript-derived SNPs (**Supplementary Note Figure 2.3**), as well as performing an all-to-all comparison among samples based on the transcript-derived SNPs (**Supplementary Note Figure 2.4**). Nine samples were removed from the dataset, of which six were clonal replicates, two did not belong to the population, and one seemed to be a technical replicate. To identify possible outliers, we examined the distribution of the remaining 135 transcriptomes and their correlations. No clear outliers could be identified based on the distribution of absolute expression levels (**Supplementary Note Figure 2.5**), however, based on the transcriptome correlations we could see that some individuals had lower correlations with the remaining samples (**Supplementary Note Figure 2.6**). A principal component analysis was performed to better visualise the set of uncorrelated samples, where the first

principal component explains 81% of the observed variation between samples and the second principal component explains 11% of the observed variation between samples (**Supplementary Note Figure 2.7**). Samples were then clustered based on the correlation of their expression profiles and we observed two clear outlier groups that were removed for downstream analyses by using a branch cut at a height of 25,000 (i.e. the first big separation between clusters), which represents the Euclidean distance between clusters based on correlation values (**Supplementary Note Figure 2.8**). This decreased the sample size of our eight-year-old population to a final set of 100 individuals. For future studies, the population size will be increased to analyse the variation between these groups and possible causes thereof.

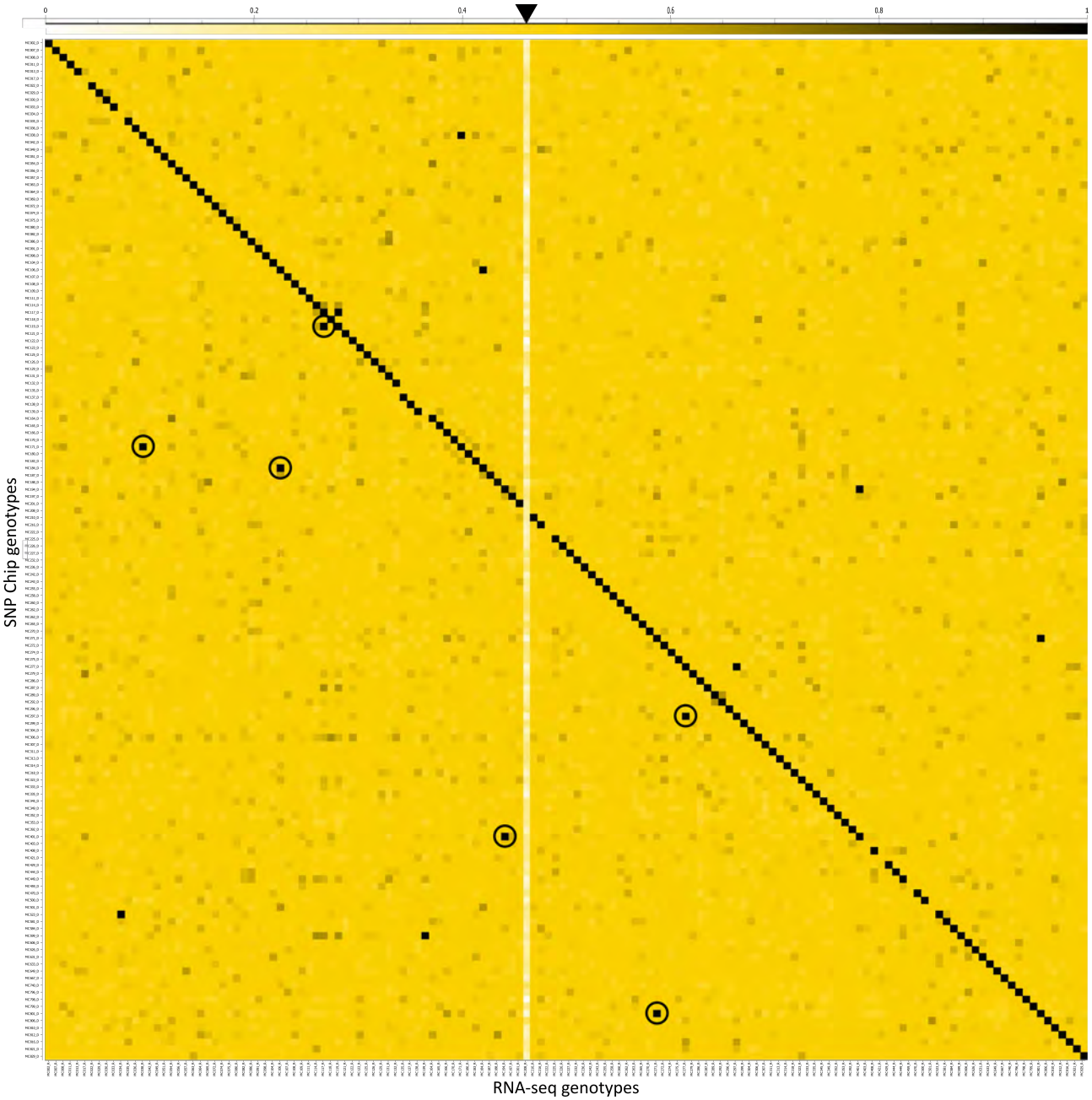
The final datasets for co-expression and eQTL analyses consisted of 24,861 genes for the three-year-old population and 25,267 genes for the eight-year-old population (**Supplementary File 2.1**). A total of 22,885 genes in these datasets are shared between the two populations, with unique genes in each population having seemingly low expression levels (**Supplementary Note Figure 2.8**).



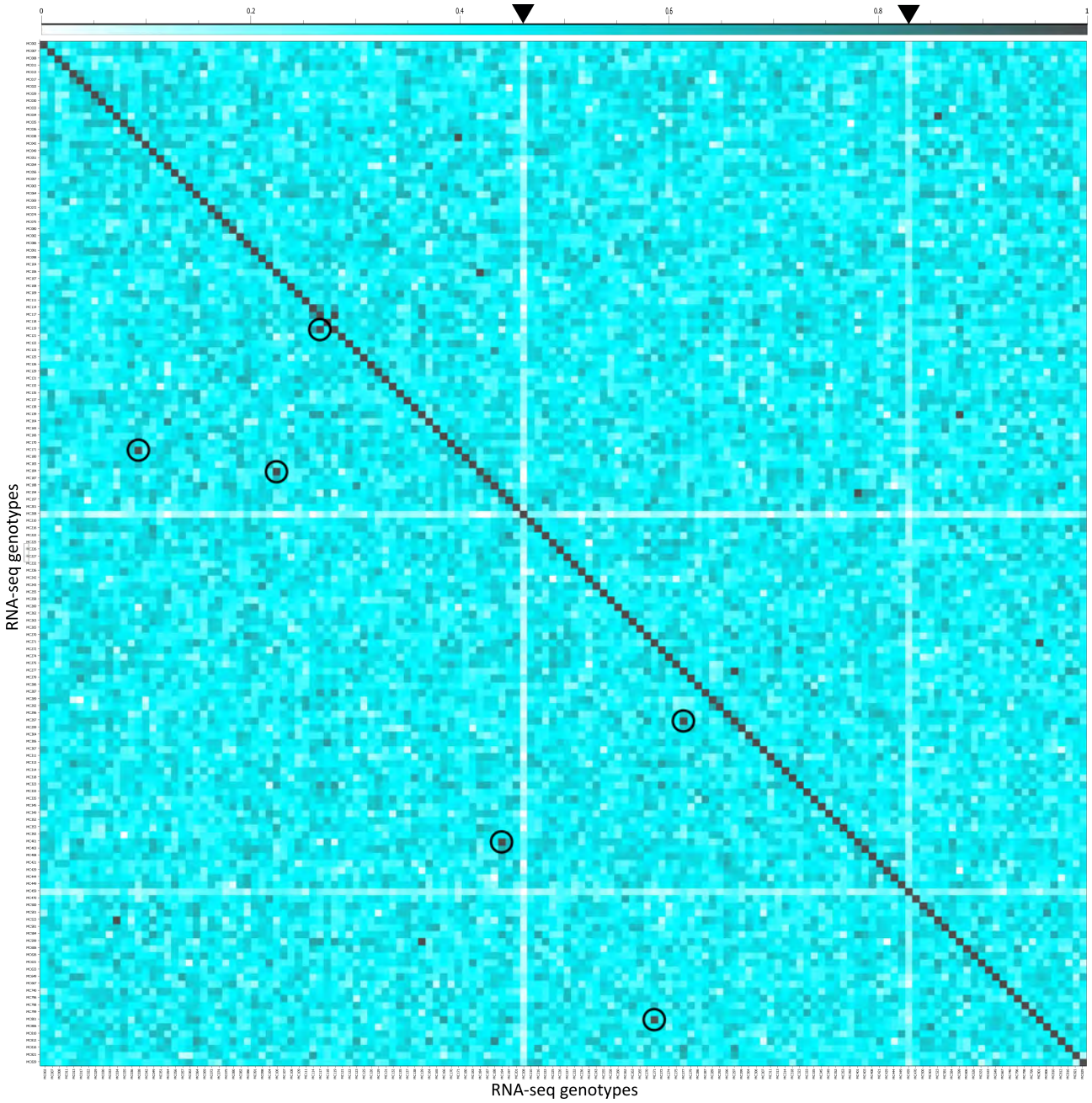
Supplementary Note Figure 2.1: Sequence quality histograms showing the mean quality across each base position (bp) in the read. Phred score assigned per base shows the quality of the identification of nucleobases generated by sequencing. All samples passed the quality test with very good quality scores > 25 (green region) over the full length of the reads.



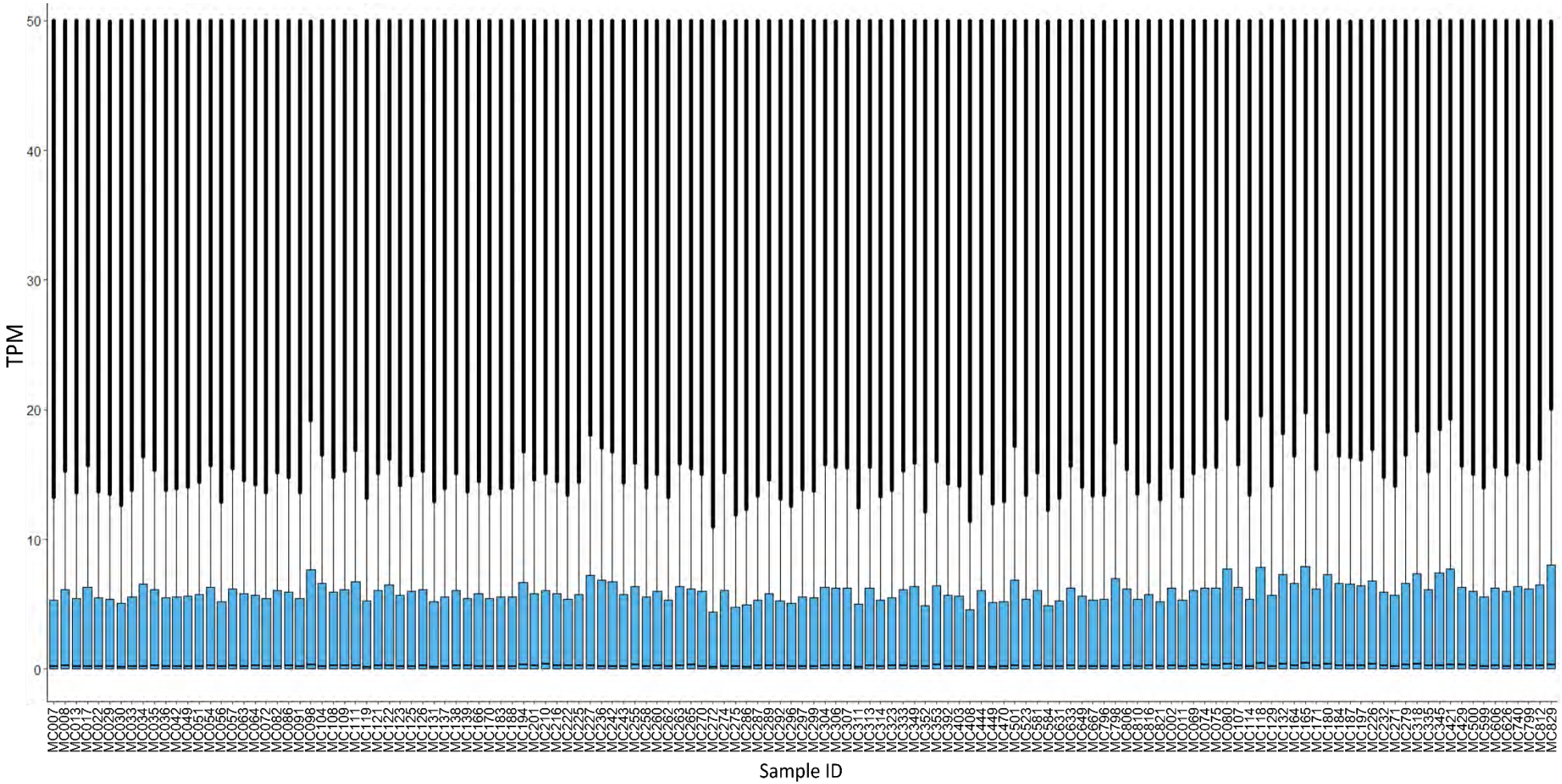
Supplementary Note Figure 2.2: Technical repeatability of four random xylem mRNA-seq samples in the eight-year-old backcross population. Pearson correlation values (*r*) were used to compare the absolute expression values between technical replicates. All four samples had a very high correlation ($r = 0.99$) between technical replicates.



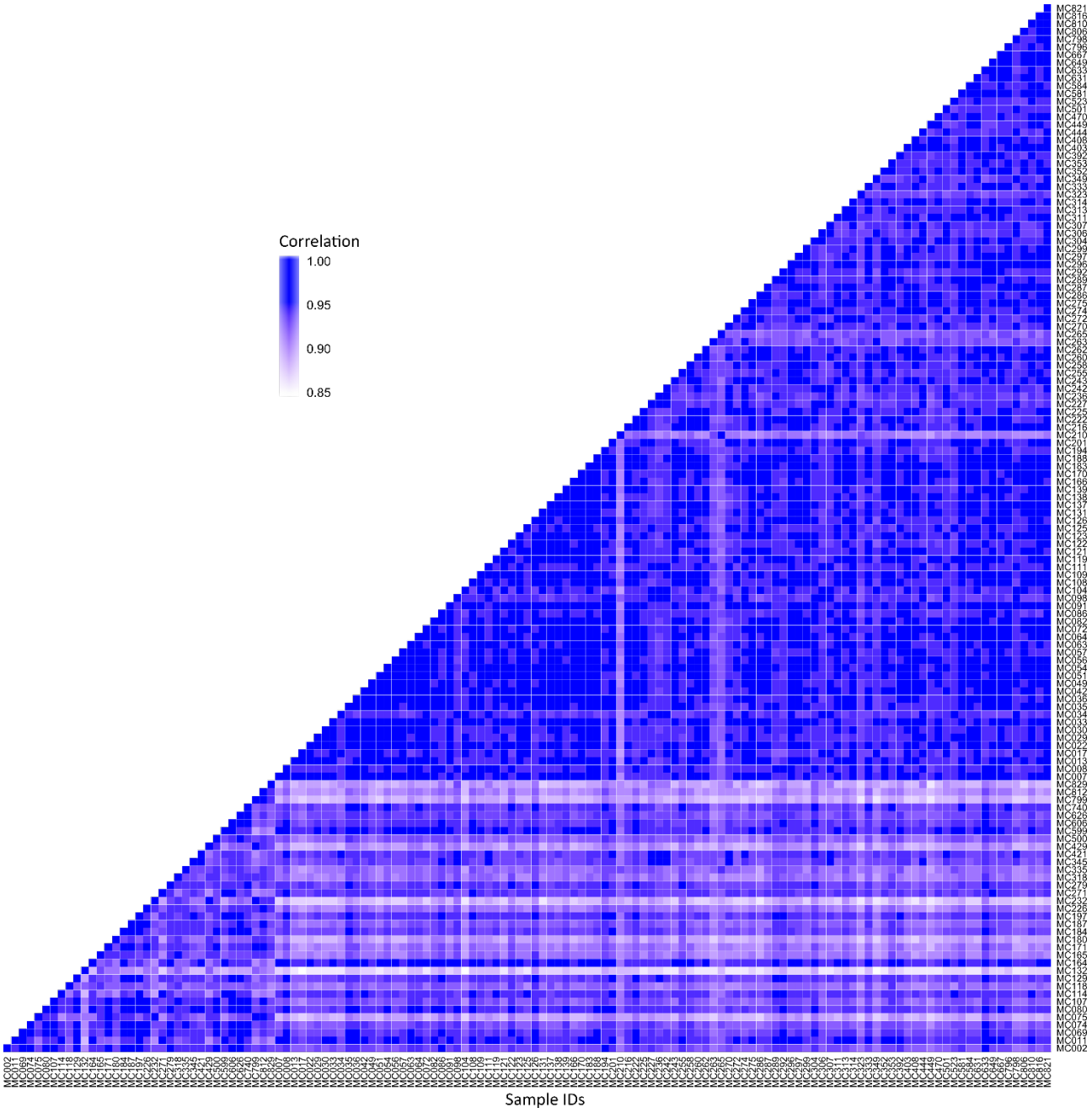
Supplementary Note Figure 2.3: Identity-by-decent (IBD) analysis results for 1,524 transcriptome-derived SNP genotypes vs. SNP chip genotypes of 144 eight-year-old individuals. Black squares indicate a full or approximately 1:1 match between the SNP chip and RNA-seq SNP genotypes of a sample. Possible clonal replicates (circles) can be identified where a sample matches both itself and another sample, as well as the RNA-seq SNP genotypes for both samples in the next figure. Yellow squares indicate an approximate 50% match, as expected from full-siblings, and white squares indicate a sample that does not belong to the backcross family (triangle). The sample that is not part of the progeny is MC208, which does not carry any of the F_1 hybrid parent alleles.



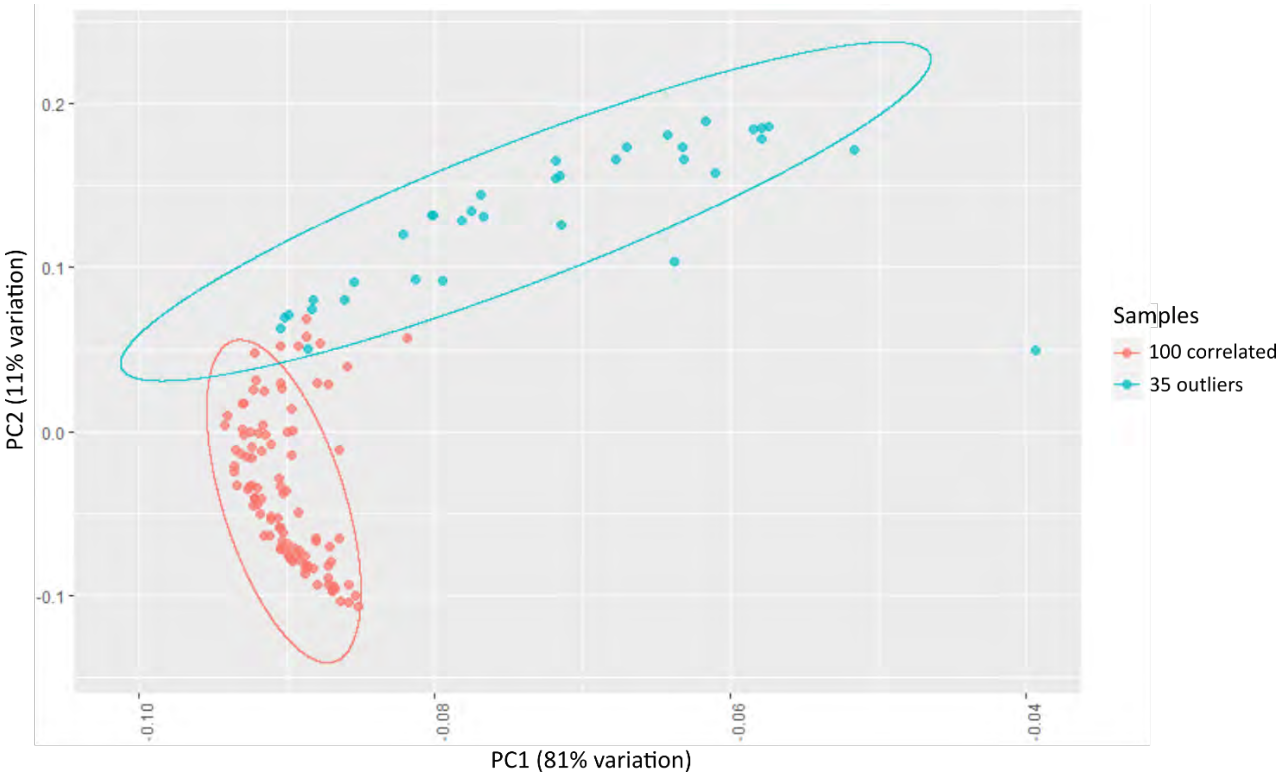
Supplementary Note Figure 2.4: Pairwise identity-by-decent (IBD) analysis results for 1,524 transcriptome-derived SNP genotypes of 144 eight-year-old individuals. Black squares indicate a full or approximately 1:1 match between the RNA-seq genotypes. Clonal replicates (circles) can be identified where a sample matches both itself and another sample, as well as the SNP chip genotypes for both samples from the previous figure. Blue squares indicate an approximate 50% match, as expected from full-siblings, and white squares indicate a sample that does not belong to the population under study (triangles). The two samples that are not part of the progeny are MC208, which does not carry any of the F_1 hybrid parent alleles, and MC459, which matches the F_1 hybrid parent (A380).



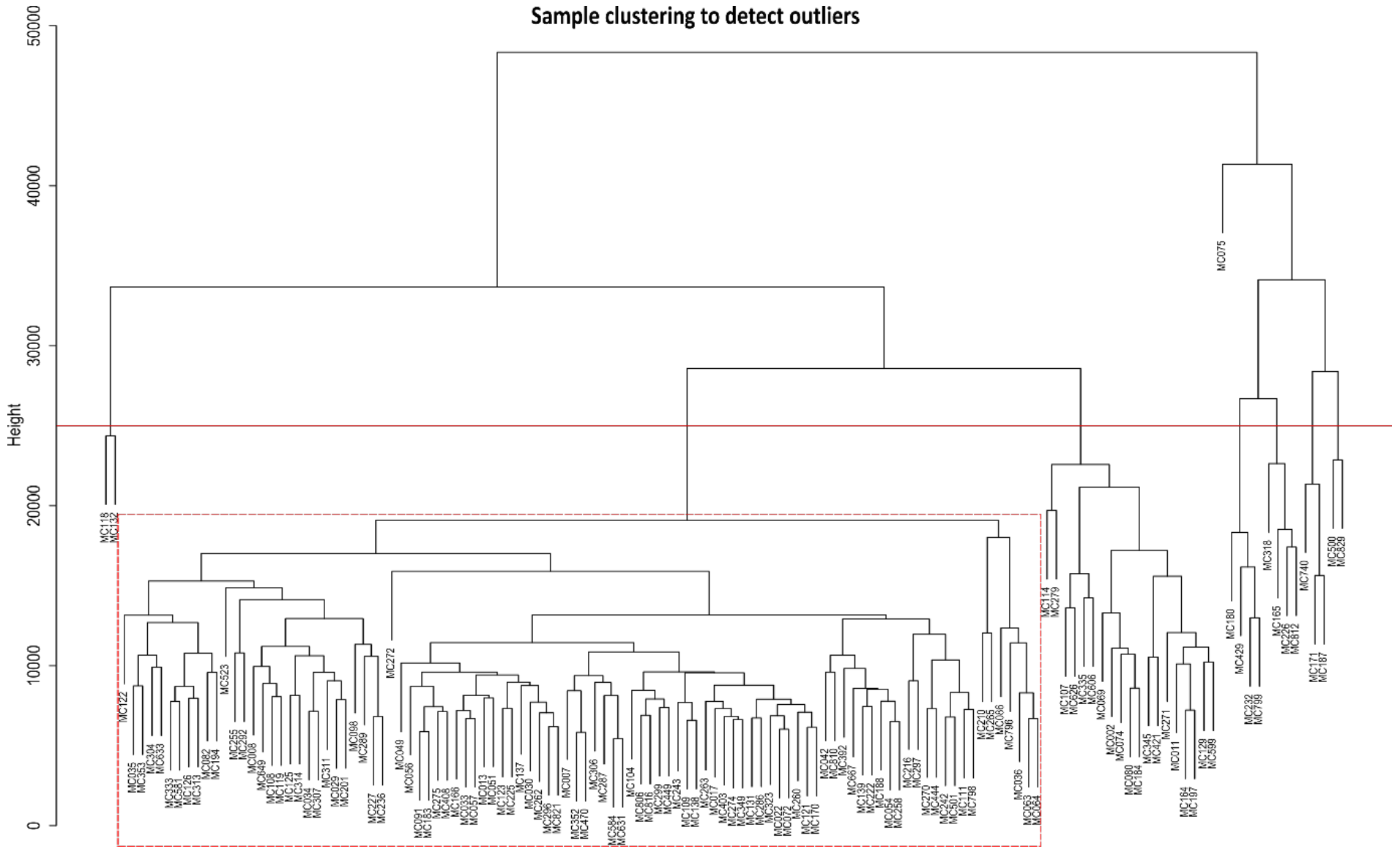
Supplementary Note Figure 2.5: Boxplots showing the distribution of TPM values of 36,349 genes for 135 individuals from the eight-year-old population. The y-axis is limited to TPM = 50 to remove noise and visualise the interquartile range (blue bars) of each sample to identify possible outliers.



Supplementary Note Figure 2.6: Heatmap showing all-by-all Spearman rank correlations of the expression levels of 36,349 genes for 135 individuals from the eight-year-old population. Gene expression profiles were compared between samples and visualised to identify outlier groups that may have an influence on downstream analyses.

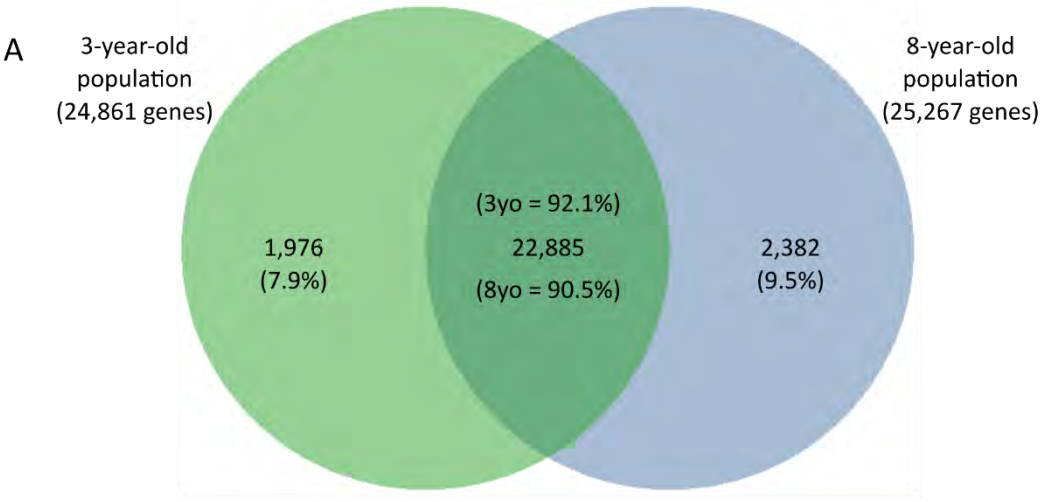


Supplementary Note Figure 2.7: Principal component analysis scatterplot of 135 individuals from the eight-year-old population, calculated from the expression levels of 36,349 genes. The first principal component (x-axis) explains 81% of the observed variation between samples and the second principal component (y-axis) explains 11% of the observed variation between samples. Ellipses represent the confidence interval for each group of samples.



Supplementary Note Figure 2.8: Dendrogram showing the clustering of 135 eight-year-old samples based on their gene expression profiles. A cut-off height (solid red line) of 25,000 (i.e. the first big separation between clusters) was used to identify outlier groups. The height represents the Euclidean distance between clusters based on correlation values. The final group of 100 samples used in downstream analyses is highlighted (dotted red line).

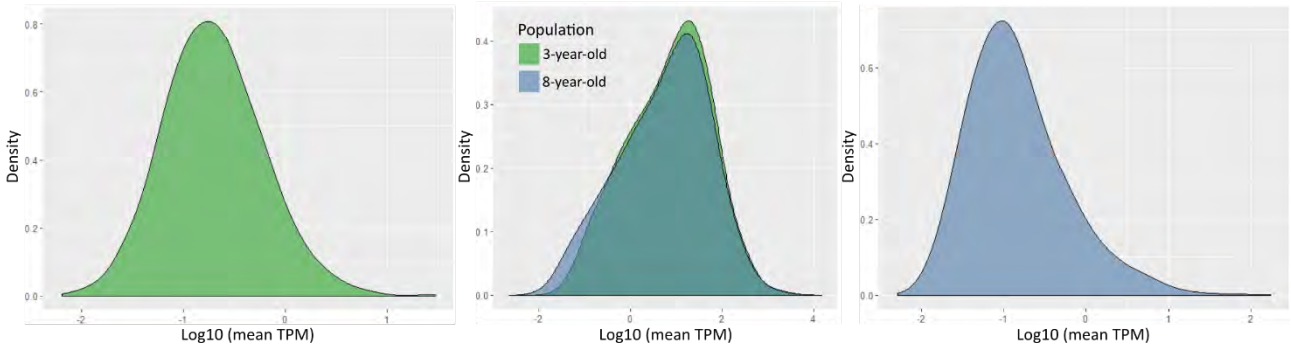
A



B

Mean TPM	0.42	43.63	0.69
Median TPM	0.05	7.43	0.04
SD (TPM)	1.5	209.9	7.4
Max TPM	76	22,674	726

C



Supplementary Note Figure 2.9: Gene overlap between three- and eight-year-old datasets. A) The number of genes shared between the three-and eight-year-old datasets (22,885) is shown in the overlapping region in the Venn diagram. **B**) A few descriptive statistics for the genes unique to each dataset and the overlapping genes are summarised in the table with corresponding colours. **C**) The log-transformed distributions of transcript abundance for genes unique to the three-year-old population (first panel), genes overlapping between the two populations (second panel), and genes unique to the eight-year-old population (last panel) are shown.

Supplementary Note Table 2.1: Summary of RNA-seq mapping quality per population. Reads were aligned to the reference genome of *E. grandis* v2.0 using STAR

Population	Number of individuals	Average read length	Average number of input reads	Average uniquely mapped reads (%)	Average mismatches per base (%)
Three-year-old	156	49 bp	25 million	86%	0.90%
Eight-year-old	144	150 bp	36 million	91%	0.93%

2.9.3 Supplementary Files

Supplementary File 2.1: TPM Values
This file contains Transcript Per Kilobase Million (TPM) values for genes expressed in at least 25% of the population. Three worksheets are presented in this file: (1) a legend for the file; (2) TPM values of 24,861 genes for 156 individuals in the three-year-old E. urophylla backcross population; and (3) TPM values of 25,267 genes for 100 individuals in the eight-year-old E. urophylla backcross population.

Supplementary File 2.2: Heritability Values
This file contains information on the broad-sense heritability (H^2) values of genes for non-clonal and clonal pairs. Four worksheets are presented in this file: (1) a legend for the file; (2) non-clonal correlations, clonal H^2 values, and average TPM values of 25,307 genes; (3) GO enrichment of genes with heritability values in the top 10% and bottom 10%; and (4) heritability values of genes involved in xylogenesis.

Supplementary File 2.3: Genetic Linkage Map
This file contains information on the genetic linkage map constructed from transcriptome data. Each marker can be represented by a gene and the physical and genetic positions are given for each marker, along with descriptive statistics of the read coverage per marker and overall coverage. The significance of distortion is indicated per marker, as well as the genotypes for three-year-old and eight-year-old individuals. See bottom of table for footnotes.

Supplementary File 2.4: eQTL Analysis Results
This file contains eQTL results per population and eQTL overlap scores between populations. Four worksheets are presented in this file: (1) a legend for the file; (2) eQTL results for three-year-old population; (3) eQTL results for eight-year-old population; and (4) eQTL overlap results between three- and eight-year-old population. See bottoms of tables for footnotes.

Supplementary File 2.5: Fisher's Test & GO Enrichment Results
This file contains results of Fisher's tests and GO enrichment analyses for the three- and eight-year-old population. Nine worksheets are presented in this file: (1) a legend for the file; (2) GO enrichment terms for gene modules and hotspots in the three-year-old population; (3) GO enrichment terms for gene modules, hotspots, split-hotspots, and module-hotspot overlaps in the eight-year-old population; (4) GO enrichment terms for gene sets unique to each population and differentially expressed genes with a log2 fold-change > 2; (5) Fisher's test results for significant overlaps between modules for the two populations and between the eight-year-old modules and hotspots; (6) a matrix showing the number of genes overlapping between modules for the two populations, with significant overlaps highlighted in yellow; (7) a matrix showing the number of GO terms overlapping between modules for the two populations, with significant overlaps highlighted in yellow; (8) a matrix showing the number of genes overlapping between modules and hotspots for the eight-year-old population, with significant overlaps highlighted in yellow; and (9) a matrix showing the number of genes overlapping between modules and split-hotspots for the eight-year-old population, with significant overlaps highlighted in yellow.

Supplementary File 2.6: Gene Information
This file contains general information on genes for both populations. Four worksheets are presented in this file: (1) a legend for the file; (2) information on 24,861 genes for the three-year-old population; (3) information on 25,267 genes for the eight-year-old population; and (4) a list of genes involved in xylogenesis and stress response, secondary cell wall transcription factors and transcription factors involved in stress response.