Norwegian University of Life Sciences

Master's Thesis 2024 60 ECTS
Faculty of Chemistry, Biotechnology and Food Sciences

Comparative transcriptomics of wood formation in angiosperms and gymnosperms

Ellen Dimmen Chapple
Bioinformatics and applied statistics

Acknowledgements

This thesis is written as part of my master's degree in bioinformatics at the Norwegian University of Life Sciences, Department of Chemistry, Biotechnology and Food Sciences.

First off, I would like to thank my main supervisor, Torgeir R. Hvidsten, for his guidance and support throughout this thesis, but also for his contagious enthusiasm. I would also like to thank my co-supervisor, Siri Birkeland, for her invaluable feedback and suggestions.

Also, a note of appreciation to my fellow BIAS-students – these past two years would have been significantly more challenging without their companionship. I would further like to thank the BIAS-group for creating a safe and welcoming academic space.

To my family, I would like to say thank you for being a great support system, and for showing an interest in my work. Finally, I would like to thank my good friend, Erik, for being an academic inspiration – but mostly for keeping me sane.

Abstract

Wood, or secondary xylem, is formed during secondary growth and serves as a structural element as well as part of a plants vascular system. Secondary xylem is largely comprised of cells with lignified secondary cell walls to which the rigidity of the tissue is attributed. Secondary growth is observed throughout the plant kingdom but is most prominent in trees. Trees are seed-producing, vascular plants that are divided into two groups based on their method of reproduction: angiosperms and gymnosperms. Most tree species belong to the group of flowering plants, the angiosperms, which are estimated to have evolved 200 million years after the first gymnosperms. Despite this ancient divergence, wood formation as a phenotype is conserved between the gymnosperms and angiosperms. Much is understood about the various processes involved in secondary growth, but little is known about the more ancient, underlaying mechanisms governing wood formation. For this thesis, high-resolution transcriptomics was used to capture the similarities in wood formation across three gymnosperm species and three angiosperm species. Orthologs conserved between the pairs of species were identified using co-expression network analysis, thereby combining similarity in molecular function with biological function. Orthologs conserved across all pairs of species were identified using sub-networks, cliques. The genes conserved across all species were involved in various processes associated with secondary growth, and certain genes were suggested as marker gene candidates for the various tissues. These genes included homologs of PHLOEM PROTEIN2 (PP2), CELL DIVISION CONTROLL2 (CDC2) and transcription factors for formation of actin filaments.

Sammendrag

Vedvev, eller sekundær xylem, blir dannet under sekundær tykkelsesvekst og har en strukturell funksjon i tillegg til at det er del av plantens vaskulære system. Sekundær xylem består hovedsakelig av celler med lignifiserte sekundære cellevegger som gir vevet sin rigiditet. Sekundær tykkelsesvekst er observert på tvers av planteriket, men er mest utbredt hos trær. Trær er frøproduserende, vaskulære planter som deles inn i to grupper basert på reproduksjonsmetode: angiospermer og gymnospermer. De fleste treartene tilhører blomster planter, angiospermer, som er estimert å ha evolvert ca. 200 millioner år etter de første gymnospermene evolverte. På tross av denne tidlige divergensen, så er veddannelse som phenotype konservert mellom gymnospermer og angiospermer. Mye av prosessene involvert i sekundær tykkelsesvekst er studert, men lite av de underliggende mekanismene som styrer veddannelse er kjent. I denne oppgaven ble genuttrykksdata med høy romlig oppløsning brukt til å fange likheter i veddannelse hos tre gymnospermarter og tre angiospermsarter. Ortologe gener konserverte mellom artspar ble indentifisert ved bruk av nettverksanalyse av samuttrykte gener, derav ble likhet i molekylær funksjon kombinert med likhet i biologisk funksjon. Ortologer som var konservert på tvers av alle artspar ble identifisert ved bruk av cliquer. Genene som var konserverte på tvers av alle artene var involvert i flere prosesser innen sekundær tykkelsesvekst, og et utvalg gener ble foreslått som markørgener for de ulike vevene. Disse genene inkluderte bl.a. homologe gener av PP2, CDC2 og transkripsjonsfaktorer for aktinfilamentdannelse.

Table of Contents

1 Introduction

Throughout human history, wood has been an invaluable resource. Today, the increasing need for sustainable and renewable materials has furthered the exploitation of wood as a multipurpose biomass, from modifying wood materials chemically or thermally to achieve desired construction properties, to generating biofuels (Ashokkumar et al., 2022; Ramage et al., 2017). The structure of wood, and thereby its properties, vary between species. Gaining insight into the genetics that produce the various wood types will allow for engineering of desired wood properties.

The earliest land plants arose around 470 million years ago (mya), and as more species competed for daylight vertical growth became a key function for survival (Gensel, 2008). Vertical growth required the development of attributes such as the ability to grow against gravity and towards light (photo- and gravitropism), and the ability for cell division to occur at specific locations (apical growth) (Reece & Campbell, 2011). In addition, vertical growth also required a sturdy stem for support and a means to transport nutrients and water, i.e. a vascular system (Zhong et al., 2019).

A common denominator for the development of both a vascular system and a sturdier stem, is cells that feature lignified secondary cell walls (SCWs)(Weng & Chapple, 2010). All plant cells are shaped and protected by a *primary* cell wall, but certain specialised cells such as tracheids (vascular) and fibres (structure) also develop a second, more rigid cell wall (Zhong et al., 2019). The SCW is formed between the plasma membrane and primary cell wall (PCW) in mature cells. A mature cell has undergone cell expansion, a morphological process in which the cells' volume increases. The higher rigidity of the SCW-forming cells is a result of the structural matrix of the second wall. Both PCW and SCW are primarily made up of cellulose and hemicelluloses, with additional components such as proteins and other smaller molecules. Cellulose, a structural isomer of starch, is a polysaccharide consisting of bundles of unbranched beta-glucose polymers (Meents et al., 2018). Hemicelluloses, such as xylans and mannans, are also classified as polysaccharides, but differ from cellulose through displaying diversity in both structure and base-sugars. The increased rigidity of the SCW, however, is achieved by an additional feature, lignin. Lignin is a complex molecule synthesised through the combining of phenylpropanoids, and acts as a binding agent, strengthening the cellulose and hemicellulose matrix (Zhong et al., 2019). Additionally, lignin has antimicrobial properties which reduces decomposition of dead cells, and its hydrophobicity creates a foundation for water conduction (University of Oslo, 2011).

Although present in multiple tissues, the primary source of SCW cells is secondary xylem (Zhong et al., 2019). Secondary xylem, commonly referred to as wood, is produced in most vascular plants during secondary growth, a process in which the plant grows laterally instead of vertically (Růžička et al., 2015). In combination with the phloem, the secondary xylem comprises the plants vascular system transporting minerals and water from the roots (secondary xylem) and biosynthesis products from the leaves (phloem). During secondary, growth, bifacial cambium, consisting of stem cells, provides new cells to the secondary xylem (inwards), and to the phloem (outwards) through two-directional differentiation (Shi et al., 2019). The secondary xylem cells gradually mature through cell expansion and the formation of the lignified secondary cell walls, before undergoing programmed cell death (Sundell et al., 2017). Mature phloem cells, on the other hand, add to the radial growth of the stem but remain alive.

Woodiness is found throughout the plant kingdom, present within almost half of all identified vascular plants (FitzJohn et al., 2014). The largest source of wood, however, are trees. Although not a monophyletic group, “tree” is a term used for predominantly tall seed-producing plants characterised by a sturdy, woody stem (trunk) and with varying type of foliage. Extant tree species belong to one of two monophyletic groups, or clades, of seed plants, namely, angiosperms or gymnosperms.

Seed-producing plant evolved around 170 million years after the first recorded land plants evolved (Sanderson, 2003). The first seed plants are the ancestors of present-day gymnosperms families. Gymnosperms are, generally, evergreen trees with foliage consisting of needles and cones, and with flaky bark, and derive their clade-name, “naked seeds”, from their seeds being stored in cones. Fast-forward 200 million years, an alternative reproductive method evolved: in contrast to the bearing seeds in cones, angiosperm encapsulate mature seeds within fruits. Angiosperms, also referred to as flowering plants, form the most abundant and diverse group of extant plants.

In addition to reproductive methods, angiosperms and gymnosperms also show differences in wood composition (Schmulsky & Jones, 2019). An important evolutionary innovation seen in Angiosperms are vessels, which may serve as more efficient for water transportation than the gymnosperm tracheids. The composition of the secondary cell walls is also different between the two clades. The relative amounts of lignin, cellulose, and hemicelluloses in the SCW vary not only between cell types, but also between different species (Meents et al., 2018). Typically, angiosperms have a higher lignin content than gymnosperms (Schmulsky & Jones, 2019) and tend to be richer in xylans while gymnosperms are richer in mannans (Berglund et al., 2020). The composition of the SCW is reflected in the wood properties, with gymnosperms often being referred to as *softwoods* due to a less dense wood matrix compared to the denser *hardwoods*, angiosperms (Schmulsky & Jones, 2019). This classification is a generalisation but is one of many aspects motivating further exploration into the similarities and differences in wood formation between gymnosperms and angiosperms.

Various aspects of wood formation have been studied. Due to the complexity of the different cellular processes within the various tissue, most research associated with wood formation has been tissue-specific, and within single species. These studies have shed light on the regulatory system coordinating the establishment of secondary cell walls (Zhong et al., 2009), genes regulating xylem differentiation and development (Kubo et al., 2005; Xu et al., 2019), and the mechanisms controlling the synthesis of cellulose (Xie et al., 2011). The changes in gene expression between tissue types is understood using transcriptomic data, which captures the products of gene expression, mRNA, within the different tissues. Furthermore, comparing expression between species can shed light on evolutionary changes across species. Identifying orthologous genes through sequence similarity allows for comparative transcriptomic studies across a range of species, potentially identifying conserved or diverging mechanisms. Most comparative studies on wood formation, however, have been focused on transferring knowledge from model plants, such as Arabidopsis thaliana, to another species (Kim et al., 2022; Wang et al., 2021), with only a few studies comparing angiosperms and gymnosperms (Jokipii-Lukkari et al., 2017; Li et al., 2021; Sundell et al., 2017). However, a study by Sundell et al. (2017) identified highly conserved gene expression across wood forming tissue between aspen (*Populus tremula*) and Norway spruce (*Picea abies*), motivating exploration into more developmental similarities between angiosperms and gymnosperms.

As a phenotype, wood formation is conserved between angiosperms and gymnosperms, and processes related to wood formation have been shown to be conserved across the two clades (Li et al., 2021). There is, however, a knowledge gap in understanding the ancestral mechanisms regulating wood formation in its entirety. Comparing multiple phases of secondary growth between

angiosperms and gymnosperm can potentially identify conserved mechanisms across the clades but can also reveal differences which might shed light on the different wood structures.

This thesis aims to outline some of the genetic similarities and differences in wood formation between angiosperms and gymnosperms using high-spatial resolution transcriptomic data and co-expression network analysis. High-spatial resolution transcriptomics is achieved through samples obtained through cryosectioning (Uggla et al., 1996) across differentiated phloem over the cambium, and into mature, lignified xylem, thus giving a detailed picture of how the expression of genes change throughout secondary growth. Linking transcriptomic data from a total of six species, both angiosperms and gymnosperms, using sequence based orthologs, can identify processes with conserved gene expression across the clades.

The aims for this study are two-fold: 1) to identify genes that are conserved between the clades, and genes that are uniquely conserved within clades, and 2) to explore if any of the conserved genes are potential marker genes for the various transition phases (between tissue types). The results achieved in this thesis will shed light over which biological processes associated with wood formation have been conserved through approx. 200 million years of evolution, as well as which processes may be unique to angiosperms and gymnosperms.

1.1 Background

1.1.1 Study system and workflow

A total of six species comprised the study system for this project: three gymnosperm, lodgepole pine (*Pinus contorta*), Scots pine (*Pinus sylvestris*), and Norway spruce (*Picea abies*), and three angiosperms, cherry (*Prunus avium*), aspen (*Populus tremula*), and birch (*Betula pendula*). All three gymnosperm species belong to the pine family (*Pinaceae*), with the two pine species belonging to the same genera (*Pinus*), thus being more closely related to each other than to Norway spruce (*Picea*) (Earle, 2024). The angiosperm species all belong to different orders but can be grouped in under the fabids clade (Stevens, 2024).

Cryosections, collected from tissues spanning differentiated phloem to SCW-forming xylem, were used to create high resolution transcriptomic data with the aim of capturing the various processes of secondary growth. To ascertain which genes showed conserved expression during secondary growth across all species, orthologous genes with conserved co-expression networks were used to identify conserved orthogroups (groups of orthologs) across pairs of species. Orthologous genes were predicted based on protein sequence, while co-expression was based on expression similarity. Orthogroups conserved specifically amongst gymnosperm or angiosperm pairs were also identified.

As a result of a several gene duplication events, plants tend to have sets of sequence similar genes that have adapted different molecular or biological functions (paralogs). Therefore, cliques, or sub-networks, were used to identify orthologous genes with a common ancestral gene within each of the conserved orthogroups, i.e. genes conserved across all six species. Gene ontology, expression profiles and genes identified in previous studies were thereafter used for a preliminary investigation into which biological processes the conserved genes were associated with.

1.1.2 Understanding biological processes

Cellular activity can be described as a network of proteins, metabolites, and molecules of all sizes interacting to activate and regulate biological processes. These processes are typically complex, involving several members, and are often studied at different levels, e.g. which genes are expressed, or which enzymes are activated or inhibited. A starting point for mapping out cellular processes is to collect data. The type of data, or omics, to be collected is determined by the research aims. For studying similarities and differences in structure and function of a genome, genomic data is of interest, while transcriptomics (RNA), proteomics (proteins) and metabolomics (metabolites) are examples of data describing cellular activity (Vailati-Riboni et al., 2017). All omics data can be used to study cellular changes, effectively describing the changes from different perspectives. Combining data means adding more levels (multi-omics) which in turns adds more detail to the map over cellular processes.

Transcriptomics

The first level that is typically studied is gene expression. Transcriptomics is the study of the transcriptome and comprises all types of RNA present within a cell at a given moment under given circumstances. Total RNA includes both coding RNA (mRNA) and non-coding RNAs (tRNA, rRNA, microRNA, lincRNA, etc). In eucaryotic cells, coding RNA holds the genetic transcript for a protein, while the non-coding RNAs are associated with various functional roles within translation, gene regulation and even epigenetics (Pikaard & Mittelsten Scheid, 2014). Studying a cell's transcriptome is of interest as it provides insight into which genes are expressed (or not expressed) under certain conditions or within different tissues. Based on the notion that sets of genes rather than single genes are active within most cellular processes, transcriptomics is commonly used for identifying co-expressed genes. Co-expressed genes show highly correlated expression across different tissues and can be considered a sub-network (module), in a larger co-expression network. Co-expression networks are a type of biological network with nodes representing genes, and edges representing co-expressed relationship. Building on the notion that similarly expressed genes are likely to be associated with the same biological processes, also referred to as 'guilt by association' (Wolfe et al., 2005), co-expressed genes can be used to infer biological function to novel genes (Emamjomeh et al., 2017). Furthermore, co-expressed genes can be applied to inter-species comparisons, thereby being a powerful tool for studying developmental and evolutionary differences between species (Birkeland et al., 2022; Ovens et al., 2021).

RNA-Seq

The past 20 years has seen an advancement in sequencing methodologies, giving rise to cheaper and more available technology. Next-generation Sequencing (NGS) techniques, such as Illumina sequencing, Ion Torrent sequencing and PacBio have sped up the sequencing process (high throughput) through automation and the capacity to run samples in parallel (Goodwin et al., 2016; Wang et al., 2009).

Methodologies for studying transcriptomic data has also evolved the last decades. Up till the 2010s microarrays were a main tool for studying expression data, or mRNA. Using hybridisation, gene expression levels can be quantified through the intensity of fluorescent labels in a high throughput

manner. However, microarrays require prior knowledge of the genes which the sampled products are to bind to. In addition, cross-hybridisation and difficulties with comparing different expression levels are limiting factors (Wang et al., 2009). Although still in use, microarrays have largely been replaced with RNA-Sequencing (RNA-Seq) as the preferred tool for transcriptomic analysis (van der Kloet et al., 2020).

With some variations, RNA-Seq technologies follow the same general steps from sample extraction to data output. The first step after RNA extraction, is library preparation where fragmented mRNA is translated into double-stranded complementary DNA (cDNA) through reverse transcriptase and used to create cDNA libraries. Before sequencing, additional steps such as amplification and adaptor ligation are performed. The sequencing step, in which the nucleotide order of each cDNA fragment is ascertained, results in reads. With NGS both short and long sequence reads can be achieved, depending on the research aims. Most RNA-Seq methods are designed for short reads mainly due to the desire to study gene expression (Stark et al., 2019), while longer reads are more useful for detecting structural differences or for de novo assembly (Mastrorosa et al., 2023). The RNA-Seq output is a FASTA file containing both the read sequence and a quality score rating how certain each nucleotide is (FASTQ).

As the reads are based on fragmented mRNA molecules, some assembly is required. At this point another advantage of using RNA-Seq becomes apparent as the reads can either be assembled by alignment to a known genome or transcriptome or be assembled de novo. With de novo assembly, no reference genome or transcriptome is required, and reads are assembled based on areas of overlap, forming contigs. For reads aligned to a reference genome or transcriptome, the number of reads overlapping the region of a gene are counted, producing read counts. Moreover, counting overlap of gene regions after alignment produces mapped reads. Alignment is not necessary for obtaining read counts and can be achieved with “raw” reads or using de novo contigs.

As the total number of reads per sample will vary due to a number of factors such as, sequencing depth, technical variability in library preparation and sequencing (McIntyre et al., 2011), or simply biological variations, it is crucial to normalise the read counts so that they are relative to the total number of reads within the sample. Within-sample normalisation takes into account variation of gene length, and there are two main methods for this: Reads Per Kilobases genes per Million mapped reads (RPKM) and Transcripts per kilobase Million (TPM). Both methods take into account the differences in gene length, but the TPM method normalises the gene length prior to normalising the sequence depth. Since the sum of TPM values for each sample will be the same, the proportion of a gene’s sequencing depth is easier to compare between samples.

Normalisation should also be performed between the samples to adjust for variations in library size between samples. This is typically done by calculating scaling factors for each sample. There are several methods for calculating scaling factors such as Total Count (TC), Median (M), DESeq2 and Quantile (Q) (Dillies et al., 2012). Furthermore, to handle the presence of extreme values as well as mean-variance dependency, variance stabilised transformation (VST) is typically used to spread out the expression values amongst genes.

Following the rise of more efficient and cheaper sequencing technologies is the large amount of data output. This has resulted in the need for efficient and precise bioinformatic tools as well as computing resources which can handle and present the data. This becomes quickly apparent when handling

large numbers of samples e.g. when comparing multiple tissue types, changes in tissues over time or when performing high-resolution analysis such as single-cell transcriptomics where the gene expression of each cell within a sample is studied separately.

Orthologs

Just as species have common ancestors, so do genes. These genes are referred to as homologous genes and are grouped into orthologs and paralogs depending on the driving evolutionary mechanisms. Paralogs are a result of gene duplication resulting from global whole genome duplication (WGD) events, or local (sub-genome) duplication events such as tandem duplication (Panchy et al., 2016). A duplication event results in a copy of a gene (or genome) which, similar to gene mutations, results in either the retention or loss of the new gene over time. In most cases, the gene copy needs to diverge functionally, at a molecular or a biological level, to be retained (Panchy et al., 2016; Zvelebil & Baum, 2008). Orthologs, on the other hand, are a result of speciation. Genes from two species are considered orthologs if they are descendants of the same gene from the last common ancestor of the two species (Koonin, 2005). These genes typically are similar in biological function as well as molecular function (sequence similarity), while paralogs are typically only sequence similar.

Protein function is defined by the proteins three-dimensional structure, which in turn depends on the order of amino acids. The chain of amino acids, polypeptide chain, is folded into a three-dimensional structure which is dictated by the various properties of the amino acids such as polarity or hydrophobicity. Amino acids are associated with a set of codons, triplets of RNA bases, meaning that each amino acid is typically specified by more than one codon. Codons associated with an amino acid differ only at the last nucleotide, e.g. both UUU and UUC are specific for phenylalanine. This flexibility allows some degree of variation in gene sequence without losing the intended amino acid order, thereby protein function. Furthermore, some amino acid substitutions are permitted without losing protein function as long as the substituted amino acid share similar properties as the intended amino acid, e.g. hydrophobicity. In this sense, protein sequences are more conserved than gene sequences are.

There are several tools for assessing homology between genes, such as OrthoFinder, OrthoMCL and InParanoid. These tools differ in how similarity scores, obtained from heuristic analysis such as BLAST or DIAMOND, are used to infer relationships between genes and whether they identify orthologs and paralogs, orthogroups, or both. When comparing multiple species, an orthogroup will contain genes that share a common ancestral gene and will therefore include both orthologs and paralogs. One method for identifying orthogroups is OrthoFinder. Using FASTA files with protein sequences for each species to be compared, OrthoFinder infers orthologous relationships using gene trees allowing the user to trace ortholog relationships, but also results in high ortholog inference accuracy (Emms & Kelly, 2019).

ComPlEx: Comparative analysis for Plant co-Expression networks

There are multiple tools available for constructing co-expression networks with algorithms based on different similarity metrics and statistical thresholds (Emamjomeh et al., 2017; Lee et al., 2020; Rao & Dixon, 2019). The common ground for co-expression networks is to first identify similar gene expression profiles (changes in gene expression across samples) by some measure before applying a threshold defining the level of similarity required between pairs of genes (Rao & Dixon, 2019). Similarity measurements are typically a correlation metric such as Pearson's or mutual information (Song et al., 2012).

Correlation metrics capture the dependency between variables, in this case change in gene expression, with Pearson measuring linear relationships and MI measuring how much knowledge is gained through the dependency of two variables and is a better method for detecting non-linear relationships. Although both metrics have their strengths and weaknesses, using Pearson's correlation allows for direct interpretation of the results (Ovens et al., 2021). In addition, considering that co-expressed genes are likely to display linear dependency, the need to include non-linear relationships is less relevant. To increase the robustness of correlation values, mutual rank (MR) is often applied. The ranking of genes sorts the relative order of correlation values for each gene with the highest correlated gene given the highest rank. Due to asymmetrical ranks, i.e. the rank of gene A to gene B, and rank of gene B to gene A are not the same, MR is applied to create a geometric average (Shekhovtsov, 2021). MR has shown to give better predictions of gene function but also enables better comparison between species (Obayashi & Kinoshita, 2009).

The power of co-expression networks becomes apparent when comparing gene expression across multiple species with multiple samples. A co-expression network identifies genes that are likely to be involved in the same biological processes. While sequence-similar genes are predicted to be similar in molecular function, they are not necessarily involved in the same biological processes. Furthermore, high sequence similarity not a guaranty of high functional similarity (Joshi & Xu, 2007). Therefore, through combining co-expression networks with homology, biological process and molecular function are combined. This also allows for the comparison without the need to align samples.

For this thesis, ComPlEx was used for identifying orthologs with conserved expression between pairs of species. ComPlEx, or Comparative analysis for Plant co-Expression networks, was built and presented by Netotea et al. (2014) with the aim of capturing conservation and divergence in gene regulation across species. The steps in ComPlEx can be summarised as following:

1) Co-expression networks: A co-expression network measuring similarity in expression between all gene pairs is inferred using Pearson's correlation on the expression data before applying MR. The size of the co-expression network for each species is defined by a density threshold.

2) Network comparison: For each ortholog pair, orthologous neighbourhoods (all co-expressed genes) are identified and compared (Figure 1 A). The statistical significance of the overlap of these neighbourhoods (by mapping one neighbourhood onto the other species using orthologs) is calculated both ways using a hypergeometric test with parameters as shown in Figure 1 B.

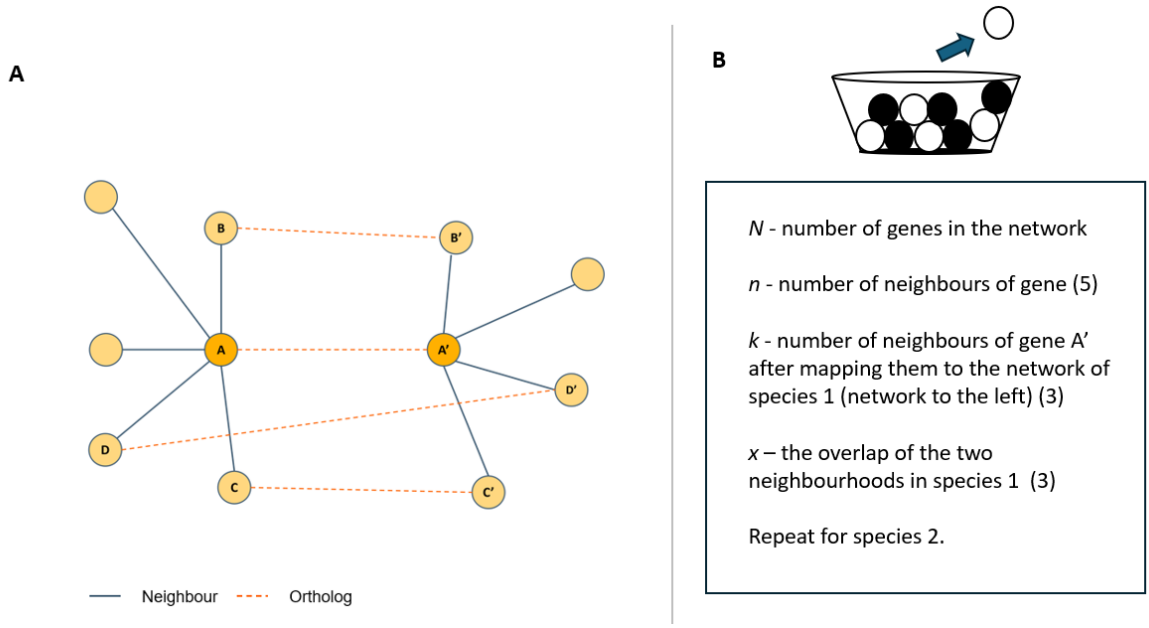


Figure 1: ***(A)*** *Visualisation of co-expression network comparison between two species. Co-expression networks for the ortholog pair A and A' (orange nodes) with their respective neighbours (yellow) indicated by blue full line. Orthologous neighbours are show with dashed yellow lines.* ***(B)*** *Parameters used for hypergeometric test.*

If there are no orthologous neighbours between the networks the hypergeometric test is not performed, and the p-value for this ortholog pair is automatically set to 1. All ortholog pairs between two species are compiled into a comparison-file, listing the gene names, number of neighbours and orthologous neighbours, and the p-values from the two-directional network comparison. As multiple hypothesis testing is performed, it is necessary to correct for false positive, i.e. type I errors. There are several corrective methods, but FDR (false discovery rate) correction methods such as the Benjamini-Hochberg or Storey-Tibshirani methods are commonly used when working with high throughput data due the sensitivity for true positives (Higdon, 2013). In effect, FDR-methods allow some false positives in exchange for retaining as many significant discoveries as possible.

1.1.3 Identifying naturally occurring groups within data sets

Co-expression networks are only one of several types of biological networks. Others biological networks include interaction networks (protein-protein interactions) signalling networks, and regulatory networks, and as with different omics data, different networks describe different aspects of cellular processes. A feature of biological (and cellular) networks that facilitate the understanding of complex processes is that they are largely scale-free. In a scale-free network the number of edges per node follows a power-law distribution meaning that most nodes will have a low number of connections, while only a few genes will display high connectivity. This contrasts to random networks where the number of edges per node follows a Poisson distribution meaning that most nodes are associated with a similar number of edges (Barabási & Oltvai, 2004). Considering biological networks as scale-free indicates the presence of natural groups within the data set. These groups are of interest to study, and the following sections will outline a couple of methods for identifying these groups.

Clustering

There are various ways of identifying natural groups within a data set, and a much-used approach within genetics is clustering. In short, clustering is a type of unsupervised machine learning that searches for naturally forming groups, or clusters, within a data set. Clustering methods such as k-means clustering assign data points full membership to a cluster and are considered hard clustering methods. Other algorithms, however, assign data points to clusters by degree of belongingness thereby allowing multiple memberships, and are also referred to as soft or fuzzy clustering methods. The basis for all clustering is a similarity measure, typically a distance metric, such as Euclidean distance. Euclidean distance is the squared distance between two points in an *m*-dimensional space, with *m* being the number of variables, or samples. If considering two variables, the distance is the length of the straightest route between two points, i.e. genes, in a two-dimensional co-ordinate system representing expression values for each sample along the respective axes.

A commonly used, and perhaps more suitable, method within genetics is hierarchical clustering. Here clusters are identified and arranged hierarchically either in an agglomerative (bottom-up) or a divisive (top-down) manner. Agglomerative clustering assumes all observations as single clusters and combines the cluster stepwise into one large cluster. The observations (assumed as a "cluster" of one) are grouped into initial cluster by merging the two closest (lowest distance) observations. These initial cluster are grouped together to form larger cluster, this time using another similarity metric as each cluster now consists of two observations. Similarity can now be the longest or shortest distance, or linkage, between members of different groups (complete or single linkage), or the overall average linkage between all members of the groups (average linkage). An alternative method is to fuse clusters which combined have the lowest within-cluster sum of squares error (SSE) (Ward's method). This process of combining cluster is continued until all cluster are combined as one large cluster which can visualised with dendrograms showing all clustering levels, or hierarchies. Although computationally demanding, there are many benefits of using hierarchical clustering such as reproducibility and the ability to observe the various hierarchies which can be cut to simplify the number of clusters.

Clusters are often used to identify co-expressed genes based on similarity in gene expression across samples and then visualised using heatmaps. This allows patterns of gene expression within the data to be identified. The expression of each gene is usually scaled and centred to accommodate for different levels of expression between genes. This means that the change, i.e. degree of up- and downregulation, in gene expression is relative to each gene and allows clusters to be clearly defined and compared.

Cliques

Co-expressed genes are assumed to be involved in the same process based on expression patterns, and as mentioned in previous sections, can be described as a network. In a larger network, these co-expressed genes appear as dense and interconnected nodes forming a sub-network. A sub-network in which all nodes are connected with each other is, within graph theory, called a clique. Identifying cliques within a network (Figure 2A) is also a method for detecting natural groups within data, with diverse applications (Collas et al., 2019; Pradhan et al., 2012). Within genetic studies, the use of cliques can also be a method for identifying, co-expressed genes (Zheng et al., 2011), but can further

be applied to study conservation of conserved co-expression between species (Oldham et al., 2006; Ovens et al., 2021), or inferring biological function (Adamcsek et al., 2006).

Two types of cliques can be identified within a network: maximal cliques and largest cliques (

Figure 2B). A maximal clique is a clique that cannot be extended to a larger clique by additional edges, and a largest clique (primarily referred to as a maximum clique) is the largest maximal clique in the network.

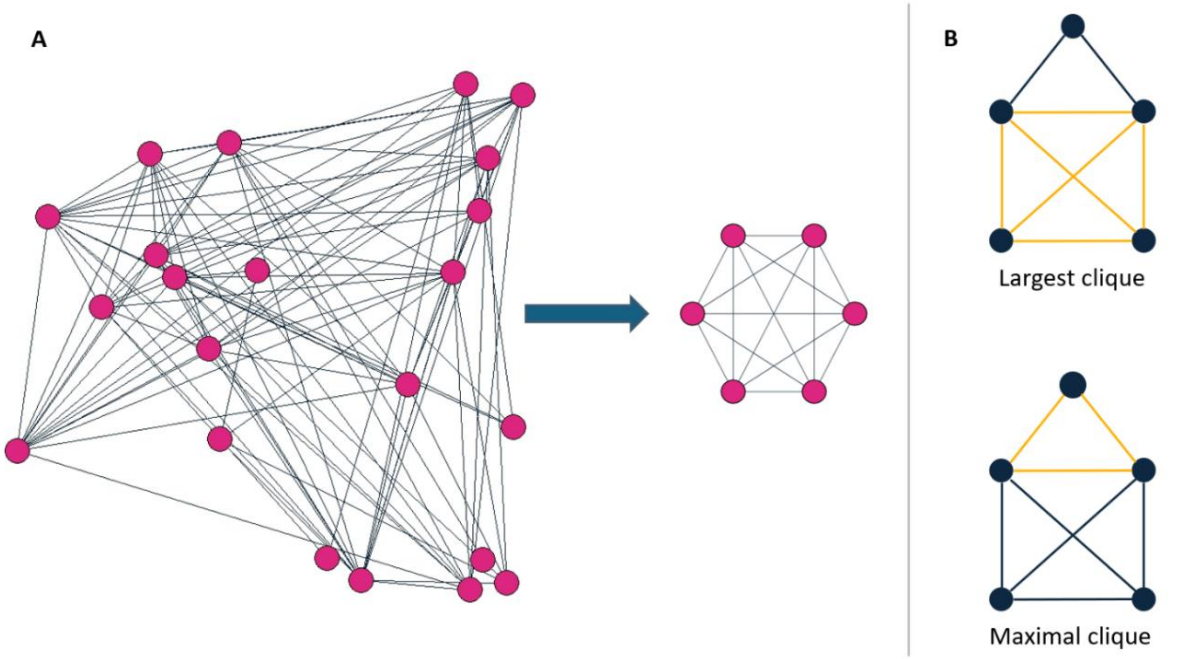


Figure 2: ***(A)*** *From a network of nodes (pink) displaying all interactions with other nodes using edges, a clique is identified. A clique of size six is shown here, but the size of a clique will depend.* ***(B)*** *Difference between a largest and maximal clique (yellow). The maximal clique cannot be extended by adding any more edges while the largest clique is the largest clique of the network.*

1.1.4 Gene ontology

Once a gene set of interest has been identified, the next natural step is to determine which processes these genes are associated with using gene ontology (GO). An ontology describes concepts, or classes, involved within one of the three different domains: biological, cellular, or molecular processes. These ontologies are associated with a GO ID for identification, and a GO term describing the role within the ontology. Ontologies are usually arranged in a directed acyclic graph (DAG) connecting parent and child terms, with the child terms usually more specific in description than the parent terms. To identify which GO terms the gene set is associated with, a gene set enrichment analysis (GSEA) is performed. An enrichment score is calculated based on how overrepresented genes annotated with various ontologies are (Consortium et al., 2023; Wolfe et al., 2005). Functional enrichment analysis (FEA) tests how enriched a gene set is in a functional category, effectively checking the overlap between the genes in the gene set and genes with functional annotation. This is performed using a hypergeometric test, also referred to as a Fisher exact test, with a p-value cutoff usually for retaining significant overlap (usually at 0,05).

2 Results

2.1 High resolution transcriptomics of six tree species

Prior to the start-up of the master's thesis, mRNA from three angiosperms (aspen, cherry, birch) and three gymnosperms (Scots pine, Norway spruce and lodgepole pine) was isolated from samples from the phloem, cambium, expanding xylem, and SCW-forming xylem, and quantified using RNA-Seq. The samples were obtained from pooling 15 µm thick sections (cryosections), collected across a segment of tree trunk, based on similarity in tissue composition. The pooled sections were divided into approximately 28 samples (spanning differentiated phloem to mature and late xylem) prior to mRNA isolation. Reads were mapped to the respective species genomes, with the exception of lodgepole pine which was mapped to *Pinus sylvestris* (Scots pine) transcripts. Samples with low read count were removed. This resulted in certain species having a non-continuous sample range. To increase the likelihood of a continuous sample range, samples from three specimens per species were collected. For visualisation, however, the specimen with most continuous sample range was used. The work in this thesis began once the read count tables were completed.

Most of the six species had a complete or a near-complete sample range. Three had a continuous sample range: aspen with 25 samples, Scots pine with 28 samples, and cherry with 27 samples. Birch and lodgepole pine both had a sample range of 1:28, with birch only lacking samples 10 (expanding xylem) and 27 (late xylem) and lodgepole pine only lacking sample 24 (late xylem). However, Norway spruce was the species with the lowest number of samples. The range of the best Norway spruce specimen was 1:27, and with samples 10 (expanding xylem), 14-16 (SCW-forming xylem), 21 (mature xylem), and 23-26 (late xylem) removed. Early heatmaps for cherry showed a diverging sample, number 17, which appeared to be an outlier. This sample was substituted with the mean of samples 16 and 18 when plotting heatmaps and expression profiles.

For both the heatmaps and the expression profiles, a separate set of expression files were used with imputed samples for Norway spruce and birch allowing more continuous and comparable figures. Sample imputation for the birch specimen was done by taking the average of the two adjacent samples for samples 10 and 27. Longer segments of imputed samples were needed for Norway spruce, but a similar approach as with birch using the mean was applied. Sample 10 was straightforward, the mean of samples 9 and 11. For samples 14-16, a mean of samples 13 and 17 were used to create sample 15, making it possible to create samples 14 and 16 from the means of adjacent samples. A similar approach was used for samples 23-26.

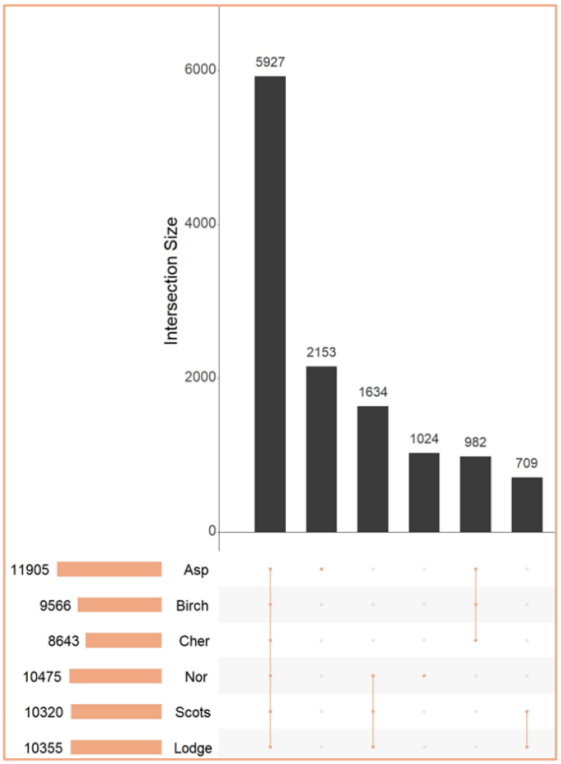
2.2 Identifying co-expressologs

Differences in wood anatomy and sampling prevented the samples to be compared directly between species. Instead, the RNA-Seq data was used to create co-expression networks for all species which were used as the basis for comparing gene co-expression across species. Species comparison was performed pairwise and was made possible using groups of orthologous genes (orthogroups) that were previously identified based on protein sequence similarity using OrthoFinder. Ortholog pairs within each orthogroup were classified as conserved orthologs, or co-expressologs, if their neighbours in the co-expression networks were conserved (FDR < 0.1).

Most orthogroups contained multiple genes per species, resulting in several co-expressologs per orthogroup. The classification of orthologs as co-expressologs was achieved using ComPlEx (Netotea et al., 2014), a tool for identifying conserved orthologs through comparing co-expression between species. Genes that are co-expressed share highly correlated expressions values across the various samples. With the assumption that high or low correlation between genes is relative to each species, ComPlEx ranks the correlated genes so that a density value is applied instead of a correlation cutoff value when co-expression networks are inferred. A density value of 0,03 was used to infer co-expression networks thereby defining the 3% most correlated genes as neighbours for each ortholog pair. Further analysis was done based only on genes with at least one expressed ortholog in at least one other species.

To gain insight into how the orthogroups were distributed across the species, UpSet plots were used to inspect the initial results (Figure 3 A). Vertical bars reflecting the number of orthogroups that contain expressed genes from all the species in the indicated species group, reveal that the species group with the largest number of orthogroups contain all six species. The second largest orthogroup is amongst the gymnosperm species, while the angiosperm species have the fifth highest number of orthogroups. Aspen (an angiosperm) and Norway spruce (a gymnosperm) are the two species with highest number of species-specific genes, while Scots pine and lodgepole pine are the pair with highest number of shared orthologs. The percentage of genes which are co-expressologs for the various species pairs ranged between 23-38% (Figure 3B). The highest number of co-expressologs relative to the number of expressed genes is generally observed between the angiosperm pairs, except for lodgepole pine and Scots pine claiming the highest relative co-expressolog ratio.

A



B

Species pair	Expressed genes	Co-expressologs
Aspen - Cherry	39,546	13,140 (33,2%)
Aspen - Birch	44,968	14,746 (32,8%)
Birch - Cherry	26,720	9810 (36,7%)
Aspen - Norway spruce	39,431	9986 (25,3%)
Aspen - Scots Pine	41,811	9841 (23,5%)
Lodgepole Pine - Aspen	42,209	9955 (23,6%)
Lodgepole Pine - Birch	29,342	7474 (25,5%)
Lodgepole Pine – Cherry	24,496	6364 (26%)
Norway Spruce - Birch	27,075	7304 (27%)
Norway Spruce – Cherry	23,026	6569 (28,5%)
Scots Pine - Cherry	24,213	6641 (27,4%)
Scots Pine – Birch	29,077	7740 (26,6%)
Norway Spruce - Scots Pine	46,054	13,795 (30%)
Lodgepole Pine - Scots Pine	57,308	21,655 (37,8%)
Lodgepole Pine - Norway Spruce	45,722	13,055 (28,6%)

Figure 3: Overview over orthologs and expressed genes. ***(A)*** *Upset plots were used to visualise the number of orthogroups with at least one expressed gene for each species (horizontal bars) as well as the number of orthogroups that contained expressed genes from all the species in the indicated species group (vertical bars).* ***(B)*** *The output from running co-expression network analysis (ComPlEx) on each species pair. Expressed genes are the number expressed ortholog pairs and co-expressologs are the number of ortholog pairs with conserved co-expression (FDR < 0.1). Co-expressologs are also shown as percentage of expressed genes.*

2.3 Sample comparisons revealed transcriptional steady states and reprogramming events

To ascertain where gene expression between species differed, sample comparison was performed for all pairs of species. Here co-expressolog were used to correlate samples between the species pairs and the correlations were then visualised as heatmaps Figure 4. Highly correlated samples are highlighted (yellow) and reflect samples with similar expression, while the darker areas (dark blue) show sample with low correlation in expression. Each species is also compared with itself using all genes in the expression data. The blocks of highly correlated samples indicate a steady state where there is little variation in from sample to sample. These steady states are likely to reflect samples within similar tissue. Between the steady state samples are transition phases which lead up to reprogramming events in which a different set of genes are expressed. The blocks of highly correlated samples reflect transitions between the different stages in secondary growth. Clearly defined blocks can be seen in the aspen-aspen comparison in which four distinct blocks are shown. The positioning

of the blocks seems to correspond to the different stages of secondary growth: phloem, expanding xylem and SCW xylem (potentially including cell death). Aspen seems to be the species with the clearest transitions between reprogramming events, which contrasts to other species such as birch and lodgepole pine where the transitions between stages are less clear. Less defined transitions might be related to sample and data quality, reducing the sample resolution. All pairs of species seem to show a diagonal pattern indicating that there is some degree of correlation in gene expression throughout the process of secondary growth. The gymnosperm pairs have higher max correlation values than the angiosperms pairs which in turn only have a slightly higher max correlation than the mixed pairs.

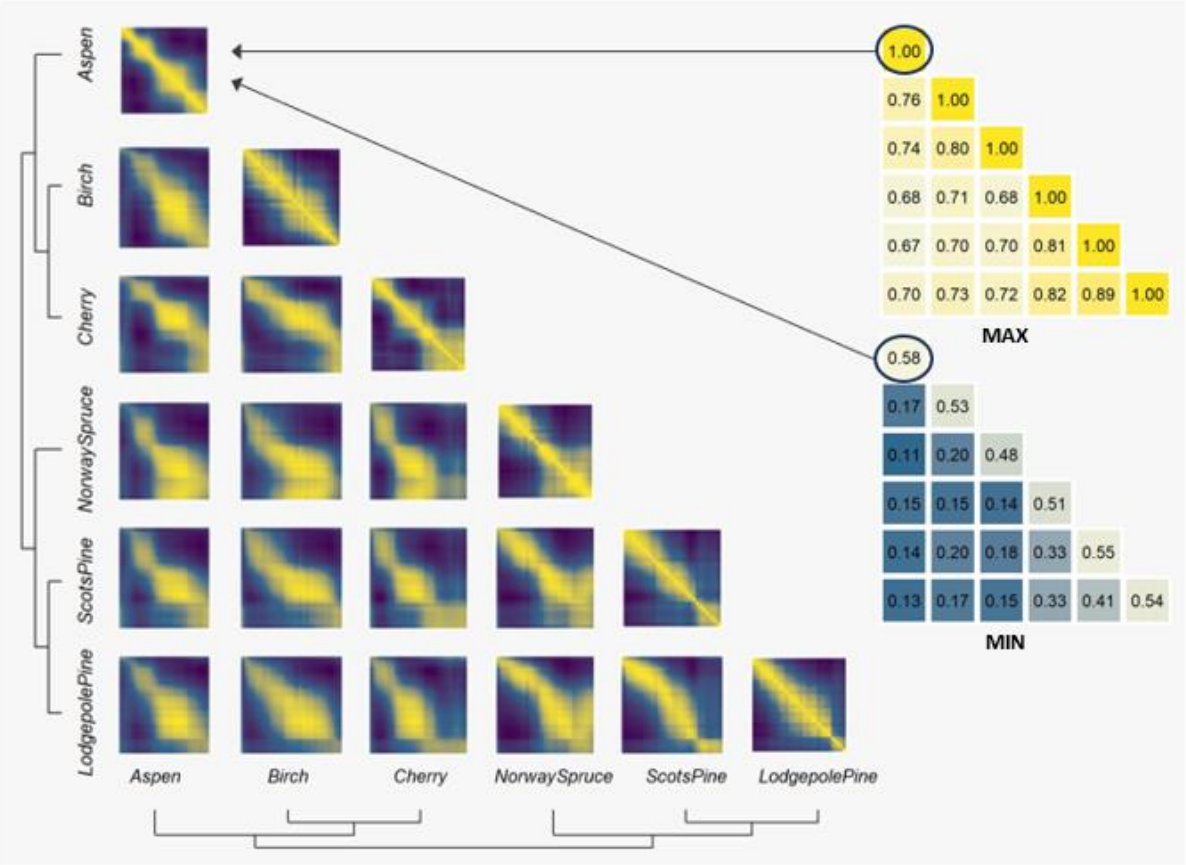


Figure 4: Heatmaps used to visualise sample correlation between all pairs using only co-expressologs. The heatmaps along the diagonal show sample correlation within the same species which serves as reference. To achieve better resolution, the heatmaps were scaled individually. The min and max correlation values for each heatmap are shown in the legends (right) with the position of each box reflecting which heatmap the min or max values belong to. As the arrows indicate, the correlation values for the aspen-aspen heatmap range from 0.58 to 1.00.

2.4 Orthologs with conserved expression

The heatmaps in Figure 4 reflect samples with correlated gene expression using co-expressologs but do not show if these co-expressologs are the same set of genes for all pairs of species. Next, the co-expressologs were analysed in the context of orthogroups in order to identify genes with conserved expression across all or many species. The species pairs consisted either of an angiosperm and a gymnosperm (cross-clade pairs), or of two angiosperms or two gymnosperms (clade-specific pairs). Together there were 15 different pairs of species. The expression profile of an orthogroup was considered fully conserved if all species pairs had at least one co-expressolog within the orthogroup. This condition was relaxed to investigate how many orthogroups were partially conserved. A partially conserved orthogroup required only co-expressologs in two out the three clade-pairs (both angiosperm and gymnosperm) and in two of the nine cross-clade pairs. Clade-specific orthogroups, i.e. orthogroups with only co-expressologs for angiosperm or gymnosperm pairs, were also identified in a similar manner but were to be specific for the clade. For instance, a gymnosperm-specific orthogroup would have no co-expressologs amongst the angiosperm pairs or the cross-clade pairs. A fully conserved orthogroup would require all three species pairs to contain co-expressologs, or two pairs to be partially conserved.

Although fully and partially conserved orthogroups only required the presence of one co-expressolog from the required number of species pairs, each orthogroup contained several co-expressologs. Plant genomes are characterised by a high number of homologous genes due to several gene duplication events (Panchy et al., 2016). This meant that some of the co-expressologs could be conserved paralogs. This raised the question of whether sets of consistent 1-1 co-expressologs were present across all species pairs within the orthogroups. By creating a network of all co-expressologs within an ortholog group, with nodes representing genes and edges representing conserved co-expression between the gene (orthologs), cliques were identified. Fully conserved orthogroups with cliques were defined as ultra conserved based on the notion that a clique would exclusively consist of orthologs and not paralogs (i.e. exactly one gene per species). Partial cliques were also identified for comparison and were “partial” as not all nodes were co-expressologs. The orthologs with complete and partial cliques represented a subgroup of the fully and partially conserved ortholog groups, respectively.

The highest number of both fully and partially conserved orthologs are observed between the angiosperms and gymnosperms, i.e. all pairs, with the second highest between gymnosperms, and fewest orthogroups conserved within angiosperms (Figure 5). This trend is reflective of the number of potential orthogroups with expressed genes for each group of species (Figure 3A). Amongst the 1096 orthogroups conserved across all 15 pairs, 65% (714 orthogroups) contain at least one clique. This ratio is higher for the clade-specific orthogroups where 86% and 84% of the ultra-conserved orthogroups contain cliques for gymnosperms and angiosperms, respectively.

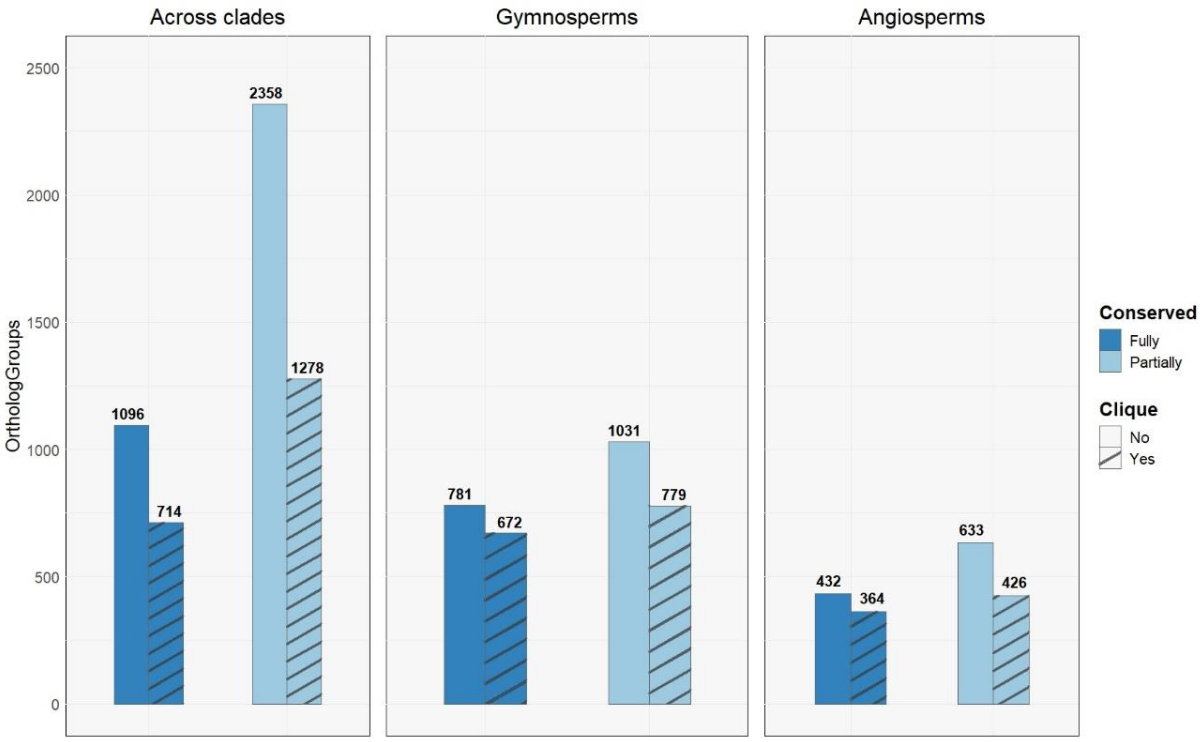


Figure 5: Overview of fully conserved and partially conserved orthogroups for all 15 pairs (across clades) and for the clade-specific pairs (gymnosperms and angiosperms). The orthologs within these groups that contain at least one clique are shown as striped bars.

Expression heatmaps visualising co-expressolog activity across the various tissue types (Supplementary figure 1, Appendix) reveal that some gene clusters show up-regulated expression within tissue-specific samples, while other clusters show an increase in expression in the transitioning samples between specific tissues. However, the varying number of unique co-expressolog members and the random order of genes (rows) per species, make cluster comparison between species challenging. Nonetheless, general observations can be made. Aspen and cherry show larger dynamic range of expression, i.e. a larger difference between high and low gene expression, compared to the other species. Another general observation is that the number of clusters within the SCW-forming xylem are fewer within the two pines than for the other species.

To ascertain the expression patterns of the ultra-conserved genes across the samples, heatmaps of the transcriptomic data were generated using only ultra-conserved genes (clique-associated genes) (Figure 6). As the heatmap rows (genes) for all species could arranged by orthogroup resulting in each row corresponding to genes within the same clique, the clusters were now comparable. Samples from the same tissue were separated by vertical lines to account for the differences in tissue size between the species.

With some variations due to the relative sample sizes of each tissue, the overall patterns are quite similar between the species heatmaps initially revealing two distinct, large clusters: one across the phloem and expanding xylem, and another spanning the expanding xylem and SCW-forming xylem tissues. These cluster can further be dived into smaller clusters. The cluster spanning the phloem and expanding xylem is likely to be associated with cambial activity in generating new phloem and xylem

cells. The clusters across the expanding and SCW-forming xylem are expressed slightly different between species with some species showing an increase in expression across both the expanding and SCW-forming xylem, as seen in birch and Scots pine, while primarily upregulated in the SCW-forming xylem in aspen and Norway spruce. These variations may, however, be a result of sample separation.

Similar heatmaps were made using ultra-conserved genes unique for the angiosperms and gymnosperms (Supplementary figures 2 and 3, Appendix). These heatmaps bear similar clustering patterns to the heatmaps in Figure 6 with clusters of increased expression amongst the lines separating the phloem and expanding xylem samples, and the lines separating the expanding xylem and SCW-forming xylem samples. One difference was the more pronounced cluster upregulated in the phloem for both angiosperms and gymnosperms which was not as evident in the heatmaps in Figure 6.

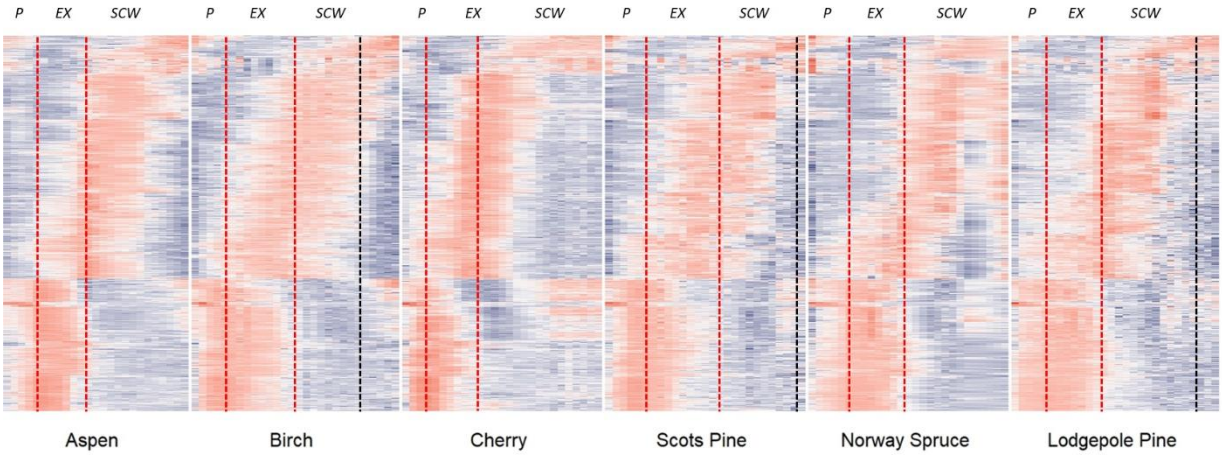


Figure 6: Heatmaps showing expression values of all ultra-conserved genes. The transcriptomics data were arranged by ortholog group resulting in each row for each heatmap corresponding to a conserved gene (co-expressolog from clique). The two red lines separate samples from the phloem (P), expanding xylem (EX), and SCW-forming xylem (SCW). The cambium is represented by the separating line between the phloem and expanding xylem. Samples from birch, Scots pine and lodgepole pile also included late-wood samples which are separated from the lignified xylem by the black line.

2.5 Gene ontology and marker gene candidates

Once the ultra-conserved genes were identified, both amongst the clade-specific species pairs and across all species pairs, the next step was to uncover which biological processes these genes were associated with. A gene ontology (GO) enrichment analysis was performed, identifying GO terms in which the genes were enriched in. The various GO terms were the grouped into higher-order, or parent, terms for better overview. The ultra-conserved genes across the clades were enriched for several GO terms, some describing potentially tissue-specific processes and others more associated with more general processes (Figure 7).

In addition to discovering which processes the ultra-conserved were associated with, one of the study aims was to suggest genes which could pose as marker genes for the various tissue types. As parent terms sometimes mask more descriptive child terms, the child terms (not visualised here) were

inspected in closer detail. The expression profiles of the ultra-conserved genes associated with potentially interesting terms were plotted for all six species and colour coded based on which clique the genes belonged to. To prevent misinterpretation, the cliques were also referenced using the aspen gene member of the clique (Potra-xxxx). In addition to GO terms, marker genes (homologs of Arabidopsis) identified in Sundell et al. (2017) were searched for amongst the ultra-conserved genes. Three of the marker genes were identified: SUS6 (Potra2n6c14105) involved in phloem differentiation, CDC2 (Potra2n6c14327) in cambial activity and BFN1 (Potra2n689s36475) associated with cell death. Functional annotation for the remaining genes was based on Arabidopsis homologs of the remaining genes that showed potential as marker gene was identified using the database PlantGenie (http://plantgenie.org).

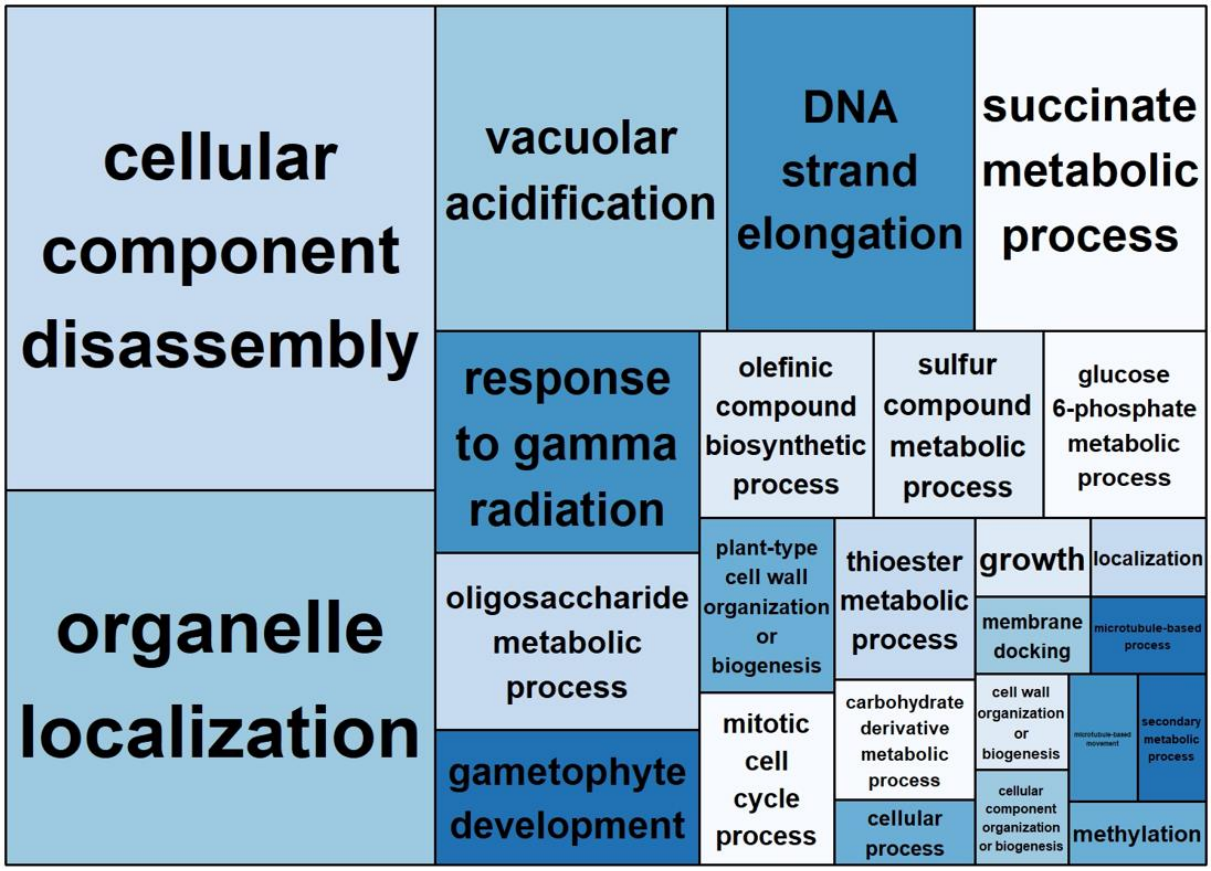


Figure 7: Higher-order gene ontology terms in which the ultra-conserved genes across all species pairs were enriched.

Many of the larger parent terms in Figure 7 such as “organelle localization” and “cellular component disassembly” seem to be quite generic terms, while other parent terms such as “gametophyte development” and “DNA strand elongation” are more specific, but not assumed to contain candidates for describing wood forming processes. “Vacuolar acidification”, relating to an intracellular pH reduction, could play an indirect role in cell growth by maintaining turgor pressure (Kaiser & Scheuring, 2020). Many of the smaller parent terms, however, are immediately associated to secondary growth: “growth”, “microtubule-based process/movement” and “plant-type cell wall organisation or biogenesis”.

One of the child terms of “growth” is “Unidimensional growth (GO:0009809)”, and the genes annotated with this term display two sets of expression profiles (Figure 8). Amongst these genes are

homologs (Potra2n6c14327) of the marker gene CDC2 (CELL DIVISION CONTROL 2). These genes stand out from the other four genes by peaking around the cambium. The other genes show less of a distinct profile, spanning across the expanding and SCW-forming xylem. This hints to the CDC2 homologs being active during cell division of the cambium, while the other sets of genes are active in cellular expansion which both phloem and xylem cells undergo, but with increased activity amongst xylem cells.

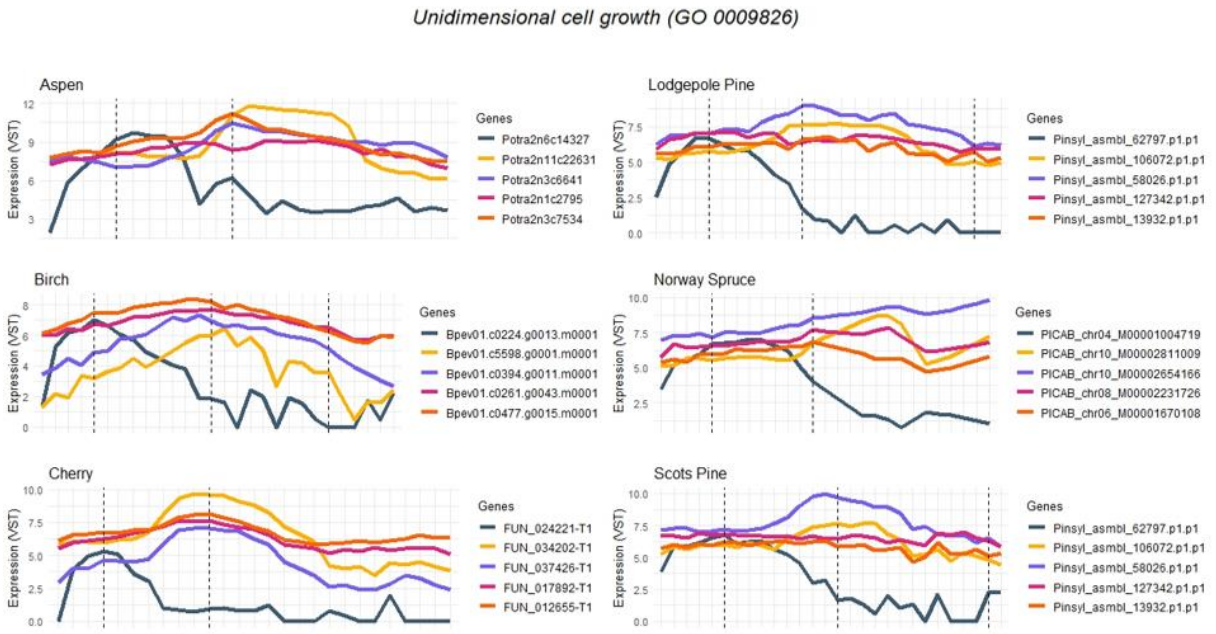


Figure 8: Expression profiles of ultra-conserved genes across all species associated unidimensional cell growth (GO 0009826). The homologs of CDC2 seen in dark blue (Potra2n6c14327).

Another GO term of interest was “Plant-type secondary cell wall biogenesis (GO: 0009834)”. This ontology is associated with five sets of genes (Figure 9) and plotting reveals two distinct expression profiles. The first profile type, displayed by only one gene (Potra2n6c14289), shows an increase in expression within the cambium and across the newly formed phloem and xylem cells, and for some species only within the expanding xylem. The remaining four genes, including the SUS6 (SUCROSE SYNTHASE 6) homolog, Potra2n7c16288, show a peak between the expanding and SCW-forming xylem. The SUS6 deviates from the profile of the SUS6 homolog described in Sundell et al. (2017) which displayed increased expression within the phloem samples.

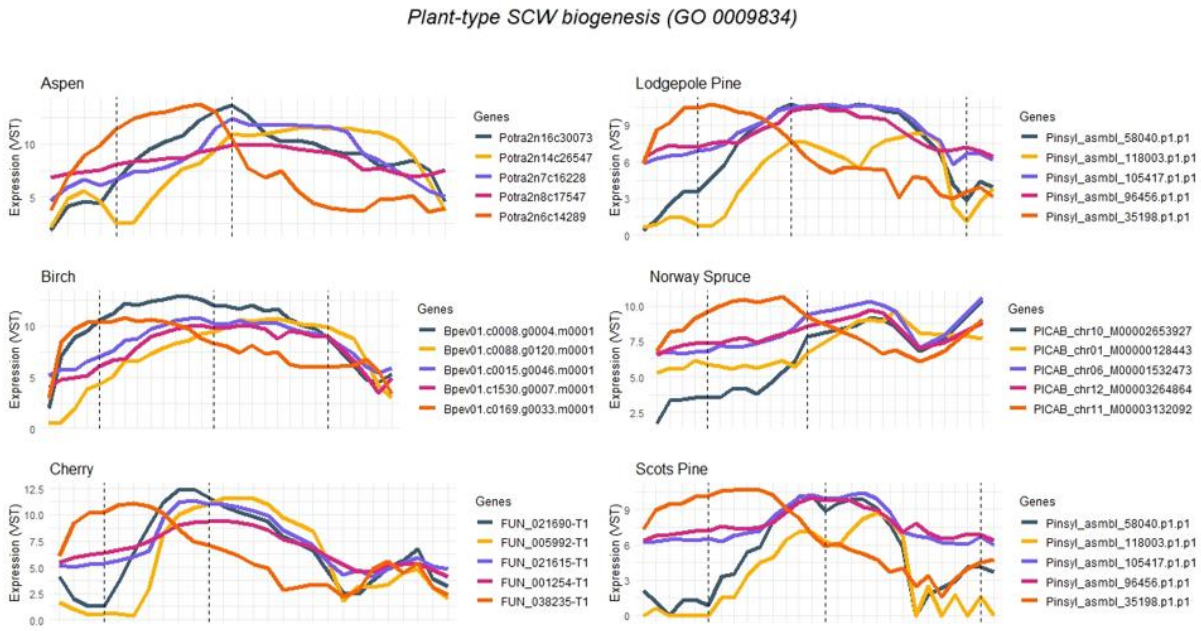


Figure 9: Expression profiles of ultra-conserved genes annotated with GO term "Plant-type SCW biogenesis GO 0009834".

During secondary growth, cells grow in one or more dimension (depending on cell type) and is a form of morphological change observed in cells. Therefore, the GO term "Cell morphogenesis (GO 0000902)" was further inspected. Three genes share this annotation, with one gene set (Potra2n4c9806) displaying a distinct curve within the expanding xylem (Figure 10). To enhance the expression profile, only this gene was plotted.

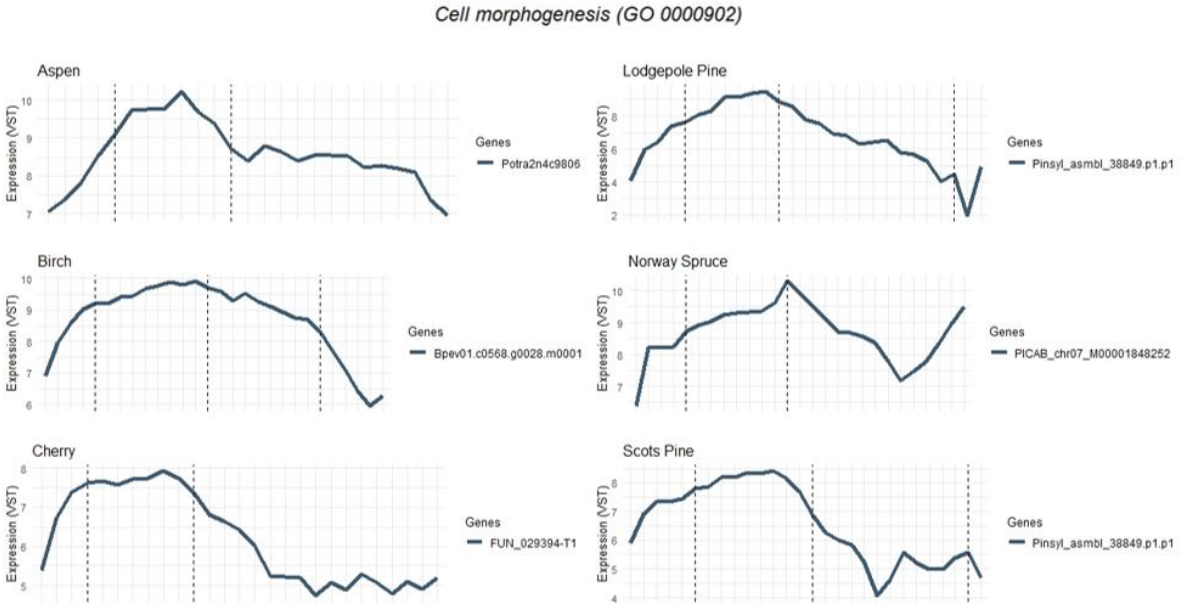


Figure 10: Expression profile of one of the ultra-conserved genes annotated with GO term "Cell morphogenesis GO 0000902".

During the assembly of the SCW, microtubules play an important role in forming the Cellulose Synthase Complexes (CSC) which creates the cellulose microfibrils, amongst other reorganizational functions (Zarra et al., 2020). Figure 11shows three expression profiles for the genes associated with "Microtubule cytoskeleton organisation (GO: 0032012)", two profiles (Potra2n2c5399 and Potra2n3c6923) with expression peaks within the expanding xylem, alternatively between the expanding and SCW-forming xylem, and the third profile (Potra2n10c21409) with relative constant expression across all samples. Amongst the two genes with more distinct profiles, Potra2n3c6923 shows a steeper increase for some species within the expanding/SCW-forming xylem.

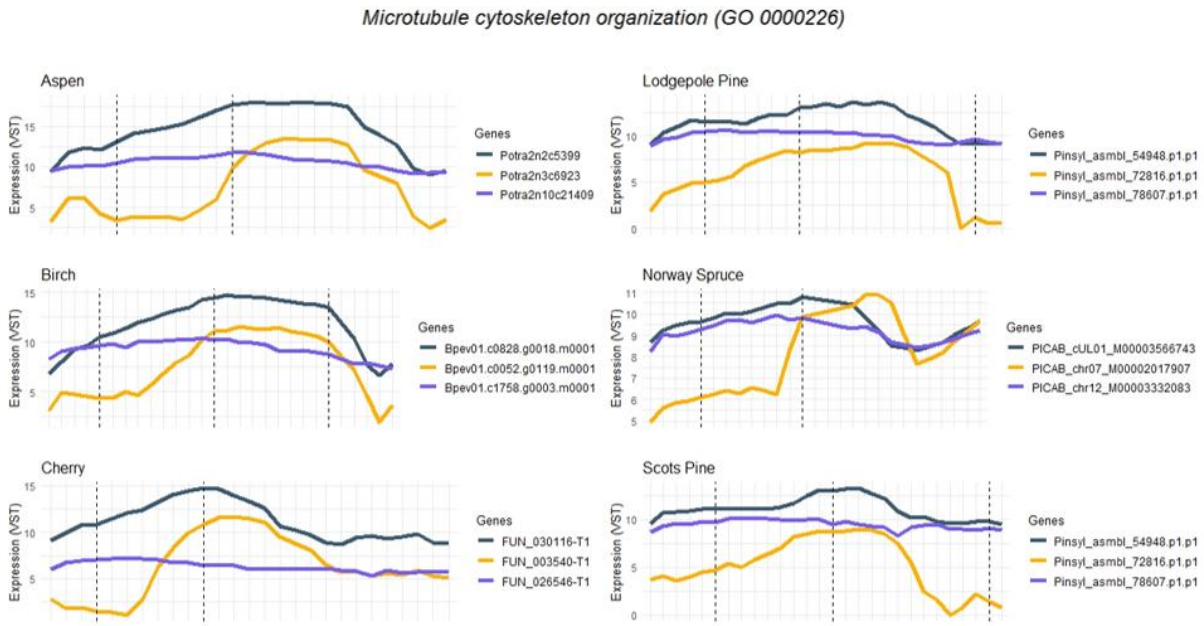


Figure 11: Expression profiles of ultra-conserved genes annotated with GO term "Microtubule cytoskeleton organisation GO 0000226".

Being key constituents of secondary cell walls, "Lignin biosynthetic process (GO: 0009809)" and "Xylan biosynthetic process (GO: 004592)" were undoubtably GO terms of interest. All five genes associated with formation of lignin, show a very correlated biphasic profile with a low peak within the phloem samples and a significantly larger peak within the SCW-forming xylem cells (Figure 12). The profiles for xylan biosynthesis also involve a set of correlated gene expressions (Supplementary figure 6, Appendix).

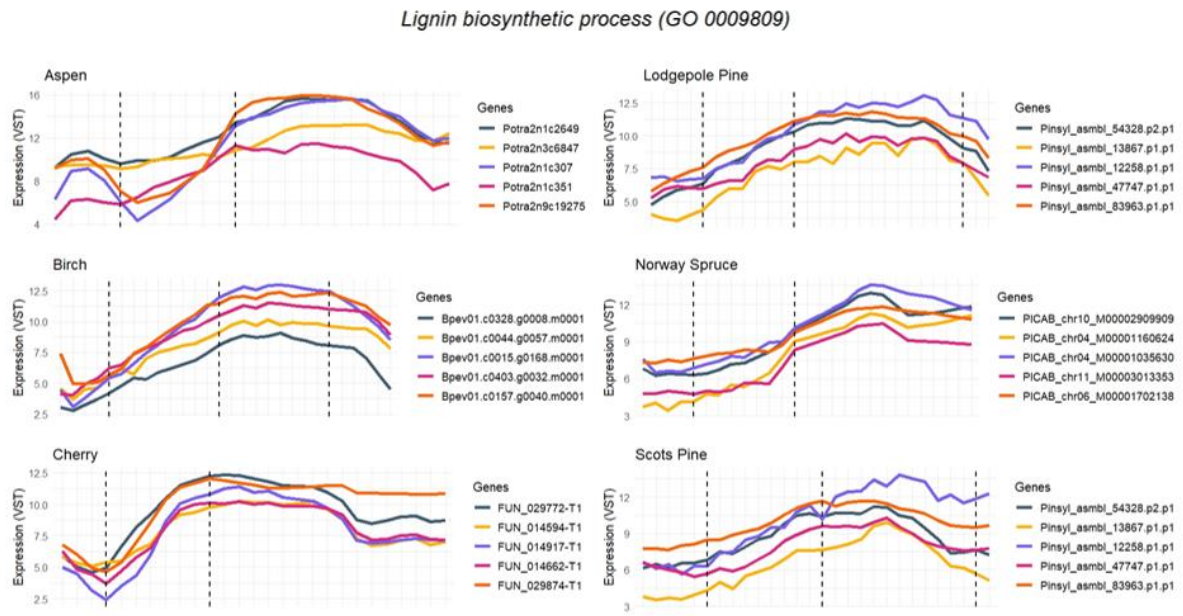


Figure 12: Expression profiles of ultra-conserved genes annotated with GO term "Lignin biosynthetic process GO 0009809".

The end of secondary growth is marked by the gradual death of xylem cells within the mature xylem. Although not identified through gene ontology, homologs of the BFN1 (BIFUNCTIONAL NUCLEASE 1) gene were also identified amongst the ultra-conserved genes. The expression profiles of these homologs seem to be descriptive of cell death within the xylem by displaying a markedly increase in expression within late SCW-forming xylem (Figure 13).

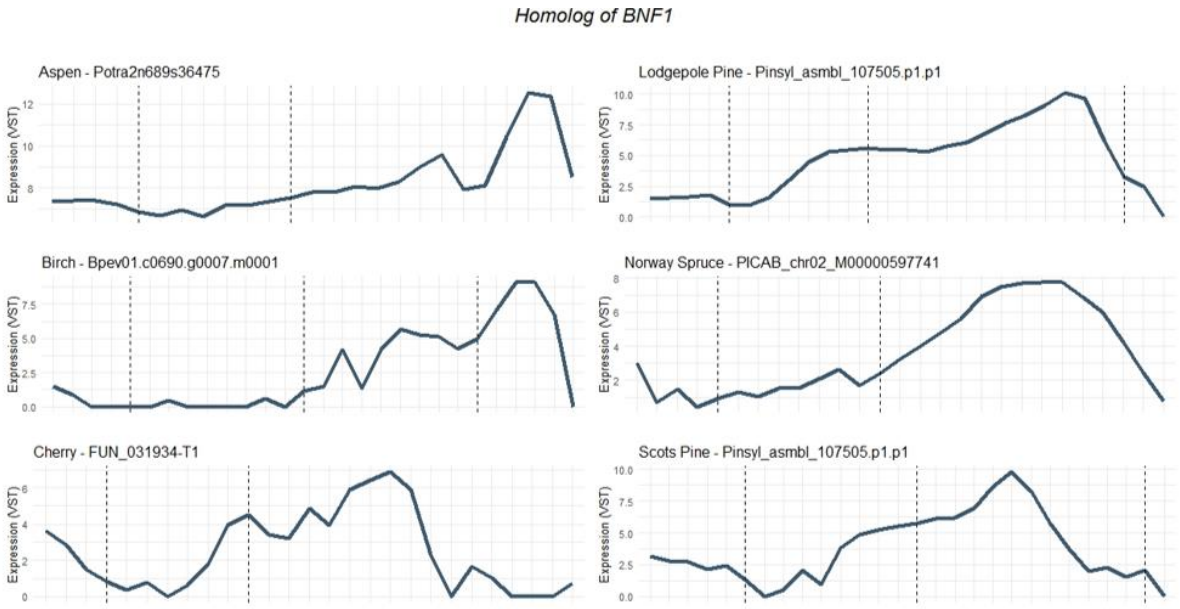


Figure 13: Expression profiles of BFN1 (Bifunctional nuclease 1) homologs.

Most of the gene clusters in Figure 6 that showed increased gene expression within the phloem were expressed across the cambial area, and into the expanding xylem as well. These were therefore more likely to be related to cambial activity. However, a small cluster of only a few genes seemed to have an increased expression specifically within the phloem. Considering no obvious GO terms associated with phloem activity were identified, phloem-specific genes were searched for manually. A small subset of genes with an expression profile indicating phloem activity were identified. One of these genes are shown in Figure 14, and has a characteristic peak uniquely within the differentiated phloem samples.

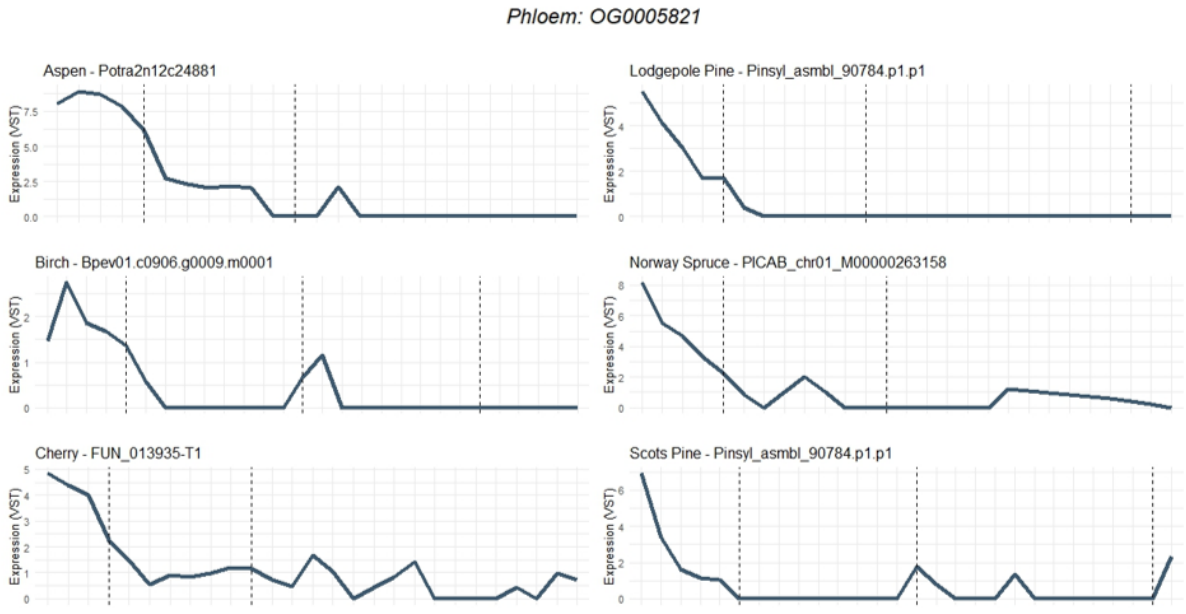


Figure 14: Expression profiles for genes associated with phloem activity.

Gene ontologies for the ultra-conserved clade-specific pairs were also identified. One of the larger parent terms for angiosperms is “lactone biosynthetic process” which is likely to be connected to another larger parent term, “sesquiterpenoid biosynthetic process” (Supplementary figure 4, Appendix). Lactones are cyclic organic while sesquiterpenoid are a class of sesquiterpene lactones ester and are associated with a variety of functions such as antimicrobial properties and fragrance and flavours (Perassolo et al., 2018). Delving into the child terms of the angiosperm-specific genes revealed terms such as “Glucuronoxylan biosynthetic process (GO: 0010417)” and “Plant-type secondary cell wall biogenesis (GO: 0009834)”.

For the gymnosperms-specific genes, the GO terms were more general with less obvious function and connection to wood formation (Supplementary figure 5, Appendix). However, “Erythrose 4-phosphate/phosphoenolpyruvate family amino acid biosynthetic process (GO: 1902223)” is indicative of the presence of intermediates of aromatic amino acid synthesis. This could be related to the synthesis of the aromatic amino acid phenylalanine, a precursor of lignin (Marchiosi et al., 2020).

3 Discussion

This thesis aimed to identify genes associated with wood formation that were conserved between angiosperms and gymnosperms. These genes were further investigated to identify genes which could pose as marker genes candidates for specific tissues or of transitional phases. Additionally, genes potentially unique to the angiosperms and gymnosperms, i.e. clade-specific genes, were also identified as it was anticipated that these genes might elucidate some of the differences in wood structure between the clades.

To achieve these aims, high-resolution transcriptomic data from samples spanning from differentiated phloem to mature xylem extracted from three gymnosperms and three angiosperms was used as basis for performing comparative co-expression network analysis. Orthogroups containing co-expressologs, i.e. orthologs with conserved neighbours (FDR < 0.1) in co-expression network, were identified between all pairs of species. An orthogroup was considered *fully* conserved if co-expressologs were identified for all 15 pairs. These fully conserved orthogroups were investigated further to identify cliques, a complete set of consistently 1-1 co-expressologs shared between the species. Orthogroups with cliques were defined as ultra-conserved orthogroups, and the genes belonging to cliques were defined as ultra-conserved genes. Using gene ontology and expression profiles, genes associated with secondary growth were identified. Ultra-conserved clade-specific genes were also identified using a similar workflow.

The GO terms that were further investigated were based on existing knowledge of the cellular processes within wood formation. In addition, three of the genes that had previously been identified as maker genes of the different tissues and transition phases by Sundell et al. (2017) were also identified, and were used to navigate the list of GO terms. These marker genes were homologs of the Arabidopsis genes CDC2 (CELL DIVISION CONTROL 2), SUS6 (SUCROSE SYNTHASE 6), and BFN1 (BIFUNCTIONAL NUCLEASE 1). The expression profiles of the genes annotated with the GO terms of interest were compare between the species to identify marker gene candidates. Marker gene candidates were required to be universal for all species, meaning that the expression profile should be recognisable in overall pattern, but also in which areas the gene expression peaked and dipped. Due to difficulties in separating samples affecting both the preciseness of the separating the samples and how well each tissue is defined, interpretation and comparison of the expression curves must be done with some margin of error.

Partially conserved genes were also identified but were not of focus for this thesis. Instead, these genes were identified to explore the composition of the orthogroups, but also to represent a wider set of genes for further analysis.

3.1 Ultra-conserved genes showed enrichment in various processes associated with wood formation

Genes involved in cell growth, the biosynthesis of lignin, secondary cell wall (SCW) formation, organisation of microtubule cytoskeleton, and cell death, were identified amongst the 714 ultra-conserved genes across angiosperms and gymnosperms. The 364 ultra-conserved angiosperm-specific genes were enriched in similar processes as the genes conserved across all species but included also genes enriched in lactone biosynthesis. One gymnosperm-specific term found among the 672 ultra-conserved genes included the GO term "Erythrose 4-phosphate/phosphoenolpyruvate family amino acid biosynthetic process (GO: 1902223)", with the remaining GO terms being more generic. This section will first discuss the genes ultra-conserved across the angiosperm and gymnosperm species, highlighting potential marker genes for each tissue, before discussing the ultra-conserved clade-specific genes.

3.1.1 Ultra-conserved genes between angiosperms and gymnosperms

Phloem

Sundell et al. (2017) identified the SUS6 gene as a marker gene for the phloem. Homologs of the SUS6 genes were identified amongst the ultra-conserved genes (Potra2n7c16228) but displayed an increased expression within the expanding xylem indicating that these genes were likely to be paralogs of SUS6 instead. This led to a manual search which identified genes with distinct profiles uniquely within the phloem. The expression profiles of this set of genes were universal and showed potential as a marker gene candidate. Furthermore, these genes are potential orthologs of the Arabidopsis thaliana gene, AT1G73040 that encodes for mannose-binding lectins. These are also referred to as Phloem Protein2 (PP2), a family of proteins associated with nutrient transport, long-distance signalling and show involvement in defensive reactions within the phloem (Dinant et al., 2003; Kehr, 2006). Phloem lectins are one of the main constituents of the phloem sap, and are known to play diverse and vital roles within both angiosperms and gymnosperms (Dinant et al., 2003).

There are some compositional differences of angiosperms and gymnosperms phloem. For instance, angiosperm phloem comprises of companion cells which are absent in gymnosperms, and also have sieve tube elements (Liesche & Schulz, 2018). Gymnosperms, on the other hand, have sieve cells. This could be one explanation to why so few phloem-specific genes were identified. However, heatmaps constructed using all co-expressologs (Supplementary figure 1, Appendix) didn't reveal large phloem-specific activity, meaning that a small subset of phloem-specific genes could be expected amongst the ultra-conserved genes.

Cambium

Cambial activity is associated with cell division, producing new cells to the phloem and expanding xylem. With no GO terms directly associated with cell division amongst the enriched ontologies, the CDC2 homolog was used as a starting point. The CDC2 homologs were amongst several genes

enriched in “Unidimensional cell growth (GO: 0009826)”. Most of these genes displayed a constant expression across the samples, while the CDC2 homologs (Potra2n6c14327) displayed increased activity across the cambium thereby corresponding with the CDC2-profile identified in Sundell et al. (2017). CDCs are known to play an important role in regulating cell division and growth within cambial cells (Cheng et al., 2024), and with the expression profiles similar across all six species, the CDC2 homologs could pose as candidate marker genes for the cambium.

Another set of genes which could be descriptive of cambial activity were amongst the genes annotated with “SCW-biogenesis (GO 0009834)”. Of the two expression profiles, the profile of Potra2n6c14289 and its orthologs displayed increased activity within the cambium for some of the species. For the other species the expression profile peaked within the expanding xylem. The variation may be due to sample separation as well as the number of samples per tissue making comparison difficult.

Expanding xylem

Multiple genes were associated with the GO term “Cell morphogenesis (GO 0000902)”, a general term referring to a change in a cell’s form or size. Within the expanding xylem, cells expand, or increase in size, and one of the genes, Potra2n4c9806, annotated with the “cell morphogenesis” showed a distinctive peak within the expanding xylem. This gene showed homology to an Arabidopsis transcription factor (TF) associated with the regulation of actin filament polymerisation. Actin filaments are one of the key elements of eukaryotic cytoskeleton and are vital in facilitating cell morphology (Tojkander et al., 2012). Based on similarities in expression profile observed for all six species, Potra2n4c9806 and its orthologs make for marker gene candidates for the expanding phloem.

As previously mentioned, the homologous SUS6 genes (Potra2n7c16228) that were identified were more likely to be SUS6 paralogs due to the overall shift in expression pattern. Studies have suggested two duplication events for the SUS gene family; first prior to divergence of angiosperms and gymnosperms, and the second within the angiosperms (Stein & Granot, 2019; Zhang et al., 2011). This could point to that the SUS6 gene identified in aspen in (Sundell et al., 2017) is a result from the second duplication event while the SUS6 gene identified across all species share a common ancestor predating divergence of angiosperms and gymnosperms. SUS genes encode proteins involved with the cleavage of sucrose, the main product of photosynthesis and are therefore commonly associated with the phloem. Additionally, SUS genes have been associated with cellulose synthesis and thereby have shown increased activity within the xylem (Stein & Granot, 2019). It is therefore possible that the SUS6 paralogs that were conserved across all six species were associated with cellulose synthesis. This could further explain the steady increase in gene expression within the expanding xylem displayed by the SUS6 paralogs (and additional genes), potentially involved in synthesising additional cellulose for SCW-construction.

SCW-forming xylem

After cessation of cell growth, secondary xylem cells develop a lignified SCW. Lignin binds cellulose and hemicelluloses, creating the more rigid SCW. Cells which develop SCW are primarily associated

with secondary xylem but are also found within the phloem (Zhang et al., 2018). This was reflected in the biphasic expression profiles of the genes with “Lignin biosynthetic process (GO:0009809)” annotation.

The biosynthesis of lignin involves the synthesis of monolignols which are subsequently oxidated and coupled into phenolic polymers (lignin) by laccases and peroxidases (Sundell et al., 2017). Monolignols are a group hydroxycinnamoyl alcohols, consisting primarily of p-coumaryl alcohols, coniferyl alcohols and sinapyl alcohols, differing slightly in structure and methylation degree. The lignin composition is therefore based on which monolignols are used as monomers which differs between plant species. One notable difference is that s lignin (from sinapyl alcohols) is primarily found amongst angiosperms than in gymnosperms (Weng & Chapple, 2010). The synthetic pathways of the various monolignols are, however, largely catalysed by the same enzymes.

Sundell et al. (2017) identified several LAC (laccase phenoloxidases) and PXR (peroxidase phenoloxidases) homologs within across wood forming aspen, as well as homologs of C4H (cinnamate 4-hydroxylase), C3H (p-coumarate 3-hydroxylase) and F5H (ferulate 5-hydroxylase). Homologs of C4H and C3H were identified amongst the ultra-conserved genes but were not amongst the ultra-conserved genes enriched in “Lignin biosynthetic process (GO:0009809)”. Instead, the enriched genes were identified as homologs encoding 4-hydroxycinnamoyl-CoA ligase (4CL), Potra2n1c307; a hydroxycinnamoyl transferase (HCT), Potra2n1c351; cinnamoyl-CoA reductase (CCR), Potra2n3c6847; cinnamyl alcohol dehydrogenase (CAD), Potra2n9c19275; and methyltransferases, Potra2n1c2649. These were all enzymes associated with monolignol synthesis (Weng & Chapple, 2010). Considering that the same set of enzymes are used for synthesising the different monolignols, it is plausible that differences in lignin composition between species would not affect the presence of these enzyme. With this in mind, all of the genes with the “Lignin biosynthetic process (GO:0009809)” annotation could be potential markers for the SCW-forming xylem.

Another set of genes that were active within the SCW-forming xylem were genes associated with the “Microtubule cytoskeleton organisation (GO 0000226)”. Localisation and patterning of the Cellulose Synthase Complex (CSC) proteins by microtubules is an important step in the construction of both PCW and SCW (Tobias et al., 2020). The expression profile of Potra2n3c6923 and its orthologs showed a distinct increase in expression within the SCW-forming xylem for most species. Based on expression alone, these genes also show potential as marker genes.

Cell death

The final step for cells within the secondary xylem, after lignification, is cell death. The fully mature cells gradually enter programmed cell death, i.e. a hollowing out of the cell, leaving only the lignified SCW (Bollhöner et al., 2012). The BFN1 gene is associated with DNA degradation in tracheary elements (Ito & Fukuda, 2002), and was identified as a marker gene for mature secondary xylem in Sundell et al. (2017). A homolog of BFN1 (Potra2n689s36475) was identified amongst the ultra-conserved genes without being enriched in any GO term. Despite some irregularities between the species, the general trend of the expression profiles for the BFN1 homologs are similar.

3.1.2 Ultra-conserved clade-specific genes

The ultra-conserved clade-specific genes were identified from orthogroups which only contained co-expressologs within the respective clade-pairs. It was of interest to see if these genes could potentially shed light on the differences in wood structure between angiosperms and gymnosperms.

The ultra-conserved angiosperm genes were enriched with a few of the same biological processes as the genes ultra-conserved across all pairs were such as cell wall biosynthesis and microtubule activity. However, the GO term "Lactone biosynthetic process (GO: 1901336)" was unique for the angiosperm species. Lactones are a class of organic cyclic molecules with a wide variety of possible functions, and the GO term is likely to be a parent term for "Stringolactone biosynthetic process (GO: 1901601)" and "Sesquiterpenoid biosynthetic process (GO: 0016106)". Stringolactones are mainly associated with plant development and promotion of symbiotic relationships between fungi and plants, but also play a role in stimulating secondary growth (Agusti et al., 2011). Sesquiterpenoids are a class of sesquiterpene lactones associated with a variety of functions such as antimicrobial properties and fragrance. What effect the stringolactones and sesquiterpenoids have on the wood structure is not clear and they do not seem to be tissue-specific (Perassolo et al., 2018).

The largest parent term for the gymnosperms was "Erythrose 4-phosphate/phosphoenolpyruvate family amino acid biosynthetic process (GO: 1902223)". Phenylalanine is an aromatic amin acid and a precursor for, amongst other components, lignin (Marchiosi et al., 2020), and erythrose 4-phosphate and phosphoenolpyruvates are intermediates in the synthesis of aromatic amino acids. This is a weak association to lignin, but could be of interest considering this was the largest parent term the ultra-conserved gymnosperm genes were enriched in.

The lack of other GO terms associated with specific wood forming processes amongst the gymnosperm genes could mean that the ultra-conserved genes have no particular or specific relevance to wood formation. The presence of wood formation terms amongst the angiosperms, and the lack thereof within gymnosperms, could be an interesting observation as this could potentially indicate that angiosperms evolved a second set of wood forming genes. Structural differences in wood structure such as the presence of vessel elements within angiosperms could be a result of evolving a second set of genes.

3.2 Using cliques to identify ultra-conserved genes

The workflow for identifying the ultra-conserved orthogroups consisted of several filtering steps, effectively removed orthologs pairs or orthogroups based on certain conditions. This section will look into the different decisions made to identify conserved genes and how altering parameters could change the outcome.

3.2.1 Significance of cliques

Comparison of co-expression networks is a powerful method for identifying conserved orthologs. By connecting molecular function, predicted through similarity in sequence, to similarity in co-

expression, distinguishes orthologs from paralogs. However, this comparison was performed for pairs of species. The identification of fully conserved orthogroups required co-expressologs within all 15 species pairs. This was a simplistic approach that didn't differentiate between orthologs and paralogs. Orthologs between two species will share the same ancestral gene, however, this gene could be a result of gene duplication which will only be detected once a third, more distant, species is included. Both the gymnosperms and angiosperms that were studied are a result of several speciation and duplication events. Therefore, to be able to filter out which homologous genes could potentially predate the divergence of angiosperms and gymnosperms, cliques were identified.

In the context of this study, a clique was defined as a set of overlapping co-expressologs within an orthogroup in which all 15 co-expressologs are made up of the same six species-specific genes, i.e. a complete set of consistently 1-1 co-expressologs. This can further be described as subnetwork in which all nodes are inter-connected, with each nodes representing orthologs and edges connecting co-expressolog members. An ortholog pair was assumed to possess similar functions based on sequence similarity, i.e. a conserved sequence. However, co-expressologs are ortholog pairs which also have conserved expression. This means that co-expressologs are co-expressed with largely the same set of genes, increasing the probability of displaying conserved process-specific function. If the same six orthologs form a clique of co-expressologs, it can be implied that these six genes form a subgroup of orthologs with consistent functional similarity compared to the others. This can be seen in Figure 10 where three sets of genes associated with the organisation of microtubule cytoskeleton display unique clique-based profile. A different perspective is that a clique will not contain paralogs conserved between two species which a co-expressolog might. Since paralogs are a result of gene duplication, they tend to adapt new functions, compared to orthologs (result of speciation) which retain function. The cliques identify homologous genes with conserved co-expression networks across all species, and therefore identify genes with a common ancestor.

3.2.2 Clique identification

The identification of cliques was achieved using the igraph package which supports the search for both largest and maximal cliques, with either weighted or unweighted edges. To allow more flexibility, the networks created were weighted using the p-values of each ortholog pair (see Materials and Methods). The process of identifying clade-specific cliques and cliques across all species was largely similar. In both instances, only the orthogroups containing expressed ortholog pairs for either all 15 pairs or all 3 pairs were selected. This was largely done to streamline the process, but also allowed the use of the same function, largest_weighted_cliques. The largest_weigthed_clique function identifies all cliques with the largest edge sum (see Materials and Methods), returning a list. Due to the large number of non-conserved ortholog pairs, a pre-filtering step removing all pairs with FDR-corrected p-value > 0,8. This reduced the number of ortholog pairs quite significantly. By doing so, the number of potential partial cliques was reduced, but so were the memory requirements. When identifying cliques across all pairs, only one of the largest weighted cliques from each orthogroup was saved. This also was a memory reducing effort but may have potentially reduce the number of orthogroups with cliques. Assuming that the cliques are identified and listed randomly, the selected clique may not necessarily be the largest (i.e. with most co-expressologs). Selecting only the first listed clique per orthogroup did have a slight effect when increasing the pre-filtering p-value cutoff from 0,8 to 0,9. This resulted in a reduction of orthogroups

with cliques. After comparing the selected cliques form the same orthogroup, it was revealed that slightly different ortholog pairs were included when increasing the p-value cutoff of which some were not co-expressologs.

This issue was overcome when identifying clade-specific cliques. Multiple cliques per orthogroup were identified, and the clique with lowest p-value sum (highest weight) was selected. This additional step effectively reduced the risk of orthogroups "losing" the clique. The initial reasoning for not performing this extra step for the orthogroups conserved across all species was that not many cliques per orthogroup were expected, however, this could potentially be added as an improvement.

3.3 Concluding remarks and future work

This study identified genes conserved *across* wood forming tissue between angiosperms and gymnosperms that could potentially belong to sets of regulatory genes, governing wood formation.

These genes were enriched in various biological processes associated with wood formation, and based on GO, previously identified marker genes, and through manual search, marker gene candidates for the various tissues were suggested. All marker gene candidates displayed were suggested based on how universal the expression profile was across all species. Homologs of PP2 were suggested as a marker gene for the phloem and was identified solely from expression profiles; based on previous findings as well as GO, homologs of CDC2 was considered highly descriptive of the cambium; for the expanding xylem a homolog of TF associated with actin filament polymerisation was suggested based on GO homologs; a cohort of homolog encoding monolignol biosynthesis enzymes showed potential as marker gene within the SCW-forming xylem based on GO, and homologs of BNF1 showed some potential as a marker for controlled cell death within the mature xylem, corresponding to previous findings.

Given the exploratory nature of this study, there are numerous aspects that warrant further investigation. This includes both a deeper understanding of the roles that the ultra-conserved genes play, but also changes in the methodologies to be considered.

Genes from only a select few GOs were studied, and many more could hold potential as marker genes. Additionally, the genes suggested marker gene candidates were largely tissue specific. It could therefore be of interest to also identify genes capturing the reprogramming events in the transition between two tissues. The genes were also conserved across all samples and are therefore likely to be regulators governing secondary growth. This raises the question of which genes would be identified from studying only a subset of samples, such as tissue specific samples. These genes would not be regulative of wood formation as an entire process but may govern specific transitions or tissue-specific processes.

To narrow the scope, the ultra-conserved genes that were highlighted and discussed in this thesis were based on their specificity to the various tissues or processes and identified primarily through GO. However, so-called master regulator genes such as NAC and VND are known to be involved in the regulation of xylem formation within both gymnosperm and angiosperm species (Jokipii-Lukkari et al., 2017; Kubo et al., 2005). Considering the central role a regulating gene has within a process,

further work could include an in-depth search into which (if any) of these master regulator genes were amongst the ultra-conserved genes.

Additionally, gene enrichment for the genes of partially conserved orthologs or associated with partial cliques could reveal other interesting GO terms. Not all ultra-conserved genes had GO-annotation therefore terms may have been left out. However, gene function may not easily be understood, especially when studying wood formation in trees due to the lengthy growth period. Many gene annotations are inferred through orthologous genes in Arabidopsis thaliana, but due to restricted amounts of secondary xylem produced during secondary growth (Bollhöner et al., 2012) many tree-specific genes cannot be inferred. In time, knock-out experiments could potentially shed light on some of these unknown genes.

Conserved orthogroups were identified based on the number of species pairs with co-expressologs. A basis for the classification of co-expressologs was the size, or density, of the co-expression network effectively determining how many neighbours were to be included and compared in the hyper geometric test (see 4.2ComPlEx: Comparative analysis for Plant co-Expression networks). However, is a density threshold of 3% too strict? Increasing the density, allowing more neighbours, could result in a higher number of conserved genes, but is this necessary to describe a genes co-expression network? Including too few neighbours will probably result in a weaker statistical basis to evaluate the conservation of networks, and too many neighbours may result in less relevant connections – or would it? Co-expressologs are ortholog pairs which are co-expressed with other conserved orthologs, thereby co-expressologs are considered conserved based on 1) similarity in amino acid sequence thus assuming some similarity in function, and 2) similarity in which genes they are co-expressed with. If by increasing the network density included less-relevant neighbours, then these neighbours are not likely to have orthologs in the ortholog genes network. This would be a balance between identifying more co-expressologs and introducing noise.

Despite numerous studies investigating various aspects of secondary growth within gymnosperms and angiosperms, very little is known about the underlaying ancestral mechanisms. The work for this thesis has contributed to this knowledge gap by identifying highly conserved genes across a broad set of gymnosperms and angiosperms. A deeper understanding of the roles which these genes play within the various wood-forming processes is beyond the scope of this thesis. However, by touching upon some of the conserved aspects between the gymnosperms and angiosperms, this study has laid the groundwork and hopes to motivate for further exploration.

4 Materials and methods

All analyses and figures were achieved using R Statistical Software (v4.3.1, Team (2023)) using personal laptop and the Orion computer cluster at NMBU. Code is available on GitHub: https://github.com/ellendim/githubEvoTree

Colour pallets used for the visualisation were assessed as colour-blind compatible (https://davidmathlogic.com/).

4.1 Sampling and RNA-extraction

The following steps were performed for three gymnosperm species, lodgepole pine (*Pinus contorta*), Scots pine (*Pinus sylvestris*), and Norway spruce (*Picea abies*), and three angiosperm species, cherry (*Prunus avium*), aspen (*Populus tremula*), and birch (*Betula pendula*).

Longitudinal sections (15 µm thick) were cut from wood blocks using a cryo-microtome and were subsequently stored at -80 ^{o}C. The sections were pooling into samples based on cell content characterising the various tissue types: phloem, cambium, expanding xylem, SCW-forming xylem, and mature xylem. Tissue characterisation was performed using light microscope. Total RNA was extracted using miRNeasy Mini Kit (Quiagen) and was sequenced using Illumina HiSeq 2000 platform (2x150bp stranded reads). Gene-based read counts were calculated using Salmon (Patro et al., 2017), and were subsequently normalised using DESeq-implemented variance stabilised transformation (VST).

4.2 Comparison files

The co-expression networks of orthologous genes were compared using ComPlEx (Netotea et al., 2014) for all 15 species pairs. The co-expression networks were based on correlating gene expressions across samples using Pearson's correlation, and subsequently ranking the genes from highest to lowest correlation (within each gene). The density setting, defining the size of the co-expression network was set at 0,03, i.e. the 3%. A hypergeometric test was performed for all genes with at least one neighbour, resulting in each ortholog pair obtaining two p-values (one for each ortholog). Genes with no neighbours were assigned a p-value of 1. The p-values were then FDR-corrected at a level of 0,05. Each pair-wise species comparison resulted in a comparison-file containing all orthologs that were expressed, including the number of neighbours and respective FDR-corrected p-values. Only the highest FDR-corrected p-value for each gene pair was used throughout this study. Orthologs with conserved expression, or co-expressologs, were identified by at FDR <0,1. Due to memory requirements all comparison-files were compiled using the Orion.

4.3 Conserved orthogroups

The number of orthogroups per species, that contained at least one ortholog, was visualised alongside the largest intercepts using the UpSetR package (v1.4.0, Gehlenborg (2019), Lex and Gehlenborg (2014)). The *upset* function required the input data frame to contain only binary values, meaning that each column (species) consisted of rows (orthogroups) with cell values of 1, if there was at least one gene within the orthogroup, or 0, if no genes were present. A similar data frame set-up was used for identifying the conserved and partially conserved orthogroups. With each column now representing a species pair, and each row containing a value of 1 if the species pair had at least one co-expressolog within the orthogroup. All orthogroups with a row sum of 15 were considered conserved across all species (across clades) and all orthogroups with at least 2 gymnosperm pairs, 2 angiosperm pairs and 2 mixed-clade pairs were considered partially conserved. Orthogroups with co-expressologs only amongst the three angiosperm or gymnosperm pairs were conserved clade-specific orthogroups if the row sum was 3 and partially conserve clade-specific orthogroups if the row sum was equal to or greater than 2.

4.4 Ultra-conserved orthogroups

Orthogroups with cliques were identified using the igraph package (v1.60.0, Csárdi and Nepusz (2006)). For each orthogroup, a weighted network was created using all orthologs as nodes connected by a weighted edge, with the weight being the max p-value (from the comparison files). Cliques involving all six species were found using the *largest_weighted_clique* function. To ensure that the cliques containing most co-expressologs were selected, the FDR-corrected p-values were negative log transformed. Once all orthogroups containing cliques were identified, non-conserved ortholog pairs (FDR > 0,1) were removed. This identified partial cliques, which were based on the same species pair requirements as for the partially conserved orthogroups, and complete cliques which only contained co-expressologs. The orthogroups with complete cliques were defined as ultra-conserved, and the genes associated with each clique were defined as ultra-conserved genes. Prior to clique identification, ortholog pairs with FDR > 0,8 were removed. Cliques involving only angiosperm or gymnosperm species, i.e. only three species, were also identified using the *largest_weighted_clique* function with negative log transformed FDR p-values. However, to streamline the process, only orthogroups containing all three clade-specific species was used. In addition, all weighted cliques containing all three species were identified. The largest weighted clique within each orthogroup was then selected. Partial and complete cliques were identified in the same manner as the cliques for all six species with FDR < 0,1.

Slice variants (genes with same annotation occurring in different orthogroups) were allowed when identifying fully and partially conserved orthogroups as well as ultra and partial cliques. However, for plotting heatmaps these were removed.

4.5 Sample comparison heatmaps

The samples between each species pair were compared based on similarity in gene expression using co-expressologs. Using Pearson's correlation, the gene expression within each sample was compared between the species. It was therefore vital that the rows, i.e. genes, in both expression data sets were ordered similarly so that each gene from a co-expressolog would appear in the same row for both species. Samples within the same species were also correlated using all genes in the transcription data set. Heatmaps were created with the ComplexHeatmap package (Gu et al., 2016). As the range of correlation varied between species pairs, each heatmap had separate colour scales to better highlight the correlation patterns. Two legends were created to reflect the correlation range for each species using the lowest and highest correlation value per matrix. The colours used to identify range were selected to signal "high" and "low" correlation without giving the misconception of positive and negative correlation.

4.6 Expression heatmaps

Heatmaps visualising gene expression patterns for each species were also created with the ComplexHeatmap package, using only co-expressologs. The genes, i.e. rows, were scaled and centred, and clustering the rows using the "ward.D2"-method. The clustering was based on a distance matrix created using Euclidean distance. Due to low read count, birch and Norway spruce had non-continuous sample ranges. Therefore, imputed expression data sets, where missing samples had been replaced with the mean of the adjacent samples, were used for a more continuous visualisation. All six heatmaps used the same colour scale, with a red-white-blue colour scale used to indicate up-regulated, neutral, and down-regulated gene expression, respectively.

Expression heatmaps using only the ultra-conserved genes were created in a similar way as the heatmaps mentioned above. However, to allow better comparison between species, the scaled and centred expressions data sets for all six species were combined prior to clustering and arranged by orthogroup.

4.7 GO enrichment analysis

A GO enrichment analysis was performed using the ultra-conserved genes to see which ontologies the genes were enriched in, and was performed using the GSEABase and GOstats packages (Falcon & Gentleman, 2007; Morgan et al., 2023). The parameters were based on annotated aspen genes with associated gene ontology (GO) terms and ID's. A hypergeometric test was performed based on parameters set using GSEAGOHyperGParams function setting ontology as "BP" (biological process) and setting a p-value cutoff of 0,05. The GO IDs were grouped into higher order terms using the web-based tool REVIGO (Supek et al., 2011), and visualised using the treemap function. Code for creating the treemaps were downloaded from the REVIGO website and further modified. The browser QuickGo (Binns et al., 2009) was used to track potential lower-order or associated terms of the higher order terms which REVIGO presented. Based on both the treemaps and manually checking for GO terms, expression profiles for genes associated with the various GO terms were plotted using the

geom_line function with colours linking genes from the various species from the same orthogroup (clique). Some GO terms were associated with more than five genes, but no more than five genes were visualised. The genes that were not visualised were either expressed similarly to the other genes and was redundant in terms of visualising or shared a similar expression profile at much lower levels obscuring the expression profiles. The marker genes from (Sundell et al., 2017) were mapped to the *Populus trichocarpa* genome which meant that the ultra-conserved genes were identified using *p. trichocarpa* genes as queries in the compiled ortholog group file of orthologous genes predicted using OrthoFinder.

5 References

Adamcsek, B., Palla, G., Farkas, I. J., Derényi, I., & Vicsek, T. (2006). CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, *22*(8), 1021-1023. https://doi.org/10.1093/bioinformatics/btl039

Agusti, J., Herold, S., Schwarz, M., Sanchez, P., Ljung, K., Dun, E. A., Brewer, P. B., Beveridge, C. A., Sieberer, T., Sehr, E. M., & Greb, T. (2011). Strigolactone signaling is required for auxin-dependent stimulation of secondary growth in plants. *Proceedings of the National Academy of Sciences*, *108*(50), 20242-20247. https://doi.org/doi:10.1073/pnas.1111902108

Ashokkumar, V., Venkatkarthick, R., Jayashree, S., Chuetor, S., Dharmaraj, S., Kumar, G., Chen, W.-H., & Ngamcharussrivichai, C. (2022). Recent advances in lignocellulosic biomass for biofuels and value-added bioproducts - A critical review. *Bioresource Technology*, *344*, 126195. https://doi.org/https://doi.org/10.1016/j.biortech.2021.126195

Barabási, A.-L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, *5*(2), 101-113. https://doi.org/10.1038/nrg1272

Berglund, J., Mikkelsen, D., Flanagan, B. M., Dhital, S., Gaunitz, S., Henriksson, G., Lindström, M. E., Yakubov, G. E., Gidley, M. J., & Vilaplana, F. (2020). Wood hemicelluloses exert distinct biomechanical contributions to cellulose fibrillar networks. *Nature Communications*, *11*(1), 4692. https://doi.org/10.1038/s41467-020-18390-z

Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'Donovan, C., & Apweiler, R. (2009). QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics*, *25*(22), 3045-3046. https://doi.org/10.1093/bioinformatics/btp536

Birkeland, S., Slotte, T., Krag Brysting, A., Gustafsson, A. L. S., Rhoden Hvidsten, T., Brochmann, C., & Nowak, M. D. (2022). What can cold-induced transcriptomes of Arctic Brassicaceae tell us about the evolution of cold tolerance? *Molecular Ecology*, *31*(16), 4271-4285. https://doi.org/https://doi.org/10.1111/mec.16581

Bollhöner, B., Prestele, J., & Tuominen, H. (2012). Xylem cell death: emerging understanding of regulation and function. *Journal of Experimental Botany*, *63*(3), 1081-1094. https://doi.org/10.1093/jxb/err438

Cheng, D.-X., Wang, X.-H., Wang, C.-L., Li, X.-Y., Ye, Z.-L., & Li, W.-F. (2024). Cambium Reactivation Is Closely Related to the Cell-Cycle Gene Configuration in Larix kaempferi. *International Journal of Molecular Sciences*, *25*(7), 3578. https://www.mdpi.com/1422-0067/25/7/3578

Collas, P., Liyakat Ali, T. M., Brunet, A., & Germier, T. (2019). Finding Friends in the Crowd: Three-Dimensional Cliques of Topological Genomic Domains [Mini Review]. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.00602

Consortium, T. G. O., Aleksander, S. A., Balhoff, J., Carbon, S., Cherry, J. M., Drabkin, H. J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N. L., Hill, D. P., Lee, R., Mi, H., Moxon, S., Mungall, C. J., Muruganugan, A., Mushayahama, T., Sternberg, P. W., Thomas, P. D., . . . Westerfield, M. (2023). The Gene Ontology knowledgebase in 2023. *Genetics*, *224*(1). https://doi.org/10.1093/genetics/iyad031

Csárdi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, *Complex Systems*, 1695. https://igraph.org

Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloë, D., Le Gall, C., Schaëffer, B., Le Crom, S., Guedj, M., Jaffrézic, F., & Consortium, o. b. o. T. F. S. (2012). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, *14*(6), 671-683. https://doi.org/10.1093/bib/bbs046

Dinant, S., Clark, A. M., Zhu, Y., Vilaine, F. o., Palauqui, J.-C., Kusiak, C., & Thompson, G. A. (2003). Diversity of the Superfamily of Phloem Lectins (Phloem Protein 2) in Angiosperms. *Plant Physiology*, *131*(1), 114-128. https://doi.org/10.1104/pp.013086

Earle, C. J. (2024, 01.01.24). *The Gymnosperm Database*. https://www.conifers.org/index.php

Emamjomeh, A., Saboori Robat, E., Zahiri, J., Solouki, M., & Khosravi, P. (2017). Gene co-expression network reconstruction: a review on computational methods for inferring functional information from plant-based expression data. *Plant Biotechnology Reports*, *11*(2), 71-86. https://doi.org/10.1007/s11816-017-0433-z

Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238. https://doi.org/10.1186/s13059-019-1832-y

Falcon, S., & Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics*, *23*.

FitzJohn, R. G., Pennell, M. W., Zanne, A. E., Stevens, P. F., Tank, D. C., & Cornwell, W. K. (2014). How much of the world is woody? *Journal of Ecology*, *102*(5), 1266-1272. https://doi.org/https://doi.org/10.1111/1365-2745.12260

Gehlenborg, N. (2019). *UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets*. In https://CRAN.R-project.org/package=UpSetR

Gensel, P. G. (2008). The Earliest Land Plants. *Annual Review of Ecology, Evolution, and Systematics*, *39*(Volume 39, 2008), 459-477. https://doi.org/https://doi.org/10.1146/annurev.ecolsys.39.110707.173526

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333-351. https://doi.org/10.1038/nrg.2016.49

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btw313

Higdon, R. (2013). Multiple Hypothesis Testing. In W. Dubitzky, O. Wolkenhauer, K.-H. Cho, & H. Yokota (Eds.), *Encyclopedia of Systems Biology* (pp. 1468-1469). Springer New York. https://doi.org/10.1007/978-1-4419-9863-7_1211

Ito, J., & Fukuda, H. (2002). ZEN1 Is a Key Enzyme in the Degradation of Nuclear DNA during Programmed Cell Death of Tracheary Elements. *The Plant Cell*, *14*(12), 3201-3211. https://doi.org/10.1105/tpc.006411

Jokipii-Lukkari, S., Sundell, D., Nilsson, O., Hvidsten, T. R., Street, N. R., & Tuominen, H. (2017). NorWood: a gene expression resource for evo-devo studies of conifer wood development. *New Phytologist*, *216*(2), 482-494. https://doi.org/https://doi.org/10.1111/nph.14458

Joshi, T., & Xu, D. (2007). Quantitative assessment of relationship between sequence similarity and function similarity. *BMC Genomics*, *8*(1), 222. https://doi.org/10.1186/1471-2164-8-222

Kaiser, S., & Scheuring, D. (2020). To Lead or to Follow: Contribution of the Plant Vacuole to Cell Growth [Mini Review]. *Frontiers in Plant Science*, *11*. https://doi.org/10.3389/fpls.2020.00553

Kehr, J. (2006). Phloem sap proteins: their identities and potential roles in the interaction between plants and phloem-feeding insects. *Journal of Experimental Botany*, *57*(4), 767-774. https://doi.org/10.1093/jxb/erj087

Kim, M. H., Bae, E. K., Lee, H., & Ko, J. H. (2022). Current Understanding of the Genetics and Molecular Mechanisms Regulating Wood Formation in Plants. *Genes (Basel)*, *13*(7). https://doi.org/10.3390/genes13071181

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet*, *39*, 309-338. https://doi.org/10.1146/annurev.genet.39.073003.114725

Kubo, M., Udagawa, M., Nishikubo, N., Horiguchi, G., Yamaguchi, M., Ito, J., Mimura, T., Fukuda, H., & Demura, T. (2005). Transcription switches for protoxylem and metaxylem vessel formation. *Genes & Development*, *19*(16), 1855-1860. https://doi.org/10.1101/gad.1331305

Lee, J., Shah, M., Ballouz, S., Crow, M., & Gillis, J. (2020). CoCoCoNet: conserved and comparative co-expression across a diverse set of species. *Nucleic Acids Research*, *48*(W1), W566-W571. https://doi.org/10.1093/nar/gkaa348

Lex, A., & Gehlenborg, N. (2014). Sets and intersections. *Nature Methods*, *11*(8), 779-779. https://doi.org/10.1038/nmeth.3033

Li, H., Chen, G., Pang, H., Wang, Q., & Dai, X. (2021). Investigation Into Different Wood Formation Mechanisms Between Angiosperm and Gymnosperm Tree Species at the Transcriptional and Post-transcriptional Level [Original Research]. *Frontiers in Plant Science*, *12*. https://doi.org/10.3389/fpls.2021.698602

Liesche, J., & Schulz, A. (2018). Phloem transport in gymnosperms: a question of pressure and resistance. *Current Opinion in Plant Biology*, *43*, 36-42. https://doi.org/https://doi.org/10.1016/j.pbi.2017.12.006

Marchiosi, R., dos Santos, W. D., Constantin, R. P., de Lima, R. B., Soares, A. R., Finger-Teixeira, A., Mota, T. R., de Oliveira, D. M., Foletto-Felipe, M. d. P., Abrahão, J., & Ferrarese-Filho, O. (2020). Biosynthesis and metabolic actions of simple phenolic acids in plants. *Phytochemistry Reviews*, *19*(4), 865-906. https://doi.org/10.1007/s11101-020-09689-2

Mastrorosa, F. K., Miller, D. E., & Eichler, E. E. (2023). Applications of long-read sequencing to Mendelian genetics. *Genome Medicine*, *15*(1), 42. https://doi.org/10.1186/s13073-023-01194-3

McIntyre, L. M., Lopiano, K. K., Morse, A. M., Amin, V., Oberg, A. L., Young, L. J., & Nuzhdin, S. V. (2011). RNA-seq: technical variability and sampling. *BMC Genomics*, *12*(1), 293. https://doi.org/10.1186/1471-2164-12-293

Meents, M. J., Watanabe, Y., & Samuels, A. L. (2018). The cell biology of secondary cell wall biosynthesis. *Annals of Botany*, *121*(6), 1107-1125. https://doi.org/10.1093/aob/mcy005

Morgan, M., Falcon, S., & Gentleman, R. (2023). *GSEABase: Gene set enrichment data structures and methods*. In https://bioconductor.org/packages/GSEABase

Netotea, S., Sundell, D., Street, N. R., & Hvidsten, T. R. (2014). ComPlEx: conservation and divergence of co-expression networks in A. thaliana, Populus and O. sativa. *BMC Genomics*, *15*(1), 106. https://doi.org/10.1186/1471-2164-15-106

Obayashi, T., & Kinoshita, K. (2009). Rank of Correlation Coefficient as a Comparable Measure for Biological Significance of Gene Coexpression. *DNA Research*, *16*(5), 249-260. https://doi.org/10.1093/dnares/dsp016

Oldham, M. C., Horvath, S., & Geschwind, D. H. (2006). Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proceedings of the National Academy of Sciences*, *103*(47), 17973-17978. https://doi.org/doi:10.1073/pnas.0605938103

Ovens, K., Eames, B. F., & McQuillan, I. (2021). Comparative Analyses of Gene Co-expression Networks: Implementations and Applications in the Study of Evolution [Review]. *Frontiers in Genetics*, *12*. https://doi.org/10.3389/fgene.2021.695399

Panchy, N., Lehti-Shiu, M., & Shiu, S. H. (2016). Evolution of Gene Duplication in Plants. *Plant Physiol*, *171*(4), 2294-2316. https://doi.org/10.1104/pp.16.00523

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417-419. https://doi.org/10.1038/nmeth.4197

Perassolo, M., Cardillo, A. B., Busto, V. D., Giulietti, A. M., & Talou, J. R. (2018). Biosynthesis of Sesquiterpene Lactones in Plants and Metabolic Engineering for Their Biotechnological Production. In V. P. Sülsen & V. S. Martino (Eds.), *Sesquiterpene Lactones: Advances in their Chemistry and Biological Aspects* (pp. 47-91). Springer International Publishing. https://doi.org/10.1007/978-3-319-78274-4_4

Pikaard, C. S., & Mittelsten Scheid, O. (2014). Epigenetic regulation in plants. *Cold Spring Harb Perspect Biol*, *6*(12), a019315. https://doi.org/10.1101/cshperspect.a019315

Pradhan, M. P., Nagulapalli, K., & Palakal, M. J. (2012). Cliques for the identification of gene signatures for colorectal cancer across population. *BMC Syst Biol*, *6 Suppl 3*(Suppl 3), S17. https://doi.org/10.1186/1752-0509-6-s3-s17

Ramage, M. H., Burridge, H., Busse-Wicher, M., Fereday, G., Reynolds, T., Shah, D. U., Wu, G., Yu, L., Fleming, P., Densley-Tingley, D., Allwood, J., Dupree, P., Linden, P. F., & Scherman, O. (2017). The wood from the trees: The use of timber in construction. *Renewable and Sustainable Energy Reviews*, *68*, 333-359. https://doi.org/https://doi.org/10.1016/j.rser.2016.09.107

Rao, X., & Dixon, R. A. (2019). Co-expression networks for plant biology: why and how. *Acta Biochimica et Biophysica Sinica*, *51*(10), 981-988. https://doi.org/https://doi.org/10.1093/abbs/gmz080

Reece, J. B., & Campbell, N. A. (2011). *Campbell biology* (9th ed.). Benjamin Cummings / Pearson.

Růžička, K., Ursache, R., Hejátko, J., & Helariutta, Y. (2015). Xylem development – from the cradle to the grave. *New Phytologist*, *207*(3), 519-535. https://doi.org/https://doi.org/10.1111/nph.13383

Sanderson, M. J. (2003). Molecular data from 27 proteins do not support a Precambrian origin of land plants. *American Journal of Botany*, *90*(6), 954-956. https://doi.org/https://doi.org/10.3732/ajb.90.6.954

Schmulsky, R., & Jones, P. D. (2019). *Forest Products and Wood Science* (7 ed.). Wiley Blackwell.

Shekhovtsov, A. (2021). How strongly do rank similarity coefficients differ used in decision making problems? *Procedia Computer Science*, *192*, 4570-4577. https://doi.org/https://doi.org/10.1016/j.procs.2021.09.235

Shi, D., Lebovka, I., López-Salmerón, V., Sanchez, P., & Greb, T. (2019). Bifacial cambium stem cells generate xylem and phloem during radial plant growth. *Development*, *146*(1). https://doi.org/10.1242/dev.171355

Song, L., Langfelder, P., & Horvath, S. (2012). Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics*, *13*(1), 328. https://doi.org/10.1186/1471-2105-13-328

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics*, *20*(11), 631-656. https://doi.org/10.1038/s41576-019-0150-2

Stein, O., & Granot, D. (2019). An Overview of Sucrose Synthases in Plants [Review]. *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.00095

Stevens, P. F. o. (2024, 14.10.2023). *Angiosperm Phylogeny Website*. Retrieved 15.02.24 from http://www.mobot.org/MOBOT/research/APweb/

Sundell, D., Street, N. R., Kumar, M., Mellerowicz, E. J., Kucukoglu, M., Johnsson, C., Kumar, V., Mannapperuma, C., Delhomme, N., Nilsson, O., Tuominen, H., Pesquet, E., Fischer, U., Niittylä, T., Sundberg, B., & Hvidsten, T. R. (2017). AspWood: High-Spatial-Resolution Transcriptome Profiles Reveal Uncharacterized Modularity of Wood Formation in Populus tremula *The Plant Cell*, *29*(7), 1585-1604. https://doi.org/10.1105/tpc.17.00153

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLOS ONE*, *6*(7), e21800. https://doi.org/10.1371/journal.pone.0021800

Team, R. C. (2023). *R: A Language and Environment for Statistical Computing*. In R Foundation for Statistical Computing. https://www.R-project.org/

Tobias, L. M., Spokevicius, A. V., McFarlane, H. E., & Bossinger, G. (2020). The Cytoskeleton and Its Role in Determining Cellulose Microfibril Angle in Secondary Cell Walls of Woody Tree Species. *Plants*, *9*(1), 90. https://www.mdpi.com/2223-7747/9/1/90

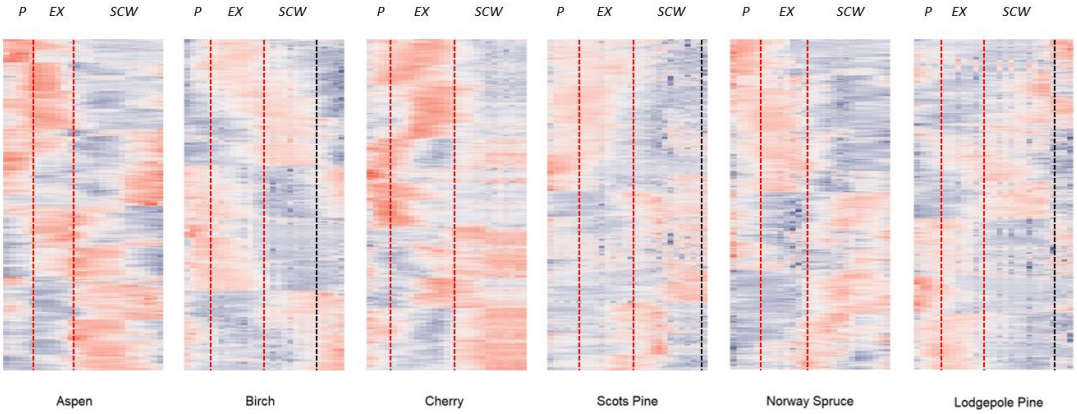
Tojkander, S., Gateva, G., & Lappalainen, P. (2012). Actin stress fibers – assembly, dynamics and biological roles. *Journal of Cell Science*, *125*(8), 1855-1864. https://doi.org/10.1242/jcs.098087

Uggla, C., Moritz, T., Sandberg, G., & Sundberg, B. (1996). Auxin as a positional signal in pattern formation in plants. *Proceedings of the National Academy of Sciences*, *93*(17), 9282-9286. https://doi.org/doi:10.1073/pnas.93.17.9282

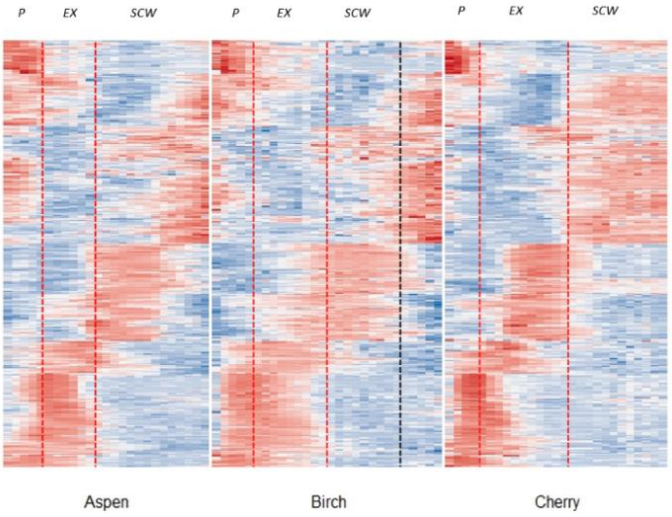
University of Oslo. (2011, 16.01.20). *Lignin*. University of Oslo. https://www.mn.uio.no/ibv/tjenester/kunnskap/plantefys/leksikon/l/lignin.html
Vailati-Riboni, M., Palombo, V., & Loor, J. J. (2017). What Are Omics Sciences? In B. N. Ametaj (Ed.), *Periparturient Diseases of Dairy Cows: A Systems Biology Approach* (pp. 1-7). Springer International Publishing. https://doi.org/10.1007/978-3-319-43033-1_1
van der Kloet, F. M., Buurmans, J., Jonker, M. J., Smilde, A. K., & Westerhuis, J. A. (2020). Increased comparability between RNA-Seq and microarray data by utilization of gene sets. *PLOS Computational Biology*, *16*(9), e1008295. https://doi.org/10.1371/journal.pcbi.1008295
Wang, D., Chen, Y., Li, W., Li, Q., Lu, M., Zhou, G., & Chai, G. (2021). Vascular Cambium: The Source of Wood Formation [Mini Review]. *Frontiers in Plant Science*, *12*. https://doi.org/10.3389/fpls.2021.700928
Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, *10*(1), 57-63. https://doi.org/10.1038/nrg2484
Weng, J.-K., & Chapple, C. (2010). The origin and evolution of lignin biosynthesis. *New Phytologist*, *187*(2), 273-285. https://doi.org/https://doi.org/10.1111/j.1469-8137.2010.03327.x
Wolfe, C. J., Kohane, I. S., & Butte, A. J. (2005). Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics*, *6*, 227. https://doi.org/10.1186/1471-2105-6-227
Xie, L., Yang, C., & Wang, X. (2011). Brassinosteroids can regulate cellulose biosynthesis by controlling the expression of CESA genes in Arabidopsis. *Journal of Experimental Botany*, *62*(13), 4495-4506. https://doi.org/10.1093/jxb/err164
Xu, C., Shen, Y., He, F., Fu, X., Yu, H., Lu, W., Li, Y., Li, C., Fan, D., Wang, H. C., & Luo, K. (2019). Auxin-mediated Aux/IAA-ARF-HB signaling cascade regulates secondary xylem development in Populus. *New Phytol*, *222*(2), 752-767. https://doi.org/10.1111/nph.15658
Zarra, I., Revilla, G., Sampedro, J., & Valdivia, E. R. (2020). Biosynthesis and Regulation of Secondary Cell Wall. In F. M. Cánovas, U. Lüttge, C. Leuschner, & M.-C. Risueño (Eds.), *Progress in Botany Vol. 81* (pp. 189-226). Springer International Publishing. https://doi.org/10.1007/124_2019_27
Zhang, D., Xu, B., Yang, X., Zhang, Z., & Li, B. (2011). The sucrose synthase gene family in Populus: structure, expression, and evolution. *Tree Genetics & Genomes*, *7*(3), 443-456. https://doi.org/10.1007/s11295-010-0346-2
Zhang, J., Xie, M., Tuskan, G. A., Muchero, W., & Chen, J.-G. (2018). Recent Advances in the Transcriptional Regulation of Secondary Cell Wall Biosynthesis in the Woody Plants [Review]. *Frontiers in Plant Science*, *9*. https://doi.org/10.3389/fpls.2018.01535
Zheng, X., Liu, T., Yang, Z., & Wang, J. (2011). Large cliques in Arabidopsis gene coexpression network and motif discovery. *Journal of Plant Physiology*, *168*(6), 611-618. https://doi.org/https://doi.org/10.1016/j.jplph.2010.09.010
Zhong, R., Cui, D., & Ye, Z.-H. (2019). Secondary cell wall biosynthesis. *New Phytologist*, *221*(4), 1703-1723. https://doi.org/https://doi.org/10.1111/nph.15537
Zhong, R., Lee, C., & Ye, Z.-H. (2009). Functional Characterization of Poplar Wood-Associated NAC Domain Transcription Factors. *Plant Physiology*, *152*(2), 1044-1055. https://doi.org/10.1104/pp.109.148270
Zvelebil, M., & Baum, J. O. (2008). *Understanding bioinformatics*. Garland Science.

6 Appendix

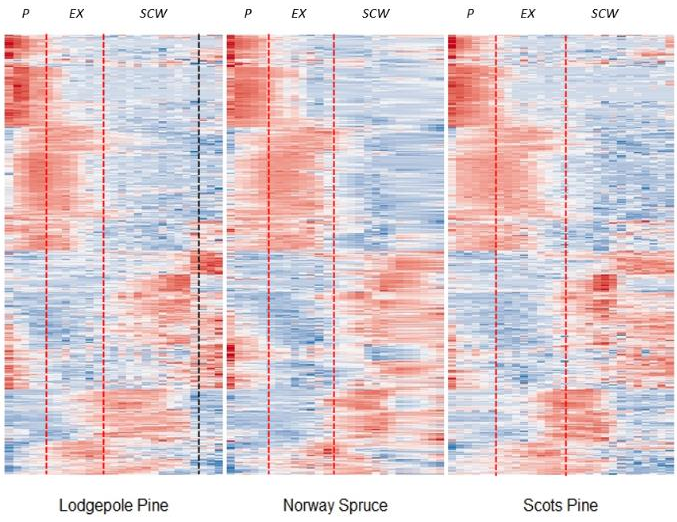
Supplementary figure 1: Expression heatmaps for all co-expressologs



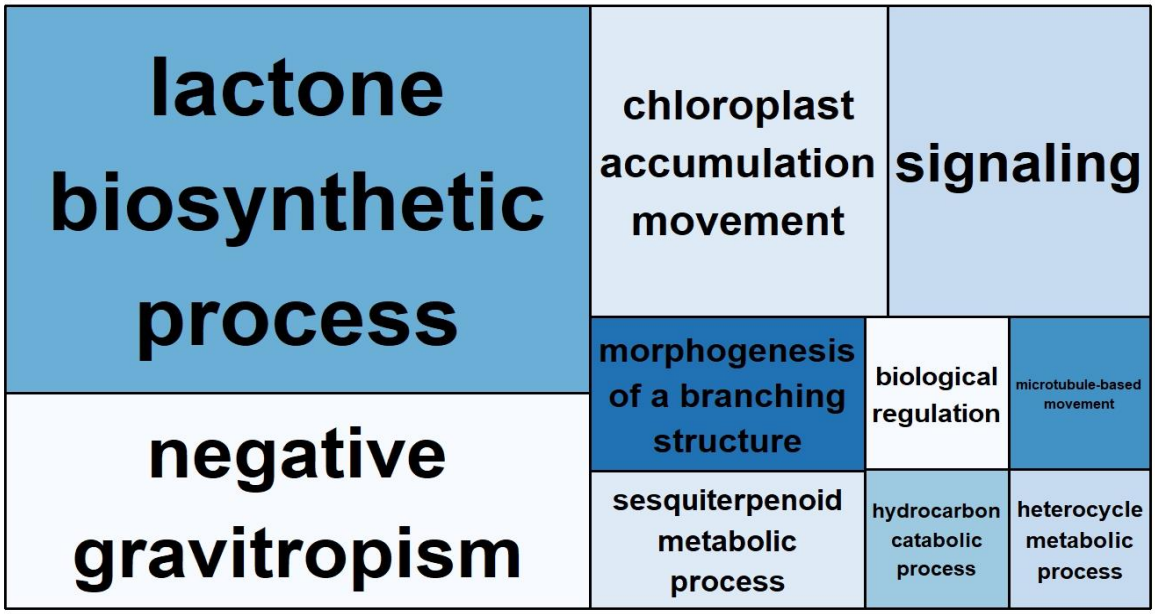
Supplementary figure 2: Ultra conserved genes secific for angiosperms



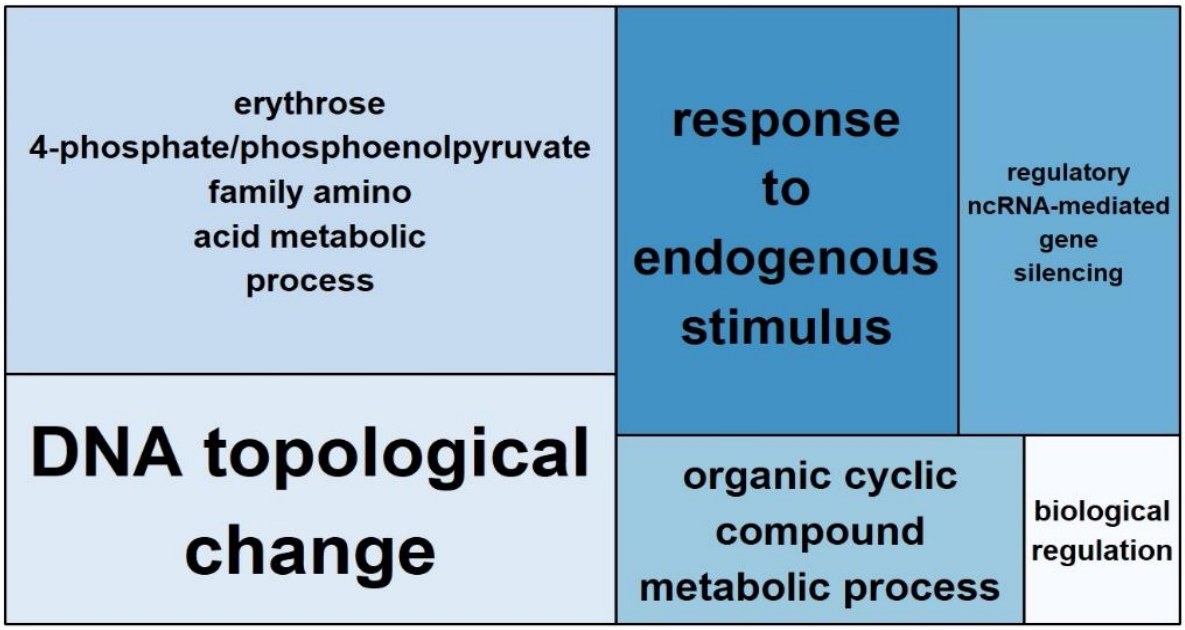
Supplementary figure 3: Ultra conserved genes secific for gymnosperms



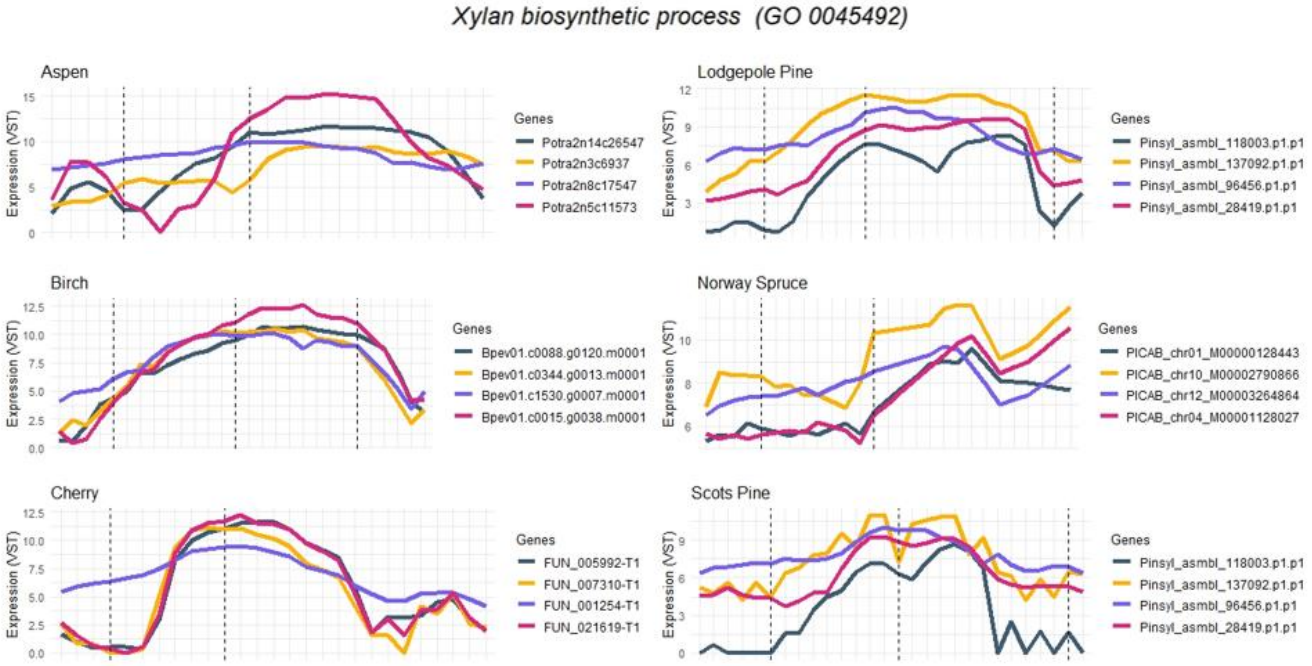
Supplementary figure 4: Gene ontology terms enriched in ultra conserved genes specific for angiosperms



Supplementary figure 5: Gene ontology terms enriched in ultra conserved genes specific for gymnosperms



Supplementary figure 6: Expression profiles of genes annotated with GO term “Xylan biosynthetic process (GO 0045492)”



Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway