

What makes a tree a tree?

Regulatory network controlling wood formation in coniferous and angiosperm forest tree species

Eduardo Rodriguez Soldado



UMEÅ UNIVERSITY

To my Zeynep, my life and light

Table of Contents

Abbreviations	i
Abstract	iii
Sammanfattning	iv
List of publications and contributions	v
Research aims and objectives	vii
Introduction	1
Wood composition	1
<i>Cellulose</i>	3
<i>Hemicellulose</i>	3
<i>Lignin</i>	4
Xylogenesis.....	5
Hardwood and Softwood.....	8
Regulation of wood formation	10
Endogenous and exogenous regulation of wood formation.....	14
Genomic and transcriptomic research	16
<i>Transcriptomic methods</i>	18
<i>Co-expression networks</i>	20
<i>Combinatorial, multi-omics methods</i>	23
Material and Methods	30
Forest Tree Species Researched in this thesis.....	30
<i>Aspen</i>	30
<i>Norway Spruce</i>	31
<i>Birch</i>	32
<i>Scots pine</i>	32
<i>Lodgepole pine</i>	33
<i>Cherry</i>	33
<i>Drimys angustifolia</i>	34
Cryosectioning.....	34
RNA and DNA extraction.....	38
DAP-seq protocol optimisation for recalcitrant woody tissues	39
Results and Discussion	42
Paper I - An updated perspective: what genes make a tree a tree?...	42

Paper II - 1000 conifer genomes: genome innovation, organisation and diversity	47
Paper III - An evo-devo resource for wood: Comparative regulomics across dicot and conifer trees	52
Paper IV - A high-throughput DNA affinity purification sequencing (DAP-seq) protocol method for recalcitrant tissues of woody species.....	58
Conclusions.....	65
Future perspectives	66
Acknowledgments.....	68
List of References.....	69

Abbreviations

ATAC-seq - Assay for Transposase-Accessible Chromatin sequencing

CESA - Cellulose synthase

cDNA - Complementary DNA

ChIP-seq - Chromatin immunoprecipitation sequencing

CRE(s) - Cis-regulatory element(s)

CSC(s) - Cellulose synthase complex(es)

CTAB - Cetyltrimethylammonium bromide

DAP-seq - DNA affinity purification sequencing

DNA - Deoxyribonucleic acid

GC-MS - Gas chromatography-mass spectrometry

GO - Gene Ontology

GT2 / GT43 - Glycosyltransferase family 2 / 43

GWAS - Genome-wide association study

HA(s) - Herbaceous annual(s)

Hi-C - High-throughput chromosome conformation capture

Iso-seq - Isoform sequencing

mRNA - Messenger RNA

NGS - Next-generation sequencing

NST - NAC Secondary wall Thickening-promoting factor

OG - Orthogroup

PCD - Programmed cell death

PCW - Primary cell wall

PTM(s) - Post-translational modification(s)

RNA - Ribonucleic acid

RNA-seq - RNA sequencing

SCW - Secondary cell wall

TE(s) - Transposable element(s)

TF(s) - Transcription factor(s)

VND - Vascular-related NAC Domain protein

WND - Wood-associated NAC domain transcription factor

WGS - Whole-genome sequencing

WP(s) - Woody perennial(s)

Abstract

What makes a tree a tree?

The capacity to form and maintain woody tissue has been key for the ecological success and economic relevance of forest trees. While fundamental cell types and developmental processes are common to most trees, there are significant differences between the two main tree lineages: angiosperms and gymnosperms.

Comparative genomic research has dramatically expanded our understanding of plant genome evolution, with several studies demonstrating that the transcriptional programmes underlying xylogenesis are largely conserved between lineages. Modern research suggests that both speciation and intraspecific variation are often the result, not only of coding sequence mutations, but also of shifts in gene expression regulation.

The aim of this thesis was to elucidate how genomic architecture and regulatory programmes govern wood development and secondary growth evolution. By combining comparative genomics with high-resolution spatial transcriptomics across angiosperm and gymnosperm species, this research establishes a multi-layered regulomic and evolutionary framework for studying wood formation.

The results identified multiple regulatory gene groups linked to wood evolution and development and generated significant genomic resources. In particular, chromosome-scale reference genomes were generated for two conifer species and an "evo-devo" resource for wood was established using a high-resolution comparative regulomic framework across wood differentiation layers in six tree species. Furthermore, a modified DNA Affinity Purification sequencing (DAP-seq) protocol was developed and optimised for mature woody tissues.

These resources can facilitate the identification of conserved and lineage-specific regulators, providing a critical blueprint for precision breeding and targeted genome engineering. Ultimately, these findings can contribute to the development of advanced materials and the transition toward a carbon-neutral bioeconomy.

Sammanfattning

Vad gör ett träd till ett träd?

Grundläggande vedceller och processer är gemensamma för de flesta träd, men det kan finnas betydande skillnader mellan de två huvudsakliga trädlinjerna: angiospermer och gymnospermer. Till exempel är asp, en angiosperm, ett lövträd, medan gran, en gymnosperm, är ett flerårigt barrträd. Barrträdet gran innehåller specialiserade hartskanaler och dess ved består av långa styva trakeidceller. Som jämförelse har lövträd två typer av starka vedfibrer och kanaler specialiserade för transport av olika ämnen.

Överraskande nog har jämförande studier visat att många av de gener som är involverade i att tillverka trä delas mellan träd och icke-träd. Det som ofta skiljer sig åt är när dessa gener aktiveras, samt hur dessa gener används och förändras av komplexa regleringsprogram som kan producera stora skillnader i vedstruktur och utveckling. Detta innebär att evolutionen ofta fungerar genom att justera genreglering snarare än att uppfinna "trädgener" uteslutande i trädarter.

I denna forskning syftade vi till att skapa en regulatorisk nätverksmodell för vedutveckling hos flera gymnosperm- och angiospermträdarter, med särskilt fokus på asp, en allmänt studerad trädmodellart, och gran, en av de dominerande gymnosperm-barrträdarterna. Vi använde den resulterande modellen för att identifiera mekanismer som skiljer båda linjerna från varandra och från andra växter. Denna avhandling ger både grundläggande insikter i utvecklingen av genreglering av vedbildning och deras bevarade mekanismer. Sammantaget kan dessa resurser hjälpa oss att hitta de regulatorer som formar vedutvecklingen och ge nya sätt att förädla eller konstruera träd för timmer, klimatbegränsning och nya träbaserade material.

List of publications and contributions

This thesis is based on the work contained in the following papers, referred to by the corresponding Roman numerals in the text:

Paper I. Birkeland, S., **Soldado, E.R.**, Ranade, S.S., M. Rosario García-Gil, Choudhary, S., Kumar, V., Tuominen, H., Mellerowicz, E.J., Street, N.R. & Hvidsten, T.R. (2025). An updated perspective: what genes make a tree a tree? *Trends in Plant Science*, Elsevier.

Paper II. Kalman, T.A., Delhomme, N., Eriksson, M.C., Hill, J., Kumar, V., Larsson, T., Mähler, N., Nandi, S., Unneberg, P., van der Valk, T., Zuccolo, A., Chen, Z.-Q., Estravis Barcala, M., **Soldado, E.R.**, Funda, T., Chaudhary, R., Birkeland, S., Olsen, R.-A., Bunikis, I., McCann, J., Tuominen, H., Canovi, C., Piombo, E., Carracedo Lorenzo, Z., van Zalen, E., Suontama, M., Hallingbäck, H.R., Mosbech, M.-B., Vassilieff, H., Schiffthaler, B., Grabherr, M., Bakker, L., Schijlen, E., Benstein, R.M., Sundström, J.F., Westrin, K.J., Emanuelsson, O., Vinnere-Pettersson, O., Hvidsten, T.R., Sherwood, E., Ingvarsson, P.K., Wu, H., Gyllensten, U., Nilsson, O., Nystedt, B. & Street, N.R. (2025). 1000 conifer genomes: genome innovation, organisation and diversity. <https://doi.org/10.21203/rs.3.rs-6502828/v1> (under revision)

Paper III. **Soldado, E.R.**, Birkeland, S., Chapple, E., Fredriksson, S., Lorenzo, Z., Kalman, T., Kumar, V., McCann, J., Hill, J., Kjendseth, Å., Tuominen, H., Mellerowicz, E. & Street, N.R. (2025). An evo-devo resource for wood: Comparative regulomics across dicot and conifer trees. <https://doi.org/10.21203/rs.3.rs-7656402/v1> (under revision)

Paper IV. **Soldado, E.R.**, Kalman, T., Kumar, V., Viljamaa, S., Niitylä, T. & Street, N.R. (2025). A high-throughput DNA affinity purification sequencing (DAP-seq) protocol method for recalcitrant tissues of woody species (manuscript)

Author's contributions to each paper:

Paper I. I generated the RNA-seq data from *Drimys angustifolia* for *de novo* assembly and contributed to the manuscript preparation and figure design.

Paper II. I assisted with sample collection, performed the sample processing from the selected tree species and generated the sample series via high-precision cryosectioning. I organised the workflow, generated the transcriptomic data of the project tree species, contributed to data interpretation and assisted in the manuscript preparation.

Paper III. I assisted with sample collection, performed the sample processing from the selected tree species, tested candidate species and generated the sample series via high-precision cryosectioning. I organised the large-scale lab workflow and generated the transcriptomic data for five of the six species investigated. I contributed to data interpretation and assisted with manuscript writing.

Paper IV. I took a leading role in the planning and generation of the DAP DNA libraries, for which I designed and conducted optimisation trials for the modified lab protocols. I generated the DAP-seq data and I produced, processed and assessed the iSeq data. I contributed to data interpretation and co-wrote the manuscript along with Teitur Kalman.

Papers I-III are reproduced with the permission of the publishers.

Additional papers not included in this thesis:

Soldado, E.R., Ranade, S., Johansson, S., Egertsdotter, U. & Street, N.R. (2025). RNA-seq dataset of initiation sites of somatic embryogenesis in Norway spruce. (manuscript).

I processed the spruce embryo samples and generated the transcriptomic data. I performed the preprocessing and biological quality assessment of the data, performed the high-throughput data analysis and identified differentially expressed genes. I designed and generated the figures and co-wrote the manuscript along with Sonali Ranade.

Research aims and objectives

The overall aim of this PhD project was to advance our understanding of the genomic and regulatory mechanisms that underpin wood formation and secondary growth evolution. To achieve this, the work combined comparative genomics, transcriptomics, comparative co-expression, genome analyses in conifers and transcription factor-DNA binding assays adapted for woody tissues.

The specific objectives of the project were:

Paper I. An updated perspective: what genes make a tree a tree?

The objective of this study was to re-evaluate the genetic basis of woodiness and perennial growth by analysing gene family evolution across a phylogenetically diverse set of woody and herbaceous angiosperms. This research aimed to identify the regulatory modules most robustly associated with tree specific traits. Furthermore, it evaluates the relative contributions of varying selection pressures on coding sequences, gene duplication events and gene losses in driving the lineage-specific innovations that underpin secondary growth and wood evolution.

Paper II. 1000 conifer genomes: genome innovation, organisation and diversity

The aim of this study was to generate chromosome-scale conifer reference genomes, alongside an extensive resequencing dataset of ~1000 Norway spruce individuals, providing a robust framework for investigating genomic innovation and conifer diversity. This research provides new insights into the drivers of genome evolution, segmental duplication and pseudogenisation, while outlining the 3D-chromatin architecture and epigenetic features of expansive conifer genomes. Furthermore, this work characterises the transcriptional divergence and sub-functionalisation of paralogous gene pairs during wood development, ultimately identifying the genomic signatures underpinning local climatic adaptation.

Paper III. An evo-devo resource for wood: Comparative regulomics across dicot and conifer trees

This research aimed to build a high-resolution comparative regulomic framework for wood formation using spatial transcriptomes across wood differentiation gradient in six tree species, three angiosperms and three gymnosperms. In this way, the study characterises both conserved and lineage-specific transcriptional programmes, facilitating the inference of core regulatory modules governing secondary growth. This work provides a foundational, open-access resource that supports comparative and evolutionary developmental research (evo-devo) on forest genomics.

Paper IV. A high-throughput DNA affinity purification sequencing (DAP-seq) protocol method for recalcitrant tissues of woody species

This method paper aimed to develop a modified DAP-seq protocol suitable for mature and lignified tissues in woody species, which were previously tested as unsuitable for this technique. This workflow was validated using dozens of transcription factors in two model tree species and consistently produced high-quality sequencing libraries and TF-DNA binding profiles with DNA extracted from both wood and leaf tissues. This modified protocol offers a robust tool for genomic analyses in woody tissues and can make DAP-seq more accessible for researchers working in forest genomics and plant secondary growth.

Introduction

Trees dominate terrestrial ecosystems, with more than half of the carbon fixed by land plants attributable to woody species. Forests alone store on the order of hundreds of billions of tonnes of carbon in woody biomass (Luo & Li, 2022) and play a central role in regulating atmospheric carbon dioxide and oxygen levels (Ye & Zhong, 2015). At the same time, wood is a renewable biological material of exceptional importance to human societies, forming the basis of construction, paper industries and the emerging bioenergy production (Groover et al., 2009; Zhong & Ye, 2013).

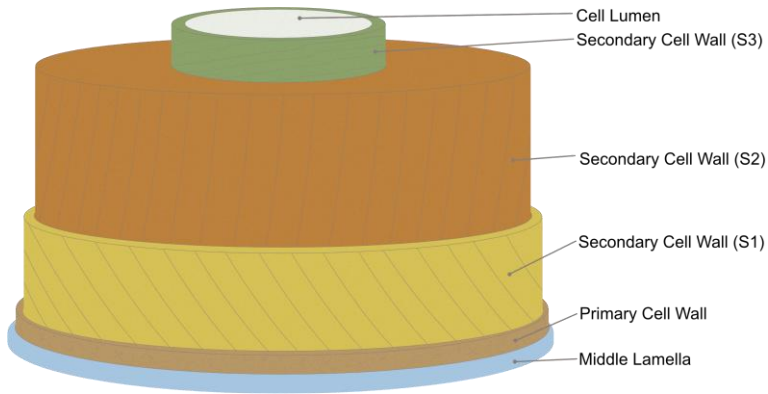
The global dominance of woody ecosystems is fundamentally rooted in the evolutionary success of vascular plants, particularly gymnosperms (softwood) and angiosperms (hardwood) (Parkinson et al., 1999). The capacity for secondary growth represents more than a mere structural adaptation, it is a fundamental physiological innovation that has defined the trajectory of land plants and the product of complex, tightly regulated molecular processes that integrate growth, cell differentiation and the biosynthesis of new cell walls.

Wood composition

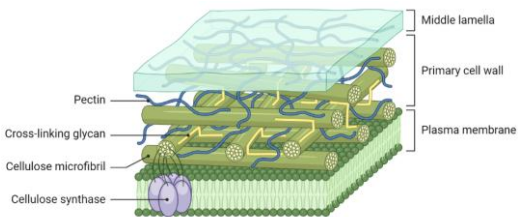
One of the most complex aspects of cell development in plant species is the precise coordination to form multiple, distinct cell wall types. Wood cell walls are organised as concentric layers deposited sequentially from the outer to the inner cell surfaces: the middle lamella, the primary cell wall (PCW) and the secondary cell wall (SCW), which itself comprises the S1, S2 and S3 layers (Fig. 1A). These layers differ in thickness, cellulose microfibril orientation and chemical composition. The outer S1 layer is relatively thin and characterised by a high microfibril angle. The S2 layer constitutes the bulk of the SCW and is enriched in cellulose with a low microfibril angle, thereby conferring most of the mechanical strength of the cell. The inner S3 layer is typically thin, with higher lignin concentration and a distinct microfibril arrangement (Fromm, 2013;

Plomion et al., 2001). During SCW formation, cellulose, hemicellulose and lignin are synthesised and deposited in the developing wall architecture (Fromm, 2013; Zhong & Ye, 2013).

A



B



C

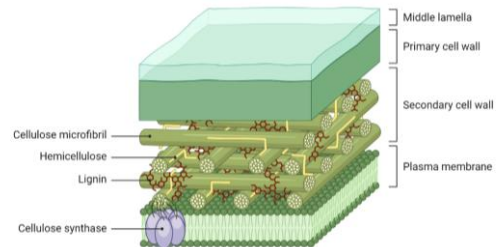


Figure 1. 3-D organisation and molecular composition of the cell wall layers in a tracheary element. (A) Concentric arrangement of cell wall layers surrounding the cell lumen: the middle lamella, PCW and the three layers of the SCW (S1, S2 and S3). (B) Molecular architecture of the PCW, with cellulose microfibrils embedded in a matrix of hemicelluloses and pectins, via cellulose synthase complexes. (C) Molecular architecture of the SCW, enriched in cellulose and hemicelluloses and extensively lignified.

Cellulose

Cellulose, an unbranched β -(1,4)-linked glucan polymer, is a major structural component of both the PCW and SCW of wood (Kollmann et al., 2013; Mutwil et al., 2008) (Fig. 1B-C). Cellulose represents roughly 40-50% of the dry wood mass and is organised into microfibrils composed of multiple glucan chains arranged around a crystalline core of approximately 3 nm in diameter. These microfibrils are responsible for the tensile strength of the cell wall (Mellerowicz & Sundberg, 2008; Tai et al., 2023).

Cellulose biosynthesis is catalysed by cellulose synthase (CESA) proteins, which assemble into cellulose synthase complexes (CSCs) that determine glucan chain polymerisation and microfibril formation. CSCs are assembled in the Golgi apparatus and delivered to the plasma membrane via vesicle trafficking, where they associate with cortical microtubules. There, they catalyse glucan chain elongation to form cellulose microfibrils extending into cell walls (Doblin et al., 2002; Luo & Li, 2022) (Fig. 1B-C). CESA proteins belong to the glycosyltransferase family 2 (GT2), are highly expressed during secondary xylem development and exhibit strong sequence conservation among higher plants, consistent with broadly shared mechanisms of cellulose biosynthesis (Sundell et al., 2017). Correspondingly, the genome of the model plant *Arabidopsis thaliana* encodes 10 CESA genes, whereas the angiosperm tree *Populus* species contain an expanded CESA gene family, comprising 18 members in some assemblies, a higher amount consistent with lineage-specific gene family diversification in trees (Hill et al., 2014; Nixon et al., 2016; Suzuki et al., 2006).

Hemicellulose

Hemicelluloses comprise diverse polysaccharides with β -(1,4)-linked backbones and variable side-chain substitutions, such as xylans, glucomannans, mannans and xyloglucans. The composition and abundance of hemicelluloses vary substantially between species and

xylem cell types, reflecting lineage- and tissue-specific wall architectures. Functionally, hemicelluloses contribute to cell wall strength by linking with both cellulose microfibrils and lignin (Li et al., 2024; Scheller & Ulvskov, 2010) (Fig. 1C).

In gymnosperms, glucomannan and xylan associate with cellulose microfibril surfaces and enhance compressive stiffness, whereas in angiosperms, xylan predominates and contributes to increased tensile extensibility through its interaction with cellulose (Berglund et al., 2020; Salmén, 2022). Hemicellulose biosynthesis is mediated by coordinated activity of multiple enzymes. Their backbone and side-chain formation is catalysed primarily by glycosyltransferases, which are highly expressed in developing xylem, such as core components of the GT43 family. Subsequent polymer modification involves methyltransferases and acetyltransferases, with acetylation occurring in the Golgi apparatus using acetyl-coenzyme A derived from cytosolic pools of acetyl-CoA, which are generated by central metabolic pathways and transported to the Golgi lumen (Lee et al., 2011; Pawar et al., 2017).

Lignin

Lignin, a complex cross-linked phenolic polymer, is the second most abundant component of wood after cellulose. It confers rigidity and hydrophobicity to the SCW and typically comprises 20-40% of dry wood mass (Blokhina et al., 2019; Koch & Schmitt, 2013; Vanholme et al., 2019). Lignin consists of three monolignol-derived subunits: p-hydroxyphenyl (H), guaiacyl (G) and syringyl (S), which are produced from the oxidative polymerisation of p-coumaryl, coniferyl and sinapyl alcohols, respectively (Li et al., 2006; Mellerowicz & Sundberg, 2008). Monolignols are synthesised via the phenylpropanoid pathway and subsequently transported across the plasma membrane into the cell wall, where polymerisation is mediated primarily by laccases and peroxidases, generating a highly heterogeneous polymer with diverse inter-unit linkages (Lin et al., 2016; Vanholme et al., 2019; Weng & Chapple, 2010) (Fig. 1C). During the lignification stage, lignin is deposited in all the cell

wall layers, conferring structural support, a process that can continue even after cell death (Kollmann et al., 2013; Pesquet et al., 2013).

From an applied perspective, lignin contributes substantially to the recalcitrance of woody biomass during bioconversion processes in industries, such as enzymatic saccharification and microbial fermentation. This has motivated extensive efforts to modify lignin content and composition in trees (Chen & Dixon, 2007; Li et al., 2010). While constitutive alteration of lignin biosynthesis frequently leads to growth and developmental defects, recent studies demonstrate that cell-type specific manipulation can alter lignin properties without compromising overall tree growth. These findings indicate that lignin biosynthesis can be specifically modified and highlight targeted engineering as a promising strategy for improving wood biomass utilisation (Bonawitz & Chapple, 2010; Luo & Li, 2022; Yang et al., 2013).

Together, the complex spatial organisation of secondary cell wall layers and the tightly coordinated deposition of cellulose, hemicelluloses and lignin define the mechanical strength and hydraulic functionality of wood tissues. At the same time, these processes are integrated within the development and differentiation of wood layers.

Xylogenesis

A tree trunk is usually covered by a protective layer of bark, while the wood beneath is the result of radial secondary growth from the cambium layer or cambial zone (Kollmann et al., 2013). Wood formation, or xylogenesis, proceeds through a series of tightly coordinated developmental transitions, beginning with cell proliferation in the vascular cambium and progressing through cell expansion, differentiation, SCW deposition and ultimately programmed cell death (PCD) (Plomion et al., 2001; Winkler & Oberhuber, 2017; Zhong & Ye, 2013) (Fig. 2). These successive stages are regulated by complex and

interconnected molecular networks that coordinate secondary growth in time and space (Luo & Li, 2022).

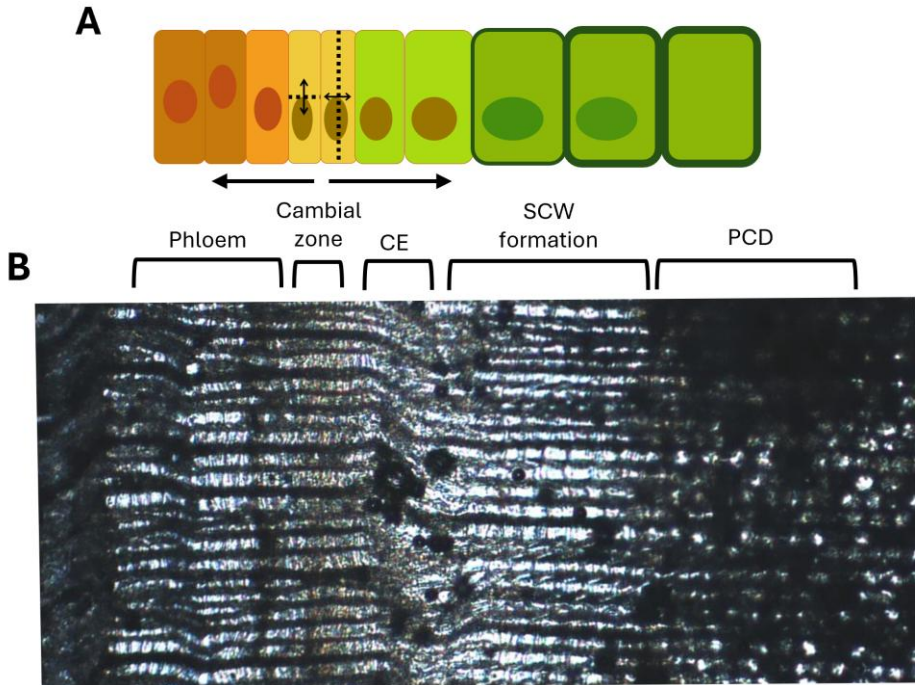


Figure 2. Cambial cell divisions and development of wood formation. (A) Schematic representation of the vascular cambium showing anticlinal divisions that expand the cambial circumference and periclinal divisions that generate xylem inwards and phloem mother cells outwards. (B) Micrograph from *Picea abies* showcasing the spatial organisation of phloem, the cambial zone and xylem, highlighting successive developmental stages of wood formation, such as cell division, cell expansion (CE), secondary cell wall (SCW) formation and programmed cell death (PCD).

While tree branches, stems and roots extend from their apical meristems, the tree grows thicker from its cambium layer (Plomion et al., 2001; Evert, 2006). The vascular cambium is a thin cylindrical tissue layer, typically 6-8 cells thick, encircling the stem and composed of

narrow cambial initial cells. The cambium drives secondary growth through periclinal divisions, producing daughter cells that differentiate outwards into secondary phloem mother cells and inwards into secondary xylem mother cells (Taiz et al., 2015). The latter subsequently differentiate into specialised wood tissues (Fig. 2A), whose specific cell types and proportion can vary between softwood and hardwood (Kollmann, et al., 2013; Plomion et al., 2001). The role of the phloem cells is the transport of nutrients and signalling molecules from photosynthetic tissues to growing, developing or storage tissues, while the xylem transports water and dissolved nutrients upwards from the roots, in addition to constituting the main structural support (De Schepper et al., 2013; Knoblauch & Oparka, 2012; Plomion et al., 2001).

Following cell division, xylem precursor cells undergo cell expansion, which determines the final size and geometry of wood cells (Fig. 2). Cell expansion is associated primarily with PCW formation and can be facilitated by cell wall-loosening proteins such as expansins, which enable controlled extension of the primary wall matrix (Gray-Mitsumune et al., 2008; Luo & Li, 2022). This wall structure consists predominantly of cellulose, hemicelluloses, pectins and structural glycoproteins, organised as a cellulose-hemicellulose network embedded within a pectin-rich matrix, with pectin playing a central role in wall plasticity during expansion (Carpita & Gibeaut, 1993).

After cell expansion is completed, SCW deposition is initiated, predominantly in xylem tissues (Fig. 2A). SCW, enriched in cellulose, hemicelluloses and lignin, is assembled through highly regulated biosynthetic and transport processes controlled by extensive transcriptional networks. The completion of secondary wall deposition triggers a final phase of intensive lignification, which provides a rigid and hydrophobic structure. In specialised xylem cell types, such as tracheids and vessel elements, this maturation process concludes with PCD, ultimately establishing the functional architecture of the mature secondary xylem (Luo & Li, 2022; Ye & Zhong, 2015) (Fig. 2B). PCD marks the final stage of secondary xylem differentiation. During PCD, proteases, nucleases and autophagy-related components are upregulated, progressively digesting the cellular contents (Luo & Li, 2022). PCD proceeds differently depending on the cell type, with vessel elements dying a few days after cambial differentiation, while fibres and

tracheids remain alive for several weeks (Bollhöner et al., 2012; Nakaba et al., 2011). Parenchyma cells, which are responsible for nutrient storage, can have a relatively long life (Kollmann et al., 2013). At later stages, tylosis formation and wood dehydration lead to the formation of heartwood at the core of the tree, which no longer contains reserve substances or living cells (Kampe & Magel, 2013). The sequential layer developments, which integrate cell proliferation, differentiation and cell death into a continuous developmental trajectory, can generate cellular structures that differ markedly between angiosperms and gymnosperms, that are described in the following section.

Hardwood and Softwood

The cellular architecture of secondary xylem shows deep evolutionary divergence between the two major tree lineages, reflecting distinct strategies for water transport and mechanical reinforcement (Plomion et al., 2001; Evert, 2006). In gymnosperms, the wood is remarkably uniform, with tracheids accounting for up to 90% of the total wood volume (Zhu & Li, 2024) (Fig. 3A-C). These slender, elongated cells fulfil multiple roles, providing both the structural framework for upright growth and the pathway for water and nutrient conduction via inter-tracheid bordered pits (Luo & Li, 2022). Radial transport is mediated by ray parenchyma (Fig. 3B-C) (Larson, 1994; Olano et al., 2013). In contrast, angiosperm wood is characterized by a high degree of functional specialisation. Wide-diameter vessel elements develop as mediators of long-distance water transport. These cells develop thickened SCWs and interconnect themselves through specialised perforation plates (Li et al., 2024)(Fig. 3D-F). Mechanical support is largely separated from transport and provided by wood fibres, which are defined by limited radial expansion and the highly lignified SCWs, contributing to increased mechanical strength at the tissue level. Angiosperms also possess a dual system of parenchymatous cells: axial parenchyma for nutrient storage and ray parenchyma for radial substance transport and storage (Evert, 2006; Larson, 1994; Plomion et al., 2001).

In addition, many conifer trees, such as *Picea* and *Pinus* species, possess specialised tubular resin canals distributed longitudinally and surrounded with resin-secreting epithelial cells (Kollmann et al., 2013) (Fig. 3A). These structures are promoted by the defence phytohormone jasmonate and provide protection against insects and pathogens, a feature largely absent in angiosperms (Martin et al., 2002; Kollmann et al., 2013). Moreover, conifer bark includes multiple specialised cell types, which move nutrients to the phloem (phloem sieve cells) and protect the cambial meristem and inner tissues from abiotic and biotic threats, such as desiccation or insect herbivores (Celedon et al., 2017).

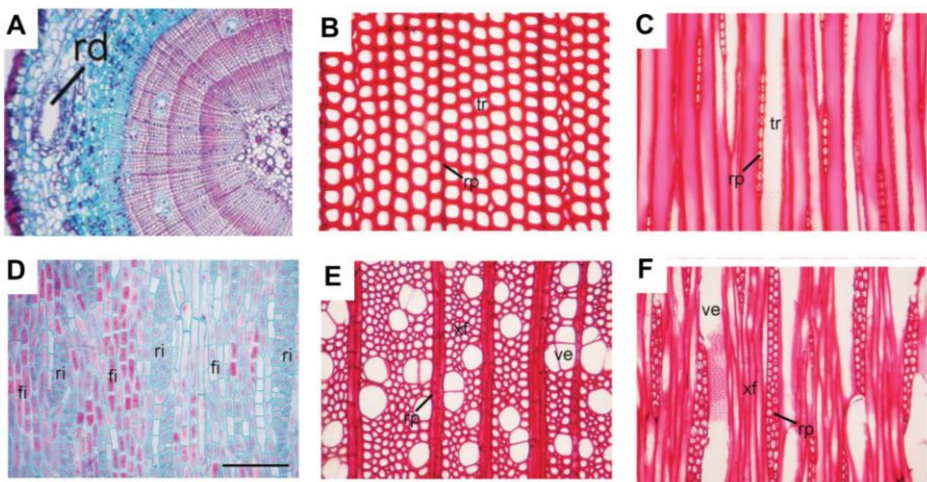


Figure 3. Comparison of gymnosperm and angiosperm wood anatomy. (A) Transverse section of a white pine gymnosperm (*Pinus strobus*) stem showing annual growth rings and resin ducts (rd). (B, C) Transverse (B) and tangential longitudinal (C) sections of *P. strobus* wood highlighting tracheids (tr) and ray parenchyma (rp). (D) Tangential longitudinal section through the cambial region of tulip tree angiosperm (*Liriodendron tulipifera*) illustrating fusiform initials (fi) and ray initials (ri). (E, F) Transverse (E) and tangential longitudinal (F) sections of *L. tulipifera* wood showing vessels (ve), xylary fibres (xf) and ray parenchyma (rp). Adapted from Ye and Zhong (2015), Journal of Experimental Botany, by permission from Oxford University Press.

Finally, molecular composition and structure can also vary significantly among the two lineages. For example, gymnosperm wood cell walls contain lignin composed almost exclusively of G-type subunits, which interact with both xylan and mannan (Kirui et al., 2022). On the other hand, angiosperm mature wood comprises a mixture of G and S subunits, with S-lignin typically predominating in the total wood volume (Musha & Goring, 1975). While G-lignin is concentrated within their highly lignified vessel elements and cell corners to support hydraulic conductivity, the more abundant fibre secondary walls are enriched in S-lignin, which interacts extensively with xylan and cellulose microfibrils (Fergus & Goring, 1970; Li et al., 2024). Ultimately, these divergent cellular and biochemical profiles are reflected in the macroscopic level. The architectural heterogeneity of angiosperm wood, specially the highly lignified fibres coupled with the synergistic interaction between S-lignin and cellulose microfibrils, results in the increased hardness and compression strength characteristic of the hardwoods (Evert, 2006; Kollmann et al., 2013). Cellular and chemical divergence between angiosperms and gymnosperms reflects not only their evolutionary history, but also distinct regulatory programmes.

Regulation of wood formation

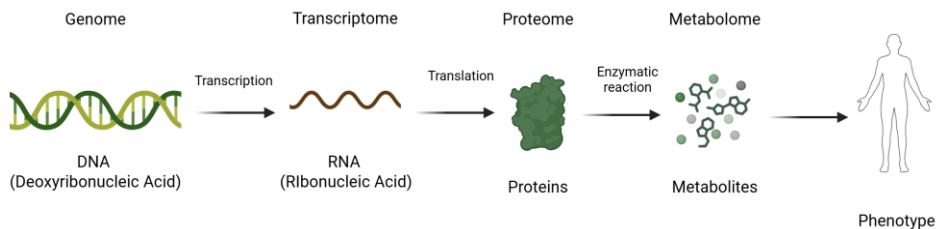


Figure 4. Summary of the central dogma of molecular biology illustrating how information encoded in the genome is transcribed into RNA and translated into proteins, which serve functions such as enzymatic reactions to synthesise metabolites, together shaping the organism's phenotype.

The study of genetics has come a long way since Watson and Crick announced their discovery of the double helix in 1953 (Watson & Crick, 1953). The central dogma of molecular biology states that, during gene expression, the genetic instruction stored in a gene encoded in DNA is transcribed into messenger RNA molecules (mRNA) (Crick, 1970). These mRNA transcript molecules are subsequently translated into proteins that carry out the structural and enzymatic functions characteristic of the organism, thereby contributing to its phenotype (Fig. 4). Accordingly, changes in DNA sequence can alter protein structure or abundance, influence biological processes and ultimately affect organismal phenotype.

Upstream of the transcription start site of a gene, there are often non-coding regions containing short, sequence-specific motifs regulating, initiating or suppressing the expression of the gene, called cis-regulatory elements (CREs). The regulation of gene transcription is mainly carried out by proteins called transcription factors (TFs), which bind the CREs, acting as TF binding sites (Badis et al., 2019). Multiple TFs can bind upstream of a single gene and, at the same time, TF binding can be affected by other proteins and DNA modifications, thus controlling when and in which cells a gene is expressed (Wang et al., 2021; Weber et al., 2016). Sometimes several TFs must form a complex to initiate the transcription of a gene or may bind to distant regions of the DNA sequence brought closer by the DNA 3D topology, enhancing or suppressing gene expression (Spitz & Furlong, 2012). Changing patterns in the binding of TFs can, thereby, affect the patterns of gene expression and, subsequently, cellular development and phenotypic variation (Liang et al., 2021) (Fig. 4). Gene expression can even be altered experimentally using mutants or transgenic lines to research the effects of gene regulation or produce new plant phenotypes for practical applications, such as crop breeding (Liu et al., 2021).

The formation of wood tissues that I described is governed by the coordinated regulation of thousands of genomic components, whose interactions control the differentiation and biochemical specialisation of distinct wood cell types (Sundell et al., 2017). Work over the past decades, primarily in *A. thaliana* (Zhong et al., 2010; Taylor-Teeples et al., 2015) and *Populus* (Lin et al., 2013), has established that SCW formation and wood development are governed by a broadly conserved

hierarchical gene regulatory network. This network integrates multiple tiers of TFs coordinating transitions from fusiform and stem-like cambial cells to dividing, expanding and differentiating xylem cells while simultaneously activating cell wall biosynthetic programmes. Although these genes generally perform in woody species functions similar to their homologues characterised in herbaceous model species, many exhibit tree-specific features (Wei & Wei, 2024; Li et al., 2024).

In *A. thaliana*, NAC domain TFs function as top-tier key regulators that activate downstream TFs and structural genes involved in cell wall biosynthesis (Mitsuda et al., 2005; Tan et al., 2018; Zhou et al., 2014). This group includes Vascular-related Domain proteins (VND1-VND7) and NAC Secondary wall Thickening promoting factors (NST1-NST3) (Wei & Wei, 2024). In trees, homologous regulators are collectively referred to as wood-associated NAC domain TFs (WNDs). These are more numerous in *Populus* than in *A. thaliana*, hinting at a more complex regulatory network in tree species. (Luo & Li, 2022; Wei et al., 2020). The first-tier regulators activate a shared set of downstream targets encompassing second-tier TFs, such as MYB46, MYB83 and MYB103, as well as genes involved in cellulose, hemicellulose and lignin biosynthesis, cytoskeleton organisation, vesicle trafficking and PCD. Together, NAC and MYB TFs form interconnected regulatory networks that integrate developmental cues and environmental signals to direct SCW formation (Zhong et al., 2008; Zhong et al., 2011). Third-tier TFs, activated by combinations of first- and second-tier TFs, selectively modulate fourth-tier structural genes, which encode enzymes for the biosynthesis of cellulose, xylan and monolignol biosynthesis (Wei & Wei, 2024).

Top-tier NAC and MYB TFs are also known to regulate some unique target genes depending on cell type, such as protein kinase genes (Luo & Li, 2022; Wei & Wei, 2024). In *A. thaliana*, NST1 and SND1 have been identified as key regulators of fibre differentiation, whereas VND6 and VND7 govern xylem vessel element specification. In *Populus*, multiple orthologues of NST1/SND1 have been identified and shown to coordinately regulate SCW in both wood and phloem fibres. Functional diversification within this family is further illustrated by WND1B, a *Populus* SND1 homologue that undergoes alternative splicing and modulates fibre cell wall thickening (Takata et al., 2019).

Some orthologous genes within the networks display divergent expression patterns and functional roles between *Populus* and *A.*

thaliana. For example, PtrWND (*Populus trichocarpa* WND) TFs accumulate broadly across vessel elements, fibres and ray parenchyma cells in the developing xylem of *Populus*, indicating lineage-specific deployment of conserved regulators (Wei & Wei, 2024). Comparative phylogenetic analyses across tree species indicate that secondary wall NAC master regulators have undergone lineage-specific expansion in angiosperms (six in *Eucalyptus* and six duplicated pairs in *Populus*), whereas gymnosperms retain fewer representatives (two in *Pinus* and *Picea*). This expansion correlates with the greater cellular and structural complexity of angiosperm wood, which comprises both vessels and fibres, in contrast to the tracheid-based xylem of gymnosperms (Li et al., 2024; Ye & Zhong, 2015).

The second-tier regulatory TFs activated by the WNDs are able to function redundantly as second-tier regulatory switches, activating downstream TFs as well as structural genes for biosynthesis. MYB46 directly activates most biosynthesis pathways for cellulose, hemicellulose and monolignols (Ko et al., 2014). In addition, like first-tier TFs, MYB46/83 also regulate other genes associated with secondary growth processes, such as PCD, cytoskeleton organisation, vesicle trafficking, signalling and monolignol transport and oxidative polymerisation (Taylor-Teeples et al., 2015; Yamaguchi et al., 2010). This regulatory architecture is largely conserved in trees but exhibits greater diversification compared to herbaceous plants. In *Populus*, second-tier MYB regulators, such as PtrMYB2, PtrMYB3, PtrMYB20, PtrMYB21 and PtrMYB74; are directly activated by PtrWNDs and fulfil roles as master-switches analogous to MYB46/83 in *A. thaliana*. However, unlike the largely redundant MYB46/83 pair, *Populus* MYBs exhibit substantial functional divergence, showing tissue-specific expression patterns and different binding affinities to their target promoters (McCarthy et al., 2010; Zhong et al., 2013). For example, PtrSND1-B1 (homolog of SND1) directly activates PtrMYB021 (homolog MYB46) and PtrMYB074 (a woody dicot-specific MYB), which control the expression of a set of wood cell wall genes (Chen et al., 2019; Lin et al., 2013; Wang et al., 2020).

In addition to transcriptional activation, SCW biosynthesis is further modulated by repression mechanisms (Luo & Li, 2022). At the network level, SCW regulatory systems incorporate extensive feed-forward and feedback loops which interconnect TF tiers, such as negative regulation and repressive mechanisms. For instance, MYB4, MYB7 and MYB32, which are directly induced by MYB46/83/58/63, can function as negative regulators by repressing SWN activity and, in some cases, their

own expression. In addition, regulatory interactions also occur among TFs of the third-tier, further adding complexity to the network. These intertwined mechanisms allow SCW formation to remain adaptive through different developmental stages and environmental conditions, which can also influence wood development (Wei & Wei, 2024).

Endogenous and exogenous regulation of wood formation

The regulatory network is dynamically modulated by both physiological programmes and environmental signals, such as drought or biotic stress. The first-tier TFs have been shown to respond to external cues and conditions (Wei & Wei, 2024). For instance, ultraviolet light can increase the expression of E2Fc, which directly regulates VND6/7 in *A. thaliana*. TCP4, which interacts with phytohormones, can trigger SCW biosynthesis and PCD in vessel cells (Mitsuda and Ohme-Takagi, 2008; Taylor-Teeples et al., 2015). This regulatory architecture enables adaptive changes of secondary growth processes and highlights the importance of spatiotemporal regulation and tissue- or cell- specific control mechanisms (Wei & Wei, 2024).

Although transcriptional regulation is the primary determinant of expression levels, gene expression profiles are also modulated by epigenetic modifications, such as DNA methylation. Methylation on promoter regions generally linked to transcriptional repression, whereas gene body methylation is often associated with stable, constitutive expression (Jaenisch & Bird, 2003; Zhang et al., 2018). In *Populus*, dynamic changes in DNA methylation accompany the transition from primary to secondary growth and are associated with altered expression of monolignol biosynthesis genes, such as PtrPAL2 and PtrC4H1 (Zhang et al., 2020). More broadly, components of the SCW regulatory network, such as WND, MYB, CESA and PAL genes, display differential methylation patterns during secondary wall formation in juvenile versus mature wood (Luo et al., 2021). These observations indicate that epigenetic mechanisms contribute to fine-tuning transcriptional networks underlying SCW biosynthesis and xylem differentiation in trees (Li et al., 2024; Wei & Wei, 2024).

Cambial activity and xylem/phloem cell differentiation are also regulated by interacting hormone gradients. It has been known since the mid-20th century that auxin has a role in promoting cambium activity (Gouwentak, 1941). Auxin concentration measurements in both angiosperms and gymnosperms have shown that the auxin gradient has a bell-shaped distribution which peaks in the cambium initials and tapers outwards, in a steep decrease toward the phloem and gradually toward the xylem (Tuominen et al., 1997; Uggla et al., 1998). Disruption of auxin signalling, through overexpression of a mutant *PttIAA3*, reduced cambial cell division in *Populus*, highlighting the central role of auxin in regulating cambial activity and vascular differentiation (Nilsson et al., 2008; Ye & Zhong, 2015).

Gibberellins are involved in cambium cell division and show a concentration gradient across the differentiating layers, peaking in the developing xylem (Israelsson et al., 2005). In addition, they seem to act in synergy with auxin (Zhu & Li, 2024). Exogenous application of gibberellin to decapitated seedlings in the absence of auxin stimulated cambial division which produced cells of abnormal shape. When auxin was applied along with gibberellin, it increased cambial division and cells differentiated normally (Taiz et al., 2015).

Cytokinins also act as an important regulator of cell proliferation in the cambium (Taiz et al., 2015). They are distributed across the vascular cambium region, with peak accumulation on the developing phloem side. However, most cytokinin receptors are expressed in the dividing cambial cells (Nieminen et al., 2008). Overexpressing a cytokinin catabolic gene reduces the cytokinin signalling and inhibits cambial activity and radial growth (Nieminen et al., 2008). The spatial domains of cytokinin and auxin signalling partially overlap within the cambial zone and increased cytokinin biosynthesis elevates local auxin levels, suggesting that coordinated hormonal interactions take place during vascular development (Luo & Li, 2022).

Ethylene seems to have a strong regulatory role in cell expansion and SCW deposition during xylem development (Hager, 2003; Sundell et al., 2017). Exogenous application of ethylene stimulates vascular cambium activity. It also intervenes in the formation of tension wood, which is developed in response to stem bending or tilting (Taiz et al., 2015). Small

peptides, such as CLAVATA3 (CLV3)/EMBRYO SURROUNDING REGIONRELATED (CLE), also play a role directing cambial activity, with some increasing cambial activity while others inhibit it (Luo & Li, 2022).

External factors can also modulate hormone action and thereby affect gene expression and wood anatomy (Fischer et al., 2019). For example, the defence-related phytohormone jasmonate induces traumatic resin duct formation and defence-related terpene biosynthesis in conifers, showing how defence signalling can reshape wood structure (Martin et al., 2002; Arbellay et al., 2014). To maximize efficiency in harsh climates, such as boreal and northern regions, metabolic cambial activity and tree dormancy is coordinated with the growing seasons (Bhalerao et al., 2016; Seo et al., 2013). The responsiveness to hormonal signals changes with season and developmental stage, with cambial sensitivity and responsiveness typically higher during the active growing season (Fromm, 2013; Fischer et al., 2019). This seasonal nature is illustrated by the annual rings of earlywood, which grows radially expanded cells, wide tracheids and vessels and thin cell walls at the start of the growing period; while latewood-derived xylem grows thicker cell walls and smaller tracheid diameters (Fromm, 2013; Jokipii-Lukkari et al., 2018). The pattern of the annual rings has stimulated studies of environmental variation through the years, with thicker rings indicative of more optimal growing conditions (Groover et al., 2009). While these endogenous and exogenous factors provide the physiological context for wood formation, a comprehensive understanding of the underlying molecular networks that regulate wood formation requires the application of high-throughput genomic and transcriptomic methodologies, as detailed in the following sections.

Genomic and transcriptomic research

Through hundreds of millions of years, species have adapted their biology to local environments, with natural selection preserving beneficial changes in gene expression or protein function that alter

phenotype and increase fitness to the environment (Carroll, 2008). Species occupying similar environments may converge on comparable phenotypes, whereas others diverge through distinct adaptive trajectories. Two species evolving from a common ancestor will share genes with a common origin, called homologs. When their function is evolutionarily conserved in both species, they are called orthologs. Sometimes, genes can undergo duplication in a species, resulting in paralogs (Koonin, 2005). The more phylogenetically separated two species are, the more challenging direct genomic comparisons become, owing not only to accumulated sequence divergence but also to lineage-specific gene duplications, losses and large-scale structural variation (Altenhoff et al., 2012). As a result, many comparative approaches focus on analysing the expression of orthologous genes under comparable baseline or perturbed conditions (Jokipii-Lukkari et al., 2017), for example, during abiotic stress or pathogen attacks. Such ortholog-based comparisons provide a common reference framework for assessing whether regulatory responses and expression programmes are conserved or have diverged over evolutionary time across lineages (He & Zhang, 2005).

However, such evolutionary change does not depend solely on structural changes in coding DNA or protein sequences. Genes are not isolated elements; they exist within a dynamic molecular environment and function as components of extensive metabolic and physical interaction networks. Some traits, such as SCW, can depend on complex pathways, involving coordinated interactions among regulatory DNA elements, proteins and metabolites (Sundell et al., 2017). DNA does not exist freely within the nucleus. It is wrapped around histone proteins to form nucleosomes, often described as bead-like or knot-like structures. The DNA-nucleosome structure can fold repeatedly and be further compacted into dense formations called chromatin. The DNA inside tightly packed chromatin is often not easily accessible for the protein complexes of the transcriptional machinery and genes within such regions, therefore, often show less activity. However, chromatin is a dynamic structure that can be changed over time and that can be altered in specific regions by other DNA-interacting proteins or by molecular modifications such as methylation, opening or closing genes for transcription (Hashimshony et al., 2003). Moreover, the 3D folding structure of the chromatin can bring into close proximity regions

separated by long distances in the linear sequence (Stuart et al., 2016). Another example of the dynamic and mutable nature of plant genomes is the transposable elements (TEs). These are mobile DNA sequences that can move or proliferate within the genome through distinct molecular mechanisms. TEs account for about 21% of the reference genome in *A. thaliana*, but this model organism is at the lower end of the TE content spectrum and most plant species have much higher numbers: 40% in rice, 60% in tomato, 80% in wheat and up to 85% in maize (Jouffroy et al., 2016; Oliver, et al., 2013; Quesneville, 2020; Zhang et al., 2021).

Transcriptomic methods

DNA sequencing methods allow the determination of the nucleotide sequence in the genome, in other words, to read the genetic code. The first widely adopted technique was developed in 1977 by Sanger and his team (Sanger et al., 1977). These technologies cannot determine the sequence of long DNA molecules such as whole chromosomes or whole bacterial genomes. Instead, they sequence DNA fragments, whose reads must be later assembled *in silico*, recreating the complete DNA sequence from individual and overlapping reads. Currently, Illumina is the most popular next-generation sequencing (NGS) technology, which has a low error rate (<1%) and high throughput, producing billions of reads per run. On the other hand, it produces relatively short DNA reads (around 150 bp), which makes assembly more difficult. The first tree genome (*Populus trichocarpa*) was published in 2006 (Tuskan et al., 2006). In the last years, new or improved chromosome and genome assemblies have been published for a variety of tree species, such as *Populus deltoides* (Bai et al., 2021), *Salix suchowensis* (Wei et al., 2020) and *Taxus baccata* (Olsson et al., 2018). By 2024, genome assemblies had been produced for 1,482 plant species, reflecting the rapid expansion of plant genomics resources (Bernal-Gallardo & de Folter, 2024).

Nevertheless, despite all somatic cells in a single organism sharing the same genetic code, different cells can perform specialised functions and develop in very distinct ways thanks to the selective expression of their genes. Knowing the genomic sequence in a cell is insufficient to

understand its biological programmes, since gene expression varies over time and between tissues. Such is the case for the mechanisms of tree secondary growth. Transcriptomics focuses on studying the many types of RNA transcripts in a cell, providing a snapshot measure of expression levels of genes at that given moment, by sequencing the protein-coding (mRNA) and non-coding RNA molecules. This can be used to profile temporal and spatial expression patterns by taking repeated measurements. Differential analysis can be performed on the results of distinct sample sets to analyse the difference in gene expression levels and identify genes that are activated or repressed in different conditions and tissues, which provides relevant insight in the regulatory mechanisms that differentiate cell types and functions (Celedon et al., 2017; Upton et al., 2023). These types of studies require careful experimental design to map the causal relationships between molecular regulations and phenotypical traits.

Early transcriptome profiling in wood showed that thousands of genes, such as biosynthetic enzymes and TFs, are tightly regulated during xylem differentiation. At the beginning of the 21st century, Hertzberg et al. (2001) used a *Populus* xylem cDNA microarray to profile developing wood, revealing that genes encoding cellulose and lignin biosynthetic enzymes, as well as numerous TFs, exhibit stage-specific expression during xylogenesis (Hertzberg *et al.*, 2001). In order to perform more precise analysis, tangential cryosectioning of the cambial region allows to sample successive developmental stages from phloem, through the cambial zone and into the differentiating xylem (Uggla et al., 1996). Early applications combined these cryosections with low-coverage cDNA microarrays to map concentration and expression gradients across the wood-forming zone (Sundell et al., 2017; Tuominen et al., 1997). With the advent of new techniques, such as RNA-seq and better sectioning workflows, cryosectioning has been improved to produce high-spatial-resolution transcriptomes, with dozens of consecutive samples per tree, that can elucidate spatially transcriptional modules across the full differentiation gradient (Sundell et al., 2017).

Currently, the primary technique for transcriptomic study is RNA-sequencing (RNA-seq). This technique consists of sequencing the mRNA molecules present in a sample and estimating their relative abundance (Dobrowolska et al., 2017). These mRNAs are reverse transcribed into

complementary DNA (cDNA), which is then sequenced (Olsson et al., 2018). Genes with higher expression generate more cDNA molecules and consequently more sequencing reads (Taiz et al., 2015). The resultant sequencing reads are then mapped against a reference genome or can be *de novo* assembled into a representation of the transcriptome. RNA-seq profiles can be compared between samples to identify differences in gene expression in different tissues, developmental stages, or experimental conditions (Soneson and Delorenzi, 2013; Sundell et al., 2017). Differentially expressed genes are commonly interpreted using Gene Ontology (GO) annotations, which classify genes into standardised functional categories describing biological processes, molecular functions and cellular components, thereby facilitating the functional interpretation of expression changes (Lu et al., 2019). As more tree genomes are sequenced and better assemblies are developed, further studies will be able to more comprehensively find orthologs of known SCW genes and discover new regulators specific to trees and non-model species. For example, Kim et al. (2021) combined short read Illumina RNA-seq and long read PacBio Iso-seq to study the developing xylem of Korean red pine (*Pinus densiflora*). They identified key genes for the biosynthesis of cellulose, xylan and lignin and discovered dozens of NAC TFs in *Pinus*. Seven PdeNAC genes were found to be highly expressed in wood-forming tissue and four, related to VND/NST, were shown to activate a SCW cellulose synthase promoter and induce ectopic wood thickening in leaf assays (Kim et al., 2021).

Co-expression networks

The large amounts of data publicly available produced by transcriptomic methods, specially RNA-seq, have enabled the inference of gene co-expression networks in plants, providing a powerful framework for identifying conserved functions and uncovering novel regulatory associations (Liesecke et al., 2019; Xu et al., 2023). Gene co-expression networks are graph representations of the transcriptional behaviour among genes and can be used to study the gene-gene similarity across many samples. These networks are composed of nodes, representing genes; and edges, representing gene pairs which, within the graph, are

connected by an edge if their co-expression similarity is above a selected threshold score. Degree centrality (the number of nodes a gene is directly connected to) can be used to highlight highly connected nodes, which are, therefore, potentially functionally important regulators, as highly connected genes often correspond to evolutionarily ancient or essential components of cellular systems (Provero, 2002; Wuchty & Stadler, 2003). At the same time, genes with similar expression patterns can form clusters, or modules, in the network corresponding to coherent, specific biological functions and can help pinpoint master regulators in such processes (Serin et al., 2016; Rao & Dixon, 2019). Co-expression networks are widely used to infer gene function under the guilt-by-association principle, whereby genes with correlated expression profiles are expected to participate in related biological processes, operate within similar cellular contexts or be activated at comparable developmental stages (Wolfe et al., 2005; Ryngajllo et al., 2011).

The earliest unsupervised approaches for constructing networks from gene-expression data were introduced by Butte & Kohane (1999). Since then, many different methods for constructing expression networks have been developed, with different criteria to establish edges and network parameters depending on the research needs. Some of the most common methods to calculate the node similarities are Pearson correlation (D'haeseleer et al., 2000) or linear modelling (Vasilevski et al., 2012). Several pioneering studies have also evaluated how best to construct such networks. For example, Liesecke et al. (2019) conducted a systematic assessment in plants, such as *A. thaliana*, *Solanum lycopersicum* and *Zea mays*, demonstrating that network quality tends to improve with increasing sample size and highlighted the importance of appropriate aggregation strategies when integrating heterogeneous datasets.

Differential expression and co-expression analyses have been widely used to identify genes associated with wood formation. Co-expression networks derived from RNA-seq data can group together genes with correlated expression profiles into modules that frequently correspond to defined biological processes, such as lignin biosynthesis or cell cycle regulation. In *Populus* and other plant species, modules enriched for SCW formation and cambial activity have been identified and genes occupying central positions within these modules often represent key

regulatory components (López-Rubio et al., 2020; Wang et al., 2018). Comparative network analyses, which align orthologous gene sets across species, can be used to examine how biological systems evolve by analysing sequence diversity and functional data across species and can further enable identifying conserved core regulators and lineage-specific regulators (Netotea et al., 2014; Wang et al., 2018; López-Rubio et al., 2020).

For tree biology, the rapid expansion of available genomes together with high-resolution transcriptomic and chromatin datasets now enables detailed comparisons of how expression programmes and regulatory networks are organised and diverge across lineages (Pai et al., 2015; Romero et al., 2012). Preserved module architectures and recurrent network motif enrichments, such as over-represented feed-forward loops, regulatory cascades and shared TF-target patterns; are more robust indicators of shared regulatory logic across species than conservation of individual gene-gene correlations or single-species network edges. Cross-species comparisons typically align modules or network components through orthogroups using specialised network-alignment techniques to identify such conserved elements and reveal common regulators of SCW across species or lineage-specific innovations associated with ecological or developmental adaptations (Netotea et al., 2014; Movahedi et al., 2011). These predicted regulatory relationships can subsequently be validated through targeted functional assays, such as CRISPR-Cas9 mutagenesis or yeast one-hybrid experiments.

Cross-species approaches depend on careful experimental and sampling design. Co-expression networks are sensitive to tissue composition and variation, mismatches between species can produce artefactual differences that do not reflect regulatory evolution and network interpretation must consider tissue composition and spatial context (Movahedi et al., 2011; Sundell et al., 2017). At the molecular level, sequence repetitions or duplicated genes may divide ancestral functions or diverge, so paralog-aware orthogroup strategies are essential for reliable inference (Wendel et al., 2016; Netotea et al., 2014). Network reconstruction also faces statistical constraints, since the number of potential regulatory interactions greatly exceeds the number of samples, making false positives likely (Bansal et al., 2007). Co-expression also

captures correlation rather than causation and therefore benefits from complementary assays (Maher et al., 2018). Because of these limitations, integrating independent method data, such as chromatin accessibility, TF-binding maps and phylogenetic information, can help improve network predictions and must be followed by experimental validation.

Combinatorial, multi-omics methods

Many biological traits, such as wood formation, are controlled by large, multi-gene regulatory pathways and their genetic basis cannot be understood from single-gene analyses alone. Furthermore, the transcriptome can also be influenced by chromatin structure, post-translational modifications (PTMs) on TFs and epigenetic regulators, such as DNA methylation patterns (Lloyd & Lister, 2022; Wakamori et al., 2020). Consequently, the study of these traits requires interdisciplinary, integrative frameworks that combine multiple data types to resolve regulatory interactions across molecular layers, incorporate protein activity and interaction assays and link regulatory variation to phenotypic and metabolic properties (ENCODE Project Consortium, 2012). The integration of such heterogeneous, genome-scale data has become feasible with the development of computational methods, extensive databases, the spread of high-throughput techniques and improved sequence alignment methods (Xu et al., 2023). These multi-omics studies use combinatorial approaches, for example, complementing the transcriptomic RNA-seq data with chromatin structure analysis to consider the topology or TF binding data to study the CREs. Some of the techniques used to complement transcriptomic data are:

Assay for Transposase-Accessible Chromatin (ATAC-seq)

Inside the nucleus, many proteins act in association with DNA, together forming chromatin complexes. Regions in which the chromatin complexes are less compacted and more unravelled can interact with

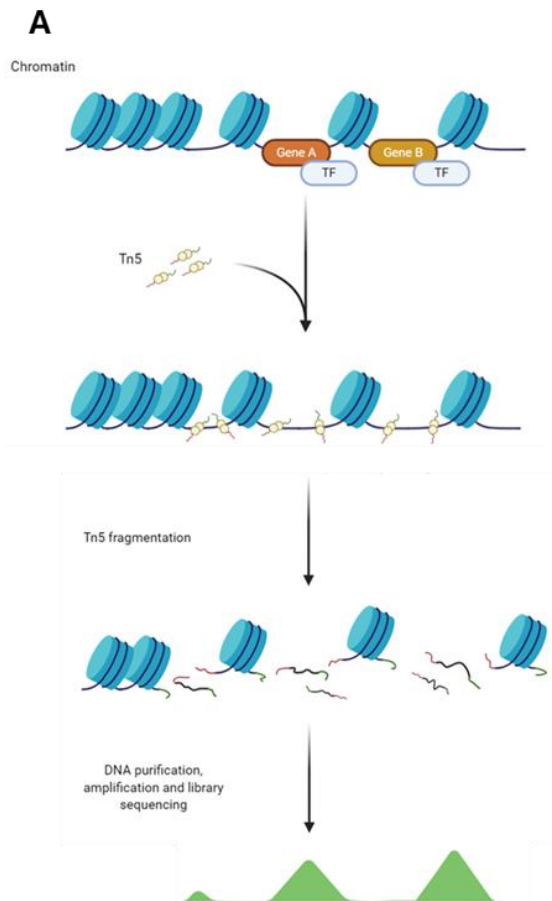
proteins, such as TFs (Pal et al., 2018). The Assay for Transposase-Accessible Chromatin (ATAC-seq) allows detecting those regions of open chromatin along the genome. Generally, TF binding sites are located within 3 kb downstream of such areas and, consequently, this technique can determine which TF binding sites are located within open chromatin (Maher et al., 2017). The protocol consists of extracting nuclei from the sample of interest and using a modified transposase that cleaves DNA and simultaneously inserts sequencing adapters. A library is elaborated through PCR and later sequenced using Illumina sequencing. Higher accessibility regions, such as open chromatin, are cleaved with higher frequency and generate more sequencing reads. Thus, the library reveals regions without nucleosomes where TFs would have access to bind to their target DNA motif (Maher et al., 2017), identifying target genes (Zander et al., 2020) (Fig. 5A).

DNA affinity purification sequencing (DAP-Seq)

The DNA affinity purification sequencing (DAP-Seq) technique allows genome-wide determination of binding interactions between specific TFs and DNA motifs. A DNA library, constructed from the genomic sample of interest and keeping its characteristic secondary modifications, is fragmented and adaptors are ligated to the ends of the resulting DNA fragments. Then, they are incubated with a set of specific, *in vitro* expressed TFs attached to an affinity tag. DNA fragments that are not bound by the TFs are washed away, while the remaining TF-DNA complexes are purified and the DNA recovered for sequencing (Fig. 5B). One reaction is performed for each TF separately. Thus, when the reads obtained are mapped to a reference genome it is possible to identify the genome-wide locations and motifs with which each TF interact. Those interactions determine the transcriptional patterns that control development, metabolism and reactions to the environment. Consequently, they provide crucial information for the investigation of gene regulatory networks and biological traits (Bartlett et al., 2017).

The data from DAP-seq can be combined with epigenetic maps, particularly methylation patterns, to assess the effect of base-level DNA modifications in TF interaction and gene regulation. The same DNA

sample with different methylation patterns can be studied for the same TF to investigate the effects of such epigenetic modifications. Moreover, other proteins can be expressed *in vitro* and used in the protocol to test their interaction with the DNA sample (Bartlett et al., 2017). The output of the DAP-seq technique is similar to the results of ChIP-seq. However, unlike ChIP-seq, DAP-seq does not require the preparation of tagged lines or gene-specific antibodies and DAP-seq can be performed as a high throughput assay for numerous TFs simultaneously. Unfortunately, DAP-seq cannot show the effect of other relevant genomic factors, such as *in vivo* protein interactions or chromatin conformation and it should be complemented with other techniques, such as ATAC-seq. DAP-seq results can also be overlapped with ChIP-seq, which may include *in vivo* binding interactions not reflected in DAP-seq (Bartlett et al., 2017; Maher et al., 2017; McLeay et al., 2012).



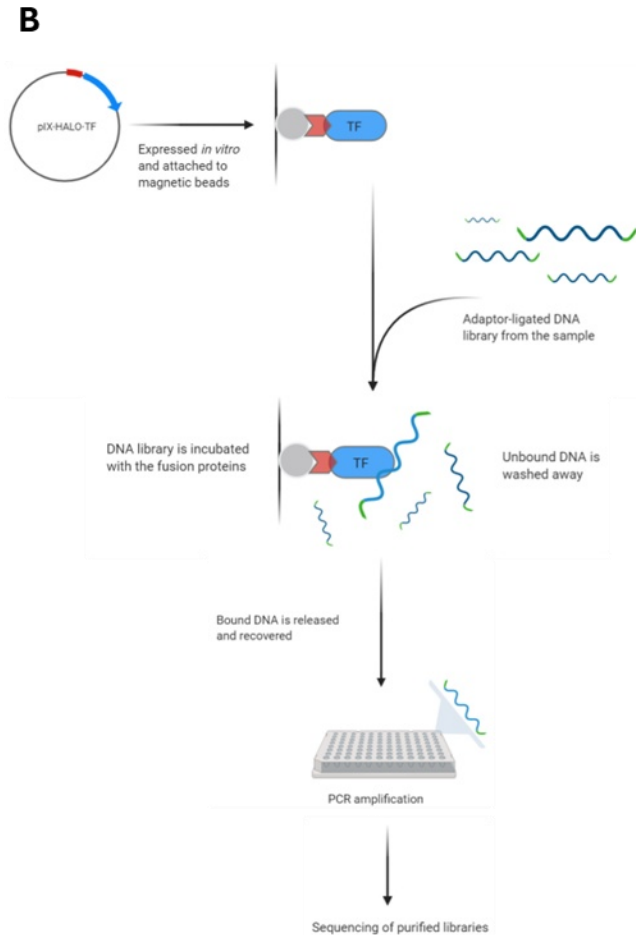


Figure 5. Outline of the two research techniques: (A) ATAC-seq and (B) DAP-seq.

It has been demonstrated that integrating co-expression networks with additional omic layers, such as DNA motifs, chromatin accessibility and TF binding, can result in integrated gene regulatory networks that outperform conventional networks, providing a powerful framework to characterize progressive cellular development and dynamic biological processes across tissues and time (Sundell et al., 2017; Taylor-Teeple et al., 2015). Such combined approaches can identify candidate key regulators, which can then be validated through targeted validation assays (O’Malley et al., 2016). First, the model must be built within a single species using co-expression networks from RNA-seq series to

pinpoint relevant modules associated with developmental stages or mechanisms. Within these modules, candidate regulators can be inferred based on their centrality, specificity and coordination with key developmental processes. Then, techniques, such as yeast one-hybrid screens, ChIP-seq or DAP-seq, can map direct TF-DNA interactions that help establish causal tissue- or process-specific regulatory relationships. Complementary open-chromatin profiling methods, such as DNase-seq and ATAC-seq, reveal the distribution and dynamics of accessible regulatory regions and provide a basis for inferring regulatory networks across organs, treatments and cell types (De Clercq et al., 2020).

These combined, multi-omic techniques have been used in multiple studies to research the mechanisms behind wood formation. In one example, Wang et al. (2018) engineered a *Populus* model by modifying 21 lignin pathway genes and measuring transcriptomes, proteomes, metabolites and wood phenotypes across ~2000 transgenic lines. Integrating these layers quantitatively allowed them to link gene expression to lignin content, wood density and many other traits. This approach predicted how manipulating specific monolignol genes would improve wood properties, outlining a path from genomics to the bioengineering of timber quality.

Taylor-Teeple et al. (2015) examined the coordination and hierarchical organisation of TFs in *A. thaliana* by constructing a protein-DNA regulatory network linked to functional adaptation under abiotic stress. Using high-resolution expression data together with selected literature, they first identified a set of 50 genes associated with SCW formation, inferred TF-target relationships and evaluated candidate regulators under stress conditions. The supervised network architecture was built with a machine-learning framework, using yeast one-hybrid-derived protein-DNA interactions to assign directionality and Pearson correlations to determine regulatory sign, whereas an unsupervised network was generated from the consensus of four inference methods (GENIE3, Inferelator, TIGRESS and ANOVERence), providing complementary, data-driven predictions of regulatory interactions (Taylor-Teeple et al., 2015). Chen et al. (2019) similarly constructed a hierarchical network in *Populus trichocarpa*: they induced a master NAC TF (PtrSND1-B1) and combined RNA-seq with ChIP-seq to map its regulatory cascade. The resulting network had four layers of TF-target

interactions and revealed 55 novel *in vivo* TF-DNA interactions controlling wood biosynthesis. Nearly 90% of tested interactions were validated across transgenic lines. These studies show how integrating expression data with TF binding assays (yeast one-hybrid, ChIP-seq) can reveal the transcriptional circuitry of wood formation in plants.

The design of comparative datasets is critical, since differences in developmental stage, tissue composition, or environmental conditions can generate expression variation that may be mistaken for evolutionary divergence. Previous work from our group demonstrated the importance of anatomically matched, high-spatial-resolution sampling: the AspWood resource, based on >20 pooled 15- μ m cryosections per tree with several replicates, provided unprecedented gene coverage and resolution, revealing a continuous progression of biological processes across the wood-forming tissues (Sundell et al., 2017). At the same time, the presence of paralogs and whole-genome duplications introduces challenges for cross-species comparisons due to mapping ambiguity. To avoid misleading inferences, comparative analyses must therefore use orthogroup-aware mappings and methods that explicitly account for paralog diversity (Sundell et al., 2017; Jokipii-Lukkari et al., 2017).

Comparative genomic and transcriptomic studies have also shed light on the evolution of wood. By mapping expression data onto phylogenies, conserved regulatory programmes can be identified. Sundell et al. (2017) demonstrated that most core processes of secondary growth, such as cell division or SCW synthesis, are co-expressed in both *Populus* and *Picea*, supporting a shared regulatory blueprint. In contrast, genes underlying vessel element formation, typical of angiosperm xylem, formed a *Populus*-specific expression module that was absent in *Picea*. These observations align with the idea that ancient SCW machinery was largely retained but modified in different lineages, for example via neo-functionalisation of duplicated genes (Zhong & Ye, 2015). In this context, Wullschleger et al. (2013) used cross-species co-expression to identify conserved NAC-MYB modules, while Lorieau et al. (2021) compared 17 tree genomes to highlight lineage-specific TF expansions. These studies are representative of evolutionary developmental (evo-devo) genomics, an integrative discipline that uses comparative genomic and transcriptomic datasets to elucidate how the molecular mechanisms governing plant development are modified across evolutionary

timescales (Carroll, 2008; Müller, 2007). By bridging the gap between ontogenetic processes and phylogenetic history, such genomics efforts are revealing how the core SCW regulatory cascade was co-opted and diversified in woody plants.

Material and Methods

In this section, I outline the workflow of the main tree species and techniques used to produce the data in each of the studies in this thesis. In particular, I focus on the methods that were modified or adapted for better results in woody samples and other tree tissues. Full details are available in the methods section of each paper.

Forest Tree Species Researched in this thesis

Aspen

The genus *Populus* comprises approximately 30 species, such as aspens, poplars (Dillen et al., 2009; Lin et al., 2018). These hardwood seasonal perennial trees exhibit extensive geographical distribution and remarkable adaptiveness, spanning from sub-tropical regions in Central Africa and Asia to the high-latitude boreal forests of the Northern Hemisphere. They are characterised by their rapid biomass accumulation and high growth rates (Klevebring et al., 2009). *Populus* species also maintain significant levels of genetic variation, driven by the production of large amounts of wind-borne pollen and frequent interspecific hybridisation (Slavov & Zhelev, 2009; Street & Tsai, 2009). Furthermore, despite their long juvenile stage, their capacity for clonal propagation and their high susceptibility to *Agrobacterium*-mediated transformation and CRISPR-Cas9 genome editing, have established them as excellent model systems to study tree biology (Lin et al., 2018; Plomion et al., 2001).

Populus trees have significant commercial value thanks to their exceptional bioproductivity and adaptability. Due to their relatively low lignin-to-cellulose ratio and thin fibres, several species, most notably the aspens *P. tremula* and *P. tremuloides*, are primary feedstocks for the pulp and paper industries (Groover et al., 2009). Beyond industrial

application, *Populus* is also a prolific source of specialised metabolites of pharmaceutical and ecological interest. For example, they synthesise various phenolic compounds through the shikimic acid pathway. Among these, of which the salicin-based phenolic glycosides from the *Salicaceae* family are the most relevant (Constabel & L. Lindroth, 2009). Moreover, their rapid root-system proliferation renders them indispensable in silviculture and ecological restoration (Dillen et al., 2009), where they can serve as effective biological agents for soil stabilisation and erosion control (Groover et al., 2009).

From a genomic perspective, *Populus* species possess a relatively compact genome (~500 Mbp) contained in a diploid set of 38 chromosomes (Lin et al., 2018; Street & Tsai, 2009), which show a low degree of linkage disequilibrium and mutation rates, comparable to most of the angiosperm families (Ingvarsson, 2009). Their genome is also well conserved across species with low somaclonal variation (Ellis et al., 2009; Fromm, 2013; Lin et al., 2018). Their genomic and physiological attributes have established *Populus* as the premier model for forest genomics. In 2006, *P. trichocarpa* was the first tree species and third plant species for which the whole genome was sequenced (Ellis et al., 2009).

Norway Spruce

Conifers represent the most phylogenetically diverse and ecologically significant lineage of gymnosperms and they serve as the primary sources of renewable biomass for the pulp and biofuel industries. Among them, Norway spruce (*Picea abies*) exhibits a vast distribution spanning from Scandinavia to the Balkans and the Ural Mountains (Bernhardsson et al., 2019). In Sweden, *P. abies* represents more than 40% of forest trees and is regarded as a cornerstone of both the national economy and forest research (Dobrowolska et al., 2017; Nystedt et al., 2013).

P. abies possesses a diploid set of 24 equally sized chromosomes containing a large genome of approximately 20 GB (Bernhardsson et al., 2019). In contrast to most angiosperms, *P. abies* genome shows higher

amounts of DNA-methylation silenced repetitive elements such as transposons, which constitute more than 70% of the *P. abies* genome (Ausin et al., 2016). Subsequently and because of limitations of the sequencing and assembly technologies in the past, the *P. abies* reference genome assembly (v1.0) was highly fragmented, representing approximately 60% (~12 GB) of the estimated genome size and showcasing the technical challenges offered by such large genomes. Yet, remarkably, the assembly's 70,736 predicted gene models have served as a robust foundation for conifer research (Bernhardsson et al., 2019; Nystedt et al., 2013).

Birch

Birch (*Betula pendula*) is a fast-growing, broadleaved tree with a relatively small and compact genome (~440 Mb), a short juvenile period and well-developed germplasm resources. These properties have encouraged its use as a tree model for functional genomics and wood biology research. Currently, a chromosome-scale reference genome and population genomic resources are available (Salojärvi et al., 2017).

A defining characteristic of *B. pendula* is its capacity for precocious flowering. While most forest trees require decades to reach reproductive maturity, birch can be induced to flower within a single year through specific long-day treatments in controlled environments. This can drastically reduce the time required for crossing and genetic analysis (Lemmetyinen et al., 2004).

Scots pine

As a representative *Pinaceae* conifer species, Scots pine (*Pinus sylvestris*) is characterised by an exceptionally expansive genome (~22-24 GB), which is predominantly composed of repetitive elements (>80%) and exhibits substantial genetic diversity across populations (Nystedt et al., 2013; Wachowiak et al., 2015). Despite the inherent computational hurdles posed by such genomic architecture, the rapid proliferation of

community-driven resources, such as dense SNP arrays (e.g.: PiSy50k), exome/transcriptome resources and large resequencing panels; has fundamentally transformed the field. Currently, these tools can facilitate sophisticated genomic studies even in the absence of a simple, contiguous reference (Kastally et al., 2022).

Lodgepole pine

Consistent with other members of the genus *Pinus*, lodgepole pine (*Pinus contorta*) possesses an enormous, repeat-rich genome (>20 GB) that poses substantial challenges for *de novo* genome assembly. These challenges are exacerbated by high levels of heterozygosity and extensive intronic and intergenic regions. *P. contorta* exhibits significant genetic structure diversity and adaptive divergence across its vast geographical range, making it the subject of extensive population and adaptive studies, such as analyses of hybrid zones with related pines. These studies, together with high-density SNP datasets, have enabled new analyses of transcriptional adaptation, providing insights into the genomic basis of local evolution (Cullingham et al., 2012; Yeaman et al., 2014; Yu et al., 2022). *P. contorta* is also relevant for Swedish forestry, particularly in northern regions, since it grows roughly 40% faster than native *P. sylvestris* (Elfving et al., 2001), offering a rapid, high-yield alternative to secure long-term timber supply and enhance carbon sequestration (Lundmark et al., 2014).

Cherry

The perennial tree cherry (*Prunus avium*) represents a convenient angiosperm model, possessing a relatively compact, well-annotated genome (~350 Mb) and supported by multiple high-quality genome assemblies and population genomics datasets. While primarily regarded as a horticultural model, its well-characterised genomic architecture provides a robust comparative reference from the *Rosaceae* group among fruit trees. Consequently, *P. avium* serves as a strong candidate for cross-species transcriptional analyses (Shirasawa et al., 2017; Wang et al., 2020).

Drimys angustifolia

Drimys angustifolia (Winteraceae) is a small, evergreen shrub from the magnoliid clade of angiosperms, native to the mountainous regions of southern Brazil. Unlike the vast majority of angiosperms, it possesses secondary xylem without vessels, a primitive structure where water transport and mechanical support are conducted solely by tracheid structures, analogous to the wood anatomy of conifers (Carlquist, 1988; Feild and Holbrook, 2000). This unique anatomical feature makes it an essential model for comparative evolutionary studies, offering insight into the evolutionary transition from tracheids to vessel elements (Marquínez et al., 2009). Due to the absence of a high-quality reference genome for this species, a *de novo* full-length transcriptome was generated in Paper I using Iso-seq data from field-collected *Drimys angustifolia* trunk tissue.

Cryosectioning

Wood blocks of *Betula pendula*, *Pinus sylvestris*, *P. abies*, *Pinus contorta* and *Prunus avium* were harvested from trees in the field, frozen on dry ice and stored at -80 °C until processing (Fig. 6). From each frozen block a prism-shaped piece spanning bark, phloem, cambium, developing xylem and mature xylem until the previous year ring was excised with a power saw and mounted in a cryo-microtome machine for serial longitudinal cryosectioning. This technique had been previously used to create high-spatial-resolution wood transcriptome atlases (Jokipii-Lukkari et al., 2017; Sundell et al., 2017).

In total, a series of 120-150 consecutive longitudinal sections (15 µm thick), covering the current year's growth and the four developmental layers, were collected from each wood piece (Uggla et al., 1996). To assign each longitudinal section to its developmental zone, transverse cross-sections were prepared frequently during sectioning and examined under a light microscope (Zeiss Axioplan 2 microscope; Carl Zeiss

Microscopy) equipped with an AxioCam HRC camera (Zeiss) (Figure 7). Only prism pieces showing clearly discernible tissue layers were used.

Prior to finalising the set of species used, we also attempted to obtain cross-sections from willow (*Salix viminalis*) and alder (*Alnus incana*) as angiosperm candidates, but they presented problems during cryosectioning. *A. incana* sections under the light microscope were stained by an orange resin which prevented distinguishing the developmental layers visually (Fig. 7F). On the other hand, *S. viminalis* developmental layers did not show straight separation limits, making it impossible to cut longitudinal sections that did not include more than one developmental layer (Fig. 7E).

Individual sections were transferred to sterile 1.5-ml tubes using RNase-free tools and stored immediately at -80 °C until extraction to preserve RNA integrity (Fig. 8). Three replicate series were collected from each tree to provide the spatial resolution and replication required for robust transcriptome analyses.



Figure 6. Sampling of wood blocks from *P. abies* stems. A chisel, mallet and knife were used to excise a rectangular block extending from the outer bark to, at least, the first growth ring, such as the principal developmental zones of phloem, cambium, expanding xylem and mature xylem. Blocks were immediately frozen at -80 °C and stored until cryosectioning.

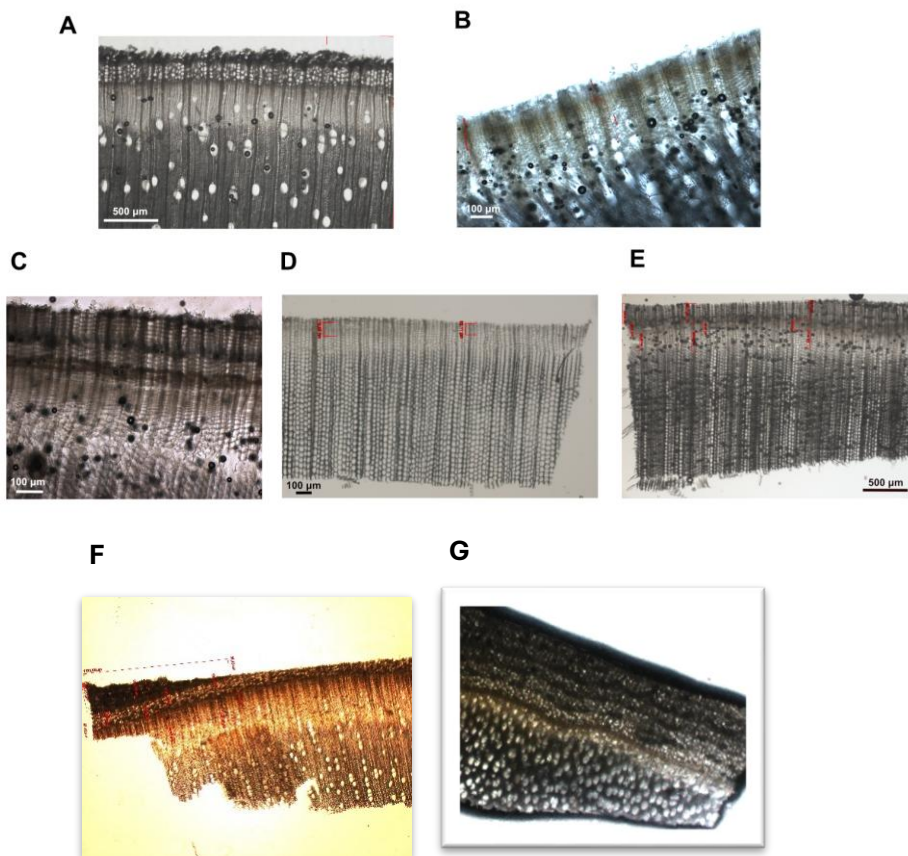


Figure 7. Transverse wood sections from the selected hardwood and softwood species. (A) Birch (*Betula pendula*), (B) cherry (*Prunus avium*), (C) Norway spruce (*P. abies*), (D) Scots pine (*Pinus sylvestris*) and (E) lodgepole pine (*Pinus contorta*). Discarded species (F) alder (*Alnus incana*) and (G) willow (*Salix viminalis*).

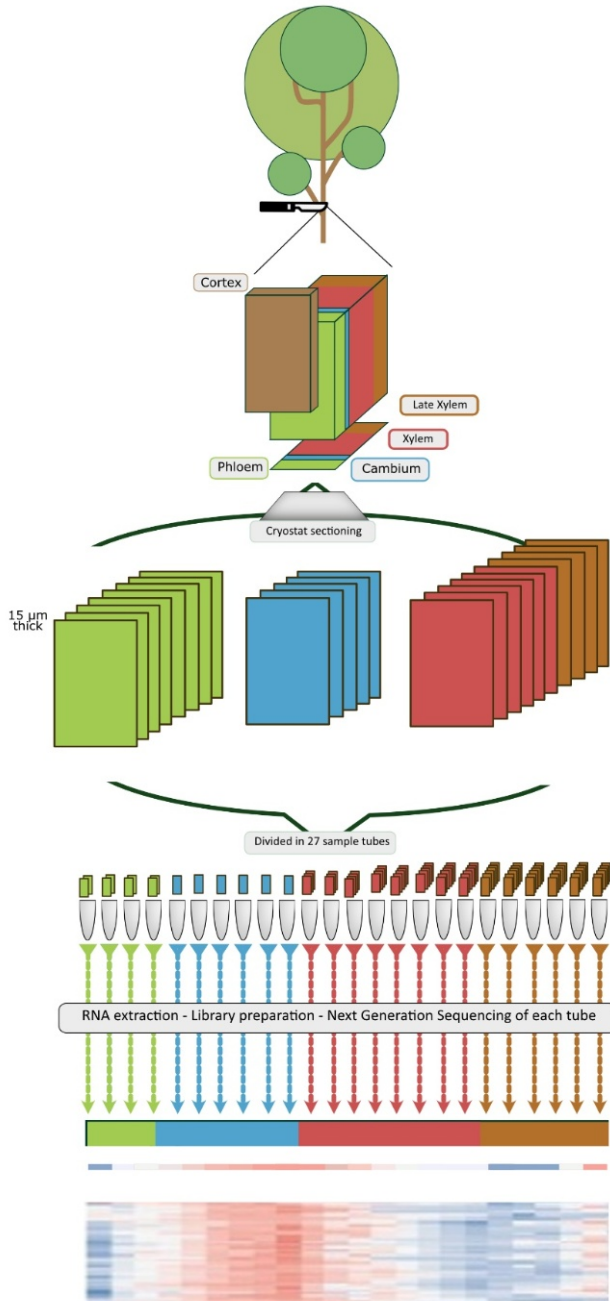


Figure 8. Sampling of wood tissue followed by cryosectioning into 15- μm layers. Consecutive sections were pooled for RNA extraction, library construction and sequencing.

RNA and DNA extraction

RNA and DNA extraction methods were adapted for high quality results in woody tissues samples. For papers I, II and III, the tissue sectioning and RNA extraction were designed to provide a high-resolution transcriptional map of the wood tissues. Some contiguous sections of the same layer were pooled together in Eppendorf tubes immediately before extraction to increase yield and improve library quality for sequencing. This was mostly performed with sections containing lower transcriptional activity and RNA content, mainly from the mature xylem. In contrast, the few sections from the active cambial region were kept separate.

Total RNA from each section or pool was extracted using a CTAB-based protocol optimised for woody tissues (adapted from Chang et al. (1993) and subsequently purified with Qiagen RNeasy Kit (Qiagen, Hilden, Germany) with modifications by Schrader et al. (2004)). The quality of the extracted RNA sample was assessed using RNA 6000 Pico Kit for 2100 Bioanalyser system (Agilent Technologies, Santa Clara, California, USA) and the high sensitivity programme in the Qubit 2.0 fluorometer (Life Technologies, Carlsbad, California, USA). Library preparation was performed using the Universal Plus mRNA-seq with NuQuant protocol (Tecan, Männedorf, Switzerland). These libraries were subsequently submitted for sequencing, producing high-spatial resolution expression data across the wood forming tissues.

For paper IV, we isolated high-molecular-weight genomic DNA from *P. tremula* and *P. abies* by adapting the nuclei isolation method from Zhang et al. (2012) and a CTAB-based nucleic-acid extraction protocol (Inglis et al., 2018, with modifications by Vikash Kumar and Zulema Carracedo Lorenzo, Umeå Plant Science Centre, 2019), which together provide a robust framework for extracting DNA from lignified tissues. The wood blocks from cryosectioning series were thawed on ice and developing xylem tissues were scraped using a scalpel previously chilled in liquid nitrogen. The scraped material was collected and powdered in liquid nitrogen using mortar and pestle. Frozen leaf material from *P. tremula* was directly ground using pre-chilled mortar and a pestle in liquid nitrogen.

Using the resulting powdered material, nuclei were isolated and the standard CTAB workflow of DNA extraction was followed. In addition,

the centrifugation speeds for nuclei recovery were optimised for each tree species according to genome size as instructed in the protocol. *P. tremula* nuclei were pelleted at ~3,500 g, whereas conifer tissues, with much larger genomes were centrifuged at ~1,900 g to prevent excessive shearing. *A. thaliana* controls were pelleted at 5,000 g as specified in the protocol. These modifications yielded high-molecular-weight DNA suitable for subsequent library preparation in paper IV.

DAP-seq protocol optimisation for recalcitrant woody tissues

DAP-seq was performed using a workflow optimised for DNA from woody, lignified tissue; with changes such as longer incubation times, performing protein-DNA binding prior to magnetic bead capture, a two-stage amplification with a customized number of cycles and an adapted data analysis. These modifications are discussed in the “Results & Discussion - Paper IV” section of this thesis and in full detail in Paper IV.

Candidate TFs were chosen by ranking co-expressed TFs upregulated in the expansion/SCW zone, retaining the top ~200 TFs with the largest expression range for DAP-seq. The TF coding sequences were cloned into a HaloTag expression vector (pIX-HALO) by Twist Bioscience (Twist Bioscience, South San Francisco, California, USA). The TF proteins were expressed *in vitro* with a Halo-tag attached using the cell-free translation system “TNT Wheat Germ Expression” (Promega, Madison, WI, USA). On average, one negative control sample (HaloTag beads with empty vector) was included for every eight TFs.

Extracted DNA from *P. abies* and *P. tremula* wood scrapings and *P. tremula* leaf tissue was fragmented via sonication into 200 bp fragments, which were end-repaired and ligated to adapters for library preparation. These DNA libraries were quantified and assessed for fragment distribution prior to TF binding. During the modified DAP protocol, high-quality DNA libraries were incubated overnight with the TF-HaloTag complexes at 4 °C with gentle rotation to enhance protein-DNA interactions and promote specific binding. Following this, Magne HaloTag magnetic beads (twice the amount used in the original

protocol), to which the HaloTag attaches, were added to the DNA-TF-HaloTag sample. This incubation was extended to 2.5 hours with gentle rotation instead of the 1 hour specified in the original protocol, as to increase the proportion of protein complexes bound to the magnetic beads. Unbound DNA fragments and proteins were washed away. The DNA fragments that had bound to the TFs were then recovered and amplified.

A two-stage PCR strategy was used to preserve library complexity and to remove small fragment contamination for a better library quality. First, a short indexing amplification of 10 cycles using the xGen DNA Library Prep MC Kit (Integrated DNA Technologies [IDT], Coralville, IA, USA) was performed, followed by a bead clean-up with AMPure XP beads (Beckman Coulter, Brea, CA, USA). Quantitative PCR (qPCR) was then performed with an aliquot of each sample to calculate the optimal number of remaining amplification cycles to avoid the formation of primer-dimers and DNA overamplification. A second, amplifying PCR was then performed based on these results, usually 2-5 cycles using the KAPA HiFi HotStart Library Amp Kit (Roche), followed by another bead clean-up. The final libraries were assessed to confirm the absence of primer-dimers and small fragments and the presence of appropriate fragment sizes with the Agilent 2100 Bioanalyser (Agilent Technologies) using High Sensitivity DNA chips (Agilent Technologies, Santa Clara, California, USA).

The final libraries were submitted for paired-end sequencing (150 bp) on an Illumina TruSeq and NovaSeq X Plus Series with a sequencing read depth of 120M. After sequencing, reads were quality trimmed and aligned to the respective reference genome using a high-sensitivity short-read mapper configured for large, repeat-rich genomes. Because multi-mapping is a major source of artefacts in these genomes, multi-mapping reads with ambiguous alignments or reads with low mapping quality were removed and PCR duplicates were marked and discarded. This conservative handling of multi-mappers ensured that downstream peak calls reflected genuine TF-DNA enrichment rather than repeat-driven noise.

Peak calling was carried out for each TF using the corresponding negative control library (empty HaloTag vector). To minimise false positives, we applied several filters not commonly performed in DAP-seq studies on smaller genomes: peaks present in the negative control, peaks recurring across many unrelated TFs and peaks overlapping annotated repeats or failing replicate reproducibility thresholds were removed.

Quality metrics such as fragment-size distribution, library complexity, duplicate rate and FRiP were used to assess data quality and only peak sets meeting these criteria were retained for further analysis. Motif assignment and regulatory inference were performed by intersecting filtered DAP peaks with gene features and, when available, ATAC-seq open chromatin regions. Motifs were identified using standard motif-discovery tools and compared against plant motif databases. Candidate target genes were linked to TFs by combining three layers of evidence: motif presence in accessible chromatin, physical binding in DAP-seq and orthogroup-informed co-expression patterns. This integrative approach provided a robust basis for prioritising high-confidence TF-target interactions for downstream functional validation.

Results and Discussion

Paper I - An updated perspective: what genes make a tree a tree?

At first glance, one can easily believe that trees, as tall, woody organisms, are their own species family, separate from other plants. However, trees are not a single biological group, but a convergent growth phenotype that has emerged in plant evolution multiple times, with wood growth evolving and disappearing repeatedly in some plant lineages, such as in angiosperms (Fig. 9A) (Luo et al., 2023). Contrary to common belief, most herbaceous plants exhibit secondary growth (Fig. 9B), albeit without the extensive biomass accumulation that forms tree trunks (Groover, 2005). In the 2005 paper “What genes make a tree a tree?” (Andrew Groover, 2005) it was argued that there are no unique “tree genes” that exist only in woody species (Groover, 2005). Instead, woody growth results from differential expression regulation of genes present in both woody and herbaceous plants. Just one year later, the first tree genome (*Populus trichocarpa*) was published (Tuskan et al., 2006), followed by many others in the following years (Fig. 9C). These new genomic resources offer the opportunity to examine the genomes of multiple woody and herbaceous species, identify the sources of tree phenotypes and address the question “What makes a tree a tree?”.

In Paper I, we used this new genomic data to perform comparative analyses within the *Rosaceae* family, which contains most of the best known forest trees and assembled genomes (Wang et al., 2009) and to provide new insights in the evolution of secondary growth. The analysis within the rosid clade was rooted using the magnoliid species *Drimys angustifolia* and the basal eudicots *Vitis vinifera* and *Macadamia integrifolia* as outgroups, providing external angiosperm references for inferring lineage-specific patterns of selection and gene family evolution. We focused on two classes of genes that could highlight the genomic patterns required for woodiness and help to identify specific agents relevant for secondary growth:

- 1) Genes exhibiting relaxed purifying selection pressure in herbaceous species derived from woody species.

2) Gene families showing decreased copy number following a transition from woody species to herbaceous species.

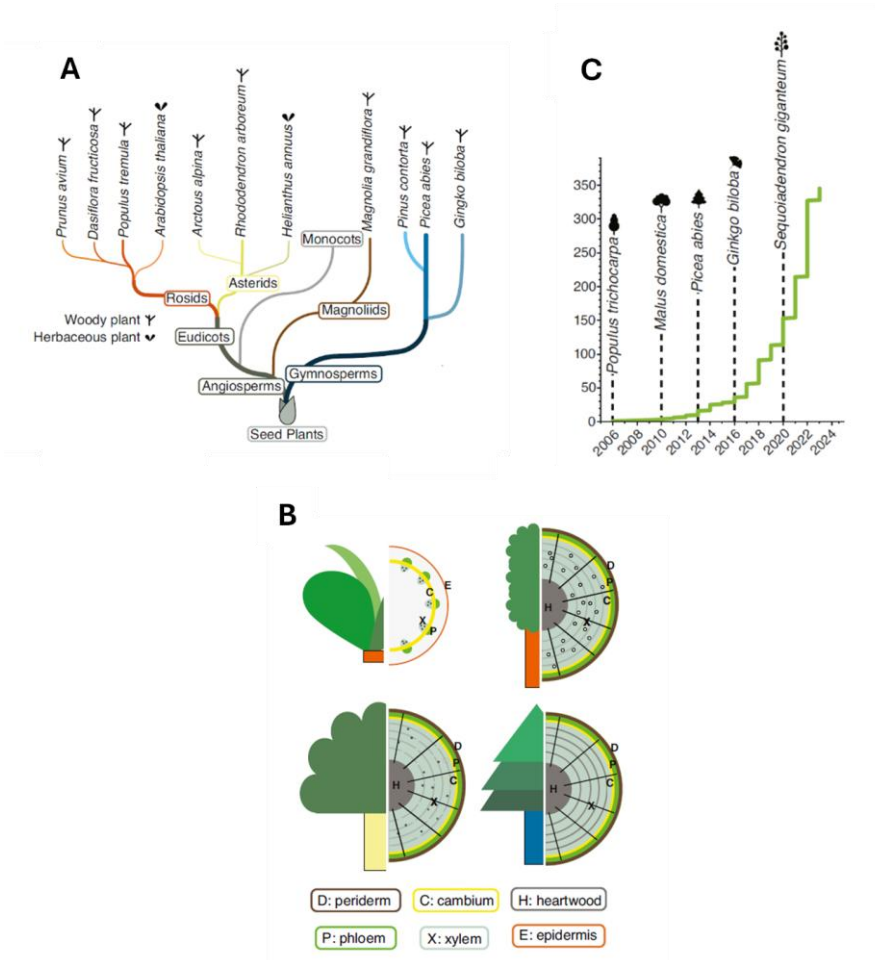


Figure 9. (A) Simplified phylogeny illustrating how the tree habit has evolved multiple times across seed plants. (B) Examples of tissues showing secondary growth in *A. thaliana* (top left), *P. tremula* (top right), *Rhododendron arboreum* (bottom left) and *P. abies* (bottom right). (C) Number of arborescent species with published genomes among angiosperms and gymnosperms. Only taxa consistently described as trees in online floras were included; shrubs were excluded to reduce ambiguity, although shrub species that occasionally attain tree form were retained. Publication data were obtained from https://plabipd.de/plant_genomes_pa.ep.

Genes under relaxed selection in herbaceous species

Genes under relaxed selection can be identified as genes that accumulate more mutations in herbaceous species than in woody species. Genes that experience weaker purifying selection in a lineage following loss of woodiness can accumulate mutations to the point that they lose their function (pseudogenisation). This suggests that the gene is not essential for an herbaceous lifestyle. Conversely, the same gene is therefore likely to be functionally important in trees, where it remains under strong purifying or even positive selection. These would be genes that help define processes that trees must maintain. We used RELAX in HyPhy (Wertheim et al., 2015) to identify such genes in transitions from woody to herbaceous phenotypes in four rosid groups. 470-1661 genes per group showed evidence of relaxed selection. More importantly, 19 genes were identified as under relaxed selection in all tested transitions, which included chromatin remodelling proteins, TFs and genes associated to wood growth (Fig. 10A).

Several genes that showed relaxed selection in all categories were involved in flowering and development, such as MYB DOMAIN PROTEIN 88 (MYB88), which is expressed consistently from the embryonic to flowering stage and affects drought resistance (Lei et al., 2015; Xie et al., 2010). Other genes highlighted in this analysis were involved in SCW and lignin formation, such as xylem-specific O-methyltransferases proteins, processes that would be much more relevant in a woody tree than an herbaceous plant (Fig. 10B).

Reduced gene families

Gene families that repeatedly lose gene copies in herbaceous lineages suggest that multiple copies of their genes are not needed for a non-woody lifestyle, whereas if they are expanded in trees they could be relevant for woodiness. We used several parallel methods to investigate the differences in gene family size in transitions from woody to herbaceous species and identify contracting gene families: statistical tests (which identified 26 gene families), a machine learning approach called random forest (101 gene families, 16 of them in common with the statistical test) and a phylogeny-aware CAFE analysis (417 gene families). The results identified families involved in SCW development and abiotic/biotic stress responses (Fig. 10C).

One family was significantly retracted in all three analyses, which consisted of leucine-rich repeat protein kinases, participate in disease defence and are expressed in wood-forming tissues (Fig. 10D). Along with the retracted families involved in stress responses, as well as stress-related relaxed genes, this may imply that a short-lived herbaceous plant does not require as many defence mechanisms as a long-lived tree that will encounter a high number of threats over time (Carlsson-Granér & Thrall, 2006).

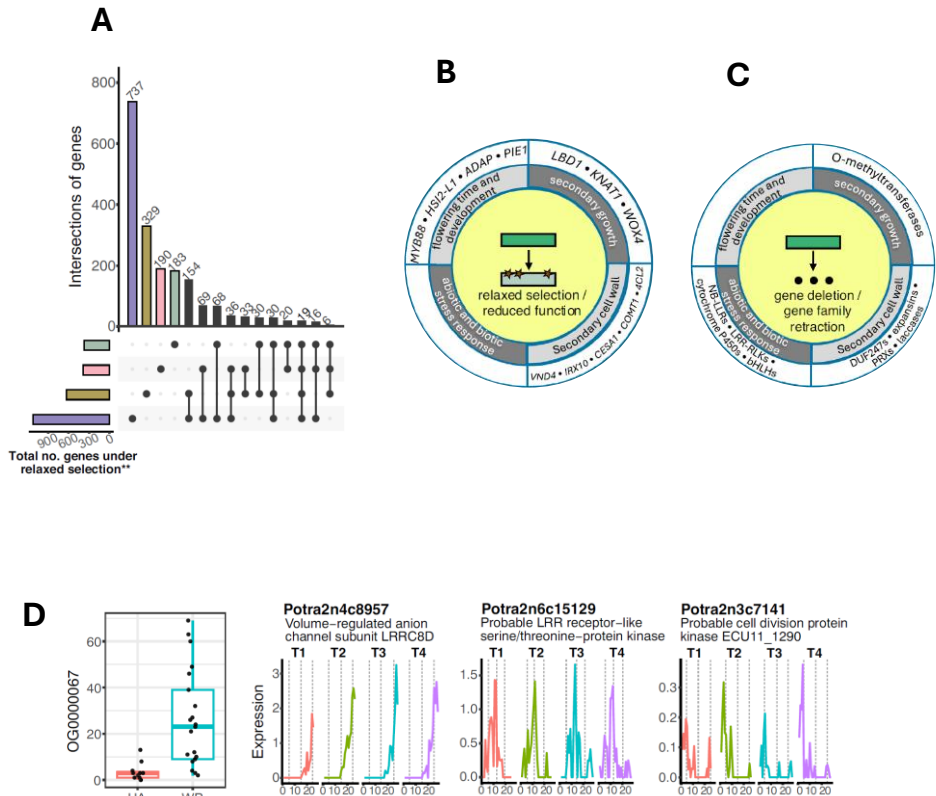


Figure 10. Comparative analysis of woodiness in the rosid clade. (A) UpSet plot showing genes inferred to be under relaxed selection in independent transitions from woodiness to herbaceousness, with intersections based on orthogroup membership. All intersections were larger than expected by chance. The bottom part indicates the total number of genes under relaxed selection, while the upper part of the plot displays the number of unique genes per order. Connected dots indicate intersections between orders. (B) A gene under purifying selection may accumulate nonsynonymous mutations once selective constraint is

relaxed, potentially diminishing or eliminating its function (pseudogenisation). (C) A gene may be deleted, often after progressive functional degradation, which can result in gene-family contraction. Expansions and contractions of gene families generally reflect multiple duplication or deletion events rather than a single change. (D) Gene family OG000067 exhibited significantly fewer copies in herbaceous annuals (HAs) than in woody perennials (WPs) based on statistical testing and random forest analysis. The copy number comparison is followed by expression profiles for three *P. tremula* genes from this family across wood-forming tissues. Sample numbers correspond to tissue types: ~1-5 = phloem, ~3-7 = dividing cambial cells (transition zone), ~6-12 = expanding xylem and ~13-28 = lignified xylem. Four natural clonal replicates (T1-T4) of a single genotype are shown (Sundell et al., 2017).

Conclusions:

Altogether, this analysis offers a more nuanced model for woodiness evolution than postulated by Andrew Groover. Many of the genes identified were TF or chromatin-remodelling genes, which supports the relevance of changes in gene regulation rather than new genes for woodiness. However, alternative factors, such as SCW growth- and defence-associated gene loss, are recurrent in the transition from woody to herbaceous species. The analysis suggests that such genes become less important in herbaceous plants with reduced lifespans. Thus, gene loss or pseudogenisation are proposed as complementary mechanisms driving the evolution and replacement of woodiness as it weakens and renders specific biochemical functions unnecessary in the new species.

A more detailed comparative analysis of the regulatory sequences among these species was not possible due to the lack of reliable alignment of noncoding regions across significant phylogenetic distances. The analysis also depended on functional annotations rooted in *A. thaliana*, which may miss lineage-specific tree functions and, since the analysis focused only on angiosperms, it remains unclear whether the regulatory features we identify extend to other tree lineages, such as gymnosperms, or instead represent angiosperm-specific aspects of SCW. These limitations can be overcome with combined approaches of complementary techniques, such as comparative regulomics and the optimised direct assays developed in this thesis. The same comparative analysis could also be extended to other members of the rosoid clade.

Following this study in the angiosperm rosid group, we used two gymnosperm conifers for a comparative analysis with RNA-sequencing and chromatin data (Paper II). Furthermore, in Paper III, we performed a comparative transcriptome analysis and studied the conserved co-expression networks between three representatives of the angiosperm family and three representatives of the gymnosperms.

Paper II - 1000 conifer genomes: genome innovation, organisation and diversity

In Paper II, we generated chromosome-scale reference genome assemblies for the two dominant conifer species of northern Eurasia, *P. abies* and *Pinus sylvestris* and produced a comprehensive population resequencing dataset comprising more than 1,000 Norway spruce individuals. This genomic resource represents a major advance for conifer research: the previous *P. abies* assembly was highly fragmented and substantially smaller than the true genome size and no reference genome was available for *P. sylvestris*. Together, these datasets provide a robust foundation for future studies in comparative genomics, evolutionary biology and forest genetics.

Conserved expression patterns

Conifer genomic studies are underrepresented in comparison to studies in angiosperms. Gymnosperm trees pose multiple challenges for research, such as slow growth, long generation times and remarkably large, repeat-rich genomes, that dwarf those of model angiosperms (Neale et al., 2017; De La Torre et al., 2014). In Paper II, I we built an extensive gene annotation of protein-coding genes (43,410 from *P. abies* and 49,387 from *P. sylvestris*) from a new multi-tissue transcriptome collection and datasets. Both species harbour thousands of very long genes (>50 kb) with multiple large introns (>15 kb). Protein-coding regions represent less than 1% of the genome, with gene space occupying around 11% of each genome (Fig. 11). We also generated the first compendium of active protein-coding transposable-element (TE) genes

and identified large numbers of recent pseudogenes, offering new insights into the gene repertoire dynamics in conifers.

Despite diverging around 130 million years ago (Kumar et al., 2022), both trees showed remarkable conservation of large-scale chromosomal structure with no major rearrangements or whole genome duplications (WGDs), unlike angiosperms. In angiosperms, WGD events occur often, after which the new gene copies specialise (sub-functionalisation) or find new roles (neo-functionalisation) and diploidy is re-established (Clark & Donoghue, 2018). However, WGD seems to be uncommon in conifers despite their expansive genome size (Jang et al., 2024; Niu et al., 2021). Instead, the large genome size of conifers is mainly attributed to the high abundance of TEs, particularly long terminal repeat retrotransposons (LTR-TEs) (Nystedt et al., 2013). These elements can generate additional gene copies through reverse transcription of mRNA, producing intronless cDNA molecules. These cDNA fragments can insert themselves back into the genome, producing genes that lack the surrounding regulatory context and are therefore prone to pseudogenisation.

Surprisingly, the analyses revealed that, although the protein-coding sequences of many 1:1 orthologs are relatively well conserved between *P. abies* and *P. sylvestris*, the non-coding regions immediately flanking those genes show little sequence similarity, probably due to frequent species-specific insertions of TEs, duplicated genes and pseudogenes. Consequently, each lineage has repeatedly remodelled the local intergenic landscape while preserving many coding sequences. A concrete example of these duplications results is the FTL (FLOWERING LOCUS T / TERMINAL FLOWER1-like) locus: *P. abies* contains a cluster with three copies of a recently duplicated FTL gene, whereas *P. sylvestris* lacks corresponding duplicates at this locus. The retained *P. abies* paralogs have diverged and show different expression patterns (Fig. 11) in altered conditions, constituting an example of a duplicated gene that developed different regulation patterns.

The analyses further showed that the most recent (since the species split) gene duplicates occur predominantly as local pairs or small clusters within short genomic distances (<10 Mb), consistent with recent segmental duplication events. In these expanded orthogroups, we found that the retained protein-coding duplicates are enriched for diverse biological processes, whereas the corresponding pseudogene copies show no enrichment for any functional categories, indicating that many duplicated sequences fail to acquire or maintain functional roles.

Together, these patterns suggest that evolutionary innovation in conifers is driven by frequent, largely random duplication events independent of function, with subsequent purifying selection retaining only those duplicates that provide beneficial functions. Based on the observed rates of lineage-specific duplication, we estimate that these processes have added approximately 1 GB of new sequence material since the split between *P. abies* and *P. sylvestris*.

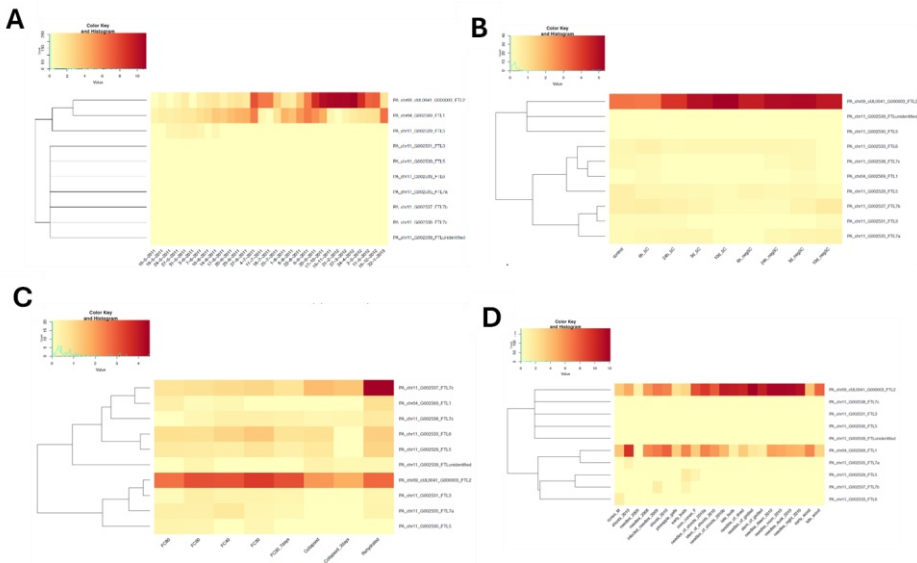


Figure 11. Expression profiles of FTL-like genes in *P. abies* across diverse environmental and tissue contexts. (A) Needles sampled across the growing season, with labels indicating collection dates. (B) Cold-stress treatments, annotated with temperature and exposure duration. (C) Drought-stress experiments, where labels denote soil water field capacity or sampling at the point of, or two days after, photosynthetic collapse. (D) Expression atlas across multiple tissues. Values represent variance-stabilized counts (VST) derived from DESeq2 and genes are clustered according to their expression similarity.

Comparative analysis for conserved and divergent expression:

To develop a comparative analysis between both tree species and investigate conserved and diverged gene expression patterns, I

performed a high-resolution workflow to profile the transcriptome of the wood-developing tissues. Utilizing precise cryosectioning, I generated a series of tissue-specific samples spanning the phloem, cambium, expanding xylem and mature xylem from both *P. abies* and *P. sylvestris*. This spatial resolution was critical to distinguish between the developmental trajectories of these two species. The co-expression results identified a substantial set of orthogroups with conserved expression patterns ($n = 5,792$), alongside a smaller group showing lineage-specific divergence ($n = 501$). Orthogroups with conserved expression tended to have more similar upstream promoter architecture than orthogroups with divergent expression. The high-resolution transcriptome and chromatin data, obtained using Hi-C chromatin contact capture and Micro-C, enabled us to test whether domain-level genome organisation contributes to transcriptional regulation during wood development. We found that genes with stage-specific expression were significantly enriched within Topological Associating Domains (TADs) bodies whereas genes expressed more constitutively across tissues are often positioned at TAD boundaries. Comparable patterns were also observed in needle tissue, indicating that this relationship between TAD structure and transcriptional regulation may be conserved across conifer organs.

The analysis of high-resolution datasets allowed for the identification of several cases where species-specific gene duplications showed signs of both sub-functionalisation and neo-functionalisation (Fig. 12). By comparing the spatial expression profiles across the wood-forming zone, we demonstrated that paralogous genes within segmentally duplicated regions exhibited divergent expression patterns and differences in the retention of functional copies, indicating ongoing pseudogenisation and regulatory divergence. Specifically, the datasets revealed that, in several expanded orthogroups, one paralog retained the ancestral expression pattern in the cambium, while its counterpart acquired a novel expression peak during secondary wall thickening. This is a sign of neo-functionalization that would remain masked in bulk tissue analysis. Together, these observations further highlight segmental duplications as an important driver of evolutionary innovation in these species, providing the raw material for the complex regulatory networks that define wood formation.

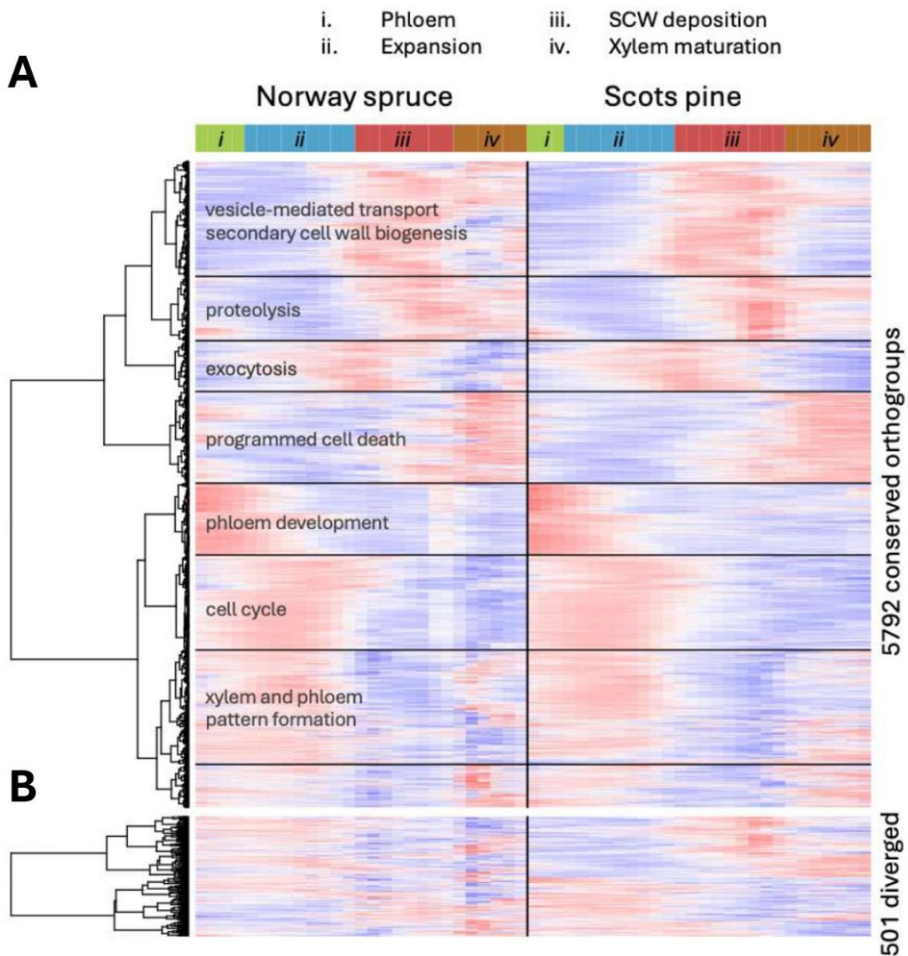


Figure 12. Evolutionary novelty through gene-expression divergence during wood development. (A) Heatmap showing expression profiles of ortholog pairs with conserved co-expression (expressologs) between *P. abies* and *P. sylvestris*. Selected significantly enriched Gene Ontology terms are highlighted. (B) Heatmap of ortholog pairs exhibiting significant expression divergence between the two species.

Paper III - An evo-devo resource for wood: Comparative regulomics across dicot and conifer trees

In Paper III, we developed a new evo-devo resource for tree research, the first high-spatial resolution transcriptomics map spanning the wood-forming tissues of six representative tree species: three dicots and three conifers, capturing 250 million years of diverging evolution between the angiosperm and gymnosperm tree lineages. By integrating spatial transcriptomics with orthology-aware network analysis, we identified key regulatory agents, conserved modules and a shared evolutionary trajectory for wood evolution. This resource is publicly available at PlantGenIE.se.

First, to achieve the required spatial resolution, I performed a high-throughput workflow using precise longitudinal cryosectioning from each replicate tree, ensuring cross-species consistency. Using those cryosection series, we generated and analysed high-spatial-resolution RNA-seq datasets covering the secondary phloem, vascular cambium, developing xylem and mature xylem of three dicot species (*P. tremula*, *Betula pendula*, *Prunus avium*) and three conifers (*P. abies*, *Pinus sylvestris*, *Pinus contorta*) (Ingvarsson & Street, 2011). Then, we mapped the RNA-seq reads to the reference genomes of each species to obtain the expression patterns of genes across the different stages of wood formation. Because no reference genome exists for *P. contorta*, its reads were aligned to the *P. sylvestris* assembly. Principal component analysis (PCA) showed that samples from the same stage (phloem - cambium - early xylem - mature xylem) clustered together consistently regardless of species, while adjacent developmental zones sit next to each other (Fig. 13A). This stage structure offers a robust framework for comparing the transcriptional programmes among species.

Next, we used this transcriptome data to identify conserved and diverged gene expression patterns across the six species. We applied an orthology-aware network analysis, in which co-expression links were defined using Pearson correlation and Mutual Rank (retaining top 3% links). To identify evolutionary conserved co-expression relationships, we used the ComPlEx framework to detect co-expressologs, pairs of orthologous genes that are co-expressed together with neighbouring genes which themselves have orthologs in each species (Fig. 13C). This conservative

definition highlights shared expression context among species and minimizes false positives, since it requires not only that the central ortholog show co-expression, but that their surrounding co-expressed genes is preserved, in other words are orthologs, across the compared species (Kalman et al., 2025). These co-expressologs were used to map orthogroup networks and extract fully connected six-member cliques.

Our analysis yielded 70,458 cliques from 2,098 orthogroups, with 2,145 unique, non-overlapping cliques. We found that 5,564 orthogroups are expressed in all six species, while 2,093 are exclusive to gymnosperms and 3,649 are exclusive to angiosperms (Fig. 13B). Together, these comparative analyses revealed that the core transcriptional program underlying wood development is remarkably conserved across the six species, despite ~250 million years of evolutionary divergence (Fig. 13D). Furthermore, the cross-species correlation matrices built using co-expressologs (Fig. 13D) showed that samples align more strongly by developmental stage than by species identity: the diagonal “transition-to-steady-state” pattern is highly similar across all species pairs, indicating that transcriptome variation is primarily driven by the shared transitions of phloem and xylem differentiation.

Despite this dominant stage structure, the matrices also retain a clear phylogenetic distinction: average correlation values for comparisons within the angiosperm group and within the conifer group are higher (~0.75), whereas between-lineage comparisons show lower overall similarity (~0.55), indicating that expression similarity is higher within evolutionary lineages than between them. In parallel, analysis of co-expressolog subnetworks (cliques) containing one gene from each species demonstrated that a substantial fraction (38% of 5,564 expressed orthogroups) exhibits conserved expression profiles across all species. This result supports deep conservation of core wood-related transcriptional programmes, rather than lineage separation. These conserved groups were enriched for cell wall and wood formation processes, such as cell cycle regulation, intracellular trafficking and polysaccharide metabolism.

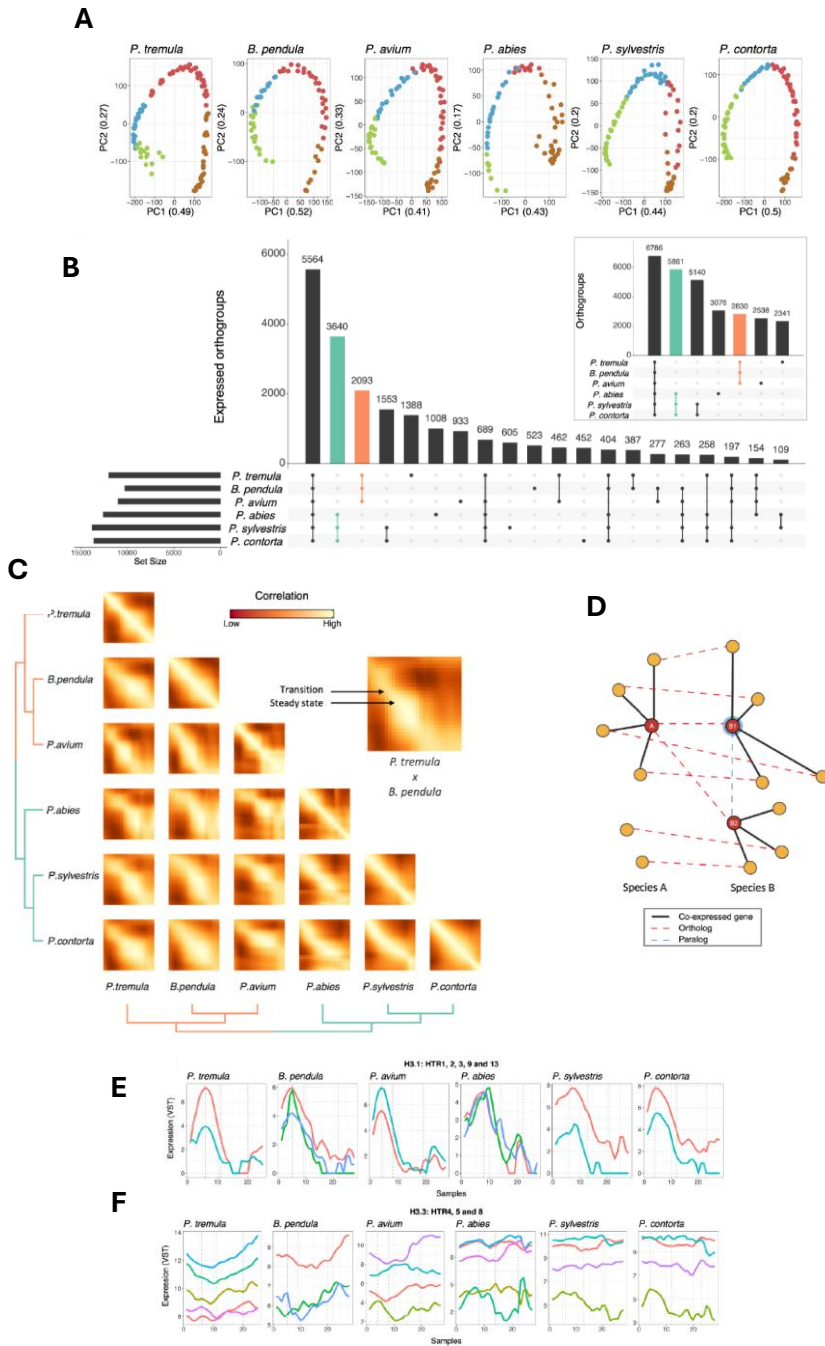


Figure 13. Comparative analysis of gene expression across the different stages of wood formation. (A) PCA of all samples from the replicate

trees, coloured according to developmental stage of wood formation. Percentage of variance explained is shown in parentheses. (B) Expressed orthogroups. UpSet plot summarizing the number of orthogroups with genes expressed in wood for each species (horizontal bars) and the number of orthogroups shared across different species subsets (vertical bars). An additional UpSet plot illustrating all orthogroups is shown in the upper-right corner. (C) Schematic of the method used to identify orthologs with conserved co-expression (co-expressologs). For each ortholog pair, a p-value is calculated that reflects the extent to which their co-expressed gene sets also consist of orthologs. In the example shown, A and B1 qualify as co-expressologs, whereas A and B2 do not. (D) Heatmaps showing sample-to-sample correlations across species pairs, computed using only co-expressologs. One biological replicate is displayed for each species. Colour scales are independently adjusted for each heatmap (species pair). (E) Genes belonging to top cliques within the H3.1 clade of the orthogroup gene tree, illustrating conserved expression across all species. (F) All expressed genes in the H3.3 clade of the orthogroup gene tree. Because H3.1 genes typically lack introns and H3.3 genes contain introns across land plants, genes that did not follow these structural criteria were removed from the analysis.

The following two orthogroups can be used as examples: H3.1 and H3.3, linked to transcriptional chromatin accessibility. H3.1-cliques showed high expression peaks restricted to the cambium of all species, representing cambial cells proliferating before differentiation (Fig. 13E). On the contrary, H3.3 orthogroups did not form co-expressolog cliques, showing instead constitutive expression across wood developmental layers and species. This lack of coordinated, stage-specific regulation explains their absence from the co-expressolog clique set and suggests a more general housekeeping or maintenance role in chromatin regulation (Fig. 13F).

Dicot and conifer wood also differ in hemicellulose acetylation and composition, offering more candidates to study those lineage divergences. Thus, we examined the expression of *Trichome Birefringence-Like (TBL)* genes, which can acetylate specific cell wall polysaccharides (Dauphin et al., 2024). We recovered nine *TBL* cliques conserved across all six species (Fig. 14B) and five cliques that are dicot-specific. The latter included *TBL* subfamilies, which are known to acetylate xylan and xyloglucan and whose expression peaks around SCW formation or later stages (Zhong et al., 2020; Zhong et al., 2017) (Fig. 14C). This suggests that, while core acetylation functions are shared, dicots have evolved specialised mechanisms for xylan modification.

Further examples are *VASCULAR-RELATED NAC-DOMAIN (VND)* orthologs, which direct SCW formation in *A. thaliana* (Nakano et al., 2015) and were found to be conserved across all six tree species, peaking during SCW as expected (Fig. 14D). In contrast, *NAC SECONDARY WALL THICKENING PROMOTING FACTOR (NST)* orthologs showed comparable expression peaks during SCW in dicots, whereas in conifers the single *NST* copy had a marked reduction in expression at this developmental stage.

To further study these regulators, we integrated ATAC-seq maps from developing xylem of *P. tremula* and *P. abies* and linked them with co-expressed gene sets (modules). From the open chromatin regions associated with each module, we identified over-represented cis-regulatory motifs of TF binding sites. These motifs were then assigned to candidate TFs by querying their orthogroups and selecting TFs whose expression profiles correlated strongly with the activity of the corresponding module ($r > 0.7$). This combined system (modules, open chromatin and TF association) resulted in a TF-module regulatory network which revealed a complex, layered regulatory structure in *P. tremula* wood, supporting the development of vessels and fibres, compared to a more streamlined system for *P. abies*, with tracheid-based wood, where *VNDs* and *MYBs* acting in parallel. This system also highlights the importance of complementing the networks with direct assays such as ATAC-seq, ChIP-seq or DAP-seq, which can provide information on chromatin and protein-genome interactions to generate a more comprehensive view of *in vivo* genome regulation. These examples represent only a small subset of the insights enabled by this evo-devo resource. These results demonstrate the resource's utility for identifying conserved and divergent biochemical and genetic mechanisms and provide a foundation for future biotechnological applications in forest genetics. To enable the future work, all expression data is publicly available at PlantGenIE.se.

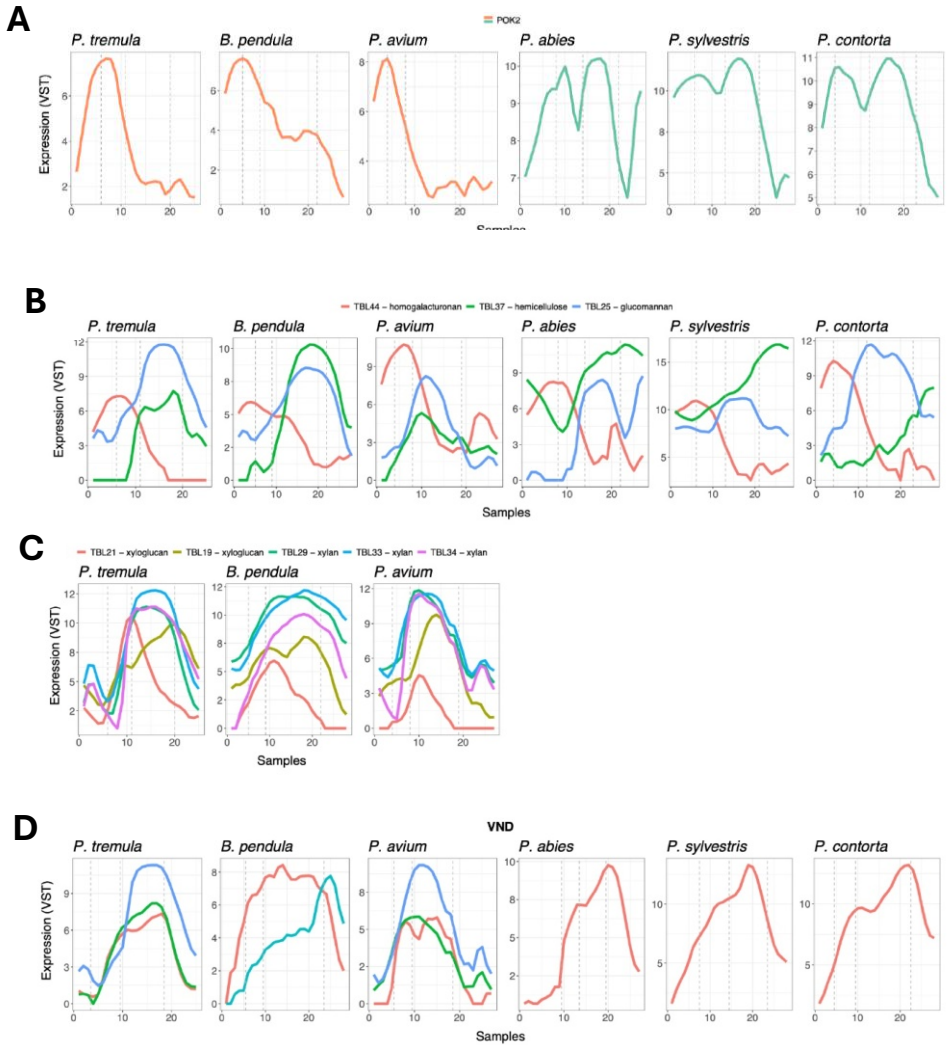


Figure 15. Genes with diverged expression between dicots and conifers. (A) Example of a gene with lineage-specific expression divergence: *phragmoplast orienting kinesin 2* (*POK2*). This gene forms separate dicot- and conifer-specific cliques and does not show conserved expression across the species examined. (B) Expression patterns of *Trichome Birefringence-Like* (*TBL*) genes with conserved expression across dicots and conifers. (C) *TBL* genes showing dicot-specific expression. (D) Conserved co-expression cliques containing *VND* master regulators in *P. tremula* and *P. abies*.

Paper IV - A high-throughput DNA affinity purification sequencing (DAP-seq) protocol method for recalcitrant tissues of woody species

Originally, the evo-devo resource of the previous paper was intended to be used to identify key regulators of wood formation in tree species and analyse them using DAP-seq, as well as CRISPR-Cas9 to generate knock-out lines of aspen. However, the initial DAP-seq trials on wood samples failed and multiple iterative experiments were performed to identify and mitigate the causes of low yield and limited scalability. Below, I summarise the DAP-seq optimisations that I investigated for woody samples and successfully tested on *P. abies* wood and *P. tremula* wood and leaves. A comprehensive description of the modified protocol is provided in Paper IV.

DNA affinity purification sequencing (DAP-seq) has become a widely used method for mapping TF binding sites at genome scale and offers an attractive alternative to ChIP-seq due to its simplicity and scalability. DAP-seq has been successfully applied in a number of annual plants, such as *A. thaliana* (O'Malley *et al.*, 2016; Bartlett *et al.*, 2017), maize (*Zea mays* L.) (Galli *et al.*, 2018; Ricci *et al.*, 2019) and rice (*Oryza sativa* L.) (Cerise *et al.*, 2021). Its use in trees remains limited and previous studies in eucalyptus (*Eucalyptus grandis* W. Hill ex Maid.), apple (*Malus domestica* Borkh.) and *Populus* spp. typically focused on a small number of TFs and relied on young tissue with low lignification or on high-yield DNA preparations from large amounts of fresh material (Brown *et al.*, 2019; Chen *et al.*, 2020; Ramos-Sánchez *et al.*, 2019; Yang *et al.*, 2019; Yao *et al.*, 2020). These choices circumvent many of the technical obstacles inherent to mature woody tissues, such as low DNA yield, high levels of enzymatic inhibitors and the presence of recalcitrant secondary metabolites. Extending DAP-seq to species such as *P. abies* and *P. tremula* introduces additional challenges. Conifer genomes are exceptionally large and repeat-rich (~20 GB for *P. abies*), which reduces the fraction of uniquely mappable reads and raises the sequencing depth needed for confident peak detection (Neale *et al.*, 2017; Treangen & Salzberg, 2012). Furthermore, only a very small proportion of fragments in the input library bind a given TF, which makes the signal highly sensitive to amplification bias and background carryover. Any reduction in library complexity, such as from degraded input or contaminants,

therefore manifests disproportionately as noise in the final dataset (Bartlett *et al.*, 2017; Landt *et al.*, 2012).

These complications are exacerbated by the biochemical properties of woody tissues. Mature secondary xylem accumulates lignin, polysaccharides and phenolic compounds that co-purify with DNA and inhibit enzymatic reactions (Ulvila *et al.*, 2020). Consequently, standard extraction protocols often require modification, typically employing CTAB buffers together with polyvinylpyrrolidone (PVP) or lithium salts to remove polysaccharides and oxidised phenolics (Porebski *et al.*, 1997; Shahan *et al.*, 2020; Lin *et al.*, 2022). Even with these adaptations, library preparation from wood often results in low DNA yield and an excess of small fragments, compromising downstream DAP-seq performance. To address these obstacles, we developed a modified DAP-seq workflow optimised specifically for recalcitrant woody tissues of *P. abies* and *P. tremula*. My objectives were to increase TF-bound DNA recovery, minimise background and ensure that final libraries were of sufficient purity and complexity for high-throughput sequencing.

The library preparation process was streamlined using the xGen DNA Library Prep MC Kit (Integrated DNA Technologies [IDT], Coralville, Iowa, USA) and AMPure XP beads (Beckman Coulter, Brea, California, USA) for bead clean-up. This was tested alongside many variations of the library preparation workflow (Fig. 15) that was presented in the original DAP protocol. In comparison, our selected library preparation method (Fig. 15I) showed higher yield and substantially reduced processing time (Fig. 16).

A major modification in the DAP protocol itself was to extend the incubation between TFs and the genomic DNA library before Halo-tag bead capture (Fig. 17). In the original protocol, TFs are immobilised on Halo-tag beads prior to their incubation with DNA, which could limit their mobility and potentially reduce binding efficiency. By first incubating TFs with DNA in solution and prolonging this step overnight at 4 °C, we enhanced the opportunity for TFs to interact with their target motifs. A second modification focused on increasing the proportion of TFs successfully captured by Halo-tag beads. Western blot analysis revealed that in many TF reactions only a minor fraction of expressed protein was recovered using the standard bead-binding duration. Extending the incubation with Halo-tag beads to 2.5 h and doubling the bead volume improved protein capture from ~10-40% to >50%, thereby increasing the downstream yield of TF-bound DNA (Fig. 17). To prevent

bead aggregation due to increased bead mass, we adjusted detergent concentration in the buffer accordingly.

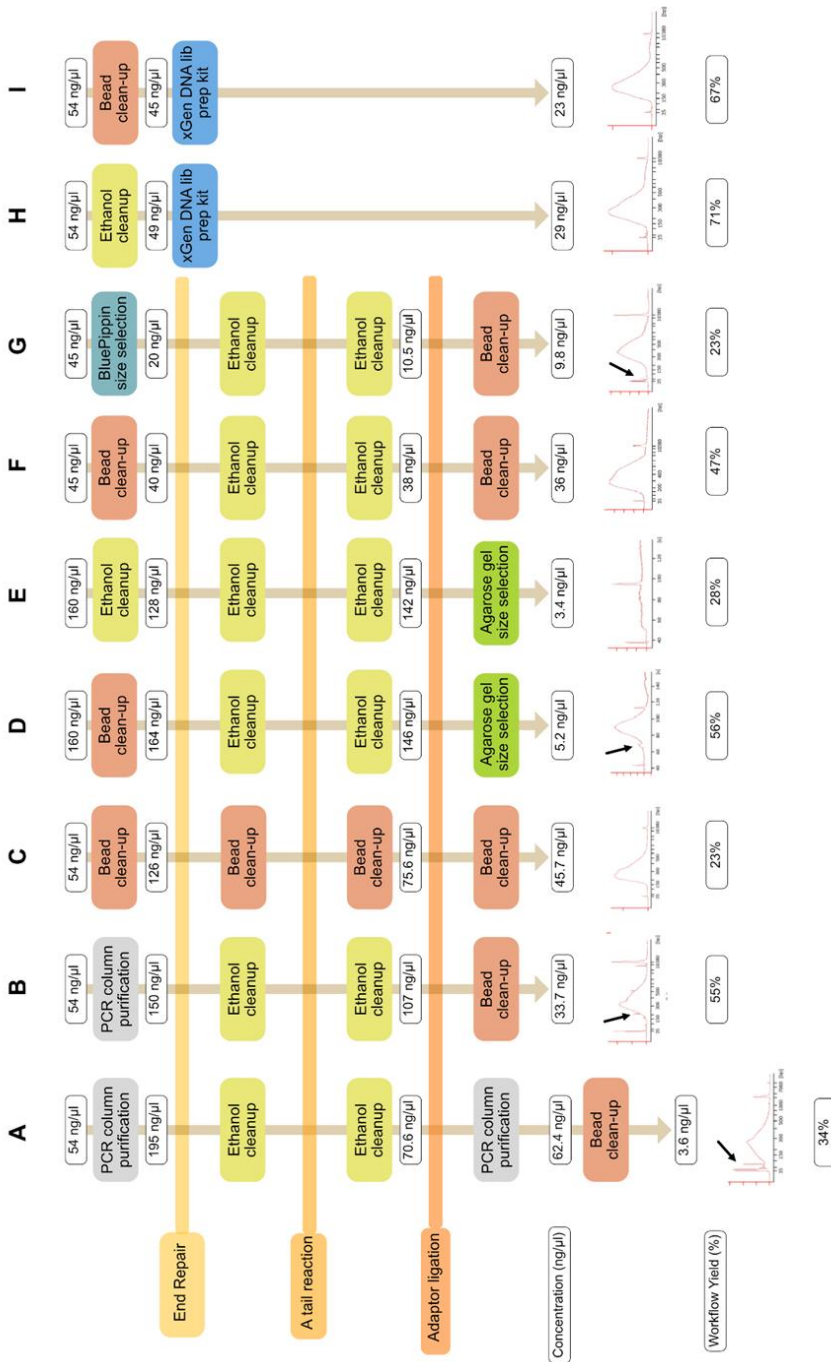


Figure 15. Comparison of DNA library preparation workflows. Each column represents a different workflow and its respective and common protocol steps. Each method was tested at least twice. The DNA concentration was measured at multiple points during the library preparation and average values are shown. A graph from a representative sample from each method, showing the DNA fragment size distribution resulting from each workflow, is shown at the end of each workflow.

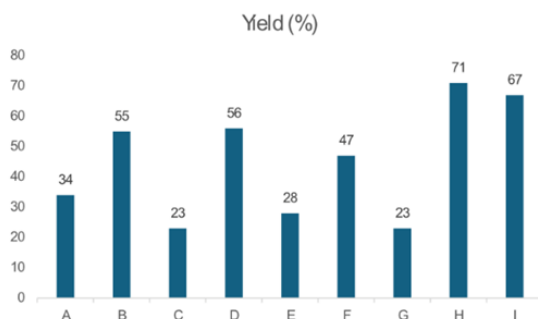


Figure 16. Average DNA library preparation yield of each tested DNA library preparation workflow. A-I correspond to the same workflows described in Figure 17.

A third set of modifications targeted DNA library purity and amplification behaviour. Early trials showed that conventional size-selection strategies failed to remove primer-dimers and small fragments (<170 bp) without also losing the limited pool of TF-bound DNA. We therefore adopted a two-stage amplification strategy: an initial indexing PCR followed by bead clean-up, a qPCR on a sample aliquot to assess quantitatively the optimal number of amplification cycles and a second enrichment step tailored to each sample based on the qPCR results (Fig. 17). This strict size selection and amplification approach substantially reduced primer-dimer carryover and maintained narrow fragment distributions suitable for sequencing.

The combined effect of these adjustments was a substantial improvement in library complexity and purity, enabling DAP-seq to be applied reproducibly to both *P. abies* wood and *P. tremula* wood and leaf

tissue. The modified protocol was validated across dozens of TFs of different families, consistently producing high-quality libraries (Fig. 18) whose sequencing profiles yielded interpretable TF-binding landscapes. These results demonstrate that, with targeted optimisation, DAP-seq can be extended effectively to woody species with large genomes and recalcitrant tissues.

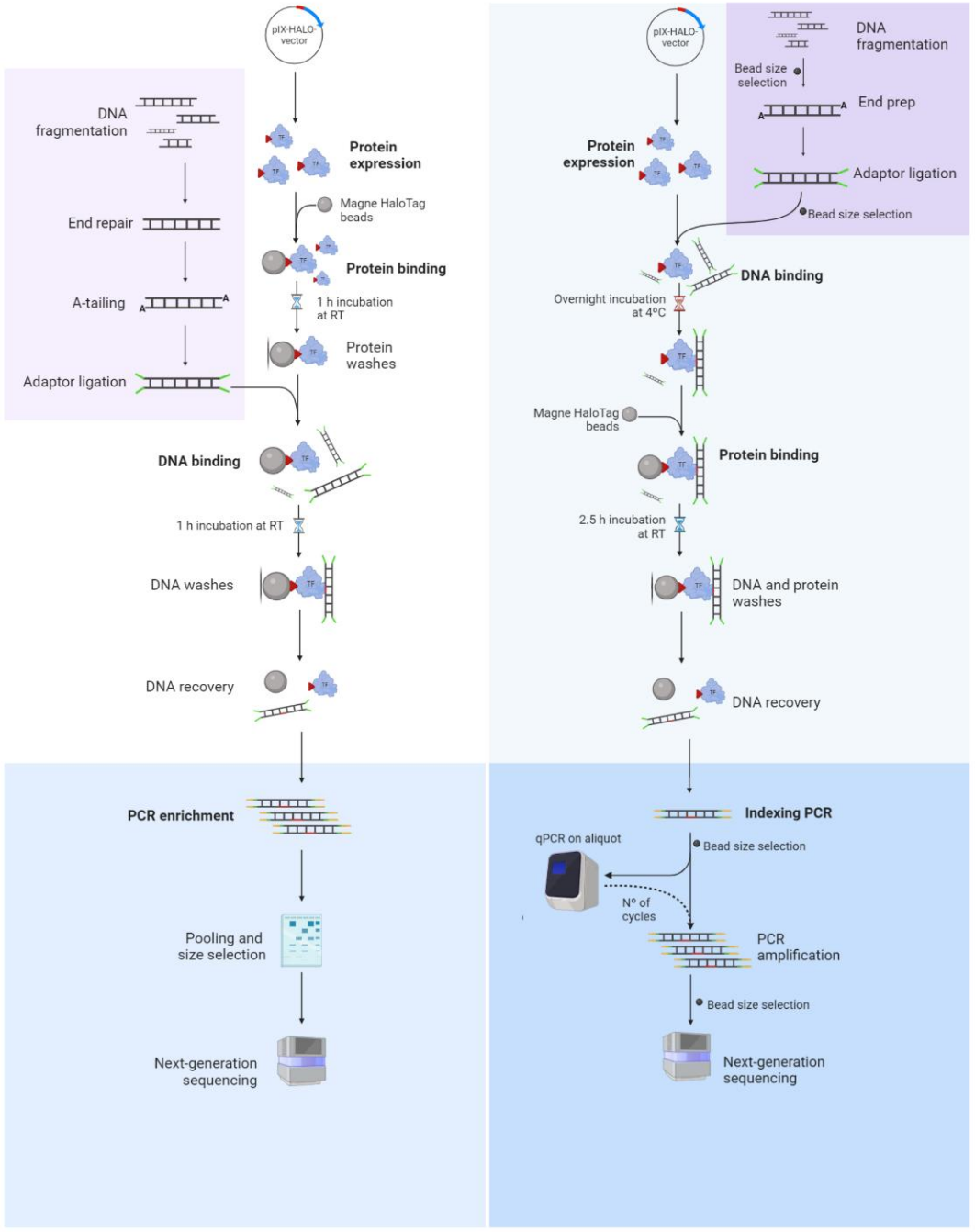


Figure 17. Comparison of original DAP workflow (left) with modified workflow (right)

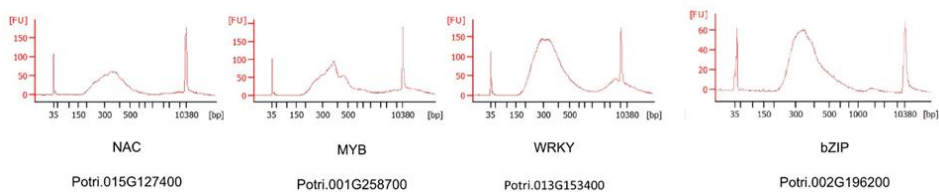
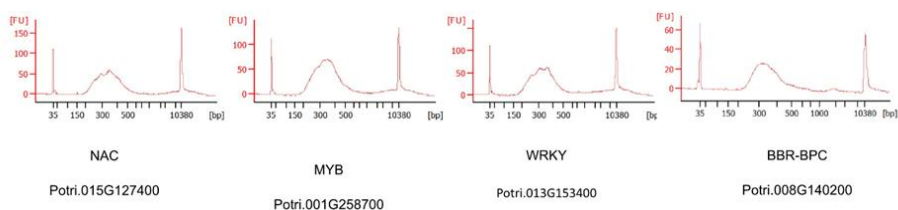
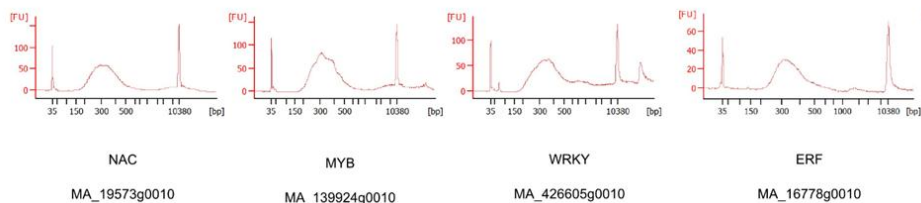
A**B****C**

Figure 18. Examples of sequencing-ready DAP libraries with the modified DAP protocol, prepared from *P. tremula* leaves (Aspen leaf DNA) and wood (Aspen wood DNA) and *P. abies* wood (Norway spruce wood DNA). The final version of the modified protocol has been used in a large-scale project to apply DAP-seq on dozens of different TFs from different families on *P. tremula* and *P. abies* DNA.

Conclusions

This thesis presents four complementary studies that advance our understanding of the regulatory basis of wood formation and provide practical tools and resources for tree genomics. Collectively, the manuscripts deliver synergistic contributions by providing high-resolution evo-devo data and comparative analyses that map conserved and lineage-specific transcriptional programmes across angiosperm and gymnosperm trees. This is further supported by a scalable, optimised protocol for assaying TF-DNA interactions in recalcitrant woody tissues.

Paper I synthesises comparative genomic analyses showing that regulatory changes, gene-family copy-number shifts and pseudogenisation all contribute to growth-form differences between herbs and trees; several candidate regulators and gene families with putative roles in longevity, stress responses and cell wall modification are highlighted for follow-up.

Paper II offers chromosome-scale references, annotations and population resequencing that contextualise regulatory inferences in the repeat-rich, duplication-prone genomes of gymnosperms. These assemblies and population datasets are essential for accurate read mapping, orthogroup assignment and interpretation of TF-binding results in conifers.

Paper III delivers the first high-spatial-resolution evo-devo resource spanning six tree species (three dicots, three conifers), with anatomically matched, 15- μm cryosection series and orthology-aware co-expression networks that reveal both conserved modules (core programmes for cambial activity, cell expansion and SCW formation) and lineage-specific rewiring. These datasets and associated network maps are made available through PlantGenIE and companion portals, enabling community reuse.

Paper IV outlines a set of targeted protocol changes that overcome key obstacles when generating DNA libraries and TF-DNA maps from mature wood: improved DNA extraction from lignified tissues, modified TF-DNA incubation and bead-binding steps, two-stage amplification with strict size selection and bespoke QC and analysis filters. Together, these changes increase library quality and reduce background, making TF-binding profiling more feasible in woody samples.

Together, these contributions create a coherent pipeline from genome to regulatory map to functional hypothesis that future work can build upon. Many of these analysis show the value of anatomically matched, high-spatial-resolution sampling and experimental design, which has enabled the study of developmental transitions (cambium - expansion - SCW formation) that would be invisible to more simple sampling. This approach strengthened comparative network inference and the identification of expressologs across deep phylogenetic distances.

As a whole, the integrated work supports a nuanced view of what makes a tree a tree. Rather than unique “tree genes”, the balance of evidence points to modified regulatory programmes, gene-family dynamics and selective retention/retraction of stress- and longevity-related loci as major drivers of woody, perennial life histories. Conserved co-expression modules could identify core transcriptional circuits for secondary growth that are shared across deep evolutionary splits, while lineage-specific modules and differences in TF neighbourhoods reveal how angiosperms and gymnosperms have tailored these programmes to their distinct anatomies and genomes.

There are several important limitations to the presented studies. Cross-species comparisons of non-coding regulatory sequences remain challenging because of difficult alignment and structural divergence; therefore, many regulatory inferences rely on expression conservation, motif enrichment and TF-binding maps rather than direct base-pair conservation. Paralogs and recent duplications require using orthogroup-aware methods to avoid misattributing regulatory roles. Finally, DAP-seq, even when optimised, analyses TF binding to naked DNA and cannot fully capture chromatin context. Integrating ATAC/ChIP and *in vivo* perturbations remains necessary to establish causality. These limitations highlight how cautious interpretation and follow-up experiments are essential.

Future perspectives

The emergence of single-cell transcriptomics and spatial multi-omics represents a paradigm shift in our ability to resolve the heterogeneous landscapes of the vascular cambium. However, as useful as these

resources are, characterizing the secondary growth system requires moving beyond mere transcriptomics. Integrating genome-wide CRISPR mutagenesis with spatial data is essential to identify the central players in plant biology (Chen et al., 2021; Li et al., 2021). On the other hand, a critical divergence remains between angiosperm models and coniferous species and the scientific community lacks robust, *in vivo* genetic transformation and whole-plant regeneration systems for gymnosperms. This increases the difficulty of functional validation of gymnosperm-specific genes, leaving our understanding of their evolutionary control of wood formation mostly speculative. Overcoming this barrier is not merely a technical necessity but a prerequisite for revolutionizing gymnosperm feedstocks for carbon sequestration and advanced bioproducts (De La Torre et al., 2014; Neale et al., 2017). Nevertheless, moving beyond descriptive transcriptomics toward a comprehensive understanding of regulatory networks will provide the predictive capacity required for the precise metabolic engineering of wood properties across both angiosperm and gymnosperm lineages.

The next frontier in wood biotechnology could lie in the rational engineering and targeted manipulation of the gene regulatory networks. At present, lignocellulosic recalcitrance is the primary barrier to efficient biomass conversion. Overcoming it requires a deep understanding of the crystalline microfibril matrix and monolignol metabolism (Li et al., 2024; Zhu & Li, 2024). The regulatory maps and high-resolution transcriptomic modules established in this thesis can serve as starting blueprints for such investigations, allowing researchers to separate biomass quality from growth rate and precisely re-engineer wood properties for the pulp, paper and biofuel industries.

Furthermore, these new scientific resources can be integrated into global climate strategies. By elucidating the fundamental mechanisms of carbon fixation and vascular differentiation across the angiosperms and gymnosperms, it could be possible to develop high-efficiency "carbon-smart" trees with enhanced sequestration capabilities (Zhu & Li, 2024). Future efforts will increasingly rely on integrative "evo-devo" models to identify conserved regulatory nodes that can be used for sustainable and robust forest production in a changing climate.

Acknowledgments

First and foremost, I wish to express my deepest gratitude to Nathaniel Street. I could not have asked for a more understanding and supportive supervisor. Every concern was answered. Every idea was considered.

I thank Vikash Kumar, my mentor in the laboratory, who taught me more about labwork and techniques than I ever learned in classes and books. I thank Torgeir Hvidsten for his accessibility and consistent support. I'm grateful to Siri Birkeland, for her extraordinary dedication to our projects, particularly for crossing borders to assist with the processing of over 400 samples. My thanks to Sonja Viljamaa for her patience and help in navigating my frequent inquiries, I must have emailed her more than any other colleague. Thanks to Teitur Kalman, for his vital work with the project data and for all the meaningful conversations about life and family.

I express my gratitude to the members of the StreetLab and the UPSC Bioinformatics Facility, present and past, for making me feel warm and inspired in the far, white North.

I give my sincere appreciation to all my colleagues at UPSC and the Swedish University of Agricultural Sciences (SLU). Thank you for cultivating such a collaborative environment.

This research was made possible through the generous financial support of the Research Council of Norway and the Trees and Crops for the Future (TC4F) strategic funding initiative from the Swedish government.

Finally, I thank my family, for helping me become who I am today.

Zeynep'im, eşim ve işğım, tanıştığımızdan beri geçen her gün ve gelecek tüm günlerimiz için teşekkürler.

Gracias, mi Anton, por sonreír, llorar, despertar y nacer.

List of References

- Altenhoff, A.M., Studer, R.A., Robinson-Rechavi, M. & Dessimoz, C., 2012. Resolving the ortholog conjecture: orthologs tend to be more conserved than paralogs. *PLOS Computational Biology*, 8(5), pp.e1002514.
- Arbellay, E., Kuster, T., Fromm, J. & Vollenweider, P., 2014. Methyl jasmonate induces formation of traumatic resin ducts in spruce trees in a dose-dependent manner. *New Phytologist*, 201(4), pp.1209-1220.
- Ausin, I., Feng, S., Yu, C., Liu, W., Kuo, H., Jacobsen, E., Zhai, J., Gallego-Bartolome, J., Wang, L., Egertsdotter, U., Street, N., Jacobsen, S. & Wang, H., 2016. DNA methylome of the 20-gigabase Norway spruce genome. *Proceedings of the National Academy of Sciences*, 113(50), pp.E8106-E8113.
- Badis, G. et al., 2009. Diversity and complexity in DNA recognition by transcription factors. *Science*, 324, pp.1720-1723.
- Bansal, M., Belcastro, V., Ambesi-Impiombato, A. & di Bernardo, D., 2007. How to infer gene networks from expression profiles. *Molecular Systems Biology*, 3, 78.
- Bartlett, A., O'Malley, R.C., Huang, S.S.C., Galli, M., Nery, J.R., Gallavotti, A. & Ecker, J.R., 2017. Mapping genome-wide transcription-factor binding sites using DAP-seq. *Nature Protocols*, 12, pp.1659-1672.
- Van Bel, M., Diels, T., Vancaester, E., Kreft, L., Botzki, A., Van de Peer, Y., Coppens, F. & Vandepoele, K., 2017. PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Research*, 46(D1), pp.D1190-D1196.
- Berglund, J., Agrawal, P.K., Paris, O., Gierlinger, N., Burgert, I. and Salmén, L. (2020). The tensile properties of xylan-cellulose composites in secondary plant cell walls. *Biomacromolecules*, 21, 418-427.
- Bernal-Gallardo, E. & de Folter, S., 2024. Plant genome sequencing: current status and future prospects. *Plant Genome*, 17(1), pp.1-12.

Bernhardsson, C., Vidalis, A., Wang, X., Scofield, D., Schiffthaler, B., Baison, J., Street, N., García-Gil, M. & Ingvarsson, P., 2019. An ultra-dense haploid genetic map for evaluating the highly fragmented genome assembly of Norway spruce (*P. abies*). *G3: Genes|Genomes|Genetics*, 9(5), pp.1623-1632.

Bhalerao, R.P. & Fischer, U., 2016.b. Environmental and hormonal control of cambial stem cell dynamics. *Journal of Experimental Botany*, 68(1), pp.79-87.

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

Blokhina, O., Laitinen, T., Hatakeyama, Y., Delhomme, N., Paasela, T., Zhao, L., Street, N., Wada, H., Kärkönen, A. & Fagerstedt, K., 2019. Ray parenchymal cells contribute to lignification of tracheids in developing xylem of Norway spruce. *Plant Physiology*, 181(4), pp.1552-1572.

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

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

Bonawitz, N.D. & Chapple, C., 2010. The genetics of lignin biosynthesis: connecting genotype to phenotype. *Annual Review of Genetics*, 44, pp.337-363.

Brown, K., Takawira, L.T., O'Neill, M.M., Mizrachi, E., Myburg, A.A. & Hussey, S.G., 2019. Identification and functional evaluation of accessible chromatin associated with wood formation in *Eucalyptus grandis*. *New Phytologist*, 223, pp.1937-1951.

Busov, V., Strauss, S. & Pilate, G., 2009. Transformation as a tool for genetic analysis in *Populus*. *Genetics and Genomics of Populus*, pp.113-133.

Butte, A.J. & Kohane, I.S., 1999. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing*, 5, pp.418-429.

Carlquist, S., 1988. *Comparative Wood Anatomy: Systematic, Ecological and Evolutionary Aspects of Dicotyledon Wood*. Springer-Verlag, Berlin.

Carlsson-Granér, U. & Thrall, P.H., 2006. The impact of host longevity on disease transmission: host-pathogen dynamics and the evolution of resistance. *Evolutionary Ecology Research*, 8, pp.659-675.

Carpita, N.C. & Gibeaut, D.M., 1993. Structural models of primary cell walls in flowering plants: consistency of molecular structure with the physical properties of the walls during growth. *The Plant Journal*, 3, pp.1-30.

Carroll, S.B., 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell*, 134(1), pp.25-36. doi: 10.1016/j.cell.2008.06.030.

Celedon, J., Yuen, M., Chiang, A., Henderson, H., Reid, K. & Bohlmann, J., 2017. Cell-type- and tissue-specific transcriptomes of the white spruce (*Picea glauca*) bark unmask fine-scale spatial patterns of constitutive and induced conifer defence. *The Plant Journal*, 92(4), pp.710-726.

Celedon, J.M. et al., 2017. Cell biology and biochemistry of conifer resin ducts. *New Phytologist*, 216(3), pp.709-724.

Cerise, M., Giaume, F., Galli, M., Khahani, B., Lucas, J., Podico, F., Tavakol, E., Parcy, F., Gallavotti, A., Brambilla, V. & Fornara, F., 2021. OsFD4 promotes the rice floral transition via florigen activation complex formation in the shoot apical meristem. *New Phytologist*, 229, pp.429-443.

Chen, F. & Dixon, R.A., 2007. Lignin modification improves fermentable sugar yields for biofuel production. *Nature Biotechnology*, 25, pp.759-761.

Chen, H., Li, Y., Wang, J.P., et al., 2019. Hierarchical transcription factor and chromatin binding network for wood formation. *Plant Cell*, 31, pp.602-626.

Chen, P., Yan, M., Li, L., He, J., Zhou, S., Li, Z., Niu, C., Bao, C., Zhi, F., Ma, F. & Guan, Q., 2020. The apple DNA-binding one zinc-finger protein MdDof54 promotes drought resistance. *Horticulture Research*, 7, pp.1-14.

Clark, J.W. & Donoghue, P.C.J., 2018. Whole-genome duplication and plant macroevolution. *Trends in Plant Science*, 23, pp.933-945.

De Clercq, I., Van de Velde, J., Luo, X., Liu, L., Storme, V., Pottier, R., Vanechoutte, D., Van Breusegem, F. & Vandepoele, K., 2020. Integrative inference of transcriptional networks in *A. thaliana* yields novel regulators involved in reactive oxygen species stress signaling. *The Plant Journal*, 102(4), pp.936-960.

ENCODE Project Consortium, 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489, 7414., pp.57-74.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258), 561-563.

Cullingham, C.I., James, P.M.A., Cooke, J.E.K. & Coltman, D.W., 2012. Characterizing the physical and genetic structure of the lodgepole pine × jack pine hybrid zone: mosaic structure and differential introgression. *Evolutionary Applications*, 5(8), pp.879-891.

Dauphin, B.G., Voiniciuc, C., Potocký, M., Wessel, G.M. & Mravec, J., 2024. TBL38 atypical homogalacturonan-acetyltransferase activity and cell-wall microdomain localization in *A. thaliana* seed mucilage secretory cells. *iScience*, 27, p.109666.

Dillen, S., Rood, S. & Ceulemans, R., 2009. Growth and physiology. *Genetics and Genomics of Populus*, pp.39-63.

Doblin, M.S., Kurek, I., Jacob-Wilk, D. & Delmer, D.P., 2002. Cellulose biosynthesis in plants: from genes to rosettes. *Plant and Cell Physiology*, 43, pp.1407-1420.

Dobrowolska, I., Businge, E., Abreu, I., Moritz, T. & Egertsdotter, U., 2017. Metabolome and transcriptome profiling reveal new insights into somatic embryo germination in Norway spruce (*P. abies*). *Tree Physiology*, 37(12), pp.1752-1766.

- D'haeseleer, P., Liang, S. & Somogyi, R., 2000. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16(8), pp.707-726.
- Elfving, B., Ericsson, T. & Rosvall, O., 2001. The introduced lodgepole pine (*Pinus contorta* var. *latifolia*) in Sweden and its comparison with native Scots pine (*Pinus sylvestris*). *Forest Ecology and Management*, 141(3), pp.203-216.
- Ellis, B., Jansson, S., Strauss, S. & Tuskan, G., 2009. Why and how *Populus* became a “model tree”. *Genetics and Genomics of Populus*, pp.3-14.
- Evert, R.F., 2006. *Esau's Plant Anatomy: Meristems, Cells and Tissues of the Plant Body: Their Structure, Function and Development*. 3rd edn. Hoboken, NJ: Wiley.
- Feild, T.S. & Holbrook, N.M., 2000. Tracheary element structure and function in Winteraceae. *International Journal of Plant Sciences*, 161(5), pp.805-812.
- Fergus, B.J. & Goring, D.A.I., 1970. The distribution of lignin in birch wood as determined by ultraviolet microscopy. *Holzforschung*, 24, pp.118-124.
- Figures 1, 4, 5 and 17 were created using BioRender.com.
- Fischer, U., Kucukoglu, M., Helariutta, Y. & Bhalerao, R.P., 2019. The dynamics of cambial stem cell activity. *Annual Review of Plant Biology*, 70, pp.293-319.
- Fromm, J., 2013. Xylem development in trees: from cambial divisions to mature wood cells. *Plant Cell Monographs*, pp.3-39.
- Fromm, J., 2013. Xylem structure and function. *Annual Plant Reviews*, 40, pp.1-30.
- Galli, M., Khakhar, A., Lu, Z., Chen, Z., Sen, S., Joshi, T., Nemhauser, J.L., Schmitz, R.J. & Gallavotti, A., 2018. The DNA binding landscape of the maize AUXIN RESPONSE FACTOR family. *Nature Communications*, 9, pp.1-11.

- Gouwentak, C.A., 1941. Growth phenomena in stems of woody plants. *Recueil des Travaux Botaniques Néerlandais*, 38, pp.1-60.
- Groover, A., Nieminen, K., Helariutta, Y. & Mansfield, S., 2009. Wood formation in *Populus*. *Genetics and Genomics of Populus*, pp.201-224.
- Groover, A.T., 2005. What genes make a tree a tree? *Trends in Plant Science*, 10(5), pp.210-214.
- Hager, A., 2003. Role of the plasma membrane H⁺-ATPase in auxin-induced elongation growth: historical and new aspects. *Journal of Plant Research*, 116, pp.483-505.
- Hashimshony, T., Zhang, J., Keshet, I., Bustin, M. & Cedar, H., 2003. The role of DNA methylation in setting up chromatin structure during development. *Nature Genetics*, 34(2), pp.187-192.
- He, X. & Zhang, J., 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*, 169(2), pp.1157-1164.
- Hidalgo, E., González-Martínez, S., Lexer, C. & Heinze, B., 2009. Conservation genomics. *Genetics and Genomics of Populus*, pp.349-368.
- Hill, J.L., Hammudi, M.B. & Tien, M., 2014. The Arabidopsis cellulose synthase complex: a proposed hexamer of CESA trimers in an equimolar stoichiometry. *Plant Cell*, 26, pp.4834-4842.
- Ietswaart, R., Gyori, B.M., Bachman, J.A., Sorger, P.K. & Churchman, L.S., 2021. GeneWalk identifies relevant gene functions for a biological context using network representation learning. *Genome Biology*, 22, 55.
- Immanen, J., Nieminen, K., Smolander, O.P., Kojima, M., Alonso Serra, J., Koskinen, P., Zhang, J., Elo, A., Mähönen, A.P., Street, N., Bhalerao, R.P. and Helariutta, Y. (2016) 'Cytokinin and auxin display distinct but interconnected distribution and signaling profiles to stimulate cambial activity', *Current Biology*, 26(15), pp. 1990-1997.
- Ingvarsson, P., 2009. Nucleotide polymorphism, linkage disequilibrium and complex trait dissection in *Populus*. *Genetics and Genomics of Populus*, pp.91-111.

Ingvarsson, P.K. & Street, N.R., 2011. Association genetics of complex traits in plants. *New Phytologist*, 189(4), pp.909-922.

One Thousand Plant Transcriptomes Initiative, 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, 574, pp.679-685.

Israelsson, M., Sundberg, B. & Moritz, T., 2005. Tissue-specific localization of gibberellins and expression of gibberellin-biosynthetic and signalling genes in wood-forming tissues in aspen. *The Plant Journal*, 44(3), pp.494-504.

Jaenisch, R. & Bird, A., 2003. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics*, 33, pp.245-254.

Jang, M.-J. et al., 2024. Haplotype-resolved genome assembly and resequencing analysis provide insights into genome evolution and allelic imbalance in *Pinus densiflora*. *Nature Genetics*, pp.1-11.

Jokipii-Lukkari, S., Delhomme, N., Schiffthaler, B., Mannapperuma, C., Prestele, J., Nilsson, O., Street, N. & Tuominen, H., 2018. Transcriptional roadmap to seasonal variation in wood formation of Norway spruce. *Plant Physiology*, 176(4), pp.2851-2870.

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

Jouffroy, O., Saha, S., Mueller, L., Quesneville, H. & Maumus, F., 2016. Comprehensive repeatome annotation reveals strong potential impact of repetitive elements on tomato ripening. *BMC Genomics*, 17.

Kalman, T.A., Van de Peer, Y., Street, N.R., Scofield, D.G., Nystedt, B., Ingvarsson, P.K., Zubair, M. & Degnan, B.M., 2025. 1000 conifer genomes: genome innovation, organisation and diversity. *Research Square*. doi:10.21203/rs.3.rs-6502828/v1.

Kampe, A. & Magel, E., 2013. New insights into heartwood and heartwood formation. *Plant Cell Monographs*, pp.71-95.

Kastally, C., Niskanen, A.K., Perry, A., Kujala, S.T., Avia, K., Cervantes, S., Haapanen, M., Kesälahti, R., Kumpula, T.A., Mattila, T.M., Ojeda, D.I., Tyrmi, J.S., Wachowiak, W., Cavers, S., Kärkkäinen, K., Savolainen, O. & Pyhäjärvi, T., 2022. Taming the massive genome of Scots pine with PiSy50k, a new genotyping array for conifer research. *The Plant Journal*, 109(5), pp.1337-1350.

Kirui, A., Dickwella Widanage, M.C., Mentink-Vigier, F., et al., 2022. Interaction of lignin with xylan and cellulose in secondary cell walls of gymnosperms. *Plant Physiology*, 188, pp.1031-1046.

Klevebring, D., Street, N., Fahlgren, N., Kasschau, K., Carrington, J., Lundeberg, J. & Jansson, S., 2009. Genome-wide profiling of *Populus* small RNAs. *BMC Genomics*, 10(1), p.620.

Knoblauch, M. & Oparka, K., 2012. The structure of the phloem—still more questions than answers. *New Phytologist*, 195(3), pp.541-542.

Ko, J.-H., Jeon, H.-W., Kim, W.-C., Kim, J.-Y. & Han, K.-H., 2014. The MYB46/MYB83-mediated transcriptional regulatory programme is a gatekeeper of secondary wall biosynthesis. *Annals of Botany*, 114, pp.1099-1107.

Koch, G. & Schmitt, U., 2013. Topochemical and electron microscopic analyses on the lignification of individual cell wall layers during wood formation and secondary changes. *Plant Cell Monographs*, pp.41-69.

Kollmann, F.F.P., Kuenzi, E.W. & Stamm, A.J., 2013. *Principles of Wood Science and Technology*. Berlin: Springer.

Koonin, E. V. (2005). Orthologs, paralogs and evolutionary genomics. *Annual Review of Genetics*, 39, 309-338.

Koutaniemi, S., Malmberg, H., Simola, L., Teeri, T. & Kärkönen, A., 2015. Norway spruce (*P. abies*) laccases: characterization of a laccase in a lignin-forming tissue culture. *Journal of Integrative Plant Biology*, 57(4), pp.341-348.

Kumar, S. et al., 2022. TimeTree 5: an expanded resource for species divergence times. *Molecular Biology and Evolution*, 39.

Kumar, V., Hainaut, M., Delhomme, N., Mannapperuma, C., Immerzeel, P., Street, N., Henrissat, B. & Mellerowicz, E., 2019. *Poplar*

carbohydrate-active enzymes: whole-genome annotation and functional analyses based on RNA expression data. *The Plant Journal*, 99(4), pp.589-609.

Lamara, M., Raheison, E., Lenz, P., Beaulieu, J., Bousquet, J. & MacKay, J., 2015. Genetic architecture of wood properties based on association analysis and co-expression networks in white spruce. *New Phytologist*, 210(1), pp.240-255.

Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P. et al., 2012. CHIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*, 22, pp.1813-1831.

Larson, P.R., 1994. *The Vascular Cambium: Development and Structure*. Berlin: Springer.

Lee, C., Teng, Q., Zhong, R. & Ye, Z.-H., 2011. Molecular dissection of xylan biosynthesis during wood formation in poplar. *Molecular Plant*, 4, pp.730-747.

Lei, Q. et al., 2015. The FOUR LIPS and MYB88 transcription factor genes are widely expressed in *A. thaliana* during development. *American Journal of Botany*, 102, pp.1521-1528.

Li, L., Lu, S. & Chiang, V., 2006. A genomic and molecular view of wood formation. *Critical Reviews in Plant Sciences*, 25, pp.215-233.

Li, P., He, Y., Xiao, L., Quan, M., Gu, M., Jin, Z., Zhou, J., Li, L., Bo, W., Qi, W., Huang, R., Lv, C., Wang, D., Liu, Q., El-Kassaby, Y.A., Du, Q. & Zhang, D., 2024. Temporal dynamics of genetic architecture governing leaf development in *Populus*. *New Phytologist*, 242(3), pp.1113-1130.

Li, W., Lin, Y.-C.J., Chen, Y.-L., Zhou, C., Li, S., De Ridder, N., Oliveira, D.M., Zhang, L., Zhang, B., Wang, J.P., Xu, C., Fu, X., Luo, K., Wu, A.-M., Demura, T., Lu, M.-Z., Zhou, Y., Li, L., Umezawa, T., Boerjan, W. & Chiang, V.L., 2024. Woody plant cell walls: Fundamentals and utilization. *Molecular Plant*. doi:10.1016/j.molp.2023.12.008.

Li, X., Weng, J.-K. & Chapple, C., 2010. Improvement of biomass through lignin modification. *The Plant Journal*, 54, pp.569-581.

Liesecke, F., De Craene, J.-O., Besseau, S., Courdavault, V., Clastre, M., Vergès, V., Papon, N., Giglioli-Guivarc'h, N., Glévarec, G., Pichon, O. & Dugé de Bernonville, T., 2019. Improved gene co-expression network quality through expression dataset down-sampling and network aggregation. *Scientific Reports*, 9, 1443.1.

Lin, Y., Wang, J., Delhomme, N., Schiffthaler, B., Sundström, G., Zuccolo, A., Nystedt, B., Hvidsten, T., de la Torre, A., Cossu, R., Hoepfner, M., Lantz, H., Scofield, D., Zamani, N., Johansson, A., Mannapperuma, C., Robinson, K., Mähler, N., Leitch, I., Pellicer, J., Park, E., Van Montagu, M., Van de Peer, Y., Grabherr, M., Jansson, S., Ingvarsson, P. & Street, N., 2018. Functional and evolutionary genomic inferences in *Populus* through genome and population sequencing of American and European aspen. *Proceedings of the National Academy of Sciences*, 115(46), pp.E10970-E10978.

Lin, Y.-C. J., Chen, Y.-L., McCarthy, R. L., et al. (2013). Functional characterization of a hierarchical transcriptional regulatory network for wood formation. *Proceedings of the National Academy of Sciences USA*, 110, 13829-13834.

Liu, L. et al., 2021. Enhancing grain-yield-related traits by CRISPR-Cas9 promoter editing of maize CLE genes. *Nature Plants*, 7, pp.287-294.

Liu, L., Ramsay, T., Zinkgraf, M., Sundell, D., Street, N., Filkov, V. & Groover, A., 2015. A resource for characterizing genome-wide binding and putative target genes of transcription factors expressed during secondary growth and wood formation in *Populus*. *The Plant Journal*, 82(5), pp.887-898.

Liu, Z., Suarez Duran, H., Harnvanichvech, Y., Stephenson, M., Schranz, M., Nelson, D., Medema, M. & Osbourn, A., 2019. Drivers of metabolic diversification: how dynamic genomic neighbourhoods generate new biosynthetic pathways in the Brassicaceae. *New Phytologist*.

Lloyd, J.P.B. & Lister, C., 2022. Epigenome plasticity in plants. *The Plant Journal*, 110

Lu, N., Zhu, T., Ouyang, F., Xia, Y., Li, Q., Jia, Z., Hu, J., Ling, J., Ma, W., Yang, G., Zhang, H., Kong, L. & Wang, J., 2019. PICEA database: a web database for *Picea* omics and phenotypic information. *Database*, 2019.

Lundmark, T., Bergh, J., Hofer, P., Lundström, A., Nordin, A., Poudel, B. C., Sathre, R., Taverna, R., & Werner, F. (2014). Potential of Environmental and Economic Co-benefits from Forest Management and Wood Use. *Ambios*, 43(Suppl 1), 82-95.

Luo, A., Xu, X., Liu, Y., Li, Y., Su, X., Li, Y., Lyu, T., Dimitrov, D., Larjavaara, M., Peng, S., Wang, Q., Zimmermann, N.E., Pellissier, L. & Wang, Z., 2023. Spatio-temporal patterns in the woodiness of flowering plants. *Global Ecology and Biogeography*, 32, pp.384-396. doi:10.1111/geb.13627.

Luo, K. & Li, L., 2022. Regulatory networks controlling secondary growth and wood formation. *Current Opinion in Plant Biology*, 65, pp.102138.

Luo, L. & Li, L., 2022. Molecular understanding of wood formation in trees. *Forestry Research*, 2, pp.5.

Luo, L., Zhu, Y. & Gui, J. et al., 2021. A comparative analysis of transcription networks active in juvenile and mature wood in *Populus*. *Frontiers in Plant Science*, 12, pp.675075.

Maher, K.A., Bajic, M., Kajala, K., Reynoso, M., Pauluzzi, G., West, D.A., Zumstein, K., Woodhouse, M., Bubb, K.L., Dorrity, M.W., Queitsch, C., Sinha, N., Brady, S.M. & Stitzer, M.C., 2018. Profiling of accessible chromatin regions across multiple plant species and cell types reveals common gene regulatory principles and new control modules. *Plant Cell*, 30, pp.15-36.

Malone, J.H. & Oliver, B., 2011. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biology*, 9, p.34.

Marquínez, X., Lohmann, L.G., Falslev, S. & Raubeson, L.A., 2009. Wood anatomy and the phylogenetic position of the vessel-less angiosperm family Winteraceae. *American Journal of Botany*, 96(8), pp.1541-1555.

Martin, D., Gershenzon, J. & Bohlmann, J., 2002. Methyl jasmonate induces traumatic resin ducts and activates terpene biosynthesis in spruce. *Plant Physiology*, 129(3), pp.1003-1018.

- Martin, D.M., Gershenzon, J. & Bohlmann, J., 2002. Induction of volatile terpene biosynthesis and resin duct formation by methyl jasmonate in Sitka spruce. *Plant Physiology*, 129(3), pp.1003-1018.
- McCarthy, R.L., Zhong, R. & Fowler, S. et al., 2010. The poplar MYB transcription factors, PtrMYB3 and PtrMYB20, are involved in the regulation of secondary wall biosynthesis. *Plant Cell Physiology*, 51, pp.1084-1090.
- McLeay, R., Lesluyes, T., Cuellar Partida, G. & Bailey, T., 2012. Genome-wide in silico prediction of gene expression. *Bioinformatics*, 28(21), pp.2789-2796.
- Mellerowicz, E.J. & Sundberg, B., 2008. Wood cell walls: biosynthesis, developmental dynamics and their implications for wood properties. *Current Opinion in Plant Biology*, 11, pp.293-300.
- Mitsuda, N. & Ohme-Takagi, M., 2008. NAC transcription factors NST1 and NST3 regulate pod shattering in a partially redundant manner by promoting secondary wall formation after the establishment of tissue identity. *The Plant Journal*, 56, pp.768-778.
- Mitsuda, N., Seki, M., Shinozaki, K. & Ohme-Takagi, M., 2005. The NAC transcription factors NST1 and NST2 of Arabidopsis regulate secondary wall thickenings and are required for anther dehiscence. *The Plant Cell*, 17, pp.2993-3006.
- Movahedi, S., Van Bel, M., Heyndrickx, K.S. & Vandepoele, K., 2011. Comparative network analysis reveals that tissue specificity and gene function are important factors influencing the mode of expression evolution in *A. thaliana* and rice. *Plant Physiology*, 156, pp.1316-1330.
- Mutwil, M., Debolt, S. & Persson, S., 2008. Cellulose synthesis: a complex complex. *Current Opinion in Plant Biology*, 11(3), pp.252-257.
- Müller, G.B., 2007. Evo-devo: extending the evolutionary synthesis. *Nature Reviews Genetics*, 8(12), pp.943-949. doi: 10.1038/nrg2219.
- Nakaba, S., Begum, S., Yamagishi, Y., Jin, H., Kubo, T. & Funada, R., 2011. Differences in the timing of cell death, differentiation and function among three different types of ray parenchyma cells in the hardwood *Populus sieboldii* × *P. grandidentata*. *Trees*, 26(3), pp.743-750.

- Nakano, Y., Yamaguchi, M., Endo, H., Rejab, N.A. & Ohtani, M., 2015. NAC-MYB-based transcriptional regulation of secondary cell wall biosynthesis in land plants. *Frontiers in Plant Science*, 6, p.288.
- Neale, D.B., Martínez-García, P.J., De La Torre, A.R., Montanari, S. & Wei, X.-X., 2017. Novel insights into tree biology and genome evolution as revealed through genomics. *Annual Review of Plant Biology*, 68, pp.457-483.
- Neale, D.B., Martínez-García, P.J., De La Torre, A.R., Montanari, S. & Wei, X.-X., 2017. Pine genomics and genetics. In: *Genetics and Genomics of Conifers*. Springer, Cham, pp.107-124.
- Netotea, S., Sundell, D., Street, N.R. & Hvidsten, T.R., 2014. ComPlex: conservation and divergence of co-expression networks in *A. thaliana*, *Populus* and *Oryza sativa*. *BMC Genomics*, 15, p.106.
- Nieminen, K., Immanen, J., Laxell, M., Kauppinen, L., Tarkowski, P., Doležal, K., Tähtiharju, S., Elo, A., Decourteix, M., Ljung, K. & Helariutta, Y., 2008. Cytokinin signaling regulates cambial development in poplar. *Proceedings of the National Academy of Sciences USA*, 105, pp.20032-20037.
- Nilsson, J., Karlberg, A., Antti, H., Lopez-Vernaza, M., Mellerowicz, E., Perrot-Rechenmann, C., Sandberg, G. & Bhalerao, R.P., 2008. Dissecting the molecular basis of the regulation of wood formation by auxin in hybrid aspen. *The Plant Cell*, 20(4), pp.843-855.
- Niu, S. et al., 2021. The Chinese pine genome and methylome unveil key features of conifer evolution. *Cell*, pp.1-14.
- Nixon, B.T., Mansouri, K., Singh, A., Du, J., Davis, J.K., Lee, J.-G., Slabaugh, E., Vandavasi, V.G., O'Neill, H., Roberts, E.M., et al., 2016. Comparative structural and computational analysis supports eighteen cellulose synthases in the plant cellulose synthesis complex. *Scientific Reports*, 6, pp.28696.
- Nystedt, B., Street, N.R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D.G., Vezzi, F., Delhomme, N., Giacomello, S., Alexeyenko, A. et al., 2013. The Norway spruce genome sequence and conifer genome evolution. *Nature*, 497, 7451., pp.579-584.

- Olano, J.M., Davis, S.J. & Mann, D.G., 2013. New Phytologist commentary: New star on the stage: amount of ray parenchyma in tree rings. *New Phytologist*, 2013.
- Olsson, S., Pinosio, S., González-Martínez, S., Abascal, F., Mayol, M., Grivet, D. & Vendramin, G., 2018. De novo assembly of English yew (*Taxus baccata*) transcriptome and its applications for intra- and inter-specific analyses. *Plant Molecular Biology*, 97(4-5), pp.337-345.
- O'Malley, R.C., Huang, S.S.C., Song, L., Lewsey, M.G., Bartlett, A., Nery, J.R., Galli, M., Gallavotti, A. & Ecker, J.R., 2016. Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell*, 165, pp.1280-1292.
- Pai, A.A., Pritchard, J.K. & Gilad, Y., 2015. The genetic and mechanistic basis for variation in gene regulation. *PLoS Genetics*, 11, e1004857.
- Pal, K., Forcato, M. & Ferrari, F., 2018. Hi-C analysis: from data generation to integration. *Biophysical Reviews*, 11(1), pp.67-78.
- Pawar, P.M.-A., Koutaniemi, S., Tenkanen, M. and Mellerowicz, E.J. (2017). Acetylation of woody cell wall polysaccharides. *Plant Physiology*, 173, 1080-1093.
- Pesquet, E., Zhang, B., Gorzsás, A., Puhakainen, T., Serk, H., Escamez, S., Barbier, O., Gerber, L., Courtois-Moreau, C., Alatalo, E., Paulin, L., Kangasjärvi, J., Sundberg, B., Goffner, D. & Tuominen, H., 2013. Non-cell-autonomous postmortem lignification of tracheary elements in *Zinnia elegans*. *The Plant Cell*, 25(4), pp.1314-1328.
- Plomion, C., Leprovost, G. & Stokes, A., 2001. Wood formation in trees. *Plant Physiology*, 127(4), pp.1513-1523.
- Porebski, S., Bailey, L.G. & Baum, B.R., 1997. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Molecular Biology Reporter*, 15, pp.8-15.
- Proost, S. et al., 2009. PLAZA: a comparative genomics resource to study gene and genome evolution in plants. *Plant Cell*, 21(12), pp.3718-3731.
- Proost, S. et al., 2015. PLAZA 3.0: an access point for plant comparative genomics. *Nucleic Acids Research*, 43(D1), pp.D974-D981.

- Provero, P., 2002. Essentiality and centrality in protein interaction networks: a graph-theoretic analysis. *BMC Bioinformatics*, 3, p.8.
- Quesneville, H., 2020. Twenty years of transposable element analysis in the *A. thaliana* genome. *Mobile DNA*, 11, p.28.
- Ramos-Sánchez, J.M., Triozzi, P.M., Alique, D., Geng, F., Gao, M., Jaeger, K.E., Wigge, P.A., Allona, I. & Perales, M., 2019. LHY2 integrates night-length information to determine timing of poplar photoperiodic growth. *Current Biology*, 29, pp.2402-2406.e4.
- Rao, X. & Dixon, R.A., 2019. Co-expression networks for plant biology: why and how. *Acta Biochimica et Biophysica Sinica*, 51(10), pp.981-988.
- Ricci, W.A., Lu, Z., Ji, L., Marand, A.P., Ethridge, C.L., Murphy, N.G., Noshay, J.M., Galli, M., Mejía-Guerra, M.K., Colomé-Tatché, M., Johannes, F., Rowley, M.J., Corces, V.G., Zhai, J., Scanlon, M.J., Buckler, E.S., Gallavotti, A., Springer, N.M., Schmitz, R.J. & Zhang, X., 2019. Widespread long-range cis-regulatory elements in the maize genome. *Nature Plants*, 5, pp.1237-1249.
- Romero, I.G., Ruvinsky, I. & Gilad, Y., 2012. Comparative studies of gene expression and the evolution of gene regulation. *Nature Reviews Genetics*, 13, pp.505-516.
- Salmén, L., 2022. Wood cell wall structure and organisation in relation to mechanics. *Cellulose*, 29, pp.681-701.
- Salojärvi, J., Smolander, O.P., Nieminen, K. et al., 2017. Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. *Nature Genetics*, 49, pp.904-912.
- Sanger, F., Nicklen, S. & Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the USA*, 74(12), pp.5463-5467.
- Scheller, H.V. & Ulvskov, P., 2010. Hemicelluloses. *Annual Review of Plant Biology*, 61, pp.263-289.
- De Schepper, V., De Swaef, T., Vandegheuchte, M. & Steppe, K., 2013. Phloem transport: a review of mechanisms and controls. *Journal of Experimental Botany*, 64(16), pp.4839-4850.

- Serin, E.A.R., Nijveen, H., Hilhorst, H.W.M. & Ligterink, W., 2016. Learning from co-expression networks: possibilities and challenges. *Frontiers in Plant Science*, 7, p.444.
- Shahan, R., Hsu, C.-W., Nolan, T.M., Cole, B.J., Taylor, I.W., Ow, L.W., Rhee, S.Y. & Brady, S.M., 2020. A single-cell *A. thaliana* root atlas reveals developmental trajectories in wild-type and cell identity mutants. *Developmental Cell*, 55, pp.781-799.
- Shirasawa, K., Isuzugawa, K., Ikenaga, M., Saito, Y., Yamamoto, T., Hirakawa, H. & Isobe, S., 2017. The genome sequence of sweet cherry (*Prunus avium*) for use in genomics-assisted breeding. *DNA Research*, 24(5), pp.499-508.
- Slavov, G. & Zhelev, P., 2009. Salient biological features, systematics and genetic variation of *Populus*. *Genetics and Genomics of Populus*, pp.15-38.
- Soneson, C. & Delorenzi, M., 2013. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, 14, 1.
- Spitz, F. & Furlong, E.E.M., 2012. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*, 13(9), pp.613-626.
- Street, N. & Tsai, C., 2009. *Populus* resources and bioinformatics. *Genetics and Genomics of Populus*, pp.135-152.
- Street, N., Jansson, S. & Hvidsten, T., 2011. A systems biology model of the regulatory network in *Populus* leaves reveals interacting regulators and conserved regulation. *BMC Plant Biology*, 11, p.13.
- Street, N., Skogström, O., Sjödin, A., Tucker, J., Rodríguez-Acosta, M., Nilsson, P., Jansson, S. & Taylor, G., 2006. The genetics and genomics of the drought response in *Populus*. *The Plant Journal*, 48(3), pp.321-341.
- Stuart, T. et al., 2016. Population-scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. *eLife*, 5, e20777.
- Sundell, D., Street, N.R., Kumar, M., Mellerowicz, E., Kucukoglu, M., Johnsson, C., Kumar, V., Mannapperuma, C., Delhomme, N., Nilsson,

O., Tuominen, H., Pesquet, E., Fischer, U., Niittylä, T., Sundberg, B. & Hvidsten, T., 2017. AspWood: high-spatial-resolution transcriptome profiles reveal uncharacterized modularity of wood formation in *P. tremula*. *The Plant Cell*, 29(7), pp.1585-1604.

Sundell, D., Mannapperuma, C., Netotea, S., Delhomme, N., Lin, Y., Sjödin, A., Van de Peer, Y., Jansson, S., Hvidsten, T. & Street, N., 2015. The Plant Genome Integrative Explorer Resource: PlantGenIE.org. *New Phytologist*, 208(4), pp.1149-1156.

Suzuki, S., Li, L., Sun, Y.-H. & Chiang, V.L., 2006. The cellulose synthase gene superfamily and biochemical functions of its members are conserved between angiosperms and gymnosperms. *Plant Physiology*, 140, pp.1233-1245.

Tai, H.-C., Chang, C.-H., Cai, W., et al., 2023. Wood cellulose microfibrils have a 24-chain core-shell nanostructure in seed plants. *Nature Plants*, 9, pp.1154-1168.

Taiz, L., Zeiger, E., Møller, I.M. and Murphy, A., 2015. *Plant Physiology and Development*. 6th edn. Sunderland, MA: Sinauer Associates

Takata, N., Saito, S., Sakamoto, T., Tanaka, H. & Moritoh, S., 2019. Poplar NAC transcription factors regulating secondary wall formation are functionally diversified. *Plant Biotechnology Journal*, 17, pp.1069-1083.

Tan, T., Endo, H., Sano, R., Kurata, T. & Yamaguchi, M., 2018. Transcription factors VND1-VND3 contribute to cotyledon xylem vessel formation. *Plant Physiology*, 176, pp.773-789.

Taylor-Teeple, M., Lin, L., de Lucas, M., Turco, G., Toal, T.W., Gaudinier, A., Young, N.F., Trabucco, G.M., Veling, M.T., Lamothe, R., Handakumbura, P.P., Xiong, G., Wang, C., Corwin, J., Tsoukalas, A., Zhang, L., Ware, D., Pauly, M., Kliebenstein, D.J., Dehesh, K., Tagkopoulos, I., Breton, G., Pruneda-Paz, J.L., Ahnert, S.E., Kay, S.A., Hazen, S.P. & Brady, S.M., 2015. An *A. thaliana* gene regulatory network for secondary cell wall synthesis. *Nature*, 517, 7536., pp.571-575.

De La Torre, A.R. et al., 2014. Insights into conifer giga-genomes. *Plant Physiology*, 166, pp.1724-1732.

- Treangen, T.J. & Salzberg, S.L., 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13, pp.36-46.
- Tuominen, H., Sitbon, F., Jacobsson, C., Sandberg, G., Olsson, O. & Sundberg, B., 1997. Altered growth and wood characteristics in transgenic hybrid aspen expressing *Agrobacterium tumefaciens* T-DNA indoleacetic acid-biosynthetic genes. *Plant Physiology*, 114, pp.1019-1029.
- Turco, G.M., Rodriguez-Medina, J., Siebert, S., et al., 2019. Molecular mechanisms driving switch behavior in xylem cell differentiation. *Cell Reports*, 28, pp.342-351.e4.
- Tuskan, G.A. et al., 2006. The genome of black cottonwood (*Populus trichocarpa*). *Science*, 313, 5793., pp.1596-1604.
- Ugglå, C., Mellerowicz, E.J. & Sundberg, B., 1998. Indole-3-acetic acid controls cambial growth in Scots pine by positional signaling. *Plant Physiology*, 117, pp.113-121.
- Uvila, J., Häggman, H. & Aronen, T., 2020. Optimisation of nuclei isolation for chromatin studies in Scots pine. *Plant Methods*, 16, 1-12.
- Upton, R.N., Correr, F.H., Lile, J., Reynolds, G.L., Falaschi, K., Cook, J.P. & Lachowicz, J., 2023. Design, execution and interpretation of plant RNA-seq analyses. *Frontiers in Plant Science*, 14, 1135.455.
- Vanholme, R., De Meester, B., Ralph, J. & Boerjan, W., 2019. Lignin biosynthesis and its integration into metabolism. *Current Opinion in Biotechnology*, 56, pp.230-239.
- Vasilevski, O., Hampton, M. & Juenger, T.E., 2012. Genome-wide association studies and linear modelling identify metabolite quantitative trait loci in *A. thaliana*. *Molecular Ecology*, 21(16), pp.4032-4048.
- Van de Velde, J., Van Bel, M., Van Eechoutte, D. & Vandepoele, K., 2016. A collection of conserved non-coding sequences to study gene regulation in flowering plants. *Plant Physiology*, pp.00821.2016.
- Wang, J.P., Matthews, M.L., Williams, C.M., et al., 2020. Improving wood properties for bioenergy through lignin modification. *Plant Biotechnology Journal*, 18, pp.1486-1497.

Watson, J.D. & Crick, F.H.C., 1953. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171, 4356., pp.737-738.

Weber, B., Zicola, J., Oka, R. & Stam, M., 2016. Plant enhancers: a call for discovery. *Trends in Plant Science*, 21(11), pp.974-987.

Wei, H., Rodriguez, K., Renneckar, S. & Vikesland, P., 2014. Environmental science and engineering applications of nanocellulose-based nanocomposites. *Environmental Science: Nano*, 1(4), pp.302-316.

Wei, S., Yang, Y. & Yin, T., 2020. The chromosome-scale assembly of the willow genome provides insight into Salicaceae genome evolution. *Horticulture Research*, 7(1).

Wei, T., Yang, H. & Yin, T., 2020. Comparative genomics analysis of lignin biosynthesis genes in woody *Populus* and herbaceous *A. thaliana*. *BMC Plant Biology*, 20, pp.1-14.

Wei, Z. & Wei, H., 2024. Deciphering the intricate hierarchical gene regulatory network: unraveling multi-level regulation and modifications driving secondary cell wall formation. *Horticulture Research*, 11(2), uhad281. doi:10.1093/hr/uhad281.

Wendel, J.F., Jackson, S.A., Meyers, B.C. & Wing, R.A., 2016. Evolution of plant genome architecture. *Genome Biology*, 17, 37.

Wertheim, J.O. et al., 2015. RELAX: detecting relaxed selection in a phylogenetic framework. *Molecular Biology and Evolution*, 32, pp.820-832.

Winkler, A. & Oberhuber, W., 2017. Cambial response of Norway spruce to modified carbon availability by phloem girdling. *Tree Physiology*, 37(11), pp.1527-1535.

Wolfe, C., Kohane, I. & Butte, A., 2005. Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics*, 6(1), p.227.

Wray, G.A., 2007. The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics*, 8(3), pp.206-216.

Wuchty, S. & Stadler, P.F., 2003. Centers of complex networks. *Journal of Theoretical Biology*, 223(1), pp.45-53.

- Xie, Z. et al., 2010. Role of the stomatal development regulators FLP/MYB88 in abiotic stress responses. *Plant Journal*, 64, pp.731-739.
- Xu, Y., Bush, S.J., Yang, X., Xu, L., Wang, B. & Ye, K., 2023. Evolutionary analysis of conserved non-coding elements subsequent to whole-genome duplication in opium poppy. *Plant Journal*, 116(6), pp.1804-1824.
- Yamaguchi, M., Goue, N., Igarashi, H. et al., 2010. Vascular-related NAC-domain6 and vascular-related NAC-domain7 effectively induce transdifferentiation into xylem vessel elements under control of an induction system. *Plant Physiology*, 153, pp.906-914.
- Yang, F., Mitra, P., Zhang, L., et al., 2013. Engineering secondary cell wall deposition in vessels and fibers of *Populus* wood. *Plant Biotechnology Journal*, 11, pp.325-335.
- Yang, H.-W., Akagi, T., Kawakatsu, T. & Tao, R., 2019. Gene networks orchestrated by MeGI: a single-factor mechanism underlying sex determination in persimmon. *The Plant Journal*, 98, pp.97-111.
- Yao, J., Shen, Z., Zhang, Y., Wu, X., Wang, J., Sa, G., Zhang, Y., Zhang, H., Deng, C., Liu, J., Hou, S., Zhang, Y., Zhang, Y., Zhao, N., Deng, S., Lin, S., Zhao, R. & Chen, S., 2020. *Populus euphratica* WRKY1 binds the promoter of H⁺-ATPase gene to enhance gene expression and salt tolerance. *Journal of Experimental Botany*, 71, pp.1527-1539.
- Ye, Z.-H. & Zhong, R., 2015. Molecular control of wood formation in trees. *Journal of Experimental Botany*, 66(14), pp.4119-4131. doi:10.1093/jxb/erv081.
- Yeaman, S., Hodgins, K.A., Suren, H., Nurkowski, K.A., Rieseberg, L.H., Holliday, J.A. & Aitken, S.N., 2014. Conservation and divergence of gene expression plasticity following c.140 million years of evolution in lodgepole pine (*Pinus contorta*) and interior spruce (*Picea glauca* × *P. engelmannii*). *New Phytologist*, 203(2), pp.578-591.
- Yue, Y., Tian, S., Wang, Y., Ma, H., Liu, S., Wang, Y. & Hu, H., 2018. Transcriptomic and GC-MS metabolomic analyses reveal the sink strength changes during *Petunia* anther development. *International Journal of Molecular Sciences*, 19(4), p.955.

Zander, M., Lewsey, M., Clark, N., Yin, L., Bartlett, A., Saldierna Guzmán, J., Hann, E., Langford, A., Jow, B., Wise, A., Nery, J., Chen, H., Bar-Joseph, Z., Walley, J., Solano, R. & Ecker, J., 2020. Integrated multi-omics framework of the plant response to jasmonic acid. *Nature Plants*, 6(3), pp.290-302.

Zhang, H., Lang, Z. & Zhu, J.-K., 2018. Dynamics and function of DNA methylation in plants. *Nature Reviews Molecular Cell Biology*, 19, pp.489-506.

Zhang, M., Ji, C., Zhu, J., Wang, X., Wang, D. & Han, W., 2017. Comparison of wood physical and mechanical traits between major gymnosperm and angiosperm tree species in China. *Wood Science and Technology*, 51(6), pp.1405-1419.

Zhang, Y., Liu, C. & Cheng, H. et al., 2020. DNA methylation and its effects on gene expression during primary to secondary growth in poplar stems. *BMC Genomics*, 21, pp.498.

Zhong, R., Cui, D. & Ye, Z.-H., 2017. Regiospecific acetylation of xylan is mediated by a group of DUF231-containing O-acetyltransferases. *Plant and Cell Physiology*, 58(12), pp.2126-2138.

Zhong, R., Cui, D., Phillips, D.R., Richardson, E.A. & Ye, Z.-H., 2020. A group of O-acetyltransferases catalyse xyloglucan backbone acetylation and can alter xyloglucan xylosylation pattern and plant growth when expressed in *A. thaliana*. *Plant and Cell Physiology*, 61(7), pp.1064-1079.

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

Zhong, R., McCarthy, R.L., Haghghat, M. & Ye, Z.-H., 2013. The poplar MYB master switches bind to the SMRE site and activate the secondary wall biosynthetic program during wood formation. *PLoS ONE*, 8(7), e69219.

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

Zhong, R., Richardson, E.A. & Ye, Z.-H., 2010. The MYB46 transcription factor is a direct target of SND1 and regulates secondary wall biosynthesis in Arabidopsis. *The Plant Cell*, 19(9), pp.2776-2792.

Zhong, R. & Ye, Z.H., 2013. Transcriptional regulation of wood formation in tree species. In: Fromm, J. (ed.) *Cellular Aspects of Wood Formation*. Plant Cell Monographs, vol. 20. Berlin: Springer, pp.141-158.

Zhong, R. & Ye, Z.-H., 2014. Complexity of the gene regulatory network controlling secondary wall biosynthesis. *Plant Science*, 229, pp.193-207.

Zhou, J., Zhong, R. & Ye, Z.-H., 2014. Arabidopsis NAC domain proteins VND1 to VND5 are transcriptional regulators of secondary wall biosynthesis in vessels. *PLoS ONE*, 9(10), pp.e105726

Zhu, Y. & Li, L., 2024. Wood of trees: Cellular structure, molecular formation and genetic engineering. *Journal of Integrative Plant Biology*, 66(2), pp.184-210.