NorWood: a gene expression resource for evo-devo studies of conifer wood development

Soile Jokipii-Lukkari¹²*, David Sundell¹*, Ove Nilsson², Torgeir R. Hvidsten³¹, Nathaniel R. Street¹ and Hannele Tuominen¹

¹Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, SE-901 87 Umeå, Sweden; ²Umeå Plant Science Centre, Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Sciences, SE-901 84 Umeå, Sweden; ³Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, 1430 Ås, Norway

Summary

- The secondary xylem of conifers is composed mainly of tracheids that differ anatomically and chemically from angiosperm xylem cells. There is currently no high-spatial-resolution data available profiling gene expression during wood formation for any coniferous species, which limits insight into tracheid development.
- RNA-sequencing data from replicated, high-spatial-resolution section series throughout the cambial and woody tissues of *Picea abies* were used to generate the NorWood.conGenIE.org web resource, which facilitates exploration of the associated gene expression profiles and co-expression networks.
- Integration within PlantGenIE.org enabled a comparative regulomics analysis, revealing divergent co-expression networks between *P. abies* and the two angiosperm species *Arabidopsis thaliana* and *Populus tremula* for the secondary cell wall (SCW) master regulator NAC Class IIB transcription factors. The SCW cellulose synthase genes (CesAs) were located in the neighbourhoods of the NAC factors in *A. thaliana* and *P. tremula*, but not in *P. abies*. The NorWood co-expression network enabled identification of potential SCW CesA regulators in *P. abies*.
- The NorWood web resource represents a powerful community tool for generating evo-devo insights into the divergence of wood formation between angiosperms and gymnosperms and for advancing understanding of the regulation of wood development in *P. abies*.

Introduction

Conifers are woody plants representing the largest lineage of the ancient seed plant group of gymnosperms (Wang & Ran, 2014). Despite the relatively small number of extant species, conifers dominate vast areas, especially in boreal forests of the northern hemisphere, playing an important role in carbon cycling and ecosystem function. Conifers are important producers of raw material for paper, solid fuels, liquid biofuels and biomaterials (Guo *et al.*, 2017; Isikgor & Becer, 2015).

The majority of conifer species are large, slow-growing trees. They share several properties with angiosperm trees, such as the formation of extensive secondary xylem (the ‘wood’), which facilitates water and nutrient transport in addition to mechanical support. However, there are striking anatomical differences in how these functions are achieved: while conifer wood consists mainly of single-celled tracheids that serve both water transport and physical support functions, angiosperm wood is composed of water-transporting multicellular conduits called vessels and libriform fibres, which provide structural support.

Secondary cell wall (SCW) properties are one of the primary determinants of xylem element function. The SCW is deposited inside the primary cell wall and is composed of polysaccharidic cellulose and hemicellulose and the polyphenolic compound lignin. Current knowledge of SCW formation derives largely from studies of *Arabidopsis thaliana*, which have identified the importance of NAC (NAM, ATAF1/2 and CUC2) domain proteins in the regulation of SCW formation. NAC SECONDARY WALL THICKENING PROMOTING FACTOR1 (NST1) and NST2 function redundantly in regulating SCW thickening in the endothecium of anthers (Mitsuda *et al.*, 2005), whereas NST1 and NST3 (also known as SND1, SECONDARY WALL-ASSOCIATED NAC DOMAIN PROTEIN1) regulate SCW thickening of xylem fibres (Mitsuda *et al.*, 2007; Zhong *et al.*, 2007). VASCULAR-RELATED NAC-DOMAIN6 (VND6) and VND7, on the other hand, control formation of vessel elements (VNDs; Kubo *et al.*, 2005; Yamaguchi *et al.*, 2008). These SCW ‘master switches’ regulate transcription of downstream transcription factors (TFs), including MYB (myeloblastosis) family proteins, which further regulate other downstream TFs and, finally, the structural SCW biosynthetic genes (Hussey *et al.*, 2013), including SCW cellulose synthase genes (*CesA*). Many of
the upstream TFs regulate both their downstream TFs and the cell wall biosynthetic genes directly (feed-forward regulation), indicating the complexity of the regulatory network (Hussey et al., 2013; Taylor-Teeple et al., 2015).

The anatomical differences between gymnosperm and angiosperm xylem elements suggest corresponding differences in the regulation of xylem formation and differentiation. A few studies have been performed to elucidate these aspects in conifer trees. R2R3 MYB TF family members, which are implicated in the biosynthesis of phenylpropanoids and flavonoids, display conserved DNA binding domains between angiosperms and gymnosperms (Bedon et al., 2007), and evidence for functional conservation has been reported for two R2R3 MYB factors in lignin biosynthesis (Bomal et al., 2008). Duval et al. (2014) performed a protein–DNA interaction study in Picea glauca for a number of xylem-expressed TFs, identifying regulatory networks that were broadly similar to those in A. thaliana. Functional conservation of the master switch NAC TFs in the Class IIB family (also referred to as WNDs (Wood-Associated NAC Domain), VNS (VND/NST/SMB) or SWNs (Secondary Wall NACs)) has not been investigated, although homologous members of the family were reported in Picea abies (Nystedt et al., 2013).

Detailed understanding of the regulatory aspects of wood development requires analyses of gene expression with high spatial resolution. This is not easily achievable in A. thaliana due to the small size of plants and the minor extent of secondary xylem formation. To overcome these limitations, angiosperm trees including members of the genus Populus and their hybrids have emerged as powerful alternative models. Several gene expression studies have been performed using cryogenic tangential cutting series (hereafter cryo-series) assayed with cDNA expression microarrays (Hertzberg et al., 2001; Schrader et al., 2004; Moreau et al., 2005; Courtois-Moreau et al., 2009) or RNA-sequencing (RNA-seq; Immanen et al., 2016). These provided the first insights into detailed patterns of gene expression across the vascular cambium, differentiating secondary phloem and xylem within a single annual growth ring. In conifers, global gene expression studies on different wood types, such as earlywood vs latewood (Paiva et al., 2008; Li et al., 2010; Raherison et al., 2015), compression wood vs normal wood (Villalobos et al., 2012) and ‘high density’ wood vs ‘low density’ wood (Stephenson et al., 2011), have provided views of the co-expression networks active during xylem development. However, these were limited either by restricted coverage of the transcriptome, specificity or dynamic range (in the case of cDNA arrays) and/or by limitations in sampling resolution.

The availability of draft conifer genome assemblies (De La Torre et al., 2014) affords new opportunities for evolutionary, evo-devo and comparative genomics studies. Here, the P. abies genome (Nystedt et al., 2013) was used together with a novel high-spatial-resolution RNA-seq gene expression resource to profile the developmental process of secondary xylem formation in mature P. abies trees growing under natural conditions. A web resource ‘NorWood’ (http://NorWood.ConGenIE.org) was implemented for data exploration, visualisation and integration with the PlantGenIE.org platform (Sundell et al., 2015). The potential of the NorWood resource is exemplified through comparative analyses of co-expression networks in A. thaliana, Populus and P. abies, demonstrating that the resource captures existing knowledge and provides novel insight into the evolutionary divergence of xylem differentiation.

Materials and Methods

Plant material

Samples were collected from 49-yr-old Norway spruce (P. abies [L.] Karst.) trees growing at a seed orchard in Hissjö (Västerbotten, Sweden). The sampled trees represented clonal copies of genotype ‘Z4006’ (see Supporting Information Notes S1, section 1.1), the genotype used for the P. abies genome sequence (Nystedt et al., 2013). Wood pieces were collected from three replicate trees at the end of July 2012 (see ‘Collection of stem samples’) when cell death of newly formed earlywood was ongoing but before latewood cells could be observed based on microscopic examination (see below). All samples were collected at noon.

Collection of stem samples with intact cambium

Samples were taken at breast height in a part of the stem free of branches or traumatic resin ducts. To preserve intact cambium the bark, phloem and part of the wood was removed from a c. 3 cm × 10 cm region around the sampling area using a carpet knife. This released the shearing forces present in the stem. Samples covering bark, phloem, cambium, developing xylem and two to three previous year rings were then taken by hammering a chisel into the stem at a 90° angle immediately above the pieces and hammering tangentially downwards. The samples used for RNA-seq were immediately frozen in liquid N2 and kept on dry ice until storage at −80°C. Stem samples used for light microscopy were kept on ice until preparation for microscopy.

Light microscopy of fresh tissues

Cell viability and the location of the cell death zone were determined by staining the fresh tissues with nitroblue tetrazolium (NBT; Berlyn & Miksche, 1976; Gahan, 1984), which in the presence of succinate indicates succinate dehydrogenase activity of living cells. Cross-sections of c. 30 μm were cut using a cylinder hand microtome and incubated for 1.5 h in 50 mM sodium phosphate buffer (pH 7.6) containing 50 mM sodium succinate and 500 mg l−1 NBT under strong light conditions. Sections were examined using an Axioplan 2 microscope (Zeiss) and micrographs captured with an Axioscope. HRC camera (Zeiss). The abundance of living xylem cells was evaluated by examining the captured images using the AXIOVISION LE software (Rel. 4.9.1; Zeiss). Data were collected from three separate sections per tree and at three positions in each section (left, middle and right).
Collecting tangential cryogenic sections

To assay gene expression at high spatial resolution across the vascular development zone, a continuous cryo-section series was taken starting in the functional phloem and ending in the latewood from the previous year. A band saw was used to rapidly trim the frozen blocks, after which specimens were mounted into OCT media (VWR, Radnor, PA, USA) at $-25^\circ$C and orientated parallel to the cambium according to Uggl{\aa} et al. (1996). Tangential 30 $\mu$m sections, the average tracheid diameter in P. abies (Havimo et al., 2008), and c. 5 $\times$ 25 mm in size were sectioned using a steel blade in an HM 505 E microtome (Microm Laborger{"a}te, Walldorf, Germany) and stored at $-80^\circ$C. Transverse sections were taken from the tissue block and used to confirm the composition of tissue types in each tangential section.

RNA extraction and amplification

Total RNA from the sections was extracted using the miRNeasy Micro Kit (Qiagen) according to the manufacturer’s instructions. The number of cryosections collected from each tree ranged from 44 to 82. Sections representing cambium or newly formed xylem were extracted individually, whereas sections from other regions of the series were pooled (Table S1) to increase the amount of isolated RNA and to even out differences among replicate trees in the width of the year ring. These are hereafter referred to as samples, which were numbered in ascending numerical order from -01 for the first sample on the phloem side. The replicate trees are referred to as T1–T3.

Homogenisation of the material was performed with metal beads in lysis buffer using a bead mill (25 Hz, 3 min; Retsch, Haan, Germany). Before purification the optional DNase treatment was performed. RNA concentration and purity were measured using a NanoDrop 2000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA) and integrity was analysed on an Agilent 2100 Bioanalyzer with Pico chips (Agilent Technologies, Waldbronn, Germany). mRNA was amplified using the MessageAmpII aRNA Amplification Kit (Thermo Fisher Scientific, Waltham, MA, USA) according to the manufacturer’s instructions. It was not possible to extract adequate RNA from samples -01 and -02 from trees T1–T3. As a result, the phloem is not represented in the current data.

Data processing and quality control

Sequencing of the amplified RNA (aRNA) was conducted at SciLifeLab (Stockholm, Sweden) using an Illumina HiSeq 2500, as in Nystedt et al. (2013). Paired-end reads ($2 \times 125$ bp, target insert size 250 bp) to a minimum sequence depth of 10 million reads per sample were generated. Reads were pre-processed as in Delhomme et al. (2014). Pre-processing steps comprised FastQC-0.11.2 (http://www.bioinformatics.babraham.ac.uk/proj/sequencing/fastqc/) for quality control, SortmRNA-2.0 to filter RNA contaminants, Trimmomatic-0.36 (Bolger et al., 2014; ILLUMINACLIP:2:30:10) for adapter removal and read trimming, STAR-2.0.3 (Dobin et al., 2013; --outSAMmapqUnique 254 --quantMode TranscriptomeSAM --outFilterMultimapNmax 100 --chimSegmentMin 1) for read alignment and HTSeq-0.6.1 (-m intersection-nonempty -s yes -t exon -i Parent; http://www-huber.embl.de/users/anders/HTSeq/) to count aligned reads. Read counts were normalised and variance stabilized transformed (VST) to obtain gene expression values using the R-package DESeq2-1.12 (Love et al., 2014). A filter requiring VST $>3$ in two samples per tree in at least two of the three replicate trees was applied, resulting in 18513 expressed genes (25% of the gene space). One sample (T2-15) was not sent for sequencing and one sample (T2-10) had insufficient reads after sequencing. Both were replaced by the average of the samples immediately before and after the missing sample.

Data availability

Sequencing data are available at the European Nucleotide Archive as accession ERP017340.

Co-expression network and network connectivity measures

A co-expression network was calculated for all 18513 expressed genes. Mutual information (MI; Daub et al., 2004) was used to calculate pairwise correlation between genes and context likelihood of relatedness (CLR) to transform these correlations into a background corrected $Z$-score (Faith et al., 2007; Netoetea et al., 2014). Briefly, MI is a popular nonparametric correlation measure that is robust to outliers, while the CLR algorithm performs a local background correction that serves to effectively remove false positive correlations. The method computes a $Z$-score for each pair of genes using a null distribution obtained from the scores between these two genes and all other genes. Further discussion of MI/CLR can be found in Netoetea et al. (2014). A network was then constructed by linking all gene-pairs with a $Z$-score above a determined CLR threshold ($Z$-score $\geq 5$). This CLR threshold was determined using a scale freeness test, based on previous analysis showing that biological networks have scale-free properties (Barab{\'{a}}si & Oltvai, 2004; MacNeil & Walhout, 2011). After applying the CLR threshold, 17942 genes were part of the largest sub-network, having at least one neighbour, and were thus included in the network. The vast majority of genes not contained in this largest subnetwork were present as disconnected nodes. Since MI measures the dependency between two expression profiles regardless of the sign of that dependency, edges in the network were annotated as positive or negative using Pearson correlation. The network was subsequently used to calculate network centrality scores for each gene: degree centrality, the number of direct neighbours (first-order neighbours) of a gene in the network; betweenness centrality, the number of shortest paths passing through a gene, where the shortest path is the smallest number of edges connecting a pair of genes; closeness centrality, the minimum total distance from a gene to all other genes; and average neighbour degree, the average number of neighbours of a gene’s direct neighbours. The betweenness and closeness centrality scores were ranked from the highest to lowest to obtain rank order scores. Network centralities were
calculated using the Python package networkx-1.9 (https://networkx.github.io/).

TF annotation, hierarchical clustering and gene ontology enrichment

TF annotations for *P. abies* were obtained from the Plant Transcription Factor Database (PlantTFDB; Jin et al., 2016). Unsupervised clustering was performed using the R package gplots-2.2.1.0 (Warnes et al., 2016) and the heatmap2 function from the R package gplots (Warnes et al., 2016). The hierarchical dendrograms were calculated using Ward’s method, and were based on Euclidian distance for the samples and Pearson correlation for genes. In total, 7,788 expressed genes with a variance of σ > 1 across all samples were included in the clustering. After visual examination of the cluster dendrogram, this was cut to define seven gene expression clusters.

In addition to the hierarchical clustering, each gene was assigned to the sample(s) in which the expression of that gene was within 4% of the maximum expression for that gene across all samples. Gene ontology (GO) enrichment was then tested for the set of genes assigned to each sample to annotate the developmental stage. As the samples did not align exactly across the three replicate cutting series, this analysis was performed using only tree T1.

GO enrichment tests were performed for all hierarchical clusters and sample clusters using Fisher’s exact test in the Python package fisher-0.1.4 (https://pypi.python.org/pypi/fisher/) and false discovery rate (FDR) correction.

NorWood implementation

The Norwood web resource (http://norwood.congenie.org) was implemented using a combination of server and user side scripting languages (HTML-5, Python-2.7, PHP-5, JQuery-1.9 and JavaScript-1.8). All expression, network and annotation data were stored and managed using a MySQL-5.7 database. PHP-MYSQL and MySQL-PYTHON-1.2.5 were used for web management. Gene profiles were plotted using the Python matplotlib-1.5.1 library. Networks were displayed using CytoscapeWeb 1.0.4 (Lopes et al., 2010) and heatmaps using the heatmap2 function in R. The gene information and functional enrichment tables were implemented using the javascript DataTables-1.9 plugin and CSS design. The web resource layout was designed using HTML-5 CSS styling. All scripts are available at https://github.com/UPSCb/NorWood.

Phylogenetic analyses of VND, NST and CesA genes

BLAST searches of *A. thaliana* NST and CesA genes were performed against the *P. abies* genome at ConGenIE.org. The *P. abies*, *A. thaliana* and *P. trichocarpa* gene family members of the identified sequences were extracted from the associated gene information pages (Gene Family tab). *P. abies* genes expressed in the NorWood cryo-section series, *Populus trichocarpa* genes expressed within the RNA-seq data at PopGenIE.org, and genes that were full-length compared to *A. thaliana* sequences were included in subsequent analyses. Two very short CesA superfamilies (MA_10429177g0020 and Potri.011G152300) were discarded. In *P. abies* some CesA gene models were truncated, and further analyses of RNA-seq read alignments, Trinity transcript assemblies available at ConGenIE.org and de novo gene prediction all confirmed two cases where a combination of CesA gene fragments could be merged to form a full-length gene model (Notes S1, section 1.2). These were then used for phylogenetic analyses. For downstream expression analysis, we used the longest of the original fragmented gene models. Phylogenetic trees were created using the GALAXY platform (Goecks et al., 2010) hosted at PlantGenIE.org. The workflow utilises MUSCLE v.3.8.31 (Edgar, 2004; maximum number of iterations: 16) for multiple alignment, and PHYLML 3.1 (Guindon et al., 2010; substitution model: WAG, aLRT test: SH-like, tree topology search operation: Nearest Neighbor Interchange) and TREEVECTOR (Pethica et al., 2010) for building and drawing phylogenetic trees, respectively. For the *NAC* gene family tree, the Class IIB subfamily was manually identified and this subset of genes was then used to create a phylogenetic tree.

Results

Gene expression analysis across *P. abies* wood revealed expression clusters associated with developmental transitions during wood formation

A high-spatial-resolution gene expression data resource was created in *P. abies* woody tissues representing developmental stages of cambial through to secondary xylem formation (Fig. 1a; Table S1). In total, 51 samples were assayed for gene expression using deep RNA-seq profiling, with an average of 15 million aligned paired-end reads per sample (Table S2). Hierarchical clustering of the gene expression data revealed expression profiles that were highly reproducible among the three replicate trees (Figs 1b, S1), supporting the high quality and reproducibility of the data. The resulting data were integrated into a web resource, ‘NorWood’, to allow community exploration and analysis of gene expression profiles and the associated co-expression network.

Hierarchical clustering of both genes and samples was performed, which identified seven gene clusters (Fig. 1b,c; Table S3) and three sample clusters (Figs 1b, S2). The gene expression profiles in clusters a-e corresponded well to the different developmental stages of wood formation, with the expression peak of cluster-a corresponding to cambial tissues, cluster-b to xylem expansion, cluster-c to the transition from xylem expansion to SCW formation, cluster-d to bulk SCW formation, cluster-e to transition between SCW formation and mature xylem (i.e. programmed cell death) and cluster-f to mature wood and the previous year’s latewood (Fig. 1c). The expression profiles of genes in cluster-z did not have a distinct developmental profile and this cluster was not considered in subsequent analyses.

GO enrichment analyses of the expression clusters confirmed that they reflected biological processes congruent with the cluster...
expression domains (Table S4). For example, GO categories related to cell division including ‘regulation of cell cycle’ (GO:0051726) and ‘DNA replication’ (GO:0006260) were enriched in cluster-a, in agreement with this cluster containing genes having highest expression in close proximity to the dividing cambial cells. Several GO categories related to chromatin status, including ‘chromatin organisation’ (GO:0006325), ‘chromatin modification’ (GO:0016568) and ‘histone modification’ (GO:0016570), were also enriched in cluster-a. The cellular components ‘cytoskeletal part’ (GO:0044430) and ‘microtubule’ (GO:0005874) were enriched in cluster-d, and the molecular functions ‘cysteine-type peptidase activity’ (GO:008233) and ‘hydrolase activity’ (GO:16787) in cluster-e, supporting the predominance of cell wall formation and cell death in determining the respective cluster expression domains. Similarly, examination of individual genes within the clusters identified numerous candidate genes related to the biological processes occurring within the cluster expression domains (Table S3). For example, cluster-a included an orthologue of A. thaliana WOX4, a known regulator of cambial development (Hirakawa et al., 2010). In cluster-f, terms related to stress and transport were enriched, including ‘response to wounding’ (GO:0009611) and ‘single-organism transport’ (GO:0044765).

GO enrichment was analysed separately in the different samples in tree T1, which was selected as the replicate with the highest sequencing data and sample quality. To identify the predominant biological processes active within each sample, expression of genes was assigned only to the sample(s) in which they were expressed within 4% of the maximum expression for that gene, after which gene ontology (GO) enrichment was calculated for each sample. Selected GO categories are represented. All significant categories are available in Supporting Information Table S5.
gene was expressed within 4% of its maximal expression. The first sample, T1-03, was composed entirely of dividing cambial cells and, accordingly, was enriched for GO categories related to cell division, including ‘DNA replication’ (GO:0006260) and ‘protein DNA complex assembly’ (GO:0065004) (Fig. 1d; Table S5). In samples within the Expanding Xylem (EX), terms including ‘cell growth’ (GO:0016049), ‘macromolecular metabolic process’ (GO:0043170), ‘protein folding’ (GO:006457) and ‘carbohydrate metabolic process’ (GO:0005975) were enriched, supporting the predominance of not only cell expansion but also biosynthesis of macromolecules, such as the cell wall carbohydrates, in these samples (Fig. 1d; Table S5). In samples from the SCW region, categories related to biosynthesis and transport of the SCW components were enriched, as would be expected (Table S5). Notably, the last samples before latewood from the previous year (T1-17 and 18) had a distinct set of enriched GO terms in comparison to the immediately surrounding samples, which could be related to their unique nature in representing the mature part of the wood, where the only living cell type is the ray cells. Also of note, the last sample (T1-20), which contained latewood from the previous year, was different from the two immediately preceding samples, suggesting that ray cells may not be the only transcriptionally active cell type in this sample, but that some tracheids could still be living and contribute to the unique gene expression pattern of this sample.

Taken together, the results of both the gene expression clustering and sample enrichment analyses show that coordinated sets of genes display distinct and spatially defined expression profiles, the functional enrichment of which represent biological transitions known to occur along the wood developmental trajectory. Further exploration and validation of this dataset is expected to reveal several novel aspects of xylem differentiation.

The NorWood web resource

An integrated set of web-accessible tools was developed as the NorWood (Norway spruce Wood) resource (http://norwood.congenie.org). These tools enable exploration and analysis of the RNA-seq data resource, with NorWood being integrated within the PlantGenIE (Plant Genome Integrative Explorer; http://plantgenie.org; Sundell et al., 2015) web platform to facilitate wider analysis using the extended range of tools and species hosted under the PlantGenIE umbrella. The NorWood web resource can be used to visualise gene expression profiles, explore the pre-calculated co-expression network and analyse enrichment of functional categories.

The co-expression network was used to calculate a number of network centrality measures based on the largest connected sub-network of 17,942 genes (Table S6). Four measures of network connectivity were calculated and are available at NorWood: degree, betweenness rank, closeness rank and average nearest neighbour degree (see the Materials and Methods section). Degree is used here as the default connectivity measure and represents a score where increasing values indicate an increasing number of connections (edges in the network) to other genes within the network. An underlying assumption is that genes of high degree play a central role within the network and that their removal, or modulation of their expression, would induce extensive knock-on effects, affecting a large number of other genes. Although correlation provides only circumstantial evidence, and such high-degree nodes are not necessarily regulators, they are strong candidates for follow-on biological studies. Similarly, genes with high betweenness can be considered potential switches linking one region of the network to another, for example a gene controlling a developmental transition from cell division to differentiation. Such network metrics can therefore be used to guide biologists to a subset of the most interesting genes for further consideration when starting with (often) considerably long lists of genes identified from other approaches such as differential expression or clustering analyses. Using network metrics can subsequently be combined with other sources of information, such as gene functional annotation, to further filter candidate gene lists. As an example, a gene with the high betweenness rank was MA_10018650g0010, which was located in cluster-b and is therefore a good candidate regulator for the transition from xylem cell expansion to xylem maturation. This was supported by expression of the closest *A. thaliana* homologue (AT3G15220), a protein kinase of unknown function that is expressed specifically in expanding xylem vessel elements of the primary root within the AtGenExpress dataset (Schmid et al., 2005), as visualised in the e-FP browser (http://bar.utoronto.ca/efp/cgi-bin/efpWeb.cgi; Winter et al., 2007) available within the AtGenIE website (http://atgenie.org/efp). Interestingly, the closest neighbour of MA_10018650g0010 within the NorWood expression network is a zinc finger TF (MA_10432012g0010) that is the most sequence-similar homologue of an *A. thaliana* gene (AT5G05660) predicted to encode the circadian clock component EARLY BIRD (Johansson et al., 2011), indicating potential involvement of circadian regulation in the onset of xylem maturation. GO enrichment tests of the top 100 genes ranked by betweenness and centrality did not identify any enriched terms (Table S6).

Exploiting NorWood integration within PlantGenIE to identify potential CesA regulators

To demonstrate the potential of the tools available at NorWood and PlantGenIE, network neighbours were extracted for the *cellulose synthase* (*CesA*) gene family members to identify their potential regulators (Fig. 2). A walk-through guide to recreating the results presented is available from the NorWood homepage. First, theGeneList tool at AtGenIE was used to identify the *A. thaliana CesA* genes (search term ‘cesa’). CesA7 (AT5G17420) was selected to extract the gene family members in *A. thaliana*, *P. trichocarpa* and *P. abies* from the Gene Family tab of the *CesA7*-associated gene information page (Fig. 2a). The phylogenetic tree of these *CesA* genes created using the PlantGenIE Galaxy workflow showed that the *P. abies* genome contains putative homologues for primary cell wall (PCW) and SCW *CesAs*, as well as a number of *CesA-like* (*CSL*) genes (Fig. 2a; Notes S1, section 1.2).
Fig. 2 Using the NorWood resource to identify putative regulators of secondary cell wall cellulose synthase genes. (a) Arabidopsis thaliana cellulose synthase A7 (CesA7; AT5G17420) was used to find and retrieve the cellulose synthase gene family at AtGenIE.org. A phylogenetic tree of the CesA gene family for A. thaliana, Picea abies (Norway spruce) and Populus trichocarpa (poplar) produced using the PlantGenIE phylogenetic tree GALAXY workflow is shown. (b) Conservation of co-expression between the expressed CesA gene family members in NorWood and their P. trichocarpa orthologues was examined using the Complex tool. (c) The network neighbourhood of CesA genes with conserved co-expression was retrieved using the NorWood dataset within the exNet tool at ConGenIE using a context-likelihood of relatedness (CLR) threshold of 5. This neighbourhood contained four genes comprising MA_183130g0010, MA_140410g0010 and MA_10429177g0010, which are orthologous to the A. thaliana secondary cell wall CesA genes, and one secondary cell wall CesA-like gene, MA_10429339g0020. Shown in this panel on the left is the network, in the centre are line plots from the exPlot tool with one coloured line representing each of the four neighbourhood genes and the network neighbourhood of these four genes on the right. In the line plots, T1–T3 represent the three replicate trees and dashed vertical lines represent cambium/expanding xylem, secondary cell wall/programmed cell death and mature xylem/late wood zones. (d) The subset of network neighbours annotated as transcription factors (TFs) were selected. The network neighbourhood of this TF subset was then selected at a CLR threshold of 5, subset to genes annotated as TFs, the expression of which were visualised as a heatmap within the NorWood resource. Shown on the left are the starting nine TFs identified in (c), in the centre is the network of TFs where the original nine are shown in green and where neighbours connected by positive edges (i.e. a positive Pearson correlation, indicated as grey edges) are shown in pink and those connected by negative edges (red edges) are shown in blue. In the heatmap expression values are scaled per gene so that expression values above the gene average are represented by red, and below average by blue. Samples appear in the order of sampling within each tree (T1–T3).
Conservation of the network neighbourhoods was analysed for *P. abies* CesA and CesA-like genes that were expressed in NorWood and compared to orthologues in *Populus*, which is the model system for forest tree genomics and also included as a resource in PlantGenIE. Five *P. abies* genes had conserved expression with *Populus* (indicated as blue network edges in Fig. 2b). These were passed as a new gene list to ConGenIE and examined in the NorWood dataset using the exNet tool (Fig. 2c), which identified four strongly co-expressed CesA genes. These four genes included one homologue for each of the *A. thaliana* CesA genes ArcCesA4, 7 and 8 (Fig. S3; MA_183130g0010, MA_140410g0010, MA_10429177g0010), which are known to control cellulose biosynthesis during SCW formation (McFarlane *et al.*, 2014), in addition to one gene (MA_10429339g0020) that was highly similar to the SCW CesA genes (see Fig. S3) and that we hereafter refer to as SCW CesA-like. All four were located in cluster-d, with highest expression during SCW formation. The network neighbourhood of these genes was extracted using the expand function in exNet (at an expansion threshold of 5) after which potential regulators were identified by selecting the subset of TFs (Fig. 2d). This TF subset contained 11 genes (Table 1) including, for example, MA_402393g0010 and MA_95898g0010, which are most sequence-similar to *A. thaliana* ANAC075 (AT4G29230), a known regulator of SCW formation (Endo *et al.*, 2014; Sakamoto & Mitsuda, 2015). Subsequently, the TF subset neighbourhood was expanded (threshold 5), which resulted in identification of 21 TFs including the homologue of *A. thaliana* KNOTTED-like (AT4G08150) and a number of MYB TFs (Table S7).

We additionally explored the co-expression and expression domains of the PCW and SCW CesAs in ConGenIE.org within the exAtlas and NorWood datasets. Proost & Mutwil (2016) reported that using the ConGenIE exAtlas dataset, it appeared that only one shared co-expression module was present containing both the PCW and the SCW CesA genes (Fig. S4). The greater spatial resolution available in the NorWood dataset resolved the expression domains of the PCW and SCW CesAs, revealing two disconnected co-expression modules. The expression profile plots of the two gene sets clearly showed that SCW CesAs were expressed in the SCW forming zone while PCW genes were more ubiquitously expressed across the developmental series, which is an interesting observation suggesting that PCW genes are perhaps more broadly expressed than expected during wood formation in *P. abies*.

NorWood provides evidence of evolutionarily diverged functions of the NAC Class IIB TFs between angiosperms and gymnosperms

Class IIB NAC TFs have been shown to serve as master regulators of the SCW in angiosperms (Nakano *et al.*, 2015). The NorWood web resource was used to explore whether there is expression-based evidence suggesting functional conservation in *P. abies*. Nystedt *et al.* (2013) identified four Class IIB family members in the *P. abies* genome assembly, of which two (MA_18939g0010 and MA_6777g0010) were highly expressed in the single wood sample of the *P. abies* exAtlas dataset (Nystedt *et al.*, 2013; see eXImage at ConGenIE.org) as well as in the NorWood dataset. In a phylogenetic tree, MA_18939g0010 clustered together with the *A. thaliana* NST1, NST2 and NST3, and is hereafter called ‘NST’. MA_6777g0010 clustered together with the *A. thaliana* VND4, VND5 and VND6, and is hereafter called ‘VND’ (Fig. 3a). To examine whether the regulation of SCW biosynthesis by the Class IIB NAC TFs is conserved between *A. thaliana*, *P. tremula* and *P. abies*, the network neighbourhoods of all family members were analysed with the prediction that central SCW biosynthetic genes, such as the SCW CesAs, would be present in all three neighbourhoods if regulation is conserved. This prediction was confirmed for *A. thaliana* and *P. tremula* using the exNet tool at AtGenIE (Fig. 3b) and PopGenIE (Fig. 3c), respectively, within the available AtGenExpress (Schmid *et al.*, 2005) and *P. tremula* exAtlas (Sundell *et al.*, 2015) co-expression networks, respectively. However, in both the *P. abies* exAtlas (Nystedt *et al.*, 2013) and the NorWood datasets, no SCW CesA genes were identified in the co-expression neighbourhoods of either VND or NST, even when using a lower co-expression threshold of 3 (Fig. 3d,e).

To identify putative targets for the VND and NST homologues in *P. abies*, their network neighbours were analysed in

---

**Table 1** Putative transcription factors identified in the network neighbourhood of *Picea abies* secondary cell wall cellulose synthase genes in NorWood

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>A. thaliana best BLASTp</th>
<th>Percentage identity</th>
<th>A. thaliana description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MA_10425867 g0020</td>
<td>AT1G69560 52.86</td>
<td>myb domain protein 105</td>
<td></td>
</tr>
<tr>
<td>MA_10434782g0020</td>
<td>AT1G65620 66.67</td>
<td>Lateral organ boundaries (LOB) domain family protein</td>
<td></td>
</tr>
<tr>
<td>MA_137453g0010</td>
<td>AT3G61250 62.07</td>
<td>myb domain protein 17</td>
<td></td>
</tr>
<tr>
<td>MA_139238g0010</td>
<td>AT4G10680 47.76</td>
<td>myb domain protein 55</td>
<td></td>
</tr>
<tr>
<td>MA_161608g0010</td>
<td>AT3G27650 68.42</td>
<td>LOB domain-containing protein 25</td>
<td></td>
</tr>
<tr>
<td>MA_30401g0010</td>
<td>AT2G45120 27.23</td>
<td>C2H2-like zinc finger protein</td>
<td></td>
</tr>
<tr>
<td>MA_402393g0010</td>
<td>AT4G29230 40.11</td>
<td>NAC domain containing protein 75</td>
<td></td>
</tr>
<tr>
<td>MA_95898g0010</td>
<td>AT1G02040 35.35</td>
<td>C2H2-like zinc finger family protein</td>
<td></td>
</tr>
<tr>
<td>MA_10426405 g0010</td>
<td>AT4G29230 61.98</td>
<td>NAC domain containing protein 75</td>
<td></td>
</tr>
</tbody>
</table>

The first-order neighbours of secondary cell wall CesA genes (MA_183130g0010, MA_140410g0010, MA_10429177g0010, which are orthologous to the Arabidopsis thaliana secondary cell wall CesA genes, and one secondary cell wall CesA-like gene, MA_10429339g0020) identified within the NorWood co-expression network using the exNet tool at ConGenIE.org at a context-likelihood of relatedness (CLR) threshold of 5. Column headers indicate: Gene ID, P. abies gene ID as available at http://congenie.org; best A. thaliana BLASTp, best A. thaliana BLASTp; percentage identity, percentage identity to the best hit A. thaliana protein calculated using Clustal 2.1; A. thaliana description, Arabidopsis thaliana gene description from TAIR 10.
Norwood, which provides far greater spatial resolution than the *P. abies* exAtlas dataset, which comprises different tissue types (Fig. 3d,e). The network neighbours of NST in NorWood included two orthologues of *A. thaliana* lignin biosynthetic Cinnamoyl-coA reductase1 (MA_137109g0010: AT1G15950; MA_166604g0010: AT1G15950) (Table S8), which is a known target of *A. thaliana* NST1 (Mitsuda et al., 2005). The network neighbours of VND included several xylem cysteine peptidase genes (MA_10425982g0010: AT4G35350; MA_10429635g0010: AT4G35350; MA_7787614g0010: AT1G20850) (Table S8), which have similarly been identified as targets of *A. thaliana* VND6 and VND7 (Zhong et al., 2010; Yamaguchi et al., 2011). Even though the *P. abies* VND and NST were found in the same network in NorWood, they were connected by negative correlation edges only (Fig. 3e). This was also evident from examination of the expression profiles of NST and VND (as well as their closest network neighbours) in NorWood: while the expression of the VND homologue peaked in the end of the

---

**Fig. 3** The network neighbourhood of the Class IIB NAC family indicates divergence in the regulation of cellulose synthase genes. (a) A phylogenetic tree of the Class IIB NAC family. Two of the four *Picea abies* genes were expressed in the NorWood dataset and are highlighted in blue and red. (b) The AtGenExpress Development (Schmid et al., 2005) network neighbourhood of the *Arabidopsis thaliana* NAC IIB subfamily was identified using the AtGenIE exNet tool at a context-likelihood of relatedness (CLR) threshold of 5. Secondary cell wall cellulose synthase genes (CesA4, 7 and 8) are indicated. (c) The *Populus tremula* exAtlas (Sundell et al., 2015) network neighbourhood of the *Populus trichocarpa* NAC IIB family was identified using the PopGenIE exNet tool at a CLR threshold of 5. (d) The *P. abies* exAtlas (Nystedt et al., 2013) network neighbourhood of the NAC IIB subfamily was identified using the ConGenIE exNet tool at a CLR threshold of 3.5. A lower CLR threshold was used to ensure a network of comparable size. (e) The NorWood network neighbourhood of the *P. abies* NAC IIB subfamily was identified using the ConGenIE exNet tool at a CLR threshold of 3.5. The NST and VND orthologues reside in different subnetworks connected with negatively correlated expression profiles (red links). No CesA genes were present in the neighbourhood of NST or of VND. (f) Gene expression profiles of *P. abies* VND and NST and heatmap representations of their respective positive network neighbourhood within the NorWood co-expression network (CLR threshold 5). For the VST and NST homologue, the expression profile is shown for the three replicate trees (T1–T3) where dashed vertical lines represent cambium/expanding xylem, secondary cell wall/programmed cell death and mature xylem/late wood zones. A heatmap of the corresponding network neighbourhood is shown below the expression profile plots with expression values scaled per gene so that expression values above the gene average are represented by red, and below average by blue. Samples appear in the order of sampling within each tree (T1–T3).
SCW zone, the NST homologue was highly expressed throughout the mature xylem (MX) and late wood (LW) zones (Fig. 3f). The differential expression profiles and network neighbourhoods strongly suggest separate regulatory roles for the two Class IIB NAC TFs in *P. abies*.

**Discussion**

High-spatial-resolution RNA-seq gene expression profiles were generated from replicated, tangential cryo-section series spanning the wood development zone in *P. abies* and deposited to an interactive, web-based tool – NorWood. While such large-scale RNA-seq data resources have an inherent value, their true potential only becomes evident and accessible to the wider community when they are made available in an intuitive, visual and interactive form. In addition, evo-devo insights require such data to be integrated within platforms that enable direct, comparative analyses of cross-species co-expression networks to facilitate identification of conserved and diverged regulatory mechanisms and co-expression profiles. While a number of excellent web resources are available, many have a focus on coding sequence-based comparisons. For example, the PLAZA (Proost et al., 2015) and Phytozome (Goodstein et al., 2012) resources provide extensive coverage of both angiosperm and gymnosperm (in the case of PLAZA) species, with cross-species gene family information allowing exploration of protein coding gene conservation. Similarly, a number of resources integrate gene expression and protein–protein interaction data, including PlaNet, PODC, CORNET and ATTED-II (Mutwil et al., 2011; De Bodt et al., 2012; Ohyanagi et al., 2015; Aoki et al., 2016). However, the majority of these are focused towards agricultural food crop species and *A. thaliana*. This limits their ability to provide insights into biological processes that are not well represented by, or are lacking in, these species, such as wood formation, seasonal dormancy and senescence or perennial growth. PlantGenIE (Sundell et al., 2015), by contrast, has a focus on tree species and seamlessly integrates sequence and expression information to investigate the conservation of co-expression networks, which represents a powerful means to uncover evolutionary divergence driven by underlying changes in gene co-expression modules. Here, the power of developing a gene expression resource profiling wood development across the annual growth ring of *P. abies* is demonstrated, revealing how the high spatial resolution available from sampling mature trees enables detailed insights into the expression domain of genes regulating the process of wood formation.

As conifers are long-lived and have to cope with changing environmental conditions, previous global gene expression analyses have often focused on resistance or acclimation responses or mechanisms (e.g. Verne et al., 2011; Yeaman et al., 2014). There have also been a number of transcriptome analyses of xylem tissues in various conifer species (e.g. Pavy et al., 2008; Li et al., 2011; Raherison et al., 2015), largely motivated by the ecological and economical value of wood. However, as the samples in these previous studies originated from multiple wood cell layers, it has been difficult to identify genes that are expressed during specific phases of tracheid development. In the current work, the spatial information on these phases was preserved in the form of RNA-seq data generated from tangential cryosections taken throughout an annual growth ring, which represented all transitions during tracheid development. A novel characteristic of the current data is that it includes samples from mature xylem and the previous year’s latewood, in which most of the cells (i.e. tracheids) are thought to be dead, with the exception of ray cells (Nakaba et al., 2008). Therefore, the transcriptomes of samples within the MX and LW zones are likely to be derived from those living rays cells only. In support of this, genes within cluster-f showed high expression specifically in the MX and LW zones and were enriched for GO terms congruent with the biological processes active within ray cells. Such insights have not previously been possible using low-resolution sampling, where mature wood was either omitted entirely or where the transcriptome of the mature part of the wood would have been combined with that of additional developmental zones within the developing xylem.

In angiosperms, recent studies have begun to resolve the transcriptional network controlling SCW formation, and NAC domain proteins, a plant-specific TF family, have been shown to play a central role in this process (e.g. Nakano et al., 2015; Taylor-Teeples et al., 2015). In conifers there is evidence that aspects of these regulatory mechanisms may differ: Rigault et al. (2011) investigated the distribution of protein domains between different species and showed that the NAC domain is one of the most underrepresented in *P. glauca* compared to studied angiosperm species (Rigault et al., 2011). In line with this, *P. abies* has only four genes encoding members of the Class IIB subfamily of NAC domain proteins in contrast to 13 genes in *A. thaliana* (Fig. 3a and Nystedt et al., 2013). The expression profiles and co-expression network neighbourhoods of the two Class IIB family genes expressed in the NorWood data were examined and found to include a single NST and a single VND homologue (Fig. 3). Surprisingly, the genes had distinctly contrasting expression profiles in *P. abies* wood; the expression of the VND gene peaked in the SCW zone whereas the NST gene was most highly expressed much later, in the MX and LW zones (Fig. 3f). When the co-expression network neighbourhoods of the *P. abies* NST and VND were examined in NorWood, no SCW CesA genes were found, whereas co-expression networks in both *A. thaliana* and *P. tremula* identified SCW CesA genes as neighbours of NST1 (Fig. 3b–e). Taken together, these results indicate a divergence in the regulation of SCW formation between the two angiosperm species and *P. abies* in addition to demonstrating the importance of high-spatial-resolution transcriptional data to resolve expression domains.

To provide a worked example, the NorWood resource was used to identify a set of candidate TFs representing potentially novel regulators of SCW CesA genes in conifers and to demonstrate how the integrated resources within PlantGenIE can be utilised to perform comparative regulomics analyses to explore their evolutionary conservation. Based on phylogenetic analysis
and examination of expression patterns, three bona fide *P. abies* SCW CesA genes were identified (Figs 2a, S3), in agreement with previous studies performed in other coniferous species (Nairn & Haselkorn, 2005; Raherison et al., 2012). Interestingly, the expression of two *P. abies* ANAC075 (AT4G29230) homologues correlated with expression of the SCW CesA genes in NorWood. Even though ANAC075 has not been directly shown to regulate expression of SCW CesAs, it has been demonstrated to induce ectopic differentiation of xylem vessel elements, accompanied by SCW deposition, when overexpressed in *A. thaliana* (Endo et al., 2014). Also, SND2, which is the closest homologue of ANAC075 in *A. thaliana*, is known to stimulate SCW formation in fibres (Hussey et al., 2011) and directly regulate expression of CesA8 (Zhong et al., 2008). These data support functional conservation of the NAC factors, such as ANAC075 and SND2, between angiosperms and gymnosperms while the Class IIB NAC factors seem to have gained central function in regulation of SCW biosynthesis only after divergence of the angiosperms and gymnosperms. Furthermore, the fact that the moss *Physcomitrella patens* possesses Class IIB NAC TFs that are functionally conserved between *P. patens* and *A. thaliana* (Xu et al., 2014) supports a view that the Class IIB NAC TFs obtained new functionalities in the angiosperms through changes in the gene expression domain rather than the protein function. Further high-spatial co-expression analyses as well as functional assays of these genes across the kingdoms are needed to confirm the evolutionary history of SCW regulation.

Here, use of the NorWood web resource is exemplified to provide novel evo-devo insights into the divergent regulation of developmental process during wood formation between *P. abies*, as a representative conifer, and the angiosperm model species *A. thaliana* and *Populus* species. The resource represents a powerful tool for hypotheses generation for subsequent validation studies and for advancing understanding of wood formation in coniferous species in addition to enabling evo-devo focused analyses. In specific cases, such comparative genomics approaches can hold particular value where candidate genes are essential, and therefore intractable to knock-out approaches, for example. More importantly, they enable discovery of novel candidates that would not be considered given current understanding.

Acknowledgements

We thank Dr Jakob Prestele and Veronica Bourquin for technical help. We thank Chanaka Mannapperuma for designing the NorWood logo. We are grateful for support from the UPSC bioinformatics platform and Nicolas Delhomme. This work was supported by funds from the Alice Wallenberg Foundation, the Kempe Foundation (SMK-1340), the Swedish Research Council VR (621-2013-4949), the Swedish Governmental Agency for Innovation Systems Vinnova (2015-02290), the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning Formas (232-2009-1698), and the Swedish Foundation for Strategic Research (SSF). N.R.S. is supported by the Trees and Crops for the Future (TC4F) project.

Author contributions

S.J.-L. performed sample collection and preparation and contributed to manuscript writing and interpretation of the results. D.S. performed the analysis of the RNA-seq data, designed and implemented the NorWood resource and contributed to writing the manuscript. O.N. contributed to the experimental design and data interpretation. T.R.H. performed comparative regulomics analyses, contributed to the design of the NorWood resource, interpretation of the results and manuscript writing. N.R.S. contributed to data interpretation and manuscript writing. H.T. conceived and implemented the study and contributed to manuscript writing.

References


**Supporting Information**

Additional Supporting Information may be found online in the Supporting Information tab for this article:

**Fig. S1** Principal component analysis plot of normalised expression values.

**Fig. S2** Gene expression heatmap indicating sample and gene clustering.

**Fig. S3** Midpoint rooted phylogenetic tree of *Picea abies*, *Arabidopsis thaliana* and *Populus trichocarpa* cellulose synthase (*CesA*) and *CesA-like* (*Csl*) subfamilies B, D, E and G.

**Fig. S4** Co-expression among primary and secondary cell wall *cellulose synthase* genes.

**Notes S1** Supplementary figures, genotype confirmation for biological replicates, phylogenetic analysis and gene model annotation refinement for *cellulose synthase* family members.

**Table S1** Section pooling and cell profiling

**Table S2** Gene expression and sequence data quality

**Table S3** Hierarchical clustering of genes

**Table S4** Gene ontology enrichment of hierarchical clusters

**Table S5** Gene ontology enrichment per sample

**Table S6** Network statistics

**Table S7** Transcription factor network neighbours of the secondary cell wall *cellulose synthase* genes in *Picea abies*

**Table S8** Secondary cell wall *cellulose synthase* A gene network neighbourhoods

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.