Comparative interrogation of the developing xylem transcriptomes of two wood-forming species: *Populus trichocarpa* and *Eucalyptus grandis*

Charles A. Hefer[1], Eshchar Mizrachi[2], Alexander A. Myburg[2], Carl J. Douglas[1] and Shawn D. Mansfield[3]

[1]Department of Botany, University of British Columbia, Vancouver, BC V6T 1Z4, Canada; [2]Department of Genetics, Forestry and Agricultural Biotechnology Institute (FABI), Genomics Research Institute (GRI), University of Pretoria, Private bag X20, Pretoria 0028, South Africa; [3]Department of Wood Science, Faculty of Forestry, University of British Columbia, Forest Sciences Centre, 4030-2424 Main Mall, Vancouver, BC V6T 1Z4, Canada

Author for correspondence:
Shawn D. Mansfield
Tel: +1 604 822 0196
Email: shawn.mansfield@ubc.ca

Key words: *Eucalyptus grandis*, gene expression, mRNA-Seq, *Populus trichocarpa*, secondary cell wall, xylem transcriptome.

Summary

• Wood formation is a complex developmental process governed by genetic and environmental stimuli. *Populus* and *Eucalyptus* are fast-growing, high-yielding tree genera that represent ecologically and economically important species suitable for generating significant lignocellulosic biomass.

• Comparative analysis of the developing xylem and leaf transcriptomes of *Populus trichocarpa* and *Eucalyptus grandis* together with phylogenetic analyses identified clusters of homologous genes preferentially expressed during xylem formation in both species.

• A conserved set of 336 single gene pairs showed highly similar xylem preferential expression patterns, as well as evidence of high functional constraint. Individual members of multigene orthologous clusters known to be involved in secondary cell wall biosynthesis also showed conserved xylem expression profiles. However, species-specific expression as well as opposite (xylem versus leaf) expression patterns observed for a subset of genes suggest subtle differences in the transcriptional regulation important for xylem development in each species.

• Using sequence similarity and gene expression status, we identified functional homologs likely to be involved in xylem developmental and biosynthetic processes in *Populus* and *Eucalyptus*. Our study suggests that, while genes involved in secondary cell wall biosynthesis show high levels of gene expression conservation, differential regulation of some xylem development genes may give rise to unique xylem properties.

Introduction

Xylem formation is a complex process that manifests from a range of cellular, molecular and developmental processes (Plomion *et al.*, 2001). The progression is influenced by external factors, such as photoperiod, nutrient availability, moisture content and temperature, and subject to internal stimuli such as phytohormones and photosynthate assimilation, and their interaction. Several thousand genes are actively involved in xylem formation and have been identified in a range of herbaceous and woody species (Prassinos *et al.*, 2005; Andersson-Gunnerås *et al.*, 2006; Mizrachi *et al.*, 2010; Rigault *et al.*, 2011; Pavy *et al.*, 2012). Despite these substantive efforts, a complete understanding of the genes and their intricate interactions, and the regulatory circuitry involved in the control of xylem formation is still lacking (Hussey *et al.*, 2013). The use of comparative studies to identify genes that are functionally conserved between diverse species sharing the woody phenotype will expand our understanding of the regulatory, metabolic and biosynthetic processes underlying xylem formation. Next-generation DNA sequencing technologies have facilitated the development of genomic resources for diverse species, enabling such comparative studies (Davidson *et al.*, 2012).

The ability to produce secondary xylem has been independently lost and gained several times in the angiosperm lineage, supporting the hypothesis that the key genes required for secondary growth are conserved among angiosperms (Kirst *et al.*, 2003; Groover, 2005; Déjardin *et al.*, 2010; Spicer & Groover, 2010; Lens *et al.*, 2012) as well as between angiosperms and gymnosperms (Pavy *et al.*, 2008). This implies that the conversion between herbaceous and woody phenotypic states depends on the ability to regulate biological processes integral to xylem development. Several comparative studies of the herbaceous plant *Arabidopsis thaliana* (Arabidopsis) and the model tree *Populus trichocarpa* have identified functional orthologs for cell wall-related transcription factors (Zhong & Ye, 2007; McCarthy *et al.*, 2012), genes active in cellulose deposition (Djerbi *et al.*, 2005), and genes active in carbohydrate synthesis and metabolism (Geisler-Lee *et al.*, 2006) and cell wall lignification (Vanholme *et al.*, 2013). Perhaps the most compelling evidence

for the level of plasticity in the pathways controlling a woody phenotype is the description of an Arabidopsis *SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1* (*SOC1*) and *FRUITFULL* (*FUL*) double mutant (described by Melzer *et al.*, 2008) which displays significant secondary xylem formation ('woody phenotype'). The identification of functional orthologs in herbaceous plants and species characterized by secondary growth such as *Populus* spp. and *Eucalyptus* spp. will ultimately aid in unraveling the molecular underpinnings of xylem formation and woody perennial growth.

Plant secondary cell walls consist mainly of cellulose, hemicellulose and lignin, and constitute the vast majority of utilizable biomass from trees. The physicochemical properties of xylem can be directly or indirectly attributed to the ratio of cellulose, hemicelluloses and lignin within the cell wall. For bioethanol production, for example, it has been shown that these biopolymers and their molecular interactions have a direct effect on the recalcitrance of the cell wall to enzymatic degradation and the release of sugars (Abramson *et al.*, 2010; Langan *et al.*, 2011; Studer *et al.*, 2011; Mansfield *et al.*, 2012). Several transcription factors have recently been shown to be directly involved in regulating secondary cell wall deposition, primarily in Arabidopsis (see reviews by Schuetz *et al.*, 2012 and Hussey *et al.*, 2013). These include a suite of *NAM/ATC/CUC* (*NAC*) domain proteins, encoded by *VASCULAR-RELATED NAC-DOMAIN6* (*VND6*), *VND7*, *SECONDARY WALL-ASSOCIATED NAC DOMAIN PROTEIN1* (*SND1*), and *NAC SECONDARY WALL THICKENING PROMOTING FACTOR1* (*NST1*) which are key regulators of secondary wall biosynthesis in different cell types (Kubo *et al.*, 2005; Mitsuda *et al.*, 2005; Zhong *et al.*, 2006; Yamaguchi *et al.*, 2008). Additionally, several *MYELOBLASTOSIS* (*MYB*) domain-containing transcription factors (e.g. *MYB46*, *MYB83*, *MYB58* and *MYB63*), *THREE-AMINO ACID LOOP EXTENSION* (*TALE*) and the homeodomain protein gene *KNOTTED ARABIDOPSIS THALIANA7* (*KNAT7*) act downstream or in parallel to regulate specific aspects of secondary wall biosynthesis (Zhong *et al.*, 2008; Ko *et al.*, 2012; Li *et al.*, 2012) and have been shown to be actively involved in regulating the deposition of the secondary cell walls (for a review of regulons involved in secondary cell wall deposition, see Hussey *et al.*, 2013).

Cellulose is synthesized by the concerted action of the *CELLULOSE SYNTHASE* (*CESA*) proteins, which form part of protein complexes in the plasma membrane that produce glucan chains coalescing to form cellulose microfibrils in primary and secondary cell walls (for a recent review of cellulose synthesis, see Mizrachi *et al.*, 2011). The transcription factors *SND2*, *SND3* and *MYB103* have been suggested to influence cellulose deposition (Zhong *et al.*, 2008) and also other related aspects of hemicellulose and lignin biosynthesis (Hussey *et al.*, 2013; Öhman *et al.*, 2013). The function of the *CESA* rosette complex depends on a multitude of other proteins, including but not limited to proteins originally identified in Arabidopsis encoded by the *COBRA* (*COB*), *COBRA-LIKE* (*COBL*), *KORRIGAN* (*KOR*), *SUCROSE SYNTHASE* (*SUSY*), *FASCICLIN-LIKE ARABINOGALACTAN* (*FLA*), *GERMIN-LIKE* and *CHITINASE-LIKE* (*CTL*) genes (also see Jansson & Douglas, 2007; Mizrachi *et al.*, 2011).

The availability of reference genomes for two angiosperm tree species that are important ecologically and commercially for their woody biomass, *Populus trichocarpa* (poplar) (Tuskan *et al.*, 2006) and *Eucalyptus grandis* (eucalyptus, eucalypts) (Myburg *et al.*, 2014), facilitates comparative studies that have the potential to provide insights into the conservation of wood development, and to identify conserved genes key to secondary cell wall biosynthesis. Whole transcriptome sequencing (mRNA-seq) is a powerful gene expression profiling approach, especially when characterizing transcripts in annotated reference genomes. Using this approach, Bao *et al.* (2013) identified over 30 000 genes (*c.* 75% of the genes annotated on the version 2.2 *Populus trichocarpa* reference genome) with evidence of active transcription in developing xylem of 20 unrelated individuals. With the availability of the *Eucalyptus grandis* genome (Myburg *et al.*, 2014), similar studies are now possible in *Eucalyptus* species, as well as comparative studies of orthologous gene expression in these two unrelated woody species with over 90 Myr of independent evolution.

In this study, we performed an in-depth analysis of the developing xylem and leaf transcriptomes of *P. trichocarpa* and *E. grandis*, using RNA-seq as a measure of transcript abundance in material isolated from multiple individuals. Our first objective was to perform genome-wide identification of clusters of xylem-expressed gene orthologs and paralogs shared between *P. trichocarpa* and *E. grandis* based on sequence similarity. Secondly, we identified genes differentially expressed in developing xylem and leaf tissue with a preference for xylem tissue in both species. Using the ortholog information, we investigated the prevalence of differential expression within clusters, and comment on the effect of cluster size on xylem and leaf gene activity. Finally, we used the cluster information to identify orthologs with conserved differential expression status in *P. trichocarpa* and *E. grandis*.

Materials and Methods

RNA-seq and gene expression

Developing stem xylem and leaf samples were collected from 3-yr-old ramets of two *Eucalyptus grandis* (W. Hill ex Maiden) clones (two ramets of TAG0014 and one ramet of TAG0079; Mondi Tree Improvement Research, Kwambonambi, KwaZulu-Natal, South Africa; 28°35′S, 32°12′E) grown in a field trial established on a flat coastal site (elevation 55 m) with deep (> 1.5 m effective rooting), uniform, well-drained sandy soils, and a mean annual rainfall of 1201 mm and mean temperature of 21°C. Sample collection was performed in late summer (1 April 2009) from actively growing trees. Similarly, developing stem xylem and leaf samples were collected from three unrelated 3-yr-old *Populus trichocarpa* (Torr. & Grey) clones (GLCA-26-1, MCGR-15-6 and VNDL-27-4) from the BC Ministry of Forests collection (the population was previously described by Geraldes *et al.*, 2011) grown in the UBC Totem experimental field (49°15′N, 123°14′W), which consists of a flat, uniform site

(80 m elevation) with moderate rooting depth (50–65 cm effective rooting) composed of well-drained upland loamy sand formed from a glacial till. The mean annual rainfall was 1118 mm with a mean annual temperature of 11°C. Immature xylem (outer glutinous 1 to 2-mm layer comprising developing xylem cell layers) was collected from breast height (1.35 m) on the main stem following bark removal. Young, rapidly expanding leaves (three to four nodes below the shoot tips) were collected from the crowns of the trees at the same time.

Transcriptome sequences were obtained by performing Illumina (Illumina Inc., San Diego, CA, USA) mRNA-seq sequencing on the collected samples following the RNA isolation protocols as described by Geraldes *et al.* (2011). Information regarding the sequencing libraries is available in Supporting Information Notes S1, and the short read data have been submitted to the Sequence Read Archive under the project ID #SRP050172.

The version 3.0 reference genome assembly for *P. trichocarpa* and the version 1.0 reference genome assembly for *E. grandis*, and the corresponding gene models were retrieved from Phytozome (http://www.phytozome.org). A total of 41 355 primary gene transcripts (gene loci) were annotated on the *P. trichocarpa* genome assembly, and 36 376 loci were annotated on the *E. grandis* assembly (version 1.1 annotation).

Gene expression values for each species were calculated by aligning the mRNA-seq reads to the respective reference genomes (using TopHat version 2.0.8 and allowing for two mismatches during alignment; Trapnell *et al.*, 2009), and differentially expressed genes (xylem/leaf, q-value < 0.05) were identified with CuffDiff (version 2.2.1; Roberts *et al.*, 2012) using the respective reference transcriptomes.

Clusters of orthologs and paralogs

Clusters of gene orthologs and paralogs were identified with OrthoMCL (version 2.0.3; Li, 2003), using a minimum input protein length of 10 amino acids, no e-value cut-off in the protein BLAST step, and an Markov Cluster Algorithm inflation value of 1.5 for clustering. OrthoMCL clusters were assigned to different classes, depending on the species composition and the number of genes within a cluster. The class designation reflects the number of gene copies in the species. For example, single copy genes with orthologs in both species were assigned to the '1' copy, '2' species ('C×S') cluster class (i.e. '1 × 2'). Clusters were similarly designated as containing multi-copy ('N') genes from a single species ('$N\times 1$', referring to paralogous genes in a single species), or containing multi-copy gene paralogs with orthologs in both species ('$N\times 2$'). Genes identified as single copy in either species, without identifiable paralogs in the same species or orthologs in the other species, were assigned to the '1 × 1' singleton cluster class.

For this analysis, genes assigned to cluster class '1 × 2' (single copy gene in both species) were considered to be orthologs of each other in the different species. Genes in clusters containing more than one member from both species (cluster class '$N\times 2$') were used to construct neighbor-joining (NJ) trees (using clustalw version 2.1; Larkin *et al.*, 2007) employing all protein sequences within the cluster (the alignment of protein sequences was performed with muscle version 3.8.31 (Edgar, 2004), and bootstrap values were calculated for the resulting trees). Combining NJ tree topology and expression data, putative functional co-expressed orthologs were identified within the cluster. Genes were considered to be functional co-expressed orthologs when genes from both species shared the same differential (xylem/leaf) expression status, and shared a bootstrap supported (> 50%) common ancestral node within the constructed phylogenetic tree.

Functional annotation

Gene ontology (GO) enrichment analyses and metabolic pathway analyses were performed with the GOstats R-package (Falcon & Gentleman, 2007). Biological process terms were used to functionally annotate gene lists. The P-value cut-off for GO term over-representation was set to 0.01. Redundant and low-level GO terms were replaced by higher level, semantically similar GO-slim terms using the REViGO web service (Supek *et al.*, 2011), and gene ontology graphs were visualized in cytoscape (version 2.8.2, www.cytoscape.org). Over-represented pathways were identified with a P-value cut-off of 0.05. Metabolic pathways were manipulated using the MapMan package (version 3.6.0RC1; Thimm *et al.*, 2004). Protein domains were identified by performing searches against the InterProScan 4 database hosted at the European Bioinformatics Institute (EBI; https://www.ebi.ac.uk/Tools/pfa/iprscan).

Test for genes under selection

To test for selective pressure on orthologous gene pairs (evolutionary rates), we calculated the rate of synonymous to nonsynonymous mutations in the DNA sequences in the two species ($K_a : K_s$ ratio). Pairwise alignment of the coding DNA sequences (excluding annotated untranslated region (UTR) sequences) was employed to calculate the $K_a : K_s$ ratio of sequence divergence between orthologs with the yn00 program from paml (Yang, 2007). Pairs of orthologs were binned into groups of low ($K_a : K_s < 0.05$), medium ($0.05 \le K_a : K_s < 0.15$), and high ($K_a : K_s \ge 0.15$) values.

Results

Clusters of orthologs and paralogs

Protein sequence clustering identified 15 925 clusters of orthologs and paralogs within and between the species (all cluster types; Table S1). In total, these clusters contained 59 892 sequences (77.07%) from the combined *P. trichocarpa–E. grandis* protein data set consisting of 77 711 sequences. Of these assigned proteins, a total of 47 489 proteins were assigned to either cluster class '$N\times 2$' or '1 × 2', containing at least one sequence from both species (12 619 clusters in total). A subset of these clusters (4622 clusters) was defined as having a single ortholog in both species, identifying 9244 genes as single copy orthologs between the species (Fig. S1; Tables 1, S1).

Table 1 Distribution of differentially expressed (DE) genes across orthologous cluster classes in *Populus trichocarpa* and *Eucalyptus grandis*

	Populus trichocarpa					*Eucalyptus grandis*				
Cluster class	Genes	DE xylem	DE leaf	Trimmed mean DE xylem FPKM[1]	Trimmed mean DE leaf FPKM[1]	Genes	DE xylem	DE leaf	Trimmed Mean DE xylem FPKM[1]	Trimmed Mean DE leaf FPKM[1]
1 × 1	8704	534	1315	22.40	58.56	9115	983	905	45.37	20.41
1 × 2	4622	601	1664	34.02	26.25	4622	788	1038	49.22	19.94
N × 1	7087	565	1134	25.61	21.03	5116	375	594	22.95	11.34
N × 2	20 922	4050	4894	30.30	20.30	17 523	3157	3050	54.47	21.58

In *Populus trichocarpa*, 49.0% and 42.7% of genes in cluster class '1 × 2' and cluster class '*N* × 2' were differentially expressed (*q*-value < 0.05), compared with 21.1% and 24.0% in cluster classes '1 × 1' and '*N* × 1', respectively. Similarly, 39.50% and 35.40% of eucalypt genes in cluster class '1 × 2' and '*N* × 2' were differentially expressed in *Eucalyptus grandis*, compared with 20.71% and 18.94% in cluster classes '1 × 1' and '*N* × 1', respectively. Trimmed mean values were calculated by removing the highest and lowest 10% of the values.
[1]Fragments per kilobase per million mapped reads.

The clustering algorithm failed to identify orthologs for 7087 paralogous poplar genes (cluster class '*N* × 1') in eucalyptus, and these were considered to be poplar specific. Similarly, there were 5116 eucalypt genes assigned to eucalypt-specific paralog-containing (cluster class '*N* × 2') clusters (Fig. S1; Tables 1, S1).

The remaining genes, *c.* 23% of the total (Fig. S1), were not assigned to clusters of paralogs/orthologs (singleton cluster class '1 × 1'). This set contained single copy genes without paralogs within a species and without detectable orthologs in the other species. Functional analysis (biological process GO terms) of genes found in the '1 × 1' cluster class revealed an over-representation of terms associated with organellar organization. Specific biological process terms included 'translation', 'response to oxidative stress' and 'protein–DNA complex subunit organization' in the poplar data set, and 'DNA-conformational change', 'nitrogen compound metabolism' and 'protein–DNA complex subunit organization' in the eucalypt data set (Table S2). The '1 × 1' cluster class probably represents genes that have been alternatively lost and retained and may contribute to unique aspects of *E. grandis* and *P. trichocarpa* biology.

Genes preferentially expressed in secondary xylem tissue

The cluster analyses were used to investigate xylem preferential expression in potential orthologs common to both species. To address the question of their functional conservation, we queried the RNA-seq data sets from both species (see the Materials and Methods section) for genes with similar expression patterns in developing xylem and leaves. Differential expression analysis (xylem/leaf) identified a total of 14 757 poplar and 10 890 eucalypt genes as differentially expressed (DE; *q*-value ≤ 0.05; Table S1). Of these, 5750 poplar and 5303 eucalypt genes showed preferential expression in developing xylem tissue. Functional annotation of the xylem preferentially expressed genes in both species revealed that similar GO terms were enriched in these data sets (*P*-value ≤ 0.01) including those associated with 'cellular localization', 'transport', 'metabolism', and 'macromolecule biosynthesis' (Fig. 1).

In cluster classes containing orthologous genes from both species (cluster classes '1 × 2' and '*N* × 2'), 49.0% and 42.74% of the poplar genes, respectively, were identified as differentially expressed. Only 21.12% of genes considered to be singletons in poplar (i.e. cluster class '1 × 1' lacking eucalypt orthologs), and 23.97% of poplar genes in cluster class '*N* × 1' lacking eucalypt orthologs, were preferentially expressed in xylem tissue. Similarly, 39.5% ('1 × 2') and 35.4% ('*N* × 2') of the eucalypt genes in multi-species clusters were preferentially expressed, compared with 20.71% and 18.94% of the genes in *E. grandis*-specific species clusters '*N* × 1' and '1 × 1' (Table 1; Fig. 2).

A total of 5216 (90.71%) of the poplar and 4320 (81.46%) of the eucalypt xylem preferentially expressed genes were assigned to clusters containing orthologs from both species. Of these, 1389 genes were assigned to clusters with only one sequence from each species (cluster class '1 × 2'), and were thus considered to be orthologous sequences in the two species. Of these 1389 genes with orthologs in both species, 336 pairs showed xylem-preferred expression in both species, and thus 672 co-expressed orthologs were identified (Table S3). High levels of xylem expression correlation ($r = 0.85$) were observed for these 336 pairs of co-expressed orthologs. Functional annotation of this set of orthologous genes identified terms associated with the 'regulation of cellular catabolism', 'vesicle-mediated transport', 'localization' and 'actin cytoskeleton organization' (Table S4), highlighting the conserved roles of these cellular processes in xylem development.

Using differential expression status as a filter (xylem and leaf preferential expression), we identified 121 pairs of orthologs in cluster class '1 × 2' where a switch in tissue preference was detected between species; that is, the genes were preferentially expressed in xylem of one species, but in leaves of the other species (Table S5). Of these, 32 *P. trichocarpa* genes preferentially expressed in xylem tissue show a preference for leaf expression in *E. grandis*. Among these are clusters associated with carbohydrate metabolism and transport (cluster PtrEgr13859 containing galactose mutarotase-like superfamily genes, cluster PtrEgr14879 containing glycosyl hydrolase family 10 genes, and cluster PtrEgr14210 containing sucrose transporters) and a set of Nodulin MtN3 family protein orthologs (cluster PtrEgr15537) previously associated with sucrose response in Arabidopsis. Similarly, 89 *E. grandis* genes with preferential expression in the xylem showed a preference for leaf expression in *P. trichocarpa*. A total

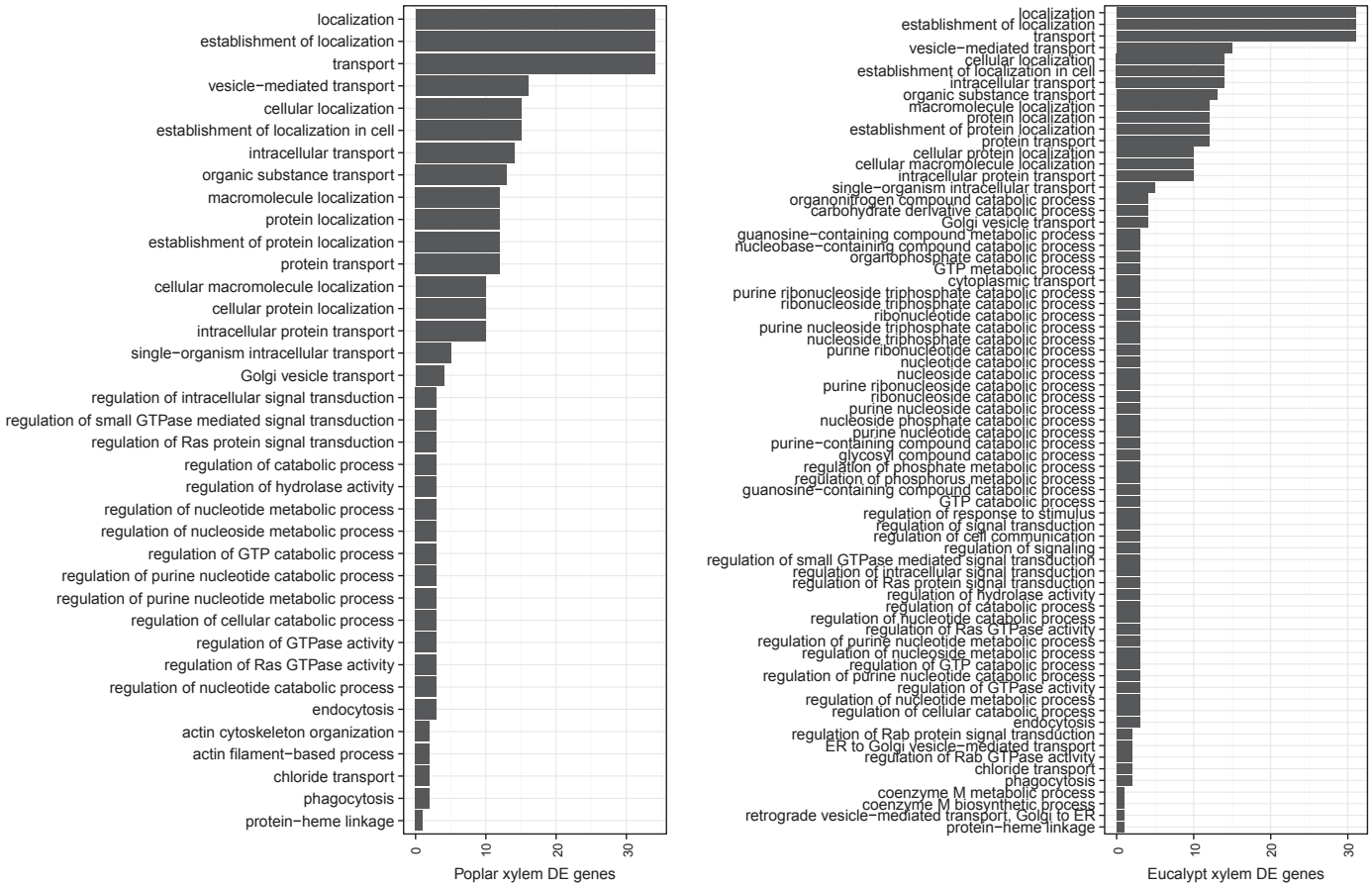

Fig. 1 Biological process terms associated with the genes differentially expressed (DE) in xylem tissue. Gene ontology (GO) terms enriched (*P*-value ≤ 0.01) in the poplar (left) and eucalypt (right) xylem preferentially expressed data sets are shown. The most prevalent GO terms in both species were associated with cellular localization, transport, metabolism, cell signalling and macromolecule biosynthesis.

of six clusters (PtrEgr14236, PtrEgr16802, PtrEgr15473, PtrEgr13277, PtrEgr15140 and PtrEgr15969) consisting of sequences containing domains of unknown function (*DUF*) were preferentially expressed in *E. grandis* xylem, while preferentially in *P. trichocarpa* leaf tissue (including the domains *DUF1995*, *PUF1589*, *DUF581*, *DUF1677*, *DUF593* and *DUF3133*). The list also contained cluster PtrEgr15077, which contains two *WRKY* DNA-binding domain transcription factors (*WRKY70*). The Arabidopsis ortholog of this cluster was identified as AT3G56400, a transcription factor involved in disease responses (Knoth *et al.*, 2007). The eucalypt ortholog of the *WRKY70* domain-containing sequence, Eucgr.G03145, had higher expression in xylem tissue (log2fold change = 1.37), while the poplar ortholog, Potri.013G090300, had higher expression in leaf tissue (log2-fold change = 1.49). Another gene encoding a *BASIC LEUCINE ZIPPER* transcription factor (Eucgr.02341, log2-fold change = 3.32; Potri.008G118300, log2-fold change = −3.00) also displayed a difference in tissue specificity between *E. grandis* (xylem-preferred) and poplar (leaf-preferred).

In the grouping consisting of multiple genes from both species (cluster class '$N \times 2$'), several clusters were identified that only contain differentially expressed genes from a single species (Table S6). These included a cluster of heat-shock protein 20 (HSP20) -like chaperones (cluster PtrEgr1185, with 18 genes preferentially expressed in *E. grandis* xylem), *LATE EMBRYOGENESIS ABUNDANT* (*LEA*) hydroxyproline-rich glucoproteins (cluster PtrEgr1086, with 15 *E. grandis* genes preferentially expressed in xylem) and genes associated with heavy metal transport and detoxification (cluster PtrEgr1091, with 12 *E. grandis* genes preferentially expressed in xylem).

Cluster class $N \times 2$ also contained genes where at least one gene from each species was preferentially expressed in the xylem (Table 2). The cluster with the most differentially expressed genes (PtrEgr1042) consists of 50 *LACCASE* (*LAC1*, *LAC2*, *LAC11* and *LAC17*) and LACCASE/DIPHENOL OXIDASE, of which 19 poplar and 10 eucalypt genes were differentially expressed in xylem tissue. Other clusters containing known gene families associated with secondary cell wall formation and xylem development included a cluster of cellulose synthase-like genes (PtrEgr1038, consisting of 53 genes of which nine *E. grandis* and nine *P. trichocarpa* genes were preferentially expressed in xylem), tubulin-related genes (PtrEgr1115, containing 27 genes, of which five *E. grandis* and 16 *P. trichocarpa* genes were preferentially expressed in xylem), and a cluster of actin genes (PtrEgr1264,

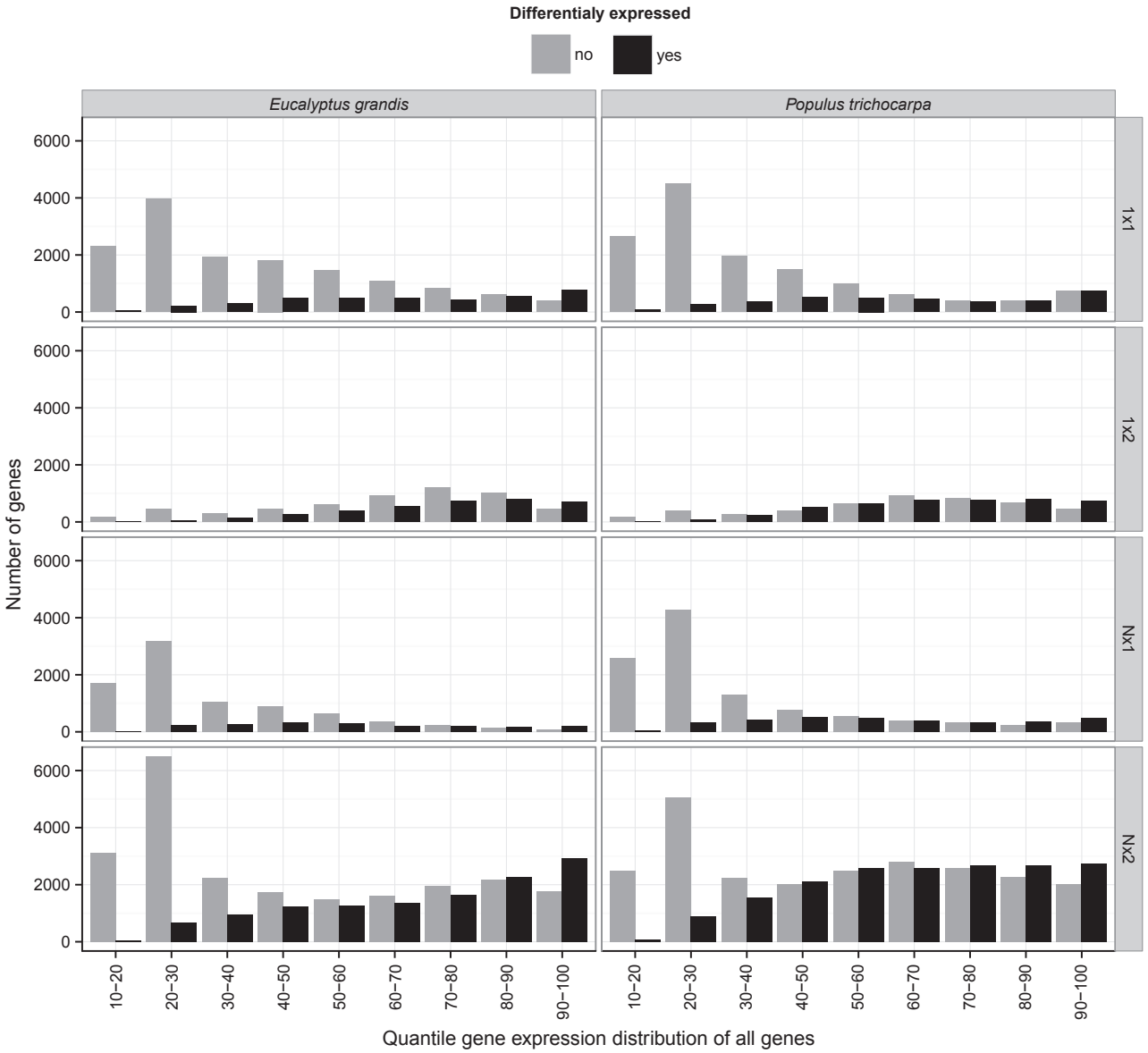

Fig. 2 Distribution of differentially expressed genes across the different orthologous cluster classes. The fragments per kilobase per million mapped reads (FPKM) distribution is divided among quantiles, with the number of genes within each FPKM quantile presented on the *x*-axis for both *Eucalyptus grandis* and *Populus trichocarpa*. Black bars indicate genes differentially expressed in either xylem or leaf, and gray bars indicate nondifferentially expressed genes. Differentially expressed genes are enriched in clusters '1 × 2' and '*N* × 2' that contain orthologs in both species. Note the low number of non-differentially expressed genes in the conserved single ortholog ('1 × 2') cluster class compared with other classes with species-specific or paralogous genes.

containing 16 genes, of which seven *E. grandis* and 16 *P. trichocarpa* genes were preferentially expressed in xylem; Table 2).

To identify functional orthologs of multi-member clusters in cluster class '*N* × 2', neighbor-joining trees were generated for all members in each cluster. The phylogenetic trees were examined for cases where potential orthologous pairs of sequences were differentially expressed in both species (see the Materials and Methods section). A total of 2109 clusters were identified containing at least one set of putative functional orthologs that were differentially expressed (Table S7). Functional annotation of the 2004 and 1486 *P. trichocarpa* and *E. grandis* genes present in these clusters, respectively, revealed GO terms associated with 'vesicle-mediated transport', 'intracellular signal transduction', 'cellular polysaccharide metabolism', 'purine-containing compound metabolism' and 'glycosyl compound metabolism' in poplar. Similar biological process terms were over-represented in the eucalypt data set, including 'microtubule-based processes', 'protein glycosylation', 'protein polymerization' and 'carbohydrate metabolism' (Table S8). A correlation value of $r = 0.563$ was observed for xylem expression between the pairs of putative functional orthologs. These gene clusters probably represent gene families that have been expanded in both species and undergone functional diversification while retaining at least one xylem preferentially expressed gene in each species.

Table 2 The 20 largest paralogous clusters (cluster class $N \times 2$) with the most differentially expressed (DE) genes (xylem preference)

Cluster ID	Cluster size	DE eucalypt genes	DE poplar genes	Cluster description
PtrEgr1042	50	10	19	Laccase
PtrEgr1115	27	5	16	Tubulin
PtrEgr1038	53	9	9	Cellulose synthase
PtrEgr1043	50	9	9	Subtilase family
PtrEgr1001	295	13	5	Disease resistance protein (*TOLL-INTERLEUKIN RECEPTOR-LIKE, NUCLEOTIDE BINDING SITE, LEUCINE-RICH REPEAT DOMAINS* class)
PtrEgr1185	20	18	0	HSP20-like chaperones
PtrEgr1086	32	15	0	Late embryogenesis abundant (LEA) hydroxyproline-rich glycoprotein family
PtrEgr1023	70	6	9	P-glycoprotein
PtrEgr1004	216	6	8	Leucine-rich repeat receptor-like protein kinase
PtrEgr1036	55	4	9	Potassium transporter family
PtrEgr1148	23	7	6	Heat-shock proteins
PtrEgr1407	12	4	8	Homeobox-leucine zipper family
PtrEgr1052	43	5	7	Galactosidase family
PtrEgr1264	16	7	5	Actin
PtrEgr1091	31	12	0	Heavy metal transport/ detoxification superfamily
PtrEgr1122	26	5	6	Myosin-like
PtrEgr1472	11	4	7	Tetratricopeptide repeat (TPR)-like superfamily
PtrEgr1032	62	9	2	Autoinhibited Ca^{2+}-ATPase
PtrEgr1000	307	3	8	S-locus lectin protein kinase family
PtrEgr1099	29	4	7	β-D-xylosidase

Laccases, tubulin and cellulose synthase protein families were among the most abundant xylem preferentially expressed genes. Species-specific xylem preferential expression was observed for heat-shock protein-like chaperones (HSP20), hydroxyproline-rich glycoproteins and heavy metal transport proteins.

Genes identified as xylem preferentially expressed orthologs (the 672 conserved single copy genes identified in cluster class '1×2', as well as the 2004 *P. trichocarpa* and 1486 *E. grandis* genes from multi-gene cluster class '$N \times 2$') were combined into a data set of co-expressed putative orthologs. In total, this data set comprised 2445 xylem-expressed ortholog pairs between species, with some genes belonging to more than one pair (Table S9). Metabolic pathways over-represented in the list of orthologous genes include the 'amino acid and nucleotide sugar metabolism' pathway, 'glycolysis/glyconeogenesis', and sugar metabolism and conversion pathways (Kyoto Encyclopedia of Genes and Genomes pathways 0040 and 0051). Within the phenylpropanoid biosynthetic pathway, the lignin biosynthetic pathway (pathway 00940) was significantly over-represented in *P. trichocarpa*, but not in *E. grandis* (Table 3).

Protein family domain analysis identified a total of 976 and 956 domains in the list of poplar and eucalypt xylem preferential expressed orthologs, respectively. The most abundant domains, Trp-Asp or WD40 repeat domains (protein family id PF00400), *PKINASE* (PF00069) and *LEUCINE-RICH REPEAT* (*LRR_8* (PF13855) and *LRR_1* (PF00560)) were shared between species (Fig. 3a,b). Several proteins containing transcription-factor DNA-binding sites, such as *MYB* DNA binding sites (PF00249), zinc finger domains (PF13639, PG13912) and homeobox domains (PF00046), were also encoded by xylem preferential genes in both species, consistent with the conserved nature of transcriptional regulation in secondary xylem tissues (Zhong *et al.*, 2010).

To test for selection pressure as a result of functional constraint, we measured the ratio of synonymous to nonsynonymous nucleotide substitution rates ($K_a : K_s$) in the coding regions of the pairs of xylem co-expressed orthologs using single nucleotide polymorphisms (SNPs) identified in these genes after sequence alignment. This analysis identified many genes under strong selective constraints. A total of 656 genes had evidence for strong purifying selection ($K_a : K_s$ ratio < 0.1), compared with 1288 genes with medium ($K_a : K_s$ ratio 0.1–0.2) and 448 genes with high ($K_a : K_s$ ratio > 0.2) $K_a : K_s$ ratios (Table S10). Consistent with their functional conservation in *E. grandis* and *P. trichocarpa*, none of these genes were found to be under strong positive selection (maximum $K_a : K_s$ was 0.428).

Species-specific gene expression

The above analysis revealed a set of 4162 genes considered to be functional orthologs in the two species (Table S9). Genes in cluster class '$N \times 2$' for which no functional orthologs were detected (based on phylogeny and expression), as well as genes in cluster class '$N \times 1$', and all genes not assigned to clusters (singleton cluster class '1×1') were considered to exhibit species-specific gene expression patterns. A total of 3410 poplar and 3481 eucalypt genes preferentially expressed in xylem were therefore considered to have no functional orthologs between these species (Table S11).

In poplar, a prominent cluster of genes considered to have species-specific xylem preferential expression encoded *FASCICLIN-LIKE ARABINOGALACTAN* proteins (*FLA*-like *AGP*; PtrEgr1159). Of the 22 poplar genes present in the cluster, 15 were preferentially expressed in the xylem (no eucalypt genes are present in cluster PtrEgr1159). Thirteen of the *LAC* proteins present in the PtrEgr1042 group (the cluster with the most differentially expressed paralogs; Table 2) did not have a xylem preferentially expressed ortholog in *E. grandis*. Clusters of tubulin-related genes, heat-shock proteins, and several kinases contained xylem preferentially expressed poplar genes, but again no eucalypt genes. Another cluster of poplar-specific xylem genes (PtrEgr1138) consisted of 24 aminotransferase-like genes, of which six were preferentially expressed in xylem, and two preferentially expressed in leaf tissue.

In *E. grandis*, the cluster with the most differentially expressed genes encoded HEAT-SHOCK (HSP20-like chaperone)

Table 3 Kyoto Encyclopedia of Genes and Genomes metabolic pathways over-represented among genes with xylem preferentially expressed orthologs in *Eucalyptus grandis* and *Populus trichocarpa*

KEGG pathway	*P. trichocarpa* *P*-value	*E. grandis* *P*-value	Term
04145	1.2E-10	3.7E-10	Phagosome
00520	3.9E-08	4.0E-06	Amino sugar and nucleotide sugar metabolism
04144	1.1E-06	5.8E-07	Endocytosis
04130	4.2E-06	7.1E-06	*SOLUBLE N-ETHYLMALEMIDE-SENSITIVE FACTOR ADAPTOR PROTEIN RECEPTORS* (*SNARE*) interactions in vesicular transport
00514	5.5E-04	6.6E-04	Other types of O-glycan biosynthesis
00010	1.5E-03	1.2E-02	Glycolysis/gluconeogenesis
00051	2.8E-03	1.1E-02	Fructose and mannose metabolism
00190	2.9E-03	9.3E-03	Oxidative phosphorylation
01100	3.2E-03	4.4E-02	Metabolic pathways
00040	4.5E-03	3.4E-02	Pentose and glucuronate interconversions
01110	5.7E-03	–	Biosynthesis of secondary metabolites
00563	9.3E-03	1.2E-02	Glycosylphosphatidylinositol (GPI)-anchor biosynthesis
00500	1.5E-02	–	Starch and sucrose metabolism
00670	1.6E-02	–	One carbon pool by folate
00941	1.7E-02	–	Flavonoid biosynthesis
00071	1.9E-02	2.6E-02	Fatty acid metabolism
00940	2.8E-02	–	Phenylpropanoid biosynthesis
00600	3.0E-02	3.6E-02	Sphingolipid metabolism
00020	3.4E-02	–	Citrate cycle (tricarboxylic acid cycle)
00053	4.2E-02	–	Ascorbate and aldarate metabolism
00270	4.8E-02	–	Cysteine and methionine metabolism
00510	–	2.5E-02	N-Glycan biosynthesis
04075	–	2.9E-02	Plant hormone signal transduction

The 1826 eucalypt and 2355 poplar genes found to be xylem-expressed orthologs were over-represented in several key metabolic pathways (*P*-value < 0.05) involved in growth and development.

proteins. Eighteen of the 19 eucalypt genes from cluster PtrEgr1185 were preferentially expressed in xylem tissue, and the single poplar gene (Potri.009G039200; log2-fold = 0.85 higher in xylem) did not display tissue-specific expression. Eucalypt-specific clusters also contained developmental genes (PtrEgr1086; *LEA* genes), disease resistance genes (PtrEgr1001), and a cluster consisting of galactinol synthase (PtrEgr1260) and cellulose synthase-like (PtrEgr1038) genes. These genes probably represent gene family clades that have been expanded in one species and lost in the other.

Protein family domain analysis identified domains present in the species-specific list of genes compared with the co-expressed gene list (Fig. 3c,d). In poplar, differentially expressed genes were found to contain asciclin (16 proteins; PF02469), *PECTIN METHYLESTERASE INHIBITOR* (*PMEI*; 16 proteins; PF04043) and *KINESIN* (12 proteins; PF0025) domains, all known to be involved in cell adhesion (Camardella *et al.*, 2001; Faik *et al.*, 2006) and cellular transport activities (Hirokawa *et al.*, 2009). Eucalypt-specific domains consisted, among others, of heat-shock protein (HSP20; PF00011) and disease response domains (*TOLL INTERLEUKIN RECEPTOR*; PF01582) together with *LATE EMBRYOGENESIS ABUNDANT* (*LEA*; PF03168) and zinc-finger domains (PF13912).

We further identified orthologous genes from both species (cluster class '$N \times 2$') that only exhibited differential expression in one of the species. These 1536 pairs of genes were used to estimate the $K_a : K_s$ ratio of species-specific regulated genes (Table S12). Low $K_a : K_s$ ratios were observed for 290 ortholog pairs, medium values for 848 pairs, and high values (maximum $K_a : K_s$ of 0.512) for 389 pairs of orthologs.

Finally, we investigated the expression of xylem preferential genes known to be involved in lignin monomer biosynthesis (Fig. 4) and cellulose and xylan biosynthesis (Fig. 5). We identified 27 xylem preferentially expressed genes in *E. grandis* and 32 genes in *P. trichocarpa* that potentially encode the 10 enzymatic steps of lignin monomer biosynthesis. For the 18 enzymatic steps involved in cellulose and xylan biosynthesis, we identified 81 xylem preferential genes in *E. grandis* and 98 genes in *P. trichocarpa*. Our analysis highlights the fact that generally a small number of genes in each species show high and specific expression in xylem tissue. *Populus trichocarpa* often has two retained xylem-expressed paralogs from the more recent whole-genome duplication. Presumably as a result of the older age of the genome-wide duplication observed in *E. grandis* (Myburg *et al.*, 2014), most gene duplicates have been lost in *E. grandis*, reverting back to single copy number for many of the lignin and carbohydrate biosynthesis genes. An example is the *CESA* gene family in *E. grandis*. *Eucalyptus grandis* is expressing only one ortholog each of the secondary cell wall-related *CESA4*, *7* and *8*, while *P. trichocarpa* has two additional xylem preferential paralogs (Fig. 5).

Discussion

Xylem formation has been studied extensively in the species *P. trichocarpa* and *E. grandis*, and several key genes that influence wood properties have been identified in both species. Xylem formation (woody perennial growth from a peripheral vascular cambium) occurred multiple times in the angiosperm lineage (Plomion *et al.*, 2001; Groover, 2005), which raises the question of whether the same functional orthologs are involved in xylem formation, and more specifically secondary cell wall development, in species that undergo secondary growth. We used genome-wide data from *P. trichocarpa* and *E. grandis*, members of distantly related genera that both contain exclusively woody species that have evolved separately for > 90 Myr, to undertake the systematic identification of orthologous genes that are preferentially expressed in the developing xylem of both species. We followed a sequence similarity-based approach to identify clusters

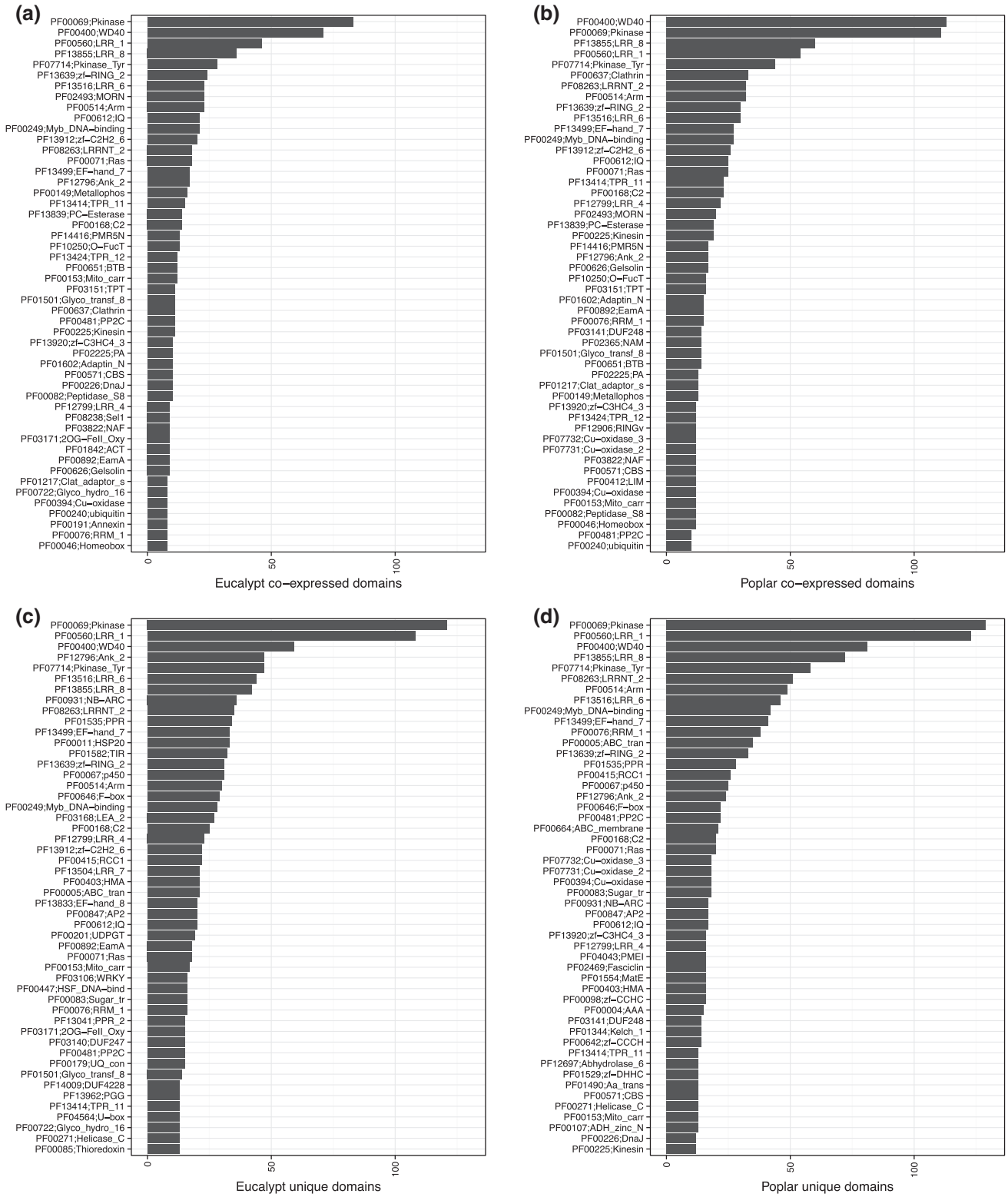

Fig. 3 Protein family domains identified in the *Populus trichocarpa* and *Eucalyptus grandis* genes preferentially expressed in developing xylem. Protein domains encoded by orthologous poplar and eucalypt genes (from clusters '1 × 2' and '*N* × 2') with xylem preferred expression are shown in (a) and (b). Protein domains encoded by xylem preferentially expressed genes lacking detectable orthologs in the alternate species (clusters '*N* × 1' and '1 × 1') are shown in (c) and (d).

of genes encoding groups of homologous proteins. Conserved sequence domains in genes from different species have been used widely to identify and group proteins together with putative functional similarities (for example the identification of carbohydrate-active enzymes in *P. trichocarpa* (Geisler-Lee *et al.*, 2006) and *E. grandis* (D. Pinard *et al.*, unpublished). The more recent

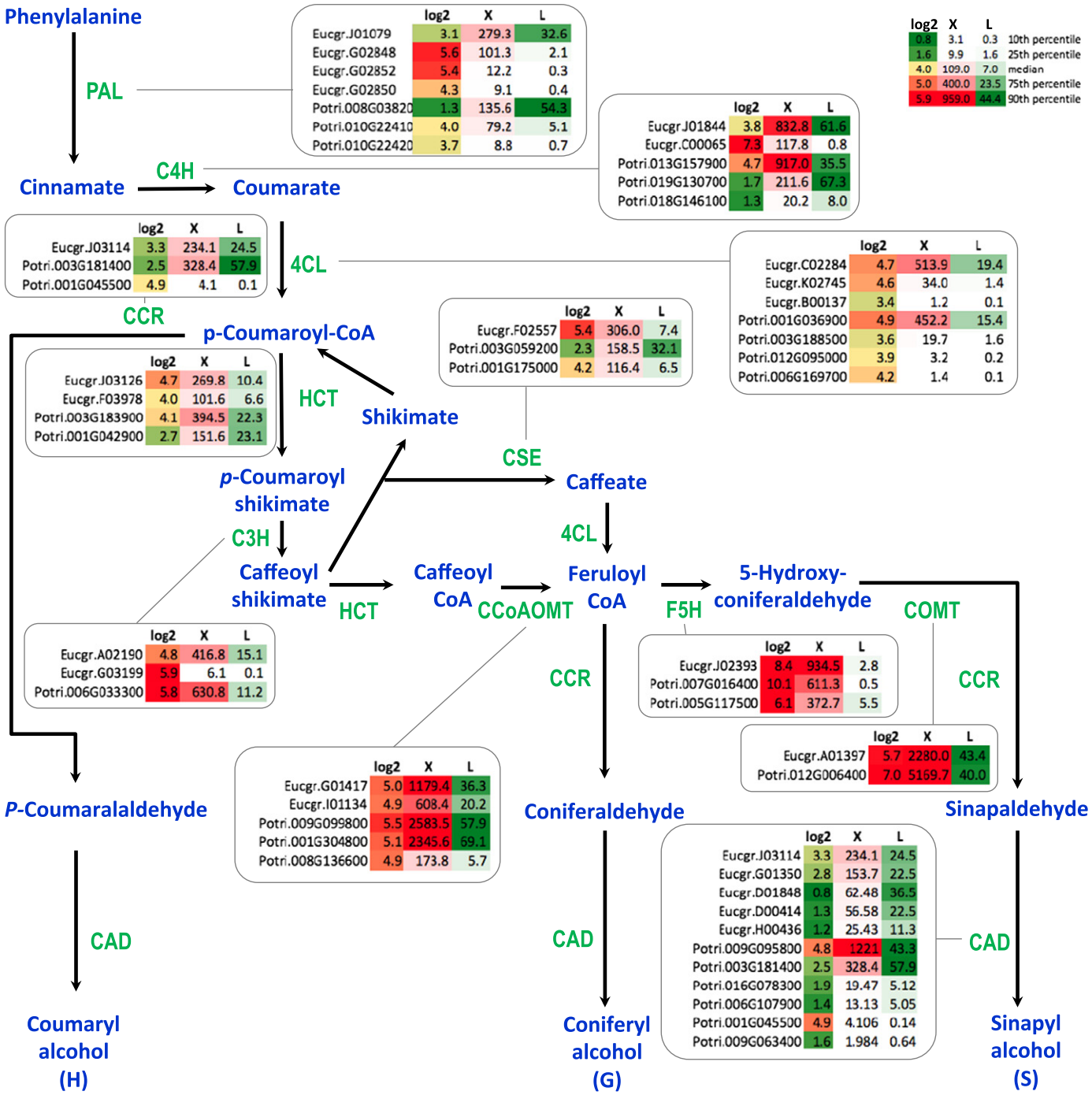

Fig. 4 Xylem preferentially expressed genes in the phenylpropanoid biosynthetic pathway leading to lignin monomer production. Gene expression ratio (log2-fold of xylem/leaf; first column) is indicated by color blocks, with warmer colors (red) indicating a preference for xylem tissue and cooler colors (green) indicating a preference for young leaf tissue. Color coding is based on the 10th and 90th percentile values for gene expression. Absolute (fragments per kilobase per million mapped reads (FPKM)) transcript abundance is shown in columns 2 (xylem (X); red) and 3 (leaf (L); green). Note that only genes with xylem preferential expression and FPKM > 1 are shown. In some cases, genes highly expressed in xylem and phloem are present but not shown here (accessible in Supporting Information Table S1). Full gene names and expression values are provided in Table S13.

whole-genome duplication event characteristic of *P. trichocarpa* (*c.* 48 Myr ago compared with *c.* 109 Myr ago in *E. grandis*) can be observed in several gene clusters where two or more copies of a *P. trichocarpa* homolog were clustered together with a single *E. grandis* ortholog (Tuskan *et al.*, 2006; Myburg *et al.*, 2014). Approximately 95% of gene copies from the 109 Myr ago whole-genome duplication in *E. grandis* have been lost (Myburg *et al.*, 2014). However, the *E. grandis* genome contains a large proportion of tandem duplicate genes (34% of annotated gene models), which underlies species-specific expansion of some gene family clades. In several instances, sequence similarity was unable to identify closest orthologs between species, and these genes were

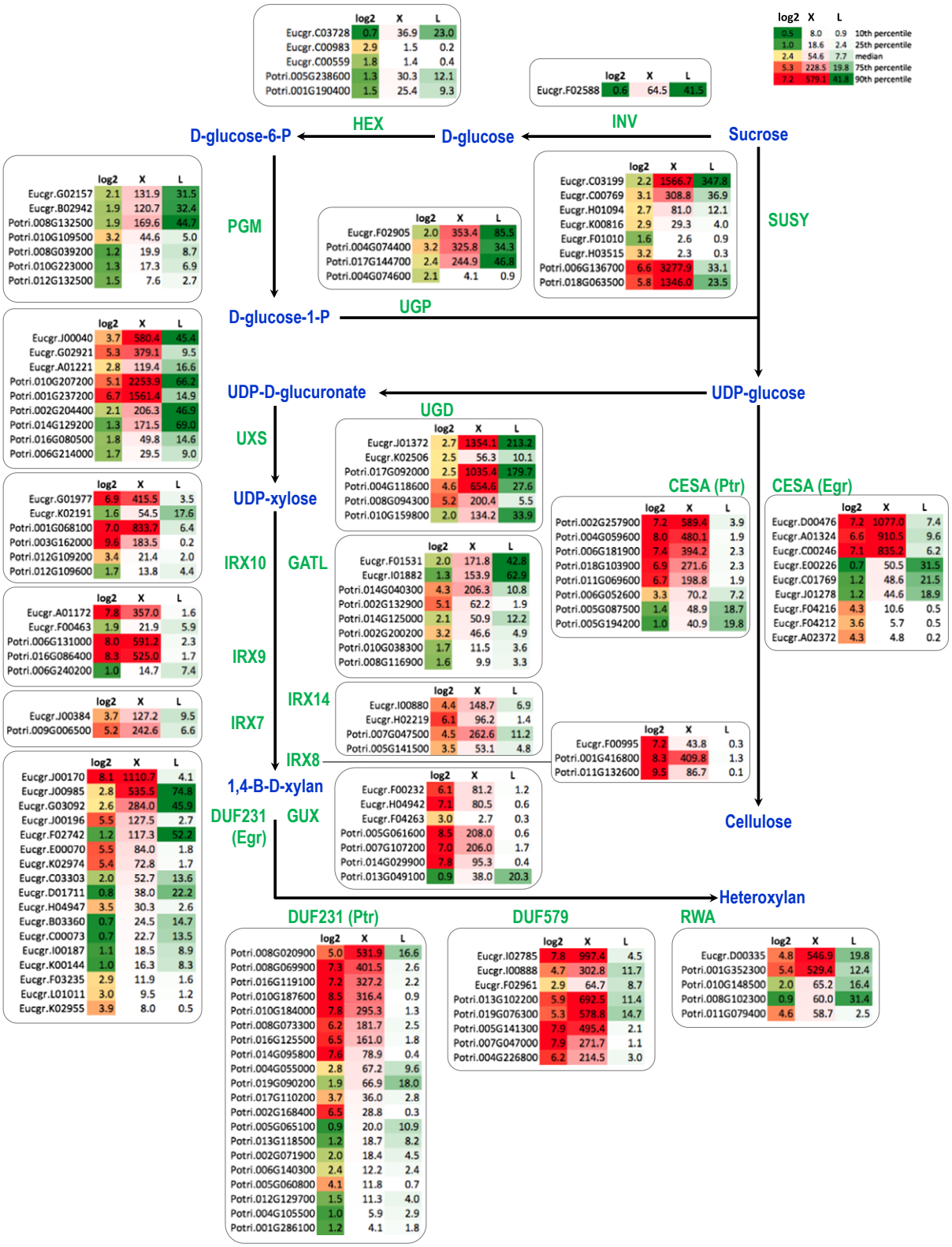

Fig. 5 Xylem preferentially expressed genes in the cellulose and xylan biosynthetic pathway. The gene expression ratio (log2-fold of xylem/leaf; first column) is indicated by color blocks, with warmer colors (red) indicating a preference for xylem tissue and cooler colors (green) indicating a preference for young leaf tissue. Color coding is based on the 10th and 90th percentile values for gene expression. Absolute (fragments per kilobase per million mapped reads (FPKM)) transcript abundance is shown in columns 2 (xylem (X); red) and 3 (leaf (L); green). Note that only genes with xylem preferential expression and FPKM > 1 are shown. Full gene names and expression values are provided in Supporting Information Table S14.

either considered to be singleton genes within a species, or a cluster of paralogs unique to one species. Together, these results support a model where gene diversity and orthology in *E. grandis* and *P. trichocarpa* are shaped by whole-genome and tandem duplication and lineage-specific gene loss.

The two tree species employed in this experiment were field-grown in different geographic conditions and subjected to varying environmental factors, including average daily temperature, soil quality and water availability, to mention only a few. However, the sites selected are native to the species, and therefore offer a true representation of the most appropriate geoclimatic conditions influencing xylem development. We controlled for the long- and short-term environmental conditions by estimating differential gene expression between tissues (primary versus secondary growth) exposed to the same environment, which should mitigate the immediate impact of environmental response at the time of sampling. Despite this approach, we observed higher levels of stress-response genes over-represented in the *E. grandis* data set compared with the *P. trichocarpa* data set. Herein, we focused on genes differentially expressed between developing xylem and young leaf tissue to identify genes and gene clusters active in xylem development in both species. Known transcriptional regulators and genes active in xylem development, such as the suite of *MYB* transcription factors and the *CESA* genes, were identified as highly expressed in developing xylem of both species (for a complete list of differentially expressed genes, see Table S1). We investigated the branch of the phenylpropanoid pathway responsible for lignin biosynthesis for genes differentially expressed in xylem tissue (Fig. 4). Lignin content and composition has been identified as one of the most important focal areas for breeding or bioengineering of woody biomass from forest trees (Boerjan *et al.*, 2003; Hinchee *et al.*, 2009; Vanholme *et al.*, 2010). Reduced lignin content and a higher syringyl to guaiacyl lignin monomer ratio have been linked to reduced recalcitrance of lignocellulosic feedstocks (Boerjan, 2005; Mansfield *et al.*, 2012; Van Acker *et al.*, 2013). Lignin biosynthesis is a well-studied molecular pathway, with at least ten known gene families involved in the process (Lu *et al.*, 2010).

Close inspection of the gene expression patterns in the lignin biosynthetic pathway in both *P. trichocarpa* and *E. grandis* identified several homologs differentially expressed between the tissues in both species (Fig. 4). Herein, the genes shown in the pathway were selected based on differential expression between tissues within each species only (q-value < 0.05). It is important to note that some of the expressed orthologs shown may not play an active role in the biosynthesis pathway branch of the phenylpropanoid pathway. For example, two *PHE AMMONIA LYASE* (*PAL*) genes (Potri.008G038200 and Potri.010G224100) have been linked to lignin biosynthesis in poplar (Hamberger *et al.*, 2007). Based on gene expression and sequence similarity, our analyses indicate that only one ortholog, Potri.010G224100, shows a strong xylem preferential expression pattern (log2-fold change = 4.0) with a similarly expressed paralog (Potri.010G224100; log2-fold change = 3.7), albeit at 50% lower level in xylem. Similarly, Hamberger *et al.* (2007) identified five *4-COUMARATE-CoA LIGASE* (*4CL*) genes active in lignin biosynthesis (Potri.001G036900, Potri.003G188500, Potri.018G094200, Potri.006G169700 and Potri.019G049500). Our results indicate that Potri.018G094200 has a preference in leaf expression, while the other *4CL* orthologs (Potri.017g033600, Potri.012g095000 and Potri.006g169600) display xylem-specific tissue expression. These genes may also contribute to other branches of the phenylpropanoid pathway, and may not be directly involved in lignin monomer biosynthesis. However, only one gene in each species (EucgrC02284 and Potri.001G036900) was highly expressed in xylem and these are the best candidates for encoding the *4CL* protein responsible for xylem lignification. Both poplars and eucalypts are inherently syringyl-rich lignin species, and it is clear that the *FERULATE 5-HYDROXYLASE* (F5H) orthologs are among the most highly expressed genes in the pathway and importantly represented by a single copy in both species (Fig. 4).

Cellulose and hemicellulose biosynthesis and deposition are also key traits for selection and manipulation in tree breeding programs. Both poplar and eucalypt trees produce significant amounts of cellulose-rich biomass in a relatively short rotation time, and the metabolic pathways leading to cellulose production have been well studied. An inspection of the cellulose biosynthetic pathway revealed that several genes showed different expression (xylem versus leaf) ratios in the two species (Fig. 5). For example, two *SUCROSE SYNTHASE* (*SUSY*) orthologs in poplar (Potri.018G063500 and Potri006G136700) were expressed in a much higher ratio in xylem (to leaf) than any of the homologs of the eucalypt *SUSY* ortholog (Eugr.C03199), although the latter was expressed at similar high levels in *E. grandis* xylem. This pattern probably reflects the more recent gene duplication in *P. trichocarpa*. A similar pattern was observed for the *UDP-D-GLUCOSE DEHYDROGENASE* (*UGD*) homologs, which convert UDP-D-glucose to UDP-D-glucuronic acid. The conversion of UDP-D-glucuronic acid to UDP-D-xylose by UDP-apiose/xylose synthase is also performed by two poplar homologs (Potri.001G237200 and Potri.010G207200) expressed in higher ratios in xylem than leaf, but not at a higher absolute expression level Fragments per Kilobase per Million mapped Reads (FPKM), compared with the eucalypt ortholog, Eucgr.G02921. Finally, we observed similarity among the *CESA* homologs differentially expressed in the two species, with three and five genes highly and differentially expressed in *E. grandis* and *P. trichocarpa*, respectively, consistent with previous reports (Djerbi *et al.*, 2005; Ranik & Myburg, 2006).

We further narrowed our focus to genes with a one-to-one ratio of orthologs present in both species, as identified by the clustering algorithms (Table S3). The 672 genes (366 orthologous pairs) show high levels of sequence similarity and also exhibit conservation in tissue expression status between species. Several of the known proteins involved in secondary cell wall formation, such as the *KNAT7* (Potri.001G112200 and Eucgr.D01935), *FLA2* (Potri.014G168100 and Eucgr.H00875) and *XYLOGLUCAN:XYLOGLYCOSYL TRANSFERASE 33* (*XET33*) (Potri.014G115000 and Eucgr.B03348) orthologs, were conserved in *P. trichocarpa* and *E. grandis*. Interestingly, a

smaller number of genes were conserved at the sequence level, but their expression profiles were switched between species (Table S5). For example, the *SUCROSE TRANSPORTER 4* (*SUT4*) orthologs (Potri.002G106900 and Eucgr.F00464) suggest that this sucrose transport enzyme is more active in *E. grandis* leaves compared with xylem tissue, and more active in *P. trichocarpa* xylem compared with leaf tissue. However, it is important to note that the expression of the *E. grandis* ortholog was 2-fold higher in *E. grandis* xylem (FPKM 36.5 versus 17.5 in *P. trichocarpa*), indicating that expression ratio and expression level should be considered for inference of gene activity in xylem. *PtSUT4* was previously shown to be functionally active in mediating apoplastic phloem loading and sucrose flux in *P. trichocarpa* (Payyavula *et al.*, 2011). The *E. grandis* ortholog was expressed at an even higher level in leaf tissue (FPKM = 58.6), suggesting active involvement in phloem loading in source tissue. A collection of genes encoding proteins with unknown function (*DUF1995, DUF581, DUF1677, DUF593, DUF3133* and *PUF1589*), as well as a cluster of *WRKY70*-like transcription factor orthologs, were preferentially expressed in eucalypt xylem tissue, while the poplar orthologs of these genes were preferentially expressed in leaf tissue. This could indicate neo-functionalization of these orthologs based on expression pattern since divergence of the two lineages (probably following gene duplication and loss of one of the gene copies in one or both species). Expression information from orthologs of these genes in other woody species may aid in resolving whether the ancestral expression pattern was in xylem, leaves, or both.

In addition to the 672 xylem preferentially expressed genes considered to be functional orthologs between species, we identified 3490 xylem-expressed putative orthologs from cluster class '$N \times 2$' using a combination of phylogenetic and differential gene expression analysis (Table S7). Genes forming cluster '$N \times 2$' were considered to be functional orthologs when paralogs from both species were placed in adjacent nodes in the neighbor-joining tree and when both were differentially expressed in xylem tissue. The remainder of the differentially expressed genes clustered into '$N \times 2$' appeared to have divergent expression patterns, and the poplar or eucalypt orthologs preferentially expressed only in xylem of one of the species may have a species-specific (or non-conserved) function in xylem formation (Table S11). Functional annotation of these genes identified cell signalling, amino acid catabolism, lipid metabolism and other regulatory functions as putative species-specific functions linked to these genes. Further investigation into the species-specific nature of these genes focusing on xylem development is needed on a gene-by-gene basis. Again, it is important to note that both orthologs may function in xylem development, but that one of the orthologs is expressed at a level not considered to be significantly different from that in leaf tissue. The abundance of annotated functions in cell signalling and hormone signal transduction pathways among these genes suggests that some of them may play important roles in the regulation of secondary xylem development.

We also estimated the evolutionary pressure acting on a set of co-expressed genes as well as the set of genes where only one species showed preferential expression. Lower $K_a : K_s$ values were observed for the set of co-expressed ortholog genes than for the set of orthologs with no expression correlation. Overall, we observe that genes that share xylem expression status are under more stringent selection pressure, confirming that this group of genes encodes conserved functions required for normal plant growth and fitness.

Conclusions

Here, we demonstrate the use of comparative transcriptome profiling to identify functional orthologs preferentially expressed during xylem formation in two ecologically and industrially important wood-producing genera. Our results indicate that a high level of gene expression and sequence similarity conservation exits for genes involved in xylem development, yet differential expression patterns were observed for genes generated by gene duplication in either species possibly underlying lineage-specific functional diversification. Furthermore, we found high levels of expression conservation between homologs active within lignin and cellulose biosynthetic pathways. However, there are unique examples where apparently orthologous pairs have switched expression patterns. Moreover, there are instances where actively expressed orthologs are not present in the other species, potentially identifying genes unique to xylem development in one species. Our study provides a comparative overview of the genetic control of genes involved in xylem development and a useful framework for future comparative studies in these economically important woody plant genera. Specifically, the identification of gene families with conserved expression in woody tissues now offers a unique means to guide research on genes and allelic variants for selection in tree breeding populations, or for targeted genetic engineering of industrially important wood traits.

Acknowledgements

This work was supported by Genome Canada Large-Scale Applied Research Project (Project 168BIO), funds to C.J.D. and S.D.M., and by a South African Department of Science and Technology (DST) grant awarded to A.A.M. *Eucalyptus grandis* leaf and xylem tissues were provided by Mondi Tree Improvement Research (Kwambonambi, South Africa). RNA-seq analysis in *E. grandis* was further supported by Mondi and Sappi through the Forest Molecular Genetics (FMG) Programme, the Technology and Human Resources for Industry Programme (THRIP, UID 80118), and the National Research Foundation (NRF, UID 71255 and 86936) of South Africa.

References

Abramson M, Shoseyov O, Shani Z. 2010. Plant cell wall reconstruction toward improved lignocellulosic production and processability. *Plant Science* **178**: 61–72.

Andersson-Gunnerås S, Mellerowicz EJ, Love J, Segerman B, Ohmiya Y, Coutinho PM, Nilsson P, Henrissat B, Moritz T, Sundberg B. 2006. Biosynthesis of cellulose-enriched tension wood in *Populus*: global analysis of transcripts and metabolites identifies biochemical and developmental regulators in secondary wall biosynthesis. *Plant Journal* **45**: 144–165.

Bao H, Li E, Mansfield SD, Cronk QCB, El-Kassaby YA, Douglas CJ. 2013. The developing xylem transcriptome and genome-wide analysis of alternative splicing in *Populus trichocarpa* (black cottonwood) populations. *BMC genomics* **14**: 359.

Boerjan W. 2005. Biotechnology and the domestication of forest trees. *Current Opinion in Biotechnology* **16**: 159–166.

Boerjan W, Ralph J, Baucher M. 2003. Lignin biosynthesis. *Annual Review of Plant Biology* **54**: 519–546.

Camardella L, Carratore V, Ciardiello MA, Servillo L, Balestrieri C, Giovane A. 2001. Kiwi protein inhibitor of pectin methylesterase. *European Journal of Biochemistry* **267**: 4561–4565.

Davidson RM, Gowda M, Moghe G, Lin H, Vaillancourt B, Shiu S-H, Jiang N, Robin Buell C. 2012. Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. *Plant Journal* **71**: 492–502.

Déjardin A, Laurans F, Arnaud D, Breton C, Pilate G, Leplé J-C. 2010. Wood formation in Angiosperms. *Comptes Rendus Biologies* **333**: 325–334.

Djerbi S, Lindskog M, Arvestad L, Sterky F, Teeri TT. 2005. The genome sequence of black cottonwood (*Populus trichocarpa*) reveals 18 conserved cellulose synthase (CesA) genes. *Planta* **221**: 739–746.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**: 1792–1797.

Faik A, Abouzouhair J, Sarhan F. 2006. Putative fasciclin-like arabinogalactan-proteins (FLA) in wheat (*Triticum aestivum*) and rice (*Oryza sativa*): identification and bioinformatic analyses. *Molecular Genetics and Genomics* **276**: 478–494.

Falcon S, Gentleman R. 2007. Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**: 257–258.

Geisler-Lee J, Geisler M, Coutinho PM, Segerman B, Nishikubo N, Takahashi J, Aspeborg H, Djerbi S, Master E, Andersson-Gunneras S *et al.* 2006. Poplar carbohydrate-active enzymes. Gene identification and expression analyses. *Plant Physiology* **140**: 946–962.

Geraldes A, Pang J, Thiessen N, Cezard T, Moore R, Zhao Y, Tam A, Wang S, Friedmann MC, Birol I *et al.* 2011. SNP discovery in black cottonwood (*Populus trichocarpa*) by population transcriptome resequencing. *Molecular Ecology Resources* **11**: 81–92.

Groover AT. 2005. What genes make a tree a tree? *Trends in Plant Science* **10**: 210–214.

Hamberger B, Ellis M, Friedmann MC, de Azevedo Souza C, Barbazuk B, Douglas CJ. 2007. Genome-wide analyses of phenylpropanoid-related genes in *Populus trichocarpa, Arabidopsis thaliana,* and *Oryza sativa*: the *Populus* lignin toolbox and conservation and diversification of angiosperm gene families. *Botany-Botanique* **85**: 1182–1201.

Hinchee M, Rottmann W, Mullinax L, Zhang C, Chang S, Cunningham M, Pearson L, Nehra N. 2009. Short-rotation woody crops for bioenergy and biofuels applications. *In Vitro Cellular & Developmental Biology – Plant* **45**: 619–629.

Hirokawa N, Noda Y, Tanaka Y, Niwa S. 2009. Kinesin superfamily motor proteins and intracellular transport. *Nature Reviews Molecular Cell Biology* **10**: 682–696.

Hussey SG, Mizrachi E, Creux NM, Myburg AA. 2013. Navigating the transcriptional roadmap regulating plant secondary cell wall deposition. *Frontiers in Plant Science* **4**: 235.

Jansson S, Douglas CJ. 2007. *Populus*: a model system for plant biology. *Annual Review of Plant Biology* **58**: 435–458.

Kirst M, Johnson AF, Baucom C, Ulrich E, Hubbard K, Staggs R, Paule C, Retzel E, Whetten R, Sederoff R. 2003. Apparent homology of expressed genes from wood-forming tissues of loblolly pine (*Pinus taeda* L.) with *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences, USA* **100**: 7383–7388.

Knoth C, Ringler J, Dangl JL, Eulgem T. 2007. *Arabidopsis WRKY70* is required for full RPP4-mediated disease resistance and basal defense against *Hyaloperonospora parasitica*. *Molecular Plant-Microbe Interactions* **20**: 120–128.

Ko J-H, Kim H-T, Hwang I, Han K-H. 2012. Tissue-type-specific transcriptome analysis identifies developing xylem-specific promoters in poplar. *Plant Biotechnology Journal* **10**: 587–596.

Kubo M, Udagawa M, Nishikubo N, Horiguchi G, Yamaguchi M, Ito J, Mimura T, Fukuda H, Demura T. 2005. Transcription switches for protoxylem and metaxylem vessel formation. *Genes & Development* **19**: 1855–1860.

Langan P, Gnanakaran S, Rector KD, Pawley N, Fox DT, Cho DW, Hammel KE. 2011. Exploring new strategies for cellulosic biofuels production. *Energy & Environmental Science* **4**: 3820–3833.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R *et al.* 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947–2948.

Lens F, Smets E, Melzer S. 2012. Stem anatomy supports *Arabidopsis thaliana* as a model for insular woodiness. *New Phytologist* **193**: 12–17.

Li E, Bhargava A, Qiang W, Friedmann MC, Forneris N, Savidge RA, Johnson LA, Mansfield SD, Ellis BE, Douglas CJ. 2012. The Class II *KNOX* gene *KNAT7* negatively regulates secondary wall formation in *Arabidopsis* and is functionally conserved in *Populus*. *New Phytologist* **194**: 102–115.

Li L. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* **13**: 2178–2189.

Lu S, Li L, Zhou G. 2010. Genetic modification of wood quality for second-generation biofuel production. *GM Crops* **1**: 230–236.

Mansfield SD, Kang K-Y, Chapple C. 2012. Designed for deconstruction – poplar trees altered in cell wall lignification improve the efficacy of bioethanol production. *New Phytologist* **194**: 91–101.

McCarthy DJ, Chen Y, Smyth GK. 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research* **40**: 4288–4297.

Melzer S, Lens F, Gennen J, Vanneste S, Rohde A, Beeckman T. 2008. Flowering-time genes modulate meristem determinacy and growth form in *Arabidopsis thaliana*. *Nature Genetics* **40**: 1489–1492.

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. *Plant Cell* **17**: 2993–3006.

Mizrachi E, Hefer CA, Ranik M, Joubert F, Myburg AA. 2010. De novo assembled expressed gene catalog of a fast-growing *Eucalyptus* tree produced by Illumina mRNA-Seq. *BMC Genomics* **11**: 681.

Mizrachi E, Mansfield SD, Myburg AA. 2011. Cellulose factories: advancing bioenergy production from forest trees. *New Phytologist* **194**: 54–62.

Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D *et al.* 2014. The genome of *Eucalyptus* grandis. *Nature* **510**: 356–362.

Öhman D, Demedts B, Kumar M, Gerber L, Gorzsás A, Goeminne G, Hedenström M, Ellis B, Boerjan W, Sundberg B. 2013. MYB103 is required for ferulate-5-hydroxylase expression and syringyl lignin biosynthesis in *Arabidopsis* stems. *Plant Journal* **73**: 63–76.

Pavy N, Boyle B, Nelson C, Paule C, Giguere I, Caron S, Parsons LS, Dallaire N, Bedon F, Berube H *et al.* 2008. Identification of conserved core xylem gene sets: conifer cDNA microarray development, transcript profiling and computational analysis. *New Phytologist* **180**: 766–786.

Pavy N, Pelgas B, Laroche J, Rigault P, Isabel N, Bousquet J. 2012. A spruce gene map infers ancient plant genome reshuffling and subsequent slow evolution in the gymnosperm lineage leading to extant conifers. *BMC Biology* **10**: 84.

Payyavula RS, Tay KHC, Tsai C-J, Harding SA. 2011. The sucrose transporter family in *Populus*: the importance of a tonoplast PtaSUT4 to biomass and carbon partitioning. *Plant Journal* **65**: 757–770.

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

Prassinos C, Ko J-H, Yang J, Han K-H. 2005. Transcriptome profiling of vertical stem segments provides insights into the genetic regulation of secondary growth in hybrid aspen trees. *Plant and Cell Physiology* **46**: 1213–1225.

Ranik M, Myburg AA. 2006. Six new cellulose synthase genes from *Eucalyptus* are associated with primary and secondary cell wall biosynthesis. *Tree Physiology* **26**: 545–556.

Rigault P, Boyle B, Lepage P, Cooke JEK, Bousquet J, MacKay JJ. 2011. A white spruce gene catalog for conifer genome analyses. *Plant Physiology* **157**: 14–28.

Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L, Trapnell C. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7**: 562–578.

Schuetz M, Smith R, Ellis B. 2012. Xylem tissue specification, patterning, and differentiation mechanisms. *Journal of Experimental Botany* **64**: 11–31.

Spicer R, Groover A. 2010. Evolution of development of vascular cambia and secondary growth. *New Phytologist* **186**: 577–592.

Studer MH, DeMartini JD, Davis MF, Sykes RW, Davison B, Keller M, Tuskan GA, Wyman CE. 2011. Lignin content in natural *Populus* variants affects sugar release. *Proceedings of the National Academy of Sciences, USA* **108**: 6300–6305.

Supek F, Bošnjak M, Škunca N, Šmuc T. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms (C Gibas, Ed.). *PLoS ONE* **6**: e21800.

Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, Selbig J, Müller LA, Rhee SY, Stitt M. 2004. MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant Journal* **37**: 914–939.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**: 1105–1111.

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A *et al.* 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**: 1596–1604.

Van Acker R, Vanholme R, Storme V, Mortimer JC, Dupree P, Boerjan W. 2013. Lignin biosynthesis perturbations affect secondary cell wall composition and saccharification yield in *Arabidopsis thaliana*. *Biotechnology for Biofuels* **6**: 46.

Vanholme R, Cesarino I, Rataj K, Xiao Y, Sundin L, Goeminne G, Kim H, Cross J, Morreel K, Araujo P *et al.* 2013. Caffeoyl shikimate esterase (CSE) is an enzyme in the lignin biosynthetic pathway. *Science* **341**: 1103–1106.

Vanholme R, Demedts B, Morreel K, Ralph J, Boerjan W. 2010. Lignin biosynthesis and structure. *Plant Physiology* **153**: 895–905.

Yamaguchi M, Kubo M, Fukuda H, Demura T. 2008. VASCULAR-RELATED NAC-DOMAIN7 is involved in the differentiation of all types of xylem vessels in *Arabidopsis* roots and shoots. *Plant Journal* **55**: 652–664.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**: 1586–1591.

Zhong R, Demura T, Ye Z-H. 2006. SND1, a NAC domain transcription factor, is a key regulator of secondary wall synthesis in fibers of Arabidopsis. *Plant Cell* **18**: 3158–3170.

Zhong R, Lee C, Ye Z-H. 2010. Evolutionary conservation of the transcriptional network regulating secondary cell wall biosynthesis. *Trends in Plant Science* **15**: 625–632.

Zhong R, Lee C, Zhou J, McCarthy RL, Ye Z-H. 2008. A battery of transcription factors involved in the regulation of secondary cell wall biosynthesis in *Arabidopsis*. *Plant Cell* **20**: 2763–2782.

Zhong R, Ye Z-H. 2007. Regulation of cell wall biosynthesis. *Current Opinion in Plant Biology* **10**: 564–572.

Supporting Information

Additional supporting information may be found in the online version of this article.

Fig. S1 OrthoMCL cluster analysis identified 15 925 clusters of genes representing a total of 59 892 annotated *Populus trichocarpa* and *Eucalyptus grandis* genes.

Table S1 A description of the 77 711 *Populus trichocarpa* and *Eucalyptus grandis* transcripts with annotations used in the study

Table S2 Biological process (BP) gene ontology terms associated with genes considered singletons in the relevant transcriptomes

Table S3 Clusters of orthologs (cluster class '1 × 2') that exhibit xylem preferential expression in both species

Table S4 Biological process (BP) gene ontology terms over-represented (P-value < 0.01) within the set of xylem orthologs (762 genes) differentially expressed in both species

Table S5 The 121 pairs of orthologs (242 genes in total) showing switching in tissue expression preferences

Table S6 Clusters containing multiple genes for both species where genes from only one of the species are preferentially expressed in xylem

Table S7 The 2109 gene clusters containing at least one set of putative functional orthologs that are both preferentially expressed in xylem tissue

Table S8 Gene ontology functional annotation of the 2004 *Populus trichocarpa* and 1486 *Eucalyptus grandis* genes from cluster class '$N \times 2$' considered to be functional orthologs after neighbor-joining phylogenetic analysis

Table S9 The 672 genes identified in cluster class '1 × 2', as well as the 2004 *Populus trichocarpa* and 1486 *Eucalyptus grandis* genes from cluster class '$N \times 2$' considered to be functional orthologous pairs between species

Table S10 The ratio of synonymous versus nonsynonymous nucleotide substitutions ($K_a : K_s$) in the coding regions of the pairs of xylem co-expressed orthologs using SNPs identified in these genes after sequence alignment

Table S11 The 3410 *Populus trichocarpa* and 3481 *Eucalyptus grandis* xylem preferentially expressed genes considered not to share a co-expressed ortholog in the other species

Table S12 $K_a : K_s$ ratio of species-specific expressed genes. Low $K_a : K_s$ ratios were observed for 290 ortholog pairs, medium values for 848 pairs, and high values (maximum $K_a : K_s$ of 0.512) for 389 pairs

Table S13 Genes preferentially expressed in xylem active in the phenylpropanoid biosynthetic pathway

Table S14 Genes preferentially expressed in xylem and active in the cellulose and xylan biosynthesis pathway

Notes S1 The number of reads sequenced within each library, the distribution of FPKM values across species, and tissue expression clustering results for all sequenced RNA-seq libraries.