Genome-wide analysis of the lignin toolbox of *Eucalyptus grandis*

Victor Carocha[1,2,3], Marçal Soler[1], Charles Hefer[4,5], Hua Cassan-Wang[1], Pedro Fevereiro [2,6], Alexander A. Myburg[7,8], Jorge A.P. Paiva[3,9] and Jacqueline Grima-Pettenati[1,*]

[1]LRSV, Laboratoire de Recherche en Sciences Végétales, UPS, CNRS, Université Toulouse 3, BP 42617 Auzeville, 31326, Castanet Tolosan, France;

[2]Instituto de Tecnologia de Química Biológica (ITQB), Biotecnologia de Células Vegetais, Av. da República, 2781-157 Oeiras, Portugal;

[3]Instituto de Investiga cão Científica e Tropical (IICT/MNE), Palácio Burnay, Rua da Junqueira, 30, 1349-007 Lisboa, Portugal;

[4]Department of Botany, University of British Columbia, 3529-6270 University Blvd, Vancouver, BC V6T 1Z4, Canada;

[5]Bioinformatics and Computational Biology Unit, Department of Biochemistry, University of Pretoria, Private Bag X20, Pretoria, South Africa;

[6]Departamento de Biologia Vegetal, Faculdade de Ciências da Universidade de Lisboa (FCUL), Campo Grande, 1749-016 Lisboa, Portugal;

[7]Department of Genetics, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Private bag X20, Pretoria 0028, South Africa;

[8]Genomics Research Institute (GRI), University of Pretoria, Private bag X20, Pretoria 0028, South Africa;

[9]iBET – Instituto de Biologia Experimental e Tecnológica, Apartado 12, 2781-901 Oeiras, Portugal

*Author for correspondence:

Jacqueline Grima-Pettenati

Tel: +33534323813

Email: grima@lrsv.ups-tlse.fr

Summary

- Lignin, a major component of secondary cell walls, hinders the optimal processing of wood for industrial uses. The recent availability of the *Eucalyptus grandis* genome sequence allows comprehensive analysis of the genes encoding the 11 protein families specific to the lignin branch of the phenylpropanoid pathway and identification of those mainly involved in xylem developmental lignification.
- We performed genome-wide identification of putative members of the lignin gene families, followed by comparative phylogenetic studies focusing on *bona fide* clades

inferred from genes functionally characterized in other species. RNA-seq and microfluid real-time quantitative PCR (RT-qPCR) expression data were used to investigate the developmental and environmental responsive expression patterns of the genes.

- The phylogenetic analysis revealed that 38 *E. grandis* genes are located in *bona fide* lignification clades. Four multigene families (shikimate *O*-hydroxycinnamoyltransferase (HCT), *p*-coumarate 3-hydroxylase (C3H), caffeate/5-hydroxyferulate *O*-methyltransferase (COMT) and phenylalanine ammonia-lyase (PAL)) are expanded by tandem gene duplication compared with other plant species. Seventeen of the 38 genes exhibited strong, preferential expression in highly lignified tissues, probably representing the *E. grandis* core lignification toolbox.
- The identification of major genes involved in lignin biosynthesis in *E. grandis*, the most widely planted hardwood crop world-wide, provides the foundation for the development of biotechnology approaches to develop tree varieties with enhanced processing qualities.

Key words: Eucalyptus, lignin biosynthesis, phenylpropanoid pathway, secondary cell wall, xylem.

Introduction

Lignin, the most complex polyphenolic heteropolymer on earth and the second most abundant biopolymer after cellulose, represents as much as 30% of the total biomass produced in the biosphere and, as a consequence of its recalcitrance to biodegradation, it is a major form of fixed carbon storage (Boerjan *et al*., 2003; Boudet *et al*., 2003). Predominantly found in the secondary cell walls of xylem, lignin has played a crucial role in the emergence of vascular plants and their transition to terrestrial habitats (Weng & Chapple, 2010). It has fundamental biological roles such as conferring mechanical rigidity, imperviousness, water conductivity and resistance to biodegradation and it has important functions in defense mechanisms (Bhuiyan *et al*., 2009). Lignin is produced by the oxidative polymerization of three *p*-hydroxycinammyl alcohol precursors also called monolignols, which differ in the degree of methoxylation in their aromatic ring (Boerjan *et al*., 2003). The three main hydroxycinnamyl alcohols (*p*-coumaryl, coniferyl and sinapyl alcohols) give rise to *p*-hydroxyphenyl (H), guaiacyl (G) and syringyl (S) units, respectively. Lignin content and composition vary greatly within and between species, among tissues and organs of the same plant (Plomion *et al*., 2001), in different developmental processes, and in response to external conditions experienced by plants (Bhuiyan *et al*., 2009).

Lignin resistance to degradation is a major obstacle for industrial processing of wood such as during pulp and paper manufacturing, where it requires aggressive and costly chemical treatments. The huge economic importance of the pulp industry has been a driving force to decipher the lignin biosynthetic pathway, which has proved more complex and reticulate than initially thought (reviewed in Grima-Pettenati & Goffner, 1999; Boudet *et al*., 2003). The topology of the pathway has been revised several times in recent decades (reviewed in Humphreys & Chapple, 2002; Boerjan *et al*., 2003; Ralph *et al*., 2004; Vanholme *et al*., 2010) and new alternative routes are still being discovered such as that involving the recently described caffeoyl shikimate esterase (CSE; Vanholme *et al*., 2013). Altogether, 11 enzymatic reactions are implicated in the synthesis of monolignols, which involves the general phenylpropanoid pathway starting with the deamination of phenylalanine and leading to the production of hydroxycinnamoyl CoA esters. The enzymes involved in this short sequence of reactions are phenylalanine ammonia-lyase (PAL), cinnamate 4-hydroxylase (C4H) and 4-coumarate:CoA ligase (4CL). Although lignin is the most abundant phenylpropanoid derived from the hydroxycinnamoyl-CoA esters, the latter are also the precursors of a wide range of end products including flavonoids, anthocyanins and condensed tannins, which vary according to species, cell type and environmental signals (Dixon & Paiva, 1995). In order to produce monolignols, hydroxycinnamoyl-CoA esters undergo successive hydroxylation and *O*-methylation of their aromatic rings (Boerjan *et al*., 2003) involving the following enzymatic activities: shikimate *O*-hydroxycinnamoyltransferase (HCT); CSE; *p*-coumarate 3-hydroxylase (C3′H); caffeoyl CoA 3-*O*-methyltransferase (CCoAOMT); ferulate 5-hydroxylase (F5H) and caffeate/5-hydroxyferulate *O*-methyltransferase (COMT). The conversion of the side-chain carboxyl to an alcohol group is catalyzed successively by cinnamoyl CoA reductase (CCR) and cinnamyl alcohol dehydrogenase (CAD), two enzymes considered to be the most specific of the monolignol biosynthesis pathway.

Eucalyptus species and their hybrids (commonly referred to as ‘eucalypts’) are among the world's leading sources of woody biomass and are the main hardwoods used for pulp and timber production given their rapid growth rate, their adaptability to diverse ecological conditions and their good-quality wood fiber (Myburg *et al*., 2007; Paiva *et al*., 2011). In particular, *Eucalyptus grandis* and *Eucalyptus globulus* are the most widely planted hardwood trees for industrial uses in tropical/subtropical areas and temperate zones, respectively. The economic importance of these two species has been increasingly reinforced since traditional interests in pulp and paper production have been extended to the emergent areas of biofuels and biomaterials. For cellulosic pulp and bioethanol production, lignin has a negative impact whereas, as a consequence of its high calorific value, it is beneficial for energy production and for recovering pulping chemicals through combustion in paper mills (Bozell *et al*., 2007). In recent decades, the huge economic importance of *Eucalyptus* wood has been a driving force to delineate the lignin pathway in this genus. Remarkably, the *CCR* gene was first cloned in *Eucalyptus gunnii* (*EguCCR*) and its identity was confirmed unambiguously by the enzymatic activity of the corresponding recombinant protein (Lacombe *et al*., 1997). *EguCCR* cDNA was then used as a probe to clone its orthologs in tobacco (*Nicotiana tabacum*) (Piquemal *et al*., 1998), *Arabidopsis thaliana* (Lauvergeat *et al*., 2001) and *Populus* (Leplé *et al*., 2007). In line with its key role in controlling lignin content and composition, *EguCCR* was later shown to co-localize with a quantitative trait locus (QTL) for S : G lignin ratio (Gion *et al*., 2011). Similar pioneering work, including proof of enzymatic activity, has been carried out for *EguCAD2* (Grima-Pettenati *et al*., 1993), which was the second *CAD* gene cloned in plants. Other lignin biosynthetic genes have been cloned in eucalypts (Poeydomenge *et al*., 1994; Gion *et al*., 2000) and located on genetic maps (Gion *et al*., 2011). However, it is the recent availability of the *E. grandis* genome (Myburg *et al*., 2014) that has provided the opportunity to perform a comprehensive genome-wide analysis of lignin biosynthetic genes, as reported in the present study. By combining a genome-wide survey of the genes putatively included in the 11 monolignol gene families with comparative phylogenetic analysis, we identified 38 *E. grandis* genes belonging to *bona fide* clades. High-throughput expression profiling using both RNA-seq and real-time quantitative PCR (RT-qPCR) technology was used to identify a core lignification gene set comprising 17 genes probably involved in developmental xylem lignification in *Eucalyptus*.

Materials and Methods

In silico identification of *E. grandis* phenylpropanoid/monolignol genes

A first survey of the phenylpropanoid genes involved in monolignol biosynthesis was conducted using *E. grandis* annotation v1.0 in phytozome v7.0, and then refined using annotation v1.1 in phytozome v8.0 (http://www.phytozome.net/eucalyptus.php). A combination of keywords and BLASTp searches (using as queries proteins from *Arabidopsis thaliana* and *Populus trichocarpa*) allowed retrieval of 175 predicted *E. grandis* protein sequences that were used to generate large comparative phylogenetic trees including protein sequences from *Populus trichocarpa*, *Vitis vinifera*, *Arabidopsis thaliana* and *Oryza sativa* (Supporting Information Figs S1–S6). To define the *bona fide* clades in *E. grandis* (i.e. clades with homologs of genes that have been experimentally verified to be involved in xylem cell lignification), we included sequences of *bona fide* enzymes proved to have true enzymatic activity/biological function. To do this, we performed an extensive literature survey, summarized in Table S1. For *E. grandis* short name gene nomenclature, we adopted a species-related prefix (Egr) followed by the multigene family abbreviation (Table S2). The numbering prioritized *E. grandis* orthologs of *bona fide* lignification genes from other species described in the literature. The remaining family members were numbered sequentially according to their position on the 11 main chromosomes. Manual curation of the retrieved *E. grandis* sequences was performed whenever necessary, resulting either in the correction of some gene models or in the elimination of truncated predictions. Gene models were plotted according to their physical position in the 11 chromosome scaffolds of *E. grandis* using mapchart 2.2 (Voorrips, 2002).

Phylogenetic analyses

Phylogenetic relationships among selected sets of predicted primary transcripts (Table S3) were analyzed separately for each multigene family. The protein sequences were first aligned using the mafft online program (Katoh *et al*., 2002) using the default settings. The trees were computed and assembled in mega5 (Tamura *et al*., 2011) using the maximum likelihood method based on the JTT matrix-based model (Jones *et al*., 1992). Bootstrap-supported consensus trees were inferred from 1000 replicates (Felsenstein, 1985). Branches with < 50% bootstrap support were collapsed. The initial tree(s) for the heuristic search was obtained automatically as follows: when the number of common sites was < 100, or less than one-fourth of the total number of sites, the maximum parsimony method was used; otherwise the BIONJ algorithm with the Markov Cluster (MCL) distance matrix was used. The trees were drawn to scale with branch lengths measured in terms of the number of substitutions per site and were graphically displayed in the form of a radiating phylogenetic tree. Protein sequence identity, similarity and global similarity matrixes were generated using the sias online tool (http://imed.med.ucm.es/Tools/sias.html). Matrixes were generated involving only the *E. grandis* family members or involving all the proteins selected for the construction of the *bona fide* protein phylogeny trees (Table S4).

Microfluid RT-qPCR expression analysis

RT-qPCR was performed using a ‘developmental tissue’ panel comprising 12 samples collected from fruit capsules, floral buds, shoot tips, roots, young leaves, mature leaves, cambium-enriched fractions, developing secondary xylem, secondary phloem, and primary and secondary stems. In addition, we used a panel of ‘contrasting xylem samples’ consisting of juvenile vs mature xylem, tension vs opposite xylem and high vs limiting nitrogen fertilization. We also included stems, leaves and roots of young eucalypt trees subjected to cold treatments. Plant material description, RNA extraction and cDNA synthesis were previously reported in Cassan-Wang *et al*. (2012), Soler *et al*. (2014) and Camargo *et al*. (2014). Transcript abundance was assessed by microfluid qPCR using the BioMark® 96.96 Dynamic Array platform (Fluidigm, San Francisco, CA, USA) as explained in Cassan-Wang *et al*. (2012). Gene-specific primer pairs (Table S5) were designed using the quantprime program (Arvidsson *et al*., 2008). A dissociation step was performed after amplification to confirm the presence of a single amplicon. Three reference genes identified in the same tissue panels by Cassan-Wang *et al*. (2012) were used for data normalization: *PP2A1* (protein phosphatase 2A subunit 1) (*Eucgr.B03386*), *PP2A3* (protein phosphatase 2A subunit 3) (*Eucgr.B03031*) and *SAND* (trafficking protein Mon1) (*Eucgr.B02502*). Data normalization was performed using the formula proposed by Pfaffl (2001) and adopting as calibrator a mix of all of the samples used in the assay. The normalized data were further processed in expander6 (Ulitsky *et al*., 2010) for use in hierarchical clustering analysis. The data treatment procedure consisted of data flooring (1 E^{-5}), log transformation and standardization (normalization of each expression pattern to have a mean of 0 and a variance of 1). Hierarchical clustering was performed using Pearson correlation and average linkage. The consistency of the results provided by the biological replicates was evaluated and the results for each sample were averaged using the geometric mean.

RNA-seq expression analysis

RNA-seq data for six tissues from three field-grown *E. grandis* individuals, and root samples prepared from young rooted cuttings, were obtained from EucGenIE (http://eucgenie.bi.up.ac.za; Mizrachi *et al*., 2010; Hefer *et al*., 2011). The absolute transcript abundance values obtained for the 38 lignin biosynthetic genes were computed from fragments per kilobase of exon per million fragments mapped (FPKM) values obtained with TopHat (Trapnell *et al*., 2009) and Cufflinks (Trapnell *et al*., 2010). Values were standardized using expander6 (Ulitsky *et al*., 2010), as described in Soler *et al*. (2014).

Results

In silico identification of phenylpropanoid genes encoding *bona fide* enzymes

Some of the 11 gene families involved in the phenylpropanoid pathway leading to monolignols, such as the *4CL*, *COMT*, *CCR*, *CCoAOMT* and *CAD* families, belong indeed to very large superfamilies resulting in erroneous annotations in genome-wide studies, as thoroughly addressed by Kim *et al*. (2004, 2007). To avoid such problems, it is necessary to identify *bona fide* members, that is, those encoding the true enzymatic reactions. To this end, we performed an extensive literature survey to identify genes from the 11 gene families proved to encode *bona fide* enzymes through biochemical characterization of their enzymatic activities and/or through forward or reverse genetic approaches (summarized in Table S1). A total of 75 genes encoding *bona fide* enzymes from several plant species were included in the phylogenetic analyses together with phenylpropanoid/lignin annotated genes from *Populus trichocarpa*, *Vitis vinifera*, *Arabidopsis thaliana* and *Oryza sativa* and the 175 putative phenylpropanoid genes retrieved from the *E. grandis* genome (Table S2). The phylogenetic reconstructions (Figs S1–S6) enabled us to delimit *bona fide* clades for each family and propose a selected subset of 38 *E. grandis bona fide* genes (Table S2).

PAL

PAL (EC: 4.3.1.5) is the first enzyme of the general phenylpropanoid pathway catalyzing the deamination of phenylalanine to produce cinnamic acid, a universal intermediate in the formation of a large variety of plant-specific phenylpropanoid derivatives. We constructed a phylogenetic tree (Fig. 1a) in which we highlighted the genes encoding *bona fide* PAL enzymes. In comparison to *A. thaliana*, *Petroselinum crispum*, tobacco and *Medicago truncatula*, where PAL is encoded by three to four members, the *E. grandis PAL* family comprises nine genes, which is three more than in *Populus*. Eight of the nine EgrPAL proteins were positioned in the *PAL bona fide* clade (Fig. 1a). Only EgrPAL2 was placed outside, being the most phylogenetically divergent member, sharing only 57–61% amino acid sequence identity with the other EgrPAL members and exhibiting a closer relationship with gymnosperm PAL proteins (Fig. 1a). The *EgrPAL2* gene was highly expressed in flowers and shoot tips, exhibiting weak expression in xylem (Fig. 1b,c). EgrPAL1 was found to be phylogenetically close to AtPAL1 and AtPAL2, and has been reported to be mainly involved in anthocyanin production (Rohde *et al*., 2004; Huang *et al*., 2010). In agreement with this putative role, this gene showed weak overall expression, mostly restricted to flower tissues.

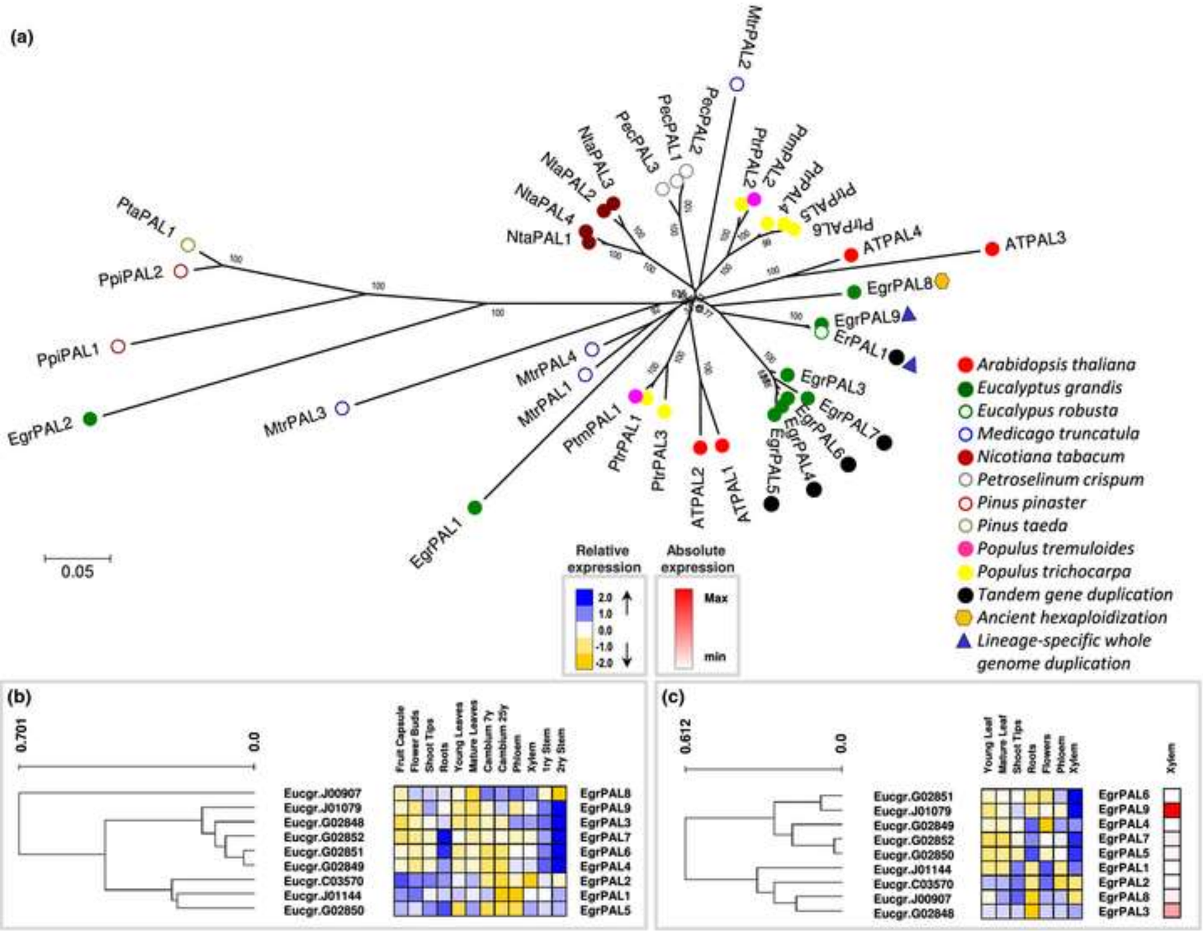


Figure 1. The phenylalanine ammonia-lyase (PAL) *bona fide* clade: comparative phylogeny and expression profiles. (a) Unrooted protein phylogenetic tree constructed with PAL *bona fide* enzymes from several species. A total of 797 nonambiguous amino acids positions were considered in the final data set. (b, c) Heatmaps of transcript accumulation patterns of *EgrPAL* genes generated by (b) microfluid real-time quantitative PCR (RT-qPCR) and (c) RNA-seq. The gene accession number and short name are indicated on the left .

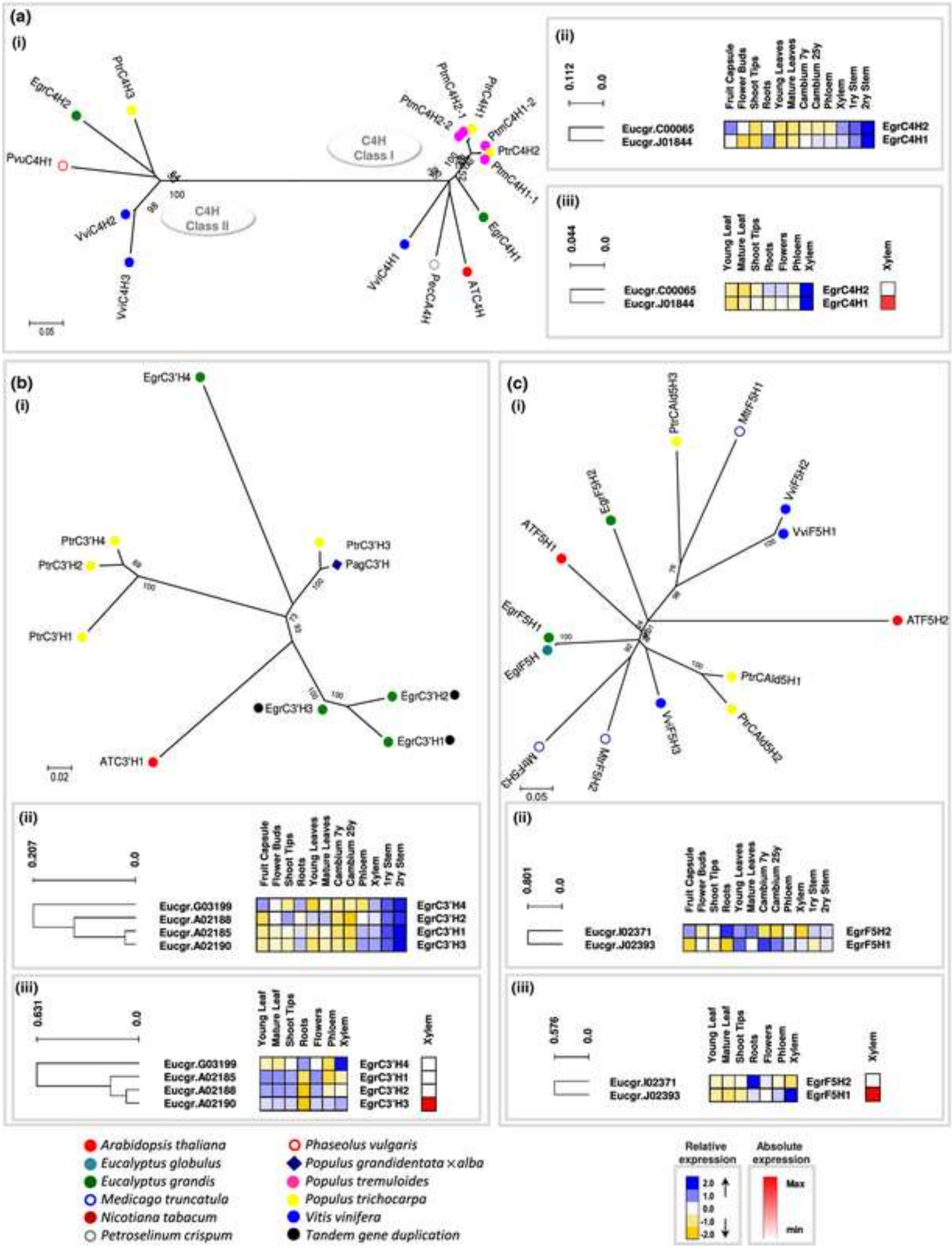


Figure 2. The cinnamate 4-hydroxylase (C4H), *p*-coumarate 3-hydroxylase (C3′H) and ferulate 5-hydroxylase (F5H) *bona fide* clades: comparative phylogeny and expression profiles. (i) Unrooted protein phylogenetic trees constructed with *bona fide* (a) C4H, (b) C3′H and (c) F5H enzymes from several species. A total of 542 (C4H), 511 (C3′H) and 565 (F5H) nonambiguous amino acid positions were considered in the final data sets. (ii, iii) Heatmaps of transcript accumulation patterns of two *EgrC4H*, four *EgrC3′H* and two *EgrF5H* genes were generated by (ii) microfluid real-time quantitative PCR (RT-qPCR) and (iii) RNA-seq.

EgrPAL8 also showed strong expression in phloem, shoot tips and flowers. *EgrPAL9* was placed in the same subclade as *AtPAL4* and was the most abundantly expressed *EgrPAL* gene, showing preferential expression in developing xylem, although not being highly specific (Fig. 1b). The *EgrPAL3* gene experienced at least two rounds of tandem duplication producing four additional lineage-specific genes, *EgrPAL4*, *5*, *6* and *7* (sharing between 97% and 99% protein sequence identity). Although all five of these *PAL* genes showed preferential expression in xylem, *EgrPAL3* was the most highly expressed member of the cluster and also displayed xylem tissue specificity (Fig. 1b,c). Based on their expression patterns, *EgrPAL3* and *9* are the *PAL* genes most likely to be involved in developmental lignification.

The hydroxylation steps

The hydroxylation steps of the monolignol pathway are catalyzed by C4H (EC: 1.14.13.11), C3′H (EC: 1.14.14.1) and F5H (EC: 1.14.13), three members of the cytochrome P450 monooxygenases superfamily, belonging to the CYP73, CYP98 and CYP84 families, respectively.

C4H, the second enzyme of the general phenylpropanoid pathway, catalyzes the 4-hydroxylation of trans-cinnamic acid into 4-hydroxy-cinnamate. With the exception of *A. thaliana*, presenting a single *C4H* gene, C4H is in general encoded by small gene families not exceeding four members. The *E. grandis* genome has two *C4H* members encoding EgrC4H1 and EgrC4H2, sharing 61% identity and belonging to classes I and II, respectively (Fig. 2a (i)). In agreement with its membership of class I, in which several members have a major role in lignin biosynthesis, *EgrC4H1* was highly and preferentially expressed in developing xylem (Fig. 2a (ii,iii)). *EgrC4H2* was also preferentially expressed in xylem, albeit to a lesser extent (sevenfold less), and its overall transcript level was lower than that of *EgrC4H1* (12-fold lower). C4H class II genes have been associated with stress responses (Costa *et al*., 2003; Raes *et al*., 2003; Lu *et al*., 2006), although involvement in vascular lignification has also been shown in *Phaseolus vulgaris* and tobacco (Nedelkina *et al*., 1999; Blee *et al*., 2001).

Together, these results suggest that *EgrC4H1* is the main *C4H* gene involved in lignin biosynthesis, but a role for *EgrC4H2,* although a less prominent role, is also likely.

C3′H catalyzes predominantly the 3-hydroxylation of 4-coumaroyl shikimate. C3H-defective *A. thaliana ref8* (reduced epidermal fluorescence8) mutants exibited lignin depleted in *meta*-hydroxylated G and S units (Franke *et al*., 2002; Abdulrazzak *et al*., 2006). The C3′H-catalyzed reaction is the first irreversible step toward S and G lignins (Anterola & Lewis, 2002). Interestingly, C3′H from *Populus* (PtrC3H3) expressed in yeast converts 4-coumaroyl shikimate, but when it was coexpressed with PtrC4H1 and/or PtrC4H2, a dramatic increase in catalytic activity and efficiency was observed. PtrC4H1, PtrC4H2, and PtrC3′H3 form all three possible heterodimers and a heterotrimer (PtrC4H1/C4H2/C3′H3) likely to be involved in monolignol biosynthesis (Chen *et al*., 2011). Whereas *A. thaliana* has a single *C3'H* gene (Raes *et al*., 2003)*,* in *E. grandis,* the *C3'H* family encompasses four members. Three (*EgrC3H1–3*) are located in a 73-kb genomic region of chromosome 1 (Fig. S7) and share 93–95% sequence identity (Table S5). As in *Populus*, this family has been expanded by lineage-specific tandem duplications (Hamberger *et al*., 2007). *EgrC3H4* is located on chromosome 7 and shares 76–80% similarity with the other three *E. grandis* genes. RNA-seq profiling highlighted marked distinctions between the four members of this family (Fig. 2b-iii). For instance, the expression profile of *EgrC3'H3* was distinct from that of *EgrC3'H1* and *2,* exhibiting a strong and preferential expression in xylem, while the other two genes were

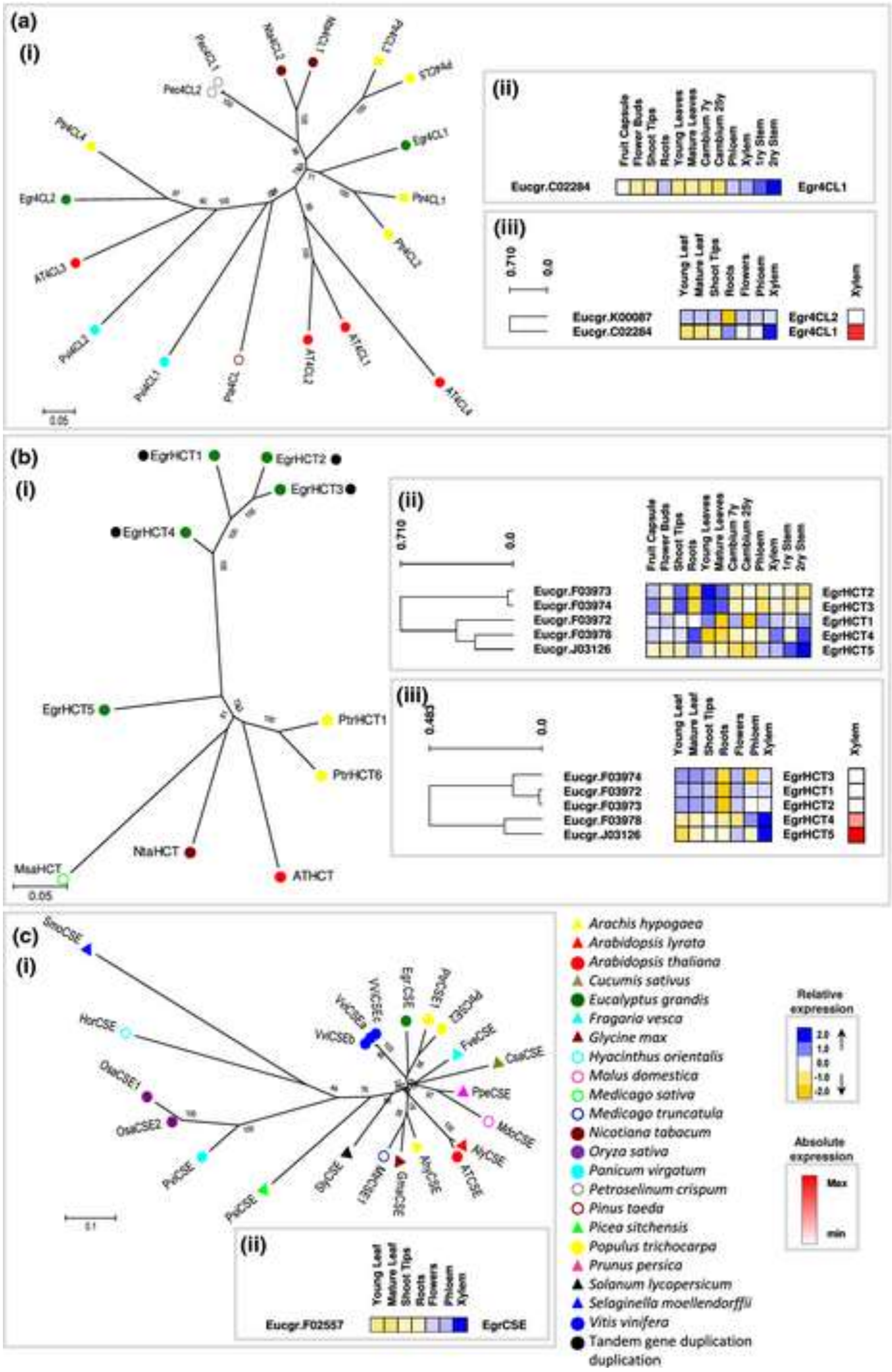


Figure 3. The 4-coumarate:CoA ligase (4CL), shikimate *O*-hydroxycinnamoyltransferase (HCT) and caffeoyl shikimate esterase (CSE) *bona fide* clades: comparative phylogeny and expression profiles. (i) Unrooted protein phylogenetic tree constructed with *bona fide* (a) 4CL, (b) HCT and (c) CSE enzymes from several species. A total of 614 (4CL), 514 (HCT) and 443 (CSE) nonambiguous amino acid positions were considered in the final data sets. (ii, iii) Heatmaps of transcript accumulation patterns of *Egr4CL, EgrHCT* and *EgrCSE* genes were generated by (ii) microfluid real-time quantitative PCR (RT-qPCR) and (iii) RNA-seq.

preferentially expressed in leaves, shoot tips and flowers. The xylem preferential expression was also shown by *EgrC3'H4*. However, *EgrC3'H3* was by far the most strongly expressed member in xylem, with a 65-fold higher expression than *EgrC3'H4* (Fig. 2b). *EgrC3'H3* and to a lesser extent *EgrC3'H4* are likely to be involved in developmental lignification in *Eucalyptus*.

F5H (EC: 1.14.13), also called coniferaldehyde/coniferyl alcohol 5-hydroxylase to better reflect its preferred substrates (Humphreys *et al*., 1999; Osakabe *et al*., 1999), is involved in the pathway leading to sinapyl alcohol and, ultimately, to S lignin. The *A. thaliana* genome harbors two *F5H* paralogs encoding AtF5H1 (CYP84A1) and AtF5H2 (CYP84A4). *AtF5H1* had been shown to be involved in lignification through analysis of the *fah1* (ferulate-5-hydroxylase-deficient) mutant which has little to no S lignin (Meyer *et al*., 1998; Marita *et al*., 1999). By contrast, the overexpression of *AtF5H1* in the mutant background produces plants displaying substantially higher proportion of S units than normal: up to *c*. 92% in *AtF5H1*-upregulated *A. thaliana* (Meyer *et al*., 1998), up to 84% in tobacco (Franke *et al*., 2000), and as high as 93.5% in hybrid *Populus* (Stewart *et al*., 2009). The gene previously termed *AtF5H2* was shown recently to be an *A. thaliana*-specific paralog of *AtF5H1*, originating from a recent duplication event that led to neofunctionalization. Indeed, the encoded enzyme (CYP84A4) is involved in the biosynthesis of alpha-pyrones and generates the catechol-substituted substrate for an extradiol ring-cleavage dioxygenase (Weng *et al*., 2012).

In *E. grandis,* we also identified two *EgrF5H* genes, both of which showed closer phylogenetic proximity to *AtF5H1*. They are located in different chromosomes (*EgrF5H2* in chromosome 9 and *EgrF5H1* in chromosome 10; Fig. S7) and encode proteins sharing 84% amino acid sequence similarity (Table S4). *EgrF5H2* showed modest overall expression, almost exclusively restricted to root tissues, whereas *EgrF5H1* was very highly and preferentially expressed (93%) in developing xylem (Fig. 2c (ii,iii)). The EgrF5H1 protein shares 99.6% similarity with its *E. globulus* ortholog recently shown to be involved in vascular lignification (García *et al*., 2014).

4CL

4CL (EC: 6.2.1.12), the third and last enzyme of the general phenylpropanoid pathway, catalyzes the formation of CoA thiol esters of 4-coumarate and other 4-hydroxycinnamates in a two-step reaction involving the formation of an adenylate intermediate (Ehlting *et al*., 2001). The 4CL family comprises the acyl:CoA synthetase (ACS) superfamily (Hamberger *et al*., 2007; de Azevedo-Souza *et al*., 2008), as shown in Fig. S1, allowing discrimination of the 4CL *bona fide* clade (Fig. 3a (i)), grouped into two classes according to Ehlting *et al*. (1999). Like *Populus* and *A. thaliana*, *E. grandis* has a single representative (*Egr4CL2*) in class II associated with flavonoid and soluble phenolic biosynthesis (Ehlting *et al*., 1999). The *Egr4CL2* gene was preferentially expressed in shoot tips, flowers and leaves, consistent with its possible function in *Eucalyptus*. In contrast to other dicots which have two to four members in class I (Hamberger *et al*., 2007), *E. grandis* has a single member (*Egr4CL1*). In agreement with the predicted role of members of class I in monolignol biosynthesis, *Egr4CL1* was strongly and preferentially expressed in developing xylem (Fig. 3a (ii,iii)).

HCT (hydroxycinnamoyl CoA:shikimate hydroxycinnamoyl transferase)

HCT (EC: 2.3.1.133) catalyzes the reactions both immediately preceding and following the insertion of the 3-hydroxyl group by C3H into monolignol precursors (Hoffmann *et al*., 2003, 2004). HCT uses *p*-coumaroyl-CoA and caffeoyl-CoA as preferential substrates to transfer an acyl group to the acceptor compound shikimic acid, yielding *p*-coumaroyl shikimate. A closely related acyl transferase, hydroxy-cinnamoyl CoA:quinate hydroxycinnamoyl transferase (HCQ), yielding *p*-coumaroyl quinate esters, is involved in chlorogenic acid biosynthesis and not lignin biosynthesis (Niggeweg *et al*., 2004; Umezawa, 2010). Because of the high similarities between HCT and HCQ, we first constructed a phylogenetic tree including both HCT and HCQ proteins (Fig. S2), allowing us to distinguish the two major clades corresponding to the *bona fide* HCQ and HCT, respectively. The five EgrHCTs belong to the latter, showing a lineage-specific expansion as compared with the other dicots where only one (*A. thaliana* and *Medicago*) or two members (*Populus*) are present (Fig. 3b (i)). *EgrHCT1* to *4* are in tandem arrangement in an 87-kb genomic region of chromosome 6 (Fig. S7), and have high amino acid sequence identities (89–96%) resulting from recent tandem gene duplication events. *EgrHCT1*, *2* and *3* exhibited very similar expression profiles (highly expressed in leaves). The fourth member of the tandem array, *EgrHCT4,* exhibited a very distinct profile, being highly and preferentially expressed in xylem, thus suggesting that functional divergence occurred after tandem duplication. Its expression profile clustered with that of *EgrHCT5* which was, however, more strongly and preferentially expressed in xylem. *EgrHCT5* is located on a different chromosome (chromosome 10; Fig. S7); the corresponding protein was phylogenetically distinct from the other four, presenting 66–69% amino acid sequence identity (Table S4), and being closer to the HCT from other species. *EgrHCT5* and *EgrHCT4* are therefore probably involved in lignin biosynthesis.

CSE

The involvement of CSE (EC: 3.1.1.-) in lignin biosynthesis was indicated very recently by an analysis of an *A. thaliana cse*-2 (caffeoyl shikimate esterase 2) knockout mutant that presented a reduced lignin content enriched in H units and depleted in S units (Vanholme *et al*., 2013). CSE hydrolyzes caffeoyl shikimate into caffeoyl-CoA (Vanholme *et al*., 2013). Together with 4CL, CSE was proposed to be involved in an alternative pathway leading to the formation of caffeoyl-CoA, bypassing the second reaction performed by HCT. We mainly used the eudicotyledon orthologs of *AtCSE* previously reported (Vanholme *et al*., 2013) to generate the CSE phylogenetic tree (Fig. 3c (i)). The *EgrCSE* gene has a closer phylogenetic relationship to the grapevine (*V. vinifera*) and the *Populus CSE* genes. The *EgrCSE* gene was found to be strongly and preferentially expressed in *E. grandis* developing xylem tissues, supporting its inferred role in lignin biosynthesis.

The methylation steps

Plants present a wide variety of *S*-adenosyl-l-Met-dependent *O*-methyltransferases (OMTs) that act on Phe-derived substrates during the production of numerous plant secondary compounds in addition to lignin (Eckardt, 2002). COMT (EC: 2.1.1.68) and CCoAOMT (EC: 2.1.1.104) are both involved in methylation steps of the monolignol pathway (Ye & Varner, 1995).

CCoAOMT catalyzes the methylation of caffeoyl CoA to produce feruloyl CoA. Functional characterization of CCoAOMT in several plant species (see Table S1) revealed that it was involved in the synthesis of G units. Focusing on the *bona fide* CCoAOMT discrete clade (Fig. 4a (i)) delimited from a superfamily tree (Fig. S3) allowed us to identify two *EgrCCoAOMT* genes, the same number as in *Populus* (Chen *et al*., 2000). *EgrCCoAOMT1* was located on chromosome 9 and *EgrCCoAOMT2* on chromosome 7 (Fig. S7). They were reported to originate from a lineage-specific whole-genome duplication event (Myburg *et al*., 2014). Both *EgrCCoAOMT* genes showed strong, preferential expression in developing

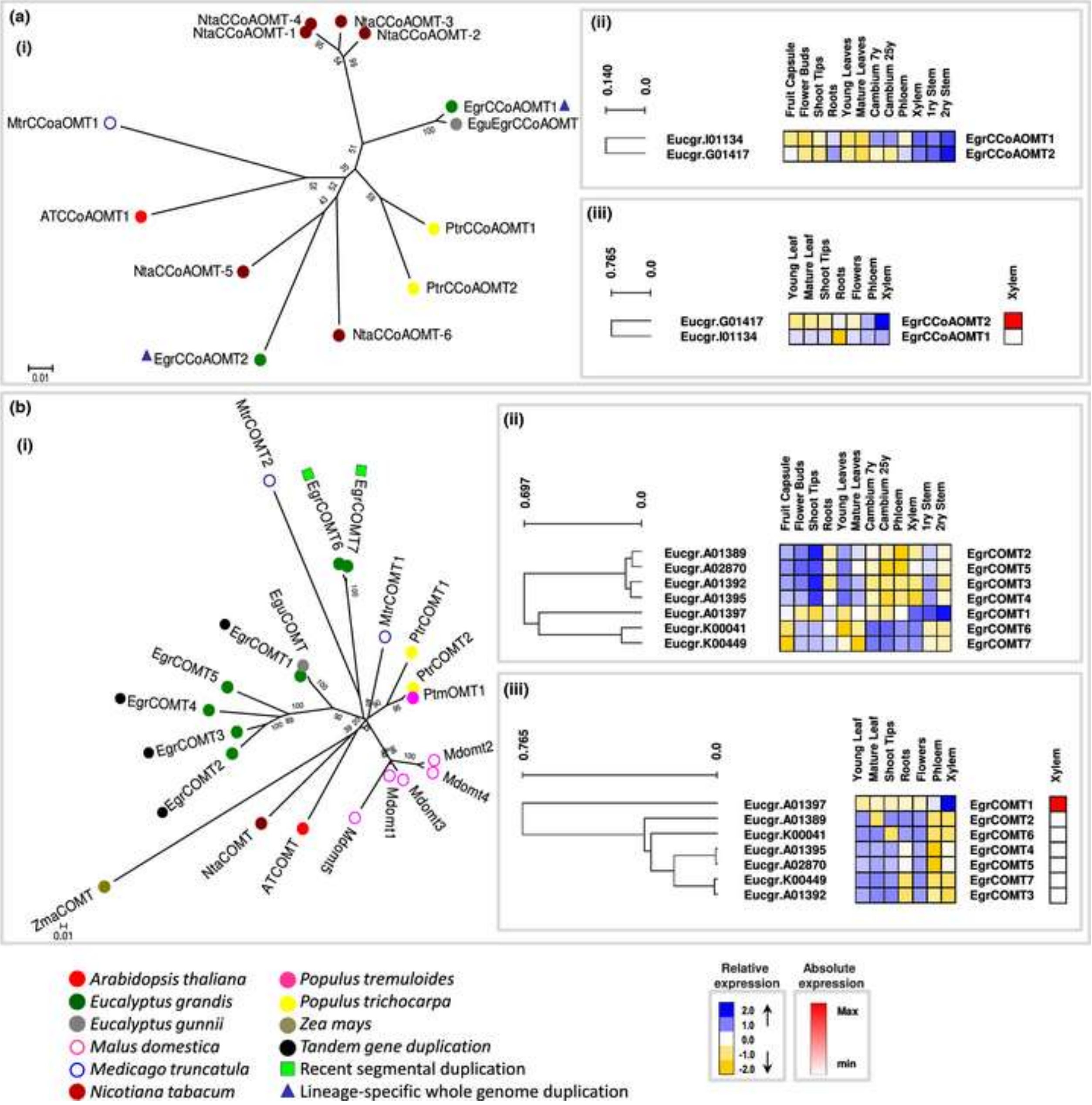


Figure 4. The caffeoyl CoA 3-*O*-methyltransferase (CCoAOMT) and caffeate/5-hydroxyferulate *O*-methyltransferase (COMT) *bona fide* clades: comparative phylogeny and expression profiles. (i) Unrooted protein phylogenetic tree constructed with *bona fide* (a) CCoAOMT and (b) COMT enzymes from several species. A total of 262 (CCoAOMT) and 365 (COMT) nonambiguous amino acid positions were considered in the final data set. (ii, iii) Heatmaps of transcript accumulation patterns of two *EgrCCoAOMT* and seven *EgrCOMT* genes were generated by (ii) microfluid real-time quantitative PCR (RT-qPCR) and (iii) RNA-seq.

xylem tissues, although *EgrCCoAOMT2* presented a twofold higher expression relative to *EgrCCoAOMT1* (Fig. 4a (iii)).

In angiosperms, COMT (EC: 2.1.1.68) was initially thought to be a bifunctional enzyme similarly involved in the synthesis of G and S precursors. However, in several plants species, COMT down-regulation led to a dramatic decrease in S units and to the incorporation of 5-hydroxyguaiacyl units (5-OH-G) into lignins. COMT is now considered to be pre-eminently involved in the synthesis of S units through the preferential methylation of 5-hydroxyconiferyl aldehyde into sinapaldehyde (Davin *et al.*, 2008). We assembled a phylogenetic tree for the OMT superfamily (Fig. S4) in which we delimited the *bona fide* COMT clade (Fig. 4b (i)). Compared with other species harboring only one or two members within the *bona fide* clade, the *E. grandis* COMT family is expanded with seven *EgrCOMT* genes resulting from distinct duplication mechanisms. Expression profiling highlighted *EgrCOMT1* as being by far the most expressed *EgrCOMT* member, with massive and highly specific expression in developing xylem. The closest ortholog to *EgrCOMT1* (99% similarity) was found to be the *E. gunnii COMT* (Poeydomenge *et al*., 1994). *EgrCOMT1*, *2*, *3* and *4* were all located in a 135-kb genomic region of chromosome 1 and resulted from tandem gene duplication events (Fig. S7) followed by functional divergence. Indeed, *EgrCOMT2*, *3* and *4* presented similar expression profiles that were very different from that of *EgrCOMT1*, exhibiting low expression in highly lignified tissues and higher expression over the other sampled tissues. *EgrCOMT2* showed very high expression in shoot tips (Fig. 4b (ii)). Together with *EgrCOMT5*, *EgrCOMT1–4* formed a distinct subclade (Fig. 4b (i)). *EgrCOMT5* was located in another region of chromosome 1, exhibiting a similar tendency although with lower expression levels. All the amino acid residues described to be involved in the catalytic activity, substrate binding and chemical interactions of COMT enzymes (Zubieta *et al*., 2002) are only fully conserved in EgrCOMT1. For the remaining proteins EgrCOMT2–5, the catalytic center amino acids (His-269, Glu-329 and Glu-297) are conserved but residue substitutions are found in the lignin monomer binding interaction residues, COMT methoxy and SAM/SAH (S-adenosylmethionine/S-adenosylhomocysteine) substrate binding pockets (data not shown). *EgrCOMT6* and *EgrCOMT7* were reported as having originated by recent segmental duplication (Myburg *et al*., 2014). Their predicted

proteins sharing 99% amino acid sequence identity (Table S4) were positioned in the same subclade as AtCOMT and NaCOMT. These two genes were found to be preferentially expressed in leaves and flowers according to *E. grandis* RNA-seq data (Fig. 4b (iii)), but were also detected in cambium, xylem and phloem tissues in RT-qPCR experiments performed with a larger, different organ and tissue panel. *EgrCOMT1* is the most likely candidate involved in developmental lignification.

The two last reductive steps

CCR (EC: 1.2.1.44) catalyzes the first committed step of the lignin-specific branch of monolignol biosynthesis by converting cinnamoyl CoA esters to their corresponding cinnamaldehydes. The first cDNA (*EguCCR*) and a genomic clone encoding CCR were isolated from *E. gunnii* and their identity confirmed by enzymatic characterization of the corresponding recombinant protein (Lacombe *et al*., 1997; Lauvergeat *et al*., 2001). The cDNA was used to clone the tobacco CCR cDNA leading to the first transgenic plants with down-regulated CCR activity, exhibiting a severe reduction (50%) in their lignin content (Piquemal *et al*., 1998). An extensive study of the CCR and CCR-like gene superfamily in land plants revealed a distribution in three major subfamilies, and highlighted the *bona fide*

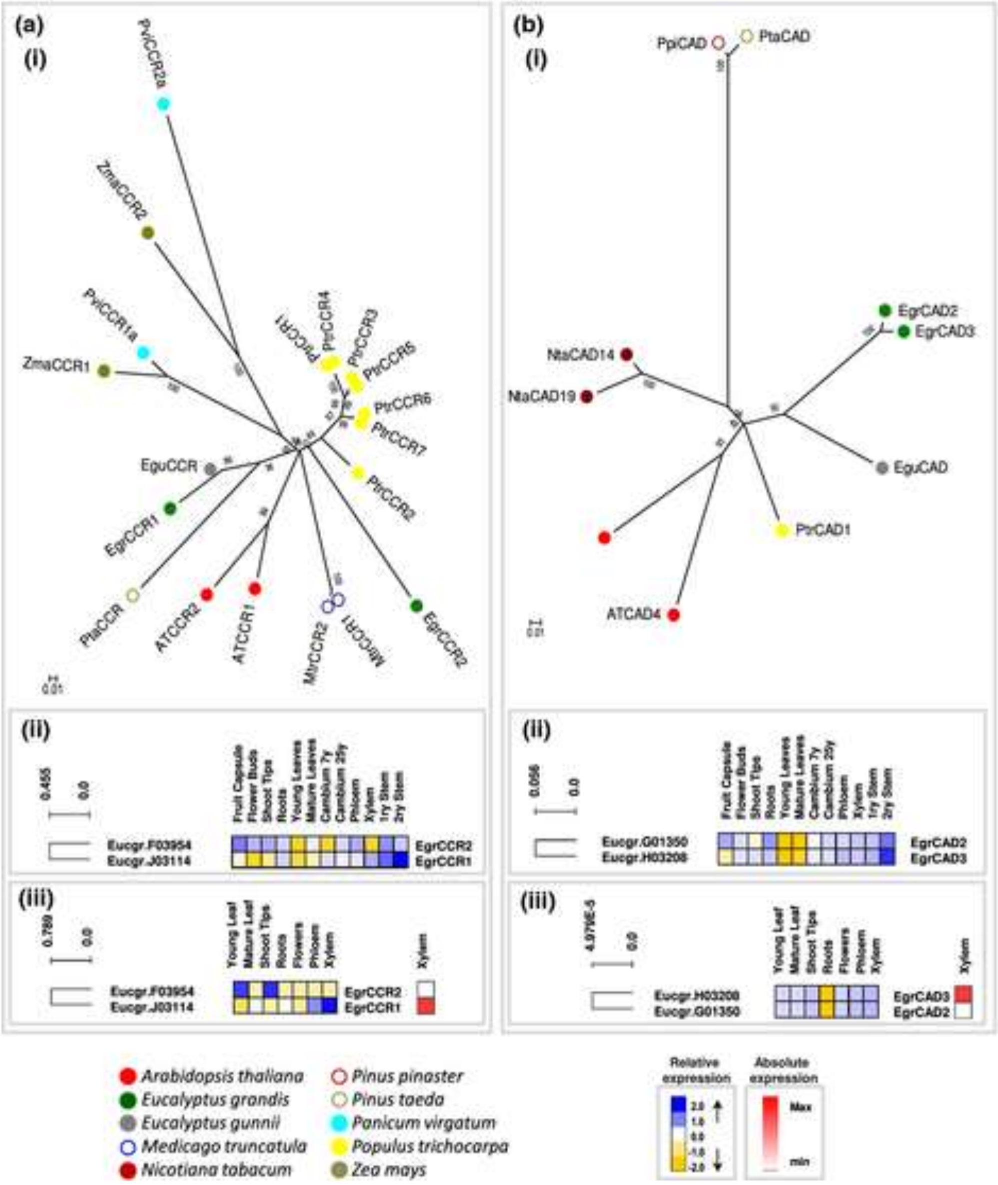


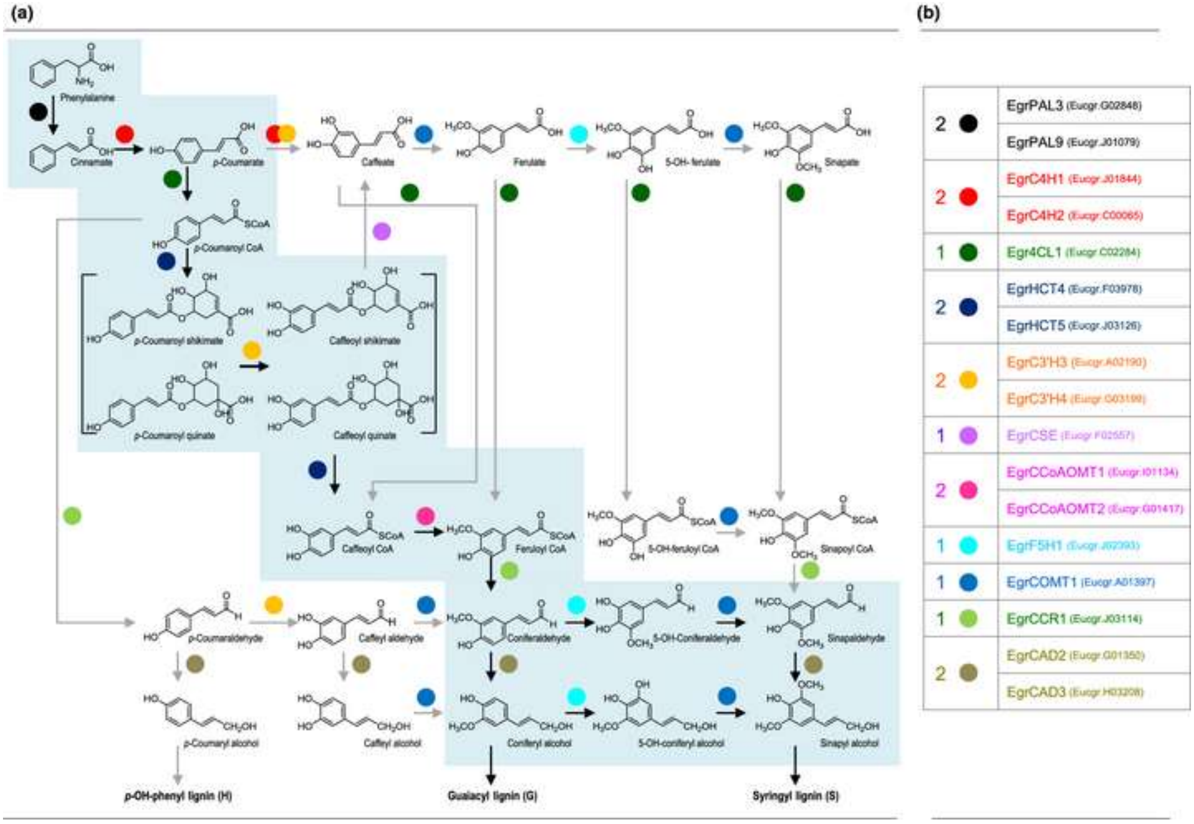
Figure 5. The cinnamoyl CoA reductase (CCR) and cinnamyl alcohol dehydrogenase (CAD) *bona fide* clades: comparative phylogeny and expression profiles. (i) Unrooted protein phylogenetic trees constructed with *bona fide* (a) CCR and (b) CAD enzymes from several species. A total of 380 (CCR) and 390 (CAD) nonambiguous amino acid positions were considered in the final data set. Heatmaps of transcript accumulation patterns of two *EgrCCR* and two *EgrCAD* genes were generated by (ii) microfluid real-time quantitative PCR (RT-qPCR) and (iii) RNA-seq.

CCR family (Barakat *et al*., 2011; this study, Fig. S5). In addition to EgrCCR1, which is closely related (91% identity) to the *E. gunnii* EguCCR1 and probably involved in lignin biosynthesis, the *bona fide* clade also includes a second *CCR* gene (*EgrCCR2*), reported for the first time in *Eucalyptus* (Fig. 5a (i)). *EgrCCR1* and *EgrCCR2* are located on distinct chromosomes (chromosome 10 and 6, respectively) and encode proteins sharing only 56% amino acid sequence identity. Both proteins harbor the CCR signature, NWYCY (Lacombe *et al*., 1997), essential for its enzymatic activity (Escamilla-Trevino *et al*., 2010). Other species have been shown to harbor two functional *CCR* genes, such as *A. thaliana* (Lauvergeat *et al*., 2001), maize (*Zea mays*) (Pichon *et al*., 1998; Tamasloukht *et al*., 2011), and switchgrass (*Panicum virgatum*) (Escamilla-Trevino *et al*., 2010). In all cases, the *AtCCR1* gene was shown to be involved in developmental lignin biosynthesis, while *AtCCR2* was poorly expressed during development but was inducible by biotic or abiotic stresses and hypothesized to play a role in defense mechanisms. Indeed, a similar situation was found in *E. grandis* where the two *EgrCCR* genes showed distinct overall expression levels and patterns. *EgrCCR1* was found to be strongly and preferentially expressed in developing xylem, in agreement with a role in developmental lignin biosynthesis. By contrast, *EgrCCR2,* which presented very low overall expression, was expressed in *E. grandis* shoot tips (RNA-seq data), whereas it could be detected in developing xylem by RT-qPCR (Fig. 5a (ii)).

CAD (EC: 1.1.1.195) catalyzes the reduction of hydroxycinnamyl aldehydes to their corresponding alcohols. *CAD* and *CAD*-like genes have been reported in numerous species as forming large multigene families containing members exhibiting a low degree of homology and different affinities to various substrates, and probably having various physiological roles (Sibout *et al*., 2003). Several authors have suggested the existence of three evolutionary CAD classes, differing in their patterns of evolution and expression (Costa *et al*., 2003; Raes *et al*., 2003; Sibout *et al*., 2005; Barakat *et al*., 2009; Guo *et al*., 2010). Class I comprises all *bona fide* CAD genes, as shown in the broad CAD/CAD-like phylogenetic tree (Fig. S6). Interestingly, in this *bona fide* clade, *A. thaliana*, *Vitis*, parsley, tobacco and *E. grandis* are represented by two CAD proteins, *Populus* being the exception with only a single *CAD* gene (*PtrCAD1*) associated with monolignol biosynthesis (Fig. 5b (i)). The two *E. grandis* genes, *EgrCAD2* and *EgrCAD3,* are located on chromosomes 7 and 8, respectively, in regions reported to result from a recent segmental duplication event (Myburg *et al*., 2014). The two genes revealed strong phylogenetic proximity (87% sequence identity at the protein level). They are also closely related to EguCAD2, which has been shown to have a true CAD enzymatic activity and to be strongly expressed in xylem tissue (Grima-Pettenati *et al*., 1993). *EgrCAD3* (Eucgr.H03208), initially present in the v1.0 annotation, was later removed in v1.1, but can still be found in the low-confidence transcripts in the current v1.1 annotation. Our expression data support the conclusion that this gene is actively transcribed. The two *EgrCAD* genes presented similar expression patterns, with preferential expression in highly lignified tissues such as secondary stem and developing xylem. *EgrCAD3* exhibited fivefold higher expression than *EgrCAD2* both in developing xylem tissues and overall.

Comparison of hierarchical clustering of RNA-seq and RT-qPCR expression profiling

The step-by-step combined analysis of the phylogeny neighborhood and of the expression profiles for the 38 *bona fide* clade members of the 11 families highlighted a subset of 17 genes as the most likely major genes involved in xylem lignification (Fig. 6). The core vascular lignification toolbox comprises *Egr PAL3* and *9*, *C4H1* and *2*, *4CL1*, *HCT4* and *5*, *SCE*, *C3H3* and *4*, *CCoAOMT1* and *2*, *F5H1*, *COMT1*, *CCR1*, and *CAD2* and *3*. In order to obtain a more global picture of the genes exhibiting strong and/or preferential expression in highly lignified tissues, given the fact that, in a few cases, we noticed distinctive expression patterns between the tissue panels used for microfluid RT-qPCR and RNA-seq, respectively, we carried out independent hierarchical clustering of the overall expression profiling data obtained in the two panels (Fig. 7; Table S6). Both heatmaps revealed large clusters of genes with preferential expression in developing xylem, cambium and/or secondary stem, allowing us to consider eight more genes (*EgrPAL5*, *6*, *7* and *8*, *C3H1* and *2*, and *COMT6* and *7*) as potential candidates for vascular lignification. However, these eight genes were not as strongly expressed in developing xylem as the 17 genes of the core toolbox. Considering their expression patterns, the remaining 13 genes are probably involved in the biosynthesis of phenylpropanoid compounds other than lignin.



2 ●	EgrPAL3 (Eucgr.G02848)
	EgrPAL9 (Eucgr.J01079)
2 ●	EgrC4H1 (Eucgr.J01844)
	EgrC4H2 (Eucgr.C00065)
1 ●	Egr4CL1 (Eucgr.C02284)
2 ●	EgrHCT4 (Eucgr.F03978)
	EgrHCT5 (Eucgr.J03126)
2 ●	EgrC3'H3 (Eucgr.A02190)
	EgrC3'H4 (Eucgr.G03199)
1 ●	EgrCSE (Eucgr.F02557)
2 ●	EgrCCoAOMT1 (Eucgr.I01134)
	EgrCCoAOMT2 (Eucgr.G01417)
1 ●	EgrF5H1 (Eucgr.J02393)
1 ●	EgrCOMT1 (Eucgr.A01397)
1 ●	EgrCCR1 (Eucgr.J03114)
2 ●	EgrCAD2 (Eucgr.G01350)
	EgrCAD3 (Eucgr.H03208)

Figure 6. The *Eucalyptus* lignification toolbox. (a) The biosynthetic pathway was adapted from Humphreys & Chapple (2002) and modified according to the recent findings of Vanholme *et al*. (2013). (b) The 17 *Eucalyptus grandis* genes encoding enzymes located in the *bona fide* clades constitute the core set of a *Eucalyptus* lignification toolbox. Reactions thought to be key steps in lignin biosynthesis are indicated with black arrows.

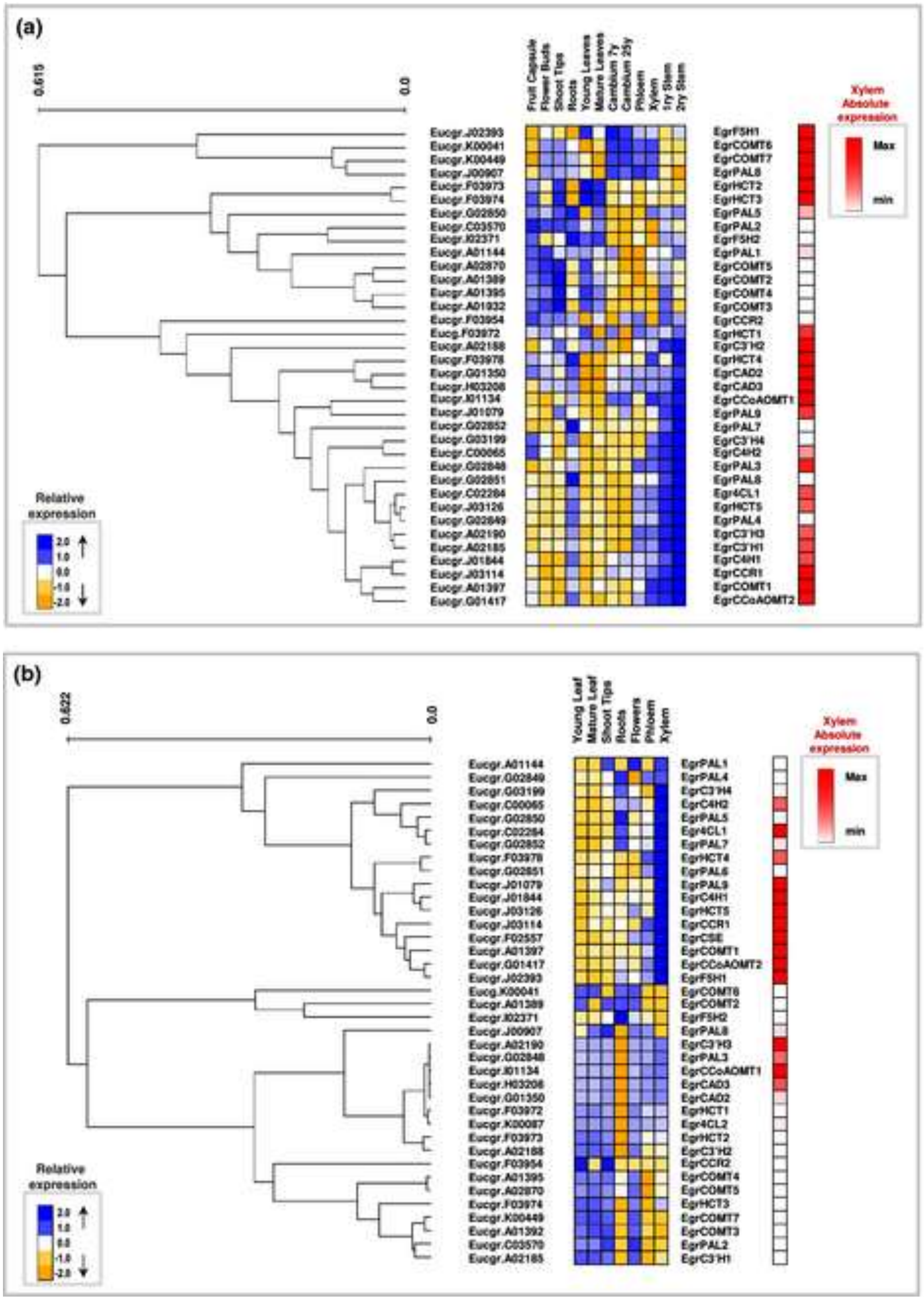


Figure 7. Heatmaps of transcript accumulation patterns of 38 *Eucalyptus* putative monolignol biosynthesis genes using (a) microfluid real-time quantitative PCR (RT-qPCR) and (b) RNA-seq. The gene accession number and short name are indicated on the left. Transcript abundance is expressed in $\log_2$ ratio, with blue colors indicating higher accumulation of transcript and yellow colors indicating lower accumulation. Absolute expression is shown using an *Excel 2010* two-color scale (minimum meaning lower than first quartile and maximum meaning higher than third quartile), with red colors indicating higher accumulation of transcript and white colors indicating lower accumulation.

Environmental and developmental expression characteristics of the 38 *bona fide* genes

Finally, we examined the response of the *bona fide* genes to several environmental cues such as gravitropic stress (tension vs opposite wood), nitrogen fertilization (high vs low) and cold treatment. We also included xylem samples at two different physiological stages (juvenile vs mature). The results are presented in Table S7. Interestingly, most of the 13 genes clearly not included in the lignification toolbox were more responsive to the environmental stimuli compared with those of the lignin toolbox, supporting a role in the synthesis of phenylpropanoids involved in plant defense. For instance, *EgrCCR2* was strongly responsive to gravitropic stimulus, nitrogen supplementation, and cold stress, whereas *EgrCCR1* transcript levels were stable under these treatments. The *EgrCOMT2–5* genes revealed strong and similar responses to cold stress in young leaf tissues, *EgrCOMT4* being the most induced gene. Among the tandem duplicated *COMT* genes (*EgrCOMT1–4*), *EgrCOMT2* was responsive to most stress conditions tested whereas *EgrCOMT1* was only moderately responsive to nitrogen fertilization. Although lignin biosynthetic genes were in general not strongly responsive to environmental stimuli, all were down-regulated in response to high nitrogen fertilization, confirming and extending recent results obtained with the same samples (Camargo *et al*., 2014). Among those lignin genes, *EgrC3H4* was the most responsive, being differentially expressed in response to all tested abiotic stimuli. It was also more expressed in mature than in juvenile wood. The nine members of the *EgrPAL* family had distinct patterns of responses to the environmental stimuli, supporting functional diversification following duplication.

Discussion

Building on the recent availability of the *E. grandis* genome (Myburg *et al*., 2014), we report here a comprehensive genome-wide analysis of the phylogenetic relationships and expression profiles of the phenylpropanoid and lignin biosynthesis gene families in *E. grandis,* highlighting the evolutionary histories of these families and those members that are likely to be involved in lignification during xylem development.

The *E. grandis* phenylpropanoid/lignin families have different evolutionary histories

Seven of the 11 families presented no traces of tandem duplications, harboring low- or even single-copy genes. According to De Smet *et al*. (2013), single-copy genes are often involved in essential housekeeping functions highly conserved across species, and result from selection pressure to preserve them as singletons. For instance, C4H and F5H have critical functions regulating lignification and determining lignin monomer composition (Franke *et al*., 2000). In addition, when proteins function as dimers or part of larger multiprotein complexes, the presence of multiple isoforms could disrupt protein stability and functionality (Cannon *et al*., 2004). C4H is indeed known to be involved in membrane multiprotein complexes with PAL (Achnine *et al*., 2004) and also with C3′H (Chen *et al*., 2011, and references therein), allowing metabolic channeling.

Although tandem duplication has been described as the most prominent mechanism contributing to multigene family expansion in *E. grandis*, shaping the size and the biological functions of many families (Hussey *et al*., 2014; Myburg *et al*., 2014; Soler *et al*., 2014), only four *bona fide* lignification families (PAL, HCT, C3′H and COMT) were shaped by tandem gene duplication events. In sharp contrast, the tandem duplication mechanism had a major impact on the non-*bona fide* clades of the 4CL, COMT, CCoAOMT and CAD families.

The frequencies of tandem duplicated genes in the *bona fide* PAL (56%), HCT (80%), C3′H (75%) and COMT (57%) clades exceeded the 34.5% observed at the genome level (Myburg *et al*., 2014). In the COMT, HCT and C3′H families, only one of the tandem duplicated genes showed preferential expression in highly lignified tissues. Gene duplication has been recognized as a primary mechanism for increasing functional diversification, and the increased expression divergence in duplicated genes can substantially contribute to morphological diversification (Wang *et al*., 2012). Indeed, the phenylpropanoid pathway leads to a wide variety of compounds such as flavonoids, lignans and hydroxycinnamate derivatives produced in specific metabolic branches and participating in diverse cellular processes underlying plant growth, development, adaptation, defense and reproduction (Tsai *et al*., 2006). In metabolic terms, this pattern might derive from cellular strategies to adjust protein synthesis according to functional needs, keeping only one or two highly expressed genes in each family, in a particular tissue, and/or in particular environmental conditions. Many of the *E. grandis COMT* and *HCT* genes expanded via tandem duplication tended to be involved in responses to environmental stimuli, consistent with earlier observations on tandem duplicated genes (Hanada *et al*., 2008). For instance, of the tandem duplicated genes *EgrCOMT1–4*, only *EgrCOMT1* is likely to be involved in developmental lignification; the other three have only residual expression in xylem and are inducible by abiotic stresses to different levels, suggesting neo- or subfunctionalization following duplication from a common ancestor. Supporting this assumption, some residue substitutions were found in the lignin monomer binding interacting residues, COMT methoxy and SAM/ASH binding pockets of EgrCOMT2–5 proteins as compared with *bona fide* COMTs. EgrCOMT2–5 could be involved in the synthesis of phenylpropanoid compounds in response to environmental cues or in as yet unknown pathways specific to *Eucalyptus*, similar to the recently demonstrated role of CYP84A4 which originates from AtHCT1 (CYP84A1) but evolved to perform a different enzymatic reaction to produce *A. thaliana*-specific alpha-pyrones (Weng *et al*., 2012).

The situation was different for the PAL family, where *EgrPAL3* experienced a high rate of tandem duplication, producing four additional lineage-specific genes, all showing preferential expression in xylem. Such a functional redundancy has been associated with increasing robustness of biological systems through functional buffering, suggesting that the eventual loss of function in one copy can be compensated for by other copies (Gu, 2003). In some cases, gene duplication has also been proposed to confer an immediate selective advantage by facilitating elevated expression (Hurles, 2004), resulting in protein dosage benefits. Thus, lineage-specific duplications of some phenylpropanoid families can lead to both functional redundancy and to divergence, probably contributing to the wide adaptability of the *Eucalyptus* species to the challenging Australian environment.

For the *bona fide CAD* and *CCoAOMT* families, duplication mechanisms other than tandem duplication seemed to be more important in family evolution. Both families present gene pairs that retained similar expression profiles, suggesting functional redundancy and substantial involvement in vascular lignification. The two *CCoAOMT* genes originated from a lineage-specific whole-genome duplication event (Myburg *et al*., 2014).

The PAL and COMT families were also affected by other duplication mechanisms. For instance, *EgrPAL1* and *EgrPAL8* were traced back to an ancient hexaploidization event shared by the core eudicots following which only a small proportion of duplicated genes have survived subsequent gene loss. *EgrPAL9* and *EgrPAL3* resulted from a lineage-specific whole-genome duplication event detected in *E. grandis* (Myburg *et al*., 2014).

Expression profiling identifies a core *Eucalyptus* vascular lignification toolbox within the *bona fide* genes

For each of the 11 *bona fide* gene families, we combined comparative phylogeny within the *bona fide* clades with individual gene developmental expression profiling in two independent data sets (RNA-seq and qRT-PCR). This approach identified 17 genes as being likely to be involved in developmental lignin biosynthesis and constituting the so-called 'core lignin toolbox', while eight more were identified by hierarchical clustering considering all of the 38 gene expression patterns. Collectively, we can consider the *E. grandis* lignin toolbox as being composed of 25 members (17 strong candidates and eight additional genes possibly involved). As many of these genes have not been reported before, this enriches our knowledge of the lignin biosynthetic pathway in eucalypts, although a role in the biosynthesis of heartwood extractives, lignans, polyphenols and condensed tannins that are present in xylem cannot be excluded. The 13 remaining genes are probably involved in the biosynthesis of phenylpropanoid compounds other than lignin. For instance, *EgrPAL1* may be involved in anthocyanin production like *EgrPAL1*, because it is phylogenetically close to *AtPAL1* and *2* (Rohde *et al*., 2004; Huang *et al*., 2010). Several of these genes were hypothesized to be involved in plant defense as inferred from their orthologs in other species. Consistent with this, we found that many of these genes (like *EgrCCR2*) were highly responsive to environmental stimuli.

In conclusion, this study provides a strong basis for future functional studies as well as for breeding and/or engineering eucalypts with improved wood properties for current uses such as pulp manufacture and also for future end uses such as biofuels and biomaterials.

Acknowledgements

This work, part of the LABEX project TULIP (ANR-10-LABX-41), was supported by grants from the Project Tree For Joules (ANR-2010-KBBE-007-01 and FCT-P-KBBE/AGR_GPL/0001/2010), the CNRS, the Toulouse III University (UPS), and the FCT project *microEGo* (PTDC/AGR-GPL/098179/2008). V.C. was supported by a FCT PhD grant (SFRH/BD/72982/2010). M.S. received a postdoctoral fellowship 'Beatriu de Pinós' from the DURSI de la Generalitat de Catalunya. J.A.P.P. acknowledges FCT for the research contract Ciência 2008 program and the postdoctoral fellowship SFRH/BPD/92207/2013. We are grateful to C. Araujo and L. Neves (Altri Florestal, Portugal), C. Marques (RAIZ, Portugal) and L. Harvengt (FCBA, France) for kindly providing and/or allowing collection of *Eucalyptus* samples, and C. Graça (IBET, IICT), N. Saidi, H. Yu and E. Camargo (all LRSV) for help with sample design/collection and RNA extraction. We also thank N. Ladouce (LRSV) and the Genotoul Genomic platform http://get.genotoul.fr/ for technical assistance with the Biomark Fluidigm RT-qPCR amplifications, and E. Mizrachi for assistance with RNA-seq data analysis. RNA-seq was funded by Mondi and Sappi through the FMG Program, the THRIP Program (UID 80118), the NRF (UID 71255 and 86936) and the DST of South Africa.

References

Abdulrazzak N, Pollet B, Ehlting J, Larsen K, Asnaghi C, Ronseau S, Proux C, Erhardt M, Seltzer V, Renou J et al. 2006. A coumaroyl-ester-3-hydroxylase insertion mutant reveals the existence of nonredundant meta-hydroxylation pathways and essential roles for phenolic precursors in cell expansion and plant growth. *Plant Physiology* 140: 30–48.

Achnine L, Blancaflor E, Rasmussen S, Dixon R. 2004. Colocalization of l-phenylalanine ammonia-lyase and cinnarnate 4-hydroxylase for metabolic channeling in phenylpropanoid biosynthesis. *Plant Cell* 16: 3098–3109.

Anterola A, Lewis N. 2002. Trends in lignin modification: a comprehensive analysis of the effects of genetic manipulations/mutations on lignification and vascular integrity. *Phytochemistry* 61: 221–294.

Arvidsson S, Kwasniewski M, Riano-Pachon D, Mueller-Roeber B. 2008. QuantPrime – a flexible tool for reliable high-throughput primer design for quantitative PCR. *BMC Bioinformatics* 9: 465.

de Azevedo-Souza C, Barbazuk B, Ralph SG, Bohlmann J, Hamberger B, Douglas CJ. 2008. Genome-wide analysis of a land plant-specific acyl: coenzyme A synthetase (ACS) gene family in *Arabidopsis*, poplar, rice and Physcomitrella. *New Phytologist* 179: 987–1003.

Barakat A, Bagniewska-Zadworna A, Choi A, Plakkat U, DiLoreto D, Yellanki P, Carlson J. 2009. The cinnamyl alcohol dehydrogenase gene family in *Populus*: phylogeny, organization, and expression. *BMC Plant Biology* 9: 26.

Barakat A, Yassin N, Park J, Choi A, Herr J, Carlson J. 2011. Comparative and phylogenomic analyses of cinnamoyl-CoA reductase and cinnamoyl-CoA-reductase-like gene family in land plants. *Plant Science* 181: 249–257.

Bhuiyan NH, Selvaraj G, Wei Y, King J. 2009. Role of lignification in plant defense. *Plant Signaling & Behavior* 4: 158–159.

Blee K, Choi J, O'Connell A, Jupe S, Schuch W, Lewis N, Bolwell G. 2001. Antisense and sense expression of cDNA coding for CYP73A15, a class II cinnamate 4-hydroxylase, leads to a delayed and reduced production of lignin in tobacco. *Phytochemistry* 57: 1159–1166.

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

Boudet AM, Kajita S, Grima-Pettenati J, Goffner D. 2003. Lignins and lignocellulosics: a better control of synthesis for new and improved uses. *Trends in Plant Science* 8: 576–581.

Bozell JJ, Holladay JE, Johnson D, White JF. 2007. *Top value-added chemicals from biomass – volume II – results of screening for potential candidates from biorefinery lignin.* Springfield, VA, USA: Available to the public from the National Technical Information Service, US Department of Commerce.

Camargo ELO, Nascimento LC, Soler M, Salazar MM, Lepikson-Neto J, Marques WL, Alves A, Teixeira PJPL, Mieczkowski P, Carazzolle MF et al. 2014. Contrasting nitrogen fertilization treatments impact xylem gene expression and secondary cell wall lignification in *Eucalyptus*. *BMC Plant Biology* 14: 256.

Cannon S, Mitra A, Baumgarten A, Young N, May G. 2004. The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana. BMC Plant Biology* 4: 10.

Cassan-Wang H, Soler M, Yu H, Camargo E, Carocha V, Ladouce N, Savelli B, Paiva J, Leplé J, Grima-Pettenati J. 2012. Reference genes for high-throughput quantitative reverse transcription-PCR qnalysis of gene expression in organs and tissues of eucalyptus grown in various environmental conditions. *Plant and Cell Physiology* 53: 2101–2116.

Chen H, Li Q, Shuford C, Liu J, Muddiman D, Sederoff R, Chiang V. 2011. Membrane protein complexes catalyze both 4- and 3-hydroxylation of cinnamic acid derivatives in monolignol biosynthesis. *Proceedings of the National Academy of Sciences, USA* 108: 21253–21258.

Chen CY, Meyermans H, Burggraeve B, De Rycke RM, Inoue K, De Vleesschauwer V, Steenackers M, Van Montagu MC, Engler GJ, Boerjan WA. 2000. Cell-specific and conditional expression of caffeoyl-coenzyme A-3-O-methyltransferase in poplar. *Plant Physiology* 123: 853–867.

Costa MA, Collins RE, Anterola AM, Cochrane FC, Davin LB, Lewis NG. 2003. An in silico assessment of gene function and organization of the phenylpropanoid pathway metabolic networks in *Arabidopsis thaliana* and limitations thereof. *Phytochemistry* 64: 1097–1112.

Davin L, Jourdes M, Patten A, Kim K, Vassao D, Lewis N. 2008. Dissection of lignin macromolecular configuration and assembly: comparison to related biochemical processes in allyl/propenyl phenol and lignan biosynthesis. *Natural Product Reports* 25: 1015–1090.

De Smet R, Adams K, Vandepoele K, van Montagu M, Maere S, Van de Peer Y. 2013. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proceedings of the National Academy of Sciences, USA* 110: 2898–2903.

Dixon RA, Paiva NL. 1995. Stress-induced phenylpropanoid metabolism. *Plant Cell* 7: 1085–1097.

Eckardt NA. 2002. Probing the mysteries of lignin biosynthesis: the crystal structure of caffeic acid/5-hydroxyferulic acid 3/5-O-methyltransferase provides new insights. *The Plant Cell* 14: 1185–1189.

Ehlting J, Buttner D, Wang Q, Douglas CJ, Somssich IE, Kombrink E. 1999. Three 4-coumarate: coenzyme A ligases in *Arabidopsis thaliana* represent two evolutionarily divergent classes in angiosperms. *Plant Journal* 19: 9–20.

Ehlting J, Shin JJK, Douglas CJ. 2001. Identification of 4-coumarate: coenzyme A ligase (4CL) substrate recognition domains. *Plant Journal* 27: 455–465.

Escamilla-Trevino LL, Shen H, Uppalapati SR, Ray T, Tang Y, Hernandez T, Yin Y, Xu Y, Dixon RA. 2010. Switchgrass (*Panicum virgatum*) possesses a divergent family of cinnamoyl CoA reductases with distinct biochemical properties. *New Phytologist* 185: 143–155.

Felsenstein J. 1985. Confidence-limits on phylogenies – an approach using the bootstrap. *Evolution* 39: 783–791.

Franke R, Humphreys JM, Hemm MR, Denault JW, Ruegger MO, Cusumano JC, Chapple C. 2002. The *Arabidopsis Ref8* gene encodes the 3-hydroxylase of phenylpropanoid metabolism. *Plant Journal* 30: 33–45.

Franke R, McMichael CM, Meyer K, Shirley AM, Cusumano JC, Chapple C. 2000. Modified lignin in tobacco and poplar plants over-expressing the *Arabidopsis* gene encoding ferulate 5-hydroxylase. *Plant Journal* 22: 223–234.

García JR, Anderson N, Le-Feuvre R, Iturra C, Elissetche J, Chapple C, Valenzuela S. 2014. Rescue of syringyl lignin and sinapate ester biosynthesis in *Arabidopsis thaliana* by a coniferaldehyde 5-hydroxylase from *Eucalyptus globulus. Plant Cell Reports* 33: 1263–1274.

Gion J-M, Carouche A, Deweer S, Bedon F, Pichavant F, Charpentier J-P, Bailleres H, Rozenberg P, Carocha V, Ognouabi N et al. 2011. Comprehensive genetic dissection of wood properties in a widely-grown tropical tree: *Eucalyptus. BMC Genomics* 12: 301.

Gion JM, Rech P, Grima-Pettenati J, Verhaegen D, Plomion C. 2000. Mapping candidate genes in *Eucalyptus* with emphasis on lignification genes. *Molecular Breeding* 6: 441–449.

Grima-Pettenati J, Feuillet C, Goffner D, Borderies G, Boudet AM. 1993. Molecular-cloning and expression of a *Eucalyptus gunnii* cDNA clone encoding cinnamyl alcohol-dehydrogenase. *Plant Molecular Biology* 21: 1085–1095.

Grima-Pettenati J, Goffner D. 1999. Lignin genetic engineering revisited. *Plant Science* 145: 51–65.

Gu X. 2003. Evolution of duplicate genes versus genetic robustness against null mutations. *Trends in Genetics* 19: 354–356.

Guo D, Ran J, Wang X. 2010. Evolution of the cinnamyl/sinapyl alcohol dehydrogenase (*CAD/SAD*) gene family: the emergence of real lignin is associated with the origin of bona fide CAD. *Journal of Molecular Evolution* 71: 202–218.

Hamberger B, Ellis M, Friedmann M, Souza C, Barbazuk B, Douglas C. 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. *Canadian Journal of Botany-Revue Canadienne De Botanique* 85: 1182–1201.

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu S-H. 2008. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiology* 148: 993–1003.

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

Hoffmann L, Maury S, Martz F, Geoffroy P, Legrand M. 2003. Purification, cloning, and properties of an acyltransferase controlling sand quinate ester intermediates in phenylpropanoid metabolism. *The Journal of Biological Chemistr*y 278: 95–103.

Hoffmann L, Besseau S, Geoffroy P, Ritzenthaler C, Meyer D, Lapierre C, Pollet B, Legrand M. 2004. Silencing of hydroxycinnamoyl-coenzyme ashikimate/quinate hydroxycinnamoyl transferase affects phenylpropanoid biosynthesis. *The Plant Cell* 16: 1446–1465.

Huang J, Gu M, Lai Z, Fan B, Shi K, Zhou Y-H, Yu J-Q, Chen Z. 2010. Functional analysis of the *Arabidopsis* PAL gene family in plant growth, development, and response to environmental stress. *Plant Physiology* 153: 1526–1538.

Humphreys J, Chapple C. 2002. Rewriting the lignin roadmap. *Current Opinion in Plant Biology* 5: 224–229.

Humphreys JM, Hemm MR, Chapple C. 1999. New routes for lignin biosynthesis defined by biochemical characterization of recombinant ferulate 5-hydroxylase, a multifunctional cytochrome P450-dependent monooxygenase. *Proceedings of the National Academy of Sciences, USA* 96: 10045–10050.

Hurles M. 2004. Gene duplication: the genomic trade in spare parts. *PLoS Biology* 2: e206.

Hussey SG, Saïdi MN, Hefer CA, Myburg AA, Grima-Pettenati J. 2014. Structural, evolutionary and functional analysis of the NAC domain protein family in *Eucalyptus. New Phytologist* 206: 1337–1350.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences* 8: 275–282.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30: 3059–3066.

Kim S, Kim M, Bedgar D, Moinuddin S, Cardenas C, Davin L, Kang C, Lewis N. 2004. Functional reclassification of the putative cinnamyl alcohol dehydrogenase multigene family in *Arabidopsis. Proceedings of the National Academy of Sciences, USA* 101: 1455–1460.

Kim S-J, Kim K-W, Cho M-H, Franceschi VR, Davin LB, Lewis NG. 2007. Expression of cinnamyl alcohol dehydrogenases and their putative homologues during *Arabidopsis thaliana* growth and development: lessons for database annotations? *Phytochemistry* 68: 1957–1974.

Lacombe E, Hawkins S, VanDoorsselaere J, Piquemal J, Goffner D, Poeydomenge O, Boudet AM, GrimaPettenati J. 1997. Cinnamoyl CoA reductase, the first committed enzyme of the lignin branch biosynthetic pathway: cloning, expression and phylogenetic relationships. *Plant Journal* 11: 429–441.

Lauvergeat V, Lacomme C, Lacombe E, Lasserre E, Roby D, Grima-Pettenati J. 2001. Two cinnamoyl-CoA reductase (CCR) genes from *Arabidopsis thaliana* are differentially expressed during development and in response to infection with pathogenic bacteria. *Phytochemistry* 57: 1187–1195.

Leplé JC, Dauwe R, Morreel K, Storme V, Lapierre C, Pollet B, Naumann A, Kang KY, Kim H, Ruel K et al. 2007. Downregulation of cinnamoyl-coenzyme A reductase in poplar: multiple-level phenotyping reveals effects on cell wall polymer metabolism and structure. *Plant Cell* 19: 3669–3691.

Lu S, Zhou Y, Li L, Chiang VL. 2006. Distinct roles of cinnamate 4-hydroxylase genes in *Populus. Plant and Cell Physiology* 47: 905–914.

Marita JM, Ralph J, Hatfield RD, Chapple C. 1999. NMR characterization of lignins in *Arabidopsis* altered in the activity of ferulate 5-hydroxylase. *Proceedings of the National Academy of Sciences, USA* 96: 12328–12332.

Meyer K, Shirley AM, Cusumano JC, Bell-Lelong DA, Chapple C. 1998. Lignin monomer composition is determined by the expression of a cytochrome P450-dependent monooxygenase in *Arabidopsis. Proceedings of the National Academy of Sciences, USA* 95: 6619–6623.

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.

Myburg A, Potts B, Marques C, Kirst M, Gion J, Grattapaglia D, Grima-Pettenati J. 2007. *Eucalyptus*. In: Kole C, ed. *Genome mapping and molecular breeding in plants*. New York, NY, USA: Springer, 115–160.

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.

Nedelkina S, Jupe S, Blee K, Schalk M, Werck-Reichhart D, Bolwell G. 1999. Novel characteristics and regulation of a divergent cinnamate 4-hydroxylase (CYP73A15) from French bean: engineering expression in yeast. *Plant Molecular Biology* 39: 1079–1090.

Niggeweg R, Michael A, Martin C. 2004. Engineering plants with increased levels of the antioxidant chlorogenic acid. *Nature Biotechnology* 22: 746–754.

Osakabe K, Tsao C, Li L, Popko J, Umezawa T, Carraway D, Smeltzer R, Joshi C, Chiang V. 1999. Coniferyl aldehyde 5-hydroxylation and methylation direct syringyl lignin biosynthesis in angiosperms. *Proceedings of the National Academy of Sciences, USA* 96: 8955–8960.

Vautrin S, Santos M, San-Clemente H, Brommonschenkel S, Fonseca P, Grattapaglia D, Song X, Ammiraju J et al. 2011. Advancing *Eucalyptus* genomics: identification and sequencing of lignin biosynthesis genes from deep-coverage BAC libraries. *BMC Genomics* 12: 137.

Pfaffl MW. 2001. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Research* 29: e45.

Pichon M, Courbou I, Beckert M, Boudet AM, Grima-Pettenati J. 1998. Cloning and characterization of two maize cDNAs encoding cinnamoyl-CoA reductase (CCR) and differential expression of the corresponding genes. *Plant Molecular Biology* 38: 671–676.

Piquemal J, Lapierre C, Myton K, O'Connell A, Schuch W, Grima-Pettenati J, Boudet AM. 1998. Down-regulation of cinnamoyl-CoA reductase induces significant changes of lignin profiles in transgenic tobacco plants. *Plant Journal* 13: 71–83.

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

Poeydomenge O, Boudet AM, Grima-Pettenati J. 1994. A cDNA encoding S-adenosyl-l-methionine:caffeic acid 3-O-methyltransferase from *Eucalyptus. Plant Physiology* 105: 749–750.

Raes J, Rohde A, Christensen J, Van de Peer Y, Boerjan W. 2003. Genome-wide characterization of the lignification toolbox in *Arabidopsis. Plant Physiology* 133: 1051–1071.

Ralph J, Lundquist K, Brunow G, Lu F, Kim H, Schatz P, Marita J, Hatfield R, Ralph S, Christensen J et al. 2004. Lignins: natural polymers from oxidative coupling of 4-hydroxyphenyl-propanoids. *Phytochemistry Reviews* 3: 29–60.

Rohde A, Morreel K, Ralph J, Goeminne G, Hostyn V, De Rycke R, Kushnir S, Van Doorsselaere J, Joseleau JP, Vuylsteke M et al. 2004. Molecular phenotyping of the pal1 and pal2 mutants of *Arabidopsis thaliana* reveals far-reaching consequences on phenylpropanoid, amino acid, and carbohydrate metabolism. *Plant Cell* 16: 2749–2771.

Sibout R, Eudes A, Mouille G, Pollet B, Lapierre C, Jouanin L, Seguin A. 2005. Cinnamyl alcohol dehydrogenase-C and -D are the primary genes involved in lignin biosynthesis in the floral stem of Arabidopsis. *Plant Cell* 17: 2059–2076.

Sibout R, Eudes A, Pollet B, Goujon T, Mila I, Granier F, Seguin A, Lapierre C, Jouanin L. 2003. Expression pattern of two paralogs encoding cinnamyl alcohol dehydrogenases in *Arabidopsis.* Isolation and characterization of the corresponding mutants. *Plant Physiology* 132: 848–860.

Soler M, Camargo ELO, Carocha V, Cassan-Wang H, Savelli B, Hefer CA, Paiva JAP, Alexander AM, Grima-Pettenati J. 2014. The *Eucalyptus grandis* R2R3-MYB transcription factor family: evidence for woody growth-related evolution and function. *New Phytologist* 206: 1364–1377.

Stewart JJ, Akiyama T, Chapple C, Ralph J, Mansfield SD. 2009. The effects on lignin structure of overexpression of ferulate 5-hydroxylase in hybrid poplar. *Plant Physiology* 150: 621–635.

Tamasloukht B, Lam MS-JWQ, Martinez Y, Tozo K, Barbier O, Jourda C, Jauneau A, Borderies G, Balzergue S, Renou J-P et al. 2011. Characterization of a cinnamoyl-CoA reductase 1 (CCR1) mutant in maize: effects on lignification, fibre development, and global gene expression. *Journal of Experimental Botany* 62: 3837–3848.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* 28: 2731–2739.

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

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Baren MJV, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* 28: 511–515.

Tsai C, El Kayal K, Harding S. 2006. Populus, the new model system for investigating phenylpropanoid complexity. *International Journal of Applied Science and Engineering* 4: 221–233.

Ulitsky I, Maron-Katz A, Shavit S, Sagir D, Linhart C, Elkon R, Tanay A, Sharan R, Shiloh Y, Shamir R. 2010. Expander: from expression microarrays to networks and functions. *Nature Protocols* 5: 303–322.

Umezawa T. 2010. The cinnamate/monolignol pathway. *Phytochemistry Reviews* 9: 1–17.

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 in *Arabidopsis. Science* 341: 1103–1106.

Vanholme R, Ralph J, Akiyama T, Lu F, Pazo JR, Kim H, Christensen JH, Van Reusel B, Storme V, De Rycke R et al. 2010. Engineering traditional monolignols out of lignin by concomitant up-regulation of F5H1 and down-regulation of COMT in *Arabidopsis. Plant Journal* 64: 885–897.

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

Wang Y, Wang X, Paterson AH. 2012. Genome and gene duplications and gene expression divergence: a view from plants. *Year in Evolutionary Biology* 1256: 1–14.

Weng J-K, Chapple C. 2010. The origin and evolution of lignin biosynthesis. *New Phytologist* 187: 273–285.

Weng JK, Li Y, Mo H, Chapple C. 2012. Assembly of an evolutionarily new pathway for α-pyrone biosynthesis in *Arabidopsis. Science* 337: 960–964.

Ye ZH, Varner JE. 1995. Differential expression of 2 O-methyltransferases in lignin biosynthesis in *Zinnia elegans. Plant Physiology* 108: 459–467.

Zubieta C, Kota P, Ferrer JL, Dixon RA, Noel JP. 2002. Structural basis for the modulation of lignin monomer methylation by caffeic acid/5-hydroxyferulic acid 3/5-O-methyltransferase. *Plant Cell* 14: 1265–1277.

Supporting Information

Fig. S1. Phylogenetic tree of the 4-coumarate:CoA ligase (4CL) multigene superfamily.

Fig. S2. Phylogenetic tree of the shikimate O-hydroxycinnamoyl transferase (HCT)/hydroxy-cinnamoyl CoA:quinate hydroxycin-namoyl transferase (HCQ) multigene family.

Fig. S3. Phylogenetic tree of the caffeoyl CoA 3-O-methyltrans-ferase (CCoAOMT) multigene superfamily.

Fig. S4. Phylogenetic tree of the caffeate/5-hydroxyferulate O-methyltransferase (COMT) multigene superfamily.

Fig. S5. Phylogenetic tree of the cinnamoyl CoA reductase (CCR) multigene superfamily.

Fig. S6. Phylogenetic tree of the cinnamyl alcohol dehydrogenase (CAD) multigene superfamily.

Fig. S7. Physical positions of the 38 Eucalyptus grandis genes involved in the phenylpropanoid pathway.

Table S1. Bibliographic survey of the bona fide genes involved in the phenylpropanoid and lignin branch pathways.

Table S2. List and characteristics of the Eucalyptus grandis genes from multigene (super) families used to assemble the large phylo-genetic trees.

Table S3. Accession numbers of genes/proteins used to assemble the phylogenetic trees.

Table S4. Identity and similarity percentages between members of each multigene family involved in the phenylpropanoid path-way.

Table S5. List of primers.

Table S6. Normalized expression data obtained using real-time quantitative PCR (RT-qPCR) and RNA-seq.

Table S7. Ratio of transcript abundance between pairs of con-trasting tissues